Algebraically constrained finite element methods for hyperbolic problems with applications in geophysics and gas dynamics

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

Der Fakultät für Mathematik der
Technischen Universität Dortmund
vorgelegt von

Hennes Hajduk

im April 2022

Dissertation

Algebraically constrained finite element methods for hyperbolic problems with applications in geophysics and gas dynamics

Fakultät für Mathematik
Technische Universität Dortmund

Erstgutachter: Prof. Dr. Dmitri Kuzmin

Zweitgutachter: Prof. Dr.-Ing. Gregor Gassner

Tag der mündlichen Prüfung: 04. Juli 2022

And if the dam breaks open many years too soon
And if there is no room upon the hill
And if your head explodes with dark forebodings too
I'll see you on the dark side of the moon

Pink Floyd, Brain Damage

Acknowledgments

First and foremost, I would like to express my sincerest gratitude to Prof. Dr. Dmitri Kuzmin for supervising this thesis and providing me with interesting research topics. His guidance and support throughout the years mean a lot to me, and I have greatly benefited from his expertise. Dmitri gave me much appreciated opportunities to travel for various scientific purposes and also allowed me to participate as co-organizer in his 2020 Moselle valley workshop. Let me also take the time to thank him for proofreading this thesis, as always providing me with invaluable feedback, and importantly, for exhibiting personal qualities as an advisor that leave nothing else to be desired.

I am also very much indebted to Dr. Christoph Lohmann and Assistant Prof. Dr. Andreas Rupp for many fruitful scientific discussions and for proofreading parts of this thesis. Moreover, I would like to thank Prof. Dr. Vadym Aizinger, Dr. Florian Frank, Dr. Manuel Quezada de Luna, and Dr. Balthasar Reuter for contributing to my scientific education and for collaborating with me. Were it not for these friends and colleagues, my work on this thesis might not have come to fruition, so thank you all for your contributions!

During my PhD studies, I had the genuine pleasure to visit and intern at Lawrence Livermore National Laboratory (LLNL). All the people who made this opportunity possible deserve my gratitude, especially Dr. Tzanio Kolev, my supervisor at LLNL. Let me also acknowledge the support I received from members of the MFEM team at LLNL, who, on multiple occasions, answered my questions regarding the code and helped me out in various other ways. I would like to take this opportunity to compliment the MFEM group on developing and maintaining such a well-designed software library and distributing it to the community as an open source toolbox.

Furthermore, I want to thank Prof. Dr. Stefan Turek for making my employment at TU Dortmund possible. By executing his position as dean in the way he does, a great working atmosphere is created, not only in his own group but in the department as a whole. I would also like to extend my gratitude to all members of the Institute of Applied Mathematics (LS III) at TU Dortmund. Special thanks go to my current and former office mates, Joshua Vedral, Jan-Phillip Bäcker, Johanna Grüll, Dr. Nikita Klyushnev and Falko Ruppenthal for fruitful discussions on mathematical topics and other matters.

Prof. Dr. Gregor Gassner is gratefully acknowledged for agreeing to act as reviewer of this thesis. Let me also express my gratitude to him and Prof. Dr. Matthias Röger for their willingness to participate as examiners during my PhD defense.

Besides the help received from people in academia, I equally value the moral support of my friends and family members. I would not have been able to complete this work without the upbringing I enjoyed under my parents, Katharina and Andreas, who contributed to this effort in a great many ways. Thank you so much for that!

Hennes Hajduk

Contents

Abstract

The research conducted in this thesis is focused on property-preserving discretizations of hyperbolic partial differential equations. Computational methods for solving such problems need to be carefully designed to produce physically meaningful numerical solutions. In particular, approximations to some quantities of interest should satisfy local and global discrete maximum principles. Moreover, numerical methods need to obey certain conservation relations, and convergence of approximations to the physically relevant exact solution should be ensured if multiple solutions may exist. Many algorithms based on the aforementioned design principles fall into the category of algebraic flux correction (AFC) schemes. Modern AFC discretizations of nonlinear hyperbolic systems express approximate solutions as convex combinations of intermediate states and constrain these states to be admissible. The main focus of our work is on monolithic convex limiting (MCL) strategies that modify spatial semi-discretizations in this way. Contrary to limiting approaches of predictor-corrector type, their monolithic counterparts are well suited for transient and steady problems alike. Further benefits of the MCL framework presented in this thesis include the possibility of enforcing entropy stability conditions in addition to discrete maximum principles.

Using the AFC methodology, we transform finite element discretizations into property-preserving low order methods and perform flux correction to recover higher orders of accuracy without losing any desirable properties. The presented methods produce physics-compatible approximations, which exhibit excellent shock capturing capabilities.

One novelty of this work is the tailor-made extension of monolithic convex limiting to the shallow water equations with a nonconservative topography term. Our generalized MCL schemes are entropy stable, positivity preserving, and well balanced in the sense that lake at rest equilibria are preserved. Another desirable property of numerical methods for the shallow water equations is the capability to handle wet-dry transitions properly. We present two new approaches to dealing with this issue.

To corroborate our computational results with theoretical investigations, we perform numerical analysis for property-preserving discretizations of the time-dependent linear advection equation. In this context, we prove stability and derive an a priori error estimate in the semi-discrete setting. We also compare the monolithic convex limiting strategy to two representatives of related flux-corrected transport algorithms.

Another highlight of this thesis is the chapter on MCL schemes for arbitrary order discontinuous Galerkin (DG) discretizations. Building on algorithms developed for continuous Lagrange and Bernstein finite elements, we extend our MCL schemes to the high order DG setting. This research effort involves the design of new AFC tools for numerical fluxes that appear in the DG weak formulation. Our limiting strategy for DG methods exploits the properties of high order Bernstein polynomials to construct sparse discrete operators leading to compact-stencil nonlinear approximations.

The proposed numerical methods are applied to various hyperbolic problems. Scalar equations are considered mainly for testing purposes and to simplify numerical analysis. Besides the shallow water system, we study the Euler equations of gas dynamics.

Chapter 1

Introduction

Many processes in nature, physics, and other application fields can be described by mathematical models based on *partial differential equations* (PDEs). Variables of these models may represent physical quantities such as density, velocity, or pressure of a fluid. The laws of physics may impose certain constraints on the main unknowns and/or derived quantities thereof. For instance, fluid density and internal energy should not become negative. In real life applications to fluid dynamics, computational methods are used to solve a continuous model problem approximately. In many cases, it is imperative that numerical approximations satisfy a subset of constraints that are known to hold for the exact solutions. However, *standard* techniques for solving PDEs either produce rather inaccurate (first order) approximations or may introduce spurious oscillations in the vicinity of steep fronts. The latter deficiency is known as the *Gibbs phenomenon* and can cause simulations to crash due to nonphysical solution values. To enforce appropriate constraints for approximate solutions, one needs to employ a *property-preserving discretization* of the continuous model problem. In essence, one should strive to design numerical methods capable of producing results that are in agreement with all important properties of the exact solution. At the same time, the approximation should be as accurate as possible under the imposed constraints.

This thesis presents some of the author's recent contributions to the development of physics-aware numerical methods. Particular emphasis is laid on nonlinear PDE systems with applications to geophysical fluid dynamics and gas flows. Scalar model problems are also studied in this work, mainly for testing purposes. Geophysical flows have numerous applications in environmental modeling and prediction of natural disasters (tsunamis, storm surges, dam breaks, etc.). These events are commonly modeled with the system of *shallow water equations*. Gas dynamics models are important for many industrial applications. For instance, they are used to study the air flow around certain objects (e. g., airfoils), to find out how they are affected by external forces, and also to improve their design. The hydrodynamic behavior of air and other gases can be reasonably well described by the *Euler equations*, which represent a simplification of the compressible *Navier–Stokes equations* for viscous flows. Both the Euler system and the inviscid shallow water equations are representatives of nonlinear *hyperbolic* systems. Such mathematical models are often challenging to solve (both analytically and numerically) for a number of reasons. For instance, steep fronts that are difficult to capture numerically may develop in a finite time, even if the solution is smooth initially. Furthermore, solutions to most hyperbolic problems are generally nonunique but in many cases there exists a unique *physical* solution. However, some computational methods may produce sequences of

nonphysical approximations. Moreover, the occurrence of inadmissible states such as negative fluid densities and internal energies in gas dynamics or negative water heights in shallow water flows, causes simulations to either break down or produce meaningless results.

The reason why the design of property-preserving methods is still an active research area can be attributed to the fact that only *monotone* schemes can be shown to satisfy all relevant constraints unconditionally. The accuracy of such schemes is limited to first order, as shown by Godunov [God59] in the linear case and by Harten et al. [Har83b] in the nonlinear case. The weaker requirement of *monotonicity preservation* implies monotonicity for linear schemes but makes it possible to design *nonlinear* discretizations that are both non-oscillatory and higher than first order accurate. Such nonlinear methods are referred to as *high resolution schemes*. In a typical method of this kind, the amount of artificial viscosity is adjusted adaptively based on the local regularity of a numerical solution. For instance, if an approximation is locally smooth, then the algorithm uses some high order baseline scheme. In the vicinity of steep fronts, a low-order method is employed instead. Existing schemes differ in the choice of constraints and algorithms for blending high and low order approximations in this manner.

Many property-preserving schemes for hyperbolic problems are based on *finite volume* (FV) space discretizations [Zal79, Swe84, Bar89, Kur07a, Noe07]. These methods evolve piecewise constant approximations to the exact solution and rely on certain reconstruction techniques to obtain discretizations that are more than first order accurate [Har87, Jia96, Aud04, Fjo11, Que21]. Some representatives of such FV methods are discussed in Section 1.1. In this work, we discretize the governing equations in space using *finite element methods* (FEM). The corresponding piecewise polynomial approximations are evolved without using reconstruction techniques. Numerical analysis of linear finite element approximations is less involved than the corresponding theory for FV schemes, and high order baseline methods are easy to derive. Property-preserving FEM are usually more difficult to construct than their FV counterparts, however, significant advances have been made in recent years [Bur07, Ric09, Kuz12a, Gue16b, Bad17]. In this thesis, we focus on the design of *algebraic flux correction* (AFC) schemes that enforce relevant constraints using *limiting techniques* (as in [Kuz05, Bar16, And17, Loh19, Paz21]). The underlying theory guarantees preservation of certain properties under suitable assumptions such as time step restrictions for explicit and semi-implicit schemes.

1.1 State of the art

Many AFC approaches and alternative high-resolution schemes were proposed in the literature over the past 50 years. We briefly review some classical algorithms and recent trends in this section. The review of the state of the art in the field of modern AFC tools for finite element discretizations is continued in Sections 3.3.1 and 5.1.

In the FV context, locally conservative adaptive schemes can be constructed using convex combinations of first and higher order numerical fluxes, as proposed by Harten and Zwas [Har72]. The choice of weights for such hybrid flux approximations may be based on smoothness indicators or nonlinear stability criteria. The first property-preserving high-resolution scheme of this kind was the *flux-corrected transport* (FCT) algorithm introduced by Boris and Book [Bor73, Boo75]. A typical implementation of FCT splits each solution update into two stages. In the first stage, the numerical solution is advanced in time using a monotone low order method. In the second stage, the accuracy is improved by adding limited antidiffusive fluxes. The FCT limiter proposed by Zalesak [Zal79] is fully multidimensional and applicable to unstructured meshes. It constrains the fluxes to preserve local maxima and minima of the low order predictor. An alternative one-step limiting strategy is adopted in *total variation diminishing* (TVD) methods [Har84]. The TVD property rules out occurrences of spurious oscillations and provides nonlinear stability, which is needed in standard proofs of convergence. A general framework for the design of TVD limiter functions was developed by Sweby [Swe84], who derived sufficient conditions for a FV scheme to be second order accurate and TVD in the 1D case. The accuracy of high-resolution schemes with TVD-type flux limiting varies between second order in smooth regions and first order at local extrema. One-dimensional TVD schemes can be generalized to 2D/3D using operator splitting on structured grids (see [LeV92, Ch. 18]) or the concept of *local extremum diminishing* (LED) schemes for unstructured grids (see [Jam93]). Such extensions often exhibit second order convergence behavior in practice, while genuine TVD schemes are at most first order accurate in multidimensions [LeV92, Thm. 18.3].

As an alternative to flux limiting of FCT and TVD type, Harten and Osher [Har87] proposed *essentially non-oscillatory* (ENO) schemes. As this name suggests, ENO approaches suppress spurious oscillations but are not TVD. Therefore, they can be more than first order accurate at local extrema and capture smooth traveling peaks much better than FCT and TVD methods. On the other hand, small undershoots/overshoots may occur in the vicinity of sharp peaks and discontinuities. The original ENO method [Har87] selects stencils for polynomial reconstructions adaptively to avoid interpolation across discontinuities and obtain flux approximations that are uniformly second order accurate for smooth data. To prevent the occurrence of Gibbs phenomena, a *minmod limiter* is applied in the process of computing the reconstructions.

Building on the ENO methodology, Liu et al. [Liu94] introduced the framework of *weighted essentially non-oscillatory* (WENO) schemes. One improvement of the latter approach over ENO is the use of convex combinations of local reconstruction polynomials with adaptively chosen nonlinear weights. By employing larger nodal stencils than in the original ENO method [Har87], WENO schemes are capable of producing approximations of increased resolution. Strategies of WENO type remain popular to this day (see, e. g., [Jia96, Shu98, Zha11, Dum14, Que21]) due to their high rates of convergence to smooth solutions and robustness in situations in which discontinuities are present. As

illustrated in [Zha11, Sec. 1], higher than second order of accuracy comes at the price of being unable to strictly enforce discrete maximum principles. Similarly to their ENO predecessors, WENO schemes may accentuate local extrema and violate global bounds.

In contrast to the aforementioned FV approaches, traditional stabilization techniques for finite elements modify the discrete weak form of the governing equations rather than the numerical fluxes of a discrete conservation law. In the classical *streamline upwind Petrov–Galerkin* (SUPG) method [Bro82] and discontinuity-capturing extensions thereof [Hug86, Cod93], the stabilization terms represent weighted residuals that vanish at the continuous level and introduce artificial viscosity at the discrete level. The amount of numerical dissipation in the streamline and crosswind direction depends on user-defined parameters. Similarly to ENO schemes, such parameter-dependent stabilized FEM approaches do not rule out the occurrence of spurious ripples at discontinuities.

Algebraic flux correction schemes for finite elements [Kuz02, Kuz05] are based on various generalizations of FCT and TVD algorithms, which can also be interpreted as nonlinear artificial viscosity methods. In contrast to SUPG, the additional dissipative terms do not represent weighted residuals. They are defined using discrete diffusion (graph Laplacian) operators and admit a conservative decomposition into adjustable numerical fluxes. Modern AFC discretizations are guaranteed to be property preserving, and the amount of numerical viscosity (usually) does not depend on free parameters.

Another promising stabilization approach based on artificial viscosity is the one proposed by Burman [Bur07]. His finite element scheme for the 1D Burgers equation employs a continuous Galerkin approximation, and the artificial viscosity operator is designed to ensure the validity of a discrete maximum principle. This property makes it possible to prove convergence to a weak solution [Bur07, Thm. 3.7]. Moreover, the limit of a convergent sequence is shown to satisfy the weak form of an entropy inequality [Bur07, Thm. 3.8]. However, Burman's approach does not guarantee the validity of a semi-discrete or fully discrete entropy inequality for any finite mesh size.

Badia and Bonilla [Bad17] discretize scalar conservation laws using a nonlinear AFC scheme in which the artificial diffusion coefficients are limited using a nodal shock detector. The amount of limiting depends on the ratio of jumps and averages of directional derivatives at the nodal points. Similarly to the sensor analyzed by Barrenechea et al. [Bar17a], the jump-average indicator employed in [Bad17] is a generalization of the one that adjusts numerical dissipation in the classical Jameson–Schmidt–Turkel (JST) scheme [Jam17]. The multidimensional AFC version proposed in [Bad17] ensures the validity of discrete maximum principles and *linearity preservation*. The latter property is an essential requirement for second order consistency and optimal convergence on general meshes. Another highlight of the methodology developed in [Bad17] is the possibility of using implicit time integrators and Newton-like solvers for a regularized version of the nonlinear discrete problem. The proposed regularization of the shock detector makes it differentiable without losing the LED property of the spatial semi-discretization.

Another framework that shares many similarities with the AFC methodology is that

of residual distribution (RD) schemes, e. g., [Abg06, Ric09, Abg17]. As pointed out by Abgrall [Abg06], many well-known finite element and finite volume methods can be interpreted as RD approaches. For steady problems, the RD formalism can be summed up as follows: Compute elementwise residuals for a baseline discretization of the PDE and decompose them into nodal contributions. The weights of these distributions can be chosen in such a way that the resulting approximations are essentially non-oscillatory [Abg06] or even LED, as in our own contributions [Haj20b, Haj20c] to the field. In contrast to classical AFC schemes, the design of property-preserving RD methods for time dependent problems is more involved than for stationary ones [Abg06, Abg17]. Moreover, RD discretizations of hyperbolic systems may fail to ensure positivity preservation.

A promising new approach to the design of property preserving schemes is the *multidimensional optimal order detection* (MOOD) methodology for a posteriori limiting [Dio13]. The MOOD procedure developed by Dumbser et al. [Dum14] combines a high order discontinuous Galerkin (DG) method with a subcell finite volume scheme using a MOOD-type troubled-cell indicator. In each time step, the approximate solution is first advanced in time using the baseline DG scheme. If either physical or numerical admissibility conditions are violated in any cell, the high order DG approximation in that cell is rejected and replaced by a WENO-FV approximation on subcells of the element. Local L^2 projections are used to ensure conservation when it comes to calculating the piecewise constant initial data for the FV update and project the result back into the high order DG space. The validity of all user-defined admissibility criteria is guaranteed, provided that the subcell FV scheme possesses the desired properties. However, the binary switch between the DG and subcell FV approximations does not provide continuous dependence on the data, which may cause convergence problems in the steady-state limit. In a similar context, Hennemann et al. [Hen21] recently designed a continuous blending strategy between a DG target scheme and a subcell FV method. Special care is taken to ensure discrete entropy stability, which is achieved by using carefully designed numerical fluxes. In contrast to the MOOD approach of Dumbser et al. [Dum14], preservation of local bounds is not enforced. More recently, Rueda-Ramírez et al. [Rue22] showed the equivalence of a subcell FV scheme to the low order method proposed by Pazner [Paz21] in the context of a localized FCT scheme for DG discretizations of hyperbolic systems. This finding makes it possible to design DG/subcell FV methods that are not only entropy stable but also bound preserving. To the best of our knowledge, the work published in [Rue22] represents the first bridge between subcell FV limiters and AFC schemes for high order DG methods.

1.2 Outline and originality of this thesis

This introductory chapter ends with a section in which we introduce some notation and commonly used symbols. In the next chapter, we review some important facts about

weak solutions to hyperbolic problems. Before presenting an excerpt on the mathematical theory, we address modeling aspects and discuss the structure of the problems under consideration. The contents of Chapter 2 do not represent any original research by the author and are to a large degree based on [LeV92, Vre94, Daf00, Fei03, Eck17]. A basic introduction to standard discretizations in space and time is given in Chapter 3. The remainder of that chapter is focused on algebraic flux correction tools for finite elements. In particular, *monolithic convex limiting* (MCL) strategies yielding bound-preserving [Kuz20a] and entropy-stable [Kuz20c, Kuz22a] schemes are reviewed and extended. These methods remain relevant throughout this thesis and are therefore discussed in detail. In Chapter 4, we generalize the MCL framework for conservation laws to a system of balance laws. Specifically, we consider the shallow water equations (SWE) with a topography source term. Our tailor-made extension of limiting strategies to this problem is provably well balanced, positivity preserving, and entropy stable. Numerical analysis for MCL space discretizations of the advection equation is presented in Chapter 5. Therein, we slightly improve upon the stability analysis and a priori error estimates from our preprint [Haj21b]. Finally, Chapter 6 is based on the author's paper [Haj21a] on MCL schemes for high order DG discretizations. These methods are applied to (systems of) conservation laws, which include the SWE and the compressible Euler equations. In addition to selected contents of [Haj21a], we present some new features and numerical examples before concluding this thesis in Chapter 7.

1.3 Notation and list of symbols

For the reader's convenience, we summarize some notational conventions for further use in this thesis. The most important symbols can be found in Tabs. 1.1 to 1.3.

Bold font lower case letters are reserved for quantities that assume values in $\mathbb{R}^d$, $d \in \{1, 2, 3\}$, i. e., vectors in physical space (e. g., velocities or forces). The components of a vector are denoted using the non-bold font for the same symbol. A subscript is used to indicate which component of the vector is referred to, e. g., $\boldsymbol{x} = (x_1, \dots, x_d)^\mathsf{T} \in \mathbb{R}^d$. Similar notation is used for matrices for which we mostly employ upper case letters, e. g., $A = (a_{ij})_{i,j=1}^n$ for square matrices or $A = (a_{ij})_{\substack{i=1,\dots,k \\ j=1,\dots,n}}$ for $(k \times n)$-matrices, where $k, n \in \mathbb{N}$. Calligraphic upper case letters such as $\mathcal{A}$, are reserved for sets.

The closure of a domain Ω w. r. t. a certain metric is denoted as $\overline{\Omega}$. The interior of a set Ω w. r. t. a certain metric is denoted as $\mathrm{int}(\Omega)$. The diameter of a domain $\Omega \subseteq \mathbb{R}^d$ is denoted as $\mathrm{diam}(\Omega) = \sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega} |x - y|$. The open ball of radius $r > 0$ around $\boldsymbol{x} \in \mathbb{R}^d$ w. r. t. the Euclidean norm is $\mathrm{B}_r(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^d : |\boldsymbol{x} - \boldsymbol{y}| < r\}$. The $(d-1)$-dimensional unit sphere centered at the origin is denoted as $\mathrm{S}_1^{d-1} = \{\boldsymbol{n} \in \mathbb{R}^d : |\boldsymbol{n}| = 1\}$.

Assignments such as $f = f(u) \in \mathbb{R}^k$, $k \in \mathbb{N}$ are used to indicate that f is an $\mathbb{R}^k$-valued function of a certain variable u. If a function depends on a single variable, its derivatives are denoted by apostrophes. In this fashion, if $f : \mathbb{R}^k \to \mathbb{R}^n$, then

$f' : \mathbb{R}^k \to \mathbb{R}^{n\times k}$, k, $n \in \mathbb{N}$ is the Jacobian and f'_{ij} is the derivative of the ith component of f w. r. t. the jth component of the variable it depends on. The divergence $\nabla\cdot$ of a vector field $\boldsymbol{a} = (a_i)_{i=1}^d = \boldsymbol{a}(\boldsymbol{x})$ is defined as

$$\nabla \cdot \boldsymbol{a} = \sum_{i=1}^{d} \frac{\partial a_i(\boldsymbol{x})}{\partial x_i}.$$

The divergence $\nabla\cdot$ of a matrix-valued function $\boldsymbol{A} = (a_1, \dots, a_d) = \boldsymbol{A}(\boldsymbol{x}) \in \mathbb{R}^{k\times d}$, $k \in \mathbb{N}$ with $a_i(\boldsymbol{x}) \in \mathbb{R}^k$, $i \in \{1, \dots, d\}$ is defined as

$$\nabla \cdot \boldsymbol{A} = \sum_{i=1}^{d} \frac{\partial a_i(\boldsymbol{x})}{\partial x_i} \in \mathbb{R}^k.$$

The bold symbol $\boldsymbol{\nabla}\cdot$ denotes the divergence operator for $(d \times d)$-valued matrices (as opposed to the general non-boldface notation for $k \neq d$).

Symbol	Description
$\lvert\cdot\rvert$	Euclidean norm for vectors in $\mathbb{R}^k$, $k \in \mathbb{N}$
$a \cdot b$	Euclidean scalar product of two vectors $a, b \in \mathbb{R}^k$, $k \in \mathbb{N}$
$A : B$	scalar product of two matrices $A : B = \operatorname{tr}(A^\mathsf{T} B)$, where tr is the trace
$\mathbb{R}_+$	set of positive real numbers $\mathbb{R}_+ = \{t \in \mathbb{R} : t > 0\}$
δ_{ij}	Kronecker delta, $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$
$\mathcal{I}$	identity matrix in $\mathbb{R}^{d\times d}$, where $d \in \{1, 2, 3\}$ is the spatial dimension
$\mathcal{I}_{k\times k}$	identity matrix in $\mathbb{R}^{k\times k}$, $k \in \mathbb{N}$
$\operatorname{supp} w$	support of a function $w \in \mathrm{C}(\Omega)$, $\operatorname{supp} w = \overline{\{\boldsymbol{x} \in \Omega : w(\boldsymbol{x}) \neq 0\}}$
$\mathrm{C}(\Omega)$	space of continuous functions
$\mathrm{C}^k(\Omega)$	space of k times continuously differentiable functions, $k \in \mathbb{N}_0 \cup \{\infty\}$
$\mathrm{L}^p(\Omega)$	Lebesgue space of measurable, pth-power integrable functions, $p \in [1, \infty)$
$\mathrm{L}^\infty(\Omega)$	Lebesgue space of measurable, essentially bounded functions
$\mathrm{L}^1_{\mathrm{loc}}(\Omega)$	space of locally integrable functions
$\mathrm{C}^1_0(\Omega)$	space of continuously differentiable functions with compact support in Ω
$\mathrm{W}^{k,p}(\Omega)$	Sobolev space of functions with kth order derivatives in $\mathrm{L}^p(\Omega)$
$\mathrm{H}^k(\Omega)$	Sobolev space $\mathrm{H}^k(\Omega) = \mathrm{W}^{k,2}(\Omega)$

Table 1.1: Important operators, symbols, and function spaces.

Symbol	Description
d	spatial dimension $d \in \{1, 2, 3\}$
m	number of unknowns in the solution vector $m \in \mathbb{N}$
Ω	spatial domain $\Omega \subseteq \mathbb{R}^d$
$\partial\Omega$	boundary of Ω
Γ	boundary segment $\Gamma \subseteq \partial\Omega$ (possibly with sub- and superscripts)
T	final time for a model problem
t	temporal variable $t \in [0, T]$
$\boldsymbol{x}$	spatial variable $\boldsymbol{x} \in \mathbb{R}^d$
$\boldsymbol{n} = \boldsymbol{n}(\boldsymbol{x})$	normal vector/outward unit normal to $\partial\Omega$
$u = u(\boldsymbol{x}, t)$	solution vector $u(\boldsymbol{x}, t) \in \mathbb{R}^m$
$u_0 = u_0(\boldsymbol{x})$	initial data $u_0(\boldsymbol{x}) \in \mathbb{R}^m$
$\hat{u} = \hat{u}(\boldsymbol{x}, t)$	external Riemann data $\hat{u}(\boldsymbol{x}, t) \in \mathbb{R}^m$, as defined in Section 2.2
$\mathbf{f} = \mathbf{f}(u)$	inviscid flux function $\mathbf{f} \in \mathrm{C}^1(\mathbb{R}^m)^{m \times d}$
$\mathbf{f}'_{\boldsymbol{n}} = \mathbf{f}'_{\boldsymbol{n}}(u)$	directional Jacobian of the inviscid flux function, $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$
$\lambda = \lambda(u)$	wave speeds of hyperbolic problems used in various contexts
$\rho = \rho(\boldsymbol{x}, t)$	density of a fluid
$\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$	velocity of a fluid, depth-averaged in the shallow water equations
$p = p(\boldsymbol{x}, t)$	pressure of a fluid
$E = E(\boldsymbol{x}, t)$	specific total energy of a fluid
$h = h(\boldsymbol{x}, t)$	total water height in the shallow water equations
$H = H(\boldsymbol{x}, t)$	free surface elevation in the shallow water equations
$b = b(\boldsymbol{x})$	bathymetry/bottom topography in the shallow water equations
$\eta = \eta(u)$	mathematical entropy of a hyperbolic PDE (system)
$\boldsymbol{q} = \boldsymbol{q}(u)$	entropy flux $\boldsymbol{q}(u) \in \mathbb{R}^{1 \times d}$ corresponding to an entropy η
$v = v(u)$	entropy variable $v(u) = \frac{\partial}{\partial u}\eta(u) \in \mathbb{R}^m$ corresponding to η
$\boldsymbol{\psi} = \boldsymbol{\psi}(u)$	entropy potential $\boldsymbol{\psi}(u) = v(u)^{\mathsf{T}}\mathbf{f}(u) - \boldsymbol{q}(u)$ corresponding $(\eta, \boldsymbol{q})$

Table 1.2: Important physical and mathematical quantities.

For the shallow water equations with nonflat bottom topography, i. e., $b \neq \text{const}$, the entropy pair $(\eta, \boldsymbol{q})$ and corresponding entropy variable v depend on b in addition to u. The entropy potential $\boldsymbol{\psi}$ remains independent of b, see Section 2.2.3 for details.

Symbol	Description
E	number of elements in a mesh
N	number of degrees of freedom
$\mathcal{K}_h$	computational mesh with spacing $h = \max_{K\in\mathcal{K}_h} \operatorname{diam}(K)$
$\boldsymbol{x}_i$	vertex of a mesh, $\boldsymbol{x}_i \in \overline{\Omega}$
K^e	mesh element/cell with index $e \in \{1, \dots, E\}$
K	generic element/cell $K \in \mathcal{K}_h$
$\mathbb{P}_p(K)$	space of polynomials of degree at most p in d variables, where $K \subset \mathbb{R}^d$
$\mathbb{Q}_p(K)$	space of polynomials of degree at most p w. r. t. each variable
$\mathbb{V}_p(K)$	polynomial space $\mathbb{P}_p(K)$ if K is a simplex, $\mathbb{Q}_p(K)$ otherwise
V_h	space of continuous piecewise (multi-)linear functions, $\mathrm{V}_h = \mathrm{V}_{h,1}(\mathcal{K}_h)$
$\varphi_i = \varphi_i(\boldsymbol{x})$	Lagrange basis function associated with vertex $\boldsymbol{x}_i$
u_h	approximate solution $u_h = u_h(\boldsymbol{x}, t) = \sum_{i=1}^N u_i(t)\varphi_i(\boldsymbol{x})$
u_i	nodal state of an approximate solution, $u_i = u_i(t) \in \mathbb{R}^m$
$\mathbf{f}_i$	inviscid flux function evaluated at $u_i \in \mathbb{R}^m$, $\mathbf{f}_i = \mathbf{f}(u_i) \in \mathbb{R}^{m\times d}$
$\mathrm{f}_{\boldsymbol{n}}(\cdot,\cdot)$	numerical flux in the space direction $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$, $\mathrm{f}_{\boldsymbol{n}}(\cdot,\cdot) \in \mathbb{R}^m$
$\mathcal{N}_i$	nodal stencil, $\mathcal{N}_i = \{j \in \{1,\dots,N\} : \operatorname{int}(\operatorname{supp}\varphi_i) \cap \operatorname{int}(\operatorname{supp}\varphi_j) \neq \emptyset\}$
$\mathcal{F}_i$	nodal boundary faces, see Definition 3.6
m_{ij}	entries of the consistent mass matrix M, see (3.8)
m_i	diagonal entries of the row sum lumped mass matrix M_{L}, see (3.24)
$\mathbf{c}_{ij}$	entries of the discrete gradient operator, see (3.20)
d_{ij}	artificial diffusion coefficients, see (3.27) for standard Rusanov values
b_i^k	integral of φ_i over a boundary face $\Gamma_k \in \mathcal{F}_i$, see (3.23)
d_i^k	Lax–Friedrichs diffusion coefficient of a boundary node, see (3.30)
$\bar{u}_{ij}$, $\bar{u}_{ij}^*$	low order and limited bar states $\bar{u}_{ij}$, $\bar{u}_{ij}^* \in \mathbb{R}^m$, $i \neq j$, see (3.28), (3.40)
$\bar{u}_i^k$	low order bar state of a boundary node, see (3.31)
f_{ij}, f_{ij}^*	raw antidiffusive fluxes and their limited counterparts, f_{ij}, $f_{ij}^* \in \mathbb{R}^m$
$\dot{u}_i^{\mathrm{L}}$	approximate low order time derivative of a nodal state, $\dot{u}_i^{\mathrm{L}} \in \mathbb{R}^m$
$u_i^{\min}$, $u_i^{\max}$	lower and upper bounds to be imposed on u_i via limiting
α_{ij}, β_{ij}	correction factors for limiting $\alpha_{ij} = \alpha_{ji} \in [0,1]$, $\beta_{ij} = \beta_{ji} \in [0,1]$
ϱ	first component of the solution $u = (\varrho, \varrho\phi_1, \dots, \varrho\phi_{m-1})$ for systems
$\phi = (\varrho\phi)/\varrho$	specific quantity corresponding to the conserved unknown $(\varrho\phi)$
$\bar{\varrho}_{ij}$, $\bar{\varrho}_{ij}^*$	low order and limited bar states of ϱ
$\overline{\phi}_{ij}$	low order bar states of the specific quantity ϕ
$\overline{(\varrho\phi)}_{ij}$	low order bar states of the conserved unknown $(\varrho\phi)$
$\overline{(\varrho\phi)}_{ij}^*$	limited bar states of the conserved unknown $(\varrho\phi)$
Δt	time step
ν	Courant–Friedrichs–Lewy parameter $\nu \in (0,1]$

Table 1.3: Symbols used in descriptions of numerical methods.

Chapter 2

Theory of hyperbolic problems

The development of numerical methods for flow problems requires a profound knowledge of the underlying mathematical theory. For instance, in order to design property-preserving discretization techniques for solving PDEs, one needs to be familiar with the structure of the problem at hand. Indeed, many concepts arising on the continuous level of a given model have discrete counterparts in computational schemes. In the hyperbolic case, for example, some methods enforce discrete entropy inequalities, which correspond to the ones satisfied by vanishing viscosity solutions.

This chapter addresses some theoretical aspects of hyperbolic equations. We begin with deriving two hyperbolic systems from the Navier–Stokes equations of fluid dynamics in Section 2.1. In Section 2.2, we discuss the properties of the scalar problem and of the systems that we investigate in this thesis. Section 2.3 closes this chapter with an excerpt on the theory of weak admissible solutions to conservation laws.

2.1 Modeling aspects

Since the models that we consider in this thesis are only valid under certain assumptions, we review the derivation of the corresponding systems of equations. Starting with the most general flow model, we neglect viscous friction and heat conduction effects. These simplifications enable us to omit all terms that depend on second order partial derivatives and to only consider first-order PDE systems. We refer to [Eck17, Sec. 5.6] for a presentation of constitutive relations for the omitted terms. Moreover, we do not discuss effects of turbulence and Reynolds averaging, nor the Boussinesq approximation (see, e. g., [Cus11, Secs. 4.1, 3.7]) since they are irrelevant for the purposes of this thesis. In Section 2.1.1, we outline the derivation of the compressible Euler equations, which govern the hydrodynamic behavior of gas flows under the above assumptions. In Section 2.1.2, we derive the shallow water equations (SWE), an important model for simulations of geophysical flows.

2.1.1 Compressible Euler equations

The compressible Navier–Stokes equations are among the most general mathematical models in fluid dynamics. For flows at low Mach numbers, they can be approximated using a hierarchy of incompressible flow models. The derivation of the Navier–Stokes system from the principles of continuum physics is discussed in detail in [Eck17, Ch. 5].

In this section, we consider the system of Euler equations, which are obtained from the Navier–Stokes model by omitting all terms that include second order derivatives.

To formulate initial boundary value problems that govern the evolution of conserved quantities in a spatial domain $\Omega \subseteq \mathbb{R}^d$, $d \in \{1, 2, 3\}$, we need to introduce some notation (cf. Section 1.3). Let $\nabla\cdot$ and $\boldsymbol{\nabla}\cdot$ denote the divergence operators for vector fields and tensor-valued quantities, respectively. We write $a \otimes b = a\, b^{\mathsf{T}} \in \mathbb{R}^{k\times n}$ for the dyadic product of two vectors $a \in \mathbb{R}^k$, $b \in \mathbb{R}^n$. If $n = k$, the Euclidean scalar product of a and b is denoted as $a \cdot b = a^{\mathsf{T}} b$. The physical variables of compressible gas dynamics include the mass density $\rho = \rho(\boldsymbol{x}, t)$, velocity $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$, specific total energy $E = E(\boldsymbol{x}, t)$, and pressure $p = p(\boldsymbol{x}, t)$ of the fluid. Furthermore, let $\boldsymbol{g}(\boldsymbol{x}, t)$ represent the density of specific external forces acting on the fluid at $\boldsymbol{x} \in \Omega$ and $t \geq 0$. If gravity is the only force to be taken into account, which is usually the case, then $\boldsymbol{g} = -g\, \mathbf{e}_d$, where g is gravitational acceleration and $\mathbf{e}_d \in \mathbb{R}^d$ is the Cartesian unit vector pointing in the upward direction. Finally, let the identity matrix be denoted as $\mathcal{I} \in \mathbb{R}^{d\times d}$. With this notation, we are now ready to present the *compressible Euler equations* [Eck17, Ch. 5]

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0, \tag{2.1a}$$

$$\frac{\partial (\rho \boldsymbol{v})}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{v} \otimes \boldsymbol{v} + p\, \mathcal{I}) = \rho \boldsymbol{g}, \tag{2.1b}$$

$$\frac{\partial (\rho E)}{\partial t} + \nabla \cdot ((\rho E + p) \boldsymbol{v}) = \rho \boldsymbol{g} \cdot \boldsymbol{v}, \tag{2.1c}$$

also known as the *Euler equations of gas dynamics*. This system models the flow of compressible gases at high speeds. Viscous friction and heat conduction effects are neglected in this model. The gravitation-induced terms on the right hand sides of (2.1b) and (2.1c) can also be omitted in many applications.

System (2.1) consists of $d + 2$ equations for $d + 3$ unknowns ρ, $\boldsymbol{v}$, E, and p. To formulate a closure for the Euler equations, we use thermodynamic relations and derive an *equation of state* for the pressure. Our presentation is based on [LeV92, Sec. 5.1.1] and [Fei03, Sec. 3.1.1]. First, we assume that the fluid obeys the ideal gas law

$$p = (c_P - c_V)\, \rho\, \theta, \tag{2.2}$$

where θ is the absolute temperature, c_V is the specific heat at constant volume, and $c_P > c_V$ is the specific heat at constant pressure. As remarked in [LeV02, Sec. 14.4] $c_P - c_V$ is equal to the universal gas constant divided by the molecular weight of the gas. In this thesis, we assume the gas to be *polytropic*, i. e., its specific internal energy (thermal energy) is given by

$$e = c_V\, \theta. \tag{2.3}$$

The specific total energy E of the fluid represents the sum of its specific internal and kinetic energies. Thus, we have

$$E = e + \frac{|\boldsymbol{v}|^2}{2}. \tag{2.4}$$

Introducing the *adiabatic constant* $\gamma = c_P/c_V > 1$, we obtain the equation of state

$$p = \frac{c_P - c_V}{c_V}\,\rho e = (\gamma - 1)\,\rho e = (\gamma - 1)\left(\rho E - \frac{\rho|\boldsymbol{v}|^2}{2}\right) \tag{2.5}$$

from (2.2)–(2.4). Thus, (2.5) provides a closure for system (2.1) valid for ideal polytropic gases. The value of γ depends on the molecular structure of fluid particles. For monatomic gases, characterized by the fact that the particles are atoms rather than molecules, we have $\gamma = \frac{5}{3}$. The adiabatic constant of diatomic gases such as the main components of air, nitrogen (N_2) and oxygen (O_2), assumes the value $\frac{7}{5} = 1.4$ (see [LeV92, Sec. 5.1.1] for details).

2.1.2 Shallow water equations

Besides the compressible Euler equations, this thesis is concerned with the system of shallow water equations, which describe the two- or three-dimensional motion of water under assumptions to be discussed below. The use of depth averaging reduces the number of velocity components and the dimension of the spatial domain by one. Shallow water flows arise in geophysics and have a wide range of applications. In particular, the SWE can be used to model flooding events, for instance, those caused by storm surges, tsunamis, or breakage of dams. Moreover, (multi-layer generalizations of) the SWE are widely used in simulation tools for large scale atmospheric and oceanic flows.

In this section, we summarize the derivation of the SWE. Our presentation is based on [Vre94, Ch. 2]. Once again, we consider the inviscid case, thus we start from the incompressible Euler equations. Furthermore, we neglect the effects of bottom friction, rotational effects (Coriolis force) and stratification. These issues are discussed in detail in [Vre94, Ch. 2] and [Cus11, Pts. II–IV].

Since we are interested in geophysical flows, the fluid under investigation is water, which can be assumed to be incompressible. That is, its density ρ is approximately constant. In this setting, the balance law (2.1c) for the total energy of water is omitted, and the incompressible Euler equations

$$\nabla \cdot \boldsymbol{v} = 0, \tag{2.6a}$$

$$\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{\nabla} \cdot \left(\boldsymbol{v} \otimes \boldsymbol{v} + \frac{p}{\rho}\mathcal{I}\right) = \boldsymbol{g} \tag{2.6b}$$

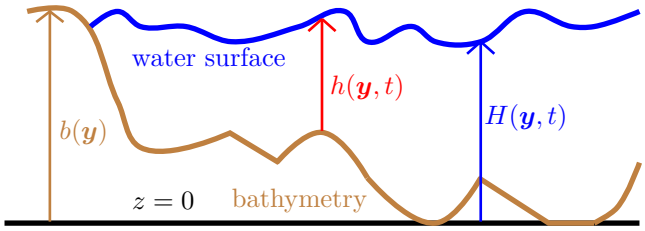


Figure 2.1: Cross section of a body of water.

are used as the starting point for the derivation of the SWE in $\mathbb{R}^d$. Contrary to the Euler equations of gas dynamics, the gravitational source term $\boldsymbol{g} = -g\,\mathbf{e}_d$, cannot be neglected here. As the name suggests, the derivation of the shallow water equations is based on the assumption that the water depth is small compared to other dimensions of the flow domain. Formally, one introduces typical horizontal and vertical length scales L and A, respectively, and assumes $L \gg A$. For geophysical applications this assumption is justified, for instance, if one considers oceanic flows. The horizontal scale may be hundreds to ten thousands of kilometers wide, while the deepest point in the Earth's ocean (Mariana trench) is a just short of being eleven kilometers below the mean sea level.

We introduce the following notation to distinguish between horizontal and vertical spatial directions. Let $\boldsymbol{x}^\mathsf{T} = (\boldsymbol{y}^\mathsf{T}, z) \in \mathbb{R}^d$ be split into the vector $\boldsymbol{y} \in \mathbb{R}^{d-1}$ of horizontal components and the vertical component z. The horizontal components of the gradient operator will be denoted by $\nabla_{\boldsymbol{y}}$. Similarly, the velocity vector $\boldsymbol{v}^\mathsf{T} = (\boldsymbol{u}^\mathsf{T}, w) \in \mathbb{R}^d$ is split into $\boldsymbol{u} \in \mathbb{R}^{d-1}$ and w. The *free surface elevation* H and bottom topography b (also called *bathymetry*) represent the upper and lower boundary of the flow domain, respectively. The total height of water h is thus equal to $H - b$, and all three quantities are independent of z. This setup is illustrated for a cross section of the flow domain in Fig. 2.1.

To simplify system (2.6) as in [Vre94, Sec. 2.4] or [Cus11, Sec. 4.3], the relative magnitude of all terms is analyzed w. r. t. the spatial length scales L and A, as well as w. r. t. a typical horizontal velocity U and a time scale T. If all terms that are small in magnitude are omitted, the vertical component of (2.6b) becomes

$$\frac{\partial p}{\partial z} = -g\rho. \tag{2.7}$$

The self-evident interpretation of the *hydrostatic pressure assumption* (2.7) is that any vertical change in the pressure is caused by the weight of water. This approximation is valid for most applications of the SWE. From (2.7) we obtain an explicit expression for the pressure using the fundamental theorem of calculus

$$p(\boldsymbol{y}, z, t) = p(\boldsymbol{y}, H(\boldsymbol{y},t), t) + \int_{H(\boldsymbol{y},t)}^{z} \frac{\partial p}{\partial \zeta}(\boldsymbol{y}, \zeta, t)\,\mathrm{d}\zeta = p(\boldsymbol{y}, H(\boldsymbol{y},t), t) - g\rho\,(z - H(\boldsymbol{y},t))\,.$$

The horizontal pressure gradient thus reads

$$\nabla_{\boldsymbol{y}} p(\boldsymbol{y}, z, t) = \nabla_{\boldsymbol{y}} p(\boldsymbol{y}, H(\boldsymbol{y}, t), t) + g\rho \nabla_{\boldsymbol{y}} H(\boldsymbol{y}, t).$$

The first term on the right is the *atmospheric pressure gradient*, which we assume to be zero. In general, it appears as a source term in the right hand side.

Let us now define boundary conditions at the top and bottom of the water column. The corresponding boundaries are formally defined by $H(\boldsymbol{y}, t) - z = 0$ and $b(\boldsymbol{y}, t) - z = 0$, respectively. Taking the material derivatives of these expressions yields

$$\frac{\partial H}{\partial t} + \boldsymbol{u} \cdot \nabla_{\boldsymbol{y}} H - w = 0 \qquad \text{at } z = H(\boldsymbol{y}, t), \tag{2.8a}$$

$$\frac{\partial b}{\partial t} + \boldsymbol{u} \cdot \nabla_{\boldsymbol{y}} b - w = 0 \qquad \text{at } z = b(\boldsymbol{y}, t). \tag{2.8b}$$

In most applications of the SWE, the bathymetry b plays the role of a parameter and is often assumed to be independent of time t. If b is known, (2.8b) constitutes the boundary condition at the bottom of the flow domain. The free surface elevation H, on the other hand, is the primary unknown of the system, which at this intermediate stage consists of (2.8a), the continuity equation (2.6a), and $d - 1$ momentum equations

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{\nabla} \cdot (\boldsymbol{u} \otimes \boldsymbol{v}) + g \nabla_{\boldsymbol{y}} H = \boldsymbol{0}. \tag{2.9}$$

For $d = 3$, this system is referred to as the three-dimensional shallow water equations.

The final step towards obtaining the classical SWE is integration over the total depth of water. To this end, we define the depth-averaged horizontal velocity as

$$\bar{\boldsymbol{u}}(\boldsymbol{y}, t) = \frac{1}{h(\boldsymbol{y}, t)} \int_{b(\boldsymbol{y},t)}^{H(\boldsymbol{y},t)} \boldsymbol{u}(\boldsymbol{y}, z, t) \, \mathrm{d}z. \tag{2.10}$$

Omitting dependencies on $\boldsymbol{y}$ and t for the sake of readability, we integrate (2.6a) using the Leibniz rule. The boundary conditions (2.8) combined with the definitions of the total water height $h = H - b$ and depth-averaged velocity (2.10) produce

$$\begin{aligned} 0 &= \int_b^H \left[\nabla_{\boldsymbol{y}} \cdot \boldsymbol{u} + \frac{\partial w}{\partial z} \right] \mathrm{d}z = \nabla_{\boldsymbol{y}} \cdot \int_b^H \boldsymbol{u} \, \mathrm{d}z - \nabla_{\boldsymbol{y}} H \cdot \boldsymbol{u}|_H + \nabla_{\boldsymbol{y}} b \cdot \boldsymbol{u}|_b + w|_H - w|_b \\ &= \frac{\partial h}{\partial t} + \nabla_{\boldsymbol{y}} \cdot (h \bar{\boldsymbol{u}}). \end{aligned}$$

Similarly, we integrate the momentum equation (2.9) over the water column, which yields

$$\boldsymbol{0} = \frac{\partial (h \bar{\boldsymbol{u}})}{\partial t} + \boldsymbol{\nabla}_{\boldsymbol{y}} \cdot \int_b^H \boldsymbol{u} \otimes \boldsymbol{u} \, \mathrm{d}z + \frac{g}{2} \nabla_{\boldsymbol{y}} h^2 + g h \nabla_{\boldsymbol{y}} b.$$

Here we have again made use of the boundary conditions (2.8) and of the definition $h = H - b$. Since the integral of $\boldsymbol{u}$ in vertical direction equals the *discharge* $h\bar{\boldsymbol{u}}$ (also referred to as *momentum*, in analogy to the Euler equations), we obtain

$$\int_b^H (\boldsymbol{u} - \bar{\boldsymbol{u}} + \bar{\boldsymbol{u}}) \otimes (\boldsymbol{u} - \bar{\boldsymbol{u}} + \bar{\boldsymbol{u}}) \,\mathrm{d}z = \int_b^H (\boldsymbol{u} - \bar{\boldsymbol{u}}) \otimes (\boldsymbol{u} - \bar{\boldsymbol{u}}) \,\mathrm{d}z + h\bar{\boldsymbol{u}} \otimes \bar{\boldsymbol{u}}. \tag{2.11}$$

If we assume the difference between $\boldsymbol{u}$ and $\bar{\boldsymbol{u}}$ to be small, the integral on the right of (2.11) is negligible.

We are now in a position to formulate the system of shallow water equations. To avoid a cumbersome notation, we omit the subscript $\boldsymbol{y}$ of the gradient operator but keep in mind that only partial derivatives w. r. t. horizontal components constitute ∇ in the SWE context. Thus, the SWE are formulated either in the one- or in the two-dimensional setting , i. e., for $d \in \{1, 2\}$. From now on, we again denote the spatial variable as $\boldsymbol{x} \in \mathbb{R}^d$. Moreover, $\boldsymbol{v}$ replaces the symbol $\bar{\boldsymbol{u}}$. Using this convention, we may denote both the depth-averaged velocity in the SWE and the velocity in the Euler equations (2.1) by the same symbol. In conclusion, the shallow water equations read

$$\frac{\partial h}{\partial t} + \nabla \cdot (h\boldsymbol{v}) = 0, \tag{2.12a}$$

$$\frac{\partial (h\boldsymbol{v})}{\partial t} + \boldsymbol{\nabla} \cdot \Big(h\boldsymbol{v} \otimes \boldsymbol{v} + \frac{g}{2} h^2 \mathcal{I} \Big) + gh\nabla b = \boldsymbol{0}. \tag{2.12b}$$

Note that for *nonflat* bottom topography, i. e., in the case $\nabla b \neq \boldsymbol{0}$, the SWE are a system of balance laws rather than conservation equations.

2.2 Structure of the problems under consideration

Having discussed the derivation of the systems of equations considered in this thesis, we present some basic results regarding their structure. In the remainder of this chapter, the equations are assumed to be in nondimensional form. The derivation thereof is illustrated in [Fei03, Sec. 1.2.23] for the compressible Navier–Stokes equations. Dimensionless forms of the Euler and shallow water systems are derived similarly. The significance of some of the concepts presented below will become clear in Section 2.3. Besides the implications for the continuous models, some of the aspects detailed in this section are also used in numerical methods. We begin with some basic definitions.

Definition 2.1 (Hyperbolicity)
Let $u = u(\boldsymbol{x}, t) \in \mathbb{R}^m$, $\mathbf{f} = (\mathrm{f}_1(u), \dots, \mathrm{f}_d(u)) \in \mathbb{R}^{m \times d}$ with $\mathrm{f}_i \in \mathrm{C}^1(\mathbb{R}^m)^m$, $i \in \{1, \dots, d\}$, $s = s(u, \nabla b) \in \mathbb{R}^m$ where $b = b(\boldsymbol{x}) \in \mathbb{R}$ is a known parameter. The system of equations

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) + s(u, \nabla b) = 0 \tag{2.13}$$

is called *hyperbolic* at $u \in \mathbb{R}^m$ if for all $\boldsymbol{n}$ on the unit sphere S_1^{d-1} the directional Jacobian

$$\mathbf{f}'_{\boldsymbol{n}}(u) := \frac{\partial}{\partial u} \left(\mathbf{f}(u)\, \boldsymbol{n}\right) \ \in \mathbb{R}^{m \times m}$$

of $\mathbf{f}$ is diagonalizable with real eigenvalues, which are also referred to as *wave speeds*. If, in addition, all eigenvalues are distinct, the system is called *strictly hyperbolic*. The solution vector u is composed of *conserved unknowns*. The function $\mathbf{f}$ is called the *inviscid flux* of system (2.13). We refer to $s(u, \nabla b)$ as the *nonconservative term*. ◊

For $m > 1$, the hyperbolicity requirement may impose certain constraints on the components of u or functions thereof. As we will see below, these constraints are important for physical admissibility of u in the case of the Euler and shallow water equations. In the abstract setting of Definition 2.1 we use the following notion.

Definition 2.2 (Largest admissible set)
The *largest admissible set* of (2.13) is defined as

$$\mathcal{A}^{\max} := \{u \in \mathbb{R}^m : (2.13) \text{ is hyperbolic at } u\}\,. \qquad \diamond$$

For the following definition, as well as for the numerical methods considered in this work, it is essential that $\mathcal{A}^{\max}$ be a convex subset of $\mathbb{R}^m$.

Definition 2.3 (Entropy pairs, functions, variables and potentials)
Let the largest admissible set $\mathcal{A}^{\max}$ of (2.13) be convex. Consider a piecewise continuously differentiable function

$$\begin{aligned} \eta : \mathcal{A}^{\max} \times \mathbb{R} &\to \mathbb{R} \\ (u, b) &\mapsto \eta(u, b) \end{aligned}$$

and its derivative w. r. t. the first argument

$$v(u, b) := \frac{\partial}{\partial u} \eta(u, b) \in \mathbb{R}^m.$$

If $\eta(\cdot, b)$ is convex for all $b \in \mathbb{R}$, and there exists a corresponding piecewise continuously differentiable vector field $\boldsymbol{q} = (q_1, \ldots, q_d) = \boldsymbol{q}(u, b) \in \mathbb{R}^{1 \times d}$ satisfying the relationships

$$\frac{\partial}{\partial u} q_k(u, b) = \mathrm{f}'_k(u)^{\mathsf{T}} v(u, b), \qquad k \in \{1, \ldots, d\}, \tag{2.14a}$$

$$\frac{\partial}{\partial b} \boldsymbol{q}(u, b) \cdot \nabla b = v(u, b) \cdot s(u, \nabla b), \tag{2.14b}$$

then $(\eta, \boldsymbol{q})$ is called an entropy pair for (2.13). The functions η, v, $\boldsymbol{q}$ are referred to as entropy function, entropy variable and entropy flux, respectively. The vector field

$$\boldsymbol{\psi}(u, b) = v(u, b)^{\mathsf{T}} \mathbf{f}(u) - \boldsymbol{q}(u, b)$$

is called the entropy potential of the entropy pair $(\eta, \boldsymbol{q})$. ◊

The motivation for introducing the quantities in Definition 2.3 will become apparent in Section 2.3. For now, we observe that if we multiply (2.13) by $v(u,b)^\mathsf{T}$ from the left, assume smoothness and use the chain rule as well as (2.14), we obtain

$$\begin{aligned} 0 &= v(u,b)^\mathsf{T}\left[\frac{\partial u}{\partial t} + \sum_{k=1}^{d} f_k'(u)\frac{\partial u}{\partial x_k} + s(u,\nabla b)\right] \\ &= \frac{\partial \eta(u,b)}{\partial u}\cdot\frac{\partial u}{\partial t} + \sum_{k=1}^{d}\frac{\partial}{\partial u}q_k(u,b)\cdot\frac{\partial u}{\partial x_k} + \frac{\partial}{\partial b}\boldsymbol{q}(u,b)\cdot\nabla b \\ &= \frac{\partial \eta(u,b)}{\partial t} + \nabla\cdot\boldsymbol{q}(u,b). \end{aligned} \tag{2.15}$$

Here we also used the assumption that the parameter b is independent of time (see Definition 2.1). The additional conservation law (2.15) holds for any convex entropy $\eta(u,b)$ and the corresponding flux $\boldsymbol{q}(u,b)$ if u is a smooth solution to problem (2.13).

2.2.1 Scalar conservation laws

Although we are mainly interested in hyperbolic systems, scalar conservation laws are considered in this thesis as well. Depending on the application, such problems may model the evolution of different physical quantities $u = u(\boldsymbol{x},t) \in \mathbb{R}^m$, $m = 1$. The choice of a particular model is determined by the formula for the inviscid flux function $\mathbf{f}(u)$. Scalar equations can serve as a stepping stone for analysis and numerical solution of more advanced problems, which is the main motivation for studying them in this work. In the scalar case, our investigations are focused on the pure conservation law

$$\frac{\partial u}{\partial t} + \nabla\cdot\mathbf{f}(u) = 0 \tag{2.16}$$

to which (2.13) reduces for $s(u,b) = 0$. No matter how $\mathbf{f} \in \mathrm{C}^1(\mathbb{R})^d$ is defined, equation (2.16) is strictly hyperbolic and the largest admissible set is $\mathcal{A}^{\max} = \mathbb{R}$.

If (2.16) is considered in a bounded domain $\Omega \subset \mathbb{R}^d$, boundary conditions need to be imposed at the inlet

$$\Gamma_-(t) := \{\boldsymbol{x} \in \partial\Omega : \mathbf{f}'(u(\boldsymbol{x},t))\cdot\boldsymbol{n}(\boldsymbol{x}) < 0\},$$

where $\boldsymbol{n}$ is the unit outward normal to $\partial\Omega$. No boundary condition is imposed at the outlet $\Gamma_+(t) := \partial\Omega \setminus \Gamma_-(t)$.

Any convex function $\eta \in \mathrm{C}^2(\mathbb{R})$ serves as an entropy for (2.16). In particular, we make use of the *square entropy* $\eta(u) = \frac{u^2}{2}$. The corresponding entropy fluxes and potentials depend on the inviscid flux. They are specified in the subsequent chapters for each scalar conservation law under consideration.

2.2.2 Euler equations of gas dynamics

The Euler equations (2.1) with the gravitational force $\boldsymbol{g} = -g\mathbf{e}_d$ can be written in the form (2.13) with $b = x_d$. The nonconservative source term is given by

$$s(u, \nabla x_d) = [0, g\rho\nabla x_d, \rho g \nabla x_d \cdot \boldsymbol{v}]^\mathsf{T}.$$

For many applications of gas dynamics, however, gravitational effects are negligible, which is why, we restrict the following discussion to the case $s(u, \nabla x_d) = 0$.

There are $m = d + 2$ equations and unknowns in the d-dimensional Euler system. The vector of conserved unknowns u and the inviscid flux written in terms of these unknowns read

$$u = \begin{bmatrix} \rho \\ \rho\boldsymbol{v} \\ \rho E \end{bmatrix}, \qquad \mathbf{f}(u) = \begin{bmatrix} (\rho\boldsymbol{v})^\mathsf{T} \\ \frac{1}{\rho}(\rho\boldsymbol{v}) \otimes (\rho\boldsymbol{v}) + (\gamma - 1)\left(\rho E - \frac{|\rho\boldsymbol{v}|^2}{2\rho}\right)\mathcal{I} \\ \left(\gamma\rho E - \frac{(\gamma-1)|\rho\boldsymbol{v}|^2}{2\rho}\right)\frac{(\rho\boldsymbol{v})^\mathsf{T}}{\rho} \end{bmatrix}.$$

Here the equation of state for the pressure (2.5) has been inserted into $\mathbf{f}$. Usually, a formulation similar to (2.1) in combination with (2.5) is chosen to avoid this cumbersome notation. In this fashion, we use the primitive variables $\rho, \boldsymbol{v}$, and p from here on out. One has to keep in mind that $\boldsymbol{v}$ and p are derived quantities, i. e., functions of the conserved unknowns. For instance, the velocity vector $\boldsymbol{v}$ is defined pointwise as the quotient of momentum $\rho\boldsymbol{v}$ and density ρ.

For $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$, the eigenvalues of the flux Jacobian $\mathbf{f}'_{\boldsymbol{n}}(u)$ are [Dol15, Sec. 8.1]

$$\lambda_1(u, \boldsymbol{n}) = \boldsymbol{v} \cdot \boldsymbol{n} - \sqrt{\gamma p/\rho}, \qquad \lambda_2(u, \boldsymbol{n}) = \ldots = \lambda_{d+1} = \boldsymbol{v} \cdot \boldsymbol{n},$$
$$\lambda_{d+2}(u, \boldsymbol{n}) = \boldsymbol{v} \cdot \boldsymbol{n} + \sqrt{\gamma p/\rho}.$$

Since the adiabatic constant γ is positive (cf. Section 2.1.1), the largest admissible set is

$$\left\{u = (\rho, \rho\boldsymbol{v}^\mathsf{T}, \rho E)^\mathsf{T} \in \mathbb{R}^{d+2} : p/\rho \geq 0\right\} = \{u \in \mathbb{R}^{d+2} : e(u) \geq 0\}, \tag{2.17}$$

where $e(u) = \frac{\rho E}{\rho} - \frac{1}{2}\frac{|\rho\boldsymbol{v}|^2}{\rho^2}$ is the internal energy (cf. Section 2.1.1). For practical purposes however, this set is too large since it allows negative densities as long as the corresponding pressure is also negative. Therefore, we restrict ourselves to the physically relevant case with nonnegative densities and choose the largest admissible set to be the subset

$$\begin{aligned} \mathcal{A}^{\max} &= \left\{u = (\rho, \rho\boldsymbol{v}^\mathsf{T}, \rho E)^\mathsf{T} \in \mathbb{R}^{d+2} : \rho \geq 0,\ e(u) \geq 0\right\} \\ &= \left\{u = (\rho, \rho\boldsymbol{v}^\mathsf{T}, \rho E)^\mathsf{T} \in \mathbb{R}^{d+2} : \rho \geq 0,\ p \geq 0\right\} \end{aligned}$$

of (2.17). In this work, we do not consider examples in which vacuum states occur, i. e., the density remains strictly positive. This assumption could also be incorporated

into the definition of the admissible set. The convexity of $\mathcal{A}^{\max}$ follows from the fact that for $\rho > 0$, the pressure p is a concave function of the vector of conserved unknowns. For $d \in \{1, 2, 3\}$, this property can easily be shown by checking that the eigenvalues of its Hessian are nonpositive.

Remark 2.4
The set (2.17) is not convex. Indeed, the internal energy of some convex combinations of the one-dimensional states $u_1 = (1, 1, 1/2)^\mathsf{T}$ and $u_2 = (-1, 0, -1/2)^\mathsf{T}$ is negative. $\diamond$

In the absence of vacuum states, the *Mach number*

$$\mathrm{M} = \mathrm{M}(u, \boldsymbol{n}) = \frac{|\boldsymbol{v} \cdot \boldsymbol{n}|}{\sqrt{\gamma p / \rho}}$$

is well defined and characterizes the behavior of compressible flows. These are called subsonic for $\mathrm{M} < 1$, sonic for $\mathrm{M} = 1$ and supersonic for $\mathrm{M} > 1$.

For hyperbolic systems of equations, the correct imposition of boundary conditions is more involved than in the scalar case. In general, the boundary type depends on the number of nonpositive eigenvalues of $\mathbf{f}'_{\boldsymbol{n}}(u)$, where $\boldsymbol{n}$ is the outward unit normal to the boundary $\partial\Omega$ of a finite domain Ω. The correct type of boundary is determined by the Mach number and the sign of the normal velocity $\boldsymbol{v} \cdot \boldsymbol{n}$. Following [Fei03, Sec. 3.3.6], we summarize the treatment of different boundary types for the compressible Euler equations in Tab. 2.1. It is possible to prescribe different sets of quantities at subsonic boundaries. We refer to [Ghi03] for an in-depth discussion of this issue.

M	$\boldsymbol{v} \cdot \boldsymbol{n}$	eigenvalues	boundary type	prescribed quantities
0	0	$\lambda_1 < 0 < \lambda_m$	reflecting wall	$\boldsymbol{v} \cdot \boldsymbol{n} = 0$
> 1	< 0	$\lambda_1 < \ldots < \lambda_m < 0$	supersonic inlet	ρ, $\boldsymbol{v}$, p
> 1	> 0	$0 < \lambda_1 < \ldots < \lambda_m$	supersonic outlet	none
≤ 1	< 0	$\lambda_1 < \lambda_{d+1} < 0 \leq \lambda_m$	subsonic inlet	ρ, $\boldsymbol{v}$
≤ 1	> 0	$\lambda_1 \leq 0 < \lambda_2 < \lambda_m$	subsonic outlet	p

Table 2.1: Types of boundary conditions for the compressible Euler equations [Fei03, Sec. 3.3.6].

The quantity $s(u) = \log(p\rho^{-\gamma})$ is called the *specific entropy* of the Euler equations. For this system, Harten [Har83a] suggests to use the entropy pair

$$(\eta(u), \boldsymbol{q}(u)) = \left(\frac{\rho s(u)}{1 - \gamma}, \frac{\rho \boldsymbol{v} s(u)}{1 - \gamma} \right)$$

and proves the convexity of η. Note that the specific entropy s is concave while the mathematical entropy η is convex. The corresponding entropy variable and potential read [Paz19]

$$v(u) = \left[\frac{\gamma - s}{\gamma - 1} - \frac{\rho |\boldsymbol{v}|^2}{2p}, \quad \frac{\rho \boldsymbol{v}^\mathsf{T}}{p}, \quad -\frac{\rho}{p} \right]^\mathsf{T}, \qquad \boldsymbol{\psi}(u) = \rho \boldsymbol{v}.$$

By using $\mathcal{A}^{\max}$ as the largest invariant set, we follow the same approach as pursued, for instance, by Shu and coauthors [Zha11, Che17]. An additional constraint that could be incorporated into the admissible set is a minimum principle that holds for the specific entropy s. This property was proven by Tadmor [Tad86]. Publications that use it as a criterion for designing numerical methods include [Kho94, Gue18a, Paz21]. Convexity of the corresponding invariant set can also be proven, see [Gue18a] and the references therein. It was demonstrated in [Kho94, Gue18a] that strict imposition of the entropy minimum principle reduces the order of accuracy of numerical methods. Therefore, we adopt a different criterion for limiting the entropy of the Euler equations in the following chapters. We remark, however, that the algorithms presented in this thesis could be modified to additionally enforce Tadmor's minimum principle on s.

2.2.3 Shallow water equations

For the shallow water equations, we have $d \in \{1, 2\}$ and $m = d + 1$. The parameter b of the nonconservative term represents the bathymetry. The solution vector, inviscid flux and nonconservative term of the SWE (2.12) read

$$u = \begin{bmatrix} h \\ h\boldsymbol{v} \end{bmatrix}, \qquad \mathbf{f}(u) = \begin{bmatrix} (h\boldsymbol{v})^\intercal \\ \frac{1}{h}(h\boldsymbol{v}) \otimes (h\boldsymbol{v}) + \frac{g}{2}h^2\mathcal{I} \end{bmatrix}, \qquad s(u, \nabla b) = \begin{bmatrix} 0 \\ gh\nabla b \end{bmatrix}.$$

The eigenvalues of the flux Jacobian $\mathbf{f}'_{\boldsymbol{n}}(u)$ are given by [LeV02, Secs. 13.1, 18.7]

$$\begin{aligned} d = 1: &\qquad \lambda_1(u, \boldsymbol{n}) = \boldsymbol{v} \cdot \boldsymbol{n} - \sqrt{gh}, \qquad \lambda_2(u, \boldsymbol{n}) = \boldsymbol{v} \cdot \boldsymbol{n} + \sqrt{gh}, \\ d = 2: &\qquad \lambda_1(u, \boldsymbol{n}) = \boldsymbol{v} \cdot \boldsymbol{n} - \sqrt{gh}, \qquad \lambda_2(u, \boldsymbol{n}) = \boldsymbol{v} \cdot \boldsymbol{n}, \qquad \lambda_3 = \boldsymbol{v} \cdot \boldsymbol{n} + \sqrt{gh} \end{aligned}$$

in the one and two-dimensional setting, respectively. The largest admissible set of the SWE is thus $\mathcal{A}^{\max} = \left\{u = (h, h\boldsymbol{v}^\intercal) \in \mathbb{R}^{d+1} : h \geq 0\right\}$, which is clearly convex.

Assuming the water height h to be nonzero, we define the quantity

$$\mathrm{Fr} = \mathrm{Fr}(u, \boldsymbol{n}) = \frac{|\boldsymbol{v} \cdot \boldsymbol{n}|}{\sqrt{gh}},$$

which is called the *Froude number* for a given flow direction $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$. It represents the ratio of inertia and pressure gradient terms [Vre94, Sec. 2.3] and characterizes the flow behavior similarly to the Mach number M in gas dynamics. For Froude numbers smaller than one, the flow is subcritical, $\mathrm{Fr} = 1$ corresponds to critical flows, and Froude numbers larger than one occur in supercritical flows. Some authors prefer to use the nomenclature sub-, and supersonic instead of -critical, just as for the characterization of compressible flows in terms of M (cf. Section 2.2.2).

The number of boundary conditions to be imposed at a point $\boldsymbol{x} \in \partial\Omega$ on the boundary of $\Omega \subset \mathbb{R}^d$ equals the number of nonpositive eigenvalues of the Jacobian $\mathbf{f}'_{\boldsymbol{n}}(u)$, where

$\boldsymbol{n} = \boldsymbol{n}(\boldsymbol{x})$ is the outward unit normal to $\partial\Omega$ at $\boldsymbol{x}$. In addition to the Froude number, the sign of the normal velocity $\boldsymbol{v} \cdot \boldsymbol{n}$ is needed to uniquely determine the boundary type. Appropriate choices of boundary conditions for the SWE are summarized in Tab. 2.2.

Fr	$\boldsymbol{v} \cdot \boldsymbol{n}$	eigenvalues	boundary type	prescribed quantities
0	0	$\lambda_1 < 0 < \lambda_m$	reflecting wall	$\boldsymbol{v} \cdot \boldsymbol{n} = 0$
> 1	< 0	$\lambda_1 < \ldots < \lambda_m < 0$	supercritical inlet	h, $\boldsymbol{v}$
> 1	> 0	$0 < \lambda_1 < \ldots < \lambda_m$	supercritical outlet	none
≤ 1	< 0	$\lambda_1 \leq \lambda_d < 0 \leq \lambda_m$	subcritical inlet	$h\boldsymbol{v}$
≤ 1	> 0	$\lambda_1 \leq 0 < \lambda_2 \leq \lambda_m$	subcritical outlet	h

Table 2.2: Types of boundary conditions for the one- and two-dimensional shallow water equations.

When it comes to finding an entropy pair, one has to be careful to account for the nonconservative term, which vanishes for problems with constant bathymetry. In general, terms depending on b appear in the entropy function, variable and flux. To avoid making a distinction between this general case and problems with flat topography, we place the origin of the Cartesian coordinate system at a point $\boldsymbol{x} \in \mathbb{R}^d$ where $b(\boldsymbol{x}) = 0$. In the case of a constant bathymetry, the obvious consequence of this convention is $b \equiv 0$. The sum of potential and kinetic energies

$$\eta(u, b) = \frac{1}{2}\left(gh^2 + h|\boldsymbol{v}|^2\right) + ghb$$

serves as an entropy for the SWE in the general case [Win17]. Convexity of η w. r. t. u can easily be shown by checking that the Hessian of η is symmetric positive definite. The entropy variable, flux, and potential

$$v(u, b) = \begin{bmatrix} g(h + b) - \frac{|\boldsymbol{v}|^2}{2} \\ \boldsymbol{v} \end{bmatrix}, \qquad \boldsymbol{q}(u, b) = \left(g(h + b) + \frac{1}{2}|\boldsymbol{v}|^2\right)h\boldsymbol{v},$$
$$\boldsymbol{\psi}(u, b) = \boldsymbol{\psi}(u) = \frac{g}{2}h^2\boldsymbol{v}$$

corresponding to η are obtained from a simple calculation.

2.3 Theory of hyperbolic conservation laws

We close this chapter with a review of some theoretical aspects. Our presentation is focused on concepts that are needed to develop reliable numerical methods for nonlinear hyperbolic problems. The main challenge is to ensure that approximations converge to physically admissible exact solutions. The existing theory of hyperbolic PDEs gives an insight into the qualitative behavior of such solutions that may represent, e. g., shocks

or rarefaction waves. This section includes a basic introduction to the method of characteristics and weak solutions. We also introduce the entropy inequality and discuss its connection to the unique vanishing viscosity solution. Most of the theorems in this section are taken from Dafermos [Daf00] and their proofs can be found therein.

In the remainder of this chapter, we restrict ourselves to the study of Cauchy problems for pure conservation laws. Thus, the influence of boundary conditions and nonconservative terms will not be addressed. In Section 2.3.1, we discuss the scalar problem in detail. Theoretical results for general systems of conservation laws are presented in Section 2.3.2. We close our review in Section 2.3.3 with a brief summary of additional difficulties arising in the study of hyperbolic problems.

Before discussing scalar problems and systems individually in Sections 2.3.1 and 2.3.2, we introduce the general Cauchy problem

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = 0 \qquad \text{in } \mathbb{R}^d \times \mathbb{R}_+, \tag{2.18a}$$

$$u = u_0 \qquad \text{in } \mathbb{R}^d. \tag{2.18b}$$

Here $u = u(\boldsymbol{x}, t) \in \mathbb{R}^m$ is the vector-valued unknown, $\mathbf{f} \in \mathrm{C}^1(\mathbb{R}^m)^{m\times d}$ is the inviscid flux function, and $u_0 \in \mathrm{L}^\infty(\mathbb{R}^d)^m$ is an initial datum.

Definition 2.5 (Classical solutions, Fei03 Def. 2.6)
Let $T \in \mathbb{R}_+ \cup \{\infty\}$. A function $u \in \mathrm{C}^1(\mathbb{R}^d \times (0,T))^m \cap \mathrm{C}(\mathbb{R}^d \times [0,T))^m$ is called a classical solution of (2.18) on $\mathbb{R}^d \times (0,T)$ if

i) $u(\boldsymbol{x}, t) \in \mathcal{A}^{\max}$ for all $(\boldsymbol{x}, t) \in \mathbb{R}^d \times (0,T)$, where $\mathcal{A}^{\max}$ is the largest admissible set of (2.18a), and

ii) u satisfies (2.18a), (2.18b) pointwise in $\mathbb{R}^d \times (0,T)$ and $\mathbb{R}^d$, respectively. ◇

2.3.1 Scalar equations

In this section, we are concerned with the Cauchy problem (2.18) for a single scalar conservation law, i. e., for $m = 1$. The theory presented below is to a large extent taken from [Daf00, Chs. 4–6] and loosely follows [LeV92, Ch. 3].

2.3.1.1 The method of characteristics and its limitations

Classical solutions of (2.18) with $m = 1$ can sometimes be computed with the method of characteristics. A characteristic is a curve $\boldsymbol{x}(t)$ parametrized by the temporal variable and satisfying the ordinary differential equation

$$\boldsymbol{x}'(t) = \mathbf{f}'(u(\boldsymbol{x}(t), t)), \qquad t \in (0, \infty). \tag{2.19}$$

Using the chain rule and (2.19), we find that u remains constant along $\boldsymbol{x}(t)$ because

$$\frac{\mathrm{d}u(\boldsymbol{x}(t),t)}{\mathrm{d}t} = \nabla u(\boldsymbol{x}(t),t)\cdot\boldsymbol{x}'(t) + \frac{\partial u(\boldsymbol{x}(t),t)}{\partial t} = \frac{\partial u(\boldsymbol{x}(t),t)}{\partial t} + \nabla\cdot\mathbf{f}(u(\boldsymbol{x}(t),t)) = 0.$$

It follows that the classical solution $u(\boldsymbol{x},t)$ to (2.18) with $m=1$ can be constructed by backtracking in time along the characteristics. Let $\boldsymbol{x}_0 = \boldsymbol{x}(0) \in \mathbb{R}^d$ be the starting point of a characteristic $\boldsymbol{x}(t)$. Then we have

$$u(\boldsymbol{x}(t),t) = u(\boldsymbol{x}(0),0) = u_0(\boldsymbol{x}_0). \tag{2.20}$$

Inserting (2.20) into (2.19), we observe that $\boldsymbol{x}'(t) = \mathbf{f}'(u_0(\boldsymbol{x}_0))$. Thus, the characteristics are straight lines in the $(d+1)$-dimensional $(\boldsymbol{x},t)$-hyperplane and satisfy the identity

$$\boldsymbol{x}(t) = \mathbf{f}'(u_0(\boldsymbol{x}_0))\,t + \boldsymbol{x}_0. \tag{2.21}$$

Relations (2.20) and (2.21) imply that classical solutions to the scalar conservation law (2.18) with a general flux function $\mathbf{f}(u)$ are implicitly defined by

$$u(\boldsymbol{x},t) = u_0(\boldsymbol{x} - \mathbf{f}'(u(\boldsymbol{x},t))\,t) = u_0(\boldsymbol{x}_0). \tag{2.22}$$

Thus, the value of a smooth solution u at $(\boldsymbol{x},t) \in \mathbb{R}^d \times \mathbb{R}_+$ can be determined by finding a fixed point $\bar{u} = u(\boldsymbol{x},t)$ of the nonlinear equation $\bar{u} = u_0(\boldsymbol{x} - \mathbf{f}'(\bar{u})\,t)$.

Example 2.6 (Linear advection equation)
If the inviscid flux is linear, i. e., $\mathbf{f}(u) = \boldsymbol{v}u$ and $\boldsymbol{v} \in \mathbb{R}^d$ is a constant velocity vector, (2.18a) is called the *linear advection equation*. In this case, the solutions to (2.19) are given by $\boldsymbol{x}(t) = \boldsymbol{v}t + \boldsymbol{x}_0$. Owing to (2.22), the solution to the advection equation reads

$$u(\boldsymbol{x},t) = u_0(\boldsymbol{x} - \boldsymbol{v}t). \tag{2.23}$$

For a smooth initial profile and $d=1$, such solutions are displayed in Fig. 2.2 along with the corresponding characteristics. Note that in the (x,t)-diagram the slope of each characteristic is $1/v$, where v is the constant 1D velocity. ◇

We have established that the characteristics are lines of constant slope in the $(\boldsymbol{x},t)$-hyperplane. For nonlinear fluxes $\mathbf{f}$, these lines may cross. The implication is that u is multivalued and, therefore, not well defined. Following [LeV92, Fig. 3.4], we sketch this scenario in Fig. 2.3 for the *one-dimensional inviscid Burgers equation*

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial (u^2)}{\partial x} = 0. \tag{2.24}$$

We observe here that intersecting characteristics correspond to a situation in which the slope of u becomes infinite. The chain rule, which we used to derive the method

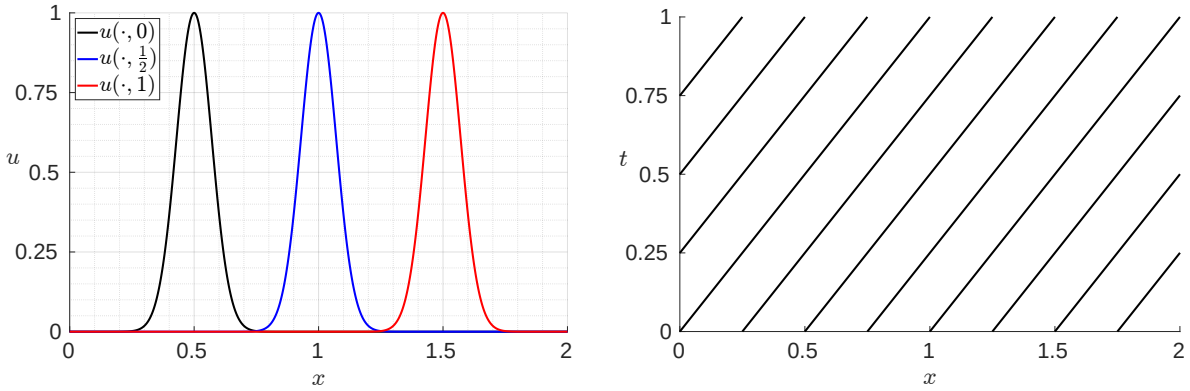


Figure 2.2: Advection equation with smooth initial data. Classical solution at $t \in \{0, \frac{1}{2}, 1\}$ (left) and the corresponding characteristics (right).

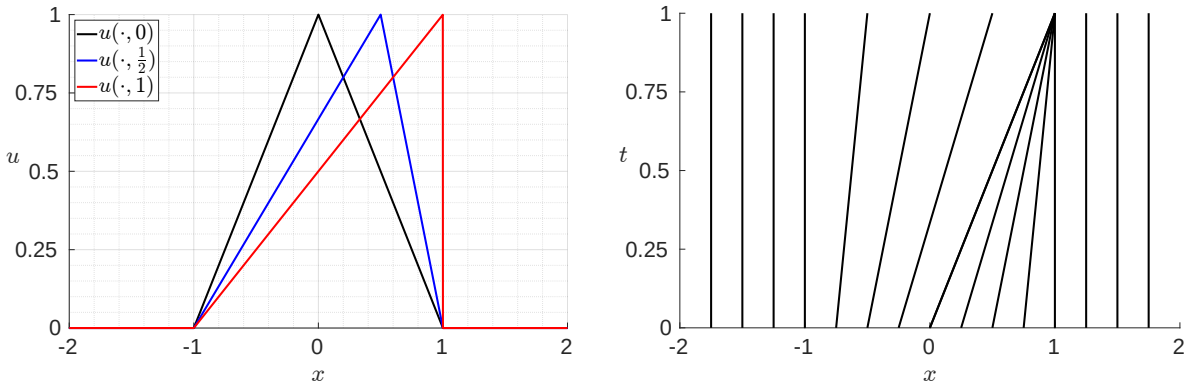


Figure 2.3: Burgers equation (2.24) with initial data $u_0(x) = \max\{0, 1 - |x|\}$. Solution at $t \in \{0, \frac{1}{2}, 1\}$ (left) and the corresponding characteristics (right).

of characteristics, is generally not applicable in this situation. Therefore, the concept of classical solutions and the implicit formula (2.22) fail at discontinuities. To develop a theory that can handle cases such as the one in Fig. 2.3, we need to introduce the concept of *weak solutions* to (2.18). To this end, we first define the functional spaces

$$\mathrm{L}^1_{\mathrm{loc}}(\Omega) := \left\{ w : \Omega \to \mathbb{R} \text{ measurable} : w|_K \in \mathrm{L}^1(\mathrm{int}(K)) \ \forall \text{ compact } K \subset \Omega \right\},$$
$$\mathrm{C}^1_0(\Omega) := \left\{ w \in \mathrm{C}^1(\Omega) : \operatorname{supp} w \subset \Omega \text{ compact} \right\}$$

for a general open set $\Omega \neq \emptyset$. Multiplying the conservation law (2.18a) by a compactly supported test function $w \in \mathrm{C}^1_0(\mathbb{R}^d \times [0,T))$, integrating the weighted residual over the space-time domain $\mathbb{R}^d \times [0,T)$ and performing integration by parts, one arrives at the following weak formulation that defines a generalized solution to (2.18).

Definition 2.7 (Weak solutions to scalar conservation laws)
For $T \in \mathbb{R}_+ \cup \{\infty\}$, a function $u \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d \times (0,T))$ is said to solve the scalar Cauchy

problem (2.18) weakly on $\mathbb{R}^d \times [0,T)$ if the identity

$$\int_0^T \int_{\mathbb{R}^d} \Big[u \, \frac{\partial w}{\partial t} + \mathbf{f}(u) \cdot \nabla w \Big] \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t + \int_{\mathbb{R}^d} u_0 \, w(\cdot, 0) \, \mathrm{d}\boldsymbol{x} = 0 \tag{2.25}$$

holds for all *test functions* $w \in \mathrm{C}_0^1(\mathbb{R}^d \times [0,T))$. Note that $w(\cdot, T) \equiv 0$ by definition. ◇

Remark 2.8
In the literature, there is no unanimous agreement regarding the regularity of w (and to some extent that of u). LeVeque's [LeV92] definition is identical to ours. Dafermos [Daf00] uses locally Lipschitz continuous test functions, while Feistauer et al. [Fei03] require test functions to be in $\mathrm{C}_0^\infty(\mathbb{R}^d \times (0,T))$. In any case, the test function needs to be compactly supported in order for the integrals in (2.25) to be well defined. ◇

Weak solutions of hyperbolic problems are thus allowed to have certain discontinuities. However, a particular relationship between the solution values along a discontinuity can be shown to hold by exploiting conservation properties. To formulate this so-called *Rankine–Hugoniot condition*, we first introduce the notion of piecewise smooth solutions. The details can be found in [Fei03, Sec. 2.3] and are omitted for brevity.

Definition 2.9 (Piecewise smooth functions, Fei03 Def. 2.14)
Let $T \in \mathbb{R}_+ \cup \{\infty\}$. A function u defined on $\mathbb{R}^d \times [0,T)$ is called piecewise smooth if there exists a finite number of smooth hypersurfaces Γ in $\mathbb{R}^d \times [0,T)$ outside which u is of class C^1 and on which u and its first derivatives have well-defined one-sided limits.◇

Remark 2.10
It can be shown that (2.23) is a weak solution of the linear advection equation in more general settings than the scenario considered in Example 2.6. In particular, (2.23) remains valid if u_0 is piecewise smooth. ◇

The following theorem is a consequence of [Fei03, Exe. 2.13, Thm. 2.15].

Theorem 2.11 (Rankine–Hugoniot jump condition, Fei03 Sec. 2.3.2)
Let u be a piecewise smooth weak solution of (2.18). Then u is a classical solution in any subdomain in which it is of class C^1. On each smooth hypersurface Γ such that u is discontinuous in the normal direction $(\boldsymbol{n}_{\boldsymbol{x}}^\mathsf{T}, n_t)$, the Rankine–Hugoniot condition

$$(u_L - u_R)\, n_t + (\mathbf{f}(u_L) - \mathbf{f}(u_R))\, \boldsymbol{n}_{\boldsymbol{x}} = 0 \tag{2.26}$$

holds for the one-sided limits u_L and u_R (see Fig. 2.4 for an illustration inspired by [Fei03, Fig. 2.3.4]). ◇

For $\boldsymbol{n}_{\boldsymbol{x}} = \mathbf{0}$, condition (2.26) implies continuity of u at $(\boldsymbol{x}, t) \in \Gamma$. Elsewhere, we can divide (2.26) by $|\boldsymbol{n}_{\boldsymbol{x}}|$. Thus, upon setting $\boldsymbol{n} := \boldsymbol{n}_{\boldsymbol{x}}/|\boldsymbol{n}_{\boldsymbol{x}}|$ and $\lambda := -n_t/|\boldsymbol{n}_{\boldsymbol{x}}|$, we obtain

$$\lambda(u_L - u_R) = (\mathbf{f}(u_L) - \mathbf{f}(u_R))\, \boldsymbol{n}. \tag{2.27}$$

The vector $\boldsymbol{n}$ and the scalar λ can be interpreted as the direction and corresponding speed of propagation of the discontinuity Γ [Fei03, Sec. 2.3].

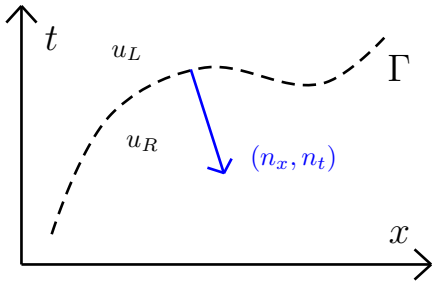


Figure 2.4: Illustration of the Rankine–Hugoniot condition along a smooth hypersurface Γ in the (x,t)-plane. The weak solution is discontinuous and attains values u_L and u_R on different sides of Γ. The vector (n_x, n_t) in the (x,t)-hyperplane is orthogonal to Γ.

Example 2.12 (LeV92 Sec. 3.5)
We consider the following *Riemann problem* for the one-dimensional Burgers equation

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial (u^2)}{\partial x} = 0 \qquad \text{in } \mathbb{R}\times\mathbb{R}_+, \qquad u_0(x) = \begin{cases} u_L & \text{if } x < 0, \\ u_R & \text{if } x > 0, \end{cases} \qquad u_{L,R} \in \mathbb{R}. \tag{2.28}$$

The Rankine–Hugoniot condition (2.27) implies that the speed of propagation in direction $n = 1$ equals

$$\lambda = \frac{1}{2}\frac{u_L^2 - u_R^2}{u_L - u_R} = \frac{u_L + u_R}{2} \tag{2.29}$$

and the speed of propagation in the opposite direction $n = -1$ equals $-\lambda$. It can be shown that for $u_L \geq u_R$, the *shock solution*

$$u(x,t) = \begin{cases} u_L & \text{if } x < \lambda t, \\ u_R & \text{if } x > \lambda t \end{cases} \tag{2.30}$$

with λ given by (2.29) is the unique weak solution of (2.28).

If $u_L < u_R$, (2.30) is also a weak solution. However, there are infinitely many other functions satisfying the weak formulation (2.25) [LeV92, Daf00]. One of them is the so-called *rarefaction wave solution*

$$u(x,t) = \begin{cases} u_L & \text{if } x \leq u_L t, \\ \frac{x}{t} & \text{if } u_L t < x < u_R t, \\ u_R & \text{if } u_R t \leq x. \end{cases} \tag{2.31}$$

♢

The characteristics for the weak solutions (2.30) and (2.31) of (2.28) are depicted in Fig. 2.5. The colored ones correspond to the shock, which satisfies the Rankine-Hugoniot condition (2.27). Note that the other characteristics in Fig. 2.5a go into the

one along which the shock propagates, while in Fig. 2.5b they emanate from it. In the former scenario, the solution remains constant along a characteristic until it bumps into the shock. When characteristics cross, the constant initial values that they carry are irretrievably lost. A parallel can be drawn to astrophysics because such loss of information occurs if mass is sucked into a black hole. Upon crossing the event horizon of a black hole, mass cannot escape its gravitational pull. In view of this analogy, the situation displayed in Fig. 2.5b corresponds to the nonphysical situation in which information escapes from the shock. As we will see below, the solutions corresponding to Figs. 2.5a and 2.5c are admissible (stable), while the ones associated with characteristics that emanate from a shock, as in Fig. 2.5b, are not admissible (unstable).

(a) (2.30) with $u_L = 1$, $u_R = 0$ (b) (2.30) with $u_L = 0$, $u_R = 1$ (c) (2.31) with $u_L = 0$, $u_R = 1$

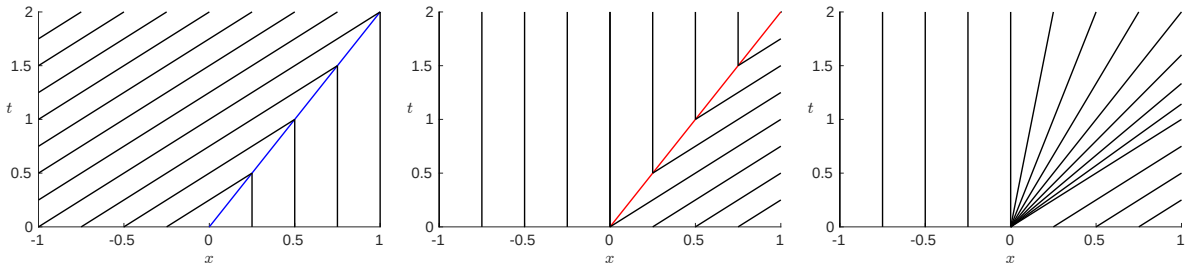


Figure 2.5: Characteristics for the Riemann problem (2.28). Left and center: shock solutions, right: a rarefaction wave solution.

2.3.1.2 The vanishing viscosity approach for scalar problems

The nonuniqueness of solutions observed in Example 2.12 motivates imposition of additional admissibility constraints on weak solutions. As pointed out in [LeV92, Sec. 3.3], the possible existence of discontinuities and characteristics that intersect each other is a consequence of neglecting the diffusive effects, which prevent formation of real discontinuities in nature. In fact, the inviscid model (2.18) should be understood as the $\varepsilon \searrow 0$ limit of the viscous Cauchy problem

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = \varepsilon \Delta u \qquad \text{in } \mathbb{R}^d \times \mathbb{R}_+, \tag{2.32a}$$

$$u = u_0 \qquad \text{in } \mathbb{R}^d. \tag{2.32b}$$

A weak solution u to the inviscid problem (2.18) is admissible if a sequence of solutions $\{u_\varepsilon\}_{\varepsilon>0}$ to (2.32) converges to u as $\varepsilon \searrow 0$. The limit $u = \lim_{\varepsilon \searrow 0} u_\varepsilon$ is referred to as *vanishing viscosity solution* (VVS). Naturally, questions regarding well-posedness of (2.32) arise just as they do for the inviscid problem (2.18). According to [Daf00, Sec. 6.3],

if $\mathbf{f}$ is sufficiently regular and $u_0 \in \mathrm{L}^\infty(\mathbb{R}^d)$, then there is a unique solution to (2.32). If, in particular, $\mathbf{f}$ is Hölder-continuous, then the solution u_ε to (2.32) is in $\mathrm{C}^1(\mathbb{R}^d \times [0,\infty))$.

Let us now revisit the concept of entropy pairs (see Definition 2.3) and their connection to (2.32). In the setting of a single scalar conservation law, we consider a convex entropy $\eta \in \mathrm{C}^2(\mathbb{R})$ and corresponding entropy flux $\boldsymbol{q} = \boldsymbol{q}(u) \in \mathbb{R}^{1\times d}$ satisfying $\boldsymbol{q}'(u) = v(u)\mathbf{f}'(u)$, where $v(u) = \eta'(u)$ is the entropy variable. Adapting the derivation of (2.15) to (2.32a), we obtain the entropy balance equation

$$\frac{\partial \eta(u)}{\partial t} + \nabla \cdot \boldsymbol{q}(u) = \varepsilon v(u)\Delta u = \varepsilon \Delta \eta(u) - \varepsilon \eta''(u)|\nabla u|^2. \tag{2.33}$$

Integrating (2.33) over a finite domain $\Omega \subset \mathbb{R}^d$ with outward unit normal $\boldsymbol{n}$ to $\partial\Omega$ yields

$$\int_\Omega \left[\frac{\partial \eta(u)}{\partial t} + \nabla \cdot \boldsymbol{q}(u)\right] \mathrm{d}\boldsymbol{x} = \int_{\partial\Omega} \varepsilon v(u)\nabla u \cdot \boldsymbol{n}\,\mathrm{d}s - \int_\Omega \varepsilon \eta''(u)|\nabla u|^2\,\mathrm{d}\boldsymbol{x}. \tag{2.34}$$

It can be argued [LeV92, Sec. 3.8], that the first integral in the right hand side of (2.34) vanishes in the limit $\varepsilon \to 0$. In particular, this statement is obviously true for solutions that are piecewise smooth in the sense of Definition 2.9. If u is discontinuous in Ω, the second integral on the right of (2.34) may not converge to zero as $\varepsilon \searrow 0$. However, due to convexity of η, the right hand side of (2.34) will be nonpositive in the limit $\varepsilon \searrow 0$. This somewhat heuristic observation justifies the use of the *entropy inequality*

$$\frac{\partial \eta(u)}{\partial t} + \nabla \cdot \boldsymbol{q}(u) \le 0 \tag{2.35}$$

as a selection criterion for weak solutions. The admissible one should satisfy a weak form of (2.35), which is derived similarly to (2.25) but using nonnegative test functions.

Definition 2.13 (Admissible weak solutions)
For $T \in \mathbb{R}_+ \cup \{\infty\}$, a weak solution u of the scalar Cauchy problem (2.18) on $\mathbb{R}^d \times [0,T)$ is said to be admissible w. r. t. an entropy pair $(\eta, \boldsymbol{q})$ if the inequality

$$\int_0^T \int_{\mathbb{R}^d} \left[\eta(u)\,\frac{\partial \psi}{\partial t} + \boldsymbol{q}(u)\cdot\nabla\psi\right] \mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_{\mathbb{R}^d} \eta(u_0)\,\psi(\cdot,0)\,\mathrm{d}\boldsymbol{x} \ge 0 \tag{2.36}$$

holds for all test functions $\psi \in \mathrm{C}_0^1(\mathbb{R}^d \times [0,T))$ such that $\psi(\boldsymbol{x},t) \ge 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$, $t \in [0,T)$. If (2.36) holds for all entropy pairs $(\eta, \boldsymbol{q})$ of (2.18), then we call u an admissible solution (or entropy solution) of (2.18). ◇

The question of how to verify the admissibility of weak solutions remains. Clearly neither the fact that these are $\varepsilon \searrow 0$ limits of viscous problems, nor the check against all convex functions η is feasible. However, it suffices to verify the entropy inequality for the family of *Kruzhkov entropy pairs*

$$\eta_\kappa(u) = |u-\kappa|, \qquad \boldsymbol{q}_\kappa(u) = \operatorname{sgn}(u-\kappa)(\mathbf{f}(u) - \mathbf{f}(\kappa)), \tag{2.37}$$

where $\kappa \in \mathbb{R}$ is the Kruzhkov parameter. A weak solution can be shown to be admissible if it satisfies entropy inequalities for all Kruzhkov entropy pairs [Daf00, Sec. 6.2]. Note that this family of entropy functions and corresponding entropy fluxes is only piecewise smooth. Nevertheless, the relationship $\boldsymbol{q}'_\kappa(u) = \eta'_\kappa(u)\,\mathbf{f}'(u)$ is satisfied for all $u \in \mathbb{R}$.

It is shown by Panov [Pan94, Thm. 1] that for scalar one-dimensional conservation laws (2.18a) with $\mathbf{f} = \mathrm{f} \in \mathrm{C}^2(\mathbb{R})$ and $\mathrm{f}'' > 0$, the admissible weak solution w. r. t. an *arbitrary* entropy pair is unique. In this setting, it is therefore sufficient to check the admissibility of solutions w. r. t. a single entropy pair.

Example 2.14
Let us briefly revisit the two weak solutions for the 1D Burgers equation found in Example 2.12. The entropy admissibility criterion

$$q(u_L) - q(u_R) - \lambda\,(\eta(u_L) - \eta(u_R)) \geq 0, \tag{2.38}$$

where $\lambda = \frac{u_L+u_R}{2}$, follows from (2.36) similarly to the Rankine–Hugoniot condition. The entropy flux corresponding to the square entropy $\eta(u) = \frac{u^2}{2}$ is $q(u) = \frac{u^3}{3}$. An algebraic manipulation of (2.38) for this entropy pair yields

$$\frac{1}{12}(u_L - u_R)^3 \geq 0.$$

Therefore, the weak solution (2.30) is admissible if and only if $u_L \geq u_R$. In this case, Panov's theorem [Pan94, Thm. 1] implies that the shock solution is admissible w. r. t. all entropy pairs. With similar arguments it can be shown that the rarefaction wave (2.31) is the unique admissible solution to (2.28) if $u_L < u_R$. ◇

Using the concepts of weak solutions and entropy inequalities, one can prove well-posedness of the scalar Cauchy problem (2.18). The following results are taken from [Daf00, Secs. 6.2–6.3] and demonstrate how this property is shown. We focus on the vanishing viscosity approach. However, there are several other methods that yield the same result for the scalar problem (see [Daf00, Ch. 6]).

Theorem 2.15 (Properties of weak admissible solutions, Daf00 Thm. 6.2.2)
Given two initial data functions $u_0, \bar{u}_0$ taking values in an interval $[a, b]$, let u and $\bar{u}$ be corresponding weak solutions that are admissible in the sense of Definition 2.13 for $T \in \mathbb{R}_+ \cup \{\infty\}$. Then there exists $\lambda = \lambda(a, b, \mathbf{f}) \geq 0$ such that for any $t \in [0, T)$, $r > 0$

$$\int_{\mathrm{B}_r(\mathbf{0})} \max\{0, u(\cdot,t) - \bar{u}(\cdot,t)\}\,\mathrm{d}\boldsymbol{x} \leq \int_{\mathrm{B}_{r+\lambda t}(\mathbf{0})} \max\{0, u_0 - \bar{u}_0\}\,\mathrm{d}\boldsymbol{x},$$
$$\|u(\cdot,t) - \bar{u}(\cdot,t)\|_{\mathrm{L}^1(\mathrm{B}_r(\mathbf{0}))} \leq \|u_0 - \bar{u}_0\|_{\mathrm{L}^1(\mathrm{B}_{r+\lambda t}(\mathbf{0}))}.$$

Furthermore, if $u_0 \leq \bar{u}_0$ almost everywhere (a. e.) in $\mathbb{R}^d$, then $u \leq \bar{u}$, a. e. in $\mathbb{R}^d \times [0, T)$. In particular the (essential) range of both u and $\bar{u}$ is contained in $[a, b]$. ◇

Theorem 2.15 has the following immediate implications.

Corollary 2.16 (Uniqueness of weak admissible solutions, Daf00 Cor. 6.2.1)
There is at most one admissible weak solution to the scalar Cauchy problem (2.18). ◇

Corollary 2.17 (Finite speed of propagation, Daf00 Cor. 6.2.2)
The value of an admissible weak solution at $(\boldsymbol{x}, t) \in \mathbb{R}^d \times \mathbb{R}_+$ depends only on $u_0|_{\mathrm{B}_{\lambda t}(\boldsymbol{x})}$, where λ is as in Theorem 2.15. ◇

The following theorem is an important auxiliary result in the process of constructing an admissible weak solution from solutions u_ε to (2.32) with $\varepsilon > 0$.

Theorem 2.18 (Admissibility of the VVS, scalar case, Daf00 Thm. 6.3.1)
For a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ with $\varepsilon_k \searrow 0$ as $k \to \infty$, assume that the corresponding viscous solutions $\{u_{\varepsilon_k}\}$ converge to a function u boundedly almost everywhere on $\mathbb{R}^d \times [0, \infty)$. Then u is an admissible weak solution of (2.18) on $\mathbb{R}^d \times [0, \infty)$. ◇

After establishing the validity of Theorem 2.18, it remains to prove that the sequence of viscous solutions does converge as required. Once this task is accomplished, we obtain the following existence and uniqueness result for scalar conservation laws.

Theorem 2.19 (Well-posedness for scalar equations, Daf00 Thm. 6.2.1)
There exists a unique admissible weak solution $u \in \mathrm{C}([0, \infty); \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d))$ to the Cauchy problem (2.18), provided that $u_0 \in \mathrm{L}^\infty(\mathbb{R}^d)$. ◇

We close our discussion of scalar conservation laws here and proceed to a brief review of the existing theory for hyperbolic systems. The theorems formulated in the next section can also be applied to scalar problems, which represent the case $m = 1$.

2.3.2 Systems of equations

Let us now consider the Cauchy problem (2.18) with $u = u(\boldsymbol{x}, t) \in \mathbb{R}^m$ and $u_0 \in \mathrm{L}^\infty(\mathbb{R}^d)^m$ for $m \geq 1$. We assume that $\mathbf{f} \in \mathrm{C}^1(\mathbb{R}^m)^{m\times d}$ and $\mathbf{f}'(u)$ is diagonalizable with real eigenvalues for all admissible states $u \in \mathcal{A}^{\max}$. No further assumptions regarding $\mathbf{f}(u)$ are made. Thus, the theory to be presented is applicable to general hyperbolic problems in $\mathbb{R}^d$. We first state some results taken from [Daf00].

2.3.2.1 The vanishing viscosity approach for hyperbolic systems

As in the scalar case, smooth solutions may develop infinite slopes. Thus, classical solutions may not exist for all times, as the following theorem suggests.

Theorem 2.20 (Breakdown of classical solutions, Daf00 Thm. 5.1.1)
Assume that the hyperbolic system (2.18) is endowed with an entropy $\eta \in \mathrm{C}^2(\mathbb{R}^m)$ and the Hessian η'' is symmetric uniformly positive definite on compact subsets of $\mathcal{A}^{\max}$. Moreover let $u_0 \in \mathrm{C}^1(\mathbb{R}^d)^m$ assume values in a compact subset of $\mathcal{A}^{\max}$, and let $\nabla u_0 \in \mathrm{H}^k(\mathbb{R}^d)^{m\times d}$ for some $k > d/2$. Then there exists a critical time $0 < T_\infty \leq \infty$, and a unique classical solution $u \in \mathrm{C}^1(\mathbb{R}^d \times [0, T_\infty))$ of (2.18) taking values in $\mathcal{A}^{\max}$. Moreover, $\nabla u \in \mathrm{C}([0, T_\infty); \mathrm{H}^k(\mathbb{R}^d))$. The interval $[0, T_\infty)$ is maximal, in the sense that

$$\limsup_{t \nearrow T_\infty} \|\nabla u(\cdot, t)\|_{\mathrm{L}^\infty(\mathbb{R}^d)} = \infty,$$

whenever $T_\infty < \infty$ and/or the range of $u(\cdot, t)$ escapes from every compact subset of $\mathcal{A}^{\max}$ as $t \nearrow T_\infty$. ◇

Thus, we again seek weak solutions $u \in \mathrm{L}^1_{\mathrm{loc}}(\mathbb{R}^d \times (0, T))^m$ of (2.18) that satisfy

$$\int_0^T \int_{\mathbb{R}^d} \Big[u \cdot \frac{\partial w}{\partial t} + \mathbf{f}(u) : \nabla w \Big] \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_{\mathbb{R}^d} u_0 \cdot w(\cdot, 0)\,\mathrm{d}\boldsymbol{x} = 0 \tag{2.39}$$

for all $w \in \mathrm{C}^1_0(\mathbb{R}^d \times [0, T))^m$. This weak formulation is obtained similarly to the scalar case. In (2.39), the colon stands for a scalar product of matrix-valued functions. It is defined by $A : B = \mathrm{tr}(A^{\mathsf{T}} B)$, where $\mathrm{tr}(\cdot)$ is the trace of a square matrix, or equivalently

$$A : B = \sum_{i=1}^{k} \sum_{j=1}^{n} a_{ij}\, b_{ij}, \qquad A = (a_{ij})_{\substack{i=1,\dots,k \\ j=1,\dots,n}}, \qquad B = (b_{ij})_{\substack{i=1,\dots,k \\ j=1,\dots,n}}, \qquad k, n \in \mathbb{N}.$$

Note that Theorem 2.11 is not restricted to scalar conservation laws. Thus, the Rankine–Hugoniot condition (2.27) applies also to piecewise smooth weak solutions of hyperbolic systems. For a given entropy pair of such a system, we can derive the scalar entropy inequality (2.35) just as in the case of a single conservation law. Thus, admissibility of weak solutions to systems is defined as in the scalar case, see Definition 2.13.

Let us now briefly discuss the vanishing viscosity approach for hyperbolic systems, as presented in [Daf00, Sec. 4.4]. For $i, j \in \{1, \dots, d\}$, let $B^{ij} = B^{ij}(u) = (b^{ij}_{kl}(u))_{k,l=1}^m$ be matrix valued functions of $u \in \mathbb{R}^m$. We consider the viscous problem

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = \varepsilon \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\Big(B^{ij}(u) \frac{\partial u}{\partial x_j} \Big) \tag{2.40}$$

for which one can show the following compatibility result.

Theorem 2.21 (Admissibility of the VVS, system case, Daf00 Thm. 4.4.1)
Let the hyperbolic system (2.18) be endowed with a designated entropy pair $(\eta, \boldsymbol{q})$. Moreover, let u_0 assume values in a compact subset of $\mathcal{A}^{\max}$, and $u_0 - \bar{u}$ be in $\mathrm{L}^2(\mathbb{R}^d)$, where $\bar{u}$ is a state in which η attains its minimum on $\mathcal{A}^{\max}$. Suppose further that for any

$\varepsilon > 0$, problem (2.40) admits a solution u_ε on $[0, T)$ that is locally Lipschitz on $\mathbb{R}^d \times (0, T)$, tends to $\bar{u}$ as $|\boldsymbol{x}| \to \infty$, and satisfies $u(\cdot, 0) = u_0$ in the strong sense. Moreover, let the u_ε assume values in a compact subset $\mathcal{B}$ of $\mathcal{A}^{\max}$, where $\mathcal{B}$ is independent of ε. Furthermore, assume that

$$\sum_{j=1}^{d} B^{ij}(u_\varepsilon) \frac{\partial u_\varepsilon}{\partial x_j} \in \mathrm{L}^2(\mathbb{R}^d \times (0, T))^m, \qquad i \in \{1, \dots, d\}$$

and

$$\sum_{i,j=1}^{d} \xi_i^{\mathsf{T}} \eta''(u) B^{ij}(u) \xi_j \geq \sum_{i=1}^{d} \Big| \sum_{j=1}^{d} B^{ij}(u) \xi_j \Big|^2 \tag{2.41}$$

holds for any $u \in \mathcal{A}^{\max}$ and all $\xi_i \in \mathbb{R}^m$, $i \in \{1, \dots, d\}$. If $u_{\varepsilon_k} \to u$ a. e. in $\mathbb{R}^d \times (0, T)$ for a sequence $\{\varepsilon_k\}_{k \in \mathbb{N}}$ with $\varepsilon_k \searrow 0$ as $k \to \infty$, then u is a weak solution of (2.18) on $[0, T)$ and satisfies the entropy admissibility condition w. r. t. $(\eta, \boldsymbol{q})$. ◇

For scalar conservation laws equipped with the square entropy $\eta(u) = \frac{u^2}{2}$, Theorem 2.21 implies the statement of Theorem 2.18. Indeed, the functions B^{ij} are scalar valued and by setting $B^{ij}(u) := \delta_{ij}$, we obtain the scalar viscous PDE (2.32a). Therefore, (2.41) holds by construction if $\eta''(u) \geq 1$. For the square entropy $\eta(u) = \frac{u^2}{2}$, we have $\bar{u} = 0$, and the regularity requirement $u_0 \in \mathrm{L}^2(\mathbb{R}^d)$ thus implies that the initial data has to decay as $|\boldsymbol{x}| \to \infty$. Since the regularity assumptions on the sequence of viscous scalar solutions hold as discussed in Section 2.3.1 and all other requirements of Theorem 2.21 are met, the inviscid limit u is an admissible weak solution.

Remark 2.22
A result similar to Theorem 2.21 is formulated in [Fei03, Thm. 2.22]. Its statement and requirements are considerably less complicated than those of Theorem 2.21. One reason for this discrepancy may be that the definitions of weak solutions in [Daf00] and [Fei03] differ slightly. In both versions of the above theorem, a main assumption is that the sequence of viscous solutions converges to a candidate solution of the inviscid problem. If this is not the case, neither admissibility criterion is applicable. ◇

2.3.2.2 Linear hyperbolic systems

Having discussed the vanishing viscosity approach for systems, we close this section with reviewing a way to derive a closed-form solution of linear hyperbolic systems as in [LeV92, Ch. 2]. Let us first consider the one-dimensional Cauchy problem

$$\frac{\partial u}{\partial t} + A \frac{\partial u}{\partial x} = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}_+, \tag{2.42a}$$

$$u = u_0 \qquad \text{in } \mathbb{R}, \tag{2.42b}$$

where $u = u(\boldsymbol{x}, t) \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times m}$. Hyperbolicity implies that the eigenvalues $\lambda_1 \leq \ldots \leq \lambda_m$ of A are real and the corresponding eigenvalues $r_1, \ldots, r_m$ are linearly independent. Thus, there exists a spectral decomposition

$$A = R\Lambda R^{-1}, \qquad R, \Lambda \in \mathbb{R}^{m\times m},$$

such that $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ and the columns of the invertible matrix $R = [r_1, \ldots, r_m]$ form an eigenvector basis of $\mathbb{R}^m$. Introducing the *characteristic variables* $w = R^{-1}u$ and multiplying (2.42a) by R^{-1} from the left yields

$$R^{-1}\frac{\partial u}{\partial t} + \Lambda R^{-1}\frac{\partial u}{\partial x} = 0,$$

or equivalently

$$\frac{\partial w}{\partial t} + \Lambda\frac{\partial w}{\partial x} = 0. \tag{2.43}$$

This linear system for the characteristic variables consists of m linear advection equations that can be solved independently from each other. The initial data for w are obtained via the transformation $w_0(x) = R^{-1}u_0(x)$. Owing to (2.23), the solution for each component w^i of w reads

$$w^i(x,t) = w_0^i(x - \lambda_i t), \qquad i \in \{1, \ldots, m\}.$$

Transforming back to conserved unknowns, we obtain the expression

$$u(x,t) = Rw(x,t) = \sum_{i=1}^{m} w_0^i(x - \lambda_i t)\, r_i$$

for the analytical solution to (2.42). Owing to the linearity of (2.42a), no admissibility conditions need to be checked, just as for the advection equation.

Using the above transformation to characteristic variables one can, for instance, derive analytical expressions for the solution of the one-dimensional wave equation, see [LeV92, Sec. 6.3]. For general linear systems, this approach is limited to the 1D case. Indeed, the multidimensional ($d > 1$) version

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{d} A_i\,\frac{\partial u}{\partial x_i} = 0$$

of (2.42a) decouples similarly to (2.43) only in the case in which the matrices A_i, $i \in \{1, \ldots, d\}$ can be *simultaneously diagonalized*, i. e., they have a common eigenvector basis. In most practical problems, however, the equations are intricately coupled and each directional Jacobian matrix has its own set of eigenvectors [LeV02, Sec. 18.5]. For instance, the multidimensional wave equation cannot be simultaneously diagonalized and solved in the above manner. This complication is a representative example of issues that arise when dealing with hyperbolic systems of equations in multidimensions.

2.3.3 Further approaches and limitations of the theory

The theory of hyperbolic systems extends far beyond what is covered in this section. Further topics on the subject include the study of shocks and rarefaction waves, other entropy conditions, as well as generalized characteristics (see [LeV92, Daf00]). Many theoretical results for systems are limited to the one-dimensional setting, while others impose strict constraints on the data (e. g., assumptions on the strength of shocks). Analytical solutions to some 1D Riemann problems for hyperbolic systems are known in closed form or as expressions that require a simple numerical calculation. For the Euler equations, the derivation of such exact solutions can be found in [Fei03, Sec. 3.1.6].

The following difficulties, among others, arise in theoretical investigations of hyperbolic problems. If, for instance, the Cauchy problem is replaced by an initial-boundary value problem, more advanced concepts such as *boundary entropy flux pairs* or *semi Kruzhkov entropy pairs* are required. Using the latter, Martin [Mar07] shows existence, uniqueness, and the validity of a maximum principle for weak admissible solutions of a scalar nonautonomous balance law. Another complication for the theory is due to possible violations of the Rankine–Hugoniot condition [LeV92, Sec. 8.4]. Recall that the assumption in Theorem 2.11 is that solutions along hypersurfaces of discontinuity have well-defined one-sided limits. This property may be violated if shock collisions occur, as they do in the one-dimensional Woodward–Colella blast wave [Woo84] example. Finally, it is remarked in [LeV92, Sec. 7.2.1] that there exist unsolvable Riemann problems. For example, there might be no solution if the magnitude of the jump between the two states in the initial condition is too large. Of course, the above issues need to be taken into account in theoretical and numerical studies of a given problem.

In conclusion, the basic theoretical considerations presented in this section are fundamental to understanding hyperbolic conservation laws. Theory alone, however, is not a practical approach for solving real-life problems. If the domain of interest has a complex geometrical shape and general initial/boundary conditions are imposed, the use of numerical methods is usually the only way to solve the Euler equations or the SWE (approximately). The results of theoretical studies for one dimensional Riemann problems can be valuable, however, when it comes to developing and testing numerical algorithms. We use existing theory in this way in subsequent chapters.

For practical purposes, it is imperative that approximations converge to the vanishing viscosity solution (as fast as possible) and preserve its qualitative properties. The remainder of this thesis is devoted to the design, analysis, and evaluation of property-preserving high-resolution schemes. Approximate solutions to scalar conservation laws are required to satisfy the maximum principle stated in Theorem 2.15. For hyperbolic systems, we use methods based on generalized admissibility criteria (see Definition 2.2). To avoid convergence to nonphysical weak solutions (such as the one sketched in Fig. 2.5b), we design algorithms that enforce entropy inequalities. Analytical and numerical studies confirm the claimed properties of the resulting approximations.

Chapter 3

Property-preserving methods for conservation laws

Having laid out the theoretical groundwork, we now turn towards the design of computational methods for solving hyperbolic problems numerically. This chapter is organized as follows. In Section 3.1, we introduce the standard finite element method (FEM) for discretizing the governing equations in space. The temporal discretization techniques used in this thesis are discussed in Section 3.2. What follows in Section 3.3 is a review of *algebraic flux correction* (AFC) schemes for enforcing discrete maximum principles and other constraints. In particular, we present Kuzmin's [Kuz20a] *monolithic convex limiting* (MCL) technique and the semi-discrete entropy correction procedures introduced by Kuzmin and Quezada de Luna [Kuz20c]. The latter algebraic fixes were first extended to hyperbolic systems by the author of this thesis in [Kuz22a]. The algorithms presented in this chapter are designed for continuous (multi-)linear finite element approximations. Numerical studies for nonlinear scalar conservation laws and the compressible Euler equations are performed in Section 3.4. In subsequent chapters, we extend the AFC framework to the shallow water equations and very high order discontinuous Galerkin discretizations of general hyperbolic problems. We also present theoretical results for AFC space discretizations of the linear advection equation in Chapter 5.

3.1 Finite element discretization

A straightforward way to discretize a PDE is to approximate all partial derivatives by difference quotients, i. e., linear combinations of function values at discrete locations in space and time. This approach is adopted in the finite difference method, which has some well-known drawbacks. In particular, it is tedious to implement for domains with complicated geometries or in situations, where local refinement may be necessary. Moreover, the regularity of solutions required to prove a priori error estimates is higher for finite difference methods than for alternative strategies, such as finite element or finite volume schemes. This drawback can, in some cases, be resolved by proving the equivalence of certain finite difference approaches to methods that require less regularity. In the past, structured grid algorithms used to be the main tool for solving multidimensional hyperbolic problems. Using this framework one can use operator splitting techniques and apply 1D property-preserving schemes [Bor73, Har83b, Swe84, Sch85, Tad87] in each spatial direction. Modern generalizations to unstructured grids are typically

based on finite volume methods or their discontinuous Galerkin (DG) counterparts [Aud04, Dio13, Che17, Moe17]. Aside from Chapter 6, we employ continuous (linear or multilinear) Lagrange finite elements in this thesis. The remainder of this section is devoted to introducing the standard Galerkin discretization in this context. To begin, we summarize the required properties of computational meshes (also referred to as *grids* or *triangulations*).

Definition 3.1 (Meshes, Ern04 Def. 1.49)
A mesh over a domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ is a collection $\mathcal{K} = \{K^1, \dots, K^E\}$ of $E \in \mathbb{N}$ compact, connected Lipschitz sets with nonempty interiors such that

$$\overline{\Omega} = \bigcup_{e=1}^{E} K^e \qquad \text{and} \qquad \operatorname{int}(K^i) \cap \operatorname{int}(K^j) = \emptyset, \qquad i \neq j, \qquad i, j \in \{1, \dots, E\}.$$

The elements in $\mathcal{K}$ are also referred to as *(mesh) cells* or *(mesh) elements*. ◊

All meshes used in our numerical experiments consist of d-dimensional simplices (intervals for $d = 1$, triangles for $d = 2$, tetrahedra for $d = 3$) or *box elements* (intervals for $d = 1$, quadrilaterals for $d = 2$, hexahedra for $d = 3$). Furthermore, we consider only *affine* and *geometrically conforming* meshes (see [Ern04, Def. 1.53, Def. 1.55]). Conceptually, it is also possible to apply the methods discussed in this thesis to prismatic elements in 3D. Nonaffine meshes consisting of cells with curved boundaries can be employed in combination with quadrature rules of appropriate order. Extensions to geometrically nonconforming meshes are expected to work similarly to [Bit13], where flux correction tools are combined with *hp*-adaptivity and arbitrary-level hanging nodes.

The boundaries of simplex and box elements that we use in this thesis consist of $(d-1)$-dimensional *faces*, which we define as follows for affine meshes.

Definition 3.2 (Element and mesh faces)
Let $\mathcal{K}$ be an affine mesh over a domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. The faces of an element $K \in \mathcal{K}$ represent the largest $(d-1)$-dimensional open sets $\Gamma_i^K \subset \partial K$ such that the outward unit normal $\boldsymbol{n}$ to $\partial K = \bigcup_{i=1}^{n_\mathrm{F}} \overline{\Gamma_i^K}$ is constant on Γ_i^K for $i \in \{1, \dots, n_\mathrm{F}\}$, where $n_\mathrm{F} = d + 1$ for simplices and $n_\mathrm{F} = 2d$ for box elements. The set of mesh faces

$$\mathcal{F} = \mathcal{F}(\mathcal{K}) := \{\Gamma \subset \overline{\Omega} : \exists K \in \mathcal{K} \text{ such that } \Gamma \text{ is a face of } K\}$$

can be split into the subsets of interior and boundary faces defined by

$$\mathcal{F}_\Omega = \mathcal{F}_\Omega(\mathcal{K}) := \{\Gamma \subset \Omega : \exists K \in \mathcal{K} \text{ such that } \Gamma \text{ is a face of } K\},$$
$$\mathcal{F}_{\partial\Omega} = \mathcal{F}_{\partial\Omega}(\mathcal{K}) := \{\Gamma \subset \partial\Omega : \exists K \in \mathcal{K} \text{ such that } \Gamma \text{ is a face of } K\},$$

respectively. Clearly, we have $\mathcal{F} = \mathcal{F}_\Omega \cup \mathcal{F}_{\partial\Omega}$ and $\mathcal{F}_\Omega \cap \mathcal{F}_{\partial\Omega} = \emptyset$. ◊

Let $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$ be a domain with Lipschitz boundary $\partial\Omega$. Following [Ern04, Def. 1.47], we assume that Ω is a finite interval if $d = 1$, a polygon if $d = 2$, and a polyhedron if $d = 3$. Let $\mathcal{K}_h = \{K^1, \dots, K^E\}$ be a mesh over Ω that is affine, geometrically conforming, and consists of $E = E(h)$ elements. We use the notation

$$h_K := \operatorname{diam}(K), \qquad K \in \mathcal{K}_h \qquad \text{and} \qquad h := \max_{K \in \mathcal{K}_h} h_K$$

for the local and global *mesh sizes*. For d-dimensional simplices K, we denote by $\mathbb{P}_p(K)$ the space of polynomials of degree $p \in \mathbb{N}_0$ in d variables. For box elements K, the space $\mathbb{Q}_p(K)$ consists of products of d one-dimensional polynomials of degree $p \in \mathbb{N}_0$ in the spatial variables. To avoid a distinction of these two cases, we define

$$\mathbb{V}_p(K) := \begin{cases} \mathbb{P}_p(K) & \text{if } K \text{ is a simplex,} \\ \mathbb{Q}_p(K) & \text{if } K \text{ is a box element.} \end{cases}$$

The scalar and vector-valued spaces of continuous linear finite elements are defined as

$$\mathrm{V}_h = \mathrm{V}_{h,1}(\mathcal{K}_h) := \left\{ w_h \in C(\overline{\Omega}) : w_h|_K \in \mathbb{V}_1(K) \ \forall K \in \mathcal{K}_h \right\}, \qquad \mathrm{V}_h^m = (\mathrm{V}_h)^m, \ m \in \mathbb{N}.$$

The *Lagrange polynomials*, which are uniquely defined by the properties

$$\varphi_i|_K \in \mathbb{V}_1(K) \qquad \forall K \in \mathcal{K}_h, \qquad \varphi_i(\boldsymbol{x}_j) = \delta_{ij} \qquad \forall i,j \in \{1,\dots,N\},$$

form a basis of V_h. The dimension $N = N(h)$ corresponds to the number of vertices $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ of $\mathcal{K}_h$. The Lagrange basis functions form a *partition of unity*, i. e.,

$$\sum_{j=1}^{N} \varphi_j \equiv 1.$$

Let us now apply the finite element method to the hyperbolic PDE (system)

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = 0 \qquad \text{in } \Omega \times \mathbb{R}_+. \tag{3.1}$$

Here $u = u(\boldsymbol{x}, t) \in \mathbb{R}^m$ is the vector of conserved unknowns and $\mathbf{f} = \mathbf{f}(u) \in \mathbb{R}^{m \times d}$ is the inviscid flux function. Additionally, we require an initial condition $u_0 = u_0(\boldsymbol{x}) \in \mathbb{R}^m$ for u and choose the external Riemann data $\hat{u} = \hat{u}(\boldsymbol{x}, t) \in \mathbb{R}^m$ in accordance with the general rules outlined in Section 2.2 for each model problem under consideration. We assume that $u(\boldsymbol{x}, t) \in \mathcal{A}^{\max}$, where $\mathcal{A}^{\max} \subseteq \mathbb{R}^m$ is the largest admissible set for (3.1) (cf. Definition 2.2). Before discussing the spatial discretization of (3.1), we introduce the following concept.

Definition 3.3 (Numerical fluxes)
Let S_1^{d-1} denote the unit sphere. A function $\mathrm{f} : \mathcal{A}^{\max} \times \mathcal{A}^{\max} \times \mathrm{S}_1^{d-1} \to \mathbb{R}^m$ is called a numerical flux (consistent with $\mathbf{f}$) if it satisfies the following assumptions

- $\mathrm{f}(u, u, \boldsymbol{n}) = \mathbf{f}(u)\,\boldsymbol{n}$ for all $u \in \mathcal{A}^{\max}$, $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$ (consistency),
- $\mathrm{f}(u_L, u_R, \boldsymbol{n}) = -\mathrm{f}(u_R, u_L, -\boldsymbol{n})$ for all $u_L, u_R \in \mathcal{A}^{\max}$, $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$ (conservation), and
- $\mathrm{f}(\cdot, u, \boldsymbol{n})$ and $\mathrm{f}(u, \cdot, \boldsymbol{n})$ are Lipschitz continuous functions on $\mathcal{A}^{\max}$ for all $u \in \mathcal{A}^{\max}$, $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$ (continuity).

Instead of $\mathrm{f}(u_L, u_R, \boldsymbol{n})$, we write $\mathrm{f}_{\boldsymbol{n}}(u_L, u_R)$ for $u_L, u_R \in \mathcal{A}^{\max}$, $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$ in most instances. ♢

Example 3.4 (Local Lax–Friedrichs flux)
Let $\varrho(A)$ denote the spectral radius of a matrix $A \in \mathbb{R}^{m\times m}$, i. e., its largest absolute eigenvalue. Then the function $\mathrm{f} : \mathcal{A}^{\max} \times \mathcal{A}^{\max} \times \mathrm{S}_1^{d-1} \to \mathbb{R}^m$ defined by

$$\mathrm{f}(u_L, u_R, \boldsymbol{n}) := \frac{1}{2}\left[(\mathbf{f}(u_L) + \mathbf{f}(u_R))\,\boldsymbol{n} + \lambda_{\boldsymbol{n}}(u_L, u_R)(u_L - u_R)\right], \tag{3.2a}$$

$$\lambda_{\boldsymbol{n}}(u_L, u_R) := \sup_{\omega\in[0,1]} \varrho(\mathbf{f}'_{\boldsymbol{n}}(\omega u_L + (1-\omega)u_R)), \qquad \mathbf{f}'_{\boldsymbol{n}}(u) := \frac{\partial}{\partial u}(\mathbf{f}(u)\boldsymbol{n}) \tag{3.2b}$$

is called the local Lax–Friedrichs flux and can be shown to satisfy the requirements of Definition 3.3. The approximation

$$\lambda_{\boldsymbol{n}}(u_L, u_R) \approx \max\left\{\varrho(\mathbf{f}'_{\boldsymbol{n}}(u_L)), \varrho(\mathbf{f}'_{\boldsymbol{n}}(u_R))\right\} \tag{3.3}$$

to the maximum wave speed (3.2b) is often employed in practice. In the case of a scalar conservation law with an *isotropic* flux, that is, for $\mathbf{f}(u) = \mathrm{f}(u)\boldsymbol{v}$, where $\boldsymbol{v} \in \mathbb{R}^d$ is a fixed vector, (3.3) is exact if $\mathrm{f} \in \mathrm{C}^2(\mathbb{R})$ and the sign of f'' does not change on the set $\{\omega u_L + (1-\omega)u_R : \omega \in [0,1]\}$. An exact formula for the wave speed (3.2b) of arbitrary scalar conservation laws can be found in [Gue17, Lem. 3.8]. Alternatively, one can use an upper bound on $\lambda_{\boldsymbol{n}}(u_L, u_R)$, which makes the numerical flux more diffusive. In general, (3.2b) cannot be evaluated exactly for hyperbolic systems. Upper bounds on the wave speeds for the Euler and shallow water equations can be found in in [Gue16a, Sec. 4] and [Aze17, Prop. 3.7], [Gue18b, Sec. 4]), respectively. ♢

We are now in a position to discuss the semi-discrete finite element spatial discretization for (3.1). First, we approximate the exact solution u by $u_h \in \mathrm{V}_h^m$ such that

$$u_h(\boldsymbol{x}, t) = \sum_{j=1}^{N} u_j(t)\varphi_j(\boldsymbol{x}), \qquad u_j(t) := u_h(\boldsymbol{x}_j, t) \in \mathbb{R}^m, \qquad j \in \{1, \dots, N\}. \tag{3.4}$$

Inserting (3.4) into (3.1), multiplying by a test function $w_h \in \mathrm{V}_h^m$, and using integration by parts, we obtain the *weak formulation* (also called *variational formulation*)

$$\int_\Omega \left[w_h \cdot \frac{\partial u_h}{\partial t} - \nabla w_h : \mathbf{f}(u_h)\right] \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} w_h \cdot \mathrm{f}_{\boldsymbol{n}}(u_h, \hat{u})\,\mathrm{d}s = 0 \tag{3.5}$$

for all $w_h \in \mathrm{V}_h^m$, where $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$ denotes the outward unit normal to $\partial\Omega$ and the dependence on time t has been suppressed. In the boundary integral, we replaced the term $\mathbf{f}(u_h)\boldsymbol{n}$ with a numerical flux, which incorporates the external Riemann data $\hat{u}$ into the formulation. For our purposes, it is convenient to write (3.5) in the *strong form*

$$\int_\Omega w_h \cdot \left[\frac{\partial u_h}{\partial t} + \nabla \cdot \mathbf{f}(u_h)\right] \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} w_h \cdot \left[\mathrm{f}_{\boldsymbol{n}}(u_h, \hat{u}) - \mathbf{f}(u_h)\boldsymbol{n}\right] \mathrm{d}s = 0, \tag{3.6}$$

which is obtained by reversing integration by parts. Again, (3.6) must hold for all $w_h \in \mathrm{V}_h^m$ or, equivalently, for a finite set of basis functions spanning the space V_h^m. In the *standard Galerkin* method, these basis functions are the same as those chosen for the representation of $u_h \in \mathrm{V}_h^m$. Let $\mathrm{e}_k = (\delta_{kl})_{l=1}^m$ denote the kth Cartesian unit vector in $\mathbb{R}^m$. Then $\varphi_i \mathrm{e}_k$, $i \in \{1, \dots, N\}$, $k \in \{1, \dots, m\}$ is a basis function for V_h^m. Substituting $\varphi_i \mathrm{e}_1, \dots, \varphi_i \mathrm{e}_m$ for w_h in (3.6), we obtain the continuous Galerkin approximation

$$\sum_{j=1}^N m_{ij} \frac{\mathrm{d}u_j}{\mathrm{d}t} = -\int_\Omega \varphi_i \nabla \cdot \mathbf{f}(u_h)\, \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \varphi_i \left[\mathbf{f}(u_h)\boldsymbol{n} - \mathrm{f}_{\boldsymbol{n}}(u_h, \hat{u})\right] \mathrm{d}s, \tag{3.7}$$

where m_{ij} are scalar-valued entries of the *consistent mass matrix*

$$M = (m_{ij})_{i,j=1}^N, \qquad m_{ij} = \int_\Omega \varphi_i\, \varphi_j \, \mathrm{d}\boldsymbol{x}, \qquad i, j \in \{1, \dots, N\}. \tag{3.8}$$

Note that (3.7) is just a shorthand notation for a set of m equations that are associated with the m components of the vector-valued *degree of freedom* $u_i \in \mathbb{R}^m$.

The mass matrix (3.8) is invertible and independent of the numerical solution u_h. Therefore, (3.7) is equivalent to a system of ordinary differential equations. An initial condition $u_h(\cdot, 0) = \mathrm{I}_h u_0$ for the semi-discrete problem can be constructed using a suitable projection or interpolation operator $\mathrm{I}_h : \mathrm{L}^\infty(\Omega)^m \to \mathrm{V}_h^m$. Then numerical time integration can be performed using the methods presented in Section 3.2.1.

3.2 Temporal discretization

Two types of temporal discretizations are employed in this thesis. In Section 3.2.1, we discuss the class of *strong stability preserving Runge–Kutta schemes*, which update u_h step by step. The other strategy, which we describe in Section 3.2.2, is to treat the time in the same way as the spatial variables. In such space-time Galerkin approaches, the time derivative acts as a linear advection term for the temporal direction and $u_h(\cdot, 0)$ becomes an inflow boundary data for the $(d+1)$-dimensional domain.

3.2.1 Strong stability preserving Runge–Kutta methods

A standard approach to discretization of time-dependent PDEs is the *method of lines* in which the space derivatives are discretized first, as in Section 3.1. In general, the result of the spatial semi-discretization is an initial value problem of the form

$$y'(t) = z(y(t), t) \qquad t \in (0, T), \tag{3.9a}$$

$$y(0) = y_0, \tag{3.9b}$$

where $y : [0, T] \to \mathbb{R}^M$ defines the vector $y(t)$ of discrete unknowns, $z : \mathbb{R}^M \times [0, T] \to \mathbb{R}^M$ is a Lipschitz continuous function, and $y_0 \in \mathbb{R}^M$ is a given initial datum.

In numerical methods for (3.9), the continuous function $y(t)$ is usually replaced by a sequence of approximations $y^0 = y_0$, $y^1 \approx y(t_1), \ldots, y^n \approx y(t^n)$ at discrete time instants $0 = t^0 < t^1 < \cdots < t^n = T$. The accuracy of such approximations depends, among other things, on the (local) time step $\Delta t = \Delta t(k) = t^{k+1} - t^k$, $k \in \{0, \ldots, n-1\}$. One of the simplest time stepping schemes is the explicit *forward Euler* method

$$y^{k+1} = y^k + \Delta t\, z(y^k, t^k), \qquad k \in \{0, \ldots, n-1\}, \tag{3.10}$$

which belongs to the family of *Runge–Kutta* (RK) schemes and is first order accurate. Higher order accuracy can be achieved by using additional RK stages to approximate $y(t)$ at the nodes of a quadrature rule for numerical time integration on $[t^k, t^{k+1}]$.

In this work, we use *strong stability preserving* (SSP) RK schemes that were developed by Shu and Osher [Shu88] to preserve the *total variation diminishing* (TVD) property of spatial semi-discretizations at the fully discrete level. For that reason, such time integrators were originally called TVD RK schemes. The name SSP was introduced by Gottlieb et al. [Got01] to reflect the following useful property [Shu88, Got11].

Definition 3.5 (SSP property, Got11 Sec. 2.2)
Let $\|\cdot\|$ be a norm on $\mathbb{R}^M$ and $z(y, t)$ a Lipschitz-continuous function such that the forward Euler update (3.10) satisfies the stability condition

$$\|y + \Delta t\, z(y, t)\| \le \|y\| \qquad \forall y \in \mathbb{R}^M$$

for all Δt such that $0 < \Delta t \le (\Delta t)_{\mathrm{FE}}$. A general RK scheme is said to be strong stability preserving w. r. t. $\|\cdot\|$ if there exists an SSP coefficient $C_{\mathrm{SSP}} > 0$ such that $\|y^{k+1}\| \le \|y^k\|$ holds for y^{k+1} generated from y^k using the time step $\Delta t \le C_{\mathrm{SSP}}\, (\Delta t)_{\mathrm{FE}}$. ◇

An explicit s-stage RK method is commonly defined using the *Butcher tableau*. The SSP property can be shown using the equivalent *Shu–Osher form* [Got11, Sec. 2.3]

$$y^{(0)} = y^k,$$

$$y^{(i)} = \sum_{l=0}^{i-1} \left[\alpha_{il} y^{(l)} + \Delta t\, \beta_{il}\, z(y^{(l)}, t^{(l)})\right], \qquad i \in \{1, \ldots, s\},$$

$$y^{k+1} = y^{(s)}.$$

The SSP property is guaranteed to hold under a time step restriction if α_{il} and β_{il} are nonnegative for all indices i and l [Got11, Thm. 2.1].

The simplest SSP RK schemes are the forward Euler (SSP1) method

$$y^{k+1} = y^{(1)} = y^k + \Delta t\, z(y^k, t^k), \tag{3.11}$$

Heun's predictor-corrector (SSP2) method

$$y^{(1)} = y^k + \Delta t\, z(y^k, t_k), \tag{3.12a}$$

$$y^{k+1} = y^{(2)} = \frac{1}{2}y^k + \frac{1}{2}\left(y^{(1)} + \Delta t\, z(y^{(1)}, t^{k+1})\right), \tag{3.12b}$$

and the Shu–Osher (SSP3) scheme [Shu88]

$$y^{(1)} = y^k + \Delta t\, z(y^k, t^k), \tag{3.13a}$$

$$y^{(2)} = \frac{3}{4}y^k + \frac{1}{4}\left(y^{(1)} + \Delta t\, z(y^{(1)}, t^{k+1})\right), \tag{3.13b}$$

$$y^{k+1} = y^{(3)} = \frac{1}{3}y^k + \frac{2}{3}\left(y^{(2)} + \Delta t\, z\left(y^{(2)}, t^k + \Delta t/2\right)\right). \tag{3.13c}$$

These methods are optimal in the sense that their SSP coefficient is $C_{\mathrm{SSP}} = 1$, and pth order accuracy is achieved with p stages [Shu88]. Note that each stage in (3.11), (3.12), and (3.13) is a convex combination of y^k and forward Euler updates. Many property-preserving discretizations of hyperbolic problems exploit this fact [Zha11, Gue16b, Moe17]. Other explicit SSP methods admit representations similar to (3.12) and (3.13) and are thus also suitable for our purposes. For instance, the four-stage third order method, referred to as SSP(4,3) RK [Got11, Prog. 6.3], has the SSP coefficient $C_{\mathrm{SSP}} = 2$. The SSP coefficient $C_{\mathrm{SSP}} = 1.508$ of the five-stage fourth order SSP(5,4) RK method is also greater than unity. Such schemes allow the use of time steps larger than those required for SSPp RK $p \in \{1, 2, 3\}$, at the cost of having to compute approximations at additional intermediate stages. Unfortunately, there exists no four-stage fourth order SSP method such that all coefficients β_{il} are nonnegative [Got98, Prop. 3.3]. Moreover, this requirement imposes a fourth order accuracy barrier on explicit SSP RK methods [Ruu02, Thm. 4.1]. The issue of negative coefficients β_{il} can be cured by using *downwinding* strategies to adjust the spatial discretization operator in a manner that ensures relevant stability properties for a given RK stage (see [Got11, Ch. 10]).

The initial value problem (3.9) can also be solved implicitly, for instance, using the *backward Euler* method

$$y^{k+1} = y^k + \Delta t\, z(y^{k+1}, t^{k+1}), \qquad k \in \{0, \ldots, n-1\}. \tag{3.14}$$

In general, some iterative procedure is needed to solve (3.14) for y^{k+1}. Under the assumption that (3.14) is satisfied exactly, the backward Euler method is unconditionally SSP. That is, no restriction needs to be imposed on the time step Δt [Got11, Ch. 7]. Unfortunately, the SSP property without restrictions on Δt can only hold for first order methods [Got11, Thm. 5.1] or RK schemes that have negative coefficients β_{il} [Got01, Sec. 6]. Therefore, the only implicit approach considered in this work is the backward Euler scheme. A detailed discussion on SSP methods including implicit and multi-step approaches can be found in the book by Gottlieb et al. [Got11].

To construct a space-time discretization that is high order accurate and SSP regardless of the time step Δt, intermediate stages and/or the final stage of a general RK time integrator can be constrained using algebraic flux correction tools similar to the ones that we use for space discretizations in this thesis. Examples of such flux limiters for explicit and implicit RK schemes can be found in [Kuz22b] and [Que21], respectively.

3.2.2 Space-time finite element formulation

As an alternative to using SSP RK methods as time integrators for (3.7), we discuss the possibility of a finite element discretization on the *space-time cylinder* $\tilde{\Omega} = \Omega \times (0, T)$, where $T > 0$ is a finite end time. This approach may be appropriate, e. g., if the exact solution changes periodically in time, or for shock-fitting purposes.

On the continuous level, the time-dependent problem (3.1) is equivalent to

$$\tilde{\nabla} \cdot \tilde{\mathbf{f}}(u) = 0 \qquad \text{in } \tilde{\Omega} \subset \mathbb{R}^{d+1}, \tag{3.15}$$

where the divergence operator $\tilde{\nabla}\cdot$ now yields the sum of partial derivatives w. r. t. the $d+1$ independent variables $x_1, \dots, x_d, t$ and $\tilde{\mathbf{f}}(u) = [\mathbf{f}(u), u] \in \mathbb{R}^{m \times (d+1)}$. The wave speeds in direction $(\boldsymbol{n}_{\boldsymbol{x}}^{\mathsf{T}}, n_t) \in \mathrm{S}_1^d$ are the eigenvalues of the Jacobian matrix

$$\frac{\partial}{\partial u}\left(\tilde{\mathbf{f}}(u)\begin{bmatrix}\boldsymbol{n}_{\boldsymbol{x}} \\ n_t\end{bmatrix}\right) = \frac{\partial}{\partial u}\left(\mathbf{f}(u)\,\boldsymbol{n}_{\boldsymbol{x}}\right) + n_t\,\mathcal{I}_{m\times m}. \tag{3.16}$$

Characteristic boundary conditions need to be imposed on the boundary of the cylindrical space-time domain $\tilde{\Omega} \subset \mathbb{R}^{d+1}$. The external state $\hat{u}$ of the original time-dependent problem provides the Riemann data for the lateral surface $\partial\Omega \times (0, T)$ of $\tilde{\Omega}$. At the "bottom" $\Omega \times \{0\}$, the outward unit normal to $\partial\tilde{\Omega}$ is given by $(\mathbf{0}^{\mathsf{T}}, -1)$. It follows that all eigenvalues of (3.16) are negative, as for a supersonic inlet of the spatial domain Ω (see Section 2.2). Therefore, we use the initial condition $\hat{u} = u_0$ as external Riemann state on $\Omega \times \{0\}$. Similarly to a supersonic outlet of Ω, no boundary condition needs to be imposed at the "top" $\Omega \times \{T\}$ of the space-time cylinder $\tilde{\Omega}$. The corresponding external state is $\hat{u} = u$. These choices of $\hat{u}$ are consistent with the boundary and initial conditions that we use for the time dependent problem (3.1).

To discretize (3.15) in space-time, we will use the Galerkin method discussed in Section 3.1 and the limiting techniques presented in the next sections. If the governing equation and/or the discretization method is nonlinear, steady-state approximate solutions u_h to (3.15) can only be obtained by employing an iterative scheme. Given a suitable initial iterate u_h^0, one may use, e. g., the forward Euler-type fixed-point iteration

$$u_h^{k+1} = u_h^k + \Delta\tau\, z_h(u_h^k, \tau^k), \tag{3.17}$$

where $\Delta\tau = \Delta\tau(k) > 0$ is a pseudo time step and $z_h(\cdot, \tau^k)$ is a consistent discretization of $\tilde{\nabla} \cdot \tilde{\mathbf{f}}(u)$ with built-in lateral boundary conditions at pseudo time $\tau^k \in [0, \infty)$. Since we are interested in marching u_h to the steady state, it usually makes little sense to employ iterative schemes corresponding to higher order temporal discretizations. We say that (3.17) has converged to $u_h = u_h^k$ if the discrete l^2 norm of the *residual* $z_h(u_h^k, \tau^k)$ becomes smaller than a prescribed tolerance $\varepsilon > 0$, i. e., $\|z_h(u_h^k, \tau^k)\|_{l^2} < \varepsilon$.

If the time interval $(0, T)$ is subdivided into M subintervals, then the cells of the computational mesh for the space-time domain $\tilde{\Omega}$ are given by $K \times [t^n, t^{n+1}]$ for $K \in \mathcal{K}_h$ and $n = 0, \ldots, M-1$. Instead of solving an N-dimensional discrete problem in each stage of an SSP RK method for (3.7), we now have to solve an NM-dimensional one in each fixed-point iteration (3.17) for the FEM discretization of (3.15). Since the cost of a solution update is much higher for (a serial implementation of) the latter approach, the SSP RK version is usually preferable. However, the use of space-time finite elements for slices of $\Omega \times \mathbb{R}_+$ may be worthwhile, e. g., if the larger size of the discrete problems leads to more efficient parallel implementations, the space-time mesh is adaptive (aligned with interfaces), and/or the number of fixed-point iterations for a steady-state computation is much smaller than M. To speed up convergence, one could replace (3.17) by an implicit pseudo time stepping scheme [Gur09] or use a quasi-Newton solver [Möl08, Bad17, Loh21] for the (flux-corrected) space-time discretization of (3.15). In this thesis, we use the space-time formulation solely to test the ability of algebraic flux correction schemes to enforce discrete maximum principles for steady hyperbolic problems. In this context, the basic iteration (3.17) turned out to be sufficient.

3.3 Algebraic flux correction schemes

The standard spatial discretization techniques discussed in Section 3.1 are, in general, unreliable, particularly for solving hyperbolic problems with nonsmooth solutions. To overcome this issue, one can use methods such as the ones mentioned in Section 1.1. In this thesis, we focus on algebraic flux correction tools for finite element approximations [Kuz12b, Bar16, And17, Gue18a, Paz21].

This section begins with a review of some literature on AFC schemes. Then we derive a property preserving low order method that represents the well-behaved part of the baseline space discretization. By construction, the difference between the residuals

of the two semi-discrete schemes admits a decomposition into numerical fluxes that can be adjusted in an adaptive manner. For this purpose, we use a monolithic convex limiting technique [Kuz20a] and a semi-discrete entropy fix [Kuz20c].

3.3.1 Literature

The flux-corrected transport (FCT) methodology introduced by Boris and Book [Bor73, Boo75] laid the foundations of what we call AFC in this thesis. The original FCT algorithm is a one-dimensional conservative finite difference scheme that consists of a *transport + diffusion* stage and an *anti-diffusion* stage. As a result of the first stage, one obtains a property-preserving low order approximation, which is usually very diffusive. This drawback can be cured using antidiffusive fluxes such that a higher order accurate approximation is recovered if these fluxes are added to the low order predictor. Since even stable high order schemes may produce unacceptable under- and overshoots, a *limiter* is applied to (potentially) offending fluxes. The corresponding correction factors are inferred from inequality constraints based on local discrete maximum principles. Importantly, the scheme remains conservative because each flux has a counterpart that has the same magnitude and opposite sign. A fully multidimensional generalization of FCT was proposed in the context of finite volume approximations and structured grids by Zalesak [Zal79]. We review his flux limiting strategy in Chapter 5.

Löhner et al. [Löh87] were the first to develop an FCT algorithm for finite element discretizations of compressible flow problems on unstructured grids. Their original method uses an element-based version of Zalesak's multidimensional limiter. In the first stage, an explicit second-order Taylor-Galerkin scheme is stabilized using a low order artificial viscosity operator that represents the difference between the consistent and lumped mass matrices. In the second stage, limited antidiffusive element contributions are added to recover second order accuracy in smooth regions. Edge-based implementations of FEM-FCT use decompositions into fluxes rather than element contributions [Kuz12b, Chs. 5-7]. In this context, an edge (of the sparsity graph) links two degrees of freedom corresponding to a pair of nonvanishing off-diagonal matrix entries.

Kuzmin and Turek [Kuz02] extended FEM-FCT to implicit time stepping and derived a *discrete upwinding* formula for edge-based artificial diffusion coefficients. Further studies have revealed interesting relationships to some classical total variation diminishing (TVD) methods [Har84] and local extremum diminishing (LED) finite volume schemes for unstructured meshes [Jam93]. The unification of seemingly different approaches under the common roof of algebraic flux correction produced many new limiting algorithms for scalar conservation laws and nonlinear systems [Kuz12b, Chs. 6–8]. In particular, FCT-type finite element schemes for the Euler equations were developed in [Kuz10b, Loh16, Dob18] using an algebraic version of the local Lax–Friedrichs method in the first stage and different choices of quantities to be limited in the second stage. These choices are further discussed in [Löh08, Sec. 9.5.1] and [Kuz12b, Ch. 7].

The edge-based AFC scheme presented by Gurris [Gur09, Sec. 5.1.6] applies TVD-type limiters to local characteristic variables, transforms back to the conservative variables, and inserts limited antidiffusive fluxes into the right hand side of the semi-discrete scheme. In contrast to FCT-like methods, such *monolithic* AFC approaches produce nonlinear problems that have well-defined residuals and steady state solutions. Another highlight of the work presented in [Gur09] is the application of AFC to coupled systems of balance laws using operator splitting to include the source terms.

A theoretical framework for FEM-AFC discretizations of scalar conservation laws was introduced by Barrenechea et al. [Bar16] in the context of monolithic schemes for stationary convection-diffusion equations. Lohmann [Loh19, Ch. 4] extended this methodology to linear hyperbolic equations. Further advances regarding the analysis of AFC methods for linear problems are discussed in Chapter 5.

Explicit schemes for nonlinear systems were analyzed by Guermond and Popov [Gue16b]. Their work provides valuable insights into the properties of the low order method for *convex limiting* procedures [Gue18a]. As an alternative to the FCT-type localized limiters developed in [Gue18a] and their scalar prototypes [Loh17b], Kuzmin [Kuz20a] introduced a monolithic convex limiting (MCL) strategy for general hyperbolic problems. Similarly to the generalized TVD limiters employed by Gurris [Gur09], it can be applied to steady and transient problems alike. Another advantage of MCL compared to FCT-type predictor-corrector schemes is the possibility of enforcing semi-discrete entropy stability via flux limiting [Kuz20c]. In Chapter 4, we extend entropy stable MCL schemes to the shallow water equations with topography. Our approach ensures positivity preservation for the water height and does not use operator splitting algorithms such as the ones proposed in [Gur09].

The development of accuracy-preserving AFC schemes for very high order baseline discretizations requires careful generalizations of existing algorithms. The approaches developed in [And17, Loh17b, Kuz20e, Haj21a] employ *Bernstein polynomials* as local basis functions for finite element approximations. The high order FCT scheme developed by Pazner [Paz21] uses a collocated discontinuous Galerkin method. As shown in [Loh17b, Kuz20e, Paz21], convex limiting for high order finite element discretizations requires sparsification of discrete operators and localization of correction procedures to subcells. The author's own contributions to the development of such AFC schemes for continuous and discontinuous finite elements can be found in [Haj20b, Haj20c, Haj21a]. Chapter 6 describes the sparse high order MCL scheme for Bernstein DG discretizations of hyperbolic systems in some detail. In the present section, we explain the underlying design philosophy in the simpler context of continuous FEM discretizations.

3.3.2 Low order method

In Section 3.1, we discretized an initial-boundary value problem for the generic system of conservation laws (3.1) using the continuous Galerkin method and (multi-)linear finite

elements. Recall that the resulting system of semi-discrete equations (3.7) is given by

$$\sum_{j=1}^{N} m_{ij} \frac{\mathrm{d}u_j}{\mathrm{d}t} = -\int_\Omega \varphi_i \nabla \cdot \mathbf{f}(u_h)\,\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \varphi_i \left[\mathbf{f}(u_h)\boldsymbol{n} - \mathrm{f}_{\boldsymbol{n}}(u_h, \hat{u})\right]\,\mathrm{d}s \tag{3.18}$$

for $i \in \{1, \dots, N\}$. This baseline space discretization can be expected to deliver second order accuracy at least for linear advection problems with smooth solutions on uniform meshes. On general meshes, the provable order of accuracy is one in the linear scalar case [Qua94, Sec. 14.3.1], although second order superconvergence may be observed in practice.

We now modify (3.18) to construct a low-order discretization that is bound preserving and entropy stable [Gue16b]. To this end, we first use the *group finite element formulation* [Fle83] in the volume integral, i. e., we linearize the inviscid flux as follows

$$\mathbf{f}(u_h) = \mathbf{f}\Big(\sum_{j=1}^{N} u_j\,\varphi_j\Big) \approx \sum_{j=1}^{N} \mathbf{f}_j\,\varphi_j =: \mathbf{f}_h = \mathbf{f}_h(u_h), \qquad \mathbf{f}_j := \mathbf{f}(u_j), \qquad j \in \{1, \dots, N\}.$$

Substituting $\mathbf{f}_h(u_h)$ for $\mathbf{f}(u_h)$ in (3.18), one obtains a second order accurate quadrature-based approximation [Bar17b]. We use a similar nodal quadrature rule for the boundary integral that appears in (3.18). Specifically, we replace u_h and $\hat{u}$ by their nodal values $u_i = u_h(\boldsymbol{x}_i)$ and $\hat{u}_i := \hat{u}(\boldsymbol{x}_i)$ in the ith equation of (3.18). This approach can be interpreted as *mass lumping* for numerical fluxes, see [Sel96, Haj20b, Kuz20a].

To properly define the quadrature-based form of our semi-discrete problem, we need to introduce the set of boundary faces associated with a node.

Definition 3.6 (Nodal boundary faces)
Let $\mathcal{F}_{\partial\Omega}$ denote the set of boundary faces of $\mathcal{K}_h$ (see Definition 3.2). Then the sets of nodal boundary faces are defined by

$$\mathcal{F}_i = \begin{cases} \emptyset & \text{if } \boldsymbol{x}_i \notin \partial\Omega, \\ \{\Gamma \in \mathcal{F}_{\partial\Omega} : \mathrm{int}(\mathrm{supp}(\varphi_i)) \cap \Gamma \neq \emptyset\} & \text{otherwise,} \end{cases} \qquad i \in \{1, \dots, N\}. \qquad \diamond$$

Following Selmin [Sel96, Sec. 5], we evaluate the Riemann data $\hat{u}$ at the boundary vertices. For nodes $\boldsymbol{x}_i$ that lie on more than one boundary segment (e. g., in domain corners), we need to define individual nodal states for each face. Thus, we denote by $\hat{u}_i^k$ the external state $\hat{u}$ of the Riemann problem associated with node $\boldsymbol{x}_i$ and the normal vector $\boldsymbol{n}_k$ orthogonal to the face $\Gamma_k \in \mathcal{F}_i$. Using the above definition of $\mathcal{F}_i$, we write the quadrature-based approximation to (3.18) as

$$\sum_{j=1}^{N} m_{ij} \frac{\mathrm{d}\tilde{u}_j}{\mathrm{d}t} = -\int_\Omega \varphi_i \nabla \cdot \mathbf{f}_h(\tilde{u}_h)\,\mathrm{d}\boldsymbol{x} + \sum_{\Gamma_k \in \mathcal{F}_i} \int_{\Gamma_k} \varphi_i \left[\mathbf{f}(\tilde{u}_i)\boldsymbol{n}_k - \mathrm{f}_{\boldsymbol{n}_k}(\tilde{u}_i, \hat{u}_i^k)\right]\,\mathrm{d}s. \tag{3.19}$$

In the next lemma, we show that discretizations (3.18) and (3.19) are conservative.

Lemma 3.7 (Conservation property of finite element methods, Sel93)
Let u_h, $\tilde{u}_h \in \mathrm{V}_h^m$ be finite element approximations satisfying (3.18) and (3.19), respectively, for $i \in \{1,\ldots,N\}$. Then the following conservation properties hold

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u_h \,\mathrm{d}\boldsymbol{x} = -\int_{\partial\Omega} \mathrm{f}_{\boldsymbol{n}}(u_h,\hat{u})\,\mathrm{d}s, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{u}_h \,\mathrm{d}\boldsymbol{x} = -\sum_{i=1}^N \sum_{\Gamma_k\in\mathcal{F}_i}\int_{\Gamma_k} \varphi_i\,\mathrm{f}_{\boldsymbol{n}_k}(\tilde{u}_i,\hat{u}_i^k)\,\mathrm{d}s. \diamond$$

Proof:
Summing (3.18) over $i \in \{1,\ldots,N\}$, using the partition of unity property $\sum_{i=1}^N \varphi_i \equiv 1$ of the Lagrange basis functions and the definition of the mass matrix (3.8), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u_h \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \nabla\cdot\mathbf{f}(u_h)\,\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega}\left[\mathbf{f}(u_h)\boldsymbol{n} - \mathrm{f}_{\boldsymbol{n}}(u_h,\hat{u})\right]\,\mathrm{d}s.$$

Thus, the use of the divergence theorem proves the global conservation property of the Galerkin approximation u_h. Similarly, from (3.19) we infer that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{u}_h \,\mathrm{d}\boldsymbol{x} =& -\sum_{j=1}^N \int_{\partial\Omega}\varphi_j\,\mathbf{f}(\tilde{u}_j)\boldsymbol{n}\,\mathrm{d}s + \sum_{i=1}^N\sum_{\Gamma_k\in\mathcal{F}_i}\int_{\Gamma_k}\varphi_i\left[\mathbf{f}(\tilde{u}_i)\boldsymbol{n}_k - \mathrm{f}_{\boldsymbol{n}_k}(\tilde{u}_i,\hat{u}_i^k)\right]\,\mathrm{d}s\\ =& -\sum_{i=1}^N\sum_{\Gamma_k\in\mathcal{F}_i}\int_{\Gamma_k}\varphi_i\,\mathrm{f}_{\boldsymbol{n}_k}(\tilde{u}_i,\hat{u}_i^k)\,\mathrm{d}s,\end{aligned}$$

which proves the global conservation property of $\tilde{u}_h$ satisfying (3.19). □

Remark 3.8
Note that in 1D the boundary faces are vertices, and the flux-lumped approximation in (3.19) is equivalent to the consistent boundary integral in (3.18). Indeed, for a node $x_i \in \partial\Omega = \{x_1, x_N\}$, the set $\mathcal{F}_i$ contains just one face $\Gamma = x_i$ and the boundary integrals reduce to point evaluations in x_i. Thus, $\tilde{u}_h|_\Gamma = \tilde{u}_i$ and $\hat{u}|_\Gamma = \hat{u}_i$. It follows that

$$\varphi_i(x_i)\left[\mathrm{f}(\tilde{u}_i)n - \mathrm{f}_n(\tilde{u}_i,\hat{u}_i)\right] = \varphi_i(x_i)\left[\mathrm{f}(\tilde{u}_h)n - \mathrm{f}_n(\tilde{u}_h,\hat{u})\right], \qquad n = \pm 1. \qquad \diamond$$

The quadrature-based approximation (3.19) is of the same order as (3.18) for (multi-) linear finite elements. Therefore, we will use (3.19) as a *target scheme* for the AFC methods to be discussed in this section. For higher order finite element spaces, quadrature errors should be compensated in the process of flux correction (as in [Kuz20e]).

We now introduce the discrete gradient operator [Kuz02]

$$\mathbf{C} = (\mathbf{c}_{ij})_{i,j=1}^N, \qquad \mathbf{c}_{ij} = \int_\Omega \varphi_i\,\nabla\varphi_j\,\mathrm{d}\boldsymbol{x}, \qquad i,j\in\{1,\ldots,N\}, \tag{3.20}$$

which obviously possesses the properties [Sel93, Sel96]

$$\sum_{j=1}^{N} \mathbf{c}_{ij} = 0 \qquad i \in \{1, \ldots, N\}, \tag{3.21}$$

$$\mathbf{c}_{ij} = -\mathbf{c}_{ji} + \int_{\partial\Omega} \varphi_i \, \varphi_j \, \boldsymbol{n} \, \mathrm{d}s, \qquad i, j \in \{1, \ldots, N\}. \tag{3.22}$$

Thus, we have $\mathbf{c}_{ij} = -\mathbf{c}_{ji}$ for all pairs of indices (i, j) for which the boundary integral in (3.22) vanishes. For further reference, we also define

$$b_i^k = \int_{\Gamma_k} \varphi_i \, \mathrm{d}s, \qquad \Gamma_k \in \mathcal{F}_i, \ i \in \{1, \ldots, N\}. \tag{3.23}$$

Next, we modify the left hand side of the quadrature-based target scheme (3.19) by employing *row sum mass lumping*. That is, we replace the consistent mass matrix M, which is defined by (3.8), with its lumped counterpart

$$M_{\mathrm{L}} = (m_i \delta_{ij})_{i,j=1}^{N}, \qquad m_i := \sum_{j=1}^{N} m_{ij} = \int_{\Omega} \varphi_i \, \mathrm{d}\boldsymbol{x}, \qquad i \in \{1, \ldots, N\}. \tag{3.24}$$

This approximation can be interpreted as second order nodal quadrature for m_{ij}.

Renaming $\tilde{u}_h$ into u_h again, introducing the *nodal stencils*

$$\mathcal{N}_i = \{j \in \{1, \ldots, N\} : \mathrm{int}(\mathrm{supp}\, \varphi_i) \cap \mathrm{int}(\mathrm{supp}\, \varphi_j) \neq \emptyset\}, \qquad i \in \{1, \ldots, N\}$$

and invoking (3.21), the lumped mass version of (3.19) can be written as

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = -\sum_{j \in \mathcal{N}_i \backslash \{i\}} (\mathbf{f}_j - \mathbf{f}_i) \, \mathbf{c}_{ij} + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k \, [\mathbf{f}_i \, \boldsymbol{n}_k - \mathrm{f}_{\boldsymbol{n}_k}(u_i, \hat{u}_i^k)]. \tag{3.25}$$

Note that mass lumping does not interfere with the conservation properties stated in Lemma 3.7 because

$$\sum_{i=1}^{N} m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{i=1}^{N} \sum_{j=1}^{N} m_{ij} \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j=1}^{N} \sum_{i=1}^{N} m_{ij} \frac{\mathrm{d}u_j}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} u_h \, \mathrm{d}\boldsymbol{x}.$$

Remark 3.9

For well-posedness of (3.25), it is essential that $m_i > 0$ for all $i \in \{1, \ldots, N\}$. This property holds for (multi-)linear Lagrange elements but is violated, for instance, if quadratic Lagrange polynomials on triangles are employed. In addition, high order finite element approximations are not necessarily bounded by their values at the nodal points. These observations illustrate why higher order Lagrange elements are not well suited for algebraic flux correction purposes [Kuz08]. Instead, one may employ bases of Bernstein polynomials as in [Abg10, Loh17b]. This approach will be discussed in Chapter 6. $\diamond$

The lumped mass approximation (3.25) to (3.19) is equivalent to a vertex centered finite volume method with centered numerical fluxes [Sel93]. The addition of artificial viscosity on the right hand side of (3.25) is the final modification we need to make to obtain a property-preserving low order method of the form [Kuz05, Gue16b]

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} \left[d_{ij}(u_j - u_i) - (\mathbf{f}_j - \mathbf{f}_i)\, \mathbf{c}_{ij} \right] + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k \left[\mathbf{f}_i\, \boldsymbol{n}_k - \mathrm{f}_{\boldsymbol{n}_k}(u_i, \hat{u}_i^k) \right]. \tag{3.26}$$

The matrix of artificial viscosity coefficients d_{ij} has the properties of a discrete Laplacian (also called *graph Laplacian* or *discrete diffusion* operator in the AFC literature).

In the terminology adopted in [Gue18a, Paz21], the fluxes $d_{ij}(u_j - u_i)$ inserted in (3.26) introduce large amounts of *graph viscosity*. Already the one-dimensional FCT schemes of Boris and Book [Bor73, Boo75] used artificial viscosity of this kind in the low order method of the first stage. In the context of FEM-FCT schemes for unstructured meshes, Löhner et al. [Löh87, Eqs. (18)–(19)] used element-level numerical viscosity. Their elementwise diffusion matrix is a discrete Laplacian defined as the difference of lumped and consistent element mass matrices. While this approach works reasonably well in practice, there is no guarantee that it introduces the correct amount of artificial viscosity. Indeed, the authors of [Löh87] remark that they observed some nonphysical artifacts. Kuzmin and Turek [Kuz02, Eq. (33)] constructed the low order method for their FEM-FCT discretization of the linear advection equation using a definition of d_{ij} that can be interpreted as edge-based *discrete upwinding* (see also [Sel93, Sec. 3.1] or [Jam93, Sec. 2.1]). This approach will be further discussed in Chapter 5.

In the low order method (3.26) for systems, we employ [Abg06, Kuz10b, Gue16b]

$$d_{ij} = \max\{\lambda_{ij}|\mathbf{c}_{ij}|, \lambda_{ji}|\mathbf{c}_{ji}|\}, \qquad \lambda_{ij} \geq \lambda_{\boldsymbol{n}}(u_i, u_j), \qquad i, j \in \{1, \ldots, N\},\ i \neq j, \tag{3.27}$$

where $\lambda_{\boldsymbol{n}}(\cdot, \cdot)$ in (3.27) is given by (3.2b) for $\boldsymbol{n} = \mathbf{c}_{ij}/|\mathbf{c}_{ij}| \in \mathrm{S}_1^{d-1}$. For this choice of d_{ij}, the addition of $d_{ij}(u_j - u_i)$ transforms the centered flux of an equivalent vertex-centered finite volume approximation into the local Lax–Friedrichs (Rusanov) flux [Sel93, Sec. 2.2]. Therefore, we call (3.26) with d_{ij} defined by (3.27) the *algebraic local Lax–Friedrichs* (LLF) method. Guermond and Popov [Gue16b] proved that this approximation is property preserving even for systems. An element-based version of the LLF scheme was proposed in [Abg06] and analyzed in [Gue16b].

Replacing sums over the nodal stencils by sums over all degrees of freedom and using the fact that $d_{ij} = d_{ji}$ for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ yields

$$\sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i \setminus \{i\}} d_{ij}(u_j - u_i) = \sum_{\substack{i,j=1 \\ i<j}}^{N} d_{ij}(u_j - u_i) + \sum_{\substack{i,j=1 \\ i<j}}^{N} d_{ji}(u_i - u_j) = 0.$$

Thus, the low order discretization (3.26) is conservative in the sense of Lemma 3.7, [Löh87, Sel93, Kuz02]. To see that it preserves other relevant properties of the entropy

solution to (3.1), we introduce the low order *bar states* [Har83b, Aud15, Gue16b]

$$\bar{u}_{ij} := \frac{u_i + u_j}{2} - \frac{(\mathbf{f}_j - \mathbf{f}_i)\,\mathbf{c}_{ij}}{2d_{ij}}, \qquad i \in \{1, \dots, N\},\ j \in \mathcal{N}_i \setminus \{i\}. \tag{3.28}$$

A bar state concept similar to (3.28) was developed by the author of this thesis [Haj21a] for lumped flux terms arising in discontinuous Galerkin discretizations. We apply this concept here as well to incorporate external Riemann data in a property-preserving manner. To this end, we restrict ourselves to using the local Lax–Friedrichs flux (3.2) for $\mathrm{f}_{\boldsymbol{n}_k}(\cdot,\cdot)$. This choice allows us to express the boundary terms in (3.26)

$$\begin{aligned} b_i^k \left[\mathbf{f}_i\,\boldsymbol{n}_k - \mathrm{f}_{\boldsymbol{n}_k}(u_i, \hat{u}_i^k)\right] &= \frac{b_i^k}{2}\left[\lambda_{\boldsymbol{n}_k}(u_i, \hat{u}_i^k)\,(\hat{u}_i^k - u_i) - (\mathbf{f}(\hat{u}_i^k) - \mathbf{f}_i)\,\boldsymbol{n}_k\right] \\ &= 2d_i^k(\bar{u}_i^k - u_i) \end{aligned} \tag{3.29}$$

in terms of the nodal quantities

$$d_i^k := \frac{b_i^k}{2}\,\lambda_{\boldsymbol{n}_k}(u_i, \hat{u}_i^k), \tag{3.30}$$

$$\bar{u}_i^k := \frac{u_i + \hat{u}_i^k}{2} - \frac{(\mathbf{f}(\hat{u}_i^k) - \mathbf{f}_i)\,\boldsymbol{n}_k}{2\lambda_{\boldsymbol{n}_k}(u_i, \hat{u}_i^k)}. \tag{3.31}$$

The so-defined bar states $\bar{u}_i^k$ of boundary nodes exhibit the same structure as their edge-based counterparts $\bar{u}_{ij}$ defined by (3.28). Using (3.29), we rewrite the low order method (3.26) in terms of (3.28) and (3.31) as follows (cf. [Gue16b, Haj21a])

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij}(\bar{u}_{ij} - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k\,(\bar{u}_i^k - u_i), \qquad i \in \{1, \dots, N\}. \tag{3.32}$$

Remark 3.10
Note that the structure of volume and boundary terms in the right hand side of (3.32) is similar, which is convenient for designing flux correction schemes. In the discontinuous Galerkin version of (3.26), the set $\mathcal{F}_i$ includes internal faces $\Gamma_k \in \mathcal{F}_\Omega$. The interfacial bar states $\bar{u}_i^k$ are defined similarly to (3.31) using the external limits $\hat{u}_i^k$ as Riemann data. Guermond et al. [Gue19, Sec. 4] combine the volume terms and boundary fluxes of such DG-AFC approaches into a single sum formulation using a unified notation for the corresponding discrete operators. Such a representation is also possible for (3.32) but obstructs the different origin of the terms that appear in the two sums. For that reason, we prefer the split form (3.32) of the low order method. In this chapter, we do not perform antidiffusive corrections for the sum over $\Gamma_k \in \mathcal{F}_i$ because such corrections have little influence on the accuracy of continuous finite element approximations. ⋄

Using the bar state form (3.32) of the LLF method (3.26), one can show that numerical approximations stay in a subset $\mathcal{A}$ of $\mathcal{A}^{\max}$. Guermond and Popov [Gue16b] prove this property for convex *invariant sets*, which are defined as follows.

Definition 3.11 (Invariant sets, Gue16b Sec. 2)
Let $\mathcal{A}^{\max} \neq \emptyset$ be the largest admissible set for (3.1). Assume that there exists a unique entropy solution u to each one-dimensional Riemann problem of the form

$$\frac{\partial u}{\partial t} + (\boldsymbol{n} \cdot \nabla)\left(\mathbf{f}(u)\,\boldsymbol{n}\right) = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}_+, \qquad u(x,0) = \begin{cases} u_L & \text{if } x < 0, \\ u_R & \text{if } x > 0, \end{cases} \tag{3.33}$$

where $(u_L, u_R) \in \mathcal{A}^{\max} \times \mathcal{A}^{\max}$ and $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$. A generalization of Corollary 2.17 to systems guarantees the existence of finite speeds λ_L and λ_R such that $u(x,t) = u_L$ for $x \leq \lambda_L t$ and $u(x,t) = u_R$ for $x \geq \lambda_R t$. A subset $\mathcal{A} \subseteq \mathcal{A}^{\max}$ is called an invariant set of (3.1) if for any initial data $(u_L, u_R) \in \mathcal{A} \times \mathcal{A}$ the spatial average

$$\bar{u}_{LR} = \frac{1}{(\lambda_R - \lambda_L)t} \int_{\lambda_L t}^{\lambda_R t} u(x,t)\,\mathrm{d}x$$

of the entropy solution $u(x,t)$ to (3.33) remains in $\mathcal{A}$ for all times $t \geq 0$. ◇

Discretizations that keep the values (or averages) of numerical solutions in an invariant set $\mathcal{A}$ of the continuous problem (3.1) are referred to as *invariant domain preserving* (IDP), *positivity preserving*, or *structure preserving* in the literature.

Guermond and Popov [Gue16b] discretize (3.32) in time using an explicit SSP RK method and show that the resulting fully discrete scheme is IDP. Additionally, they prove that a fully discrete entropy inequality holds for every entropy pair of the hyperbolic problem [Gue16b, Sec. 4]. Their IDP analysis of the LLF method can be readily extended to SSP RK time discretizations of other semi-discrete schemes that can be expressed in terms of admissible bar states similarly to (3.32). If a given scheme has the same structure as (3.32), then the proof requires the IDP property of the corresponding bar states. We first prove it for the low order method in the following lemma.

Lemma 3.12 (Admissibility of the bar states, Har83b, Gue16b)
Let $(u_L, u_R) \in \mathcal{A} \times \mathcal{A}$, where $\mathcal{A}$ is a convex invariant set of (3.1). Assume that for all $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$ the one dimensional Riemann problem (3.33) has a unique entropy solution u. Then the bar states (3.28) and (3.31) corresponding to (u_L, u_R) are also in $\mathcal{A}$. ◇

Proof:
For hyperbolic systems, the property that $\bar{u}_{ij} \in \mathcal{A}$ for $u_i, u_j \in \mathcal{A}$ can be shown by integrating (3.1) over space-time control volumes, using the definition of an invariant set and exploiting convexity. See [Gue16b, Lem. 2.1 and Sec. 3.3] for details. For scalar conservation laws, definition (3.27) of d_{ij} and the mean value theorem imply that $\bar{u}_{ij} \in \mathcal{A}$ is a convex combination of u_i and u_j [Kuz20a]. Virtually the same arguments can be used to verify that $\bar{u}_i^k \in \mathcal{A}$ provided that $u_i, \hat{u}_i^k \in \mathcal{A}$ [Haj21a, Thm. 1]. □

Remark 3.13
The statement of Lemma 3.12 is valid for general hyperbolic systems. To avoid the computation of $\lambda_{\boldsymbol{n}}(u_L, u_R)$ defined by (3.2b), overestimation of the maximum wave speed was proposed by Harten et al. [Har83b, Sec. 3 B]. Computable bounds for $\lambda_{\boldsymbol{n}}(u_L, u_R)$ can be found in [Gue16a] for the Euler equations and in [Aze17, Gue18b] for the SWE. The wave speeds of the two systems are bounded by $\boldsymbol{v} \cdot \boldsymbol{n} \pm a$, where a denotes the *speed of sound* $\sqrt{\gamma p/\rho}$ or the *celerity* $\sqrt{gh}$, respectively (cf. Section 2.2). Thus, the spectral radius of the directional Jacobian is given by $|\boldsymbol{v} \cdot \boldsymbol{n}| + a$. We use this formula to construct bounds for $\lambda_{ij} = \lambda_{\mathbf{c}_{ij}/|\mathbf{c}_{ij}|}(u_i, u_j)$ as in [Kuz10b, Sec. 6] for the Euler equations and in [Wu21, Sec. 3.1] for the SWE. That is, we set (cf. [Sel93, Sec. 3.1])

$$\lambda_{ij} = \max\left\{ \frac{|\boldsymbol{v}_i \cdot \mathbf{c}_{ij}|}{|\mathbf{c}_{ij}|} + a_i, \frac{|\boldsymbol{v}_j \cdot \mathbf{c}_{ji}|}{|\mathbf{c}_{ji}|} + a_j \right\}. \tag{3.34}$$

As illustrated in [Gue16a, Appendix B], such definitions are not necessarily upper bounds for the actual wave speeds $\lambda_{\mathbf{c}_{ij}/|\mathbf{c}_{ij}|}(u_i, u_j)$. Nevertheless, we observed no practical problems when using the two-state wave speed estimate (3.34). In fact, this choice of λ_{ij} guarantees positivity preservation for the systems we are interested in. Wu et al. [Wu21, Thm. 3.1] proved that (3.34) ensures nonnegativity of the water height of the SWE. Positivity preservation for the density and internal energy of the Euler system follows from a combination of [Lin22, Lem. 5.1] and [Zha17, Lem. 6]. We conclude that the IDP property for particular systems can sometimes be shown under weaker assumptions than those made in Lemma 3.12. ◇

At this stage, we depart from the fully discrete approach pursued by Guermond et al. [Gue16b, Gue18a] and instead continue the discussion in the semi-discrete setting following Kuzmin [Kuz20a]. A summary on FCT limiters that could alternatively be used in combination with the low order method (3.26) can be found in Section 5.2.3.

3.3.3 Definition of raw antidiffusive fluxes

The difference between the residuals of the target scheme (3.19) and of the low order LLF method is caused by mass lumping on the left hand side and addition of dissipative fluxes $d_{ij}(u_j - u_i)$ on the right hand side of (3.26). To recover (3.19) from (3.26) in nodes that lie in regions of smoothness, we use raw antidiffusive fluxes $f_{ij} \in \mathbb{R}^m$ such that $f_{ij} = -f_{ji}$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ and $f_{ij} = 0$ if $j \notin \mathcal{N}_i \setminus \{i\}$. In the FCT context, the fluxes f_{ij} are called *phoenical* ("like a phoenix") if the high order solution can be resurrected exactly in the second stage [Boo75]. In general AFC schemes, the fluxes f_{ij} are used to write a target scheme in the equivalent form

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} \left[d_{ij}(u_j - u_i) - (\mathbf{f}_j - \mathbf{f}_i)\, \mathbf{c}_{ij} + f_{ij} \right] + \sum_{\Gamma_k \in \mathcal{F}_i} 2 d_i^k (\bar{u}_i^k - u_i). \tag{3.35}$$

Such algebraic splittings (3.35) of semi-discrete schemes offer many possibilities to modify f_{ij} similarly to antidiffusive components of numerical fluxes in finite volume schemes. For example, a high order dissipative component may be added to f_{ij} or the magnitude of f_{ij} may be reduced using a flux limiter. The scheme remains conservative if the modified fluxes f_{ij}^* and f_{ji}^* satisfy $f_{ij}^* = -f_{ji}^*$ for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$.

To design an algebraic flux correction scheme based on (3.35), we need to define the raw antidiffusive fluxes f_{ij}, formulate additional constraints for f_{ij}^*, and devise an algorithm for enforcing these constraints. We begin with the first task, the choice of fluxes f_{ij} such that (3.35) is stable and second order accurate. Clearly, f_{ij} must compensate the diffusive flux $d_{ij}(u_j - u_i)$, which can be accomplished by using

$$f_{ij} = d_{ij}(u_i - u_j).$$

This definition of f_{ij} is phoenical w. r. t. the lumped mass version of (3.19). We use it in time marching schemes for steady problems. In the context of Galerkin approximations to time dependent problems, mass lumping introduces significant phase errors (see [Tho16]) and should be avoided. The consistent mass version (3.19) can be recovered using

$$f_{ij} = m_{ij}\left(\frac{\mathrm{d}u_i}{\mathrm{d}t} - \frac{\mathrm{d}u_j}{\mathrm{d}t}\right) + d_{ij}(u_i - u_j). \tag{3.36}$$

Indeed, the sum of additional terms depending on m_{ij} can be rewritten as

$$\sum_{j \in \mathcal{N}_i \setminus \{i\}} m_{ij}\left(\frac{\mathrm{d}u_i}{\mathrm{d}t} - \frac{\mathrm{d}u_j}{\mathrm{d}t}\right) = \sum_{j=1}^{N} m_{ij}\left(\frac{\mathrm{d}u_i}{\mathrm{d}t} - \frac{\mathrm{d}u_j}{\mathrm{d}t}\right) = m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} - \sum_{j=1}^{N} m_{ij}\frac{\mathrm{d}u_j}{\mathrm{d}t},$$

which recovers the consistent mass matrix in the left hand side of (3.35). Definition (3.36) is phoenical w. r. t. (3.19) if the solution of the linear system

$$\sum_{j=1}^{N} m_{ij}\dot{u}_j^{\mathrm{G}} = -\sum_{j \in \mathcal{N}_i \setminus \{i\}} (\mathbf{f}_j - \mathbf{f}_i)\,\mathbf{c}_{ij} + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k(\bar{u}_i^k - u_i), \qquad i \in \{1, \ldots, N\} \tag{3.37}$$

is used to calculate the time derivatives $\frac{\mathrm{d}u_i}{\mathrm{d}t} = \dot{u}_i^{\mathrm{G}}$ for (3.36). In principle, this is a viable option. However, it corresponds to an unstabilized Galerkin (hence superscript G) target scheme, which tends to produce spurious oscillations. Evidence for the occurrence of such ripples is provided by the test problems studied in Sections 3.4.3.2 and 5.5.2.

To stabilize the raw antidiffusive fluxes (3.36), the oscillatory time derivatives $\dot{u}_i^{\mathrm{G}}$ may be replaced by their low order (hence superscript L) counterparts

$$\dot{u}_i^{\mathrm{L}} = \frac{1}{m_i}\sum_{j \in \mathcal{N}_i \setminus \{i\}} [d_{ij}(u_j - u_i) - (\mathbf{f}_j - \mathbf{f}_i)\,\mathbf{c}_{ij}] + \frac{1}{m_i}\sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k(\bar{u}_i^k - u_i) \tag{3.38}$$

for $i \in \{1, \ldots, N\}$. Note that the right hand side of (3.38) corresponds to that of the low order method (3.26) and there is no need to invert the consistent mass matrix.

Remark 3.14
The need for stabilization of f_{ij} arises due to the use of a continuous finite element approximation. For a conforming discretization of the linear advection equation, a priori error analysis guarantees the suboptimal convergence rate of $\mathcal{O}(h^p)$ if polynomial bases of degree p and general meshes are employed [Qua94, Sec. 14.3.1]. Optimal convergence behavior for $p = 1$ is usually observed if the fluxes f_{ij} of the target scheme (3.35) are defined using the low order time derivatives (3.38) [Kuz20a]. As shown in [Haj21b], the replacement of $\dot{u}_i^{\mathrm{G}}$ by $\dot{u}_i^{\mathrm{L}}$ generates fourth order background dissipation. Higher order finite element methods require different stabilization techniques because the amount of artificial diffusion introduced by the low order time derivatives is too large to preserve optimal orders of accuracy. The list of available stabilization techniques for FEM is too long to be properly covered here. Stabilization of target fluxes in AFC schemes for linear problems is discussed in [Loh17b, Sec. 5]. State of the art strategies for high order discretizations of nonlinear conservation laws can be found in [Kuz20d].

Guermond et al. [Gue18a] define the raw antidiffusive fluxes (3.36) using nodal time derivatives corresponding to a truncated series approximation to the solution of (3.37). For stabilization purposes, they include nonlinear high order dissipation of the form $d_{ij}^{\mathrm{EV}}(u_j - u_i)$, where d_{ij}^{EV} is an entropy viscosity coefficient. The magnitude of d_{ij}^{EV} depends on the residual of the discretized identity $\eta'(u)^{\mathsf{T}}\nabla \cdot \mathbf{f}(u) = \nabla \cdot \boldsymbol{q}(u)$ for a specific entropy pair $(\eta, \boldsymbol{q})$. Details on this algorithm can be found in [Gue18a, Sec. 3.4]. To the best of our knowledge, AFC schemes using such residual-based flux stabilization have so far only been tested in the context of continuous (multi-)linear finite elements. ◇

In principle it is also possible to include correction terms compensating quadrature errors in the antidiffusive fluxes. Such schemes will be discussed in Chapter 6.

3.3.4 Monolithic convex limiting

3.3.4.1 Design philosophy

Having discussed the possible definitions of raw (unlimited) antidiffusive fluxes for the phoenical splitting (3.35) of the target scheme, we now present the monolithic convex limiting strategy proposed by Kuzmin [Kuz20a]. This semi-discrete alternative to fractional-step FCT algorithms is derived from a flux-corrected version of the LLF method (3.32). Rewriting (3.35) in terms of the bar states (3.28) and replacing the fluxes f_{ij} with their limited counterparts $f_{ij}^* \in \mathbb{R}^m$ (to be specified below), we obtain

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij}\Big(\bar{u}_{ij} - u_i + \frac{f_{ij}^*}{2d_{ij}}\Big) + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k(\bar{u}_i^k - u_i) \tag{3.39}$$

for $i \in \{1, \ldots, N\}$. Introducing the limited bar states [Kuz20a]

$$\bar{u}_{ij}^* := \bar{u}_{ij} + \frac{f_{ij}^*}{2d_{ij}}, \qquad i \in \{1, \ldots, N\},\ j \in \mathcal{N}_i \setminus \{i\}, \tag{3.40}$$

the flux-corrected semi-discrete scheme (3.39) can be written as

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij}(\bar{u}_{ij}^* - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k \, (\bar{u}_i^k - u_i), \qquad i \in \{1, \dots, N\}. \tag{3.41}$$

Note that (3.41) has the same structure as the bar state form (3.32) of the low order LLF scheme. If we discretize (3.41) in time using an explicit SSP Runge–Kutta method, each stage is a forward Euler update of the form [Gue16b, Kuz20a]

$$\begin{aligned} \tilde{u}_i = \Big[1 - \frac{\Delta t}{m_i}\Big(\sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij} + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k \Big)\Big] u_i \\ + \frac{\Delta t}{m_i}\Big(\sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij} \bar{u}_{ij}^* + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k \, \bar{u}_i^k \Big), \qquad i \in \{1, \dots, N\}. \end{aligned} \tag{3.42}$$

Here $\tilde{u}_i$ is the updated solution and $\Delta t > 0$ is the current time step. Below we discuss ways of enforcing various admissibility constraints for the limited bar states (3.40). In general, we constrain the states $\bar{u}_{ij}^*$ to lie in a certain convex set $\mathcal{A}_i \subseteq \mathcal{A}^{\max}$ that also contains u_i and the boundary term bar states $\bar{u}_i^k$. Thus, if the time step Δt is small enough to satisfy the Courant–Friedrichs–Lewy (CFL)-like condition [Gue16b, Kuz20a]

$$1 - \frac{\Delta t}{m_i}\Big(\sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij} + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k \Big) \geq 0, \tag{3.43}$$

then the updated solution $\tilde{u}_i$ is a convex combination of elements in $\mathcal{A}_i$ due to (3.42). Hence, convexity of $\mathcal{A}_i$ implies $\tilde{u}_i \in \mathcal{A}_i$ for all $i \in \{1, \dots, N\}$.

Similarly to the approach presented in [Gue18a], we use condition (3.43) to calculate suitable values of the time step Δt for SSP RK methods. Our version of the algorithm for adaptive time step control based on (3.43) consists of the following steps:

- Pick a fixed CFL parameter $\nu \in (0, 1]$ during the initialization step.
- In stage s of every Runge–Kutta cycle, use $u^{(s-1)}$ to compute

$$\tau^{(s)} := \min_{i \in \{1, \dots, N\}} \frac{m_i}{\sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij} + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k}.$$

- After performing the first RK stage update, set the time step to $\Delta t = \nu \tau^{(1)}$.
- If it turns out that $\tau^{(s)} < \Delta t$ for some s, then reject the intermediate results of all RK stages and repeat the whole RK cycle using the time step $\Delta t = \nu \tau^{(s)}$.

The question of how to define the limited antidiffusive fluxes f_{ij}^* for the semi-discrete scheme (3.41) remains. We already mentioned that these fluxes must satisfy $f_{ij}^* = -f_{ji}^*$ and ensure that $\bar{u}_{ij}^* \in \mathcal{A}_i \subseteq \mathcal{A}^{\max}$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. The remainder of our presentation on MCL is devoted to the construction of limited antidiffusive fluxes, which meet these requirements. Our approach adjusts (individual components of) the antidiffusive fluxes step by step until all relevant constraints are satisfied by f_{ij}^*.

In the process of multistage flux limiting for f_{ij}, we first enforce *numerical admissibility conditions*, which are meant to eliminate/reduce spurious oscillations, e. g., due to Gibbs phenomena. Next, we constrain the antidiffusive fluxes to ensure the IDP property. In other words, we impose *physical admissibility conditions*. Finally, we further limit the bound-preserving MCL fluxes f_{ij}^* if they violate a semi-discrete entropy stability condition based on Tadmor's theory [Tad03, Sec. 3], see also [Tad87, Kuz20c].

3.3.4.2 Scalar conservation laws

Let us begin with a discussion on numerical admissibility conditions for $m = 1$, i. e., for a scalar conservation law. As a general rule, the admissible set $\mathcal{A}_i \subset \mathbb{R}$ for the degree of freedom $i \in \{1, \dots, N\}$ must contain both the low order bar states $\bar{u}_{ij}$ and their limited counterparts $\bar{u}_{ij}^*$ for all $j \in \mathcal{N}_i \setminus \{i\}$. By default, we constrain $\bar{u}_{ij}^*$ to be contained in the interval $[u_i^{\min}, u_i^{\max}]$ with local bounds defined by [Kuz20a, Sec. 4]

$$u_i^{\min} := \min_{j \in \mathcal{N}_i} u_j, \qquad u_i^{\max} := \max_{j \in \mathcal{N}_i} u_j. \tag{3.44}$$

The mean value theorem implies that there exist $\xi \in [\min\{u_i, u_j\}, \max\{u_i, u_j\}]$ and, owing to (3.27), $\mu := \mathbf{f}'(\xi) \cdot \mathbf{c}_{ij}/d_{ij} \in [-1, 1]$ such that the low order bar states satisfy

$$\bar{u}_{ij} = \frac{u_i + u_j}{2} - \frac{(u_j - u_i)\, \mathbf{f}'(\xi) \cdot \mathbf{c}_{ij}}{2 d_{ij}} = \frac{1+\mu}{2} u_i + \frac{1-\mu}{2} u_j.$$

Therefore, $\bar{u}_{ij} \in [\min\{u_i, u_j\}, \max\{u_i, u_j\}]$. It follows that $\bar{u}_{ij} \in [u_i^{\min}, u_i^{\max}]$ for $u_i^{\min}$ and $u_i^{\max}$ defined by (3.44). Let us now constrain the limited bar states in such a manner that $u_i^{\min} \le \bar{u}_{ij}^* \le u_i^{\max}$ will hold for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. Using (3.40) and taking into account that similar conditions must hold for $f_{ji}^* = -f_{ij}^*$, local maximum principles for $\bar{u}_{ij}^*$ and $\bar{u}_{ji}^*$ can be converted into the flux constraints

$$f_{ij}^{\min} \le f_{ij}^* \le f_{ij}^{\max} \tag{3.45}$$

with local bounds defined by

$$f_{ij}^{\min} = -f_{ji}^{\max} := \max\left\{2 d_{ij} u_i^{\min} - \bar{w}_{ij}, \bar{w}_{ji} - 2 d_{ij} u_j^{\max}\right\} \le 0, \tag{3.46a}$$

$$f_{ij}^{\max} = -f_{ji}^{\min} := \min\left\{2 d_{ij} u_i^{\max} - \bar{w}_{ij}, \bar{w}_{ji} - 2 d_{ij} u_j^{\min}\right\} \ge 0. \tag{3.46b}$$

In practical implementations, the products $\bar{w}_{ij} := 2d_{ij}\bar{u}_{ij}$ are calculated directly instead of $\bar{u}_{ij}$ to avoid division by diffusion coefficients d_{ij} that may approach zero [Kuz20a].

The scalar version of Kuzmin's [Kuz20a] monolithic convex limiter yields

$$f_{ij}^* = \begin{cases} \min\left\{f_{ij},\ f_{ij}^{\max}\right\} & \text{if } f_{ij} \geq 0, \\ \max\left\{f_{ij},\ f_{ij}^{\min}\right\} & \text{if } f_{ij} \leq 0 \end{cases} \tag{3.47}$$

for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. Note that (3.47) is equivalent to

$$f_{ij}^* = \max\left\{f_{ij}^{\min}, \min\left\{f_{ij},\ f_{ij}^{\max}\right\}\right\} = \min\left\{f_{ij}^{\max}, \max\left\{f_{ij},\ f_{ij}^{\min}\right\}\right\}. \tag{3.48}$$

That is, the MCL method simply trims the flux f_{ij} to satisfy the box constraints (3.45).

Remark 3.15
Since the low order bar states correspond to averaged solutions of a Riemann problem, a natural alternative to definition (3.44) is given by (cf. [Gue18a])

$$u_i^{\min} := \min_{j \in \mathcal{N}_i \setminus \{i\}} \bar{u}_{ij}, \qquad u_i^{\max} := \max_{j \in \mathcal{N}_i \setminus \{i\}} \bar{u}_{ij}. \tag{3.49}$$

We tested this approach for nonlinear conservation laws in 1D for which we did not observe any significant differences to the numerical results obtained with (3.44). The disadvantage of using (3.49) is that the computation of low order bar states $\bar{u}_{ij}$ can no longer be avoided in general. Hence, we use (3.44) for scalar conservation laws. ◇

The reader will have noticed that so far we did not specify at which time instant the bounds (3.44) or (3.49) are evaluated. Owing to the monolithic limiting approach, the scheme under investigation is still in semi-discrete form. In explicit RK methods for (3.41), the bar states and local bounds are computed from the nodal values at the current stage. Similarly, nonlinear iterations for approximating the update of an implicit method use the nodal values of the current iterate. This approach distinguishes the MCL approach from FCT algorithms in which the local bounds $u_i^{\min}$ and $u_i^{\max}$ for the second stage are usually computed using a low order predictor [Bor73, Kuz02, Loh17b]. Some FCT limiters, however, define the bounds using the current solution instead of [And17, Gue18a, Paz21] or in addition to [Zal79, Löh87] the low order values.

3.3.4.3 Systems of conservation laws

The question of how to define numerical admissibility conditions is more involved for hyperbolic systems than it is for scalar problems. Appropriate choices are certainly peculiar to particular systems of equations. Much more so than for scalar conservation laws, the best choice of local bounds for limiting depends on the specific problem setting. Before presenting the approach pursued in this chapter, we list a few suggestions from

the literature. A natural first step is to treat each component of the degree of freedom $u_i \in \mathbb{R}^m$ as a scalar unknown and limit it separately. Even though this approach produces the least diffusive results, Löhner et al. [Löh87] advise against using such independent limiters because they lead to spurious oscillations in all quantities of interest. Instead, the authors of [Löh87] propose *synchronized limiters* such that $f_{ij}^* = \alpha_{ij} f_{ij}$ for $f_{ij} \in \mathbb{R}^m$, where $\alpha_{ij} \in [0, 1]$ is a scalar correction factor. Different ways to choose α_{ij} are discussed in [Löh87, p. 1099] for the compressible Euler equations. Examples of modern synchronized flux correction schemes can be found in [Loh16, Gue18a, Paz21].

An alternative to independent or synchronized limiting is the *sequential* limiting technique, originally proposed in [Dob18] and further developed in [Haj19, Kuz20a]. This approach is designed to enforce numerical admissibility constraints not for the conserved unknowns but for quantities that are derived from them. Particular representatives of such derived quantities are, for instance, the components of the velocity vector, which are defined as the quotients of momentum $\rho\boldsymbol{v}$ components and density ρ in case of the Euler equations, and as the quotients of discharge $h\boldsymbol{v}$ components and water height h for the SWE. Another important derived quantity of the Euler system is the specific total energy E. It is defined as the quotient of the conserved unknown ρE and density ρ. Vector-valued unknowns, such as the velocity, are limited by imposing local maximum principles on each component separately. Alternative strategies, in particular ones that guarantee rotational invariance of results, are explored in [Haj19].

Below we discuss the sequential limiting approach, as presented in the context of the Euler equations in [Kuz20a, Sec. 5.1] for products of the form $\varrho\phi$, where ϱ is a positive scalar conserved unknown and ϕ is a specific quantity of the fluid. The requirement $\varrho > 0$ is met both for the SWE, where $\varrho = h$ and for the Euler equations, where $\varrho = \rho$. The challenge is to limit the fluxes of $\varrho\phi$ in a way that guarantees preservation of local bounds for ϕ. We first discuss the general concept, before specifying appropriate bounds for limiting. To ease the presentation, we consider a fixed set of nodes $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ with corresponding low order bar states $\bar{u}_{ij}$ and unlimited antidiffusive fluxes f_{ij}. The components of these quantities w. r. t. the conserved unknowns ϱ and $\varrho\phi$ are denoted as $\bar{\varrho}_{ij}$, $\overline{(\varrho\phi)}_{ij}$, and f_{ij}^{ϱ}, $f_{ij}^{\varrho\phi}$, respectively.

In the first step of the sequential approach, the flux f_{ij}^{ϱ} is limited in exactly the same way as for scalar problems. Thus, we obtain limited bar states

$$\bar{\varrho}_{ij}^* = \bar{\varrho}_{ij} + \frac{f_{ij}^{\varrho,*}}{2d_{ij}}, \tag{3.50}$$

where $f_{ij}^{\varrho,*}$ is the limited counterpart of f_{ij}^{ϱ}. Next, we define the low order bar states of the derived quantity ϕ via

$$\bar{\phi}_{ij} := \frac{\overline{(\varrho\phi)}_{ij} + \overline{(\varrho\phi)}_{ji}}{\bar{\varrho}_{ij} + \bar{\varrho}_{ji}}. \tag{3.51}$$

Note that while there may be pairs of nodes for which $\bar{u}_{ij} \neq \bar{u}_{ji}$, we certainly have $\bar{\phi}_{ij} = \bar{\phi}_{ji}$ by definition. To ensure consistency, (3.50) should be accounted for in the process of limiting the bar states of unknowns depending on ϱ, i. e.,

$$\overline{(\varrho\phi)}_{ij}^{*} = \overline{(\varrho\phi)}_{ij} + \frac{f_{ij}^{\varrho\phi,*}}{2d_{ij}} = \bar{\varrho}_{ij}^{*}\bar{\phi}_{ij} + \frac{g_{ij}^{\varrho\phi,*}}{2d_{ij}}, \tag{3.52}$$

where $g_{ij}^{\varrho\phi,*}$ is a limited counterpart of

$$g_{ij}^{\varrho\phi} = f_{ij}^{\varrho\phi} + 2d_{ij}\left(\overline{(\varrho\phi)}_{ij} - \bar{\varrho}_{ij}^{*}\bar{\phi}_{ij}\right). \tag{3.53}$$

To control ϕ_i by limiting $f_{ij}^{\varrho\phi,*}$, we impose the numerical admissibility constraints

$$\bar{\varrho}_{ij}^{*}\phi_i^{\min} \leq \overline{(\varrho\phi)}_{ij}^{*} \leq \bar{\varrho}_{ij}^{*}\phi_i^{\max} \tag{3.54}$$

with generic bounds $\phi_i^{\min}, \phi_i^{\max}$ on ϕ_i to be specified below. To derive a limiter that enforces (3.54), we first observe that $g_{ij}^{\varrho\phi} = -g_{ji}^{\varrho\phi}$. This property is proved in the last formula on p. 13 of [Kuz20a]. Let us remark that in this reference, the last term appearing before the final identity should contain the limited fluxes for the variable ϱ instead of the unlimited ones. Inserting (3.53) into (3.52), converting (3.54) into flux constraints for $g_{ij}^{\varrho\phi,*}$ as in the scalar case, and imposing similar constraints on the companion flux $g_{ji}^{\varrho\phi,*} = -g_{ij}^{\varrho\phi,*}$, we define

$$g_{ij}^{\varrho\phi,*} = \begin{cases} \min\left\{g_{ij},\ 2d_{ij}\min\left\{\bar{\varrho}_{ij}^{*}(\phi_i^{\max} - \bar{\phi}_{ij}),\ \bar{\varrho}_{ji}^{*}(\bar{\phi}_{ij} - \phi_j^{\min})\right\}\right\} & \text{if } g_{ij} \geq 0, \\ \max\left\{g_{ij},\ 2d_{ij}\max\left\{\bar{\varrho}_{ij}^{*}(\phi_i^{\min} - \bar{\phi}_{ij}),\ \bar{\varrho}_{ji}^{*}(\bar{\phi}_{ij} - \phi_j^{\max})\right\}\right\} & \text{if } g_{ij} \leq 0. \end{cases} \tag{3.55}$$

This formula corrects a typo ($\min \leftrightarrow \max$, $i \leftrightarrow j$) in [Kuz20a, Eq. (85)] and exploits the fact that $\bar{\phi}_{ij} = \bar{\phi}_{ji}$. Reversing (3.53), we finally obtain

$$f_{ij}^{\varrho\phi,*} = g_{ij}^{\varrho\phi,*} - 2d_{ij}\left(\overline{(\varrho\phi)}_{ij} - \bar{\varrho}_{ij}^{*}\bar{\phi}_{ij}\right).$$

The corresponding bar state can then be computed from (3.52). The following property of the sequential limiter is important for the design of additional limiters.

Lemma 3.16 (Compatibility of sequential MCL with synchronized limiters) Suppose that the component $\bar{\varrho}_{ij}^{*}$ and the product variable components (3.52) of the limited bar state $\bar{u}_{ij}^{*}$ satisfy the numerical admissibility constraints

$$\varrho_i^{\min} \leq \bar{\varrho}_{ij}^{*} \leq \varrho_i^{\max}$$

and (3.54), respectively, for $i \in \{1, \ldots, N\},\ j \in \mathcal{N}_i \setminus \{i\}$. Assume further that the bounds for ϕ are defined in such a way that the following constraints hold

$$\bar{\varrho}_{ij}\phi_i^{\min} \leq \overline{(\varrho\phi)}_{ij} \leq \bar{\varrho}_{ij}\phi_i^{\max}, \qquad i \in \{1, \ldots, N\},\ j \in \mathcal{N}_i \setminus \{i\}. \tag{3.56}$$

Let $\alpha_{ij} = \alpha_{ji} \in [0,1]$ be a synchronized correction factor for

$$\bar{u}_{ij}^{**} := \bar{u}_{ij} + \frac{\alpha_{ij} f_{ij}^{*}}{2d_{ij}}. \tag{3.57}$$

Then the local maximum principles

$$\varrho_i^{\min} \le \bar{\varrho}_{ij}^{**} \le \varrho_i^{\max}, \qquad \bar{\varrho}_{ij}^{**}\phi_i^{\min} \le \overline{(\varrho\phi)}_{ij}^{**} \le \bar{\varrho}_{ij}^{**}\phi_i^{\max}, \qquad i \in \{1,\dots,N\},\ j \in \mathcal{N}_i \setminus \{i\}$$

hold for the components $\bar{\varrho}_{ij}^{**}$, and $\overline{(\varrho\phi)}_{ij}^{**}$ of the bar state $\bar{u}_{ij}^{**}$. ◊

Proof:
By design, the sequential limiter enforces the inequalities

$$\max\Big\{2d_{ij}(\varrho_i^{\min} - \bar{\varrho}_{ij}), 2d_{ij}(\bar{\varrho}_{ji} - \varrho_j^{\max})\Big\} \le f_{ij}^{\varrho,*} \le \min\Big\{2d_{ij}(\varrho_i^{\max} - \bar{\varrho}_{ij}), 2d_{ij}(\bar{\varrho}_{ji} - \varrho_j^{\min})\Big\}.$$

These estimates remain true if $f_{ij}^{\varrho,*}$ is replaced by $\alpha_{ij} f_{ij}^{\varrho,*}$ for $\alpha_{ij} \in [0,1]$. The situation is slightly more involved for derived unknowns. By (3.50), (3.52), and (3.54), we have

$$\begin{aligned} f_{ij}^{\varrho\phi,*} \le\ & \min\left\{2d_{ij}\left(\bar{\varrho}_{ij}^{*}\phi_i^{\max} - \overline{(\varrho\phi)}_{ij}\right),\ 2d_{ij}\left(\overline{(\varrho\phi)}_{ji} - \bar{\varrho}_{ji}^{*}\phi_j^{\min}\right)\right\} \\ =\ & \min\left\{2d_{ij}\left(\bar{\varrho}_{ij}\phi_i^{\max} - \overline{(\varrho\phi)}_{ij}\right) + f_{ij}^{\varrho,*}\phi_i^{\max},\ 2d_{ij}\left(\overline{(\varrho\phi)}_{ji} - \bar{\varrho}_{ji}\phi_j^{\min}\right) - f_{ji}^{\varrho,*}\phi_j^{\min}\right\}. \end{aligned}$$

Multiplying this inequality by $\alpha_{ij} \in [0,1]$ and invoking (3.56) yields

$$\begin{aligned} \alpha_{ij} f_{ij}^{\varrho\phi,*} \le\ & \min\Big\{2d_{ij}\left(\bar{\varrho}_{ij}\phi_i^{\max} - \overline{(\varrho\phi)}_{ij}\right)\alpha_{ij} + \alpha_{ij} f_{ij}^{\varrho,*}\phi_i^{\max}, \\ & \qquad 2d_{ij}\left(\overline{(\varrho\phi)}_{ji} - \bar{\varrho}_{ji}\phi_j^{\min}\right)\alpha_{ij} - \alpha_{ij} f_{ji}^{\varrho,*}\phi_j^{\min}\Big\} \\ \le\ & \min\left\{2d_{ij}\left(\bar{\varrho}_{ij}^{**}\phi_i^{\max} - \overline{(\varrho\phi)}_{ij}\right),\ 2d_{ij}\left(\overline{(\varrho\phi)}_{ji} - \bar{\varrho}_{ji}^{**}\phi_j^{\min}\right)\right\}. \end{aligned} \tag{3.58}$$

In a similar way, we find that $\alpha_{ij} f_{ij}^{\varrho\phi,*}$ is bounded below as follows

$$\max\left\{2d_{ij}\left(\bar{\varrho}_{ij}^{**}\phi_i^{\min} - \overline{(\varrho\phi)}_{ij}\right),\ 2d_{ij}\left(\overline{(\varrho\phi)}_{ji} - \bar{\varrho}_{ji}^{**}\phi_j^{\max}\right)\right\} \le \alpha_{ij} f_{ij}^{\varrho\phi,*}. \tag{3.59}$$

Inserting (3.58) and (3.59) into (3.57) yields the claimed inequalities. □

We conclude the presentation of the sequential approach by specifying the local bounds for flux limiting based on numerical admissibility conditions. Recall that in the scalar case the choice of bounds (3.44) implies that all low order bar states $\bar{u}_{ij}$ $j \in \mathcal{N}_i \setminus \{i\}$ are contained in $[u_i^{\min}, u_i^{\max}]$ for $i \in \{1,\dots,N\}$ by the mean value theorem. For systems, we need to ensure that the low order bar states are admissible. Otherwise the limiter may not be able to enforce the imposed constraints. For the main unknown ϱ we use

$$\varrho_i^{\min} := \min_{j \in \mathcal{N}_i \setminus \{i\}} \bar{\varrho}_{ij}, \qquad \varrho_i^{\max} := \max_{j \in \mathcal{N}_i \setminus \{i\}} \bar{\varrho}_{ij}. \tag{3.60}$$

We choose local bounds for derived quantities in a way that ensures the validity of condition (3.56) in Lemma 3.16. Since in this chapter our numerical experiments for systems are restricted to the 1D case in which $\bar{\phi}_{ij} = \overline{(\varrho\phi)}_{ij}/\bar{\varrho}_{ij}$ for all $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$, we may define the local bounds for ϕ as follows

$$\phi_i^{\min} := \min_{j\in\mathcal{N}_i\setminus\{i\}} \bar{\phi}_{ij}, \qquad \phi_i^{\max} := \max_{j\in\mathcal{N}_i\setminus\{i\}} \bar{\phi}_{ij}. \tag{3.61}$$

To preserve the validity of (3.56) in the multidimensional case, a modification of (3.61) for boundary nodes is required. Definition (3.60) differs from the one proposed in [Kuz20a, Sec. 5.1], where the main unknown is constrained exactly as in the scalar case. The bounds of numerical admissibility conditions can be further adjusted, e. g., using smoothness indicators, alternative data sets and/or the principle of linearity preservation. Since all approximations obtained with (3.60)–(3.61) were of satisfactory quality in our 1D numerical experiments, we did not attempt to construct better bounds.

3.3.5 Invariant domain preservation

Thus far, we have only enforced numerical admissibility conditions through limiting. In general, there is no guarantee, however, that the resulting approximations will remain in an invariant domain of the given system. On the other hand, the low order method produces approximations that are provably IDP, owing to the fact that the low order bar states belong to convex invariant sets. This argument allows us to perform invariant domain fixes for our flux-corrected scheme if necessary. In this section, we discuss the corresponding modifications for all problems under consideration. First, we show that no IDP corrections are needed in the case of scalar conservation laws.

Lemma 3.17 (Invariant domain preservation for scalar conservation laws)
Let $\tilde{u}_i,\ i \in \{1,\dots,N\}$ be the result of a forward Euler update (3.42) for the MCL scheme using formula (3.47) to constrain $\bar{u}_{ij}^*$. Suppose that the time step Δt satisfies the CFL-like condition (3.43). Then $\tilde{u}_i \in \mathcal{A}_i \subseteq [u^{\min}, u^{\max}]$, where

$$\begin{aligned} u^{\min} &:= \min\Big\{ \min_{j\in\{1,\dots,N\}} u_h(\boldsymbol{x}_j,0),\ \inf_{t\in\mathbb{R}_+} \min_{\boldsymbol{x}_j\in\partial\Omega} \hat{u}(\boldsymbol{x}_j,t)\Big\}, \\ u^{\max} &:= \max\Big\{ \max_{j\in\{1,\dots,N\}} u_h(\boldsymbol{x}_j,0),\ \sup_{t\in\mathbb{R}_+} \max_{\boldsymbol{x}_j\in\partial\Omega} \hat{u}(\boldsymbol{x}_j,t)\Big\} \end{aligned}$$

are the global bounds to be preserved (cf. Theorem 2.15). ◊

Proof:
By construction, the limited bar states satisfy $u_i^{\min} \le \bar{u}_{ij}^* \le u_i^{\max}$ for all $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. Moreover, $\bar{u}_i^k \in [u^{\min}, u^{\max}]$. From (3.42) and (3.43), we deduce that

$$\tilde{u}_i = \Big[1 - \frac{\Delta t}{m_i}\Big(\sum_{j\in\mathcal{N}_i\setminus\{i\}} 2d_{ij} + \sum_{\Gamma_k\in\mathcal{F}_i} 2d_i^k\Big)\Big]u_i + \frac{\Delta t}{m_i}\Big(\sum_{j\in\mathcal{N}_i\setminus\{i\}} 2d_{ij}\bar{u}_{ij}^* + \sum_{\Gamma_k\in\mathcal{F}_i} 2d_i^k\,\bar{u}_i^k\Big)$$

$$\leq \Big[1 - \frac{\Delta t}{m_i}\Big(\sum_{j\in\mathcal{N}_i\backslash\{i\}} 2d_{ij} + \sum_{\Gamma_k\in\mathcal{F}_i} 2d_i^k\Big)\Big]u^{\max} + \frac{\Delta t}{m_i}\Big(\sum_{j\in\mathcal{N}_i\backslash\{i\}} 2d_{ij}u^{\max} + \sum_{\Gamma_k\in\mathcal{F}_i} 2d_i^k\, u^{\max}\Big)$$
$$= u^{\max}, \qquad i \in \{1,\dots,N\}. \tag{3.62}$$

The lower bound $u^{\min}$ for $\tilde{u}_i$ is derived analogously. □

Similarly, we observe that no IDP fix is needed for the shallow water equations because the water height remains nonnegative under the numerical admissibility conditions.

Lemma 3.18 (Invariant domain preservation for the SWE)
Consider the SWE with flat topography (see Section 2.2.3). Suppose that the water heights of the discrete initial condition and external Riemann data are nonnegative in Ω and on $\partial\Omega\times[0,\infty)$, respectively. Furthermore, let Δt satisfy the CFL-like condition (3.43). Then the water heights $\tilde{h}_i$ produced by forward Euler stages (3.42) of the sequential MCL approximation remain nonnegative for all $i\in\{1,\dots,N\}$. ◇

Proof:
Let $\bar{h}_{ij}^*$ be the water height component of $\bar{u}_{ij}^*$. Since the low order bar states $\bar{u}_{ij}$ and $\bar{u}_i^k$ are in the invariant set $\mathcal{A}_i\subseteq\mathcal{A}^{\max}$, their water heights are nonnegative. Thus, by construction of the water height limiter, we have $0\leq h_i^{\min}\leq\bar{h}_{ij}^*$. An estimate similar to (3.62) now shows that the water height computed from (3.42) remains nonnegative.

Contrary to scalar problems and the SWE, the Euler equations require at least one IDP fix. Recall that the largest (physical) admissible set of this system (cf. Section 2.2.2) is $\mathcal{A}^{\max}=\{(\rho,\rho\boldsymbol{v}^\mathsf{T},\rho E)\in\mathbb{R}^{d+2}:\rho\geq 0,\ e\geq 0\}$. While positivity of ρ can be shown using the arguments of Lemma 3.18, nonnegativity of the internal energy e (cf. Section 2.1.1) is not guaranteed by the scheme presented thus far. In fact, the MCL approximation applied to many classical benchmarks of compressible flow problems does produce negative internal energies in some nodes, which causes simulations to break down. On the other hand, the low order approximation is provably IDP under the usual time step restriction (3.43). Thus, nonnegativity of e can be enforced by further reducing the magnitude of the antidiffusive fluxes that violate physical admissibility conditions for the bar states. In the context of FEM-FCT schemes for the Euler equations, nonnegative local bounds were imposed on the pressure $p=(\gamma-1)\rho e$ in [Loh16] using a Zalesak-like limiter. For MCL schemes, we use the Lipschitz-continuous IDP fix proposed in [Kuz20a, Sec. 5.1]. We summarize it here for completeness.

Applying the sequential MCL limiter based on numerical admissibility conditions, we obtain the prelimited bar states

$$\bar{u}_{ij}^* = \bar{u}_{ij} + \frac{f_{ij}^*}{2d_{ij}} =: \bar{u}_{ij} + \frac{1}{2d_{ij}}\left[f_{ij}^{\rho,*},(\boldsymbol{f}_{ij}^{\rho\boldsymbol{v},*})^\mathsf{T},f_{ij}^{\rho E,*}\right]^\mathsf{T}.$$

If the corresponding internal energy $\bar{e}_{ij}^*$ becomes negative, then $\bar{u}_{ij}^* \notin \mathcal{A}^{\max}$ and an IDP fix is required. Let $\alpha_{ij} = \alpha_{ji} \in [0,1]$ be synchronized correction factors such that

$$\bar{u}_{ij}^{**} = \bar{u}_{ij} + \frac{\alpha_{ij} f_{ij}^*}{2d_{ij}} \in \mathcal{A}^{\max}, \qquad i \in \{1,\dots,N\},\ j \in \mathcal{N}_i \setminus \{i\}. \tag{3.63}$$

Then the IDP property can be shown as for the low order method, which corresponds to the choice $\alpha_{ij} = 0$ for all $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. Kuzmin [Kuz20a, Sec. 5.1] derived two formulas that ensure positivity preservation for the internal energy of the bar state $\bar{u}_{ij}^{**}$ with nontrivial α_{ij}. Similarly to the FCT limiter developed in [Loh16], his estimates linearize quadratic inequality constraints for $\alpha_{ij} \in [0,1]$ using the property $\alpha_{ij}^2 \le \alpha_{ij}$. A potential disadvantage of synchronized limiting in (3.63) is the lack of continuous dependence on the data. The IDP fix proposed in [Kuz20a, Eq. (92)] adapts the formula for α_{ij} to ensure Lipschitz-continuity of $\alpha_{ij} f_{ij}^*$ and thus preserve a desirable property of f_{ij}^*. The resulting values of α_{ij} may be smaller than necessary to satisfy $\bar{u}_{ij}^{**} \in \mathcal{A}^{\max}$ but intentional underestimation of α_{ij} does not significantly degrade the overall accuracy compared to the non-Lipschitz version [Kuz20a, Eq. (90)] of the energy fix.

Let $\bar{w}_{ij} = (\bar{w}_{ij}^{\rho}, (\bar{\boldsymbol{w}}_{ij}^{\rho\boldsymbol{v}})^{\mathsf{T}}, \bar{w}_{ij}^{\rho E})^{\mathsf{T}} = 2d_{ij}\bar{u}_{ij}$ denote the scaled low order bar states. Then the Lipschitz-continuous version [Kuz20a, Eq. (92)] of the IDP fix reads

$$\alpha_{ij} = \begin{cases} \frac{Q_{ij}}{R_{ij}} & \text{if } R_{ij} > Q_{ij}, \\ 1 & \text{otherwise,} \end{cases} \tag{3.64}$$

where

$$\begin{aligned} Q_{ij} = Q_{ji} = {} & \min\left\{ \bar{w}_{ij}^{\rho}\bar{w}_{ij}^{\rho E} - \frac{1}{2}|\bar{\boldsymbol{w}}_{ij}^{\rho\boldsymbol{v}}|^2,\ \bar{w}_{ji}^{\rho}\bar{w}_{ji}^{\rho E} - \frac{1}{2}|\bar{\boldsymbol{w}}_{ji}^{\rho\boldsymbol{v}}|^2 \right\}, \\ R_{ij} = R_{ji} = {} & \max\{|\bar{\boldsymbol{w}}_{ij}^{\rho\boldsymbol{v}}|,\ |\bar{\boldsymbol{w}}_{ji}^{\rho\boldsymbol{v}}|\}|\boldsymbol{f}_{ij}^{\rho\boldsymbol{v},*}| + \max\{|\bar{w}_{ij}^{\rho}|,\ |\bar{w}_{ji}^{\rho}|\}|f_{ij}^{\rho E,*}| \\ & + \max\{|\bar{w}_{ij}^{\rho E}|,\ |\bar{w}_{ji}^{\rho E}|\}|f_{ij}^{\rho,*}| + \max\left\{0,\ \frac{1}{2}|\boldsymbol{f}_{ij}^{\rho\boldsymbol{v},*}|^2 - f_{ij}^{\rho,*} f_{ij}^{\rho E,*}\right\} \end{aligned}$$

for $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. The IDP property can be verified as follows.

Lemma 3.19 (Invariant domain preservation for the Euler equations)
Consider the forward Euler stage (3.42) of the sequential MCL scheme for the compressible Euler equations. Replace the prelimited bar states $\bar{u}_{ij}^*$ with their IDP counterparts (3.63) that use the correction factors (3.64). Let $u_h(\boldsymbol{x},0) \in \mathcal{A}^{\max}$ for all $\boldsymbol{x} \in \Omega$ and $\hat{u}(\boldsymbol{x},t) \in \mathcal{A}^{\max}$ for all $(\boldsymbol{x},t) \in \partial\Omega \times [0,\infty)$. If the CFL-like condition (3.43) is satisfied, then $\tilde{u}_i \in \mathcal{A}^{\max}$ for all $i \in \{1,\dots,N\}$. ◇

Proof:
Positivity preservation for the density is shown as in the proof of Lemma 3.18. By definition of α_{ij} in (3.64), we have $\bar{u}_{ij}^{**} \in \mathcal{A}^{\max}$. In particular, the internal energy of the

bar state $\bar{u}_{ij}^{**}$ is guaranteed to be nonnegative [Kuz20a, Sec. 5.1]. Moreover, $\bar{u}_i^k \in \mathcal{A}^{\max}$. Owing to (3.43), the result $\tilde{u}_i$ for each forward Euler stage is again a convex combination of states belonging to $\mathcal{A}^{\max} = \{(\rho, \rho\boldsymbol{v}^\mathsf{T}, \rho E) \in \mathbb{R}^{d+2} : \rho \geq 0,\ e \geq 0\}$. The claim follows from convexity of $\mathcal{A}^{\max}$. □

Remark 3.20
In addition to the nonnegativity of density and internal energy, Guermond et al. [Gue18a] enforce a minimum principle for the logarithmic specific entropy s (see Section 2.2.2). We believe that a limiter similar to the IDP fix for the internal energy can be constructed to satisfy this constraint as well. The definition of the convex admissible set $\mathcal{A}^{\max}$ needs to be modified accordingly. If no closed-form expression can be derived for the IDP correction factor α_{ij} as in [Kuz20a], a line search will need to be performed as in [Gue18a]. We do not recommend imposition of tight local bounds on nonlinear derived quantities such as e and s. The corresponding constraints may be violated even if the conserved unknowns are linear functions and all bar states belong to $\mathcal{A}^{\max}$. Therefore, it is usually impossible to achieve second order accuracy without using smoothness indicators [Kho94, Dob18, Gue18a].

In our experience, a well-tuned combination of numerical admissibility conditions with IDP fixes that enforce global bounds is the best limiting strategy for hyperbolic systems [Haj20c]. The bounds of local maximum principles for conserved unknowns and ratios ϕ thereof may be relaxed using smoothness sensors. However, all physical constraints built into the definition of $\mathcal{A}^{\max}$ are nonnegotiable. They must be enforced even in smooth regions, where all numerical admissibility criteria are satisfied. ◇

3.3.6 Semi-discrete entropy fix

For certain hyperbolic problems, even approximations obtained with bound-preserving methods may converge to weak solutions that violate entropy conditions. The MCL scheme presented thus far is no exception and may require an *entropy fix*.

A variety of entropy conservative/stable schemes for hyperbolic problems can be found in the literature. Some of them employ numerical fluxes that ensure entropy stability of spatial semi-discretizations [Har83b, Fjo11, Che17, Win17, Wu21]. Other methods are designed to satisfy discrete entropy (in-)equalities, see for instance [Hen21]. Such strategies are promising because fully discrete entropy stability is a prerequisite to proving Lax–Wendroff-type theorems, which state that converging bounded sequences of approximations converge to entropy solutions [Krö94, Thms. 3.14 and 4.11].

Let us attempt to summarize some important results on the topic of entropy-aware numerical methods in a few sentences. Jiang and Shu [Jia94] derived discrete cell entropy inequalities w. r. t. the square entropy for discontinuous Galerkin discretizations of scalar conservation laws. In the context of continuous finite element methods, the corresponding inequality is satisfied as an identity [Tad87, Kuz20d]. Gassner [Gas13]

derived a criterion for the artificial viscosity of numerical fluxes in DG discretizations to ensure entropy/energy stability in the semi-discrete case. Kuzmin and Quezada de Luna [Kuz20c] combine bound-preserving flux correction schemes with a limiter that enforces nodal entropy inequalities. Generalizations of this approach to higher order finite element spaces and strategies that achieve fully discrete entropy stability can be found in [Kuz20d] and in our paper [Kuz22a], respectively. The method proposed by Berthon et al. [Ber20] performs a somewhat similar fully discrete entropy fix by adjusting the amount of artificial viscosity in Godunov-type methods.

We conclude our summary of the literature by commenting on the necessity to employ algorithms that are entropy stable in the fully discrete case. For practical purposes, semi-discrete entropy fixes are usually sufficient as long as the target discretization is based on entropic fluxes (see [Kuz22a] for details). Further corrections aimed at achieving fully discrete entropy stability have marginal effects on the quality of approximations to discontinuous weak solutions but may degrade the rates of convergence to smooth ones. In fact, one-dimensional three-point schemes satisfying local fully discrete entropy inequalities cannot be second order accurate in the explicit case [Sch85]. Implicit flux-corrected schemes can circumvent this order barrier [Kuz22a, Sec. 6.2] but are, of course, computationally more expensive. Therefore, we chose to enforce only semi-discrete entropy inequalities in this thesis.

Let us now discuss how to ensure entropy stability by modifying the MCL scheme

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij}(\bar{u}_{ij}^{**} - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} 2d_i^k \, (\bar{u}_i^k - u_i), \qquad i \in \{1, \dots, N\}, \tag{3.65}$$

which, at this stage, guarantees preservation of local bounds and invariant domains. For the purposes of entropy limiting, we depart from our previous strategy of writing the scheme in terms of bar states and reformulate (3.65) as

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} \Big[d_{ij}(u_j - u_i) - (\mathbf{f}_j - \mathbf{f}_i)\, \mathbf{c}_{ij} + \alpha_{ij} f_{ij}^* \Big] \tag{3.66a}$$

$$+ \sum_{\Gamma_k \in \mathcal{F}_i} \frac{b_i^k}{2} \Big[\lambda_{\boldsymbol{n}_k}(u_i, \hat{u}_i^k) \, (\hat{u}_i^k - u_i) - (\mathbf{f}(\hat{u}_i^k) - \mathbf{f}_i)\, \boldsymbol{n}_k \Big], \tag{3.66b}$$

where we invoked (3.63), (3.28), and (3.29). To keep the following presentation concise, we exploit the fact that the volume terms on the right of (3.66a) and the boundary fluxes in (3.66b) have similar structure. First, we introduce the set $\mathcal{S} := \{1, \dots, N\} \cup \mathcal{B}$, where $\mathcal{B} := \{N+1, \dots, \hat{N}\}$ contains the indices of all *ghost nodes* $\{\hat{\boldsymbol{x}}_j\}_{j=N+1}^{\hat{N}}$ corresponding to boundary vertices. Here we use the convention that for each vertex $\boldsymbol{x}_i$, there is exactly one ghost node per boundary segment $\Gamma_k \in \mathcal{F}_i$. Thus, each pair of indices $(i, j) \in \{1, \dots, N\} \times \mathcal{B}$ can be associated with at most one boundary segment Γ_k. We exploit this observation to define an index mapping $\omega : \{1, \dots, N\} \times \mathcal{B} \to \mathbb{N}$ such that

$k = \omega(i,j)$ is the index of a boundary segment Γ_k and the fictitious degree of freedom u_j represents the external state $\hat{u}_i^k$ for the approximate Riemann solver.

Using the above notation, we define volume and boundary term fluxes as follows

$$g_{ij} := \begin{cases} d_{ij}(u_j - u_i) + \alpha_{ij} f_{ij}^* & \text{if } j \in \mathcal{N}_i \setminus \{i\}, \\ \frac{1}{2}\, b_i^k\, \lambda_{\boldsymbol{n}_k}(u_i, \hat{u}_i^k)(\hat{u}_i^k - u_i) & \text{if } j \in \mathcal{B} \text{ and } \omega(i,j) = k, \\ 0 & \text{otherwise} \end{cases}$$

for $i \in \{1, \dots, N\},\ j \in \mathcal{S}$. A similar generalization yields

$$\mathbf{f}_j := \begin{cases} \mathbf{f}_j & \text{if } j \in \{1, \dots, N\}, \\ \mathbf{f}(\hat{u}_i^k) & \text{if } j \in \mathcal{B} \text{ and } \omega(i,j) = k, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{c}_{ij} := \begin{cases} \mathbf{c}_{ij} & \text{if } j \in \{1, \dots, N\}, \\ \frac{1}{2} b_i^k\, \boldsymbol{n}_k & \text{if } j \in \mathcal{B} \text{ and } \omega(i,j) = k, \\ 0 & \text{otherwise} \end{cases}$$

for $i \in \{1, \dots, N\},\ j \in \mathcal{S}$. Using these conventions, we may write (3.66) as

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{S}} \left[g_{ij} - (\mathbf{f}_j - \mathbf{f}_i)\, \mathbf{c}_{ij} \right] = \sum_{j \in \mathcal{S}} \left[g_{ij} - (\mathbf{f}_j + \mathbf{f}_i)\, \mathbf{c}_{ij} \right] + 2\mathbf{f}_i \sum_{j \in \mathcal{S}} \mathbf{c}_{ij}. \tag{3.67}$$

Similar notation is used in [Gue19, Sec. 4.3] to write volume terms and numerical fluxes of bound-preserving DG discretizations in a unified compact form.

We are now in a position to formulate entropy stability conditions for (3.67). Let $(\eta, \boldsymbol{q})$ be an entropy pair of the given hyperbolic system, $v(u) = \eta'(u)$ the corresponding entropy variable, and $\boldsymbol{\psi}(u) = v(u)^\mathsf{T} \mathbf{f}(u) - \boldsymbol{q}(u)$ the associated entropy potential (see Definition 2.3). Evaluations of these quantities in the nodal states of the discrete solution u_h are denoted using a subscript $i \in \{1, \dots, \hat{N}\}$. For instance, η_i is the shorthand notation for $\eta(u_i)$. A scheme of the form (3.67) is entropy stable in the sense of Tadmor [Tad87, Tad03, Kuz20c] if the following inequalities hold

$$\frac{(v_i - v_j)^\mathsf{T}}{2} \left[g_{ij} - (\mathbf{f}_j + \mathbf{f}_i)\, \mathbf{c}_{ij} \right] \le (\boldsymbol{\psi}_j - \boldsymbol{\psi}_i) \cdot \mathbf{c}_{ij}, \qquad i \in \{1, \dots, N\},\ j \in \mathcal{S}. \tag{3.68}$$

Our intention is to enforce (3.68) through limiting. Before presenting such limiter-based entropy fixes, let us discuss the implications of Tadmor's condition for scheme (3.67).

Theorem 3.21 (Local semi-discrete entropy inequality, Kuz20c, Thm. 1)
Suppose that condition (3.68) holds for all $i \in \{1, \dots, N\}$ and $j \in \mathcal{S}$. Then a solution to (3.67) satisfies the semi-discrete entropy inequalities

$$m_i \frac{\mathrm{d}\eta_i}{\mathrm{d}t} \le \sum_{j \in \mathcal{S}} \left(G_{ij} - (\boldsymbol{q}_j - \boldsymbol{q}_i) \cdot \mathbf{c}_{ij} \right) \tag{3.69}$$

for all $i \in \{1, \ldots, N\}$. The numerical fluxes G_{ij} that appear in (3.69) are given by

$$G_{ij} := \frac{(v_i + v_j)^\mathsf{T}}{2} g_{ij} + \frac{(v_i - v_j)^\mathsf{T}}{2} (\mathbf{f}_i - \mathbf{f}_j)\, \mathbf{c}_{ij}, \qquad i \in \{1, \ldots, N\},\ j \in \mathcal{S}. \qquad \diamond$$

Proof:
Following [Tad03, Kuz22a], we multiply (3.67) by the nodal entropy variable v_i, which is split into its symmetric part $\frac{1}{2}(v_i + v_j)$ and the antisymmetric remainder $\frac{1}{2}(v_i - v_j)$. The rest of the proof is as in [Fjo11]. We refer to [Kuz22a, Sec. 4.1] for a version that is directly applicable to the AFC scheme (3.67). □

Remark 3.22
The fluxes G_{ij} modify $\sum_{j \in \mathcal{S}} (\boldsymbol{q}_j - \boldsymbol{q}_i) \cdot \mathbf{c}_{ij} \approx m_i (\nabla \cdot \boldsymbol{q})_i$ in the same manner as the fluxes $g_{ij} = d_{ij}(u_j - u_i) + \alpha_{ij} f_{ij}^*$ modify the centered approximation to $(\nabla \cdot \mathbf{f})_i$ in the flux-corrected version (3.41) of the target scheme (3.19). $\diamond$

The implications of Theorem 3.21 are most striking in settings in which boundary terms vanish or have no influence on the finite element approximation. To discuss this case, we introduce the following concept.

Definition 3.23 (Boundary indifference for AFC schemes)
We say that the flux-corrected finite element discretization (3.41) of (3.1) is boundary indifferent for a given $u_h(t) \in \mathrm{V}_h^m$ if $\mathbf{c}_{ij} = -\mathbf{c}_{ji}$ or $u_i(t) = u_j(t)$ for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$, and, furthermore, $\hat{u}_i^k(t) = u_i(t)$ for all $i \in \{1, \ldots, N\},\ \Gamma_k \in \mathcal{F}_i$. $\diamond$

Example 3.24
For (initial-)boundary value problems with periodic Riemann data $\hat{u}$, continuous finite element discretizations of (3.1) are boundary indifferent regardless of u_h. In general, periodic boundary conditions are formulated using identity mappings for certain nodes on $\partial\Omega$. If all boundaries of the domain Ω are periodic, then $\mathcal{F}_{\partial\Omega} = \emptyset$ and therefore the sum over $\Gamma_k \in \mathcal{F}_i$ vanishes in (3.41). The matrix $\mathbf{C}$ becomes fully skew symmetric and all of its diagonal entries are zero in the fully periodic case. The identities $\mathbf{c}_{ij} = -\mathbf{c}_{ji}$ and $G_{ij} = -G_{ji}$ are also satisfied for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ on nonperiodic 1D meshes consisting of more than one element. There are many examples in which u_h coincides with the discretized Riemann data $\hat{u}_h$ on (subsets of) $\partial\Omega$. For instance, this must be the case for a supersonic outlet by definition. If $u_h = \hat{u}_h$ on the remaining boundaries of $\partial\Omega$ as well, then the approximation is boundary indifferent. $\diamond$

Corollary 3.25 (Global semi-discrete entropy inequality)
Let the assumptions of Theorem 3.21 be fulfilled. Additionally, assume that the spatial semi-discretization (3.41) is boundary indifferent w. r. t. $u_h(t)$ for all $t \geq 0$. Then $u_h(t)$ satisfies the global semi-discrete entropy inequality

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \Big(\sum_{i=1}^N \eta_i\, \varphi_i \Big)\, \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \Big(\sum_{i=1}^N \boldsymbol{q}_i\, \varphi_i \Big) \cdot \boldsymbol{n}\, \mathrm{d}s \leq 0. \tag{3.70}$$

$\diamond$

Proof:
Following [Kuz20c, Sec. 3], we sum (3.69) over $i \in \{1, \dots, N\}$ using (3.21), which yields

$$\sum_{i=1}^{N} m_i \frac{\mathrm{d}\eta_i}{\mathrm{d}t} \le \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i \setminus \{i\}} \Big(G_{ij} - (\boldsymbol{q}_j - \boldsymbol{q}_i) \cdot \mathbf{c}_{ij} \Big) = \sum_{i=1}^{N} \Big[\sum_{\substack{j \in \mathcal{N}_i \\ i<j}} G_{ij} + \sum_{\substack{j \in \mathcal{N}_i \\ i>j}} G_{ij} - \sum_{j=1}^{N} \boldsymbol{q}_j \cdot \mathbf{c}_{ij} \Big],$$

where all boundary terms cancel by assumption. Furthermore, boundary indifference implies that the numerical fluxes G_{ij} are skew symmetric. Thus, an application of the divergence theorem concludes the proof. □

Remark 3.26
Integration of the strong form entropy inequality $\frac{\partial \eta(u)}{\partial t} + \nabla \cdot \boldsymbol{q}(u) \le 0$ over Ω yields

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \eta(u) \, \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \boldsymbol{q}(u) \cdot \boldsymbol{n} \, \mathrm{d}s \le 0. \tag{3.71}$$

Hence, (3.70) is a semi-discrete analogue of the inequality constraint (3.71) for global entropy production in Ω. In the case of fully periodic domains, all boundary integrals vanish and (3.71) reduces to $\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \eta(u) \, \mathrm{d}\boldsymbol{x} \le 0$, while (3.70) reduces to [Kuz20c]

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \Big(\sum_{i=1}^{N} \eta_i \, \varphi_i \Big) \, \mathrm{d}\boldsymbol{x} \le 0. \qquad \diamond$$

We have thus established that Tadmor's condition is indeed a valuable criterion for proving semi-discrete entropy inequalities. The question of how to enforce (3.68) in the context of flux correction schemes remains. Clearly, FCT-type algorithms are not suited for entropy limiting based on the semi-discrete inequality constraint (3.68) for $g_{ij} = d_{ij}(u_j - u_i) + \alpha_{ij} f_{ij}^*$. An alternative approach for enforcing entropy inequalities in fully discrete FCT-type difference schemes was proposed by Kivva [Kiv22]. In his fully discrete method, a global optimization problem is solved in every time step to find correction factors that yield bound-preserving and entropy-stable approximations.

Instead of enforcing fully discrete entropy stability using optimization-based FCT, we exploit the semi-discrete nature of the bound-preserving MCL scheme (3.66) and employ a limiter that enforces Tadmor's entropy stability condition (3.68) in addition to discrete maximum principles. First, we notice that removal of the term $\alpha_{ij} f_{ij}^*$ in (3.66) reproduces the low order method (3.26) again. Entropy stability for $\alpha_{ij} = 0$ follows from arguments used by Chen and Shu [Che17, Sec. 3.5] to prove it for DG schemes that use Godunov-type approximate Riemann solvers such as the local Lax–Friedrichs flux.

Lemma 3.27 (Entropy stability of the low order method, Che17 Sec. 3.3)
Let $(\eta, \boldsymbol{q})$ be an entropy pair for (3.1). Assume that the assumptions of Lemma 3.12 are satisfied. In addition, for arbitrary space directions $\boldsymbol{n} \in \mathrm{S}_1^{d-1}$, let the one-dimensional entropy inequalities

$$\frac{\partial \eta(u)}{\partial t} + (\nabla \cdot \boldsymbol{n}) \, \boldsymbol{q}(u) \cdot \boldsymbol{n} \le 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}_+$$

hold for the unique admissible solution u. If $\eta \in \mathrm{C}^2(\overline{\mathcal{A}^{\max}})$, then Tadmor's entropy stability condition (3.68) is satisfied for the numerical fluxes $g_{ij} = d_{ij}(u_j - u_i)$. ◇

Proof:
Adapting the proof techniques of Chen and Shu [Che17, Thm. 3.6 and Cor. 3.2] to the AFC version (3.26) of the local Lax–Friedrichs method, the validity of the claim can be readily established using similar arguments. □

Let us now discuss a way to enforce Tadmor's entropy stability condition (3.68). Since it holds for $g_{ij} = d_{ij}(u_j - u_i)$ by Lemma 3.27, it also holds for fluxes of the form $g_{ij} = d_{ij}(u_j - u_i) + \beta_{ij}\alpha_{ij}f_{ij}^*$ if $\beta_{ij} = \beta_{ji} \in [0,1]$ is sufficiently small. The corresponding criterion for calculation of the entropy correction factors β_{ij} for the final limited antidiffusive fluxes $\beta_{ij}\alpha_{ij}f_{ij}^*$ of the MCL scheme is given by [Kuz20c]

$$\frac{(v_i - v_j)^\mathsf{T}}{2}\Big[d_{ij}(u_j - u_i) + \beta_{ij}\alpha_{ij}f_{ij}^* - (\mathbf{f}_j + \mathbf{f}_i)\,\mathbf{c}_{ij}\Big] \le (\boldsymbol{\psi}_j - \boldsymbol{\psi}_i)\cdot\mathbf{c}_{ij} \tag{3.72}$$

for $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$, or equivalently,

$$\frac{\beta_{ij}}{2}R_{ij} \le Q_{ij} - \frac{d_{ij}}{2}P_{ij}, \qquad i \in \{1,\dots,N\},\ j \in \mathcal{N}_i \setminus \{i\},$$

where

$$P_{ij} := (v_i - v_j)^\mathsf{T}(u_j - u_i) = P_{ji}, \qquad R_{ij} := \alpha_{ij}(v_i - v_j)^\mathsf{T}f_{ij}^* = R_{ji}, \tag{3.73a}$$

$$Q_{ij} := \Big[(\boldsymbol{\psi}_j - \boldsymbol{\psi}_i) + \frac{1}{2}(v_i - v_j)^\mathsf{T}(\mathbf{f}_j + \mathbf{f}_i)\Big]\cdot\mathbf{c}_{ij} \tag{3.73b}$$

for $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. In addition to (3.72), the correction factors of the semi-discrete entropy fix must satisfy the symmetry condition $\beta_{ij} = \beta_{ji}$. Under these linear constraints, the optimal value of β_{ij} is given by [Kuz20c, Eq. (43)]

$$\beta_{ij} = \begin{cases} \dfrac{2\min\{Q_{ij},Q_{ji}\} - d_{ij}P_{ij}}{R_{ij}} & \text{if } R_{ij} > 2\min\{Q_{ij},Q_{ji}\} - d_{ij}P_{ij}, \\ 1 & \text{otherwise.} \end{cases} \tag{3.74}$$

Note that $2\min\{Q_{ij},Q_{ji}\} - d_{ij}P_{ij}$ is nonnegative for all $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ by Lemma 3.27. We close this section by pointing out that practical calculation of the correction factors β_{ij} is further discussed in Remark 6 of our paper [Kuz22a].

3.4 Numerical examples

We now illustrate the performance of bound-preserving and entropy-stable schemes for a variety of test problems. First we consider two nonlinear scalar conservation laws before moving on to the Euler equations of gas dynamics.

The strategies under investigation include the algebraic local Lax–Friedrichs scheme, i. e., the low order method (LOW), the bound-preserving monolithic convex limiting approach without entropy fix (MCL), as well as the MCL algorithm enhanced by the semi-discrete entropy limiter (MCL-SDE). If bound-preserving flux correction is disabled and only semi-discrete entropy stability is enforced, the scheme is referred to as SDE.

In most examples, we employ a uniform one-dimensional mesh consisting of 128 elements. Unless stated otherwise, the SSP2 RK method is used for temporal discretization in combination with adaptive time stepping as discussed in Section 3.3.4.1. The CFL parameter $\nu = 1$ is used for scalar equations whereas, by default, we set $\nu = 0.5$ for the Euler equations.

For simplicity, we interpolate u_0 in V_h^m to obtain discrete initial data (although this approach is, strictly speaking, not possible if the initial profiles are discontinuous). Compared to interpolations, projections have the benefit of being conservative in the case of exact integration. The lumped L^2 projection is a suitable approach to use in real-life applications because the resulting loss of accuracy is usually dominated by the discretization error. The consistent L^2 projection should generally not be employed as an alternative because it produces oscillatory approximations at discontinuities. If initial profiles obtained in that manner are evolved, bound-preserving limiters such as the ones discussed in this chapter are not capable of curing these spurious features. In the worst case, the L^2 projected initial condition can even be outside the largest admissible set of a system of conservation laws. Instead of either type of L^2 projection, one can use FCT constrained initialization, see for instance [Kuz10b, Sec. 8–9]. To enforce IDP properties for the discrete initial data, one can employ FCT-type limiters such as the one in [Gue18a].

3.4.1 Burgers equation

Some theoretical concepts of hyperbolic conservation laws were already illustrated for the inviscid Burgers equation (2.24). Let us now perform numerical studies for the problem

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial (u^2)}{\partial x} &= 0 \qquad && \text{in } \Omega \times (0,T), \\ u &= u_0 \qquad && \text{in } \Omega \times \{0\}, \end{aligned}$$

where the spatial domain $\Omega \subset \mathbb{R}$ is equipped with periodic boundaries. Since the flux function $\mathrm{f}(u) = \frac{u^2}{2}$ is convex, we may safely employ the wave speeds $\lambda_n(u_i, u_j) = \max\{|\mathrm{f}'(u_i)|, |\mathrm{f}'(u_j)|\} = \max\{|u_i|, |u_j|\}$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. To enforce entropy stability, we use the entropy pair $(\eta(u), q(u)) = (\frac{u^2}{2}, \frac{u^3}{3})$ with the corresponding entropy potential $\psi(u) = \frac{u^3}{6}$. If $u_0 \in \mathrm{C}^1(\overline{\Omega})$, the exact solution to this problem can be found by the method of characteristics (cf. Section 2.3.1, in particular, (2.22)). Hence,

we solve the nonlinear equation $u = u_0(x - ut)$ via Newton's method to obtain the solution value in (x, t). However, unless u_0 is monotonically increasing, there exists a critical time [LeV92, Sec. 3.3]

$$t_c = -\frac{1}{\min_{x \in \overline{\Omega}} u_0'(x)}$$

after which this approach is no longer valid because a shock starts to develop.

$1/h$	LOW	EOC	MCL	EOC	MCL-SDE	EOC
32	3.18E-02		4.03E-03		5.84E-03	
64	1.65E-02	0.95	1.37E-03	1.56	1.62E-03	1.85
128	9.25E-03	0.83	3.83E-04	1.83	3.93E-04	2.05
256	4.96E-03	0.90	9.81E-05	1.96	9.74E-05	2.01
512	2.57E-03	0.95	2.40E-05	2.03	2.47E-05	1.98

Table 3.1: Convergence history for the one-dimensional Burgers equation. The $\|\cdot\|_{\mathrm{L}^1(\Omega)}$ errors at $T = 0.1$ and the corresponding EOC for $\Omega = (0, 1)$, $u_0(x) = \sin(2\pi x)$.

$1/h$	LOW	EOC	MCL	EOC	MCL-SDE	EOC
32	2.05E-01		2.55E-02		2.98E-02	
64	1.03E-01	1.00	7.19E-03	1.82	7.59E-03	1.97
128	5.29E-02	0.96	1.79E-03	2.00	1.86E-03	2.03
256	2.71E-02	0.97	4.56E-04	1.98	4.58E-04	2.02
512	1.37E-02	0.99	1.12E-04	2.03	1.14E-04	2.00

Table 3.2: Convergence history for the one-dimensional Burgers equation. The $\|\cdot\|_{\mathrm{L}^1(\Omega)}$ errors at $T = 0.5$ and the corresponding EOC for $\Omega = (0, 2\pi)$, $u_0(x) = 0.5 + \sin(x)$.

In the first example, we evolve the initial condition $u_0(x) = \sin(2\pi x)$ in the spatial domain $\Omega = (0, 1)$ [Kuz20c, Sec. 7.3]. Here the critical time is $t_c = \frac{1}{2\pi}$ and the shock remains stationary at location $x = 0.5$. We solve this problem up to the final time $T = 0.1 < t_c$ and perform convergence analysis for LOW, MCL and MCL-SDE, employing a hierarchy of uniform meshes for spatial discretization. The obtained $\mathrm{L}^1(\Omega)$ errors at the final time and the corresponding experimental orders of convergence (EOC) are reported in Tab. 3.1. Next, we repeat this test in $\Omega = (0, 2\pi)$ and replace the previously employed initial condition with $u_0(x) = 0.5 + \sin(x)$ [Kur00, Sec. 6.2]. In this example, the shock develops at $t_c = 1$ and propagates to the right. The results of this convergence test for end time $T = 0.5$ are reported in Tab. 3.2. In both examples, we observe optimal first order rates for the low order method and second order of accuracy for the flux-limited schemes with and without the entropy fix.

Next, we increase the end times in both simulations, to run them longer than the respective critical times t_c. Snapshots for both cases are displayed in Fig. 3.1. Here we

(a) $\Omega = (0, 1)$, $T = 0.5$, and $u_0(x) = \sin(2\pi x)$. (b) $\Omega = (0, 2\pi)$, $T = 2$, and $u_0(x) = 0.5 + \sin(x)$.

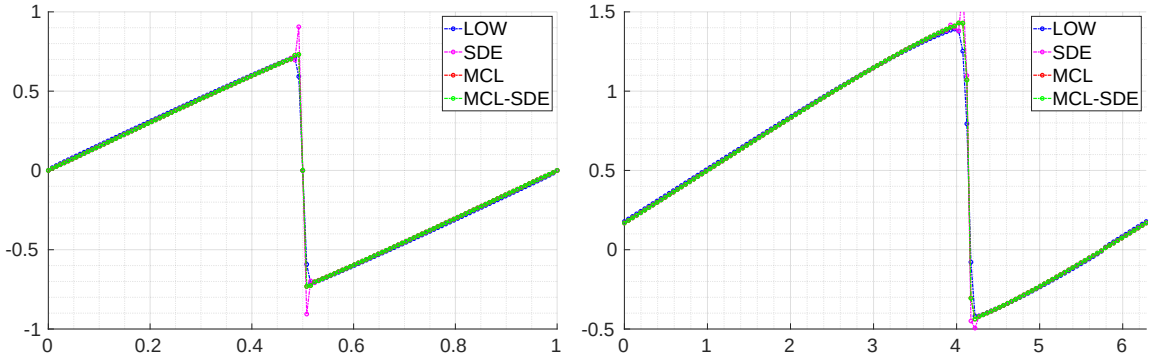


Figure 3.1: Approximations to the one-dimensional Burgers equation obtained with adaptive SSP2 RK time stepping and $\nu = 1$ on uniform meshes consisting of 128 elements.

additionally present approximations obtained with the SDE scheme, i. e., with the bound-preserving limiter disabled. The corresponding curves exhibit over- and undershoots around the shocks, which are suppressed by the standard MCL limiter. Aside from these ripples the approximations are satisfactory. In contrast, the target scheme without any limiting introduces more severe oscillations, which propagate and increase in magnitude, causing the simulation to blow up before the respective final time T. For this test problem, we see no benefit in using the MCL-SDE scheme rather than the merely bound-preserving but not necessarily entropy stable MCL method. Thus, we consider a more delicate scalar problem in the following section.

3.4.2 KPP problem

We study a benchmark proposed by Kurganov et al. [Kur07b, Sec. 5.3] that is commonly known as the KPP problem. In this 2D test case, we solve the scalar conservation law with the nonconvex flux function $\mathbf{f}(u) = (\sin(u), \cos(u))^\mathsf{T}$. The spatial domain is chosen as $\Omega = (-2, 2) \times (-2.5, 1.5)$ and the initial condition reads

$$u_0(\boldsymbol{x}) = \begin{cases} u^{\max} := \frac{7}{2}\pi & \text{if } |\boldsymbol{x}| < 1, \\ u^{\min} := \frac{1}{4}\pi & \text{if } |\boldsymbol{x}| \geq 1. \end{cases}$$

According to [Kur07b], the unique vanishing viscosity solution exhibits a spiral structure that many high-resolution schemes fail to capture correctly.

For the end time $T = 1$, we may treat all boundaries as outlets. In our experiments, we overestimate the wave speeds by setting $\lambda_{ij} = 1$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ [Kuz20c]. Sharper estimates can be found in [Gue17]. Since $\mathbf{f}$ is not isotropic, setting λ_{ij} to $\max\{|\mathbf{f}'(u_i) \cdot \boldsymbol{n}|, |\mathbf{f}'(u_j) \cdot \boldsymbol{n}|\}$ produces approximations that violate global discrete maximum principles.

The numerical results displayed in this section are visualized with the open source C++ software GLVis. In Fig. 3.2 we present the LOW approximation obtained on a uniform quadrilateral mesh of 1024^2 elements. SSP2 RK time stepping with constant time step $\Delta t = 2^{-10}$ is employed to match the spatial accuracy and to satisfy the CFL condition (3.43). Since fully discrete entropy inequalities w. r. t. *every* entropy pair hold for the low order method [Gue16b, Thm. 4.7], we may expect the profiles in Fig. 3.2 to capture the qualitative behavior of the vanishing viscosity solution reasonably well.

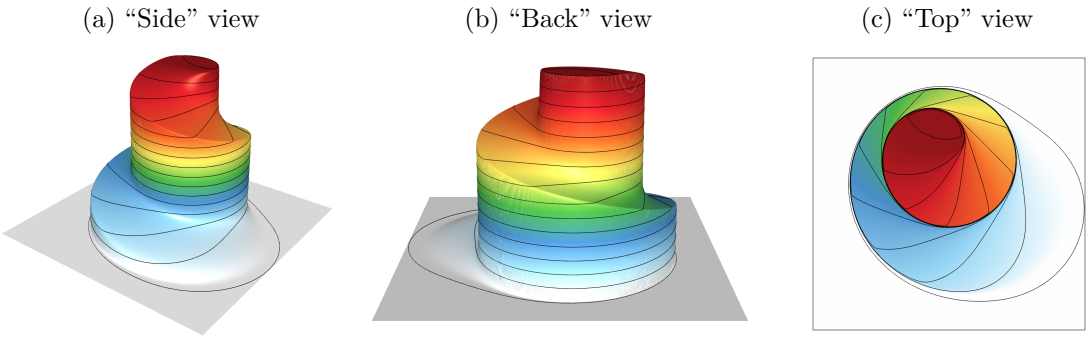


Figure 3.2: LOW approximation at $T = 1$ to the KPP problem [Kur07b] obtained with SSP2 RK time stepping and $\Delta t = 2^{-10}$ on a uniform quadrilateral mesh consisting of 1024^2 elements.

Let us now investigate the performance of the flux-limited schemes under consideration by choosing a much coarser uniform mesh with 128^2 quadrilateral elements. In accordance with the CFL condition (3.43), we employ a constant time step of $\Delta t = 2^{-7}$ for SSP2 RK time stepping. In Fig. 3.3 we display the MCL results without any entropy fixes as well as two variants of MCL-SDE approximations. The first one enforces entropy stability w. r. t. the quadratic entropy $\eta(u) = \frac{u^2}{2}$ for which the entropy flux and potential of the KPP problem read

$$\boldsymbol{q}(u) = (u\sin(u) + \cos(u),\, u\cos(u) - \sin(u)), \qquad \boldsymbol{\psi}(u) = (-\cos(u),\, \sin(u)),$$

respectively. Secondly, we use a Kruzhkov entropy pair (2.37) with corresponding entropy potential $\boldsymbol{\psi}_\kappa(u) = \operatorname{sign}(u-\kappa)\,\mathbf{f}(\kappa)$. In all experiments performed in this section, the Kruzhkov parameter is set to $\kappa = \frac{1}{2}(u^{\min} + u^{\max}) = \frac{15}{4}\pi$.

Remark 3.28
Note that Lemma 3.27 does not apply to Kruzhkov entropy pairs because of their insufficient regularity. Thus, we continually check whether the entropy production bounds $\min\{Q_{ij}, Q_{ji}\} - d_{ij}P_{ij}/2$ (see (3.73)) become negative for any pair of nodes, which could pose difficulties for entropy limiting. In all numerical experiments of this section, the inequality $\min\{Q_{ij}, Q_{ji}\} - d_{ij}P_{ij}/2 \geq 0$ was always satisfied up to machine precision for the Kruzhkov entropy with $\kappa = \frac{15}{4}\pi$. ♢

(a) MCL without entropy fix (b) MCL-SDE, $\eta(u) = \frac{u^2}{2}$ (c) MCL-SDE, $\eta(u) = |u - \frac{15}{4}\pi|$

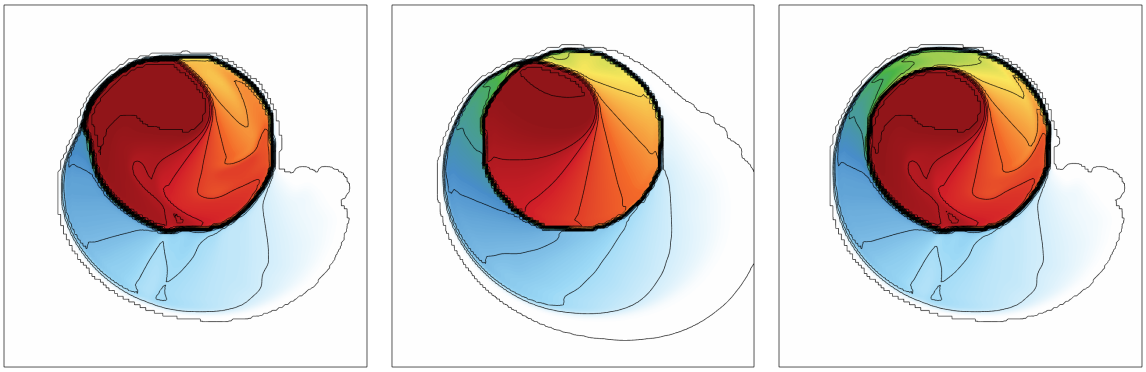


Figure 3.3: Flux-corrected approximations at $T = 1$ to the KPP problem [Kur07b] obtained with SSP2 RK time stepping and $\Delta t = 2^{-7}$ on a uniform quadrilateral mesh consisting of 128^2 elements.

None of the profiles in Fig. 3.3 is in satisfactory agreement with the approximation displayed in Fig. 3.2c. Specifically, the gap between the spiral-shaped shocks visible in Fig. 3.2c cannot be made out in Figs. 3.3a and 3.3b. Based on the profile shown in Fig. 3.3c this issue seems to be resolved if the entropy fix is performed w. r. t. this Kruzhkov entropy pair. By taking a close look at the contour lines of this approximation, we can tell that this scheme does not capture the exact solution very well either. None of the results in Fig. 3.3 can be improved by merely increasing the spatial and/or temporal resolution. We illustrate this fact by displaying the MCL-SDE results for the square entropy obtained on a sequence of refined meshes in Fig. 3.4. The kink visible in the approximations does not vanish, even upon further refinement.

This issue can be fixed by employing additional stabilization techniques. For instance, entropy viscosity can be introduced into the high order scheme as in [Gue17, Kuz20c]. In [Kuz22a] we alternatively enforce Tadmor's condition w. r. t. *entropy dissipative bounds* (see [Kuz22a, p. 8] for details). To further diversify the available options, we follow another strategy here. Note that, thus far, we have enforced entropy stability of flux-corrected approximations w. r. t. a single entropy pair. But what if the MCL-SDE approximations in Figs. 3.3b and 3.3c do in fact converge to respective weak solutions that are entropic w. r. t. *only a specific* entropy pair? The limits of such approximations would be different from the vanishing viscosity solution, which satisfies entropy inequalities w. r. t. *every* entropy pair. This argument suggests that one would have to enforce entropy inequalities for every Kruzhkov entropy, which is not feasible. We may however perform entropy fixes for $n \in \mathbb{N}$ entropy pairs and choose the smallest of the n correction factors β_{ij} for each pair of nodes. To test this approach, we repeat the MCL-SDE run with *both* of the above entropy pairs. This time we employ three unstructured triangular meshes with mesh sizes $h \in \{0.08, 0.04, 0.02\}$. For time stepping

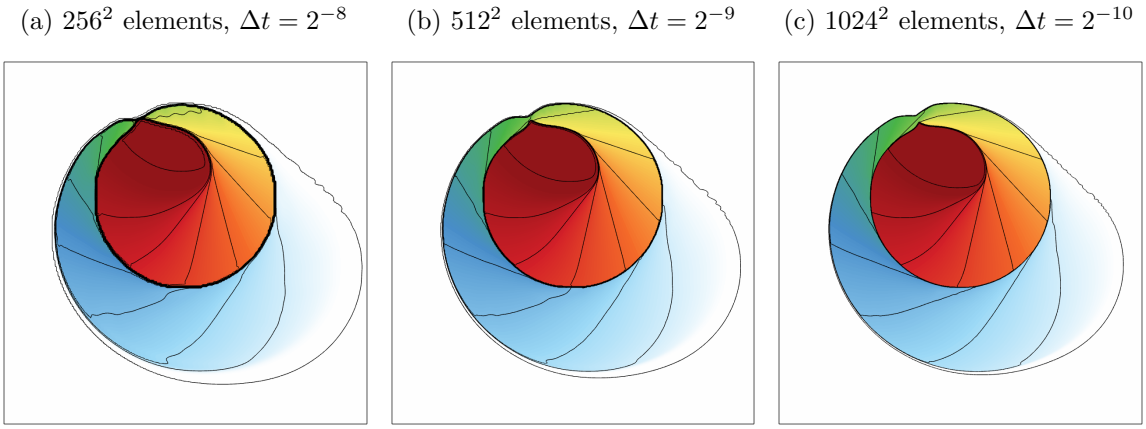


Figure 3.4: MCL-SDE approximations at $T = 1$ to the KPP problem [Kur07b], where entropy limiting is performed w. r. t. $\eta(u) = \frac{u^2}{2}$. Solutions obtained with SSP2 RK time stepping and constant time steps on uniform quadrilateral meshes.

we use the SSP2 RK method with a total of 125, 250, and 500 time steps of constant size, respectively. In Fig. 3.5 we display the results obtained with SDE and MCL-SDE schemes. Since the bound preserving limiter is disabled in the former, oscillations in the vicinity of shocks do arise but the entropy fixes lead to well-separated spiral shock patterns in all cases. Similar results are obtained if structured triangular or quadrilateral meshes are employed. Moreover, variations of the temporal discretization do not seem to affect the results as long as the CFL condition (3.43) is satisfied.

A conclusion that can be drawn from this section is that for certain problems, it is imperative to enforce entropy stability in addition to using a bound-preserving limiter. For fixes (3.74) based on Kruzhkov entropy pairs (2.37), the Kruzhkov parameter κ must be an element of the interval $[u^{\min}, u^{\max}]$. Otherwise both sides of Tadmor's condition (3.68) are zero (if the approximations are bound-preserving) and the antidiffusive fluxes are not modified. For problems satisfying the conditions of Panov's theorem [Pan94, Thm. 1], it suffices to perform a single entropy fix at most. For general scalar conservation laws it may be necessary to employ these corrections w. r. t. more than one entropy pair. It is unclear a priori, how many fixes are required, and, in particular, which entropy pairs should be considered. Therefore, entropy limiting for a single entropy pair with entropy-dissipative bounds as in [Kuz22a, Sec. 6.2] seems to be a better option for such problems.

3.4.3 Euler equations of gas dynamics

We now apply the schemes under consideration to classical test problems for the compressible Euler equations (cf. Section 2.2.2). In the below tests, the ratio of specific heats/adiabatic constant is set to $\gamma = 1.4$. Our experiments are restricted to the 1D

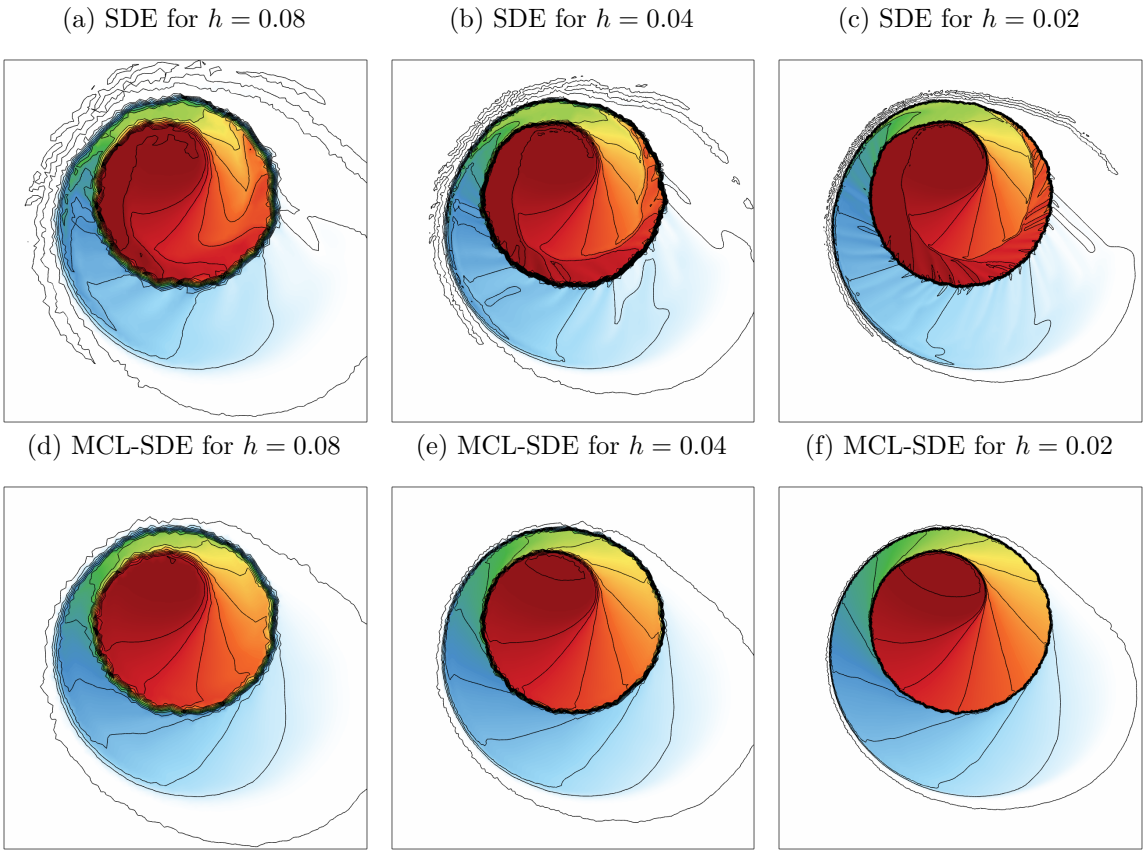


Figure 3.5: SDE and MCL-SDE approximations at $T = 1$ to the KPP problem [Kur07b], where entropy limiting is performed w. r. t. $\eta(u) = \frac{u^2}{2}$ and $\eta(u) = |u - \frac{15}{4}\pi|$. Solutions obtained with SSP2 RK time stepping and constant time steps on unstructured triangular meshes.

case. For numerical studies of the MCL scheme applied to the two-dimensional Double Mach reflection problem [Woo84], we refer to [Kuz20a, Sec. 6.6].

3.4.3.1 Sod's shock tube

A very common test case for the Euler equations is the shock tube problem studied by Sod [Sod78]. This benchmark constitutes a classical Riemann problem and is a relatively mild example. The spatial domain is $\Omega = (0, 1)$ and the initial condition expressed in conservative variables reads

$$u_0(x) = \begin{cases} (1, 0, 2.5)^\mathsf{T} & \text{if } x < 0.5, \\ (0.125, 0, 0.25)^\mathsf{T} & \text{if } x > 0.5. \end{cases}$$

This Riemann problem has a semi-analytical solution, in the sense that one can derive a closed form expression for $u(x, t)$ that depends on a parameter, which can only be

approximated numerically (see [Fei03, Sec. 3.1.6] or [Tor09, Chap. 4] for details).

We solve Sod's shock tube numerically up to time $T = 0.25$, which constitutes a time instant before the flow reaches either of the domain boundaries. Indeed, no boundary treatment is necessary in this example although the classical setting suggests to employ reflecting walls at $x \in \{0, 1\}$. Fig. 3.6 displays the LOW, MCL, and MCL-SDE approximations obtained for this setup. We also report the accumulated $\mathrm{L}^1(\Omega)$ error at the final time

$$e_T^1 := \|u(T) - u_h(T)\|_{\mathrm{L}^1(\Omega)^3}.$$

In other words, we sum the errors in each component of the vector of conserved unknowns.

(a) LOW, e_T^1=1.08E-01 (b) MCL, e_T^1=2.74E-02 (c) MCL-SDE, e_T^1=3.09E-02

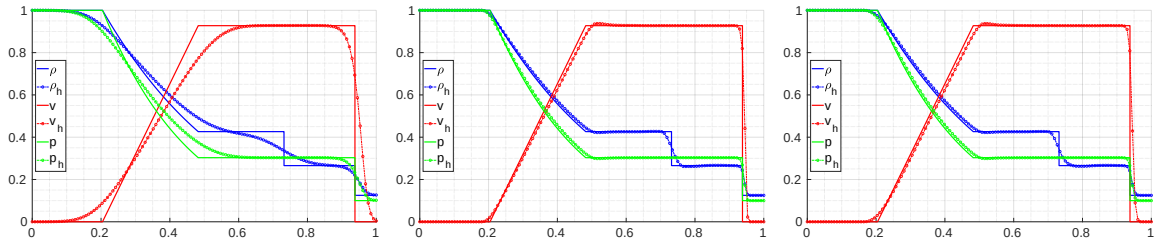


Figure 3.6: Sod's shock tube problem for the Euler equations [Sod78]. Density, velocity, and pressure profiles at $T = 0.25$ obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 128 elements.

The low order method produces quite diffusive profiles for each displayed field, while the MCL and MCL-SDE approximations are in very good agreement with the exact solutions. There is no qualitative difference between the approximations with and without entropy fix. Only the numerical error is slightly higher in the entropy-limited scheme. For these high resolution schemes, there seems to be an overshoot in the velocity profile on the right of the rarefaction wave. Actually, this feature can also be observed in the low order solution, on very fine meshes [Gue16b, Sec. 5.3.2]. Thus, it is not a spurious artifact but a result of the Eulerian solution approach (as opposed to a Lagrangian framework). Moreover, convergence to the exact solution is not inhibited.

Next, we modify the CFL parameter ν and report the total number of employed time steps `#TS`, the number of repeated Runge–Kutta stages `#RK`, as well as the total number of Euler steps `#Euler=2#TS+#RK` in Tab. 3.3. Here the factor 2 is due to the employed SSP2 RK method. Among the three CFL parameters under investigation, the most economical choice is $\nu = 0.9$ because it requires the smallest number of forward Euler updates. Of course, smaller values for ν may be employed, since they can be expected to produce approximations of increased temporal accuracy.

Either of the flux-corrected profiles shown in Fig. 3.6 are satisfactory for this test problem. The slight smearing of the contact discontinuity (jump in the blue curve,

	$\nu = 0.5$			$\nu = 0.9$			$\nu = 1$		
scheme	#TS	#RK	#Euler	#TS	#RK	#Euler	#TS	#RK	#Euler
LOW	274	0	548	152	1	305	138	109	385
MCL	280	0	560	156	2	314	141	84	366
MCL-SDE	280	0	560	156	2	314	141	89	371

Table 3.3: Sod's shock tube problem for the Euler equations [Sod78]. Number of employed time steps, repeated Runge–Kutta stages and total number of forward Euler updates for different values of the CFL parameter ν.

located at roughly $x = 0.73$) may be worthy of improving. This task can be achieved by increasing the spatial and temporal resolution. A full convergence history for each method is presented in Tab. 3.4. We observe convergence rates of at least $\sqrt{h}$, which

$1/h$	LOW	EOC	MCL	EOC	MCL-SDE	EOC
32	2.36E-01		9.84E-02		1.10E-01	
64	1.65E-01	0.52	5.37E-02	0.87	6.00E-02	0.87
128	1.08E-01	0.61	2.74E-02	0.97	3.09E-02	0.96
256	6.79E-02	0.67	1.41E-02	0.96	1.61E-02	0.94
512	4.17E-02	0.70	6.89E-03	1.04	7.96E-03	1.02

Table 3.4: Convergence history of Sod's shock tube problem for the Euler equations [Sod78]. The $\|\cdot\|_{\mathrm{L}^1(\Omega)}$ errors at $T = 0.25$ and the corresponding EOC.

is the optimal order to be expected. Here the benefit of employing the high-resolution schemes MCL or MCL-SDE over the low order method is clearly visible in the improved convergence rates.

We conclude our experiments of Sod's shock tube problem with a remark on the need for employing an IDP fix to enforce nonnegativity of the pressure/internal energy. Since this is a relatively mild test problem, all of the runs in this section can be performed without using such a limiter. However, the pressure fix is activated a few times in the course of the simulation. The occurrence of negative internal energies in the bar states does not automatically imply a violation of the nonnegativity constraint for the updated solution, because the convex combination of multiple bar states may still produce nonnegative nodal pressures. In general, however, we need to employ such an IDP fix. Therefore, all approximations for Sod's shock tube in this section were computed with the pressure fix. The next test problem is one where MCL and MCL-SDE simulations break down unless a pressure fix is employed.

3.4.3.2 Woodward–Colella blast wave problem

Another classical benchmark for the Euler equation is the one-dimensional blast wave problem proposed by Woodward and Colella [Woo84]. As for Sod's shock tube, the

spatial domain $\Omega = (0, 1)$ has reflecting walls at $x \in \{0, 1\}$. Here the initial condition expressed in conserved unknowns reads

$$u_0(x) = \begin{cases} (1, 0, 2500)^\mathsf{T} & \text{if } x < 0.1, \\ (1, 0, 0.025)^\mathsf{T} & \text{if } 0.1 < x < 0.9, \\ (1, 0, 250)^\mathsf{T} & \text{if } 0.9 < x \end{cases}$$

and is evolved up to end time $T = 0.038$. Strong shock waves develop from the discontinuities present in the initial profile and travel towards the domain center. By the end time, a collision of these waves occurs. Therefore, no closed form expression for the exact solution is available.

We solve the blast wave problem numerically with LOW, MCL, and MCL-SDE schemes using a fine uniform mesh consisting of 1 000 elements. This mesh resolves the initial condition accurately everywhere but at the points $x = 0.1$ and $x = 0.9$. As for the previous example, no repetition of any Runge–Kutta stages needs to be performed if the CFL parameter is set to $\nu = 0.5$. Numerical solutions for density, velocity and pressure are displayed in Fig. 3.7, along with respective reference solutions that were computed with a finite volume method on a very fine mesh. We observe satisfactory agreement of flux-corrected approximations with the respective reference solutions. Again, the low order method produces profiles that are drastically more diffusive than their limited counterparts. This behavior is most prominent in the density and pressure profiles and less pronounced for the velocity. Once more, the MCL and MCL-SDE results are almost indistinguishable from each other, which suggests that no entropy fixes are necessary for this problem.

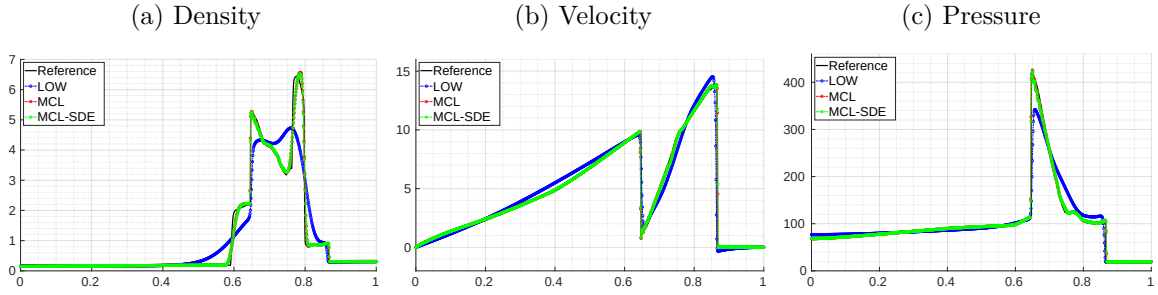


Figure 3.7: One-dimensional blast wave problem for the Euler equations [Woo84]. Approximations at $T = 0.038$ obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 1 000 elements.

Since the value ranges of the conserved unknowns are quite different, we present the $\mathrm{L}^1(\Omega)$ errors in density, momentum, and total energy separately, along with the total number of time steps `#TS` needed for each run. These results can be found in Tab. 3.5.

Next, we study a few variants of the schemes under investigation for the blast wave problem. First, we repeat the above experiment with backward Euler time stepping and

	LOW	MCL	MCL-SDE
$\|\rho(T) - \rho_h(T)\|_{\mathrm{L}^1(\Omega)}$	2.66E-01	5.30E-02	5.35E-02
$\|(\rho v)(T) - (\rho v)_h(T)\|_{\mathrm{L}^1(\Omega)}$	3.99E-01	1.29E-01	1.31E-01
$\|(\rho E)(T) - (\rho E)_h(T)\|_{\mathrm{L}^1(\Omega)}$	8.69E00	1.66E00	1.71E00
`#TS`	8720	8628	8628

Table 3.5: One-dimensional blast wave problem for the Euler equations [Woo84]. Errors in each conserved unknown and number of employed time steps.

leave the rest of the setup unmodified. In particular, we continue to employ adaptive time stepping with $\nu = 0.5$. Given the solution at the previous time step, the backward Euler updated solution is computed by solving the nonlinear equations

$$\tilde{u}_i = u_i + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i \backslash \{i\}} 2\tilde{d}_{ij}(\tilde{\bar{u}}_{ij} - \tilde{u}_i) + \frac{\Delta t}{m_i} \sum_{\Gamma_k \in \mathcal{F}_i} 2\tilde{d}_i^k(\tilde{\bar{u}}_i^k - \tilde{u}_i) \tag{3.75}$$

for $i \in \{1, \dots, N\}$. Here all quantities featuring a $\sim$ refer to values at the new time level. To solve the implicit equation (3.75), we implement the forward Euler type fixed point iteration

$$u_i^{(n+1)} = u_i + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij}^{(n)}(\bar{u}_{ij}^{(n)} - u_i^{(n)}) + \frac{\Delta t}{m_i} \sum_{\Gamma_k \in \mathcal{F}_i} 2(d_i^k)^{(n)}((\bar{u}_i^k)^{(n)} - u_i^{(n)}) \tag{3.76}$$

for $i \in \{1, \dots, N\}$. The iteration (3.76) is stopped once the residual of (3.75) combined for all degrees of freedom becomes less than 10^{-8} in magnitude in the discrete l^2 norm. In the process of calculating numerical solutions, we continually check whether the time step is in accordance with the CFL condition (3.43). With our choice of $\nu = 0.5$ this prerequisite is never violated in any of the schemes under consideration. Thus, all iterates in (3.76) are admissible and therefore, so are the approximate solutions produced by the backward Euler method.

It should be remarked that, in this example, we employ implicit time stepping solely for demonstrative purposes. Clearly, our nonlinear iteration (3.76) is very basic, and significantly more expensive than explicit time stepping. Indeed, our (serial, non-vectorized) Matlab implementation of the explicit methods required approximately 265 seconds on a laptop to compute the profiles of the three schemes shown in Fig. 3.7, while for backward Euler more than 45 minutes were needed on the same machine. The slow convergence of (3.76) motivates the use of Newton-like solvers, and suitable preconditioners for iterations. To employ the former in the context of high-resolution schemes, one needs to modify the merely Lipschitz-continuous schemes to depend smoothly on the unknowns (see [Bad17, Loh21]). We also refer to [Gur09, Chap. 8] for a detailed discussion on advanced nonlinear iterations for implicit schemes in combination with algebraic flux correction techniques.

The density profiles of our implicit run are shown in Fig. 3.8a along with a zoom to the region around the left peak in the profiles. These results look quite similar to those in Fig. 3.7 with the exception, that the leftmost discontinuity is significantly more smeared by the implicit approach. This increase in diffusivity is an unsurprising drawback of backward Euler time discretizations.

Next, we modify some of the setup of the numerical experiments with which the results in Fig. 3.7 were obtained. In Fig. 3.8b we show the LOW profile obtained on a very fine uniform mesh consisting of 10^4 elements. This curve is compared to those of MCL, and MCL-SDE schemes on the original mesh with 1 000 elements but with the bound-preserving limiter disabled. In other words, the MCL scheme only constrains the flux-corrected bar states to have nonnegative pressures, while the MCL-SDE scheme additionally enforces the semi-discrete entropy inequality. No bounds for numerical admissibility are enforced in either case.

(a) Backward Euler approximations (b) LOW refined & MCL-(SDE) w/o local bounds

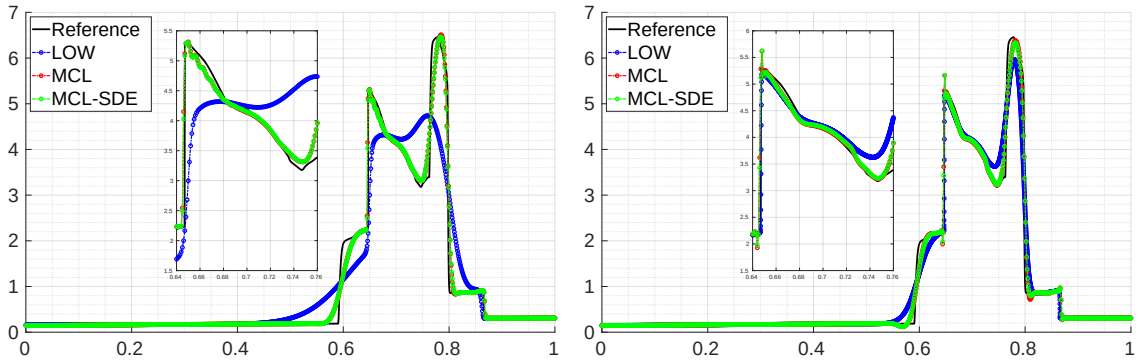


Figure 3.8: One-dimensional blast wave problem for the Euler equations [Woo84]. Density profiles for variants of LOW, MCL, and MCL-SDE at $T = 0.038$ obtained with adaptive SSP RK time stepping and $\nu = 0.5$ on uniform meshes.

We observe significant improvement in the low order solution compared to the one obtained on the coarser mesh. Nevertheless smearing around discontinuities is still more pronounced than in the flux-corrected profiles obtained on a mesh with 10 times fewer elements. If bound-preserving limiting is disabled we observe spurious oscillations around the discontinuities but, otherwise, the approximations are acceptable. Both the MCL and MCL-SDE profiles now require a total of 8746 time steps to reach the final simulation time. The fact that it is possible to obtain these profiles suggests that the only flux-correction necessary to solve the blast wave problem is the IDP pressure fix. Oscillations in the density, although present, do not cause the simulations to break down. Nevertheless, one should generally enforce local bounds to prevent these Gibbs phenomena in the vicinity of steep fronts. We remark that without the pressure fix for the flux-corrected bar states, both high resolution schemes produce inadmissible states,

which shows that these methods are not IDP if the pressure fix is disabled.

Interestingly enough, the small scale oscillations, which can be spotted in the zoomed region of Fig. 3.8a are not present in either the green or red curves in Fig. 3.8b. Indeed, the over- and undershoots are more local to the discontinuities if the bound-preserving limiter is disabled. This observation suggests that our local bounds for limiting (3.60)–(3.61) impose too restrictive constraints in this example. Recall that we define local bounds based solely on the low order bar states, a strategy that may be inappropriate here. In fact, admissibility of the bar states is proven by invoking the theory of simple Riemann problems in which no shock collisions occur (cf. Lemma 3.12).

Let us perform a final test with the MCL scheme using different definitions of raw antidiffusive fluxes. Our default option $f_{ij} = m_{ij}(\dot{u}_i^{\mathrm{L}} - \dot{u}_j^{\mathrm{L}}) + d_{ij}(u_i - u_j)$, where $\dot{u}_i^{\mathrm{L}}$ are the low order time derivatives (3.38) is compared with $f_{ij} = m_{ij}(\dot{u}_i^{\mathrm{G}} - \dot{u}_j^{\mathrm{G}}) + d_{ij}(u_i - u_j)$, where $\dot{u}_i^{\mathrm{G}}$ are the nodal Galerkin time derivatives defined by (3.37). In Fig. 3.9 we present the results of this test in which the problem setup is the same as in our first simulation for the blast wave problem.

(a) $f_{ij} = m_{ij}(\dot{u}_i^{\mathrm{L}} - \dot{u}_j^{\mathrm{L}}) + d_{ij}(u_i - u_j)$ (b) $f_{ij} = m_{ij}(\dot{u}_i^{\mathrm{G}} - \dot{u}_j^{\mathrm{G}}) + d_{ij}(u_i - u_j)$

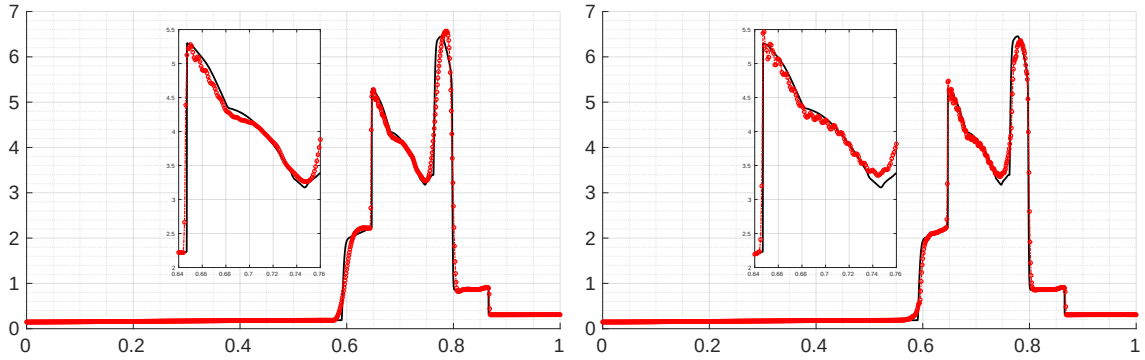


Figure 3.9: One-dimensional blast wave problem for the Euler equations [Woo84]. MCL density profiles at $T = 0.038$ obtained using different choices of f_{ij} with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 1 000 elements.

Both density profiles exhibit oscillations, which are clearly visible in the zoomed regions. However, these are significantly more pronounced if the unstabilized Galerkin version of antidiffusive fluxes is employed. Similar unsatisfactory profiles are obtained if the approximate time derivatives correspond to the lumped Galerkin method (same as $\dot{u}_i^{\mathrm{G}}$ but with inversion of M_{L} instead of the consistent mass matrix M) and with the lumped mass version $f_{ij} = d_{ij}(u_i - u_j)$. This experiment demonstrates the benefit of using the low order nodal time derivatives (3.38). As an alternative to our approach, one may employ the Galerkin time time derivative as long as some other form of stabilization is introduced into the algorithm. The use of low order time derivatives is an inexpensive strategy because the consistent mass matrix does not need to be inverted.

Chapter 4

Limiting for the shallow water equations with nonflat topography

In the previous chapter, we discussed monolithic convex limiting strategies and related concepts that were developed in [Kuz20a, Kuz20c, Kuz22a] for conservation laws. We now extend these techniques to a system of balance laws. In particular, we consider the shallow water equations (SWE) with a nonconservative topography term. This important nonlinear system of partial differential equations reads (cf. Section 2.2.3)

$$\frac{\partial}{\partial t}\begin{bmatrix} h \\ h\boldsymbol{v} \end{bmatrix} + \nabla \cdot \begin{bmatrix} h\boldsymbol{v}^\mathsf{T} \\ h\boldsymbol{v} \otimes \boldsymbol{v} + \frac{g}{2}h^2\mathcal{I} \end{bmatrix} + \begin{bmatrix} 0 \\ gh\nabla b \end{bmatrix} = 0 \qquad \text{in } \Omega \times \mathbb{R}_+. \tag{4.1}$$

Additional theoretical and numerical challenges arise when it comes to solving (4.1) instead of the system of conservation laws corresponding to the case $b \equiv \text{const}$. We refer to [Bou04, Ch. 3] and references cited therein for a review of the theory of balance laws and some aspects of the shallow water equations with topography. A summary on finite volume schemes for general hyperbolic problems with nonconservative terms can be found in [LeV02, Ch. 17]. Another useful reference on the shallow water equations is the paper by Delestre et al. [Del13]. In particular, it presents numerical benchmarks and shows how exact solutions to some of the test problems can be obtained.

We begin this chapter by specifying its main objectives in Section 4.1. Next, in Section 4.2, we review some of the literature on property-preserving schemes for the SWE. Our algebraic limiters for system (4.1) are introduced in Section 4.3. In Section 4.4, we combine them with algorithmic strategies for handling wet-dry transitions. The one-dimensional numerical examples presented in Section 4.5 conclude this chapter.

4.1 Objectives

The goal of this chapter is to generalize the bound-preserving and entropy-stable MCL schemes to the inhomogeneous SWE (4.1). One key requirement that we deemed essential in the development of our algorithms is that they represent generalizations of the corresponding schemes from Section 3.3 for the flat bottom case. Another desirable property of numerical methods for balance laws is their *well-balancedness*. For some hyperbolic systems, there exist certain steady states, i. e., solutions $u(\boldsymbol{x})$ that are independent of t because the flux and source terms are in equilibrium. A numerical method for solving such a system is called well balanced if it captures the simplest steady

states exactly in the discrete setting. Recall that in Section 2.1.2 we used the chain rule to rewrite the term $gh\nabla(h+b)$ as $\frac{g}{2}\nabla h^2 + gh\nabla b$. This decomposition suggests that system (4.1) admits the so-called *lake at rest* steady state solution

$$\boldsymbol{v} = \mathbf{0}, \qquad h\nabla(h+b) = 0. \tag{4.2}$$

This configuration corresponds to a still body of water that is unperturbed by external forces, such as in- and outflows through domain boundaries. Note that the second identity in (4.2) does not imply that the free surface elevation $H = h + b$ has to be a global constant, as is the case for a classical lake at rest. In fact, (4.2) allows variations in H, as long as the water height h is zero at the same physical location. This case corresponds to a so-called *dry state* that occurs whenever the topography b exceeds the water depth h. For an island that rises from a body of water, every point on the surface of the island represents a dry state.

Besides lake at rest configurations, other types of equilibria exists for the SWE. In the absence of friction and/or Coriolis forces, (4.1) admits so-called *moving water equilibria* steady states. In 1D, such configurations occur if the discharge hv as well as the expression $\frac{1}{2}v^2 + g(h+b)$ remain constant [Bou04, Ch. 3], [Kur07a]. A two-dimensional analogue of the moving water equilibrium reads

$$\nabla\cdot(h\boldsymbol{v}) = 0, \qquad (\boldsymbol{\nabla v})\,\boldsymbol{v} + g\nabla(h+b) = \mathbf{0}.$$

These relations are derived from (4.1) by employing the product rule. They remain valid for discontinuous solutions with steady shocks. While lake at rest scenarios can be preserved with simple numerical treatments (see, e. g., [Aud04, Kur07a, Fjo11, Aze17, Ber19]), moving water equilibria require advanced well-balancing techniques (see for instance [Noe07]). The incorporation of such approaches into our flux correction schemes is a topic of its own and will not be attempted in this work. Instead, we focus on well-balancing w. r. t. lake at rest configurations. Nevertheless, some numerical examples of moving water equilibria are solved numerically in this chapter.

Another important aspect of numerical methods for the SWE and related models is the need to deal with *wetting and drying* scenarios (see, e. g., [Ric09, Bar15, Vat15]) in which simulations may crash if no special measures are implemented. In many examples of practical interest, there exist islands rising from the body of water but the interface between these islands and the water surface is moving. The dry land masses are then modeled by allowing the bottom topography to exceed the values of the free surface elevation at the same location (cf. Fig. 2.1). Even in the case that the resolution is sufficient to capture the interface, it can be quite difficult to accurately resolve the moving shoreline with numerical methods. In this chapter, we present two new ways of dealing with this issue and compare our results with some existing approaches.

4.2 Literature

Before presenting our generalization of the MCL methodology to system (4.1), let us mention some existing approaches that, in our opinion, represent important contributions to the field. Our brief discussion of these methods is by no means a comprehensive review of all property-preserving schemes proposed in the SWE literature.

Many well-balanced methods use the *hydrostatic reconstruction* technique developed by Audusse et al. [Aud04]. Originally proposed in the context of finite volume methods, it yields approximations that preserve lake at rest scenarios, ensure nonnegativity of water heights under standard CFL conditions, and satisfy a semi-discrete entropy inequality. Hydrostatic reconstructions achieve these properties by properly balancing flux and source terms. However, even the original low order hydrostatic reconstruction method does not satisfy fully discrete entropy inequalities [Aud04, Sec. 2.2]. This issue is addressed by Berthon et al. [Ber19], who increase the amount of artificial viscosity to construct a method that is entropy stable in the fully discrete sense. However, the final numerical example of their study indicates that their scheme still violates the fully discrete entropy inequality in the presence of nonflat bathymetry and dry states.

Noelle et al. [Noe07] transform to *equilibrium variables* and use equilibrium-limited reconstructions in their high order finite volume methods. The main focus of their work is on exact preservation of moving water equilibria in addition to lakes at rest. They show that their method captures such states exactly if all stationary shocks are located at cell interfaces and Roe's numerical flux is employed. Additionally, they prove a Lax–Wendroff-type theorem [Noe07, Thms. 3.14 and 3.17, respectively].

Another type of well-balanced finite volume discretizations of the SWE is the family of *central-upwind schemes*. The ones presented by Kurganov and Petrova [Kur07a] are well balanced for the lake at rest and positivity preserving for the water height. As in [Aud04], these properties are achieved by performing compatible reconstructions for the conserved unknowns and the bathymetry. To ensure positivity preservation for the water heights, the algorithm is enhanced with a generalized minmod limiter. Additionally, a modification for numerical treatment of wetting and drying is proposed in [Kur07a]. In Section 4.5, we test this approach and some new alternatives in the context of our flux-limited finite element schemes. In contrast to the central-upwind methods presented in [Kur07a], the applicability of AFC tools is not restricted to Cartesian grids.

Fjordholm et al. [Fjo11] present well-balanced and entropy-conservative/-stable finite volume schemes for the SWE with topography. Contrary to [Aud04], their approach does not rely on reconstructions of the bathymetry. Instead, a transformation to equilibrium variables is used to ensure well-balancedness for moving water equilibria. Moreover, Fjordholm et al. generalize Tadmor's entropy stability condition to the case of nonflat topography and use it to design numerical fluxes. It is admitted in [Fjo11] that oscillations around discontinuities may produce negative water heights. This

shortcoming could be cured by employing a positivity-preserving limiter.

In [Ric09], Ricchiuto and Bollermann design residual distribution schemes for the shallow water equations. As in our case, linear continuous finite elements are used in the baseline discretization. The method preserves lake at rest scenarios and guarantees nonnegativity of the water height under CFL-like constraints. Some similarities and differences of algebraic flux correction schemes and residual distribution methods were already discussed in Section 1.1. In the SWE context, the latter approach introduces some complications, as mentioned in the conclusions of [Ric09].

Wintermeyer et al. [Win17] discretize the SWE using high order discontinuous Galerkin spectral element methods. A proof of entropy conservation/stability is provided for suitable choices of numerical fluxes. Well-balancedness w. r. t. lake at rest scenarios is achieved despite difficulties caused by the presence of metric terms in the case of curvilinear elements. Although the entropy-stable DG scheme developed in [Win17] is not bound preserving, it seems to be well suited for algebraic flux correction. Thus, we envisage that under- and overshoots can be avoided using MCL-type approaches.

Azerad et al. [Aze17] present a property-preserving finite element method that is well balanced w. r. t. the lake at rest. Their scheme incorporates hydrostatic reconstructions into a nodal continuous Galerkin formulation. Detailed analysis is performed in [Aze17] for a low order version, which is a generalization of the algebraic Lax–Friedrichs method to the case of a nonflat bottom. Second order of accuracy is recovered by adjusting the numerical viscosity. Extensions [Gue18b] of the schemes in [Aze17] are based on FCT-type limiting and incorporate a regularized friction term into the model. The AFC methodology that we propose in the present chapter differs from the one developed in [Aze17] in the formulation of the low order method and in the limiting strategy. Adopting the MCL design philosophy, we incorporate discretized bathymetry gradients into the bar states of a well-balanced and positivity-preserving semi-discrete scheme. Our algorithm provides all desired properties and does not use hydrostatic reconstructions. Indeed, our strategy of adjusting bathymetry sources is based solely on limiting.

4.3 Algebraic flux correction schemes

Let us now extend the schemes discussed in Section 3.3 to the inhomogeneous hyperbolic system (4.1) step by step. As before, we first choose a target discretization suitable for flux correction procedures. Next, we derive a low order method for which all desirable properties (conservation, numerical and physical admissibility, entropy stability) are guaranteed. Then we recover the target scheme by including raw antidiffusive fluxes into the algorithm. Finally, these fluxes are limited in a way that ensures preservation of both local and global bounds, as well as semi-discrete entropy stability.

The most important considerations regarding the low order method and flux-corrected schemes were already discussed in the context of general systems of conservation laws.

For brevity, we focus solely on aspects that need to be modified and refer to Section 3.3 for the rest. In particular, the treatment of boundary terms does not present any additional difficulties because these terms are the same for discretizations of (4.1) with constant and spatially variable bathymetry b. Therefore, we omit all boundary terms in the following presentation but remark that they generally need to be incorporated into the algorithm. This task is achieved in exactly the same manner as in Section 3.3.

Conceptually, the only difference compared to the case of a flat topography is the presence of the nonconservative term $gh\nabla b$ in the momentum equation. The consistent Galerkin discretization of this term produces the nodal contribution

$$g \sum_{k \in \mathcal{N}_i} h_k \int_\Omega \varphi_i \, \varphi_k \nabla b \, \mathrm{d}\boldsymbol{x} \tag{4.3}$$

to the ith component of the momentum equation. In this formula, h_k denotes the value of the discretized water height at node $\boldsymbol{x}_k$. It is common [Kur07a] to approximate b by its piecewise (multi)-linear continuous interpolant $b_h \in \mathrm{V}_h$, defined by

$$b_h(\boldsymbol{x}) = \sum_{j=1}^{N} b_j \varphi_j(\boldsymbol{x}), \qquad b_j := b(\boldsymbol{x}_j).$$

If the bathymetry b is discontinuous in node $\boldsymbol{x}_j$, one may set b_j equal to any of the one-sided limits in cells containing $\boldsymbol{x}_j$ or to an average of these limits. Alternatively, a projection can be used to obtain b_h from b.

In Section 3.3.2, we approximated the inviscid flux $\mathbf{f}(u_h)$ using a group finite element formulation to derive a quadrature rule for the corresponding integral. Discretization (4.3) of the source term must be approximated similarly for our method to be well balanced. With this goal in mind, we replace (4.3) by the quadrature-based version

$$g \sum_{j \in \mathcal{N}_i \backslash \{i\}} \frac{h_i + h_j}{2} (b_j - b_i) \, \mathbf{c}_{ij}, \tag{4.4}$$

which is similar to what is done, for instance, in [Aud04, Eq. (3.8)], [Kur07a, Eq. (2.6)], and [Fjo11, Eq. (2.7)]. Note that if $h_k \equiv \text{const}$ for $k \in \mathcal{N}_i$, then (4.4) equals (4.3) with b replaced by b_h. In this sense, (4.4) is similar to the quadrature rule based on the group finite element approximation of $\mathbf{f}(u_h)$. The approximation (4.4) to (4.3) is second order accurate if h_k is constant for $k \in \mathcal{N}_i$ and first order accurate otherwise.

Remark 4.1
In principle, it is possible to compensate the quadrature error due to the source term approximation (4.4) in the process of flux correction. If this is desired, one needs to decompose the difference between (4.3) and (4.4) into edge contributions, add the corresponding correction terms $\boldsymbol{f}_{ij}^b$ to the raw antidiffusive fluxes $\boldsymbol{f}_{ij}^{h\boldsymbol{v}} = -\boldsymbol{f}_{ji}^{h\boldsymbol{v}}$ of the

momentum equation, and modify the limiting formula because $\boldsymbol{f}_{ij}^b \neq -\boldsymbol{f}_{ji}^b$ in general. At an early stage of developing our method, we performed a preliminary study that showed the feasibility of this approach. In the final version, we disregard first order quadrature errors caused by using (4.4) instead of (4.3) because in real-life applications such errors are likely to be negligible compared to measurement errors in the bathymetry data. ♢

Inserting (4.4) into the semi-discrete momentum equation and approximating other terms as in (3.19), we obtain the quadrature-based target scheme

$$\sum_{j=1}^{N} m_{ij} \frac{\mathrm{d}h_j}{\mathrm{d}t} = -\sum_{j\in\mathcal{N}_i\setminus\{i\}} ((h\boldsymbol{v})_j - (h\boldsymbol{v})_i)^{\mathsf{T}} \mathbf{c}_{ij}, \tag{4.5a}$$

$$\sum_{j=1}^{N} m_{ij} \frac{\mathrm{d}(h\boldsymbol{v})_j}{\mathrm{d}t} = -\sum_{j\in\mathcal{N}_i\setminus\{i\}} \left[\mathbf{f}_j^{h\boldsymbol{v}} - \mathbf{f}_i^{h\boldsymbol{v}} + g\frac{h_i+h_j}{2}(b_j-b_i)\,\mathcal{I}\right] \mathbf{c}_{ij}, \tag{4.5b}$$

where

$$\mathbf{f}_i^{h\boldsymbol{v}} = \frac{1}{h_i}(h\boldsymbol{v})_i \otimes (h\boldsymbol{v})_i + \frac{g}{2}h_i^2\,\mathcal{I}, \qquad i \in \{1,\dots,N\}.$$

It is worth checking at this stage whether (4.5) is well balanced for lake at rest configurations (4.2) and clarify the meaning of well-balancedness in this context. If $(h\boldsymbol{v})_i = \mathbf{0}$ for all $i \in \{1,\dots,N\}$, then (4.5b) reduces to

$$\begin{aligned}\sum_{j=1}^{N} m_{ij} \frac{\mathrm{d}(h\boldsymbol{v})_j}{\mathrm{d}t} &= -\sum_{j\in\mathcal{N}_i\setminus\{i\}} \frac{g}{2}[h_j^2 - h_i^2 + (h_i+h_j)(b_j-b_i)]\,\mathbf{c}_{ij} \\ &= -\sum_{j\in\mathcal{N}_i\setminus\{i\}} \frac{g}{2}(h_j+h_i)[h_j - h_i + b_j - b_i]\,\mathbf{c}_{ij}.\end{aligned}$$

Assuming for now that $h_i \geq 0$ for all $i \in \{1,\dots,N\}$ (a condition that we enforce later on), we see that the right hand side of this expression is zero if and only if

$$h_i = h_j = 0 \qquad \text{or} \qquad H_i = H_j \qquad \forall i \in \{1,\dots,N\},\ j \in \mathcal{N}_i \setminus \{i\}, \tag{4.6}$$

where $H_i := h_i + b_i$, $i \in \{1,\dots,N\}$ are the coefficients of the discrete free surface elevation $H_h \in \mathrm{V}_h$. If (4.6) holds, the discharge is unperturbed, and thus the right hand side of the continuity equation (4.5a) is zero.

4.3.1 Low order method

As in the case of a system of conservation laws, we need to modify (4.5) to obtain a property-preserving semi-discretization. To this end, we perform row sum mass lumping

and include Rusanov (local Lax–Friedrichs) artificial dissipation. However, the presence of the nonconservative term makes matters more involved. Special care needs to be taken, for instance, to ensure semi-discrete entropy stability, positivity preservation for water heights, and well-balancedness. To construct a low order method that meets all of our requirements, let us begin with the straightforward generalization

$$m_i \frac{\mathrm{d}h_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} \Big[d_{ij}(h_j - h_i) - ((h\boldsymbol{v})_j - (h\boldsymbol{v})_i)^{\mathsf{T}} \mathbf{c}_{ij} \Big], \tag{4.7a}$$

$$\begin{aligned} m_i \frac{\mathrm{d}(h\boldsymbol{v})_j}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} \Big[& d_{ij}((h\boldsymbol{v})_j - (h\boldsymbol{v})_i) - (\mathbf{f}_j^{h\boldsymbol{v}} - \mathbf{f}_i^{h\boldsymbol{v}})\, \mathbf{c}_{ij} \\ & - \frac{g}{2}(h_i + h_j)(b_j - b_i)\, \mathbf{c}_{ij} \Big] \end{aligned} \tag{4.7b}$$

of the algebraic Lax–Friedrichs method (3.26) applied to the SWE with flat bathymetry. As before, the artificial viscosity coefficients d_{ij} are defined by (3.27). We will modify (4.7) step by step until we are able to prove the desired properties.

A first observation regarding (4.7) is that this scheme does not preserve the lake at rest (4.2) if the given velocity is zero, the free surface elevation is constant ($h_i + b_i = H = h_j + b_j$ for all pairs of nodes) but $h_i \neq h_j$ for some $j \neq i$. In this scenario, the flux $d_{ij}(h_j - h_i)$ of the semi-discrete continuity equation (4.7a) will disturb the equilibrium and produce nonphysical waves. This issue can be resolved by replacing (4.7a) with

$$m_i \frac{\mathrm{d}h_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} \Big[d_{ij}(H_j - H_i) - ((h\boldsymbol{v})_j - (h\boldsymbol{v})_i)^{\mathsf{T}} \mathbf{c}_{ij} \Big]. \tag{4.8}$$

This discretization preserves the lake at rest in the case $H \equiv \text{const}$. However, the theory that was used to derive the low order method in Section 3.3.2 does not carry over to systems of balance laws. For the SWE with flat bottom, nonnegativity of water heights follows from the fact that the low order bar states are averaged exact solutions of the Riemann problem [Gue16b]. If source terms are included, the so-defined intermediate states may fail to stay in the admissible set $\mathcal{A}^{\max}$ of the homogeneous SWE. Thus, we need to enforce the nonnegativity constraint for h_i by modifying the discretization of the continuity equation. To this end, we notice that $H_j - H_i = h_j - h_i + (b_j - b_i)$ and introduce a bathymetry limiter $\alpha_{ij}^b \in [0, 1]$ that transforms (4.8) into

$$\begin{aligned} m_i \frac{\mathrm{d}h_i}{\mathrm{d}t} &= \sum_{j \in \mathcal{N}_i \backslash \{i\}} \Big[d_{ij}(h_j - h_i + \alpha_{ij}^b (b_j - b_i)) - ((h\boldsymbol{v})_j - (h\boldsymbol{v})_i)^{\mathsf{T}} \mathbf{c}_{ij} \Big] \\ &= \sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij} \Big[\bar{h}_{ij} - h_i + \frac{\alpha_{ij}^b}{2}(b_j - b_i) \Big] = \sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij}(\bar{h}_{ij}^b - h_i). \end{aligned} \tag{4.9}$$

Here $\bar{h}_{ij}$ is the first component of the usual low order bar state (3.28) and

$$\bar{h}_{ij}^b := \bar{h}_{ij} + \frac{\alpha_{ij}^b}{2}(b_j - b_i). \tag{4.10}$$

The correction factor α_{ij}^b is used to ensure that $\bar{h}_{ij}^b \geq 0$. This condition holds for $\alpha_{ij}^b = 0$ by definition of $\bar{h}_{ij}^b$. However, the largest admissible value of $\alpha_{ij}^b \in [0,1]$ should be employed for consistency reasons. To maintain the conservation property of the semi-discrete continuity equation, we impose the usual symmetry condition $\alpha_{ij}^b = \alpha_{ji}^b$. Note that for $b_j - b_i \geq 0$, the use of $\alpha_{ij}^b = 1$ in (4.10) cannot produce negative $\bar{h}_{ij}^b$ provided that $\bar{h}_{ij} \geq 0$. In this case, however, the limiter may need to act to enforce the condition $\bar{h}_{ji}^b \geq 0$. These considerations lead to the definition

$$\alpha_{ij}^b = \begin{cases} \min\left\{1, \frac{2\bar{h}_{ji}}{b_j - b_i}\right\} & \text{if } b_i - b_j < 0, \\ 1 & \text{if } b_i - b_j = 0, \\ \min\left\{1, \frac{2\bar{h}_{ij}}{b_i - b_j}\right\} & \text{if } b_i - b_j > 0. \end{cases} \tag{4.11}$$

This approach to enforcing the nonnegativity of water heights in the low order method is equivalent to the correction procedure proposed by Audusse et al. [Aud15, Sec. 2.2]. The authors of [Aud15] also impose the conservation and positivity requirements, which yields a 1D finite volume version of our water height limiter based on (4.11).

It is worth checking how the limiter (4.11) behaves for lake at rest configurations.

Lemma 4.2 (Well-balancedness of the positivity-preserving limiter)
Let $u_h \in \mathrm{V}_h^{d+1}$ be a lake at rest solution, i. e., assume that (4.6) holds in addition to $h_i \geq 0$ and $(h\boldsymbol{v})_h = \mathbf{0}$. Then for any $i \in \{1, \ldots, N\}$ and $j \in \mathcal{N}_i \setminus \{i\}$ we have

i) $\alpha_{ij}^b(b_j - b_i) = 0$ if $h_i = h_j = 0$ and

ii) $\alpha_{ij}^b = 1$ if $H_i = H_j$.

In either case, the application of the bathymetry limiter (4.11) does not perturb the lake at rest state because the right hand side of (4.9) is zero for the given data. ◊

Proof:
i) For $h_i = h_j = 0$ and $(h\boldsymbol{v})_h \equiv \mathbf{0}$, the maximum wave speed λ_{ij} is actually not well defined because the calculation of nodal velocities requires division by zero. The wetting and drying models that we present in Section 4.4 ensure that $\boldsymbol{v}_i$ remains finite as h_i goes to zero. If this additional requirement is met, any positive upper bound for λ_{ij} will produce $d_{ij} > 0$ and $\bar{h}_{ij} = \bar{h}_{ji} = \frac{1}{2}(h_i + h_j) = 0$ in the case under consideration. The claim then follows from the definition (4.11) of the correction factor α_{ij}^b.

ii) If $b_i > b_j$ then $2\bar{h}_{ij} = h_i + h_j = 2H_i - (b_i + b_j)$ and $\bar{h}_{ij} - (b_i - b_j)/2 = H_i - b_i = h_i \geq 0$ or, equivalently, $2\bar{h}_{ij}/(b_i - b_j) \geq 1$. Similarly, for $b_i < b_j$ we deduce $\bar{h}_{ji} - (b_j - b_i)/2 = h_j \geq 0$ and $2\bar{h}_{ji}/(b_j - b_i) \geq 1$.

If case i) applies to a pair of nodes $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ then the corresponding term on the right hand side of (4.9) is zero. Otherwise, in case ii), the right hand side term reduces to $d_{ij}(H_j - H_i) = 0$ because $\alpha_{ij}^b = 1$. □

At this stage, one may be tempted to use the spatial semi-discretization consisting of (4.9) and (4.7b) as a low order method for algebraic flux correction. While this version is already usable, it does not yet ensure semi-discrete entropy stability for general bathymetry. For this reason, we modify the momentum equation (4.7b) as follows

$$m_i \frac{\mathrm{d}(h\boldsymbol{v})_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} \Big[d_{ij} \Big((h\boldsymbol{v})_j - (h\boldsymbol{v})_i + \frac{\boldsymbol{v}_i + \boldsymbol{v}_j}{2} \alpha_{ij}^b (b_j - b_i) \Big) - (\mathbf{f}_j^{h\boldsymbol{v}} - \mathbf{f}_i^{h\boldsymbol{v}})\, \mathbf{c}_{ij} - g \frac{h_i + h_j}{2} \alpha_{ij}^b (b_j - b_i)\, \mathbf{c}_{ij} \Big]. \tag{4.12}$$

The term $\frac{1}{2}(\boldsymbol{v}_i + \boldsymbol{v}_j)\alpha_{ij}^b\, d_{ij}(b_j - b_i)$ is included for entropy stabilization purposes. For consistency reasons, we apply the correction factor α_{ij}^b to all bathymetry fluxes.

Remark 4.3
Even though the bathymetry plays the role of a parameter in the SWE model, the correction factor α_{ij}^b adjusts the source term contribution to the momentum equation. The consistency error introduced in this way is acceptable because $\alpha_{ij}^b \neq 1$ is used only for (neighbors of) dry states. A similar concept is used in the popular hydrostatic reconstruction approach [Aud04, Aze17, Ber19] in which topography values are locally adjusted to guarantee nonnegativity of water heights and well-balancedness. ⋄

Let us now discuss how to generalize Tadmor's entropy stability condition (3.68) to our setting and verify it for the low order method. Recall that an entropy pair of the SWE with nonflat topography is given by (cf. Section 2.2.3)

$$\eta(u, b) = \frac{1}{2}\left(gh^2 + h|\boldsymbol{v}|^2\right) + ghb, \qquad \boldsymbol{q}(u, b) = \Big(g(h + b) + \frac{1}{2}|\boldsymbol{v}|^2\Big) h\boldsymbol{v}. \tag{4.13}$$

As of now, our proofs of entropy stability are, in fact, limited to this entropy pair. The entropy variable and potential corresponding to (4.13) read

$$v(u, b) = \begin{bmatrix} g(h + b) - \frac{1}{2}|\boldsymbol{v}|^2 \\ \boldsymbol{v} \end{bmatrix} = v(u, 0) + \begin{bmatrix} gb \\ \mathbf{0} \end{bmatrix}, \qquad \boldsymbol{\psi}(u, b) = \boldsymbol{\psi}(u) = \frac{g}{2} h^2 \boldsymbol{v}. \tag{4.14}$$

A generalized version of Tadmor's entropy stability condition (3.68) was derived by Fjordholm et al. [Fjo11, Sec. 2.1] in the context of finite volume methods for structured grids. Adapting this generalization to our continuous FEM setting, we arrive at

$$\frac{d_{ij}}{2} P_{ij} \leq (\boldsymbol{\psi}_j - \boldsymbol{\psi}_i) \cdot \mathbf{c}_{ij} + \frac{(v(u_i, 0) - v(u_j, 0))^\mathsf{T}}{2} (\mathbf{f}_j + \mathbf{f}_i)\, \mathbf{c}_{ij} + g \left[\frac{h_i + h_j}{2} \frac{\boldsymbol{v}_i + \boldsymbol{v}_j}{2} - \frac{(h\boldsymbol{v})_i + (h\boldsymbol{v})_j}{2} \right]^\mathsf{T} \mathbf{c}_{ij} \alpha_{ij}^b (b_j - b_i) =: Q_{ij}, \tag{4.15}$$

where

$$P_{ij} := \begin{bmatrix} g[h_i - h_j + \alpha_{ij}^b(b_i - b_j)] - \frac{|\boldsymbol{v}_i|^2 - |\boldsymbol{v}_j|^2}{2} \\ \boldsymbol{v}_i - \boldsymbol{v}_j \end{bmatrix}^\mathsf{T} \begin{bmatrix} h_j - h_i + \alpha_{ij}^b(b_j - b_i) \\ (h\boldsymbol{v})_j - (h\boldsymbol{v})_i + \frac{\boldsymbol{v}_i + \boldsymbol{v}_j}{2}\alpha_{ij}^b(b_j - b_i) \end{bmatrix}.$$

Inequality (4.15) imposes the upper bound $(\boldsymbol{\psi}_j - \boldsymbol{\psi}_i) \cdot \mathbf{c}_{ij}$ on the rates of entropy production/dissipation due to low order fluxes and source terms. We now show that the parameters d_{ij} of our low order method (4.9), (4.12) can be chosen sufficiently large (overestimating the maximum speed if necessary) to ensure the validity of (4.15).

Lemma 4.4 (Entropy stability of the low order method)
There exist coefficients $d_{ij} \geq 0$ such that condition (4.15) holds for the numerical fluxes of the semi-discrete low order method defined by (4.9) and (4.12). ⋄

Proof:
Clearly, (4.15) holds for sufficiently large $d_{ij} > 0$ if $P_{ij} = Q_{ij} = 0$ or $P_{ij} < 0$ and Q_{ij} is finite. By definition, we have $P_{ij} = P_{ji}$. A simple calculation reveals that

$$\begin{aligned} P_{ij} =& - g[h_i - h_j + \alpha_{ij}^b(b_i - b_j)]^2 - \frac{|\boldsymbol{v}_i|^2 - |\boldsymbol{v}_j|^2}{2}(h_j - h_i) + (\boldsymbol{v}_i - \boldsymbol{v}_j)^\mathsf{T}((h\boldsymbol{v})_j - (h\boldsymbol{v})_i) \\ =& - g[h_i - h_j + \alpha_{ij}^b(b_i - b_j)]^2 - \frac{1}{2}\left[|\boldsymbol{v}_i|^2(h_j + h_i) + |\boldsymbol{v}_j|^2(h_j + h_i)\right] + \boldsymbol{v}_i^\mathsf{T}\boldsymbol{v}_j\,(h_i + h_j) \\ =& - g[h_i - h_j + \alpha_{ij}^b(b_i - b_j)]^2 - \frac{h_i + h_j}{2}|\boldsymbol{v}_i - \boldsymbol{v}_j|^2 \leq 0. \end{aligned}$$

It is also easy to verify that $Q_{ij} = Q_{ji} = 0$ if $P_{ij} = 0$, which completes the proof. □

To ensure entropy stability of the low order method (4.9), (4.12) in practice, we verify whether the coefficients d_{ij} defined by (3.27) are large enough to satisfy $\frac{d_{ij}}{2}P_{ij} \leq Q_{ij}$. If this is not the case, we set $d_{ij} = d_{ji} = 2\min\{0, Q_{ij}, Q_{ji}\}/P_{ij}$. In practical applications, such adjustments seem to be necessary only in the vicinity of dry states. For such configurations, our approach of increasing the artificial viscosity does not significantly reduce the time step if an appropriate wetting and drying treatment is adopted.

Remark 4.5
As an alternative to adjusting the diffusion coefficients d_{ij}, condition (4.15) can be satisfied by further reducing the value of α_{ij}^b. Indeed, (4.15) reduces to Tadmor's usual entropy stability condition (3.68) for $\alpha_{ij}^b = 0$. A formula to compute such α_{ij}^b can be derived similarly to the IDP pressure fix for the Euler equations, see Section 3.3.5. ⋄

Let us now generalize the setting of Theorem 3.21 to derive local semi-discrete entropy inequalities. First, we rewrite the low order method (4.9), (4.12) as follows

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} [g_{ij} - (\mathbf{f}_j + \mathbf{f}_i)\,\mathbf{c}_{ij} + s_{ij}] + 2\mathbf{f}_i \sum_{j \in \mathcal{N}_i \backslash \{i\}} \mathbf{c}_{ij},$$

where

$$g_{ij} = d_{ij} \begin{bmatrix} h_j - h_i + \alpha_{ij}^b (b_j - b_i) \\ (h\boldsymbol{v})_j - (h\boldsymbol{v})_i + \frac{\boldsymbol{v}_i + \boldsymbol{v}_j}{2} \alpha_{ij}^b (b_j - b_i) \end{bmatrix}, \qquad s_{ij} = \begin{bmatrix} 0 \\ -g \frac{h_i + h_j}{2} \alpha_{ij}^b (b_j - b_i)\, \mathbf{c}_{ij} \end{bmatrix}.$$

Thus, $g_{ij} = -g_{ji}$ for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ and $s_{ij} = s_{ji}$ if $\mathbf{c}_{ij} = -\mathbf{c}_{ji}$.

Theorem 4.6 (Local semi-discrete entropy inequality)
Consider the low order method (4.9), (4.12) satisfying $\frac{d_{ij}}{2} P_{ij} \leq \min\{Q_{ij}, Q_{ji}\}$ for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. Define

$$\begin{aligned} G_{ij} &:= \frac{(v_i + v_j)^\mathsf{T}}{2} g_{ij} + \frac{(v_i - v_j)^\mathsf{T}}{2} [(\mathbf{f}_i - \mathbf{f}_j)\, \mathbf{c}_{ij} + s_{ij}], \\ W_{ij} &:= \frac{g}{2} (1 - \alpha_{ij}^b)(b_i - b_j)[d_{ij}(h_j - h_i + \alpha_{ij}^b (b_j - b_i)) - ((h\boldsymbol{v})_i + (h\boldsymbol{v})_j)^\mathsf{T} \mathbf{c}_{ij}]. \end{aligned}$$

Then for all $i \in \{1, \ldots, N\}$ the semi-discrete entropy inequalities

$$m_i \frac{\mathrm{d}\eta_i}{\mathrm{d}t} \leq \sum_{j \in \mathcal{N}_i \setminus \{i\}} [G_{ij} + W_{ij} - (\boldsymbol{q}_j - \boldsymbol{q}_i) \cdot \mathbf{c}_{ij}] \tag{4.16}$$

hold w. r. t. the entropy pair $(\eta, \boldsymbol{q})$ defined by (4.13). ◊

Proof:
We essentially generalize the proof of Theorem 3.21. To exploit (4.15), we split the nodal entropy variable v_i into $\frac{1}{2}(v_i + v_j)$ and $\frac{1}{2}(v_i - v_j)$ as in [Tad03, Fjo11, Kuz20c]. The fluxes g_{ij} are antisymmetric, whereas the sources s_{ij} are symmetric by definition. Multiplication by $\frac{1}{2}(v_i + v_j)$ preserves these properties, whereas multiplication by $\frac{1}{2}(v_i - v_j)$ swaps them for fluxes and sources in the following identity

$$\begin{aligned} m_i v_i^\mathsf{T} \frac{\mathrm{d}u_i}{\mathrm{d}t} = &\sum_{j \in \mathcal{N}_i \setminus \{i\}} \left(\frac{(v_i + v_j)^\mathsf{T}}{2} [g_{ij} - (\mathbf{f}_j + \mathbf{f}_i)\, \mathbf{c}_{ij}] + \frac{(v_i - v_j)^\mathsf{T}}{2} s_{ij} \right) + 2 v_i^\mathsf{T} \mathbf{f}_i \sum_{j \in \mathcal{N}_i \setminus \{i\}} \mathbf{c}_{ij} \\ &+ \sum_{j \in \mathcal{N}_i \setminus \{i\}} \left(\frac{(v_i - v_j)^\mathsf{T}}{2} [g_{ij} - (\mathbf{f}_j + \mathbf{f}_i)\, \mathbf{c}_{ij}] + \frac{(v_i + v_j)^\mathsf{T}}{2} s_{ij} \right). \end{aligned} \tag{4.17}$$

Before applying (4.15) to the last term on the right hand side of (4.17), we need to account for the influence of α_{ij}^b in the continuity equation. To this end, we recall the definition of entropy variables for the SWE (4.14) and split the symmetric terms in (4.17) as follows

$$\begin{aligned} &\frac{(v_i - v_j)^\mathsf{T}}{2} [g_{ij} - (\mathbf{f}_j + \mathbf{f}_i)\, \mathbf{c}_{ij}] + \frac{(v_i + v_j)^\mathsf{T}}{2} s_{ij} \\ &= \frac{d_{ij}}{2} \begin{bmatrix} g[h_i - h_j + \alpha_{ij}^b (b_i - b_j)] - \frac{1}{2}(|\boldsymbol{v}_i|^2 - |\boldsymbol{v}_j|^2) \\ \boldsymbol{v}_i - \boldsymbol{v}_j \end{bmatrix}^\mathsf{T} \begin{bmatrix} h_j - h_i + \alpha_{ij}^b (b_j - b_i) \\ (h\boldsymbol{v})_j - (h\boldsymbol{v})_i + \frac{\boldsymbol{v}_i + \boldsymbol{v}_j}{2} \alpha_{ij}^b (b_j - b_i) \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
&+ \frac{d_{ij}}{2}\left[g(1-\alpha_{ij}^b)(b_i-b_j)\right]\left[h_j - h_i + \alpha_{ij}^b(b_j-b_i)\right] \\
&- \frac{(v(u_i,0)-v(u_j,0))^\mathsf{T}}{2}(\mathbf{f}_j+\mathbf{f}_i)\,\mathbf{c}_{ij} - \frac{(\boldsymbol{v}_i+\boldsymbol{v}_j)^\mathsf{T}}{2}\,\mathbf{c}_{ij}\,g\frac{h_i+h_j}{2}\alpha_{ij}^b(b_j-b_i) \\
&- \frac{g\alpha_{ij}^b(b_i-b_j)}{2}((h\boldsymbol{v})_i+(h\boldsymbol{v})_j))^\mathsf{T}\mathbf{c}_{ij} - \frac{g(1-\alpha_{ij}^b)(b_i-b_j)}{2}((h\boldsymbol{v})_i+(h\boldsymbol{v})_j)^\mathsf{T}\mathbf{c}_{ij}.
\end{aligned}
$$

All terms on the right hand side of this identity can be found in W_{ij} or in the entropy stability condition (4.15), which we invoke in the next step. Recalling the definitions of P_{ij} and Q_{ij} for (4.15), we estimate the sum of symmetric terms using the inequality

$$
\frac{(v_i-v_j)^\mathsf{T}}{2}[g_{ij} - (\mathbf{f}_j+\mathbf{f}_i)\,\mathbf{c}_{ij}] + \frac{(v_i+v_j)^\mathsf{T}}{2}s_{ij} \le (\boldsymbol{\psi}_j - \boldsymbol{\psi}_i)\cdot\mathbf{c}_{ij} + W_{ij}.
$$

In combination with (4.17), this stability estimate implies

$$
\begin{aligned}
m_i\frac{\mathrm{d}\eta(u_i)}{\mathrm{d}t} \le{}& \sum_{j\in\mathcal{N}_i\setminus\{i\}}\left(\frac{(v_i+v_j)^\mathsf{T}}{2}[g_{ij}-(\mathbf{f}_j+\mathbf{f}_i)\,\mathbf{c}_{ij}] + \frac{(v_i-v_j)^\mathsf{T}}{2}s_{ij}\right) + 2v_i^\mathsf{T}\mathbf{f}_i\sum_{j\in\mathcal{N}_i\setminus\{i\}}\mathbf{c}_{ij} \\
&+ \sum_{j\in\mathcal{N}_i\setminus\{i\}}\left[(v_j^\mathsf{T}\mathbf{f}_j - \boldsymbol{q}_j - v_i^\mathsf{T}\mathbf{f}_i + \boldsymbol{q}_i)\cdot\mathbf{c}_{ij} + W_{ij}\right] \\
={}& \sum_{j\in\mathcal{N}_i\setminus\{i\}}\left[G_{ij} + W_{ij} - (\boldsymbol{q}_j-\boldsymbol{q}_i)\cdot\mathbf{c}_{ij}\right].
\end{aligned}
$$

The last identity is obtained exactly as in the proof of Theorem 3.21 (see [Kuz22a, Sec. 4.1] for details). □

The consistency errors W_{ij} can be attributed to the occurrence of dry or almost dry areas, which require the use of $\alpha_{ij}^b < 1$. For such states, even the validity of the continuous entropy inequality is questionable because the momentum equation of the SWE model does not describe the underlying physics correctly. In particular, the absence of friction terms becomes an issue. This argument justifies the presence of W_{ij} in (4.16).

Corollary 4.7 (Global semi-discrete entropy inequality)
Let the assumptions of Theorem 4.6 be fulfilled and assume that the spatial semi-discretization (4.9), (4.12) is boundary indifferent (see Definition 3.23) w. r. t. $u_h(t)$ for all $t \ge 0$. If, in addition, $\alpha_{ij}^b = 1$ for all $i \in \{1,\dots,N\}$, $j \in \mathcal{N}_i \setminus \{i\}$, then the following semi-discrete entropy inequality holds

$$
\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega\Big(\sum_{i=1}^N \eta_i\,\varphi_i\Big)\,\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega}\Big(\sum_{i=1}^N \boldsymbol{q}_i\,\varphi_i\Big)\cdot\boldsymbol{n}\,\mathrm{d}s \le 0. \qquad \diamond
$$

Proof:
By assumption, we have $W_{ij} = 0$ and $(v_i - v_j)^\mathsf{T} s_{ij} = -(v_j - v_i)^\mathsf{T} s_{ji}$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. Therefore, the claim follows as in the proof of Corollary 3.25. □

We conclude the discussion of the low order method by formulating the bar state form of the momentum equation (4.12), which reads

$$m_i \frac{\mathrm{d}(h\boldsymbol{v})_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij} \Big(\overline{(h\boldsymbol{v})}^b_{ij} - (h\boldsymbol{v})_i \Big).$$

Similarly to the bar state (4.10) of the water height, the bar state of the discharge

$$\begin{aligned} \overline{(h\boldsymbol{v})}^b_{ij} &= \overline{(h\boldsymbol{v})}_{ij} + \frac{\boldsymbol{v}_i + \boldsymbol{v}_j}{4} \alpha^b_{ij} (b_j - b_i) - \frac{g \frac{h_i + h_j}{2} \alpha^b_{ij} (b_j - b_i)\, \mathbf{c}_{ij}}{2d_{ij}} \\ &= \frac{(h\boldsymbol{v})_i + (h\boldsymbol{v})_j + \frac{\boldsymbol{v}_i + \boldsymbol{v}_j}{2} \alpha^b_{ij} (b_j - b_i)}{2} - \frac{\Big(\mathbf{f}^{h\boldsymbol{v}}_j - \mathbf{f}^{h\boldsymbol{v}}_i + g \frac{h_i + h_j}{2} \alpha^b_{ij} (b_j - b_i)\, \mathcal{I} \Big) \mathbf{c}_{ij}}{2d_{ij}} \end{aligned}$$

consists of symmetric and skew-symmetric terms.

Remark 4.8
A similar concept based on intermediate states is employed in the work of Audusse et al. [Aud15]. We already mentioned the fact that their limiter for the water height [Aud15, Sec. 2.2] is equivalent to (4.11). A minor difference between our low order method and the well-balanced scheme of Audusse et al. is that their finite volume method employs intermediate states based on the HLL Riemann solver [Har83b] instead of local Lax–Friedrichs-type bar states. More importantly, our discretization of the momentum equation includes an entropy-stabilizing term, which is missing in [Aud15]. The absence of this term might be the reason why no conclusive evidence regarding the validity of discrete entropy inequalities could be provided in [Aud15, Sec. 2.4]. ◇

4.3.2 Monolithic convex limiting

Having derived the low order method (4.9), (4.12), we now discuss the MCL methodology for the SWE with topography. As in the case of conservation laws, we first need to define the raw antidiffusive fluxes $f_{ij} = -f_{ji} \in \mathbb{R}^{d+1}$, $f_{ij} := (f^h_{ij}, (\boldsymbol{f}^{h\boldsymbol{v}}_{ij})^\mathsf{T})^\mathsf{T}$ with which the target scheme (4.5) can be recovered from the low order method. Most of the considerations discussed in Section 3.3.3 apply here as well. However, we have to include additional terms due to modifications that make our low order method property preserving for nonflat topography. A straightforward computation shows that if $\alpha^b_{ij} = 1$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$, then (4.5) can be recovered via

$$m_i \frac{\mathrm{d}h_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} \Big[2d_{ij} (\bar{h}^b_{ij} - h_i) + f^h_{ij} \Big],$$

$$m_i \frac{\mathrm{d}(h\boldsymbol{v})_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \backslash \{i\}} \Big[2d_{ij} \Big(\overline{(h\boldsymbol{v})}^b_{ij} - (h\boldsymbol{v})_i \Big) + \boldsymbol{f}^{h\boldsymbol{v}}_{ij} \Big],$$

where

$$\begin{aligned} f^h_{ij} &= m_{ij} \big(\dot{h}_i - \dot{h}_j \big) + d_{ij} \left[h_i - h_j + \alpha^b_{ij} (b_i - b_j) \right], \\ \boldsymbol{f}^{h\boldsymbol{v}}_{ij} &= m_{ij} \left((\dot{h\boldsymbol{v}})_i - (\dot{h\boldsymbol{v}})_j \right) + d_{ij} \Big[(h\boldsymbol{v})_i - (h\boldsymbol{v})_j + \frac{\boldsymbol{v}_i + \boldsymbol{v}_j}{2} \alpha^b_{ij} (b_i - b_j) \Big]. \end{aligned}$$

In our fully discrete scheme, we once more define the dotted quantities as low order time derivatives, which are computed from (4.9) and (4.12), respectively. For steady state problems, these quantities are set to zero as discussed in Section 3.3.3.

We are now in a position to present the generalized sequential limiting technique with which we obtain flux-corrected counterparts $f^{h,*}_{ij}$ and $\boldsymbol{f}^{h\boldsymbol{v},*}_{ij}$ of f^h_{ij} and $\boldsymbol{f}^{h\boldsymbol{v}}_{ij}$. In the first step of the sequential MCL algorithm, we limit the water height using

$$h^{\min}_i := \min_{j \in \mathcal{N}_i \backslash \{i\}} \bar{h}^b_{ij}, \qquad h^{\max}_i := \max_{j \in \mathcal{N}_i \backslash \{i\}} \bar{h}^b_{ij}$$

as local bounds of numerical admissibility conditions, which imply global positivity preservation for the water height if the bathymetry correction factor α^b_{ij} defined by (4.11) is applied. The limiting formula for the raw antidiffusive fluxes f^h_{ij} becomes

$$f^{h,*}_{ij} = \begin{cases} \min \Big\{ f^h_{ij},\ 2d_{ij} \min \Big\{ h^{\max}_i - \bar{h}^b_{ij}, \bar{h}^b_{ji} - h^{\min}_j \Big\} \Big\} & \text{if } f^h_{ij} \geq 0, \\ \max \Big\{ f^h_{ij},\ 2d_{ij} \max \Big\{ h^{\min}_i - \bar{h}^b_{ij}, \bar{h}^b_{ji} - h^{\max}_j \Big\} \Big\} & \text{if } f^h_{ij} \leq 0. \end{cases} \tag{4.18}$$

The corresponding flux-corrected bar states in the continuity equation can be written as

$$\bar{h}^{b,*}_{ij} = \bar{h}^b_{ij} + \frac{f^{h,*}_{ij}}{2d_{ij}} = \bar{h}_{ij} + \frac{\alpha^b_{ij}(b_j - b_i)}{2} + \frac{f^{h,*}_{ij}}{2d_{ij}} = \bar{h}^*_{ij} + \frac{\alpha^b_{ij}(b_j - b_i)}{2}.$$

Next, we need to limit $\boldsymbol{f}^{h\boldsymbol{v}}_{ij}$ in a way that ensures the validity of numerical admissibility conditions for individual velocity (rather that discharge) components. To construct local bounds for this step, we first define the velocity bar states as

$$\bar{\boldsymbol{v}}_{ij} := \frac{\overline{(h\boldsymbol{v})}^b_{ij} + \overline{(h\boldsymbol{v})}^b_{ji}}{\bar{h}^b_{ij} + \bar{h}^b_{ji}} = \frac{2d_{ij} \Big(\overline{(h\boldsymbol{v})}_{ij} + \overline{(h\boldsymbol{v})}_{ji} \Big) - g \frac{h_i + h_j}{2} \alpha^b_{ij} (b_j - b_i)(\mathbf{c}_{ij} - \mathbf{c}_{ji})}{2d_{ij}(\bar{h}_{ij} + \bar{h}_{ji})} = \bar{\boldsymbol{v}}_{ji},$$

which represents a generalization of (3.51). Note that the components $\frac{1}{2}\alpha^b_{ij}(b_j - b_i)$ and $\frac{1}{4}(\boldsymbol{v}_i + \boldsymbol{v}_j)\alpha^b_{ij}(b_j - b_i)$ of the bar states $\bar{h}^b_{ij}$, $\overline{(h\boldsymbol{v})}^b_{ij}$ and the corresponding antisymmetric components of the bar states $\bar{h}^b_{ji}$, $\overline{(h\boldsymbol{v})}^b_{ji}$ cancel out upon summation in the numerator and denominator of the second ratio. The symmetric source terms add up in the numerator.

Let $\boldsymbol{v}_i^{\min}$, $\boldsymbol{v}_i^{\max} \in \mathbb{R}^d$ be vectors containing local bounds to be imposed on individual components of the nodal velocity (alternative limiting strategies for vector fields are discussed in [Haj19]). Inequalities involving vectors should be understood componentwise. As before, we limit the bar states of the momentum equation as follows

$$\bar{h}_{ij}^* \boldsymbol{v}_i^{\min} \le \overline{(h\boldsymbol{v})}_{ij}^{b,*} := \overline{(h\boldsymbol{v})}_{ij}^{b} + \frac{\boldsymbol{f}_{ij}^{h\boldsymbol{v},*}}{2d_{ij}} = \bar{h}_{ij}^* \bar{\boldsymbol{v}}_{ij} + \frac{\boldsymbol{g}_{ij}^{h\boldsymbol{v},*}}{2d_{ij}} \le \bar{h}_{ij}^* \boldsymbol{v}_i^{\max}, \tag{4.19}$$

where $\boldsymbol{g}_{ij}^{h\boldsymbol{v},*}$ is a limited counterpart of the flux

$$\boldsymbol{g}_{ij}^{h\boldsymbol{v}} = \boldsymbol{f}_{ij}^{h\boldsymbol{v}} + 2d_{ij}\Big(\overline{(h\boldsymbol{v})}_{ij}^{b} - \bar{h}_{ij}^* \bar{\boldsymbol{v}}_{ij}\Big).$$

It is easy to verify that $\boldsymbol{g}_{ij}^{h\boldsymbol{v}} + \boldsymbol{g}_{ji}^{h\boldsymbol{v}} = \boldsymbol{0}$ by construction. This property must be preserved by the flux limiter. From (4.19) we derive the flux constraints for

$$\boldsymbol{g}_{ij}^{h\boldsymbol{v},*} = \begin{cases} \min\Big\{\boldsymbol{g}_{ij}^{h\boldsymbol{v}},\ 2d_{ij}\min\Big\{\bar{h}_{ij}^*\left(\boldsymbol{v}_i^{\max} - \bar{\boldsymbol{v}}_{ij}\right),\ \bar{h}_{ji}^*\Big(\bar{\boldsymbol{v}}_{ij} - \boldsymbol{v}_j^{\min}\Big)\Big\}\Big\} & \text{if } \boldsymbol{g}_{ij}^{h\boldsymbol{v}} \ge 0, \\ \max\Big\{\boldsymbol{g}_{ij}^{h\boldsymbol{v}},\ 2d_{ij}\max\Big\{\bar{h}_{ij}^*\left(\boldsymbol{v}_i^{\min} - \bar{\boldsymbol{v}}_{ij}\right),\ \bar{h}_{ji}^*\Big(\bar{\boldsymbol{v}}_{ij} - \boldsymbol{v}_j^{\max}\Big)\Big\}\Big\} & \text{if } \boldsymbol{g}_{ij}^{h\boldsymbol{v}} \le 0. \end{cases}$$

Note that this formula applies a scalar limiter of the form (3.55) to each component of the flux vector $\boldsymbol{g}_{ij}^{h\boldsymbol{v}}$. Finally, we obtain the flux-corrected momentum bar states via

$$\boldsymbol{f}_{ij}^{h\boldsymbol{v},*} = \boldsymbol{g}_{ij}^{h\boldsymbol{v},*} - 2d_{ij}\Big(\overline{(h\boldsymbol{v})}_{ij}^{b} - \bar{h}_{ij}^* \bar{\boldsymbol{v}}_{ij}\Big), \qquad \overline{(h\boldsymbol{v})}_{ij}^{b,*} = \overline{(h\boldsymbol{v})}_{ij}^{b} + \frac{\boldsymbol{f}_{ij}^{h\boldsymbol{v},*}}{2d_{ij}}.$$

What remains is to choose feasible bounds $\boldsymbol{v}_i^{\min}$, $\boldsymbol{v}_i^{\max}$. As before, we should include the generalized velocity bar states $\bar{\boldsymbol{v}}_{ij}$ in their definition. To prove that a subsequent entropy fix based on (4.15) cannot cause violations of (4.19), we need to extend the bounds by including the states $\overline{(h\boldsymbol{v})}_{ij}^{b}/\bar{h}_{ij}$ (cf. Lemma 3.16). We remark that violations of local bounds for velocity components are not critical and can be interpreted as automatic adjustment of numerical admissibility conditions. To strictly enforce local bounds without extending them in one way or another, the entropy fix should be applied to the raw antidiffusive fluxes *before* passing them to the bound-preserving limiter. Our numerical experiments confirmed that this approach is viable but we did not pursue it further for two reasons. First, we prefer our approach to be consistent with the algorithm for conservation laws if the bottom is flat. Second, we found that the entropy limiter for systems did not significantly modify the approximations obtained with the sequential MCL scheme. This observation suggests that entropy stabilization may be unnecessary (at least in the present context regarding approximations to the SWE) if the scheme yields provably bound-preserving solutions. On the other hand, using a single correction factor for all components of the given fluxes, as our entropy limiter does, may produce overly diffusive results if such corrections are performed prior to enforcing numerical admissibility constraints. Therefore, we extend the velocity bounds

in such a way that additional limiting does not destroy any of the constraints enforced previously.

The statement of Lemma 3.16 remains valid if the velocity bounds are defined by

$$\boldsymbol{v}_i^{\min} := \min_{j\in\mathcal{N}_i\backslash\{i\}} \min\left\{\bar{\boldsymbol{v}}_{ij},\ \frac{\overline{(h\boldsymbol{v})}_{ij}^{b}}{\bar{h}_{ij}}\right\}, \qquad \boldsymbol{v}_i^{\max} := \max_{j\in\mathcal{N}_i\backslash\{i\}} \max\left\{\bar{\boldsymbol{v}}_{ij},\ \frac{\overline{(h\boldsymbol{v})}_{ij}^{b}}{\bar{h}_{ij}}\right\}.$$

Note that, contrary to the sequential approach for the SWE with flat bottom, the states $\overline{(h\boldsymbol{v})}_{ij}^{b}/\bar{h}_{ij}$ are neither symmetric nor antisymmetric in general. Therefore, they need to be computed for all bar states, even the ones corresponding to pairs of interior nodes.

4.3.3 Semi-discrete entropy fix

At the current design stage, the semi-discrete bound-preserving scheme reads

$$\begin{aligned} m_i\frac{\mathrm{d}h_i}{\mathrm{d}t} &= \sum_{j\in\mathcal{N}_i\backslash\{i\}}\left[2d_{ij}(\bar{h}_{ij}^{b}-h_i)+f_{ij}^{h,*}\right] = \sum_{j\in\mathcal{N}_i\backslash\{i\}} 2d_{ij}(\bar{h}_{ij}^{b,*}-h_i),\\ m_i\frac{\mathrm{d}(h\boldsymbol{v})_i}{\mathrm{d}t} &= \sum_{j\in\mathcal{N}_i\backslash\{i\}}\left[2d_{ij}\Big(\overline{(h\boldsymbol{v})}_{ij}^{b}-(h\boldsymbol{v})_i\Big)+\boldsymbol{f}_{ij}^{h\boldsymbol{v},*}\right] = \sum_{j\in\mathcal{N}_i\backslash\{i\}} 2d_{ij}\Big(\overline{(h\boldsymbol{v})}_{ij}^{b,*}-(h\boldsymbol{v})_i\Big). \end{aligned}$$

To enforce a semi-discrete entropy inequality, we employ limiting coefficients $\beta_{ij} = \beta_{ji} \in [0,1]$ and entropy limited fluxes $f_{ij}^{**} = \beta_{ij}f_{ij}^{*} = \beta_{ij}(f_{ij}^{h,*}, (\boldsymbol{f}_{ij}^{h\boldsymbol{v},*})^{\mathsf{T}})^{\mathsf{T}}$. Our approach represents a straightforward generalization of the entropy limiter used for conservation laws. We adjust the Rusanov coefficients d_{ij} if necessary to guarantee that the low order method corresponding to $f_{ij}^{*}=0$ satisfies the entropy stability condition $\frac{d_{ij}}{2}P_{ij} \le Q_{ij}$, i. e., (4.15). The flux-corrected scheme with f_{ij}^{*} replaced by f_{ij}^{**} is entropy stable if

$$\frac{d_{ij}}{2}P_{ij} + \frac{\beta_{ij}}{2}R_{ij} \le Q_{ij}, \tag{4.20}$$

where

$$R_{ij} := \begin{bmatrix} g[h_i - h_j + \alpha_{ij}^{b}(b_i-b_j) - \frac{1}{2}(|\boldsymbol{v}_i|^2 - |\boldsymbol{v}_j|^2)] \\ \boldsymbol{v}_i - \boldsymbol{v}_j \end{bmatrix}^{\mathsf{T}} f_{ij}^{*} = R_{ji}.$$

Thus, we enforce (4.20) similarly to (3.74) by setting

$$\beta_{ij} = \begin{cases} \dfrac{2\min\{Q_{ij},Q_{ji}\} - d_{ij}P_{ij}}{R_{ij}} & \text{if } R_{ij} > 2\min\{Q_{ij},Q_{ji}\} - d_{ij}P_{ij},\\ 1 & \text{otherwise.} \end{cases} \tag{4.21}$$

Since the Rusanov coefficients d_{ij} are chosen large enough for $2\min\{Q_{ij},Q_{ji}\} \ge d_{ij}P_{ij}$ to hold, (4.21) produces $\beta_{ij} = \beta_{ji} \in [0,1]$.

The generalization of our monolithic limiting strategies for the SWE with topography is now complete. Written in terms of the flux-corrected bar states

$$\bar{h}_{ij}^{b,**} = \bar{h}_{ij}^{b} + \frac{\beta_{ij} f_{ij}^{h,*}}{2d_{ij}}, \qquad \overline{(h\boldsymbol{v})}_{ij}^{b,**} = \overline{(h\boldsymbol{v})}_{ij}^{b} + \frac{\beta_{ij} \boldsymbol{f}_{ij}^{h\boldsymbol{v},*}}{2d_{ij}},$$

the resulting semi-discrete method reads

$$\begin{aligned} m_i \frac{\mathrm{d}h_i}{\mathrm{d}t} &= \sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij} \Big(\bar{h}_{ij}^{b,**} - h_i \Big), \\ m_i \frac{\mathrm{d}(h\boldsymbol{v})_i}{\mathrm{d}t} &= \sum_{j \in \mathcal{N}_i \backslash \{i\}} 2d_{ij} \Big(\overline{(h\boldsymbol{v})}_{ij}^{b,**} - (h\boldsymbol{v})_i \Big). \end{aligned}$$

By construction, this finite element method is provably well balanced, bound preserving, and entropy stable. We summarize its properties in the following theorem.

Theorem 4.9 (Properties of flux correction schemes for (4.1)**)**
The low order method and the flux-corrected schemes presented in this section

i) reduce to the corresponding algorithms discussed in Section 3.3 if applied to the shallow water equations with flat topography,

ii) are well balanced for the lake at rest in the sense of Lemma 4.2,

iii) produce nonnegative water heights under the CFL condition (3.43), and

iv) satisfy the semi-discrete entropy inequalities (4.16) w. r. t. the entropy pair (4.13) if the correction factors β_{ij} are either zero (low order method) or calculated using (4.21). In the flux-corrected version of the scheme, the numerical fluxes G_{ij} and consistency errors W_{ij} appearing in (4.16) are replaced with

$$\begin{aligned} G_{ij}^{*} &:= G_{ij} + \frac{\beta_{ij}}{2} \Big(\Big[g(h_i + b_i + h_j + b_j) - \frac{1}{2}(|\boldsymbol{v}_i|^2 + |\boldsymbol{v}_j|^2) \Big] f_{ij}^{h,*} + (\boldsymbol{v}_i + \boldsymbol{v}_j)^{\mathsf{T}} \boldsymbol{f}_{ij}^{h\boldsymbol{v},*} \Big), \\ W_{ij}^{*} &:= W_{ij} + \frac{g}{2}(1 - \alpha_{ij}^{b})(b_i - b_j)\beta_{ij} f_{ij}^{h,*}, \end{aligned}$$

respectively. Moreover, the statement of Corollary 4.7 remains valid. ◊

Proof:
i) For $b \equiv \text{const}$ the modified bar states of the water height, discharge, and velocity reduce to those of the SWE without topography term. The antidiffusive fluxes are also limited in the same way. Since $b \equiv 0$ is used for flat bathymetry by convention (see Section 2.2.3), the entropy limiter (4.21) is equivalent to (3.74) (see also Remark 4.5).

ii) The low order nodal time derivatives that we use in the definition of antidiffusive fluxes are easily checked to be zero (in fact, the consistent Galerkin time derivatives vanish too). Owing to Lemma 4.2, the application of α_{ij}^b does not interfere with the well-balancedness property. In a similar fashion, we observe that the raw antidiffusive fluxes f_{ij}^h and $\boldsymbol{f}_{ij}^{h\boldsymbol{v}}$ are exactly zero. Thus, the the scheme reduces to the low order method (4.9), (4.12), which is well balanced w. r. t. the lake at rest as shown earlier.

iii) The low order bar states $\bar{h}_{ij}^b$ are nonnegative by definition (4.11) of α_{ij}^b for the bathymetry fix. Since the bounds $h_i^{\min}$ are defined by these states, they are also nonnegative. If the flux $f_{ij}^{h,*}$ is nonnegative, then so is $\bar{h}_{ij}^{b,*}$. Otherwise, the validity of

$$\bar{h}_{ij}^{b,*} = \bar{h}_{ij}^b + \frac{f_{ij}^{h,*}}{2d_{ij}} \geq \bar{h}_{ij}^b + \frac{2d_{ij}(h_i^{\min} - \bar{h}_{ij}^b)}{2d_{ij}} = h_i^{\min} \geq 0$$

for the MCL scheme follows from (4.18). Under the CFL condition (3.43), the nodal state h_i is thus a convex combination of nonnegative states.

iv) To rigorously prove the claim, one has to adapt the arguments used for the low order fluxes in the proof of Theorem 4.6 by including the antisymmetric term $\beta_{ij} f_{ij}^*$. The semi-discrete entropy inequality is derived using the same splitting of the entropy variable v_i. The presence of limited antidiffusive fluxes produces three additional terms, two of which modify the numerical fluxes G_{ij} and consistency errors W_{ij} in the claimed manner. The third additional term $\frac{\beta_{ij}}{2} R_{ij}$ is absorbed into the modified entropy stability condition (4.20) that is enforced by the entropy limiter (4.21). □

4.4 Wetting and drying algorithms

Before moving on to numerical examples, we still need to specify how to handle dry and almost dry regions numerically. To this end, we first discuss some wetting and drying algorithms selected from the literature. Then we present our own approach.

Loosely speaking, Ricchiuto and Bollermann [Ric09, Sec. 4.3] set the velocity to zero if the water height is smaller than a prescribed tolerance for which they use the square of the normalized mesh size $(h/|\Omega|)^2$. Admittedly, their wetting and drying approach is more involved. In particular, it incorporates information on the topography slope in wet-dry transition regions. We have not tested this part of their algorithm but ran experiments with the version that sets the velocity to zero in dry regions. Besides the fact that this approach interferes with the principle of continuous dependence on the data, this nodal fix does not perform very well for our schemes.

A velocity fix that does guarantee continuous dependence on data is given by

$$\tilde{\boldsymbol{v}} = \frac{2h(h\boldsymbol{v})}{h^2 + \max\{h, \varepsilon\}^2}, \tag{4.22}$$

where $\varepsilon \ll 1$. Azerad et al. [Aze17] use $\varepsilon = 10^{-16} \max_{\boldsymbol{x}\in\overline{\Omega}} h_0(\boldsymbol{x})$ in this formula. A problem with this approach is that if the water height approaches zero but the discharge does not, the use of (4.22) produces velocities with magnitudes that tend to infinity, resulting in unrealistic CFL conditions and a blowup of kinetic energy.

Kurganov and Petrova suggest a similar fix [Kur07a, Eqs. (2.17), (2.21)] in which the velocity is computed via

$$\tilde{\boldsymbol{v}} = \frac{\sqrt{2}h(h\boldsymbol{v})}{\sqrt{h^4 + \max\{h,\varepsilon\}^4}} \tag{4.23}$$

and the parameter ε is set equal to the (normalized) mesh size. In our experience, this choice introduces significant approximation errors because the mesh size is usually much larger than the thickness of a water layer that can be considered as dry. On the other hand, this fix seems to be quite robust in practice. Importantly, Kurganov and Petrova [Kur07a, Eq. (2.21)] emphasize the need for adjusting the discharge by setting $(h\boldsymbol{v}) = h\tilde{\boldsymbol{v}}$ after calculating the velocity via (4.23).

Many more algorithms for wetting and drying processes exist besides the ones already mentioned. Most of them work in a fashion similar to the approaches discussed above. There are also schemes that can not directly be applied in the context of continuous finite elements. For instance, Vater et al. [Vat15] employ slope limiters to handle wetting and drying scenarios.

Let us now discuss a new nodal velocity correction based on the entropy of the shallow water system. Here we restrict ourselves to the case of a flat topography because it is currently unclear to us whether an extension to the general case is feasible for our MCL schemes. The underlying idea is based on the observation that unbounded velocities, which may occur in dry or nearly dry areas, result in blow ups of the kinetic energy and, therefore, of the entropy. On the other hand, entropy analysis of the bar state form for the SWE with flat bathymetry provides an upper bound for the nodal entropy. Violations of this bound (and the resulting lack of discrete entropy stability in practice) are caused not by the discretization but by the numerically unstable calculation of nodal velocities for the next step. Thus, for entropy stability reasons, the magnitudes of nodal velocities should stay bounded in the vicinity of dry states. In physics, this property is enforced by viscous friction, which is missing in our model.

Recall once more that for $b \equiv 0$, an entropy for the shallow water equations is the sum of potential and kinetic energies (cf. (4.13)). By convexity, the entropy of the state $\tilde{u}_i = (\tilde{h}_i, \widetilde{(h\boldsymbol{v})}_i)$ produced by a forward Euler update satisfies the estimate

$$\begin{aligned}\eta(\tilde{u}_i) &= \eta\Big(\Big(1 - \frac{\Delta t}{m_i}\sum_{j\in\mathcal{N}_i\setminus\{i\}} 2d_{ij}\Big)u_i + \frac{\Delta t}{m_i}\sum_{j\in\mathcal{N}_i\setminus\{i\}} 2d_{ij}\bar{u}_{ij}^{**}\Big) \\ &\le \Big(1 - \frac{\Delta t}{m_i}\sum_{j\in\mathcal{N}_i\setminus\{i\}} 2d_{ij}\Big)\eta(u_i) + \frac{\Delta t}{m_i}\sum_{j\in\mathcal{N}_i\setminus\{i\}} 2d_{ij}\eta(\bar{u}_{ij}^{**}) =: \eta_i^{\max}\end{aligned} \tag{4.24}$$

under the CFL condition (3.43). The value $\eta_i^{\max}$ can now be used to prohibit the occurrence of unbounded velocities that would lead to a violation of (4.24). Invoking the definition (4.13) of η, we enforce (4.24) by adjusting the nodal velocities as follows

$$\tilde{\boldsymbol{v}}_i = \begin{cases} \frac{\widetilde{(h\boldsymbol{v})}_i}{\tilde{h}_i} & \text{if } |(\widetilde{h\boldsymbol{v}})_i| \le h_i Q_i, \\ \frac{Q_i}{|\widetilde{(h\boldsymbol{v})}_i|}(\widetilde{h\boldsymbol{v}})_i & \text{if } |(\widetilde{h\boldsymbol{v}})_i| > h_i Q_i, \end{cases} \qquad \text{where} \qquad Q_i := \sqrt{\frac{2\eta_i^{\max}}{\tilde{h}_i} - g\tilde{h}_i}.$$

We then follow Kurganov and Petrova [Kur07a, Eq. (2.21)] and overwrite the nodal discharge by $\tilde{h}_i\tilde{\boldsymbol{v}}_i$. The approach presented here for a forward Euler update directly carries over to other SSP RK methods, which are convex combinations of forward Euler steps. Unfortunately, the entropy-based approach interferes with the well-balancedness property for the lake at rest unless the topography is flat.

To keep our scheme well balanced, we developed a wetting and drying algorithm that is based on the theory of laminar boundary layers (see for instance [Sch17]). As suggested by the above discussion of our entropy-based approach, a particular challenge for realistic treatment of wetting and drying processes is to obtain a physically correct model for the velocities in wet-dry transition regions. According to the boundary layer theory, viscous friction effects should not be neglected in these areas. For the SWE in particular, a bottom friction term should be incorporated into the system. A derivation of the viscous SWE including bottom friction can be found in [Ger00]. In essence, a nonconservative term $\sigma\boldsymbol{v}$ is added on the left hand side of the momentum equation. Here $\sigma > 0$ is the bottom friction coefficient, which may generally depend on the solution and parameters of the SWE. Particular models for σ are discussed, for instance, in [Vre94, Sec. 2.7] and [Cus11, Sec. 9.8]. Physical intuition tells us that wet-dry transitions occur in a boundary layer of thickness $0 < \delta \ll 1$. According to [Sch17, Ch. 2], one may assume that inertial and viscous forces are in equilibrium and contributions of the material derivative can be neglected in the boundary layer, which in our case implies

$$gh\nabla H + \sigma\boldsymbol{v} = 0.$$

For nodes belonging to wet-dry zones, i. e., for $h_i \le \delta$, we use this identity to compute a nodal boundary layer velocity $\boldsymbol{v}_i^{\mathrm{BL}}$ via the lumped L^2 projection

$$m_i\boldsymbol{v}_i^{\mathrm{BL}} = -\frac{g}{\sigma}\, h_i \sum_{j\in\mathcal{N}_i}^{N} H_j\, \mathbf{c}_{ij}. \tag{4.25}$$

Then the nodal velocity $\tilde{\boldsymbol{v}}_i = (h\boldsymbol{v})_i/h_i$ is adjusted as follows

$$\tilde{\boldsymbol{v}}_i = \frac{(h\boldsymbol{v})_i}{\max\{h_i,\delta\}} + \max\left\{0, \frac{\delta - h_i}{\delta}\right\}\boldsymbol{v}_i^{\mathrm{BL}}. \tag{4.26}$$

Finally, we overwrite the discharge by $h_i\tilde{\boldsymbol{v}}_i$ as in the energy-based version and in the algorithm proposed by Kurganov and Petrova [Kur07a, Eq. (2.21)].

Note that formula (4.26) ensures a continuous transition between $\tilde{\boldsymbol{v}}_i$ for $h_i \in [0, \delta]$ and $\boldsymbol{v}_i^{\mathrm{SWE}} := (h\boldsymbol{v})_i / h_i$ for $h_i \geq \delta$. Similarly to wall function models for turbulent flows, it uses the inviscid SWE model for $h_i > \delta$ but adapts (the solution of) the momentum equation in the boundary layer, where viscous friction effects are dominant and some assumptions behind the derivation of the shallow water equations are invalid.

In all of the numerical experiments below, we set the bottom friction parameter and the boundary layer thickness to $\sigma = 10$ and $\delta = 10^{-3}$, respectively. These values are chosen according to [Ger00] and the boundary layer theory [Sch17, Ch. 2]. Note that our boundary layer has a thickness of 1 millimeter for the SWE without nondimensionalization. In our opinion this constitutes a reasonable value for which a nodal state can be considered as almost dry and friction should come into play.

Barros et al. [Bar15, Sec. 3] propose an approach that models the impact of bottom friction on wetting and drying in a different way. The underlying idea is to treat the sea bed as a porous medium. Based on precomputed water depths, the authors of [Bar15] distinguish between wet, dry, and transitional regions. The water height and the bottom friction coefficient are adjusted in the latter two regimes. Since we consider the SWE without a bottom friction term (our own fix based on boundary layer theory only adjusts the velocity and discharge), this approach cannot be directly pursued here.

4.5 Numerical examples

Let us now apply the generalized flux correction schemes for the shallow water equations with topography to various one-dimensional benchmarks. By default, we employ a uniform mesh consisting of 128 elements and adaptive SSP2 RK time stepping with CFL parameter $\nu = 0.5$. In particular, we consider classical steady state examples and various dam break problems, before testing our algorithms for an idealized parabolic lake. The same acronyms as in Section 3.4 are used to abbreviate the methods under investigation.

4.5.1 Steady problems

We investigate the well-balancedness of our schemes by applying them to an exact lake at rest configuration as well as moving water equilibria. By default, we employ the raw antidiffusive fluxes $f_{ij} = d_{ij}(u_j - u_i)$ for $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$ in this section. This choice is suitable for steady and weakly time-dependent problems.

4.5.1.1 Lake at rest

In our first test, we set $\Omega = (0, 1)$, $g = 1$, and $b(x) = \max(0, 0.25 - 5(x - 0.5)^2)$. Our initial conditions read $v_0 \equiv 0$ and $h_0(x) = \max\{0.2\mathrm{H}(0.5-x)+0.1\mathrm{H}(x-0.5), b(x)\}$, where

H is the Heaviside function. This test problem represents a lake at rest configuration, and is essentially the same as in [Lia09, Sec. 4.2] but many similar benchmarks exist in the literature. In our case, there are two bodies of water with different depths that are separated from each other by a land mass. Boundary conditions are realized as reflecting walls. However, this choice does not affect the numerical results.

We solve this problem numerically using the boundary layer-based wetting and drying approach (4.25)–(4.26) and display the results in Fig. 4.1. All methods clearly preserve the lake at rest scenario, which is why the free surface elevation profiles in Fig. 4.1a are perfectly on top of each other. Note that the oscillations observable in discharge and velocity are of the order of machine precision and do not amplify in the course of the simulation. The stability of the approach becomes evident by realizing that a total of 22898 time steps was performed with each scheme to reach the very large end time $T = 100$.

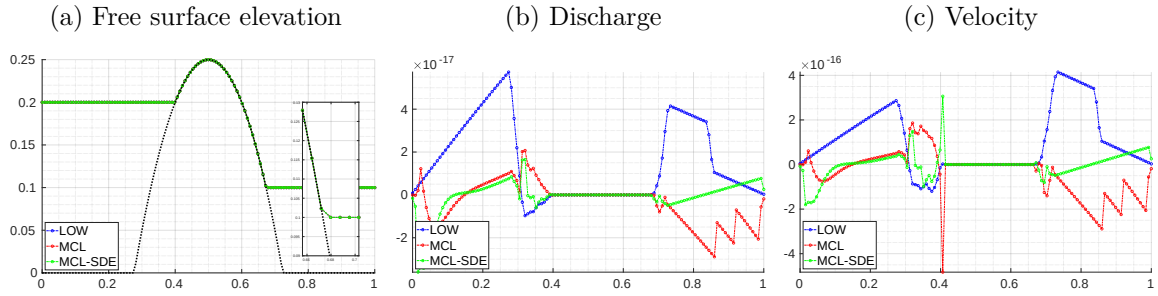


Figure 4.1: A lake at rest for the shallow water equations. Approximations at $T = 100$ obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 128 elements.

In this test, the employed combination of mesh and initial conditions does not capture the shorelines exactly. Thus, the well-balancedness condition (4.6) does not apply here. As a consequence, elements containing a wet-dry transition can only resolve the free surface in these cells by introducing an artificial slope in the discrete water heights. One can see this artifact in the zoomed region of Fig. 4.1a. It was our intention to show that, in practice, our methods remain well balanced for such practical examples, even if (4.6) does not hold. We also ran a similar experiment where (4.6) is satisfied and well-balancedness is guaranteed by Lemma 4.2. For such problems our schemes preserve the lake at rest configuration up to machine precision without introducing any nonphysical slopes in the water height. The discharge and velocity profiles obtained in this fashion are similar to those in Fig. 4.1.

4.5.1.2 Moving water equilibria

Next, we study three classical steady benchmarks [Vaz99, Sec. 5.3], [Del13, Sec. 3.1] as well a supercritical modification thereof. In all cases, the spatial domain is $\Omega = (0, 25)$

and the bathymetry is set to $b(x) = \max\{0, 0.2 - 0.05(x-10)^2\}$. At first, we assume that no nondimensionalization has been performed, thus we use $g = 9.81$.

In the first example, we employ $(h_0, (hv)_0) \equiv (2, 0)$ as initial condition and prescribe $(hv)_{\text{in}} = 4.42$ at the subcritical inlet on the left and $h_{\text{in}} = 2$ at the subcritical outlet on the right. In fact the flow is subcritical everywhere and the treatment of boundaries is in accordance with Tab. 2.2. The exact solution for this setup can be computed as discussed in [Del13, Sec. 3.1]. As a result of the bump in the bathymetry, there appears a corresponding one in the free surface elevation.

Next, we consider a transcritical flow example without a shock, which is obtained with the initial and left boundary data $(h_0, (hv)_0) \equiv (0.66, 0)$ and $(hv)_{\text{in}} = 1.53$, respectively. In this example, the type of the right boundary changes in time and is determined numerically, by computing the eigenvalues of the flux Jacobian for the internal state at the boundary. The external boundary state is then set according to Tab. 2.2 based on the available boundary data $h_{\text{in}} = 0.66$ and $(hv)_{\text{in}} = 1.53$. In this example, the flow becomes supercritical (this behavior is referred to as *torrential* flow) at the bathymetry bump and to the right of it but remains subcritical at the left domain boundary.

Another transcritical example is obtained by setting initial and boundary data as $(h_0, (hv)_0) \equiv (0.33, 0)$, $(hv)_{\text{in}} = 0.18$ on the left and $h_{\text{in}} = 0.33$ on the right, respectively. Again, the region around the bathymetry bump becomes torrential, and this time, a steady shock forms. In this example however, the flow is subcritical, not only on the left of the area with elevated topography but also in the post-shock region. Thus, the treatment of boundary conditions is in accordance with Tab. 2.2. Respective reference solutions obtained with the SWASHES software [Del13] on uniform meshes of 1 000 cells are displayed in Fig. 4.2 for both transcritical examples.

The obtained free surface elevations for the three test problems are displayed in Fig. 4.2. With the employed resolution, one can clearly see that the LOW profiles do not quite attain the exact values in regions where the exact solutions are constant. As expected, LOW also smears the shock in Fig. 4.2c significantly. On the other hand, the agreement of the flux-corrected approximations with the respective exact or reference solutions is satisfactory.

All three of the above examples are classical steady benchmarks. Thus, we check, whether the approximations converge to steady states. Unfortunately, this is only the case for the low order method, not for the flux-limited schemes. A variety of reasons for this lack of convergence can be imagined. In these examples it can be due to the fact, that our schemes are not exactly well balanced for moving water equilibria.

In our final steady example, we modify the above configurations by assuming the system to be in nondimensional form. Thus, we set $g = 1$. As initial condition we use $(h_0, (hv)_0) \equiv (1, 2.1)$, which corresponds to supercritical flow. Thus, supercritical in- and outlet boundary conditions are prescribed at $x = 0$ and $x = 25$, respectively. Again, the bathymetry bump produces a corresponding feature in the free surface elevation. Contrary to the subcritical case displayed in Fig. 4.2a, the bump is pointing upwards in

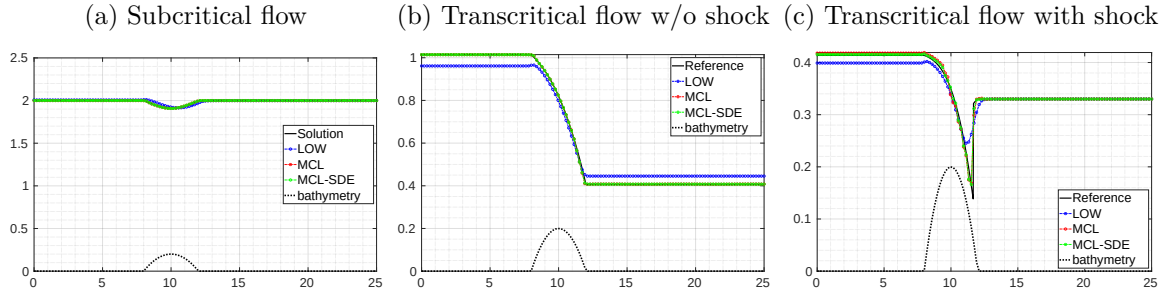


Figure 4.2: Moving water equilibria for the shallow water equations [Vaz99]. Approximations to the free surface elevation at $T = 400$ (a), $T = 200$ (b), and $T = 800$ (c) obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 128 elements.

this example. The exact solution to this problem can be derived as in [Del13, Sec. 3.1].

In this example, all three schemes under investigation do converge to the steady states displayed in Fig. 4.3. These profiles were obtained on a uniform mesh consisting of 128 elements. To rule out that these are isolated instances, we increased the spatial and temporal resolutions by factors of two and four. The steady state residual in each run eventually drops below the threshold of 10^{-12}, although our schemes are not exactly well balanced for moving water equilibria. It is quite remarkable, that the displayed oscillatory discharge profiles represent discrete steady states. Nevertheless, a combination of AFC with strategies that guarantee well-balancedness w. r. t. to moving water equilibria (for instance, as in [Noe07]) is an important topic for future research.

If we include the term $m_{ij}(\dot{u}_i^{\mathrm{L}} - \dot{u}_j^{\mathrm{L}})$ in the antidiffusive fluxes f_{ij}, LOW and MCL-SDE methods still converges to steady state for all three resolutions under consideration, while the MCL method without entropy fix does not. Therefore, it may be a good idea to employ an entropy fix in practical computations.

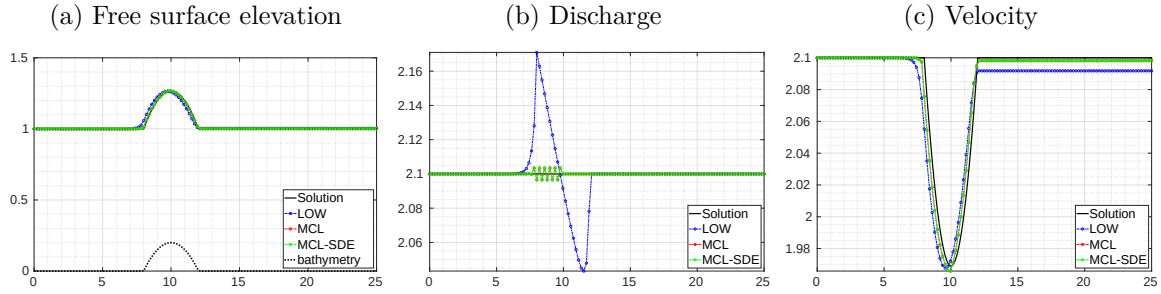


Figure 4.3: Supercritical moving water equilibrium for the shallow water equations. Approximations at steady state obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 128 elements.

The fact that the low order discharge in Fig. 4.3 looks quite different from its flux-corrected counterparts is due to the term $\frac{1}{2}(\boldsymbol{v}_i + \boldsymbol{v}_j)\alpha_{ij}^b(b_j - b_i)$ present in the low order

method. Its influence on the numerical approximations is reduced in the process of flux limiting. A preliminary version of our low order method without this term did not produce the slight phase errors visible in the low order approximations in Fig. 4.3a and Fig. 4.2a. It is somewhat interesting that these deviations from the respective exact solutions are opposite in the subcritical and supercritical cases.

4.5.2 Dam breaks

Having studied some problems with steady state solutions in the previous section, we now perform experiments for the generalized Riemann problem

$$\frac{\partial}{\partial t}\begin{bmatrix} h \\ hv \end{bmatrix} + \frac{\partial}{\partial x}\begin{bmatrix} hv \\ hv^2 + \frac{g}{2}h^2 \end{bmatrix} + \begin{bmatrix} 0 \\ gh\frac{\partial b}{\partial x} \end{bmatrix} = 0 \qquad \text{in } \Omega \times (0,T),$$
$$h_0(x) = \begin{cases} h_L & \text{if } x < x_0, \\ h_R & \text{if } x > x_0, \end{cases} \qquad v_0 \equiv 0,$$

where $\Omega \subset \mathbb{R}$. We use values $h_L > h_R$ in the three below tests. This setup corresponds to an idealized dam located at $x_0 \in \Omega$ that is removed at time $t = 0$. As a result a water wave propagates into the region $x > x_0$, while a rarefaction wave travels in the opposite direction.

4.5.2.1 Wet dam break over flat topography

First, we set $g = 1$ and consider an example with flat bottom topography, i. e., $b \equiv 0$. Thus, we may apply the standard limiting techniques for conservation laws, instead of their generalized versions for the SWE with a topography source term. If both h_L and h_R are positive, the generalized Riemann problem is referred to a wet dam break. Such tests represent relatively mild test cases, which are similar to Sod's shock tube problem [Sod78] for the Euler equations. A difference is that there is one fewer unknown in the system, and the exact solution does not feature any contact discontinuities.

We equip the spatial domain $\Omega = (0, 1)$ with reflecting wall boundaries (although other options are feasible). In our first test, the dam is located at $x_0 = 0.5$ and the two values for the water height are set to $h_L = 1$ and $h_R = 0.1$. As end time we choose $T = 0.3$. The exact solution to this problem can be found in [Del16, Sec. 4.1.1].

First, we perform a convergence study of LOW, MCL, and MCL-SDE schemes on a series of uniform meshes. The rates of convergence observed in Tab. 4.1, are similar to those of the shock tube problem for the Euler equations (see Tab. 3.4), and are optimal for such examples.

In Fig. 4.4, we display the approximations obtained on a uniform mesh consisting of 128 elements. Just as for the Euler equations, the low order profiles are significantly more diffusive than they tend to be in approximations to some scalar problems. Similarly

$1/h$	LOW	EOC	MCL	EOC	MCL-SDE	EOC
32	7.93E-02		3.28E-02		3.66E-02	
64	4.98E-02	0.67	1.67E-02	0.97	1.89E-02	0.95
128	3.00E-02	0.73	8.47E-03	0.98	9.59E-03	0.98
256	1.77E-02	0.76	4.28E-03	0.99	4.85E-03	0.98
512	1.06E-02	0.75	1.94E-03	1.14	2.24E-03	1.11

Table 4.1: Convergence history of the wet dam break for the shallow water equations. The $\|\cdot\|_{\mathrm{L}^1(\Omega)}$ errors at $T = 0.3$ and the corresponding EOC.

to the shock tube example, there are some non-IDP-violating under- and overshoots in the flux-corrected solutions on the right of the rarefaction wave.

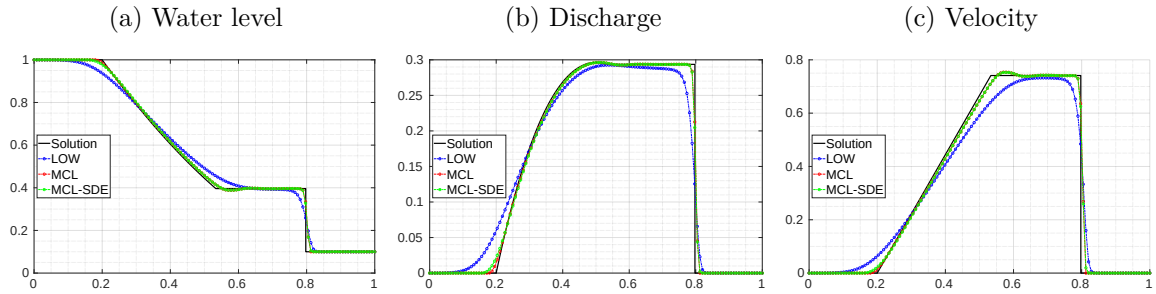


Figure 4.4: Wet dam break for the shallow water equations. Approximations at $T = 0.3$ obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 128 elements.

Let us briefly study a few variations of time stepping schemes and artificial viscosities. For completeness we run this problem with the SSP3 RK method and leave the rest of the setup unchanged. The results for the water heights in Fig. 4.5a are indistinguishable from the ones in Fig. 4.4a. In this example, setting $\nu = 0.67$ was sufficient for the CFL condition (3.43) to be satisfied at all times while values $\nu \geq 0.68$ required some repetitions of individual RK stages.

The theory presented in Section 3.3 suggests that we may also employ forward Euler (SSP1 RK) time stepping and may even set the CFL parameter ν to one. Indeed, the resulting approximations do not violate the IDP property, and one may assume the results to be reliable. We investigate the validity of this assumption, first by using our nodal approximation (3.34) to the wave speeds, and, alternatively, the guaranteed maximum wave speed (GMS) proposed in [Aze17, Prop. 3.7], [Gue18b, Sec. 4]. In either case, oscillations and incorrect approximations in the left part of the rarefaction waves are visible in Figs. 4.5b and 4.5c. The fact that even the use of the GMS wave speed is not sufficient to prevent these nonphysical effects, implies that they are not a result of an incorrect wave speed approximation. A reduction of the CFL parameter ν masks this issue in the sense that the amplitude of the oscillations becomes smaller. The spurious

approximations observed in this study originate from the combination of forward Euler time stepping with continuous finite element methods [Kuz12a, Sec. 4]. As we will see in Chapter 6, schemes based on discontinuous approximation spaces can be safely employed in combination with forward Euler time stepping. The fact that the low order method remains stable can be attributed to its equivalence to the vertex-centered finite volume scheme of local Lax–Friedrichs type [Sel93]. For the shallow water equations we tested the GMS wave speed [Aze17, Prop. 3.7], [Gue18b, Sec. 4] for multiple examples. Although the low order method is derived based on assumptions that encourage the use of GMS instead of our approximation (3.34), we encountered no example in which the use of GMS is actually necessary. This observation was recently confirmed by Wu et al. [Wu21, Thm. 3.1] who show that (3.34) preserves the IDP property for the SWE.

(a) SSP3 RK with $\nu = 0.67$ (b) SSP1 RK, $\nu = 1$ with (3.34) (c) SSP1 RK, $\nu = 1$ with GMS

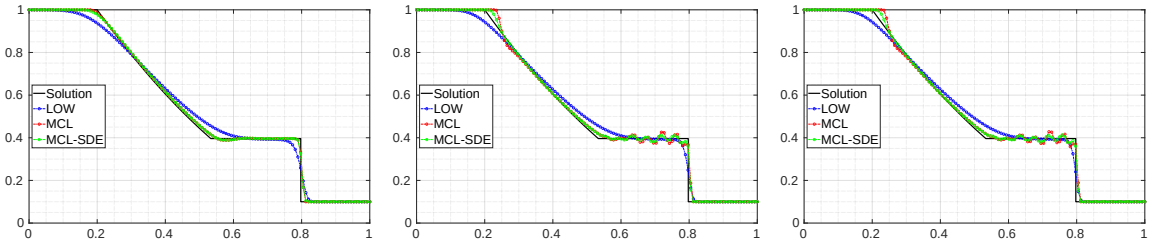


Figure 4.5: Wet dam break for the shallow water equations. Approximations at $T = 0.3$ obtained with adaptive SSP RK time stepping on a uniform mesh consisting of 128 elements.

Using SSPp RK time stepping with $p \in \{2, 3\}$, we observe satisfactory agreement of MCL and MCL-SDE profiles with the exact solutions, not only for the conserved unknowns but also for the velocity. The situation may be different if the test problem features dry or nearly dry states, which is why we consider such an example next.

4.5.2.2 Dry dam break over flat topography

Let us now set h_R to zero, the end time to $T = 0.15$, and leave the rest of the setup from the previous example unchanged. Here the exact solution does not feature a shock wave, only a rarefaction wave is produced as a result of the dam break [Del16, Sec. 4.1.2].

This example already requires some form of treatment to correctly capture the wet-dry transition. If no such approach has been implemented, one can simply run this example by setting h_R to a very small value, for instance 10^{-12}. Instead of following this approach, we compare the results obtained with our new friction- and entropy-based wetting and drying algorithms in Fig. 4.6. The approximations obtained with the existing schemes of Azerad et al. [Aze17], Kurganov and Petrova [Kur07a] as well as the one by Ricchiuto and Bollermann [Ric09] are shown in Fig. 4.7.

All wetting and drying approaches produce acceptable numerical solutions for the water height, whereas approximations for the discharge close to the dry region are

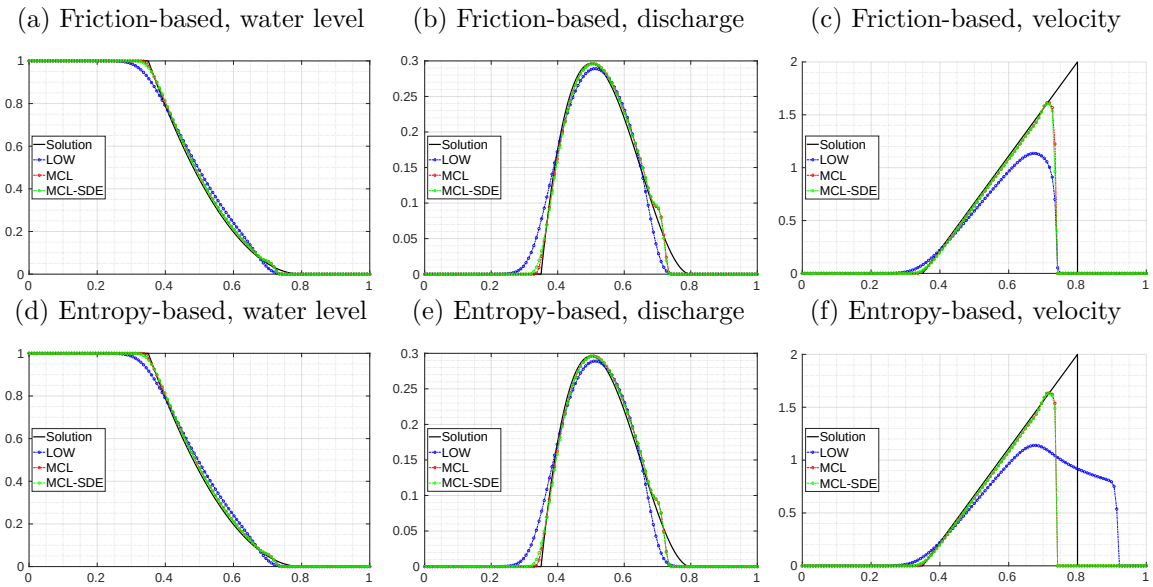


Figure 4.6: Dry dam break for the shallow water equations with new wetting and drying strategies. Approximations at $T = 0.15$ obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 128 elements.

somewhat underresolved. Improvements can only be obtained with refined meshes and time steps. Among the five approaches for wetting and drying, the [Kur07a]-based fix (4.23) produces the most pronounced kink in the discharge and a similar artifact is visible in the corresponding water levels. Moreover, this fix produces the smallest velocities among all considered approaches. All other wetting and drying algorithms produce satisfactory results for this test problem. The somewhat significant differences in the velocities, particularly for the low order solution are unsurprising to us because the calculation of $v = (hv)/h$ is quite sensitive to small water heights, which occur in almost dry areas. Again, refinement is needed to obtain more accurately resolved velocity profiles.

4.5.2.3 Wet dam break over a bump

Next, we study a dam break problem proposed in [Win15, Sec. 5.6]. It involves a nonflat bottom topography. The spatial domain $\Omega = (0, 20)$ is again equipped with reflecting wall boundaries and the gravitational constant is $g = 1$. The bottom topography, and initial conditions read

$$b(x) = \begin{cases} \sin(0.25\pi x) & \text{if } |x - x_0| < 2, \\ 0 & \text{otherwise,} \end{cases} \qquad h_0(x) = \begin{cases} 1.6 - b(x) & \text{if } x < x_0, \\ 1.05 - b(x) & \text{if } x > x_0, \end{cases}$$

where $x_0 = 10$ and $v_0 \equiv 0$.

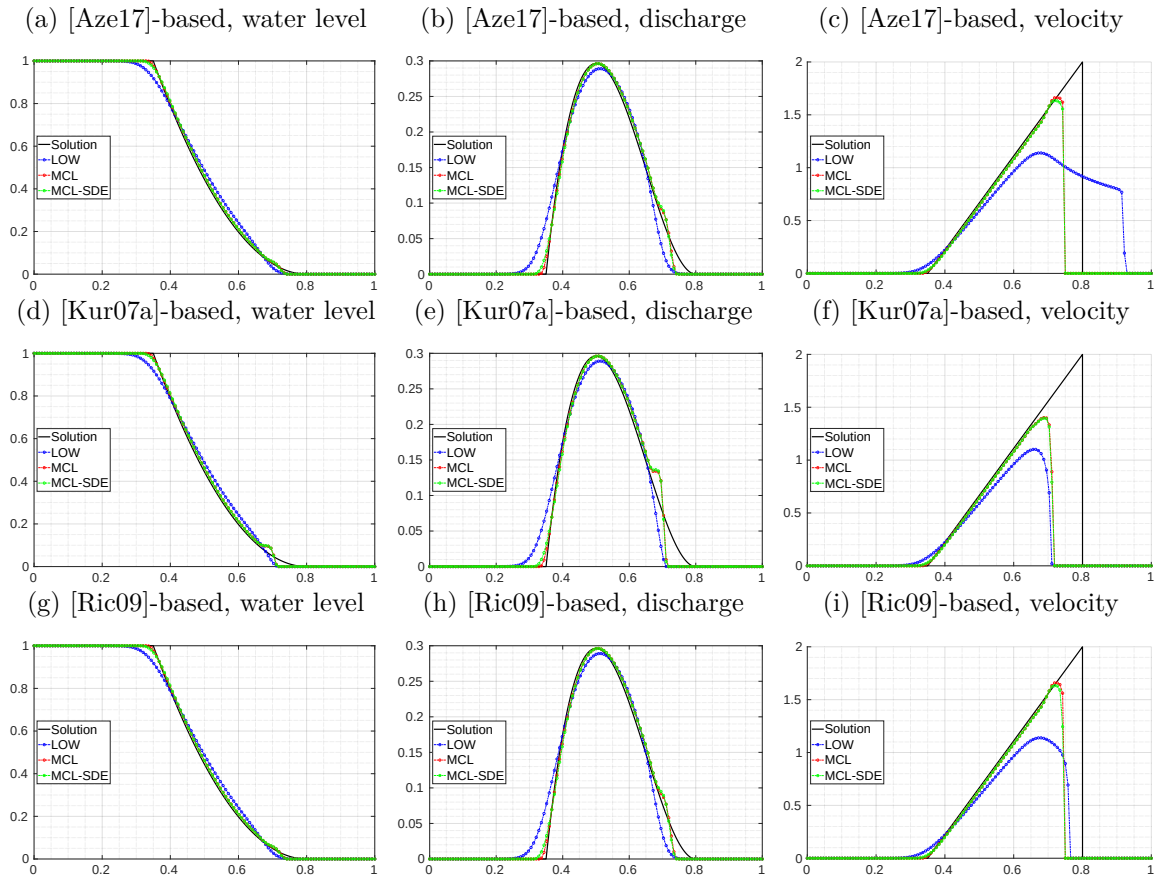


Figure 4.7: Dry dam break for the shallow water equations with wetting and drying strategies from the literature. Approximations at $T = 0.15$ obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 128 elements.

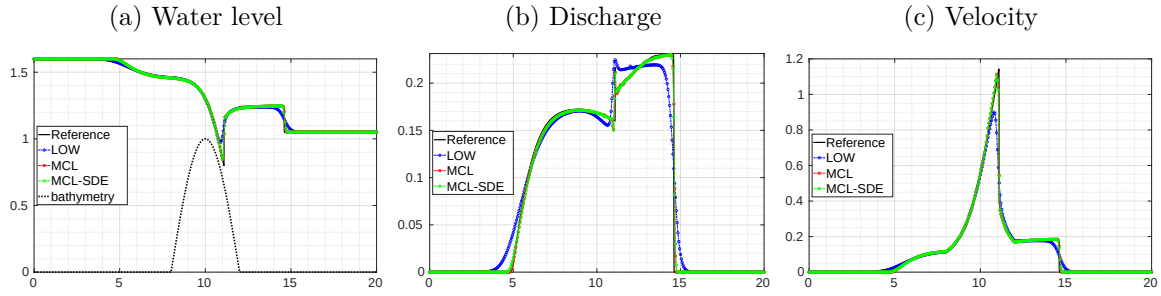


Figure 4.8: Dam break over a bump for the shallow water equations [Win15]. Approximations at $T = 4.5$ obtained with adaptive SSP2 RK time stepping and $\nu = 0.5$ on a uniform mesh consisting of 400 elements.

To facilitate a comparison of our results with the ones in [Win15], we solve this problem up to time $T = 4.5$ on a mesh consisting of 400 elements. A reference solution

is obtained with a finite volume method on a fine mesh consisting of $E = 10^4$ elements. Even though the initial water height on the right of the dam is quite small, our friction-based wetting and drying algorithm is never activated in this problem. The results of this study are displayed in Fig. 4.8, where we observe excellent agreement with our reference solutions. The obtained profiles also agree well with the ones in [Win15, Sec. 5.6] with the exception that the peaks in the velocity profiles are slightly lower in our results. This issue requires further investigations and comparisons with the methods in [Win15].

4.5.3 Oscillating surface in a parabolic lake

In the final numerical example of this chapter, we apply our schemes to one of Thacker's oscillatory lakes with a parabolic basin [Tha81]. Such benchmarks are challenging tests for wetting and drying algorithms. We use the same setup as in Vater et al. [Vat15, Sec. 4.4], where $\Omega = (-5000, 5000)$, $g = 9.81$, and $b(x) = h_0(x/a)^2$ with $h_0 = 10$ and $a = 3000$. In the absence of friction, the exact solution is periodic and reads [Tha81, Lia09, Vat15]

$$
\begin{aligned}
x_\pm(t) &= -\frac{B}{\omega}\cos(\omega t) \pm a, \qquad B = 5, \qquad \omega = \frac{\sqrt{2gh_0}}{a},\\
H(x,t) &= \begin{cases} h_0 - \frac{B^2}{4g}(1+\cos(2\omega t)) - \frac{Bx}{a}\sqrt{\frac{2h_0}{g}}\cos(\omega t) & \text{if } x_-(t) \le x \le x_+(t),\\ b(x) & \text{otherwise,}\end{cases}\\
v(x,t) &= \begin{cases} \frac{Ba\omega}{\sqrt{2h_0 g}}\sin(\omega t) & \text{if } x_-(t) \le x \le x_+(t),\\ 0 & \text{otherwise.}\end{cases}
\end{aligned}
$$

We employ a CFL parameter of $\nu = 0.05$ in combination with our friction-based wetting and drying approach to solve this problem numerically up to end time $T = 3000$. Larger CFL parameters lead to either repetitions of single Runge–Kutta stages or increases of Rusanov diffusion coefficients for nodes around the wet-dry transitions. For $\nu = 0.05$, all schemes remain stable without the need for employing either of these adjustments, even in the case of adaptive time stepping. Fig. 4.9 displays LOW, MCL and MCL-SDE water levels at three different times along with the initial condition for illustrative purposes. From Fig. 4.9b we can make out that the low order profile is trailing the exact solution and its flux-corrected counterparts. Agreement of the flux-limited profiles with the exact water levels is again satisfactory.

We also tested whether we can employ other wetting and drying algorithms in this example. With the fixes from [Aze17] and [Ric09] our simulations crash. The fix from [Kur07a] produces profiles similar to the ones in Fig. 4.9. It is actually possible to employ a larger CFL parameter ν with this wetting and drying approach. This observation motivates further tests and adjustments of our friction-based strategy. Specifically, nonlinear friction models should be considered and the parameters δ and σ may need to

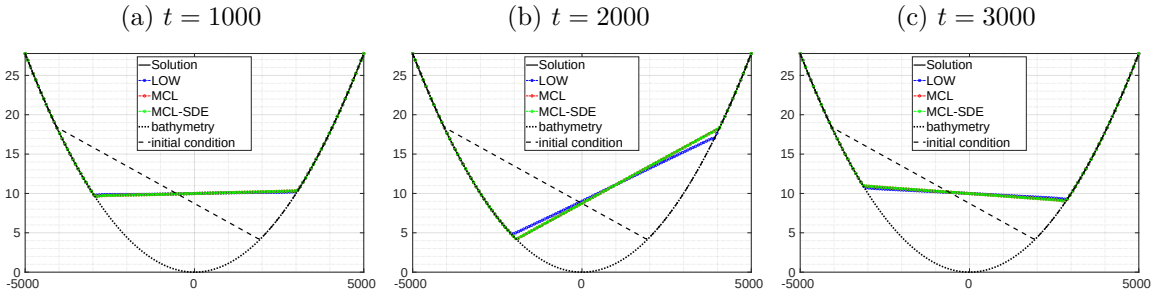


Figure 4.9: Oscillating surface in a parabolic lake for the shallow water equations [Tha81]. Approximations to the free surface elevation at various times obtained with adaptive SSP2 RK time stepping and $\nu = 0.05$ on a uniform mesh consisting of 128 elements.

be adjusted. Since these studies should include multidimensional test cases, we have not yet conducted further research in this direction.

Chapter 5

Analysis of monolithic convex limiting for advection problems

Thus far we focused on computational aspects of monolithic convex limiting schemes. Provable properties of the bound-preserving MCL strategy were not yet fully explored and comparisons of MCL to FCT-type methods were also postponed. These topics will be addressed now in the context of linear time-dependent advection problems.

This chapter is based on our preprint [Haj21b] and is organized as follows. We first review the rather sparse literature on theoretical investigations of AFC methods in Section 5.1. Subsequently, in Section 5.2, we summarize important aspects of flux correction schemes based on MCL and FCT methodologies for linear advection problems. This discussion is followed by the presentation of our stability and error analysis in Sections 5.3 and 5.4, respectively. Illustrative numerical examples in Section 5.5 conclude this chapter.

5.1 Literature

Theoretical properties of flux correction schemes are less commonly studied than their computational aspects. Besides the few results already mentioned in Chapter 3, AFC schemes for linear scalar problems are analyzed in the following references.

Barrenechea et al. [Bar16] show the solvability of a nonlinear algebraic system originating from the AFC discretization of a steady scalar convection-diffusion-reaction equation. Moreover, the authors prove a discrete maximum principle and derive an a priori error estimate. Under the assumption that the sequence of employed meshes consists of *Delaunay triangulations*, the error in the energy norm can be shown to behave as $h^{\frac{1}{2}}$. This rate is, in fact, optimal in the setting under consideration because no assumptions on the correction factors of the limiter are made (see for instance the experiment in [Bar16, Sec. 8.1]). The analysis in [Bar16] suggests that, in general, the approximation under consideration is inconsistent if diffusive terms are present.

Similarities between an edge-based diffusion approach and AFC schemes are explored by Barrenechea et al. [Bar17a]. Therein, the authors introduce a nonlinear stabilization term into the discretization of a steady convection-diffusion-reaction equation. For this scheme they again show solvability, prove a maximum principle, and derive an a priori error estimate with rate $\frac{1}{2}$. No consistency errors arise from the viscous term, contrary to the AFC scheme considered in the earlier work [Bar16]. It is shown that the stabilization

approach under consideration is equivalent to an AFC strategy if the correction factors of the latter are defined appropriately. Since a rigorous error estimate can be obtained for the nonlinear stabilization approach, the authors advertise the use of this scheme instead of the original AFC method, which is inconsistent for elliptic problems.

Under the assumption that the limiter under consideration is *linearity preserving*, Barrenechea et al. [Bar18, Sec. 3.3] derive an error estimate with improved linear convergence rate. However, this result does not directly carry over to the inviscid case due to the presence of the diffusivity parameter in a denominator, see [Bar18, p. 667].

The theoretical results obtained for elliptic equations in [Bar16] are adapted to linear hyperbolic PDEs by Lohmann [Loh19]. In particular, the issues of well-posedness and discrete maximum principles are addressed for steady and unsteady problems. Moreover, the $\mathrm{L}^2(\Omega)$ error of the steady problem is proven to be $\mathcal{O}(h^{\frac{1}{2}})$ in the inviscid case as well.

In the context of AFC schemes for time-dependent problems, the first a priori error estimates are the ones obtained by Jha and Ahmed [Jha21]. Their results generalize the steady-state analysis performed in [Bar16] to a transient convection-diffusion-reaction equation. The AFC schemes analyzed in [Jha21] are based on FCT approaches that are fully discrete and employ implicit time stepping.

Contrary to [Jha21], we perform our stability and error analysis in the semi-discrete setting, which is made possible by the monolithic limiting strategy. Following [Bar16], we allow all correction factors to be zero, and thus our a priori error estimate with convergence rate $\frac{1}{2}$ in the $\mathrm{L}^2(\Omega)$ norm is optimal. Before stating and proving our own theoretical results, we need to discuss further properties of flux correction schemes for time-dependent linear advection problems. For fully discrete analysis of these methods, we refer the reader to Lohmann [Loh19, Ch. 4], who studied both FCT and monolithic approaches in detail. A discrete maximum principle for the MCL methodology applied to steady nonlinear hyperbolic problems can be found in [Kuz20a, Appendix].

5.2 Algebraic flux correction schemes

In this section, we summarize limiting strategies based on MCL and FCT for linear transport problems. For details on the former, we refer the reader to Section 3.3.

5.2.1 Model problem and low order method

First, we define the continuous model problem to be discretized using AFC in this chapter. Let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ be a polyhedral domain and $\boldsymbol{v} \in \mathbf{C}(\overline{\Omega} \times \mathbb{R}_+)^d$ a known velocity field. We define the time-dependent in- and outflow boundaries of Ω as

$$\Gamma_-(t) := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{v}(\boldsymbol{x}, t) \cdot \boldsymbol{n}(\boldsymbol{x}) < 0\}, \qquad \Gamma_+(t) := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{v}(\boldsymbol{x}, t) \cdot \boldsymbol{n}(\boldsymbol{x}) \geq 0\}.$$

In what follows, we suppress the dependence of $\Gamma_\pm(t)$ on time t. The initial-boundary value problem for the linear advection equation reads

$$\frac{\partial u}{\partial t} + \boldsymbol{v} \cdot \nabla u = 0 \qquad \text{in } \Omega \times \mathbb{R}_+, \tag{5.1a}$$

$$u = \hat{u} \qquad \text{on } \Gamma_- \times \mathbb{R}_+, \tag{5.1b}$$

$$u = u_0 \qquad \text{in } \Omega, \tag{5.1c}$$

where $\hat{u}$ is a given inflow boundary profile and u_0 is an initial datum. For analytical purposes, we assume that the velocity field is *solenoidal*, i. e., $\nabla \cdot \boldsymbol{v} = 0$ in Ω, which allows us to interpret (5.1a) as a hyperbolic conservation law with flux function $\mathbf{f}(u, \boldsymbol{x}, t) = \boldsymbol{v}(\boldsymbol{x}, t)u$. Let us remark that the flux correction tools discussed in this chapter can also be applied to problem (5.1) in the case of more general velocities.

With regard to the continuous weak formulation of (5.1), we follow Di Pietro and Ern [DiP12, Chs. 2–3]. In particular, we introduce the *graph space* [DiP12, Def. 2.1]

$$\mathrm{V} := \{w \in \mathrm{L}^2(\Omega) : \boldsymbol{v} \cdot \nabla w \in \mathrm{L}^2(\Omega)\}$$

and define a weak solution to (5.1) as follows.

Definition 5.1 (Weak solutions to the linear advection equation)
A function $u \in \mathrm{C}(\mathbb{R}_+; \mathrm{V}) \cap \mathrm{C}^1(\mathbb{R}_+; \mathrm{L}^2(\Omega))$ is a weak solution to (5.1) if $u(\cdot, 0) = u_0$ almost everywhere in Ω and

$$\int_\Omega w\, \partial_t u \,\mathrm{d}\boldsymbol{x} + a(u, w) = b(w) \qquad \forall w \in \mathrm{V},\ t \in \mathbb{R}_+, \tag{5.2}$$

where $\partial_t u := \dfrac{\partial u}{\partial t}$ and

$$a(\cdot,\cdot) : \mathrm{V} \times \mathrm{V} \to \mathbb{R}, \qquad a(u, w) := \int_\Omega w\, \boldsymbol{v} \cdot \nabla u \,\mathrm{d}\boldsymbol{x} - \int_{\Gamma_-} w\, u\, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s, \tag{5.3}$$

$$b(\cdot) : \mathrm{V} \to \mathbb{R}, \qquad b(w) := -\int_{\Gamma_-} w\, \hat{u}\, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s. \tag{5.4}$$

♢

Formulation (5.2)–(5.4) is derived similarly to the general strong form (3.6). For the linear advection equation, the local Lax–Friedrichs flux (3.2) reduces to the *upwind flux*

$$\mathrm{f}_{\boldsymbol{n}}(u, \hat{u}) = \frac{\boldsymbol{v} \cdot \boldsymbol{n}}{2}(u + \hat{u}) + \frac{|\boldsymbol{v} \cdot \boldsymbol{n}|}{2}(u - \hat{u}) = \begin{cases} \boldsymbol{v} \cdot \boldsymbol{n}\, u & \text{on } \Gamma_+, \\ \boldsymbol{v} \cdot \boldsymbol{n}\, \hat{u} & \text{on } \Gamma_-. \end{cases}$$

Thus, only boundary integrals over the inlet Γ_- appear in (5.3) and (5.4).

Remark 5.2
In this work, we assume that a unique solution u in the sense of Definition 5.1 and [DiP12] exists. For settings similar to ours, the validity of this assumption can be rigorously proven (see for instance [Daf00]) but for general velocities this is not a trivial task. In principle, one can invoke the method of characteristics and use an energy estimate to show well-posedness. However, rigorous existence and uniqueness results regarding solutions of (5.2) are typically obtained under additional assumptions. For details on these issues, we refer the reader to [DiP12, Sec. 3.1.1] and the references therein. ♢

The low order method that is employed in this chapter is the algebraic Lax–Friedrichs scheme (3.26) adapted to linear advection problems. In the AFC literature, this linear version is called the *discrete upwinding* method because of its equivalence to the node-centered upwind finite volume scheme [Kuz02, Sec. 6]. Let us now review the main steps of deriving this low order method. First, we discretize (5.2) in space using continuous linear finite elements as in Section 3.1 and adopt notation similar to that introduced in Chapter 3. For $i \in \{1, \dots, N\}$, we obtain the spatial semi-discretization

$$\int_\Omega \varphi_i \frac{\partial u_h}{\partial t} \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \varphi_i \, \boldsymbol{v} \cdot \nabla u_h \,\mathrm{d}\boldsymbol{x} - \int_{\Gamma_-} \varphi_i \, (\hat{u} - u_h) \, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s. \tag{5.5}$$

To construct the low order method, we proceed as in Section 3.3.2, i. e., perform row sum mass lumping, use a lumped approximation of boundary terms, and add diffusive fluxes of the form $d_{ij}(u_j - u_i)$. However, we refrain from using the group finite element formulation for the flux $\mathbf{f}(u_h, \boldsymbol{x}, t) = \boldsymbol{v}(\boldsymbol{x}, t)u_h$ because the dependence on $\boldsymbol{x}$ and t may produce bound-violating bar states, as noticed in [Kuz20a]. Since we use the consistent Galerkin discretization (5.5) as target scheme, the definition of artificial viscosity coefficients d_{ij} needs to be adapted slightly as well. The low order counterpart of (5.5) reads

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} (d_{ij} - a_{ij})(u_j - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \dots, N\}, \tag{5.6}$$

where $\hat{u}_i^k$ denotes the upwind value $\hat{u}(\boldsymbol{x}_i)$ corresponding to $\Gamma_k \in \mathcal{F}_i$ and

$$m_i := \int_\Omega \varphi_i \,\mathrm{d}\boldsymbol{x}, \qquad a_{ij} := \int_\Omega \varphi_i \, \boldsymbol{v} \cdot \nabla \varphi_j \,\mathrm{d}\boldsymbol{x}, \qquad b_i^k := -\int_{\Gamma_k} \varphi_i \, \min\{0, \boldsymbol{v} \cdot \boldsymbol{n}\} \,\mathrm{d}s.$$

Note that $b_i^k \geq 0$. The modified Rusanov coefficients are defined by

$$d_{ij} = \max\{|a_{ij}|, |a_{ji}|\}, \qquad i \in \{1, \dots, N\}, \; j \in \mathcal{N}_i \setminus \{i\}. \tag{5.7}$$

We may also write (5.6) in the bar state form [Gue16b], [Kuz20a, Sec. 4.1]

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i \setminus \{i\}} 2d_{ij}(\bar{u}_{ij} - u_i) + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \dots, N\}, \tag{5.8}$$

where

$$\bar{u}_{ij} = \begin{cases} \dfrac{u_i + u_j}{2} - \dfrac{a_{ij}(u_j - u_i)}{2d_{ij}} & \text{if } a_{ij} \neq 0, \\ \dfrac{u_i + u_j}{2} & \text{if } a_{ij} = 0, \end{cases} \qquad i \in \{1, \ldots, N\},\ j \in \mathcal{N}_i \setminus \{i\}. \tag{5.9}$$

Remark 5.3
As shown in Section 3.3.2, the low order method that uses the group finite element approximation is bound preserving because $\bar{u}_{ij}$ is a convex combination of u_i and u_j in the scalar case. Definition (5.7) of d_{ij} ensures the same property for (5.9) because

$$\min\{u_i, u_j\} \leq \bar{u}_{ij} \leq \max\{u_i, u_j\} \qquad \Leftrightarrow \qquad |a_{ij}| \leq d_{ij}.$$

The classical version of the discrete upwinding method uses [Kuz02, Kuz12a]

$$d_{ij} = \max\{a_{ij}, 0, a_{ji}\}, \qquad i \in \{1, \ldots, N\},\ j \in \mathcal{N}_i \setminus \{i\}. \tag{5.10}$$

If $\nabla \cdot \boldsymbol{v} = 0$, this definition is equivalent to (5.7), unless both nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ lie on $\partial\Omega$. This fact follows from integration by parts and omission of the resulting boundary integral. The validity of discrete maximum principles for nodal values can be shown for (5.10) using alternative proof techniques [Loh19, Sec. 4.3.2]. However, individual bar states $\bar{u}_{ij}$ of the discrete upwinding method based on (5.10) may violate the local maximum principle $\min\{u_i, u_j\} \leq \bar{u}_{ij} \leq \max\{u_i, u_j\}$. ◇

5.2.2 Monolithic convex limiting

For time-dependent advection problems, we employ raw antidiffusive fluxes defined by $f_{ij} = m_{ij}(\dot{u}_i - \dot{u}_j) + d_{ij}(u_i - u_j)$. Here $\dot{u}_h = \sum_{i=1}^N \dot{u}_i \varphi_i$ is a suitable approximation to the time derivative (given by the low order nodal values $\dot{u}_i^{\mathrm{L}}$ defined by (3.38) in practice). Furthermore, the limited bar states are constrained using formula (3.47). As in Section 3.3.4.2, we use the local bounds (3.44) for flux limiting, see [Kuz20a, Sec. 4].

Let us now rewrite the bar state form (5.8) of the semi-discrete MCL scheme in a formulation that is more amenable to theoretical investigations. Despite the fact that using MCL, the fluxes f_{ij}^* can be calculated directly via (3.47), we introduce correction factors $\alpha_{ij}(u_h) = \alpha_{ji}(u_h) \in [0, 1]$ defined by $\alpha_{ij}(u_h) = f_{ij}^*/f_{ij}$ if $f_{ij} \neq 0$ and $\alpha_{ij}(u_h) = 1$ otherwise. The dependence of correction factors on the discrete solution makes AFC schemes nonlinear. Using the definition of f_{ij}, the semi-discrete MCL scheme

$$\begin{aligned} m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = & \sum_{j \in \mathcal{N}_i \setminus \{i\}} \left[(1 - \alpha_{ij}(u_h))\, d_{ij}(u_j - u_i) - a_{ij}(u_j - u_i) + \alpha_{ij}(u_h)\, m_{ij}(\dot{u}_i - \dot{u}_j) \right] \\ & + \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \ldots, N\} \end{aligned} \tag{5.11}$$

can equivalently be written as

$$\sum_{i=1}^{N} w_i\, m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} + a_h(u_h, w_h) + d_h(u_h; u_h, w_h) - m_h(u_h; \dot{u}_h, w_h) = b_h(w_h) \tag{5.12}$$

for all $w_h \in \mathrm{V}_h$ given by $w_h = \sum_{j=1}^{N} w_j \varphi_j$. The bilinear and linear forms

$$\begin{aligned} a_h(u_h, w_h) &:= \int_\Omega w_h\, \boldsymbol{v} \cdot \nabla u_h \,\mathrm{d}\boldsymbol{x} - \sum_{i=1}^{N} w_i\, u_i \int_{\Gamma_-} \varphi_i\, \boldsymbol{v} \cdot \boldsymbol{n} \,\mathrm{d}s, \\ b_h(w_h) &:= -\sum_{i=1}^{N} w_i \sum_{\Gamma_k \in \mathcal{F}_i} \hat{u}_i^k \int_{\Gamma_k} \varphi_i \, \min\{0, \boldsymbol{v} \cdot \boldsymbol{n}\} \,\mathrm{d}s \end{aligned}$$

are associated with the (stabilized) Galerkin finite element discretization corresponding to $\alpha_{ij} = 1$ for all $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. The nonlinear forms [Bar16, Loh19, Jha21]

$$d_h(u_h; v_h, w_h) = \sum_{i=1}^{N} w_i \sum_{j \in \mathcal{N}_i \setminus \{i\}} (1 - \alpha_{ij}(u_h))\, d_{ij}(v_i - v_j), \tag{5.13}$$

$$m_h(u_h; v_h, w_h) = \sum_{i=1}^{N} w_i \sum_{j \in \mathcal{N}_i \setminus \{i\}} \alpha_{ij}(u_h)\, m_{ij}(v_i - v_j) \tag{5.14}$$

in (5.12) are due to algebraic flux correction.

Lemma 5.4 (Scalar product properties of nonlinear forms, Bar16)
For arbitrary $u_h, v_h, w_h \in \mathrm{V}_h$, the nonlinear forms (5.13) and (5.14) satisfy

$$\begin{aligned} d_h(u_h; v_h, v_h) \geq 0, \qquad & d_h(u_h; v_h, w_h)^2 \leq d_h(u_h; v_h, v_h)\, d_h(u_h; w_h, w_h), \\ m_h(u_h; v_h, v_h) \geq 0, \qquad & m_h(u_h; v_h, w_h)^2 \leq m_h(u_h; v_h, v_h)\, m_h(u_h; w_h, w_h). \end{aligned} \qquad \diamond$$

Proof:
Proofs of these statements for $d_h(\cdot; \cdot, \cdot)$ can be found in [Loh19, p. 113], see also [Bar16, Lem. 3.1 and Sec. 6]. The same arguments apply to $m_h(\cdot; \cdot, \cdot)$. □

5.2.3 Flux-corrected transport algorithms

The monolithic convex limiting strategy discussed so far is a relatively new technique. Another family of schemes producing similar bound-preserving approximations are flux-corrected transport algorithms (see, e. g., [Bor73, Zal79, Löh87, Kuz12b, Loh17b, Gue18a, Haj20b, Paz21]). Many aspects discussed in the context of the MCL methodology have a lot in common with corresponding components of FCT schemes. For instance, the low order method and the definition of raw antidiffusive fluxes are essentially the

same in both approaches. The key difference is that FCT is a predictor-corrector limiting strategy that manipulates the finite element solution rather than the (semi-)discrete problem of the baseline discretization. The FCT-constrained version has no semi-discrete counterpart and no well-defined steady state residual. While FCT-like limiters usually perform very well in applications to strongly time-dependent problems, they tend to inhibit convergence of time marching schemes for steady state computations, and the accuracy of a quasi-stationary result is affected by the pseudo time step. Additionally, the fully discrete nature of FCT rules out the use of semi-discrete entropy fixes developed for MCL (see Section 3.3.6). An alternative entropy stabilization techniques for FCT-type schemes [Kiv22] was already mentioned in Section 3.3.6.

To allow a comparison of MCL and FCT schemes for advection problems, we briefly present two FCT algorithms for continuous finite elements. The first one can be found in [Kuz12a, Sec. 6.4.2] and uses the classical Zalesak limiter [Zal79]. The second approach is referred to as localized FCT [Loh19, Sec. 4.4]. We do not discuss the design principles behind these schemes here and refer the reader to the above references for details.

The first step of an FCT algorithm needs to be property preserving. Thus, we perform a forward Euler-type update using the low order method

$$u_i^{\mathrm{L}} = u_i + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i \backslash \{i\}} (d_{ij} - a_{ij})(u_j - u_i) + \frac{\Delta t}{m_i} \sum_{\Gamma_k \in \mathcal{F}_i} b_i^k (\hat{u}_i^k - u_i), \qquad i \in \{1, \ldots, N\}$$

to obtain the predictor $u_h^{\mathrm{L}} \in \mathrm{V}_h$. This stage is bound preserving for small enough time steps Δt. The validity of local maximum principles for u_i^{L} can be shown following the bar state analysis in Section 3.3, and, in particular, in Section 3.3.2. In FCT schemes for advection problems we may employ discrete upwinding coefficients defined by (5.10) instead of (5.7). Contrary to the MCL scheme, the bounds for the FCT constraints should be based on the values of the predictor [Kuz02, Sec. 2]. Thus, we set

$$u_i^{\min} := \min_{j \in \mathcal{N}_i} u_j^{\mathrm{L}}, \qquad u_i^{\max} := \max_{j \in \mathcal{N}_i} u_j^{\mathrm{L}}. \tag{5.15}$$

As additional input for the corrector step, we need to compute the raw antidiffusive fluxes $f_{ij} = m_{ij}(\dot{u}_i - \dot{u}_j) + d_{ij}(u_i - u_j)$, which are defined as in the MCL approach. The purpose of the local extremum diminishing corrector step

$$\tilde{u}_i = u_i^{\mathrm{L}} + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i \backslash \{i\}} \alpha_{ij} f_{ij}, \qquad i \in \{1, \ldots, N\}$$

is to recover as much accuracy as possible without violating the local bounds (5.15).

At this stage, only the question of how to define the correction factors remains. Zalesak's multidimensional limiter [Zal79, Kuz12a] calculates them as follows:

1. Compute the sums of raw antidiffusive fluxes f_{ij} for individual nodes

$$P_i^- := \sum_{j \in \mathcal{N}_i \backslash \{i\}} \min\{0, f_{ij}\}, \qquad P_i^+ := \sum_{j \in \mathcal{N}_i \backslash \{i\}} \max\{0, f_{ij}\}.$$

2. Compute bounds for the corresponding sums of limited antidiffusive fluxes

$$Q_i^- := \frac{m_i}{\Delta t}(u_i^{\min} - u_i^{\mathrm{L}}), \qquad Q_i^+ := \frac{m_i}{\Delta t}(u_i^{\max} - u_i^{\mathrm{L}}).$$

3. Calculate nodal undershoot and overshoot limiters for flux correction

$$R_i^- := \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \qquad R_i^+ := \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}.$$

4. Limit the fluxes f_{ij} and $f_{ji} = -f_{ij}$ using the correction factor

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ \min\{R_i^-, R_j^+\} & \text{if } f_{ij} < 0. \end{cases}$$

As an alternative to Zalesak's limiter, one can employ localized FCT algorithms [Loh17a, Gue18a, Loh19, Paz21]. Instead of limiting sums of positive and negative antidiffusive fluxes under worst-case assumptions, FCT schemes of this kind distribute the bounds $Q_i^{\pm}$ between pairs of nodes and limit each flux *independently* to satisfy

$$\frac{\tilde{m}_{ij}}{\Delta t}(u_i^{\min} - u_i^{\mathrm{L}}) \le \alpha_{ij} f_{ij} \le \frac{\tilde{m}_{ij}}{\Delta t}(u_i^{\max} - u_i^{\mathrm{L}}), \qquad i \in \{1, \dots, N\},\ j \in \mathcal{N}_i \setminus \{i\}. \quad (5.16)$$

The particular choice of the weights $\tilde{m}_{ij} > 0$ might influence the quality of approximations. The FCT constraints (5.16) imply $\tilde{u}_i \in [u_i^{\min}, u_i^{\max}]$ provided that

$$\sum_{j \in \mathcal{N}_i \setminus \{i\}} \tilde{m}_{ij} \le m_i \qquad \forall i \in \{1, \dots, N\}. \quad (5.17)$$

The least restrictive version of a localized FCT scheme is obtained if the inequality in (5.17) holds as identity. Following Lohmann [Loh19, Sec. 4.4], we set

$$\tilde{m}_{ij} := \frac{m_{ij}\, m_i}{m_i - m_{ii}} > m_{ij}, \qquad i \in \{1, \dots, N\},\ j \in \mathcal{N}_i \setminus \{i\}.$$

The bound-preserving correction factors are then given by

$$\alpha_{ij} = \begin{cases} \min\left\{1, \frac{\tilde{m}_{ij}}{\Delta t}\frac{u_i^{\max} - u_i^{\mathrm{L}}}{f_{ij}}, \frac{\tilde{m}_{ji}}{\Delta t}\frac{u_j^{\mathrm{L}} - u_j^{\min}}{f_{ij}}\right\} & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ \min\left\{1, \frac{\tilde{m}_{ij}}{\Delta t}\frac{u_i^{\min} - u_i^{\mathrm{L}}}{f_{ij}}, \frac{\tilde{m}_{ji}}{\Delta t}\frac{u_j^{\mathrm{L}} - u_j^{\max}}{f_{ij}}\right\} & \text{if } f_{ij} < 0. \end{cases}$$

Another representative of localized FCT limiters, the scheme proposed in [Gue18a], uses $\tilde{m}_{ij} = m_i / |\mathcal{N}_i \setminus \{i\}|$, where $|\cdot|$ denotes the cardinality of a set.

Remark 5.5
Both MCL and localized FCT schemes impose flux constraints of the form $f_{ij}^{\min} \le f_{ij} \le f_{ij}^{\max}$ and produce limited antidiffusive fluxes f_{ij}^* that can be calculated via (cf. (3.48))

$$f_{ij}^* = \max\left\{f_{ij}^{\min},\ \min\left\{f_{ij},\ f_{ij}^{\max}\right\}\right\}.$$

The difference between the two approaches lies in the definition of the bounding fluxes $f_{ij}^{\min} \le 0$ and $f_{ij}^{\max} \ge 0$. In the FCT version, the distributed bounds

$$\begin{aligned} f_{ij}^{\min} = -f_{ji}^{\max} &:= \max\left\{\frac{\tilde{m}_{ij}}{\Delta t}(u_i^{\min} - u_i^{\mathrm{L}}), \frac{\tilde{m}_{ji}}{\Delta t}(u_j^{\mathrm{L}} - u_j^{\max})\right\}, \\ f_{ij}^{\max} = -f_{ji}^{\min} &:= \min\left\{\frac{\tilde{m}_{ij}}{\Delta t}(u_i^{\max} - u_i^{\mathrm{L}}), \frac{\tilde{m}_{ji}}{\Delta t}(u_j^{\mathrm{L}} - u_j^{\min})\right\} \end{aligned}$$

are inversely proportional to Δt. Therefore, the limiting constraints become less/more restrictive for smaller/larger time steps. ◇

5.3 Energy estimate

Let us now derive an energy estimate for approximations obtained via (5.12). In the proof of this stability result, we rely on the assumption that the following requirement is satisfied.

Definition 5.6 (Compatibility condition for $(u_h, \dot{u}_h)$, Haj21b)
Let $\lambda := \|\boldsymbol{v}\|_{\mathbf{L}^\infty(\Omega\times\mathbb{R}_+)^d}$ be the maximum velocity and $\dot{u}_h, u_h \in \mathrm{V}_h$ be given functions. Define the nonlinear forms $d_h(\cdot;\cdot,\cdot)$ and $m_h(\cdot;\cdot,\cdot)$ as in (5.13) and (5.14), respectively. Suppose that there exists a constant $\gamma \in (0,1)$ such that

$$\frac{\gamma h}{\lambda} m_h(u_h;\dot{u}_h,\dot{u}_h) \le (1-\gamma) d_h(u_h;u_h,u_h) - m_h(u_h;\dot{u}_h,u_h). \tag{5.18}$$

Then we say that $\dot{u}_h \in \mathrm{V}_h$ is compatible with $u_h \in \mathrm{V}_h$. ◇

The ratio h/λ has physical units $[h]/[\lambda] = \mathrm{m}/(\mathrm{ms}^{-1}) = \mathrm{s}$. It is used in inequality (5.18) to ensure that all terms have the same units for $[\dot{u}_h] = \mathrm{s}^{-1}[u_h]$.

Before presenting our energy estimate, we need to prove the following technical result.

Lemma 5.7
Any function $v_h \in \mathrm{V}_h$ defined by $v_h = \sum_{i=1}^N v_i\varphi_i$ satisfies the identity

$$v_h^2 - \sum_{i=1}^N v_i^2\varphi_i = -\sum_{\substack{i,j=1\\ i<j}}^N (v_i - v_j)^2 \varphi_i\,\varphi_j. \qquad \diamond$$

Proof:
Invoking the partition of unity property of basis functions, we obtain

$$\begin{aligned} v_h^2 - \sum_{i=1}^N v_i^2\,\varphi_i &= \sum_{i=1}^N v_i^2\varphi_i\,(\varphi_i - 1) + \sum_{\substack{i,j=1\\ i\neq j}}^N v_i\,v_j\,\varphi_i\,\varphi_j = -\sum_{\substack{i,j=1\\ i\neq j}}^N v_i^2\,\varphi_i\,\varphi_j + \sum_{\substack{i,j=1\\ i\neq j}}^N v_i\,v_j\,\varphi_i\,\varphi_j \\ &= \sum_{\substack{i,j=1\\ i<j}}^N v_i\,(v_j - v_i)\,\varphi_i\,\varphi_j + \sum_{\substack{i,j=1\\ j<i}}^N v_i\,(v_j - v_i)\,\varphi_i\,\varphi_j \\ &= \sum_{\substack{i,j=1\\ i<j}}^N (v_i - v_j)(v_j - v_i)\,\varphi_i\,\varphi_j. \end{aligned}$$

□

Theorem 5.8 (Semi-discrete energy estimate)
Assume that there is a finite time $T > 0$ such that $\boldsymbol{v}(\cdot,t) \in \mathbf{W}^{1,\infty}(\Omega)$ and $\nabla\cdot\boldsymbol{v}(\cdot,t) = 0$ in Ω for all $t \in (0,T)$. Let $u_h(\cdot,t)$ and $\dot{u}_h(\cdot,t)$ satisfy (5.12) and, additionally, the compatibility condition (5.18) with a constant $\gamma \in (0,1)$ for all $t \in (0,T)$. Then the following estimate holds for the solution $u_h(\cdot,T)$ of the semi-discrete problem (5.12)

$$\begin{aligned} &\sum_{i=1}^N m_i\,u_i(T)^2 + \int_0^T \int_{\Gamma_+} u_h^2\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s\,\mathrm{d}t - \int_0^T \sum_{\substack{i,j=1\\ i<j}}^N (u_i - u_j)^2 \int_{\Gamma_-} \varphi_i\,\varphi_j\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s \\ &\quad - \frac{1}{2}\int_0^T \sum_{i=1}^N u_i^2 \int_{\Gamma_-} \varphi_i\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s + 2\gamma\int_0^T \Big[\frac{h}{\lambda} m_h(u_h;\dot{u}_h,\dot{u}_h) + d_h(u_h;u_h,u_h)\Big]\,\mathrm{d}t \\ &\qquad \le \sum_{i=1}^N m_i\,u_i(0)^2 + 2\int_0^T \sum_{i=1}^N \sum_{\Gamma_k\in\mathcal{F}_i} \int_{\Gamma_k} \varphi_i\,(\hat{u}_i^k)^2\,\max\{0, -\boldsymbol{v}\cdot\boldsymbol{n}\}\,\mathrm{d}s\,\mathrm{d}t. \end{aligned} \tag{5.19}$$

♢

Proof:
Testing (5.12) with $w_h = u_h$, we use the compatibility condition (5.18), the identity $u_h\,\boldsymbol{v}\cdot\nabla u_h = \frac{1}{2}\nabla\cdot(\boldsymbol{v}\,u_h^2)$, the divergence theorem and Young's inequality to show that

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^N m_i \frac{\mathrm{d}(u_i)^2}{\mathrm{d}t} + \frac{1}{2}\int_{\partial\Omega} u_h^2\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s - \sum_{i=1}^N u_i^2 \int_{\Gamma_-} \varphi_i\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s \\ &\qquad + \frac{\gamma h}{\lambda} m_h(u_h;\dot{u}_h,\dot{u}_h) + \gamma d_h(u_h;u_h,u_h) \\ &\quad \le \sum_{i=1}^N u_i\,m_i\frac{\mathrm{d}u_i}{\mathrm{d}t} + \int_\Omega u_h\,\boldsymbol{v}\cdot\nabla u_h\,\mathrm{d}\boldsymbol{x} - \sum_{i=1}^N u_i^2 \int_{\Gamma_-} \varphi_i\,\boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s \\ &\qquad + d_h(u_h;u_h,u_h) - m_h(u_h;\dot{u}_h,u_h) \\ &\quad = b_h(u_h) = -\sum_{i=1}^N \sum_{\Gamma_k\in\mathcal{F}_i} \int_{\Gamma_k} \varphi_i\,u_i\,\hat{u}_i^k\,\min\{0,\boldsymbol{v}\cdot\boldsymbol{n}\}\,\mathrm{d}s \end{aligned}$$

$$\leq -\sum_{i=1}^{N} \frac{u_i^2}{4} \int_{\Gamma_-} \varphi_i \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s - \sum_{i=1}^{N} \sum_{\Gamma_k \in \mathcal{F}_i} \int_{\Gamma_k} \varphi_i \, (\hat{u}_i^k)^2 \, \min\{0, \boldsymbol{v} \cdot \boldsymbol{n}\} \, \mathrm{d}s.$$

Multiplying by factor 2 and combining the integrals over Γ_-, we write this inequality as

$$\sum_{i=1}^{N} m_i \frac{\mathrm{d}(u_i)^2}{\mathrm{d}t} + \int_{\Gamma_+} u_h^2 \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s + \int_{\Gamma_-} \boldsymbol{v} \cdot \boldsymbol{n} \left(u_h^2 - \sum_{i=1}^{N} u_i^2 \, \varphi_i \right) \mathrm{d}s - \frac{1}{2} \sum_{i=1}^{N} u_i^2 \int_{\Gamma_-} \varphi_i \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s$$
$$+ \frac{2\gamma h}{\lambda} m_h(u_h; \dot{u}_h, \dot{u}_h) + 2\gamma d_h(u_h; u_h, u_h) \leq 2 \sum_{i=1}^{N} \sum_{\Gamma_k \in \mathcal{F}_i} \int_{\Gamma_k} \varphi_i \, (\hat{u}_i^k)^2 \, \max\{0, -\boldsymbol{v} \cdot \boldsymbol{n}\} \, \mathrm{d}s.$$

Employing Lemma 5.7 and integrating in time produces (5.19). □

Note that as a consequence of Lemma 5.4 and of the nonnegativity of basis functions, all terms appearing on the left hand side of inequality (5.19) are nonnegative. To guarantee that the assumptions of Theorem 5.8 are satisfied in practice, we developed a scheme that enforces (5.18) for user defined values of γ, see [Haj21b, Sec. 3.3]. In our experience, failure to apply this limiter has no negative practical effects, however.

Remark 5.9
The reader may wonder what significance is attached to Theorem 5.8. Since the *fully discrete* MCL scheme produces locally bound-preserving approximations, it is stable by design. Preservation of global bounds in the *semi-discrete* setting can be shown as in [Kuz22b] under the assumption that a solution exists. The semi-discrete MCL scheme represents a nonlinear system of ordinary differential equations. Well-posedness of such initial value problems can be shown by invoking the Picard–Lindelöf theorem, which guarantees the existence of solutions on finite time intervals. Once local existence is established, we exploit a global existence and uniqueness result for ordinary differential equations [Ama90, Thm. 7.6]. According to this theorem, solutions that cannot be extended to arbitrary times must in fact blow up, which, in our case, is prevented by Theorem 5.8. It follows that the semi-discrete MCL scheme (5.12) possesses a unique solution that exists for all times $t \geq 0$. ♢

5.4 Error analysis

Compared to the energy estimate derived in the previous section, our error analysis is rather involved. In particular, we need to make additional assumptions on the data of the continuous problem (5.2) as well as on the mesh sequences. These aspects are discussed in Section 5.4.1. Subsequently, in Section 5.4.2, we recall some auxiliary results from the literature on numerical analysis of finite element methods including AFC schemes. Finally, in Section 5.4.3, we state, prove, and discuss the main result of this chapter, which is a semi-discrete a priori error estimate for MCL approximations.

Throughout this section, the letter C (possibly with a subscript) denotes a generic positive constant that is independent of the mesh size h. Moreover, we assume that $h \leq 1$ and therefore $h^p \leq h^q$ for $p \geq q$.

5.4.1 Preliminaries

As everywhere in this thesis, we consider only meshes that are affine and geometrically conforming triangulations of $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. In this section, we additionally restrict ourselves to simplicial meshes, which allows us to exploit the linearity of finite element approximations inside mesh cells.

The a priori error estimate that we present in Section 5.4.3 is valid only for *quasi-uniform* families of meshes, i. e., there has to exist $C > 0$ such that [DiP12, Sec. 3.1.2]

$$h = \max_{K \in \mathcal{K}_h} h_K \leq C \min_{K \in \mathcal{K}_h} h_K,$$

where $h_K = \operatorname{diam}(K)$. As is standard in finite element analysis, we also assume *shape-regularity* of $(\mathcal{K}_h)_{h>0}$. For this requirement to be satisfied, there has to exist $C > 0$ such that $Ch_K \leq r_K$, where r_K is the radius of the largest open ball that fits into K [DiP12, Sec. 1.4.1]. Additionally, we assume that the mesh faces, which are simplices in $\mathbb{R}^{d-1}$, are also shape regular in this sense. Our final assumption regarding the mesh sequence is that there exists $C > 0$ such that $h \leq C\tilde{h}$, where $\tilde{h} = \min_{\Gamma \in \mathcal{F}_{\partial\Omega}} \operatorname{diam}(\Gamma)$ and $\mathcal{F}_{\partial\Omega}$ is the set of boundary faces (cf. Definition 3.2). We do not need to assume *contact regularity* of the mesh sequence [DiP12, Def. 1.38] as Di Pietro and Ern do because this requirement is automatically satisfied for simplicial triangulations.

Following [Bar16, Loh19], we assume $\mathrm{H}^2(\Omega)$ regularity of the exact solution $u(\cdot, t)$ for all $t \geq 0$. We also require the time derivative $\partial_t u$ to have this regularity. Specifically, we restrict our investigations to exact solutions of (5.2) that satisfy

$$u \in \mathrm{W}^{1,\infty}(\mathbb{R}_+; \mathrm{H}^2(\Omega)), \qquad u|_{\Gamma_-} \in \mathrm{L}^\infty(\mathbb{R}_+; \mathrm{H}^2(\Gamma_-)).$$

For simplicity, we set $u_h(\cdot, 0)$ equal to the continuous interpolant $\mathrm{I}_h u_0 \in \mathrm{V}_h$ of $u_0 \in \mathrm{C}(\overline{\Omega})$. The interpolation operator $\mathrm{I}_h : \mathrm{C}(\overline{\Omega}) \to \mathrm{V}_h$ is defined by

$$w \mapsto w_h := \sum_{i=1}^{N} w(\boldsymbol{x}_i)\, \varphi_i.$$

Also for simplicity, we assume that the boundary data $\hat{u}$ is linear on every boundary face $\Gamma \in \mathcal{F}_-$, where $\mathcal{F}_- = \mathcal{F}_-(t) := \{\Gamma \in \mathcal{F}_{\partial\Omega} : \Gamma \cap \Gamma_- \neq \emptyset\}$. This assumption corresponds to a particular choice of the quadrature rule for boundary integrals.

5.4.2 Auxiliary statements

To prepare the ground for the derivation of our error estimate, we first summarize a few important ingredients of its proof, beginning with some standard inequalities. Then we discuss aspects that are peculiar to algebraic flux correction schemes. Most of the AFC results were originally proven by Barrenechea et al. [Bar16].

Lemma 5.10 (Interpolation error estimate for volume integrals)
Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. Then there exists $C > 0$ such that

$$\|w - \mathrm{I}_h w\|_{\mathrm{L}^2(\Omega)} + h|w - \mathrm{I}_h w|_{\mathrm{H}^1(\Omega)} \leq Ch^2 |w|_{\mathrm{H}^2(\Omega)} \qquad \forall w \in \mathrm{H}^2(\Omega). \qquad \diamond$$

Proof:
See [Ern04, Sec. 1.5.1, in particular Ex. 1.111]. □

Lemma 5.11 (Interpolation error estimate for boundary integrals)
Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ and let $\Gamma \subset \partial K$ be a face of $K \in \mathcal{K}_h$. Then there exists $C > 0$ such that

$$\|w - \mathrm{I}_h w\|_{\mathrm{L}^2(\Gamma)} \leq Ch_K^{3/2} |w|_{\mathrm{H}^2(K)} \qquad \forall w \in \mathrm{H}^2(K). \qquad \diamond$$

Proof:
The claim follows from the *continuous trace inequality* [DiP12, Lem. 1.49] in combination with Lemma 5.10. □

Lemma 5.12 (Discrete trace inequality, DiP12 Lem. 1.46)
Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ and let $\Gamma \subset \partial K$ be a face of $K \in \mathcal{K}_h$. Then there exists $C > 0$ such that

$$\|v_h\|_{\mathrm{L}^2(\Gamma)} \leq Ch_K^{-1/2} \|v_h\|_{\mathrm{L}^2(K)} \qquad \forall v_h \in \mathbb{P}_1(K). \qquad \diamond$$

Lemma 5.13 (Inverse inequality, DiP12 Lem. 1.44)
Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ and let $K \in \mathcal{K}_h$. Then there exists $C > 0$ such that

$$|v_h|_{\mathrm{H}^1(K)} \leq Ch_K^{-1} \|v_h\|_{\mathrm{L}^2(K)} \qquad \forall v_h \in \mathbb{P}_1(K). \qquad \diamond$$

Lemma 5.14 (Bar16)
Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. Define $\Gamma_{ij} := \{\mu \boldsymbol{x}_i + (1-\mu)\boldsymbol{x}_j : \mu \in [0, 1]\}$ for a pair of mesh vertices $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ $i \in \{1, \ldots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. Let $K \in \mathcal{K}_h$ with $\Gamma_{ij} \subset \partial K$. Then there exists $C > 0$ such that

$$|v_h(\boldsymbol{x}_i) - v_h(\boldsymbol{x}_j)| \leq Ch_K^{1-d/2} |v_h|_{\mathrm{H}^1(K)} \qquad \forall v_h \in \mathbb{P}_1(K). \qquad \diamond$$

Proof:
The claim follows from Taylor expansion, linearity, and shape regularity, see [Bar16, Pf. of Lem. 7.3] or [Loh19, Ineq. (4.90)] for details. □

Lemma 5.15 (Bar16, Jha21)
Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$. Then there exist constants $C_1 = C_1(d) > 0$ and $C_2 = C_2(d, \boldsymbol{v}) > 0$ such that

$$m_{ij} \le C_1 h^d, \qquad d_{ij} \le C_2 h^{d-1}, \qquad i \in \{1,\dots,N\},\ j \in \mathcal{N}_i \setminus \{i\}. \qquad \diamond$$

Proof:
We have $\operatorname{supp}(\varphi_i\varphi_j) \subseteq \Omega_{ij} := \{\boldsymbol{x} \in \overline{\Omega} : \exists \mu \in [0,1] : |\boldsymbol{x} - (\mu \boldsymbol{x}_i + (1-\mu)\boldsymbol{x}_j)| \le h\}$, and due to shape regularity, there exists $C = C(d) > 0$ such that $|\Omega_{ij}| \le C h^d$. Therefore

$$m_{ij} = \int_{\Omega_{ij}} \varphi_i\,\varphi_j \,\mathrm{d}\boldsymbol{x} \le \|\varphi_i\|_{\mathrm{L}^2(\Omega_{ij})} \|\varphi_j\|_{\mathrm{L}^2(\Omega_{ij})} \le \|1\|^2_{\mathrm{L}^2(\Omega_{ij})} = |\Omega_{ij}| \le C h^d.$$

The estimate for d_{ij} is obtained similarly by invoking (5.7), factoring out the maximum velocity λ and using the inverse inequality, i. e., Lemma 5.13, see [Bar16, Pf. of Lem. 7.3] or [Loh19, Pf. of Thm. 4.72] for details. □

Lemma 5.16 (Bar16)
Let $(\mathcal{K}_h)_{h>0}$ be a shape-regular family of meshes over $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$. Then there exist constants $C_1 = C_1(d) > 0$ and $C_2 = C_2(d, \boldsymbol{v}) > 0$ such that

$$m_h(v_h; \mathrm{I}_h w, \mathrm{I}_h w) \le C_1 h^2 \|w\|^2_{\mathrm{H}^2(\Omega)}, \qquad d_h(v_h; \mathrm{I}_h w, \mathrm{I}_h w) \le C_2 h \|w\|^2_{\mathrm{H}^2(\Omega)}$$

for all $v_h \in \mathrm{V}_h$, $w \in \mathrm{H}^2(\Omega)$. $\diamond$

Proof:
The estimate for $d_h(\cdot;\cdot,\cdot)$ is obtained by invoking Lemmata 5.10, 5.14 and 5.15, see [Bar16, Lem. 3.1] or [Loh19, Ineq. (4.122)]. The estimate for $m_h(\cdot;\cdot,\cdot)$ follows similarly.□

5.4.3 A priori error estimate

To state our main result, we need to define some auxiliary quantities. For $t \ge 0$, let $\vartheta_h(\cdot,t) = \sum_{i=1}^N \vartheta_i(t)\,\varphi_i \in \mathrm{V}_h$ be the *discrete error* $\vartheta_h(\cdot,t) := \mathrm{I}_h u(\cdot,t) - u_h(\cdot,t)$ and define

$$\begin{aligned} q(T) := \sum_{\substack{i,j=1\\ i<j}}^N m_{ij}(\vartheta_i(T) - \vartheta_j(T))^2 + \int_0^T \Bigg[\int_{\Gamma_+} \vartheta_h^2\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s - \sum_{i=1}^N \vartheta_i^2 \int_{\Gamma_-} \varphi_i\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s \\ - \sum_{\substack{i,j=1\\ i<j}}^N (\vartheta_i - \vartheta_j)^2 \int_{\Gamma_-} \varphi_i\,\varphi_j\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s + \gamma d_h(u_h; u_h, u_h) + \frac{\gamma h}{\lambda} m_h(u_h; \dot{u}_h, \dot{u}_h) \Bigg] \mathrm{d}t, \\ z(T) := \int_0^T \left[\|\partial_t u\|^2_{\mathrm{H}^2(\Omega)} + \|u\|^2_{\mathrm{H}^2(\Omega)} + \|u\|^2_{\mathrm{H}^2(\Gamma_-)} + |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} \right] \mathrm{d}t. \end{aligned}$$

Theorem 5.17 (Semi-discrete a priori error estimate)
Let the assumptions made in Section 5.4.1 be satisfied. Assume that there is a finite time $T > 0$ such that $\boldsymbol{v}(\cdot,t) \in \mathbf{W}^{1,\infty}(\Omega)$ and $\nabla \cdot \boldsymbol{v}(\cdot,t) = 0$ in Ω for all $t \in (0,T)$. Let $u_h(\cdot,t)$ and $\dot{u}_h(\cdot,t)$ satisfy (5.12) and, additionally, the compatibility condition (5.18) with a constant $\gamma \in (0,1)$ independent of h for all $t \in (0,T)$. Then there exist positive constants $C_1 = C_1(d,\boldsymbol{v})$, $C_2 = C_2(d,\boldsymbol{v},\gamma)$, and $C_3 = C_3(d)$ such that the estimate

$$\|u(\cdot,T) - u_h(\cdot,T)\|_{\mathrm{L}^2(\Omega)} \le C_3 h^2 |u(\cdot,T)|_{\mathrm{H}^2(\Omega)} + \sqrt{y(T) + C_1 \int_0^T \mathrm{e}^{C_1(T-t)}\, y(t)\,\mathrm{d}t} \quad (5.20)$$

holds for the exact solution $u(\cdot,T)$ of the continuous problem (5.2), the exact solution $u_h(\cdot,T)$ of the semi-discrete problem (5.12), and $y(T) := C_2 h\, z(T) - q(T)$. ◊

Corollary 5.18 (Convergence order of the semi-discrete MCL scheme)
Under the assumptions of Theorem 5.17, the a priori error estimate

$$\|u(\cdot,T) - u_h(\cdot,T)\|_{\mathrm{L}^2(\Omega)} \le C_3 h^2 |u(\cdot,T)|_{\mathrm{H}^2(\Omega)} + \sqrt{\mathrm{e}^{C_1 T}\, C_2 h\, \|z\|_{\mathrm{L}^\infty(0,T)}} \le C_4\, h^{\frac{1}{2}} \quad (5.21)$$

holds with a constant $C_4 = C_4(C_1, C_2, C_3, T, u, \hat{u}) > 0$, which behaves as $\mathrm{e}^{C_1 T/2}$. ◊

Proof of Corollary 5.18:
Since q and z are nonnegative functions, we may use the estimate $y(T) \le C_2 h\, z(T)$ in (5.20). The claim follows by calculating the integral of the exponential function. □

Proof of Theorem 5.17:
This proof combines recent results on AFC schemes [Bar16, Loh19] with a new way of proving a priori error estimates for nonconforming discretizations of the advection equation [Rup21]. A particular similarity of the approach developed in [Rup21] to our theory is that both apply to semi-discrete formulations.

We introduce the *interpolation error* $\Theta(t) = \Theta(u,h;t) := u(\cdot,t) - \mathrm{I}_h u(\cdot,t)$ and subtract (5.12) from (5.2). Setting $w = w_h = \vartheta_h$, we obtain the error equation

$$\overbrace{\int_\Omega \vartheta_h \frac{\partial u}{\partial t}\,\mathrm{d}\boldsymbol{x} - \sum_{i=1}^N \vartheta_i\, m_i \frac{\mathrm{d}u_i}{\mathrm{d}t}}^{\Xi_1} + \overbrace{a(u,\vartheta_h) - a_h(u_h,\vartheta_h)}^{\Xi_2}$$
$$= \underbrace{b(\vartheta_h) - b_h(\vartheta_h)}_{\Xi_3} + \underbrace{d_h(u_h;u_h,\vartheta_h) - m_h(u_h;\dot{u}_h,\vartheta_h)}_{\Xi_4}.$$

Recall that the identity $m_i = \sum_{j=1}^N m_{ij}$ holds for row sum mass lumping. Using this decomposition of m_i and the identities $u = \Theta + \vartheta_h + \mathrm{I}_h u - \vartheta_h$, $u_h = \mathrm{I}_h u - \vartheta_h$, we find that

$$\Xi_1 = \int_\Omega \vartheta_h \frac{\partial \Theta}{\partial t}\,\mathrm{d}\boldsymbol{x} + \int_\Omega \vartheta_h \frac{\mathrm{d}\vartheta_h}{\mathrm{d}t}\,\mathrm{d}\boldsymbol{x} + \sum_{i=1}^N \vartheta_i \frac{\mathrm{d}}{\mathrm{d}t}\Big(\sum_{j=1}^N m_{ij}[(\mathrm{I}_h u)_j - \vartheta_j] - m_i\,[(\mathrm{I}_h u)_i - \vartheta_i] \Big)$$

$$= \int_\Omega \vartheta_h \frac{\partial \Theta}{\partial t} \,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{i,j=1}^N \vartheta_i\, m_{ij} \frac{\mathrm{d}}{\mathrm{d}t}\left[(\mathrm{I}_h u)_j - (\mathrm{I}_h u)_i - (\vartheta_j - \vartheta_i)\right]$$
$$= \int_\Omega \vartheta_h \frac{\partial \Theta}{\partial t} \,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{\substack{i,j=1\\ i<j}}^N (\vartheta_i - \vartheta_j)\, m_{ij} \frac{\mathrm{d}}{\mathrm{d}t}\left[(\mathrm{I}_h u)_j - (\mathrm{I}_h u)_i - (\vartheta_j - \vartheta_i)\right].$$

Arguing as in the proof of Theorem 5.8, we invoke the divergence theorem, Lemma 5.7 as well as the identities $u = \Theta + \mathrm{I}_h u$ and $u_h = \mathrm{I}_h u - \vartheta_h$, which yields

$$\begin{aligned}
\Xi_2 = &\int_\Omega \vartheta_h\, \boldsymbol{v}\cdot\nabla\Theta \,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\partial\Omega} \vartheta_h^2\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s - \int_{\Gamma_-} \vartheta_h\, \Theta\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s\\
&- \int_{\Gamma_-} \vartheta_h\, \mathrm{I}_h u\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s + \sum_{i=1}^N \vartheta_i\,(u(\boldsymbol{x}_i) - \vartheta_i)\int_{\Gamma_-} \varphi_i\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s\\
= &\int_\Omega \vartheta_h\, \boldsymbol{v}\cdot\nabla\Theta \,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\Gamma_+} \vartheta_h^2\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s - \int_{\Gamma_-} \vartheta_h\, \Theta\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s\\
&- \frac{1}{2}\sum_{\substack{i,j=1\\ i<j}}^N (\vartheta_i - \vartheta_j)^2 \int_{\Gamma_-} \varphi_i\,\varphi_j\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s - \frac{1}{2}\sum_{i=1}^N \vartheta_i^2 \int_{\Gamma_-} \varphi_i\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s\\
&+ \int_{\Gamma_-} \Big(\sum_{i=1}^N \vartheta_i\,\varphi_i\,(u(\boldsymbol{x}_i) - \mathrm{I}_h u)\Big)\, \boldsymbol{v}\cdot\boldsymbol{n}\,\mathrm{d}s.
\end{aligned}$$

As in [Kna03, Thm. 3.43], we exploit transformation to the reference element, shape-regularity, and the equivalence of norms in finite dimensional spaces to show that

$$\sum_{i=1}^N \int_\Gamma (v_i\,\varphi_i)^2 \,\mathrm{d}s \le \sum_{i=1}^N \int_\Gamma v_i^2\,\varphi_i \,\mathrm{d}s \le C\|v_h\|^2_{\mathrm{L}^2(\Gamma)} \qquad \forall v_h \in \mathrm{V}_h,\ \Gamma \in \mathcal{F}_{\partial\Omega}. \tag{5.22}$$

To derive an estimate for Ξ_3, we rewrite the boundary integrals as a sum of integrals over faces. On each face $\Gamma \in \mathcal{F}_-$, we use the estimate $|\hat{u}_i^k - \hat{u}| \le C h_\Gamma |\nabla \hat{u}|$, where $h_\Gamma = \mathrm{diam}(\Gamma)$. In addition, we invoke Young's inequality, estimate (5.22), Lemma 5.12, and incorporate $\lambda = \|\boldsymbol{v}\|_{\mathbf{L}^\infty(\Omega\times\mathbb{R}_+)^d}$ into the constant C, which yields

$$\begin{aligned}
\Xi_3 &= \sum_{i=1}^N \sum_{\Gamma_k \in \mathcal{F}_i} \int_{\Gamma_k} \vartheta_i\varphi_i\,(\hat{u}_i^k - \hat{u})\,\min\{0, \boldsymbol{v}\cdot\boldsymbol{n}\}\,\mathrm{d}s\\
&\le C \sum_{\Gamma\in\mathcal{F}_-} \sum_{i=1}^N \int_\Gamma \Big[h_\Gamma(\vartheta_i\varphi_i)^2 + \frac{1}{h_\Gamma}|\hat{u}_i^k - \hat{u}|^2\Big]\,\mathrm{d}s\\
&\le C \sum_{\Gamma\in\mathcal{F}_-} h_\Gamma \left(\|\vartheta_h\|^2_{\mathrm{L}^2(\Gamma)} + |\hat{u}|^2_{\mathrm{H}^1(\Gamma)}\right) \le C\|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch|\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)}.
\end{aligned} \tag{5.23}$$

For the nonlinear terms in Ξ_4, we use Lemma 5.4, Young's inequality, the compatibility condition (5.18) with constant $\gamma \in (0,1)$, and Lemma 5.16 to deduce

$$\Xi_4 = d_h(u_h; u_h, \mathrm{I}_h u) - d_h(u_h; u_h, u_h) + m_h(u_h; \dot{u}_h, u_h) - m_h(u_h; \dot{u}_h, \mathrm{I}_h u)$$

$$
\begin{aligned}
&\leq \frac{\gamma}{2} d_h(u_h; u_h, u_h) + \frac{1}{2\gamma} d_h(u_h; \mathrm{I}_h u, \mathrm{I}_h u) - \gamma d_h(u_h; u_h, u_h) \\
&\quad - \frac{\gamma h}{\lambda} m_h(u_h, \dot{u}_h, \dot{u}_h) + \frac{\gamma h}{2\lambda} m_h(u_h; \dot{u}_h, \dot{u}_h) + \frac{\lambda}{2\gamma h} m_h(u_h; \mathrm{I}_h u, \mathrm{I}_h u) \\
&\leq -\frac{\gamma}{2} d_h(u_h; u_h, u_h) - \frac{\gamma h}{2\lambda} m_h(u_h, \dot{u}_h, \dot{u}_h) + Ch \|u\|^2_{\mathrm{H}^2(\Omega)},
\end{aligned}
$$

where the factor $1/\gamma$ was incorporated into the constant C. Combining the above identities for Ξ_1 and Ξ_2 with the inequalities for Ξ_3 and Ξ_4 produces the estimate

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + \sum_{\substack{i,j=1 \\ i<j}}^{N} m_{ij} \frac{\mathrm{d}}{\mathrm{d}t} (\vartheta_i - \vartheta_j)^2 + \int_{\Gamma_+} \vartheta_h^2 \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s - \sum_{i=1}^{N} \vartheta_i^2 \int_{\Gamma_-} \varphi_i \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s \\
&\qquad - \sum_{\substack{i,j=1 \\ i<j}}^{N} (\vartheta_i - \vartheta_j)^2 \int_{\Gamma_-} \varphi_i \, \varphi_j \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s + \gamma d_h(u_h; u_h, u_h) + \frac{\gamma h}{\lambda} m_h(u_h, \dot{u}_h, \dot{u}_h) \\
&\leq -2 \int_\Omega \vartheta_h \frac{\partial \Theta}{\partial t} \, \mathrm{d}\boldsymbol{x} + 2 \sum_{\substack{i,j=1 \\ i<j}}^{N} (\vartheta_i - \vartheta_j) \, m_{ij} \left[(\mathrm{I}_h \partial_t u)_i - (\mathrm{I}_h \partial_t u)_j \right] \\
&\qquad - 2 \int_\Omega \vartheta_h \, \boldsymbol{v} \cdot \nabla \Theta \, \mathrm{d}\boldsymbol{x} + 2 \int_{\Gamma_-} \vartheta_h \, \Theta \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} \\
&\qquad - 2 \int_{\Gamma_-} \Big(\sum_{i=1}^{N} \vartheta_i \varphi_i \, (u(\boldsymbol{x}_i) - \mathrm{I}_h u) \Big) \, \boldsymbol{v} \cdot \boldsymbol{n} \, \mathrm{d}s + Ch \|u\|^2_{\mathrm{H}^2(\Omega)} =: (\star). \qquad (5.24)
\end{aligned}
$$

The terms on the right hand side of inequality (5.24) are now bounded using standard arguments. Specifically, we make use of the assumptions on the mesh and of Young's inequality, apply Lemma 5.10 to $\Theta = u - \mathrm{I}_h u$ and $\partial_t \Theta$, invoke Lemmata 5.11 through 5.15 and argue as in the derivation of (5.23) to obtain

$$
\begin{aligned}
(\star) &\leq \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^4 |\partial_t u|^2_{\mathrm{H}^2(\Omega)} + Ch^2 \sum_{K \in \mathcal{K}_h} |\vartheta_h|_{\mathrm{H}^1(K)} |\mathrm{I}_h \partial_t u - \partial_t u + \partial_t u|_{\mathrm{H}^1(K)} \\
&\quad + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^2 |u|^2_{\mathrm{H}^2(\Omega)} + \lambda \sum_{\Gamma \in \mathcal{F}_-} \Big(h_\Gamma \|\vartheta_h\|^2_{\mathrm{L}^2(\Gamma)} + \frac{1}{h_\Gamma} \|\Theta\|^2_{\mathrm{L}^2(\Gamma)} \Big) + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad + Ch |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} + C \sum_{\Gamma \in \mathcal{F}_-} h_\Gamma \int_\Gamma \Big[\sum_{i=1}^{N} (\vartheta_i \varphi_i)^2 + |\nabla (\mathrm{I}_h u - u + u)|^2 \Big] \, \mathrm{d}s + Ch \|u\|^2_{\mathrm{H}^2(\Omega)} \\
&\leq \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^4 |\partial_t u|^2_{\mathrm{H}^2(\Omega)} + Ch \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^3 |\partial_t u|^2_{\mathrm{H}^2(\Omega)} + Ch |\partial_t u|^2_{\mathrm{H}^1(\Omega)} \\
&\quad + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^2 |u|^2_{\mathrm{H}^2(\Omega)} + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + Ch^2 |u|^2_{\mathrm{H}^2(\Omega)} + C \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad + Ch |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} + C \sum_{\Gamma \in \mathcal{F}_-} h_\Gamma \left(\|\vartheta_h\|^2_{\mathrm{L}^2(\Gamma)} + \|u\|^2_{\mathrm{H}^2(\Gamma)} \right) + Ch \|u\|^2_{\mathrm{H}^2(\Omega)} \\
&\leq C_1 \|\vartheta_h\|^2_{\mathrm{L}^2(\Omega)} + C_2 h \left(\|\partial_t u\|^2_{\mathrm{H}^2(\Omega)} + \|u\|^2_{\mathrm{H}^2(\Omega)} + \|u\|^2_{\mathrm{H}^2(\Gamma_-)} + |\hat{u}|^2_{\mathrm{H}^1(\Gamma_-)} \right).
\end{aligned}
$$

We now integrate in time observing that, by our definition of the discrete initial data, we have $\vartheta_h(0) \equiv 0$. At this stage, we recall the previously given definitions of q and z, which enables us to write the resulting inequality as

$$\|\vartheta_h(T)\|^2_{\mathrm{L}^2(\Omega)} + q(T) \leq C_1 \int_0^T \|\vartheta_h(t)\|^2_{\mathrm{L}^2(\Omega)} \,\mathrm{d}t + C_2 h\, z(T).$$

Using Grönwall's Lemma as in [Dol15, Lem. 1.9], we obtain

$$\|\vartheta_h(T)\|^2_{\mathrm{L}^2(\Omega)} + q(T) + C_1 \int_0^T \mathrm{e}^{C_1(T-t)}\, q(t)\,\mathrm{d}t \leq C_1 \int_0^T \mathrm{e}^{C_1(T-t)}\, C_2 h\, z(t)\,\mathrm{d}t + C_2 h\, z(T).$$

The triangle inequality applied to $u - u_h = \Theta + \vartheta_h$ then yields the error estimate (5.20) by Lemma 5.10. □

We conclude the theoretical discussion with a few remarks regarding the derived error estimate. Let us first point out that in the general setting with unspecified correction factors α_{ij} our result is indeed optimal (cf. Section 5.5.1). If we set all correction factors equal to zero, we obtain the low order method, which cannot be expected to be more than $\frac{1}{2}$ order accurate in general.

A drawback of our current approach is that the constant on the right hand side of the a priori error estimate (5.21) depends exponentially on the time T. Kučera and Shu [Kuč18] demonstrate that exponentially increasing constants can be avoided in some situations. They discretize the advection equation using discontinuous Galerkin methods and derive an error estimate without invoking Grönwall's inequality. It would be interesting to investigate the merit of their approach for the purposes of our analysis.

Let us briefly remark that we assumed all integrals appearing in the bilinear and linear forms $a_h(\cdot,\cdot)$ and $b_h(\cdot)$ to be evaluated exactly. In fact, even the energy estimate stated in Theorem 5.8 was derived under this assumption. For polynomial velocities one can indeed employ a quadrature rule of sufficiently high order to accurately compute all integrals. For general velocities, the theory presented in this chapter needs to be adapted to include quadrature errors. As is common for linear finite elements [Ern04, Thm. 8.5], we recommend to employ quadrature rules that are exact for polynomials in $\mathbb{P}_2$ and $\mathbb{P}_3$ for volume and boundary integrals, respectively.

Admittedly, a major limitation of Theorem 5.17 is the fact that the estimate is valid only for problems with exact solutions of very high regularity. In particular, the assumption that $\partial_t u$ is H^2 in space is restrictive. In our opinion, the adaptation of the proofs in [Bar16, Loh19] to the time-dependent setting necessitates this regularity. One can argue that if the exact solution is smooth enough for Theorem 5.17 to be applicable, a limiter may not even be needed and we could instead employ a stabilized Galerkin method. Since this strategy does not guarantee the validity of discrete maximum principles, AFC schemes provide an appealing alternative. Therefore, theoretical investigations of these methods should be undertaken. It is hoped that our results may serve as a stepping stone for further efforts in this direction.

5.5 Numerical examples

Let us now corroborate the theoretical results of this chapter with numerical experiments. Recall that our stability and convergence proofs rely on the compatibility condition (5.18). In general, this condition is not fulfilled by the standard MCL approach. However, if $\dot{u}_h$ is set to zero, i. e., if the mass lumping error is not compensated, (5.18) holds due to Lemma 5.4. To distinguish between the standard MCL scheme and the lumped-mass version, we employ the acronyms MCL-L and MCL-0, respectively. Here the letter L stands for *low order time derivatives*, while 0 stands for *zero time derivatives*. Other methods used for comparative purposes are specified below.

In the following sections, we verify that approximations converge at least as fast as the provable rate of $\frac{1}{2}$. Moreover, we stress the need for stabilization by the use of low order time derivatives and present a comparison of results obtained with MCL and FCT. Finally, we perform an a posteriori check to see for which values of the parameter γ the compatibility condition (5.18) is satisfied by the MCL-L scheme.

5.5.1 Experimental orders of convergence

In this section, we solve the one-dimensional advection equation with constant velocity $v = 1$. The spatial domain $\Omega = (0, 1)$ has periodic boundaries. Thus, at each time instant $T \in \mathbb{N}_0$, the exact solution coincides with the initial condition. In this example, we use $u_0(x) = \exp(-100(x - 0.5)^2)$.

We study the experimental orders of convergence for discrete upwinding (LOW), MCL-L, and MCL-0 schemes using SSP2 RK time stepping and CFL parameter $\nu = 0.5$. While values as large as $\nu = 1$ can safely be employed without causing violations of maximum principles, smaller values may be necessary to observe certain rates of convergence. Alternatively, SSP3 RK time stepping can be used to improve temporal accuracy. As discussed in Section 4.5.2.1, SSP1 RK time stepping should not be employed in combination with flux correction schemes based on continuous finite elements (see also [Kuz12a, Sec. 4]). In this study, we employ sequences of nested meshes with generally nonuniform mesh size h obtained by randomly perturbing the positions of the interior mesh vertices of the coarsest grid. The relative mesh sizes $\min_{K \in \mathcal{K}_h} h_K/h$ of the three sequences are 1 (uniform), ≈ 0.69 (mildly perturbed), and ≈ 0.087 (severely perturbed), respectively. We present the $\mathrm{L}^2(\Omega)$ errors at the final time $T = 1$ and the corresponding EOC in Tabs. 5.1 to 5.3. The observed rates of convergence are in accordance with our expectations. As suggested by Corollary 5.18, discrete upwinding converges at least with the rate of $\frac{1}{2}$. Actually, the low order method becomes first order accurate on very fine uniform meshes. Our preferred MCL-L scheme produces second order accurate results in this test. If no correction of the mass lumping error is performed, the order of accuracy deteriorates, while still exceeding the provable rate of $\frac{1}{2}$. In this example, the influence

of mesh perturbations on the results is insignificant. A decay in the convergence rate of MCL for a steady problem was observed on perturbed 2D meshes in [Kuz20a, Sec. 6.1].

$1/h$	LOW	EOC	MCL-L	EOC	MCL-0	EOC
32	2.21E-01		6.92E-02		9.93E-02	
64	1.75E-01	0.34	2.07E-02	1.74	4.46E-02	1.16
128	1.26E-01	0.47	4.65E-03	2.16	1.65E-02	1.44
256	8.18E-02	0.62	1.12E-03	2.06	5.29E-03	1.64
512	4.84E-02	0.76	2.76E-04	2.02	1.65E-03	1.68

Table 5.1: Convergence history for the one-dimensional advection equation on a sequence of uniform periodic meshes. The $\|\cdot\|_{\mathrm{L}^2(\Omega)}$ errors at $T = 1$ and the corresponding EOC for $u_0(x) = \mathrm{e}^{-100(x-0.5)^2}$.

$1/h$	LOW	EOC	MCL-L	EOC	MCL-0	EOC
32	2.21E-01		7.06E-02		9.97E-02	
64	1.75E-01	0.34	2.22E-02	1.67	4.51E-02	1.15
128	1.26E-01	0.47	4.95E-03	2.17	1.70E-02	1.40
256	8.23E-02	0.62	1.16E-03	2.09	5.48E-03	1.64
512	4.88E-02	0.75	2.87E-04	2.01	1.71E-03	1.68

Table 5.2: Convergence history for the one-dimensional advection equation on a sequence of mildly perturbed periodic meshes. The $\|\cdot\|_{\mathrm{L}^2(\Omega)}$ errors at $T = 1$ and the corresponding EOC for $u_0(x) = \mathrm{e}^{-100(x-0.5)^2}$.

$1/h$	LOW	EOC	MCL-L	EOC	MCL-0	EOC
32	2.24E-01		1.01E-01		1.16E-01	
64	1.82E-01	0.30	4.85E-02	1.05	5.65E-02	1.03
128	1.36E-01	0.43	1.25E-02	1.96	2.33E-02	1.28
256	9.08E-02	0.58	2.98E-03	2.07	7.72E-03	1.59
512	5.51E-02	0.72	7.26E-04	2.04	2.46E-03	1.65

Table 5.3: Convergence history for the one-dimensional advection equation on a sequence of severely perturbed periodic meshes. The $\|\cdot\|_{\mathrm{L}^2(\Omega)}$ errors at $T = 1$ and the corresponding EOC for $u_0(x) = \mathrm{e}^{-100(x-0.5)^2}$.

5.5.2 On the stabilizing effect of low order time derivatives

In Section 3.4.3.2, we presented an example where the consistent Galerkin time derivative $\dot{u}_h^{\mathrm{G}}$ is employed to correct the mass lumping error through the antidiffusive fluxes. This approach produces spurious ripples in the solution profiles and should therefore be

avoided. The consequences of lacking stabilization can best be observed in the context of the advection equation, which is why we compare the results of computations with consistent Galerkin and stabilized approximations in this section.

We consider the same setup as in the previous section with the exception that the initial condition u_0 is replaced by [Haj21a]

$$u_0(x) = \begin{cases} 1 & \text{if } 0.2 \leq x \leq 0.4, \\ \exp(10)\exp(\frac{1}{0.5-x})\exp(\frac{1}{x-0.9}) & \text{if } 0.5 < x < 0.9, \\ 0 & \text{otherwise.} \end{cases} \tag{5.25}$$

This profile features discontinuities as well as a C^∞ region. In Fig. 5.1 we display standard continuous Galerkin approximations obtained with four different combinations of time stepping schemes and CFL parameters on a uniform, a mildly perturbed and a severely perturbed mesh with 128 elements in each case. Spurious ripples that are not local to the vicinity of the discontinuities can be observed in all profiles. Although limiters can remove these oscillations, the quality of approximations obtained in this fashion is usually poor compared to solutions obtained with flux limiters applied to a stabilized target discretization.

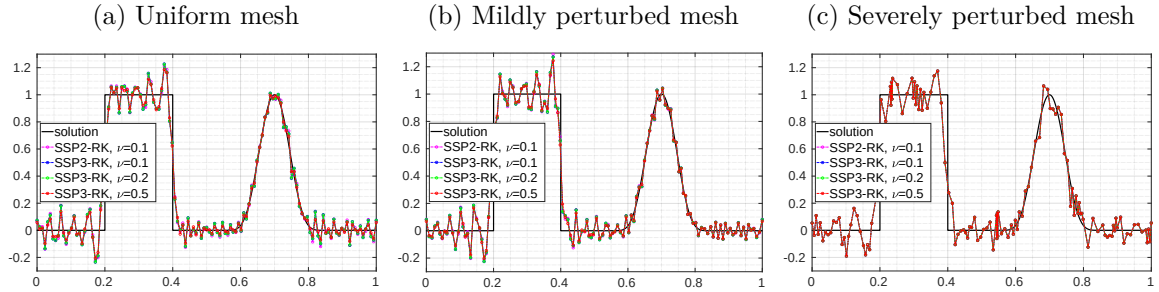


Figure 5.1: One-dimensional advection equation with initial condition (5.25). Consistent Galerkin approximations at $T = 1$ obtained with SSP RK time stepping on periodic meshes consisting of 128 elements.

Next, we compute approximations of the stabilized target scheme, i. e., (5.11) with $\alpha_{ij} = 1$ for all $i \in \{1, \dots, N\}$, $j \in \mathcal{N}_i \setminus \{i\}$. These are compared to the profiles obtained with the LOW, MCL-L, MCL-0, and MCL-L-SDE schemes. The latter incorporates the semi-discrete entropy fix from Section 3.3.6 for the entropy pair $(\eta(u), q(u)) = (u^2/2, vu^2/2)$ with corresponding entropy potential $\psi(u) = q(u)$. Although the issue of nonuniqueness of weak solutions does not arise for the linear advection equation, we found it instructive to illustrate the effect that this limiter has on the approximations. In practice, we advise against applying entropy fixes in discretizations of linear conservation laws. SSP2 RK time stepping with CFL parameter $\nu = 1$ is employed in combination with all spatial semi-discretizations. The results of this study are displayed in Fig. 5.2.

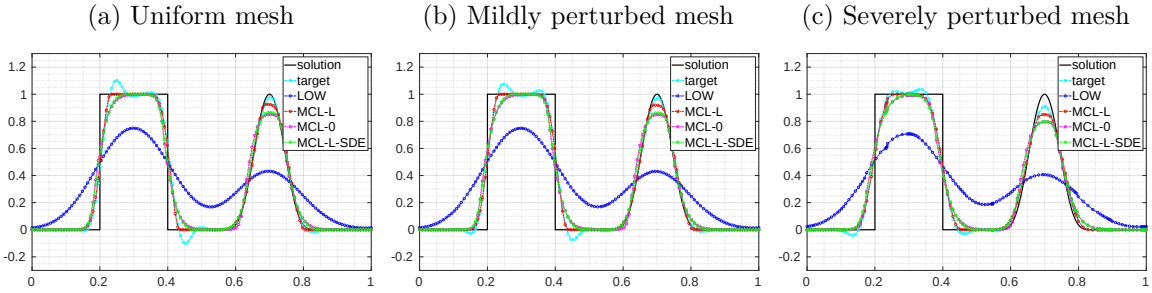


Figure 5.2: One-dimensional advection equation with initial condition (5.25). Stabilized Galerkin approximations at $T = 1$ obtained with SSP2 RK time stepping and $\nu = 1$ on periodic meshes consisting of 128 elements.

We observe significant improvements in the solution quality for the unlimited target scheme compared to the consistent Galerkin approximations displayed in Fig. 5.1. Numerical results obtained with LOW, MCL-L and MCL-0 exhibit behavior similar to that observed in Section 5.5.1 In particular, the MCL-0 scheme produces a nonsymmetric profile in the left part of the domain, which can be attributed to dispersive errors [Tho16]. The same is true for the MCL-L-SDE scheme, which suggests that the entropy fix has an adverse effect on the solution quality of approximations to linear advection problems.

5.5.3 Comparison of MCL with FCT

Let us now compare the MCL-L scheme with the two FCT strategies discussed in Section 5.2.3. As test problem, we select the 2D solid body rotation benchmark [Zal79, LeV96, Kuz12a] in which $\Omega = (0,1)^2$, $\boldsymbol{v}(x,y) = 2\pi\,(0.5-y, x-0.5)^\mathsf{T}$, $\hat{u} = 0$ and

$$u_0(x,y) = \begin{cases} u_0^{\text{cone}}(x,y) & \text{if } r(x,y;0.5,0.25) \le r_0, \\ u_0^{\text{bump}}(x,y) & \text{if } r(x,y;0.25,0.5) \le r_0, \\ 1 & \text{if } r(x,y;0.5,0.75) \le r_0 \ \wedge\ (|x-0.5| \ge 0.025 \ \vee\ y \ge 1-r_0)\,, \\ 0 & \text{otherwise,} \end{cases}$$

where $r(x,y;x_0,y_0) := \sqrt{(x-x_0)^2+(y-y_0)^2}$, $r_0 = 0.15$, and

$$u_0^{\text{cone}}(x,y) = 1 - \frac{r(x,y;0.5,0.25)}{r_0},$$

$$u_0^{\text{bump}}(x,y) = \frac{1}{4}\left(1+\cos\left(\frac{\pi\, r(x,y;0.25,0.5)}{r_0}\right)\right).$$

In this example, a cone, a smooth bump and a slotted cylinder rotate around the domain center. At each time instant $T \in \mathbb{N}_0$, the exact solution is equal to the initial condition, which is shown in Fig. 5.3. The numerical results displayed in this section are visualized with the open source C++ software GLVis.

(a) Side view of $u_h(\cdot, 0)$ and the mesh

(b) Top view of $u_h(\cdot, 0)$, contour lines

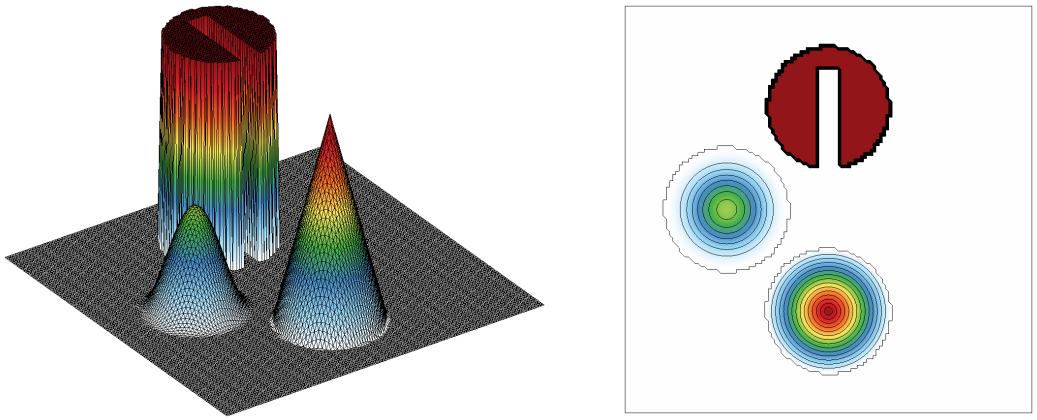


Figure 5.3: Exact initial condition of the solid body rotation [LeV96] interpolated in V_h for a uniform triangular mesh consisting of $2 \cdot 128^2$ elements.

We solve this problem numerically using triangular meshes and $h = c/128$, where $c = \sqrt{2}$ for uniform grids and $c = 1$ for unstructured ones. For time stepping we employ the SSP2 RK method with constant time steps $\Delta t = 5 \cdot 10^{-4}$ and $\Delta t = 3.125 \cdot 10^{-4}$, respectively. Numerical results are displayed in Fig. 5.4 and the approximate $\mathrm{L}^2(\Omega)$ errors e_T^2 at the final time $T = 1$ are presented in the captions along with the maximum solution value $u_h^{\max}$ for each approximation. The minimum value of each approximation is zero up to machine precision.

All three schemes under investigation yield results of similar quality. In this example Zalesak's algorithm produces the most accurate results, followed by localized FCT and then the MCL scheme. We remark that the situation may be different if the employed spatial and temporal discretization parameters are modified. Compared to MCL, the value of Δt has more influence on the accuracy of FCT schemes (see Remark 5.5). Since the differences in the results of this section are marginal, all three schemes can be recommended for applications of time-dependent advection problems.

5.5.4 A posteriori compatibility check

The compatibility condition (5.18) turned out to be an invaluable tool for our theoretical investigations. Unfortunately, we are unable to prove that the MCL-L scheme automatically produces compatible pairs $(u_h, \dot{u}_h)$ under suitable assumptions. However, it is easy to check for which values of $\gamma \in (0, 1)$ condition (5.18) is fulfilled *a posteriori*. Indeed, (5.18) is equivalent to

$$\gamma \leq \frac{d_h(u_h; u_h, u_h) - m_h(u_h; \dot{u}_h, u_h)}{d_h(u_h; u_h, u_h) + \frac{h}{\lambda} m_h(u_h; \dot{u}_h, \dot{u}_h)} \tag{5.26}$$

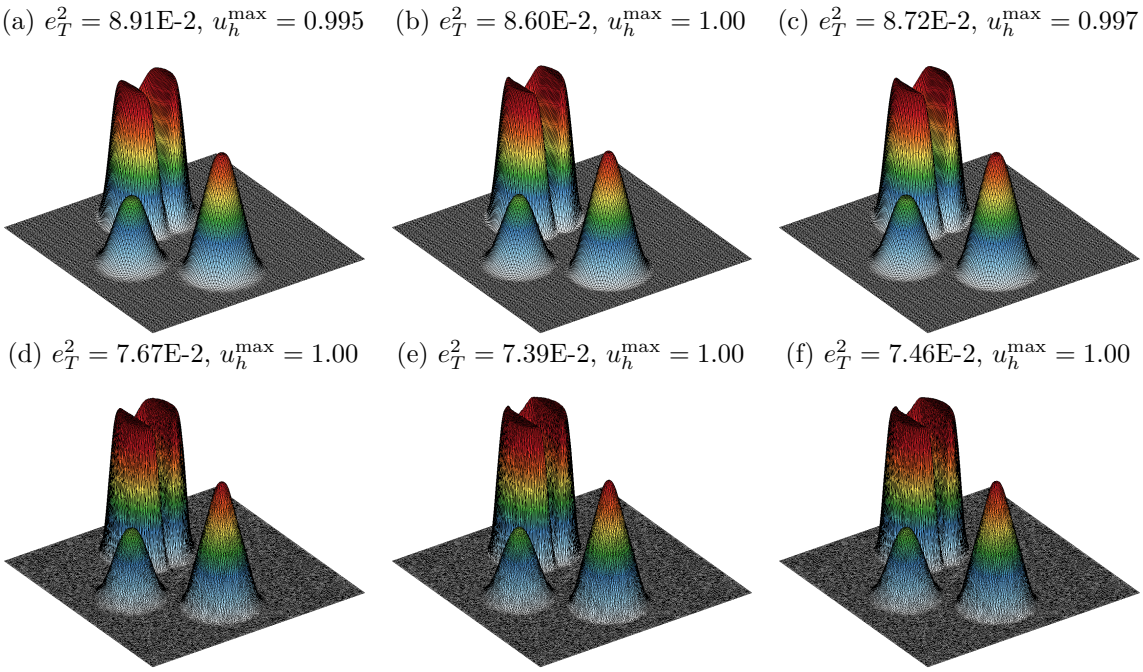


Figure 5.4: Solid body rotation for the 2D advection equation [LeV96]. MCL-L [Kuz20a] (left), Zalesak's FCT [Zal79] (center), and localized FCT [Loh19] (right) approximations at $T = 1$. Solutions obtained on uniform (top row) and unstructured (bottom row) triangular meshes with SSP2 RK time stepping using $\Delta t = 5 \cdot 10^{-4}$ and $\Delta t = 3.125 \cdot 10^{-4}$, respectively.

if the denominator in the right hand side of (5.26) is nonzero (it is nonnegative due to Lemma 5.4). If the numerator in (5.26) is also nonnegative, this criterion yields an upper bound on γ. Having calculated these a posteriori bounds via (5.26), one can check how they behave upon mesh refinement. Two issues that lead to a violation of compatibility can occur in practice. First, (5.26) can produce a negative upper bound on γ, a case that is not covered by our theory. Secondly, γ may approach zero upon mesh refinement, which would cause the constant in the leading order term of our error estimate to approach infinity. We found the former concern to be valid on perturbed one-dimensional meshes. Using smaller time steps does not resolve incompatibility issues, which seem to be caused by triangulations of bad quality. In such instances, our stability and error estimates are not applicable to MCL-L but remain valid for the LOW and MCL-0 schemes, as well as for the method proposed in [Haj21b, Sec. 3.3].

Having performed the described a posteriori check for various test problems, we conjecture that compatibility of $(u_h, \dot{u}_h)$ holds for the MCL-L scheme on uniform meshes. Below we report the results of our experiments in which we compute the values of the right hand side of (5.26). First, we consider four one-dimensional test problems. In each case, the spatial domain $\Omega = (0, 1)$ is equipped with periodic boundaries and the velocity is $v = 1$. The first and second tests are the same as in Sections 5.5.1 and 5.5.2.

In the third and fourth tests, the final time is $T = 0.5$ and the initial conditions read

$$u_0(x) = \begin{cases} 0.5\left(1 + \cos\left(\frac{\pi}{0.15}(x - 0.25)\right)\right) & \text{if } |x - 0.25| < 0.15, \\ 0 & \text{otherwise,} \end{cases}$$

and $u_0(x) = \max\{0, 1 - 10|x - 0.2|\}$, respectively.

We solve each of these problems on a hierarchy of uniform meshes using SSP2 RK time stepping with CFL parameters $\nu \in \{1, 0.1\}$. The largest value of γ for which (5.18) is satisfied during the whole simulation is presented in Tab. 5.4.

$1/h$	Test 1	Test 2	Test 3	Test 4
32	0.58	0.65	0.62	0.67
64	0.54	0.54	0.56	0.56
128	0.52	0.53	0.52	0.55
256	0.51	0.52	0.51	0.53
512	0.50	0.52	0.51	0.52

$1/h$	Test 1	Test 2	Test 3	Test 4
32	0.62	0.66	0.64	0.66
64	0.56	0.58	0.58	0.60
128	0.53	0.55	0.54	0.57
256	0.52	0.53	0.52	0.57
512	0.51	0.52	0.51	0.55

Table 5.4: Maximum values of γ over all time steps for four 1D examples. Results obtained on uniform meshes with SSP2 RK time stepping and CFL parameters $\nu = 1$ (left) and $\nu = 0.1$ (right).

Additionally, we repeat the solid body rotation test [LeV96] from Section 5.5.3 on sequences of uniform ($c = \sqrt{2}$) and unstructured ($c = 1$) triangular meshes and compute a posteriori values for γ via (5.26). The results of this study are summarized in Tab. 5.5, where `#TS` refers to the total number of employed time steps.

c/h	Uniform meshes	`#TS`	Unstructured meshes	$\min_{K\in\mathcal{K}_h} h_K/h$	`#TS`
32	$\gamma = 0.59$	500	$\gamma = 0.48$	0.32	625
64	$\gamma = 0.54$	1000	$\gamma = 0.45$	0.31	1250
128	$\gamma = 0.49$	2000	$\gamma = 0.46$	0.29	3200
256	$\gamma = 0.49$	4000	$\gamma = 0.47$	0.28	6400
512	$\gamma = 0.48$	8000	$\gamma = 0.47$	0.26	12500

Table 5.5: Maximum values of γ over all time steps for the solid body rotation [LeV96]. Results obtained on uniform and unstructured meshes with SSP2 RK time stepping and constant time steps.

We observe slightly larger maximum values of γ on coarse girds than on fine meshes. The use of smaller time steps seems to have marginal influence on the results. In all cases, we have $\gamma > 0.4$, which is consistent to the value that we used in [Haj21b] to enforce (5.18). Contrary to the 1D case, (5.26) does not produce negative values for γ even on unstructured meshes in 2D. This observation leads us to believe that the low order time derivative $\dot{u}_h$ is compatible to u_h even on nonuniform meshes that are regular in some sense. In fact, the only violations of (5.18) that we observed were obtained on nonuniform one-dimensional grids, which are of limited interest for practical purposes.

Chapter 6

Algebraic flux correction tools for discontinuous Galerkin methods

Thus far we performed algebraic flux correction only in the context of continuous Galerkin methods and (multi-)linear finite elements. However, the use of higher order spaces and/or nonconforming elements is feasible as well. To illustrate the applicability of AFC strategies to baseline discretizations of this kind, we present an extension of MCL to high order discontinuous Galerkin (DG) methods. Our scheme employs *Bernstein polynomial* basis functions as, e. g., in [Abg10, And17, Loh17b, Kuz20e]. Alternative choices of nodal bases lead to similar flux correction schemes [Paz21, Rue22].

This chapter is based on the author's paper [Haj21a]. In Section 6.1 we briefly motivate the use of DG methods and review scarce literature on algebraic limiters for these discretizations. Our generalization of MCL to high order nonconforming elements is introduced in Section 6.2 and numerical results are presented in Section 6.3.

6.1 Motivation and state of the art

Discontinuous Galerkin schemes can be interpreted as a bridge between classical finite element and finite volume methods. The DG approach pieces together weak formulations of conservation laws on individual mesh cells. Coupling is achieved by using numerical fluxes in integrals over element boundaries. Inside each cell, polynomials of degree $p \in \mathbb{N}_0$ are employed as test and trial functions. Contrary to standard finite elements, no continuity requirements across element interfaces are imposed in the DG setting. For $p > 0$, higher order test functions are included in the DG spaces. In contrast to FV schemes, no reconstructions are required to obtain well-resolved approximations. For many hyperbolic problems with smooth exact solutions, unconstrained DG methods deliver results of optimal accuracy, as shown, e. g., in [Paz19, Sec. 4.4]. These advantages of the DG approach are somewhat diminished by the presence of additional terms in weak formulations and additional degrees of freedom on internal boundaries of mesh cells. Indeed, the total number of unknowns is significantly larger for DG methods than for standard finite elements (of the same order) and FV methods on the same grid. We remark, however, that a fair comparison of alternative discretization strategies should be based on accuracy achieved with the same *total number of unknowns*. In such experiments, the quality of DG approximations may be comparable to that of numerical results obtained with continuous Galerkin schemes on finer grids [Haj20b, Sec. 8.1].

Using the characteristic functions of mesh cells as test functions in the DG weak formulation, one can derive evolution equations for cell averages of the numerical solution. These equations represent discrete conservation laws and exhibit the same structure as cell-centered FV schemes, but traces of the DG solution (rather than reconstructions from cell averages) are used to define numerical fluxes for $p > 0$. For applications in which the local conservation property is desired but the overhead cost of using a DG-$\mathbb{P}_1$/-$\mathbb{Q}_1$ discretization is unacceptable, the so-called *enriched Galerkin* (EG) method, originally proposed by Becker et al. [Bec03], may be a good alternative baseline discretization (see, e. g., [Rup21]). In the standard EG approach, the piecewise (multi-)linear continuous Galerkin space is enriched by piecewise constant basis functions. Flux correction schemes for such EG methods were developed in [Kuz20b]. For finite elements of degree $p > 1$, polynomial enrichments of degree up to $p - 1$ are needed to obtain stable EG algorithms for hyperbolic problems [Rup21]. A simple calculation shows that the total number of unknowns in the EG version is still smaller than that of a full DG method but much larger than in the case of a continuous approximation. Therefore, savings offered by the EG approach (as compared to full DG) become less significant as p increases.

In conclusion, standard DG schemes represent suitable high order target discretizations for hyperbolic problems. Therefore, we extend some of the flux correction tools discussed so far to the DG setting. Before doing so, we briefly review the literature on AFC schemes and some other limiters for DG discretizations.

A very popular class of limiting techniques for FV and DG methods are so-called *slope limiters* (see, e. g., [Bar89, Kuz10a, Zha11, Vat15, Haj19]). Such schemes adjust the gradients (and high order derivatives, if any) of polynomial approximations/reconstructions on mesh cells to preserve local bounds based on cell averages of the numerical solution in a neighborhood of the element. Approximations obtained with slope limiters are bound preserving as long as the cell averages stay in the admissible range. If necessary, discrete maximum principles for cell averages can be enforced using a flux limiter [Moe17]. In contrast to *algebraic* flux limiting techniques, which we favor in this thesis, slope limiters represent *geometric* postprocessing tools. Similarly to FCT algorithms, they can be interpreted as predictor-corrector methods and may inhibit convergence to steady state solutions. A comparison of slope and flux limiters for one-dimensional problems can be found in [LeV92, Ch. 16]. Any flux limiter for the 1D linear advection equation can be easily converted into an equivalent slope limiter and vice versa. However, this is generally not true for multidimensional and nonlinear problems.

Other promising strategies for nonlinear stabilization of DG methods include weighted essentially non-oscillatory (WENO) schemes, a posteriori subcell limiters, as well as shock capturing techniques based on artificial viscosity. A brief description of these approaches and corresponding references can be found in Section 1.1.

To the best of our knowledge, Anderson et al. [And17] were the first to apply algebraic flux correction tools in the context of DG methods. Their element-based FCT limiters are designed for high order Bernstein finite element discretizations of the linear advection

equation. To avoid the costly calculation of element matrices, matrix-free AFC schemes were developed by the author and his collaborators in [Haj20b, Haj20c] using residual distribution (RD) ideas to design the underlying low order method and a monolithic alternative to FCT. The applicability of our matrix-free RD-DG approaches is currently restricted to scalar linear problems. The sparse DG version that we present in this chapter is designed for systems and offers similar computational efficiency. Therefore, the methods developed in [Haj20b, Haj20c] are not included in this thesis.

As mentioned by Guermond et al. [Gue19], the convex limiting procedures presented in [Gue18a] are technically applicable to DG discretizations. The corresponding generalization as outlined in [Gue19] without presenting any numerical results is of FCT type. The monolithic alternative developed by the author in [Haj21a] supports the use of arbitrary-order Bernstein finite elements and constrains DG boundary terms in a different manner, as explained in Remark 3.10.

Pazner [Paz21] developed an algebraic FCT scheme for high order DG discretizations of nonlinear systems. In his method, collocated Gauss–Lobatto quadrature is employed to obtain nodal connectivity corresponding to cross-stencil patterns. To reduce the computational cost and the levels of numerical dissipation, Pazner replaces the discrete gradient operator of the standard collocated DG method with a *sparsified* one. This idea was first used by Lohmann et al. [Loh17b] in the context of continuous, high order Bernstein finite elements for the linear advection equation. It was later adapted to discretizations of general conservation laws by Kuzmin and Quezada de Luna [Kuz20e], whose sparsification approach is also employed in this chapter. The development of FCT-type convex limiting tools for collocated DG discretizations of hyperbolic systems was initiated in [Paz21] and continued in [Rue22]. These recent developments indicate that collocated nodal bases represent a promising alternative to Bernstein polynomials when it comes to designing sparse DG-AFC schemes for high order finite elements.

6.2 Algebraic flux correction schemes

In this section, we extend the MCL methodology [Kuz20a, Kuz20e] to DG discretizations of the hyperbolic PDE (system)

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = 0 \qquad \text{in } \Omega \times \mathbb{R}_+, \tag{6.1}$$

where $u = u(\boldsymbol{x}, t) \in \mathbb{R}^m$, $\mathbf{f}(u) \in \mathbb{R}^{m \times d}$, $m \in \mathbb{N}$, $d \in \{1, 2, 3\}$. For well-posedness, we require initial conditions $u_0 = u_0(\boldsymbol{x}) \in \mathbb{R}^m$ and external Riemann data $\hat{u} = \hat{u}(\boldsymbol{x}, t) \in \mathbb{R}^m$.

In the following sections, we first present the baseline DG scheme for discretizing (6.1) in space. Then we introduce our DG extensions of the low order method and of the monolithic convex limiting strategy. The design principles behind these generalizations are the same as in the case of continuous finite elements. To avoid repetition, we refer the reader to Section 3.3 for an introduction to basic ingredients of AFC schemes.

6.2.1 Target discretization

Let $\mathcal{K}_h = \{K^1, \ldots, K^E\}$ be an affine and geometrically conforming mesh consisting of $E = E(h)$ elements such that $\bigcup_{e=1}^{E} K^e = \overline{\Omega}$. For $p \in \mathbb{N}_0$ and $K \in \mathcal{K}_h$, let $\mathbb{V}_p(K)$ denote the local polynomial space for the DG finite element approximation ($\mathbb{V}_p = \mathbb{P}_p$ if K is a simplex, $\mathbb{V}_p = \mathbb{Q}_p$ for quadrilaterals and hexahedra, cf. Section 3.1). The corresponding scalar- and vector-valued DG spaces of degree p over the mesh $\mathcal{K}_h$ are defined by

$$\mathrm{W}_{h,p} = \left\{ w_h \in \mathrm{L}^2(\Omega) : w_h|_K \in \mathbb{V}_p(K) \ \forall K \in \mathcal{K}_h \right\}, \qquad \mathrm{W}_{h,p}^m = (\mathrm{W}_{h,p})^m, \ m \in \mathbb{N}.$$

Let $N^e = N(p, K^e) = \dim \mathbb{V}_p(K^e)$. If all elements are topologically equivalent, then the number of local degrees of freedom $N^e \equiv N$ is independent of e and, therefore, $\dim \mathrm{W}_{h,p} = NE$. A basis of $\mathrm{W}_{h,p}$ can be chosen such that the support of each basis function is restricted to a single mesh cell. Thus, for every element K^e we select a *local* basis $\{\varphi_1^e, \ldots, \varphi_N^e\}$ of $\mathbb{V}_p(K^e)$ and express the discrete solution as follows

$$u_h(\boldsymbol{x}, t) = \sum_{e=1}^{E} u_h^e(\boldsymbol{x}, t)\, \chi_{K^e}(\boldsymbol{x}), \qquad u_h^e(\boldsymbol{x}, t) = \sum_{i=1}^{N} u_i^e(t)\, \varphi_i^e(\boldsymbol{x}), \qquad u_i^e(t) \in \mathbb{R}^m,$$

where χ_{K^e} denotes the characteristic function of K^e. Multiplying (6.1) by a test function $w_h^e \in \mathbb{V}_p(K^e)^m$, integrating the weighted residual over K^e, and imposing flux boundary conditions on ∂K^e as for $\partial\Omega$ in Section 3.1, we obtain the strong form

$$\int_{K^e} w_h^e \cdot \left[\frac{\partial u_h^e}{\partial t} + \nabla \cdot \mathbf{f}(u_h^e) \right] \mathrm{d}\boldsymbol{x} + \int_{\partial K^e} w_h^e \cdot [\mathrm{f}_{\boldsymbol{n}^e}(u_h^e, \hat{u}_h^e) - \mathbf{f}(u_h^e)\, \boldsymbol{n}^e] \, \mathrm{d}s = 0 \tag{6.2}$$

of the DG discretization. Therein, $\boldsymbol{n}^e$ is the unit outward normal to ∂K^e, $\mathrm{f}_{\boldsymbol{n}^e}(\cdot,\cdot)$ is a numerical approximation to the corresponding normal flux, and

$$\hat{u}_h^e(\boldsymbol{x}, t) := \begin{cases} \hat{u}(\boldsymbol{x}, t) & \text{if } \boldsymbol{x} \in \partial K^e \cap \partial\Omega, \\ \lim_{\varepsilon \searrow 0} u_h(\boldsymbol{x} + \varepsilon \boldsymbol{n}^e, t) & \text{otherwise} \end{cases} \tag{6.3}$$

is the external state of the weakly imposed boundary/interface condition. Strictly speaking, $\hat{u}_h^e(\boldsymbol{x}, t)$ is not well-defined in the vertices of K^e but this issue is unimportant for formulation (6.2) and the discussions that follow. The DG solution u_h^e must satisfy (6.2) for all $w_h^e \in \mathbb{V}_p(K^e)^m$.

The global weak formulation corresponding to (6.2) can be obtained by summation over all elements. Contrary to continuous Galerkin methods, the integrals over interfaces of adjacent mesh cells do not cancel out in the DG version of the variational formulation. Therefore, numerical fluxes need to be evaluated for each interface. In our target scheme (6.2), volume and boundary integrals are approximated with standard quadrature rules employed in DG discretizations (see [Ern04, Sec. 8.1]). To avoid cumbersome notation

involving quadrature, we stick to formulation (6.2) for presentation purposes. For additional information on the mathematical and computational aspects of standard DG discretizations, particularly ones for hyperbolic problems, we refer the reader to the books by Di Pietro and Ern [DiP12] and Dolejší and Feistauer [Dol15].

6.2.2 Low order method

The DG discretization discussed so far is valid for any particular basis of $\mathrm{W}_{h,p}$. For algebraic flux correction purposes, however, we require a nodal basis $\{\varphi_1^e, \ldots, \varphi_N^e\}$ in each element K^e. In this chapter, we employ Bernstein (also called Bézier) polynomials as basis functions, following [Abg10, And17, Loh17b, Kuz20e]. On the 1D reference element $[0, 1]$, the Bernstein basis functions are defined by

$$\mathrm{B}_k^p(x) = \binom{p}{k} (1-x)^{p-k} x^k, \qquad k \in \{0, \ldots, p\}.$$

Bases for quadrilaterals and hexahedra are obtained from products of 1D Bernstein polynomials in each spatial variable. To define the Bernstein basis on the reference simplex $\mathrm{conv}\{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_d\}$ in $\mathbb{R}^d$, where $\mathbf{e}_1, \ldots, \mathbf{e}_d$ are the Cartesian unit vectors, we employ *barycentric coordinates* $\Lambda_0, \ldots, \Lambda_d$ such that

$$\Lambda_0(x_1, \ldots, x_d) = 1 - \sum_{j=1}^{d} x_j, \qquad \Lambda_k(x_1, \ldots, x_d) = x_k, \qquad k \in \{1, \ldots, d\}.$$

The corresponding Bernstein polynomials read [Lai07]

$$\mathrm{B}_{k_0,\ldots,k_d}^p(\boldsymbol{x}) = \frac{p!}{\prod_{j=0}^d k_j!} \prod_{j=0}^{d} (\Lambda_j(\boldsymbol{x}))^{k_j}, \qquad k_0, \ldots, k_d \in \mathbb{N}_0, \qquad \sum_{j=0}^{d} k_j = p.$$

Note that for $p = 1$, Bernstein and Lagrange polynomials are identical. The local basis functions $\{\varphi_1^e, \ldots, \varphi_N^e\}$ are obtained by transforming the Bernstein polynomials $\{\mathrm{B}_i^p\}_{i=1}^N$ defined on a reference element $\hat{K}$ to the physical cell K^e. Alternatively, the local bases can be obtained directly, for instance, by using barycentric coordinates for K^e instead of those for $\hat{K}$ in the case of simplices. A number of useful mathematical properties can be shown for Bernstein polynomial approximations [Lai07, Ch. 2]. We exploit some of them when it comes to algebraic flux correction for the baseline DG method. For instance, the Bernstein basis functions φ_i^e satisfy

$$\varphi_i^e(\boldsymbol{x}) \in [0, 1] \qquad \forall i \in \{1, \ldots, N\},\ \boldsymbol{x} \in K^e, \qquad \text{and} \qquad \sum_{j=1}^{N} \varphi_j^e(\boldsymbol{x}) \equiv 1 \qquad \forall \boldsymbol{x} \in K^e$$

for any $p \in \mathbb{N}_0$. These properties imply [Loh17b]

$$\min_{j\in\{1,\dots,N\}} u_j^e \le \sum_{j=1}^{N} u_j^e\, \varphi_j^e(\boldsymbol{x}) \le \max_{j\in\{1,\dots,N\}} u_j^e. \tag{6.4}$$

That is, any value that a polynomial $u_h^e \in \mathbb{V}_p(K^e)$ may attain in K^e is bounded by the largest and smallest Bernstein coefficient of u_h^e. In the context of algebraic flux correction, (6.4) suggests that u_h^e can be constrained by imposing appropriate bounds on the local degrees of freedom. Another useful property of the Bernstein basis is the fact that each basis function is associated with a unique point $\boldsymbol{x}_i^e \in K^e$. In these *nodes*, which are distributed uniformly within the element K^e, the corresponding basis functions attain their respective maxima [Lai07, Thm. 2.5], see Fig. 6.1 for an illustration.

(a) Bernstein basis for $\mathbb{P}_3([0,1])$ (b) Nodes of a $\mathbb{Q}_3$ element (c) Nodes of a $\mathbb{P}_3$ element

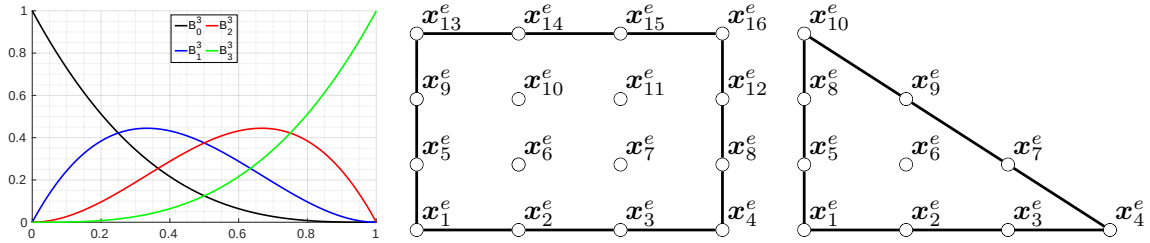


Figure 6.1: Cubic Bernstein basis functions in 1D and lexicographically numbered nodes of (bi-)cubic Bernstein finite elements in 2D.

Having discussed the most important aspects of the employed basis, let us now come back to the discretization of (6.1). The semi-discrete formulation (6.2) is property preserving only if piecewise constant basis functions are employed in combination with suitable numerical fluxes. This case corresponds to a first order finite volume discretization, which is not addressed here. To obtain a property-preserving low order method for second and higher order spaces, we modify the target scheme similarly to our approach in Section 3.3.2. Let $\mathcal{F}(K^e)$ be the set of faces of element K^e (cf. Definition 3.2). For $i,j \in \{1,\dots,N\}$ and $e \in \{1,\dots,E\}$, we define

$$m_{ij}^e = \int_{K^e} \varphi_i^e\, \varphi_j^e \,\mathrm{d}\boldsymbol{x}, \qquad m_i^e = \int_{K^e} \varphi_i^e \,\mathrm{d}\boldsymbol{x}, \qquad \mathbf{c}_{ij}^e = \int_{K^e} \varphi_i^e\, \nabla\varphi_j^e \,\mathrm{d}\boldsymbol{x},$$
$$b_{i,k}^e = \int_{\Gamma_k^e} \varphi_i^e \,\mathrm{d}s, \qquad \Gamma_k^e \in \mathcal{F}(K^e), \qquad \mathbf{f}_i^e = \mathbf{f}(u_i^e).$$

Note that, as a consequence of employing the Bernstein basis, the coefficient $b_{i,k}^e$ is nonnegative and equals zero unless the node $\boldsymbol{x}_i^e$ lies on the boundary face Γ_k^e.

Employing the group finite element formulation, row sum mass lumping, and a lumped approximation to the DG flux terms in (6.2), we obtain

$$m_i^e \frac{\mathrm{d}u_i^e}{\mathrm{d}t} = -\sum_{\substack{j=1\\ j\neq i}}^{N} (\mathbf{f}_j^e - \mathbf{f}_i^e)\, \mathbf{c}_{ij}^e + \sum_{\Gamma_k^e \in \mathcal{F}(K^e)} b_{i,k}^e \Big[\mathbf{f}_i^e\, \boldsymbol{n}_k^e - \mathrm{f}_{\boldsymbol{n}_k^e}(u_i^e, \hat{u}_{i,k}^e) \Big] \tag{6.5}$$

for $e \in \{1, \ldots, E\}$, $i \in \{1, \ldots, N\}$. Here $\boldsymbol{n}_k^e$ is the outward unit normal to $\Gamma_k^e \subset \partial K^e$ and $\hat{u}_{i,k}^e$ is the unique nodal value of $\hat{u}_h$ determined by (6.3) corresponding to node $\boldsymbol{x}_i^e$ and Γ_k^e. We remark that (6.5) approximates (6.2) using nodal (collocated) quadrature.

At this stage, we could proceed as in Section 3.3.2 by incorporating Rusanov (local Lax–Friedrichs) dissipation into (6.5). Instead, we first introduce the *sparsified* discrete gradient operator $\tilde{\mathbf{C}}^e = (\tilde{\mathbf{c}}_{ij}^e)_{i,j=1}^N = (\tilde{\mathrm{C}}_1, \ldots, \tilde{\mathrm{C}}_d)^\mathsf{T}$, defined by [Loh17b, Kuz20e]

$$\tilde{\mathrm{C}}_k^e = M_{\mathrm{L}}^e\, (M^e)^{-1} \mathrm{C}_k^e, \qquad M_{\mathrm{L}}^e = \mathrm{diag}(m_1^e, \ldots, m_N^e), \qquad M^e = (m_{ij}^e)_{i,j=1}^N.$$

By construction, the element matrices $\tilde{\mathrm{C}}_k^e$ have zero row sums. Moreover, the columns of $\tilde{\mathrm{C}}_k^e - \mathrm{C}_k^e$, where $\mathbf{C}^e = (\mathbf{c}_{ij}^e)_{i,j=1}^N = (\mathrm{C}_1^e, \ldots, \mathrm{C}_d^e)^\mathsf{T}$, sum to zero [Kuz20e, Sec. 4]. Thus, replacing $(\mathbf{f}_j^e - \mathbf{f}_i^e)\, \mathbf{c}_{ij}^e$ in (6.5) with $(\mathbf{f}_j^e - \mathbf{f}_i^e)\, \tilde{\mathbf{c}}_{ij}^e$ does not lead to violations of local conservation properties. The benefit of introducing $\tilde{\mathbf{C}}^e$ lies in its sparsity pattern. One can show that the entries of $\tilde{\mathbf{C}}^e$ are nonzero if and only if the corresponding nodes are nearest neighbors within the element. This remarkable property of Bernstein elements was first observed by Lohmann et al. [Loh17b, Appendix B] in 1D and later generalized to the multidimensional case by Kuzmin and Quezada de Luna [Kuz20e, Appendix]. In fact, the sparse element matrices $\tilde{\mathbf{C}}^e$ may have even more zero entries than this analysis suggests. Indeed, for non-simplicial affine meshes, the entries of $\tilde{\mathbf{C}}^e$ can be shown to be zero for nodes, which are closest neighbors in diagonal directions [Haj21a]. This observation leads to more efficient implementations because of the reduced number of nodes in the stencils. Moreover, the scheme now bears a remarkable resemblance to the collocated cross-stencil DG-FCT scheme proposed by Pazner [Paz21].

Remark 6.1

The condition number of the Bernstein element mass matrices M^e grows exponentially with p [Lyc00]. Thus, their inversion should be avoided when it comes to computing the sparsified operators $\tilde{\mathbf{C}}^e$ for high order spaces. An explicit formula for the entries of $\tilde{\mathbf{C}}^e = \tilde{\mathrm{C}}_1^e$ in 1D can be found in [Loh17b, Eq. B.4]. For multidimensional non-simplicial elements, we use a tensor product version of this formula. In the case of simplices, the degree elevation rule for Bernstein polynomials [Lai07, Sec. 2.15] enables us to compute the entries of $\tilde{\mathbf{C}}^e$ directly as in [Kuz20e] and [Haj21a]. In any case, we generate sparse matrices on the reference element and perform standard transformations to obtain $\tilde{\mathbf{C}}^e$. A Matlab code for computing sparsified gradient operators of arbitrary order on the reference triangle can be found in the author's GitHub repository [Haj20a]. Let us further remark that for $\mathbb{P}_1$ elements in 1D, an application of the sparsifier $M_L^e (M^e)^{-1}$ to $\mathbf{C}^e = \mathrm{C}_1^e$ produces $\tilde{\mathrm{C}}_1^e = \mathrm{C}_1^e$, as one can easily verify. ⋄

Following [Kuz20e], we now introduce *sparsified Rusanov dissipation* coefficients

$$\tilde{d}_{ij}^e := \begin{cases} \max\{|\tilde{\mathbf{c}}_{ij}^e|\lambda_{ij}^e, |\tilde{\mathbf{c}}_{ji}^e|\lambda_{ji}^e\} & \text{if } j \in \tilde{\mathcal{N}}_i \setminus \{i\}, \\ 0 & \text{otherwise,} \end{cases} \qquad \lambda_{ij}^e := \lambda_{\tilde{\mathbf{c}}_{ij}^e/|\tilde{\mathbf{c}}_{ij}^e|}(u_i^e, u_j^e), \tag{6.6}$$

where $\lambda_{\boldsymbol{n}}(\cdot,\cdot)$ is the maximum speed defined by (3.2b) and $\tilde{\mathcal{N}}_i \subseteq \{1,\dots,N\}$ is the set of indices $j \in \{1,\dots,N\}$ such that $j = i$ or $\min\{|\tilde{\mathbf{c}}_{ij}^e|, |\tilde{\mathbf{c}}_{ji}^e|\} > 0$. Since we restrict ourselves to cases in which all elements have the same topology and polynomial bases of the same degree $p \in \mathbb{N}$ are used in every cell, the index sets $\tilde{\mathcal{N}}_i$ are independent of the element number e. Adding algebraic Rusanov fluxes to (6.5) and employing the local Lax–Friedrichs flux in boundary terms, we obtain the low order method

$$m_i^e \frac{\mathrm{d}u_i^e}{\mathrm{d}t} = \sum_{j \in \tilde{\mathcal{N}}_i \setminus \{i\}} \Big[\tilde{d}_{ij}^e (u_j^e - u_i^e) - (\mathbf{f}_j^e - \mathbf{f}_i^e)\, \tilde{\mathbf{c}}_{ij}^e \Big] \tag{6.7a}$$

$$+ \sum_{\Gamma_k^e \in \mathcal{F}(K^e)} b_{i,k}^e \Big[\lambda_{i,k}^e \frac{\hat{u}_{i,k}^e - u_i^e}{2} - \frac{\hat{\mathbf{f}}_{i,k}^e - \mathbf{f}_i^e}{2}\, \boldsymbol{n}_k^e \Big], \tag{6.7b}$$

where $\lambda_{i,k}^e := \lambda_{\boldsymbol{n}_k^e}(u_i^e, \hat{u}_{i,k}^e)$ and $\hat{\mathbf{f}}_{i,k}^e := \mathbf{f}(\hat{u}_{i,k}^e)$. Note that the sum on the right hand side of (6.7a) is over the compact stencil $\tilde{\mathcal{N}}_i \setminus \{i\}$. Here we observe the main advantage of applying the sparsifier $M_{\mathrm{L}}^e\,(M^e)^{-1}$ to the elementwise discrete gradient operator $\mathbf{C}^e$. Indeed, by (6.6), the Rusanov flux $\tilde{d}_{ij}^e(u_j^e - u_i^e)$ vanishes if nodes i and j are not nearest neighbors. Without sparsification, fluxes $d_{ij}^e(u_j^e - u_i^e)$ associated with $j \notin \tilde{\mathcal{N}}_i$ tend to have a devastating effect on the approximation quality [Haj20b]. Interestingly enough, the quadrature error introduced via lumping for boundary terms does not seem to increase significantly in higher order methods. This observation is consistent to the results we obtained with our scalar prototype DG-AFC schemes [Haj20b, Haj20c].

To conclude this section, let us present the equivalent bar state form

$$m_i^e \frac{\mathrm{d}u_i^e}{\mathrm{d}t} = \sum_{j \in \tilde{\mathcal{N}}_i \setminus \{i\}} 2\tilde{d}_{ij}^e (\bar{u}_{ij}^e - u_i^e) + \sum_{\Gamma_k^e \in \mathcal{F}(K^e)} 2 d_{i,k}^e (\bar{u}_{i,k}^e - u_i^e) \tag{6.8}$$

of (6.7), where $d_{i,k}^e = \frac{b_{i,k}^e}{2}\lambda_{i,k}^e$, and [Har83b, Gue16b, Kuz20a, Haj21a]

$$\bar{u}_{ij}^e = \frac{u_i^e + u_j^e}{2} - \frac{(\mathbf{f}_j^e - \mathbf{f}_i^e)\, \tilde{\mathbf{c}}_{ij}^e}{2\tilde{d}_{ij}^e}, \qquad \bar{u}_{i,k}^e = \frac{u_i^e + \hat{u}_{i,k}^e}{2} - \frac{(\hat{\mathbf{f}}_{i,k}^e - \mathbf{f}_i^e)\, \boldsymbol{n}_k^e}{2\lambda_{i,k}^e}. \tag{6.9}$$

As in [Har83b, Gue16b] and elsewhere in this thesis, the diffusion coefficients $\tilde{d}_{ij}^e$ and $d_{i,k}^e$ are defined in a way that allows an interpretation of the bar states (6.9) as averages of exact solutions to Riemann problems (cf. Lemma 3.12). Thus, the states $\bar{u}_{ij}^e$ and $\bar{u}_{i,k}^e$ stay in every convex invariant set of (6.1) that contains u_i^e, u_j^e, and $\hat{u}_{i,k}^e$. A fully discrete

counterpart of the low order method (6.8) employing an SSP time stepping scheme is therefore invariant domain preserving (IDP) under a CFL-like time step restriction [Gue19].

A few further remarks regarding the low order DG discretization (6.8) are in order.

Remark 6.2
Contrary to the bar states arising in continuous Galerkin discretizations, we have $\bar{u}_{ij}^e \neq \bar{u}_{ji}^e$ in general. This is a consequence of employing the DG baseline discretization in which the elementwise discrete gradient operator $\mathbf{C}^e$ is skew symmetric only for pairs of nodes that do not both lie on the element boundary. Furthermore, sparsification produces operators $\tilde{\mathbf{C}}^e$ that may not be skew symmetric even for such pairs of nodes. On the other hand, an *interfacial bar state* $\bar{u}_{i,k}^e$ always equals the bar state of the node with the same physical location in an adjacent element. This property follows from the use of identical wave speeds $\lambda_{i,k}^e$ in both bar states and the fact that the unit outward normal to the boundary of the neighbor element is $-\boldsymbol{n}_k^e$. ◊

Remark 6.3
One can write (6.8) in a unified notation for volumetric and interfacial bar states by disguising the different nature of the underlying approximations, as we did in Section 3.3.6 and [Haj21a, Rem. 5] (see also [Gue19, Sec. 4]). Since the volume and flux terms are treated differently in practice, we avoid using compressed single-sum representations. ◊

6.2.3 Monolithic convex limiting

In this section, we introduce an MCL strategy that recovers the accuracy of the standard DG discretization without sacrificing the bound-preserving properties of the low order method. As usual, we first define the raw antidiffusive fluxes. Then we present the limiter for the bar states and choose the local bounds. Many features to be discussed were originally proposed by Kuzmin and Quezada de Luna [Kuz20e] in the context of high order continuous Galerkin schemes for scalar problems. The novelty of what we present in this section is the extension to DG discretizations and hyperbolic systems.

6.2.3.1 Definition of raw antidiffusive fluxes

In AFC schemes for continuous Galerkin discretizations, raw antidiffusive fluxes associated with volume integrals should compensate the effects of Rusanov dissipation and mass lumping errors (for time-dependent problems). In high order DG extensions that we propose in this chapter, it is also essential to compensate errors due to the use of the group finite element approximation, quadrature-based lumping for boundary terms, and sparsification of the discrete gradient operator. Moreover, the decomposition of

correction terms into fluxes should preserve the compact sparsity pattern of the low order method.

To derive an array of fluxes that comply with the above design principles and are well suited for AFC, we first define the *nodal contributions* [Kuz20e]

$$f_i^e = \sum_{j=1}^{N} m_{ij}^e(\dot{u}_i^e - \dot{u}_j^e) - \sum_{j\in\tilde{\mathcal{N}}_i\setminus\{i\}} \tilde{d}_{ij}^e(u_j^e - u_i^e) \tag{6.10a}$$

$$+ \sum_{j\in\tilde{\mathcal{N}}_i\setminus\{i\}} (\mathbf{f}_j^e - \mathbf{f}_i^e)\,\tilde{\mathbf{c}}_{ij}^e + \int_{K^e} \mathbf{f}(u_h^e)\,\nabla\varphi_i^e\,\mathrm{d}\boldsymbol{x} - \int_{\partial K^e} \varphi_i^e\,\mathbf{f}_i^e\,\boldsymbol{n}^e\,\mathrm{d}s, \tag{6.10b}$$

$$f_{i,k}^e = \int_{\Gamma_k^e} \varphi_i^e \left[\lambda_{i,k}^e \frac{u_i^e - \hat{u}_{i,k}^e}{2} + \frac{\mathbf{f}_i^e + \hat{\mathbf{f}}_{i,k}^e}{2}\,\boldsymbol{n}_k^e - \mathrm{f}_{\boldsymbol{n}_k^e}(u_h^e, \hat{u}_h^e)\right]\mathrm{d}s, \tag{6.10c}$$

where $\mathrm{f}_{\boldsymbol{n}_k^e}(\cdot,\cdot)$ is a user-defined numerical flux that possesses the properties required by Definition 3.3. The external state $\hat{u}_h^e$ is defined by (6.3). For time-dependent problems, the vector of time derivatives $\dot{u}^e = (\dot{u}_1^e, \ldots, \dot{u}_N^e)^\mathsf{T}$ in element K^e is given by the solution of the linear system corresponding to the semi-discrete local problem

$$\int_{K^e} w_h^e \cdot [\dot{u}_h^e + \nabla\cdot\mathbf{f}(u_h^e)]\,\mathrm{d}\boldsymbol{x} + \int_{\partial K^e} w_h^e \cdot [\mathrm{f}_{\boldsymbol{n}^e}(u_h^e, \hat{u}_h^e) - \mathbf{f}(u_h^e)\,\boldsymbol{n}^e]\;\mathrm{d}s = 0$$

for all $w_h^e \in \mathbb{V}_p(K^e)^m$, $e \in \{1,\ldots,E\}$. Contrary to our approach in Section 3.3.3, this definition of $\dot{u}_h^e$ corresponds to the high order DG target scheme because no additional stabilization needs to be introduced in this context.

Remark 6.4

Entries of the vector $\dot{u}^e$ can be calculated either by directly inverting the ill-conditioned Bernstein mass matrix or by first computing a representation of $\dot{u}_h^e$ w. r. t. another basis and then obtaining the Bernstein coefficients of $\dot{u}_h^e$. In either case, the linear systems that need to be solved are quite ill-conditioned for high order spaces [Lyc00], even though these are just element matrices. Algorithms specifically designed to invert Bernstein mass matrices of simplicial DG methods can be found in [Kir17, Ain19]. The condition number also grows with the spatial dimension. Thus, in the case of nonsimplicial elements, we calculate solutions of linear systems, by inverting 1D mass matrices and exploiting the tensor-product structure of elements. ♢

Let us now prove a few auxiliary results regarding the nodal contributions (6.10). The following lemma ensures that the DG target scheme (6.2) can indeed be recovered from the sparse low order approximation (6.7) using f_i^e and $f_{i,k}^e$.

Lemma 6.5 (Haj21a Lem. 1)

Let f_i^e and $f_{i,k}^e$ be defined by (6.10). Then the semi-discrete problem

$$m_i^e \frac{\mathrm{d}u_i^e}{\mathrm{d}t} = \sum_{j\in\tilde{\mathcal{N}}_i\setminus\{i\}} \left[\tilde{d}_{ij}^e(u_j^e - u_i^e) - (\mathbf{f}_j^e - \mathbf{f}_i^e)\,\tilde{\mathbf{c}}_{ij}^e\right] + f_i^e$$

$$+ \sum_{\Gamma_k^e \in \mathcal{F}(K^e)} \Big[b_{i,k}^e \Big(\lambda_{i,k}^e \frac{\hat{u}_{i,k}^e - u_i^e}{2} - \frac{\hat{\mathbf{f}}_{i,k}^e - \mathbf{f}_i^e}{2} \boldsymbol{n}_k^e \Big) + f_{i,k}^e \Big] \tag{6.11}$$

for $i \in \{1, \ldots, N\}$, $e \in \{1, \ldots, E\}$ is equivalent to the DG target scheme (6.2). ◊

Proof:
The sums in (6.10a) compensate the mass lumping error and the sparsified Rusanov dissipation of the low order method (6.7). The boundary integrals (6.10c) are designed to recover the non-lumped version of the interfacial numerical flux in (6.2). Using integration by parts, we find that (6.11) reduces to

$$\begin{aligned} \sum_{j=1}^N m_{ij}^e \frac{\mathrm{d}u_j^e}{\mathrm{d}t} = & - \int_{K^e} \varphi_i^e \, \nabla \cdot \mathbf{f}(u_h^e) \, \mathrm{d}\boldsymbol{x} + \int_{\partial K^e} \varphi_i^e \, \mathbf{f}(u_h^e) \, \boldsymbol{n}^e \, \mathrm{d}s \\ & - \int_{\partial K^e} \varphi_i^e \, \mathbf{f}_i^e \, \boldsymbol{n}^e \, \mathrm{d}s + \sum_{\Gamma_k^e \in \mathcal{F}(K^e)} \int_{\Gamma_k^e} \varphi_i^e \, [\mathbf{f}_i^e \, \boldsymbol{n}_k^e - \mathrm{f}_{\boldsymbol{n}_k^e}(u_h^e, \hat{u}_h^e)] \, \mathrm{d}s, \end{aligned}$$

which is equivalent to the semi-discrete DG formulation (6.2). □

Lemma 6.6 (Kuz20e Sec. 4, Haj21a Lem. 2)
The nodal contributions f_i^e defined by (6.10a)–(6.10b) satisfy

$$\sum_{i=0}^N f_i^e = 0 \qquad \forall e \in \{1, \ldots, E\}. \qquad \diamond$$

Proof:
The terms in (6.10a) add up to zero due to the symmetry of mass matrices and Rusanov dissipation operators. For the remainder (6.10b), we use the zero row sum properties of $\tilde{\mathbf{C}}^e$ and $\mathbf{C}^e$, the zero column sum property of $\tilde{\mathbf{C}}^e - \mathbf{C}^e$, integration by parts, and the fact that the local basis functions form a partition of unity on K^e. In this fashion, we obtain [Kuz20e]

$$\begin{aligned} \sum_{i=1}^N f_i^e &= \sum_{i=1}^N \sum_{j \in \tilde{\mathcal{N}}_i} \mathbf{f}_j^e \, \tilde{\mathbf{c}}_{ij}^e + \sum_{i=1}^N \int_{K^e} \mathbf{f}(u_h^e) \, \nabla \varphi_i^e \, \mathrm{d}\boldsymbol{x} - \sum_{j=1}^N \int_{\partial K^e} \varphi_j^e \, \mathbf{f}_j^e \, \boldsymbol{n}^e \, \mathrm{d}s \\ &= \sum_{i,j=1}^N \mathbf{f}_j^e \, (\tilde{\mathbf{c}}_{ij}^e - \mathbf{c}_{ij}^e + \mathbf{c}_{ij}^e) - \sum_{j=1}^N \int_{\partial K^e} \varphi_j^e \, \mathbf{f}_j^e \, \boldsymbol{n}^e \, \mathrm{d}s \\ &= \sum_{j=1}^N \mathbf{f}_j^e \Big[\sum_{i=1}^N (\tilde{\mathbf{c}}_{ij}^e - \mathbf{c}_{ij}^e) - \sum_{i=1}^N \mathbf{c}_{ji}^e + \int_{\partial K^e} \varphi_j^e \, \boldsymbol{n}^e \, \mathrm{d}s \Big] - \sum_{j=1}^N \int_{\partial K^e} \varphi_j^e \, \mathbf{f}_j^e \, \boldsymbol{n}^e \, \mathrm{d}s = 0. \quad \square \end{aligned}$$

Lemma 6.7 (Haj21a Lem. 3)
Let $K^{e'}$ be the cell that shares the boundary face $\Gamma_k^e = \Gamma_{k'}^{e'}$ with K^e and let $\boldsymbol{x}_{i'}^{e'}$ be the node in $K^{e'}$ with the same physical location as node $\boldsymbol{x}_i^e \in K^e$. Then the corresponding antidiffusive DG fluxes (6.10c) satisfy $f_{i,k}^e = -f_{i',k'}^{e'}$. ◊

Proof:
The basis functions φ_i^e and $\varphi_{i'}^{e'}$ coincide on Γ_k^e. Moreover, the bracketed term in (6.10c) consists of numerical fluxes, which must have opposite signs by Definition 3.3. □

For algebraic flux correction purposes, we require raw antidiffusive *fluxes* f_{ij}^e instead of *nodal* contributions (6.10a)–(6.10b). Thus, we distribute f_i^e between pairs of nodes within the element. Importantly, we employ the nodal stencil of the low order method. In other words, we seek $f_{ij}^e \in \mathbb{R}^m$ satisfying

$$\sum_{j \in \tilde{\mathcal{N}}_i \setminus \{i\}} f_{ij}^e = f_i^e, \qquad i \in \{1, \dots, N\}, \qquad f_{ij}^e = -f_{ji}^e, \qquad i, j \in \{1, \dots, N\}, \tag{6.12}$$

and $f_{ij}^e = 0$ if $j \notin \tilde{\mathcal{N}}_i \setminus \{i\}$. The existence of such decompositions is a consequence of Lemma 6.6. In general, however, many alternative splittings of f_i^e may provide the above properties, and the specific choice can influence the numerical results. In this work, we pursue the same approach as in [Kuz20e] to obtain the f_{ij}^e. First, the sparsified Rusanov dissipation is split from f_i^e because it already exhibits the desired compact-stencil structure. Thus, we define the nodal contributions

$$g_i^e := f_i^e + \sum_{j \in \tilde{\mathcal{N}}_i \setminus \{i\}} \tilde{d}_{ij}^e (u_j^e - u_i^e).$$

Note that the statement of Lemma 6.6 remains valid if f_i^e is replaced by g_i^e therein.

We obtain the raw antidiffusive fluxes by solving linear systems $Av^e = g^e$ for each element, where $g^e = (g_i^e)_{i=1}^N$ and $A = (a_{ij})_{i,j=1}^N$ is a sparse graph Laplacian to be defined below. Given the solution vector $v^e = (v_i^e)_{i=1}^N$ of this linear system, we set

$$f_{ij}^e := a_{ij}(v_j^e - v_i^e) - \tilde{d}_{ij}^e (u_j^e - u_i^e), \qquad i \in \{1, \dots, N\},\ j \in \tilde{\mathcal{N}}_i \setminus \{i\}.$$

In view of our design principles for f_{ij}^e, we must have $a_{ij} = 0$ for $j \notin \tilde{\mathcal{N}}_i$. Moreover, the matrix A must be symmetric and have zero row sums. In our code, we construct A using the sparse consistent mass matrix $\tilde{M} = (\tilde{m}_{ij})_{i,j=1}^N$ of a piecewise (multi)-linear continuous approximation on the reference element. The nodes of this subcell discretization are the same as the Bernstein nodes of the high order space (see Fig. 6.1). For non-simplicial elements, we set $\tilde{m}_{ii} := \tilde{m}_{ii} + \tilde{m}_{ij}$, $\tilde{m}_{ij} := 0$ if node j is one of the closest diagonal neighbors of node i. The so-defined partially lumped matrix $\tilde{M}$ has the same cross-stencil sparsity pattern as $\tilde{D}^e = (\tilde{d}_{ij}^e)_{i,j=1}^N$. The entries of A are defined by

$$a_{ij} = -\tilde{m}_{ij}, \qquad i, j \in \{1, \dots, N\},\ i \neq j, \qquad a_{ii} = \sum_{j \in \tilde{\mathcal{N}}_i \setminus \{i\}}^{N} \tilde{m}_{ij}, \qquad i \in \{1, \dots, N\}.$$

Note that A is a discrete Laplacian operator in the sense that it is symmetric positive semi-definite with negative off-diagonal entries and zero row sums. Therefore, the solution of the linear system $Av^e = g^e$ is unique up to a constant, which does not influence the values of f_{ij}^e. To fix the constant and avoid singularity, we overwrite the first equation of the linear system by the zero sum condition $\sum_{j=1}^N v_j^e = 0$. It is easy to verify that the presented approach produces raw antidiffusive fluxes f_{ij}^e satisfying the above design criteria [Kuz20e].

6.2.3.2 Outline of the bound-preserving limiting strategy

Let us now adapt the monolithic convex limiting strategy designed for continuous finite elements [Kuz20a, Kuz20e] to the context of high order DG methods. In fact, our limiting techniques for scalar problems and systems are relatively easy to extend to the DG setting conceptually because the AFC decomposition of the target scheme exhibits similar structure. For instance, (6.7) resembles the low order method (3.26) (cf. also (3.29)) for a continuous Galerkin approximation using (multi-)linear Lagrange elements. The representation of the high order scheme in sparse form (6.11), where the nodal contributions f_i^e admit decomposition (6.12) into subcell fluxes f_{ij}^e is again similar to algebraic splittings that we used in previous chapters. Therefore, flux limiting for f_{ij}^e can be performed using limiters developed for second order continuous finite elements. A peculiarity of the DG version is the presence of interfacial fluxes $f_{i,k}^e = -f_{i',k'}^{e'}$. Recall that these additional terms recover the non-lumped version of boundary integrals in which the normal flux is defined by an arbitrary Riemann solver. Thus, a limiting strategy for $f_{i,k}^e$ needs to be devised.

Our extension of MCL to high order DG methods using the above decomposition of the target scheme into a sparse low order method and antidiffusive fluxes between nearest neighbors works as follows. First, we invoke (6.12). Next, similarly to our approach for continuous Lagrange finite elements in Section 3.3.4, we write the compact-stencil DG scheme (6.11) in the bar state form

$$m_i^e \frac{\mathrm{d}u_i^e}{\mathrm{d}t} = \sum_{j \in \tilde{\mathcal{N}}_i \backslash \{i\}} 2\tilde{d}_{ij}^e (\bar{u}_{ij}^{e,*} - u_i^e) + \sum_{\Gamma_k^e \in \mathcal{F}(K^e)} 2d_{i,k}^e (\bar{u}_{i,k}^{e,*} - u_i^e) \tag{6.13}$$

with flux-corrected bar states

$$\bar{u}_{ij}^{e,*} = \bar{u}_{ij}^e + \frac{\alpha_{ij}^e f_{ij}^{e,*}}{2\tilde{d}_{ij}^e}, \qquad \bar{u}_{i,k}^{e,*} = \bar{u}_{i,k}^e + \frac{\alpha_{i,k}^e f_{i,k}^{e,*}}{2d_{i,k}^e}.$$

The limited counterparts $f_{ij}^{e,*}$ and $f_{i,k}^{e,*}$ of f_{ij}^e and $f_{i,k}^e$, respectively, are constrained to preserve certain local bounds $u_i^{e,\min}$, $u_i^{e,\max}$ for numerical admissibility. We discuss how to choose these bounds in Section 6.2.3.3 below. Importantly, the limited fluxes must satisfy $f_{ij}^{e,*} = -f_{ji}^{e,*}$ and $f_{i,k}^{e,*} = -f_{i',k'}^{e',*}$ to ensure conservation. The indices of $f_{i',k'}^{e',*}$ are

defined as in Lemma 6.7. If necessary, IDP correction factors $\alpha_{ij}^e = \alpha_{ji}^e \in [0,1]$ and $\alpha_{i,k}^e = \alpha_{i',k'}^{e'} \in [0,1]$ can be applied to $f_{ij}^{e,*}$ and $f_{i,k}^{e,*}$, respectively, to enforce physical admissibility conditions in flux-limited DG methods for systems.

For brevity, we summarize our limiting approach by referring to the corresponding formulas in Sections 3.3.4 and 3.3.5. A self-contained presentation in which only the DG case is addressed, can be found in [Haj21a]. The original references on sequential FCT-type limiting for (multi)-linear elements [Dob18] and MCL schemes for continuous Galerkin approximations [Kuz20a, Kuz20e] may also be consulted.

For scalar problems, we limit f_{ij}^e and $f_{i,k}^e$ using formulas similar to (3.46)–(3.47). In the limiter for $f_{i,k}^e$, we can make use of the fact that $\bar{u}_{i,k}^e = \bar{u}_{i',k'}^{e'}$, see Remark 6.2. For the shallow water equations and the Euler system, we employ the sequential limiting strategy discussed in Section 3.3.4.3. Specifically, we limit antidiffusive fluxes of derived unknowns using formulas similar to (3.55) for f_{ij}^e and $f_{i,k}^e$. In the limiter for the Euler equations, we additionally enforce positivity of the internal energy via the fix (3.64), which yields the aforementioned IDP correction factors α_{ij}^e and $\alpha_{i,k}^e$.

6.2.3.3 Definition of local admissible bounds

Compared to our earlier considerations in Section 3.3.4, the question of how to choose local bounds for limiting is much more intricate if high order spaces and discontinuous Galerkin methods are employed. For the linear advection equation, we used elementwise bounds in a similar context [Haj20b, Haj20c]. For high order discretizations of nonlinear conservation laws, however, we found that element-stencil bounds are too wide because they allow spurious oscillations within the admissible range of element-global constraints. Therefore, we use localized subcell-stencil bounds, which are based on values of the solution in the nearest neighbors of each node [Loh17b]. In accordance with our earlier work [Haj20b, Haj20c], we define numerical admissibility conditions using identical local bounds for all DG nodes that have the same physical location.

For scalar equations, our approach works as follows. We first compute the minima and maxima of the Bernstein coefficients over the compact stencils $\tilde{\mathcal{N}}_i$ within the element K^e. For interior nodes $\boldsymbol{x}_i^e \in \mathrm{int}(K^e)$ of the element, these values represent the local bounds used for limiting the Bernstein coefficients u_i^e. For $\boldsymbol{x}_i^e \in \partial K^e$, we extend the nodal bounds using the corresponding minima and maxima of nodes with the physical location $\boldsymbol{x}_i^e$ in *all* adjacent elements. Let us briefly clarify our strategy with an illustration.

Example 6.8 (Local bounds for limiting in the scalar 1D case)
Consider the patch of one-dimensional $\mathbb{P}_1$ elements displayed in Fig. 6.2a. The local bounds for the nodes with global indices I and J are given by

$$\begin{aligned} u_I^{\min} &= \min\{u_1^{e-1}, u_2^{e-1}, u_1^e, u_2^e\}, \qquad & u_I^{\max} &= \max\{u_1^{e-1}, u_2^{e-1}, u_1^e, u_2^e\}, \\ u_J^{\min} &= \min\{u_1^e, u_2^e, u_1^{e+1}, u_2^{e+1}\}, \qquad & u_J^{\max} &= \max\{u_1^e, u_2^e, u_1^{e+1}, u_2^{e+1}\}. \end{aligned}$$

The same bounds $u_I^{\min}$, $u_I^{\max}$ are used to limit the Bernstein coefficients u_2^{e-1} and u_1^e in our approach. Similarly, we use $u_J^{\min}$, $u_J^{\max}$ for u_2^e and u_1^{e+1} alike.

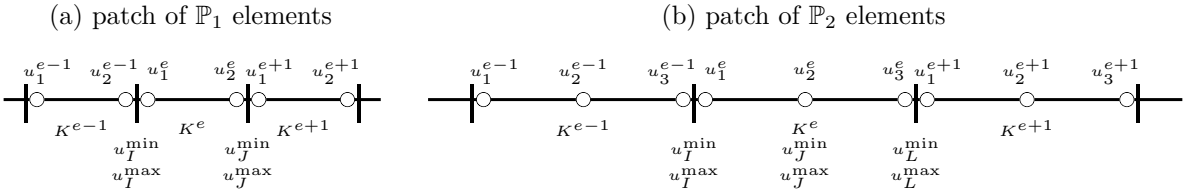


Figure 6.2: A possible definition of local bounds for limiting in the scalar 1D case.

For $\mathbb{P}_2$ elements displayed in Fig. 6.2b, our definition of local bounds yields

$$\begin{aligned} u_I^{\min} &= \min\{u_2^{e-1}, u_3^{e-1}, u_1^e, u_2^e\}, & u_I^{\max} &= \max\{u_2^{e-1}, u_3^{e-1}, u_1^e, u_2^e\}, \\ u_J^{\min} &= \min\{u_1^e, u_2^e, u_3^e\}, & u_J^{\max} &= \max\{u_1^e, u_2^e, u_3^e\}, \\ u_L^{\min} &= \min\{u_2^e, u_3^e, u_1^{e+1}, u_2^{e+1}\}, & u_L^{\max} &= \max\{u_2^e, u_3^e, u_1^{e+1}, u_2^{e+1}\}. \end{aligned} \qquad \diamond$$

Since we adopt the sequential limiting approach for hyperbolic systems, local bounds on the main unknown ϱ and derived quantities ϕ are required. In accordance with our experience regarding continuous Galerkin discretizations, the admissible bounds for systems are based on the low order bar states. For the main unknown ϱ, we employ

$$\varrho_i^{e,\min} := \min\left\{ \min_{j\in\tilde{\mathcal{N}}_i\setminus\{i\}} \bar{\varrho}_{ij}^e, \min_{\Gamma_k^e\in\mathcal{F}_i^e} \bar{\varrho}_{i,k}^e \right\}, \qquad \varrho_i^{e,\max} := \max\left\{ \max_{j\in\tilde{\mathcal{N}}_i\setminus\{i\}} \bar{\varrho}_{ij}^e, \max_{\Gamma_k^e\in\mathcal{F}_i^e} \bar{\varrho}_{i,k}^e \right\},$$

where $\mathcal{F}_i^e$ is the set of all element faces Γ_k^e such that $\boldsymbol{x}_i^e \in \overline{\Gamma_k^e}$, and $\bar{\varrho}_{ij}^e$, $\bar{\varrho}_{i,k}^e$ are the components of $\bar{u}_{ij}^e$ and $\bar{u}_{i,k}^e$ corresponding to ϱ. Similarly, for derived unknowns, we set the local bounds to

$$\begin{aligned} \phi_i^{e,\min} &:= \min\left\{ \min_{j\in\tilde{\mathcal{N}}_i\setminus\{i\}} \bar{\phi}_{ij}^e, \min_{\Gamma_k^e\in\mathcal{F}_i^e} \bar{\phi}_{i,k}^e, \frac{(\varrho\phi)_i}{\varrho_i} \right\}, \\ \phi_i^{e,\max} &:= \max\left\{ \max_{j\in\tilde{\mathcal{N}}_i\setminus\{i\}} \bar{\phi}_{ij}^e, \max_{\Gamma_k^e\in\mathcal{F}_i^e} \bar{\phi}_{i,k}^e, \frac{(\varrho\phi)_i}{\varrho_i} \right\}, \end{aligned}$$

where we choose to additionally incorporate the nodal state $(\varrho\phi)_i/\varrho_i$ into the numerically admissible sets. Again, $\bar{\phi}_{ij}^e$, $\bar{\phi}_{i,k}^e$ are the low order bar states for ϕ corresponding to $\bar{u}_{ij}^e$ and $\bar{u}_{i,k}^e$, respectively. As in the scalar case, we construct bounds for nodes on cell boundaries using the local minima and maxima from all elements that meet at the nodal point. In our implementation, we accomplish this task by employing the data structures of a continuous Galerkin method that uses the same nodal basis on individual cells.

In some applications, preservation of global bounds may suffice. Nonnegativity of a scalar conserved unknown u can be preserved using the one-sided MCL limiter

$$f_{ij}^{e,*} = \max\left\{-2\tilde{d}_{ij}^e\,\bar{u}_{ij}^e, \min\left\{f_{ij}^e, 2\tilde{d}_{ij}^e\,\bar{u}_{ji}^e\right\}\right\}, \tag{6.14}$$

which guarantees the validity of the limiting constraints

$$0 \leq \bar{u}_{ij}^e + \frac{f_{ij}^{e,*}}{2\tilde{d}_{ij}^e}, \qquad 0 \leq \bar{u}_{ji}^e + \frac{f_{ji}^{e,*}}{2\tilde{d}_{ij}^e}, \qquad f_{ji}^{e,*} = -f_{ij}^{e,*}$$

for volumetric bar states. A formula similar to (6.14) is used to limit the states $\bar{u}_{i,k}^{e,*}$.

6.3 Numerical results

Having completed our discussion of bound-preserving limiting strategies, we now present numerical results obtained with DG discretizations. The hyperbolic problems under investigation include multidimensional versions of Burgers equation, in addition to the shallow water system and the Euler equations of gas dynamics. For scalar problems, we employ solely the local Lax–Friedrichs flux. For systems, we use the Harten–Lax–van Leer (HLL) Riemann solver [Har83b, Eq. (3.15)]

$$\mathrm{f}_{\boldsymbol{n}}(u_L, u_R) = \begin{cases} \mathbf{f}(u_L) & \text{if } 0 \leq s_L, \\ \dfrac{s_R \mathbf{f}(u_L) - s_L \mathbf{f}(u_R) + s_L s_R (u_R - u_L)}{s_R - s_L} & \text{if } s_L < 0 < s_R, \\ \mathbf{f}(u_R) & \text{if } s_R \leq 0 \end{cases} \tag{6.15}$$

by default. The wave speeds s_L and s_R in (6.15) are estimated as follows

$$s_L = \min\{\boldsymbol{v}_L \cdot \boldsymbol{n} - a_L, \boldsymbol{v}_R \cdot \boldsymbol{n} - a_R\}, \qquad s_R = \max\{\boldsymbol{v}_L \cdot \boldsymbol{n} + a_L, \boldsymbol{v}_R \cdot \boldsymbol{n} + a_R\}.$$

Here $\boldsymbol{v}_{L,R}$ denotes the velocity of the fluid. For the SWE, $a_{L,R} = \sqrt{g h_{L,R}}$ is the celerity, whereas for the Euler equations $a_{L,R} = \sqrt{\gamma p_{L,R}/\rho_{L,R}}$ is the speed of sound. We refer to [Tor09, Ch. 10] for an in-depth discussion of the HLL flux, modifications thereof, and improved wave speed estimates. In our experience, the influence of numerical fluxes on the approximation quality is marginal for second and higher order DG schemes compared to the case of piecewise constant approximations. Since (6.15) is an entropy stable numerical flux [Che17, Cor. 3.2], we expect that no entropy correction is required for MCL approximations. Other numerical fluxes such as Roe's approximate Riemann solver [Tor09, Ch. 11], might produce stationary entropy shocks at sonic points or other entropy-violating solutions. To resolve such issues, one can use the fully discrete entropy fixes developed in [Kuz22a]. For some scalar problems, even the use of entropy stable numerical fluxes can produce approximations that do not converge to the unique vanishing viscosity solution [Kur07b]. In our experience [Kuz22a], such issues do not seem to arise for discretizations of the SWE and of the Euler equations, which is why we choose not to discuss entropy fixes any further in the DG context.

For steady problems, we use pseudo-time stepping with mass lumping in the unconstrained DG schemes. Therefore, the time derivative term from the definition of antidiffusive volumetric contributions (6.10a)–(6.10b) is omitted in such examples.

Temporal discretization is performed using explicit SSP RK time stepping schemes with $p \in \{1, 2, 3\}$ stages. By default, we use the third-order Shu–Osher method [Shu88], labeled SSP3 RK. At the beginning of each p-stage update, the time step is chosen adaptively to satisfy the CFL-like condition

$$\Delta t = \min_{e\in\{1,\dots,E\}} \min_{i\in\{1,\dots,N\}} \frac{\nu\, m_i^e}{\sum_{j\in\tilde{\mathcal{N}}_i\setminus\{i\}} 2\tilde{d}_{ij}^e + \sum_{\Gamma_k^e\in\mathcal{F}(K^e)} 2d_{i,k}^e}, \tag{6.16}$$

where $\nu \in (0, 1]$ is a user-prescribed CFL-parameter. For scalar problems we use $\nu = 1$, whereas for systems we choose $\nu = 0.5$ to avoid the need for repetition of individual Runge–Kutta stages.

In all examples below, discrete initial conditions are obtained via consistent $\mathrm{L}^2(\Omega)$ projections. Depending on the application, this strategy may produce oscillatory profiles. Alternatively, we may evaluate the continuous function u_0 in every node and use these values as Bernstein coefficients of the discrete initial data. This strategy produces second-order accurate approximations to u_0 [Lai07, Thm. 2.45] but does not qualify to be interpreted as a projection P in the sense that $P^2 = P$. As a consequence of (6.4), approximations obtained in this manner preserve bounds of the exact initial conditions.

In the following sections, we present results obtained with unconstrained discontinuous Galerkin schemes (DG), the low order method (LOW), as well as its bound-preserving counterpart (MCL). These algorithms correspond to spatial semi-discretizations (6.2), (6.8), and (6.13), respectively. We append $-\mathbb{P}_p$ and $-\mathbb{Q}_p$ with $p \in \mathbb{N}_0$ to the acronyms to specify which DG baseline discretization is employed. The implementation of all methods discussed in this chapter is based on the open source C++ library MFEM (see also [And21]). Among other features, MFEM supports the use of arbitrary order Bernstein elements. The numerical results displayed in this section are visualized using the open source C++ software GLVis that comes with MFEM.

6.3.1 Burgers equation

In this section, we study two variants of the multidimensional Burgers equation. First, we consider the isotropic extension of the 1D conservation law (2.24) to multiple space dimensions. Subsequently, we solve (2.24) using a space-time discretization approach.

6.3.1.1 Isotropic extension of Burgers equation to 2D

Let us consider the scalar conservation law

$$\frac{\partial u}{\partial t} + \frac{1}{2}\nabla\cdot(\boldsymbol{v}u^2) = 0 \qquad \text{in } \Omega\times(0,T),$$

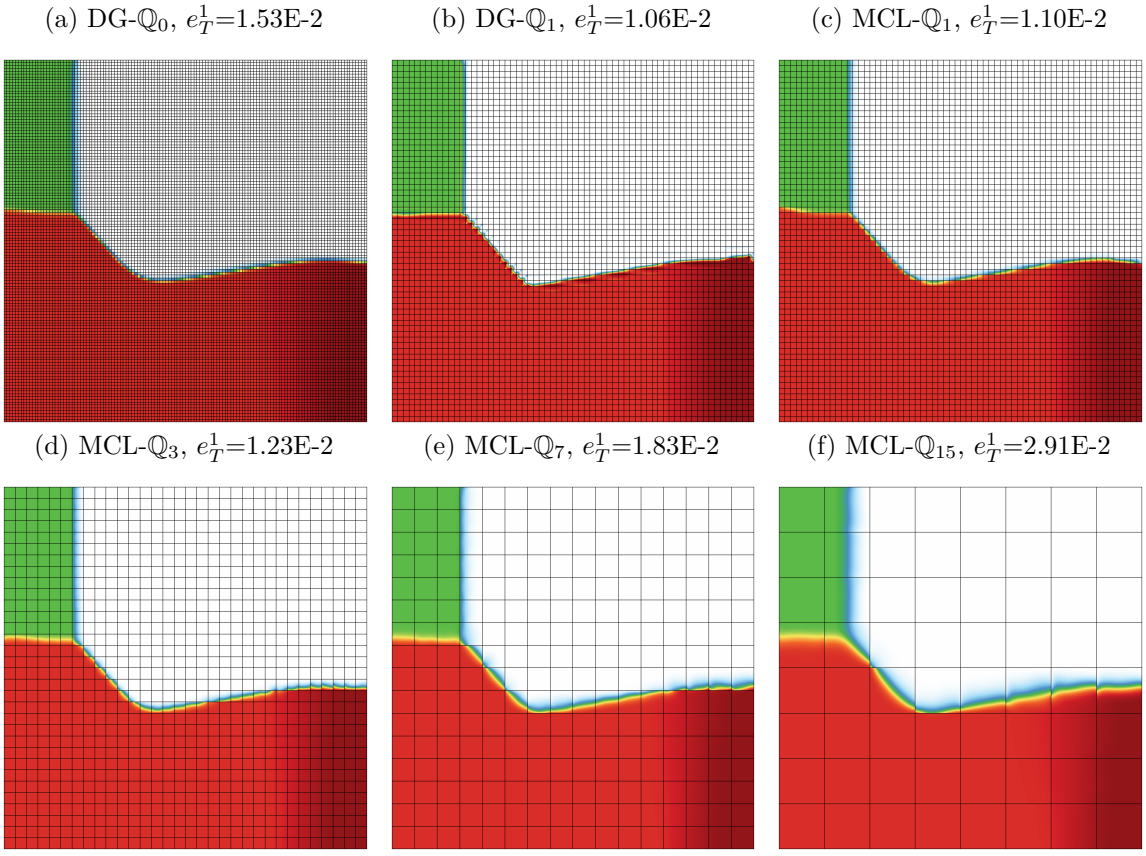


Figure 6.3: Burgers equation in 2D with initial condition (6.17). DG-$\mathbb{Q}_0$, DG-$\mathbb{Q}_1$, and MCL-$\mathbb{Q}_p$ approximations using $p \in \{1, 3, 7, 15\}$ at $T = 0.5$. Numerical solutions obtained with SSP3 RK time stepping and $\Delta t = 10^{-3}$ on uniform quadrilateral meshes with 128^2 DOFs.

where $\boldsymbol{v} = (1, 1)^\mathsf{T}$, $\Omega = (0, 1)^2$, $T = 0.5$, and

$$\Gamma_- = \Gamma_-(t) := \{\boldsymbol{x} \in \partial\Omega : \mathbf{f}'(u(\boldsymbol{x}, t)) \cdot \boldsymbol{n} < 0\} = \{\boldsymbol{x} \in \partial\Omega : u(\boldsymbol{x}, t)\, \boldsymbol{v} \cdot \boldsymbol{n} < 0\}.$$

As initial data, we use the piecewise constant function

$$u_0(\boldsymbol{x}) = \begin{cases} -1 & \text{if } x > 0.5 \text{ and } y > 0.5, \\ -0.2 & \text{if } x < 0.5 \text{ and } y > 0.5, \\ 0.5 & \text{if } x < 0.5 \text{ and } y < 0.5, \\ 0.8 & \text{if } x > 0.5 \text{ and } y < 0.5. \end{cases} \tag{6.17}$$

The inflow boundary data $\hat{u}$ is defined using the exact solution, which can be found in [Gue14, Sec. 5.1]. In this example, the location of the points separating the in- and outlets along the vertical domain boundaries is changing in time.

We solve this problem with DG schemes of varying polynomial degree whilst keeping the total number of degrees of freedom (DOFs) constant at 128^2. Uniform meshes with 128^2 and 64^2 square elements are employed to compute DG-$\mathbb{Q}_0$ and DG-$\mathbb{Q}_1$ approximations, respectively. In addition, we compute MCL-$\mathbb{Q}_p$ solutions for $p \in \{1, 3, 7, 15\}$. We use the values of the exact solution to determine whether or not a boundary point belongs to the inlet Γ_-. The results displayed in Fig. 6.3 are obtained with SSP3 RK time stepping using a constant time step of $\Delta t = 10^{-3}$. The unconstrained DG-$\mathbb{Q}_1$ scheme produces spurious oscillations at the shocks and the corresponding approximation violates global maximum principles. Based on the approximate $\mathrm{L}^1(\Omega)$ errors e_T^1 at the final time, the MCL-$\mathbb{Q}_1$ scheme produces the most accurate approximation among the flux limited schemes. Since the exact solution to this problem at any time instant is composed of states that are at most piecewise linear [Gue14, Sec. 5.1], this result is unsurprising. For such problems, there is no benefit in using high order finite elements because they do not converge faster than piecewise $\mathbb{Q}_1$ approximations as the mesh is refined. We note that the MCL-$\mathbb{Q}_3$ result is still more accurate than the DG-$\mathbb{Q}_0$ approximation on a mesh with 16 times more elements, while DG-$\mathbb{Q}_0$ outperforms MCL-$\mathbb{Q}_7$ and MCL-$\mathbb{Q}_{15}$. The latter produces the least accurate approximation, which is due to the fact that the shocks at the final time are not as well aligned with the mesh edges as in the case of the lower order approximations on finer grids. Still, the overall shape of the exact solution remains recognizable in this very high order approximation on a rather coarse mesh.

(a) Boundary type based on u_h, $\lambda_{ij} = \max\{|u_i|, |u_j|\}|\boldsymbol{v} \cdot \boldsymbol{n}|$ (b) Boundary type based on u_h, $\lambda_{ij} = \max\{|u_i|, |u_j|\}|\boldsymbol{n}|_1$ (c) Boundary type based on $\hat{u}$, $\lambda_{ij} = \max\{|u_i|, |u_j|\}|\boldsymbol{v} \cdot \boldsymbol{n}|$

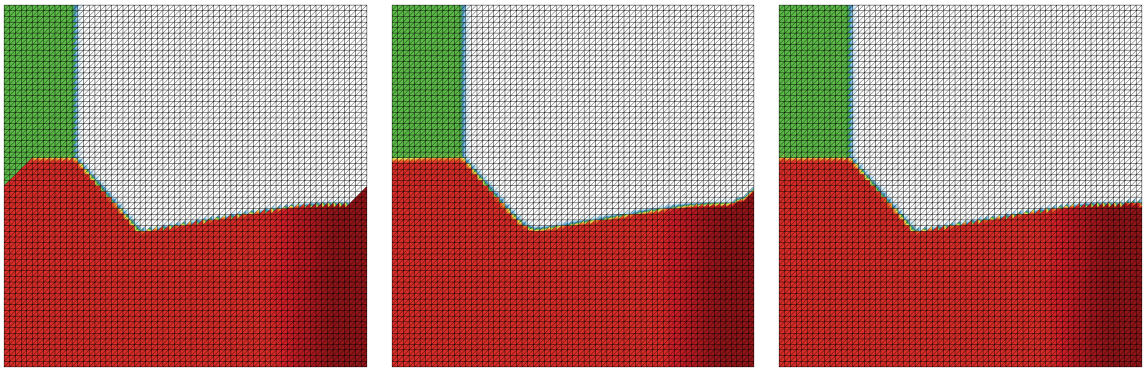


Figure 6.4: Burgers equation in 2D with initial condition (6.17). Variants of MCL-$\mathbb{P}_1$ approximations at $T = 0.5$ obtained with SSP3 RK time stepping and $\Delta t = 10^{-3}$ on uniform triangular meshes using $2 \cdot 128^2$ DOFs.

In this example, the flux function $\mathbf{f}(u) = \frac{u^2}{2}\boldsymbol{v}$ is isotropic. Thus, we may safely use our usual wave speed estimate $\lambda_{ij} = \max\{|u_i|, |u_j|\}|\boldsymbol{v} \cdot \boldsymbol{n}|$. Note that for quadrilateral meshes aligned with the coordinate axis, this expression can only become zero if $u_i = u_j = 0$

because none of the element normals is orthogonal to $\boldsymbol{v} = (1,1)^\mathsf{T}$. For general meshes this statement is not valid. In principle, this issue should not lead to problems even if some wave speeds are zero. However, such sonic points can produce wrong numerical solutions if the boundary type is determined based solely on the sign of *interior* solution values. We illustrate this issue in Fig. 6.4. The MCL-$\mathbb{P}_1$ solutions displayed therein are obtained on structured triangular meshes such that $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ is zero for exactly one face per element. Figs. 6.4a and 6.4b display approximations in which the boundary type is determined based on the interior state. In the simulation that produced Fig. 6.4b, we overestimated the wave speed by setting $\lambda_{ij} = \max\{|u_i|, |u_j|\}|\boldsymbol{n}|_1$, where $|\cdot|_1$ is the l^1 norm. Neither approximation captures the exact solution at the vertical boundaries correctly. For the approximation in Fig. 6.4b, this issue can best be observed if a larger end time is employed. As we can see in Fig. 6.4c, implementations in which boundary types are determined based on the inflow profile do not produce these spurious features and should therefore be preferred in practice.

6.3.1.2 Space-time Burgers equation

The one-dimensional Burgers equation (2.24) equipped with suitable initial and boundary conditions is equivalent to the 2D boundary value problem

$$\begin{aligned} \nabla \cdot \left[\tfrac{1}{2}u^2, \quad u\right] &= 0 \qquad \text{in } \Omega, \\ u &= \hat{u} \qquad \text{on } \Gamma_-. \end{aligned}$$

In this section, we use a space-time discretization approach (see Section 3.2.2) to obtain a solution to the steady problem from which one can recover a solution to the transient one-dimensional Burgers equation. In our numerical experiment, $\Omega = (0,1)^2$ and $\Gamma_- = (0,1) \times \{0\} \cup \{0\} \times (0,1)$, provided that $u \geq 0$ on Γ_-. Our boundary data

$$\hat{u} = \begin{cases} 2 & \text{if } |x - 0.125| \leq 0.075, \\ 1 & \text{if } |x - 0.275| \leq 0.075, \\ 0 & \text{otherwise} \end{cases}$$

is inspired by [Möl08, Sec. 3.5.2]. Applying the theory presented in Section 2.3.1 to this problem, we deduce that the solution for small y contains one rarefaction wave and two shocks. The faster waves catch up and merge with the slower moving ones. Once the rarefaction wave reaches the only remaining shock front, the pre-shock value begins to decay, which in turn reduces the shock speed. A closed form expression of the exact solution reads

$$u(x,y) = \begin{cases} 1 & \text{if } 2(x-0.35) \leq y \leq \frac{2}{3}(x-0.2), \\ 2 & \text{if } \max\left\{\frac{2}{3}(x-0.2), x-0.275\right\}\Big\} \leq y \leq 0.5(x-0.05), \\ \frac{x-0.05}{t} & \text{if } 0.05 \leq x \text{ and } \max\left\{0.5(x-0.05), \frac{10}{9}(x-0.05)^2\right\} \leq y, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Approximation and mesh (b) Contour lines, e_T^1=1.23E-2 (c) Steady state residuals

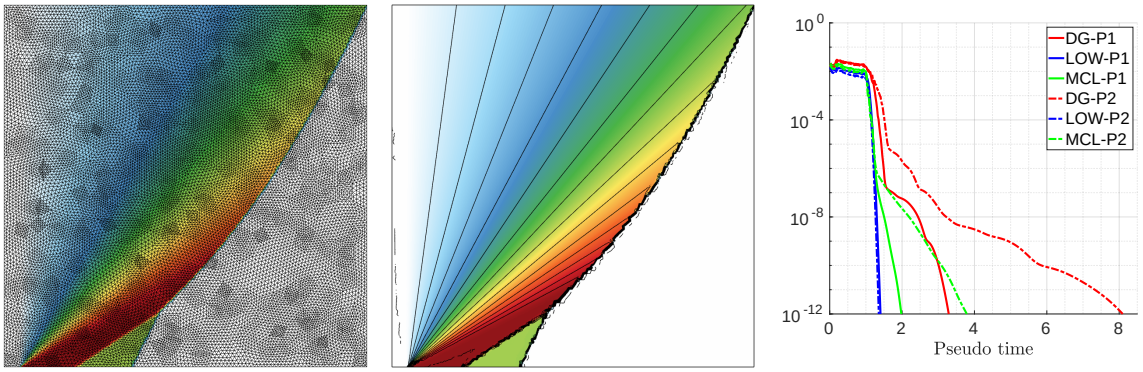


Figure 6.5: Space-time Burgers equation [Möl08]. MCL-$\mathbb{P}_2$ approximation at steady state (a)–(b) obtained with adaptive SSP1 RK pseudo-time stepping and $\nu = 1$ on an unstructured triangular mesh. Convergence history of steady state residuals for various targets (c).

We solve this problem numerically on an unstructured triangular mesh with 33 984 elements and 17 197 vertices. A comparative study of DG, LOW and MCL schemes is performed for $\mathbb{P}_1$ and $\mathbb{P}_2$ spaces. Simulations are cold-started from $u_h \equiv 0$ and marched to steady state using forward Euler pseudo-time stepping with mass lumping. We use the l^2 norm of the steady state residuals as convergence criterion and the value 10^{-12} as threshold. In contrast to standard FEM, combinations of DG with the forward Euler method do not lead to instabilities such as the ones observed in Section 4.5.2.1.

The MCL-$\mathbb{P}_2$ result displayed in Fig. 6.5 clearly reproduces the parabolic shape of the shock front for large enough values of y. Note that the contour lines of the solution in the space-time domain correspond to the characteristics of the time-dependent 1D Burgers equation (2.24). For all schemes under investigation, pseudo-time integrators converge to steady state solutions. This process takes a long time for the lumped DG-$\mathbb{P}_2$ scheme, which produces spurious oscillations of significant magnitude at the shocks. The non-lumped DG scheme converges faster but also produces oscillatory approximations. No ripples are observed in the LOW and MCL results. The fact that the pseudo-time stepping scheme for MCL does converge even on unstructured meshes is a clear advantage over FCT methods. It is unsurprising that the flux-limited schemes require more pseudo-time iterations than the respective low order methods because limiters introduce severe nonlinearities into the algorithm. Improved iterative solvers for MCL discretizations of steady state problems can be designed building on approaches developed in [Bad17] and [Loh21].

6.3.2 Shallow water equations

We consider the two-dimensional shallow water equations with a flat bottom topography, i. e., (4.1) with $d = 2$ and $b \equiv 0$. Experiments are performed for a transient smooth test

problem as well as for a steady example involving shocks.

6.3.2.1 Vorticity advection

Let us first study a benchmark from [Fjo09, Sec. 4.3.1], which is referred to as vorticity advection in the original publication. In this example, the gravitational constant is set to $g = 1$. The initial water height and velocity are given by

$$h_0(x,y) = 1 - \frac{c_1^2}{4c_2 g}\,\mathrm{e}^{-2c_2(x^2+y^2)}, \qquad \boldsymbol{v}_0(x,y) = M\begin{bmatrix}\cos(\alpha)\\ \sin(\alpha)\end{bmatrix} + c_1\mathrm{e}^{-c_2(x^2+y^2)}\begin{bmatrix}y\\ -x\end{bmatrix}.$$

Compared to [Fjo09, Sec. 4.3.1], we modify the parameters

$$M = \sqrt{2}, \qquad c_1 = -0.1, \qquad c_2 = 0.005, \qquad \alpha = \frac{\pi}{4}$$

of this test problem to represent an example that more closely resembles a similar benchmark for the Euler equations [Shu98, Sec. 5.1], which we study in Section 6.3.3.1. Furthermore, in our approach, the spatial domain $\Omega = (-50, 50)^2$ is equipped with periodic boundaries. Our reason for this modification is that at any time $t = 100n$ with $n \in \mathbb{N}_0$, the exact solution to this problem coincides with the initial condition.

We solve the vorticity advection problem with the MCL-$\mathbb{P}_1$ scheme using the HLL flux (6.15) and adaptive SSP2 RK time stepping. Additionally, we test the MCL-$\mathbb{P}_2$ scheme equipped with HLL and local Lax–Friedrichs (LxF) fluxes. For $\mathbb{P}_2$ discretizations, we use adaptive SSP3 RK time stepping. In this study, we enforce only the positivity of the water height and no local bounds because the exact solution is smooth. The meshes used in this study are periodic and consist of uniform simplicial elements. In Tab. 6.1 we present the $\mathrm{L}^1(\Omega)$ errors for the water height at the final time $T = 100$ and the corresponding experimental orders of convergence (EOC).

$100\sqrt{2}/h$	$\mathbb{P}_1$ with HLL	EOC	$\mathbb{P}_2$ with HLL	EOC	$\mathbb{P}_2$ with LxF	EOC
16	5.37E-03		5.04E-04		6.20E-04	
32	1.05E-03	2.35	4.42E-05	3.51	1.09E-04	2.51
64	1.95E-04	2.43	4.37E-06	3.34	1.58E-05	2.78
128	3.77E-05	2.37	5.03E-07	3.12	2.07E-06	2.94
256	7.92E-06	2.25	5.82E-08	3.11	2.84E-07	2.86

Table 6.1: Convergence history for the vorticity advection problem [Fjo09, Sec. 4.3.1]. The $\|\cdot\|_{\mathrm{L}^1(\Omega)}$ errors in the water height of the two-dimensional shallow water model at $T = 100$ and the corresponding EOC for MCL schemes enforcing only positivity of the water height.

In this example, optimal convergence rates can be observed if the HLL flux is used and slightly less than third order of accuracy is obtained with the local Lax–Friedrichs flux for $\mathbb{P}_2$ elements. Let us point out that for higher than second order spaces, the

errors in the momentum components do not converge with optimal rates. More accurate wave speed estimates in the HLL flux (see [Tor09, Ch. 10]) or better Riemann solvers might cure this unsatisfactory behavior. We also remark that our standard version of the DG-MCL schemes degrades the order of convergence because the local bounds for limiting are too tight. In fact, the example presented in this section does not require enforcement of local bounds because the solution is smooth. Therefore, a smoothness indicator should be designed to automatically detect this situation and disable the limiter. Examples of such sensors can be found in [Dio13, Loh17b, Dob18, Haj20c, Paz21].

6.3.2.2 Supercritical flow in a constricted channel

In this benchmark proposed in [Zie95], we set $\omega = \tan\left(\frac{\pi}{36}\right)$ and consider the domain

$$\Omega = \{(x, y) \in (-10, 80) \times (0, 40) : \omega \max\{0, x\} < y < 40 - \omega \max\{0, x\}\},$$

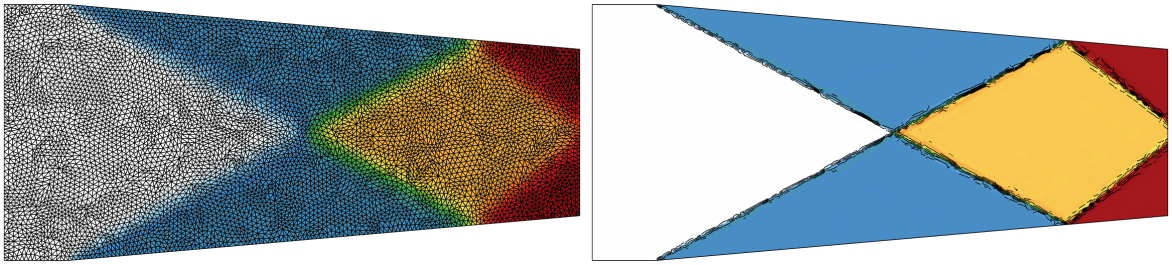
which represents a 40 m wide and 90 m long channel. Ten meters to the right of the supersonic inlet $\{(-10, y) \in \mathbb{R}^2 : y \in (0, 40)\}$, the lateral channel boundaries are symmetrically constricted with an angle of $5° = \pi/36$. The right boundary is a supersonic outlet. Reflecting boundary conditions are imposed on the horizontal lateral walls and along the channel constrictions. Setting the gravitational constant to $g = 0.16$, we obtain a supercritical flow regime with an inlet Froude number of 2.5. We use $u = (1, 1, 0)$ as spatially uniform initial condition and inlet boundary data. The profiles of the exact solution exhibit oblique shock waves that form at the channel constrictions, meet in the center of the domain, and are reflected at the opposite channel walls. A symmetric steady state flow pattern consisting of piecewise constant states separated by steady shock fronts is assumed soon after [Zie95].

We employ an unstructured, nonsymmetric triangular mesh consisting of $E = 12\,620$ elements and march numerical solutions to steady state with forward Euler pseudo time stepping. For the unconstrained DG schemes we use the fixed time step $\Delta t = 10^{-2}$. Computations are terminated when the l^2 norm of the residual for all variables becomes less than 10^{-12}. In Fig. 6.6, we display approximations for the water height at the steady states of DG-$\mathbb{P}_0$, DG-$\mathbb{P}_1$, LOW-$\mathbb{P}_1$ and MCL-$\mathbb{P}_1$ schemes. Additionally, we present the result of the MCL-$\mathbb{P}_1$ simulation with limiting for the interfacial DG fluxes disabled. In other words, the lumped version of boundary integrals and the local Lax–Friedrichs Riemann solver are used in this approach, which we refer to as MCLV-$\mathbb{P}_1$. The appended letter V indicates that flux correction is restricted to volumetric terms.

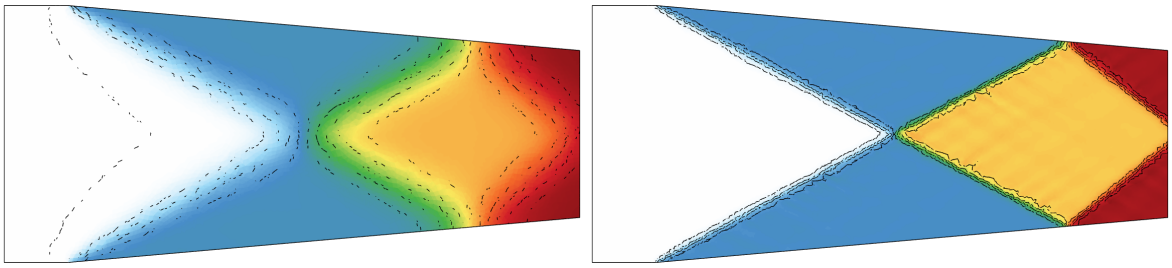
The unconstrained DG schemes suffer from their usual deficiencies, i. e., low resolution at shocks for the DG-$\mathbb{P}_0$ version and Gibbs phenomena for the DG-$\mathbb{P}_1$ version. The low order $\mathbb{P}_1$ approximation is unacceptably diffusive. The MCL results exhibit crisp resolution of shocks and good symmetry preservation properties even on the employed nonsymmetric, unstructured mesh that is also not aligned with the shocks. Interestingly enough, the deactivation of the interfacial flux limiter does not seem to degrade the

overall accuracy. Thus, for some problems it may be worthwhile to employ lumped Lax–Friedrichs fluxes instead of our interfacial bar state limiter.

(a) DG-$\mathbb{P}_0$ and mesh, $u_h \in [1.000, 1.835]$ (b) DG-$\mathbb{P}_1$ and contour lines, $u_h \in [0.867, 1.984]$



(c) LOW-$\mathbb{P}_1$ and contour lines, $u_h \in [1.000, 1.831]$ (d) MCLV-$\mathbb{P}_1$ and contour lines, $u_h \in [1.000, 1.868]$



(e) MCL-$\mathbb{P}_1$ and contour lines, $u_h \in [1.000, 1.847]$ (f) Convergence to steady state

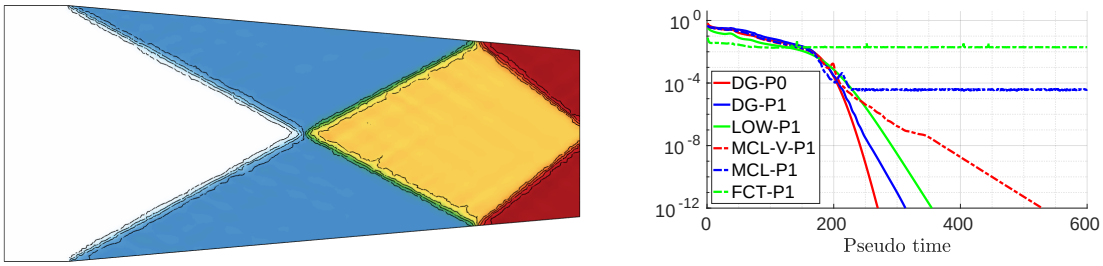


Figure 6.6: Constricted channel flow for the shallow water equations [Zie95]. DG-$\mathbb{P}_0$ and $\mathbb{P}_1$ approximations to the water height at steady state (a)–(d) and at $T = 600$ (e) obtained with SSP1 RK pseudo time stepping on an unstructured triangular mesh with 12 620 elements. Convergence history of steady state residuals for monolithic schemes and convergence indicator (6.18) for the SL method (f).

For comparison purposes, we also studied the steady-state convergence behavior of the DG-$\mathbb{P}_1$ scheme equipped with a geometric slope limiter. Specifically, we used the open source Matlab code FESTUNG (see [Fra20, Reu21]) and applied the vertex-based slope limiter from Kuz10a to each component of the vector of conserved unknowns. More sophisticated versions of this scheme exist [Dob18, Haj19] but for the problem under consideration, these strategies yield similar results. Since slope-limited (SL) algorithms constrain fully discrete schemes, they do not have a well-defined steady state residual.

As a convergence indicator for the SL method, we therefore use

$$\sum_{e=1}^{E} \frac{\|\tilde{u}_h^e - u_h^e\|_{\mathrm{L}^2(K^e)}^2}{|K^e|}, \tag{6.18}$$

where $\tilde{u}_h$ and u_h denote solutions at two consecutive pseudo-time steps.

In this example, the MCLV-$\mathbb{P}_1$ scheme does converge to a steady state solution, while the residual of the standard MCL-$\mathbb{P}_1$ method stagnates after reaching the order of 10^{-4}. Fig. 6.6f shows the convergence history for each DG and MCL scheme in addition to the value of (6.18) for the SL algorithm. Let us remark that for monolithic schemes, steady state residuals are better indicators of convergence than quantities like (6.18) because the latter approach encourages cancellation effects. For the SL scheme studied here, this is not an issue because fluctuations around a steady state can be observed even visually. The rather disappointing lack of convergence of the standard MCL-$\mathbb{P}_1$ scheme requires further investigations. It can possibly be cured by using a more advanced iterative solver for the nonlinear steady-state problem. We performed experiments with alternative definitions of local bounds and also considered the non-lumped local Lax–Friedrichs flux instead of the HLL Riemann solver but to no avail. If only nonnegativity of the water height is enforced instead of local bounds as in the sequential limiter, a steady state is reached. In this example, however, enforcing only the IDP property corresponds to essentially just using the standard DG-$\mathbb{P}_1$ method.

6.3.3 Euler equations of gas dynamics

In the final experiments of this thesis we apply our DG schemes to the compressible Euler equations. First, we study a benchmark with a smooth solution. Then we consider a modification of the classical Sod shock tube problem. Finally, we solve an example with a forward facing step. The adiabatic constant is set to $\gamma = 1.4$ in all problems.

6.3.3.1 Isentropic vortex

One of the test problems proposed in [Shu98, Sec. 5.1] evolves a smooth vortex in the domain $\Omega = (-5, 5)^2$ with periodic boundaries. The initial conditions are set as follows

$$\rho_0(x,y) = \theta_0(x,y)^{\frac{1}{\gamma-1}}, \qquad \boldsymbol{v}_0(x,y) = \begin{bmatrix}1\\1\end{bmatrix} + \frac{\varepsilon}{2\pi}\,\mathrm{e}^{0.5(1-(x^2+y^2))}\begin{bmatrix}-y\\x\end{bmatrix},$$
$$p_0(x,y) = \theta_0(x,y)^{\frac{\gamma}{\gamma-1}}, \qquad \theta_0(x,y) = 1 - \frac{(\gamma-1)\,\varepsilon^2}{8\gamma\pi^2}\mathrm{e}^{1-(x^2+y^2)},$$

where $\varepsilon = 5$. At any time $t = 10n$ with $n \in \mathbb{N}_0$, the exact solution to this problem coincides with the initial condition. The density profiles in this example look similar to those of the water height for the vorticity advection example studied in Section 6.3.2.1.

We solve this problem numerically with the MCL-$\mathbb{Q}_1$ scheme using the HLL flux and SSP2 RK time stepping. Additionally, we test the MCL-$\mathbb{Q}_2$ scheme with either HLL or local Lax–Friedrichs (LxF) fluxes in combination with SSP3 RK time stepping. Similarly to our approach in Section 6.3.2.1, we only enforce nonnegativity of the density, and, additionally, nonnegativity of the internal energy for individual bar states. Since the solution is smooth, we do not rely on limiters to enforce local bounds. The meshes used in this study are periodic and, this time, consist of uniform quadrilateral elements. In Tab. 6.2 we present the $\mathrm{L}^1(\Omega)$ errors for the density at the final time $T = 10$ and the corresponding experimental orders of convergence.

$10/h$	$\mathbb{Q}_1$ with HLL	EOC	$\mathbb{Q}_2$ with HLL	EOC	$\mathbb{Q}_2$ with LxF	EOC
16	5.37E-03		6.06E-04		5.29E-04	
32	9.32E-04	2.53	2.60E-05	4.54	5.07E-05	3.38
64	1.58E-04	2.56	2.35E-06	3.47	7.84E-06	2.69
128	2.90E-05	2.44	2.70E-07	3.12	1.22E-06	2.68
256	6.28E-06	2.21	3.31E-08	3.03	1.79E-07	2.78

Table 6.2: Convergence history for the isentropic vortex problem [Shu98, Sec. 5.1]. The $\|\cdot\|_{\mathrm{L}^1(\Omega)}$ errors in the density of the two-dimensional Euler equations at $T = 10$ and the corresponding EOC for MCL schemes enforcing only positivity of density and internal energy.

Virtually the same conclusions as in Section 6.3.2.1 can be drawn from this experiment. Thus, the behavior of DG schemes observed for the shallow water equations carries over to problems in gas dynamics. Moreover, the orders of accuracy do not seem to depend on the geometry of elements since here we use quadrilaterals instead of simplices that were employed in Section 6.3.2.1. Once more, the use of the HLL flux is worthwhile for higher order spaces, although, again, the convergence rates deteriorate for components of the solution vector other than the density. Again, enforcement of local bounds by the MCL limiter leads to larger error values and decreased EOCs. This issue can be resolved using a well-designed smoothness indicator to relax the bounds.

6.3.3.2 Modified Sod shock tube problem

Let us now apply the high order property-preserving DG methods to a one-dimensional Riemann problem studied in [Tor09, Sec. 6.4]. The domain $\Omega = (0, 1)$ has a supersonic inlet and a reflecting wall boundary on the left and right, respectively. The initial condition expressed in conserved variables is given by

$$u_0(x) = \begin{cases} (1, 0.75, 89/32) & \text{if } x < 0.25, \\ (0.125, 0, 0.25) & \text{if } x > 0.25. \end{cases}$$

The corresponding initial states of the primitive variables are similar to those of Sod's shock tube problem [Sod78] with the exception that the velocity is nonzero for the left

initial state. This modification to the problem we solved in Section 3.4.3.1 produces a sonic point within the rarefaction wave region. Depending on the employed algorithm, numerical solutions may contain nonphysical entropy shocks at that location (see [Tor09, Kuz22a]).

We solve the modified shock tube problem up to end time $T = 0.25$ using MCL-$\mathbb{P}_p$ with $p \in \{1, 3, 7, 15, 31\}$ whilst keeping the number of DOFs for each unknown constant at 128. Uniform meshes and adaptive SSP3 RK time stepping are employed in this study, the results of which can be found in Fig. 6.7. Here the HLL flux is used but the results obtained with the local Lax–Friedrichs flux are similar. A reference solution is obtained with the DG-$\mathbb{P}_0$ method on a fine mesh consisting of $E = 10^4$ elements.

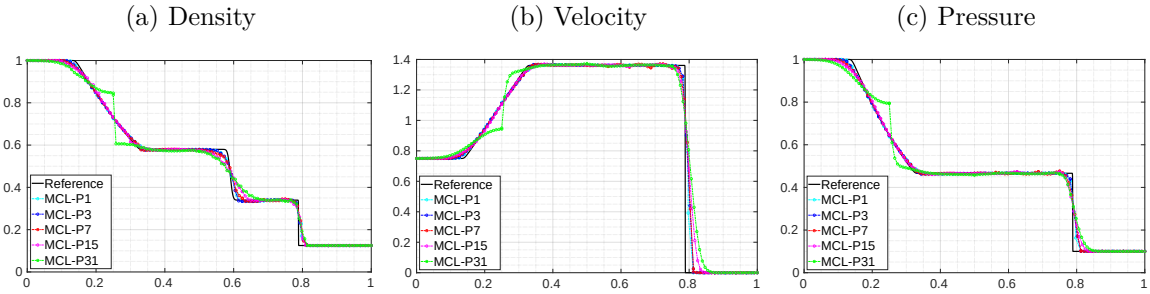


Figure 6.7: Modified Sod shock tube problem for the Euler equations [Tor09, Sec. 6.4]. MCL-$\mathbb{P}_p$ approximations at $T = 0.25$ using 128 DOFs per conserved unknown and $p \in \{1, 3, 7, 15, 31\}$. Solutions obtained with adaptive SSP3 RK time stepping and $\nu = 0.5$ on uniform meshes.

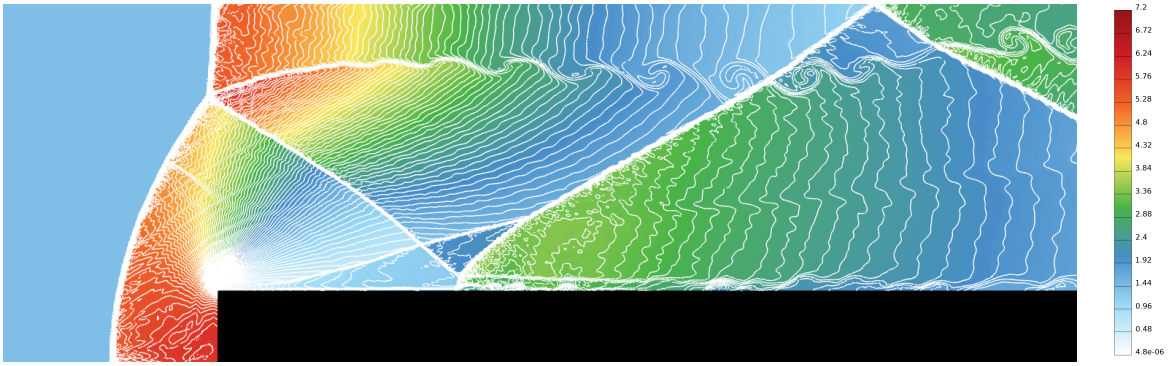
We observe an entropy shock at the sonic point for the very high order space $\mathbb{P}_{31}$. The employed mesh consists of only four elements in this case, which does not seem to be sufficient to produce accurate results for the modified shock tube problem. We remark however, that the entropy shock disappears on finer meshes. The other approximations are to a large degree satisfactory apart from some minor oscillations visible in the profiles of the MCL-$\mathbb{P}_7$ scheme. Increased amounts of dissipation in the higher order spaces lead to more significant smearing of the contact discontinuity. This issue can to some degree be resolved by employing the HLLC flux [Tor09, Ch. 10] instead of HLL or local Lax–Friedrichs Riemann solvers. We conclude that the use of very high order spaces such as $\mathbb{P}_{31}$ on rather coarse meshes does seem to be a good idea. It would be interesting to investigate whether the occurrence of entropy shocks could be prevented by using an entropy limiter such as the one presented in Section 3.3.6.

6.3.3.3 Forward facing step

Finally, we apply our property-preserving DG schemes to a two-dimensional problem that was studied in depth by Woodward and Colella [Woo84]. The spatial domain $\Omega = (0, 3) \times (0, 1) \setminus [0.6, 3] \times [0, 0.2]$ represents a wind tunnel with a forward facing step that acts as an obstacle for the fluid flow. The initial values of the conserved

variables for the diatomic gas are set to $u_0 \equiv (1.4, 4.2, 0, 8.8)^\mathsf{T}$ in the whole domain. This initial condition corresponds to a Mach 3 flow. The boundary condition imposed at the supersonic inlet $\{0\} \times (0, 1)$ is given by $u(0, y, t) \equiv u_0$ for all $y \in (0, 1)$ and $t \geq 0$. Since the flow remains supersonic, no boundary condition is needed at the outlet $\{3\} \times (0.2, 1)$. All other domain boundaries are reflecting walls.

(a) MCL-$\mathbb{P}_1$ with SSP2 RK on a mesh consisting of 163 586 elements and 82 436 vertices, $\rho_h \in [0.090, 6.6]$



(b) MCL-$\mathbb{P}_2$ with SSP3 RK on a mesh consisting of 81 585 elements and 41 245 vertices, $\rho_h \in [0.097, 6.9]$

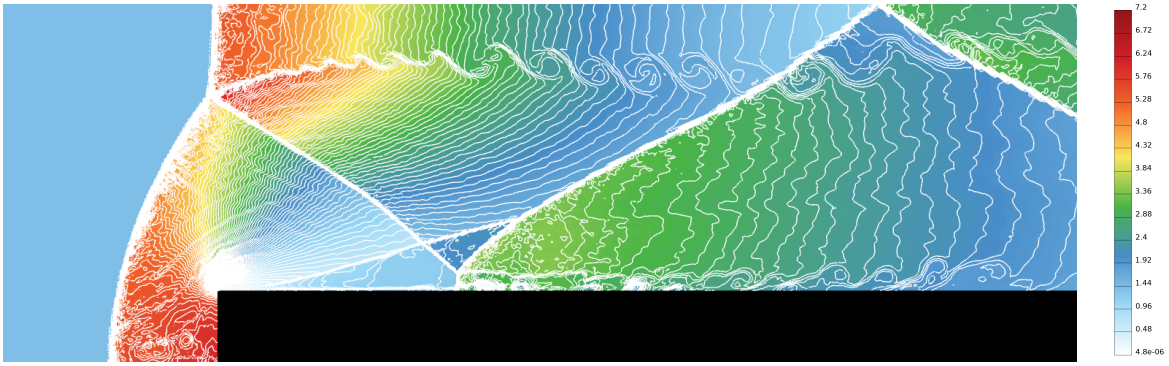
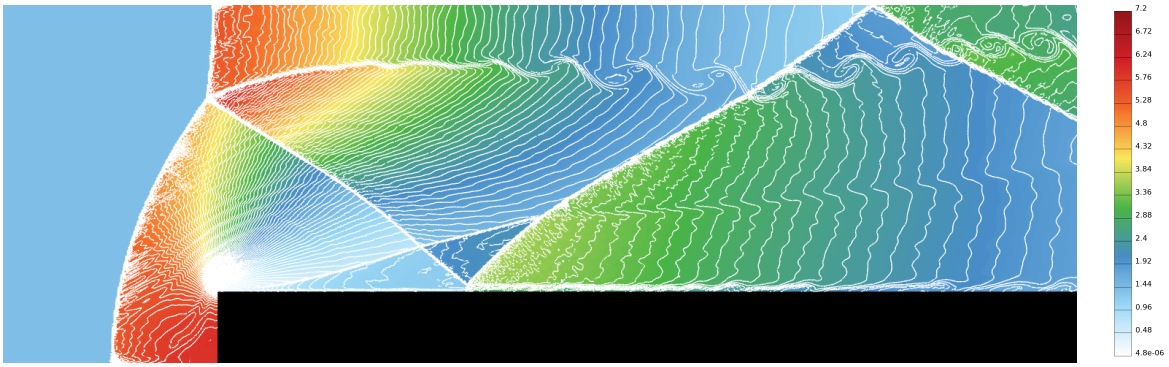


Figure 6.8: Forward facing step for the Euler equations [Woo84]. Density distribution at $T = 4$ obtained using local Lax–Friedrichs fluxes and adaptive time stepping with $\nu = 0.5$ on unstructured, locally refined, triangular meshes created with gmsh [Geu09]. Plots show 100 contour lines.

The reflection of the gas at the step produces a bow shock, which propagates towards the upper wall, where it is reflected. More reflections occur further to the right in the wind tunnel. By the final time $T = 4$, a triple point above the step develops from the reflected bow shock. A Kelvin–Helmholtz instability emanates from this triple point and produces vortical features in the downstream region. The re-entrant corner of the domain is the center of a rarefaction wave. Owing to the imposed reflecting wall boundary conditions at the step, the exact velocity at the inward-pointing corner is zero. To capture this behavior reasonably well in numerical simulations, one should employ local mesh refinement in this region [Woo84, Hen21].

(a) MCL-$\mathbb{P}_1$ with SSP2 RK on a mesh consisting of 163 586 elements and 82 436 vertices, $\rho_h \in [0.069, 7.1]$



(b) MCL-$\mathbb{P}_2$ with SSP3 RK on a mesh consisting of 81 585 elements and 41 245 vertices, $\rho_h \in [0.061, 7.0]$

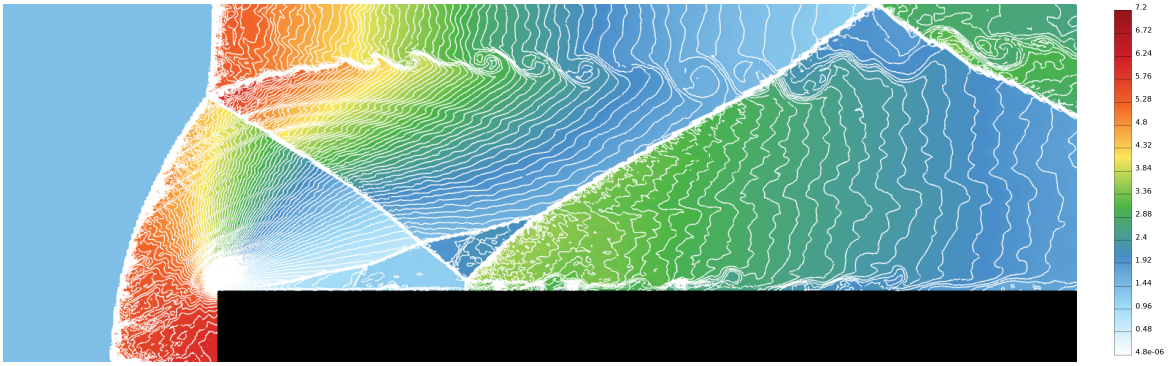


Figure 6.9: Forward facing step for the Euler equations [Woo84]. Density distribution at $T = 4$ obtained using HLL fluxes and adaptive time stepping with $\nu = 0.5$ on unstructured, locally refined, triangular meshes created with gmsh [Geu09]. Plots show 100 contour lines.

We solve the forward facing step problem numerically using MCL-$\mathbb{P}_1$ with SSP2 RK as well as MCL-$\mathbb{P}_2$ with SSP3 RK. For this test problem, we perform a comparative study of local Lax–Friedrichs and HLL Riemann solvers for interfacial boundary terms. The unstructured triangular meshes employed in our experiments were created with gmsh [Geu09] and exhibit an increased resolution of about $h/4$ around the re-entrant corner of the step. Special care was taken to ensure that the total number of DG unknowns is almost the same in the $\mathbb{P}_1$ and $\mathbb{P}_2$ discretizations. Snapshots of density distributions are displayed in Figs. 6.8 and 6.9 for approximations obtained with local Lax–Friedrichs and and HLL fluxes, respectively.

These profiles capture all flow features described in [Woo84] correctly. In particular, the Kelvin–Helmholtz instability emanating from the triple point can be observed. Moreover, the resolution with which discontinuities are captured is satisfactory for the employed meshes. For this test problem, the differences in results obtained with local Lax–Friedrichs and HLL fluxes seem to be insignificant. Among the four approaches, the MCL-$\mathbb{P}_1$ scheme with HLL fluxes produces the sharpest resolution of the bow shock layer.

This behavior is unsurprising because the best strategy to resolve discontinuities is to use low order approximations on fine meshes (and the least diffusive Riemann solver among the available ones). On the other hand, slightly more pronounced vortical features can be seen in the MCL-$\mathbb{P}_2$ profiles. Thus, different choices of baseline discretizations may be appropriate for different applications. The computational performance of a given method should also be taken into account when it comes to making decisions regarding the type of numerical approximations. Let us again stress the need for relaxation of local bounds for flux limiting constraints in smooth regions to obtain better than second order convergence rates with high order schemes [Zha11, Sec. 1]. We expect to see more pronounced vortical features in numerical solutions to the forward facing step problem if such an approach is incorporated into our method.

Chapter 7

Conclusions

This thesis presents some of the author's contributions to the development of algebraic flux correction schemes. Two main fields of applications are considered, geophysical fluid flows and gas dynamics. The corresponding mathematical models are the shallow water equations and the Euler equations of gas dynamics, respectively. After a brief introduction and a review of some of the mathematical theory on hyperbolic PDEs, we focused on numerical methods for solving these problems. In particular, we discussed property-preserving schemes based on the monolithic convex limiting methodology for (systems of) conservation laws [Kuz20a, Kuz20c]. We generalized this strategy to an important nonlinear system of balance laws, the shallow water equations with a topography source term. In addition, we presented our own contributions to the numerical analysis of AFC schemes. Moreover, we extended limiting techniques designed for continuous finite elements to arbitrary order discontinuous Galerkin discretizations.

7.1 Summary

In Chapter 3 we presented an in-depth review of recently developed property-preserving numerical methods for hyperbolic conservation laws. After a brief introduction to continuous finite elements and temporal discretization techniques, we focused on algebraic flux correction schemes. In particular, we addressed important properties of the low order method using the theory developed in [Har83b, Gue16b]. Next, the monolithic convex limiting technology, developed in [Kuz20a] was discussed and a new way to handle the weakly imposed boundary conditions was proposed. Additionally, we presented the sequential limiting approach for hyperbolic systems that was originally developed in [Dob18, Kuz20a]. These techniques were enhanced by the option of performing an additional semi-discrete entropy fix based on Tadmor's entropy stability condition [Tad87, Tad03, Kuz20c]. Sequential limiting for the shallow water equations and enforcement of entropy stability for hyperbolic systems was first pursued by this author [Haj19, Haj21a, Kuz22a]. The numerical results presented in Chapter 3 illustrate the need for each component of the algorithm. For a scalar problem with nonconvex flux, both bound-preserving and entropy-stabilizing limiting were found to be beneficial. For the compressible Euler equations, we observed that the entropy fix has only little influence on the approximate solution. We also demonstrated numerically that we obtain optimal convergence rates for smooth solutions of scalar problems. Stabilization of continuous Galerkin discretizations is achieved solely by using low order nodal time

derivatives in the definition of raw antidiffusive fluxes.

Subsequently, in Chapter 4, we extended the algebraic flux correction schemes for conservation laws discussed in Chapter 3 to the system of shallow water equations with a nonconservative topography term. For configurations with a flat bottom, this term vanishes and the generalized schemes coincide with the original ones. We also developed two new approaches for modeling wetting and drying processes. The first one is an entropy-based velocity fix restricted to flat topographies. The second fix uses a boundary layer approximation and works in the general case.

Furthermore, we presented stability and a priori error analysis for a property-preserving discretization of the advection equation in Chapter 5. Our theoretical findings were corroborated by the conducted numerical experiments. Besides deriving energy and error estimates, we compared the results obtained with monolithic AFC schemes and two representatives of FCT algorithms.

Finally, in Chapter 6 we extended the MCL methodology to discontinuous Galerkin discretizations of arbitrary order. While other high-resolution schemes may produce results of similar quality, our AFC methodology for DG methods belongs to the first *general-purpose* limiting approaches that *provably* guarantee preservation of invariant domains and local bounds even for high order finite element discretizations of nonlinear systems. It turns out that the DG baseline discretization can be adapted to make it bound preserving by means very similar to the ones used for standard finite elements. In fact, most ideas employed in Chapter 6 were actually proposed in the context of high order continuous Galerkin schemes [Kuz20e]. A novelty of our approach is the limiting of interfacial DG fluxes. Introducing low order interfacial bar states that are similar to those arising from boundary terms in continuous Galerkin schemes, we designed a monolithic flux limiter for higher order DG methods. Our approach makes it possible to employ numerical fluxes other than a lumped local Lax–Friedrichs Riemann solver. In particular, we used the HLL flux in our numerical experiments. This flexibility in the choice of Riemann solvers turned out to be quite valuable because we were not able to achieve third order of accuracy with DG-$\mathbb{P}_2/\mathbb{Q}_2$ schemes and local Lax–Friedrichs fluxes. Unfortunately, the more complicated nonlinear nature of the scheme with interfacial flux limiting resulted in a lack of convergence to steady state in one example.

7.2 Outlook

Several interesting avenues for future research currently present themselves. First, it is certainly desirable to apply the presented methods to more complicated problems. One could, for instance, include diffusive terms to model viscous effects. Monolithic convex limiting strategies were successfully applied to scalar convection-diffusion equations by Quezada de Luna and Ketcheson [Que21], who modified the usual bar states to include diffusive fluxes. In the context of inviscid flow problems, flux correction of MCL type

can be extended to more sophisticated hyperbolic systems such as the equations of magnetohydrodynamics (MHD). The ideal MHD system represents a generalization of the Euler equations, which includes a conservation law for a divergence-free magnetic field. The need to keep this vector field (approximately) solenoidal poses a significant additional difficulty and is, therefore, an important topic of its own.

With respect to the MCL scheme for the shallow water equations, we believe that the next step is to incorporate friction terms and, possibly, Coriolis forces into the discrete formulations. Moreover, well-balancing for steady states more complicated than the lake at rest remains to be achieved. Although we were able to demonstrate convergence to one such steady state solution without being exactly well balanced, it is nonetheless a desirable property to capture moving water equilibria exactly.

Our experience with MCL for the shallow water system with bathymetry indicates that similar limiting approaches can be developed, e. g., for the Euler equations with gravity. For this system, well-balancedness is less of an issue than for SWE. Since the gravitational source term is a potential force, its discrete counterpart can be simply decomposed into numerical fluxes and incorporated into the low order bar states. Bound-preserving flux limiting can then be performed in much the same way as for the topography source term of the shallow water equations. Enforcing entropy stability for the Euler equations with gravity is an open problem. The modifications we needed to make in the low order method to enforce entropy stability for the SWE suggest that this task may not be easy. Indeed, the usual entropy pair for the Euler equations is more complicated than the sum of the potential and kinetic energies, which we used as an entropy for SWE.

Moreover, a variety of open problems regarding the theory of AFC schemes remains besides the aspects that were already mentioned in Section 5.4.3. It is certainly desirable to avoid the compatibility condition that we used in our proofs. This task could be accomplished either by proving its validity under certain assumptions or by finding another approach to bound the terms involving low order time derivatives. Another issue worth investigating is whether the results on solvability and discrete maximum principles proven for steady problems [Bar16, Loh19] directly carry over to MCL discretizations. We believe that they do but have not investigated this issue, and focused on the analysis of time-dependent problems instead. Furthermore, it would be interesting to analyze the theoretical properties of AFC schemes based on target discretizations other than piecewise linear continuous finite elements. In particular, higher order finite elements and/or nonconforming spaces could be studied. Our analysis of semi-discrete AFC problems may also serve as a stepping stone for theoretical investigations of fully discrete schemes. We do not expect any major difficulties to arise in such studies but additional technicalities may need to be dealt with. To the best of our knowledge, the theory of AFC schemes is currently restricted to linear PDEs. An extension to nonlinear scalar conservation laws or even systems could be performed by adapting existing analysis for finite volume schemes to the finite element context. However, theoretical investigations

of AFC methods for nonlinear problems are certain to be significantly more involved than the analysis presented in Chapter 5.

The property-preserving DG schemes discussed in Chapter 6, can be generalized to systems of balance laws in the same way as continuous finite element discretizations of the SWE in Chapter 4. One issue that was not addressed in the DG context is entropy limiting. The results from earlier chapters indicate that enforcing entropy stability might not to be necessary for DG (at least in practice) if entropy stable numerical fluxes are employed. Nevertheless, combining our high order property-preserving DG schemes with the entropy limiters designed in [Kuz20d] for continuous finite elements and scalar nonlinear problems would constitute an interesting further development. For the KPP problem [Kur07b] that we solved in Section 3.4.2, the bound-preserving DG schemes discussed in Chapter 6 produce entropy-violating results. A preliminary study that was not included in this thesis has been conducted by the author and showed the feasibility of entropy limiting for this particular benchmark.

With the advent of programming techniques for *graphics processing units* (GPUs), the use of high order spaces has become increasingly popular [And21]. The field of property-preserving methods is no exception to this development [Noe07, Zha11, Dum14, Loh17b, Hen21]. The high order spatial accuracy of finite element discretizations that we use as target schemes may be lost if a low order time integrator is employed. In this thesis, we used SSP time stepping schemes, which are at most fourth-order accurate (in the explicit case) [Ruu02, Thm. 4.1]. As shown in [Que21, Kuz22a, Kuz22b], convex limiting techniques can be used to constrain arbitrary Runge–Kutta time discretizations in a way that makes them property preserving. However, extension of such schemes to high order AFC discretizations of hyperbolic systems has not yet been undertaken.

Finally, the incorporation of smoothness indicators [Dio13, Dum14, Hen21] into the algorithms discussed in this work is yet to be accomplished. In the context of algebraic flux correction schemes, such methods can be found, for instance, in [Loh17b, Dob18, Gue18a, Haj20c, Paz21]. They are supposed to deactivate flux limiters or relax local bounds in smooth regions. Note that the underlying smoothness criteria are often somewhat heuristic, and the resulting algorithms may fail to prevent nonphysical behavior. It is therefore essential to enforce global bounds even if violations of tight local bounds are allowed by the smoothness indicator. Examples of fail-safe limiters based on this design philosophy can be found in [Haj20c, Paz21].

References

[Abg06] R. Abgrall (2006) *Essentially non-oscillatory residual distribution schemes for hyperbolic problems* J. Comput. Phys. **214**: 773–808 doi: 10.1016/j.jcp.2005.10.034

[Abg10] R. Abgrall, J. Trefilík (2010) *An example of high order residual distribution scheme using non-Lagrange elements* J. Sci. Comput. **45**: 3–25 doi: 10.1007/s10915-010-9405-y

[Abg17] R. Abgrall, S. Tokareva (2017) *Staggered grid residual distribution scheme for Lagrangian hydrodynamics* SIAM J. Sci. Comput. **39**: A2317–A2344 doi: 10.1137/16M1078781

[Ain19] M. Ainsworth, S. Jiang, M. A. Sanchéz (2019) *An $\mathcal{O}(p^3)$ hp-version FEM in two dimensions: Preconditioning and post-processing* Comput. Method. Appl. M. **350**: 766–802 doi: 10.1016/j.cma.2019.03.020

[Ama90] H. Amann (1990) *Ordinary Differential Equations* De Gruyter doi: 10.1515/9783110853698

[And17] R. Anderson, V. Dobrev, T. Kolev, D. Kuzmin, M. Quezada de Luna, R. Rieben, V. Tomov (2017) *High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation* J. Comput. Phys. **334**: 102–124 doi: 10.1016/j.jcp.2016.12.031

[And21] R. Anderson, J. Andrej, A. Barker, J. Bramwell, J.-S. Camier, J. Cerveny, V. Dobrev, Y. Dudouit, A. Fisher, T. Kolev, W. Pazner, M. Stowell, V. Tomov, I. Akkerman, J. Dahm, D. Medina, S. Zampini (2021) *MFEM: A modular finite element methods library* Comput. Math. Appl. **81**: 42–74 doi: 10.1016/j.camwa.2020.06.009

[Aud04] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, B. Perthame (2004) *A Fast and Stable Well-Balanced Scheme with Hydrostatic Reconstruction for Shallow Water Flows* SIAM J. Sci. Comput. **25**: 2050–2065 doi: 10.1137/s1064827503431090

[Aud15] E. Audusse, C. Chalons, P. Ung (2015) *A simple well-balanced and positive numerical scheme for the shallow-water system* Commun. Math. Sci. **13**: 1317–1332 doi: 10.4310/CMS.2015.v13.n5.a11

[Aze17] P. Azerad, J.-L. Guermond, B. Popov (2017) *Well-Balanced second-order approximation of the shallow water equation with continuous finite elements* SIAM J. Numer. Anal. **55**: 3203–3224 doi: 10.1137/17M1122463

[Bad17] S. Badia, J. Bonilla (2017) *Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization* Comput. Method. Appl. M. **313**: 133–158 doi: 10.1016/j.cma.2016.09.035

[Bar89] T. J. Barth, D. C. Jespersen (1989) *The design and application of upwind schemes on unstructured meshes* in *Proc. of the 27th Aerospace Sciences Meeting* doi: 10.2514/6.1989-366

[Bar15] M. Barros, P. Rosman, J. Telles (2015) *An effective wetting and drying algorithm for numerical shallow water flow models* J. Braz. Soc. Mech. Sci. **7**: 803–819 doi: 10.1007/s40430-014-0211-6

[Bar16] G. R. Barrenechea, V. John, P. Knobloch (2016) *Analysis of algebraic flux correction schemes* SIAM J. Numer. Anal. **54**: 2427–2451 doi: 10.1137/15M1018216

[Bar17a] G. R. Barrenechea, E. Burman, F. Karakatsani (2017) *Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes* Numer. Math. **135**: 521–545 doi: 10.1007/s00211-016-0808-z

[Bar17b] G. R. Barrenechea, P. Knobloch (2017) *Analysis of a group finite element formulation* Appl. Numer. Math. **118**: 238–248 doi: 10.1016/j.apnum.2017.03.008

[Bar18] G. R. Barrenechea, V. John, P. Knobloch, R. Rankin (2018) *A unified analysis of algebraic flux correction schemes for convection–diffusion equations* SeMA J. **75**: 655–685 doi: 10.1007/s40324-018-0160-6

[Bec03] R. Becker, E. Burman, P. Hansbo, M. G. Larson (2003) *A reduced P^1-discontinuous Galerkin method* Chalmers Finite Element Center Preprint 2003-13, Chalmers University of Technology https://www.researchgate.net/publication/37445460_A_reduced_P1-discontinuous_Galerkin_method

[Ber19] C. Berthon, A. Duran, F. Foucher, K. Saleh, J. D. D. Zabsonré (2019) *Improvement of the hydrostatic reconstruction scheme to get fully discrete entropy inequalities* J. Sci. Comput. **80**: 924–956 doi: 10.1007/s10915-019-00961-y

[Ber20] C. Berthon, A. Duran, K. Saleh (2020) *An easy control of the artificial numerical viscosity to get discrete entropy inequalities when approximating hyperbolic systems of conservation laws* in *Continuum Mechanics, Applied Mathematics and Scientific Computing: Godunov's Legacy* 29–36 Springer doi: 10.1007/978-3-030-38870-6_5

[Bit13] M. Bittl, D. Kuzmin (2013) *An hp-adaptive flux-corrected transport algorithm for continuous finite elements* Computing **95**: 27–48 doi: 10.1007/s00607-012-0223-y

[Boo75] D. L. Book, J. P. Boris, K. Hain (1975) *Flux-corrected transport II: Generalizations of the method* J. Comput. Phys. **18**: 248–283 doi: 10.1016/0021-9991(75)90002-9

[Bor73] J. P. Boris, D. L. Book (1973) *Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works* J. Comput. Phys. **11**: 38–69 doi: 10.1016/0021-9991(73)90147-2

[Bou04] F. Bouchut (2004) *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources* Birkhäuser

[Bro82] A. N. Brooks, T. J. R. Hughes (1982) *Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations* Comput. Method. Appl. M. **32**: 199–259 doi: 10.1016/0045-7825(82)90071-8

[Bur07] E. Burman (2007) *On nonlinear artificial viscosity, discrete maximum principle and hyperbolic conservation laws* BIT Numer. Math. **47**: 715–733 doi: 10.1007/s10543-007-0147-7

[Che17] T. Chen, C.-W. Shu (2017) *Entropy stable high order discontinuous Galerkin methods with suitable quadrature rules for hyperbolic conservation laws* J. Comput. Phys. **345**: 427–461 doi: 10.1016/j.jcp.2017.05.025

[Cod93] R. Codina (1993) *A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation* Comput. Method. Appl. M. **110**: 325–342 doi: 10.1016/0045-7825(93)90213-H

[Cus11] B. Cushman-Roisin, J.-M. Beckers (2011) *Introduction to Geophysical Fluid Dynamics* Elsevier Science & Technology 2nd ed.

[Daf00] C. M. Dafermos (2000) *Hyperbolic Conservation Laws in Continuum Physics* Springer 1st ed. doi: 10.1007/978-3-662-22019-1

[Del13] O. Delestre, C. Lucas, P.-A. Ksinant, F. Darboux, C. Laguerre, T.-N.-T. Vo, T. François, S. Cordier (2013) *SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies* Int. J. Numer. Meth. Fl. **72**: 269–300 doi: 10.1002/fld.3741

[Del16] O. Delestre, C. Lucas, P.-A. Ksinant, F. Darboux, C. Laguerre, T.-N.-T. Vo, F. James, S. Cordier (2016) *SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies* Preprint, arXiv: 1110.0288v7 [math.NA]

[Dio13] S. Diot, R. Loubère, S. Clain (2013) *The Multidimensional Optimal Order Detection method in the three-dimensional case: very high-order finite volume method for hyperbolic systems* Int. J. Numer. Meth. Fl. **73**: 362–392 DOI: 10.1002/fld.3804

[DiP12] D. A. Di Pietro, A. Ern (2012) *Mathematical Aspects of Discontinuous Galerkin Methods* Springer DOI: 10.1007/978-3-642-22980-0

[Dob18] V. Dobrev, T. Kolev, D. Kuzmin, R. Rieben, V. Tomov (2018) *Sequential limiting in continuous and discontinuous Galerkin methods for the Euler equations* J. Comput. Phys. **356**: 372–390 DOI: 10.1016/j.jcp.2017.12.012

[Dol15] V. Dolejší, M. Feistauer (2015) *Discontinuous Galerkin Method* Springer DOI: 10.1007/978-3-319-19267-3

[Dum14] M. Dumbser, O. Zanotti, R. Loubère, S. Diot (2014) *A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws* J. Comput. Phys. **278**: 47–75 DOI: 10.1016/j.jcp.2014.08.009

[Eck17] C. Eck, H. Garcke, P. Knabner (2017) *Mathematical modeling* Springer 1st ed. DOI: 10.1007/978-3-319-55161-6

[Ern04] A. Ern, J.-L. Guermond (2004) *Theory and Practice of Finite Elements* Springer DOI: 10.1007/978-1-4757-4355-5

[Fei03] M. Feistauer, J. Felcman, I. Straškraba (2003) *Mathematical and Computational Methods for Compressible Flow* Oxford University Press

[Fjo09] U. S. Fjordholm (2009) *Structure preserving finite volume methods for the shallow water equations* Master's thesis University of Oslo https://www.duo.uio.no/bitstream/handle/10852/10904/1/thesis.pdf

[Fjo11] U. S. Fjordholm, S. Mishra, E. Tadmor (2011) *Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography* J. Comput. Phys. **230**: 5587–5609 DOI: 10.1016/j.jcp.2011.03.042

[Fle83] C. Fletcher (1983) *The group finite element formulation* Comput. Method. Appl. M. **37**: 225–244 DOI: 10.1016/0045-7825(83)90122-6

[Fra20] F. Frank, B. Reuter, V. Aizinger, H. Hajduk, A. Rupp (2020) *FESTUNG: The Finite Element Simulation Toolbox for UNstructured Grids, Version 1.0* https://github.com/FESTUNG

[Gas13] G. J. Gassner (2013) *A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods* SIAM J. Sci. Comput. **35**: A1233–A1253 DOI: 10.1137/120890144

[Ger00] J.-F. Gerbeau, B. Perthame (2000) *Derivation of viscous Saint-Venant system for laminar shallow water; numerical validation* Research Report INRIA RR-4084 inria–00072549 https://hal.inria.fr/inria-00072549/

[Geu09] C. Geuzaine, J.-F. Remacle (2009) *Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities* Int. J. Numer. Methods Eng. **79**: 1309–1331 DOI: 10.1002/nme.2579

[Ghi03] J.-M. Ghidaglia, F. Pascal (2003) *On boundary conditions for multidimensional hyperbolic systems of conservation laws in the finite volume framework* Report CMLA, Ens de Cachan http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.209.135&rep=rep1&type=pdf

[GLVis] *GLVis: OpenGL Finite Element Visualization Tool* https://glvis.org

[God59] S. K. Godunov (1959) *A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics* Mat. Sb. **47(89)**: 271–306 http://mi.mathnet.ru/eng/msb4873

[Got98] S. Gottlieb, C.-W. Shu (1998) *Total variation diminishing Runge-Kutta schemes* Math. Comput. **67**: 73–85 DOI: 10.1090/S0025-5718-98-00913-2

[Got01] S. Gottlieb, C.-W. Shu, E. Tadmor (2001) *Strong Stability-Preserving High-Order Time Discretization Methods* SIAM Rev. **43**: 89–112 DOI: 10.1137/S003614450036757X

[Got11] S. Gottlieb, D. Ketcheson, C.-W. Shu (2011) *Strong stability preserving Runge–Kutta and multistep time discretizations* World Scientific DOI: 10.1142/7498

[Gue14] J.-L. Guermond, M. Nazarov (2014) *A maximum-principle preserving C^0 finite element method for scalar conservation equations* Comput. Method. Appl. M. **272**: 198–213 DOI: 10.1016/j.cma.2013.12.015

[Gue16a] J.-L. Guermond, B. Popov (2016) *Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations* J. Comput. Phys. **321**: 908–926 DOI: 10.1016/j.jcp.2016.05.054

[Gue16b] J.-L. Guermond, B. Popov (2016) *Invariant domains and first-order continuous finite element approximation for hyperbolic systems* SIAM J. Numer. Anal. **54**: 2466–2489 DOI: 10.1137/16M1074291

[Gue17] J.-L. Guermond, B. Popov (2017) *Invariant domains and second-order continuous finite element approximation for scalar conservation equations* SIAM J. Numer. Anal. **55**: 3120–3146 DOI: 10.1137/16M1106560

[Gue18a] J.-L. Guermond, M. Nazarov, B. Popov, I. Tomas (2018) *Second-order invariant domain preserving approximation of the Euler equations using convex limiting* SIAM J. Sci. Comput. **40**: A3211–A3239 DOI: 10.1137/17M1149961

[Gue18b] J.-L. Guermond, M. Quezada de Luna, B. Popov, C. E. Kees, M. W. Farthing (2018) *Well-balanced second-order finite element approximation of the shallow water equations with friction* SIAM J. Sci. Comput. **40**: A3873–A3901 DOI: 10.1137/17M1156162

[Gue19] J.-L. Guermond, B. Popov, I. Tomas (2019) *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems* Comput. Method. Appl. M. **347**: 143–175 DOI: 10.1016/j.cma.2018.11.036

[Gur09] M. Gurris (2009) *Implicit finite element schemes for compressible gas and particle-laden gas flows* Ph.D. thesis TU Dortmund University

[Haj19] H. Hajduk, D. Kuzmin, V. Aizinger (2019) *New directional vector limiters for discontinuous Galerkin methods* J. Comput. Phys. **384**: 308–325 DOI: 10.1016/j.jcp.2019.01.032

[Haj20a] H. Hajduk (2020) *Preconditioned gradient matrix on the reference simplex* https://github.com/HennesHajduk/PrecMatSimplex

[Haj20b] H. Hajduk, D. Kuzmin, T. Kolev, R. Abgrall (2020) *Matrix-free subcell residual distribution for Bernstein finite element discretizations of linear advection equations* Comput. Method. Appl. M. **359**: 112658 DOI: 10.1016/j.cma.2019.112658

[Haj20c] H. Hajduk, D. Kuzmin, T. Kolev, V. Tomov, I. Tomas, J. N. Shadid (2020) *Matrix-free subcell residual distribution for Bernstein finite elements: Monolithic limiting* Comput. Fluids **200**: 104451 DOI: 10.1016/j.compfluid.2020.104451

[Haj21a] H. Hajduk (2021) *Monolithic convex limiting in discontinuous Galerkin discretizations of hyperbolic conservation laws* Comput. Math. Appl. **87**: 120–138 DOI: 10.1016/j.camwa.2021.02.012

[Haj21b] H. Hajduk, A. Rupp, D. Kuzmin (2021) *Analysis of algebraic flux correction for semi-discrete advection problems* Preprint, arXiv: 2104.05639 [math.NA]

[Har72] A. Harten, G. Zwas (1972) *Self-adjusting hybrid schemes for shock computations* J. Comput. Phys. **9**: 568–583 DOI: 10.1016/0021-9991(72)90012-5

[Har83a] A. Harten (1983) *On the symmetric form of systems of conservation laws with entropy* J. Comput. Phys. **49**: 151–164 DOI: 10.1016/0021-9991(83)90118-3

[Har83b] A. Harten, P. D. Lax, B. van Leer (1983) *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws* SIAM Rev. **25**: 35–61 DOI: 10.1137/1025002

[Har84] A. Harten (1984) *On a class of high resolution total-variation-stable finite-difference-schemes* SIAM J. Numer. Anal. **21**: 1–23 DOI: 10.1137/0721001

[Har87] A. Harten, S. Osher (1987) *Uniformly high-order accurate nonoscillatory schemes. I* SIAM J. Numer. Anal. **24**: 279–309 DOI: 10.1007/978-3-642-60543-7_11

[Hen21] S. Hennemann, A. M. Rueda-Ramírez, F. J. Hindenlang, G. J. Gassner (2021) *A provably entropy stable subcell shock capturing approach for high order split form DG for the compressible Euler equations* J. Comput. Phys. **426**: 109935 DOI: 10.1016/j.jcp.2020.109935

[Hug86] T. J. R. Hughes, M. Mallet (1986) *A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective-diffusive systems* Comput. Method. Appl. M. **58**: 329–336 DOI: 10.1016/0045-7825(86)90153-2

[Jam93] A. Jameson (1993) *Computational algorithms for aerodynamic analysis and design* Appl. Numer. Math. **13**: 383–422 DOI: 10.1016/0168-9274(93)90096-A

[Jam17] A. Jameson (2017) *Origins and further development of the Jameson–Schmidt–Turkel scheme* AIAA J. **55**: 1487–1510 DOI: 10.2514/1.J055493

[Jha21] A. Jha, N. Ahmed (2021) *Analysis of flux corrected transport schemes for evolutionary convection-diffusion-reaction equations* Preprint, arXiv: 2103.04776 [math.NA]

[Jia94] G.-S. Jiang, C.-W. Shu (1994) *On a cell entropy inequality for discontinuous Galerkin methods* Math. Comput. **62**: 531–531 DOI: 10.1090/S0025-5718-1994-1223232-7

[Jia96] G.-S. Jiang, C.-W. Shu (1996) *Efficient implementation of weighted ENO schemes* J. Comput. Phys. **126**: 202–228 DOI: 10.1006/jcph.1996.0130

[Kho94] B. Khobalatte, B. Perthame (1994) *Maximum principle on the entropy and second-order kinetic schemes* Math. Comput. **62**: 119–131 DOI: 10.1090/S0025-5718-1994-1208223-4

[Kir17] R. C. Kirby (2017) *Fast inversion of the simplicial Bernstein mass matrix* Numer. Math. **135**: 73–95 DOI: 10.1007/s00211-016-0795-0

[Kiv22] S. Kivva (2022) *Entropy stable flux correction for scalar hyperbolic conservation laws* J. Sci. Comput. **91**: 10 DOI: 10.1007/s10915-022-01792-0

[Kna03] P. Knabner, L. Angermann (2003) *Numerical Methods for Elliptic and Parabolic Partial Differential Equations* Springer DOI: 10.1007/b97419

[Krö94] D. Kröner, M. Rokyta (1994) *Convergence of upwind finite volume schemes for scalar conservation laws in two dimensions* SIAM J. Numer. Anal. **31**: 324–343 DOI: 10.1137/0731017

[Kuč18] V. Kučera, C.-W. Shu (2018) *On the time growth of the error of the DG method for advective problems* IMA J. Numer. Anal. **39**: 687–712 DOI: 10.1093/imanum/dry013

[Kur00] A. Kurganov, E. Tadmor (2000) *New high-resolution central schemes for nonlinear conservation laws and convection–diffusion equations* J. Comput. Phys. **160**: 241–282 DOI: 10.1006/jcph.2000.6459

[Kur07a] A. Kurganov, G. Petrova (2007) *A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system* Commun. Math. Sci. **5**: 133–160 DOI: 10.4310/CMS.2007.v5.n1.a6

[Kur07b] A. Kurganov, G. Petrova, B. Popov (2007) *Adaptive semidiscrete central-upwind schemes for nonconvex hyperbolic conservation laws* SIAM J. Sci. Comput. **29**: 2381–2401 DOI: 10.1137/040614189

[Kuz02] D. Kuzmin, S. Turek (2002) *Flux correction tools for finite elements* J. Comput. Phys. **175**: 525–558 DOI: 10.1006/jcph.2001.6955

[Kuz05] D. Kuzmin, M. Möller (2005) *Algebraic Flux Correction II. Compressible Euler Equations* in *Flux-Corrected Transport: Principles, Algorithms, and Applications* 145–192 Springer DOI: 10.1007/3-540-27206-2_7

[Kuz08] D. Kuzmin (2008) *On the design of algebraic flux correction schemes for quadratic finite elements* J. Comput. Appl. Math. **218**: 79–87 DOI: 10.1016/j.cam.2007.04.045

[Kuz10a] D. Kuzmin (2010) *A vertex-based hierarchical slope limiter for p-adaptive discontinuous Galerkin methods* J. Comput. Appl. Math. **233**: 3077–3085 DOI: 10.1016/j.cam.2009.05.028

[Kuz10b] D. Kuzmin, M. Möller, J. N. Shadid, M. Shashkov (2010) *Failsafe flux limiting and constrained data projections for equations of gas dynamics* J. Comput. Phys. **229**: 8766–8779 DOI: 10.1016/j.jcp.2010.08.009

[Kuz12a] D. Kuzmin (2012) *Algebraic flux correction I. Scalar conservation laws* in *Flux-Corrected Transport: Principles, Algorithms, and Applications* 145–192 Springer 2nd ed. DOI: 10.1007/978-94-007-4038-9_6

[Kuz12b] D. Kuzmin, R. Löhner, S. Turek (2012) *Flux-Corrected Transport: Principles, Algorithms, and Applications* Springer 2nd ed. DOI: 10.1007/978-94-007-4038-9

[Kuz20a] D. Kuzmin (2020) *Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws* Comput. Method. Appl. M. **361**: 112804 DOI: 10.1016/j.cma.2019.112804

[Kuz20b] D. Kuzmin, H. Hajduk, A. Rupp (2020) *Locally bound-preserving enriched Galerkin methods for the linear advection equation* Comput. Fluids **205**: 104525 DOI: 10.1016/j.compfluid.2020.104525

[Kuz20c] D. Kuzmin, M. Quezada de Luna (2020) *Algebraic entropy fixes and convex limiting for continuous finite element discretizations of scalar hyperbolic conservation laws* Comput. Method. Appl. M. **372**: 113370 DOI: 10.1016/j.cma.2020.113370

[Kuz20d] D. Kuzmin, M. Quezada de Luna (2020) *Entropy conservation property and entropy stabilization of high-order continuous Galerkin approximations to scalar conservation laws* Comput. Fluids **213**: 104742 DOI: 10.1016/j.compfluid.2020.104742

[Kuz20e] D. Kuzmin, M. Quezada de Luna (2020) *Subcell flux limiting for high-order Bernstein finite element discretizations of scalar hyperbolic conservation laws* J. Comput. Phys. **411**: 109411 DOI: 10.1016/j.jcp.2020.109411

[Kuz22a] D. Kuzmin, H. Hajduk, A. Rupp (2022) *Limiter-based entropy stabilization of semi-discrete and fully discrete schemes for nonlinear hyperbolic problems* Comput. Method. Appl. M. **389**: 114428 DOI: 10.1016/j.cma.2021.114428

[Kuz22b] D. Kuzmin, M. Quezada de Luna, D. I. Ketcheson, J. Grüll (2022) *Bound-preserving flux limiting for high-order explicit Runge–Kutta time discretizations of hyperbolic conservation laws* J. Sci. Comput. **91**: 21 DOI: 10.1007/s10915-022-01784-0

[Lai07] M. Lai, L. Schumaker (2007) *Spline Functions on Triangulations* Cambridge University Press DOI: 10.1017/CBO9780511721588.003

[LeV92] R. J. LeVeque (1992) *Numerical methods for conservation laws* Birkhäuser DOI: 10.1007/978-3-0348-8629-1

[LeV96] R. J. LeVeque (1996) *High-resolution conservative algorithms for advection in incompressible flow* SIAM J. Numer. Anal. **33**: 627–665 DOI: 10.1137/0733033

[LeV02] R. J. LeVeque (2002) *Finite Volume Methods for Hyperbolic Problems* Cambridge University Press DOI: 10.1017/CBO9780511791253

[Lia09] Q. Liang, F. Marche (2009) *Numerical resolution of well-balanced shallow water equations with complex source terms* Adv. Water Resour. **32**: 873–884 DOI: 10.1016/j.advwatres.2009.02.010

[Lin22] Y. Lin, J. Chan, I. Tomas (2022) *A positivity preserving strategy for entropy stable discontinuous Galerkin discretizations of the compressible Euler and Navier-Stokes equations* Preprint, arXiv: 2201.11816 [math.NA]

[Liu94] X.-D. Liu, S. Osher, T. Chan (1994) *Weighted Essentially Non-Oscillatory Schemes* J. Comput. Phys. **115**: 200–212 DOI: 10.1006/jcph.1994.1187

[Loh16] C. Lohmann, D. Kuzmin (2016) *Synchronized flux limiting for gas dynamics variables* J. Comput. Phys. **326**: 973–990 DOI: 10.1016/j.jcp.2016.09.025

[Loh17a] C. Lohmann (2017) *Eigenvalue range limiters for tensors in flux-corrected transport algorithms* MultiMat, September 18–22, 2017, Santa Fe, USA https://custom.cvent.com/F6288ADDEF3C4A6CBA5358DAE922C966/files/e4c3aedf74394eb1a33e141f57f33b2e.pdf

[Loh17b] C. Lohmann, D. Kuzmin, J. N. Shadid, S. Mabuza (2017) *Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements* J. Comput. Phys. **344**: 151–186 DOI: 10.1016/j.jcp.2017.04.059

[Loh19] C. Lohmann (2019) *Physics-Compatible Finite Element Methods for Scalar and Tensorial Advection Problems* Springer Spektrum DOI: 10.1007/978-3-658-27737-6

[Loh21] C. Lohmann (2021) *An algebraic flux correction scheme facilitating the use of Newton-like solution strategies* Comput. Math. Appl. **84**: 56–76 DOI: 10.1016/j.camwa.2020.12.010

[Löh87] R. Löhner, K. Morgan, J. Peraire, M. Vahdati (1987) *Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier–Stokes equations* Int. J. Numer. Meth. Fl. **7**: 1093–1109 DOI: 10.1002/fld.1650071007

[Löh08] R. Löhner (2008) *Applied Computational Fluid Dynamics Techniques: An Introduction Based on Finite Element Methods* John Wiley & Sons DOI: 10.1002/9780470989746

[Lyc00] T. Lyche, K. Scherer (2000) *On the p-norm condition number of the multivariate triangular Bernstein basis* J. Comput. Appl. Math. **119**: 259–273 DOI: 10.1016/S0377-0427(00)00383-6

[Mar07] S. Martin (2007) *First order quasilinear equations with boundary conditions in the L^∞ framework* J. Differ. Equ. **236**: 375–406 DOI: 10.1016/j.jde.2007.02.007

[Matlab] *MATLAB* The MathWorks Inc. https://mathworks.com/products/matlab

[MFEM] *MFEM: Modular Finite Element Methods [Software]* https://mfem.org

[Moe17] S. A. Moe, J. A. Rossmanith, D. C. Seal (2017) *Positivity-preserving discontinuous Galerkin methods with Lax–Wendroff time discretizations* J. Sci. Comput. **71**: 44–70 DOI: 10.1007/s10915-016-0291-9

[Möl08] M. Möller (2008) *Adaptive high-resolution finite element schemes* Ph.D. thesis TU Dortmund University http://hdl.handle.net/2003/25933

[Noe07] S. Noelle, Y. Xing, C.-W. Shu (2007) *High-order well-balanced finite volume WENO schemes for shallow water equation with moving water* J. Comput. Phys. **226**: 29–58 DOI: 10.1016/j.jcp.2007.03.031

[Pan94] E. Y. Panov (1994) *Uniqueness of the solution of the Cauchy problem for a first order quasilinear equation with one admissible strictly convex entropy* Math. Notes **55**: 517–525 DOI: 10.1007/BF02110380

[Paz19] W. Pazner, P.-O. Persson (2019) *Analysis and Entropy Stability of the Line-Based Discontinuous Galerkin Method* J. Sci. Comput. **80**: 376–402 DOI: 10.1007/s10915-019-00942-1

[Paz21] W. Pazner (2021) *Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting* Comput. Method. Appl. M. **382**: 113876 doi: 10.1016/j.cma.2021.113876

[Qua94] A. Quarteroni, A. Valli (1994) *Numerical Approximation of Partial Differential Equations* Springer doi: 10.1007/978-3-540-85268-1

[Que21] M. Quezada de Luna, D. I. Ketcheson (2021) *Maximum principle preserving space and time flux limiting for Diagonally Implicit Runge–Kutta discretizations of scalar convection-diffusion equations* Preprint, arXiv: 2109.08272 [math.NA]

[Reu21] B. Reuter, H. Hajduk, A. Rupp, F. Frank, V. Aizinger, P. Knabner (2021) *FESTUNG 1.0: Overview, usage, and example applications of the MATLAB / GNU Octave toolbox for discontinuous Galerkin methods* Comput. Math. Appl. **81**: 3–41 doi: 10.1016/j.camwa.2020.08.018

[Ric09] M. Ricchiuto, A. Bollermann (2009) *Stabilized residual distribution for shallow water simulations* J. Comput. Phys. **228**: 1071–1115 doi: 10.1016/j.jcp.2008.10.020

[Rue22] A. M. Rueda-Ramírez, W. Pazner, G. J. Gassner (2022) *Subcell limiting strategies for discontinuous Galerkin spectral element methods* Preprint, arXiv: 2202.00576 [math.NA]

[Rup21] A. Rupp, M. Hauck, V. Aizinger (2021) *A subcell-enriched Galerkin method for advection problems* Comput. Math. Appl. **93**: 120–129 doi: 10.1016/j.camwa.2021.04.010

[Ruu02] S. J. Ruuth, R. J. Spiteri (2002) *Two barriers on strong-stability-preserving time discretization methods* J. Sci. Comput. **17**: 211–220 doi: 10.1023/A:1015156832269

[Sch85] M. E. Schonbek (1985) *Second-order conservative schemes and the entropy condition* Math. Comput. **44**: 31–38 doi: 10.1090/S0025-5718-1985-0771028-7

[Sch17] H. Schlichting, K. Gersten (2017) *Boundary-Layer Theory* Springer 9th ed. doi: 10.1007/978-3-662-52919-5

[Sel93] V. Selmin (1993) *The node-centred finite volume approach: Bridge between finite differences and finite elements* Comput. Methods Appl. Mech. Engrg. **102**: 107–138 doi: 10.1016/0045-7825(93)90143-L

[Sel96] V. Selmin, L. Formaggia (1996) *Unified construction of finite element and finite volume discretizations for compressible flows* Int. J. Numer. Methods Eng. **39**: 1–32 DOI: 10.1002/(SICI)1097-0207(19960115)39:1<1::AID-NME837>3.0.CO;2-G

[Shu88] C.-W. Shu, S. Osher (1988) *Efficient implementation of essentially non-oscillatory shock-capturing schemes* J. Comput. Phys. **77**: 439–471 DOI: 10.1016/0021-9991(88)90177-5

[Shu98] C.-W. Shu (1998) *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws* in *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations* Lecture Notes in Mathematics 325–432 Springer DOI: 10.1007/BFb0096355

[Sod78] G. A. Sod (1978) *A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws* J. Comput. Phys. **27**: 1–31 DOI: 10.1016/0021-9991(78)90023-2

[Swe84] P. K. Sweby (1984) *High resolution schemes using flux limiters for hyperbolic conservation laws* SIAM J. Numer. Anal. **21**: 995–1011 DOI: 10.1137/0721062

[Tad86] E. Tadmor (1986) *A minimum entropy principle in the gas dynamics equations* Appl. Numer. Math. **2**: 211–219 DOI: 10.1016/0168-9274(86)90029-2

[Tad87] E. Tadmor (1987) *The numerical viscosity of entropy stable schemes for systems of conservation laws. I* Math. Comput. **49**: 91–103 DOI: 10.1090/S0025-5718-1987-0890255-3

[Tad03] E. Tadmor (2003) *Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems* Acta Numer. **12**: 451–512 DOI: 10.1017/S0962492902000156

[Tha81] W. Thacker (1981) *Some exact solutions to the nonlinear shallow-water wave equations* J. Fluid Mech. **107**: 499–508 DOI: 10.1017/S0022112081001882

[Tho16] T. Thompson (2016) *A discrete commutator theory for the consistency and phase error analysis of semi-discrete C^0 finite element approximations to the linear transport equation* J. Comput. Appl. Math. **303**: 229–248 DOI: 10.1016/j.cam.2016.02.042

[Tor09] E. F. Toro (2009) *Riemann Solvers and Numerical Methods for Fluid Dynamics* Springer 3rd ed. DOI: 10.1007/b79761

[Vat15] S. Vater, N. Beisiegel, J. Behrens (2015) *A limiter-based well-balanced discontinuous Galerkin method for shallow-water flows with wetting and drying: One-dimensional case* Adv. Water Resour. **85**: 1–13 DOI: 10.1016/j.advwatres.2015.08.008

[Vaz99] M. E. Vázquez-Cendón (1999) *Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry* J. Comput. Phys. **148**: 497–526 DOI: 10.1006/jcph.1998.6127

[Vre94] C. B. Vreugdenhil (1994) *Numerical Methods for Shallow-Water Flow* Springer Science+Business Media DOI: 10.1007/978-94-015-8354-1

[Win15] A. R. Winters, G. J. Gassner (2015) *A comparison of two entropy stable discontinuous Galerkin spectral element approximations for the shallow water equations with non-constant topography* J. Comput. Phys. **301**: 357–376 DOI: 10.1016/j.jcp.2015.08.034

[Win17] N. Wintermeyer, A. R. Winters, G. J. Gassner, D. A. Kopriva (2017) *An entropy stable nodal discontinuous Galerkin method for the two dimensional shallow water equations on unstructured curvilinear meshes with discontinuous bathymetry* J. Comput. Phys. **340**: 200–242 DOI: 10.1016/j.jcp.2017.03.036

[Woo84] P. Woodward, P. Colella (1984) *The numerical simulation of two-dimensional fluid flow with strong shocks* J. Comput. Phys. **54**: 115–173 DOI: 10.1016/0021-9991(84)90142-6

[Wu21] X. Wu, N. Trask, J. Chan (2021) *Entropy stable discontinuous Galerkin methods for the shallow water equations with subcell positivity preservation* Preprint, arXiv: 2112.07749 [math.NA]

[Zal79] S. T. Zalesak (1979) *Fully multidimensional flux-corrected transport algorithms for fluids* J. Comput. Phys. **31**: 335–362 DOI: 10.1016/0021-9991(79)90051-2

[Zha11] X. Zhang, C.-W. Shu (2011) *Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments* Proc. R. Soc. A **467**: 2752–2776 DOI: 10.1098/rspa.2011.0153

[Zha17] X. Zhang (2017) *On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier–Stokes equations* J. Comput. Phys. **328**: 301–343 DOI: 10.1016/j.jcp.2016.10.002

[Zie95] O. C. Zienkiewicz, P. Ortiz (1995) *A split-characteristic based finite element model for the shallow water equations* Int. J. Numer. Meth. Fl. **20**: 1061–1080 DOI: 10.1002/fld.1650200823