Reconstruction based error detection for robust approximation of partial differential equations

University of Reading: Department of Mathematics and Statistics

Thesis submitted for the degree of Doctor of Philosophy

Georgios Sialounas

November, 2022

Abstract

In this work we present a framework for the construction of robust a posteriori estimates for classes of finite difference schemes. We are motivated by the relative lack of such frameworks compared to existing ones for other numerical discretisation methods, such as finite elements and finite volumes.

The framework we propose is based on the use of reconstructions, which are obtained by post-processing the finite difference solution. The post-processed object is a key ingredient in obtaining a posteriori bounds using the relevant stability framework of the problem. The resulting bounds are fully computable and allow us to establish a posteriori error control over the problem at hand.

In the first part of the thesis we motivate and investigate the behaviour of our framework using model ODE, elliptic and hyperbolic problems. We use our framework to obtain reconstructions which are used to compute a posteriori error estimates. We validate the numerical behaviour of these estimates using solutions of varying regularity.

In the second part of the thesis we focus on hyperbolic conservation laws in one spatial dimension and we deal with scalar problems as well as systems. Hyperbolic conservation laws are widely used in the modelling of physical phenomena. The numerical modelling of conservation laws, which arises due to the frequent lack of explicit solutions, is challenging, largely due to the complex behaviour these problems exhibit, such as shock formation even with smooth initial conditions.

In this setting, we present a framework which is applicable to general non-linear conservation laws. We investigate its numerical behaviour and showcase our results by using popular finite difference discretisations for a range of problems.

We demonstrate that the the framework can produce optimal estimates, capable of tracking features of interest and act as refinement/coarsening indicators.

Dedication

To my family, Loucas, Katerina, Myrianthi, Christina and Michalis who supported me during all these years and to my girlfriend, Petrina, who patiently stood by me.

Declaration

I hereby confirm that this is my own work and that all material from other sources has been properly and fully acknowledged. This work has not and will not be submitted in whole or in part to another University for consideration for any other qualification.

Acknowledgements

First and foremost, I would like to express my eternal gratitude to my supervisor, Tristan Pryer, for his insight, patience and support throughout the past years. His guidance enabled me to explore and cultivate my mathematical interests throughout this research journey. This project would not have been possible without him.

In words of gratitude far fewer than they deserve, I would like to thank my parents, Katerina and Loukas, my aunt, Myrianthi and my girlfriend, Petrina for all their love: your support shall be appreciated and fondly remembered.

Thank you also to the EPSRC, the MPE CDT, the University of Reading and Imperial College London both for their financial support as well as their administrative help throughout this amazing journey I had over the last years.

Lastly, I would like to thank my cherished friends Renos Karamanis and Nicolas Miscourides for the inspired lunch time discussions (before Covid) and for their words of encouragement and support throughout my PhD.

Contents

Chapter 1

Introduction

Partial Differential Equations (PDEs) are one of the most versatile and indispensable tools in the scientific arsenal. It is impossible to overstate their importance on account of their utility in virtually all aspects of science and engineering which underpin modern human society.

Despite their widespread applicability and the extensive study that they were the subject of, several PDEs elude analytical solutions to this day. The importance of these problems has been a motivating factor for finding alternative methods for approximating their solution, thereby motivating and establishing the numerical approximation of PDEs as a rich field of mathematical research.

Several numerical solution approaches evolved over the year. Finite Differences (FD), Finite Volume (FV) and Finite Element (FE) methods are but a few such methodologies. These methods make it possible to simulate PDEs on computers and use them to model situations of high industrial and scientific interest, such as stresses in structures and fluid flows around aircraft wings and in reactors to name but a few.

In all matters of numerical approximations of PDEs it is important and indeed useful to have an indication of the error incurred in the approximation procedure. The importance of knowledge of the error is self-explanatory: we cannot know whether the numerical method is performing correctly otherwise. The reason that knowledge of the error is useful, especially if it is local, robust and available in real time, is that it can be used to increase the approximating power of the numerical approximation where it is most necessary. This last consideration motivates the

study of a posteriori error estimation.

A posteriori error estimates are error statements which can be computed without knowledge of the true solution and which give an indication of the error behaviour. If they are local in nature, they can be used to refine or coarsen the resolution of the approximation. They are particularly useful in flow-related problems and generally in problems where solutions demonstrate a complex structure locally. In such cases a posteriori error estimates can automate the procedure of detecting high errors and modifying the approximation structure appropriately to accommodate local behaviour.

Our interest in this thesis is a posteriori error estimation in the context of Finite Difference (FD) methods. Amongst the three aforementioned numerical approximation methods, FD methods were the earliest to appear. They involve the approximation of derivatives using difference quotients. They are not as flexible as FV or FE methods if domains are geometrically complex. However, they are still widely used in many problems of practical importance.

Relevantly to our interests, FD methods, on account of a lack of a variational formulation of the problem and the point-wise nature of the FD solution, do not receive as much interest with regard to a posteriori error estimation, often relying on heuristics (e.g. gradient indicators or a priori physical knowledge of the problem) in order to define/detect regions of interest in the solution's domain.

The main contribution from this work is a framework for facilitating a posteriori error estimation for classes of frequently used FD schemes. One of the main tools in this endeavour is the reconstruction of the numerical solution (see [Mak07, GMP15, GP17]). This is a post-processing of the point-wise FD solution that performs two key functions in our case. Firstly, it facilitates an alternative error interpretation which is more amenable to a posteriori error control and secondly, it enables us to utilise the stability framework of the PDE in deriving an a posteriori error bound. In this way we are able to derive optimal error bounds for classes of frequently used FD schemes for a variety of problems as well as to utilise existing a posteriori bounds from the literature.

This thesis is broadly divided in two parts. In the first part we motivate, with illustrative examples, the framework for obtaining a posteriori error estimates using

the reconstruction approach. The model problems we present as examples include an ODE, an elliptic PDE and a linear hyperbolic PDE. In each case we use reconstructions to post-process the point-wise FD solution to facilitate a posteriori error control using the stability framework of the underlying problem. We then benchmark the behaviour of the resulting a posteriori error estimate and where possible, we compare it with existing estimates from the literature.

The first part leads us conveniently into a discussion on a framework for obtaining and using reconstruction-based a posteriori error bounds. This is largely our interest in the second part, with the focus shifted entirely to application on hyperbolic conservation laws. In particular, in the second part, we start with a discussion on a framework for obtaining optimal a posteriori error estimates for hyperbolic conservation laws and we test our framework with scalar problems as well as systems, in a linear as well as in a non-linear setting, in one spatial dimension.

Hyperbolic conservation laws are widely used in several scientific and engineering disciplines, such as electromagnetics, civil, aeronautical and mechanical engineering to name but a few areas. The frequent lack of analytical solutions means that numerical approximation is the only avenue for treating these problems. In this context, there are significant challenges that must be addressed, especially in the non-linear cases. The key reason for this is the rich and complex solution features that they exhibit, such as propagating discontinuities, and their tendency to form shocks in finite time and even with smooth initial conditions. This behaviour must be accounted for in deriving appropriate numerical schemes to avoid artificial and spurious numerical behaviour.

In the spirit of deriving schemes, an important concern is the accurate description of this highly localised behaviour whilst ensuring the economical use of computational resources. In this regard, the numerical analyst is also incentivized to allocate computational resources optimally in the domain of interest. This is how a posteriori error estimation arises in importance as a driver for facilitating adaptivity in hyperbolic problems

1.1 Literature review

In this section we will provide the context in which we place our contributions in a chapter-based literature review format.

1.1.1 Chapter 2: A posteriori analysis for conservative linear multistep methods

Differential Equations (DEs) are an important mathematical tool in numerous aspects of science and engineering ([FLQ03], [Chr09], [vRC15], [HPC$^+$00]) as well as in the social sciences ([Bro07]), finance ([EKPQ97]), epidemiology ([Het00], [LM95]) and economics. They form the basis for large areas of research for their own sake but also with respect to their applications ([Eva10], [Olv00], [Arn74]) . However, it often happens in problems of substantial interest, such as the Navier Stokes equations, that explicit/analytical solutions are unavailable.

The lack of explicit/analytical solutions motivates, in large part, the research interest behind numerical methods for approximating DEs, leading to a number of different techniques for the numerical solution of DEs. Since the historic publication of the paper by Courant, Friedrichs and Lewy on partial difference equations for mathematical physics in 1928 (see [CFL67] for a 1967 re-publication), the field of numerical analysis of DEs has flourished and enriched with numerous techniques (see [Tho90] for a historical account).

Numerical methods for the approximation of DEs include Finite Difference (FD) methods ([RM94], [MM05]), Finite Volume (FV) methods ([L$^+$02], [VM07]) and Finite Element (FE) methods ([Arg54], [ZTZ05],[Joh12]) to name but a few. Naturally, an important concern in the process of approximating the DE is the error incurred in the process. It is imperative that a user of a numerical method has some guarantee (e.g. a set of conditions for instance) that the error incurred in approximating some data or a differential operator will remain bounded. If not, the results produced by the numerical method are not reliable. It is not surprising then that error control constitutes an active area of mathematical research in its own right.

Early considerations on error control were concerned with roundoff error, which is a concern whenever we approximate using a computer. An important paper in

this area was a 1947 paper of von Neumann and Goldstine (see [VNG47]), which is concerned with error control for the inversion of matrices. Although this is not related to this chapter, it serves as a detailed account of the importance of error control in numerical methods, particularly those implemented on early machines. An example of such a machine is the differential analyser.

The differential analyser is a mechanical device/analogue computer which was used to solve differential equations by integration. Naturally, research pertaining to solution of DEs using such machines was also concerned with the error incurred in the process of approximation using difference operators. We note the 1910 paper by L.F. Richardson ([Ric11]) and the 1937 paper of Hartree and Womersley on the mechanical solution of certain PDEs ([HW37]). These works informed later attempts for analytical treatments of errors which are still in use today. In particular, we refer to the 1947 paper by Crank and Nicolson on the numerical treatment of errors incurred in the FD approximation of heat conduction type (PDEs) ([CN47]), and the 1950 paper of Charney, Fjortoff and von Neumann on the numerical integration of baroclinic vorticity equations [CFVN90]. Both of these papers voice concern over the fact that error in one time-step may propagate, increase and pollute computations at later time-steps. Both papers included analytical treatments of the error incurred in a FD approximations of the underlying problem. The work in the latter served as the foundation for what eventually became known as as von Neumann stability Analysis.

Statements which provide some sort of guarantee on the behavior of a numerical algorithm without requiring the evaluation of quantities produced by the algorithm are called a priori. These statements are useful in having peace of mind with regard to properties such as stability, convergence, constistency and error bounds. They inform the user whether the algorithm is even worth running in the first place. However, they are not of practical importance once the algorithm runs. In particular, they often cannot be evaluated as they involve the exact solution and/or its derivative and limits as the step-size goes to zero.

A user of the numerical algorithm may often want to have an indication of the behaviour of the error, potentially at a local level. This is particularly the case if the solution has local features, such as shocks, steep gradients, or oscillatory

behaviour. If such features are of smaller scale than the initially specified time-step, this behaviour may go unnoticed. Alternatively, choosing very small time-steps for problems which do not feature steep gradients or erratic behaviour for the most part may lead to large computational expense. This is the motivation behind a posteriori error control.

A posteriori error bounds are error statements which make use of quantities which are explictly computable at run-time. They give the user an indication of the behaviour of the error possibly at a local level, during run-time. This facilitates adaptive control of the step-size.

Adaptive methods and a posteriori error control is of research interest also for ordinary differential equations for their own sake, for their applications and also as a conceptual step in adaptive error control for PDEs, especially where evolution problems are concerned. The intent, as described by [EEHJ95], is to use feedback from computations to inform adaptive methods in an effort to: 1) estimate and control different sources of error - in particular data, modelling and computation errors - 2) improve precision where necessary 3) make efficient use of computational resources. In that work, the authors introduce a framework based on Finite Elements (FE) for the adaptive control of error. They then apply this framework to obtain a posteriori error estimates on a range of ODE and PDE problems.

The literature on a posteriori error estimation for ODEs is extensive and stems, as noted above, from a wide variety of motivations. We note, as an example, [Est95], who constructs a rigorous theory of global error control by combining a priori and a posteriori bound. The author states that the main source of motivation behind this paper is a wider effort to construct a theory of adaptive control for approximations of PDEs. This paper is one of many that are part of this project, which includes [Joh88], [EJ87], [EJ91] and subsequently [EL93], [EJ95b] and[EJ95a]. This work is largely concerned with adaptive time-step control and error estimates for ODEs (stiff problems) as well as adaptive error control for parabolic problems. [EL93] was an example of using a posteriori error estimation for ODEs as a stepping stone for a posteriori error control for PDEs.

In the context of ODEs we note a sequence of works on multi-adaptive Galerkin methods for ODEs (continuous and discontinuous Galerkin) for ODEs in [Log03],

[Log04a], [Log04b],[JL04]. As an example of application for such methods, the author notes the problem of modelling multiple bodies in orbit (comets, planets, satellites etc.) with different orbit durations, hence requiring various time steps of different size.

The work described so far caters to a posteriori error estimation for FE discretisations, which do possess a variational formulation, unlike FD discretisations. In addition to this challenge, a posteriori error estimation for FD techniques also has to address the pointwise nature of the discrete solution. In Chapter 2 we obtain an a posteriori error bound for an ODE problem using energy arguments, with the bound being independent from a specific choice of discretisation. In order to address the issues we identified we use the reconstruction technique.

In summary, the reconstruction is a means to obtain a 'post-processed' form of the numerical solution, endowing it in the process with desirable characteristics. In our case it also enables the calculation of a posteriori bounds. The characteristics one desires from a reconstruction vary but in general they involve the restoration of optimality to suboptimal error estimates. It is used with precisely this intent i.e to restore optimality to the suboptimal posteriori error estimates for FE semi-discretisations of parabolic problems in [MN03]) and for fully discrete linear parabolic problems in [LM06]. Since then, the reconstruction technique has been used in multiple works to obtain optimal a posteriori error estimates (see e.g. [GMP15], [GP17], [LP12], [AMN06], [MN06]; see also [Mak07] for a review).

It is worth briefly commenting upon the use of the reconstruction technique in the aforementioned works and the work of Zadunaisky in [Zad76] (see also associated works: [Zad66b], [Zad66a], [Zad70] and [Zad72]).

Briefly, [Zad76] is concerned with the errors propagated due to integration of systems of ODEs in the context of astronomy problems. A polynomial is constructed over consecutive time intervals using the approximate solution, which is subsequently used to formulate a "pseudo-problem" - essentially a perturbed version of the original problem which possesses a known solution. The integration routine for the original problem is applied to the pseudo-problem instead, obtaining pseudo-numerical solution. The author, assuming that the numerical solutions to the original and pseudo-problems are "close" expects that the errors incurred in the two problems

will likewise be close. With this justification, they adopt the pseudo-error as a good approximation to the error in the original problem.

In our case, we aim at obtaining optimal order a posteriori bounds and at extending these results to PDEs. We note that we demonstrate optimality numerically rather than prove it in this chapter. In addition, we emphasize that we desire local control over the error. Therefore, the reconstruction must be constructible using local considerations our construction procedure reflects these considerations (see also [AMN09, §1] and [MN06, §1] for discussions on the same issue). We also note that we do obtain reconstructions with WENO-based techniques (see [Shu98]), which actually utilize wider stencils.

In particular, we examine a reconstruction for a model second order IVP problem, discretised using the well-known Leapfrog scheme. Briefly, the Leap-frog scheme is an explicit, two-step method that is popular for use with (amongst other things) for second order wave-type PDEs. This method is also known by other names in other fields. In molecular dynamics, where it is popular in its use as an integration scheme, it is known as the Verlet method ([Ver67]). It is also known as the Störmer method, because variants of it have been used by Carl Störmer to calculate the trajectories of ionized particles in the Earth's magnetic field ([Stö07]). In fact, the leapfrog method has been in use since earlier than these works and examples can even be found in Newton's Principia. A more thorough historical account of the method can be found in [HLW03].

The Leapfrog integration method posesses favourable long-term numerical properties. Specifically, it conserves first integrals (e.g. total linear and angular momentum in N-body systems) and features linear error growth (see [HLW03, §5: Fig. 5.3]) (see [HLW03] for a detailed discussion). In addition to these properties, as explained by [GLMV16], the Leapfrog scheme is the only explicit, two-step, second-order accurate method for the time integration of second order problems.

Briefly, in [GLMV16] the authors perform an a posteriori analysis of second order implicit and explicit two-step schemes for wave-type problems. They derive optimal a posteriori estimates for controlling the error of the temporal discretization. In this chapter, we follow their work closely and we compare our results with those obtained using their method, applied to a second order ODE model problem.

1.1.2 Chapter 3: Simple a posteriori control of finite difference discretisations of elliptic problems

In this section we review FD methods and variations thereof for elliptic problems, with a focus of different a posteriori error estimation techniques. We emphasize that, although we do not use the methods we review, we nonetheless feel that it is necessary to include pertinent material, especially with regard to a posteriori error estimation for variations of FD methods, in order to endow the reader with a comprehensive picture of the field and, therefore, of the context of our contribution.

Finite differences date back centuries - they were introduced by Brook Taylor in 1717 (see [Tay17]). FD discretisations of elliptic problems are correspondingly widespread and they constitute part of many undergraduate and graduate curricula on numerical methods for differential equations. However, a posteriori error control for elliptic problems discretised using classical FD methods has not received as much attention compared to the FE counterparts. This is in part is due to the lack of a variational structure of FD methods and the pointwise nature of the FD approximation.

The popularity of the FD approximations warrants an in-depth look into a posteriori error estimation for such methods. To this extent, in this section, we will present a brief overview of work that resulted in different versions of FD methods, with the emphasis of our review placed upon a posteriori error bounds.

Mimetic Finite Difference (MFD) method

The MFD method is used to produce discrete approximations to PDEs on unstructured polygonal and polyhedral meshes. An aim of the method is that the obtained discretisations preserve or, more appropriately, "mimic" important properties of the underlying mathematical and physical systems which they approximate (see [LMS14]).

The properties depend on the underlying problem. On the mathematical side, these properties include symmetry, positivity, duality and self-adjointness of the discrete operators, maximum principles and asymptotic limits amongst others. On the physical side, amongst other properties, for fluids problems, the MFD method aims to conserve discrete analogies of energy, mass on momentum and, for incompressible

flows, to preserve divergence free conditions.

A posteriori error estimates for the MFD method for elliptic problems were derived in [dV08], [BdVM08] and [AdVLV13]. We briefly summarise the results of these works.

In [dV08] a local, residual-based posteriori error indicator for the MFD method is derived and presented for diffusion-type problems on polyhedral meshes. This estimate incorporates a post-processing scheme for the scalar variable (which is the pressure in the model problem treated in this paper) in order to show convergence in a stronger norm. Addtionallly, the post-processed variable features in the a posteriori bound. This is related to what we introduce in Chapter 3, as, conceptually, we also use a post-processed numerical solution in order to improve convergence behaviour, and ultimately state bounds.

In [BdVM08], the framework of [dV08] is used to develop an error estimate for elliptic (steady diffusion) problems in mixed form with homogeneous and non-homogeneous Dirichlet boundary conditions. The error estimate is used to implement adaptive refinement. The obtained error indicator resulted in optimal rates of convergence in a mesh-dependent norm (see [BdVM08, §2 and §4]).

In [AdVLV13], a hierarchical-type a posteriori error estimator is presented for the MFD discretisation of elliptic problems (see [AO11, ZGK83, Ban96, DLY89] for more information on hierarchical a posteriori estimators). The posteriori estimate is optimal with respect to the discrete energy norm for the tested problems (see [AdVLV13, §4,5]).

Lastly, it is worth noting the use of reconstructions as a post-processing tool in the context of the MFD method. Briefly, reconstructions in the MFD method map mesh functions to continuum functions and enable the use of Finite Element machinery with MFD discritezations. It can be used in posteriori error estimates to improve convergence behaviour (see [dVLM14, Ch.5]) and as a post-processing tool to improve accuracy of the numerical approximation to the solution (see [CM08] for an example based on a diffusion problem in mixed form).

Generalised Finite Difference (GFD) method

The GFD method is an evolution of the traditional FD method that can be applied to irregular grids of points. The evolution of the use of FD methods for approximating PDEs on irregular grid started from the 50s. In [Jen72], a basis was proposed for a practical (i.e. practically possible given the computational capabilities of the time) FD method on irregular grids. This work addressed challenges related to approximating PDEs in non-rectangular domains and in particular, the problems of obtaining difference coefficients and implementing the boundary conditions on curved boundaries. The proposed method, based on a six-point scheme (star), obtained FD formulas for approximating up to second order derivatives.

A series of ensuing works were aimed at addressing the two main disadvantages from the approach of [Jen72]; namely, singular derivative coefficient matrices and limited accuracy of the obtained derivatives (see e.g. [PK75], [WTDS75], approaches and references therein). Interested readers are also advise to check [LO80] and [Lis84] for improvements and important contributions to the method.

These advances enabled the implementation of adaptivity in the context of the GFD method. In particular, we note the work of [Ork98] on an adaptive multigrid GFD method (see also references on related work of Orkisz in [BUGA03]. In [BUGA03], an h adaptive method is described for second order PDEs. The adaptive method therein, is facilitated by an a posteriori error estimate (see [BUGA03, eq. (16)]) which is obtained for a given node in a stencil by calculating (a weighted combination of the) additional Taylor expansion terms relative to those used to obtain the scheme. Specifically, an extra, fourth order Taylor expansion is used to obtain the estimate by calculating its difference with the original, second order Taylor expansion used to obtain the scheme. Then, additional nodes are added/subtracted from the stencil accordingly. The reader should note that these estimates are indicators/estimators of the errors and not necessarily error bounds like they are in our case.

The concept of obtaining posteriori estimates using a higher order Taylor expansion compared to the one used for the scheme is also used extensively in the GFD literature. In [UBAG05], it is used to to obtain a posteriori estimate for a 3D problem. In [BUGA08] the estimate of [BUGA03] is compared with earlier work from

[Ork98] on the basis of adaptivity for an elliptic problem and found to be of similar efficiency but better computationally. In [UBU+18] the estimate of [BUGA03] is used to drive h-adaptivity in 2D and 3D second order PDEs. In [GUB+18] adaptive refinement is used to improve the solutions of GFD approximations for elliptic problems.

Virtual Element Method

The Virtual Element Method (VEM) is a numerical framework for approximating PDEs. We include the VEM in this review section because it can be regarded as an evolution of the MFD method (see [BdVBC+13] for the basic principles of VEM and its connection to MFD). Nevertheless, it should be noted that in its latest iterations, the VEM grew to more closely resemble FE methods rather than FD method.

In particular the VEM resembles a generalisation of FE methods on polygons. It can be used with general polygonal/polyhedral elements ([BdVBMR14], [BdVBMR16], [BFM14] [CMS17]) and it utilises approximation spaces of arbitrary global regularity (see [dVM14]).

The mesh flexibility allowed by the VEM with makes it attractive for the implementation of adaptivity, as [CGPS17, §1]) explain. This is because of the ease with which refinement, coarsening and mesh distortion are treated by the VEM. We will briefly summarise these. Firstly, since general polytopal meshes are admissible on account of the polynomial subspaces included in the VEM space, unlike standard FE methods, maximum angle conditions or mesh-distortion considerations are not problematic. Furthermore, there is no need to introduce additional degrees of freedom for hanging nodes resulting from the refinement of neighbouring elements. This is because co-planar element interfaces are acceptable and therefore hanging nodes are treated as new nodes. Coarsening is also trivial.

A residual-based posteriori error estimate is developed in [DVM15] for a VEM approximation of the Poisson problem with (piecewise) constant coefficients. In [CGPS17], an a posteriori error analysis is presented for the VEM applied to general elliptic problems from [CMS17]. The resulting error estimate is of residual type and relies on the degrees of freedom and the element-wise polynomial projection of the VEM solution. The estimate is shown to be equivalent to the error between the true

solution and the VEM approximation in the energy norm.

In [BdVCN+21], the foundation was constructed for building a rigorous theory for Adaptive Virtual Element Methods (AVEMs). The analysis pertains to triangular meshes in 2D with a systematic refinement procedure such that shape regularity and optimal complexity are preserved. This work focuses in addressing the presence of the stabilisation term which prevents the equivalence of the residual error estimator with the error estimate in the energy norm. The authors demonstrate that, under a set of appropriate assumptions, the stabilisation term can be made arbitrarily small relative to the error estimate (see [BdVCN+21, §1,4]). The estimate is found to behave optimally (see [BdVCN+21, §8]).

1.1.3 Chapter 4: Automated Error control for the transport equation

A posteriori error estimates for linear hyperbolic problems

This chapter is intended as a motivation for obtaining a general framework of a posteriori error estimates for FD schemes for hyperbolic conservation laws. In this context, it is useful to include some material on a posteriori error estimation, even for other numerical methods (e.g. FV or FE). We do this for two reasons.

Firstly, a posteriori error estimation for FD schemes is ultimately the end purpose for which we are intending the material in this chapter. Secondly, the inclusion of material from FE and FV a posteriori error estimation enables as to set the tone for the chapter itself.

A posteriori error estimates for hyperbolic conservation is, justifiably, a large area of research. Hyperbolic conservation laws are widely used in practice. They feature solutions which can develop discontinuities in finite time and it is in this context that a posteriori error estimates are particularly useful: they inform the user where the approximability of the underlying numerical scheme must be improved locally in order to better capture the effect of the discontinuity.

A posteriori error estimates for evolution problems (including conservation laws) are generally obtained using duality and energy methods as [MN06, §1] notes. We note that while a posteriori estimates do not necessarily have to be produced with

a particular numerical method in mind (see e.g. [CG95]), it is often the case that a posteriori error estimates have been produced for specific numerical method.

Briefly, the duality methods rely on the stability of a backward dual problem (see [JS95]; see also [SH96] for a review for duality based estimates for the FE method). In [HS01] and [BO96] the authors consider a posteriori error analysis for *hp*-dG FE approximations to first order hyperbolic problems see also [HRS00] for work on a posteriori error analysis of stabilised finite element approximations of transport problems.

Energy methods involve testing the error representation formula with the error (or some integral/derivative of the error) and deriving an a posteriori error bound using the stability framework of the PDE (see e.g. [NSV00] for a discussion in the context of evolution equations in general and [LP12] for energy estimates for parabolic problems).

In [MN06], an a posteriori error estimate is constructed for time discretisations by the dG method for both linear and non-linear evolution problems. In this work, the a posteriori error analysis relies on the reconstructions to derive optimal order a posteriori error estimates. In [GHM14] an a posteriori error bound is presented for a first order linear hyperbolic problem, discretized by the dG method. In this case, the a posteriori error bound is based on a reconstruction in the spirit of [MN06]. It is worth noting that both these works use the concept of reconstruction in the derivation of a posteriori estimates. A discussion of reasons for using reconstructions in the context of evolution problems can be found in [Mak07].

In our case, the reconstruction facilitates an alternative error interpretation which allows us to compare a post-processor of the numerical solution with the exact solution. In turn, this error interpretation allows us to use the PDE's stability framework; something we could not have done otherwise on account of the point-wise nature of the FD method. This enables us to produce optimal order a posteriori estimates for the problems under consideration.

It is interesting to compare our use of reconstructions with the work of [GM00] as this work is also concerned with a posteriori error estimates for FD schemes for scalar conservation laws. In this case the numerical approximation produced by the FD scheme is interpreted to be constant between interval mid-points. In a sense one

may describe this as an implicitly defined piece-wise constant reconstruction

In [BCL13] the authors present a unified approach for constructing local a posteriori error estimates for FE approximations, including conforming, non-conforming and DG. Their approach is based on $\mathrm{H}^{\mathrm{div}}(\Omega)$-reconstructed fluxes, an idea first proposed in [LW04], where it is presented in the context of conforming Raviart-Thomas elements.

Parasitic waves

Parasitic waves are numerical artefacts that arise when a solution to an evolution problem travels over an abrupt change in the underlying numerical model. They manifest as high frequency numerical dispersion with wavelengths comparable to the local grid size. Various types of changes in the underlying numerical discretisation or in the PDE itself may cause parasite generation. Examples include, step changes in mesh resolution, step changes in PDE coefficients or changes in the PDE model itself (e.g. from advection to advection-diffusion).

In this area, it is worth making a note of the works of Robert Vichnevetsky. Firstly, in [Vic81b], the author studies the spurious reflection phenomena that arise at the mesh-size change interface for numerical approximation of hyperbolic conservation laws using finite differences. In this work, the author uses the time-Fourier transform of the numerical solution (see also [Vic81a, Vic87]) to obtain an analytical solution of the reflection that occurs at the fine-coarse mesh interface. The author is also able to derive other useful properties using the Fourier transform, such as phase and group velocities for the actual and the parasitic components of the solution. A dedicated study on the propagation properties of semi-discretisations of hyperbolic equations by the same author can be found in [Vic80], while the Fourier analysis aspect of numerical approximations to hyperbolic equations is treated in [VB82].

The behaviour of parasitic waves is also investigated in [Tre82] in the context of a broader study of the importance of group velocity in FD schemes for time-dependent problems. Specifically, the author studies the generation and propagation of parasitic waves by FD schemes and in particular, their generation and transmission at interfaces. In the numerical experiments presented in [Tre82, §3], the generation of parasitic waves arising from coefficient change is demonstrated. The author

used the concept of group velocity to obtain useful descriptions of their propagation characteristics, such as the speed at which they propagate.

On the topic of parasites, more recent attempts include that of [FR04] and [LT11]. In [FR04], the authors address the issue of spurious modes excited at the interface of non-uniformities in the context of the advection and wave equations. For the advection equation in particular, they compare the dispersion relations between the central difference and box-scheme spatial semi-discretisations (see [AM04] for details on the box scheme). They show that the box-scheme does not suffer from the same spurious modes at mesh interfaces that the central difference scheme does on account of its monotonic dispersion relation with respect to the wave-number.

In [LT11], the authors investigate the propagation behaviour for central difference schemes on non-uniform staggered grids and extend results to shallow water equations. They show that, asymptotically, there is no reflection in the limit as the grid becomes slowly varying as long as the waves of the relevant frequencies are well-resolved. In addition, they propose how to tailor the difference scheme to minimise spurious wave reflections.

1.1.4 Chapter 5: Postprocessing in finite difference schemes

A posteriori error estimation for FD schemes for conservation laws

Hyperbolic conservation laws ubiquitously arise in many physical applications. Inviscid compressible flows are well described by Euler's equations which have meteorological applications, for example. A major difficulty in designing numerical schemes for hyperbolic conservation laws is that they can form shocks in finite time. There has been considerable activity in this area based on various numerical techniques, such as finite difference, volume and element approaches [CCL95, KR94, L$^+$02]. The formation and tracking of these discontinuities is a significant challenge.

A substantial body of work has accumulated over the years in applications of FD schemes for hyperbolic problems, resulting in several noteworthy contributions (see [LeV92], [JT97] for overviews). Early examples include Godunov's scheme ([God59]), the Lax-Friedrichs (LxF) scheme ([Lax54]), the two-step Lax-Wendroff scheme (see the recent work of [LVW21]), as well as the works of van Leer (see ([VL73], [VL74], [VL77a], [VL77b] and [VL79])). The works of [NT90], who use

the LxF solver in conjunction with MUSCL-type interpolants to compensate for the excessive LxF viscosity are also of note. Two classes of FD schemes that are of particular importance in the context of hyperbolic conservation laws are the Essentially Non-Oscillatory schemes (see [HEOC87], [SO88], [SO89]) and the Weighted ENO schemes ([LOC94], [JS96], [JT98]; see [Shu98] and references therein for an overview). ENO and WENO schemes combine high orders of approximation in smooth regions and non-oscillatory behaviour in the vicinity of discontinuities.

A posteriori error estimation aims to provide the user with local computational control over the error incurred in approximating a partial differential equation (PDE) with a given numerical scheme. A posteriori error estimates for hyperbolic problems have received considerable attention, particularly for discontinuous Galerkin (dG) finite element methods [Joh90, JS95, DMO07, GMP15, GP17, DGPR19, AO11, Ver13] and finite volume (FV) methods [CCL94, CG14, BHO18, SCR16, SL18]. (see also [GHM14]; see also [SH96],[HS01] for relevant material)

By comparison, finite difference (FD) schemes have seen less interest with regard to a posteriori estimates. This is predominantly due to the problem lacking a variational structure, something quite crucial for typical a posteriori techniques to be applied. Indeed, goal-oriented a posteriori estimates have been derived for the Lax-Wendroff scheme by proving the method is equivalent to a finite element scheme [CET14]. It is worth noting that there are the approaches that work for general numerical schemes approximating scalar conservation laws [CG95] and a posteriori estimates derived for FD schemes with local error estimation based on Richardson extrapolation [BO84, ABF88]. The estimates are used to facilitate mesh adaptivity.

ENO and WENO schemes

Although the WENO scheme itself is not the focal point of this chapter, it is necessary to make a special note for this scheme as WENO interpolating polynomials are key to our reconstruction procedure for general non-linear hyperbolic conservation laws. Additionally, ENO and WENO schemes are a cornerstone in the area of numerical approximation of hyperbolic conservation laws.

Essentially Non-Oscillatory schemes were first designed in [HEOC87]. The work in [HEOC87] was set in a FV context. Since then, this area of research expanded

significantly. In [SO88] and [SO89] the ENO schemes were extended to FD methods and since then they became widely used. At the heart of ENO schemes there is an automatic stencil selection procedure, whereby intervals are successively added to the computational stencil. The selection of additional intervals is based on the local smoothness of the relevant function: if an interval contains a discontinuity, then it is not selected. It is this procedure that ensures uniformly high orders of accuracy in smooth regions and sharp, essentially-non oscillatory transition over the discontinuity.

Despite their utility and widespread usage, ENO interpolation and reconstruction have a few shortcomings (see [Shu20, §3]) for a detailed discussion). Very briefly, some of these shortcomings are the following. Firstly, while many candidate stencils are considered, ultimately, only one is chosen. This is a sensible approach near discontinuities but in smooth regions, the other candidate stencils can be put to good use to help increase the order of accuracy. Secondly, the divided difference approach may result in a left-biased stencil. This could be the case, for example, for functions whose derivatives are strictly monotonically increasing since their divided differences will reflect this behaviour. This bias can result in stability and accuracy issues in solving time-dependent hyperbolic problems (see [RM90]). Thirdly, from a computational perspective, the procedure for the adaptive stencil choice in ENO schemes contains several `if` statements which are inefficient on certain machines. For a complete list of issues associated with ENO procedures, the reader is directed to [Shu98, §2.2.2].

WENO schemes are an extension of the Essentially Non-Oscillatory (ENO) procedure. WENO schemes seek to maintain the advantages of the ENO procedure, such as uniformly high order accuracy and non-oscillatory transitions over discontinuities, while simultaneously addressing the shortcomings. The key idea in WENO schemes is to use a convex combination of candidate stencils by assigning a weight to each stencil (termed a non-linear weight) and then combining all stencils to obtain the result. This is in contrast with the ENO procedure where only one stencil is chosen.

There are two important considerations when choosing the weights. Firstly, it must be ensured that, when the solution is smooth in all candidate stencils, the

weights are worked out such that they highest possible order of accuracy is obtained for the combined stencil. Secondly, when one candidate stencil out of a set contains a discontinuity while the rest do not, then the weight for this stencil should be very small so its effect on the combined approximation to the numerical flux is small. These two considerations ensure the uniformly high order and the non-oscillatory behaviour for WENO schemes.

WENO procedures were introduced in the context of FV schemes for hyperbolic conservation laws in [LOC94]. In this work a $(k+1)\,th$ order WENO scheme was produced from the same stencils that would produce a kth order ENO scheme. In [JS96] a general framework was presented for constructing $(2k-1)\,th$ order WENO approximations from a kth order ENO stencil. In the same paper, a fifth order FD WENO scheme was constructed for multi-dimensional conservation laws using this framework. This scheme has been been very popular ever since.

ENO and WENO schemes have been extensively implemented in several and diverse fields. These include simulations of turbulent flows ([SZA+19]), studies of shock waves, explosive flows and chemically reactive flows, respectively [OZ19], [WSHN13], [CSKO19], aerodynamics and magnetohydrodynamics, [LRKK19] and [FIDSG19], atmospheric and climate sciences ([LC19]) and fluid structure interaction ([NBOT19]).

In this work we are interested in the application of FD WENO schemes on non-uniform grids. In this regard, we make extensive use of the results of [JSB+19], where the WENO approximation is extended to non-uniform grids. In this chapter we present this procedure in the context of an interpolant of functions of varying regularity and in the following chapters we use this strategy to facilitate mesh adaptivity in hyperbolic conservation laws.

1.1.5 Chapter 6: Automated error control for linear hyperbolic systems

Systems of hyperbolic conservation laws are of high importance in the physical sciences (fluid mechanics, electromagnetics, acoustics, earthquake engineering to name but a few). In cases where the perturbations that are being propagated are small, these are modelled by linear systems (see [LeV07, §10]). The frequent lack of

analytical solutions of hyperbolic problems makes their numerical approximation an important field of study. In addition, because these problems often possess features such as complex solution structures and shocks, an important consideration is to ensure that computational resources be used optimally and focused in parts of the numerical domain where such features are present. In turn, this makes a posteriori error estimates, which are often used to drive adaptivity, an important field of study in the context of hyperbolic conservation laws in and of itself.

A posteriori error estimation for hyperbolic conservation conservation laws is perhaps not as developed as the corresponding areas for elliptic PDEs (see e.g. [SH03] for a discussion on potential explanations).

Linear systems

In this chapter we are interested in symmetric positive linear systems. These systems, which are also known as Friedrichs systems (see [Fri58]) cover a wide class of problems (wider than we consider here). This class of problems includes not only hyperbolic systems but also symmetric elliptic problems. The treatment of these problems together did not arise as an intention to treat them both in the same framework, but rather to treat problems which may change from one type of PDE to the other. In the words of the author of the paper (K.O. Friedrichs) which later gave this class of problems their name, a unified treatment of these problems poses challenges as they use different tools. An example of a problem which changes from one to the other type is the modelling of transonic flow (see [W$^+$95]). Briefly, in regions of the domain where the flow is subsonic, the PDE that governs the flow is elliptic whereas in regions where the flow becomes supersonic, the PDE is hyperbolic.

Numerical schemes

There is a rich literature of FD, FV and FE schemes for hyperbolic conservation laws (see [Swe84, Swe89, CJST06, LeV92, L$^+$02, SH96]). The methods involved for non-linear problems certainly work for linear ones as well. In the case of linear problems, aside from high approximability, an important consideration is the behaviour of the numerical method in the case of discontinuous solutions.

Numerical methods which would work well with smooth solutions may blow up

in the presence of discontinuities. Numerical artefacts such as diffusion (first order methods) and dispersion (numerical oscillations) may be exacerbated in the presence of discontinuities. Furthermore, even if there is convergence this may be of lower order than the formal order of convergence for the method. An example is the Lax-Friedrichs FD scheme, which is formally first order accurate but converges as $\mathcal{O}\left(h^{1/2}\right)$ to the exact solution for the advection equation with a discontinuous initial condition.

A posteriori error estimation

A posteriori error estimates for systems of linear conservation laws can follow as corollaries from a posteriori estimates for non-linear problems, particularly in the context of FE (both cG and dG) and FV discretisations. We are reviewing these in the next chapter so in this chapter we will focus towards results obtained specifically for linear systems and in particular on Friedrichs system.

Interest in the a posteriori error analysis of Friedrichs systems arises naturally from a need to drive adaptivity and to ensure economical use of computational resources in such problems (see [SH96], [AABM00]).

We note in particular work of [HMSW99], in which local and global a posteriori error bounds are obtained for a steady problem. An overview of a posteriori error analysis for FE approximations to hyperbolic problems, including for linear systems, can be found in [SH03], in which a discussion is presented on a posteriori error estimates based on hyperbolic duality arguments. This work includes a summary of the Johnson paradigm of a posteriori error estimation (see [Joh93],[EEHJ95]; see also [JS95]).

1.1.6 Chapter 7: A posteriori error analysis for non-linear hyperbolic problems

A note on non-linear hyperbolic problems

In this chapter we consider the Cauchy problem for non-linear scalar conservation laws and systems of conservation laws in one spatial dimension. The mathematical treatment of these problems has to account for two difficulties (see also [Daf05, §4]):

shock formation and non-uniqueness of weak solutions.

Firstly, the solutions to these problems form shocks in finite time: if one visualizes the solution as a propagating wave, then the wave profile would grow steeper in time and eventually form discontinuities which would propagate. Therefore, such solutions are examined in a weak sense rather than in the classical sense.

Secondly, weak solutions are not necessarily non-unique - in fact a problem may have infinitely many of them (see [Daf05, §4.4]). This means that appropriate admissibility criteria must be established to eliminate undesirable or physically irrelevant weak solutions from consideration. Briefly, the matter of existence and uniqueness of weak solutions has been settled for non-linear scalar problems in many spatial variables and partially for systems in one spatial variable. Two widely used admissibility criteria for weak solutions stem from the notions of entropy and vanishing viscosity.

The vanishing viscosity technique involves the addition of a diffusive term with a small coefficient to the hyperbolic system. The justification for this is the observation that the physical problem which induces the conservation law possesses some degree (however small) of dissipation ([Smo12, §15.D]). Unsurprisingly the exact structure of the added term is informed from the thermoelastic properties of the underlying physical system (see [Daf05, §4.6]). In this regard, shocks in physical systems would in essence manifest as very steep gradients in the considered quantity. Then, assuming that the viscous term was added in such away so as to ensure well-posedness of the viscous problem, one obtains the solution of the original, non-dissipative hyperbolic system as the limit of the viscous term tending to zero of the solutions to the viscous problem. Convergence in this case should be defined in an appropriate sense. A solution obtained in this way is said to satisfy the viscosity admissibility criterion (see e.g. [DiP83] for a convergence theorem using the vanishing viscosity method applied to the isentropic equations of gas dynamics).

Entropy-related admissibility conditions trace their physical origin back to the second law of thermodynamics, which states that entropy of a system must be non-decreasing with time. Such conditions are used to identify the physically relevant weak solution (see [LeV92]) by demanding that entropy increases across a physically admissible shock ([L$^+$02, §11.13]) - a property that can be used to identify such

shocks. There are a number of entropy conditions, such as, for instance, the Lax and Oleinik entropy conditions (see [Lax57, Lax73] and [Ole57]; see also [L$^+$02, §11.13] for a brief review).

An alternative entropy approach is the definition of an entropy function. The entropy function satisfies a "companion conservation law", which becomes an inequality -in weak form- in the presence of discontinuities (see [LeV92, §3.8.1]). This criterion is called the entropy inequality. It should be noted that the existence of an entropy function is not guaranteed in general: specifically, while in the case of scalar problems, conveniently, any convex function is an entropy this is not the case for systems. In general, for systems of conservation laws, the existence of entropy functions is a property of the system and while it is certainly possible to find an entropy function, it is not as easy as the scalar case [GR13], where every convex function is an entropy function. A famous example is the family of entropy functions known as Kruzhkov entropies (see [Kru70]). It should be noted that in [Kru70], the author gave a characterisation of admissible weak solutions and contributed existence, uniqueness and stability results. Incidentally, the doubling of variables technique from that paper provides a natural way of establishing a posteriori error control in the $\mathrm{L}^1(\Omega)-$norm for scalar problems in several variables (see [CG95], [Ohl09]).

A note on the numerical discretisation of non-linear problems

The numerical study of non-linear hyperbolic problems is an expansive and important field of study that has accumulated over decades. A review of several contributions in the numerical discretisation methods for non-linear hyperbolic problems can be found in [CJST06] (see also references therein). This work includes reviews on FD, FV and FE techniques amongst others. The work of [GR13] contains an extensive exposition of FD and FV schemes of systems of conservation laws with schemes for up to two spatial variables.

There is a extensive literature with regard to numerical discretisation methods for hyperbolic problems. While we do not intent by any means to provide an exhaustive literature review we will refer to some review articles which offer a more thorough coverage of the subject matter. For the theory of convergence of approximate solutions (using Finite Differences and viscosity solutions) of hyperbolic systems in

one dimension we refer readers to [GR13] (see also the related work [GR91] by the same authors on theory for scalar problems). An exposition on FD methods for hyperbolic conservation laws can be found in [LeV92]; a survey of frequently used FD schemes for systems of non-linear hyperbolic conservation laws can be found in [Sod78] (see also references therein). With regard to FV methods for conservation laws we refer readers to [LeV07] and references therein. Lastly for DG schemes for hyperbolic problems we refer readers to [HW07, §5], [CKS12] and references therein. A more recent review of several numerical methods can be found in [Hes17].

A substantial body of work has accumulated over the years in applications of FD schemes for hyperbolic problems, resulting in several noteworthy contributions (see [LeV92], [JT97] for overviews). Early examples include Godunov's scheme ([God59]), the Lax-Friedrichs (LxF) scheme ([Lax54]), the two-step Lax-Wendroff scheme (see the recent work of [LVW21]), as well as the works of van Leer (see ([VL73], [VL74], [VL77a], [VL77b] and [VL79])). The works of [NT90], who use the LxF solver in conjunction MUSCL-type interpolants to compensate for the excessive LxF viscosity are also of note.

We must also refer to high resolution methods. An early work in this area is the flux-corrected transport method of [BB73], which can also be viewed as a flux-limiter method. In these methods, the objective is to facilitate high accuracy without introducing spurious oscillations. This is facilitated by a linear combination of low-order and a high order flux, with the coefficient of the correction being referred to as the flux limiter. We also refer readers to the work of [Swe84], where a large class of these methods is studied and conditions are derived to guarantee the TVD property and second order accuracy (see also [LeV92, §16]).

Two classes of FD schemes that are of particular importance in the context of hyperbolic conservation laws are the Essentially Non-Oscillatory schemes (see [HEOC87], [SO88], [SO89]) and the Weighted ENO schemes ([LOC94], [JS96], [JT98]; see [Shu98] and references therein for an overview). ENO and WENO schemes combine high orders of approximation in smooth regions and non-oscillatory behaviour in the vicinity of discontinuities.

A posteriori error estimation for non-linear problems

In the numerical treatment of non-linear hyperbolic problems a major challenge is the reliable computational representation of the localised structures that we alluded to previously that typically arise in these problems, such as propagating shocks, contact discontinuities and rarefactions. These phenomena highlight the need for efficient, adaptive computational meshes, which ensure the economical allocations of computational resolution to capture this behaviour. Adaptivity in this context is driven either by ad hoc heuristics (e.g. by physically motivated considerations, problem geometry etc.) or -preferably- by rigorous a posteriori error estimates.

A posteriori error estimates for non-linear scalar conservation laws in many dimensions were derived in [KO00] in the $\mathrm{L}^1(\Omega)-$norm for FV schemes (see also [CH99] for related error estimates using the Kruzkov framework). We also refer to [GM00] for a one-dimensional scalar problem with results based on Kruzkov-type estimates (see [Kru70], [BP98]). A posteriori error estimates for general numerical discretisations for the non-linear scalar problem are derived in [CG95] (see also [Coc99] and references therein). We also note the Dual Weighted Residual (DWR) approach to a posteriori error estimation for hyperbolic problems (see [Sül99, HS01, HH03]; see also [Laf04], [HH03] for non-linear problems).

We note the utility of the doubling of variables technique introduced in [Kru70] in obtaining error estimates for non-linear scalar conservation laws for Finite Volume and discontinuous Galerkin schemes (see [Ohl09]). This technique is used both in a priori and a posteriori error estimates for scalar problems. In the case of a priori estimates we refer readers to [Vil94, CCL94, CG96, CG97, CGY98], which pertain to a priori error estimates fof FV methods for scalar multi-dimensional conservation laws.

In the case of a posteriori error estimates for non-linear scalar problems in one or multiple dimensions we refer readers to the work of [KO00], [OV06] for derivation of results using FV schemes. We also refer reader to [DMO07] for the derivation of an a posteriori error estimate for mesh adaptivity with the Runge-Kutta Discontinuous Galerkin (RK-DG) method (see also [Ohl09] and references therein).

In the context of a posteriori error control for non-linear hyperbolic systems we note the work of [GMP15], who extended the reconstruction technique of [Mak07],

originally for parabolic problems, and used it in conjunction with the relative entropy framework of Dafermos and Diperna to derive optimal a posteriori error estimates (pre-shock) for a DG discretisation of a one dimensional system of non-linear hyperbolic conservation laws - a result we utilize in this chapter (see [Daf78],[Daf79] and [DiP79]; see also [SV16] on relative entropy for hyperbolic systems). We also note related work by [GP17], where a posteriori error estimates were derived using the entropy framework and suitable reconstructions to facilitate model adaptivity, as well as the work of [DG16], which is based upon and extends the work of [GMP15].

1.2 Thesis structure

The rest of the Thesis is structured as follows. In Chapter 2, inspired by the work of [GLMV16], we obtain reconstructions using various approaches from the discrete solution of an ODE problem discretised using a well known multi-step method. We use these reconstructions to establish a posteriori error control and we compare their performance on the basis of the convergence behaviour of the estimate. This chapter is used to motivate and form the basis for the work in later chapters.

In Chapter 3 we use reconstructions to facilitate an a posteriori analysis of a central difference discretisation of our model elliptic problem. In this regard the reconstruction facilitates an alternative error interpretation and enables us to use the underlying stability framework to obtain the a posteriori estimate. We compare this a posteriori estimate with a classical estimate obtained for a linear Lagrange finite element discretisation of the same problem.

In Chapter 4 we perform an a posteriori error analysis for finite difference methods for the transport equation. This chapter sets the tone and motivates the rest of the thesis by using linear transport to raise the issues that we desire to address in the subsequent chapters. We use simple reconstruction operators in order to highlight the issue of optimality. Subsequently, we validate and benchmark the performance of a reconstruction based a posteriori estimate using several tests.

In Chapter 5 we introduce the numerical schemes we will use for the subsequent chapters. An important component of this chapter pertains to Weighted Essentially Non-Oscillatory (WENO) schemes. WENO schemes are an important class of

schemes in the context of hyperbolic conservation laws on account of their desirable properties, such as high order approximability in smooth regions and non-oscillatory behaviour in the vicinity of shocks. WENO interpolation, a procedure which is a part of WENO schemes, is an important component of the framework we present for obtaining reconstructions of the FD solution. Specifically, it enables us to obtain reconstructions of high polynomial order.

In this chapter we also perform benchmark tests to evaluate the numerical convergence behaviour of the WENO interpolant, using functions of varying regularity, in order to demonstrate its suitability as the spatial component of the reconstructions in subsequent sections.

In Chapter 6 we extend the results of Chapter 4 to linear, symmetric hyperbolic systems in one spatial dimension. We use the WENO interpolation, presented in Chapter 5, to obtain a reconstruction for a model problem, which we then use to facilitate a posteriori error control for the FD discretisation we examine. The behaviour of the a posteriori error estimate is validated numerically.

In Chapter 7 we examine non-linear conservation laws in one spatial dimension. We approximate both scalar and systems problems with well-known and frequently-used FD schemes and we show a posteriori error estimates in different cases. In the scalar case, we numerically test different a posteriori error bounds with the intention of combining them into a single bound that is optimal in both the pre-shock and post-shock regimes. The pre-shock estimate is based on a relative entropy framework, see [GMP15], while the post-shock bound is based on a Kruzkov framework, see [CG95] (see also [DMO07, Ohl09] for relevant reviews). In the systems case we use the relative entropy framework (see [GMP15]) to show a posteriori error bounds for general systems in the regime prior to shock formation. We also use the residual obtained from the reconstruction as a refinement criterion in an adaptive setting. We validate our results using a range of numerical tests.

Finally, in Chapter 8 we conclude the thesis by summarising our contributions and identifying avenues for further research.

Chapter 2

A posteriori analysis for conservative linear multistep methods

Abstract

In this chapter we perform an a posteriori analysis of a model setup of an ordinary differential equation discretised with a well known linear multistep method. We are able to prove an a posteriori upper bound using the energy norm of the problem and a post processing of the discrete solution, inspired by [GLMV16]. We compare three different methods of forming the a posteriori bound introducing appropriate reconstructions of the discrete solution, forming the basis of the work in the subsequent chapters.

2.1 Introduction

In this chapter we obtain and examine reconstructions of a numerical solution to a differential equation. Specifically, we present a simple framework for obtaining reconstructions. We highlight the use of the framework through a reconstruction-based a posteriori error estimate for an illustrative ODE initial value problem, which is approximated using linear multistep methods. We assess the behaviour of the estimator on the basis of convergence characteristics and we compare its performance with an existing estimate for the same problem from the literature, [GLMV16].

2.1.1 Motivation

Our motivation in this chapter is the development of a framework for constructing reliable, optimal, reconstruction based a posteriori error estimates. Let us briefly expand upon the concepts of reconstructions and a posteriori error estimates for the sake of clarity of exposition.

Firstly, a posteriori error estimates are computable error bounds that enable the user to exert local control over the error. The importance of local error control is that it facilitates the implementation of adaptivity using the a posteriori estimate as a refinement/coarsening criterion. In addition, global error control enables a guaranteed use of knowledge of the accuracy of the approximation. Secondly, reconstructions are mathematical objects that can be viewed as post-processors of the discrete approximation to the solution. They are constructed such that they have certain desirable properties (e.g. convergence characteristics). We will expand upon desirable characteristics in the relevant sections.

The ODE model problem serves as a convenient stepping stone for extending the framework to PDEs, which we do in later chapters. In particular, it will pave the way for constructing the temporal component of the reconstructions in PDE problems in subsequent chapters.

There is extensive research in FD methods for ODEs as well as a posteriori error control for ODEs. However, there is not as much interest in the intersection of the two areas. Possible reasons for this may be that FD schemes lack the variational formulation that finite element schemes possess naturally for example. The second challenge is that FD approximations are only defined pointwise in the domain of interest. A lot of the literature on a posteriori error estimates requires globally defined objects. Because of these two facts, it is difficult to obtain a posteriori error estimates for FD schemes and the work done usually caters to FE techniques.

In this chapter we endeavour to use reconstructions as an avenue to compute robust post-processors for FD solutions that are easily usable for establishing a posteriori error control. We then obtain an estimate using the stability framework of the underlying problem in the energy norm. We construct different estimators using our framework and compare them to an existing estimate for this problem from the literature [GLMV16].

2.1.2 Chapter contribution

In this chapter we perform an a posteriori analysis of a second order initial value problem discretized by a well-used two step explicit method: the Leap-frog scheme. We obtain an a posteriori error bound in the problem's energy norm. In order to compute the bound we present and use reconstruction approaches which utilise the numerical solution to construct globally defined interpolants of the solution in the spatial variable. We compare the performance of a bound constructed from these reconstruction approaches to the performance of a bound constructed using the approach of [GLMV16]. The comparison between the different approaches is on the basis of effectivity and rate of convergence.

The rest of this chapter is structured as follows: in §2.2 we introduce our model problem. In §2.2.3 we present the numerical discretization of the problem using the Leap-frog scheme. In §2.2.5 we present the re-formulation of the scheme, summarized from [GLMV16, §2]. The re-formulation is required in order to obtain the reconstruction with the framework of [GLMV16], which will be the benchmark case against which we will compare our results.

In §2.3 we present the a posteriori error bound that we will use to compare the behaviour of reconstructions obtained using our framework with that of the framework of [GLMV16]. We also describe and present the frameworks in detail in this section. We provide an illustrative example in order to illucidate the concept of reconstruction. In particular, in §2.3.8 we present the reconstruction framework fo [GLMV16], in §2.3.10 we present the framework we will be using. As an alternative for obtaining higher order reconstructions we also present a framework based on the WENO reconstruction in §2.3.13. In §2.4 we present numerical experiments based on the reconstructions we propose and the one from [GLMV16]. Finally, we discuss the results in 2.5.

2.2 Setup

In this section we will include some preliminary material required to present the problem we will examine. Despite the fact that the model problem we will consider is an ODE problem, we will nonetheless include considerations for the PDE problem

in the setup phase so as not to have to re-introduce notation later on.

Let $\Omega \subset \mathbb{R}^n$ denote an open bounded set, let $\partial\Omega$ denote the boundary of the set and consider and let $T \in \mathbb{R}^+$. Also, let $f : \Omega \times (0,T] \to \mathbb{R}$, $u_0, v_0 : \Omega \to \mathbb{R}$ denote given functions and let $u : \overline{\Omega} \times [0,T] \to \mathbb{R}$ denote the unknown function. Also, consider a second order differential operator $\mathcal{A}$ such that

$$\mathcal{A}u = -\sum_{i,j=1}^{n} \left(a^{ij}(x,t)\, u_{x_i}\right)_{x_j} + \sum_{i=1}^{n} b^i u_{x_i} + c(x,t)\, u \tag{2.1}$$

for given coefficients $a^{i,j}, b^i, c$ where $i,j = 1,\dots,n$. Now, consider an Initial/Boundary Value Problem (IBVP)

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} + \mathcal{A}u &= f \quad \text{in } \Omega \times (0,T] \\ u &= 0 \quad \text{on } \partial\Omega \times [0,T] \\ u &= u_0 \quad \text{on } \Omega \times \{t=0\} \\ \frac{\partial u}{\partial t} &= v_0 \quad \text{on } \Omega \times \{t=0\} \end{aligned} \tag{2.2}$$

2.2.1 Definition (Second order hyperbolic problem)**.** We say that the partial differential operator $\partial_t^2 + \mathcal{A}$ is (uniformly) hyperbolic if there exists a constant $\theta > 0$ such that

$$\sum_{i,j=1}^{n} a^{ij}(x,t)\, \xi_i \xi_j \geq \theta \left|\xi\right|^2. \tag{2.3}$$

for all $(x,t) \in \Omega \times (0,T]$, $\xi \in \mathbb{R}^n$.

2.2.2 The model problem and notation

We denote by $(\mathrm{H}, \langle\cdot,\cdot\rangle)$ a Hilbert space equipped with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|_{\mathrm{H}}$. We consider the interval $[0,T] \subset \mathbb{R}$, $T > 0$ and denote by $\mathcal{A}$ a positive definite, self-adjoint, linear operator on $D(\mathcal{A})$ -the domain of $\mathcal{A}$-, which is dense in H, i.e. $\overline{D(\mathcal{A})} = \mathrm{H}$, such that $\mathcal{A} : D(\mathcal{A}) \to \mathrm{H}$. Let $\Phi = (\phi_1, \phi_2)$, $\Psi = (\psi_1, \psi_2) \in D(\mathcal{A}) \times \mathrm{H}$. We define the following bilinear form on $[D(\mathcal{A}) \times \mathrm{H}]$:

$$\langle\langle \Phi, \Psi \rangle\rangle := \langle \mathcal{A}^{1/2}\phi_1, \mathcal{A}^{1/2}\psi_1 \rangle + \langle \phi_2, \psi_2 \rangle. \tag{2.4}$$

Note that (2.4) is also the standard energy inner product on $\left[D\left(\mathcal{A}^{1/2}\right) \times \mathrm{H}\right]^2$ and that it induces the energy norm

$$|||\Phi||| := \left(\left\| \mathcal{A}^{1/2}\phi_1 \right\|^2 + \|\phi_2\|^2 \right)^{1/2} \tag{2.5}$$

With the notation in place we can introduce the model problem we are interested in. Specifically, we seek a solution u,

$$\begin{aligned} u: \ [0,T] &\to \mathbb{R} \\ t &\mapsto u(t) \end{aligned} \tag{2.6}$$

to the linear second order hyperbolic problem given by

$$\begin{aligned} \frac{\mathrm{d}^2 u(t)}{\mathrm{d}t^2} + \mathcal{A}u(t) = 0 \quad \text{for} \quad t \in (0,T]\,, \\ u(0) = u_0, \\ \frac{\mathrm{d}u(t)}{\mathrm{d}t}(0) = v_0. \end{aligned} \tag{2.7}$$

where u_0, $v_0 \in \mathrm{H}$ are given functions. In the analysis and examples we present in this section, and in particular in the derivation of the a posteriori bound, it will be helpful to express (2.7) as a system of equations. To do this, we introduce the auxiliary variable $v(t)$, which we define as

$$v(t) := \frac{\mathrm{d}u(t)}{\mathrm{d}t}. \tag{2.8}$$

In order to facilitate exposition, we will denote $\frac{\mathrm{d}(\cdot)(t)}{\mathrm{d}t}$ as $(\cdot)'$ and $\frac{\mathrm{d}^2(\cdot)(t)}{\mathrm{d}t^2}$ as $(\cdot)''$. We use v to write (2.7) as a system of equations:

$$\begin{aligned} u'(t) - v(t) = 0 \quad t \in (0,T] \\ v'(t) + \mathcal{A}u(t) = 0 \quad t \in (0,T] \\ u(0) = u_0 \\ v(0) = v_0. \end{aligned} \tag{2.9}$$

2.2.3 Numerical methods

In this section we present the discretisation of our domain and the numerical method we will use to approximate (2.9). We will closely follow ([GLMV16]) in the notation we use and in the formulation of our numerical approximation. Firstly, we uniformly partition the temporal domain, $[0,T]$ by choosing $0 = t^0 < \cdots < t^N = T$, with constant step size τ. We denote by $u^n := u(t^n)$ the exact solution to (2.7) and we denote by $U^n \in D(\mathcal{A})$, $n = 0, \ldots, N$, the numerical approximation to u^n.

Following the approach of [GLMV16], we approximate the model problem using the Leap-frog scheme. This is an explicit, numerical discretisation scheme for second

order problems. It has desirable conservative properties and second order accuracy. Furthermore, it can be easily formulated as a system of equations, which is useful in this context as we have formulated our model problem as a system of equations as well (see (2.9)). Firstly, we discretise $u'(t)$ and $u''(t)$ by

$$\partial U^{n+1} := \frac{U^{n+1} - U^n}{\tau}, \quad n = 1, \dots, N-1. \tag{2.10}$$

and

$$\partial^2 U^{n+1} := \frac{\partial U^{n+1} - \partial U^n}{\tau} = \frac{U^{n+1} - 2U^n + U^{n-1}}{\tau^2}, \quad n = 1, \dots, N-1. \tag{2.11}$$

We seek approximations $U^{n+1} \in D(\mathcal{A})$ to u^{n+1} such that

$$\begin{aligned} \partial^2 U^{n+1} + \mathcal{A}U^n = 0, \quad n = 1, \cdots, N-1, \\ U^0 = u_0. \end{aligned} \tag{2.12}$$

Note that this method requires two initial conditions: U^0 and U^1. The value for U^1, is obtained as follows

$$\frac{\partial U^1 - v_0}{\tau} + \frac{1}{2}\mathcal{A}U^0 = 0. \tag{2.13}$$

After the first step we can use (2.12) to obtain subsequent values for U^n. The numerical scheme (2.12) can be re-formulated as a numerical discretisation for (2.9) using staggered grids for u and v. This is the approach followed by [GLMV16] as it is useful for the analysis. In order to re-formulate (2.12) as a system we introduce the auxiliary variable

$$V^{n+1/2} := \partial U^{n+1} \quad \text{for} \quad n = 0, \cdots, N-1. \tag{2.14}$$

This serves as an approximation to v half time steps $t^{n+1/2} := \frac{1}{2}(t^n + t^{n+1})$, i.e. $v^{n+1/2} := v\big(t^{n+1/2}\big)$. We define

$$V^{-1/2} := 2v_0 - V^{1/2} \tag{2.15}$$

and use this to define

$$U^{-1} := U^0 - \tau V^{-1/2}. \tag{2.16}$$

Finally, we define

$$\partial V^{n+1/2} := \frac{V^{n+1/2} - V^{n-1/2}}{\tau}, \quad \text{for} \quad n = 0, \cdots, N-1. \tag{2.17}$$

U^n and $V^{n+1/2}$ are used to reformulate the leap-frog scheme as a system of equations on a staggered grid (see Defn. 2.2.4). The reader should note that the re-formulated version - which we present below - and the original numerical discretization are equivalent.

2.2.4 Definition (Leap-frog scheme for (2.9))**.** The Leap-frog scheme for (2.9), formulated as a system of equations on a staggered grid is given by

$$\begin{aligned} \partial U^{n+1} - V^{n+1/2} &= 0, \\ \partial V^{n+1/2} + \mathcal{A}U^n &= 0, \\ U^0 &= 0 \quad \text{and} \\ V^0 &= 1, \end{aligned} \tag{2.18}$$

for $n = 0, \cdots, N-1$.

2.2.5 Re-formulation of the Leap-frog scheme

In order to carry out the analysis of the scheme and obtain residuals and the a posteriori bounds, [GLMV16] define interpolants for V and U on staggered grids. We will introduce all of these interpolants here in order to simplify and expedite our exposition.

2.2.6 Definition. (Interpolants U and V) We denote by U and V respectively the piecewise linear interpolants of the sequences $\{U^n\}_{n=-1}^{N}$ at points $\{t^n\}_{n=-1}^{N}$, $t^{-1} := -\tau$ and $\left\{V^{n+1/2}\right\}_{n=-1}^{N-1}$ at points $\left\{t^{n+1/2}\right\}_{n=-1}^{N-1}$, $t^{-1/2} := -\tau/2$:

$$\begin{aligned} U &:[-\tau, T] \to D(\mathcal{A}) \\ V &:\left[-\tfrac{\tau}{2}, t^{N-1/2}\right] \to D(\mathcal{A}) \end{aligned} \tag{2.19}$$

In order to avoid confusion we provide a graphical depiction for $\{U^n\}_{n=-1}^{N}$, U, $\left\{V^{n+1/2}\right\}_{n=-1}^{N-1}$ and V in Figure 2.1 to elucidate their construction.

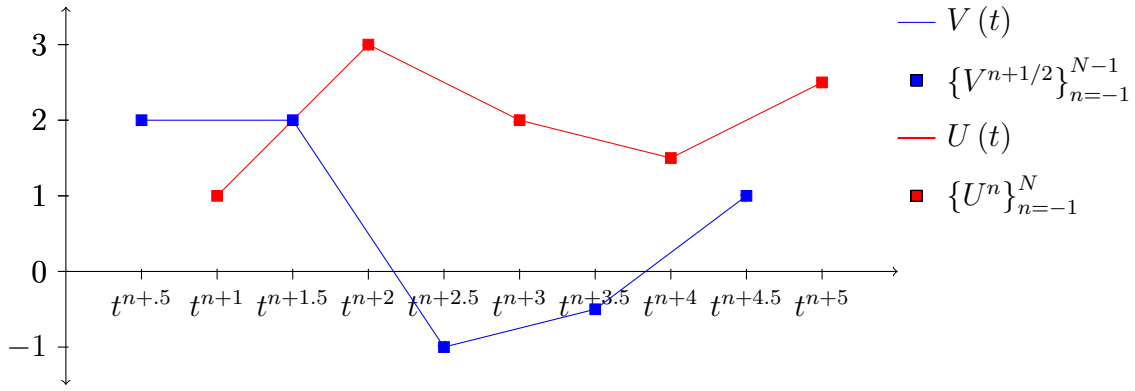


Fig. 2.1. An illustration of $\{U^n\}_{n=-1}^N$, U, $\{V^{n+1/2}\}_{n=-1}^{N-1}$ and V.

The interpolants U and V from (2.19) are subsequently used to define

$$\begin{aligned} U^{n+1/2} &:= U\big(t^{n+1/2}\big) = \frac{1}{2}\big(U^n + U^{n+1}\big) \\ V^n &:= V(t^n) = \frac{1}{2}\big(V^{n-1/2} + V^{n+1/2}\big) \end{aligned} \qquad \text{for } n = 0, \dots, N-1. \tag{2.20}$$

$U^{n+1/2}$ and V^n are used in (2.18) as follows

$$\begin{aligned} \partial U^{n+1} - \frac{1}{2}\big(V^{n+1} + V^n\big) &= V^{n+1/2} - \frac{1}{2}\left(V^{n+3/2} + 2V^{n+1/2} + V^{n-1/2}\right) \\ \partial V^{n+1/2} + \frac{1}{2}\big(U^{n+1/2} + U^{n-1/2}\big) &= -\mathcal{A}U^n + \frac{1}{2}\mathcal{A}\left(U^{n+1} + 2U^n + U^{n-1}\right), \end{aligned} \tag{2.21}$$

which simplifies to

$$\begin{aligned} \partial U^{n+1} - \frac{1}{2}\big(V^{n+1} + V^n\big) &= -\frac{1}{4}\left(V^{n+3/2} - 2V^{n+1/2} + V^{n-1/2}\right) \\ \partial V^{n+1/2} + \frac{1}{2}\big(U^{n+1/2} + U^{n-1/2}\big) &= -\frac{1}{4}\mathcal{A}\left(U^{n+1} - 2U^n + U^{n-1}\right), \end{aligned} \tag{2.22}$$

for $n = 0, \dots, N-1$. Notice that in going from (2.18) to (2.22), we have incurred residuals (the r.h.s. of (2.22)). We formally define these (piecewise constant) residuals as follows:

$$\begin{aligned} R_U(t)\,|_{\left(t^{n-1/2}, t^{n+1/2}\right]} \equiv R_U^n &:= \frac{1}{4}\mathcal{A}\left(U^{n+1} - 2U^n + U^{n-1}\right) \\ R_V(t)\,|_{(t^n, t^{n+1}]} \equiv R_V^{n+1/2} &:= -\frac{1}{4}\left(V^{n+3/2} - 2V^{n+1/2} + V^{n-1/2}\right). \end{aligned} \tag{2.23}$$

2.2.7 Remark (Order of the residuals R_U and R_V)**.** As [GLMV16, §2.1] point out, since the Leap-frog method is second order for both U^n and $V^{n+1/2}$, we have both $R_U^n = O(\tau^2)$ and $R_V^{n+1/2} = O(\tau^2)$ as $\tau \to 0$, provided that the underlying solution is sufficiently regular. In this case, we consider smooth solutions which are regular enough for optimality.

Hence, (2.22) can be viewed as a second order perturbation of (2.18). Proceeding from this point, [GLMV16] write (2.22) as a perturbation of the original problem written as a system of equations i.e. (2.9). They do this by introducing additional interpolants.

2.2.8 Definition. (Interpolants U_1 and V_1) Let $U^{n+1/2}$ and V^n be defined as in (2.20). Denote by U_1 the piecewise linear interpolant of $\left\{U^{n+1/2}\right\}_{n=-1}^{N-1}$ at $\left\{t^{n+1/2}\right\}_{n=-1}^{N-1}$ by V_1 the piecewise linear interpolant of $\{V^n\}_{n=0}^{N-1}$ at $\{t^n\}_{n=0}^{N-1}$ such that

$$\begin{aligned} U_1 &: [0,T] \to D(\mathcal{A}) \\ V_1 &: \left[0, t^{N-1}\right] \to D(\mathcal{A}) \end{aligned} \tag{2.24}$$

Additionally, we define the following interpolators:

1. $\tilde{I}_0$: the piecewise constant midpoint interpolator on $\left\{\left(t^{n-1/2}, t^{n+1/2}\right]\right\}_{n=0}^{N-1}$.
2. I_0: the piecewise constant midpoint interpolator on $\{(t^{n-1}, t^n]\}_{n=1}^{N-1}$.

The interpolants U_1 and V_1 (see Defn. 2.2.8) and U and V (see Defn. 2.2.6) are used to write (2.22) as

$$\begin{aligned} U' - I_0 V_1 &= R_V \\ V' + \mathcal{A}\tilde{I}_0 U_1 &= R_U \end{aligned} \tag{2.25}$$

The system (2.25) serves as the starting point in obtaining the reconstruction of [GLMV16].

2.3 Reconstructions and a posteriori bounds

In this section we motivate the use of reconstructions as post processors of numerical solutions in the context of a posteriori error estimation. We present the relevant a-posteriori bound from [GLMV16], which we will use to benchmark the numerical behavior of the estimator. We will firstly present the reconstruction of [GLMV16]. Then, we will present the reconstruction derived using the framework we propose.

2.3.1 Remark (Remark on a posteriori estimators)**.** Let u and U be the exact and the numerical solution respectively to the problem at hand. A posteriori error control involves establishing an estimate of the form

$$||u - U|| \leq \eta(U)\,, \tag{2.26}$$

where $\eta(U)$ is called the a posteriori estimator. The a posteriori estimator should be explicitly (and, preferably, easily) computable. This implies that $\eta(U)$ should depend on available and explicitly computable quantities, such as the numerical solution and given problem data. Furthermore, it is desirable for $\eta(U)$ to converge optimally; that is, with the same order as the error for the chosen numerical scheme.

2.3.2 Remark. In the context of the temporal FD discretisation (2.18), the numerical solution U which is produced by the scheme is only defined point-wise in the temporal variable. This poses challenges with regard to a posteriori error estimation, because the norms involved in the estimator compare (temporally) globally defined objects. Hence, rather than $\|u - U\|$ we examine an alternative interpetation of the error, namely $\left\|u - \widehat{U}\right\|$, where $\widehat{U}$ is a *reconstruction* of the numerical solution U. In addition, rather than $\eta(U)$, we examine $\eta\big(\widehat{U}\big)$.

We will briefly explain why we use $\widehat{U}$ rather than U both for the (alternative) error estimation $\left\|u - \widehat{U}\right\|$ and for the a posteriori error estimator, with reference to a similar discussion in [AMN09, §1]. The reader should note that the work in [AMN09] pertains to Galerkin methods, whereby the numerical solution is globally defined in time.

The reconstruction, $\widehat{U}$, can be designed to be globally continuous and such that applying the PDE operator to it will result in an explicitly computable quantity. In turn, this $\widehat{U}$ satisfies a perturbed PDE of similar form to the original problem, with the difference being the presence of an explicitly and easily computable residual. We will shortly demonstrate the residual's construction with an example. This residual is then utilized to obtain optimal a posteriori bounds using the stability framework of the PDE (see Lem. 2.3.3). In summary, using the reconstruction $\widehat{U}$ rather than the numerical solution U, where the (pointwise) U is obtained by the FD discretization, allows us to obtain optimal bounds and to use a posteriori estimates which pertain to globally defined objects, of which there are more in the literature.

Recall the problem

$$\begin{aligned} v'(t) + \mathcal{A}u(t) &= 0 \quad t \in (0, T] \\ u'(t) - v(t) &= 0 \quad t \in (0, T] \\ u(0) &= 0 \\ v(0) &= 1 \end{aligned} \tag{2.27}$$

and consider the perturbed problem

$$\begin{aligned} \widehat{V}'(t) + \mathcal{A}\widehat{U}(t) &=: -\mathcal{R}_1(t) \quad t \in (0,T] \\ \widehat{U}'(t) - \widehat{V}(t) &=: -\mathcal{R}_2(t) \quad t \in (0,T] \\ \widehat{U}(0) &= 0 \\ \widehat{V}(0) &= 1, \end{aligned} \tag{2.28}$$

where $\mathcal{R}_1$ and $\mathcal{R}_2$ are the resulting residuals. We define the errors

$$\begin{aligned} \widehat{e}_V &:= v - \widehat{V} \quad \text{and} \\ \widehat{e}_U &:= u - \widehat{U}. \end{aligned} \tag{2.29}$$

Taking the difference between (2.27) and (2.28) we obtain the error equations

$$\begin{aligned} \widehat{e}_V' + \mathcal{A}\widehat{e}_U &= \mathcal{R}_1, \\ \widehat{e}_U' - \widehat{e}_V &= \mathcal{R}_2. \end{aligned} \tag{2.30}$$

We will now formally state the a posteriori error bound we will be using, which is from [GLMV16].

2.3.3 Lemma. *(see [GLMV16, §3: Thm 3.1]) Let (u,v) denote the solution of (2.9), $\widehat{e}_U := u - \widehat{U}$ and $\widehat{e}_V := v - \widehat{V}$ and let $\mathcal{R}_1$ and $\mathcal{R}_2$ be defined as in (2.28). Then, the following a posteriori error estimate holds*

$$\sup_{t\in[0,t^N]} |||(\hat{e}_U, \hat{e}_V)(t)|||^2 \leq 2\,|||(\hat{e}_U, \hat{e}_V)(0)|||^2 + 4\bigg(\int_0^t |||(\mathcal{R}_2, \mathcal{R}_1)|||\,\mathrm{d}t\bigg)^2 =: \eta(t)^2\,. \tag{2.31}$$

Proof. The estimate follows by applying energy arguments to (2.30). The starting point is

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,|||(\hat{e}_U, \hat{e}_V)|||^2 = \langle\!\langle (\hat{e}_U', \hat{e}_V')\,, (\hat{e}_U, \hat{e}_V)\rangle\!\rangle. \tag{2.32}$$

Then, using the definition of the bilinear form $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ (see (2.4)) and (2.30) we obtain

$$\begin{aligned} \langle\!\langle (\hat{e}_U', \hat{e}_V')\,, (\hat{e}_U, \hat{e}_V)\rangle\!\rangle &= \langle \mathcal{A}\hat{e}_U', \hat{e}_U\rangle + \langle \hat{e}_V', \hat{e}_V\rangle \\ &= \langle \mathcal{A}\hat{e}_V, \hat{e}_U\rangle + \langle \mathcal{A}\mathcal{R}_2, \hat{e}_U\rangle - \langle \mathcal{A}\hat{e}_U, \hat{e}_V\rangle + \langle \mathcal{R}_1, \hat{e}_V\rangle \\ &= \langle \mathcal{A}\mathcal{R}_2, \hat{e}_U\rangle + \langle \mathcal{R}_1, \hat{e}_V\rangle. \end{aligned} \tag{2.33}$$

We then use the Cauchy-Schwarz inequality to obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,|||(\hat{e}_U, \hat{e}_V)|||^2 \leq |||(\mathcal{R}_2, \mathcal{R}_1)|||\,|||(\hat{e}_U, \hat{e}_V)|||\,. \tag{2.34}$$

Now, we integrate (2.34) from 0 to τ, with $0 \leq \tau \leq t^N$, such that

$$|||(\hat{e}_U, \hat{e}_V)(\tau)||| = \sup_{t\in[0,t^N]} |||(\hat{e}_U, \hat{e}_V)(t)|||\,, \tag{2.35}$$

which leads us to

$$\begin{aligned}\frac{1}{2}\,|||(\hat{e}_U, \hat{e}_V)(\tau)|||^2 &\leq \frac{1}{2}\,|||(\hat{e}_U, \hat{e}_V)(0)|||^2 + \int_0^\tau |||(\mathcal{R}_2, \mathcal{R}_1)(t)|||\,|||(\hat{e}_U, \hat{e}_V)(t)|||\,\mathrm{d}t \\ &\leq \frac{1}{2}\,|||(\hat{e}_U, \hat{e}_V)(0)|||^2 + |||(\hat{e}_U, \hat{e}_V)(\tau)||| \int_0^\tau |||(\mathcal{R}_2, \mathcal{R}_1)(t)|||\,\mathrm{d}t,\end{aligned} \tag{2.36}$$

where we have used (2.35). Using Cauchy's inequality with ϵ (A.1) with $\epsilon = 1$, $a = \int_0^\tau |||(\mathcal{R}_2, \mathcal{R}_1)(t)|||\,\mathrm{d}t$ and $b = |||(\hat{e}_U, \hat{e}_V)(\tau)|||$ for the product term on the rhs and multiplying by two throughout yields the required result. □

2.3.4 Remark (The error at $t = 0$)**.** The constant $2\,|||(\hat{e}_U, \hat{e}_V)(0)|||^2$ in (7.32) may be non-zero depending on the choice of reconstruction. In particular, for the reconstruction of [GLMV16] (see Defn. 2.3.9) this term is not zero.

We will provide an example in order to illucidate the concepts of a-posteriori error estimation, reconstruction and optimality. For the purposes of this example, we will take $\mathcal{A} := I$. Recall that we denote by $\{U\}_n^N$ and $\left\{V^{n+1/2}\right\}_n^N$ the numerical approximations to $u(t^n)$ and $v\big(t^{n+1/2}\big)$, the solutions to (2.9), obtained using the Leap-frog scheme for the system (2.18).

In order to utilize the a posteriori bound from [GLMV16], we need a globally defined interpretation of the numerical solution, whereas the solutions produced by the Leap-frog scheme are only defined pointwise. A simple approach to address this issue is to use the linear Lagrange interpolants of each of the two sequences, which will serve, in this case, as a simple example of a reconstruction. We will denote these by $\widehat{U}$ and $\widehat{V}$ and they are defined as follows:

$$\begin{aligned}\widehat{U}(t) &:= U^n + \frac{t - t^n}{\tau}\big(U^{n+1} - U^n\big)\,, \quad t \in \big(t^n, t^{n+1}\big]\,, \quad n = 0, \ldots, N-1 \\ \widehat{V}(t) &:= V^{n-1/2} + \frac{t - t^{n-1/2}}{\tau}\big(V^{n+1/2} - V^{n-1/2}\big)\,, \quad t \in \big(t^{n-1/2}, t^{n+1/2}\big]\,, \quad n = 0, \ldots, N.\end{aligned} \tag{2.37}$$

2.3.5 Remark. We note that these reconstructions will lead to a sub-optimal estimate. Nonetheless, we still use them because they elucidate the concept of reconstruction and are very simple to code and use as a benchmark for comparison (with higher order reconstructions) during numerical experiments.

2.3.6 Definition (EOC and EI)**.** To test the validity and robustness of our estimate we will examine the *estimated order of convergence* (EOC) of the estimate and the *effectivity index* (EI).

Consider two sequences $a_i(t)$ and h_i which converge to zero from above we define the EOC for these to be

$$EOC(a_i(t); h_i) := \frac{\log(a_{i+1}(t)/a_i(t))}{\log(h_{i+1}/h_i)}. \tag{2.38}$$

We define the EI at a time t to be the ratio of the estimator and the error at that time, that is:

$$EI(t) := \frac{\mathscr{E}(t)}{\left\|u - \widehat{U}\right\|_X}, \tag{2.39}$$

for some norm X. This allows us to quantify how effective a bound the estimator is over time.

2.3.7 Remark. Generally speaking, the closer to one the EI is (bearing in mind it is also greater than one), the closer to the error the estimate is is. This in turn means that the estimate with the smaller EI is sharper than the one with the larger EI. A small EI (close to one) combined with a high EOC are desirable characteristics for an estimate.

We now have everything we need in order to benchmark the behaviour of the bound

$$\eta_1(t) := \left(2\left|\left|\left|(\hat{e}_U, \hat{e}_V)(0)\right|\right|\right|^2 + 4\left(\int_0^t \left|\left|\left|(\mathcal{R}_2, \mathcal{R}_1)\right|\right|\right| \mathrm{d}t\right)^2\right)^{1/2} \tag{2.40}$$

for the Leap-frog discretisation (2.18) of (2.9) with $\mathcal{A} := I$ and initial conditions $(u_0, v_0) := (0, 1)$. The simulations are conducted using a sequence of time-steps given by $\tau = \frac{2^{-m}}{10}, m = 7, \ldots 10$. The exact solution to this problem is

$$(u(t), v(t)) := (\sin(t), \cos(t)). \tag{2.41}$$

The results are shown in Fig.2.2. Notice that the chosen reconstruction lacks the approximability required to result in an optimal estimate: instead, the estimate converges slower than the error.

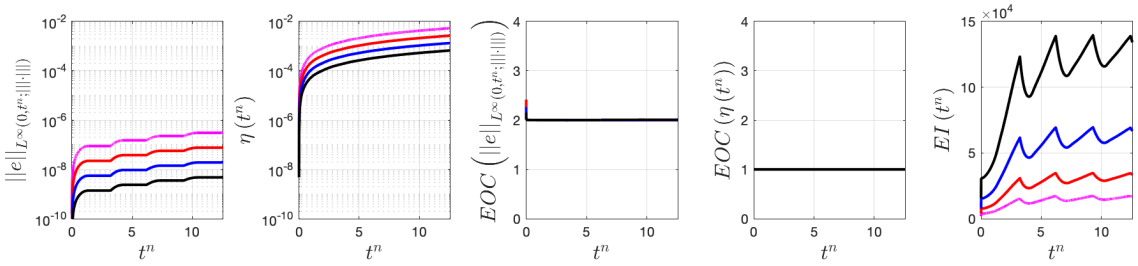


Fig. 2.2. Errors and asymptotic convergence rates for the linear Lagrange interpolant for the Leapfrog approximation (2.18) to (2.9). Notice that the estimate η, given by (2.31), is suboptimal as it converges with a lower rate than the error.

Next, we address the the sub-optimality in the bound. We will use two different methods: that of [GLMV16], which we present in the next section and one which we will introduce in 2.3.10. We will compare the results on the Basis of EOC and EI.

2.3.8 Reconstruction of [GLMV16]

In this section we firstly present the reconstruction from ([GLMV16]).

2.3.9 Definition (Reconstruction from [GLMV16])**.** Let $\{U^n\}_{n=0}^{N}$, $\left\{V^{n+1/2}\right\}_{n=-1}^{N-1}$ denote the numerical approximations to the solution of (2.9) produced by the Leapfrog scheme (2.18). Let U and V denote the piecewise linear interpolants in Defn. 2.2.6 and let U_1 and V_1 denote the piecewise linear interpolants in Defn. 2.2.8. Lastly, let R_U, R_V denote the residuals defined in (2.23). Then, the reconstructions for $\widehat{U}$ and $\widehat{V}$ in [GLMV16, §3.1] are given by

$$\begin{aligned} \widehat{U} &= U^n + \int_{t_n}^{t} (V_1 + R_V)\,\mathrm{d}t,\ t \in (t_n, t_{n+1}] \quad \text{and} \\ \widehat{V} &= V^{n-1/2} + \int_{t^{n-1/2}}^{t} (-\mathcal{A}U_1 + R_U)\,\mathrm{d}t,\ t \in \left(t_{n-1/2}, t_{n+1/2}\right]. \end{aligned} \tag{2.42}$$

We will compare the behavior $\widehat{U}$ and $\widehat{V}$ defined in (2.42) with that of (2.37). The basis for comparison will be the performance of the bound (2.40). Naturally, the residuals $\mathcal{R}_1$ and $\mathcal{R}_2$ will have a different form than the corresponding ones in (2.30) as they depend on different reconstructions. Specifically, for R_U, R_V defined in (2.23), U_1, V_1 defined in, and $\widehat{U}$, $\widehat{V}$ defined in (2.42), the residuals $\mathcal{R}_1$ and $\mathcal{R}_2$ are

given by

$$\begin{aligned}\mathcal{R}_1 &:= -\mathcal{A}\big(\widehat{U} - U_1\big) - R_U,\\ \mathcal{R}_2 &:= \widehat{V} - V_1 - R_V.\end{aligned} \tag{2.43}$$

2.3.10 Reconstruction using our framework

In this section we will present a quadratic reconstruction which is different from (2.42), with the intention of comparing the two in the numerical experiments section. In addition to the nodal equivalence conditions, we will leverage information from the finite diffence schemes to obtain a reconstruction of a higher order. Firstly, we introduce the space of piecewise polynomials.

2.3.11 Definition (Space of piecewise polynomials)**.** Let $\mathbb{P}^q\big(\big[t^{n-1}, t^n\big], \mathbb{R}\big)$ denote the space of real-valued polynomials of degree q in the interval $[t^{n-1}, t^n]$. Then, we define

$$\mathbb{V}_q := \Big\{w : [0,T] \to \mathbb{R} : w\|_{[t^{n-1},t^n]} \in \mathbb{P}^q\Big(\big[t^{n-1}, t^n\big], \mathbb{R}\Big)\Big\} \tag{2.44}$$

to be the space of piecewise polynomials of degree q.

Next, we introduce the reconstruction

2.3.12 Definition. Quadratic Reconstruction for the numerical solution of (2.18) We define the reconstruction, $\big(\widehat{U}, \widehat{V}\big)$, of the numerical solution, (U, V), of (2.18) on a staggered grid consists of the functions $\widehat{U} \in \mathbb{V}_2$, $\widehat{V} \in \mathbb{V}_2$ which satisfy the following set of conditions:

$$\begin{aligned}&\widehat{U}(t^n) = U^n\\ &\widehat{U}''|_{(t^n,t^{n+1})}(t) = -U^n, \quad \forall t \in \big(t^n, t^{n+1}\big)\\ &\widehat{V}\big(t^{n+1/2}\big) = V^{n+1/2},\\ &\widehat{V}'|_{\left(t^{n-1/2},t^{n+1/2}\right]}\big(t^{n+1/2}\big) = -\widehat{U}\big(t^{n+1/2}\big)\end{aligned} \tag{2.45}$$

2.3.13 WENO Reconstruction

In the previous sections we have seen a second order reconstruction, which converges at an optimal rate for the given scheme. In this section, we will present an alternative method of obtaining temporal reconstructions, using WENO interpolation (see [JSB+19], [LSZ09]), which may serve as an avenue for obtaining higher order temporal reconstructions. In the remainder of this section, we will introduce the

WENO interpolant and walk through its construction for this problem. We will then use it to obtain reconstructions $\widehat{U}$ and $\widehat{V}$. We will compare the performance of a reconstruction obtained in this way with the results from the preceding sections.

We consider a uniform partition of the temporal variable $0 = t^0 < \cdots < t^N = T$ with constant time-step τ. Consider a function $u(t)$ with a set of point values $\{u^n := u(t^n)\}$ at times $\{t^n\}$. We want to construct a third order WENO intepolating polynomial in an interval $[t^n, t^{n+1}]$ by using the 4-point stencil

$$S := \left\{t^{n-1}, \ldots, t^{n+2}\right\} \tag{2.46}$$

The interpolant is obtained as a convex combination of polynomials which are constructed on two 3-point sub-stencils, S_1 and S_2 of S, which are given by

$$\begin{aligned} S_1 &:= \left\{t^{n-1}, t^n, t^{n+1}\right\}, \\ S_2 &:= \left\{t^n, t^{n+1}, t^{n+2}\right\}. \end{aligned} \tag{2.47}$$

The polynomials are Lagrange interpolants over the sub-stencils:

$$\begin{aligned} p_1(x) &:= U^{n-1}\frac{(t-t^n)(t-t^{n+1})}{2\tau^2} + U^n\frac{(t-t^{n-1})(t-t^{n+1})}{\tau^2} + U^{n+1}\frac{(t-t^{n-1})(t-t^n)}{2\tau^2} \quad \text{and} \\ p_2(x) &:= U^{n}\frac{(t-t^{n+1})(t-t^{n+2})}{2\tau^2} + U^{n+1}\frac{(t-t^{n})(t-t^{n+2})}{\tau^2} + U^{n+2}\frac{(t-t^{n})(t-t^{n+1})}{2\tau^2} \end{aligned} \tag{2.48}$$

for $t \in [t^n, t^{n+1}]$. A polynomial approximation to $u(t)$, $p(t)$, can be obtained as a convex combination of the $p^{(i)}$. The WENO approach is such that $p(t)$ is a high order approximation in intervals where $u(t)$ is smooth. $p(t)$ is obtained as a weighted sum of the $p^{(i)}$ with the (linear) weights γ_1 and γ_2, each corresponding to a sub-stencil of the large stencil:

$$\begin{aligned} \gamma_1(t) &:= -\frac{t-t^{n+2}}{t^{n+2}-t^{n-1}} \quad \text{and} \\ \gamma_2(t) &:= \frac{t-t^{n-1}}{t^{n+2}-t^{n-1}}. \end{aligned} \tag{2.49}$$

The linear weights are positive and satisfy

$$\sum_i \gamma_i = 1. \tag{2.50}$$

Interested readers can find details on the construction of these weights in ([CFR05]) and [LSZ09]. If the solution is discontinuous inside a sub-stencil, we would like that stencil to have little contribution to ensure the non-oscillatory behaviour of the

scheme. This is achieved by using the non-linear weights $\omega_i(t)$, which are obtained from the $\gamma_i(t)$ as follows:

$$\omega_j(t) := \frac{\alpha_j(t)}{\sum_{i=1}^{2} \alpha_i(t)}, \quad \alpha_i(t) := \frac{\gamma_i(t)}{\epsilon + \beta_i}, \tag{2.51}$$

where the β_i are the *smoothness indicators* for the sub-stencil to which they pertain. They are an indication of how non-smooth the solution is in the corresponding sub-stencil. If the solution is smooth in the sub-stencil S_j, then the relevant β_j is small and the relevant ω_j is close to the γ_j in S_j. If instead the solution has a discontinuity in S_j, then the β_j is large, leading to a small ω_j and ensuring the non-oscillatory behaviour.

The β_i which are used in this paper are given in [JSB+19]. For a constant time-step τ, they are are defined as

$$\begin{aligned} \beta_1 &:= 4\Big(\big|y'_{j+1} - y'_j\big| - \big|y'_j - y'_{j-1}\big|\Big)^2 \quad \text{and} \\ \beta_2 &:= 4\Big(\big|y'_{j+2} - y'_{j+1}\big| - \big|y'_{j+1} - y'_j\big|\Big)^2 . \end{aligned} \tag{2.52}$$

The calculation of the y'_i is presented in detail in [JSB+19, §3.3.2]. Finally, the WENO approximation to $u(t)$ in the interval $[t^n, t^{n+1}]$ based on the stencil $S = S_1 \cup S_2 = \{t^{n-1}, t^n, t^{n+1}, t^{n+2}\}$ can be obtained as

$$p(t) := \omega_1 p_1(t) + \omega_2 p_2(t) . \tag{2.53}$$

Now we can define the spatio-temporal reconstruction in terms of the WENO approximation.

2.3.14 Definition (WENO temporal reconstruction)**.** The WENO temporal reconstruction, $\widehat{U}$, of the numerical solution, U^n, of (2.18) is obtained as the WENO interpolant (2.53) for $t \in [t^n, t^{n+1}]$ on the stencil $S := \{t^{n-1}, t^n, t^{n+1}, t^{n+2}\}$. The WENO temporal reconstruction $\widehat{V}$ is obtained in the same way for $t \in \big[t^{n-1/2}, t^{n+1/2}\big]$. In this case, the process must be modified appropriately to reflect the fact that the interpolant is over a staggered grid and the stencil is $S := \big\{t^{n-3/2}, t^{n-1/2}, t^{n+1/2}, t^{n+3/2}\big\}$.

2.3.15 Remark (WENO reconstruction for $\widehat{V}$)**.** The reader should note that in order to calculate the a posteriori estimate (2.40) using the WENO reconstructions, one to store four consecutive values for U^n and five for $V^{n+1/2}$. Specifically, we require $\{U^{n-1}, U^n, U^{n+1}, U^{n+2}\}$, $\big\{V^{n-3/2}, V^{n-1/2}, V^{n+1/2}, V^{n+3/2}\big\}$ and

$\left\{V^{n-1/2}, V^{n+1/2}, V^{n+3/2}, V^{n+5/2}\right\}$. The reason for the two sets of values for $V^{n+1/2}$ is because the FD solution for $V^{n+1/2}$ is defined on the staggered grid, whereas the estimator is calculated (by choice) on the integer valued grid.

2.3.16 Remark (Initial calculation of the a posteriori estimate)**.** The reader will notice that the WENO reconstruction requires four nodal values for U^n and five for $V^{n+1/2}$. In the first few time-steps these values are not available. Hence, for the first two steps, we utilitze a different reconstruction, namely Defn. 2.3.9.

2.3.17 Remark (Positivity of linear weights)**.** In order to avoid treating negative weights (cf. [LSZ09]), we always evaluate the polynomial (2.53) in the middle interval, $[t^n, t^{n+1}]$ of the 4-point stencil $S := \{t^{n-1}, t^n, t^{n+1}, t^{n+2}\}$. The practical implication is that we must calculate U^{n+2}, $V^{n+3/2}$ and $V^{n+5/2}$. Furthermore, in order to calculate the interpolant we must store four time-steps worth of values for U^n and five for $V^{n+1/2}$.

2.4 Numerical Experiments

In this section we run numerical benchmarking experiments for the reconstructions we introduced in §2.3.8, §2.3.10 and §2.3.13. We will use the same model problem and the same initial conditions as we did for the sub-optimal, linear reconstruction we presented as an example in (2.37). Specifically, we use $\mathcal{A} := I$ and initial conditions $(u_0, v_0) = (0, 1)$. All the simulations are conducted using a sequence of time-steps given by $\tau = \frac{2^{-m}}{10}, m = 7, \ldots 10$. The results are shown in Fig. 2.3 for the reconstruction obtained using Defn. 2.3.9, Fig. 2.4 for Defn. 2.3.12 and lastly, Fig. 2.5 for a reconstruction obtained using Defn. 2.3.14.

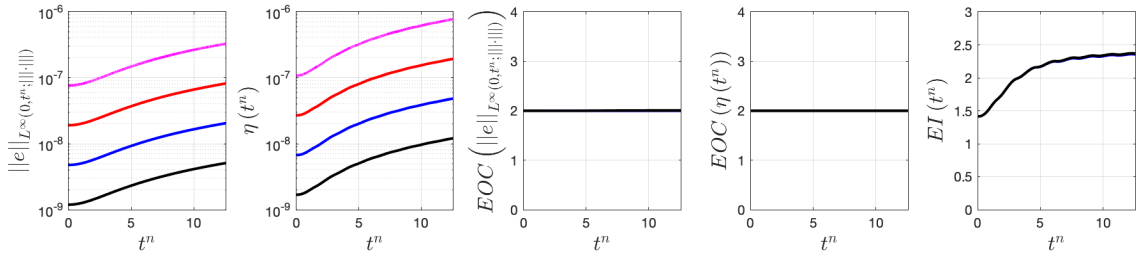


Fig. 2.3. Errors and asymptotic convergence rates for the reconstruction of ([GLMV16]), given by (2.42), for the Leap-frog approximation (2.18) to (2.9). The a posteriori estimate, η, (2.31), is optimal.

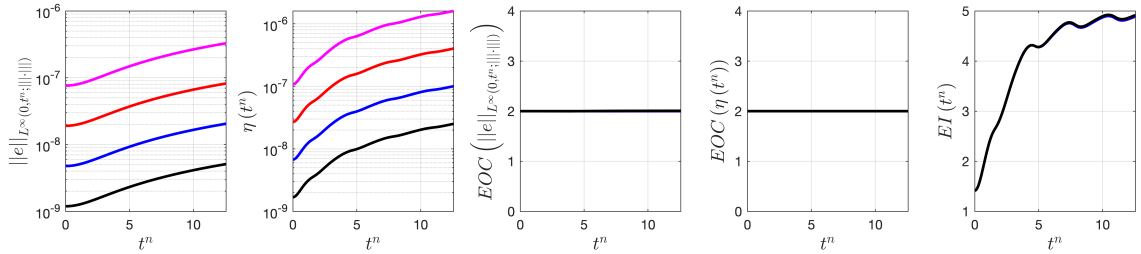


Fig. 2.4. Errors and asymptotic convergence rates for reconstruction obtained from Defn. 2.3.12 for the Leap-frog approximation, (2.18), to (2.9). The estimate, η, (2.31) is also optimal, as in Fig. 2.3, albeit with a higher (worse) EI.

The results for this test are shown in Fig.2.5. Notice that the a posteriori bound converges optimally.

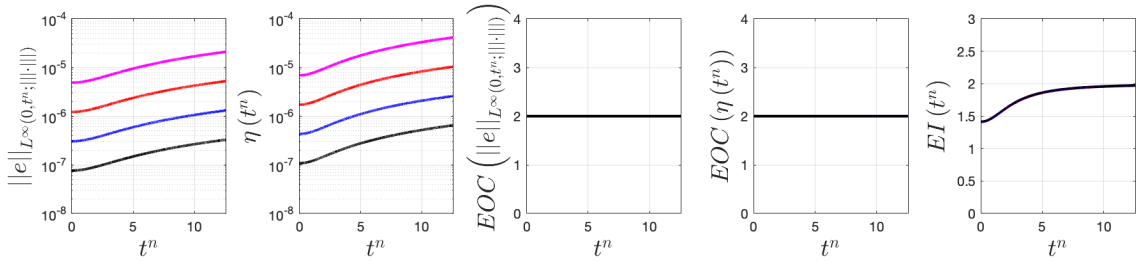


Fig. 2.5. Errors and asymptotic convergence rates for the WENO reconstruction from Defn. 2.3.14 for the Leap-frog approximation (2.18) to (2.9). The estimate, η, given by (2.31) is optimal and has a lower EI than in Fig. 2.3.

2.5 Discussion

We constructed the a posteriori bound of [GLMV16, Thm. 3.1] and used it to compare different reconstructions approaches with regard to the temporal aspect of the problem using an IVP ODE model problem. Specifically, we compared the reconstruction of [GLMV16], (2.42) with three different reconstructions: a linear and a quadratic reconstruction obtained using the framework in Defn. 2.3.12 and a WENO reconstruction obtained as explained in Defn. 2.3.14.

The comparison between different methods of constructing the bounds was conducted the following grounds: the evolution of the L^∞-norm of the errors defined in (2.29), the evolution of the estimator defined in (2.40), a comparison of the Experimental Order of Convergence (EOC) of these quantities, (2.38) and lastly, a comparison of the Effectivity Index (EI) (2.39).

The first numerical test we run is the first order, linear reconstruction given by (2.37). The results are shown in Fig. 2.2. We can immediately see that the a posteriori bound is one order sub-optimal relative to the order of the error for the scheme; the estimate converges with order one whereas the scheme converges with order two.

In the second and third numerical tests we construct the a posteriori estimate using the frameworks of [GLMV16] (see Defn. 2.3.9) and the framework we introduce in this chapter (see Defn. 2.3.12). The results are shown in Figures 2.3 and 2.4 respectively for the two tests. In both cases the error estimates are optimal. However, the estimate constructed using Defn. 2.3.9 results in a lower EI (overall) compared to the one using Defn. 2.3.12.

Despite this difference in effectivity, there is merit in using the quadratic reconstruction described in the framework from Defn. 2.3.12. In particular, by incorporating additional information from the FD scheme (such as another FD quotient at the unused temporal sub-interval endpoint) , this framework gives us an avenue for constructing a posteriori estimates up to order three in the temporal component (for solutions which possess sufficient regularity). Indeed, we use this in later chapters to construct optimal estimates for the temporal component of reconstructions for FD schemes which are up to order three in time.

The last framework we examine for obtaining reconstructions is Defn. 2.3.14.

In this case, we use the WENO polynomial construction process to obtain the construction. The results are shown in Fig. 2.5. We can see that the estimate is of optimal order and the EI is slightly improved compared to the other two optimal reconstructions. More importantly, this method can be used - at least in principle- to obtain reconstructions of optimal order for schemes higher than order three. In later chapters we will use WENO reconstructions for the spatial component of the reconstruction process for hyperbolic problems. It is also important to note that the computational stencil is wider than a single sub-interval, meaning that the construction process is not localizable to a single interval.

Chapter 3

Simple a posteriori control of finite difference discretisations of elliptic problems

Abstract

In this chapter we perform an a posteriori analysis of a model elliptic problem discretised by a central finite difference scheme. The analysis is based on a reconstruction of the discrete solution, which facilitates an alternative error interpretation. We use this interpretation to construct robust a posteriori error bounds and compare the performance of such bounds with classical bounds obtained for linear Lagrange finite element discretisation of the same problems.

3.1 Introduction

In this chapter we shift our focus to the spatial component of the reconstruction process. Continuing the work from the previous chapter, we use the reconstruction approach to construct computable a posteriori bounds for a reconstruction of the FD approximation to an elliptic model problem.

The performance of the bound will be assessed with regard to its convergence behaviour using numerical experiments with solutions of varying regularity. The reconstruction-based a posteriori bound for the FD discretisation will be compared with an a posteriori bound obtained for a Finite Element (FE) formulation of the

same problem using linear Lagrange elements.

3.1.1 Motivation

Our motivation in this chapter is to derive and examine robust a posteriori error bounds for FD schemes without using an equivalent Finite Element-based formulation. In this regard, we adapt the idea in Chapter 2 and create a framework for constructing a posteriori error bounds for FD discretisations of elliptic problems using the idea of reconstructions.

The novelty in this work is that it enables the user of the method to utilise existing bounds for elliptic problems in the context of FD discretisations through the use of an alternative error interpretation, which we will specify in this chapter. The alternative error interpretation is based upon the post-processing of the pointwise FD solution and, crucially, it does not rely upon recasting the FD discretisation as an equivalent FE discretisation.

3.1.2 Chapter contribution

The main contribution from this chapter is a framework for obtaining robust a posteriori error bounds for the central FD discretisation of the model elliptic problem - the Poisson equation. The bound is based on a reconstruction - a continuously defined post-processed object which is a function of the point-wise defined FD solution. The post-processor enables us to provide an alternative, "globally-defined" interpretation of the error which facilitates the use of a posteriori error bounds which require globally defined objects, such as classical residual error bounds. We demonstrate how to obtain a reconstruction using this framework. We then use the reconstruction to obtain computable bounds.

We emphasise the fact that we do not recast the FD problem as an equivalent FE problem (see e.g. [CET14]). However, we note that there is an interesting connection between the FE discretisation of our problem using linear triangular Lagrange elements on a uniform, regular mesh with the 5-point FD discretisation on the same problem on the same grid (i.e. same, uniform node placement). In particular, this connection is that the resulting discrete operator is the same up to a constant for the two discretisations. This in turn means that it is computationally inexpen-

sive to draw comparisons between the performance of the bound obtained using the approach we introduce, with a classical posteriori bound for the Finite Element method with linear triangular elements as described. The additional computational expenses we incur for the FE discretisation is the computation of the right hand side and the multiplication of the FD operator to obtain the FE discrete operator, which simply involves multiplication with a constant. However, since the left hand side operator is available, this simplifies the process. In this way, we can, relatively inexpensively, create a benchmark set of examples, where the solution behaviour is well understood, to enable insights to be made on more complicated problems in later chapters.

The rest of the chapter is structured as follows. In §3.2 we introduce the notation we adopt throughout the rest of the chapter and present the model problem we will be using as a test case. We also present and prove the bound we will test in later sections. In §3.3 we present the Finite Difference (FD) numerical discretisation we will use for our problem. In §3.4 we present the reconstruction procedure for the model problem. In §3.3.4 we present the FE discretisation of the model problem. We use the FE discretisation in deriving the main result of the section, Lemma 3.5.4 as well as to produce classical a posteriori bounds, which we use for comparisons in the subsequent section. In §3.6 we test the convergence behaviour of an a posteriori bound constructed using the reconstruction for a range of test cases of varying regularity. The chapter is concluded in §3.7.

3.2 Setup and preliminaries

In this section we introduce notation and present the tools we will use in subsequent sections. We will present the model problem in strong and weak form as well as the necessary Sobolev spaces machinery we will need in this regard.

3.2.1 Spaces of Continuous functions

We denote by $\boldsymbol{\alpha}$ the multi-index $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_N) \in \mathbb{N}_0^N$ and we denote by $|\boldsymbol{\alpha}|$ the length of $\boldsymbol{\alpha}$ given by $|\boldsymbol{\alpha}| := \sum_{i=1}^{N} \alpha_i$. We use the multi-index notation for expository

convenience with regard to derivatives:

$$\mathrm{D}^{\boldsymbol{\alpha}} := \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_N}\right)^{\alpha_N} \tag{3.1}$$

For some domain $\Omega \subset \mathbb{R}^N$, we denote by $\mathrm{C}^k(\Omega)$ the set of all continuous real-valued functions, say u, on Ω such that $\mathrm{D}^{\boldsymbol{\alpha}} u$ is continuous on Ω for all $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ for which $|\boldsymbol{\alpha}| \leq k$. Let $\overline{\Omega}$ denote the closure of Ω. We denote by $\mathrm{C}^k\big(\overline{\Omega}\big)$ the set of all functions $u \in \mathrm{C}^k(\Omega)$ for which $\mathrm{D}^{\boldsymbol{\alpha}} u$ can be extended to a continuous function on $\overline{\Omega}$ for all $\boldsymbol{\alpha}$ for which $|\boldsymbol{\alpha}| \leq k$. We denote by $\mathrm{C}_0^k(\Omega)$ the set of functions $u \in \mathrm{C}^k(\Omega)$ whose support is a bounded subset of Ω. Furthermore, we denote by $\mathrm{C}_0^\infty(\Omega)$ the set of functions $u \in \mathrm{C}_0^k(\Omega)$ $\forall k \geq 0$, i.e. the set of infinitely differentiable functions with compact support.

3.2.2 Elliptic model problem

Now, let $\Omega \subset \mathbb{R}^2$ denote a polygonal, bounded, connected domain with boundary $\partial\Omega$. We consider the linear second order partial differential equation (PDE) with homogeneous Dirichlet boundary conditions:

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{3.2}$$

where $f : \overline{\Omega} \to \mathbb{R}$.

Next, we give a Maximum Principle result, which is used to prove uniqueness of solutions to (3.2). This is not the topic of this chapter but we do use a discrete version of the Maximum Principle in 3.3 to prove uniqueness of solution for the FD discretisation of (3.2) we will introduce later.

3.2.3 Theorem (Maximum Principle)**.** *Assume that $u \in \mathrm{C}^2(\Omega) \cap \mathrm{C}^0\big(\overline{\Omega}\big)$. Then, if $\Delta u \geq 0$ (resp. $\Delta u \leq 0$) in a bounded domain Ω, then the maximum (resp. the minimum) is achieved on the boundary $\partial\Omega$.*

3.2.4 Lebesgue and Sobolev spaces

Sobolev spaces are the required setting for a lot of the results of relevance to this chapter, in particular with regard to weak formulation of the model problem and the

FE discretisation. We firstly introduce the notation pertaining to Lebesgue spaces, since we will not only be using it here but also in later chapters (see [ES11]):

$$\mathrm{L}^p(\Omega) = \left\{\phi\colon \int_\Omega |\phi|^p < \infty\right\} \quad \text{for } p \in [1,\infty) \text{ and } \mathrm{L}^\infty(\Omega) = \left\{\phi\colon \operatorname{ess\,sup}_{x\in\Omega} |\phi(x)| < \infty\right\}. \tag{3.3}$$

which are equipped with corresponding norms

$$\|u\|_{\mathrm{L}^p(\Omega)} = \begin{cases} (\int_\Omega |u|^p)^{1/p} \text{ for } p \in [1,\infty), \\ \operatorname{ess\,sup}_{x\in\Omega} |u(x)| \text{ for } p = \infty. \end{cases} \tag{3.4}$$

We make extensive use of $\mathrm{L}^2(\Omega)$, which is equipped with the inner product

$$\langle u, v\rangle_{\mathrm{L}^2(\Omega)\times\mathrm{L}^2(\Omega)} := \int_\Omega uv. \tag{3.5}$$

We also consider the Sobolev spaces

$$\mathrm{W}^{k,p}(\Omega) = \{\phi \in \mathrm{L}^p(\Omega) : \mathrm{D}^\alpha\phi \in \mathrm{L}^p(\Omega) \quad \text{for} \quad |\alpha| \le k\}, \tag{3.6}$$

where the derivatives $\mathrm{D}^{\boldsymbol{\alpha}}$ are understood in the weak sense. The Sobolev spaces are equipped with norms and semi-norms given by

$$\|u\|_{\mathrm{W}^{k,p}(\Omega)} := \left(\sum_{|\boldsymbol{\alpha}|\le k} \|\mathrm{D}^{\boldsymbol{\alpha}} u\|^p_{\mathrm{L}^p(\Omega)}\right)^{1/p} \text{ and } |u|_{\mathrm{W}^{k,p}(\Omega)} := \left\|\mathrm{D}^k u\right\|^p_{\mathrm{L}^p(\Omega)}. \tag{3.7}$$

3.2.5 Remark. We use the notation $\mathrm{W}^{k,p}_0(\Omega)$ to denote the closure of $\mathrm{C}^\infty_0(\Omega)$ in $\mathrm{W}^{k,p}(\Omega)$. This is the set of functions $u \in \mathrm{W}^{k,p}(\Omega)$ for which there exist $u_m \in \mathrm{C}^\infty_0(\Omega)$ such that $u_m \to u$ in $\mathrm{W}^{k,p}(\Omega)$.

3.2.6 Remark. For Sobolev spaces, we are interested in the case $p = 2$. For this reason, we will use the notation $\mathrm{H}^k(\Omega)$ and $\mathrm{H}^k_0(\Omega)$ to denote the spaces $\mathrm{W}^{k,2}(\Omega)$ and $\mathrm{W}^{k,2}_0(\Omega)$, respectively. We will equip the spaces $\mathrm{H}^k(\Omega)$ with norms and semi-norms

$$\begin{aligned} \|u\|^2_{\mathrm{H}^k(\Omega)} &:= \sum_{|\boldsymbol{\alpha}|\le k} \|\mathrm{D}^{\boldsymbol{\alpha}} u\|^2_{\mathrm{L}^2(\Omega)} \\ |u|^2_{\mathrm{H}^k(\Omega)} &:= \sum_{|\boldsymbol{\alpha}|= k} \|\mathrm{D}^{\boldsymbol{\alpha}} u\|^2_{\mathrm{L}^2(\Omega)} \end{aligned} \tag{3.8}$$

We will also make use of the dual Sobolev spaces. We will denote the dual of $\mathrm{H}^k_0(\Omega)$ by $\mathrm{H}^{-k}(\Omega)$. In particular, we will use $\mathrm{H}^{-1}(\Omega)$. In this regard, we denote the

duality pairing between $\mathrm{H}_0^1(\Omega)$ and $\mathrm{H}^{-1}(\Omega)$ as $\langle \cdot \,|\, \cdot \rangle_{\mathrm{H}^{-1}(\Omega)\times \mathrm{H}_0^1(\Omega)}$. The space $\mathrm{H}^{-k}(\Omega)$ is equipped with the norm

$$\|u\|_{\mathrm{H}^{-k}(\Omega)} := \sup_{0\neq\phi\in \mathrm{H}_0^k(\Omega)} \frac{\langle u \,|\, \phi\rangle_{\mathrm{H}^{-k}(\Omega)\times \mathrm{H}_0^k(\Omega)}}{|\phi|_{\mathrm{H}^k(\Omega)}} \tag{3.9}$$

Now that we have the necessary Lebesgue and Sobolev space machinery we will re-cast the model problem (3.2) into its weak form, which is the form we will utilise throughout this chapter.

3.2.7 Definition (Linear and Bilinear forms)**.** We will use the short-hand $a(\cdot,\cdot) : \mathrm{H}_0^1(\Omega) \times \mathrm{H}_0^1(\Omega) \to \mathbb{R}$ and $l(\cdot) : \mathrm{H}_0^1(\Omega) \to \mathbb{R}$ to denote the bilinear and linear forms respectively that we will be using in the FE formulation we will introduce in this section. These are defined as follows:

$$\begin{aligned} a(u,v) &:= \int_\Omega \nabla u \cdot \nabla v \mathrm{d}x, \\ l(v) &:= \int_\Omega f v \mathrm{d}x, \end{aligned} \tag{3.10}$$

where $f : \overline{\Omega} \to \mathbb{R}$.

3.2.8 Definition. (Weak formulation and discretisation) Let $a(\cdot,\cdot) : \mathrm{H}_0^1(\Omega)\times\mathrm{H}_0^1(\Omega) \to \mathbb{R}$ and $l(\cdot) : \mathrm{H}_0^1(\Omega) \to \mathbb{R}$ be defined as in (3.10). The weak formulation of (3.2) is to find $u \in \mathrm{H}_0^1(\Omega)$ such that

$$a(u,v) = l(v) \quad \forall v \in \mathrm{H}_0^1(\Omega)\,. \tag{3.11}$$

3.2.9 Lemma. *(Lax-Milgram [Cia02]) Consider a bilinear functional $a(\cdot,\cdot) : \mathrm{H}_0^1(\Omega)\times \mathrm{H}_0^1(\Omega) \to \mathbb{R}$ which satisfies the following:*

1. *$\exists c_0 > 0$ such that $c_0 \|v\|^2_{\mathrm{H}_0^1(\Omega)} \leq a(v,v) \quad \forall v \in \mathrm{H}_0^1(\Omega)$,*
2. *$\exists c_1 > 0$ such that $|a(v,w)| \leq c_1 \|v\|_{\mathrm{H}_0^1(\Omega)} \|w\|_{\mathrm{H}_0^1(\Omega)} \quad \forall v,w \in \mathrm{H}_0^1(\Omega)$.*

Also let $l(\cdot) : \mathrm{H}_0^1(\Omega) \to \mathbb{R}$ denote a linear functional such that

$$\exists c_2 > 0 \ s.t. \ |l(v)| \leq c_2 \|v\|_{\mathrm{H}_0^1(\Omega)} \quad \forall v \in \mathrm{H}_0^1(\Omega)\,. \tag{3.12}$$

Then, there exists a unique $u \in \mathrm{H}_0^1(\Omega)$ such that

$$a(u,v) = l(v) \quad \forall v \in \mathrm{H}_0^1(\Omega)\,. \tag{3.13}$$

3.2.10 Remark. (Existence and uniqueness of weak solutions to (3.11)) The variational formulation (3.11) possesses a unique solution as it satisfies the conditions of the Lax-Milgram theorem .

3.3 Numerical Methods and Discretisation

In this section we will present the domain discretisation of Ω using a uniform grid. We will also present the FE and FD discretisations of (3.2).

3.3.1 Domain discretisation

3.3.2 Definition. (Discretisation of Ω) We let $\Omega := [0,1]^2$ denote the unit square and we denote its boundary by $\partial\Omega$. We define a grid over Ω by choosing points $(x_i, y_j) \in \Omega$ such that $0 = x_0 < \ldots, x_M = 1$ and $0 = y_0 < \cdots < y_M = 1$ (see Fig. 3.1). We will use uniform grid spacing $h := 1/M$ in both directions. We will denote by Ω_h the set of interior grid-points:

$$\Omega_h := \{(ih, jh) : i, j = 1, \ldots, M-1\} \tag{3.14}$$

and by $\partial\Omega_h$ the set of boundary grid-points:

$$\partial\Omega_h := \{(ih, jh) : i, j = 0, M\}. \tag{3.15}$$

We will define $\overline{\Omega}_h := \Omega_h \cup \partial\Omega_h$. Lastly we denote by I_{ij} the quadrilateral $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$, $i, j = 0, \ldots, M-1$.

3.3.3 Remark. Note that the results can be extended a mesh with non-uniform spacings.

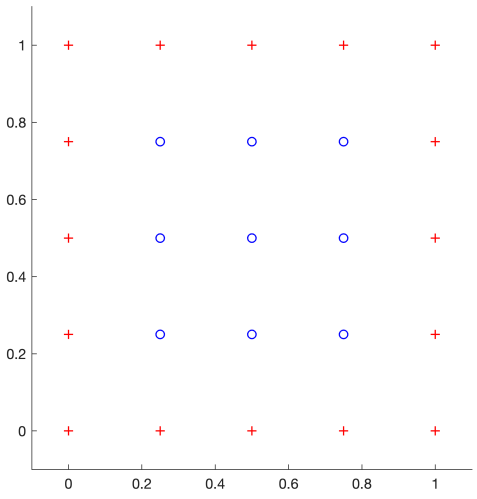


Fig. 3.1. A uniform grid over Ω. We denote by Ω_h the set of interior grid-points (circles) and by $\partial\Omega_h$ the set of boundary grid points (plus sign).

3.3.4 Finite Element approximation

In this section we present a Finite Element approximation of the model problem. Our motivation for doing this is that in the specific problem we are considering, the FE method posed over a regular diagonal mesh (as shown in Fig. 3.2) with linear Lagrange elements, results in the same discrete operator (up to a constant power of h) as the one obtained from the central FD discretisation on the same grid (i.e. same node placement). In this sense, since the left hand side operator is available (up to a constant) by using either discretisation as a starting point, and since no additional computational expense is incurred from re-derivation, it is interesting to compare the behaviour of classical and well-used FE bounds with bounds based on the reconstruction we will obtain from the FD solution. We caution the reader that the two discretisations are NOT the same since neither the left nor the right hand sides are the same for the two different discretisation routes.

3.3.5 Remark. We will use the uniform grid in Defn. 3.3.2 in order to pose the triangulation, $\mathcal{T}$ we will use for the FE discretisation of (3.11). It is important for the FE discretisation to have the same left-hand side matrix (up to a constant power of h) as the FD discretisation.

3.3.6 Definition. (Triangulation of Ω) We define a triangulation, $\mathcal{T}$, of $\overline{\Omega}$ by using the grid defined in Defn. 3.3.2 (see also Fig. 3.2). We will use using uniform triangular elements, denoted K. We denote by ∂K the boundary of the element K. We denote by $\mathcal{E}$ the set of edges associated with $\mathcal{T}$. We note that the elements K of $\mathcal{T}$ intersect along complete edges, at vertices or not at all. We denote by h_E the length of edge $E \in \mathcal{E}$. We denote by h_K the diameter of the element K, which is defined to be the length of is longest side. We denote by ρ_K the radius of the largest ball contained in K. Lastly, we denote by $\Omega_{h,K}$ the set of vertices of element K.

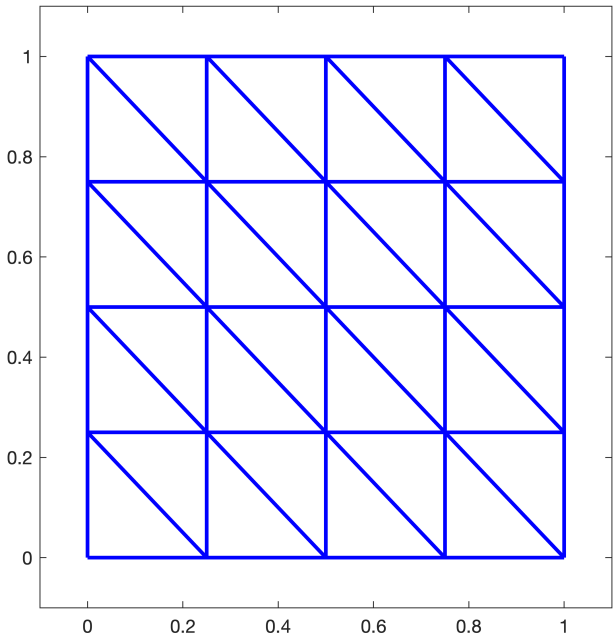


Fig. 3.2. Regular diagonal mesh.

3.3.7 Definition (Finite Element space)**.** Let $\mathbb{P}^q(K)$ denote the space of polynomials of total degree q over the triangle $K \in \mathcal{T}$. We define the FE subspace of order q associated with the triangulation $\mathcal{T}$ of Ω defined in Defn. 3.3.6 as

$$\mathbb{V}_q^h := \left\{ w \in \mathrm{C}^0\big(\overline{\Omega}\big) : \overline{\Omega} \to \mathbb{R} : \forall K \in \mathcal{T},\, w|_K \in \mathbb{P}^q(K) \right\}. \tag{3.16}$$

3.3.8 Definition. (Finite Element discretisation) Let $\mathbb{V}_1^h(\Omega)$ be the FE subspace of $\mathrm{H}_0^1(\Omega)$ as specified in Defn. 3.3.7. Then, the finite element solution to (3.11) is the $u_h \in \mathbb{V}_1^h(\Omega)$ such that

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in \mathbb{V}_1^h(\Omega)\,. \tag{3.17}$$

3.3.9 Lemma. *(Galerkin orthogonality [EG21, Lemma 26.12]) Let $u \in \mathrm{H}_0^1(\Omega)$ and $u_h \in \mathbb{V}_1^h(\Omega)$ denote the solutions to (3.11) and (3.17) respectively. Then,*

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in \mathbb{V}_1^h(\Omega)\,. \tag{3.18}$$

3.3.10 Lemma. *(Céa's Lemma) The finite element approximation, u_h, of u, the weak solution of (3.2), is the best approximation to u amongst all $v_h \in \mathbb{V}^h(\Omega)$ in the $\mathrm{H}_0^1(\Omega)$ norm. That is,*

$$\|u - u_h\|_{\mathrm{H}_0^1(\Omega)} \leq \|u - v_h\|_{\mathrm{H}_0^1(\Omega)} \quad \forall v_h \in \mathbb{V}^h(\Omega)\,. \tag{3.19}$$

Since we will be using a bound pertaining to a finite element approximation of the model problem we will be examining in this chapter, we will introduce some relevant notation and results in this section.

3.3.11 Remark. (Weak solutions to (3.17)) The FE discretisation,(3.17) can be shown by the Lax-Milgram theorem (see [EG13, Lem. 2.2]) to admit unique solutions. Furthermore, by Céa's lemma ([EG13, Lem. 2.28]), the solution $u_h \in \mathbb{V}_1^h(\Omega)$ to (3.17) is the best possible approximation amongst all $v_h \in \mathbb{V}_1^h(\Omega)$ to the solution $u \in \mathrm{H}_0^1(\Omega)$ of (3.11).

3.3.12 Proposition. *(Trace inequality (see [Ver13, §3.3]))For every element $K \in \mathcal{T}$, every edge E of K and every function $v \in \mathrm{H}^1(K)$, the following inequality holds,*

$$\|v\|_{\mathrm{L}^2(E)} \leq c_1 h_K^{1/2} \|v\|_{\mathrm{L}^2(K)} + c_2 h_K^{-1/2} \|\nabla v\|_{\mathrm{L}^2(K)} \tag{3.20}$$

where the constants c_1 and c_2 depend only upon the shape parameter of the triangulation (see [Ver13, Ch.1,3] for more details).

3.3.13 Lemma. *(Interpolation error bound [SM03]) Suppose that $u \in \mathrm{H}^2(\Omega)$. The linear Lagrange interpolant $\mathcal{I}_h u \in \mathbb{V}_1^h(\Omega)$ of u satisfies the following error bounds:*

$$\begin{aligned} \|u - \mathcal{I}_h u\|_{\mathrm{L}^2(\Omega)} &\leq c_1 h^2 |u|_{\mathrm{H}^2(\Omega)} \quad \text{and} \\ |u - \mathcal{I}_h u|_{\mathrm{H}_0^1(\Omega)} &\leq c_2 h |u|_{\mathrm{H}^2(\Omega)} . \end{aligned} \tag{3.21}$$

3.3.14 Lemma. *(A priori error bound for the model elliptic problem) Suppose that $u \in \mathrm{H}^2(\Omega) \cap \mathrm{H}_0^1(\Omega)$ solves and let $C > 0$. Then, the following a priori error bound holds for (3.11) where u_h is the solution to (3.17):*

$$\|u - u_h\|_{\mathrm{H}_0^1(\Omega)} \leq Ch |u|_{\mathrm{H}^2(\Omega)} . \tag{3.22}$$

Proof. The result follows by applying Céa's lemma with $v_h := \mathcal{I}_h u$ in combination with the interpolation error bounds for $\mathcal{I}_h u$ from Lem. 3.3.13. □

In what follows, we are interested in an (explicitly computable) a posteriori bound for $\|\nabla(u - u_h)\|_{\mathrm{L}^2(\Omega)}$.

3.3.15 Definition. (Jumps) We use the notation $[\![\cdot]\!]$ to denote the jump operator. With reference to Fig. 3.2, let $E \in \mathcal{E}$ denote an (interior) edge shared by adjacent patches I and I' and let $\boldsymbol{n}_I$ and $\boldsymbol{n}_{I'}$ denote the outward normals on E for each of I

and I' respectively. Consider a point $\boldsymbol{x} := (x, y) \in E$. We define the jump across E at $\boldsymbol{x}$ of a piecewise continuous function, u and its derivative $\partial \boldsymbol{u}/\partial n$ in the direction of the outward normal at E as follows:

$$[\![u]\!]_E := u\boldsymbol{n}|_I + u\boldsymbol{n}|_{I'}. \tag{3.23}$$

3.3.16 Lemma. *(A posteriori error bound for the model elliptic problem (see [Ver13, Thm. 1.4.6])) Let $u \in \mathrm{H}_0^1(\Omega)$ solve (3.11) and $u_h \in \mathbb{V}_1^h(\Omega) \subset \mathrm{H}_0^1(\Omega)$ solve (3.17) for some $f \in \mathrm{L}^2(\Omega)$. We define the quantity*

$$\eta_K^2 := h_K^2 \left\| f + \Delta u_h \right\|_{\mathrm{L}^2(K)}^2 + \frac{1}{2} \sum_{E \in \partial K} h_E \left\| [\![\nabla u_h]\!]_E \right\|_{\mathrm{L}^2(E)}^2 \quad \textit{for } K \in \mathcal{T}. \tag{3.24}$$

Then, for some $C > 0$ the following error bound holds (see [AO11, Ch. 2]):

$$\left\| \nabla(u - u_h) \right\|_{\mathrm{L}^2(\Omega)} \leq \eta(u_h; f) := C \left(\sum_{K \in \mathcal{T}} \eta_K^2 \right)^{1/2}. \tag{3.25}$$

3.3.17 Corollary. *(One-dimensional a posteriori bound) In the case where the domain is the unit interval discretised by a (uniform) partition $0 = x_0 < \cdots < x_M = 1$, the a posteriori error bound (3.25) simplifies by choosing*

$$\eta_j^2 := h^2 \left\| f \right\|_{\mathrm{L}^2([x_j, x_{j+1}])}^2 + \frac{h}{2} \sum_{j,j+1} \left| \left[\!\!\left[\frac{\partial u_h}{\partial x} \right]\!\!\right]_{x_j} \right|^2 \quad \textit{for } j = 0, \ldots, M-1. \tag{3.26}$$

3.3.18 Finite Differences approximation

In this section we present the FD approximation we will be using for (3.2). We will use the FD solution produced by this numerical scheme in order to obtain the reconstruction.

We will approximate the model problem, (3.2), using the well-known central FD scheme. Firstly, we denote by $U_{i,j}$ the numerical solution obtained by the chosen FD discretisation at the grid-point (x_i, y_j):

$$\begin{array}{rccc} U: & \mathbb{R}^2 & \to & \mathbb{R} \\ & \overline{\Omega}_h & \mapsto & U_{i,j} \end{array} \tag{3.27}$$

and by Δ_h we denote the difference operator we use to approximate the Laplacian operator (in 2D):

$$\begin{array}{rccc} \Delta_h: & \mathbb{R} & \to & \mathbb{R} \\ & U_{i,j} & \mapsto & \Delta_h U_{i,j} \end{array} . \tag{3.28}$$

We now define our FD approximation of (3.2).

3.3.19 Definition. (FD approximation to (3.2)) We seek a function $U : \overline{\Omega}_h \to \mathbb{R}$ which satisfies

$$\begin{aligned} -\Delta_h U_{i,j} &= f_{i,j} \quad \text{in} \quad \Omega_h \\ U_{i,j} &= 0 \quad \text{on} \quad \partial\Omega_h, \end{aligned} \tag{3.29}$$

where, for $1 \leq i, j \leq M-1$, the discrete operator Δ_h is defined as:

$$\Delta_h U_{i,j} := \frac{U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4U_{i,j}}{h^2}. \tag{3.30}$$

3.3.20 Remark (One dimensional model problem)**.** In one dimension, the model problem reduces to seeking a function $U : \overline{\Omega}_h \to \mathbb{R}$ which satisfies

$$\begin{aligned} -\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} &= f_j \quad \text{for} \quad 1 \leq j \leq M-1, \\ U_j &= 0 \quad \text{for} \quad j = 0, M. \end{aligned} \tag{3.31}$$

The FD approximation (3.29) can be written as a system of $(M-1)^2$ equations which can be solved with an appropriate iterative/direct method. We will now provide some classical results to demonstrate that the discrete problem, (3.29), has a unique solution. In addition we will provide an error bound between $U_{i,j}$ and the exact solution, u, of (3.2). We briefly go through some relevant stability, consistency and convergence results pertaining to the elliptic model problem.

In order to obtain the error result we will utilise the truncation error for the scheme (3.29).

3.3.21 Definition. (Truncation Error) We obtain the truncation error for the FD approximation to Δu by applying the discrete operator to Δ_h to the exact solution u of (3.2) and comparing the two quantities. Applying the discrete operator Δ_h, defined in (3.30) and noting that U agrees with u on $\partial\Omega_h$, results in

$$\Delta_h u(x_i, y_j) := \frac{u(x_{i+1}, y_j) + u(x_{i-1}, y_j) + u(x_i, y_{j+1}) + u(x_i, y_{j-1}) - 4u(x_i, y_j)}{h^2}. \tag{3.32}$$

Let $u \in \mathrm{C}^4\big(\overline{\Omega}\big)$ and define $M_4 := \max\left\{ ||\partial^4 u/\partial x^4||_{\mathrm{L}^\infty\left(\overline{\Omega}\right)}, ||\partial^4 u/\partial y^4||_{\mathrm{L}^\infty\left(\overline{\Omega}\right)} \right\}$. Then, it holds that

$$||\Delta_h u - \Delta u||_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)} \leq \frac{h^2}{6} M_4. \tag{3.33}$$

In order to prove the uniqueness result we will use a discrete maximum principle.

3.3.22 Theorem (Discrete Maximum principle [MM05, Ch 6: Lemma 6.1])**.** *Let* $V : \overline{\Omega}_h \to \mathbb{R}$ *denote a mesh function on* $\overline{\Omega}_h$ *which satisfies*

$$\Delta_h V \geq 0 \quad on \quad \Omega_h. \tag{3.34}$$

Then, $\max_{\Omega_h} V \leq \max_{\partial\Omega_h} V$ *with equality iff* V *is constant.*

3.3.23 Remark. (Discrete Minimum principle)The analogous discrete minimum principle holds as folows: we let $V : \overline{\Omega}_h \to \mathbb{R}$ which satisfies

$$\Delta_h V \leq 0 \quad \text{on} \quad \Omega_h. \tag{3.35}$$

Then $\min_{\Omega_h} V \geq \min_{\partial\Omega_h} V$.

3.3.24 Theorem. *The FD approximation, defined in (3.29), is uniquely solvable.*

Proof. Suppose that there exist two solutions $\tilde{U}_1$ and $\tilde{U}_2$ for the FD aproximation (3.29). Then, their difference satisfies

$$\begin{aligned} -\Delta_h\big(\tilde{U}_1 - \tilde{U}_2\big)_{i,j} &= 0 \quad \text{in} \quad \Omega_h \\ \big(\tilde{U}_1 - \tilde{U}_2\big)_{i,j} &= 0 \quad \text{on} \quad \partial\Omega_h. \end{aligned} \tag{3.36}$$

By the discrete maximum principle (see Thm. 3.3.22) and its discrete minimum counterpart (see Rem. 3.3.23) it holds that

$$0 = \min_{\partial\Omega_h}\big(\tilde{U}_1 - \tilde{U}_2\big)_{i,j} \leq \min_{\Omega_h}\big(\tilde{U}_1 - \tilde{U}_2\big)_{i,j} \leq \max_{\Omega_h}\big(\tilde{U}_1 - \tilde{U}_2\big)_{i,j} \leq \max_{\partial\Omega_h}\big(\tilde{U}_1 - \tilde{U}_2\big)_{i,j} = 0 \tag{3.37}$$

Hence $\tilde{U}_i \equiv \tilde{U}_2$. □

3.3.25 Theorem. *The solution* U *to (3.29) satisfies*

$$||U||_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)} \leq \frac{1}{8}\,||f||_{\mathrm{L}^\infty(\Omega_h)}\,. \tag{3.38}$$

Proof. The proof is through an application of the maximum principle. We introduce the function $\phi := \left[(x-1/2)^2 + (y-1/2)^2\right]/4$. This function satisfies $\Delta_h\phi = 1$ on Ω_h and $0 \leq \phi \leq 1/8$ on $\overline{\Omega}_h$. We let $M := \|\phi\|_{\mathrm{L}^\infty(\Omega_h)}$. Then

$$\Delta_h(U + M\phi)_{i,j} = \Delta_h U_{i,j} + M \geq 0. \tag{3.39}$$

Then, through an application of the maximum principle we obtain

$$\max_{\Omega_h} U_{i,j} \leq \max_{\Omega_h}(U_{i,j} + M\phi_{i,j}) \leq \max_{\partial\Omega_h}(U_{i,j} + M\phi_{i,j}) \leq \frac{1}{8}M. \tag{3.40}$$

By using the same argument on $-U_{i,j}$ in combination with the minimum principle counterpart we obtain the required result. □

Now we can prove a global error estimate for the problem.

3.3.26 Theorem. *(from [MM05, §6.2]) Let $u \in \mathrm{C}^4\big(\overline{\Omega}_h\big)$ denote the solution of (3.2) and let U denote the solution of the discrete approximation (3.29) to (3.2). Let $T_{i,j}$ denote the truncation error and let this be bounded from above by T for all $0 \leq i, j \leq M$ by defining T, as follows*

$$T := \frac{h^2}{12}(M_{xxxx} + M_{yyyy}) \tag{3.41}$$

where

$$M_{xxxx} := \left|\left|\frac{\partial^4 u}{\partial x^4}\right|\right|_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)} \quad \textit{and} \quad M_{yyyy} := \left|\left|\frac{\partial^4 u}{\partial y^4}\right|\right|_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)}. \tag{3.42}$$

Then, the following bound holds:

$$||u - U||_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)} \leq \frac{h^2}{96}(M_{xxxx} + M_{yyyy})\,. \tag{3.43}$$

Proof. Let $T_{i,j}$ denote the truncation error of for the FD scheme at the grid-points (x_i, y_j), $0 \leq i, j, \leq M$. Then, let T, where $\max_{i,j} |T_{i,j}| \leq T := \frac{h^2}{12}(M_{xxxx} + M_{yyyy})$, denote the maximum truncation error, where

$$M_{xxxx} := \left|\left|\frac{\partial^4 u}{\partial x^4}\right|\right|_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)} \quad \text{and} \quad M_{yyyy} := \left|\left|\frac{\partial^4 u}{\partial y^4}\right|\right|_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)}. \tag{3.44}$$

see also (3.33). Let U denote the vector containing the $(M-1)^2$ entries $U_{i,j}$ (i.e. excluding the boundary values), arranged in lexicographical order. Also, let Δ_h denote the discrete 5-point central difference operator defined in (3.32). We define the global error as

$$e_{i,j} := U_{i,j} - u_{i,j} \tag{3.45}$$

and consider difference relations at the interior mesh points, Ω_h of $\overline{\Omega}_h$, given by

$$\begin{aligned} \Delta_h U_{i,j} + f_{i,j} &= 0, \\ \Delta_h u_{i,j} + f_{i,j} &= T_{i,j}, \end{aligned} \tag{3.46}$$

in order to obtain the global-truncation error relation

$$\Delta_h e_{i,j} = -T_{i,j}. \tag{3.47}$$

We will obtain the bound on $e_{i,j}$ by using the discrete maximum principle (Thm. 3.3.22). In order to obtain a function which satisfies the condition in Thm 3.3.22 we define a comparison function

$$\Phi_{i,j} := \left(x_i - \frac{1}{2}\right)^2 + \left(y_j - \frac{1}{2}\right)^2, \tag{3.48}$$

for which

$$\Delta_h \Phi_{i,j} = 4. \tag{3.49}$$

Now, we define on $\overline{\Omega}_h$

$$\psi_{i,j} := e_{i,j} + \frac{1}{4} T \Phi_{i,j}. \tag{3.50}$$

We apply the discrete operator to ψ to obtain

$$\begin{aligned} \Delta_h \psi_{i,j} &= \Delta_h e_{i,j} + \frac{1}{4} T \Delta_h \Phi_{i,j} \\ &= -T_{i,j} + T \\ &\geq 0 \quad \forall (x_i, y_j) \in \Omega_h. \end{aligned} \tag{3.51}$$

Notice that $\Delta_h \psi_{i,j} \geq 0$ which implies that we can use the discrete maximum principle (see Thm. 3.3.22) to prove that $\psi_{i,j}$ - where $(x_i, y_j) \in \Omega_h$ - cannot be greater than its neighbours. We apply the maximum principle recursively until we arrive at the boundary. This implies that a maximum value of $\psi_{i,j}$ must be attained at a boundary point $(x_i, y_j) \in \partial\Omega_h$.

At the boundary, the exact and approximate solutions u and U agree and therefore $e_{i,j} = 0$ for $(x_i, y_j) \in \partial\Omega_h$. In addition, the comparison function $\Phi_{i,j}$ has a maximum value of $1/2$ which is also attained at the boundary of our unit square domain. Hence, we obtain from (3.50) that

$$\psi_{i,j} \leq \frac{1}{8} T \quad \forall (x_i, y_j)\,. \tag{3.52}$$

Now, noting that $T, \Phi \geq 0$ in Ω_h, we obtain from (3.50) that

$$U_{i,j} - u(x_i, y_k) = e_{i,j} \quad \leq \psi_{i,j} \leq \frac{1}{8} T = \frac{h^2}{96} (M_{xxxx} + M_{yyyy})\,. \tag{3.53}$$

This bound is one-sided. In order to obtain the required result we need to repeat the process, only this time we will define

$$\psi_{i,j} := -e_{i,j} + \frac{1}{4} T \Phi_{i,j}, \tag{3.54}$$

in order to show that

$$-e_{i,j} \leq \frac{1}{8} T \leq \frac{h^2}{96} (M_{xxxx} + M_{yyyy})\,.. \tag{3.55}$$

Then, (3.53) in combination with (3.55) yield the desired result:

$$||u - U||_{\mathrm{L}^\infty(\overline{\Omega}_h)} \leq \frac{h^2}{96} (M_{xxxx} + M_{yyyy})\,. \tag{3.56}$$

□

3.3.27 Corollary. *(Asymptotic error behaviour of (3.29)) If* $u \in \mathrm{C}^2\big(\overline{\Omega}\big)$ *then*

$$\lim_{h\to 0} ||u - U||_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)} = 0 \tag{3.57}$$

and if $u \in \mathrm{C}^4\big(\overline{\Omega}\big)$*, then*

$$||u - U||_{\mathrm{L}^\infty\left(\overline{\Omega}_h\right)} \leq \frac{h^2}{48} M_4, \tag{3.58}$$

where

$$M_4 := \max\left\{\left|\left|\partial^4 u/\partial x^4\right|\right|_{\mathrm{L}^\infty\left(\overline{\Omega}\right)}, \left|\left|\partial^4 u/\partial y^4\right|\right|_{\mathrm{L}^\infty\left(\overline{\Omega}\right)}\right\}. \tag{3.59}$$

3.3.28 Remark. In fact, the FD scheme would converge for less strict requirement than $u \in \mathrm{C}^2(\Omega)$, i.e. even if $u \in \mathrm{H}^1(\Omega)$ only. We do not demonstrate this and neither can we show it readily from FD error analysis. We advise interested readers to consult [JS13, §2] for more details on the convergence analysis of elliptic boundary value problems. We do note however, that the FE formulation does converge for $u \in \mathrm{H}^1(\Omega)$. Therefore, since the two discretisations result in the same discrete operator (up to an appropriate constant) for the specific choices for the grid and triangulation (see Defn. 3.3.6 and Defn. 3.3.2), we would expect the FD formulation to also converge (albeit to a different discrete solution since the right hand sides are different).

3.4 Reconstruction

Having defined both the FE and FD approximations to our model problem, we can proceed with the definition of the reconstruction. We will go over the reconstruction procedure and demonstrate how a reconstruction is used in posteriori error estimation for our problem.

3.4.1 Reconstruction Procedure

In this section we present conditions we use in order to obtain the polynomial reconstruction, $\widehat{U}$. We use both nodal values of the numerical solution, U_j, as well as information from the numerical scheme.

3.4.2 Definition (Space of the reconstruction)**.** Let $\mathbb{Q}^q(I_{ij})$ denote the space of polynomials of degree q in each variable over the quad $I_{ij} := [x_j, x_{j+1}] \times [y_j, y_{j+1}]$.

We define the space of the reconstruction $\widehat{U}$ to be the space of piecewise polynomials of degree q in each variable over quads in Ω

$$\mathbb{U}_q^h(\Omega) := \left\{ w : \Omega \to \mathbb{R} : w|_{I_{i,j}} \in \mathbb{Q}^q(I_{i,j}) \right\}. \tag{3.60}$$

We will give an illustrative example of the reconstruction procedure in 1D in order to motivate the subsequent material.

3.4.3 Definition (Reconstruction in 1D)**.** In one dimension, the reconstruction, $\widehat{U} \in \mathbb{U}_3^h(\Omega)$, of the numerical solution U of (3.29) (see Defn. 3.4.8) is the unique function that satisfies

$$\begin{aligned} \widehat{U}(x_j) &= U_j \text{ for } j = 0, \ldots, M, \\ \frac{\mathrm{d}^2}{\mathrm{d}x^2}\widehat{U}(x_j) &= \frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} \text{ for } j = 1, \ldots, M-1, \\ \frac{\mathrm{d}^2}{\mathrm{d}x^2}\widehat{U}(x_j) &= -f(x_j) \text{ for } x_j \in \partial\Omega. \end{aligned} \tag{3.61}$$

3.4.4 Remark (Polynomial degree of the reconstruction)**.** The conditions in Defn. (3.4.3) lead to a piece-wise polynomial reconstruction of order three i.e. $\widehat{U} \in \mathbb{U}_3^h(\Omega)$. One can choose to obtain a reconstruction of lower polynomial order by using only a subset of the conditions (3.71). For instance, nodal equivalence alone would result in a reconstruction $\widehat{U} \in \mathbb{U}_1^h(\Omega)$. This would effectively be the bilinear-interpolant of the FD solution or, equivalently, the FE solution for continuous linear Lagrange elements.

3.4.5 Remark (Reconstruction Residual)**.** We obtain the reconstruction residual, R by substituting $\widehat{U}$ in (3.2):

$$R(x) := f + \Delta\widehat{U}. \tag{3.62}$$

The reconstruction residual R is the quantity of interest that we use in the computation of a posteriori bounds in this chapter. A desirable property for a $\widehat{U}$ to possess is that it leads to a residual that, in turn, leads to an a posteriori error estimate of optimal order. An optimal estimate is one that converges at the same rate as error for the underlying numerical scheme.

3.4.6 Lemma (Asymptotic convergence rate for the reconstruction residual in 1D)**.** *Let $\{U_j\}_{j=0}^M$ denote the the central difference approximation of u, the solution of*

(3.2) with $f \in \mathrm{C}^2(\overline{\Omega})$. Suppose $\widehat{U} \in \mathbb{U}_3^h(\Omega)$ is the piecewise cubic interpolant of the nodal values of $\{U_j\}_{j=0}^M$, obtained from Defn. 3.4.3 and let $R(x)$ be defined as in (3.62). Then, as $h \to 0$,

$$\|R\|_{\mathrm{L}^2(\Omega)} \leq \tfrac{1}{8} h^2 \max_j \|f''\|^2_{\mathrm{L}^\infty([\![x_j, x_{j+1}]\!])}. \tag{3.63}$$

(see also [BSB+01]).

Proof. We begin by defining $\widehat{U}(x)$ as a

$$\widehat{U}(x) := c_0(x - x_j)^3 + c_1(x - x_j)^2 + c_2(x - x_j) + c_3, \tag{3.64}$$

where the constants c_i, $0 = 1, \ldots, 3$ are defined by the conditions in Defn. 3.4.8. Hence, the quantity

$$\frac{\mathrm{d}^2\widehat{U}}{\mathrm{d}x^2} := \frac{x_{j+1} - x}{h} f_j + \frac{x - x_j}{h} f_{j+1}, \tag{3.65}$$

is, by construction, the piecewise linear Lagrange interpolant of $\{f_j\}_{j=0}^M$. With this result at hand, we proceed as follows

$$\begin{aligned} \|R\|^2_{\mathrm{L}^2(\Omega)} &= \sum_j \|R\|^2_{\mathrm{L}^2([\![x_j x_{j+1}]\!])} \\ &= \sum_j \int_{x_j}^{x_{j+1}} |R(s)|^2 \, \mathrm{d}s. \end{aligned} \tag{3.66}$$

We now use the fact that $R := \widehat{U}'' - f$ is a continuous function on all closed intervals $[x_j, x_{j+1}]$ to obtain

$$\|R\|^2_{\mathrm{L}^2(\Omega)} = \sum_j \int_{x_j}^{x_{j+1}} |R(s)|^2 \, \mathrm{d}s \leq h \sum_j \|R\|^2_{\mathrm{L}^\infty([\![x_j, x_{j+1}]\!])}. \tag{3.67}$$

Finally, since $f \in \mathrm{C}^2([\![x_j, x_{j+1}]\!])$ and since $\widehat{U}''$ is the piecewise linear Lagrange interpolant of f at points $0 = x_0 < x_1 < \cdots < x_M = 1$. Then, the following error bound holds ([SM03, Thm 11.1])

$$\left\| f - \widehat{U}'' \right\|_{\mathrm{L}^\infty([\![0,1]\!])} \leq \tfrac{1}{8} h^2 \left\| f'' \right\|_{\mathrm{L}^\infty([\![0,1]\!])} \tag{3.68}$$

We substitute the result (3.68 in (3.67) to obtain

$$\|R\|^2_{\mathrm{L}^2(\Omega)} \leq h \sum_j \|R\|^2_{\mathrm{L}^\infty([\![x_j, x_{j+1}]\!])} \leq h \sum_j \tfrac{1}{64} h^4 \left\| f'' \right\|^2_{\mathrm{L}^\infty([\![x_j, x_{j+1}]\!])} \tag{3.69}$$

Noting that $h := 1/M$ and that the sum is over M intervals we obtain

$$\|R\|^2_{\mathrm{L}^2(\Omega)} \leq \tfrac{1}{64} h^4 \max_j \|f''\|^2_{\mathrm{L}^\infty([\![x_j, x_{j+1}]\!])}. \tag{3.70}$$

Finally, we take square roots on both sides to conclude the proof. □

3.4.7 Remark. Although we used $f \in \mathrm{C}^2\big(\overline{\Omega}\big)$ in Lemma 3.4.6, in reality we obtain optimally converging a posteriori bounds for less regular f, such as for example $f \in \mathrm{H}^2(\Omega)$. We will demonstrate this practically with a numerical example.

3.4.8 Definition (Reconstruction in 2D)**.** The reconstruction , $\widehat{U} \in \mathbb{U}_3^h(\Omega)$, of the numerical solution U of (3.29) is the unique function that satisfies

$$\begin{aligned}
\widehat{U}(x_i, y_j) &= U_{i,j} \quad \text{for } i,j = 1,\ldots,M \\
\partial_{xx}\widehat{U}(x_i, y_j) &= \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} \quad \text{for } i,j = 2,\ldots,M-1 \\
\partial_{yy}\widehat{U}(x_i, y_j) &= \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} \quad \text{for } i,j = 2,\ldots,M-1 \\
\partial_{xy}\widehat{U}(x_i, y_j) &= \frac{1}{2h}\left(\frac{U_{i+1,j+1} - U_{i-1,j+1}}{2h} - \frac{U_{i+1,j-1} - U_{i-1,j-1}}{2h}\right) \quad \text{for } i,j = 1,\ldots,M-1
\end{aligned} \tag{3.71}$$

3.4.9 Remark (Reconstruction operator on the boundary)**.** In the reconstruction conditions, we note that special attention must be paid to the treatment of the boundary. In the particular (square) domain, for homogeneous Dirichlet boundary conditions, the reader will notice that the central difference approximations to the second and first derivatives cannot be defined at the boundary as the required points do not exist. In order to overcome this we split the boundary in two sets of points, corner and non-corner points and treat them separately.

In order to compute the difference quotients required for Defn. 3.4.8 at boundary sides - where we lack the required nodes - we use the relation (3.29). Specifically, using (3.29), we can compute$(h^{-2})(U_{i,j+1} - 2U_{i,j} + U_{i,j-1})$ for the vertical portion of the boundary and $(h^{-2})(U_{i+1,j} - 2U_{i,j} + U_{i-1,j})$ for the vertical portion of the boundary. For the mixed derivatives, we use one-sided differences at all boundary points. The second derivatives at the corner points are zero.

3.4.10 Remark (Curved boundary)**.** We note that if the boundary is curved then a different approach would be required. In such a case we could consider using a lower order reconstruction that only makes use of nodal values for the boundary patches.

3.4.11 Remark (General Operator)**.** The convergence rate in Lemma 3.4.6 was demonstrated for the model problem (3.2). A similar result can be shown to hold

for a more general problem. In particular, consider the problem

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\,\frac{\mathrm{d}u}{\mathrm{d}x}\right)+q(x)\,u=f(x)\,. \tag{3.72}$$

A central finite difference discretisation for this problem with Dirichlet boundary conditions and $p \in \mathrm{C}^3\big(\overline{\Omega}\big)$, $q \in \mathrm{C}^2\big(\overline{\Omega}\big)$ $f \in \mathrm{C}^2\big(\overline{\Omega}\big)$ would lead to second order convergence for the error. In this case, the reader should note that the conditions for the reconstruction are not as obvious as they are for (3.2).

3.5 A posteriori error analysis

In this section we use the reconstruction to obtain optimal a posteriori error bounds.

3.5.1 Lemma. *Suppose $z \in \mathrm{H}^1_0(\Omega)$ is the weak solution of the following perturbed problem:*

$$\begin{aligned} -\Delta z &= R \quad \textit{in} \quad \Omega \\ z &= 0 \quad \textit{on} \quad \partial\Omega. \end{aligned} \tag{3.73}$$

for some $R \in \mathrm{H}^{-1}(\Omega)$. Then, the following result holds:

$$\|\nabla z\|_{\mathrm{L}^2(\Omega)} = ||R||_{\mathrm{H}^{-1}(\Omega)}\,. \tag{3.74}$$

Proof. To begin, note in the case $R \equiv 0$, $\Delta z = 0$ in Ω, with $z = 0$ on $\partial\Omega$, so $z \equiv 0$ in $\overline{\Omega}$ and the result is trivially true.

Now, for $z \neq 0$, we recall the definition of the $\mathrm{H}^{-1}(\Omega)$-norm of the residual R:

$$||R||_{\mathrm{H}^{-1}(\Omega)} := \sup_{\phi\in\mathrm{H}^1_0(\Omega)} \frac{\langle R\,|\,\phi\rangle_{\mathrm{H}^{-1}(\Omega)\times\mathrm{H}^1_0(\Omega)}}{||\nabla\phi||_{\mathrm{L}^2(\Omega)}}. \tag{3.75}$$

Testing (3.73) with z, we obtain

$$||\nabla z||^2_{\mathrm{L}^2(\Omega)} = \langle \nabla z, \nabla z\rangle_{\mathrm{L}^2(\Omega)\times\mathrm{L}^2(\Omega)} \;= \langle R\,|\,z\rangle_{\mathrm{H}^{-1}(\Omega)\times\mathrm{H}^1_0(\Omega)} \tag{3.76}$$

Hence, since $z \neq 0$, we have that

$$\begin{aligned} ||\nabla z||_{\mathrm{L}^2(\Omega)} &= \frac{\langle R\,|\,z\rangle_{\mathrm{H}^{-1}(\Omega)\times\mathrm{H}^1_0(\Omega)}}{||\nabla v||_{\mathrm{L}^2(\Omega)}} \\ &\leq \sup_{\phi\in\mathrm{H}^1_0(\Omega)} \frac{\langle R\,|\,\phi\rangle_{\mathrm{H}^{-1}(\Omega)\times\mathrm{H}^1_0(\Omega)}}{||\nabla\phi||_{\mathrm{L}^2(\Omega)}} \\ &= ||R||_{\mathrm{H}^{-1}(\Omega)}\,. \end{aligned} \tag{3.77}$$

Furthermore, by a Cauchy–Schwartz inequality

$$\begin{aligned} ||R||_{\mathrm{H}^{-1}(\Omega)} &= \sup_{\phi \in \mathrm{H}^1_0(\Omega)} \frac{\langle R \,|\, \phi \rangle_{\mathrm{H}^{-1}(\Omega)\times \mathrm{H}^1_0(\Omega)}}{||\nabla \phi||_{\mathrm{L}^2(\Omega)}} \\ &= \sup_{\phi \in \mathrm{H}^1_0(\Omega)} \frac{\langle \nabla z, \nabla \phi \rangle_{\mathrm{L}^2(\Omega)\times \mathrm{L}^2(\Omega)}}{||\nabla \phi||_{\mathrm{L}^2(\Omega)}} \\ &\leq \sup_{\phi \in \mathrm{H}^1_0(\Omega)} \frac{||\nabla z||_{\mathrm{L}^2(\Omega)} \, ||\nabla \phi||_{\mathrm{L}^2(\Omega)}}{||\nabla \phi||_{\mathrm{L}^2(\Omega)}} \\ &= ||\nabla z||_{\mathrm{L}^2(\Omega)} \end{aligned} \tag{3.78}$$

Finally, combining (3.77) and (3.78) we obtain the desired result. □

We will now use Lemma 3.5.1 to establish stability and results for the model problem (3.2).

3.5.2 Corollary (Stability for the model elliptic problem)**.** *Let u be the weak solution of (3.2) and let $v \in \mathrm{H}^1_0(\Omega)$ solve the following perturbed problem for a prescribed f and a residual $R \in \mathrm{H}^{-1}(\Omega)$:*

$$\begin{aligned} -\Delta v &= f - R \text{ in } \Omega \\ v &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{3.79}$$

Then, the error between the two functions $e := u - v$ satisfies the bound

$$||\nabla e||_{\mathrm{L}^2(\Omega)} \leq \|R\|_{\mathrm{H}^{-1}(\Omega)} \,. \tag{3.80}$$

Ideally we would like to use (3.80) as an posteriori bound. However, as it involves the H^{-1}-norm, it is not immediately obvious how to do this. In order to overcome this hurdle we examine two alternative strategies. In this spirit we try to find quantities which behave like the required norm of R but which can be computed directly or for which have known or derivable bounds. The first avenue we will explore is when $R \in \mathrm{L}^2(\Omega)$.

3.5.3 Corollary. *Suppose that in addition to the assumptions in Corollary 3.5.2 we have that $R \in \mathrm{L}^2(\Omega)$. Then, the following bound holds:*

$$||\nabla e||_{\mathrm{L}^2(\Omega)} \leq C_p \, ||R||_{\mathrm{L}^2(\Omega)} \,. \tag{3.81}$$

Proof. Since we have $R \in \mathrm{L}^2$, we have

$$\langle R \,|\, \phi \rangle_{\mathrm{H}^{-1}(\Omega)\times \mathrm{H}^1_0(\Omega)} = \int_\Omega R\phi \mathrm{d}x. \tag{3.82}$$

Hence, from the definition of the H^{-1} −norm we have

$$||R||_{\mathrm{H}^{-1}(\Omega)} = \sup_{\phi\in\mathrm{H}_0^1(\Omega)} \frac{\langle R \,|\, \phi\rangle_{\mathrm{H}^{-1}(\Omega)\times\mathrm{H}_0^1(\Omega)}}{||\nabla\phi||_{\mathrm{L}^2(\Omega)}} \leq \sup_{\phi\in\mathrm{H}_0^1(\Omega)} \frac{||R||_{\mathrm{L}^2(\Omega)}\,||\phi||_{\mathrm{L}^2(\Omega)}}{||\nabla\phi||_{\mathrm{L}^2(\Omega)}}. \tag{3.83}$$

We now use the Poincaré inequality on $||\phi||_{\mathrm{L}^2(\Omega)}$:

$$||\phi||_{\mathrm{L}^2(\Omega)} \leq C_p\,||\nabla\phi||_{\mathrm{L}^2(\Omega)}\,, \tag{3.84}$$

where the constant C_p depends upon the domain Ω. Substituting this in (3.83) gives

$$||\nabla e||_{\mathrm{L}^2(\Omega)} \leq ||R||_{\mathrm{H}^{-1}(\Omega)} \leq \sup_{\phi\in\mathrm{H}_0^1(\Omega)} \frac{C_p\,||R||_{\mathrm{L}^2(\Omega)}\,||\nabla\phi||_{\mathrm{L}^2(\Omega)}}{||\nabla\phi||_{\mathrm{L}^2(\Omega)}} = C_p\,||R||_{\mathrm{L}^2(\Omega)}\,. \tag{3.85}$$

□

There are interesting connections to point out between the FD discretisation (3.29) and the FE discretisation (3.17). Firstly, if the two methods are based over the same Cartesian grid, with the FE discretisation consisting of linear, triangular, Lagrange elements over a regular diagonal partition, then the matrix operator on the left hand side of both methods will be the same (up to a constant factor).

This could be used to directly make use of the estimate (3.25) in our FD computations. However, it is not our goal here, instead we want to use it as a benchmark to test our reconstruction based approach.

It also is going to prove essential in computations when f is not as smooth as required for the conditions of Theorem 3.4.6 to hold.

Now we derive an alternative approach for when $R \notin \mathrm{L}^2(\Omega)$. This could happen for example if the problem data, $f \notin \mathrm{L}^2(\Omega)$. This is also the situation for the specific reconstruction since $\widehat{U} \notin \mathrm{C}^1(\Omega)$.

3.5.4 Lemma. *Let $\widehat{U} \in \mathbb{V}_3^h(\Omega) \cap \mathrm{C}^0(\Omega)$ denote the bicubic reconstruction of the FD solution, $\{U_{ij}\}_{i,j=0}^M$ to (3.2), obtained using the FD discretisation (3.29). Let $\mathrm{H}^{-1}(\Omega) \ni R := f + \Delta\widehat{U}$ denote the resulting residual and let $v \in \mathrm{H}_0^1(\Omega)$ be the unique weak solution of*

$$\int_\Omega \nabla v\cdot\nabla\phi \mathrm{d}x = \int_\Omega R\phi \mathrm{d}x \quad \forall\phi\in\mathrm{H}_0^1(\Omega)\,. \tag{3.86}$$

Now, let $v_h \in \mathbb{V}_1^h$ be the finite element approximation of v satisfying

$$\int_\Omega \nabla v_h\cdot\nabla\phi \mathrm{d}x = \int_\Omega R\phi \mathrm{d}x \quad \forall\phi\in\mathbb{V}_1^h(\Omega)\,. \tag{3.87}$$

Then, the following error a posteriori error estimate holds

$$\left|\left|\nabla u - \nabla \widehat{U}\right|\right|_{\mathrm{L}^2(\Omega)} = ||R||_{\mathrm{H}^{-1}(\Omega)} \leq \left(\eta(v_h; R)^2 + ||\nabla v_h||^2_{\mathrm{L}^2(\Omega)}\right)^{1/2}, \tag{3.88}$$

where η is defined in (3.25).

Proof. From Lem. 3.5.1 we obtained the following result for (3.73):

$$||R||_{\mathrm{H}^{-1}(\Omega)} = ||\nabla v||_{\mathrm{L}^2(\Omega)}\,. \tag{3.89}$$

In addition, because of Galerkin orthogonality (i.e. because the error $v - v_h$ is orthogonal to $\mathbb{V}^1_h(\Omega)$), we have that

$$\begin{aligned} ||R||^2_{\mathrm{H}^{-1}(\Omega)} &= ||\nabla v||^2_{\mathrm{L}^2(\Omega)} \\ &= ||\nabla(v - v_h) + \nabla v_h||^2_{\mathrm{L}^2(\Omega)} \\ &= ||\nabla(v - v_h)||^2_{\mathrm{L}^2(\Omega)} + ||\nabla v_h||^2_{\mathrm{L}^2(\Omega)} \end{aligned} \tag{3.90}$$

Finally, from Lemma 3.3.16 we know that

$$||\nabla(v - v_h)||_{\mathrm{L}^2(\Omega)} \leq \eta(v_h; R)\,, \tag{3.91}$$

with η defined in (3.25). Combining these results gives us:

$$\begin{aligned} ||R||^2_{\mathrm{H}^{-1}(\Omega)} &\leq \eta(v_h; R)^2 + \|\nabla v_h\|^2_{\mathrm{L}^2(\Omega)} \\ &= C\left(\sum_{K\in\mathcal{T}} h_K^2 \left|\left|f + \Delta\widehat{U} + \Delta v_h\right|\right|^2_{\mathrm{L}^2(K)} + \frac{1}{2}\sum_{E\in\mathcal{E}} h_E \left|\left|[\![\nabla v_h]\!]\right|\right|^2_{\mathrm{L}^2(E)}\right) \\ &\quad + ||\nabla v_h||^2_{\mathrm{L}^2(\Omega)}\,. \end{aligned} \tag{3.92}$$

The required result is obtained by applying square roots to both sides of the inequality. □

3.5.5 Remark. The reader will note that in order to obtain the result in Lemma (3.5.4), in addition to the FD discretisation and solution of the model problem, we solve an additional FE problem, (3.87), with the residual from the FD solution and post-processing as the right-hand side. Admittedly, this extra step is a practical disadvantage in obtaining the corresponding bound. However, if the finite element approximation space consists of a regular diagonal mesh, the stiffness matrix used in the FD computation can be reused, as well as any preconditioners, or LU factorisations.

In the next lemma, we present a bound which does not require this additional step but which is only optimal under the conditions of Theorem 3.4.6. In general it is of a lower order than the one we just derived.

3.5.6 Remark (Lemma for the bound involving reconstruction jumps)**.** Let $v \in \mathrm{H}_0^1(\Omega)$ solve (3.73) for $R \in \mathrm{H}^{-1}(\Omega)$ given by

$$R := f + \Delta\widehat{U}, \tag{3.93}$$

where $f \in \mathrm{L}^2(\Omega)$. Since $\widehat{U} \in \mathbb{U}_3^h(\Omega) \cap \mathrm{C}^0(\Omega)$ we have that $\Delta\widehat{U} \in \mathrm{H}^{-1}(\Omega)$. Hence the Poincaré argument we used in Theorem 3.5.3 is not immediately applicable. However, since the $\widehat{U}$ lives on the same mesh as the finite element approximation, we can reuse parts of the a posteriori argument. Indeed, it can be shown that the following bound holds:

$$||R||_{\mathrm{H}^{-1}(\Omega)} \leq C_p \left(\sum_{K \in \mathcal{T}} \left\| f + \Delta\widehat{U} \right\|_{\mathrm{L}^2(K)}^2 + \sum_{E \in \mathcal{E}} h_E^{-1/2} \left|\left| \left[\!\left[\nabla\widehat{U} \right]\!\right]_E \right|\right|_{\mathrm{L}^2(E)}^2 \right)^{1/2} \tag{3.94}$$

Proof. We omit the proof as it is not the focus of the work, however, it is similar to the a residual posteriori argument from classical finite element techniques. □

3.6 Numerical Verification

In this section we present numerical experiments carried out to benchmark the behaviour of the a posteriori error bounds (3.88) and (3.94) for the model problem, (3.2) and the classical FE a posteriori bound (3.25). In all tests, we will discretise the problem using a central finite difference approximation given by (3.29). We will use the FD solution, U_{ij}, to obtain a reconstruction , $\widehat{U}$ using Defn. 3.4.8 in 2D and Cor. 3.4.3 in 1D. The reconstruction is used in the process of obtaining both of the a posteriori error bounds as well as to facilitate the alternative error interpretation, which we define as

$$e := u - \widehat{U}. \tag{3.95}$$

For the sake of brevity and clarity we will adopt the following notation when referring to the bounds (3.88) and (3.94):

$$\begin{aligned} B_1 &:= \left(\sum_{K \in \mathcal{T}} h_K^2 \left|\left| f + \Delta\widehat{U} + \Delta v_h \right|\right|_{\mathrm{L}^2(K)}^2 + ||\nabla v_h||_{\mathrm{L}^2(\Omega)}^2 + \frac{1}{2} \sum_{E \in \mathcal{E}} h_E \, ||[\![\nabla v_h]\!]_E||_{\mathrm{L}^2(E)}^2 \right)^{1/2}, \\ B_2 &:= \left(\sum_{K \in \mathcal{T}} \left\| f + \Delta\widehat{U} \right\|_{\mathrm{L}^2(K)}^2 + \sum_{E \in \mathcal{E}} h_E^{-1} \left|\left| \left[\!\left[\nabla\widehat{U} \right]\!\right]_E \right|\right|_{\mathrm{L}^2(E)}^2 \right)^{1/2} \end{aligned} \tag{3.96}$$

In the one-dimensional case our domain of consideration is the unit interval $\Omega := [0, 1]$. We use equidistant spacing for the $M + 1$ grid points: $x_{j+1} - x_j = h$ for $j = 0, \ldots, M - 1$.

We consider test cases of solutions with varying regularities and in each case we obtain the corresponding right hand side, f, using a method of manufactured solutions. In all cases we approximate quantities involving integrals using a Gauss-quadrature. In order to compute polynomials of order three exactly, we use a quadrature rule of at least two points in each direction, which is formally order four.

3.6.1 Two-dimensional tests

In this section we examine in a 2D setting the asymptotic convergence rate for the gradient error (3.95) of our finite difference post-processor 3.4.8, and the a posteriori error bounds (3.88) and (3.94) for the model problem (3.2). The comparison of the convergence characteristics is carried out on a sequence of approximations on uniform grids with discretisation parameter $h = 2^{-m}$, $m = 4, \ldots, 8$ (see Fig. 3.1).

In addition, we carry out a companion set of experiments using an FE discretisation of the same model problem (3.2) using linear Lagrange elements (3.17) on a regular diagonal mesh (see Fig. 3.2). In these tests, we use the well-used, classical FE bound (3.25). These tests will serve as an intuitive benchmark of the relative performance of the reconstruction in the a posteriori bound.

Test 1- 2D

In this test we benchmark the behaviour of the reconstruction for two smooth solutions:

$$u(x, y) = \sin(2\pi x) \sin(2\pi y). \tag{3.97}$$

The results are shown in Figs. 3.3 and 3.4 respectively for the FD discretisation and the FE discretisation we use for comparison.

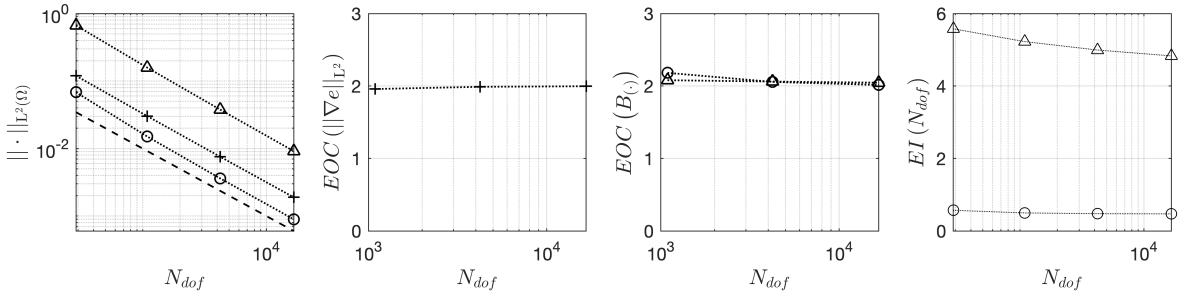


Fig. 3.3. Asymptotic convergence rate for the gradient error (3.95) for the finite difference post-processor 3.4.8 (plus sign), and the a posteriori error bounds (3.88) (circles) and (3.94) (triangles) for the model problem (3.2). The exact solution (3.97) is smooth. Both bounds are optimal, with (3.88) being sharper.

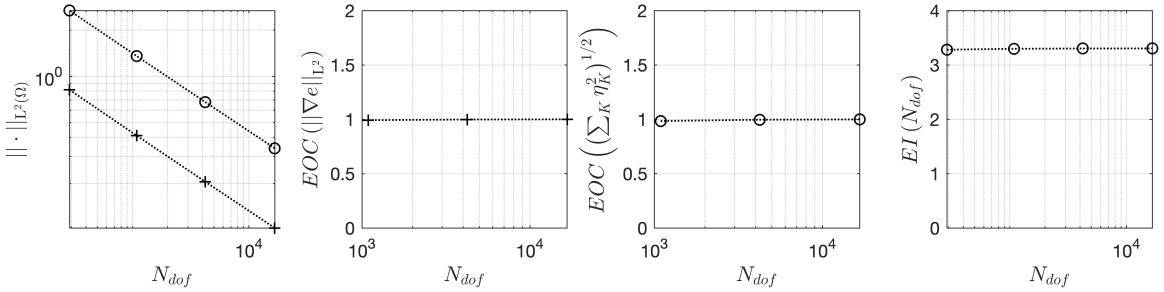


Fig. 3.4. Asymptotic convergence rates for $e := \|\nabla(u - u_h)\|_{\mathrm{L}^2(\Omega)}$ (plus sign) of the model problem (3.2) discretised using linear Lagrange elements (3.17), and of the bound (3.25) (circles). The exact solution (3.97) is smooth. The bound is optimal.

Test 2

In this test we use an $\mathrm{H}^2(\Omega) \setminus \mathrm{H}^3(\Omega)$ exact solution

$$u(x) := \begin{cases} \frac{1}{4}\cos\left(8\pi\left|\boldsymbol{x} - \frac{1}{2}\right|^2\right) + 1 & \text{if } |\boldsymbol{x} - 1/2|^2 \leq \frac{1}{8} \\ 0 & \text{otherwise.} \end{cases} \tag{3.98}$$

where $\boldsymbol{x} = (x, y) \in \mathbb{R}^2$. The test case (3.98) results in a source term given by

$$f(x) := \begin{cases} -8\pi\sin\left(8\pi\left|\boldsymbol{x} - \frac{1}{2}\right|^2\right) - 64\pi^2\cos\left(8\pi\left|\boldsymbol{x} - \frac{1}{2}\right|^2\right)|\boldsymbol{x} - 1/2|^2 & \text{if } |x - 1/2| \leq \frac{1}{8} \\ 0 & \text{otherwise.} \end{cases} \tag{3.99}$$

The results are shown in Fig. 3.5 and 3.6 for the FD and FE approximations respectively. In this test, the residual converges sub-optimally compared to the error.

Notice that, in Fig. 3.5, the residual $\|R\|_{\mathrm{L}^2(\Omega)}$ converges sub-optimally compared to the error, whereas the bound (3.88) converges optimally.

3.6.2 Remark. We note that the order of convergence in the error $\|\nabla e\|_{\mathrm{L}^2(\Omega)}$ in this case is $\mathcal{O}\big(h^{3/2}\big)$. The reason for that is the lower regularity of the solution. However, our focus here is to examine the robustness of the post-processor based bound, (3.88). We note that in this regard the bound behaves optimally, also with regard to reflecting the loss of approximability due to the lower regularity of the solution.

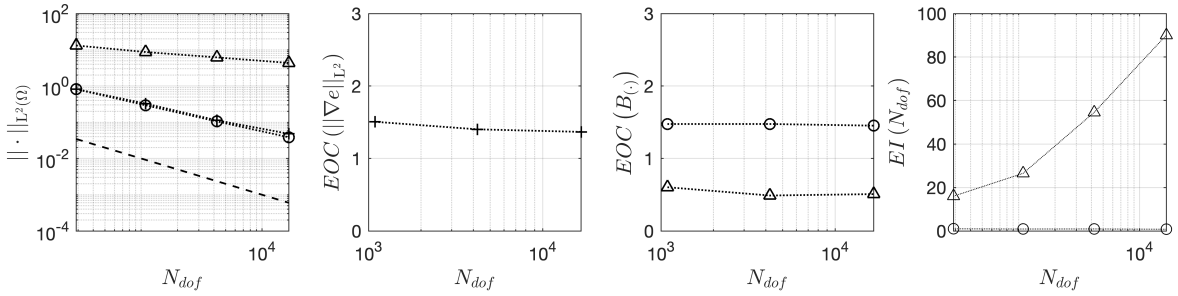


Fig. 3.5. Asymptotic convergence rate for the gradient error (3.95) of the FD postprocessor 3.4.8 (plus sign), and the of the error bounds (3.88) (circles) and (3.94) (triangles) for the model problem (3.2). The exact solution (3.98) is in $\mathrm{H}^2(\Omega)\backslash\mathrm{H}^3(\Omega)$. The bound (3.88) is optimal whereas bound (3.94), is sub-optimal.

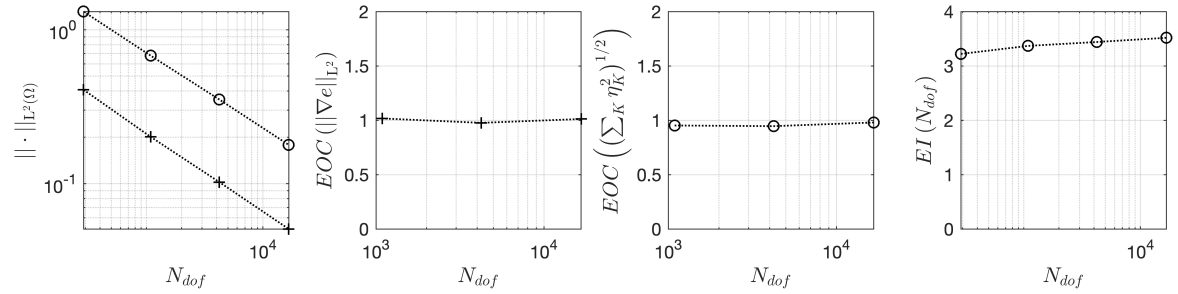


Fig. 3.6. Asymptotic convergence rates for $e := \|\nabla(u - u_h)\|_{\mathrm{L}^2(\Omega)}$ (plus sign) of the FE discretisation of (3.2) using linear Lagrange elements (3.17) and of the bound (3.25) (circles). The exact solution (3.98) is in $\mathrm{H}^2(\Omega)\backslash\mathrm{H}^3(\Omega)$. The bound is optimal.

3.6.3 One-dimensional tests

In this section we examine in a 1D setting the asymptotic convergence rate for the gradient error (3.95) of the finite difference post-processor 3.4.8, and the a posteriori

error bounds (3.88) and (3.94) for the model problem (3.2). The comparison of the convergence characteristics is carried out on a sequence of approximations on uniform grids with discretisation parameter $h = 2^{-m}$, $m = 4, \ldots, 8$ (see Fig. 3.1).

In the same fashion as we did in the 2D tests, we carry out a companion set of experiments using the same problem but approximated with a Finite Element discretisation with linear Lagrange elements (3.17), posed over the same uniform grid as the FD discretisation (i.e. same discrete operator up to a constant). For the a posteriori error estimate we use the classical FE bound (3.25).

Test 3: $\mathrm{C}^1(\Omega) \setminus \mathrm{C}^2(\Omega)$ **solution**

In this test we use an $\mathrm{C}^1(\Omega) \setminus \mathrm{C}^2(\Omega)$ exact solution

$$u(x) := \begin{cases} x^2/2 & \text{if } x \leq 0.25 \\ -x^2/2 + x - 1/16 & \text{if } 0.25 < x \leq 0.75 \\ x^2/2 - x + 1/2 & \text{if } 0.75 < x \leq 1, \end{cases} \tag{3.100}$$

which results in a step-function source term:

$$f(x) := \begin{cases} 1 & \text{if } x \leq 0.25 \\ -1 & \text{if } 0.25 < x \leq 0.75 \\ 1 & \text{if } 0.75 < x \leq 1. \end{cases} \tag{3.101}$$

The results are shown in Fig. 3.7.

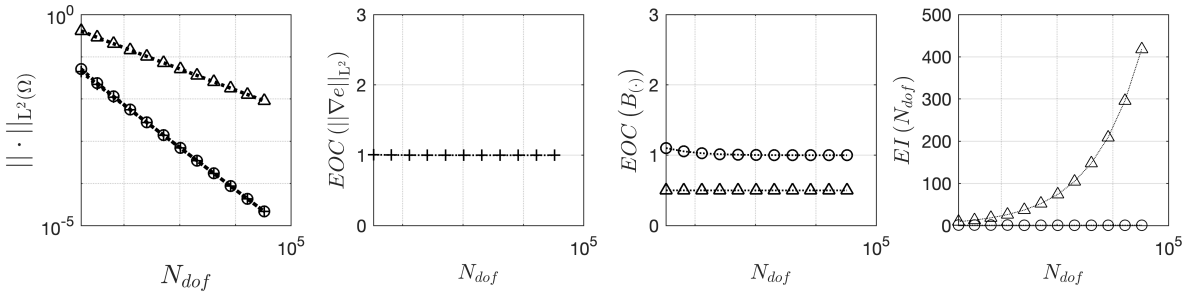


Fig. 3.7. Asymptotic convergence rate for the error (3.95) of our FD postprocessor 3.4.8 (plus sign), and for the bounds (3.88) (circles) and (3.94) (triangles) for the model problem (3.2). The exact solution (3.100) is $\mathrm{C}^1(\Omega) \setminus \mathrm{C}^2(\Omega)$. The bound (3.88) is optimal with $EI \sim 10$ (the constant is not included) while (3.94) is sub-optimal.

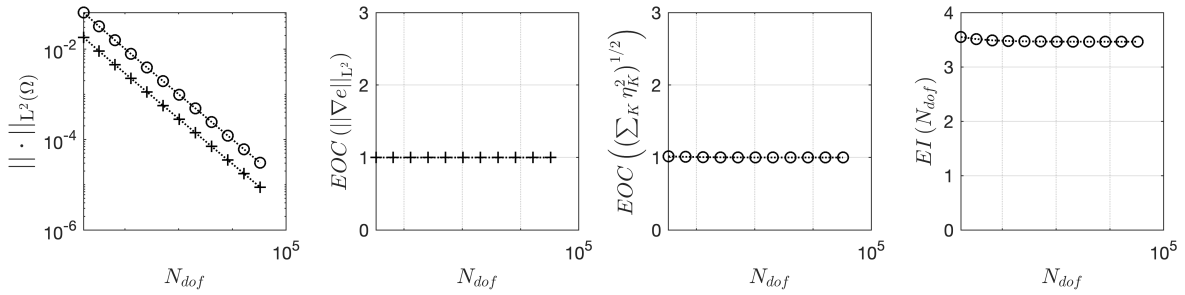


Fig. 3.8. Asymptotic convergence rates for the error $e := \|\nabla(u - u_h)\|_{\mathrm{L}^2(\Omega)}$ (plus sign) of the FE discretisation of the model problem (3.2) using linear Lagrange elements (3.17) and of the FE bound (3.25) (circles). The exact solution (3.100) is in $\mathrm{C}^1(\Omega) \setminus \mathrm{C}^2(\Omega)$. The bound is optimal.

Test 4: $\mathrm{C}^0(\Omega) \setminus \mathrm{C}^1(\Omega)$ **solution**

In this case, we will use a hat function exact solution given by

$$u(x) := \frac{3}{8} - \frac{3}{4}\left|x - \frac{1}{2}\right|. \tag{3.102}$$

One can easily verify that, in this case, the required f we obtain using the method of manufactured solutions is a Dirac delta distribution:

$$\delta\big(x - \tfrac{1}{2}\big) := \begin{cases} \infty & \text{if } x = \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{3.103}$$

As it is difficult to numerically represent (3.103), we work instead with a numerical approximation which converges to (3.103) in some appropriate sense, which we will define shortly. In order to present this result we firstly introduce the necessary notation. We adopt the same notation as [HNS16], as we use their results to construct the numerical approximation for (3.103).

We denote by $\mathcal{D}(\Omega)$ the dual space of $\mathrm{C}_0^\infty(\Omega)$. The Dirac delta is an element of $\mathcal{D}(\Omega)$ and it is defined as

$$\delta(\phi) := \phi(0) \quad \text{for} \quad \phi \in \mathrm{C}_0^\infty(\Omega)\,. \tag{3.104}$$

In particular, $\delta \in \mathrm{H}^{-s}(\Omega)$ for $s > 0$. Following the approach of [HNS16], we approximate the Dirac delta using a sequence of distributions $\widetilde{\delta}_h \in \mathcal{D}(\Omega)$. The $\widetilde{\delta}_h$ are parametrised by their support, which we identify with $h > 0$, and possess the

property that $\widetilde{\delta}_h \to \delta$ in a suitable sense as $h \to 0$. The $\widetilde{\delta}_h$ are constructed using suitably regular functions $\delta_h \in \mathrm{H}_0^s(\Omega)$. Then, $\widetilde{\delta}_h \in \mathrm{H}^{-s}(\Omega)$, $s > 0$ is defined as

$$\widetilde{\delta}_h(\phi) := \langle \delta_h, \phi \rangle_{\mathrm{L}^2(\Omega)\times\mathrm{L}^2(\Omega)} \quad \forall \phi \in \mathrm{H}_0^s(\Omega)\,. \tag{3.105}$$

We omit details about the construction of the δ_h and requirements to guarantee required convergence rates and instead direct interested readers to [HNS16, §2,3].

3.6.4 Definition (Weak-* convergence)**.** The sequence, $\tilde{\delta}_h$ of distributions converges in distribution to δ as $h \to 0$ iff

$$\tilde{\delta}_H(\phi) \to \delta(\phi) = \phi(0) \quad \forall \phi \in \mathrm{C}_0^\infty(\Omega) \quad \text{as} \quad h \to 0. \tag{3.106}$$

In the numerical example we examine in this section, we use

$$\delta_h(x) := \begin{cases} \frac{1}{2h}, & \left|x - \frac{1}{2}\right| \le h \\ 0 & \text{otherwise.} \end{cases} \tag{3.107}$$

Having identified an appropriate approximation $\widetilde{\delta}_h$ to δ, we solve (3.2) with $f = \delta_h$, which is given by by (3.107). The results are shown in Fig. (3.9).

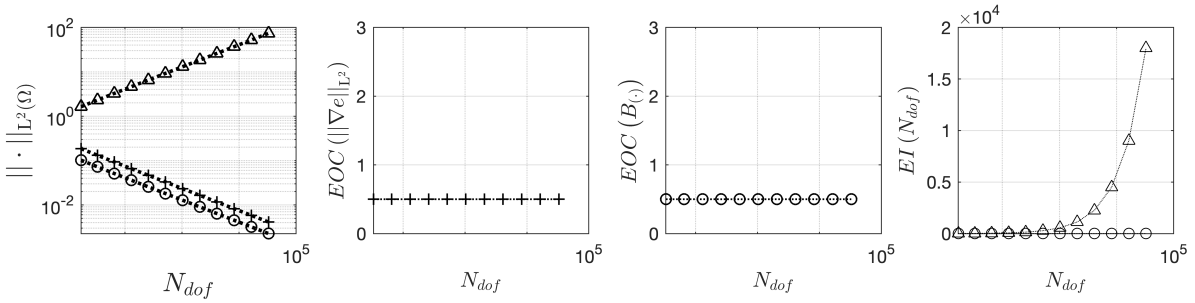


Fig. 3.9. Asymptotic convergence rates for the gradient error $e := \|\nabla(u - u_h)\|_{\mathrm{L}^2(\Omega)}$ (plus sign) of the FE discretisation of the model problem (3.2) using linear Lagrange elements (3.17) and of the bound (3.25) (circles). The exact solution (3.102) is in $\mathrm{C}^0(\Omega) \setminus \mathrm{C}^1(\Omega)$. The bound (3.88) is optimal with $EI \sim 5$ (constant not included), while (3.94) diverges.

3.7 Conclusion

In this chapter we introduced a methodology for obtaining reconstructions for the FD discretisation of the elliptic model problem with homogeneous Dirichlet boundary

conditions. We evaluated the performance of a quadratic reconstruction when it is used to obtain an a posteriori error bound in a range of numerical experiments where the underlying solutions have different regularity.

We demonstrated in practice that this framework can be used to obtain optimal a posteriori error estimates (provided the solution possesses sufficient regularity) which reflect the behaviour of the underlying error, measured in the $\mathrm{H}_0^1(\Omega)$-norm. We compared the performance of a bound obtained using a post-processor of an FD solution with well-used bounds based on FE solutions of the same problems. We found that bounds constructed using our framework compare favourably with existing FE bounds.

In particular, for smooth solutions, we are able to achieve order 2 convergence in the gradient norm for our post-processed solution. This is one order higher than the analogous FE computation at very little additional computational expense.

In addition, we proved results which allowed us to investigate the behaviour of the residual in the $\mathrm{H}^{-1}(\Omega)$ $-$norm - a quantity which cannot be practically computed. However, in the process of doing so we had to solve an additional FE problem in order to compute required quantities. This resulted in an optimal bound with at the expense of small extra effort. A bound which does not require the extra FE solution was also computed but it was sub-optimal for solutions less regular than $\mathrm{C}^4(\Omega)$.

Chapter 4

Automated error control for the transport equation

Abstract

In this chapter, we present an a posteriori error bound for a class of fundamental finite difference methods for the transport equation. This is based on a simple to evaluate reconstruction operator that yields optimal a posteriori upper bounds for the fully discrete solution. We are also able to show global lower bounds for this bound. We validate the analysis with some numerical experimentation and examine the ability to detect parasites, a numerical artefact produced when a wave passes over a non-uniformity in the mesh.

4.1 Introduction

In this chapter we shift our focus to hyperbolic problems. We will use this chapter to motivate the subsequent work on hyperbolic problems using the linear transport equation in one dimension as an illustrative example. In this regard, we will use an easy-to-evaluate reconstruction operator for the model problem.

We will examine the performance of the a posteriori estimate constructed in this way using a number of numerical experiments. In addition, we will test the ability of the bound to detect numerical parasites - numerical artefacts which propagate at the wrong speed/direction, polluting the computation. We will validate our analysis with some simple numerical experiments.

4.1.1 Motivation

Our motivation in this chapter is to derive and implement an a posteriori error bound for a class of fundamental FD methods using reconstructions. In particular, the a posteriori error bound is derived using the stability framework of the PDE, thereby decoupling from the chosen discretisation method. Then, the bound is obtained in practice by using a reconstruction framework in the hyperbolic setting.

We will demonstrate this by using fundamental FD methods for the linear transport equation, obtaining a bound for the fully discrete solution.

In this regard, the novelty in this chapter - which will be made concrete in the following chapters - are the foundational ideas for a framework for constructing a posteriori error estimates based on FD solutions of hyperbolic conservation laws. The underlying motivation is that, while FD schemes for conservation laws are widely used in practice, a posteriori error bounds for FD schemes receive less attention when compared to the FV and FE counterparts. In addition, we will demonstrate the ability of such an estimate to detect parasitic waves. This may be considered as the start of an effort to come up with a strategy for removing these from computations.

4.1.2 Chapter contribution

In this chapter we demonstrate in practice the reconstruction procedure from the previous chapter in the process of obtaining an a posteriori bound for linear advection. We will use the procedure to construct optimal bounds for this problem and we will prove that the bound is optimal for a simple numerical scheme in one dimension. We will use the simple test cases to demonstrate the ability of a bound constructed in this way to reliably indicate areas requiring refinement. We conclude the chapter with an implementation of mesh adaptivity in one spatial dimension, using the bound as an indicator.

The rest of this chapter is structured as follows. In §4.3 we present our linear advection model problem an a posteriori error bound. In §4.4 we demonstrate the construction of bounds for two well-known used numerical schemes for linear advection. We prove that the bound is optimal for the Forward-Time Backward-Space (FTBS) scheme. In §4.5 we demonstrate the use of the bound with simple

numerical examples for these two schemes. In §4.6 we provide the details of the adaptive strategy for the linear advection problem and we demonstrate the use of the a posteriori bound in the implementation of adaptivity. We conclude the chapter in §4.7

4.2 Preliminaries and problem setup

4.2.1 Definition. (Bochner spaces) We use the following the notation for time-dependent Sobolev (Bochner spaces):

$$\mathrm{L}^{\infty}\big(0,T;\mathrm{H}^{k}(\Omega)\big) := \left\{ u : [0,T] \to \mathrm{H}^{k}(\Omega) : \sup_{t\in[0,T]} \|u(t)\|_{\mathrm{H}^{k}(\Omega)} < \infty \right\}. \tag{4.1}$$

4.2.2 Definition. (Periodic boundary conditions) Let Ω denote the one dimensional unit interval, i.e. $\Omega := [x_0, x_M]$ with $x_0 < x_M$. Then, periodic boundary conditions are implemented by identifying the points x_0 and x_M as being the same point.

4.3 Hyperbolic model problem

We consider the linear transport equation

$$\begin{aligned} u_t + u_x &= 0 \quad \text{in } \Omega \times (0,T] \\ u(x,0) &= u_0(x) \quad \text{in } \Omega \times \{0\} \end{aligned} \tag{4.2}$$

with periodic boundary conditions (see Defn. 4.2.2).

4.3.1 Remark (The solution to (4.2))**.** Let $u_0 \in \mathrm{C}^1(\mathbb{R}\times\mathbb{R}^+)$ in (4.2). The 1D transport equation, (4.2), admits a solution

$$u(x,t) = u_0(x-t)\,. \tag{4.3}$$

As [LeV92] notes, the solution $u(x,t)$ to (4.2) at a point (x,t) depends only on the initial data $u_0(x)$ at a single point, x_0, which is the point through which the characteristic line through (x,t) passes.

For the particular problem, (4.2), characteristic lines do not cross as they have a constant gradient. Hence, discontinuities in the intial data will be transferred along characteristics and will only affect the value of the solution along the specific

characteristic curve that passes through them. This allows for non-smooth solutions to these problem, for which the PDE does not make sense. The reason is that discontinuous solutions will not be differentiable at the point of discontinuity.

In order for such solutions to make sense we should expand our definition of solution to that of weak solutions. This will be introduced in later chapters. For now, for expositions sake, we will make the underlying assumption that the solution is classical. This is true under appropriate regularity conditions on the initial condition.

4.3.2 Lemma (Stability and error control for the linear advection equation)**.** *Let u be a classical solution of the initial boundary value problem*

$$\begin{aligned} u_t + u_x &= 0 \quad \text{in } \Omega \times (0,T] \\ u(x,0) &= u_0(x) \quad \text{in } \Omega \times \{0\} \end{aligned} \tag{4.4}$$

with periodic boundary conditions. Suppose also that v is a classical solution of a perturbed balance law, specifically for some $R \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))$

$$\begin{aligned} v_t + v_x &= -R \quad \text{in } \Omega \times (0,T] \\ v(x,0) &= v_0(x) \quad \text{in } \Omega \times \{0\}, \end{aligned} \tag{4.5}$$

also with periodic boundary conditions. Then, the error between the two functions, $e := u - v$, satisfies the following bound for all $t \in [0,T]$:

$$\|e(t)\|^2_{\mathrm{L}^2(\Omega)} \leq \omega(t)\left[\|e(0)\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \|\delta(s)R(s)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s\right], \tag{4.6}$$

where

$$\omega(t) = \begin{cases} \exp(t) & \text{for } t \leq 1 \\ t\exp(1) & \text{for } t \geq 1. \end{cases} \tag{4.7}$$

and

$$\delta(s) = \begin{cases} 1 & \text{for } s \leq 1 \\ \sqrt{s} & \text{for } s \geq 1. \end{cases} \tag{4.8}$$

Proof. We defer the proof of this result until Chapter 6, where we prove it for linear systems. The scalar linear case that we examine in this chapter follows easily (in particular, see Lem. 6.4.1 and Cor. 6.4.3). □

4.4 Fundamental numerical methods and a posteriori bounds

We partition the domain Ω (see Defn. 4.2.2) by choosing $0 = x_0 < \cdots < x_M = 1$. We denote the spatial mesh size $h_j := x_{j+1} - x_j$ for $0 \leq j \leq M-1$ and we use I_j to denote the sub-interval $[x_j, x_{j+1}]$ of Ω.

In the temporal variable, we partition $[0, T]$ into sub-intervals with endpoints given by $0 = t^0 < \cdots < t^N = T$. The time-step is defined by $\tau^n := t^{n+1} - t^n$. We denote by U_j^n as an approximation to $u(x_j, t^n)$, the solution of (4.2) given by either one of two different schemes. The two schemes we consider are both posed over a uniform temporal and spatial partition, that is $\tau^n \equiv \tau$ for all n and $h_j \equiv h$ for all j and are both classical schemes in the study of conservation laws. We consider an upwinding forward-time backward-space (FTBS)

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{\tau}{h}\big(U_j^n - U_{j-1}^n\big) \text{ for } n = 0, \ldots, N-1 \text{ and } j = 0, \ldots, M-1 \\ U_j^0 &= u_0(x_j) \text{ for } j = 0, \ldots, M \end{aligned} \tag{4.9}$$

and a Crank-Nicolson central-space (CNCS)

$$\begin{aligned} U_j^{n+1} = U_j^n - \frac{1}{2}\Big(\frac{\tau}{2h}\big(U_{j+1}^n - U_{j-1}^n\big) + \frac{\tau}{2h} + \big(U_{j+1}^{n+1} - U_{j-1}^{n+1}\big)\Big) \text{ for } \quad j = 0, \ldots, M-1. \\ U_j^0 = u_0(x_j) \text{ for } j = 0, \ldots, M, \end{aligned} \tag{4.10}$$

for $n = 0, \ldots, N-1$

4.4.1 Definition (Truncation error for the FTBS scheme)**.** We define the Truncation error for the FTBS scheme at (t, x), $T(t, x)$, as

$$\begin{aligned} T(t,x) &:= \frac{u(t+\tau, x) - u(x,t)}{\tau} + \frac{u(t,x) - u(t, x-h)}{h} \\ &= \frac{1}{2}(\tau u_{tt}(x, \eta) + h u_{xx}(\xi, t))\,, \end{aligned} \tag{4.11}$$

for $\xi \in (x-h, x)$, $\eta \in (t, t+\tau)$.

4.4.2 Definition (Truncation error for the BTFS scheme)**.** We define the Truncation error for the BTFS scheme at (t, x), $T(t, x)$, as

$$\begin{aligned} T(t,x) &:= \frac{u(t+\tau, x) - u(x,t)}{\tau} + \frac{u(t+\tau, x+h) - u(t+\tau, x)}{h} \\ &= (\tau - h)\, u_{tx}, \end{aligned} \tag{4.12}$$

for $\xi \in (x+h, x)$, $\eta \in (t, t+\tau)$.

4.4.3 Remark. (Consistency and truncation error) Formally, both the FTBS and CNCS scheme are consistent and have truncation errors $\mathrm{O}(\tau + h)$ and $\mathrm{O}(\tau^2 + h^2)$ respectively.

4.4.4 Remark. (Stability) The FTBS scheme is stable, conditional upon the ratio τ/h being less or equal to one. The CNCS scheme is unconditionally stable.

4.4.5 Remark. (Convergence) For the model problem (4.2), the FTBS scheme is consistent and conditionally stable and the CNCS scheme is consistent and unconditionally stable. Hence, according to the Lax equivalence theorem (see Thm. A.0.2) both schemes will converge as long as the parameters τ and h are appropriately chosen.

In order to make use of the abstract bounds given in §4.3 we must have an interpretation of the numerical approximation $\{U_j^n\}_j^n$, which is only defined as point values over the space-time domain. The most intuitive post-processing is to apply a bilinear Lagrange interpolant in space-time.

4.4.6 Corollary (An a posteriori bound)**.** *Let $\widehat{U}$ be a continuous reconstruction of the finite difference approximation $\{U_j^n\}_j^n$. Then, in view of Lemma 4.3.2, we have the a posteriori bound:*

$$\left\|(u-\widehat{U})(t)\right\|_{\mathrm{L}^2(\Omega)}^2 \leq \omega(t)\left[\left\|(u-\widehat{U})(0)\right\|_{\mathrm{L}^2(\Omega)}^2 + \int_0^t \|\delta(s)R(s)\|_{\mathrm{L}^2(\Omega)}^2 \,\mathrm{d}s\right] =: \omega(t)\,\mathscr{E}(t)^2, \tag{4.13}$$

where

$$R := -\widehat{U}_t - \widehat{U}_x \tag{4.14}$$

is the discrete residual of the reconstruction.

Note that given the numerical solution, the right hand side of (4.13) is fully computable. It can even be shown to be fully robust when u_0 is a sufficiently smooth initial condition as we will indicate in the following result.

4.4.7 Lemma (Asymptotic convergence rate for the reconstruction residual)**.** *Let $\{U_j^n\}_j^n$ be the FTBS approximation of u, the solution of (4.2) with $u_0 \in \mathrm{C}^2(\Omega)$. Suppose $\widehat{U}$ is the piecewise bilinear interpolant of the nodal values of $\{U_j^n\}_j^n$ and let $\omega(t)\,\mathscr{E}(t)^2$ be defined in (4.13), then*

$$\omega(t)\,\mathscr{E}(t)^2 \leq C\big(\tau^2 + h^2\big)\,\|u_0\|_{\mathrm{C}^2(\Omega)}^2\,. \tag{4.15}$$

Proof. We begin by defining, $\widehat{U}^t$ nodally as

$$\widehat{U}^t_j(t) := U^n_j + \frac{U^{n+1}_j - U^n_j}{\tau}(t - t^n)\,, \quad \text{for} \quad t \in \left[t^n, t^{n+1}\right] \text{ and } j \in [0, M-1], \tag{4.16}$$

which represents the interpolant in the temporal direction. This allows us to write $\widehat{U}$ as

$$\widehat{U}(x,t) := \widehat{U}^t_j(t) + \frac{\widehat{U}^t_{j+1}(t) - \widehat{U}^t_j(t)}{h}(x - x_j) \quad \text{for} \quad (x,t) \in [x_j, x_{j+1}] \times \left[t^n, t^{n+1}\right]. \tag{4.17}$$

Since $\widehat{U}$ is bilinear on a space-time slab we can compute R explicitly as

$$-R = \partial_t \widehat{U} + \partial_x \widehat{U} = \partial_t \widehat{U}^t_j + \frac{\partial_t \widehat{U}^t_{j+1} - \partial_t \widehat{U}^t_j}{h}(x - x_j) + \frac{\widehat{U}^t_{j+1} - \widehat{U}^t_j}{h}. \tag{4.18}$$

Now consider the residual component of the estimator. Since δR is linear over each spatial interval, we see that

$$\begin{aligned} \|\delta(s)R(s)\|^2_{\mathrm{L}^2(\Omega)} &= \sum_{j=0}^{M-1} \int_{x_j}^{x_{j+1}} |\delta(s)R(s,x)|^2 \,\mathrm{d}x \\ &= \sum_{j=0}^{M-1} h \left|\delta(s)R(s, x_{j+1/2})\right|^2, \end{aligned} \tag{4.19}$$

Now

$$\int_0^T \|\delta(s)R(s)\|^2_{\mathrm{L}^2(\Omega)} \,\mathrm{d}s = \sum_{n=0}^{N-1} \int_{t^n}^{t^{n+1}} \sum_{j=0}^{M-1} h \left|\delta(s)R(s, x_{j+1/2})\right|^2 \mathrm{d}s. \tag{4.20}$$

We recall from Lem. 4.3.2 that

$$\delta(s) = \begin{cases} 1 & \text{for } s \le 1 \\ \sqrt{s} & \text{for } s \ge 1 \end{cases} \tag{4.21}$$

hence we may bound (4.20) as follows:

$$\begin{aligned} \int_0^T \|\delta(s)R(s)\|^2_{\mathrm{L}^2(\Omega)} \,\mathrm{d}s &\le \max(1,T) \sum_{n=0}^{N-1} \int_{t^n}^{t^{n+1}} \sum_{j=0}^{M-1} h \left|R(s, x_{j+1/2})\right|^2 \mathrm{d}s \\ &\le \max(1,T)\, \tau h \sum_{n=0}^{N-1} \sum_{j=0}^{M-1} \left|R(t^{n+1/2}, x_{j+1/2})\right|^2. \end{aligned} \tag{4.22}$$

Using (4.18) we see that

$$-R(t^{n+1/2}, x_{j+1/2}) = \frac{U^{n+1}_j - U^n_j}{\tau} + \frac{U^n_{j+1} - U^n_j}{h} + \left(\frac{1}{2\tau} + \frac{1}{2h}\right)\left(U^{n+1}_{j+1} - U^n_{j+1} - \left(U^{n+1}_j - U^n_j\right)\right), \tag{4.23}$$

which simplifies to

$$
\begin{aligned}
-R(t^{n+1/2}, x_{j+1/2}) &= \frac{1}{2\tau}\big(U_{j+1}^{n+1} - U_{j+1}^{n} + U_{j}^{n+1} - U_{j}^{n}\big) + \frac{1}{2h}\big(U_{j+1}^{n+1} - U_{j}^{n+1} + U_{j+1}^{n} - U_{j}^{n}\big) \\
&= \frac{1}{2}\Bigg(\underbrace{\frac{U_{j+1}^{n+1} - U_{j+1}^{n}}{\tau} + \frac{U_{j+1}^{n} - U_{j}^{n}}{h}}_{A}\Bigg) + \frac{1}{2}\Bigg(\underbrace{\frac{U_{j}^{n+1} - U_{j}^{n}}{\tau} + \frac{U_{j+1}^{n+1} - U_{j}^{n+1}}{h}}_{B}\Bigg).
\end{aligned}
\tag{4.24}
$$

We note that term A in (4.24) is the FTBS discretisation (4.9) evaluated at the $(j+1)$-th node and it is therefore equal to zero. Hence, (4.24) simplifies to

$$
-R(t^{n+1/2}, x_{j+1/2}) = \frac{1}{2}\left(\frac{U_{j}^{n+1} - U_{j}^{n}}{\tau} + \frac{U_{j+1}^{n+1} - U_{j}^{n+1}}{h}\right). \tag{4.25}
$$

Notice that this corresponds to a backward time, forward space discretisation. In a sense this is the "opposite" discretisation of the method we study. This allows us to relate the residual, (4.25), to the truncation error of the BTFS scheme (see Defn. 4.4.2) and to subsequently use this in our calculation.

Therefore, to assess the contribution of the remaining terms, we relate them to quantities for which we have known bounds, such as the truncation error. In this spirit we utilise the relation between the BTFS scheme and its truncation error, (4.12) to (4.25):

$$
\frac{U_{j}^{n+1} - U_{j}^{n}}{\tau} + \frac{U_{j+1}^{n+1} - U_{j}^{n+1}}{h} = \frac{u_{j}^{n+1} - u_{j}^{n}}{\tau} + \frac{u_{j+1}^{n+1} - u_{j}^{n+1}}{h} - T_{j}^{n}. \tag{4.26}
$$

On account of (4.9), (4.26) is zero. We add the right-hand side of (4.26), evaluated at (t^n, x_j) and (t^{n+1}, x_{j+1}) to (4.24) to obtain:

$$
\begin{aligned}
-R(t^{n+1/2}, x_{j+1/2}) = \frac{1}{2}&\left(\frac{U_{j}^{n+1} - U_{j}^{n}}{\tau} + \frac{U_{j+1}^{n+1} - U_{j}^{n+1}}{h}\right) \\
&- \frac{1}{2}\left(\frac{u_{j}^{n+1} - u_{j}^{n}}{\tau} + \frac{u_{j+1}^{n+1} - u_{j}^{n+1}}{h}\right) + \frac{1}{2}T_{j}^{n},
\end{aligned}
\tag{4.27}
$$

where T_j^n is the truncation error, defined in (4.12), at (t^n, x_j). Then, (4.27) simplifies to

$$
-R(t^{n+1/2}, x_{j+1/2}) = \frac{1}{2}\left(\frac{e_{j}^{n+1} - e_{j}^{n}}{\tau} + \frac{e_{j+1}^{n+1} - e_{j}^{n+1}}{h}\right) + \frac{1}{2}T_{j}^{n}. \tag{4.28}
$$

Examining (4.28), we see that the term in brackets is essentially a truncation error. We denote the maximum of this quantity over (t^n, x_j) as T^e. In the same vein as in Defn. 4.4.2, we will assume that there exists a constant $C_e > 0$ such that the following bound holds:

$$
|T^e| := \max_n \max_j \left|\left(\frac{e_{j}^{n+1} - e_{j}^{n}}{\tau} + \frac{e_{j+1}^{n+1} - e_{j}^{n+1}}{h}\right)\right| \leq C_e(\tau + h)\left\|u_0\right\|_{\mathrm{C}^2(\Omega)}. \tag{4.29}
$$

Combining (4.29) with the known bounds we have for the truncation error, T_j^n (see Defn. 4.4.1), we obtain a bound for the residual $|R|$.

Firstly, since $u \in \mathrm{C}^2([0,T]\times\Omega)$, $|u_{tt}|$ and $|u_{xx}|$ are bounded, for the particular problem, by $\|u_0\|_{\mathrm{C}^2([0,T]\times\Omega)}$ for all $(t,x) \in [0,T]\times\Omega$. We define the maximum truncation error for all $t = [0,T]$ to be

$$T^u := \max_n \max_j \left|T_j^n\right|, \tag{4.30}$$

for which the following bound holds (see also Defn. 4.4.1)

$$|T^u| \leq C_u(\tau+h)\,\|u_0\|_{\mathrm{C}^2(\Omega)}\,, \tag{4.31}$$

for some $C_u \in \mathbb{R}^+$. Finally, we combine (4.29) and (4.31) to obtain

$$\begin{aligned}\left|R\big(t^{n+1/2}, x_{j+1/2}\big)\right| &\leq \frac{1}{2}(|T^u| + |T^e|)\\ &\leq \frac{1}{2}\max\{C_e, C_u\}(\tau+h)\,\|u_0\|_{\mathrm{C}^2(\Omega)}\,.\end{aligned} \tag{4.32}$$

For brevity, we define $C := \frac{1}{2}\max\{C_e, C_u\}$, which simplifies (4.32) to

$$\left|R\big(t^{n+1/2}, x_{j+1/2}\big)\right| \leq C(\tau+h)\,\|u_0\|_{\mathrm{C}^2(\Omega)}\,. \tag{4.33}$$

Now, we substitute (4.33) in (4.20) we obtain

$$\begin{aligned}\int_0^T \|\delta(s)R(s)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s &\leq \max(1,T)\,\tau h \sum_{n=0}^{N-1}\sum_{j=0}^{M-1}\left|R(t^{n+1/2}, x_{j+1/2})\right|^2\\ &\leq C^2\max(1,T)\,\tau h(\tau+h)^2\,\|u_0\|^2_{\mathrm{C}^2(\Omega)}\sum_{n=0}^{N-1}\sum_{j=0}^{M-1}1.\end{aligned} \tag{4.34}$$

Now we use the fact that

$$\sum_{j=0}^{M-1} 1 = M = \frac{1}{h} \quad \text{and} \quad \sum_{n=0}^{N-1} 1 = N = \frac{1}{\tau}, \tag{4.35}$$

to evaluate the summations in the second line of (4.34), thereby simplifying it to

$$\int_0^T \|\delta(s)R(s)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s \leq C^2\max(1,T)(\tau+h)^2\,\|u_0\|^2_{\mathrm{C}^2(\Omega)}\,. \tag{4.36}$$

Finally, knowing from interpolation theory that $\left\|\big(u-\widehat{U}\big)(0)\right\|_{\mathrm{L}^2(\Omega)} = \mathcal{O}(h^2)$ and combining this with (4.36) concludes the proof. □

4.4.8 Remark. Combining the results of Corollary 4.4.6 and Lemma 4.4.7 implies the a posteriori bound is fully robust for FTBS, at least with smooth initial data. We will subsequently demonstrate this with numerical examples.

4.4.9 Remark (Asymptotic convergence rate for the reconstruction residual)**.** Let $\{U_j^n\}_j^n$ be the CNCS approximation of u, the solution of (4.2) with $u_0 \in \mathrm{C}^2(\Omega)$. Suppose $\widehat{U}$ is the piecewise linear in time and piecewise quadratic in space interpolant which is constructed using the conditions (4.41) and let $\omega(t)\,\mathscr{E}(t)^2$ be defined in (4.13), then

$$\sqrt{\omega(t)\,\mathscr{E}(t)^2} = \mathrm{O}(\tau^2 + h^2). \tag{4.37}$$

4.5 Numerical experiments

In this section we test the validity and asymptotic behaviour of our estimate.

4.5.1 Test 1: Smooth initial condition

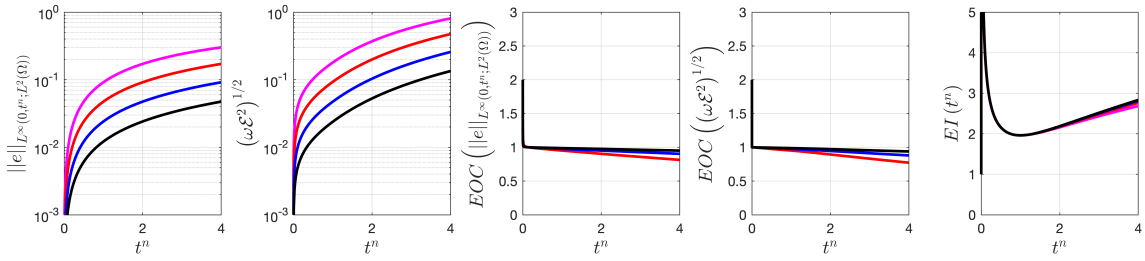
We illustrate the asymptotic behaviour of the a posteriori bound by considering the solution of (4.2) with initial condition

$$u_0(x) = \sin(2\pi x)\,, \tag{4.38}$$

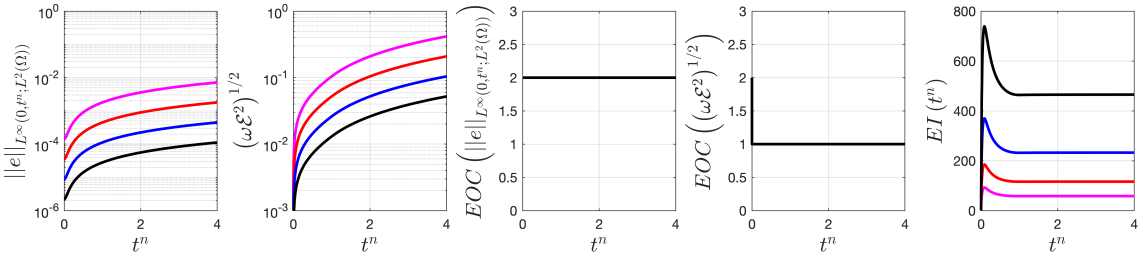
and periodic boundary conditions. In this case the solution of this problem is given by

$$u(x,t) = \sin(2\pi(x-t))\,. \tag{4.39}$$

Using the bilinear Lagrange interpolant as a reconstruction operator, we examine the behaviour of the bound for both the FTBS scheme, where we have an optimal estimate, and the CNCS scheme, where we have only an upper bound. We conduct the simulations over a family of meshes with discretisation parameter $h = 2^{-m}, m = 4, \ldots 7$, with a time-step $\tau = \frac{h}{10}$. The results are shown in Figure 4.1. As indicated in Lemma 4.4.7 the asymptotic convergence rate of the estimate matches that of the error for the FTBS scheme with a favourable $EI(4) < 3$, but the Lemma does not naturally extend to the CNCS scheme. The reason for this is that the naive bilinear interpolant lacks the approximability to achieve the optimal convergence rate of the residual.



(a) FTBS (4.9).



(b) CNCS (4.10).

Fig. 4.1. Errors and asymptotic convergence rates for the a posteriori bound given in Corollary 4.4.6 using the bilinear interpolant of the FTBS (4.9) and CNCS (4.10) approximations of (4.2) with initial condition (4.38) and periodic boundary conditions. The estimate is optimal for the FTBS scheme and suboptimal for CNCS.

4.5.2 Remark (Suboptimal bound for CNCS)**.** The bound given in Corollary 4.4.6 is suboptimal for the CNCS scheme. The reason for this is that the bilinear Lagrange interpolant we used for the reconstruction simply does not have the approximability required. Since the CNCS scheme is formally of order two, we must incorporate this information into the reconstruction we use.

We do this by building information from within the finite difference spatial discretisation directly into the post-processor. Indeed, within each space-time patch $[t^n, t^{n+1}] \times [x_j, x_{j+1}]$, we can augment $\widehat{U}$ such that it is defined as a patch-wise biquadratic interpolant by constructing it in two steps as follows. Firstly, we construct the interpolant in the temporal direction as the unique piecewise quadratic polynomial in time which satisfies

$$\begin{aligned} \widehat{U}_j^t(t^n) &= U_j^n \\ \widehat{U}_j^t\big(t^{n+1}\big) &= U_j^{n+1} \quad \text{and} \\ \partial_t \widehat{U}_j^t(t^n) &= -\frac{1}{2h}\big(U_{j+1}^n - U_{j-1}^n\big)\,. \end{aligned} \tag{4.40}$$

Note that we have used the superscript t to denote time dependence for this inter-

mediate step of the reconstruction, $\widehat{U}^t$, and the sub-script j to denote the spatial position, x_j, on the grid where we construct the temporal polynomial. Once the $\widehat{U}^t_j$, $j = 0, \ldots, M-1$ are computed, we use them to obtain the unique patch-wise bi-quadratic polynomial which satisfies

$$\begin{aligned} \widehat{U}(x_j, t^n) &= \widehat{U}^t_j(t^n)\,, \\ \widehat{U}(x_{j+1}, t^n) &= \widehat{U}^t_{j+1}(t^n) \quad \text{and} \\ \partial_x \widehat{U}(x_j, t^n) &= \frac{\widehat{U}^t_{j+1}(t^n) - \widehat{U}^t_{j-1}(t^n)}{2h}. \end{aligned} \tag{4.41}$$

In this way we are closer to the spirit of a finite difference method. The argument from Lemma 4.4.7 can be modified to apply in this case. Indeed, one can show that applying the same argument the interpolant given by (4.41) yields $\sqrt{w\mathscr{E}^2} = \mathrm{O}(\tau^2 + h^2)$, which is optimal for the CNCS scheme (provided that the solution possesses the requisite regularity to allow the estimate to achieve this optimal rate).

To illustrate this asymptotic convergence properties we have replicated the same experiment as in Figure 4.1 for the quadratic reconstruction. This is shown in Figure 4.2. It can be seen that the additional information provided by appropriately increasing the order of data representation allows us to achieve optimal convergence rates of the a posteriori bound and favourable effectivities.

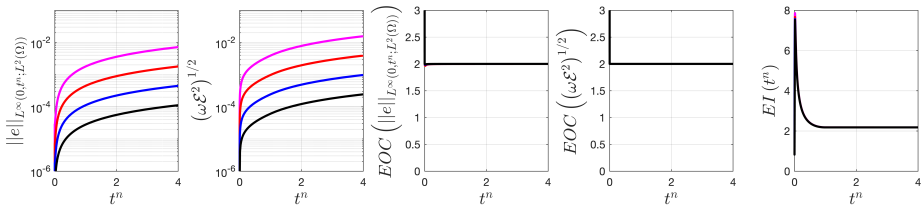


Fig. 4.2. Errors and asymptotic convergence rates for the linear in time, quadratic in space Hermite interpolant of CNCS (4.10) approximations of (4.2) with initial condition (4.38) and periodic boundary conditions. The estimate is now optimal for the CNCS scheme with favourable effectivity of $EI(4) \sim 2.2$.

The ideas presented in this section form an intuitive way to obtain the reconstruction. It is possible to generalise this quite naturally to other spatio-temporal discrestisations as we will present in the forthcoming sections.

The last test in this section is presented in order to motivate adaptivity in the context of finite difference schemes for conservation laws. We highlight additional

challenges that arise in implementing adaptivity, even for scalar examples in one spatial dimension.

4.5.3 Test 2: Parasite detection in 1D

Grid and discretisation

Consider a piecewise uniform grid which has a mesh-size change at $x = 0.5$:

$$x_j - x_{j-1} = \begin{cases} h & \text{if} \quad x_j \quad \leq 0.5 \\ 2h & \text{if} \quad x_j \quad > 0.5. \end{cases} \tag{4.42}$$

Such a grid is shown in Fig. 4.3. Let the set $\{u_j(t)\}_{j=0}^{M}$ denote the numerical solution on time-continuous lines produced by the following central difference semi-discretisation posed over the grid (4.42):

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} = \begin{cases} \frac{u_{j+1}-u_{j-1}}{2h} & \text{for} \quad x_j < 0.5, \\ \frac{u_{j+1}-u_{j-1}}{3h} & \text{for} \quad x_j = 0.5, \\ \frac{u_{j+1}-u_{j-1}}{4h} & \text{for} \quad x_j > 0.5. \end{cases} \tag{4.43}$$

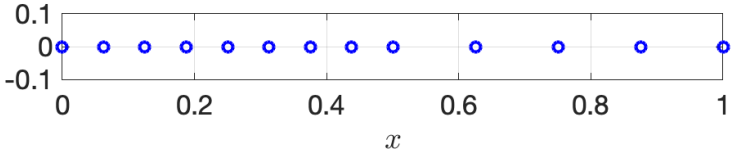


Fig. 4.3. A piecewise uniform grid given by (4.42) with a doubling of the mesh size for $x \geq 0.5$.

In this test we investigate the resulting behaviour of the a posteriori bound (4.13) in the presence of parasitic waves and we verify the capability of the bound we construct to detect such waves. Firstly, we will give some background information on the propagation characteristics of parasites and then we will describe the relevant numerical experiments.

Propagation characteristics of parasitic waves

In this section we provide some background on propagation properties of parasitic waves for the case they are generated by a mesh non-uniformity (see Fig. 4.3). The material we will present in this section is a summary of [Vic80, §4].

4.5.4 Definition. (Fourier Transform in the time-domain) We denote the time-domain Fourier transform (FT) of u_j by $\widehat{u}_j$

$$\widehat{u}_j(\xi) = \int_{-\infty}^{\infty} u_j(t) \exp(-i\xi t)\, \mathrm{d}t, \tag{4.44}$$

and we let $\{\widehat{u}_j(\xi)\}$ denote the Fourier transforms of the semi-discrete numerical solutions $\{u_j(t)\}$ which are obtained by (4.43).

4.5.5 Definition. (Fourier transform of the spatial semi-discretisation) In order to obtain information on the propagation characteristics of the $\{u_j(t)\}$, we apply the Fourier Transform to (4.43) to obtain

$$i\xi\widehat{u}_j = -\left(\frac{\widehat{u}_{j+1} - \widehat{u}_{j-1}}{2h}\right), \tag{4.45}$$

which can be re-written as

$$\widehat{u}_{j+1} + 2i\xi h\widehat{u}_j - \widehat{u}_{j-1} = 0. \tag{4.46}$$

4.5.6 Definition. (Space-shift Operator) We define the space-shift operator $\mathbb{E}$ through the relation

$$u_{j+1} = \mathbb{E}u_j, \tag{4.47}$$

and we denote by $\widehat{\mathbb{E}}$ its image in the Fourier domain.

We solve (4.46) by seeking solutions for which the ratio

$$\widehat{\mathbb{E}} = \frac{\widehat{u}_{j+1}}{\widehat{u}_j} \tag{4.48}$$

is independent of j. In order to find solutions that satisfy (4.48) we use $\widehat{\mathbb{E}}$ to write (4.46) as

$$\left(\widehat{\mathbb{E}} + 2i\xi h - \widehat{\mathbb{E}}^{-1}\right)\widehat{u}_n = 0 \tag{4.49}$$

and we impose that $\widehat{\mathbb{E}}$ satisfies the relation

$$\widehat{\mathbb{E}}^2 + 2i\xi h\widehat{\mathbb{E}} - 1 = 0. \tag{4.50}$$

This equation has two solutions for $|\xi h| < 1$:

$$\begin{aligned} \widehat{\mathbb{E}}_1 = &\quad -i\xi h + \sqrt{1-(\xi h)^2}, \\ \widehat{\mathbb{E}}_2 = &\quad -i\xi h - \sqrt{1-(\xi h)^2}. \end{aligned} \tag{4.51}$$

4.5.7 Remark. The cases $|\xi h| \geq 1$ are treated in [Vic81b] and interested readers can consult this source for more information. However, we are not interested in them in this section.

4.5.8 Remark. In [Vic81a], $\widehat{\mathbb{E}}_{1,2}$ are referred to as cell transfer functions. They are used in the derivation of the propagation characteristics of solutions, $\{u_j(t)\}$, of (4.43) and in particular of the wavelength, phase velocity and group velocity.

The existence of two cell transfer functions reflects the fact that the set of solutions $\{u_j(t)\}$ to (4.43) can be decomposed into two types of solutions, say $\{p_j(t)\}$ and $\{q_j(t)\}$:

$$\{u_j(t)\} = \{p_j(t)\} + \{q_j(t)\}\,. \tag{4.52}$$

Then, $\widehat{\mathbb{E}}_1$ corresponds to solutions of the type $\{p(t)\}$ while $\widehat{\mathbb{E}}_2$ corresponds to $\{q(t)\}$. These two different types of solutions, $\{p(t)\}$ and $\{q(t)\}$ have different propagation characteristics. In particular, we are most interested in wavelength, group velocity and phase velocity.

4.5.9 Remark. Phase velocity is the velocity at which a wave propagates in a medium. Group velocity, in a context where, say, sinusoidal wave forms of different frequencies are superimposed, is the speed at which the entire *pattern* travels. The two can be different.

4.5.10 Proposition. *(Propagation characteristics of $\{p_j(t)\}$ and $\{q_j(t)\}$) Let λ denote the wavelength, let C_p denote phase velocity and C_g denote group velocity.*

Solutions of the the type $\{p_j(t)\}$ have the following propagation characteristics:

$$C_p^p(\xi) = \frac{\xi h}{\arcsin(\xi h)}, \quad C_g^p(\xi) = \sqrt{1-(\xi h)^2} \quad \textit{and} \quad \lambda^p(\xi) = \frac{2\pi h}{\arcsin(\xi h)} \tag{4.53}$$

Solutions of the the type $\{q_j(t)\}$ have the following propagation characteristics:

$$C_p^q(\xi) = \frac{\xi h}{\pi - \arcsin(\xi h)}, \quad C_g^q(\xi) = -\sqrt{1-(\xi h)^2} \quad \textit{and} \quad \lambda^q(\xi) = \frac{2\pi h}{\pi - \arcsin(\xi h)} \tag{4.54}$$

4.5.11 Remark. In the context we are considering, solutions of the type $\{p(t)\}$ have positive phase and group velocities and wavelengths in the range $(4h, \infty)$, whereas solutions of the type $\{q(t)\}$ have positive phase velocity, negative group velocity and wavelengths in the range of $(2h, 4h)$ (see [Vic81a, §3 and §5]).

4.5.12 Remark. Suppose that a smooth wave-form is travelling over a grid and the spacing changes from fine to coarse abruptly. Then, it is shown in [Vic81b, §6] that reflection occurs at the interface between the fine and coarse mesh portions, and that the reflected component of the solution contains mostly the type $\{q(t)\}$.

Recall that solutions of type $\{q_j(t)\}$ contain smaller wavelengths and are characterized by negative group velocity (see Pro. 4.5.10). Hence, once reflection occurs at the interface, the reflected components of the waveform will appear as a spurious oscillatory wave-train which travels in the opposite direction than that of the ongoing solution. This spurious oscillation, which is a numerical artefact, is what we refer to by the term parasite.

In order to demonstrate the concept of parasites in practice, as well as our a posteriori estimates capability to detect it, we use the CNCS scheme (4.10) on a piecewise non-uniform grid given by

$$h_j = x_{j+1} - x_j = \begin{cases} 2^{-9} & \text{if} \quad x_{j+1} > 1/2, \\ 2^{-10} & \text{otherwise} \end{cases}. \tag{4.55}$$

4.5.13 Remark (Truncation error of the central difference quotient)**.** The truncation error of the standard central difference approximation to the first derivative,

$$u_x(x_j, t^n) \approx \frac{U^n_{j+1} - U^n_{j-1}}{h_{j+1} + h_j}, \tag{4.56}$$

on a uniform grid is order 2 globally. On a non-uniform grid, it is locally only order 1 wherever $[\![h_j]\!] := h_{j+1} - h_j \neq 0$. In order to ensure that the spatial discretisation is second order on a non-uniform grid, we use the modified quotient:

$$u_x(x_j, t^n) \approx \left(\frac{1}{h_{j+1}} - \frac{1}{h_j + h_{j+1}}\right) U^n_{j+1} + \left(\frac{1}{h_j} - \frac{1}{h_{j+1}}\right) U^n_j + \left(\frac{1}{h_j + h_{j+1}} - \frac{1}{h_j}\right) U^n_{j-1}, \tag{4.57}$$

which is order two globally on a non-uniform grid as well and simplifies to (4.56) on a uniform grid.

The CNCS scheme (4.10) is modified in the spatial component with the new quotient, (4.57) and becomes

$$
\begin{aligned}
U_j^{n+1} &= U_j^n \\
&\quad - \frac{\tau}{2}\left[\left(\frac{1}{h_{j+1}} - \frac{1}{h_j + h_{j+1}}\right) U_{j+1}^n + \left(\frac{1}{h_j} - \frac{1}{h_{j+1}}\right) U_j^n + \left(\frac{1}{h_j + h_{j+1}} - \frac{1}{h_j}\right) U_{j-1}^n\right] \\
&\quad - \frac{\tau}{2}\left[\left(\frac{1}{h_{j+1}} - \frac{1}{h_j + h_{j+1}}\right) U_{j+1}^{n+1} + \left(\frac{1}{h_j} - \frac{1}{h_{j+1}}\right) U_j^{n+1} + \left(\frac{1}{h_j + h_{j+1}} - \frac{1}{h_j}\right) U_{j-1}^{n+1}\right] \\
&\quad \text{for } n = 0, \ldots, N-1 \text{ and } j = 0, \ldots, M-1. \\
U_j^0 &= u_0(x_j) \text{ for } j = 0, \ldots, M.
\end{aligned}
\tag{4.58}
$$

4.5.14 Remark. (Time-step) In the modified CNCS scheme, (4.58) there are two spatial step-sizes present in the domain we are considering. We note that we couple the temporal step, τ to the spatial step of the fine domain: $\tau = 0.1 \times 2^{-10}$.

In order to demonstrate the effect of the parasite, and the ability of the a posteriori bound to detect it, we localise the bound and plot $\|R\|_{\mathrm{L}^2(I_j)}$, with

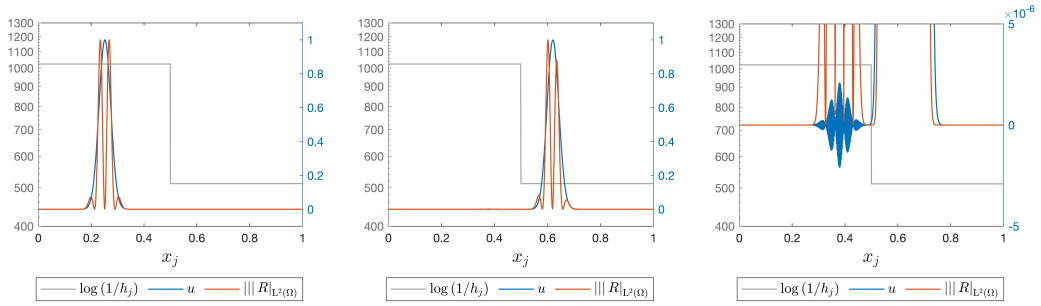
$$R = -\widehat{U}_t - \widehat{U}_x \tag{4.59}$$

with the reconstruction $\widehat{U}$ obtained using the conditions (4.41). We use an initial condition

$$u_0(x) = \exp\left(-1000(x - 0.25)^2\right). \tag{4.60}$$

as it is narrow and helps us show what the parasite looks like, and periodic boundary conditions.

The results are shown in Fig. 4.4. The parasitic wave is the highly oscillatory wave (blue colour) in the right-most plot (magnified). It forms as the solution travels over an abrupt grid-spacing change (see (4.55)) and travels in the opposite direction from the solution. Notice that the residual we have constructed using the conditions (4.40) and (4.41) correctly detects and tracks the parasite as can be seen by the small bumps in the second plot (orange colour). In general, we note that parasitic waves may rapidly pollute the computation.



(a) Before parasite forms. **(b)** After parasite forms. **(c)** Magnified Parasite.

Fig. 4.4. A parasitic wave (blue oscillatory waveform), forming as the solution travels over an abrupt mesh change, given by (4.55). We plot the solution and the normalized L^2-norm of the local residual (4.59) before and after parasite formation to demonstrate that the residual can be used to detect and track the parasite.

4.5.15 Remark. The reader should note that implementing mesh adaptivity in the presence of an existing grid discontinuity may exacerbate parasite formation and propagation. A potential solution would be to implement model adaptivity simultaneously (see [GP17]) in order to "dampen" parasites out or to selectively remove them.

4.6 Adaptivity

4.6.1 Adaptive Algorithm

Our adaptive algorithm is of SOLVE → ESTIMATE → MARK → REFINE type.

4.6.2 Remark (Maximum number of refinements)**.** For the purposes of this paper we allow a maximum of four refinement levels relative to the initial, uniform triangulation.

There are potentially several ways to compare the performance of a numerical solution on a uniform and an adaptive mesh. In this case we will use what we will refer to as an *equivalent uniform mesh*.

4.6.3 Definition (Equivalent Uniform Mesh)**.** We define a uniform mesh to be equivalent to an adaptive mesh if it has the same cumulative number of degrees of

freedom, which we define as

$$\sum_{n=0}^{N} N_{dof}(t^n)\,. \tag{4.61}$$

We find this number by firstly running the simulation on the adaptive mesh and recording the number of dof at each time-step. We then average this over the number of time steps and set the resulting value to be the number of degrees of freedom for the equivalent uniform mesh.

4.6.4 Marking

The criterion for marking cells for refinement/coarsening is based is a maximum strategy. We refine cells wherein the value of the local indicator is larger than some multiple of the maximum value of the local residual and coarsen cells where it is lower. We do nothing in cases where the local value of the indicator falls in between the two values. Lastly, we do not coarsen a cell marked if its sibling is marked for refinement at the same time-step.

This strategy is described in detail in [SS05, §1.5] and it is modified for time-dependent problems for the purposes of these experiments. Briefly, let η_S denote the local residual term $\|R\|_{\mathrm{L}^2(I_j)}$ in a 'cell' $S := I_j = [x_j, x_{j+1}]$ for $S \in S_K$, where S_K is the initial parent triangulation. We set two predefined tolerances γ_r and γ_c for refining and coarsening respectively. We mark a cell for refinement if

$$\eta_S \geq \gamma_r \max_{S'} \eta_{S'},\ S' \in S_k \tag{4.62}$$

and we mark for coarsening coarsen if

$$\eta_S \leq \gamma_c \max_{S'} \eta_{S'},\ S' \in S_k. \tag{4.63}$$

The reader should note that there are several ways of choosing γ_c and γ_r. We chose them empirically as $\gamma_c = .05e-10$ and $\gamma_r = 0.5e-8$. We summarize the marking-refinement/coarsening process in Algorithm 1.

Algorithm 1 Mesh Adaptivity

Require: Maximum number of refinement levels relative to initial triangulation, refinement parameter γ_r and coarsening parameter γ_c.

while $t^n < T$ **do**

set $S_K^{(0)} = S_k$, the initial grid for the time-step t^n.

Solve (4.9) on $S_K^{(0)}$ and compute the local indicator η_S for all $S \in S_K^{(0)}$.

Set $\eta_{\max} := \max_{S' \in \mathcal{S}_k} \eta_{S'}$.

for $S \in S_K^{(0)}$ **do**

if $\eta_S > \gamma_r \eta_{\max}$ **then**

Mark S for refiment

end if

end for

for $S \in S_K^{(0)}$ **do**

if $\eta_S < \gamma_c \eta_{\max}$ **then**

if $S \notin S_K$ **and** the siblings of S are not marked for refinement **then**

Mark S for coarsening

end if

end if

end for

for $S \in S_K^{(0)}$ **do**

if S is marked for refinement **then**

Create two children of S

Prolong $\boldsymbol{U}$ over S and assign corresponding values to children nodes

Prolong the grid $\{x_j\}$ assigning corresponding values to children nodes

elseif S is marked for coarsening

Restrict $\boldsymbol{U}$ by assigning the relevant values from S and its sibling to their parent

Restict the grid $\{x_j\}$ by assigning corresponding values from S and its sibling to their parent

Delete S and its sibling

end if

end for

Recompute both the error and the bound from (4.13) and record them as the values for time step t^n

do $n := n + 1$

end while

4.6.5 Adaptive experiments

In this section we describe the numerical experiments we run to test the a posteriori bounds as criteria for adaptivity. We use a linear advection problem with a discontinuous initial condition and periodic boundary conditions. In this way we benchmark the performance of the indicator in a more challenging setting than in previous sections.

4.6.6 Remark (Grid-spacing for the adaptive mesh)**.** We start the simulations at the coarsest available refinement level. Then, the indicator detects where this is insufficient and refines locally, which essentially pertains to the vicinity of the discontinuities. This is the reason for the small but abrupt increase in the number of dofs in the left plot in Fig. 4.6 during the first few time steps of the simulation.

4.6.7 Remark (Time-step for the adaptive mesh)**.** In order to maintain numerical stability, in the absence of an adaptive mechanism for the time-step, we couple the time-step to the smallest spatial step present in the computation, to maintain numerical stability.

Linear advection

We run a benchmarking experiment experiment using the linear advection equation

$$u_t + u_x = 0 \tag{4.64}$$

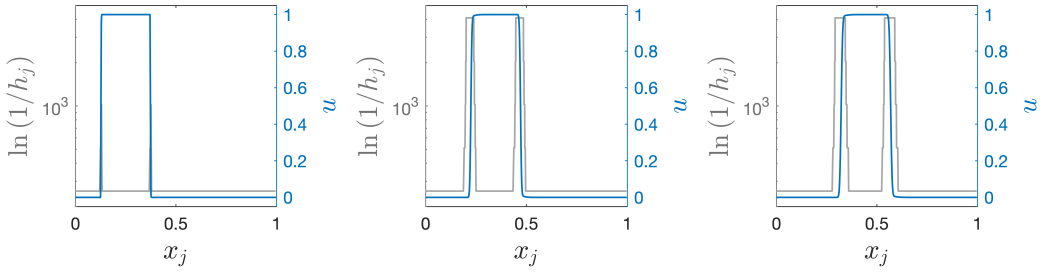
over a domain $\Omega = [0, 1]$ with $T = 0.5$ using periodic boundary conditions and a discontinuous initial condition given by

$$u_0(x) = \begin{cases} 1 & \text{if} \quad |x - .25| \leq 0.125 \\ 0 & \text{if} \quad |x - .25| > 0.125. \end{cases} \tag{4.65}$$

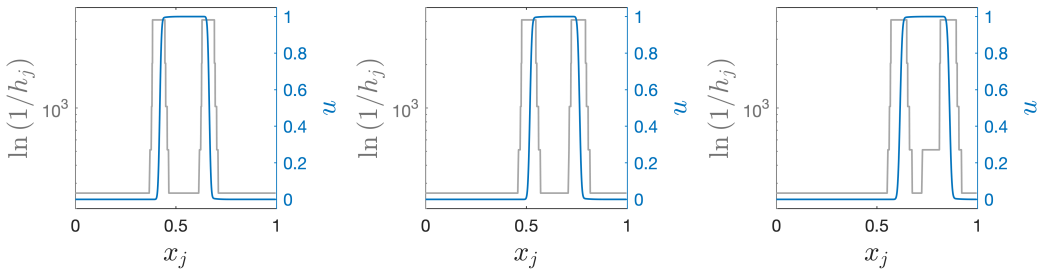
We discretise the problem using a Forward-Time Backward Space scheme given by (4.9) for both the adaptive and the uniform case. The residual in this test case is constructed using a bilinear Lagrange interpolant (see (4.14)). The simulation starts at the coarsest refinement level out of a maximum allowable three refinement, with a uniform grid with spatial step $h = 2^{-5}$ and a temporal step $\tau = \frac{2^{-7}}{10}$ which remains constant throughout the simulation. The grid-spacing for the equivalent uniform

grid (see Rem. 4.6.3) corresponds to 720 degrees of freedom, i.e. $0 = x_0 < \cdots < x_{719} = 1$.

In Fig. 4.5, we plot snapshots of the solution (blue line) and the logarithm of the reciprocal of the grid spacing (grey line). The evolution of the error and the number of DOFs is shown in 4.6 alongside a comparison with an equivalent uniform grid. Notice that the residual reliable detects and tracks regions of refinement/coarsening interest, such as in the vicinity of discontinuities (for refinement) and away from them (for coarsening). This is evident from an examination of the evolution of the grid spacing, shown as the grey line.



(a) $t = 0\,s$ **(b)** $t \approx 0.1\,s$ **(c)** $t \approx 0.2\,s$



(d) $t \approx 0.3\,s$ **(e)** $t \approx 0.4\,s$ **(f)** $t \approx 0.5\,s$

Fig. 4.5. Evolution of the solution, u, (blue line) and of the logarithm of the reciprocal of the local grid-spacing (grey line) for the advection problem (4.2), discretised with a FTBS spatio-temporal discretization (see (4.9)). The residual, (4.14), reliably detects and tracks regions where refinement/coarsening is required.

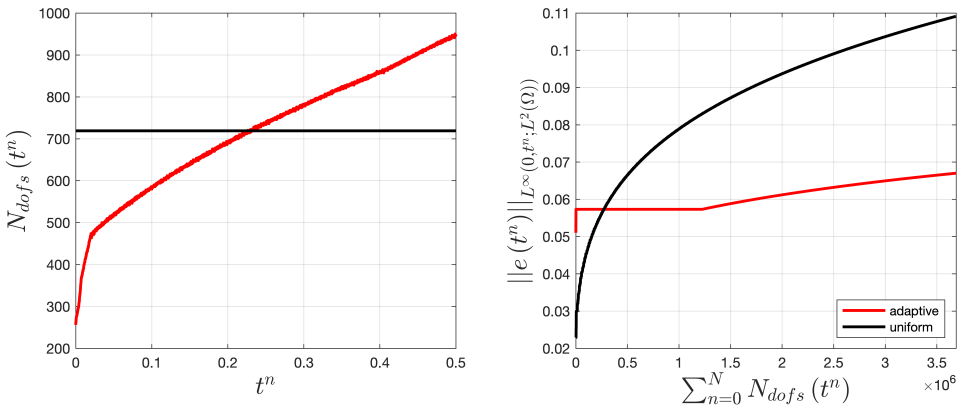


Fig. 4.6. A comparison of the performance of the adaptive grid with an equivalent uniform grid (see Defn. 4.6.3), which in this case has 720 dofs. The adaptive grid consistently maintains a lower level of error throughout the major part of the simulation compared to the equivalent uniform grid.

4.7 Conclusion

In this section we presented an a posteriori error bound for a class of well used finite difference schemes for the transport equation in one spatial dimension. This is intended as motivation for a more general framework for constructing a posteriori error bounds for FD schemes for hyperbolic conservation laws in the chapters that follow.

The a posteriori bound was constructed using a simple reconstruction of the FD numerical solution obtained using the FTBS and CNCS scheme. We touched upon the issue of optimality and how this depends on the particular set of conditions used to obtain the reconstruction. Specifically, we showed, in a simple example, how we can incorporate information from the chosen FD scheme in order to obtain an optimal bound for a higher order FD approximation.

Additionally, we demonstrated the capability of a bound constructed in this way to detect parasites: numerical artefacts that are generated as a result of the presence of discontinuities in the underlying numerical approximation. In this case the discontinuity in question is an abrupt change in mesh spacing.

Lastly, we demonstrated that the bound can be used as a driver for adaptivity using a numerical experiment with a discontinuous initial condition. We found that the

reconstruction-based bound performed favourably compared to a grid with uniform spacing, with the comparison being on the basis of the two experiments (adaptive and non-adaptive) having the same number of cumulative degrees of freedom.

Chapter 5

Postprocessing in finite difference schemes

Abstract

In this chapter, we introduce a class of Weighted Essentially Non-Oscillatory (WENO) schemes for systems of conservation laws. These schemes form the basis of the a posteriori analysis that will appear in later chapters through the WENO reconstruction operator. While the WENO schemes are fundamentally based upon this operator, it is not explicitly written down in the literature, rather it is typically eliminated in the scheme derivation. We use this chapter to give an explicit representation of the operator which we will then use as a fundamental component of the a posteriori bounds we construct.

5.1 Introduction

In this chapter we expand upon the material in Chapter 4 by explaining how to obtain post-processors of FD solutions for non-linear problems. As we explained in previous chapters, the post-processor is necessary in order to obtain the a posteriori bound.

In addition, we expand upon the issue of optimality of the a posteriori bound obtained from the post-processor that was touched upon in the previous chapter. Specifically, we do this by showing how to obtain polynomial reconstructions of solutions obtained by higher order schemes both temporally and spatially. We also show how to obtain post processors for non-linear problems.

We describe explicitly and in detail the WENO component of the reconstruction procedure. We evaluate the performance of this as an interpolant using various functions of varying regularity in order to demonstrate the approximability of this interpolant.

5.1.1 Motivation

Our motivation in this chapter is to extend the procedure introduced in Chapter 4 and formalize it into a framework for obtaining reconstructions for hyperbolic conservation laws in one-dimension.

We emphasize to the reader the fact that now we are changing strategy with regard to obtaining reconstructions. We will now endeavour to directly build information from the numerical scheme in the reconstruction. This will be reflected from the the very beginning of the procedure, at the framework level.

The framework will contain mechanisms for obtaining spatial reconstruction for higher order schemes. This will enable us to build optimal a posteriori error bounds for conservation laws, at least in the pre-shock regime, without resorting to unfavourable temporal to spatial step coupling. We note that in the temporal component the framework we present in this chapter is limited to third order but it can be extended to higher orders in the way shown in Chapter 2.

We will present the reconstruction procedure in the context of WENO schemes on non-uniform grids. The reason for this preference is twofold. Firstly, WENO schemes have high orders of approximability in space and good behaviour in the presence of discontinuities. Secondly, presenting the procedure for non-uniform grids will enable us to use the a posteriori estimate as a driver for mesh adaptivity in the numerical tests we will run in later chapters.

5.1.2 Chapter contribution

In this chapter we present the framework for obtaining reconstructions from solutions of general finite difference schemes approximating systems of non-linear conservation laws. The novelty is that the reconstructions from this framework will enable the construction of reliable a posteriori estimate for general FD schemes. The result is quite general, in that we assume nothing on the exact solution although the final

estimate is conditional, in that it holds only under some conditions on the numerical solution. The framework has inherent mechanisms to construct robust estimates of high-order, at least in the pre-shock regime, enabling the user to obtain optimal bounds for high order FD schemes using WENO interpolation [LSZ09, JSB+19]. This will be demonstrated using appropriate numerical experiments in the next chapter.

The rest of this chapter is structured as follows. In §5.2 we present the model problem we will be using throughout this chapter. In §5.3 we present the spatial and temporal discretisations we will be using to approximate the model problem. In §5.3.9 we present WENO schemes which will be used as spatial discretisations in some of the examples we consider in later chapters. In §5.3.17 we present the procedure for obtaining the spatial component of the WENO reconstruction (which we present in later chapters). In §5.4 we carry out some numerical experiments to benchmark the behaviour of the spatial component of the reconstruction by using it as an interpolant of functions of varying regularity. We conclude the chapter in §5.5.

5.2 Hyperbolic systems model problem

In this section we will present the model problem we will use throughout this chapter as well as for the following chapters.

5.2.1 Definition (One-dimensional system of conservation laws)**.** We consider problems of the form

$$\begin{aligned} \boldsymbol{u}_t + \partial_x \boldsymbol{f}(\boldsymbol{u}) &= \boldsymbol{0}, \\ \boldsymbol{u}(x,0) &= \boldsymbol{u}_0(x)\,, \end{aligned} \qquad \text{for } (x,t) \in \Omega \times (0,\infty) \tag{5.1}$$

with $\boldsymbol{u} = (u_1, \ldots, u_p)^T$ and $\boldsymbol{f}(\boldsymbol{u}) = (f_1(\boldsymbol{u}), \ldots, f_p(\boldsymbol{u}))^T$ and complemented with periodic boundary conditions. In particular,

$$\begin{aligned} \boldsymbol{u}: \quad \mathbb{R} \times \mathbb{R}^+ &\to \mathbb{R}^p \\ (x,t) &\mapsto \boldsymbol{u}(x,t) \end{aligned} \tag{5.2}$$

and the flux function $\boldsymbol{f}$

$$\begin{aligned} \boldsymbol{f}: \quad \mathbb{R}^p &\to \mathbb{R}^p \\ \boldsymbol{u}(x,t) &\mapsto \boldsymbol{f}(\boldsymbol{u}(x,t)) \end{aligned} \tag{5.3}$$

5.3 Numerical methods and discretisation

In this section we present the spatial and temporal discretisations we will use to approximate (5.1). We will use boldface notation for vector-valued quantities in order to distinguish them from scalar quantities.

5.3.1 Spatial discretisation

We will denote by $\boldsymbol{U}_j^n$ the numerical approximation to $\boldsymbol{u}(x_j, t^n)$. It is well known that numerical schemes for non-linear conservation laws may converge to functions which are not weak-solutions of the original problem (see [LeV92, §12.1]). We address this problem by expressing the method in conservation form. We use a consistent numerical flux function $\boldsymbol{F}$, which takes $p+q+1$ arguments:

$$\begin{array}{rccl} \big(\boldsymbol{U}_{j-p+1}^n, \cdots, \boldsymbol{U}_{j+q}^n\big): & \boldsymbol{F}_j^n & \mapsto & \boldsymbol{F}\big(\boldsymbol{U}_{j-p}^n, \cdots, \boldsymbol{U}_{j+q}^n\big) \\ & \boldsymbol{F}(\boldsymbol{v}, \cdots, \boldsymbol{v}) & = & \boldsymbol{f}(\boldsymbol{v})\,, \end{array} \tag{5.4}$$

where p and q are simply used to determine the width of the computational stencil. We use $\boldsymbol{F}$ to approximate $\partial_x \boldsymbol{f}$ such that

$$\partial_x \boldsymbol{f}(\boldsymbol{u}) \approx \frac{1}{h}\big(\boldsymbol{F}\big(\boldsymbol{U}_{j-p}^n, \cdots, \boldsymbol{U}_{j+q}^n\big) - \boldsymbol{F}\big(\boldsymbol{U}_{j-p-1}^n, \cdots, \boldsymbol{U}_{j+q-1}^n\big)\big)\,. \tag{5.5}$$

We can then use a method-of-lines approach in the discretisation of (5.1) by requiring

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{U}_j = \frac{1}{h}\big(\boldsymbol{F}\big(\boldsymbol{U}_{j-p}^n, \cdots, \boldsymbol{U}_{j+q}^n\big) - \boldsymbol{F}\big(\boldsymbol{U}_{j-p-1}^n, \cdots, \boldsymbol{U}_{j+q-1}^n\big)\big) \quad \forall\, j = 0, \ldots, M. \tag{5.6}$$

For clarity, we provide illustrative examples of $\boldsymbol{F}$ for the Lax-Friedrichs and the Lax-Wendroff scheme.

5.3.2 Remark (Conservation form for the Lax-Friedrichs scheme)**.** The Lax-Friedrichs scheme can be written in conservation form, (5.6), by defining the numerical flux function $\boldsymbol{F}$ as

$$\boldsymbol{F}(\boldsymbol{U}_j, \boldsymbol{U}_{j+1}) := \frac{h}{2\tau}(\boldsymbol{U}_j - \boldsymbol{U}_{j+1}) + \frac{1}{2}(\boldsymbol{f}(\boldsymbol{U}_j) + \boldsymbol{f}(\boldsymbol{U}_{j+1}))\,. \tag{5.7}$$

The Lax-Friedrichs flux is formally $\mathcal{O}(h)$.

5.3.3 Remark (Conservation form for the Lax-Wendroff scheme)**.** The Lax-Wendroff scheme can be written in conservation form by using the Richtmayer two-stage

method. We define the numerical flux function $\boldsymbol{F}$ as

$$\boldsymbol{F}(\boldsymbol{U}_j, \boldsymbol{U}_{j+1}) := \boldsymbol{f}\Big(\boldsymbol{U}_{j+1/2}^{n+1/2}\Big), \tag{5.8}$$

where

$$\begin{aligned} \boldsymbol{U}_{j+1/2}^{n+1/2} &:= \frac{1}{2}\big(\boldsymbol{U}_{j+1}^{n} + \boldsymbol{U}_{j}^{n}\big) - \frac{\tau}{2h}\big(\boldsymbol{f}\big(\boldsymbol{U}_{j+1}^{n}\big) - \boldsymbol{f}\big(\boldsymbol{U}_{j}^{n}\big)\big) \\ \boldsymbol{U}_{j-1/2}^{n+1/2} &:= \frac{1}{2}\big(\boldsymbol{U}_{j}^{n} + \boldsymbol{U}_{j-1}^{n}\big) - \frac{\tau}{2h}\big(\boldsymbol{f}\big(\boldsymbol{U}_{j}^{n}\big) - \boldsymbol{f}\big(\boldsymbol{U}_{j-1}^{n}\big)\big). \end{aligned} \tag{5.9}$$

The conservation form of the scheme is then given by:

$$\boldsymbol{U}_{j}^{n+1} := \boldsymbol{U}_{j}^{n} - \frac{\tau}{h}\Big(\boldsymbol{f}\Big(\boldsymbol{U}_{j+1/2}^{n+1/2}\Big) - \boldsymbol{f}\Big(\boldsymbol{U}_{j-1/2}^{n+1/2}\Big)\Big). \tag{5.10}$$

The Lax-Wendroff scheme is formally $\mathcal{O}(\tau^2 + h^2)$.

5.3.4 Remark. The aforementioned schemes have various limiters which can be applied to them. These can be fully accounted for in the framework we propose.

5.3.5 Temporal discretisation

We approximate the temporal variable using both implicit and explicit temporal discretisations. For example, using various 1-stage Runge-Kutta methods given by the Butcher tableau

$$\begin{array}{c|c} \theta & \theta \\ \hline & 1 \end{array}. \tag{5.11}$$

In general, we will be using Strong-Stability Preserving Runge-Kutta (SSP-RK) methods ([GST01]).

5.3.6 Definition. (RK methods) Let $\boldsymbol{U}^{(j)}$ denote the j^{th} predictor stage of an $m-$stage RK and likewise, let $\boldsymbol{F}^{(j)}$ denote the numerical flux function (see (5.4)) computed at the j^{th} stage. Then, the $m-$stage RK method is given as follows:

$$\begin{aligned} \boldsymbol{U}^{(0)} &= \boldsymbol{U}^{n}, \\ \boldsymbol{U}^{(i)} &= \boldsymbol{U}^{n} + \tau \sum_{k=1}^{s} \frac{a_{i,k}}{h}\Big(\boldsymbol{F}_{j}^{(k)} - \boldsymbol{F}_{j-1}^{(k)}\Big) \\ \boldsymbol{U}^{n+1} &= \boldsymbol{U}^{n} + \tau \sum_{i=1}^{s} \frac{b_i}{h}\Big(\boldsymbol{F}_{j}^{(i)} - \boldsymbol{F}_{j-1}^{(i)}\Big) \end{aligned} \tag{5.12}$$

We will represent RK methods using Butcher tableaus. Butcher tableaus are given in the well-known general format;

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} . \tag{5.13}$$

5.3.7 Definition. (Discrete Total Variation (TV) and Total Variation Diminishing (TVD) Schemes) The discrete TV of the numerical solution, $\left\{U_j^n\right\}_j^n$ at the $n-$th time-step is defined as

$$TV(U^n) := \sum_{j=1}^{M} \left| U_{j+1}^n - U_j^n \right| . \tag{5.14}$$

A difference scheme is then said to possess the TVD property or to be TVD if

$$TV\left(U^{n+1}\right) \leq TV(U^n) . \tag{5.15}$$

SSP-RK schemes were developed in [SO88], [Shu88] and further explored in [GS98]. Originally, they were called TVD time discretisations [Shu02]. In the context of non-linear hyperbolic conservation laws, which are known for developing discontinuities even for smooth initial conditions, SSP-RK methods possess an advantage over classical RK methods. In particular, as demonstrated in [GS98, §2], even if the chosen discretisation for the spatial derivative is free of spurious oscillations, if the RK method is non-TVD, spurious oscillations can still occur, which highlights the need for TVD time discretisations.

SSP-RK methods arose as the natural extension of the TVD property to high order time-discretisation. Suppose that the spatial discretisation (see subsequent sections) is such that, with a suitable CFL condition on the temporal step, the resulting forward Euler discretisation possesses the TVD property (see Defn. 5.3.7). Then, under suitable conditions for the coefficients in (5.12), the RK method can have the TVD property, with a potentially different time step restriction (see [Shu02, Lem. 2.1]).

5.3.8 Remark. We also make use of Strong-Stability Preserving Runge-Kutta (SSPRK) methods and in particular of the one given in the following Butcher tableau

(see [GST01, Pro. 4.1]):

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1/2 & 1/4 & 1/4 & 0 \\ \hline & 1/6 & 1/6 & 2/3 \end{array} \tag{5.16}$$

5.3.9 WENO Schemes

In this section we briefly summarise the details behind Weighted Essentially Non-Oscillatory (WENO) schemes, (cf. [JS96], [Shu98]), which pertain to our implementation. ENO/WENO schemes have been used with great success in several areas and in particular in the discretisation of hyperbolic conservation laws and convecion-diffusion equations (see e.g. [Shu20] for an overview).

They have several features which contributed to their widespread success. These include a mechanism of computational stencil construction that can achieve arbitrarily high accuracy in regions where the solution is smooth, essentially non-oscillatory behavior in the vicinity of discontinuities (no artificial, numerical overshoots and undershoots) and proven ability to simulate complex smooth solution structures (see [Shu98] for more details). In the present work we only consider WENO schemes.

There are two, closely related procedures that we use in the context of WENO schemes: reconstruction and approximation. The reconstruction procedure is used to form $\boldsymbol{F}$ in the spatial discretisation of (5.5) while the approximation procedure is used to obtain the spatial component of the post-processor $\widehat{U}$, such as, for example, in Lemma 4.3.2 (compare with conditions (4.41)).

5.3.10 Remark. In order to avoid cofusion, it is emphasized that the ENO/WENO reconstruction procedure, as that is defined in [Shu98], refers to the procedure used to formulate the ENO/WENO numerical scheme. The reconstruction procedure we have developed refers to the procedure used to obtain the post-processor which is subsequently used for a posteriori error computations.

The readers should note that this procedure, as well as the characteristics of the resulting reconstruction, are built using information from the scheme and are therefore inherently linked with it.

5.3.11 Definition (WENO reconstruction problem from [Shu98])**.** Given the cell

averages of a function $v(x)$:

$$\bar{v}_j := \frac{1}{h_j}\int_{x_j}^{x_{j+1}} v(\xi)\,\mathrm{d}\xi, \quad j = 0, \ldots, M-1, \tag{5.17}$$

find a polynomial $p_j(x)$ of degree at most $k-1$ for each cell I_j such that it is a $k-th$ order accurate approximation to $v(x)$ inside I_j:

$$p_j(x) = v(x) + \mathcal{O}\big(h^k\big), \quad x \in I_j, \quad j = 0, \ldots, M-1. \tag{5.18}$$

The procedure for obtaining the polynomial p_j from $\bar{v}_j$ can be found in [Shu98, Procedure 2.2] and in [Shu20, §.2.2]. This procedure is used both for finite volumes and for finite differences with a minor difference. In the context of finite volumes, we use the cell averages $\bar{v}_j$ to obtain a high order approximation to v. We then substitute this in an expression for $\boldsymbol{F}$ to calculate the numerical flux. In contrast, in the context of finite differences, the computational variables are point values rather than cell averages. In this case, the values $\boldsymbol{f}(\boldsymbol{U}_j)$, $j = 1, \ldots, M$, are used to obtain a high-order accurate approximation to $\boldsymbol{F}$ and subsequently $\partial_x \boldsymbol{f}$ in (5.4) and (5.5). This is the approach we will be using throughout this work.

We present this procedure below for a uniform scheme for simplicity. The reader should note that in the shallow water numerical experiment in §4.6.5, the WENO scheme used is derived for a non-uniform scheme over an adaptive grid. As a result, all geometry-related quantities - such as the sub-stencil polynomials, the smoothness indicators and the resulting weights - are no longer pre-computable constants. Instead, they have to be re-computed at every time-step. The reader can find the detailed procedure on how to derive the scheme on a non-uniform grid in [Shu98].

5.3.12 Remark (Order of WENO schemes on non-uniform grids)**.** The reader is advised that, as is pointed out in [Shu98], WENO schemes for finite differences which are posed on non-uniform grids cannot be higher than order two.

5.3.13 The WENO-3 scheme

In one spatial dimension, the WENO reconstruction procedure for the third order finite difference WENO scheme is used to approximate f_{j_x} on the cell $I_j := [x_j, x_{j+1}]$. The cell I_j is chosen to be the central cell of the computational stencil

$S := \{I_{j-1}, I_j, I_{j+1}\}$. The approximation is then obtained as a convex combination of polynomials over two sub-stencils of S, namely

$$\begin{aligned} S_1 &:= \{I_{j-1}, I_j\} \quad \text{and} \\ S_2 &:= \{I_j, I_{j+1}\}\,. \end{aligned} \tag{5.19}$$

The polynomials $p_1(x)$and $p_2(x)$ are the ENO reconstructions of $\boldsymbol{f}(u)$ on the substencils S_1 and S_2 respectively. The numerical flux at x_j, denoted by F_j, is obtained as the combination

$$F_j := w_1 p_1(x_j) + w_2 p_2(x_j)\,, \tag{5.20}$$

where w_1 and w_2 are the non-linear weights (see (5.33)) corresponding to $S_1$and S_2, which have to satisfy the conditions

$$w_l \geq 0, \quad \sum_{l=1}^{2} w_l = 1. \tag{5.21}$$

Finally, the WENO approximation to the flux derivative is obtained using

$$\partial_x f_j \approx \frac{1}{h_{j-1}}(F_j - F_{j-1})\,. \tag{5.22}$$

5.3.14 Remark. For exposition, we consider the case of scalar f to highlight the main ideas.

5.3.15 Remark (Flux-splitting)**.** In problems where WENO schemes are implemented using finite differences one should ensure upwinding and stability (see [Shu98]). This can be achieved in various ways. In this paper, this was done by applying the chosen finite difference method to a *flux-splitting* $f^{\pm}(u)$ of $f(u)$. In particular,

$$f(u) = f^{+}(u) + f^{-}(u)\,. \tag{5.23}$$

where

$$\frac{d}{du} f^{+}(u) \geq 0 \quad \text{and} \quad \frac{d}{du} f^{-}(u) \leq 0. \tag{5.24}$$

A simple flux-splitting is the Lax-Friedrichs splitting, which is given by

$$f^{\pm}(u) := \frac{1}{2}(f(u) \pm \alpha u)\,, \tag{5.25}$$

For 1D scalar conservation laws

$$\alpha := \max_u |f'(u)|\,. \tag{5.26}$$

For hyperbolic systems of conservation laws, f' is a Jacobian (with real eigenvalues). In this case upwinding is slightly more involved. The reader should note that in that case one can either perform a characteristic decomposition (which is the more robust approach) or use flux-splitting on a component-by-component basis. We opt for the latter route as it is sufficient for the purposes of this study. In this case, α is calculated using the eigenvalues, λ_i, of the Jacobian f':

$$\alpha := \max_u \max_i |\lambda_i(u)| \,. \tag{5.27}$$

The reader should also note that when coding WENO schemes, the stencil used for $f^+(u)$ is biased to the left, while the stencil for $f^-(u)$ is biased one point to the right.

We will demonstrate the process for obtaining f^+ as an example. The ENO sub-stencil polynomials for f^+, for a third order WENO scheme are given by

$$\begin{aligned} p_1(U) &= \frac{1}{2}(-f(U_{j-1}) + 3f(U_j)) \,, \\ p_2(U) &= \frac{1}{2}(f(U_j) + f(U_{j+1})) \end{aligned} \tag{5.28}$$

Next, we construct the weights w_1 and w_2. Suppose we wanted to create a reconstruction for a function $v(x)$, which is piecewise smooth in sub-stencils S_1 and S_2. There are constants d_r, $r = 1, 2$ such that

$$v_{j+1} = d_1 v_{j+1}^{(1)} + d_2 v_{j+1}^{(2)} = v(x_{j+1}) + \mathcal{O}\big(h^{2k-1}\big) \,, \tag{5.29}$$

where $v_{j+1}^{(i)}$ is the reconstruction of $v(x)$ in the sub-stencil S_i evaluated at x_{j+1}. More specifically, for f^+,

$$d_1 = \frac{1}{3}, \quad d_2 = \frac{2}{3}. \tag{5.30}$$

If $v(x)$ is smooth, the nonlinear weights w_i should be very close to d_i. If, instead, $v(x)$ has a discontinuity in some stencil, the w_i from that stencil should be close to zero to avoid spurious oscillatory behaviour. This is accomplished by using smoothness indicators β_i, where

$$\beta_i := \sum_{l=1}^{k-1} \int_{x_j}^{x_{j+1}} h^{2l-1} \left(\frac{\partial^l p_i(x)}{\partial x^l} \right)^2 \mathrm{d}x. \tag{5.31}$$

This is simply a sum of scaled $\mathrm{L}^2(\Omega)$ norms of the derivatives of p_i. The factor h^{2l-1} ensures that β_i scales like an $\mathrm{L}^2(\Omega)$ −norm over polynomials. In the case of the third

order WENO scheme over a uniform grid,

$$\beta_1 = (v_{j-1} - v_j)^2, \quad \beta_2 = (v_j - v_{j+1})^2. \tag{5.32}$$

We can now obtain the weights w_i, which are given by

$$w_i := \frac{\alpha_i}{\sum_{s=0}^{k-1} \alpha_s}, \text{ with } \alpha_i = \frac{d_i}{(\epsilon + \beta_i)^2}. \tag{5.33}$$

The constant $\epsilon \ll 1$ is a small constant to ensure the denominator does not vanish. In experiments we use $\epsilon = 10^{-6}$. We repeat this process to obtain f^- noting that in this case the entire computational stencil is biased one position to the right.

5.3.16 Remark (Choice of nonlinear weights)**.** The choice of nonlinear weights is very important. As is demonstrated in [JS96], an appropriate choice of nonlinear weights can upgrade the order of accuracy of (5.20) in smooth regions relative to an ENO scheme with a stencil S_1 or S_2. Furthermore, because these weights are designed to reflect the smoothness of the reconstruction polynomial in the relevant stencil, they are also used to facilitate the non-oscillatory property of the WENO scheme.

The Smoothness Increasing Accuracy Preserving (SIAC) filtering is a comparable concept. This is a post-processing technique which has been used to reduce error oscillations and recover smoothness in the solution and its derivatives in the context of the Discontinuous Galerkin method (see [DGPR19]). Note that SIAC is complicated to implement in multiple spatial dimensions. However, the WENO scheme is relatively simple.

5.3.17 WENO approximation

In this section we present the procedure we will use in later chapters for obtaining the spatial component of the reconstruction. This is based on the WENO interpolation procedure of [JSB+19]. An advantage of this interpolant is that all its aspects (sub-stencil polynomials, linear and non-linear weights) have been modified for use on non-uniform grids. In addition, it has the other advantages of WENO interpolants. These include high orders of approximation in region where the solution is smooth and essentially non-oscillatory behaviour in the vicinity of discontinuities.

Consider a function $y(x)$ with a set of point values $\{y_j\}$ at locations $\{x_j\}$, where the grid is not necessarily uniform. We want to construct a third order WENO interpolating polynomial in an interval $[x_j, x_{j+1}]$ by using the 4-point stencil

$$S := \{x_{j-1}, \ldots, x_{j+2}\} \tag{5.34}$$

The interpolant is obtained as a convex combination of polynomials which are constructed on two 3-point sub-stencils, S_1 and S_2 of S, which are given by

$$\begin{aligned} S_1 &:= \{x_{j-1}, x_j, x_{j+1}\}, \\ S_2 &:= \{x_j, x_{j+1}, x_{j+2}\}. \end{aligned} \tag{5.35}$$

The polynomials are Lagrange interpolants over the sub-stencils:

$$p_1(x) := y_{j-1}\frac{(x-x_j)(x-x_{j+1})}{h_{j-1}(h_{j-1}+h_j)} + y_j\frac{(x-x_{j-1})(x-x_{j+1})}{h_{j-1}h_j} + y_{j+1}\frac{(x-x_{j-1})(x-x_j)}{(h_{j-1}+h_j)\,h_j}$$

and

$$p_2(x) := y_j\frac{(x-x_{j+1})(x-x_{j+2})}{h_j(h_j+h_{j+1})} + y_{j+1}\frac{(x-x_j)(x-x_{j+2})}{h_j h_{j+1}} + y_{j+2}\frac{(x-x_j)(x-x_{j+1})}{(h_j+h_{j+1})\,h_{j+1}} \tag{5.36}$$

for $x \in [x_j, x_{j+1}]$. A polynomial approximation to $u(x)$, $p(x)$, can be obtained as a convex combination of the $p^{(i)}$. The WENO approach is such that $p(x)$ is a high order approximation in intervals where $u(x)$ is smooth. $p(x)$ is obtained as a weighted sum of $p^{(1)}$ with the (linear) weights γ_1 and γ_2, each corresponding to a sub-stencil of the large stencil:

$$\begin{aligned} \gamma_1(x) &:= -\frac{x-x_{j+2}}{x_{j+2}-x_{j-1}} \quad \text{and} \\ \gamma_2(x) &:= \frac{x-x_{j-1}}{x_{j+2}-x_{j-1}}. \end{aligned} \tag{5.37}$$

The linear weights are positive and satisfy

$$\sum_i \gamma_i = 1. \tag{5.38}$$

Interested readers can find details on the construction of these weights in ([CFR05]) and [LSZ09]. If the solution is discontinuous inside a sub-stencil, we would like that stencil to have little contribution to ensure the non-oscillatory behaviour of the scheme. This is achieved by using the non-linear weights $\omega_i(x)$, which are obtained from the $\gamma_i(x)$ as follows:

$$\omega_j(x) := \frac{\alpha_j(x)}{\sum_{i=1}^{2}\alpha_i(x)}, \quad \alpha_i(x) := \frac{\gamma_i(x)}{\epsilon+\beta_i}, \tag{5.39}$$

where the β_i are the *smoothness indicators* for the sub-stencil to which they pertain. They are an indication of how non-smooth the solution is in the corresponding sub-stencil. If the solution is smooth in the sub-stencil S_j, then the relevant β_j is small and the relevant ω_j is close to the γ_j in S_j. If instead the solution has a discontinuity in S_j, then the β_j is large, leading to a small ω_j and ensuring the non-oscillatory behaviour.

The β_i which are used in this paper are given in [JSB+19] and are defined as

$$\begin{aligned}\beta_1 :=& (h_j + h_{j+1})^2 \left(\frac{\left|y'_{j+1} - y'_j\right|}{h_j} - \frac{\left|y'_j - y'_{j-1}\right|}{h_{j-1}} \right)^2 \quad \text{and} \\ \beta_2 :=& (h_{j-1} + h_j)^2 \left(\frac{\left|y'_{j+2} - y'_{j+1}\right|}{h_{j+1}} - \frac{\left|y'_{j+1} - y'_j\right|}{h_j} \right)^2 .\end{aligned} \tag{5.40}$$

The calculation of the y'_i is presented in detail in [JSB+19, §3.3.2]. Finally, the WENO approximation to $u(x)$ in the interval $[x_j, x_{j+1}]$ based on the stencil $S = S_1 \cup S_2 = \{x_{j-1}, x_j, x_{j+1}, x_{j+2}\}$ can be obtained as

$$p(x) := \omega_1 p_1(x) + \omega_2 p_2(x) \,. \tag{5.41}$$

5.3.18 Remark (Boundary conditions)**.** We implement periodic boundary conditions by identifying the points $j = 0$ and $j = M$ as the same point. In the case of non-periodic boundary conditions - Neumann boundary conditions in particular - the computational domain is extended using an appropriate number of ghost nodes. For WENO schemes the number of ghost nodes depends on the WENO sub-stencil width. The values for the ghost nodes can be obtained by extrapolation based on the local WENO polynomials.

5.4 Numerical tests

In this section we conduct numerical tests to investigate the convergence of the spatial component of the WENO interpolant. In order to facilitate the comparison we define an alternative error interpretation in the same fashion as in previous sections, i.e. by evaluating the error between a function and an appropriate reconstruction. In this case, we consider the WENO interpolant which is defined in (5.41).

In the numerical experiments in this section, the sequence of approximations is carried out on grids with discretisation parameter $h = 2^{-m}$, $m = 8, \ldots, 19$. We will

examine the EOC of the $\mathrm{L}^2(\Omega)$ −norm of the error for the different test cases we specify below.

5.4.1 Test 1: Sinusoidal function

In this test we use the WENO interpolant to interpolate the sinusoidal function

$$u(x) = \sin(40\pi x) \quad \text{for} \quad x \in [0,1]\,. \tag{5.42}$$

The results are shown in Fig.5.1. The $\mathrm{L}^2(\Omega)$ −error converges with order 3.5.

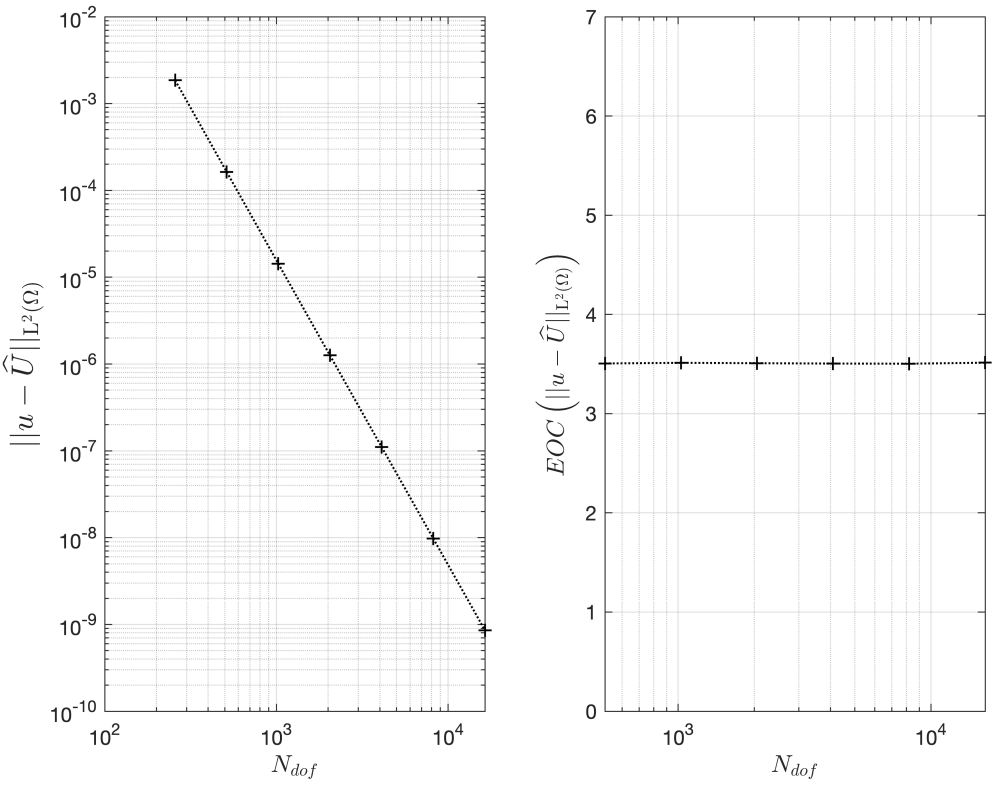


Fig. 5.1. In this plot we examine the asymptotic convergence rate for the $\mathrm{L}^2(\Omega)$ − error of the spatial WENO interpolant (5.41) for a smooth, sinusoidal function given by (5.42).

5.4.2 Test 2: Hat function

In this test we use the WENO interpolant to interpolate the hat function

$$u(x) = \begin{cases} x & x \in [0, 0.25)\,, \\ \frac{1}{2} - x & x \in [0.25, 0.75)\,, \\ x - 1 & x \in [0.75, 1]\,. \end{cases} \tag{5.43}$$

The results are shown in Fig. 5.2. Notice that, since $u \notin \mathrm{C}^1(\Omega)$, the convergence rates drop significantly to order 1.5.

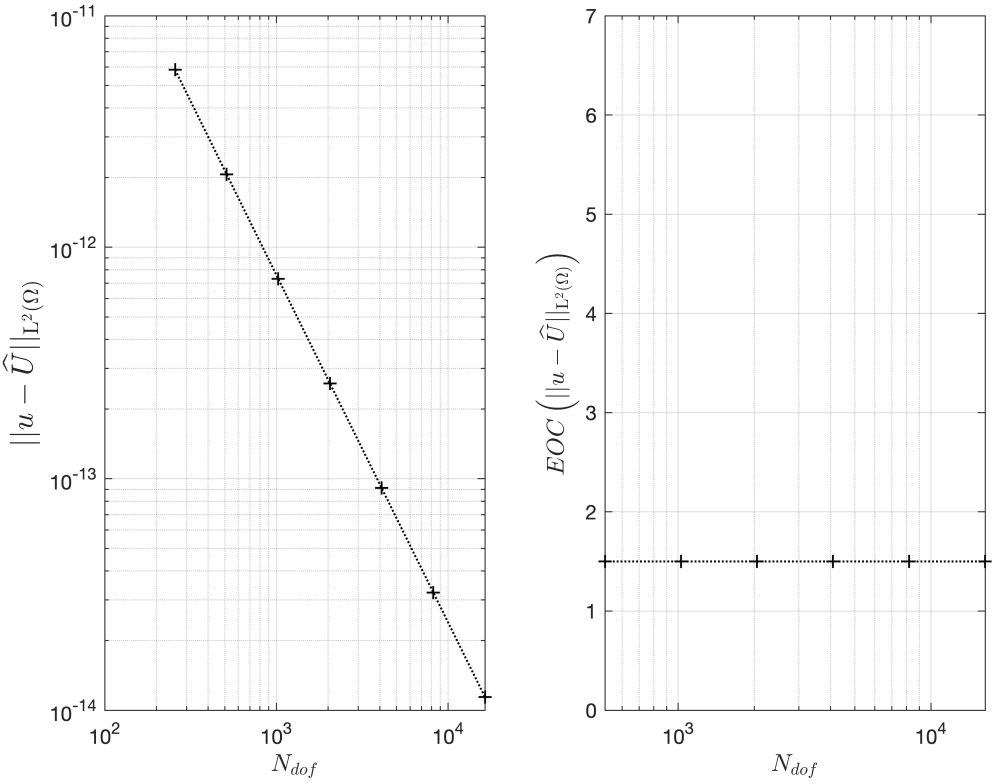


Fig. 5.2. In this plot we examine the asymptotic convergence rate for the $\mathrm{L}^2(\Omega)-$ error of the spatial WENO interpolant (5.41) for a sawtooth function given by (5.43). In this case, we note the high approximability offered by the WENO interpolant, evidenced by the fact that, at the coarsest discretisation level, the error is already of the order of 10^{-11}.

5.4.3 Test 3: Step function

In this test we use the WENO interpolant to interpolate the step function

$$u(x) = \begin{cases} 1 & \text{for} \quad \left| x - \frac{1}{2} \right| \leq \frac{1}{4} \\ 0 & \text{otherwise.} \end{cases} \tag{5.44}$$

The results are shown in Fig. 5.3. Notice that in this case the convergence rate drops to order 0.5.

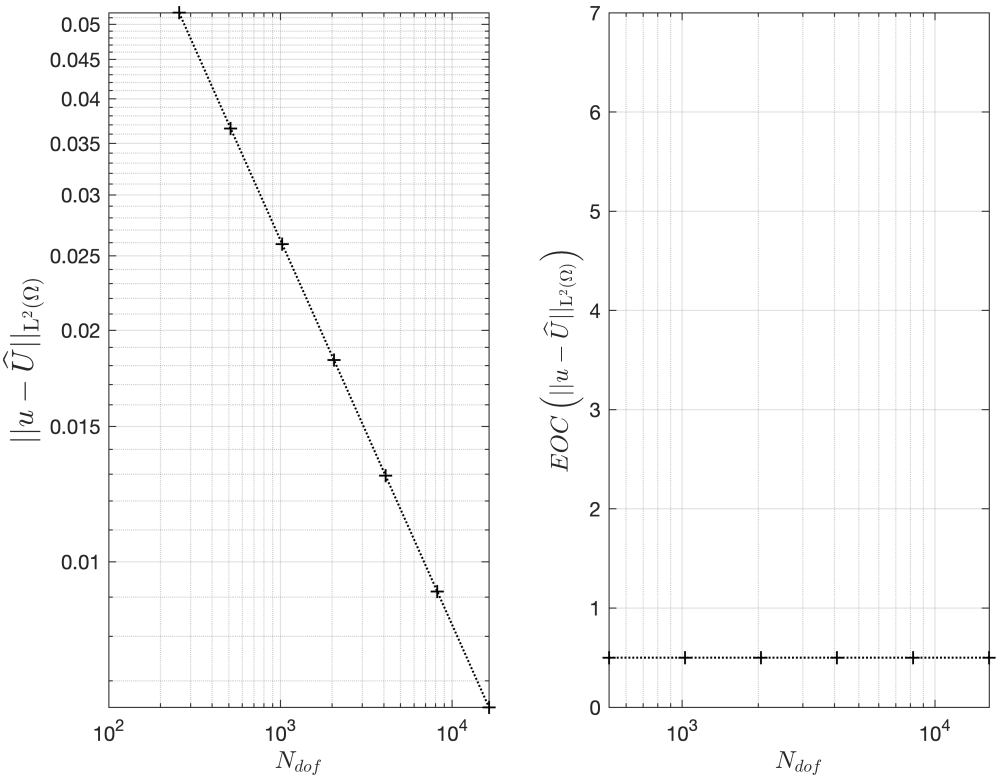


Fig. 5.3. In this plot we examine the asymptotic convergence rate for the $\mathrm{L}^2(\Omega)-$ error of the spatial WENO interpolant (5.41) for a sawtooth function given by (5.44).

5.5 Conclusion

In this section we presented the WENO interpolation procedure which allows us to post-process the FD solution in the temporal and spatial components, for general problems. We concluded the chapter with numerical experiments where we demonstrated the high approximability offered by the WENO interpolant for functions of varying regularity.

Chapter 6

Automated error control for linear hyperbolic systems

Abstract

In this chapter, we extend the results of Chapter 4 to general linear systems. We are able to show a posteriori error control for general classes of schemes, including the WENO schemes from Chapter 5. We validate these results numerically and examine the impact of low regularity entropy solutions on the robustness of the bounds.

6.1 Introduction

In this chapter we consider linear systems of conservation laws with symmetric coefficient matrices in one spatial dimension. We derive a simple a posteriori error bound based on the stability framework of the PDE and use the reconstruction procedure to compute it.

We illustrate the reconstruction procedure using the one-dimensional wave equation in system form as a model problem. We discretise this using the Leapfrog scheme on staggered grids. The performance of the a posteriori bound in this way is evaluated on the basis of optimality.

6.1.1 Motivation

A linear system allows us to demonstrate the framework of the reconstruction and its utility in obtaining a posteriori error bounds. It is also a convenient intermediate

stage for extending the results presented so far to systems of non-linear conservation laws. We distinguish between linear and non-linear cases. We address linear systems in this chapter and defer non-linear ones to the following chapter. In this regard, we focus our attention to symmetric hyperbolic problems in this chapter. We derive an appropriate a posteriori bound based on the stability framework of the problem. The procedure for obtaining the bound is, as in previous chapters, independent from the numerical discretisation used to obtain the approximation - in this case the Leapfrog scheme.

6.1.2 Chapter contribution

In this chapter we derive a residual based a posteriori error estimate for symmetric linear systems of hyperbolic conservation laws. This estimate is based on the stability framework of the PDE and is not dependent on the chosen numerical discretisation method. We then use the reconstruction framework introduced in the previous chapter to compute the quantities involved in this estimate. We demonstrate the use of the framework using as a model problem the one dimensional wave equation in the form of a first order system of conservation laws. The system is discretised using a Leapfrog scheme on grids which are staggered in both time and space. We demonstrate numerically that the obtained estimate is optimal in the example we consider, where we use a smooth initial condition. The rest of the chapter is structured as follows. In §6.2 we set up the preliminaries we require for the rest of the chapter. In §6.3 we introduce the model problem we will be using. In §6.4 we present an a posteriori bound for symmetric linear systems of linear conservation laws. In §6.5 we present the Leapfrog discretisation for our problem and the reconstruction we will use to evaluate the bound. In §6.6 we present the numerical experiments we ran to benchmark the bounds behaviour. We conclude the chapter in §6.7.

6.2 Preliminaries and Problem Setup

In this section we present the model problem we will be using throughout the rest of the chapter, along with relevant notation. We will specify the structure of the linear

symmetric hyperbolic system we consider in this chapter. We will use the latter in obtaining the a posteriori bound in subsequent sections. We include some results which pertain to solutions of systems of conservation laws, such as admissibility criteria. Despite the fact that we do not make use of these in the present chapter, the a posteriori estimate makes reference to entropy solutions. As such, we will introduce the material we need in order to present entropy solutions to conservation laws, despite the fact that we do not use the entropy framework yet.

6.2.1 Remark. (Convention for matrix derivatives) Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $\boldsymbol{u} := \boldsymbol{u}(\boldsymbol{x}) : \mathbb{R}^m \to \mathbb{R}^n$ such that $\boldsymbol{u} \in \mathrm{C}^1(\mathbb{R}^m; \mathbb{R}^n)$. We adopt the following convention for matrix differentiation

$$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} := \begin{bmatrix} \frac{\partial u_1}{\partial x_1} & \cdots & \frac{\partial u_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial u_m}{\partial x_1} & \cdots & \frac{\partial u_m}{\partial x_n} \end{bmatrix}. \tag{6.1}$$

6.2.2 Remark. (Derivative of a field) Derivatives of a field, q, are denoted $\mathrm{D}q := (\partial_{u_1} q(\mathbf{u}), \ldots, \partial_{u_d} q(\mathbf{u}))$. The matrix of second derivatives is

$$\mathrm{D}^2 q(\boldsymbol{u}) := \begin{bmatrix} \partial_{u_1,u_1} q(\boldsymbol{u}) & \ldots & \partial_{u_1,u_d} q(\boldsymbol{u}) \\ \vdots & \ddots & \vdots \\ \partial_{u_d,u_1} q(\boldsymbol{u}) & \ldots & \partial_{u_d,u_d} q(\boldsymbol{u}) \end{bmatrix}. \tag{6.2}$$

6.2.3 Definition (Linear, one-dimensional system of first order PDEs)**.** Let $T \in \mathbb{R}^+$, $t \in (0, T]$ and let $\Omega := [0, 1]$. We consider the linear system given by

$$\begin{aligned} \frac{\partial \boldsymbol{u}}{\partial t} + \frac{\partial}{\partial x}(A(x,t)\,\boldsymbol{u}) = 0 \quad \text{in} \quad \Omega \times (0, T]\,, \\ \boldsymbol{u}(x, 0) = \boldsymbol{u}_0(x) \quad \text{in} \quad \Omega \times \{0\} \end{aligned} \tag{6.3}$$

in the unknown $\boldsymbol{u} : \Omega \times [0, T] \to \mathbb{R}^m$ with $\boldsymbol{u} := (u_1, \ldots, u_m)^T$, $A \in \mathrm{C}^1(\Omega \times [0, T]\,; \mathbb{R}^{m \times m})$ and $\boldsymbol{u}_0 \in \mathrm{C}^1(\Omega; \mathbb{R}^{m \times m})$ a given initial condition. Throughout this chapter we will use homogeneous Dirichlet boundary conditions: $\boldsymbol{u}(0, t) = \boldsymbol{u}(1, t) = 0$.

6.2.4 Definition (Hyperbolic system)**.** The system of equations given in (6.3) is called hyperbolic if the matrix $A(x, t)$ is diagonalizable for each $x \in \mathbb{R}$ and $t \geq 0$, i.e. if it has m real eigenvalues and the corresponding eigenvectors $\{\boldsymbol{r}_k(x, t)\}_{k=1}^m$ form a basis of $\mathbb{R}^m$. Furthermore, the system is called strictly hyperbolic if the eigenvalues of $A(x, t)$ are not only real but also distinct for all (x, t).

6.2.5 Remark. Examples of equations which are/can be written in the form (6.3) or (6.4) include Maxwell's equations of electrodynamics and the linearized Euler equations. In this chapter we will use the one-dimensional wave equation as an example for demonstrating our framework (see Ex. 6.3.2).

6.3 Model Problem

6.3.1 Definition. (Linear, symmetric hyperbolic system with constant coefficients) Let $\boldsymbol{u} := (u_1, \ldots, u_m)^T : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times m}$ be a constant, symmetric matrix. Then, the linear system in (6.3) reduces to

$$\frac{\partial \boldsymbol{u}}{\partial t} + A \frac{\partial \boldsymbol{u}}{\partial x} = 0. \tag{6.4}$$

6.3.2 Example (The wave equation)**.** A simple example of (6.4) with symmetric A is the wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}. \tag{6.5}$$

It can be easily shown that (6.5) can be written as a first-order system of conservation laws by introducing a variable v which is related to u as follows:

$$\begin{aligned} u_t + v_x &= 0 \\ v_t + u_x &= 0, \end{aligned} \tag{6.6}$$

with appropriate boundary and initial conditions. In this case, $\boldsymbol{u} = (u, v)^T$ and A is given as

$$A := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{6.7}$$

6.3.3 Remark. In general, conservation laws do not admit classical solutions: their solutions can develop discontinuities in finite time, even if $\boldsymbol{u}_0$ is smooth. This motivates the use of weak solutions. Weak solutions are particularly important with regard to (5.1) when $\boldsymbol{f}$ is non-linear. We will use them to motivate entropy solutions, which we will in turn refer to in the a posteriori error result we present in this chapter. Hence, we will introduce the weak solution results and definitions we need here.

6.3.4 Definition. (Weak derivative) Let X denote a Banach space and let $\boldsymbol{u} \in \mathrm{L}^1(0,T;X)$. We say that $\boldsymbol{v} \in \mathrm{L}^1(\llbracket 0,T \rrbracket;X)$ is the weak derivative of $\boldsymbol{u}$, denoted by

$$\boldsymbol{u}' = \boldsymbol{v} \tag{6.8}$$

if $\forall \phi \in \mathrm{C}_c^\infty(0,T)$

$$\int_0^T \phi'(t)\,\boldsymbol{u}(t)\,\mathrm{d}t = -\int_0^T \phi(t)\,\boldsymbol{v}(t)\,\mathrm{d}t. \tag{6.9}$$

6.3.5 Definition. (Weak solution for (6.4) [GR13]) Let $\boldsymbol{u}_0(x) \in \mathrm{L}^\infty(\Omega;\mathbb{R}^m)$. Then, we say that the function $\boldsymbol{u}(x,t) \in \mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega;\mathbb{R}^m))$ is said to be a weak solution to (6.4) if $\boldsymbol{u}(x,0) = \boldsymbol{u}_0(x)$ for $x \in \Omega$ and the following equality holds for all $\boldsymbol{\phi} \in \mathrm{C}^1\big(0,T;\mathrm{C}^1(\Omega;\mathbb{R}^m)\big)$

$$\int_0^\infty \int_{\mathbb{R}} \boldsymbol{u} \cdot \frac{\partial \boldsymbol{\phi}}{\partial t} + A\boldsymbol{u} \cdot \frac{\partial \boldsymbol{\phi}}{\partial x} \mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} \boldsymbol{u}_0(x) \cdot \boldsymbol{\phi}(x,0)\,\mathrm{d}x = 0. \tag{6.10}$$

6.3.6 Remark. Weak solutions to (5.1) are not necessarily unique. For this reason admissibility criteria are used to remove inappropriate solutions (e.g. physically irrelevant ones) from consideration. There are multiple admissibility criteria, such as the Rankine-Hugoniot condition, the Lax shock condition and the entropy conditions to name but a few. Entropy is more relevant to our case.

6.3.7 Definition (Entropy/entropy-flux pair)**.** The pair (η, q) is an entropy/entropy-flux pair associated with the conservation law (5.1) iff η is convex and

$$\mathrm{D}q = \mathrm{D}\eta \mathrm{D}f. \tag{6.11}$$

6.3.8 Definition (Entropy solution)**.** A function $\boldsymbol{u} \in \mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega;\mathbb{R}^m))$ is an entropy solution of (5.1) with an associated entropy/entropy-flux pair (η, q) if

$$\int_0^\infty \int_\Omega \boldsymbol{u} \cdot \partial_t \phi + \boldsymbol{f}(\boldsymbol{u}) \cdot \partial_x \phi \,\mathrm{d}x\mathrm{d}t + \int_\Omega \boldsymbol{u}_0 \cdot \phi(\cdot,0)\,\mathrm{d}x = 0 \quad \forall \phi \in \mathrm{C}^1\big(0,T;\mathrm{C}^1(\Omega;\mathbb{R}^m)\big)$$

and

$$\int_0^\infty \int_\Omega \eta(\boldsymbol{u})\,\partial_t \phi + q(\boldsymbol{u})\,\partial_x \phi \,\mathrm{d}x\mathrm{d}t + \int_\Omega \eta(\boldsymbol{u}_0)\,\phi(\cdot,0)\,\mathrm{d}x \geq 0 \quad \forall \phi \in \mathrm{C}^1\big(0,T;\mathrm{C}^1(\Omega;\mathbb{R}^m)\big) \tag{6.12}$$

It can be verified that strong solutions of (5.1) also satisfy the additional conservation law

$$\partial_t \eta(\boldsymbol{u}) + \partial_x q(\boldsymbol{u}) = 0. \tag{6.13}$$

We have now defined an entropy solution for (6.4). The rest of this section includes a series of results showing that an entropy solution, as is defined in Defn. 6.3.8, exists and it is unique for the problem under consideration.

6.3.9 Vanishing Viscosity method

An important approach for obtaining a unique solution to (6.4), is the vanishing viscosity method. We follow the exposition of [Eva10] to briefly present the method here.

Firstly, (6.4) is approximated by the following problem through the addition of artificial viscosity:

$$\begin{aligned} \frac{\partial \boldsymbol{u}^\epsilon}{\partial t} + A\frac{\partial \boldsymbol{u}^\epsilon}{\partial x} &= \epsilon \frac{\partial^2 \boldsymbol{u}^\epsilon}{\partial x^2} \quad \text{in } \Omega \times (0,T] \\ \boldsymbol{u}^\epsilon(x,0) &= \boldsymbol{u}_0^\epsilon(x) \quad \text{on } \Omega \times \{t=0\}, \end{aligned} \tag{6.14}$$

where $0 < \epsilon \leq 1$ and $\boldsymbol{u}_0^\epsilon$ is the mollification of $\boldsymbol{u}_0$, obtained as shown in Defn. A.0.4. The approximate problem (6.14) has a unique solution $\boldsymbol{u}^\epsilon$ for each $\epsilon > 0$, which converges to 0 as $|x| \to \infty$. The purpose of this approximation is to obtain a solution $\boldsymbol{u}$ of (6.4) as the limit of a sequence of solutions $\boldsymbol{u}^\epsilon$ of the approximating problem (6.14) as $\epsilon \to 0$. This is encapsulated in Theorem 6.3.10.

6.3.10 Theorem. *(Existence of approximate solutions to (6.14)) For each $\epsilon > 0$, there exists a unique solution $\boldsymbol{u}^\epsilon$ of (6.14) with*

$$\boldsymbol{u}^\epsilon \in \mathrm{L}^2\big(0,T;\mathrm{H}^3(\mathbb{R};\mathbb{R}^m)\big), \quad \dot{\boldsymbol{u}}^\epsilon \in \mathrm{L}^2\big(0,T;\mathrm{H}^1(\mathbb{R};\mathbb{R}^m)\big). \tag{6.15}$$

Proof. See [Eva10, §7.3: Theorem 1]. □

Our intention is to obtain solutions $\boldsymbol{u}$ of (4.2) as limits of the solutions $\boldsymbol{u}^\epsilon$ of the parabolic problem (6.14) as the coefficient of the viscous term, ϵ, goes to 0. In that regard, the next two results are helpful.

6.3.11 Theorem. *(Energy estimate) There exists a constant C, depending only on the spatial dimensions of Ω and the coefficients, such that*

$$\max_{0\leq t\leq T}\Big(\|\boldsymbol{u}^\epsilon(t)\|_{\mathrm{H}^1(\mathbb{R};\mathbb{R}^m)} + \|\dot{\boldsymbol{u}}^\epsilon(t)\|_{\mathrm{L}^2(\mathbb{R};\mathbb{R}^m)}\Big) \leq C\,\|\boldsymbol{u}_0\|_{\mathrm{H}^1(\mathbb{R};\mathbb{R}^m)} \tag{6.16}$$

for each $0 < \epsilon \leq 1$.

Proof. See [Eva10, §7.3: Thm 2] □

6.3.12 Remark. The significance of Thm. 6.3.11 is that it provides a bound for the sequence $\boldsymbol{u}^\epsilon$, $0 < \epsilon \leq 1$ as $\epsilon \to 0$. This bound is utilized to obtain a (weakly)

convergent subsequence, $\boldsymbol{u}^{\epsilon_k}$, of $\boldsymbol{u}^{\epsilon}$. More specifically, there exists a subsequence $\epsilon_k \to 0$ and a function $\boldsymbol{u} \in \mathrm{L}^2\big(0,T;\mathrm{H}^1(\mathbb{R};\mathbb{R}^m)\big)$ such that the weak derivative $\dot{\boldsymbol{u}} \in \mathrm{L}^2\big(0,T;\mathrm{L}^2(\mathbb{R};\mathbb{R}^m)\big)$ such that

$$\begin{aligned} \boldsymbol{u}^{\epsilon_k} &\rightharpoonup \boldsymbol{u} \quad \text{weakly in } \mathrm{L}^2\big(0,T;\mathrm{H}^1(\mathbb{R};\mathbb{R}^m)\big) \\ \dot{\boldsymbol{u}}^{\epsilon_k} &\rightharpoonup \dot{\boldsymbol{u}} \quad \text{weakly in } \mathrm{L}^2\big(0,T;\mathrm{L}^2(\mathbb{R};\mathbb{R}^m)\big) \end{aligned} \tag{6.17}$$

and $\boldsymbol{u}(0) = \boldsymbol{u}_0$.

This is encapsulated in [Eva10, §7.3: Thm 3], which tells us that there does exist weak solution to (4.2). Furthermore, this solution is unique ([Eva10, §7.3: Thm 4]).

6.4 A posteriori error bound for a linear system

In this section we present two a posteriori error bounds for linear systems of hyperbolic conservation laws. In the first one we will assume that the matrix in one spatial dimension. This bound is the systems analogue to the scalar bound for linear advection we derived in Lemma 4.3.2.

The result in Lemma 6.4.3 serves as a gateway to more interesting test cases. We will start looking at such cases by assuming that $A(x,t)$ is non-constant, symmetric and differentiable.

6.4.1 Lemma (Stability and error control for a non-constant, symmetric linear system of equations)**.** *Let $A \in \mathrm{C}^1(\mathbb{R}\times[0,T]\,;\mathbb{R}^{m\times m})$ be a symmetric matrix and define $C := \|A_x\|_{\mathrm{L}^\infty}$. Also, let $\boldsymbol{u}$ be an entropy solution of the initial boundary value problem*

$$\begin{aligned} \boldsymbol{u}_t + A\boldsymbol{u}_x &= 0 \quad \text{in } \Omega\times(0,T] \\ \boldsymbol{u}(x,0) &= \boldsymbol{u}_0(x) \quad \text{in } \Omega\times\{0\} \end{aligned} \tag{6.18}$$

with periodic boundary conditions and suppose $\boldsymbol{v}$ is an entropy solution of the perturbed problem for some $\boldsymbol{R} \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega;\mathbb{R}^m))$

$$\begin{aligned} \boldsymbol{v}_t + A\boldsymbol{v}_x &= -\boldsymbol{R} \quad \text{in } \Omega\times(0,T] \\ \boldsymbol{v}(x,0) &= \boldsymbol{v}_0(x) \quad \text{in } \Omega\times\{0\}, \end{aligned} \tag{6.19}$$

also with periodic boundary conditions. Then, the error between the two functions, $\boldsymbol{e} := \boldsymbol{u} - \boldsymbol{v}$, satisfies the following bound for all $t \in [0,T]$:

$$\|\boldsymbol{e}(t)\|^2_{\mathrm{L}^2(\Omega)} \le \omega(t)\left[\|\boldsymbol{e}(0)\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \|\delta(s)\boldsymbol{R}(s)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s\right], \tag{6.20}$$

where

$$\omega(t) = \begin{cases} \exp(2C)\exp(t) & \text{for } t \leq 1 \\ t\exp(2C+1) & \text{for } t \geq 1. \end{cases} \tag{6.21}$$

and

$$\delta(s) = \begin{cases} 1 & \text{for } s \leq 1 \\ \sqrt{s} & \text{for } s \geq 1. \end{cases} \tag{6.22}$$

Proof. Subtracting (6.18) from (6.19) we have the following error equation for e

$$\begin{aligned} \boldsymbol{e}_t + A\boldsymbol{e}_x &= \boldsymbol{R} \quad \text{in } \Omega \times (0,T] \\ \boldsymbol{e}(x,0) &= (\boldsymbol{u}_0 - \boldsymbol{u}_0)(x) \quad \text{in } \Omega \times \{0\}. \end{aligned} \tag{6.23}$$

Testing (6.23) with $\boldsymbol{e}$ we see

$$\begin{aligned} \int_\Omega \boldsymbol{e}\cdot\boldsymbol{R} &= \int_\Omega \boldsymbol{e}\cdot\boldsymbol{e}_t + \boldsymbol{e}\cdot(A\boldsymbol{e}_x) \\ &= \int_\Omega \boldsymbol{e}\cdot\boldsymbol{e}_t + \frac{1}{2}((A\boldsymbol{e})\cdot\boldsymbol{e})_x - \frac{1}{2}((A_x\boldsymbol{e})\cdot\boldsymbol{e}), \end{aligned} \tag{6.24}$$

Then, since the domain is periodic, it follows that

$$\int_\Omega \boldsymbol{R}\cdot\boldsymbol{e} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{e}\|^2_{\mathrm{L}^2(\Omega)} - \frac{1}{2}\int_\Omega (A_x\boldsymbol{e})\cdot\boldsymbol{e}, \tag{6.25}$$

which is re-arranged to obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{e}\|^2_{\mathrm{L}^2(\Omega)} = \int_\Omega \boldsymbol{R}\cdot\boldsymbol{e} + \frac{1}{2}\int_\Omega (A_x\boldsymbol{e})\cdot\boldsymbol{e}, \tag{6.26}$$

Let us look at the two terms on the right seperately. Firstly, concerning the first term, we apply an identical argument as we did in the proof of Lemma 6.4.3. That is, we firstly apply the Cauchy-Schwarz inequality, followed by the Cauchy inequality to obtain

$$\int_\Omega (\delta\boldsymbol{R})\cdot\left(\delta^{-1}\boldsymbol{e}\right) \leq \|\delta\boldsymbol{R}\|_{\mathrm{L}^2(\Omega)}\left\|\delta^{-1}\boldsymbol{e}\right\|_{\mathrm{L}^2(\Omega)} \leq \frac{1}{2}\|\delta\boldsymbol{R}\|^2_{\mathrm{L}^2(\Omega)} + \frac{1}{2}\left\|\delta^{-1}\boldsymbol{e}\right\|^2_{\mathrm{L}^2(\Omega)}, \tag{6.27}$$

for any $\delta \in \mathrm{C}^0([0,T],\mathbb{R}^+)$. Next, we obtain an upper bound for the second term on the r.h.s. of (6.26) using the Cauchy-Schwarz inequality along with the fact that $A \in \mathrm{C}^1([0,T];\mathbb{R}^{m\times m})$ (see also [Eva10]):

$$\frac{1}{2}\int_\Omega (A_x\boldsymbol{e})\cdot\boldsymbol{e} \leq \left|\frac{1}{2}\int_\Omega (A_x\boldsymbol{e})\cdot\boldsymbol{e}\right| \leq C\|\boldsymbol{e}\|^2_{\mathrm{L}^2(\Omega)}, \tag{6.28}$$

where $C = \|A_x\|_{\mathrm{L}^\infty}$. Now we can combine this results to obtain an upper bound for (6.26) as follows

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{e}\|^2_{\mathrm{L}^2(\Omega)} \leq \frac{1}{2}\|\delta\boldsymbol{R}\|^2_{\mathrm{L}^2(\Omega)} + \frac{1}{2}\left\|\delta^{-1}\boldsymbol{e}\right\|^2_{\mathrm{L}^2(\Omega)} + C\|\boldsymbol{e}\|^2_{\mathrm{L}^2(\Omega)}. \tag{6.29}$$

Finally, we use Gronwall's inequality to realise the bound

$$\|\boldsymbol{e}(t)\|^2_{\mathrm{L}^2(\Omega)} \leq \exp\left(\int_0^t \delta(s)^{-2} + 2C\,\mathrm{d}s\right)\left[\|\boldsymbol{e}(0)\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \|\delta(s)\boldsymbol{R}(s)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s\right]. \tag{6.30}$$

Choosing

$$\delta(s) = \begin{cases} 1 & s \leq 1 \\ \sqrt{s} & s \geq 1 \end{cases} \tag{6.31}$$

concludes the proof. □

6.4.2 Remark. In order to avoid confusion, we note that the equations (6.3), (6.18) and (6.19) may be considered as a conservation law, a Hamilton-Jacobi type equation and a balance law respectively. Our intent in Lemma 6.4.1 is the posteriori estimate in a more general format. We simplify this result in Cor. 6.4.3 below to make it specific to our case of interest: namely, a linear system of conservation laws with a constant, symmetric coefficient matrix.

6.4.3 Corollary (Stability and error control for a constant, symmetric linear system of equations)**.** *Let the conditions of Lem. 6.4.1 with $A \in \mathbb{R}^{m\times m}$ constant and symmetric. Then, the error bound (6.20) holds with $C = 0$ in (6.21).*

6.5 Numerical methods and discretisation

In this section we will present the spatial and temporal discretisations for our numerical experiments, as well as the Leapfrog scheme we will use to approximate the wave equation as a system (see the model problem (6.4)). We will express the model problem - the one-dimensional wave equation- in the form (6.4). We will then discretise the system using the Leapfrog scheme on staggered grids.

6.5.1 Temporal and spatial domain discretisation

We discretise the spatial domain $\Omega := [0, 1]$ by choosing a uniform partition with constant spatial step-size h and points $0 = x_0 < \cdots < x_M = 1$ such that $x_{j+1} - x_j =$

h for all $j = 0, \ldots, M$. Next, we uniformly partition the temporal domain $[0, T]$ into N by defining $0 = t^0 < \cdots < t^N = T$ with constant step-size τ, where N is chosen such so that the desired CFL condition is satisfied. We denote by $\boldsymbol{u}_j^n := \boldsymbol{u}(t^n, x_j)$ the exact solution to (6.4) and by $\boldsymbol{U}_j^n$ we denote the approximation to $\boldsymbol{u}_j^n$ obtained by the chosen numerical scheme.

The test problem we will use is the one dimensional wave equation with periodic boundary conditions and prescribed initial conditions on u and its temporal derivative:

$$\begin{aligned} u_{tt} &= u_{xx} \\ u(x,0) &= u_0(x) \\ u_t(x,0) &= v_0(x)\,, \end{aligned} \tag{6.32}$$

We express (6.32) as a one-dimensional system of first order advection equations by introducing a variable v which is related to u in (6.32) as follows: This is written as a system

$$\begin{aligned} v_t + u_x &= 0 \\ u_t + v_x &= 0. \end{aligned} \tag{6.33}$$

The test problem, (6.33), is now in the form of the model problem (6.4), with $\boldsymbol{u} = (u, v)^T$ and

$$A := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{6.34}$$

6.5.2 Numerical scheme

Central difference Leapfrog

The model problem (6.32) can be approximated using a central difference FD discretisation:

$$\frac{U_j^{n+1} - 2U_j^n + U_j^{n-1}}{\tau^2} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} \tag{6.35}$$

for $n = 1, \ldots, N$. Notice that the scheme uses three time levels U^{n+1}, U^n and U^{n-1}. In order to initiate it we need the value of U^{-1}, which is unavailable. Instead, we make use of the given initial conditions, $u_0(x_j)$ and $v_0(x_j)$ from (6.32). We use a central difference discretisation to obtain the value of U^{-1}:

$$\frac{U_j^1 - U_j^{-1}}{2\tau} = v_0(x_j)\,, \tag{6.36}$$

which gives us

$$U_j^{-1} = U_j^1 - 2\tau v_0(x_j)\,. \tag{6.37}$$

We can substitute (6.37) in (6.35) to obtain U^1 as follows

$$U_j^1 = U_j^0 + \tau v_0(x_j) + \frac{\tau^2}{2h^2}\big(U_{j+1}^0 - 2U_j^0 + U_{j-1}^0\big)\,. \tag{6.38}$$

Once we have computed U^1 we can compute U^{-1} from (6.35), by substituting the obtained U^1 values back in the scheme. In a similar fashion to what we did in §2.2.3, we will formulate the leapfrog scheme for u and v using staggered grids. The staggering will be in both space and time. Specifically, with reference to (6.32) the approximation to the variable u, will be on integer space and time points (t^n, x_j) and will be denoted as U_j^n. In the same vein, the approximation to v will be defined on mid-points of space-time slabs, $\big(t^{n+1/2}, x_{j+1/2}\big)$ and will be denoted as $V_{j+1/2}^{n+1/2}$.

Leapfrog scheme on staggered grids

We now have all the information we need in order to re-formulate the scheme (6.35) into a Leapfrog scheme on staggered grids. In this way we can solve (6.32) as a first order system of conservation laws given by (6.33), for which we already have presented a framework for obtaining reconstructions. We will use the Leapfrog scheme posed over staggered grids, which is second order accurate and has several advantages, explained in Chapter 2 and in [GLMV16]. The temporal component of the scheme is in fact the same as that presented in §2.2.3 and the spatial component is also staggered.

In the notation of (6.35) and using V_j^n to denote the numerical approximation to v in (6.33), we define the Leapfrog scheme on staggered grids (in both space and time) as follows:

$$\begin{aligned} \frac{V_{j+1/2}^{n+1/2} - V_{j+1/2}^{n-1/2}}{\tau} + \frac{U_{j+1}^n - U_j^n}{h} &= 0 \\ \frac{U_j^{n+1} - U_j^n}{\tau} + \frac{V_{j+1/2}^{n+1/2} - V_{j-1/2}^{n+1/2}}{h} &= 0 \end{aligned} \tag{6.39}$$

for $n = 0, \ldots, N$, $j = 0, \ldots, M$, with periodic boundary conditions which are implemented by identifying the end-points of the domain, i.e $x = 0$ and $x = 1$.

Reconstruction

In the reconstruction procedure the first step is to obtain the temporal component, $\widehat{\boldsymbol{U}}^t = \left(\widehat{U}^t, \widehat{V}^t\right)$, from the numerical solution $\boldsymbol{U} = (U, V)$ produced by the numerical scheme (6.39). Once the temporal component of the reconstruction is available, it is used to obtain the full spatio-temporal reconstruction.

The procedure for obtaining the temporal component of the WENO reconstruction, $\widehat{\boldsymbol{U}}^t = \left(\widehat{U}^t, \widehat{V}^t\right)$, from the FD solution obtained from the numerical scheme, (6.39), is explained in Defn. 2.3.14. Since the definition pertains to the same temporal discretisation as the one we are using in this chapter, that is Leapfrog on grids staggered in time, the exact same procedure is followed as in in Chapter 2.

6.5.3 Remark. In particular, the reader should note that, because the grids are staggered in time, we need to store an additional value of $V^{n+1/2}$, as we explain in Rem 2.3.15 in greater detail.

Once the temporal component of the reconstruction, $\widehat{\boldsymbol{U}}^t = \left(\widehat{U}^t, \widehat{V}^t\right)$, is available we use it to obtain the full spatio-temporal reconstruction, $\widehat{\boldsymbol{U}} = \left(\widehat{U}, \widehat{V}\right)$. The full reconstruction is obtained as explained in Defn. 7.4.7. The reader should note that, because the grids are staggered in space as well, we need to account for this when obtaining the reconstruction.

6.5.4 Remark. Once again, attention must be paid during implementation to the fact that the staggering is also in space. This is not an issue during the calculation of the error; it is of concern only during the calculation of the residual computation, which involves quantities which are defined on different grids.

6.6 Numerical verification

In this section we conduct numerical tests in order to investigate the behaviour of the a posteriori error bound given in Cor. 6.4.3. In the same fashion as in the previous chapters, we use the reconstruction, $\widehat{\boldsymbol{U}}$, for two purposes. Firstly, to facilitate an alternative error interpretation and secondly to obtain a computable a posteriori error bound using the stability framework of the PDE. We define the alternative error interpretation as

$$\boldsymbol{e} := \boldsymbol{u} - \widehat{\boldsymbol{U}}. \tag{6.40}$$

The a posteriori error bound we will compute is given by

$$\eta(t) := \left(\omega(t)\left[\|\boldsymbol{e}(0)\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \|\delta(s)\boldsymbol{R}(s)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s\right]\right)^{1/2}, \tag{6.41}$$

where $\omega(t)$ and $\delta(t)$ are given in Cor. 6.4.3. We assess the results based on the EI and the EOC of the alternative error interpretation (6.40) and the a posteriori bound (6.41).

We run a benchmarking experiment for (6.33), discretised by the Leapfrog scheme, (6.39) on a staggered grid, in the domain $\Omega = [0, 1]$ with periodic boundary conditions, an initial condition given by

$$\begin{aligned} u_0(x) &= 15\sin(2\pi x)\,, \\ v_0(x) &= \frac{\partial u}{\partial t}(0, x) = 30\pi\sin(2\pi x) \end{aligned} \tag{6.42}$$

and an exact solution given by

$$u(x, t) = 15\sin(\pi x)[\cos(2\pi t) + \sin(2\pi t)]\,. \tag{6.43}$$

For an exact solution, $u(x, t)$, given by (6.43), we obtain (after a few calculations) $v(x, t)$ in (6.32) as

$$u(x, t) = 15\cos(\pi x)[\cos(2\pi t) - \sin(2\pi t)]\,. \tag{6.44}$$

The simulations are conducted over a family of grids parametrised by $h = 2^{-m}$, $m = 9, \ldots, 12$ and a time-step $\tau = h/10$. The spatial and temporal grids are staggered for the two variables in (6.39), with $\left\{U_j^n\right\}$ obtained at integer space-time points and $\left\{V_{j+1/2}^{n+1/2}\right\}$ obtained at the space-time slabs' mid-points. The results are shown in Figure 6.1.

Notice that the estimate behaves optimally as it converges at the same rate as the underlying error in the chosen norm. The somewhat strange initial behaviour of the error and estimate, which is rather more apparent in the EOC plots (third and fourh plots from the right in Fig. 6.1) may potentially be the result of a parasite in time. This may in turn be due due to the choice of the particular norm.

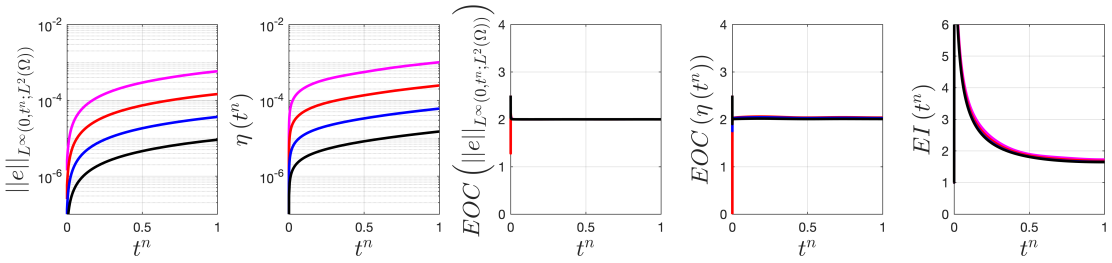


Fig. 6.1. Errors and asymptotic convergence rates for the error, (6.40), and the a posteriori bound, (6.41), using the WENO reconstruction (see Defns. 2.3.14 and 7.4.7), of the FD solution of (6.33), discretised with the Leapfrog scheme (6.39) over staggered grids. Notice that the estimate is optimal for the Leapfrog scheme with favourable effectivity of $EI(1) \sim 2$.

6.7 Conclusion

In this chapter we derive an a posteriori error bound for a linear system of conservation laws with a constant and symmetric coefficient matrix. We benchmark the behaviour of this bound by using as a model problem the one-dimensional wave equation in system form, discretised by a Leapfrog scheme on staggered grids.

In order to construct an a posteriori error bound we firstly define an alternative error interpretation to overcome the difficulty caused by the point-wise nature of the FD solution. We do this by using the WENO reconstruction procedure that we introduced in Chapters 2 and 5 for the temporal and spatial components of the reconstruction respectively. The WENO reconstruction is used in the computation of both the error as well as of the a posteriori bound.

We find that the a posteriori bound evaluated in this way is optimal, in that it converges with the same EOC as the bound. In addition, a bound constructed in this way has a favourable effectivity index of approximately two (see also Fig. 6.1).

Chapter 7

A posteriori error analysis for non-linear hyperbolic problems

Abstract

In this chapter, we examine non-linear conservation laws. We approximate the solutions with the finite difference schemes considered in previous chapters and show a posteriori bounds in different cases. We are able to use the relative entropy framework to prove a posteriori upper bounds for general systems and a Kruzkov framework to prove upper bounds for scalar conservation laws.

7.1 Introduction

In this chapter we use the framework introduced in Chapter 5 to obtain reconstruction-based a posteriori error bounds for non-linear problems. We demonstrate the use of the framework in both a scalar example and a systems example.

In the scalar case, we use Burgers equation as a model problem. The reconstruction is used to obtain a optimal a posteriori error bounds both in the pre-shock and post-shock regime. In the pre-shock regime we use the entropy framework of [GP17] to obtain a bound in the $\mathrm{L}^2(\Omega)-$norm. In the post-shock regime we examine the use of a bound from [CG95], which is based on the Kruzhkov framework (see [OV06],[Ohl09]). We lay the groundwork for using the two bounds in conjunction with a view of achieving a result which is optimal in both the pre-shock and post-shock regimes. This currently remains a future challenge.

In the systems case we use the shallow water equation as a model problem. In this case, we use the a posteriori error bound to drive adaptive mesh refinement.

The a posteriori error analysis is carried out using the stability framework of the PDE (see [GMP15]) and it therefore yields a bound which is usable regardless of the chosen numerical discretisation technique. The construction of the estimate for computational purposes is done using reconstruction techniques. Similar techniques have been used for dG methods (see e.g. [Mak07, GMP15, GP17, DGPR19]).

We conduct a range of numerical experiments to benchmark the behaviour of the estimates and to demonstrate the concept of optimal order for smooth solutions. We showcase relevant results for widely used schemes for both linear and non-linear examples.

7.1.1 Motivation

In this chapter we apply the framework to non-linear problems for FD schemes. FD schemes are in widespread use for hyperbolic conservation laws. However, as pointed out in previous chapters, a posteriori error control for FD schemes has not received as much attention as for the FV and FE counterpart. In this spirit we introduced a framework for obtaining reconstructions from FD solutions. Reconstructions enable a posteriori error control for FD discretisations, albeit for an alternative error interpretation.

The novelty in this chapter is the application to non-linear hyperbolic problems. These problems exhibit behaviour which makes it challenging both to pose numerical discretisation schemes as well as to establish a posteriori error control when compared to linear problems. An example of such behaviour in non-linear problems is that shocks form in finite time, even for smooth initial conditions.

For scalar problems, such as Burgers equations, we examine both $\mathrm{L}^1(\Omega)$ stability results as well as relative entropy $\mathrm{L}^2(\Omega)$ stability results, with the intention of combining them in future work to derive an a posteriori bound which is robust both prior and subsequent to shock formation. We believe this will be the first such bound in the literature.

7.1.2 Chapter contribution

In this Chapter we use reconstructions to compute a posteriori error estimates derived in [GMP15] for non-linear examples for two examples in one spatial dimension: a non-linear scalar example, Burgers equation, and a non-linear system, the shallow water equations.

In the scalar example we examine the behaviour of two a-posteriori error estimates - a relative entropy-based a posteriori estimate from [GMP15] with an a posteriori estimate based on Kruzkov's doubling of variables technique from [CG95]. Our intention is to examine the possibility of combining the two to obtain an estimate that is optimal both prior and following shock-formation for the scalar problem in one dimension.

In the shallow water example we use the a posteriori error estimate to facilitate an adaptive algorithm and we benchmark the results against an equivalent uniform mesh - a mesh that possesses the same number of cumulative degrees of freedom as the adaptive mesh over the simulated time.

The rest of the chapter is structured as follows. In §7.2 we set up the preliminaries and present the chosen notation. In §7.3 we present the model problems we will address. In §7.4 we present the domain discretisation, the numerical methods we will be using as well as the relevant reconstructions. In §7.5 we present the relevant a posteriori error bounds. In §7.6 we present numerical experiments to validate the behaviour of the bound. Finally, we conclude the chapter in §7.7.

7.2 Setup

In this section we present preliminary material and results we need in order to set up model problems and a posteriori error results in later sections. We present the general format of both the scalar problem and the system and we define theoretical results and concepts that are necessary in this context. In particular, we define the concepts of weak solutions, admissibility criteria, entropy/ entropy flux pairs and entropy solutions. We will recall notation from previous sections as necessary.

7.2.1 Scalar conservation law

In this section we present results for the non-linear scalar problem

$$\begin{aligned} u_t + \partial_x f(u) &= 0, \\ u(x,0) &= u_0(x)\,, \end{aligned} \qquad \text{for } (x,t) \in \Omega \times (0,\infty) \tag{7.1}$$

complemented with periodic boundary conditions.

7.2.2 Remark. (Non-uniqueness of weak solutions) As alluded to in the previous chapter, It is possible that multiple weak solutions, u may exist for a specific problem (7.1). In order to address the issue of non-uniqueness one must incorporate admissibility criteria for choosing the appropriate weak solution. Such criteria arise naturally through the second law of thermodynamics and can be realized, as [Daf05] points out, either by requiring that admissible solutions satisfy entropy inequalities or by adding small amounts of diffusion (vanishing viscosity solutions). In the case of the latter choice, small amounts of diffusions have far greater effect on solutions containing shocks than on smooth solution.

7.2.3 Definition. (Vanishing viscosity solution) To build upon some of the results in the previous chapter, consider the perturbed problem, parametrized by (a small parameter) $\epsilon > 0$:

$$u_t^\epsilon + f(u^\epsilon)_x = \epsilon u_{xx}^\epsilon \tag{7.2}$$

with initial data $u^\epsilon(x,t) := u_0^\epsilon(x)$, which satisfies $\lim_{\epsilon \to 0^+} u_0^\epsilon(x) = u_0(x)$. Then, we say that u is a viscosity solution of (7.1) if it can be obtained as the limit

$$u = \lim_{\epsilon \to 0^+} u^\epsilon \tag{7.3}$$

of solutions u^ϵ to the perturbed problem (7.2)

In order to relate the concept of vanishing viscosity with entropy solutions, we consider a convex function $\eta : \mathbb{R} \to \mathbb{R}$ and we use this to construct a function $q : \mathbb{R} \to \mathbb{R}$ which is defined as

$$q(u) := \int_0^u f'(s)\,\eta'(s)\,\mathrm{d}s. \tag{7.4}$$

7.2.4 Lemma. *Let u be a smooth solution of (7.1) and suppose that η and q are an entropy and entropy-flux pair that satisfies the relation*

$$q' = \eta' f'. \tag{7.5}$$

Then, the following result holds:

$$\eta'(u)\, u_t + q'(u)\, u_x = 0. \tag{7.6}$$

Proof.

$$\begin{aligned} \frac{\partial}{\partial t}\eta(u) + \frac{\partial}{\partial x}q(u) &= \eta'(u)\, u_t + q'(u)\, u_x \\ &= \eta'(u)(u_t + f'(u)) \\ &= 0. \end{aligned} \tag{7.7}$$

□

7.2.5 Remark. (Entropies for scalar conservation laws) In the case of scalar conservation laws, any convex function η is an entropy for (7.1) ([GR13]). The resulting entropy flux q is then obtained using the relation (7.4).

Let (η, q) be an entropy/entropy flux pair with $\eta \in \mathrm{C}^2(\mathbb{R} \times \mathbb{R}^+)$ and let u^ϵ be a vanishing viscosity solution to (7.1) (see Defn. 7.2.3). We multiply (7.2) by $\eta'(u^\epsilon)$ and obtain

$$\eta'(u^\epsilon)\, u^\epsilon_t + \eta'(u^\epsilon)\, f'(u^\epsilon)\, u^\epsilon_x = \epsilon \eta'(u^\epsilon)\, u^\epsilon_{xx}. \tag{7.8}$$

We then use the relation (7.5) to obtain

$$\begin{aligned} \eta'(u^\epsilon)\, u^\epsilon_t + q'(u^\epsilon)\, u^\epsilon_x &= \epsilon \eta'(u^\epsilon)\, u^\epsilon_{xx} \\ \eta(u^\epsilon)_t + q(u^\epsilon)_x &= \epsilon \eta(u^\epsilon)_{xx} - \epsilon \eta''(u^\epsilon)(u^\epsilon_x)^2 . \end{aligned} \tag{7.9}$$

Since η is a convex function, $-\eta''(u^\epsilon)(u^\epsilon_x)^2$ is non-positive and hence

$$\eta(u^\epsilon)_t + q(u^\epsilon)_x \leq \epsilon \eta(u^\epsilon)_{xx} . \tag{7.10}$$

Since u^ϵ is a vanishing viscosity solution to (7.1) (see Defn.7.2.3), passing to the limit $\epsilon \to 0$ we have that u, the weak solution to (7.1) satisfies the inequality

$$\partial_t \eta(u) + \partial_x(u) \leq 0. \tag{7.11}$$

in the sense of distributions. We state this formally in Lem. 7.2.6.

7.2.6 Lemma. *(Entropy condition (see [GR13, Thm. 3.3])) Assume that (7.1) admits an entropy $\eta(u)$ with an associated entropy flux $q(u)$. Let $(u^\epsilon)_\epsilon$ be a sequence of sufficiently smooth solutions of (7.2) such that*

$$\begin{aligned} &\|u_\epsilon\|_{\mathrm{L}^\infty(\mathbb{R}\times(0,+\infty])} \leq C, \quad C > 0, \\ &u_\epsilon \to u \, as \, \epsilon \to 0 \, a.e. \, in (x,t) \, \mathbb{R} \times (0, +\infty) , \end{aligned} \tag{7.12}$$

where $C > 0$ is constant and independent of ϵ. Then, u is a weak solution of (7.1) and satisfies the entropy condition

$$\partial_t \eta(u) + \partial_x q(u) \leq 0 \tag{7.13}$$

in the sense of distributions i.e.

$$\int_0^\infty \int_\Omega \eta(u)\, \partial_t \phi + q(u)\, \partial_x \phi \, \mathrm{d}x\mathrm{d}t \geq 0 \quad \forall \phi \in C_c^1(\mathbb{R} \times [0, \infty)) \,,\ \phi \geq 0. \tag{7.14}$$

The inequality (7.14) is called the entropy condition.

We now have the tools to give an existence and uniqueness theorem for entropy solutions to the scalar conservation law (7.1).

7.2.7 Theorem. *(Existence and uniqueness of entropy solution for (7.1); [GR13, Theorem 3.4]) Assume that the $u_0 \in \mathrm{L}^\infty(\mathbb{R})$. Then, the problem (7.1) has a unique entropy solution $u \in \mathrm{L}^\infty(\mathbb{R} \times (0, T))$. This solution satisfies for almost all $t \geq 0$*

$$\|u(\cdot, t)\|_{\mathrm{L}^\infty(\mathbb{R})} \leq \|u_0\|_{\mathrm{L}^\infty(\mathbb{R})} \,. \tag{7.15}$$

Moreover, if u and v are the entropy solutions of (7.1) associated with initial conditions u_0 and v_0 respectively, we have

$$u_0 \geq v_0 \Longrightarrow u(\cdot, t) \geq v(\cdot, t) \ \ a.e. \tag{7.16}$$

Finally, if $u_0 \in \mathrm{L}^\infty(\mathbb{R}) \cap \mathrm{BV}(\mathbb{R})$, then $u(\cdot, t) \in \mathrm{BV}(\mathbb{R})$ with

$$TV(u(\cdot, t)) \leq TV(u_0) \,. \tag{7.17}$$

7.2.8 Remark. (Uniqueness of entropy solution for the scalar problem (7.1)) A uniqueness proof of the entropy solutions (up to a set of measure 0) to (7.1) can be found in [Eva10, §3.4 Thm 3].

7.2.9 Systems of conservation laws

In this section we will present relevant results for one-dimensional systems of nonlinear conservation laws, recalling notation from previous sections as necessary.

7.2.10 Definition (One-dimensional system of conservation laws)**.** We consider problems of the form

$$\begin{aligned} \boldsymbol{u}_t + \partial_x \boldsymbol{f}(\boldsymbol{u}) &= \mathbf{0}, \\ \boldsymbol{u}(x,0) &= \boldsymbol{u}_0(x)\,, \end{aligned} \qquad \text{for } (x,t) \in \Omega \times (0,\infty) \tag{7.18}$$

with $\boldsymbol{u} = (u_1, \ldots, u_m)^T$ and $\boldsymbol{f}(\boldsymbol{u}) = (f_1(\boldsymbol{u}), \ldots, f_m(\boldsymbol{u}))^T$ and complemented with periodic boundary conditions. In particular,

$$\begin{aligned} \boldsymbol{u}: \quad \mathbb{R} \times \mathbb{R}^+ &\to \mathbb{R}^m \\ (x,t) &\mapsto \boldsymbol{u}(x,t) \end{aligned} \tag{7.19}$$

and the flux function $\boldsymbol{f}$

$$\begin{aligned} \boldsymbol{f}: \quad \mathbb{R}^m &\to \mathbb{R}^m \\ \boldsymbol{u}(x,t) &\mapsto \boldsymbol{f}(\boldsymbol{u}(x,t)) \end{aligned} \tag{7.20}$$

7.2.11 Remark. (Piecewise smooth solutions to 7.18) In the context of weak solutions, the system of conservation laws (7.18) may also admit solutions which feature jump discontinuities (see [LeF02]).

7.2.12 Remark. (Jump) Consider a single curve which splits the space-time domain in two pieces, which we will denote by Ω^+ and Ω^-. Also suppose that this curve is $\mathrm{C}^1(\mathbb{R} \times \mathbb{R}^+)$. We will denote the shock by $x = \gamma(t)$, where $\gamma : t \to \gamma(t) \in \mathrm{C}^1(\mathbb{R} \times [t^1, t^2])$ for some $t^1 < t^2$ and we let $\Gamma := \{(x,t) \in \mathbb{R} \times \mathbb{R}^+ : x = \gamma(t)\}$. We use this to define the limits of u on each side of Γ:

$$\boldsymbol{u}^{\pm}(x,t) := \lim_{x \to \gamma(t)^{\pm}} \boldsymbol{u}(x,t)\,. \tag{7.21}$$

Let $s(t) := \gamma'(t)$ denote the speed of the shock and let $\boldsymbol{\nu} := (\nu^t, \nu^x)$ denote the normal to the curve:

$$\left(\nu^t, \nu^x\right) := (-s(t)\,, 1) \tag{7.22}$$

Lastly let $\boldsymbol{u}^+ \in \mathrm{C}^1(\Omega^+)$ and $\boldsymbol{u}^- \in \mathrm{C}^1(\Omega^-)$. The relation between $\boldsymbol{u}$ and $\boldsymbol{f}(\boldsymbol{u})$ across the curve is given by the *Rankine-Hugoniot* jump condition

7.2.13 Theorem. *(Rankine-Hugoniot jump condition [GR13, Thm. 2.1]) Let* $\boldsymbol{u} : \mathbb{R} \times \mathbb{R}^+ \to \Omega$ *be a piecewise* $\mathrm{C}^1(\mathbb{R} \times \mathbb{R}^+)$ *function, with the discontinuity given by the curve described in the remark above. Then* $\boldsymbol{u}$ *is a weak solution of (7.18) on* $\mathbb{R} \times \mathbb{R}^+$ *if and only if the following two conditions are satisfied:*

1. $\boldsymbol{u}^{\pm}$ *are classical solutions of (7.18) in* $\Omega^{\pm}$ *respectively.*

2. $\boldsymbol{u}$ *satisfies the following jump condition along the discontinuity*

$$\left(\boldsymbol{u}^{+}-\boldsymbol{u}^{-}\right)\nu^{t}+\left(\boldsymbol{f}\left(\boldsymbol{u}^{+}\right)-\boldsymbol{f}\left(\boldsymbol{u}^{-}\right)\right)\nu^{x}=\mathbf{0}. \tag{7.23}$$

The jump relation (7.23) is called the Rankine-Hugoniot condition.

7.2.14 Remark. We will use the shorthand $[\![\cdot]\!]$ to denote the jump in a certain quantity. Hence, we can write the Rankin-Hugoniot condition, (7.23), as

$$s\,[\![\boldsymbol{u}]\!]=[\![\boldsymbol{f}(\boldsymbol{u})]\!]\,. \tag{7.24}$$

7.2.15 Remark. The Rankine-Hugoniot condition is a means for determining whether a shock-wave is indeed a weak solution for (7.18).

7.2.16 Remark. (Existence of entropy pairs for systems of conservation laws) In contrast with scalar conservation laws where every convex entropy η gives rise to an entropy flux q, it is possible that systems of conservation laws possess only a single entropy pair, the existence of which may be a special property of the specific case in question ([GR13]).

7.2.17 Theorem. *([GR13, Thm 3.1]) Let* $\eta:\Omega\to\mathbb{R}$ *be a strictly convex function. A necessary and sufficient condition for* η *to be an entropy for the system (7.18) is that the* $p\times p$ *matrices* $\eta''(\boldsymbol{u})\,\boldsymbol{f}'(\boldsymbol{u})$ *are symmetric.*

As [GR13] point out, the symmetrization of (7.18) can be accomplished by introducing new dependent variables $\boldsymbol{v}$, i.e. $\boldsymbol{u}=\boldsymbol{u}(\boldsymbol{v})$ s.t. $\boldsymbol{u}'$ is symmetric positive definite and $\boldsymbol{f}'(\boldsymbol{u})\,\boldsymbol{u}'(\boldsymbol{v})$ are symmetric. This is formalized as the following theorem

7.2.18 Theorem. *[GR13, Thm. 3.2] A necessary and sufficient condition for the system (7.18) to possess a strictly convex entropy* η *is that there exists a change of dependent variables* $\boldsymbol{u}=\boldsymbol{u}(\boldsymbol{v})$ *that symmetrizes (7.18).*

7.3 Model problems

7.3.1 Scalar model problem: Burgers equation

We consider (7.1) for the one-dimensional Burgers equation given by

$$\begin{aligned} \partial_t u + \partial_x\left(\frac{u^2}{2}\right) = 0, \quad \text{in} \quad \Omega \times (0,T], \\ u(x,0) = u_0(x) \quad \text{in} \quad \Omega \times \{0\} \end{aligned} \tag{7.25}$$

and coupled with periodic boundary conditions.

7.3.2 Remark. Periodic boundary conditions for Burgers equation The reader should note that, in this problem, we use periodic boundary conditions as a computational convenience. It is important to emphasize that if the computation is allowed to run for long enough it will eventually become polluted. This will happen once the solution reaches the periodic boundary.

7.3.3 System model problem: shallow water equations

We consider the one dimensional shallow water equations given by

$$\begin{aligned} \eta_t &+ (\eta v)_x &= 0 \\ (\eta v)_t &+ \left(\eta v^2 + \tfrac{1}{2} g \eta^2\right)_x &= 0, \end{aligned} \tag{7.26}$$

in $\Omega \times (0,T]$ equipped with the initial conditions

$$\begin{aligned} h(x,0) &= \begin{cases} h_0 & \text{for} \quad x \leq x_0 \\ h_1 & \text{for} \quad x > x_0, \end{cases} \\ v(x,0) &= v_0(x) \end{aligned} \tag{7.27}$$

and coupled with free outflow boundary conditions.

7.4 Numerical discretisation and reconstruction

In this section we present the temporal and spatial approximation of the domains in which we conduct our numerical experiments as well as the numerical FD approximations for the corresponding to the scalar and system model problem (7.1) and (7.18) respectively. Subsequently, we will present the relevant reconstructions.

7.4.1 Scalar model problem

Spatial and temporal domain discretisation

We partition the domain $\Omega = [-\pi, \pi]$ uniformly by choosing points $-\pi = x_0 < \cdots < x_M = \pi$ and we denote the step-size by h. We denote by I_j the sub-interval $[x_j, x_{j+1}]$ of Ω, $j = 0, \ldots, M-1$. In the temporal variable we partition the interval $[0, T]$ into sub-intervals with end-points $0 = t^0 < \cdots < t^N = T$, where N is chosen such that a CFL condition is satisfied. We denote by $u_j^n := u(t^n, x_j)$ the exact solution to (7.1) and by U_j^n we denote the approximation to u_j^n, obtained by the chosen numerical scheme.

Numerical Scheme

The numerical scheme we use is the SSP3-WENO3 scheme. We specify the temporal component of this scheme in the Butcher tableau given in (5.16) (see [GST01, §4:(4.2)]) and the spatial component is specified in §5.3.13 (see [Shu98, Table 2.1] and [JSB+19, §3]).

7.4.2 Reconstruction

The reconstruction is obtained using the FD solution of (7.1), which is obtained as explained in §5.3.

We illustrate the reconstruction procedure in Fig. 7.1. Firstly, we use Defn. 7.4.5 to obtain $\widehat{\boldsymbol{U}}^t(t)$ from $\boldsymbol{U}$, where the latter is produced by the chosen numerical discretisation to (4.2). This is the solid red line in the middle plot. Then, we use Defn. 7.4.7 to obtain $\widehat{\boldsymbol{U}}^{ts}(x,t)$ from $\widehat{\boldsymbol{U}}^t(t)$. This is the transparent red surface in the third plot.

The reconstruction procedure will now be explained in detail. We obtain a polynomial reconstruction, which we will denote by $\widehat{\boldsymbol{U}}^{ts}$, by using the nodal values of $\boldsymbol{U}$ as well as the temporal and spatial approximations of the partial derivatives of the equation. The reader should note that we interpolate firstly in time and subsequently in space, because the temporal component of (7.31) is linear while the spatial one may be non-linear.

In the exposition that follows, we will use the superscripts t and s to represent

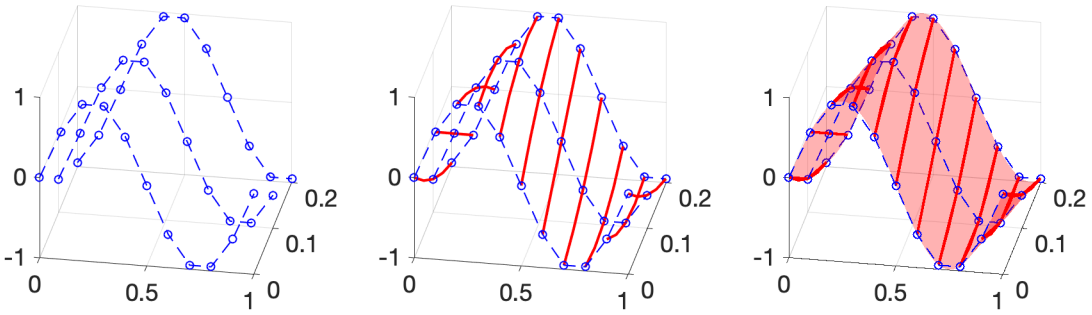


Fig. 7.1. Example of a reconstruction procedure. (Left) $\boldsymbol{U}^n$ for $n = 0, 1, 2$ produced by a Finite Difference scheme in 1D, with periodic boundary conditions and a sinusoidal initial condition (blue dashed lines). (Middle) Temporal reconstruction step $\widehat{\boldsymbol{U}}^t(t)$, depicted as solid, red lines. (Right) Spatio-temporal reconstruction step $\widehat{\boldsymbol{U}}^{ts}(x,t)$, depicted as a transparent, red surface.

that the function in question is either time or space dependent only. We will also use the supercript ts to denote dependence on both time and space.

7.4.3 Definition (Space of the spatial reconstruction step)**.** Let $\mathbb{P}^q([\![x_j, x_{j+1}]\!])$ denote the space of polynomials of degree q over the sub-interval $[x_j, x_{j+1}]$. We define the space of the spatial step of the reconstruction,

$$\mathbb{V}_q^s := \left\{ w : [0, L] \to \mathbb{R} : w|_{[\![x_j, x_{j+1}]\!]} \in \mathbb{P}^q([\![x_j, x_{j+1}]\!]) \right\}, \tag{7.28}$$

to be the space of piecewise polynomials of degree q over $[0, L]$. The superscript s indicates the space dependence.

7.4.4 Definition (Space of the temporal reconstruction step)**.** We define the space of the temporal step of the reconstruction as the space of piecewise polynomials of degree 3 over $[0, T]$ such that

$$\mathbb{V}_3^t(0, T; \mathrm{L}^\infty(\Omega)) := \left\{ g : [0, T] \to V : g|_{[\![t^n, t^{n+1}]\!]} \in \mathbb{P}^3\left([\![t^n, t^{n+1}]\!], \mathrm{L}^\infty(\Omega)\right) \right\}. \tag{7.29}$$

Here, $\mathbb{P}^3([\![t^n, t^{n+1}]\!], \mathrm{L}^\infty(\Omega))$ is the space of functions which are polynomials of degree q in time and belong to $\mathrm{L}^\infty(\Omega)$ in space.

7.4.5 Definition (Temporal reconstruction)**.** The temporal reconstruction, $\widehat{\boldsymbol{U}}^t \in \mathbb{V}_3^t(0, T; \mathrm{L}^\infty(\Omega))$of the numerical solution $\boldsymbol{U}$, is the unique function satisfying

$$\begin{aligned} \widehat{\boldsymbol{U}}^t(t^n) &= \boldsymbol{U}_j^n \quad \text{and} \\ \partial_t \widehat{\boldsymbol{U}}^t(t^n) &= -\frac{1}{h}\left(\boldsymbol{F}\left(\boldsymbol{U}_{j-p}^n, \cdots, \boldsymbol{U}_{j+q}^n\right) - \boldsymbol{F}\left(\boldsymbol{U}_{j-p-1}^n, \cdots, \boldsymbol{U}_{j+q-1}^n\right)\right). \end{aligned} \tag{7.30}$$

for $n = 0, \dots, N$.

7.4.6 Remark (Order of the temporal reconstruction)**.** The procedure presented in Defn. 7.4.5 limits the temporal component of the reconstruction to third order. The reason for this is that the four conditions specified in (7.30) allow us to obtain reconstructions of polynomial order up to and including three. A possibility for increasing the order of the temporal reconstruction is by obtaining a WENO interpolant for the temporal component. We have demonstrated this approach in §2.3.13.

Once we obtain the temporal reconstruction we use it to define the full spatio-temporal reconstruction. The procedure used to obtain the spatial component is based on the WENO interpolation procedure which is derived and presented in detail in [JSB+19].

7.4.7 Definition (Spatio-temporal reconstruction)**.** Let $\widehat{\boldsymbol{U}}^t$ be a temporal reconstruction of the numerical solution $\boldsymbol{U}$ of (5.1). The spatio-temporal reconstruction $\widehat{\boldsymbol{U}}^{ts} \in \mathbb{V}_3^t\big(0, T; \mathbb{V}_q^s\big)$ in the interval $[x_j, x_{j+1}]$ is given by the WENO interpolant of $\widehat{\boldsymbol{U}}^t$, which is defined in (5.41).

7.4.8 Remark (Order of the reconstruction)**.** The conditions presented in Defn. 7.4.5 result in a polynomial which is of third order in the temporal variable. In contrast Defn. 7.4.7 can be used to obtain spatial reconstructions of arbitrary order in space, simply by using a higher order WENO interpolant. The limiting factor in the order of the full spatio-temporal reconstruction will be the lowest order between the spatial and temporal steps. In this case, this will be the order of the temporal component (order three).

7.4.9 Definition (Reconstruction)**.** The reconstruction, $\widehat{\boldsymbol{U}}^{ts}$, of the numerical solution, $\boldsymbol{U}$, to (5.1) is a function that satisfies

$$\begin{aligned} \widehat{\boldsymbol{U}}_t^{ts} + \partial_x \boldsymbol{f}\big(\widehat{\boldsymbol{U}}\big)^{ts} &=: -\boldsymbol{R} \quad \text{in } \Omega \times (0, T] \\ \widehat{\boldsymbol{U}}^{ts}(\boldsymbol{x}, 0) &= \boldsymbol{u}_0(\boldsymbol{x}) \quad \text{in } \Omega \times \{0\} \end{aligned} \tag{7.31}$$

and the relevant boundary conditions, such that the reconstruction residual, $\boldsymbol{R}$, is well-defined and explicitly computable. Furthermore, $\widehat{\boldsymbol{U}}$ should lead to an optimal a posteriori error estimate. An estimate is optimal when it converges at the same rate as the chosen numerical scheme.

7.4.10 System model problem

Spatial and temporal domain discretisation

We partition the domain $\Omega = [0, 32\pi]$ uniformly by choosing points $0 = x_0 < \cdots < x_M = 32\pi$, denoting the step-size by h. We denote by I_j the sub-interval $[x_j, x_{j+1}]$ of Ω, $j = 0, \ldots, M-1$. In the temporal variable we partition the interval $[0, T]$ into sub-intervals with end-points $0 = t^0 < \cdots < t^N = T$, where N is chosen such that a CFL condition is satisfied. We denote by $\boldsymbol{u}_j^n := \boldsymbol{u}(t^n, x_j)$ the exact solution to (7.1) and by $\boldsymbol{U}_j^n$ we denote the approximation to $\boldsymbol{u}_j^n$, obtained by the chosen numerical scheme.

7.4.11 Remark. In what follows we remind users that the term reconstruction is used to refer to both the procedure that is used to formulate the WENO scheme (described in §5.3.13 and in [SO88, §2]) as well as to the procedure that we utilize to post-process the FD solution.

Numerical Scheme

The numerical scheme we use for this problem is the SSP3-WENO3 scheme specified in (5.16) for the temporal component and in §5.3.13 for the spatial component.

7.4.12 Remark. (WENO schemes for system) As we noted in Rem. 5.3.15, there are two approaches for formulating the WENO scheme for systems: the component-wise approach and the characteristic decomposition approach. In the component-wise approach the reconstruction procedure is applied to each component of the solution to obtain the scheme, as was done in the scalar case.

In more challenging problems, e.g., highly non-linear problems or problems involving multiple shocks and discontinuities, the simple component-wise approach may not prevent spurious oscillations (see [Shu20, §4.1.2]). In this case, a more robust (albeit more computationally expensive) approach, is the characteristic decomposition procedure, which is described in detail in [Shu98, Procedure 2.10]. Briefly, this procedure involves the application of the reconstruction procedure to each component of the characteristic variables and subsequently the transformation back to physical space. The reader should note that we do not use this approach here.

7.5 A posteriori error bounds

In this section we present the a posteriori error results we will use in this chapter. These include an a posteriori error bound for Burgers' equation specifically which holds in the pre-shock regime, a Kruzkov entropy based bound from [Ohl09] which also holds in the post-shock regime and a relative-based error bound from [GMP15] for systems of conservation law in one spatial dimension which admit a convex entropy.

The first way is to facilitate an alternative error interpretation. This alternative error - measured in an appropriate, easily computable norm - is amenable to rigorous a posteriori error control, which is in turn facilitated by an appropriate stability framework. The second way the reconstruction is used, is to apply to it the differential equation in order to obtain a reconstruction residual. The reconstruction residual is then used in the computation of the a posteriori error bound.

7.5.1 Theorem (a posteriori error control for non-linear systems of 1D conservation laws from [GMP15])**.** *Let $V \subset \mathbb{R}^d$ be a convex state space. Let $\boldsymbol{f} \in \mathrm{C}^2\big(V, \mathbb{R}^d\big)$ satisfy (6.13) and let $\boldsymbol{u}$ be an entropy solution of (5.1) with periodic boundary conditions. Let $\widehat{\boldsymbol{U}}^{ts}$ take values in $\mathcal{D}$ (which is a convex, compact subset of the state space, V). Then for $0 \leq t \leq T$ the error between $\boldsymbol{u}$ and $\widehat{\boldsymbol{U}}^{ts}$ is given by*

$$\begin{aligned}\left\|\boldsymbol{u}(\cdot,t)-\widehat{\boldsymbol{U}}^{ts}(\cdot,t)\right\|^2_{\mathrm{L}^2(I)} \leq & C_{\underline{\eta}}^{-1}\left(\|\boldsymbol{R}\|^2_{\mathrm{L}^2(I\times(0,t))}+C_{\overline{\eta}}\left\|\boldsymbol{u}_0-\widehat{\boldsymbol{U}}^{ts}_0\right\|^2_{\mathrm{L}^2(I)}\right) \\ & \times \exp\left(\int_0^t \frac{\left(C_{\overline{\eta}}C_{\overline{\boldsymbol{f}}}\left\|\partial_x\widehat{\boldsymbol{U}}^{ts}(\cdot,s)\right\|_{\mathrm{L}^\infty(I)}+C^2_{\overline{\eta}}\right)}{C_{\underline{\eta}}}\mathrm{d}s\right)\end{aligned} \tag{7.32}$$

The constants $C_{\underline{\eta}}$ and $C_{\overline{\eta}}$ represent the minimum and maximum absolute eigenvalues of $\mathrm{D}^2\eta$. Furthermore, $C_{\overline{\boldsymbol{f}}} := \left(\sum_i C_{\overline{\boldsymbol{f}_i}}\right)^{1/2}$, where $C_{\overline{\boldsymbol{f}_i}}$ is an upper bound for the absolute values of the eigenvalues of the ith component of $\boldsymbol{f}$.

Proof. See [GMP15, Thm.5.5] □

7.5.2 Remark. The reader should note that the Gibbs phenomenon does not cause problems in the blow-up rate Thm. 7.5.1.

7.5.3 Lemma (Stability and error control for Burgers' equation with periodic boundary conditions)**.** *Let the conditions of Lemma 7.5.1 hold with $f(u) := \frac{1}{2}u^2$, i.e.*

the scalar Burgers' equation. Suppose the initial value problem for u and $\widehat{U}^{ts}$ are coupled with periodic boundary data. Then, the error between the two functions, $e := u - \widehat{U}^{ts}$, satisfies the following bound for all $t \in [0,T]$:

$$\begin{aligned}\|e(t)\|^2_{\mathrm{L}^2(\Omega)} &\leq \exp\left(\int_0^t \left(\left\|\partial_x \widehat{U}(s)\right\|_{\mathrm{L}^\infty(\Omega)} + 1\right) \mathrm{d}s\right)\left[\|e(0)\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \|R(s)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s\right] \\ &=: \omega_b(t)\,\mathscr{E}_b^2\big(t;\mathrm{L}^2(\Omega)\big)\,,\end{aligned} \tag{7.33}$$

where

$$\omega_b(t) := \exp\left(\int_0^t \left(\left\|\partial_x \widehat{U}(s)\right\|_{\mathrm{L}^\infty(\Omega)} + 1\right) \mathrm{d}s\right) \tag{7.34}$$

and

$$-R := \partial_t \widehat{U} + \partial_x\left(\frac{\widehat{U}^2}{2}\right). \tag{7.35}$$

Proof. Let $u(x,t)$ solve

$$\begin{aligned} u_t + \partial_x\left(\frac{u^2}{2}\right) &= 0 \\ u(x,0) &= u_0(x)\end{aligned} \tag{7.36}$$

with periodic boundary conditions. Let v denote the reconstruction, $\widehat{U}$, which satisfies the perturbed PDE

$$\begin{aligned} v_t + \partial_x\left(\frac{v^2}{2}\right) &= -R \\ v(x,0) &= v_0(x)\,.\end{aligned} \tag{7.37}$$

Then, the error $e := u - v$ satisfies

$$\begin{aligned} e_t + \partial_x\left(\frac{(u+v)(u-v)}{2}\right) &= R \\ e(x,0) &= e_0(x)\,.\end{aligned} \tag{7.38}$$

We use the fact that $u + v = u + v - v + v = e + 2v$ to rewrite this as

$$\begin{aligned} e_t + \partial_x\left(\frac{(e+2v)(e)}{2}\right) &= R \\ e(x,0) &= e_0(x)\,.\end{aligned} \tag{7.39}$$

We test with (7.39) e to obtain

$$\begin{aligned} \int Re &= \int ee_t + \frac{1}{2}e\partial_x\big(e^2 + 2ve\big) \\ \int Re &= \int ee_t + \int \frac{1}{3}\partial_x\big(e^3\big) + \int e\partial_x(ve)\end{aligned} \tag{7.40}$$

We note that

$$\frac{1}{2}\partial_x\big(ve^2\big) = vee_x + \frac{1}{2}e^2 v_x \tag{7.41}$$

and that

$$\begin{aligned} e\partial_x(ve) &= vee_x + e^2 v_x \\ &= \frac{1}{2}\partial_x\big(ve^2\big) + \frac{1}{2}e^2 v_x. \end{aligned} \tag{7.42}$$

Hence

$$\frac{1}{2}\frac{\mathrm{d}}{dt}\left\|e\right\|^2_{\mathrm{L}^2(\Omega)} = \int Re + \int \frac{1}{2}e^2 v_x \tag{7.43}$$

We use the Cauchy-Schwarz and Holder's inequality to obtain

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{dt}\left\|e\right\|^2_{\mathrm{L}^2(\Omega)} &\leq \left\|R\right\|_{\mathrm{L}^2(\Omega)}\left\|e\right\|_{\mathrm{L}^2(\Omega)} + \frac{1}{2}\left\|e\right\|^2_{\mathrm{L}^2(\Omega)}\left\|v_x\right\|_{\mathrm{L}^\infty(\Omega)} \\ &\leq \frac{1}{2}\left\|R\right\|^2_{\mathrm{L}^2(\Omega)} + \frac{1}{2}\left\|e\right\|^2_{\mathrm{L}^2(\Omega)} + \frac{1}{2}\left\|e\right\|^2_{\mathrm{L}^2(\Omega)}\left\|v_x\right\|_{\mathrm{L}^\infty(\Omega)}. \end{aligned} \tag{7.44}$$

Finally now use Gronwall's inequality to obtain the desired result:

$$\left\|e(t)\right\|^2_{\mathrm{L}^2(\Omega)} \leq \exp\!\left(\int_0^t \Big(\left\|v_x(s)\right\|_{\mathrm{L}^\infty(\Omega)} + 1\Big)\,\mathrm{d}s\right)\left[\left\|e(0)\right\|^2_{\mathrm{L}^2(\Omega)} + \int_0^t \left\|R(s)\right\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s\right]. \tag{7.45}$$

□

7.5.4 Remark. Note that the result given in Lemma 7.5.3 loses its robustness in the post-shock regime due to the presence of the reconstruction's derivative, which blows up upon shock formation, in the exponential accumulation factor, see (7.33). The next result on a posteriori error control is from [CG95]. It pertains to a posteriori error control for a scalar, non-linear hyperbolic problem and it makes use of the Kruzkov doubling of variable technique (see [Kru70],[Kuz76]). This estimate retains robustness in the post-shock regime. We introduce it in order to combine it with the a posteriori error estimate for the scalar problem and present an estimate that is optimal in both the pre-shock and post-shock regimes.

7.5.5 Theorem. *($\mathrm{L}^1(\Omega)$ error bound - [CG95, Thm. 2.1]) Let u be an entropy solution of (7.1) and denote by $\widehat{U}^{ts}$ the reconstruction of the numerical solution $\big\{U_j^n\big\}_j^n$ (see Defn.7.4.7). Then,*

$$\begin{aligned} \Big\|u(t) - \widehat{U}^{ts}(t)\Big\|_{\mathrm{L}^1(\Omega)} &\leq \left\|u_0 - v_0\right\|_{\mathrm{L}^1\left(0,t;\mathrm{L}^1(\Omega)\right)} + \int_0^t \left\|R(s)\right\|_{\mathrm{L}^1(\Omega)}\,\mathrm{d}s \\ &=: \omega_a(t)\,\mathcal{E}_a\big(t;\mathrm{L}^1(\Omega)\big), \end{aligned} \tag{7.46}$$

where

$$\omega_a(t) = 1 \quad \forall t \tag{7.47}$$

and

$$R := \partial_t \widehat{U}^{ts} + \partial_x\big(f\big(\widehat{U}^{ts}\big)\big) \tag{7.48}$$

is the discrete residual of the reconstruction.

Proof. See [CG95, Appendix]. □

7.5.6 Remark. By combining Lemma 7.5.3 and Theorem 7.5.5 we obtain an a posteriori error estimate that is optimal in both the pre-shock and post-shock regime for the scalar problem, (7.1), in one spatial dimension.

7.5.7 Theorem (Robust bound pre and post shock)**.** *Let the conditions of Lemma 7.5.3 and Theorem 7.5.5 hold. Then, the following error estimate holds:*

$$\|e\|_{\mathrm{L}^1(\Omega)} \leq \min\{\omega_b(t)\mathscr{E}_b\big(t, \mathrm{L}^1(\Omega)\big)\,, |\Omega|^{1/2}\,\omega_a(t)\mathscr{E}_a\big(t, \mathrm{L}^2(\Omega)\big)\}, \tag{7.49}$$

where the subscript $(\cdot)_b$ corresponds to the bound variables of Lem. 7.5.3 in the pre-shock regime and the subscript $(\cdot)_a$ corresponds to the post-shock regime of Thm. 7.5.5.

Proof Notice that for any $w \in \mathrm{L}^2(\Omega)$

$$\|w\|_{\mathrm{L}^1(\Omega)} = \int_\Omega |w| \leq |\Omega|^{1/2}\,\|w\|_{\mathrm{L}^2(\Omega)}\,. \tag{7.50}$$

The result follows from the fact that the $\mathrm{L}^2(\Omega)$ estimate, Lem. 7.5.3, contains an exponential term which contains the derivative of the reconstruction. □

7.5.8 Remark. The presence of an exponential term involving the spatial derivative of the reconstruction in the $\mathrm{L}^2(\Omega)$ estimate in Lem. 7.5.3 causes the estimate to blow up in the presence of shocks. At that point, taking the minimum between the two results will mean that the $\mathrm{L}^1(\Omega)$ result will be chosen.

7.6 Numerical verification

In this section we will study the asymptotic behaviour of the a posteriori bound and compare and contrast it with the behaviour of the error, defined as

$$\boldsymbol{e} := \boldsymbol{u} - \widehat{\boldsymbol{U}}, \tag{7.51}$$

for a linear example, the transport problem, and two non-linear examples: Burgers' equation and the shallow water equations. The reason for including the linear example is to investigate the behaviour of the estimates in situations where we can conveniently examine solutions of different regularity before proceeding to the non-linear problems.

The tests in this section are a preliminary step before the next section, in which the a posteriori bound is used as a refinement/coarsening criterion in adaptive tests. We will firstly present the bounds we will be testing for the non-linear problems we benchmark in this section, along with the numerical schemes we will use.

7.6.1 Remark (a posteriori bound for non-linear problems)**.** Some of the test cases we examine in this section are non-linear so the post-processor from Lemma 4.3.2 is not appropriate. Instead, we will use a different a posteriori estimate which is appropriate for nonlinear systems of hyperbolic conservation laws.

7.6.2 Remark (Reconstruction residual)**.** The reconstruction residual, $\boldsymbol{R}$, is used to compute the smooth post-processor that bounds the error of the problem from above in Thm. 4.3.2. We obtain $\boldsymbol{R}$ by substituting $\widehat{\boldsymbol{U}}^{ts}$ in (5.1):

$$-\widehat{\boldsymbol{R}}(x,t) := \partial_t \widehat{\boldsymbol{U}}^{ts}(x,t) + \partial_x \boldsymbol{f}\Big(\widehat{\boldsymbol{U}}^{ts}(x,t)\Big) \tag{7.52}$$

The reader may notice that $\widehat{\boldsymbol{U}}^{ts}(x,t)$ may be more difficult to compute than the solution. Nonetheless, this is a worthwhile endeavour from a computational perspective because of the utility of a (locally) coarser grid. In this case, $\widehat{\boldsymbol{U}}^{ts}(x,t)$ can be used in an a posteriori estimate estimate, a driver for adaptivity. In turn this can lead to reduction of the necessary degrees of freedom in the computational domain where appropriate. If the regions of the domain where high resolution are necessary are highly localised, it may be the case that the extra computational expense incurred in the computation of $\widehat{\boldsymbol{U}}^{ts}(x,t)$ is justified by the reduction in computational resource usage.

We test the bound (7.33) for three different schemes - Lax-Friedrichs, Lax-Wendroff and SSP3-WENO - with uniform temporal and spatial discretizations, i.e. $\tau^n := \tau \ \forall n$ and $h_j := h \ \forall j$. The reconstruction residual, (7.52), will be obtained by using Defn. 7.4.5 for the temporal component and Defn. 7.4.7 for the spatial component.

7.6.3 Test 1: Advection equation

In this section, the model problem we test is the advection equation with periodic boundary conditions, see (4.2). The objective of the test is to compare the performance of the a posteriori error bound of [CG95] with the a posteriori error bound we presented in Cor. 4.4.6.

We discretise the model problem, (4.2), using a simple FTBS scheme, (4.9), a Lax-Friedrichs (LxF) scheme (see (5.6)-(5.7)) as well as a SSP3 - WENO3 scheme, for which we provide the temporal discretisation as a Butcher tableaux in Tbl. 5.16 and the spatial discretisation in §5.3.13.

In the FTBS and LxF tests we obtain the reconstruction, $\widehat{U}^{ts}$ by using a piecewise bilinear interpolant of the numerical solution, $\{U_j^n\}_j^n$. In the SSP3-WENO3 tests, the reconstruction is obtained as a Hermite-in-time, WENO3-in space interpolant, as explained in Defns. 7.4.5 and 7.4.7.

We test three initial conditions: a smooth initial condition given by

$$u_0(x) = \sin(2\pi x), \tag{7.53}$$

a piecewise linear continuous initial condition given by

$$u_0(x) = \begin{cases} 1 - 4\left|x - \frac{1}{4}\right| & \text{for} \quad |x| <= 0.25 \\ 0 \quad \text{otherwise,} \end{cases} \tag{7.54}$$

and a discontinuous initial condition given by

$$u_0(x) = \begin{cases} 1 & \text{for} \quad \left|x - \frac{1}{4}\right| \leq \frac{1}{8}, \\ 0 & \text{otherwise.} \end{cases} \tag{7.55}$$

We will use periodic boundary conditions in all tests. In all these cases the exact solution is given by

$$u(x,t) = u_0(x-t). \tag{7.56}$$

The results from each test will be presented in figures consisting of three sub-figures, each sub-figure corresponding to an initial condition. Each sub-figure will in turn be composed of two rows of plots: in the top row the spatial components of the error are calculated in the $\mathrm{L}^1(\Omega)-$norm and the estimate is given in Thm. 7.5.7. In the bottom row the spatial components of the error are calculated in the $\mathrm{L}^2(\Omega)-$norm

and the estimate used is given in Cor. 4.4.6. The simulations are conducted over a family of meshes with discretisation parameter $h = 2^{-m}$, $m = 7, \ldots, 10$, with a temporal step $\tau = h/10$.

Discussion

The results are shown in Fig. 7.2 for the FTBS scheme, in Fig. 7.3 for the Lax-Friedrichs scheme and in Fig. 7.4 for the SSP3-WENO3 scheme. The reader should note that in each of the figure's subplots, the top row pertains to the $\mathrm{L}^1(\Omega)$ estimate, (7.46), while the bottom row corresponds to the $\mathrm{L}^2(\Omega)$ estimate from Cor. 4.4.6.

We observe that, in the case of the FTBS scheme, Fig. 7.2, the $\mathrm{L}^1(\Omega)$ bound, (7.46), is robust and converges optimally for all three initial conditions, (7.53), (7.54) and (7.55). The $\mathrm{L}^2(\Omega)$ estimate from Cor.4.4.6 is robust for the smooth and piecewise linear initial conditions, but it loses robustness for the step initial condition (7.55).

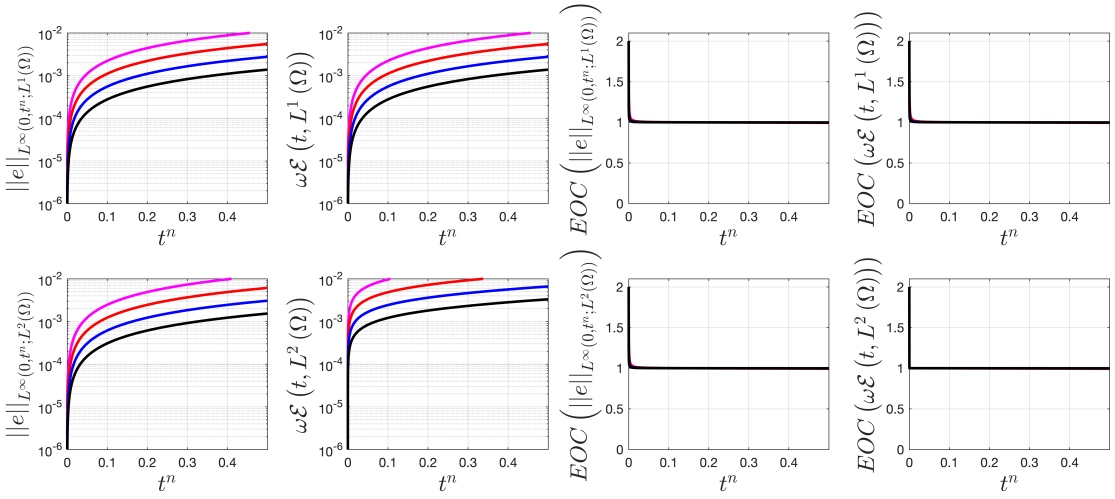
In the case of the Lax-Friedrichs scheme, Fig. 7.3, we observe behaviour similar to the FTBS scheme in the smooth and piecewise linear cases. The notable difference between the two is in the non-robustness of the $\mathrm{L}^1(\Omega)$ bound in the case of the Lax-Friedrichs scheme with the step initial condition. In contrast, in the corresponding FTBS result we observe robustness. This behaviour was unexpected as the routines for the error calculation, the reconstruction and the residual computation are the same for both cases. The distinguishing feature between the two is the spatial discretisation, which nonetheless yields the same error behaviour (asymptotically) for the two schemes. Hence, it is reasonable to expect similar behaviour for the estimates, which is not the case as can be seen in Figs. 7.2c with 7.3c.

In deciphering the difference in behaviour between LxF and FTBS in the step-initial condition test, we note that the feature that stand out first is the spatial discretisation, which is the only essential difference between the two cases. We believe that, possibly, the reason for the difference may lie in the fact that the bilinear reconstruction we employ may not possess the requisite approximability to accommodate the underlying LxF flux as well as it does the FTBS flux. This does not manifest in the smooth and hat test-cases but it does become apparent in the case of the relatively more challenging step-initial condition.

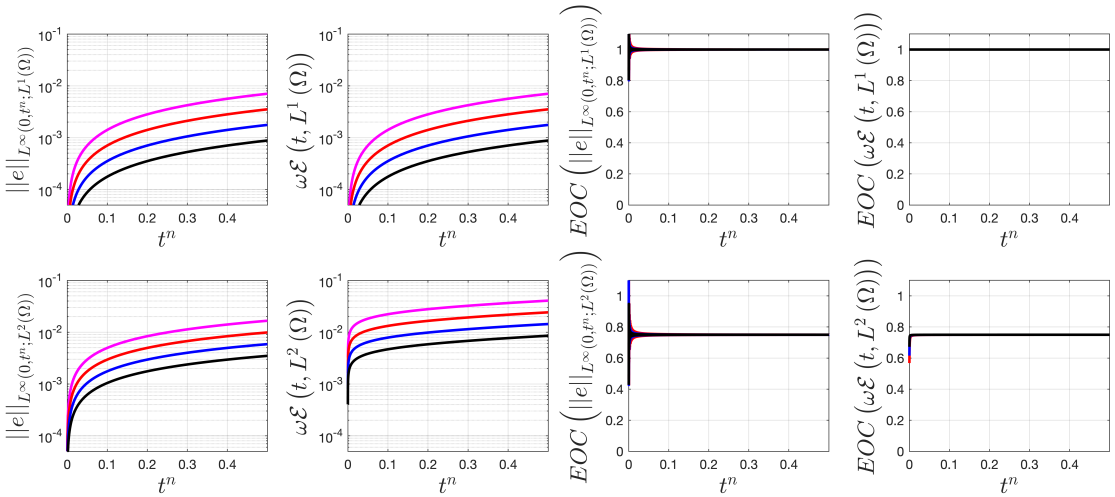
In particular, as we can see in Fig. 7.3c, the estimate actually diverges. This

results by a lack of balance of relative contribution to the residual between the temporal and spatial component of the reconstruction. This may be the result of the fact that, by using a bilinear interpolant, the information from the LxF flux is not incorporated in the reconstruction. In contrast, in the case of the FTBS flux, which involves a backward difference of the numerical solution, the bilinear interpolant does incorporate this information to some degree (though coincidentally and due to the form of the particular scheme).

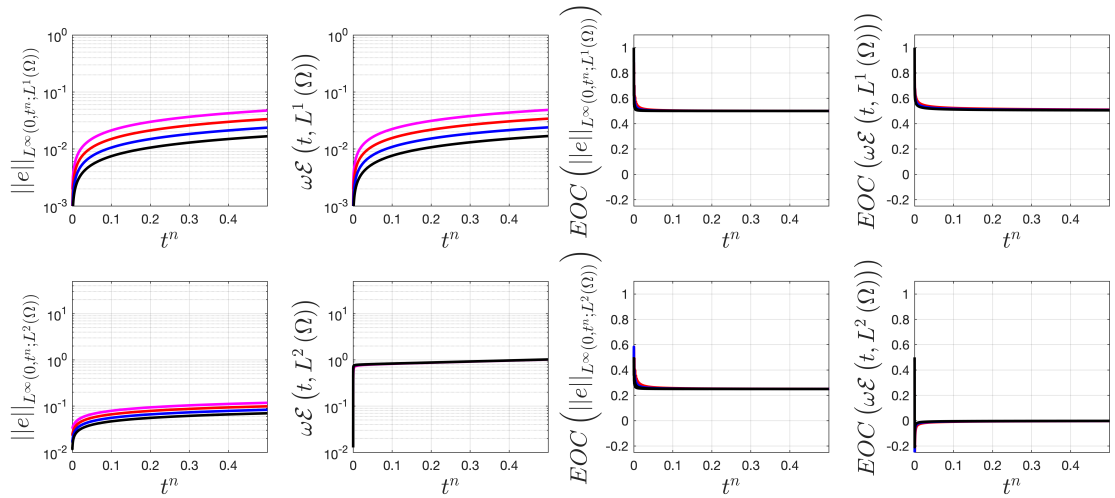
Lastly, for the SSP3-WENO3 scheme, Fig. 7.4, we observe that both estimates are robust and converge for the smooth initial condition. For the piecewise linear initial condition, the $\mathrm{L}^1(\Omega)$ estimate is robust while the $\mathrm{L}^2(\Omega)$ bound seems to be slightly sub-optimal, although this may be because the resolution is not fine enough for the bound to enter the asymptotic convergence regime. However, in the case of the step initial condition, while the $\mathrm{L}^1(\Omega)$ estimate is optimal, the $\mathrm{L}^2(\Omega)$ one loses optimality.



(a) Smooth initial condition, (7.53).

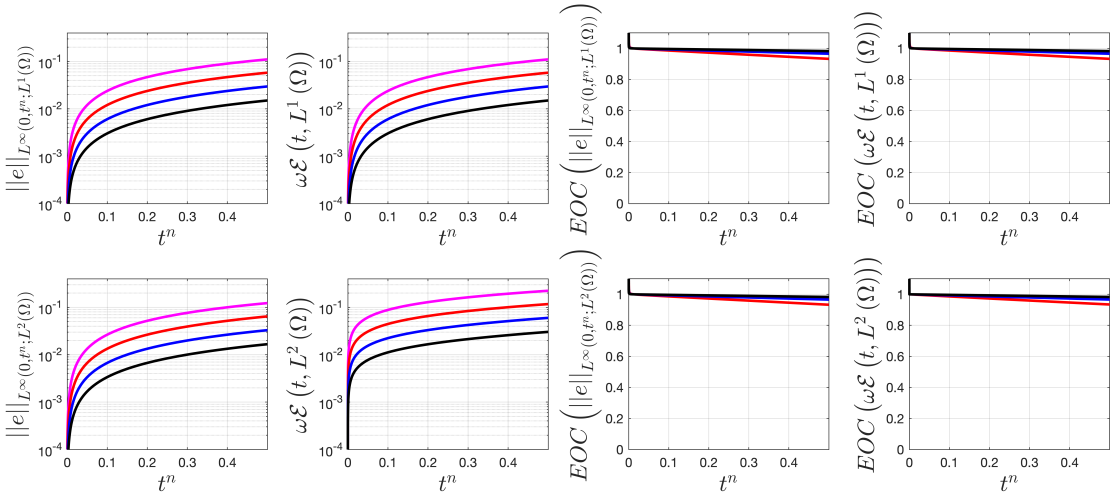


(b) Hat initial condition, (7.54) .

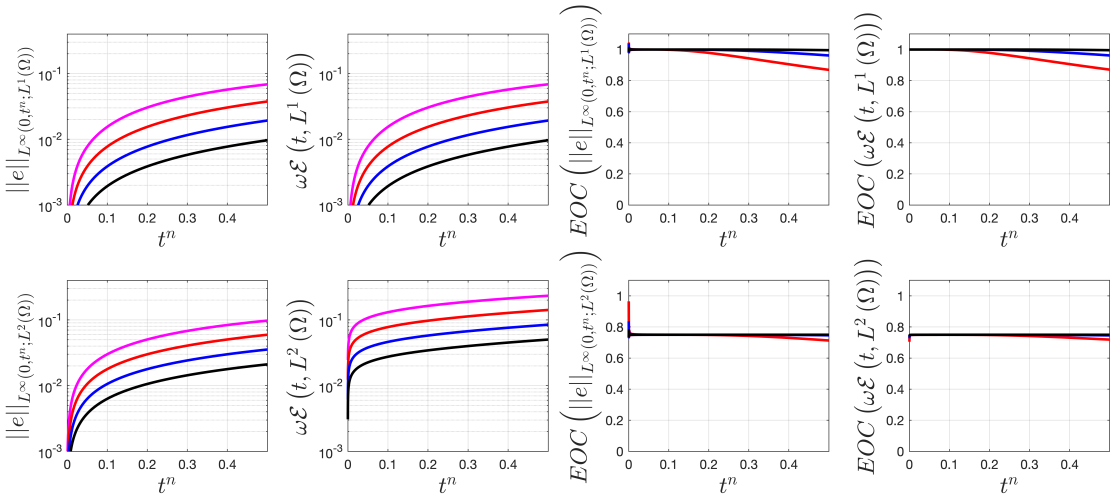


(c) Step initial condition, (7.55).

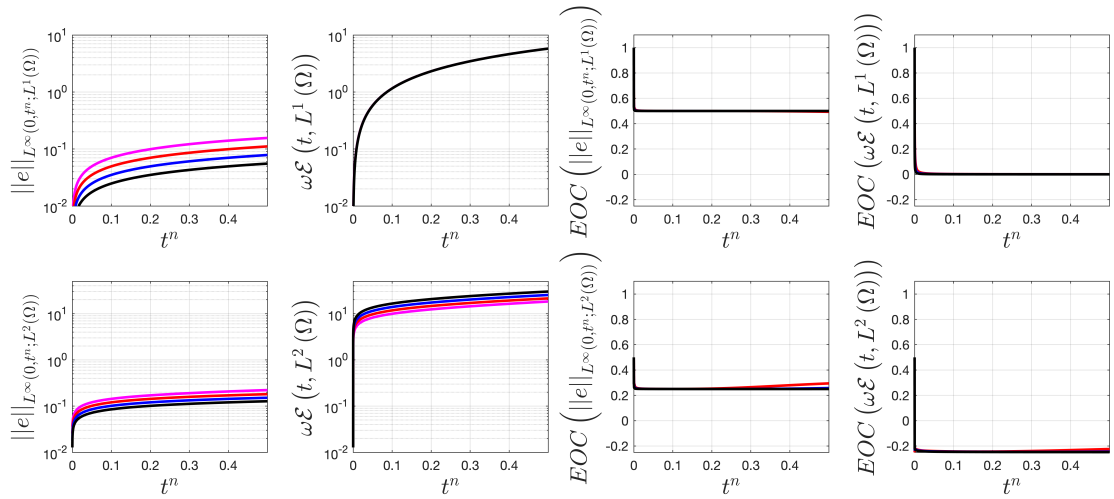
Fig. 7.2. Errors and asymptotic convergence rates for an estimate constructed using a bilinear interpolant for the FTBS scheme, (4.9), for the transport equation. Both estimates are optimal for smooth and continuous initial conditions, as shown in Figs. 7.2a, 7.2b, but the $L^2(\Omega)$ estimate from Cor. 4.4.6 loses optimality for discontinuous initial conditions, see Fig. 7.2c.



(a) Smooth initial condition, (7.53).

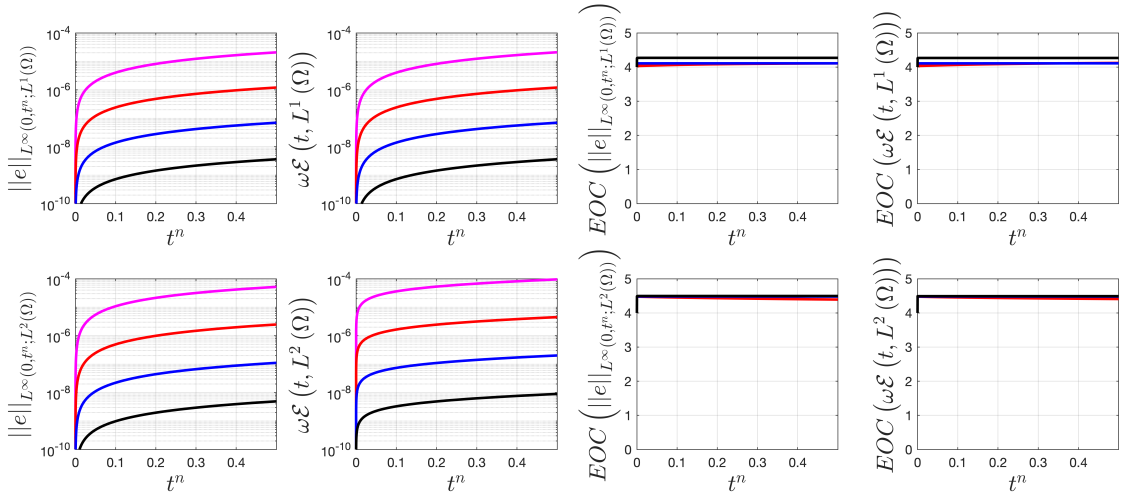


(b) Hat initial condition, (7.54) .

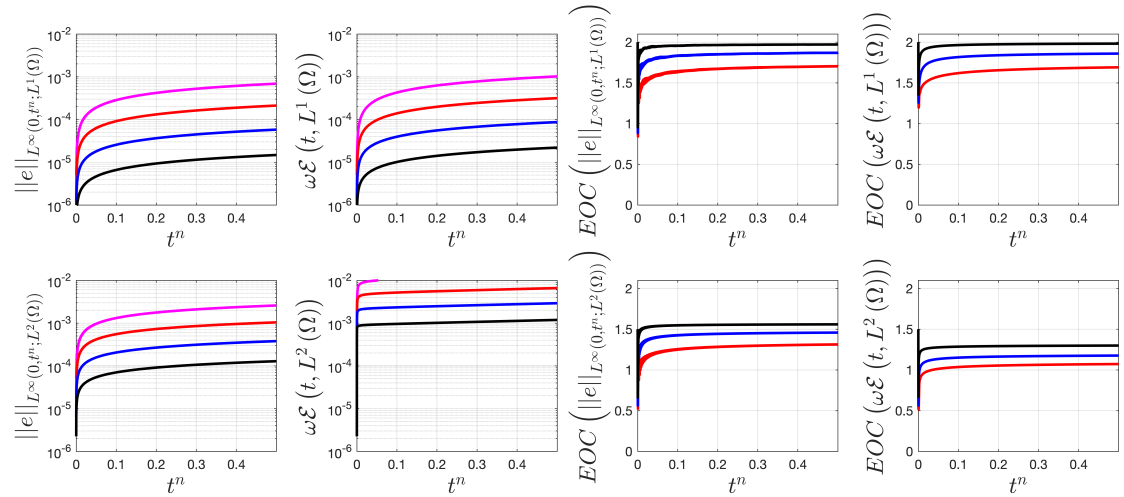


(c) Step initial condition, (7.55).

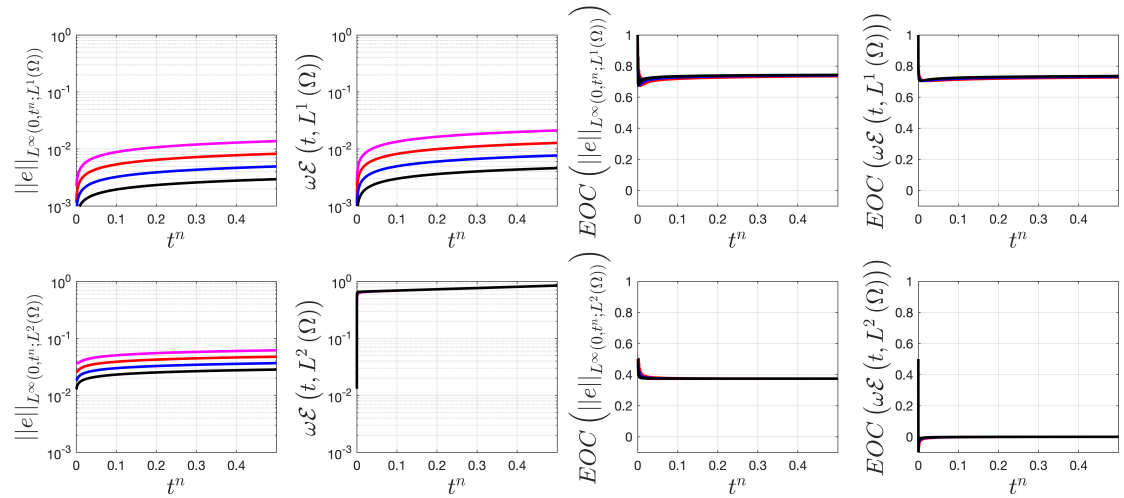
Fig. 7.3. Errors and asymptotic convergence rates for an estimate constructed using a bilinear interpolant for the LxF scheme, (see (5.6)-(5.7)), for the transport equation. Both estimates are optimal for smooth and continuous initial conditions, as shown in Figs. 7.3a, 7.3b, but the $L^2(\Omega)$ estimate from Cor. 4.4.6 loses optimality and diverges for discontinuous initial conditions, see Fig. 7.3c.



(a) Smooth initial condition, (7.53).



(b) Hat initial condition, (7.54) .



(c) Step initial condition, (7.55).

Fig. 7.4. Errors and asymptotic convergence rates for an estimate constructed using a Hermite-in-time, WENO-in-space interpolant for the SSP3-WENO3 scheme, (Tbl. 5.16 and §5.3.13), for the transport equation. Both estimates are optimal for smooth and continuous initial conditions, as shown in Figs. 7.4a, 7.4b, but the $L^2(\Omega)$ estimate from Cor. 4.4.6 loses optimality for discontinuous initial conditions, see Fig. 7.4c.

7.6.4 Test 2: Scalar Inviscid Burgers' equation.

In this section, we benchmark the performance of the a posteriori estimate (7.33) on the basis of a non-linear scalar problem. The model problem we use is the one-dimensional inviscid Burgers' equation with periodic boundary conditions and a prescribed initial condition:

$$\begin{aligned} \partial_t u + \partial_x\left(\frac{u^2}{2}\right) &= 0, \\ u(x,0) &= u_0(x) \end{aligned} \qquad \text{for } (x,t) \in [-\pi,\pi]\times(0,T]. \tag{7.57}$$

We discretise the model problem, (4.2), a Lax-Friedrichs (LxF) scheme (see (5.6)-(5.7)), a Lax-Wendroff (LxW) scheme, (5.10), well as a SSP3 - WENO3 scheme, (see Tbl. 5.16 for the temporal discretisation and §5.3.13 for the spatial discretisation).

In the LxF tests we obtain the reconstruction, $\widehat{U}^{ts}$ as a piecewise bilinear interpolant of the numerical solution, $\{U_j^n\}_j^n$. In the LxW and SSP3-WENO3 tests, the reconstruction is obtained as a Hermite-in-time, WENO3-in space interpolant, as explained in Defns. 7.4.5 and 7.4.7.

We use two different initial conditions. We use a sinusoidal initial condition given by

$$u_0(x) = -\sin(x), \tag{7.58}$$

and a piecewise linear initial condition given by

$$u_0(x) = \begin{cases} 1-|x| & \text{for} \quad |x|\leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7.59}$$

In the case of (7.58), the exact solution in the pre-shock regime can be represented as an infinite sum of Bessel functions (see [GMP15]):

$$u(x,t) = -2\sum_{k=1}^{\infty}\frac{J_k(kt)}{kt}\sin(kx), \tag{7.60}$$

where J_k denotes the kth Bessel function. This is a decaying sequence, so we can approximate the solution by truncating it (we truncate at $k = 100$). In the post-shock regime, we do not have an analytic formula for the solution.

In the case of the piecewise linear initial condition, (7.59) we obtain the exact solution using the method of characteristics. In the pre-shock regime, the exact

solution is given by

$$\begin{cases} 0 \quad \text{for} \quad x < -1, \\ \frac{1+x}{1+t} \quad \text{for} \quad -1 \leq x < t, \\ \frac{1-x}{1-t} \quad \text{for} \quad t \leq x < 1, \\ 0 \quad \text{otherwise}, \end{cases} \tag{7.61}$$

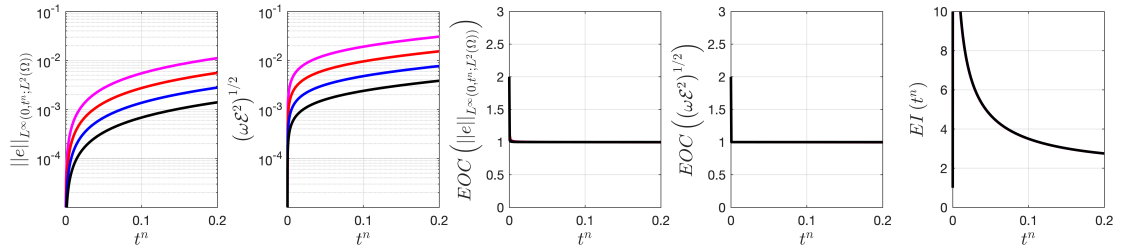
while in the post-shock regime, i.e $t \geq 1$, it is given by

$$u(x,t) = \begin{cases} 0 \quad \text{for} \quad x < -1, \\ \frac{1+x}{1+t} \quad \text{for} \quad -1 \leq x \leq \sqrt{2(1+t)} - 1, \\ 0 \quad \text{otherwise}. \end{cases} \tag{7.62}$$

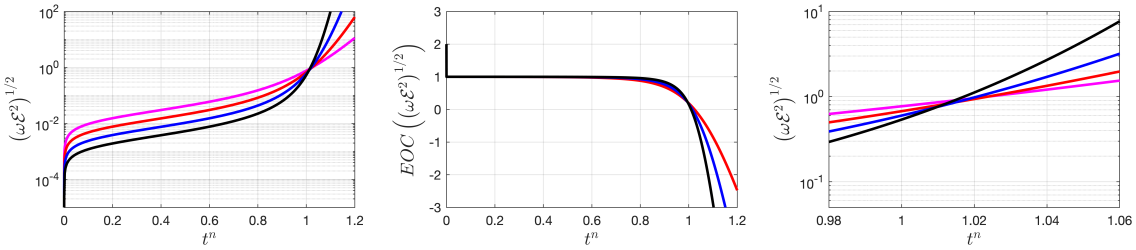
The tests we perform aim to investigate the behaviour of the estimates prior and subsequent to shock formation.

In Figs. 7.5, 7.6 and 7.7 we show the behaviour of the error, $e := u - \widehat{U}^{ts}$ and the estimate (7.33) before and after shock formation, using the sinusoidal initial condition (7.58). In Fig. 7.8 we decouple the residual component and the exponential accumulation factor in (7.33) and investigate them separately.

All simulations are conducted over a family of meshes with discretisation parameter $h = 2^{-m}$, $m = 11, \dots 14$, with a temporal step $\tau = h/10$. The results in each figure will pertain to individual schemes and they will be presented in two rows of sub-figures (within the same figure): a top row corresponding to the pre-shock behaviour and a bottom row corresponds to the post-shock behaviour.

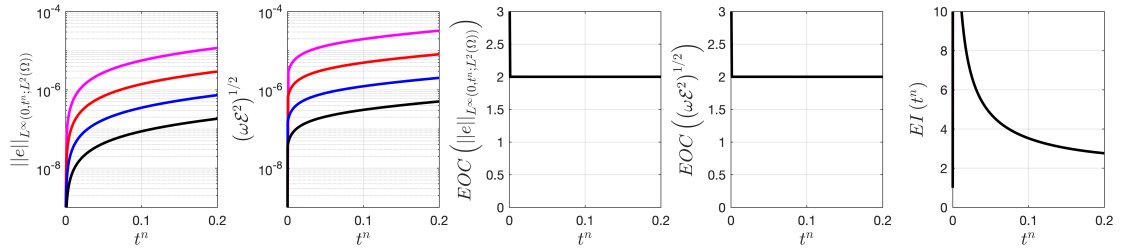


(a) Pre-shock.

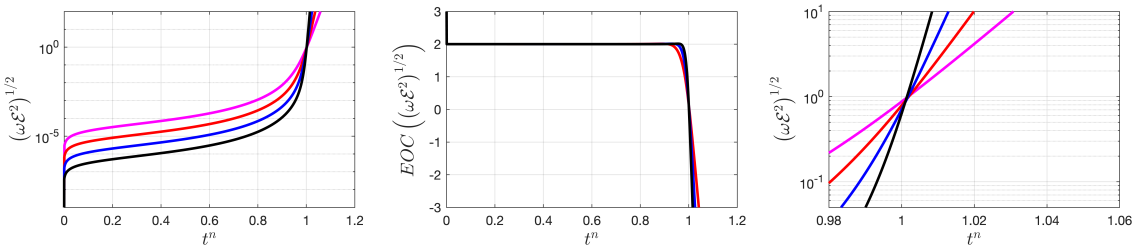


(b) Post-shock.

Fig. 7.5. Errors and asymptotic convergence rates for an estimate constructed using a bilinear reconstruction for the Lax-Friedrichs scheme, (5.7), for Burgers' equation with sinusoidal initial conditions. The estimate is optimal prior to shock formation and blows up once the shock forms due to the exponential factor in (7.33).

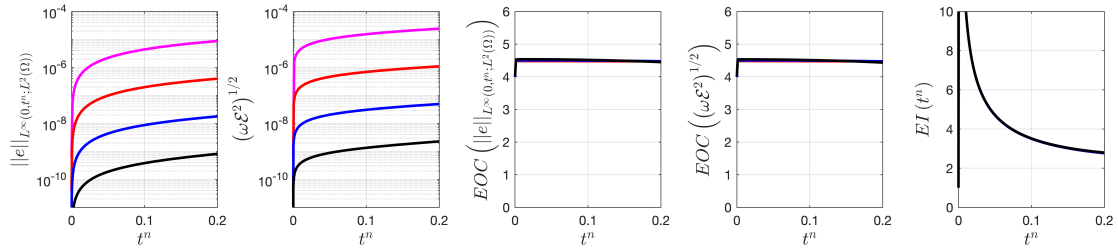


(a) Pre-shock.

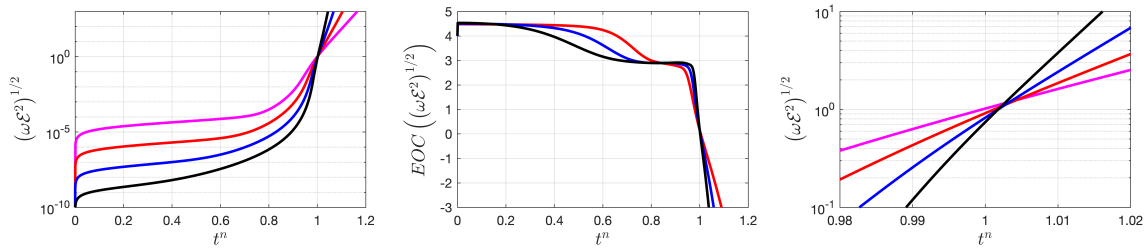


(b) Post-shock.

Fig. 7.6. Errors and asymptotic convergence rates for a Hermite-in-time-WENO3-in space reconstruction for the Lax-Wendroff scheme, (5.10), for Burgers' equation with sinusoidal initial conditions. The estimate is optimal prior to shock-formation and blows up post-shock because of the exponential factor in (7.33).



(a) Pre-shock.



(b) Post-shock.

Fig. 7.7. Errors and asymptotic convergence rates for a Hermite-WENO3 reconstruction for the SSP3-WENO3 discretisation of Burgers' equation with a sinusoidal initial condition. Notice that the estimate loses optimality gradually in the interval $0.2 \leq t \leq 0.8$. This is because the solution itself is losing regularity, only this time the scheme and the residual are both of sufficiently high order to capture this. At $t \approx 1$ the shock forms and the exponential factor in (7.33) blows up.

In order to explain the behaviour of the estimate (7.33) we decouple the time accumulation factor from the residual component of the post-processor (see Fig. 7.8). Notice that at $t \approx 1$ the exponential factor blows up rapidly as the spatial derivative of the reconstruction, $\partial_x \widehat{U}^{ts}$ blows up. The reason for this is that at $t = 1$ the solution starts forming a shock. This explains the behaviour for $t \geq 1$.

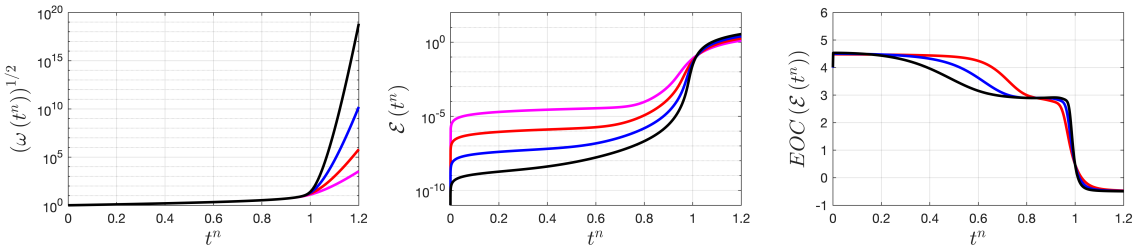


Fig. 7.8. Decoupling for the post-processor for the SSP3-WENO3 approximations of Burgers' equation with a sinusoidal initial condition. The blow-up in the time accumulation term is because of the spatial derivative in the exponential factor.

7.6.5 Test 2: Comparison of estimates

In this section we compare the performance of the estimates (7.33) and (7.46) using Burgers equation as a model problem. We use a sinusoidal initial condition given by (7.58) and a piecewise linear initial condition given by (7.59). We have analytical expressions for the exact solutions for both initial conditions in the pre-shock regime but in the post-shock regime we only have an exact solution for (7.59).

The motivation behind the tests in this section is to investigate and compare the pre-shock and post-shock performance of the two estimates. The numerical discretisation of the problems will be done using the Lax-Friedrichs and SSP3-WENO3 schemes we introduced earlier. The respective reconstructions are obtained using a bilinear interpolant for the Lax-Friedrichs scheme and a Hermite-WENO3 interpolant for the SSP3-WENO3 scheme. The results for all experiments will be arranged in figures of two-subplots corresponding to the two initial conditions. In turn, each subplot contains two rows of results. In the top row the spatial components of the error are calculated in the $\mathrm{L}^1(\Omega)$ −norm and the estimate is given in Thm. 7.5.7. In the bottom row the spatial components of the error are calculated in the $\mathrm{L}^2(\Omega)$ −norm and the estimate used is given in (7.33).

In all cases the simulations are conducted over a family of meshes with discretisation parameter $h = 2^{-m}, m = 9, \ldots, 12$, with a temporal step $\tau = h/10$.

The results for the Lax-Friedrichs scheme are shown in Figs. 7.9 and 7.11 for the pre-shock and post-shock regimes respectively. The results for the SSP3-WENO3 scheme are shown in Figs. 7.10 and 7.12 for the pre-shock and post-shock regime respectively.

Discussion

In discussing our results for this section, we draw some comparisons, where relevant, with the analogous results from the transport problem in §7.6.3. The reason is that in case of the piecewise linear initial condition for the Burgers problem, (7.59), the solution transitions from being piecewise linear in the pre-shock regime to being discontinuous in the post-shock regime. Since we have solutions of this regularity in the tranport problem (albeit, for two different initial conditions) we will use the behaviour of the error and the estimate in those cases as a reference in order to

assess whether the numerical behaviour in the Burgers pre and post shock regimes for the piecewise linear initial condition is reasonable.

In the case of the Lax Friedrichs scheme, in the pre-shock regime, Fig. 7.9, we observe that both estimates are robust for both the smooth and the piecewise linear initial conditions, Figs. 7.9a and 7.9b respectively. Furthermore, the behaviours of both the error and the estimate are consistent with our observations in the Lax-Friedrichs scheme for the linear transport problem, when we use initial conditions of the same regularity as in the pre-shock regime of the Burgers solution (see Figs. 7.3a-7.3b).

In the post-shock regime, we will only examine the results pertaining to the piecewise linear initial condition, Fig. 7.11b, where we have an expression for the solution. We observe that the error converges in both the $\mathrm{L}^1(\Omega)$ and $\mathrm{L}^2(\Omega)$ cases, with $\mathcal{O}\big(h^{3/4}\big)$ and $\mathcal{O}\big(h^{3/8}\big)$ respectively. In addition, both estimates lose robustness. These results are unexpected on two levels:

1. Firstly, with regard to the EOC of the errors (for both $\mathrm{L}^1(\Omega)$ and $\mathrm{L}^2(\Omega)$), our observations in Fig. 7.11b are not consistent with our results in the transport problem for solutions possessing the same regularity and discretised by the Lax-Friedrichs scheme. Specifically, with reference to Fig. 7.3c the error for the step-initial condition converges in the $\mathrm{L}^1(\Omega)$ case as $\mathcal{O}\big(h^{1/2}\big)$ and in the $\mathrm{L}^2(\Omega)$ as $\mathcal{O}\big(h^{1/4}\big)$.

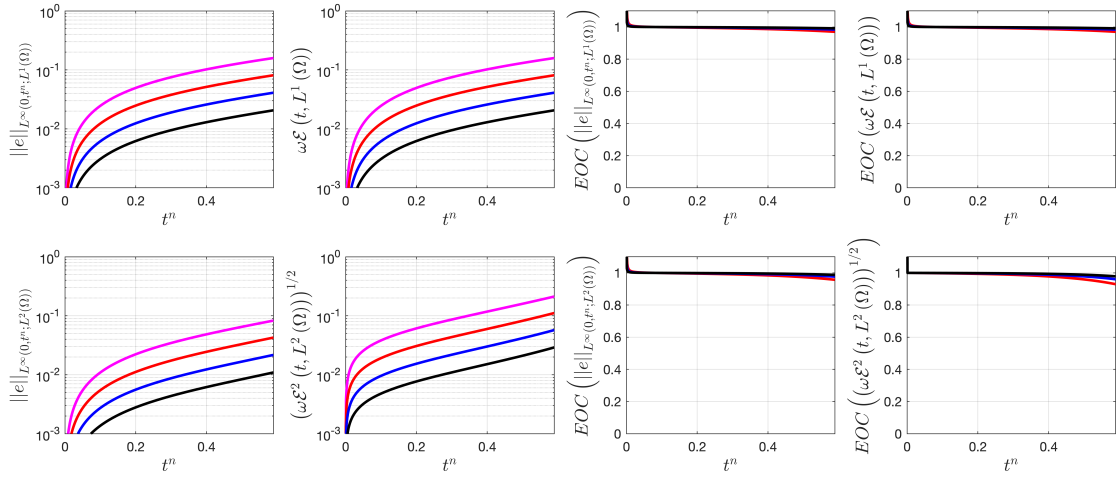
2. Secondly, with regard to the EOC of the estimates, with reference to Fig. 7.11b, we observe that both estimates lose robustness in the post-shock regime. The reason this result is unexpected is the same reason that the corresponding Lax-Friedrichs result for transport was unexpected as well (compare with Fig. 7.3c). That is, because the $\mathrm{L}^1(\Omega)$ estimate loses robustness in this setting whereas it is fully robust in a very similar setting. The setting we refer to is of course the transport problem with the step initial condition, discretised by the FTBS scheme and with the exact same reconstruction routine as in the Burgers problem.

The last component of the results in this section pertains to the SSP3-WENO3 discretisation for the Burgers model problem, (7.57). In the pre-shock regime, for

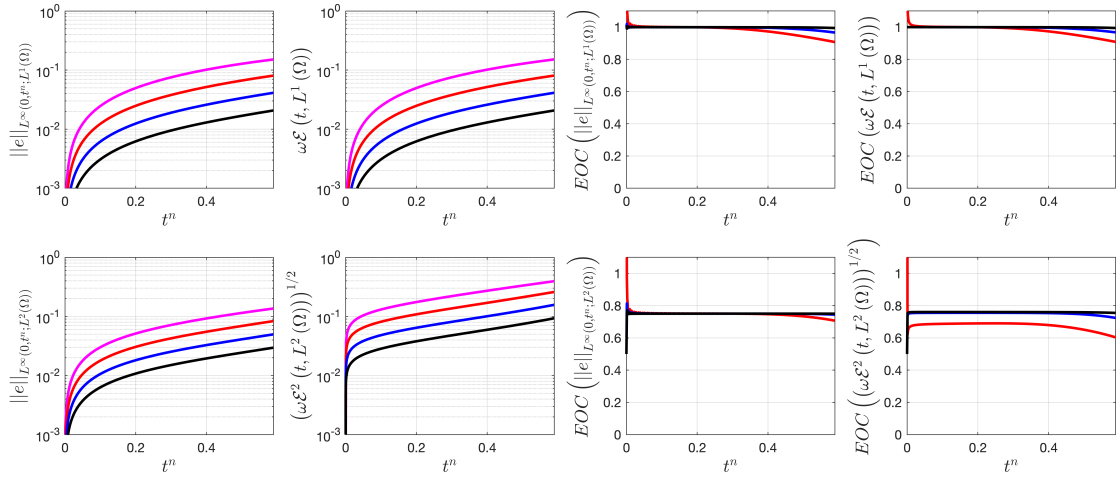
the smooth initial condition, (7.58), with reference to Fig. 7.10a, we observe high orders of convergence for the error and robustness in both the $\mathrm{L}^1(\Omega)$ and $\mathrm{L}^2(\Omega)$ estimates, (7.46) and (7.33) respectively. For the piecewise linear initial condition, (7.59), with reference to Fig. 7.10b we observe that the $\mathrm{L}^1(\Omega)$ estimate, (7.46), is robust while the $\mathrm{L}^2(\Omega)$ estimate, (7.33), appears to be sub-optimal compared to the EOC of the corresponding error, although the estimate's EOC may increase upon further refinement.

In the post-shock regime, both estimates lose robustness for the smooth initial condition (see Fig. 7.12a). This is not unexpected for the $\mathrm{L}^2(\Omega)$ estimate, (7.33) on account of the spatial derivative term in the exponential accumulation factor which goes to infinity when shocks appear. Unfortunately, we cannot ascertain whether this behaviour is justifiable for the $\mathrm{L}^1(\Omega)$ estimate, (7.46), as we do not possess an expression of the solution in the post-shock regime.

In the case of the piecewise linear initial condition, with reference to Fig. 7.12b, we observe that the errors converge with $\mathcal{O}(h)$ for the $\mathrm{L}^1(\Omega)$ case and $\mathcal{O}\big(h^{1/2}\big)$ for the $\mathrm{L}^2(\Omega)$ case. However, both of the respective estimates lose robustness. This behaviour is consistent with our observation in the Lax-Friedrichs case. The source of this behaviour remains an open question at this point.

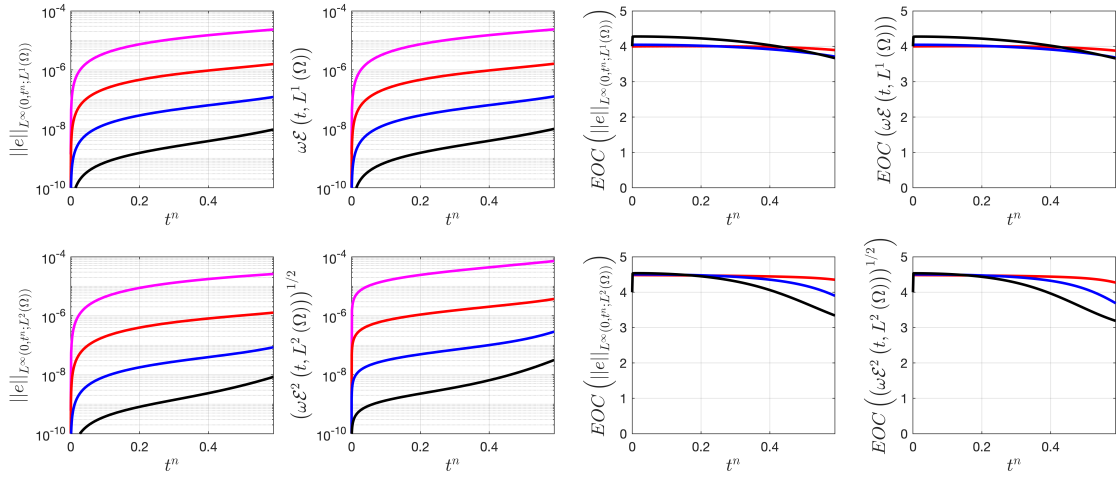


(a) Pre-shock: smooth initial condition, (7.58).

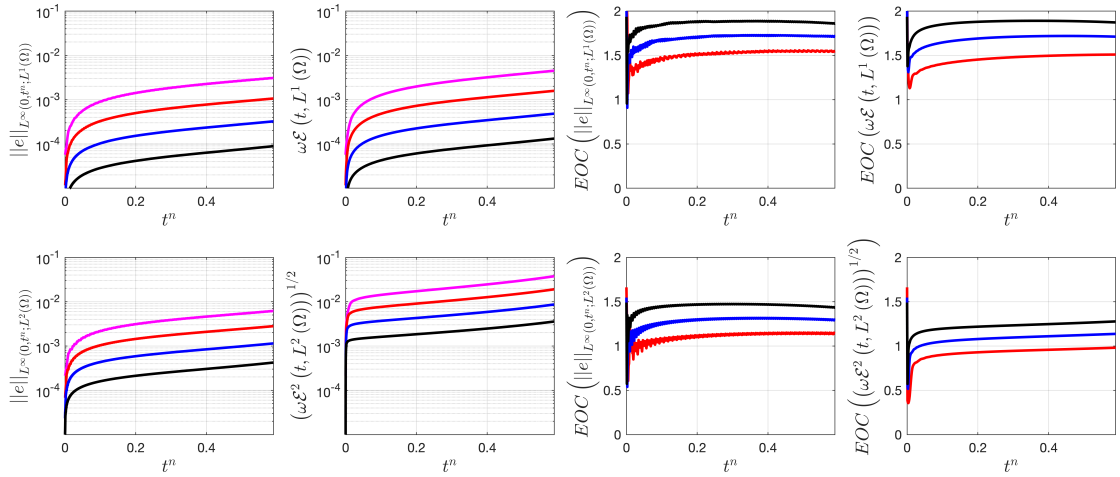


(b) Pre-shock: hat initial condition, (7.59) .

Fig. 7.9. Errors and asymptotic convergence rates for an estimate constructed using a bilinear interpolant for the Lax-Friedrichs scheme, (5.7), for Burgers' equation. Both estimates are optimal for the smooth and continuous initial conditions, with the $\mathrm{L}^2(\Omega)$ estimate converging more slowly.

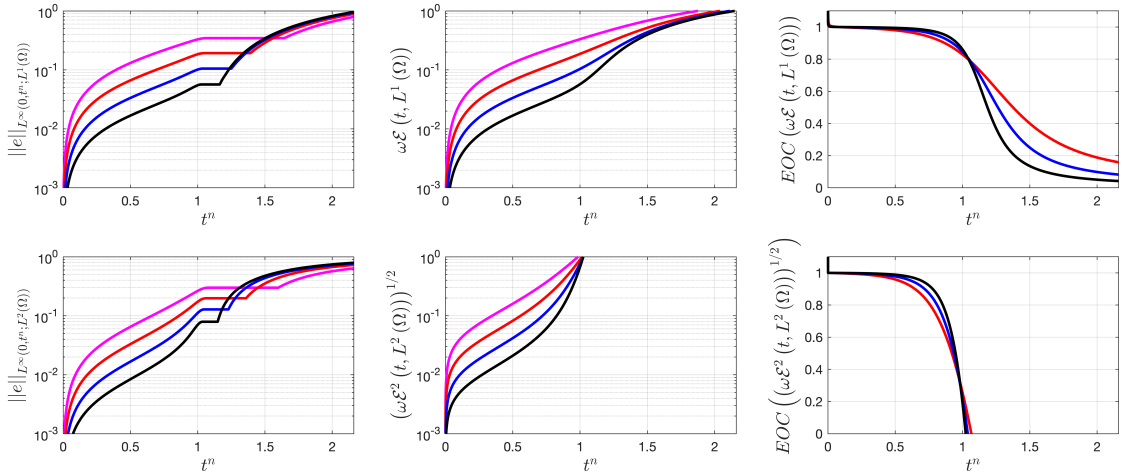


(a) Pre-shock: smooth initial condition, (7.58).

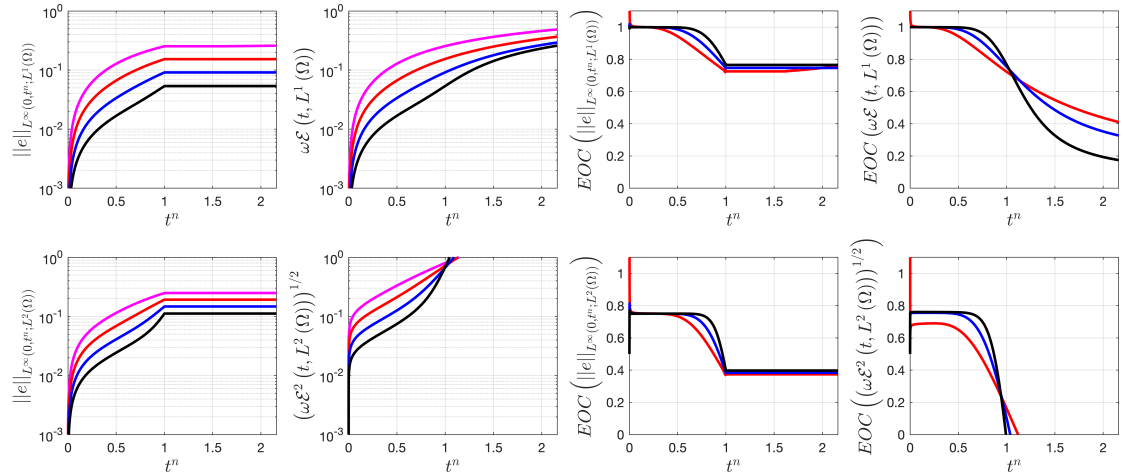


(b) Pre-shock: hat initial condition, (7.59) .

Fig. 7.10. Errors and asymptotic convergence rates for an estimate constructed using a Hermite-WENO reconstruction for the SSP3-WENO3 scheme, (Tbl. 5.16 and §5.3.13), for Burgers' equation. Both estimates are optimal for the smooth initial condition, as shown in Figs. 7.10a. The estimate (7.33) loses optimality for the hat initial condition (see 7.10b).

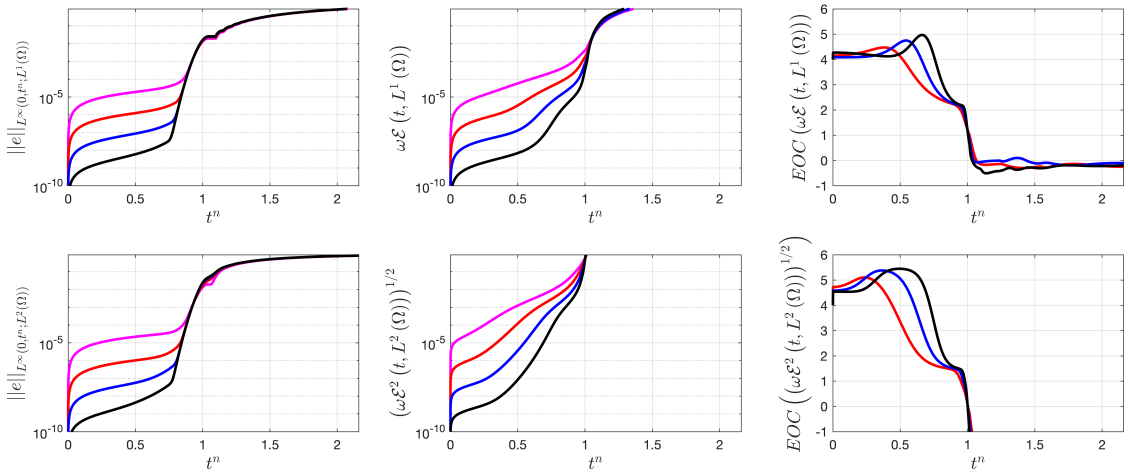


(a) Post-shock: smooth initial condition, (7.58).

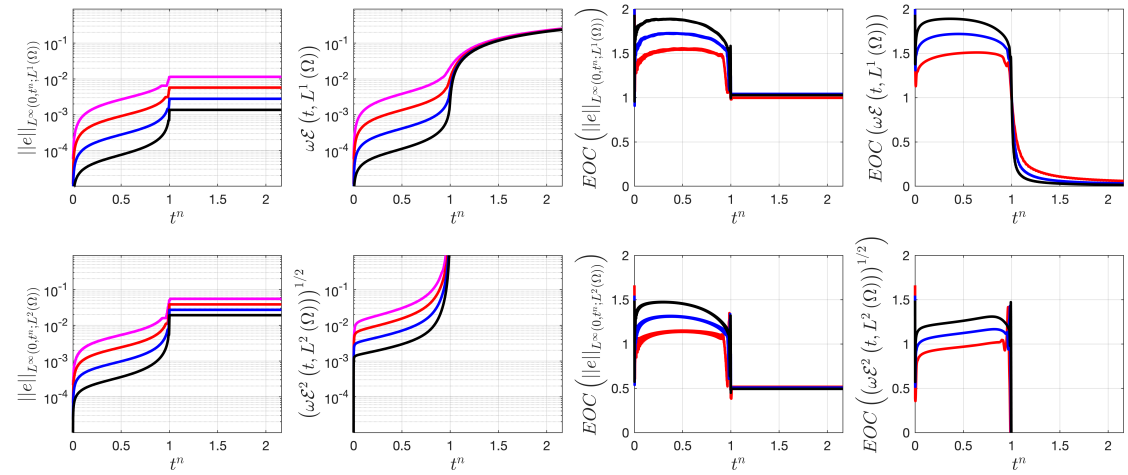


(b) Post-shock: hat initial condition, (7.59) .

Fig. 7.11. Errors and asymptotic convergence rates in the post-shock regime, $t \geq 1$ for an estimate constructed using a bilinear interpolant for the Lax-Friedrichs scheme, (5.7), for Burgers equation. Both estimates lose optimality in the post-shock regime. In particular, the $\mathrm{L}^2(\Omega)$ estimate diverges for both initial conditions in the post-shock regime.



(a) Post-shock: smooth initial condition, (7.58).



(b) Post-shock: hat initial condition, (7.59) .

Fig. 7.12. Errors and asymptotic convergence rates in the post-shock regime, $t \geq 1$ for an estimate constructed using a Hermite-WENO3 reconstruction for the SSP3-WENO3 scheme, (Tbl. 5.16 and §5.3.13), for Burgers' equation.

7.6.6 Test 3: Shallow Water equations

We conclude the numerical testing of our framework with with a systems example. We use the shallow water equations as a model problem. There are two tests in this section, a benchmarking test where we assess the behaviour of the bound using a sinusoidal initial condition, and an adaptive experiment where use the estimate to drive adaptivity. The model problem we test is given by

$$\begin{aligned} \eta_t &+ (\eta v)_x &&= 0 \\ (\eta v)_t &+ \left(\eta v^2 + \tfrac{1}{2} g \eta^2\right)_x &&= 0, \end{aligned} \tag{7.63}$$

equipped with the initial conditions

$$h(x,0) = \begin{cases} h_0 & \text{for} \quad x \leq x_0 \\ h_1 & \text{for} \quad x > x_0, \end{cases} \tag{7.64}$$
$$v(x,0) = v_0(x)\,.$$

over a domain $\Omega = [0, 32\pi]$ and $T \approx 10$. We will use free outflow boundary conditions.

7.6.7 Remark. In order to avoid confusion we note that the set of initial conditions (7.64) with the free outflow boundary conditions will be used in the adaptive test later in the chapter. An additional set of sinusoidal initial conditions and a periodic boundary condition will be used in the benchmarking of the residual, as we clarify below.

Benchmarking of the estimate

In the first instance we simply want to validate the numerical behaviour of the scheme and of the residual component, $\boldsymbol{R}$, (see (7.31)) of the estimate, (7.32) for a smooth initial condition. Specifically, we want to confirm that, in the case a high order scheme is used with smooth initial conditions, the reconstruction possesses the requisite approximability to produce a high-order estimate. Hence, in this experiment our focus is exclusively the EOC of the global residual not the error of the scheme. In this context we define the following constants: $d = 2.0$, $\lambda = 32\pi$, $k = 2\pi/\lambda$, $\omega = \sqrt{gk\tanh(kd)}$ and $a = 1/10$. The initial condition we will use to benchmark the behaviour of the residual is given by

$$\begin{aligned} h(x,0) &= d + a\sin(kx)\,, \\ v(x,0) &= a\omega\frac{\cosh(kh)}{\sinh(kd)}\sin(-kx)\,. \end{aligned} \tag{7.65}$$

We will use periodic boundary conditions for this experiment. The results are shown in Fig. 7.13. Notice that the reconstruction possesses the high order approximability which is required in order to construct an optimal a posteriori estimate for a high order scheme (provided the solution possesses sufficient regularity).

7.6.8 Remark. The reader will notice that the set of initial and boundary conditions in (7.65) is different compared to what we prescribe in (7.64). The reason

is that, in this instance, we are interested in confirming that the reconstruction possesses sufficient approximability to construct an optimal estimate (which necessitates a high order residual) for a high order scheme. We do this by using a smooth (albeit unknown) solution and examining the behaviour of the residual under these conditions. Our desired result is to confirm high order convergence of the residual when a high order scheme is used.

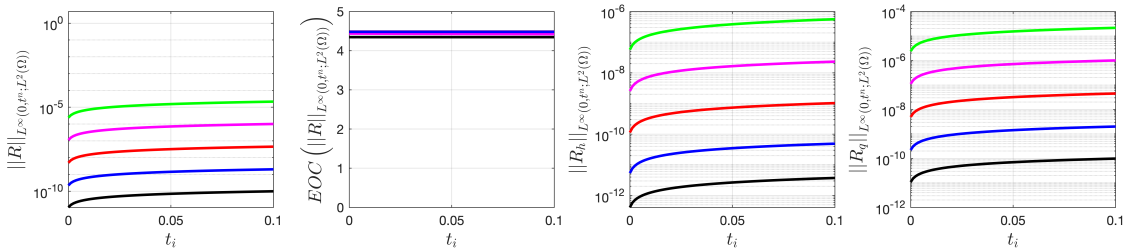


Fig. 7.13. Bench marking for the shallow-water equations with a sinusoidal initial condition, (7.65) and periodic boundary conditions. We use an SSP3-WENO3 scheme with $h = 2^{-m}, m = 9, \dots 12$, with a timestep $\tau = \frac{h}{10}$. The bound is constructed using a Hermite-WENO3 interpolant. Observe that the residual maintains a high order of convergence throughout the simulation.

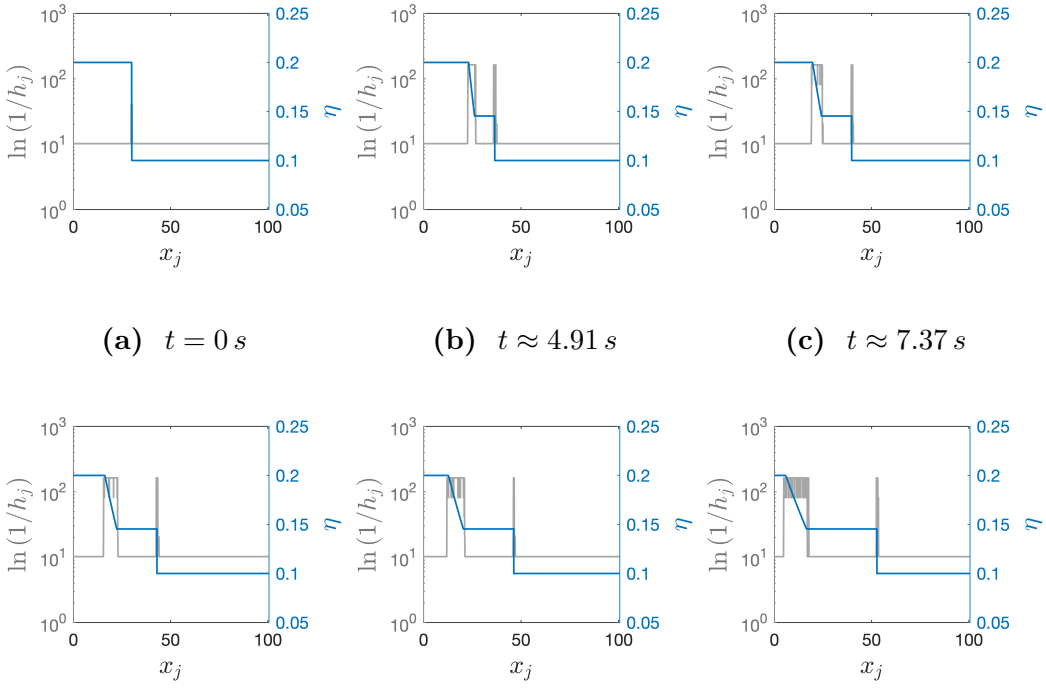
Adaptive experiment

In this case we conduct a numerical experiment to test the residual term $\boldsymbol{R}$ in (7.32) as a local refinement criterion for the shallow water equations using a dam-break initial condition and free outflow conditions. We use $v_0(x, 0) = 0$, $x_0 = 30$, $h_0 = 0.2$ and $h_1 = 0.1$ for the initial condition. This problem has an exact solution and this can be found in [DLK+13, §4.1.1]. We discretise the problem using an SSP3-WENO3 spatio-temporal discretisation and we use a Hermite-WENO3 spatio-temporal reconstruction to obtain the residual, $\boldsymbol{R}$ (see (7.52)).

We start the simulation with the coarsest possible mesh, which is given by $h = 32\pi \times 2^{-10}$ and we use a constant time-step $\tau = \frac{32\pi \times 2^{-12}}{10}$ throughout the simulation. The equivalent uniform mesh is found to have 2183 degrees of freedom, i.e. $0 = x_0 < \cdots < x_{2182} = 32\pi$. The adaptive simulation is shown in Fig. 7.14.

Notice that the estimate is able to accurately track regions of refinement interest. Also, we note that the adaptive mesh consistently performs better than the

equivalent uniform mesh throughout the major part of the simulation, as can be seen in Fig. 7.15.



(a) $t = 0\,s$ **(b)** $t \approx 4.91\,s$ **(c)** $t \approx 7.37\,s$

(d) $t \approx 9.82\,s$ **(e)** $t \approx 12.27\,s$ **(f)** $t \approx 17.18\,s$

Fig. 7.14. Evolution of the surface elevation (blue line) for the shallow water dam break problem, using and SSP3-WENO3 spatio temporal discretization. The logarithm of the reciprocal of the local grid-spacing is overlaid with a grey line. We use a Hermite-WENO3 spatio-temporal reconstruction to construct the residual, $\boldsymbol{R}$ (see (7.52)). Notice that the residual reliably detects regions where refinement is required - such as in the vicinity of the shock and the rarefaction) and where coarsening is appropriate.

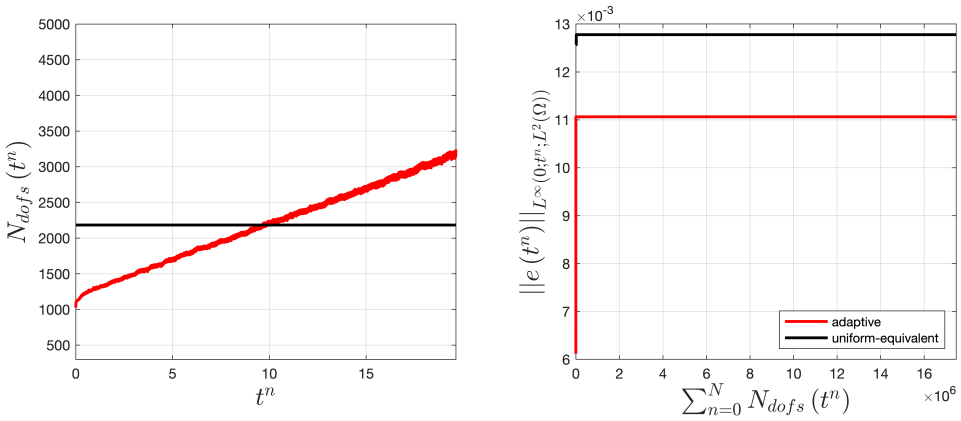


Fig. 7.15. A comparison of the performance of the adaptive grid for the shallow water dam break problem with an equivalent uniform grid (see Defn. 4.6.3), which in this case has 2183 dofs. The adaptive grid consistently maintains a lower level of error than the equivalent uniform grid throughout the major part of the simulation.

7.7 Conclusions

The main contribution from this chapter is the presentation and numerical testing of a framework for constructing reliable, optimal a posteriori error estimates for classes of Finite Difference schemes in the context of scalar and systems of non-linear hyperbolic conservation law in one spatial dimension. The framework is generally applicable: it does not depend on the specific choice of the underlying FD scheme.

The methodology incorporates both the numerical solution and information from the FD scheme itself, thereby facilitating the construction of high order a posteriori estimates using reconstruction techniques. This is a desirable property as it enables the user to construct optimal estimates for high order FD schemes. In addition, in the scalar case we examine the possibility of combining existing a posteriori estimates that are individually optimal in the pre-shock and post-shock regime with the intention of combining them into a single bound that is optimal in both regimes. These consist of a $\mathrm{L}^1(\Omega)$ based on a Kruzkov framework from [CG95] and a relative entropy based $\mathrm{L}^2(\Omega)$ bound from [GMP15].

We demonstrate that the obtained estimates possess desirable characteristics in the pre-shock regime using a range of numerical tests with both linear and non-linear, scalar and vectorial examples. In these tests, both estimates constructed

using the presented framework show optimal convergence characteristics while the solution is smooth, but lose robustness in the post-shock regime. In the case of the relative entropy bound we attribute the loss of robustness to the presence of the spatial derivative of the reconstruction in an exponential time accumulation factor in the estimate: this term blows up in the presence of shocks. However, in the case of the $\mathrm{L}^1(\Omega)$ estimate, the question of the loss of robustness in the post-shock regime for the non-linear problem remains open. Hence, the question of the combination of the bounds into a single bound that is robust throughout both regimes is also open.

Lastly, we show that residuals constructed using the methodology we propose can be reliably used as local mesh refinement criteria. We demonstrate this capability by using the residual to accurately track figures of refinement interest in the dam break scenario for the shallow water equations in one spatial dimension.

Chapter 8

Conclusion

In this thesis we examined the topic of a posteriori error detection for approximations of PDEs using FD methods. After motivating the topic of a posteriori error estimation with an ODE model problem discretised by a linear multistep method, we shifted our attention exclusively on FD discretisations of PDEs, with a focus on hyperbolic conservation laws for the last four chapters. Our main contribution is the development and extensive testing of a framework based on reconstructions for establishing error control over the problems by using their stability framework.

We extensively tested this framework numerically with a range of PDE problems, using several frequently-used numerical schemes to discretise them. We showed that this framework can be used to construct reliable, robust a posteriori error bounds. We showed that under appropriate conditions the estimates behave optimally, that is they converge at the same rate as the numerical scheme. We demonstrate how to incorporate these conditions in the construction of the estimate and we also identify scenarios in which the estimates lose optimality, by using examples from non-linear hyperbolic conservation laws after shocks develop.

8.1 Part 1

The first part of the thesis consists of Chapters 2, 3 and 4.

In 2, inspired by the work in [GLMV16], we proposed a framework to construct an a posteriori error bound for the same linear multistep method as the one examined in that paper and then we tested this framework on a second order ODE

problem. Specifically, in [GLMV16], the time discretisation was a Leapfrog scheme on a staggered grid. We performed numerical tests which showed that the a posteriori error estimate computed using the proposed framework behaved optimally, that is it converged at the same rate as the error for underlying discretisation.

Subsequently, in Chapter 3 we performed an a posteriori error analysis of a model elliptic problem. We used a reconstruction-based framework to obtain an a posteriori error bound of the FD solution. We compared the performance of a bound obtained using this framework with a classical bound for FE elements, for an FE method which is nodally equivalent to our FD discretisation. We showed that the bound obtained based on the reconstruction of the FD solution compares favourably with the classical FE bound.

In Chapter 4, the last chapter of the first part of the thesis, we use a reconstruction based framework to facilitate a posteriori error control for a class of FD schemes for the linear advection equation in one spatial dimension. The behaviour of the a posteriori bounds resulting from the proposed reconstruction operators was validated in a number of numerical tests, in which the estimates were shown to behave optimally. In addition, the estimates were shown to reliably have the capability of tracking features of interest, in this case parasitic, highly oscillatory waves (see [Vic81b]), which arise when numerical solutions encounter mesh non-uniformities.

Lastly, the performance of the estimates as a criterion for adaptivity was evaluated and found to be favourable compared to a mesh with the same number of cumulative degrees of freedom. This chapter sets the tone for the second part of the thesis, which focuses on hyperbolic conservation laws.

8.1.1 Optimal a posteriori error estimation for fourth order discretisations

A direction for further research is the generalization of the proposed framework to higher orders. As an example, [Yos90] presents a fourth order version of the Leapfrog. In this case, the WENO approach in §2.3.13 for constructing the a posteriori error estimate can be used to obtain an a posteriori error estimate that would be optimal for the higher order discretisation.

8.2 Part 2

The second part of the thesis consists of Chapters 5, 6 and 7. In this part, we shift our focus to hyperbolic conservation laws. In this context, in part two, we present and numerically test a reconstruction-based framework for constructing a posteriori error estimates that is intended for general hyperbolic problems in one spatial dimension.

In Chapter 5 we present the class of temporal and spatial discretisations we use for the hyperbolic problems in the rest of the thesis (see [GST01] and [Shu98] respectively). We also present the WENO interpolation process, [JSB+19], which we use in the framework we present in Chapter 7 for obtaining reconstructions of general conservation laws in one spatial dimension. We test the WENO interpolation procedure in functions of varying regularity to show its suitability as an interpolant on account of both its high order accuracy for smooth solutions and non-oscillatory behaviour in the presence of discontinuities.

Chapter 6 is a extension of Chapters 2 and 4 in both the model problem it treats, i.e. the wave equation in system form as well as in the numerical scheme it uses to discretise it (in the temporal variable), the Leapfrog scheme,[GLMV16]. We conduct a numerical experiment with a smooth initial condition where we confirm that the reconstruction-based framework we present can be used to obtain optimal a posteriori estimates from the FD solution.

Finally, in Chapter 7 we focus on non-linear hyperbolic conservation laws. In this chapter we use the relative entropy framework to show a posteriori error bounds for general systems (see [GMP15]) and a Kruzkov framework for upper bounds for scalar problem, see [CG95]. We then present a framework for obtaining reconstructions for non-linear problems, both scalar and systems, see §7.4.2. These reconstructions are used to compute the aforementioned a posteriori estimates.

In both scalar and systems problems, the resulting estimates are validated using classes of well-used FD schemes, see §7.4. In the scalar case we use Burgers equation as a model problem. We demonstrate using numerical experiments that the framework can be used to obtain optimal a posteriori estimates in the pre-shock regime. In the post-shock regime, we show that the framework can be used to construct a robust estimate, (see [CCL95, Thm. 2.1]) using the Kruzkov framework.

In the systems case, we use the shallow water equations as a model problem. In this example, we use the a posteriori error estimate as a driver for mesh adaptivity, demonstrating the capacity of the estimator to identify and track features of interest which require higher resolution.

8.2.1 Incorporation of limiters

In the framework that we have presented in this work, we have not incorporated flux limiters, [Swe84, BB73, BBH75] and this is a possible avenue for further research.

8.2.2 Optimal a posteriori estimate in the pre/post shock regime for scalar problem

We have examined individually the behaviour of two a posteriori error estimates, a $\mathrm{L}^1(\Omega)$ estimate that is based on a Kruzkov framework from [CG95] and a $\mathrm{L}^2(\Omega)$ estimate based on a relative entropy framework from [GMP15] that is optimal in the pre-shock regime. Our intention is to combine these two estimates into a single estimate that is optimal in both the pre- and post-shock regime. At the moment the $\mathrm{L}^1(\Omega)$ estimate loses robustness in the post-shock regime. An avenue of further research is to locate the source of this behaviour and then either use the bound of [CG95], if appropriate, or produce an alternative solution that can function robustly in the post-shock regime (see e.g. [Ohl09]).

8.2.3 Mesh adaptivity

In the adaptive examples that we have presented in Chapters 4 and 7, for $h-$ *adaptivity* was chosen for refining and coarsening. There are additional choices we could implement, such as r, that is relocating nodes in the mesh, (see [HR10]) and hr adaptivity, which is a combination of both methods (see [PPG$^+$05]). There is utility in examining these alternatives, and $r-$adaptivity in particular. Doing so would facilitate a more direct comparison between the performance of the adaptive method and the uniform mesh case as the number of dofs would be the same at every time step. Briefly, such a simulation would involve using the a posteriori estimate to relocate nodes.

8.2.4 Model Adaptivity

Dynamic model adaptation, involves adaptively choosing from a hierarchy of models the most appropriate one to solve in different parts of the domain. In a part of the domain where the physics are particularly complex, a more physically descriptive PDE model can be chosen. In contrast, in parts of the domain where the complex model can be well approximated by a simpler one, the simpler one can be solved, resulting in computational savings.

The choice of model can be done a priori with knowledge of the physics in specific parts of the domain. Alternatively, it can be done in real time, which can be achieved through an a posteriori estimate. The latter is the route that would be of interest to us for further research. Model adaptivity can be coupled with mesh adaptivity, see [GP17] (see also [BE03, BE04] for theory on multi-modelling and its coupling to mesh adaptivity; see [CBvB05, AGQ06] for applications in advection-diffusion equations and flow related problems).

8.2.5 Neural networks and deep learning

In Chapter 4 we presented the issue of parasite formation in the context of non-stationary solutions encountering mesh non-uniformities. A solution to such problems in practice is to add viscosity locally. Mechanisms for adding viscosity vary. They include heuristics, such as gradient indicators, rigorous a posteriori error estimators (e.g. in the style of [GP17]) or machine learning based approaches, [DHR20], for example as a means of determining the optimal amount of viscosity that should be added. A direction of further research could be to compare approaches based on neural networks with a dynamic model adaptive approach based on a posteriori error estimation.

Appendix A

Useful results

A.0.1 Lemma (Cauchy inequality with ϵ)**.** *Let $a, b \in \mathbb{R}$ and $\epsilon > 0$. Then*

$$ab \leq \epsilon a^2 + \frac{1}{4\epsilon} b^2. \tag{A.1}$$

Proof. See [Eva98, Appendix B.2]. □

A.0.2 Theorem. *(Lax equivalence theorem (see [LeV92, §10.5])) For a consistent difference approximation to a well-posed linear evolution problem, stability is necessary and sufficient for convergence.*

A.0.3 Definition. (Standard mollifier from [Eva10]) We define the standard mollfier, $\eta \in \mathrm{C}^\infty(\mathbb{R})$, as follows

$$\eta(x) := \begin{cases} C \exp\left(\frac{1}{|x|^2+1}\right) & \text{for } |x| < 1 \text{ and} \\ 0 & \text{for } |x| \geq 1, \end{cases} \tag{A.2}$$

where the constant $C > 0$ is chosen such that $\int_{\mathbb{R}} \eta \mathrm{d}x = 1$. We use η to obtain functions η_ϵ for $\epsilon > 0$, which we define as

$$\eta_\epsilon(x) := \frac{1}{\epsilon} \eta\left(\frac{x}{\epsilon}\right). \tag{A.3}$$

The functions η_ϵ are in $\mathrm{C}^\infty(\mathbb{R})$ and satisfy

$$\int_{\mathbb{R}} \eta_\epsilon \mathrm{d}x = 1. \tag{A.4}$$

The support of the η_ϵ is a subset of the open ball of radius ϵ in $\mathbb{R}$, centered at $x = 0$.

A.0.4 Definition. (Mollification of a function) Let $f : U \to \mathbb{R}$ be a locally integrable function. Then, we define its mollification, f^ϵ as follows:

$$f^\epsilon := \eta_\epsilon * f \quad \text{in } U_\epsilon, \tag{A.5}$$

where

$$f^\epsilon(x) = \int_u \eta_\epsilon(x-y)\, f(y)\, \mathrm{d}y = \int_{B(0,\epsilon)} \eta_\epsilon(y)\, f(x-y)\, \mathrm{d}y \tag{A.6}$$

A.0.5 Proposition. *(Derivative of a matrix-vector product) Let $A \in \mathbb{R}^{m\times n}$ denote a constant matrix and let $\boldsymbol{x}$ denote an $n \times 1$ vector and $\boldsymbol{y} = A\boldsymbol{x}$ denote an $m \times 1$ vector. Then,*

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = A \tag{A.7}$$

Proof. Let a_{ij}, $1 \le i \le m$, $1 \le j \le n$ denote the entry of matrix A in row i and column j. Then,

$$y_i = \sum_{k=1}^{n} a_{ik} x_k. \tag{A.8}$$

Hence,

$$\frac{\partial y_i}{\partial x_j} = a_{ij}, \tag{A.9}$$

which is precisely the matrix A. □

A.0.6 Proposition. *(Derivative of a matrix-vector product with respect to a scalar) Let x denote a scalar independent variable. Let $A \in \mathrm{C}^1(\mathbb{R}; \mathbb{R}^{m\times m})$ and $\boldsymbol{x} \in \mathrm{C}^1(\mathbb{R}; \mathbb{R}^m)$. We define*

$$\boldsymbol{y} := A\boldsymbol{z}. \tag{A.10}$$

Then,

$$\frac{\partial \boldsymbol{y}}{\partial x} = \frac{\partial A}{\partial x}\boldsymbol{z} + A\frac{\partial \boldsymbol{z}}{\partial x}. \tag{A.11}$$

A.0.7 Proposition. *Consider the m-component vectors (i.e. $m\times 1$ column vectors) $\boldsymbol{y}$, $\boldsymbol{x}$ and let both of them be functions of an underlying vector, n-component vector $\boldsymbol{z}$. Let the scalar $\alpha = \alpha(\boldsymbol{z})$ be defined as the product*

$$\alpha := \boldsymbol{y}^T\boldsymbol{x}. \tag{A.12}$$

Then,

$$\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}^T\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} + \boldsymbol{y}^T\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}. \tag{A.13}$$

Proof. Let x_i, y_i, $1 \leq i \leq m$, denote the ith entry of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. Also, let z_k, $1 \leq k \leq n$ denote the kth entry of vector $\boldsymbol{z}$. We have

$$\alpha = \sum_{i=1}^{m} x_i y_i. \tag{A.14}$$

Using this notation,

$$\frac{\partial \alpha}{\partial z_k} = \sum_{j=1}^{n} \left(x_j \frac{\partial y_j}{\partial z_k} + y_j \frac{\partial x_j}{\partial z_k} \right), \quad k = 1, \ldots, n. \tag{A.15}$$

Hence,

$$\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}^T \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} + \boldsymbol{y}^T \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}} \tag{A.16}$$

□

We will use the result in Proposition A.0.7 to obtain the derivative of the product $\boldsymbol{e}^T A \boldsymbol{e}$ in the case when $\boldsymbol{e}$ and A depend on x.

A.0.8 Proposition. *We denote the independent variable by x. We consider a matrix $A \in \mathrm{C}^1(\mathbb{R}; \mathbb{R}^{m \times m})$ and a column vector $\boldsymbol{e} \in \mathrm{C}^1(\mathbb{R}; \mathbb{R}^m)$. Define the product*

$$\alpha(x) := \boldsymbol{e}(x)^T A \boldsymbol{e}(x)\,. \tag{A.17}$$

Then

$$\frac{\partial \alpha}{\partial x} = \boldsymbol{e}^T \left(A^T \frac{\partial \boldsymbol{e}}{\partial x} + A \frac{\partial \boldsymbol{e}}{\partial x} + \frac{\partial A}{\partial x} \boldsymbol{e} \right) \tag{A.18}$$

Proof. Let $\boldsymbol{v} := A\boldsymbol{e}$. Then, we can write (A.17) as

$$\alpha = \boldsymbol{e}^T \boldsymbol{v}. \tag{A.19}$$

We can use Proposition A.0.7 obtain

$$\frac{\partial \alpha}{\partial x} = \boldsymbol{v}^T \frac{\partial \boldsymbol{e}}{\partial x} + \boldsymbol{e}^T \frac{\partial \boldsymbol{v}}{\partial x}. \tag{A.20}$$

Substituting back $\boldsymbol{v} = A\boldsymbol{e}$ gives us

$$\begin{aligned} \frac{\partial \alpha}{\partial x} &= (A\boldsymbol{e})^T \frac{\partial \boldsymbol{e}}{\partial x} + \boldsymbol{e}^T \frac{\partial (A\boldsymbol{e})}{\partial x} \\ &= \boldsymbol{e}^T A^T \frac{\partial \boldsymbol{e}}{\partial x} + \boldsymbol{e}^T \left(\frac{\partial A}{\partial x} \boldsymbol{e} + A \frac{\partial \boldsymbol{e}}{\partial x} \right) \end{aligned} \tag{A.21}$$

which simplifies to

$$\frac{\partial \alpha}{\partial x} = \boldsymbol{e}^T \left(A^T \frac{\partial \boldsymbol{e}}{\partial x} + A \frac{\partial \boldsymbol{e}}{\partial x} + \frac{\partial A}{\partial x} \boldsymbol{e} \right) \tag{A.22}$$

□

Bibliography

[AABM00] B Achchab, A Agouzal, J Baranger, and JF Maitre. A posteriori error estimates in finite element methods for general friedrichs' systems. *Computer methods in applied mechanics and engineering*, 184(1):39–47, 2000.

[ABF88] David C Arney, Rupak Biswas, and Joseph E Flaherty. A posteriori error estimation of adaptive finite difference schemes for hyperbolic systems. Technical report, US Army Armanent Research Development and Engineering Center, Watervliet NY, 1988.

[AdVLV13] Paola F Antonietti, Lourenco Beirao da Veiga, Carlo Lovadina, and Marco Verani. Hierarchical a posteriori error estimators for the mimetic discretization of elliptic problems. *SIAM Journal on Numerical Analysis*, 51(1):654–675, 2013.

[AGQ06] Valery Agoshkov, Paola Gervasio, and Alfio Quarteroni. Optimal control in heterogeneous domain decomposition methods for advection-diffusion equations. *Mediterranean Journal of Mathematics*, 3(2):147–176, 2006.

[AM04] Uri M Ascher and Robert I McLachlan. Multisymplectic box schemes and the korteweg–de vries equation. *Applied Numerical Mathematics*, 48(3-4):255–269, 2004.

[AMN06] Georgios Akrivis, Charalambos Makridakis, and Ricardo Nochetto. A posteriori error estimates for the crank–nicolson method for parabolic equations. *Mathematics of computation*, 75(254):511–531, 2006.

[AMN09] Georgios Akrivis, Charalambos Makridakis, and Ricardo H Nochetto. Optimal order a posteriori error estimates for a class of runge–kutta and galerkin methods. *Numerische Mathematik*, 114(1):133–160, 2009.

[AO11] Mark Ainsworth and J Tinsley Oden. *A posteriori error estimation in finite element analysis*, volume 37. John Wiley & Sons, 2011.

[Arg54] John H Argyris. Energy theorems and structural analysis: A generalized discourse with applications on energy principles of structural analysis including the effects of temperature and non-linear stress-strain relations. *Aircraft Engineering and Aerospace Technology*, 1954.

[Arn74] Ludwig Arnold. Stochastic differential equations. *New York*, 1974.

[Ban96] Randolph E Bank. Hierarchical bases and the finite element method. *Acta numerica*, 5:1–43, 1996.

[BB73] Jay P Boris and David L Book. Flux-corrected transport. i. shasta, a fluid transport algorithm that works. *Journal of computational physics*, 11(1):38–69, 1973.

[BBH75] David L Book, Jay P Boris, and K Hain. Flux-corrected transport ii: Generalizations of the method. *Journal of Computational Physics*, 18(3):248–283, 1975.

[BCL13] Roland Becker, Daniela Capatina, and Robert Luce. Reconstruction-based a posteriori error estimators for the transport equation. In *Numerical Mathematics and Advanced Applications 2011*, pages 13–21. Springer, 2013.

[BdVBC+13] L Beirão da Veiga, Franco Brezzi, Andrea Cangiani, Gianmarco Manzini, L Donatella Marini, and Alessandro Russo. Basic principles of virtual element methods. *Mathematical Models and Methods in Applied Sciences*, 23(01):199–214, 2013.

[BdVBMR14] L Beirão da Veiga, Franco Brezzi, Luisa Donatella Marini, and Alessandro Russo. The hitchhiker's guide to the virtual element method. *Mathematical models and methods in applied sciences*, 24(08):1541–1573, 2014.

[BdVBMR16] L Beirão da Veiga, Franco Brezzi, Luisa Donatella Marini, and Alessandro Russo. Virtual element method for general second-order elliptic problems on polygonal meshes. *Mathematical Models and Methods in Applied Sciences*, 26(04):729–750, 2016.

[BdVCN+21] L Beirao da Veiga, C Canuto, RH Nochetto, G Vacca, and M Verani. Adaptive vem: Stabilization-free a posteriori error analysis. *arXiv e-prints*, pages arXiv–2111, 2021.

[BdVM08] Lourenço Beirão da Veiga and Gianmarco Manzini. An a posteriori error estimator for the mimetic finite difference approximation of elliptic problems. *International journal for numerical methods in engineering*, 76(11):1696–1723, 2008.

[BE03] Malte Braack and Alexandre Ern. A posteriori control of modeling errors and discretization errors. *Multiscale Modeling & Simulation*, 1(2):221–238, 2003.

[BE04] Malte Braack and Alexandre Ern. Coupling multimodelling with local mesh refinement for the numerical computation of laminar flames. *Combustion Theory and Modelling*, 8(4):771, 2004.

[BFM14] Franco Brezzi, Richard S Falk, and L Donatella Marini. Basic principles of mixed virtual element methods. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(4):1227–1240, 2014.

[BHO18] Timothy Barth, Raphaèle Herbin, and Mario Ohlberger. Finite volume methods: foundation and analysis. *Encyclopedia of Computational Mechanics Second Edition*, pages 1–60, 2018.

[BO84] Marsha J Berger and Joseph Oliger. Adaptive mesh refinement for hyperbolic partial differential equations. *Journal of computational Physics*, 53(3):484–512, 1984.

[BO96] Kim S Bey and J Tinsley Oden. hp-version discontinuous galerkin methods for hyperbolic conservation laws. *Computer Methods in Applied Mechanics and Engineering*, 133(3-4):259–286, 1996.

[BP98] François Bouchut and Benoıt Perthame. Kruzkov's estimates for scalar conservation laws revisited. *Transactions of the American Mathematical Society*, 350(7):2847–2870, 1998.

[Bro07] Courtney Brown. *Differential equations: A modeling approach*. Number 150. Sage, 2007.

[BSB+01] Ivo Babuska, Theofanis Strouboulis, Ivo Babuška, John Robert Whiteman, et al. *The finite element method and its reliability*. Oxford university press, 2001.

[BUGA03] JJ Benito, F Urena, L Gavete, and R Alvarez. An h-adaptive method in the generalized finite differences. *Computer methods in applied mechanics and engineering*, 192(5-6):735–759, 2003.

[BUGA08] Juan José Benito, Francisco Ureña, Luis Gavete, and Beatriz Alonso. A posteriori error estimator and indicator in generalized finite differences. application to improve the approximated solution of elliptic pdes. *International Journal of Computer Mathematics*, 85(3-4):359–370, 2008.

[CBvB05] JM Cnossen, H Bijl, and EH van Brummelen. Model-error estimation for goal-oriented model adaptation in flow-simulations. In *Proceedings of the Finite Volumes for Complex Applications IV (Marrakech, Morocco, July 2005)*, pages 173–183. Wiley-ISTE, 2005.

[CCL94] Bernardo Cockburn, Frédéric Coquel, and Philippe LeFloch. An error estimate for finite volume methods for multidimensional conservation laws. *mathematics of computation*, 63(207):77–103, 1994.

[CCL95] Bernardo Cockburn, Frédéric Coquel, and Philippe G LeFloch. Convergence of the finite volume method for multidimensional conservation laws. *SIAM Journal on Numerical Analysis*, 32(3):687–705, 1995.

[CET14] James B Collins, Don Estep, and Simon Tavener. A posteriori error estimation for the lax–wendroff finite difference scheme. *Journal of Computational and Applied Mathematics*, 263:299–311, 2014.

[CFL67] Richard Courant, Kurt Friedrichs, and Hans Lewy. On the partial difference equations of mathematical physics. *IBM journal of Research and Development*, 11(2):215–234, 1967.

[CFR05] Elisabetta Carlini, Roberto Ferretti, and Giovanni Russo. A weighted essentially nonoscillatory, large time-step scheme for hamilton–jacobi equations. *SIAM Journal on Scientific Computing*, 27(3):1071–1091, 2005.

[CFVN90] Jules G Charney, Ragnar Fjortoft, and John Von Neumann. Numerical integration of the barotropic vorticity equation. In *The Atmosphere - A Challenge*, pages 267–284. Springer, 1990.

[CG95] Bernardo Cockburn and Huiing Gau. A posteriori error estimates for general numerical methods for scalar conservation laws. *Mat. Aplic. Comp*, 14(1):37–47, 1995.

[CG96] Bernardo Cockburn and Pierre-Alain Gremaud. A priori error estimates for numerical methods for scalar conservation laws. part i: The general approach. *Mathematics of computation*, 65(214):533–573, 1996.

[CG97] Bernardo Cockburn and Pierre-Alain Gremaud. A priori error estimates for numerical methods for scalar conservation laws. part ii: Flux-splitting monotone schemes on irregular cartesian grids. *Mathematics of computation*, 66(218):547–572, 1997.

[CG14] Qingshan Chen and Max Gunzburger. Goal-oriented a posteriori error estimation for finite volume methods. *Journal of Computational and Applied Mathematics*, 265:69–82, 2014.

[CGPS17] Andrea Cangiani, Emmanuil H Georgoulis, Tristan Pryer, and Oliver J Sutton. A posteriori error estimates for the virtual element method. *Numerische mathematik*, 137(4):857–893, 2017.

[CGY98] Bernardo Cockburn, Pierre-Alain Gremaud, and Jimmy Xiangrong Yang. A priori error estimates for numerical methods for scalar conservation laws part iii: Multidimensional flux-splitting monotone schemes on non-cartesian grids. *SIAM journal on numerical analysis*, 35(5):1775–1803, 1998.

[CH99] Claire Chainais-Hillairet. Finite volume schemes for a nonlinear hyperbolic equation. convergence towards the entropy solution and error estimate. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(1):129–156, 1999.

[Chr09] Marios Christou. Fully nonlinear computations of waves and wave-structure interaction. 2009.

[Cia02] Philippe G Ciarlet. *The finite element method for elliptic problems*. SIAM, 2002.

[CJST06] Bernardo Cockburn, Claes Johnson, C-W Shu, and Eitan Tadmor. *Advanced numerical approximation of nonlinear hyperbolic equations: lectures given at the 2nd session of the Centro Internazionale Matematico Estivo (CIME) held in Cetraro, Italy, June 23-28, 1997*. Springer, 2006.

[CKS12] Bernardo Cockburn, George E Karniadakis, and Chi-Wang Shu. *Discontinuous Galerkin methods: theory, computation and applications*, volume 11. Springer Science & Business Media, 2012.

[CM08] Andrea Cangiani and Gianmarco Manzini. Flux reconstruction and solution post-processing in mimetic finite difference methods. *Com-*

puter Methods in Applied Mechanics and Engineering, 197(9-12):933–945, 2008.

[CMS17] Andrea Cangiani, Gianmarco Manzini, and Oliver J Sutton. Conforming and nonconforming virtual element methods for elliptic problems. *IMA Journal of Numerical Analysis*, 37(3):1317–1354, 2017.

[CN47] John Crank and Phyllis Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 43, pages 50–67. Cambridge University Press, 1947.

[Coc99] Bernardo Cockburn. A simple introduction to error estimation for nonlinear hyperbolic conservation laws. In *The Graduate Student's Guide to Numerical Analysis' 98*, pages 1–45. Springer, 1999.

[CSKO19] S Chen, Q Sun, I Klioutchnikov, and H Olivier. Numerical study of chemically reacting flow in a shock tube using a high-order point-implicit scheme. *Computers & Fluids*, 184:107–118, 2019.

[Daf78] Constantine M Dafermos. The second law of thermodynamics and stability. Technical report, BROWN UNIV PROVIDENCE RI LEFSCHETZ CENTER FOR DYNAMICAL SYSTEMS, 1978.

[Daf79] Constantine M Dafermos. Stability of motions of thermoelastic fluids. *Journal of Thermal Stresses*, 2(1):127–134, 1979.

[Daf05] Constantine M Dafermos. *Hyperbolic conservation laws in continuum physics*, volume 3. Springer, 2005.

[DG16] Andreas Dedner and Jan Giesselmann. A posteriori analysis of fully discrete method of lines discontinuous galerkin schemes for systems of conservation laws. *SIAM Journal on Numerical Analysis*, 54(6):3523–3549, 2016.

[DGPR19] Andreas Dedner, Jan Giesselmann, Tristan Pryer, and Jennifer K Ryan. Residual estimates for post-processors in elliptic problems. *arXiv preprint arXiv:1906.04658*, 2019.

[DHR20] Niccolo Discacciati, Jan S Hesthaven, and Deep Ray. Controlling oscillations in high-order discontinuous galerkin schemes using artificial viscosity tuned by neural networks. *Journal of Computational Physics*, 409:109304, 2020.

[DiP79] Ronald J DiPerna. Uniqueness of solutions to hyperbolic conservation laws. *Indiana University Mathematics Journal*, 28(1):137–188, 1979.

[DiP83] Ronald J DiPerna. Convergence of the viscosity method for isentropic gas dynamics. *Communications in mathematical physics*, 91(1):1–30, 1983.

[DLK+13] Olivier Delestre, Carine Lucas, Pierre-Antoine Ksinant, Frédéric Darboux, Christian Laguerre, T-N-Tuoi Vo, Francois James, and Stéphane Cordier. Swashes: a compilation of shallow water analytic solutions for hydraulic and environmental studies. *International Journal for Numerical Methods in Fluids*, 72(3):269–300, 2013.

[DLY89] Peter Deuflhard, Peter Leinen, and Harry Yserentant. Concepts of an adaptive hierarchical finite element code. *IMPACT of Computing in Science and Engineering*, 1(1):3–35, 1989.

[DMO07] Andreas Dedner, Charalambos Makridakis, and Mario Ohlberger. Error control for a class of runge–kutta discontinuous galerkin methods for nonlinear conservation laws. *SIAM Journal on Numerical Analysis*, 45(2):514–538, 2007.

[dV08] L Beirao da Veiga. A residual based error estimator for the mimetic finite difference method. *Numerische Mathematik*, 108(3):387–406, 2008.

[dVLM14] Lourenço Beirao da Veiga, Konstantin Lipnikov, and Gianmarco Manzini. *The mimetic finite difference method for elliptic problems*, volume 11. Springer, 2014.

[dVM14] Lourenco Beirão da Veiga and Gianmarco Manzini. A virtual element method with arbitrary regularity. *IMA Journal of Numerical Analysis*, 34(2):759–781, 2014.

[DVM15] L Beirao Da Veiga and G Manzini. Residual a posteriori error estimation for the virtual element method for elliptic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(2):577–599, 2015.

[EEHJ95] Kenneth Eriksson, Don Estep, Peter Hansbo, and Claes Johnson. Introduction to adaptive methods for differential equations. *Acta numerica*, 4:105–158, 1995.

[EG13] Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*, volume 159. Springer Science & Business Media, 2013.

[EG21] Alexandre Ern and Jean-Luc Guermond. *Finite Elements II*. Springer, 2021.

[EJ87] Kenneth Eriksson and Claes Johnson. Error estimates and automatic time step control for nonlinear parabolic problems, i. *SIAM journal on numerical analysis*, 24(1):12–23, 1987.

[EJ91] Kenneth Eriksson and Claes Johnson. Adaptive finite element methods for parabolic problems i: A linear model problem. *SIAM Journal on Numerical Analysis*, 28(1):43–77, 1991.

[EJ95a] Kenneth Eriksson and Claes Johnson. Adaptive finite element methods for parabolic problems iv: Nonlinear problems. *SIAM Journal on Numerical Analysis*, 32(6):1729–1749, 1995.

[EJ95b] Kenneth Eriksson and Claes Johnson. Adaptive finite element methods for parabolic problems v: Long-time integration. *SIAM journal on numerical analysis*, 32(6):1750–1763, 1995.

[EKPQ97] Nicole El Karoui, Shige Peng, and Marie Claire Quenez. Backward stochastic differential equations in finance. *Mathematical finance*, 7(1):1–71, 1997.

[EL93] Donald Estep and Stig Larsson. The discontinuous galerkin method for semilinear parabolic problems. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 27(1):35–54, 1993.

[ES11] LawrenceC. Evans and CharlesK. Smart. Adjoint methods for the infinity laplacian partial differential equation. *Archive for Rational Mechanics and Analysis*, 201(1):87–113, 2011.

[Est95] Donald Estep. A posteriori error bounds and global error control for approximation of ordinary differential equations. *SIAM Journal on Numerical Analysis*, 32(1):1–48, 1995.

[Eva98] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[Eva10] Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010.

[FIDSG19] L Freret, Lucian Ivan, Hans De Sterck, and Clinton PT Groth. High-order finite-volume method with block-based amr for magnetohydrodynamics flows. *Journal of Scientific Computing*, 79(1):176–208, 2019.

[FLQ03] Luca Formaggia, Daniele Lamponi, and Alfio Quarteroni. One-dimensional models for blood flow in arteries. *Journal of engineering mathematics*, 47(3):251–276, 2003.

[FR04] Jason Edward Frank and Sebastian Reich. *On spurious reflections, nonuniform grids and finite difference discretizations of wave equations*. Citeseer, 2004.

[Fri58] Kurt Otto Friedrichs. Symmetric positive linear differential equations. *Communications on Pure and Applied Mathematics*, 11(3):333–418, 1958.

[GHM14] Emmanuil H Georgoulis, Edward Hall, and Charalambos Makridakis. Error control for discontinuous galerkin methods for first order hyperbolic problems. In *Recent developments in discontinuous Galerkin finite element methods for partial differential equations*, pages 195–207. Springer, 2014.

[GLMV16] Emmanuil H Georgoulis, Omar Lakkis, Charalambos G Makridakis, and Juha M Virtanen. A posteriori error estimates for leap-frog and cosine methods for second order evolution problems. *SIAM Journal on Numerical Analysis*, 54(1):120–136, 2016.

[GM00] Laurent Gosse and Charalambos Makridakis. Two a posteriori error estimates for one-dimensional scalar conservation laws. *SIAM Journal on Numerical Analysis*, 38(3):964–988, 2000.

[GMP15] Jan Giesselmann, Charalambos Makridakis, and Tristan Pryer. A posteriori analysis of discontinuous galerkin schemes for systems of hyperbolic conservation laws. *SIAM Journal on Numerical Analysis*, 53(3):1280–1303, 2015.

[God59] Sergei Konstantinovich Godunov. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Matematicheskii Sbornik*, 89(3):271–306, 1959.

[GP17] Jan Giesselmann and Tristan Pryer. A posteriori analysis for dynamic model adaptation in convection-dominated problems. *Mathematical Models and Methods in Applied Sciences*, 27(13):2381–2423, 2017.

[GR91] Edwige Godlewski and Pierre-Arnaud Raviart. *Hyperbolic systems of conservation laws*. Ellipses, 1991.

[GR13] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118. Springer Science & Business Media, 2013.

[GS98] Sigal Gottlieb and Chi-Wang Shu. Total variation diminishing runge-kutta schemes. *Mathematics of computation of the American Mathematical Society*, 67(221):73–85, 1998.

[GST01] Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM review*, 43(1):89–112, 2001.

[GUB+18] Luis Gavete, Francisco Ureña, Juan Jose Benito, Miguel Ureña, and Maria Lucia Gavete. Solving elliptical equations in 3d by means of an adaptive refinement in generalized finite differences. *Mathematical Problems in Engineering*, 2018, 2018.

[HEOC87] Ami Harten, Bjorn Engquist, Stanley Osher, and Sukumar R Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes, iii. In *Upwind and high-resolution schemes*, pages 218–290. Springer, 1987.

[Hes17] Jan S Hesthaven. *Numerical methods for conservation laws: From analysis to algorithms*. SIAM, 2017.

[Het00] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[HH03] Ralf Hartmann and Paul Houston. Adaptive discontinuous galerkin finite element methods for nonlinear hyperbolic conservation laws. *SIAM Journal on Scientific Computing*, 24(3):979–1004, 2003.

[HLW03] Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration illustrated by the störmer–verlet method. *Acta Numerica*, 12:399–450, 2003.

[HMSW99] Paul Houston, John A Mackenzie, Endre Süli, and Gerald Warnecke. A posteriori error analysis for numerical approximations of Friedrichs systems. *Numerische Mathematik*, 82(3):433–470, 1999.

[HNS16] Bamdad Hosseini, Nilima Nigam, and John M Stockie. On regularizations of the dirac delta distribution. *Journal of Computational Physics*, 305:423–447, 2016.

[HPC+00] Giles W Hunt, Mark A Peletier, Alan R Champneys, Patrick D Woods, M Ahmer Wadee, Chris J Budd, and Gabriel James Lord. Cellular buckling in long structures. *Nonlinear Dynamics*, 21(1):3–29, 2000.

[HR10] Weizhang Huang and Robert D Russell. *Adaptive moving mesh methods*, volume 174. Springer Science & Business Media, 2010.

[HRS00] Paul Houston, Rolf Rannacher, and Endre Süli. A posteriori error analysis for stabilised finite element approximations of transport problems. *Computer methods in applied mechanics and engineering*, 190(11-12):1483–1508, 2000.

[HS01] Paul Houston and Endre Süli. hp-adaptive discontinuous galerkin finite element methods for first-order hyperbolic problems. *SIAM Journal on Scientific Computing*, 23(4):1226–1252, 2001.

[HW37] Douglas Rayner Hartree and John R Womersley. A method for the numerical or mechanical solution of certain types of partial differential equations. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 161(906):353–366, 1937.

[HW07] Jan S Hesthaven and Tim Warburton. *Nodal discontinuous Galerkin methods: algorithms, analysis, and applications*. Springer Science & Business Media, 2007.

[Jen72] Paul S Jensen. Finite difference techniques for variable grids. *Computers & Structures*, 2(1-2):17–29, 1972.

[JL04] Johan Jansson and Anders Logg. *Multi-adaptive Galerkin methods for ODEs V: Stiff problems*. Chalmers Finite Element Centre, Chalmers University of Technology, 2004.

[Joh88] Claes Johnson. Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations. *SIAM Journal on Numerical Analysis*, 25(4):908–926, 1988.

[Joh90] Claes Johnson. Adaptive finite element methods for diffusion and convection problems. *Computer Methods in Applied Mechanics and Engineering*, 82(1-3):301–322, 1990.

[Joh93] Claes Johnson. A new paradigm for adaptive finite element methods. In *Proceedings of MAFELAP*, volume 93, 1993.

[Joh12] Claes Johnson. *Numerical solution of partial differential equations by the finite element method*. Courier Corporation, 2012.

[JS95] Claes Johnson and Anders Szepessy. Adaptive finite element methods for conservation laws based on a posteriori error estimates. *Communications on Pure and Applied Mathematics*, 48(3):199–234, 1995.

[JS96] Guang-Shan Jiang and Chi-Wang Shu. Efficient implementation of weighted eno schemes. *Journal of computational physics*, 126(1):202–228, 1996.

[JS13] Boško S Jovanović and Endre Süli. *Analysis of Finite Difference Schemes: For Linear Partial Differential Equations with Generalized Solutions*, volume 46. Springer Science & Business Media, 2013.

[JSB+19] Gioele Janett, Oskar Steiner, Ernest Alsina Ballester, Luca Belluzzi, and Siddhartha Mishra. A novel fourth-order weno interpolation technique-a possible new tool designed for radiative transfer. *Astronomy & Astrophysics*, 624:A104, 2019.

[JT97] B Cockburn C Johnson and C-W Shu E Tadmor. Advanced numerical approximation of nonlinear hyperbolic equations. 1997.

[JT98] Guang-Shan Jiang and Eitan Tadmor. Nonoscillatory central schemes for multidimensional hyperbolic conservation laws. *SIAM Journal on Scientific Computing*, 19(6):1892–1917, 1998.

[KO00] Dietmar Kröner and Mario Ohlberger. A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multi dimensions. *Mathematics of Computation*, 69(229):25–39, 2000.

[KR94] Dietmar Kröner and Mirko Rokyta. Convergence of upwind finite volume schemes for scalar conservation laws in two dimensions. *SIAM journal on numerical analysis*, 31(2):324–343, 1994.

[Kru70] Stanislav N Kružkov. First order quasilinear equations in several independent variables. *Mathematics of the USSR-Sbornik*, 10(2):217, 1970.

[Kuz76] NN Kuznetsov. Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation. *USSR Computational Mathematics and Mathematical Physics*, 16(6):105–119, 1976.

[L+02] Randall J LeVeque et al. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge university press, 2002.

[Laf04] Marc Laforest. A posteriori error estimate for front-tracking: systems of conservation laws. *SIAM journal on mathematical analysis*, 35(5):1347–1370, 2004.

[Lax54] Peter D Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Communications on pure and applied mathematics*, 7(1):159–193, 1954.

[Lax57] Peter D Lax. Hyperbolic systems of conservation laws ii. *Communications on pure and applied mathematics*, 10(4):537–566, 1957.

[Lax73] Peter D Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, volume 11. SIAM, 1973.

[LC19] Cheng Li and Xi Chen. Simulating nonhydrostatic atmospheres on planets (snap): Formulation, validation, and application to the jovian atmosphere. *The Astrophysical Journal Supplement Series*, 240(2):37, 2019.

[LeF02] Philippe G LeFloch. *Hyperbolic Systems of Conservation Laws: The theory of classical and nonclassical shock waves*. Springer Science & Business Media, 2002.

[LeV92] Randall J LeVeque. *Numerical methods for conservation laws*, volume 132. Springer, 1992.

[LeV07] Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

[Lis84] Tadeusz Liszka. An interpolation method for an irregular net of nodes. *International Journal for Numerical Methods in Engineering*, 20(9):1599–1612, 1984.

[LM95] Michael Y Li and James S Muldowney. Global stability for the seir model in epidemiology. *Mathematical biosciences*, 125(2):155–164, 1995.

[LM06] Omar Lakkis and Charalambos Makridakis. Elliptic reconstruction and a posteriori error estimates for fully discrete linear parabolic problems. *Mathematics of computation*, 75(256):1627–1658, 2006.

[LMS14] Konstantin Lipnikov, Gianmarco Manzini, and Mikhail Shashkov. Mimetic finite difference method. *Journal of Computational Physics*, 257:1163–1227, 2014.

[LO80] Tadeusz Liszka and Janusz Orkisz. The finite difference method at arbitrary irregular grids and its application in applied mechanics. *Computers & Structures*, 11(1-2):83–95, 1980.

[LOC94] Xu-Dong Liu, Stanley Osher, and Tony Chan. Weighted essentially non-oscillatory schemes. *Journal of computational physics*, 115(1):200–212, 1994.

[Log03] Anders Logg. Multi-adaptive galerkin methods for odes i. *SIAM Journal on Scientific Computing*, 24(6):1879–1902, 2003.

[Log04a] Anders Logg. Multi-adaptive galerkin methods for odes ii: Implementation and applications. *SIAM Journal on Scientific Computing*, 25(4):1119–1141, 2004.

[Log04b] Anders Logg. Multi-adaptive galerkin methods for odes iii: Existence and stability. 2004.

[LP12] Omar Lakkis and Tristan Pryer. Gradient recovery in adaptive finite-element methods for parabolic problems. *IMA Journal of Numerical Analysis*, 32(1):246–278, 2012.

[LRKK19] AM Lipanov, IG Rusyak, SA Korolev, and SA Karskanov. Numerical solution of the problem of flow past projected bodies for determining their aerodynamic coefficients. *Journal of Engineering Physics and Thermophysics*, 92(2):477–485, 2019.

[LSZ09] Yuan-yuan Liu, Chi-wang Shu, and Meng-ping Zhang. On the positivity of linear weights in weno approximations. *Acta Mathematicae Applicatae Sinica, English Series*, 25(3):503–538, 2009.

[LT11] David Long and John Thuburn. Numerical wave propagation on non-uniform one-dimensional staggered grids. *Journal of Computational Physics*, 230(7):2643–2659, 2011.

[LVW21] Richard Liska, Pavel Váchal, and Burton Wendroff. Lax-wendroff methods on highly non-uniform meshes. dedicated to the memory of blair swartz (1932–2019). *Applied Numerical Mathematics*, 163:167–181, 2021.

[LW04] Robert Luce and Barbara I Wohlmuth. A local a posteriori error estimator based on equilibrated fluxes. *SIAM Journal on Numerical Analysis*, 42(4):1394–1414, 2004.

[Mak07] Charalambos Makridakis. Space and time reconstructions in a posteriori analysis of evolution problems. In *ESAIM: Proceedings*, volume 21, pages 31–44. EDP Sciences, 2007.

[MM05] Keith W Morton and David Francis Mayers. *Numerical solution of partial differential equations: an introduction*. Cambridge university press, 2005.

[MN03] Charalambos Makridakis and Ricardo H Nochetto. Elliptic reconstruction and a posteriori error estimates for parabolic problems. *SIAM journal on numerical analysis*, 41(4):1585–1594, 2003.

[MN06] Charalambos Makridakis and Ricardo H Nochetto. A posteriori error analysis for higher order dissipative methods for evolution problems. *Numerische Mathematik*, 104(4):489–514, 2006.

[NBOT19] Koji Nishiguchi, Rahul Bale, Shigenobu Okazawa, and Makoto Tsubokura. Full eulerian deformable solid-fluid interaction scheme based on building-cube method for large-scale parallel computing. *International Journal for Numerical Methods in Engineering*, 117(2):221–248, 2019.

[NSV00] Ricardo H Nochetto, Giuseppe Savaré, and Claudio Verdi. A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 53(5):525–589, 2000.

[NT90] Haim Nessyahu and Eitan Tadmor. Non-oscillatory central differencing for hyperbolic conservation laws. *Journal of computational physics*, 87(2):408–463, 1990.

[Ohl09] Mario Ohlberger. A review of a posteriori error control and adaptivity for approximations of non-linear conservation laws. *International Journal for Numerical Methods in Fluids*, 59(3):333–354, 2009.

[Ole57] Olga Arsen'evna Oleinik. Discontinuous solutions of non-linear differential equations. *Uspekhi Matematicheskikh Nauk*, 12(3):3–73, 1957.

[Olv00] Peter J Olver. *Applications of Lie groups to differential equations*, volume 107. Springer Science & Business Media, 2000.

[Ork98] J Orkisz. Finite difference method (part iii). *Handbook of Computational Solid Mechanics*, pages 336–432, 1998.

[OV06] Mario Ohlberger and Julien Vovelle. Error estimate for the approximation of nonlinear conservation laws on bounded domains by the finite volume method. *Mathematics of Computation*, 75(253):113–150, 2006.

[OZ19] Junfeng Ou and Zhigang Zhai. Effects of aspect ratio on shock-cylinder interaction. *Acta Mechanica Sinica*, 35(1):61–69, 2019.

[PK75] Nicholas Perrone and Robert Kao. A general finite difference method for arbitrary meshes. *Computers & Structures*, 5(1):45–57, 1975.

[PPG+05] MD Piggott, CC Pain, GJ Gorman, PW Power, and AJH Goddard. h, r, and hr adaptivity with applications in numerical ocean modelling. *Ocean modelling*, 10(1-2):95–113, 2005.

[Ric11] Lewis Fry Richardson. Ix. the approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 210(459-470):307–357, 1911.

[RM90] AM Rogerson and E Meiburg. A numerical study of the convergence properties of eno schemes. *Journal of Scientific Computing*, 5(2):151–167, 1990.

[RM94] Robert D Richtmyer and Keith W Morton. Difference methods for initial-value problems. *Malabar*, 1994.

[SCR16] Matteo Semplice, Armando Coco, and Giovanni Russo. Adaptive mesh refinement for hyperbolic systems based on third-order compact weno reconstruction. *Journal of Scientific Computing*, 66(2):692–724, 2016.

[SH96] Endre Suli and Paul Houston. Finite element methods for hyperbolic problems: a posteriori error analysis and adaptivity. 1996.

[SH03] Endre Süli and Paul Houston. Adaptive finite element approximation of hyperbolic problems. In *Error estimation and adaptive discretization methods in computational fluid dynamics*, pages 269–344. Springer, 2003.

[Shu88] Chi-Wang Shu. Total-variation-diminishing time discretizations. *SIAM Journal on Scientific and Statistical Computing*, 9(6):1073–1084, 1988.

[Shu98] Chi-Wang Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In *Advanced numerical approximation of nonlinear hyperbolic equations*, pages 325–432. Springer, 1998.

[Shu02] Chi-Wang Shu. A survey of strong stability preserving high order time discretizations. *Collected lectures on the preservation of stability under discretization*, 109:51–65, 2002.

[Shu20] Chi-Wang Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes. *Acta Numerica*, 29:701–762, 2020.

[SL18] Matteo Semplice and Raphaël Loubère. Adaptive-mesh-refinement for hyperbolic systems of conservation laws based on a posteriori stabilized high order polynomial reconstructions. *Journal of Computational Physics*, 354:86–110, 2018.

[SM03] Endre Süli and David F Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003.

[Smo12] Joel Smoller. *Shock waves and reaction—diffusion equations*, volume 258. Springer Science & Business Media, 2012.

[SO88] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of computational physics*, 77(2):439–471, 1988.

[SO89] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes, ii. In *Upwind and High-Resolution Schemes*, pages 328–374. Springer, 1989.

[Sod78] Gary A Sod. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of computational physics*, 27(1):1–31, 1978.

[SS05] Alfred Schmidt and Kunibert G Siebert. Design of adaptive finite element software. *Lecture Notes in Computational Science and Engineering. Springer-Verlag, Berlin*, 2005.

[Stö07] Carl Störmer. Sur les trajectoires des corpuscules électrisés dans l'espace. applications à l'aurore boréale et aux perturbations magnétiques. *Radium (Paris)*, 4(1):2–5, 1907.

[Sül99] Endre Süli. A posteriori error analysis and adaptivity for finite element approximations of hyperbolic problems. In *An introduction to recent developments in theory and numerics for conservation laws*, pages 123–194. Springer, 1999.

[SV16] Denis Serre and Alexis F Vasseur. About the relative entropy method for hyperbolic systems of conservation laws. *Contemp. Math. AMS*, 658:237–248, 2016.

[Swe84] Peter K Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM journal on numerical analysis*, 21(5):995–1011, 1984.

[Swe89] Peter K Sweby. "tvd" schemes for inhomogeneous conservation laws. In *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications*, pages 599–607. Springer, 1989.

[SZA+19] Armin Shahmardi, Sagar Zade, Mehdi N Ardekani, Rob J Poole, Fredrik Lundell, Marco E Rosti, and Luca Brandt. Turbulent duct flow with polymers. *Journal of Fluid Mechanics*, 859:1057–1083, 2019.

[Tay17] Brook Taylor. *Methodus incrementorum directa et inversa*. Innys, 1717.

[Tho90] Vidar Thomée. Finite difference methods for linear parabolic equations. *Handbook of numerical analysis*, 1:5–196, 1990.

[Tre82] Lloyd N Trefethen. Group velocity in finite difference schemes. *SIAM review*, 24(2):113–136, 1982.

[UBAG05] F Ureña, JJ Benito, R Alvarez, and L Gavete. Computational error approximation and h-adaptive algorithm for the 3-d generalized finite difference method. *International Journal for Computational Methods in Engineering Science and Mechanics*, 6(1):31–39, 2005.

[UBU+18] Miguel Ureña, Juan José Benito, Francisco Ureña, Ángel García, Luis Gavete, and Luis Benito. Adaptive strategies to improve the application of the generalized finite differences method in 2d and 3d. *Mathematical Methods in the Applied Sciences*, 41(17):7115–7129, 2018.

[VB82] Robert Vichnevetsky and John B Bowles. *Fourier analysis of numerical approximations of hyperbolic equations*. SIAM, 1982.

[Ver67] Loup Verlet. Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical review*, 159(1):98, 1967.

[Ver13] Rüdiger Verfürth. *A posteriori error estimation techniques for finite element methods*. OUP Oxford, 2013.

[Vic80] R Vichnevetsky. Propagation properties of semi-discretizations of hyperbolic equations. *Mathematics and computers in simulation*, 22(2):98–102, 1980.

[Vic81a] Robert Vichnevetsky. Energy and group velocity in semi discretizations of hyperbolic equations. 1981.

[Vic81b] Robert Vichnevetsky. Propagation through numerical mesh refinement for hyperbolic equations. 1981.

[Vic87] Robert Vichnevetsky. Wave propagation and reflection in irregular grids for hyperbolic equations. *Applied Numerical Mathematics*, 3:133–166, 1987.

[Vil94] J-P Vila. Convergence and error estimates in finite volume schemes for general multidimensional scalar conservation laws. i. explicite monotone schemes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 28(3):267–295, 1994.

[VL73] Bram Van Leer. Towards the ultimate conservative difference scheme i. the quest of monotonicity. In *Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics*, pages 163–168. Springer, 1973.

[VL74] Bram Van Leer. Towards the ultimate conservative difference scheme. ii. monotonicity and conservation combined in a second-order scheme. *Journal of computational physics*, 14(4):361–370, 1974.

[VL77a] Bram Van Leer. Towards the ultimate conservative difference scheme iii. upstream-centered finite-difference schemes for ideal compressible flow. *Journal of Computational Physics*, 23(3):263–275, 1977.

[VL77b] Bram Van Leer. Towards the ultimate conservative difference scheme. iv. a new approach to numerical convection. *Journal of computational physics*, 23(3):276–299, 1977.

[VL79] Bram Van Leer. Towards the ultimate conservative difference scheme. v. a second-order sequel to godunov's method. *Journal of computational Physics*, 32(1):101–136, 1979.

[VM07] Henk Kaarle Versteeg and Weeratunge Malalasekera. *An introduction to computational fluid dynamics: the finite volume method*. Pearson education, 2007.

[VNG47] John Von Neumann and Herman H Goldstine. Numerical inverting of matrices of high order. *Bulletin of the American Mathematical Society*, 53(11):1021–1099, 1947.

[vRC15] Maarten van Reeuwijk and John Craske. Energy-consistent entrainment relations for jets and plumes. *Journal of Fluid Mechanics*, 782:333–355, 2015.

[W+95] G Warnecke et al. A second order finite difference error indicator for adaptive transonic flow computations. *Numerische Mathematik*, 70(2):129–161, 1995.

[WSHN13] Cheng Wang, Chi-Wang Shu, Wenhu Han, and Jianguo Ning. High resolution weno simulation of 3d detonation waves. *Combustion and Flame*, 160(2):447–462, 2013.

[WTDS75] MJ Wyatt, TAYLOR, G DAVIES, and C SNELL. A new difference based finite element method. *Proceedings of the Institution of Civil Engineers*, 59(3):395–409, 1975.

[Yos90] Haruo Yoshida. Construction of higher order symplectic integrators. *Physics letters A*, 150(5-7):262–268, 1990.

[Zad66a] PE Zadunaisky. A method for the estimation of errors propagated in the numerical solution of a system of ordinary differential equations. In *The Theory of Orbits in the Solar System and in Stellar Systems*, volume 25, page 281, 1966.

[Zad66b] Pedro E Zadunaisky. Motion of halley's comet during the return of 1910. *The Astronomical Journal*, 71:20, 1966.

[Zad70] Pedro E Zadunaisky. On the accuracy in the numerical computation of orbits. In *Periodic Orbits, Stability and Resonances*, pages 216–227. Springer, 1970.

[Zad72] PE Zadunaisky. On the determination of nongravitational forces acting on comets. In *Symposium-International Astronomical Union*, volume 45, pages 144–151. Cambridge University Press, 1972.

[Zad76] Pedro E Zadunaisky. On the estimation of errors propagated in the numerical integration of ordinary differential equations. *Numerische Mathematik*, 27(1):21–39, 1976.

[ZGK83] Olgierd C Zienkiewicz, JP De SR Gago, and Don W Kelly. The hierarchical concept in finite element analysis. *Computers & Structures*, 16(1-4):53–65, 1983.

[ZTZ05] Olgierd Cecil Zienkiewicz, Robert Leroy Taylor, and Jian Z Zhu. *The finite element method: its basis and fundamentals*. Elsevier, 2005.