Doctoral Thesis

Compact third-order limiter functions for finite volume methods

Author(s):
Čada, Miroslav

Publication Date:
2009

Permanent Link:
https://doi.org/10.3929/ethz-a-005907275

Rights / License:
In Copyright - Non-Commercial Use Permitted
Compact Third-Order Limiter Functions for Finite Volume Methods

A dissertation submitted to the
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
MIROSLAV ČADA
Dipl. Geo Physiker, Ludwig-Maximilians University, Munich
born January 14, 1976
citizen of Slovak Republic

accepted on the recommendation of
Prof. Dr. Manuel Torrilhon, examiner
Prof. Dr. Rolf Jeltsch, co-examiner

2009
Acknowledgment

This research work has been carried out during my employment as a research and teaching assistant in the Seminar for Applied Mathematics (SAM) at the ETH Zurich. This project was initiated and supervised by Prof. Dr. Manuel Torrilhon and Prof. Dr. Rolf Jeltsch.

I would like to thank my supervisors for the opportunity to undertake research in such a pleasant and engaging work environment. The opportunities and support for travel for various conference participation they provided, enabled me to extend my professional horizon as a researcher. I gladly acknowledge Manuel Torrilhon for his continuous suggestions, support, comments and continued interest in this research project.

My special thanks goes to all my colleagues of SAM. I enjoyed the cooperation and discussions and the collective exam correction periods. I wish to thank, Paolo, Daniel, Christian, Marcel, Bastian, Claude, Roman, Norbert, Christoph, Harish, Vincent, Robert, Holger, Kersten, Sibylle, Gisela, Lukas and Patrick.

Finally, I would like to thank all my friends, my family and especially Caroline. This work could not have been done without her continuous support, encouragement and understanding.
Abstract

In this thesis we consider finite volume methods for the numerical solution of conservation laws modeled by systems of nonlinear hyperbolic partial differential equations. We are in particular interested in high-order accurate numerical approximations of conservation laws.

Numerical approximations of nonlinear hyperbolic partial differential equation are particularly challenging, since their solution develops jump discontinuities in finite time. In order to achieve high-order accurate numerical approximation to nonlinear smooth functions, we introduce new high-order accurate reconstruction functions for the spatial approximation of hyperbolic partial differential equations. We therefore employ and generalize the idea of non-oscillatory double-logarithmic reconstruction of Artebrant and Schroll [5]. The proposed spatial approximation is based upon a new class of limiter functions, which recover efficiently and with high-order accuracy smooth, as well as discontinuous functions.

The result is a three-point accurate scheme with a compact three point stencil. The interface values between two neighboring cells are obtained by a single nonlinear limiter function. In contrast to classical limiter functions, the new limiters handle discontinuities as well as local extrema within the standard semi-discrete TVD-MUSCL [77] framework utilizing an explicit SSP Runge-Kutta time marching scheme [18]. The shape preserving properties of the reconstruction are significantly improved, resulting in sharp, accurate and symmetric shock capturing.

An extensive analysis of existing limiters, reveals the major drawbacks associated with second-order accurate TVD reconstructions. Smearing of discontinuities, clipping and squaring effects of smooth extrema associated with classical second-order TVD limiters are examined. Stability analysis confirms, that squaring effects, originally associated with compressive limiter are essentially local instabilities, which trigger exponential growth of certain wave modes. The proposed new class of third-order accurate limiter completely avoids these defects, producing perfectly symmetric results.

Although the new algorithm employs a three-stage, third-order accurate time marching scheme, its computational efficiency is enhanced due to large allowable Courant number $\nu$. As indicated by the von Neumann stability analysis, large time steps, satisfying $\nu \lesssim 1.6$ are admissible and in some cases even preferable. Utilizing third-order time integration, the dissipation of high frequency waves gets reduced with large Courant numbers $1.0 \leq \nu \lesssim 1.6$. Furthermore third-order accurate schemes get unstable for
high frequency waves for $\nu \lesssim 1.6$, but remain stable for small frequencies. In contrast to second-order accurate schemes, where basically all Fourier modes get unstable, resulting eventually in wave dispersion.

To test the proposed scheme and to verify the theoretical analysis, we conducted a series of numerical experiments for a variety of hyperbolic partial differential equations. We calculated numerical approximation for the linear advection equation, the nonlinear Burger’s equation, the nonlinear Buckley–Leverett equation, the Euler equation (in 1d and 2d) of ideal gas dynamics and the ideal magneto-hydrodynamic-dynamic equations. Furthermore we also considered balance equations with stiff relaxation sources. We tested the scheme on the shallow water equations, the Broadwell system of kinetic gas theory and Grad’s 13 moment system of extended thermodynamics of rarefied gas dynamics.

The proposed scheme was compared to several well established high-order accurate methods, such as Artebrant and Schroll’s LDLR [5], Marquina’s LHHR [42], ENO / WENO [60] and several TVD schemes. The results indicate the superiority of the proposed limiter over classical second-order TVD limiters and their modifications. The new scheme also compares favorably with third-order methods such as LDLR, LHHR and ENO / WENO. A major part of the results of this work can be found in condensed form in [78–80].
Zusammenfassung

In dieser Arbeit betrachten wir Finite Volumen Verfahren für die numerische Lösung von Erhaltungsgleichungen. Erhaltungsgleichungen werden durch Systeme von nichtlinearen hyperbolischen partiellen Differentialgleichungen modelliert. Wir sind insbesondere an Verfahren interessiert, die Erhaltungsgleichungen mit hoher Genauigkeit approximieren.


Obwohl wir für die Zeitintegration ein drei stufiges Runge-Kutta Verfahren verwenden, wird die Kon-
vergenzgeschwindigkeit des Verfahrens durch grosse Courant-Zahlen \( \nu \) verbessert. Die von Neumann-Stabilitäts-Analyse deutet an, dass grosse Zeitschritte \( \nu \lesssim 1.6 \) zulässig und in einigen Fällen sogar wünschenswert sind. Bei der Verwendung von Integrationsverfahren mit dritter Ordnung Genauigkeit im der Zeit, wird die Dissipation von hochfrequenten Wellen bei grossen Courant-Zahlen \( 1 \leq \nu \lesssim 1.6 \) reduziert. Im Gegensatz zu Verfahren zweiter Ordnung räumliche und zeitlicher Genauigkeit, sind Verfahren dritter Ordnung dissipativ. Das führt bei Instabilität, \( \nu > 1.6 \), zur Verstärkung von hochfrequenten, aber nicht niederfrequenten Wellen.


Contents

Aknowledgment iii
Abstract vii
Zusammenfassung viii
List of Figures xv

1 Numerical Methods for Nonlinear Conservation Laws 1
  1.1 Nonlinear Conservation Laws ................................................. 1
  1.2 Finite Volume Methods .......................................................... 2
  1.3 Godunov’s Method ................................................................. 4
  1.4 Thesis Outline ................................................................. 6

2 High-Order Shock Capturing Methods 9
  2.1 Different Concepts of High-Order FV Schemes ............................... 9
  2.2 Spatial Reconstruction .......................................................... 12

3 New Perspective on van Leer’s MUSCL Scheme 15
  3.1 Limiter Functions ................................................................. 15
  3.2 Boundedness Criteria ........................................................... 16
  3.3 Local Truncation Error Analysis ............................................. 19
    3.3.1 Monotone Initial Data .................................................... 20
    3.3.2 Initial Data With Local Extrema ....................................... 23

4 New Third-Order Accurate Limiters 27
List of Figures

3.1 Monotonicity regions. ................................................................. 18
3.2 Limiter plots in Sweby’s second-order limiter region. ....................... 23

4.1 Shape of the logarithmic limiter. .................................................... 31
4.2 Piecewise linear third-order limiter. ................................................. 32
4.3 Plot of error constants for TVD second-order accurate limiter. .............. 35
4.4 Plot of error constants of three different version of \( \Phi(\theta, \alpha, 2, 2) \). ................................................................. 37
4.5 Plot of smooth limiter functions. ...................................................... 38
4.6 Plot of modified smooth limiter functions. ....................................... 38
4.7 Plot of error constants for different versions of the smooth Koren-limiter. .... 39
4.8 Plot of error constants for different versions of the GPR-1/3 limiter. .......... 40
4.9 Plot of error constants for different versions of the logarithmic limiter. .... 40
4.10 Sketch of a smooth initial profile and its discrete representation at different times, i.e., different locations on a uniform grid. ......................................................... 40
4.11 Sketch of new limiter function including the asymptotic region. ............ 43

6.1 Stability region of explicit second-order Runge-Kutta scheme. .............. 55
6.2 Stability region of explicit three-stage and four-stage third-order Runge-Kutta scheme. .... 55
6.3 Stability region of explicit second-order Heun scheme. .......................... 56
6.4 Dispersion relation for various spatial reconstructions. .......................... 57
6.5 Linear advection of a square wave. ................................................. 59

7.1 Double logarithmic plots of the error vs. number of grid-cells for the advection equation for the advection equation with \( u_0(x) = \sin(\pi x) \). ................................................................. 62
7.2 Solution of the advection equation with a square wave (left) and a sin-wave (right) as initial conditions. ................................................................. 62
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3</td>
<td>Double logarithmic plots of the error vs. number of grid-cells for the advection equation with $u_0(x) = \sin^4(\pi x)$.</td>
</tr>
<tr>
<td>7.4</td>
<td>Advection of discontinuities: Minmod limiter, van Leer’s smooth limiter, superbee limiter and MCD limiter.</td>
</tr>
<tr>
<td>7.5</td>
<td>Advection of discontinuities: ModGPR-1/3, $\Pi_{GPR-1/3}$, log-limiter, $\phi_{test}$ and with Koren-limiter.</td>
</tr>
<tr>
<td>7.6</td>
<td>Advection of discontinuities: LimO3, ENO3, LDLR and LHHR.</td>
</tr>
<tr>
<td>7.7</td>
<td>Total variation of Burger’s equation.</td>
</tr>
<tr>
<td>7.8</td>
<td>Total variation of the Buckley–Leverett equation.</td>
</tr>
<tr>
<td>9.1</td>
<td>EOC density profile of a non-linear plane wave for the Euler equations.</td>
</tr>
<tr>
<td>9.2</td>
<td>EOC density profile of a non-linear plane wave for the Euler equations with different Courant numbers.</td>
</tr>
<tr>
<td>9.3</td>
<td>Density profile of shock-blast interaction problem for Euler equations.</td>
</tr>
<tr>
<td>9.4</td>
<td>Energy and pressure distribution of shock-blast interaction problem for Euler equations.</td>
</tr>
<tr>
<td>9.5</td>
<td>Density distribution of Shu–Osher shock-acoustic test case for Euler equations.</td>
</tr>
<tr>
<td>9.6</td>
<td>Density distribution of Shu–Osher shock-acoustic test case for Euler equations.</td>
</tr>
<tr>
<td>9.7</td>
<td>Density profile of shock-blast interaction problem for Euler equations.</td>
</tr>
<tr>
<td>9.8</td>
<td>Density distribution of Shu–Osher shock-acoustic test case for Euler equations.</td>
</tr>
<tr>
<td>9.9</td>
<td>EOC of LimO3 for the 1D MHD-system.</td>
</tr>
<tr>
<td>9.10</td>
<td>EOC of LDLR for the 1D MHD-system.</td>
</tr>
<tr>
<td>9.11</td>
<td>Convergence studies of the almost co-planar MHD Riemann problem.</td>
</tr>
<tr>
<td>10.1</td>
<td>EOC density profile for the 2d Euler equations.</td>
</tr>
<tr>
<td>10.2</td>
<td>Shock-Bubble experimental setup for 2D Euler equation.</td>
</tr>
<tr>
<td>10.3</td>
<td>Emulated Schlieren pictures of shock-bubble simulation.</td>
</tr>
<tr>
<td>10.4</td>
<td>Zoom of the shock-bubble Schlieren image.</td>
</tr>
<tr>
<td>10.5</td>
<td>Zoom of the shock-bubble Schlieren images.</td>
</tr>
<tr>
<td>10.6</td>
<td>Euler-Four-Shocks problem: van Leer’s smooth limiter.</td>
</tr>
<tr>
<td>10.7</td>
<td>Euler-Four-Shocks problem: LimO3, Koren-limiter.</td>
</tr>
<tr>
<td>10.8</td>
<td>Euler-Four-Contacts problem: van Leer’s smooth limiter.</td>
</tr>
<tr>
<td>10.9</td>
<td>Euler-Four-Contacts problem: LimO3, Koren-limiter.</td>
</tr>
</tbody>
</table>
10.10 Euler double Mach reflection experiment: van Leer’s smooth limiter, LimO3. ..... 101

11.1 Comparison of numerical solution of the shallow water model. .................. 108
11.2 Numerical solution of the Broadwell model: Plot of density and momentum. .... 109
11.3 Numerical solution of the Broadwell model: Plot of the variable $Z$ and $e$. ....... 110
11.4 Numerical solution of Grad’s 13 moment equation: Plot of the density field and velocity field. ................................................................. 112
11.5 Numerical solution of Grad’s 13 moment equation: Plot of the stress field and heat flux field. ................................................................. 112
Chapter 1

Numerical Methods for Nonlinear Conservation Laws

In the present work we are interested in high-order accurate numerical approximation of the solution of systems of conservation laws. Solutions to conservation laws have typically smooth structures interspersed with discontinuities. An accurate prediction of such interactions is of importance in many computational fluid dynamic (CFD) applications, such as aircraft design, stellar formation, and weather simulations, to name only a few. The main task is to develop algorithms, that are highly accurate for smooth regions in both time and space, and that have sharp transition where large gradients or discontinuities appear.

In this introductory chapter we settle the terms and definitions, commonly used in the literature of computational methods for hyperbolic partial differential equations (PDE). The aim is to become familiar with the notation and to understand the challenge of constructing numerical methods for conservation laws. Although conservation laws are mostly studied for systems, numerical schemes are generally derived for scalar linear equations. Therefore, without loss of generality, we reduce most of our explanations to scalar conservation laws.

1.1 Nonlinear Conservation Laws

Conservation laws describe the continuum dynamics of a physical system. They are balance equations which describe the relation between the rate of change of a physical quantity in a bounded domain and its convective or diffusive flux through the boundary of the domain. In their integral form, the conservation law for a spatial domain $\Omega$ and a quantity $u(x,t)$ is

$$
\frac{d}{dt} \int_{\Omega} u(x,t) \, dx + \int_{\partial \Omega} f(u(x,t)) \cdot n \, dS = \int_{\Omega} s(u(x,t), x,t) \, dV.
$$

(1.1)
The volume of \( u(x,t) : \mathbb{R}^d \times \mathbb{R}^+ \rightarrow \mathbb{R}^p \) can only change in time by the dynamics of the flux \( f(u(x,t)) : \mathbb{R}^p \rightarrow \mathbb{R}^p \) across the boundary \( \partial \Omega \) and by productive / dissipative processes associated with the source term \( s(u(x,t)) : \mathbb{R}^p \rightarrow \mathbb{R}^p \).

Assuming \( \Omega \) to be an infinitesimal volume and applying Gauss’s theorem, we obtain the \textit{differential form} of the conservation law

\[
\partial_t u(x,t) + \nabla \cdot f(u(x,t)) = s(u(x,t)) \quad \text{in} \quad \mathbb{R}^d \times \mathbb{R}^+ \tag{1.2}
\]

\[
u(x,t = 0) = u_0(x) \quad \text{in} \quad \mathbb{R}^d, \tag{1.3}
\]

where the initial data \( u_0(x) : \mathbb{R}^d \rightarrow \mathbb{R}^p \), is either a piecewise smooth function with compact support or a periodic function. This Cauchy initial value problem is a typical conservation law, characteristic for mathematical modeling in continuum physics. In the following we will mainly be concerned with the approximation of the spatial operator \( \nabla \cdot f(u(x,t)) \) and the time discretization \( \partial_t u(x,t) \). For simplicity we will drop the source term \( s(u(x,t)) \) in further considerations. In Chapter 11 we will deal with balance laws with stiff source terms. There we will explicitly utilize the system (1.2).

A system of conservation laws (1.2) is called \textit{hyperbolic} if the Jacobian matrix \( A = \nabla_u f(u) \), \( A \in \mathbb{R}^{p \times p} \) has real eigenvalues and is diagonalizable, i.e., its eigenvectors span \( \mathbb{R}^p \). The system (1.2) is \textit{linear} if the Jacobian does not depend on the values \( u(x,t) \), otherwise it is called \textit{quasilinear} or in the context of conservation laws nonlinear.

Although in most practical applications systems of nonlinear conservation laws are solved, numerical methods are typically studied using simpler model equations. To analyze the accuracy and stability of a scheme, one usually uses the scalar one-dimensional linear advection equation

\[
u_t + a \nu_x = 0, \quad \text{with} \quad a = \text{const}, \tag{1.4}
\]

which has the analytical solution

\[
u(x,t) = u_0(x - a t). \tag{1.5}
\]

This equation will serve for theoretical, as well as numerical calculations throughout this thesis.

### 1.2 Finite Volume Methods

The solution of a linear hyperbolic system maintains the regularity of the initial data for all times. We are interested, however, in nonlinear conservation laws. An essential feature of the solution of nonlinear hyperbolic equations is that, in general, their solution lose regularity of the initial data after finite time. Gradients of \( u(x,t) \) can blow up in finite time, even when the initial data is arbitrary smooth. Beyond some critical time, the \textit{strong} solution ceases to exist and solutions in the \textit{weak} sense must be considered. We do not intend to explain the different notions of solutions in detail. We referred to classical textbooks,
1.2. Finite Volume Methods

E.g., [16, 32]. Yet it should be made clear the solutions calculated numerically with a finite volume (FV) method are, in general, weak solutions.

We introduce the grid points \( x_i = i \Delta x, \) \( i = \ldots, -1, 0, 1, \ldots \) to be the discretization of the spatial domain, with the uniform step size \( \Delta x = x_{i+1/2} - x_{i-1/2} \) for all \( i \). The time is also uniformly discretized with time levels \( t^n = n \Delta t \) for \( n = 0, 1, 2, \ldots \), where \( \Delta t = t^{n+1} - t^n \) represents the temporal step size. The uniform computational region is partitioned by finite volumes \( C^n_i = [x_{i-1/2}, x_{i+1/2}] \times [t^n, t^{n+1}] \).

The grid cells \( C_i \) are the discrete representations of the infinitesimal volumes \( \Omega \).

Integrating the conservation law (1.2) over the control volume \( C^n_i \), we obtain the standard fully-discrete finite volume scheme

\[
\bar{u}^{n+1}_i = \bar{u}^n_i + \frac{\Delta t}{\Delta x} \left( F^n_{i-1/2} - F^n_{i+1/2} \right) \tag{1.6}
\]

for the cell average \( \bar{u}^n_i \) at time \( t^n \)

\[
\bar{u}^n_i = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} u^n(x) \, dx, \tag{1.7}
\]

and the interface flux \( F^n_{i+1/2} \)

\[
F^n_{i+1/2} = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(u(x_{i+1/2}), t) \, dt. \tag{1.8}
\]

A different formulation can be obtained by integrating only with respect to the spatial variable, i.e., over the cell \( C_i = [x_{i-1/2}, x_{i+1/2}] \). Then we obtain the semi-discrete finite volume scheme

\[
\frac{d}{dt} \bar{u}_i = L_i(\bar{u}^n) = \frac{1}{\Delta x} \left( F^n_{i-1/2} - F^n_{i+1/2} \right). \tag{1.9}
\]

In the semi-discrete formulation, the flux function \( F^n_{i+1/2} \) represents an approximation of the flux value \( f(u(x_{i+1/2}), t^n) \) at one point \( n \) in time. In contrast, the interface flux \( F^n_{i+1/2} \) in the fully discretized formulation is an approximation to a time average of the flux value \( f(u(x_{i+1/2}), t) \) (see eq. (1.8)).

The semi-discrete FV scheme (1.9) yields a system of ordinary differential equations (ODE)s, which can be marched forward in time by employing a variety of explicit or implicit time integration routines. This approach is also called the method of lines. We will discuss the differences of both schemes in the upcoming chapter, especially when considering high-order accurate methods.

At this point we would like to introduce the following concept:

**Definition 1.2.1 (Conservative Form)** A numerical method of the form (1.6) or (1.9) is conservative if the flux function \( F^n_{i+1/2} \) or \( F^n_{i+1/2} \) is a function of the neighboring values

\[
F^n_{i+1/2} = \mathcal{F}(\bar{u}_{i-L}, \cdots, \bar{u}_i, \cdots, \bar{u}_{i+R}), \quad \text{with} \quad L, R > 0 \tag{1.10}
\]
where \( F \) is called the numerical flux, which is Lipschitz continuous and consistent with the physical flux \( f(u) \)

\[
F(u, \cdots, u) = f(u).
\]  

(1.11)

If the numerical scheme can be written in this form, the method mimics the true solution, i.e., it converges to a weak solution of the conservation law (Theorem of Lax and Wendroff (1960) [34]). Lipschitz continuity of the numerical flux \( F \) is essential for the existence of weak solutions. An important class of weak solutions, considered throughout this thesis, are approximations of piecewise smooth functions, which contain jump discontinuities. These approximations are piecewise classical solutions with discontinuities separating the smooth regions. They are weak solutions if and only if the Rankine-Hugoniot jump conditions are satisfied,

\[
[u]s = [f(u)],
\]  

(1.12)

where \( s \) is the characteristic speed of the discontinuity, and

\[
[u] = u^{(+)} - u^{(-)} \quad \text{and} \quad [f(u)] = f(u^{(+)}) - f(u^{(-)})
\]  

(1.13)

are the jumps across the discontinuity with \( u^{(-)} \) and \( u^{(+)} \) being the left and right value of \( u \), respectively.

Note that piecewise smooth solutions that satisfy eq. (1.12) are not unique and additional conditions, such as the entropy condition have to be introduced to single out the physically relevant weak solution. This, however, goes beyond the scope of this work and the interested reader is referred to, e.g., [16, 32].

### 1.3 Godunov’s Method

In the previous section we have seen that one has to find a proper choice for the numerical flux function \( F \), in order to obtain a numerical approximation of the solution at the next time level \( t^{n+1} \). A very popular method, which is essentially the base of many modern schemes, is Godunov’s method [17]. This scheme considers the physical character of hyperbolic PDEs, in particular the propagation of information along characteristics [38]. It is based on the consecutive exact solution of a sequence of Riemann problems located at the cell interfaces.

A Riemann problem is an initial value problem for which the initial data are piecewise constant.

\[
\begin{align*}
  u_t + f(u)_x &= 0, \\
  u(x, t = 0) &= \begin{cases} 
    u^{(-)} & \text{if } x < 0 \\
    u^{(+)} & \text{if } x > 0
  \end{cases},
\end{align*}
\]  

(1.14)

where \( u^{(-)} \) and \( u^{(+)} \) represent the left and right state of the initial data, respectively. In the Godunov scheme, we consider the Riemann problem localized at the cell interfaces \( x_{i+1/2} \) at every time \( t^n \).

\[
\begin{align*}
  u(x, t^n) &= \begin{cases} 
    \bar{u}_i^n & \text{if } x_{i-1/2} \leq x < x_{i+1/2}, \\
    \bar{u}_{i+1}^n & \text{if } x_{i+1/2} < x \leq x_{i+3/2}
  \end{cases}.
\end{align*}
\]  

(1.15)
Each local Riemann problem is then solved exactly, in order to obtain an approximate solution $\bar{u}_{i}^{n+1}$ at the next time level $t^{n+1}$. In conservative form the approximation reads

$$\bar{u}_{i}^{n+1} = \bar{u}_{i}^{n} + \frac{\Delta t}{\Delta x} \left( \mathcal{F}^{\text{God}}(\bar{u}_{i-1}^{n}, \bar{u}_{i}^{n}) - \mathcal{F}^{\text{God}}(\bar{u}_{i}^{n}, \bar{u}_{i+1}^{n}) \right),$$  \hspace{1cm} (1.16)$$

with the numerical flux function

$$\mathcal{F}^{\text{God}}(\bar{u}_{i}^{n}, \bar{u}_{i+1}^{n}) = \frac{1}{\Delta t} \int_{t_{n}}^{t_{n+1}} f(u(x_{i+1/2}, t)) dt.$$  \hspace{1cm} (1.17)$$

Note that the numerical flux $\mathcal{F}^{\text{God}}$ is a two argument function, whereas the physical flux only depends on the point value at the cell boundary. If the Riemann problem does not cell-wise interact, i.e., the waves issued from two neighboring Riemann problems do not intersect, then the function $u(x_{i+1/2}, t)$ is a similarity solution, constant along each ray $\frac{x-x_{i+1/2}}{t}$

$$u(x_{i+1/2}, t) \equiv u^{*}(\bar{u}_{i}^{n}, \bar{u}_{i+1}^{n}) = \begin{cases} 
\bar{u}_{i}^{n} & \text{if } s > 0 \\
\bar{u}_{i+1}^{n} & \text{if } s < 0.
\end{cases}$$  \hspace{1cm} (1.18)$$

Consequently the integration (1.17) is trivial and we can obtain a simple expression for the Godunov flux in the case of convex conservation laws, i.e, $f''(u) > 0$ [32]

$$\mathcal{F}^{\text{God}}(u^{-}, u^{+}) = \begin{cases} 
\min_{u^{-} \leq u \leq u^{+}} f(u) & \text{if } u^{-} < u^{+} \\
\max_{u^{-} \geq u \geq u^{+}} f(u) & \text{if } u^{-} > u^{+}.
\end{cases}$$  \hspace{1cm} (1.19)$$

The values $u^{-}, u^{+}$ denote $\bar{u}_{i}^{n}, \bar{u}_{i+1}^{n}$, for the Riemann problem located at the right cell boundary $x_{i+1/2}$, respectively $\bar{u}_{i-1}^{n}, \bar{u}_{i}^{n}$, for the Riemann problem located at the left cell boundary $x_{i-1/2}$. Here they are constant cell averages, yet generally they could be approximated by high-order functions. Ben-Artzi and Falcovitz [6] generalized the concept of Riemann problem by admitting initial conditions that are linear function separated by discontinuities. Note that the solution of the Riemann problem depends on the speed of the discontinuity $s$ (1.12), i.e., on the direction characteristic information is transported. This is the upwind character of the Godunov’s method. Note that the neat formulation of (1.18), respectively (1.19) is only applicable in the scalar ($p = 1$) case. When applied to linear systems, the solution of the Riemann problem can be written as

$$u^{*}(\bar{u}_{i}, \bar{u}_{i+1}) = \bar{u}_{i} + \sum_{j: \lambda_{j} < 0}^{p} \alpha_{j}^{i+1/2} r_{j},$$  \hspace{1cm} (1.20)$$

$$= \bar{u}_{i+1} + \sum_{j: \lambda_{j} > 0}^{p} \alpha_{j}^{i+1/2} r_{j}.$$  \hspace{1cm} (1.21)$$

The vector $\alpha_{j}^{i+1/2}$ is obtained, by solving the linear system of equation

$$[\bar{u}]_{i+1/2} = \bar{u}_{i+1} - \bar{u}_{i} = \sum_{j}^{p} \alpha_{j}^{i+1/2} r_{j},$$  \hspace{1cm} (1.22)$$
where \( r_j \) are the right eigenvectors of the matrix \( A \) that defines the linear flux \( f(u) = Au \).

For large \( t \), the waves issued from two neighboring Riemann problems might intersect. Hence the solution doesn’t remain constant at the cell boundary. The cell-wise interaction of the Riemann problems is avoided by a time step size restriction in the scheme (1.16). Since the wave speeds are bounded by the eigenvalues \( \lambda_j \) of \( f'(u) \), the time step should be limited such that [38]

\[
|\lambda_j| \frac{\Delta t}{\Delta x} \leq 1 \quad j = 1, \ldots, p.
\]  

(1.23)

The maximum of this quantity is the so-called Courant-Friedrichs-Lewy (CFL) [11] condition

\[
\rho(A) \frac{\Delta t}{\Delta x} \leq 1,
\]

(1.24)

where

\[
\rho(A) \equiv \max_{1 \leq j \leq p} |\lambda_j(A)|.
\]

(1.25)

is the spectral radius of the Jacobian \( A = \nabla_u f(u), A \in \mathbb{R}^{p \times p} \).

We would like to remark that the Godunov flux (1.19), more precisely the Godunov scheme (1.16), belongs to the family of monotone schemes. Monotone numerical methods in conservation form converge to a physically relevant weak solution [67]. Moreover these schemes do not produce any spurious oscillations in the presence of discontinuities. Yet they are only of first-order accuracy, generating numerical diffusion (smearing) at discontinuities. In order to reduce numerical diffusion, high-order accurate methods have been developed. These methods are the focus of the next chapter.

### 1.4 Thesis Outline

In the next chapter (Chapter 2), we discuss high-order shock capturing methods. We introduce two standard, yet different FV discretization for hyperbolic PDE. Furthermore a general approach for constructing third-order accurate interpolants is introduced.

In Chapter 3 we introduce van Leer’s MUSCL scheme and discuss limiter functions in general. We investigate the accuracy of limiters and repeat Harten’s boundedness criteria. Furthermore we derive a general form for piecewise linear TVD limiter functions, which are at least of second-order accuracy away from discontinuities and local extrema.

Chapter 4 is essentially the main chapter. We derive a new class of third-order accurate limiter functions, which are eventually based upon non-polynomial reconstruction functions. We therefore discuss the concept on non-polynomial and nonlinear local variation bounded reconstructions, originally introduced by Marquina [42]. We introduce a new smooth logarithmic limiter function and its piecewise linear counterpart. We analyze the accuracy degeneration of limiters at local extrema and introduce an new concept to
avoid this phenomenon. In addition we examine the relevancy of several limiter extensions and discuss compressive schemes.

In Chapter 5 we link the suggested limiter to previous work on smooth limiter functions. We eventually derive another smooth-third order accurate limiter, which is shock capturing, local variation bounded and also recovers smooth extrema to full accuracy.

In Chapter 6 we discuss strong stability preserving Runge-Kutta time marching schemes. We conduct a von Neumann analyses and a dispersion analyses for schemes, employing different time integrators combined with different spatial reconstructions. We introduce the concept of modified equations in the frequency domain and discuss its connection to compressive schemes.

In Chapter 7, 9 and 10 we conduct a series of numerical experiments for scalar equations, for 1d systems and for 2d systems, respectively. In Chapter 10 we also discuss FV schemes and their accuracy in 2d. Euler equations, characteristic based reconstruction as well as approximative Riemann solvers are discussed in Chapter 8.

In Chapter 11 we test the new scheme on stiff relaxation systems, such as Grad 13 moments systems of extended thermodynamics. We discuss the concept of implicit explicit time integration schemes and under resolved FV schemes. Finally we conclude the results of this work in Chapter 12.
Chapter 2

High-Order Shock Capturing Methods

In the previous chapter we only discussed first-order accurate schemes. Such methods are generally considered too inaccurate for practical applications. The philosophy of high-order methods has been clearly pointed out by Leonard (1993) [37], saying, that “... a modern simulation would use a higher-order non-oscillatory method ... The result might be almost identical, but the philosophy is significantly different: we (rather than the grid-dependent differencing scheme) are controlling the transport model.”

Godunov’s scheme is a linear scheme that preserves the monotonicity of the solution. Higher order methods utilize essential nonlinearities so that higher than first order accuracy at smooth regions of the solution can be obtained. At the same time these schemes attain monotone resolution of discontinuities.

2.1 Different Concepts of High-Order FV Schemes

The two main principles of high-order FV methods are the improvement of the spatial accuracy coupled with a more accurate time marching scheme. A more accurate spatial approximation is recovered from employing high-order approximation of the cell interface values $u^{(-)}$, $u^{(+)}$ in the numerical flux (1.19) at the cell boundaries $x_{i\pm 1/2}$. Similar to the previous discussion in Section 1.2, we can distinguish between two different concepts of high-order FV methods, namely the full-discrete FV scheme (1.6) and the semi-discrete FV scheme (1.9).

For first-order accuracy, the full-discrete formulation (1.6) can be obtained from (1.9) by the forward Euler integration method

$$
\frac{d}{dt}\bar{u}_i = \frac{1}{\Delta t} (\bar{u}_{i+1} - \bar{u}_i) + O(\Delta t^2).
$$

Full-discrete schemes of higher than first-order accuracy cannot be simply recovered from a higher order time marching scheme, because they employ a time averaged flux function $F_{i+1/2}$ (1.8). Furthermore they become quite complicated for nonlinear equations, employing solvers for general Riemann problems (see, e.g., [6,69,70,76]). For linear problems this approach leads to a whole family of very high order schemes,
which are particularly effective for calculating long evolution problems (see, e.g., [56, 68]). To illustrate the difficulty arising for nonlinear hyperbolic systems, we approximate the time averaged flux function \( F_{i+1/2}^n \) with a simple quadrature rule (midpoint rule), yielding

\[
F_{i+1/2}^n = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(u(x_{i+1/2}), t) \, dt = f(u(x_{i+1/2}, t^n + \Delta t/2)) + \frac{1}{24} \Delta t^2 f''_{i+1/2} + O(\Delta t^4) \tag{2.2}
\]

with the abbreviation

\[
f''_{i+1/2} = f''(\bar{u}) (\bar{u})^2 + f' (\bar{u}) \bar{u}_{tt}, \quad \text{and} \quad \bar{u} = u(x_{i+1/2}, t^n + \Delta t/2). \tag{2.3}
\]

Then we can write the fully-discrete FV scheme (1.6) utilizing (2.2) as

\[
\bar{u}_{i+1}^{n+1} = \bar{u}_i^n + \frac{\Delta t}{\Delta x} \left( F_{i-1/2}^n - F_{i+1/2}^n \right) = \bar{u}_i^n + \frac{\Delta t}{\Delta x} \left( f(u(x_{i-1/2}, t^n + \Delta t/2)) - f(u(x_{i+1/2}, t^n + \Delta t/2)) \right). \tag{2.4}
\]

We neglected deliberately the term \( \frac{1}{24} \Delta t^2 f''_{i+1/2} \) in the above scheme to simplify the conceptual ideas behind this approach. This scheme is in conservation form and the only task is now to find an appropriate approximation for the physical fluxes \( f(u(x_{i+1/2}, t^n + \Delta t/2)) \) in the phase space \((x, t)\). The simplest way to do this is to assume that the function is analytic, so one can expand it to arbitrary order with the Taylor series. Consequently one considers also the function \( f''_{i+1/2} \) (2.3) in the numerical integration (2.2). Afterwards one eliminates the time derivatives in the Taylor expansion by space derivatives from the conservation law (1.2). Finally the cell centered space derivatives are approximated with their discrete cell neighbors.

These schemes originate from the linear advection equation (1.4). They make use of the fact, that higher time derivatives are replaced by expressions involving higher spatial derivatives:

\[
u_t = -au_x, \quad u_{tt} = -au_{xt} = a^2 u_{xx}. \tag{2.5}
\]

To derive the scheme, we utilize a Taylor expansion for the values \( u(x_{i+1/2}, t^n + \Delta t/2) \) in (2.4) up to second-order,

\[
\bar{u}_{i+1}^{n+1} = \bar{u}_i^n + a \frac{\Delta t}{\Delta x} (-\Delta x u_x - \Delta t u_{xt}) \tag{2.6}
\]

and use (2.5) to exchange mixed derivatives

\[
\bar{u}_{i+1}^{n+1} = \bar{u}_i^n - a \Delta t u_x + (a \Delta t)^2 u_{xx}. \tag{2.7}
\]

When we approximate the spatial derivatives by central differences, we yields the well known Lax-Wendroff method

\[
\bar{u}_{i+1}^{n+1} = \bar{u}_i^n - a \frac{\Delta t}{2\Delta x} \left( \bar{u}_{i+1}^n - \bar{u}_{i-1}^n \right) + \frac{1}{2} a^2 \left( \frac{\Delta t}{\Delta x} \right)^2 \left( \bar{u}_{i-1}^n - 2\bar{u}_i^n + \bar{u}_{i+1}^n \right). \tag{2.8}
\]
2.1. Different Concepts of High-Order FV Schemes

Obviously the choice of discrete approximations for \( u_x \) and \( u_{xx} \) changes the numerical scheme. In higher than second-order fully-discrete FV methods it is necessary to approximate the numerical flux function to higher-than second-order accuracy (see eq. (2.2)). Which is already an elaborated task for linear systems, turns out to be significantly more complicated for nonlinear systems. In fact one cannot simply generalize this approach for any conservation law, instead for every system one has to derive a new scheme (see, e.g., [56, 69]). Therefore, we will not focus on this approach.

An alternative and significantly simpler approach is to use the semi-discrete FV scheme (see eq. (1.9))

\[
\frac{d}{dt} \bar{u}_i = L_i(\bar{u}^n) = \frac{1}{\Delta x} \left( F(\hat{u}_{i-\frac{1}{2}}, \hat{u}_{i+\frac{1}{2}}) - F(\hat{u}_{i+\frac{1}{2}}, \hat{u}_{i+\frac{1}{2}}) \right),
\]

with the numerical flux \( F(\hat{u}^-, \hat{u}^+) \). This could be, for example, the Godunov flux (1.19)

\[
F(\hat{u}_{i+\frac{1}{2}}, \hat{u}_{i+\frac{1}{2}}) = f(u^*(\bar{u}_i^n, \bar{u}_{i+1}^n)),
\]

but also some other two argument flux function (which we will introduce later). The time averaging of the flux is then attained by a time integration routine (ODE solver), which for high-order schemes consist of repeated stages of the forward Euler method (2.1). Since we calculate the fluxes in every stage over the time interval \([t^n, t^{n+1}]\) they are accurately integrated, i.e., averaged. Note that we can use Godunov’s flux function because \( u(x_{i+1/2}, t) \) is constant in time in the solution of the Riemann problem (1.18).

This approach has the major advantage that time and space accuracy is decoupled, which allows a separate treatment of both approximations. We will first focus on the spatial reconstruction, before we apply an appropriate time marching scheme. The evolution of \( \bar{u}_i^n \) is governed by the left and right limits \( \hat{u}_{i+\frac{1}{2}}^{(\pm)} \), the interface values, of the reconstructed function \( \hat{u}(x) \). The cell interface value \( \hat{u}_{i+\frac{1}{2}} = \hat{u}(\bar{u}_i^n, \bar{u}_{i+1}^n) \) denotes the intermediate value at \( x_{i+1/2} \) of the Riemann problem solution with initial data \( \bar{u}_i^n, \bar{u}_{i+1}^n \). The approximation of the interface values \( \hat{u}^- \) and \( \hat{u}^+ \) from the known cell mean values is the essential reconstruction task and determines the scheme’s order of accuracy. The term reconstruction is used, to clarify the difference between high order interpolation, which generates spurious oscillations near discontinuities. High order reconstructions are designed to be essentially or even totally free of oscillations.

Note that in semi-discrete schemes (2.9), the high-order approximation of the flux function is shifted to the problem of accurately reconstructing the cell boundary values \( \hat{u}^-, \hat{u}^+ \), respectively. In contrast to fully-discrete schemes, we do not encounter any particular difficulties for nonlinear systems, since the spatial approximation of cell boundaries is not depending on temporal informations of the system (compare to swapping of temporal derivatives in eq. (2.7)). The nonlinearities are handled by an approximative Riemann solver, which will be discussed in Section 8.3. Therefore, semi-discrete schemes are quite general and a popular choice for high-order numerical approximations of hyperbolic PDEs.
Chapter 2. High-Order Shock Capturing Methods

2.2 Spatial Reconstruction

In Section 1.2 we derived the Godunov scheme, which is based upon constant spatial reconstruction. The cell interface values are \( \hat{u}_{i+1/2}^{-} = \bar{u}_{i}^n \) and \( \hat{u}_{i+1/2}^{+} = \bar{u}_{i+1}^n \), respectively. In combination with forward Euler time integrator (2.1), we obtain a first-order accurate FV scheme.

To obtain higher order accuracy, we have to improve the reconstruction of the left and right limits \( \hat{u}_{i+1/2}^{(\pm)} \) and employ a higher-order time-stepping method. We focus in this work only on local spatial reconstruction function. Thus we utilize a three-point stencil, where the interface values are defined as

\[
\hat{u}_{i+1/2}^{-} \equiv L(\bar{u}_{i-1}, \bar{u}_{i}, \bar{u}_{i+1}), \quad (2.11)
\]

and capital \( L \) and capital \( R \) denote the left and right cell interface approximations, respectively. The motivation for choosing a compact stencil is twofold. First we want to remain in a standard and very successful framework introduced originally by van Leer in 1979 [77] for second-order schemes. These schemes are widely used in CFD community (see, e.g., [21, 71]). Secondly compact stencils are computationally very efficient, reducing the amount of storage and exchange of information to a minimum. An additional advantage of having a compact stencil, is the simplified implementation of boundary conditions.

In this research work, we are particularly interested in globally third-order accurate spatial reconstruction of the cell boundary values. The principle used to obtain third-order of spatial accuracy is based upon Taylor expansion.

**Lemma 2.2.1** Given two functions \( f, g \in C^3[a, b] \), with \( \Delta x = b - a \). Assume that the integrals hold

\[
\int_a^b f(x)dx = \int_a^b g(x)dx, \quad (2.12)
\]

and that the lateral gradients agree up to second-order

\[
(f - g)'(a) = O(\Delta x^2) = (f - g)'(b), \quad (2.13)
\]

then \( (f - g)(x) = O(\Delta x^3) \) for all \( x \in [a, b] \)

**Proof.** We define \( w(x) \equiv (f - g)(x) \). A Taylor expansion around \( x' \in [a, b] \) yields

\[
w(x) = w(x') + w'(x')(x - x') + \frac{1}{2} w''(x')(x - x')^2 + O(|x - x'|^3), \quad (2.14)
\]

for an arbitrary \( x \in [a, b] \). By continuity of \( f, g \) and because of \( \int_a^b w(x)dx = 0 \), we know that there exist a \( x' \in [a, b] \), such that

\[
w(x') = 0. \quad (2.15)
\]
The mean value theorem implies that

\[ w''(x') = \frac{w'(a) - w'(b)}{\Delta x}, \quad (2.16) \]

for \( x' \in [a, b] \) and \( \Delta x = b - a \). With assumption (2.13) it follows that \( w''(x') = O(\Delta x) \). A Taylor expansion of \( w'(x') \) around \( a \) yields

\[ w'(x') = w'(a) + O(|x' - a|^2) = O(\Delta x^2), \quad (2.17) \]

since \( w'(a) = O(\Delta x^2) \) and \( w''(a) = O(\Delta x) \). Consequently we get

\[ w(x) \equiv (f - g)(x) = O(\Delta x^3), \quad (2.18) \]

which concludes the proof. ■

Three data points are sufficient for constructing a unique quadratic interpolation function with the Ansatz function

\[ p_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2. \quad (2.19) \]

To derive the coefficients \( a_i, b_i \) and \( c_i \) we impose the conditions set in Lemma 2.2.1. Therefore the quadratic function \( p_i(x) \), integrated over the cell \( C_i = [x_{i-1/2}, x_{i+1/2}] \) has to recover the cell average \( \bar{u}_i \) itself and its left and right derivatives located at cell interfaces have to be approximated to second-order \( O(\Delta x^2) \).

Given the set of cell averages \( \{\bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1}\} \) we require

\[ \frac{1}{\Delta x} \int_{C_i} p_i(x)dx = \bar{u}_i \quad (2.20) \]

\[ p'_i(x_i - \Delta x/2) = \frac{\delta_{i-1/2}}{\Delta x} \quad (2.21) \]

\[ p'_i(x_i + \Delta x/2) = \frac{\delta_{i+1/2}}{\Delta x}, \quad (2.22) \]

where

\[ \delta_{i+\frac{1}{2}} = \bar{u}_{i+1} - \bar{u}_i \quad \text{and} \quad \delta_{i-\frac{1}{2}} = \bar{u}_i - \bar{u}_{i-1} \quad (2.23) \]

are the right and left differences across the cell interfaces, respectively. Conditions (2.20, 2.21, 2.22) lead to the following coefficients

\[ a_i = \bar{u}_i - \frac{1}{24}(\delta_{i+1/2} - \delta_{i-1/2}) \quad (2.24) \]

\[ b_i = \frac{1}{2\Delta x}(\delta_{i+1/2} + \delta_{i-1/2}) \]

\[ c_i = \frac{1}{\Delta x^2}(\delta_{i+1/2} - \delta_{i-1/2}), \]
We are in particular interested in evaluating the polynomial (2.19) at the cell boundaries

\[
\hat{u}_{i+\frac{1}{2}}^{(-)} \equiv p_i(x_{i+1/2}) = \frac{5}{6} \bar{u}_i - \frac{1}{6} \bar{u}_{i-1} + \frac{1}{3} \bar{u}_{i+1}
\]  

(2.25)

\[
\hat{u}_{i-\frac{1}{2}}^{(+)} \equiv p_i(x_{i-1/2}) = \frac{5}{6} \bar{u}_i - \frac{1}{6} \bar{u}_{i+1} + \frac{1}{3} \bar{u}_{i-1}.
\]  

(2.26)

Note that the Ansatz function must not be necessarily a polynomial. In fact one can use any function \( r_i(x) \), for which holds \( r_i(x) \in C^3[x_{i-1/2}, x_{i+1/2}] \) for all \( i \).

The idea of high-order polynomial reconstruction goes back to van Leer [76, 77]. Van Leer published a sequence of papers in which he introduces different local reconstruction functions, including the quadratic interpolation (2.25, 2.26). However he focused more on linear interpolation, developing second-order accurate schemes. Although high-order methods give much better accuracy on smooth regions of the solution than first-order monotone schemes, they generate spurious oscillations near discontinuities. Van Leer designed a scheme, which is second-order accurate on smooth regions of the solution, yet does not produce any spurious oscillations at jump discontinuities. Such methods, known as shock capturing schemes, are the focus of the following chapter.

For the sake of completeness we should mention that quadratic reconstruction functions have been employed by Harten et al. [19, 20] in essentially nonoscillatory (ENO) schemes. Previous to that, Agarwal [1] utilized a quadratic interpolation for finite-difference upwind scheme. Both schemes, however, are not local and employ five, respectively four data points.
Chapter 3

New Perspective on van Leer’s MUSCL Scheme

Up to now we considered spatial approximations under the assumption that the solution $u(x, t)$ is arbitrary smooth. Thus utilizing high-order reconstruction functions. We know that this assumption does not hold in the presence of jump discontinuities and that higher-order interpolation will effectively introduce spurious oscillations. The ultimate goal of high-order FV schemes is to reduce numerical diffusion and diminish spurious oscillations. This leads to the development of van Leer’s Monotone Upstream-Centered Scheme for Conservation Laws (MUSCL) [77] and its semi-discrete version by Osher [43].

3.1 Limiter Functions

The classical second-order TVD-MUSCL scheme assumes a piecewise linear interpolation from the average values $\bar{u}_i(t)$

\[
\hat{u}_{i+1/2}^{(-)} = \bar{u}_i + \frac{\Delta x}{2} \sigma_i, \\
\hat{u}_{i-1/2}^{(+)} = \bar{u}_i - \frac{\Delta x}{2} \sigma_i,
\]

(3.1)

where the slope $\sigma_i \approx u'_i$ can either be expressed via upwind, downwind or centered finite-differences to recover the Beam-Warming, Lax-Wendroff or Fromm method, respectively (see e.g., [21, 38, 71]). In order to avoid spurious oscillation and to control the introduced spatial variation, the slope $\sigma_i$ has to be exchanged by a so-called limiter using e.g., the downwind slope

\[
\sigma_i = \left( \frac{\delta_{i+1/2} - \delta_{i-1/2}}{\Delta x} \right) \phi(\theta_i),
\]

(3.2)

where

\[
\theta_i = \frac{\delta_{i-1/2}}{\delta_{i+1/2}}, \quad \delta_{i+1/2} \neq 0
\]

(3.3)
is a local smoothness measure. The function \( \phi(\theta) \) represents the limiter. Since \( \phi(\theta) \) is an operator applied on the spatial gradients, this procedure is also called slope limiting.

**Remark 3.1.1 (Five-point scheme)** Although the reconstruction of the cell interface values is local, utilizing only a three point stencil (see (2.11)), the final scheme essentially employs five points. For the upwind direction \((a > 0)\), the spatial accuracy depends on the discrete set of cell mean values \( \{ \bar{u}_{i-2}, \bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1} \} \). In the downwind direction \((a < 0)\), the schemes utilizes \( \{ \bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1}, \bar{u}_{i+2} \} \). Hence this type of methods is often called five-point schemes [77].

Theoretical studies and derivations of slope-limiters has been an ongoing research, which started with van Leer’s MUSCL scheme in 1977 and was carried on by several researches from mathematics to engineering (see, e.g., Waterson and Deconinck [85] for a comprehensive review). The basic design ideas behind limiters are consistency (or accuracy) and stability (or boundedness).

### 3.2 Boundedness Criteria

High-order FV schemes produce spurious oscillation for discontinuous data. An appropriate reconstruction function should achieve fully bounded behavior in the presence of steep gradients, i.e., fulfill the monotonicity restriction

\[
\min(\bar{u}_i, \bar{u}_{i+1}) \leq \bar{u}_{i+1/2} \leq \max(\bar{u}_i, \bar{u}_{i+1}). \tag{3.4}
\]

In other words, the reconstruction value of a local extremum is nondecreasing for a local minimum and nonincreasing for a local maximum and no new local extrema in \( x \) are created [19]. The data \( u(x, t) \) is monotone in \([x_i, x_{i+1}]\) if it is increasing or decreasing in \([x_i, x_{i+1}]\). These properties imply that the total variation (TV) of the function \( u(x, t) \) in \( x \) is diminishing in \( t \), i.e.,

\[
TV(u(\cdot, t^2)) \leq TV(u(\cdot, t^1)), \quad \text{for all } t^2 \geq t^1. \tag{3.5}
\]

Here TV denotes the total variation function, which for differentiable functions \( v(x) \), reads

\[
TV(v) = \int_{-\infty}^{\infty} |v'(x)| \, dx. \tag{3.6}
\]

For a discrete solution \( \bar{u}^n \) the total variation can be calculated by

\[
TV(\bar{u}^n) = \sum_{i=1}^{N} |\bar{u}_{i+1}^n - \bar{u}_i^n|. \tag{3.7}
\]

This is a simple way to measure how oscillatory, i.e., how much variation is introduced by a reconstruction function and if the total variation is preserved as time integration evolves. These ideas motivated Harten [19] to introduce the concept of total variation diminishing (TVD) schemes.

**Definition 3.2.1 (TVD Schemes)** A FV scheme in conservation form Eq. (2.9) is TVD if

\[
TV(\bar{u}^{n+1}) \leq TV(\bar{u}^n) \quad \text{for all } n. \tag{3.8}
\]
Harten also proposed in [19] sufficient conditions for a fully discrete scheme to be TVD.

**Theorem 3.2.1 (Harten 1983)** A scheme of the form

\[ \bar{u}_{i}^{n+1} = \bar{u}_{i}^{n} + D_{i+1/2}^{n} \frac{\delta_{i+1/2}^{n}}{2} - C_{i-1/2}^{n} \frac{\delta_{i-1/2}^{n}}{2}, \]  

(3.9)

where the coefficients \( D_{i+1/2}^{n} \) and \( C_{i-1/2}^{n} \) are arbitrary values (which may depend on \( \bar{u}^{n} \) in a nonlinear way) and \( \delta_{i\pm1/2}^{n} = \delta_{i\pm1/2}^{n} \) at time \( t^{n} \), is TVD if the inequalities

\[ D_{i+1/2}^{n} \geq 0, \quad C_{i-1/2}^{n} \geq 0, \quad C_{i+1/2}^{n} + D_{i+1/2}^{n} \leq 1 \]  

(3.10)

are satisfied for all \( i \).

**Proof.** We compute the total variation for \(-\infty \leq i \leq N \leq \infty\)

\[ \text{TV}(\bar{u}^{n+1}) = \sum_{i=-\infty}^{\infty} |\delta_{i+1/2}^{n+1}| = \sum_{i=-\infty}^{\infty} |\bar{u}_{i+1}^{n+1} - u_{i}^{n+1}| \]  

(3.11)

\[ \leq \sum_{i=-\infty}^{\infty} (1 - (C_{i+1/2}^{n} + D_{i+1/2}^{n})) |\delta_{i+1/2}^{n}| + \sum_{i=-\infty}^{\infty} C_{i-1/2}^{n} |\delta_{i-1/2}^{n}| \]

\[ + \sum_{i=-\infty}^{\infty} D_{i+3/2}^{n} |\delta_{i+3/2}^{n}| \]

\[ = \sum_{i=-\infty}^{\infty} |\delta_{i+1/2}^{n}| = \text{TV}(\bar{u}^{n}) \]

Consequently condition (3.8) holds for all \( n \) and the scheme is TVD. ■

It is obvious that a first-order monotone scheme is TVD. For high-order schemes Harten’s TVD properties (3.10) give direct bounds on limiter functions \( \phi(\theta) \). A high-order limiter diminishes the total variation if the following conditions hold (see, e.g., [16, 21, 38, 66, 71]):

\[ 0 \leq \phi(\theta) \leq 2 \theta, \quad 0 \leq \phi(\theta) \leq 2 \quad \text{and} \quad \phi(\theta) = 0 \quad \text{if} \quad \theta \leq 0. \]  

(3.12)

These conditions define a region called TVD region inside the \( \phi(\theta) - \theta \) plane. For \( \theta \leq 0 \) the reconstruction reduces to the constant cell average itself and smooth functions for which \( \theta \leq 0 \) holds cannot be recovered accurately. These bounds give no direct estimates on the convergence order of the scheme for smooth data. Note that reconstructions utilizing three-data points can recover second-order, as well as third-order accuracy. Yet limiter functions which lie inside the TVD region will not produce spurious oscillations on discontinuous data. We should remark that Harten’s TVD theorem cannot be generalized for nonlinear systems. In general the physical solution for a system is not TVD in any sense [38]. Hence we cannot expect the numerical solution to be. Indeed numerical examples show that some classical TVD limiters fail to produce physical solutions for hyperbolic systems. A major disadvantage of TVD limiters is the fact, that they cannot approximate smooth non-monotone data to full accuracy.
Chapter 3. New Perspective on van Leer’s MUSCL Scheme

Fig. 3.1 shows plots of two different monotonicity region. The left plot illustrates Harten’s TVD-region for nonlinear limiter functions. The right plot displays an extended monotonicity region derived by Spekreijse [64] and Dubois [12] (see also references therein). This region includes also monotone linear interpolants for $\theta < 0$. The bounds for a limiter function to lie inside the extended monotonicity region are:

\[
0 \leq \phi(\theta) \leq \beta \theta \quad \text{if} \quad 0 \leq \theta \leq 1
\]
\[
0 \leq \phi(\theta) \leq \gamma \quad \text{if} \quad \theta > 0
\]
\[
0 \leq \phi(\theta) \leq (\gamma - 2) \theta \quad \text{if} \quad \theta \leq 0,
\]

with $1 \leq \beta \leq 2$ and $1 \leq \gamma \leq 2$. The extended monotonicity region indicates a region where the interpolant, is monotone, although the reconstructed function is not necessarily monotone. It is based upon the following conditions [12]

\[
\max(0, \theta) \leq \phi(\theta) \leq 1 \quad \text{if} \quad -\infty < \theta \leq 1
\]
\[
1 \leq \phi(\theta) \leq \min(\theta, 2) \quad \text{if} \quad \theta \geq 2
\]
\[
\phi(\theta) \leq \theta \phi(\theta^{-1}) \quad \text{if} \quad \theta \geq 1.
\]

The first two conditions are simply Harten’s TVD bounds for $\theta > 0$. The last condition, in fact, allows to consider also quadratic interpolants. Thus generalizing the concept of limiting for high-order schemes (see following Section 3.3.1, Remark 3.3.1).

For linear interpolation, homogeneity of the interpolant reduces to simple symmetry requirements for the slope $u_i' \approx \sigma_i \equiv \sigma(\delta_{i-1/2}, \delta_{i+1/2})$ (3.2). If the lateral slope satisfies

\[
\sigma(\delta_{i-1/2}, \delta_{i+1/2}) = \hat{\sigma} \left( \frac{\delta_{i-1/2}}{\delta_{i+1/2}} \right) \delta_{i+1/2} \equiv \phi(\theta) \delta_{i+1/2} \equiv \phi(\theta^{-1}) \delta_{i-1/2},
\]

then every reconstruction of the form (3.1) utilizing (3.2) will only recover second-order accurate cell interface values. In the following sections, we will see that neglecting condition (3.15), we can derive a third-order accurate interpolant employing a limiter function.
3.3 Local Truncation Error Analysis

The local truncation error, defined as

\[ E_{\Delta x}(i, t) \equiv L_i(\bar{u}(t)) - L_{\text{ex}}(u(\cdot, t), i), \quad \text{for all } i, \]

(3.16)

represents the difference between the discrete operator applied to the true solution \( L_i(\bar{u}(t)) \) and exact operator applied to the true solution

\[ L_{\text{ex}}(u(\cdot, t), i) \equiv \frac{1}{\Delta x} \left( f(u(x_{i-1/2}, t)) - f(u(x_{i+1/2}, t)) \right) = \frac{d}{dt} \bar{u}_i(t). \]

(3.17)

For a scheme to be consistent, one requires that

\[ \lim_{\Delta x \to 0} E_{\Delta x}(i, t) = 0 \quad \text{for } \frac{\Delta t}{\Delta x} \text{ fixed.} \]

(3.18)

A scheme for which holds \(|E_{\Delta x}(i, t)| = O(\Delta x^q)\) is of \(q\)-th order accuracy in space. For simplicity we will reduce our accuracy analysis on the scalar linear advection with \(a = 1\) (see eq. (1.4)). We assume the initial data to be continuous and arbitrarily differentiable. The truncation error \( E_{\Delta x}(i, t) \) (3.16) then yields

\[ E_{\Delta x}(i, t) = \frac{1}{\Delta x} \left( \hat{u}^{(-)}_{i-\frac{1}{2}} - \hat{u}^{(-)}_{i+\frac{1}{2}} - (u(x_{i+1/2}) - u(x_{i-1/2})) \right). \]

(3.19)

Since our reconstruction entirely depends on cell interface differences (see eqs. (3.2, 3.3) we expand \( \delta_{i+1/2} \) around \( x_i - \xi \Delta x \) with \( \xi \in [-\frac{1}{2}, \frac{1}{2}] \) (see Appendix A for details):

\[ \delta_{i-1/2} = u'(x_{i-\xi}) \Delta x + \left( \xi - \frac{1}{2} \right) u''(x_{i-\xi}) \Delta x^2 \]

(3.20)

\[ + \frac{1}{24} (5 + 12(-1 + \xi) \xi) u^{(3)}(x_{i-\xi}) \Delta x^3 + O(\Delta x^4) \]

In contrast to conventional analysis we do not constrain the data to be located at the cell-center. It is crucial to understand that extended limiter conditions derived upon cell-centered Taylor series are too restrictive. A smooth extrema for example, although initially located at the cell center, will eventually as time iterations proceed be relocated inside a computational cell. The purpose of this analysis is to get accuracy conditions on limiter function for smooth data.

For \( \xi = \frac{1}{2} \), the lateral differences eq. (3.20, 3.21) yield

\[ \delta_{i-1/2} = u'(x_{i-1/2}) \Delta x + \frac{1}{12} u'''(x_{i-1/2}) \Delta x^3 + O(\Delta x^5) \]

(3.22)

\[ \delta_{i+1/2} = u'(x_{i-1/2}) \Delta x + u''(x_{i-1/2}) \Delta x^2 + \frac{7}{12} u'''(x_{i-1/2}) \Delta x^3 + O(\Delta x^4) \]

(3.23)

Note that in (3.22) the even derivatives vanish.
3.3.1 Monotone Initial Data

In the following we will derive the truncation error for monotone initial data, i.e., \( u'(x_{i-\xi}) \neq 0 \) (see Appendix A for details). For small \( \Delta x \) a series expansion of the local smoothness measure \( \theta = \frac{\delta_{i+1/2} - \delta_{i-1/2}}{\delta_{i+1/2}} \) (3.3) yields

\[
\theta_{i}^{(\xi)} = \theta_{i-1}^{(\xi)} = 1 - \frac{u''(x_{i-\xi})}{u'(x_{i-\xi})} \Delta x + O(\Delta x^2). \tag{3.24}
\]

Substituting eq. (3.24) into \( \phi(\theta) \) and expanding around one, we obtain

\[
\phi_{i}^{(\xi)} = \phi_{i-1}^{(\xi)} = \phi(1) - \frac{u''(x_{i-\xi})}{u'(x_{i-\xi})} \phi'(1) \Delta x + O(\Delta x^2). \tag{3.25}
\]

Note that we formally have to Taylor expand all above equations up to \( O(\Delta x^3) \) to get a consistent truncation error of up to third-order. Yet for readability we try to restrict some of the expressions to a single line.

Inserting the later expression into eq. (3.19), the truncation error for monotone data, i.e., \( u'(x_{i-\xi}) \neq 0 \) utilizing reconstruction \( \tilde{u}_{i\pm1/2}^{(\xi)} \) (3.1) reads

\[
E_{\Delta x}(i - \xi, t) = \frac{1}{2} (\phi(1) - 1) u''(x_{i-\xi}) \Delta x - \frac{1}{6} (1 + 3\xi(\phi(1) - 1) - 3\phi'(1)) u^{(3)}(x_{i-\xi}) \Delta x^2 + C(\xi) \Delta x^3 + O(\Delta x^4), \tag{3.26}
\]

with

\[
C(\xi) = -\frac{1}{48} u^{(4)}(x_{i-\xi}) (-3 + 8\xi - 12\xi^2 + 3\phi(1) + 12\xi^2\phi(1) + 12\phi'(1) - 24\xi\phi'(1))
- \frac{u''(x_{i-\xi}) (-u''(x_{i-\xi})^2 + 2u'(x_{i-\xi})u^{(3)}(x_{i-\xi})) \phi''(1)}{4u'(x_{i-\xi})^2}. \tag{3.27}
\]

It is obvious that for a method to be of high-order, i.e., at least of second-order accuracy the limiter function \( \phi \) must pass Lipschitz continuously through the point \( \phi(1) = 1 \). Then the dependence on the parameter \( \xi \) is completely canceled out for monotone profiles. For a method to be of third-order accuracy the limiter function must fulfill

\[
\phi(1) = 1, \quad \text{and} \quad \phi'(1) = \frac{1}{3}. \tag{3.28}
\]

A compact scheme in limiter formulation cannot achieve more than third-order accuracy, since

\[
C(\xi) = -\frac{u^{(4)}(x_{i-\xi})}{12} \frac{u''(x_{i-\xi}) (-u''(x_{i-\xi})^2 + 2u'(x_{i-\xi})u^{(3)}(x_{i-\xi})) \phi''(1)}{4u'(x_{i-\xi})^2}. \tag{3.29}
\]

We should mention that we do not have any geometrical interpretation of the convergence function \( C(\xi) \). Because of the high convergence rate this function is also of rather less interest. However, we will see that for non-monotone data the convergence constants will be of more interest.
3.3. Local Truncation Error Analysis

We can find a whole set of linear second-order accurate functions depending on $\phi'(1)$. Popular choices for second-order accurate reconstructions are:

- monotonized symmetric-upwind central differences, also known as Fromm’s scheme [15]
  \[
  \phi(\theta) = \frac{1 + \theta}{2},
  \]
  (3.30)

- linear-upwind interpolation
  \[
  \phi(\theta) = \theta
  \]
  (3.31)

- simple central differences, interface values are approximated without upwind information
  \[
  \phi(\theta) = 1
  \]
  (3.32)

- quadratic-upwind reconstruction, i.e., Leonard’s QUICK scheme [35, 36]
  \[
  \phi(\theta) = \frac{3 + \theta}{4}.
  \]
  (3.33)

Yet there exists only one unique reconstruction function
\[
\phi(\theta) = \frac{2 + \theta}{3},
\]
(3.34)
for which we get third-order spatial accuracy. This function is a parabola through the discrete set of cell mean values $\{\bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1}\}$. With eq. (3.34) we recover the quadratic reconstruction of Section 2.2. We should remark that traditionally most limiters employ only second-order accurate reconstructions (see, e.g., [21, 38, 66, 71] and references therein). We summarize these results in the following remarks.

**Remark 3.3.1 (Limiter Accuracy)**

i) The functions above, i.e., eqs. (3.30, 3.31, 3.32, 3.33, 3.34), can be written in the compact $\kappa$-form
\[
\phi(\theta, \kappa) = \frac{1}{2} ((1 - \kappa) \theta + (1 + \kappa))
\]
(3.35)
utilizing the reconstruction formulation
\[
\hat{u}_{i+\frac{1}{2}} = \bar{u}_i + \frac{1}{2} \phi(\theta_i) \delta_{i+\frac{1}{2}}
\]
(3.36)
\[
\hat{u}_{i-\frac{1}{2}} = \bar{u}_i - \frac{1}{2} \phi(\theta_{i-1}) \delta_{i-\frac{1}{2}}.
\]
For $\kappa \in [-1, 1]$, the limiter $\phi(\theta, \kappa)$ is a monotone increasing function, for which holds $\phi(0, \kappa) = 0$ and $\phi(1, \kappa) = 1$.

ii) The local truncation error of $\phi(\theta, \kappa)$ for continuous data reads:
\[
E_{\Delta x}^\kappa(i - \xi, t) = -a \frac{1}{12} \left( (1 - 3 \kappa) u_{(3)}^{(3)}(x_{i-\xi}) \right) \Delta x^2 + O(\Delta x^3),
\]
(3.37)
with $a \in \mathbb{R}$. 

iii) Second-order accurate schemes, based upon a piecewise linear reconstruction satisfy the symmetry requirement
\[ \phi(\theta^{-1}) = \phi(\theta) \theta^{-1}, \quad \forall \theta \in \mathbb{R} \setminus \{0\}. \] (3.38)

Therefore, the reconstruction formulation (3.36) can be significantly simplified, yielding
\[ \hat{u}_{i+\frac{1}{2}}^{-} = \bar{u}_i + \frac{1}{2} \phi(\theta_i) \delta_{i+\frac{1}{2}} \] (3.39)
\[ \hat{u}_{i-\frac{1}{2}}^{+} = \bar{u}_i - \frac{1}{2} \phi(\theta_i) \delta_{i+\frac{1}{2}}. \]

Obviously no evaluation of \( \theta^{-1} \) is needed.

iv) For finite-differences, i.e., point values \( \{u_{i-1}, u_i, u_{i+1}\} \), the quadratic-upwind interpolation (3.33) approximates the interface values by
\[ \hat{u}_{i+\frac{1}{2}}^{-} = \frac{3}{4} u_i - \frac{1}{8} u_{i-1} + \frac{3}{4} u_{i+1} \]
\[ = u \left( x_{i+1/2} \right) + \frac{1}{16} u^{(3)} \left( x_{i+1/2} \right) \Delta x^3. \]

Thus yielding locally third-order accuracy. Yet the interpolation for the whole FV-cell is only second-order accurate, since \( \phi'(1) = 1/4 \). (compare (3.28)).

v) With the compact \( \kappa \)-formulation (3.35) we can derive a family of limiters
\[ \phi^K(\theta) = \max \left( 0, \min \left( 2\theta, \frac{1}{2} (1 - \kappa) \theta + (1 + \kappa) \right), 2 \right). \] (3.40)

This is a set of functions bounded by Harten’s TVD conditions.

The \( \kappa \)-parameter formulation goes back to van Leer’s MUSCL scheme [77]. Yet van Leer utilizes a different reconstruction procedure, which is not based upon a single limiter function \( \phi^K(\theta) \) (3.40). Furthermore van Leer’s MUSCL scheme is a fully-discrete scheme, in which the numerical flux is only approximated to second-order accuracy (see discussion in Chapter 2).

In Fig. 3.2 (left) we envisage the different spatial reconstruction inside Sweby’s second-order TVD region [66]. This is simply the region bounded by Harten’s monotonicity constrains and the function
\[ \phi^{MM}(\theta) = \max(0, \min(\theta, 1)), \] (3.41)
which is the \textit{minmod} limiter. In the right plot in Fig. 3.2 one can see different piecewise-linear second-order accurate limiter function, such as Roe’s \textit{suberbee} limiter
\[ \phi^{SB}(\theta) = \max(0, \min(2\theta, 1), \min(\theta, 2)), \] (3.42)
and van Leer’s \textit{monotonized central difference} (MCD) limiter employing the continuous reconstruction (3.30)
\[ \phi^{MCD}(\theta) = \max(0, \min(2\theta, \frac{1}{2} (1 + \theta), 2)). \] (3.43)
The MCD limiter has a smooth counterpart, namely van Leer’s smooth limiter

\[
\phi^{\text{VL}}(\theta) = \frac{\theta + |\theta|}{1 + \theta}.
\]  

(3.44)

This is simply a harmonic mean function, which is zero for negative values of \(\theta\). Smooth limiter function are eventually more diffusive than their piecewise linear counterparts. A simple analysis shows that for small deviations from one, i.e., for \(\theta = 1 - \Delta \theta\) with \(\Delta \theta = \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \Delta x\) (see eq. (3.24)) we recover

\[
|\phi^{\text{VL}}(1 - \Delta \theta) - \phi^{\text{MCD}}(1 - \Delta \theta)| = \frac{1}{4} \left( \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \right)^2 \Delta x^2 + \mathcal{O}(\Delta \theta^3).
\]  

(3.45)

Obviously on coarse meshes and for \(\frac{u''(x_i - \xi)}{u'(x_i - \xi)} \ll 1\), smooth limiter function recover only first-order accuracy.

Note that neither the minmod nor the superbee limiter belong to the family of \(\phi^\kappa\) limiter functions (3.40). Whereas the monotonized central difference limiter is recovered for \(\kappa = 0\).

### 3.3.2 Initial Data With Local Extrema

So far we examined the spatial reconstruction accuracy for monotone data. In the presence of a local extremum, i.e., \(u'(x_i - \xi) = 0\) and \(u''(x_i - \xi) \neq 0\), the smoothness measure \(\theta\) yields

\[
\theta_i^{(\xi)} = \frac{2 \xi - 1}{1 + 2 \xi} + C_1(\xi) \frac{u^{(3)}(x_i - \xi)}{u''(x_i - \xi)} \Delta x + \mathcal{O}(\Delta x^2), \quad \xi \neq -\frac{1}{2}
\]  

(3.46)

\[
\theta_{i-1}^{(\xi)} = \frac{3 - 2 \xi}{1 - 2 \xi} + C_2(\xi) \frac{u^{(3)}(x_i - \xi)}{u''(x_i - \xi)} \Delta x + \mathcal{O}(\Delta x^2), \quad \xi \neq \frac{1}{2}
\]  

(3.47)

\[\begin{array}{c}
\text{Figure 3.2: Limiter plots in Sweby’s second-order limiter region. Different reconstructions: } \phi_{(1)}(\theta) = \theta, \phi_{(2)}(\theta) = \frac{1}{2}(1 + \theta), \phi_{(3)}(\theta) = \frac{1}{3}(2 + \theta), \phi_{(4)}(\theta) = \frac{1}{4}(3 + \theta) \text{ and } \phi_{(5)}(\theta) = 1 \text{ (left). Classical piecewise linear second-order limiter: } \phi^{\text{SB}} (3.42), \phi^{\text{MCD}} (3.43) \text{ and } \phi^{\text{MM}} (3.41) \text{ (right).}
\end{array}\]
Taylor expanding $\phi(\theta)$ for $\Delta x \to 0$, we recover

$$
\phi_i^{(\xi)} = \phi \left( \frac{2\xi - 1}{1 + 2\xi} \right) + C_1(\xi) \frac{u''(x_i - \xi)}{u''(x_i - \xi)} \phi' \left( \frac{2\xi - 1}{1 + 2\xi} \right) \Delta x + O(\Delta x^2) \quad (3.48)
$$

$$
\phi_{i-1}^{(\xi)} = \phi \left( \frac{3 - 2\xi}{1 - 2\xi} \right) + C_2(\xi) \frac{u''(x_i - \xi)}{u''(x_i - \xi)} \phi' \left( \frac{3 - 2\xi}{1 - 2\xi} \right) \Delta x + O(\Delta x^2) \quad (3.49)
$$

with the abbreviations

$$
C_1(\xi) = \frac{5 - 12\xi^2}{6(1 + 2\xi)^2} \quad \text{and} \quad C_2(\xi) = \frac{7 - 24\xi + 12\xi^2}{6(1 - 2\xi)^2}. \quad (3.50)
$$

Obviously smooth extrema with one vanishing lateral derivative, i.e., $\xi \to \pm \frac{1}{2}$ can’t be resolved accurately using a slope limiting approach. TVD reconstructions will eventually degenerate to first-order spatial accuracy, since

$$
\lim_{\xi \to -\frac{1}{2}} \phi \left( \theta_i^{(\xi)} \right) = \lim_{\xi \to +\frac{1}{2}} \phi \left( \theta_i^{(\xi)} \right) = 0 \quad (3.51)
$$

This is a principle design effect of limiter functions based upon a slope ratio $\theta$ as smoothness measure. Accuracy reduction for smooth extrema is often called *extrema-clipping*. The truncation error for non-monotone data then reads

$$
E_{\Delta x}(i - \xi, t) = \frac{1}{4} \left( 2 + (2\xi - 1) \phi(a_1) - (1 + 2\xi) \phi(a_2) \right) u''(x_i - \xi) \Delta x
$$

$$
- \frac{1}{4} \left( \frac{2}{3} - 2\xi - (5 - 12\xi + 12\xi^2) \phi(a_1) + (5 + 12\xi + 12\xi^2) \phi(a_2) \right)
$$

$$
- C_2(\xi) \phi'(a_1) + C_1(\xi) \phi'(a_2) \right) u''(x_i - \xi) \Delta x^2 + O(\Delta x^3) \quad (3.52)
$$

with the abbreviation

$$
a_1 = \frac{3 - 2\xi}{1 - 2\xi} \quad \text{and} \quad a_2 = \frac{2\xi - 1}{1 + 2\xi}. \quad (3.53)
$$

In contrast to the truncation error for monotone data, the linear term not only depends on the limiter function, but it also depends explicitly on the location $\xi$ of the extremum inside a FV-cell. It is a simple calculation, yet an important result, that for quadratic interpolation, i.e., $\phi(\theta) = \frac{1}{3} (2 + \theta) \quad (3.34)$, the error (3.52) truncates at $O(\Delta x^3)$. In other words, we can always use a limiter reconstruction based upon (3.36), as long as we ensure that in the presence of an extremum we employ the reconstruction (3.34). This is the only possibility to recover third-order spatial accuracy.

However if we assume an extremum located directly for every time-step at the cell-center, i.e., $\xi = 0, \forall \Delta t > 0$, the error analysis yields:

$$
E_{\Delta x}(i, t) = \frac{1}{4} \left( 2 - \phi(-1) - \phi(3) \right) u''(x_i) \Delta x - \frac{1}{4} \left( \frac{2}{3} + 5\phi(-1) - 5\phi(3) \right)
$$

$$
+ C_1(0) \phi'(-1) - C_2(0) \phi'(3) \right) u''(x_i) \Delta x^2 + O(\Delta x^3).
$$
3.3. Local Truncation Error Analysis

Hence one might assume that if
\[ \phi(-1) + \phi(3) = 2 \] (3.54)
holds we would recover second-order spatial accuracy at local extrema. This condition for second-order accuracy was originally proposed by Dubois [12]. Based on this condition, he derived a limiter, which utilized a linear reconstruction (3.33) for monotone data. Numerical convergence studies, however, envisaged an accuracy degeneration for local extrema.

It is clear from the previous error analysis that this condition is only seldom applicable. Local extrema although initially located at a cell center, will eventually be relocated during their evolution inside the cell. Consequently condition (3.54) will no longer be valid. Furthermore restricting the analysis only to second-order accurate schemes is a pure waste of potential, since we know that third-order spatial accuracy can also be obtained with (3.34).

We should remark that result (3.52) represents a generalization of a theorem originally derived by Osher and Chakravarthy in 1984 [44]. The theorem says, that a semi-discrete FV-scheme (2.9) is at most first-order accurate at non-sonic critical points of the data \( u \). (A sonic point \( u^* \) is one such that \( f'(u^*) = 0 \).) Dubois [12] explicitly claims, that if a limiter satisfies \( \phi(-1) + \phi(3) = 2 \) this is not the case and second-order accuracy is recovered. Our analysis confirms Osher and Chakravarthy’s result. Moreover, with (3.52) we get quantitative information on the reconstruction accuracy of any particular limiter for non-monotone data.
Chapter 4

New Third-Order Accurate Limiters

In the previous chapter (Chapter 3) we considered only second-order accurate schemes, which utilize TVD limiters. In the following, we will focus on third-order accurate schemes, which in contrast to TVD schemes approximate smooth data accurately. Before we examine some limiter functions, which were particularly designed to improve the accuracy for smooth data, we will discuss different concepts of spatial reconstruction.

4.1 Non-Polynomial Reconstruction Functions

Marquina [42] has suggested a different concept for the approximation of cell interface values in FV schemes. The main idea is to use nonlinear non-polynomial reconstruction functions, which are of high-order accuracy, are essentially non oscillatory at discontinuities and do not utilize limiters. Thus the reconstruction recovers high-order accuracy also at local extrema. For this purpose he introduces the concept of local total variation bounded (LTVB) reconstructions [42].

Definition 4.1.1 (Local Total Variation Bounded) The local variation (LV) of a piecewise smooth function \( f_i(x) \) in a cell \( C_i \) is given by

\[
LV(f_i) = TV(f(x))|_{C_i}. 
\] (4.1)

The function \( f_i \) is locally total variation bounded in \( C_i \), if there exists a constant \( K > 0 \), independent of the cell size \( \Delta x \), such that for all \( i \)

\[
LV(f_i) \leq K \Delta x. 
\] (4.2)

This concept allows to quantify the variation in a cell \( C_i \), introduced by a particular reconstruction function. It is in fact a weaker property, than total variation boundedness (TVB) [42]. We say a scheme is TVB in \( 0 \leq t < T \) (\( T \) fixed), if there exist a constant \( M \) independent of \( \Delta x \), such that

\[
TV(\bar{u}^n) \leq M TV(\bar{u}^0). 
\] (4.3)

If \( M = 1 \), the scheme is TVD (see Definition 3.2.1 (TVD Schemes)).
The interesting case is to estimate the local variation of a reconstruction function in the presence of a jump discontinuity. For jump discontinuities one of the lateral derivatives of the reconstruction function is an unbounded function of \( O(1/\Delta x) \). Thus we obtain (see (2.21, 2.22))

\[
 f'_i(x_i - \Delta x/2) = \frac{\delta_{i-1/2}}{\Delta x} = O(1/\Delta x) \quad \text{and} \quad f'_i(x_i + \Delta x/2) = \frac{\delta_{i+1/2}}{\Delta x} = O(1). \quad (4.4)
\]

The local variation of an unlimited linear interpolation function

\[
 l_i(x) = \bar{u}_i + u'_i(x - x_i),
\]

yields

\[
 LV(l_i) = TV(l(x))|_{C_i} = \int_{C_i} |l'_i(x)|dx = \Delta x |u'_i|. \quad (4.5)
\]

Identically we obtain the local variation of a quadratic reconstruction (2.19) assuming monotone data with \( \delta_{i-1/2} \delta_{i+1/2} > 0 \)

\[
 LV(p_i) = \frac{|\delta_{i-1/2}| + |\delta_{i+1/2}|}{2}. \quad (4.6)
\]

Consequently in the presence of jump discontinuities the local variation of both, the linear and the quadratic interpolation is of \( O(1) \). Schemes utilizing pure polynomial reconstruction functions produce spurious oscillations at discontinuities. The LVB condition (4.2) appears to be necessary for a method to own enough local smoothing to deal with discontinuities and to be eventually total variation bounded. To avoid oscillation in the presence of discontinuities we have to either apply limiters, which diminish the total variation in the vicinity of discontinuities (see Chapter 3) or find different LVB approximation functions.

Marquina [42] proposed to approximate the cell interface values with a local hyperbolic reconstruction (LHR) function

\[
 h_i(x) = a_i + \frac{b_i}{x - x_i + c_i}. \quad (4.7)
\]

Employing accuracy conditions (2.20, 2.21, 2.22) (see Chapter 2), one can derive unique values \( a_i, b_i \) and \( c_i \), such that \( h_i(x) \) is local and of third-order accuracy for smooth data. The hyperbolic reconstruction is local, since the values \( a_i, b_i \) and \( c_i \) solely depend on \( \bar{u}_i \) and \( \delta_{i\pm 1/2} \). At discontinuities, the reconstruction is essentially variational bounded, i.e.,

\[
 LV(h_i) = O(\sqrt{\Delta x}). \quad (4.8)
\]

However, Marquina’s hyperbolic reconstruction still has to preprocess the lateral derivatives \( \delta_{i\pm 1/2}/\Delta x \) in the presence of jump discontinuities. Although the variation of the hyperbola tends to zero, it is in fact not fast enough (in terms of spatial resolution \( \Delta x \)) for practical calculations. This preprocessing, which in fact is very similar to limiting, affects the accuracy of the scheme for smooth extrema, such that it degenerates to \( O(\Delta x^{3/2}) \).
Motivated by the work of Marquina on local non-polynomial reconstruction functions, Artebrant and Schroll [5] developed a local double logarithmic reconstruction (LDLR) using the Ansatz function
\[ r_i(x) \sim a_i + b_i \ln(x - x_i + c_i) + d_i \ln(x - x_i + c_i) \] (4.9)
to approximate the interface values. Identically to the construction of a unique parabola \( p_i(x) \) (2.19) and hyperbola (4.7), one can construct a local function \( r_i(x) \) with uniquely defined values \( a_i, b_i, c_i \) and \( d_i \).

The LDLR is, in contrast to Marquina’s hyperbolic reconstruction, essentially of third-order away from discontinuities without the explicit use of limiters and without direct preprocessing procedures of lateral derivatives. Due to the logarithmic nature its total variation scales, in the presence of jump discontinuities, with
\[ \text{LV}(r_i) = O(\Delta x^q | \ln(\Delta x)|), \quad q > 0. \] (4.10)
Hence, at discontinuities, LDLR recovers the cell mean value faster, in terms of mesh resolution \( \Delta x \), than Marquina’s LHR. The exponent \( q \) will be discussed later. Yet it is immediately clear, that the larger \( q \), the faster LDLR recovers zero variation. Still a local smoothness measure as argument of the logarithm has to be built in. We should remark, that the construction of accurate reconstruction functions is only based on Lemma 2.2.1. In fact both, LDLR and Marquina’s LHR, utilize logarithmic functions. Yet LDLR depends on four parameters, resulting in an under-determined system of algebraic equations (compare accuracy conditions (2.20, 2.21, 2.22), Chapter 2). The free parameter in LDLR is constructed in such a way that, the reconstruction can distinguish between smooth and discontinuous data. We will discuss this issue in details in Section 5.3.

### 4.2 Logarithmic Limiter

Since LDLR is local and uses only three data point, it turns out that the whole reconstruction procedure can be significantly simplified and written in a convenient limiter formulation. The complete derivation is shown in Appendix B.

After some algebraic reformulation we finally find
\[ \hat{u}^{(-)}_{i+\frac{1}{2}} = \bar{u}_i + \frac{1}{2} \phi^{\log}(\theta_i, q) \delta_{i+\frac{1}{2}} \] (4.11)
\[ \hat{u}^{(+)}_{i-\frac{1}{2}} = \bar{u}_i - \frac{1}{2} \phi^{\log}(\theta_i, q) \delta_{i-\frac{1}{2}}. \] (4.12)

Where we employ only a single limiter function, the so called log-limiter
\[ \phi^{\log}(\theta_i, q) = \frac{2p((p^2 - 2p\theta_i + 1) \ln p - (1 - \theta_i)(p^2 - 1))}{(p^2 - 1)(p - 1)^2} \] (4.13)
with
\[ p = p(\theta_i, q) = 2 \frac{|\theta_i|^q}{1 + |\theta_i|^{2q}}. \] (4.14)
Chapter 4. New Third-Order Accurate Limiters

In contrast to the classical second-order TVD-MUSCL reconstruction (3.39), we have to use the limiter also with inverse input and hence also with reverse lateral differences $\delta_{i \pm 1/2}$. This optimizes the use of the local stencil in the TVD-MUSCL framework, resulting in third-order accuracy away from discontinuities and smooth extrema (see Remark 3.3.1). The exponent $q$ in eq. (4.14) controls the amount of the total variation of the smooth limiter. It also appears as an exponent in the convergence estimate of the total variation of LDLR at jump discontinuities (4.10).

The limiter function (4.13) has three removable singularities at $p(\pm 1) = 1$ and $p = 0$. The first two can be eliminated by setting $\phi_{\text{log}}(1, q) = 1$ and $\phi_{\text{log}}(-1, q) = 1/3$. This ensures that accuracy is not degenerated for data for which $\theta = \pm 1 + \mathcal{O}(\Delta x)$ holds. These singularities appear also in the original LDLR algorithm, where Artebrant and Schroll [5] suggest to interpolate around the singularities using three data point. Thus they employ a quadratic reconstruction for monotone data. The singularity for $\theta = 0$ is avoided by constant extrapolation.

Comparing the log-limiter with the unlimited quadratic reconstruction $\phi_{\text{poly}} = \frac{1}{3}(2 + \theta)$ (3.34), one can identify the following characteristic properties.

**Remark 4.2.1 (Properties of the Log-Limiter)**

i) For smooth monotone data we obtain

$$|\phi_{\text{log}}(1 - \Delta \theta, q) - \phi_{\text{poly}}(1 - \Delta \theta)| = \mathcal{O}(\Delta \theta^4),$$

with $\Delta \theta = \frac{u''(x_i)}{w'(x_i)} \Delta x$, for $\xi = 0$ (see (3.24)).

ii) For local extrema, which lateral slopes have identical values but different signs, we obtain

$$|\phi_{\text{log}}(-1 + \Delta \theta, q) - \phi_{\text{poly}}(-1 + \Delta \theta)| = \mathcal{O}(\Delta \theta^4),$$

with $\Delta \theta = \frac{5}{6} \frac{u'(x_i)}{w'(x_i)} \Delta x$, for $\xi = 0$ (see (3.46)).

iii) For jump discontinuities the log-limiter yields

$$\lim_{\theta \to 0} \phi_{\text{log}}(\theta, q) = 0 \quad \text{and} \quad \lim_{\theta \to \pm \infty} \phi_{\text{log}}(\theta, q) = 0$$

for $q > 0$.

iv) For $q = 0$, $p(0, 0) = 1$ for all $\theta \in \mathbb{R} \setminus \{0\}$ the log-limiter yields

$$\phi_{\text{log}}(\theta, 0) = \frac{2 + \theta}{3}.$$ (4.18)

The log-limiter recovers for $\theta = \pm 1 + \mathcal{O}(\Delta x)$ third-order accuracy, deviating from the unlimited quadratic interpolation (3.34) only at fourth-order. Typically smooth limiters converge only quadratically to their unlimited interpolation function (compare with (3.45)). The nonlinear limiter function $\phi_{\text{log}}$ smoothly extends into the range of $\theta < 0$ and vanishes for $\theta \to \pm \infty$ and $\theta \to 0$, which is essential for limiting discontinuities. Note that in the presence of discontinuities the local variation of the log-limiter scales as (4.10).
4.3 Simplified Third-Order Limiter

Remark 4.2.2 (Local Variation Control Parameter) The parameter $q$ is associated with the amount of locally introduced variation in the log-limiter (4.13), respectively eq. (4.14). Indeed, the smaller $q$ the more variation is produced, i.e., the function $\phi^{\log}(\theta, q)$ moves out of Harten’s TVD-region (see Fig. 4.2, right), recovering eventually for $q = 0$ the quadratic interpolation (see (4.18)). The larger $q$ the less variation is produced, the log-limiter is inside Harten’s TVD-region, hence discontinuities are smeared out within more cells.

The new limiter is simpler and more efficient than the original LDLR reconstruction and seems to go beyond the limiter categories described in [85]. We want to emphasize, that to our knowledge this is the first time that a whole family of smooth third-order accurate limiters has been derived.

4.3 Simplified Third-Order Limiter

The logarithmic limiter $\phi^{\log}(\theta, q)$ is derived from a double logarithmic Ansatz function, which is conservative, third-order accurate and essentially local variation bounded. Both, the LDLR and the logarithmic limiter function are eventually complicated and computationally expensive, although limiter (4.13) has already only one logarithmic evaluation. Furthermore the LDLR is very sensitive to the parameter $q$ and Artebrant and Schroll [5] suggested to set $q = 1.4$ to ensure stability (see discussion above). Bearing the main features of the log-limiter in mind, we can construct a piecewise-linear limiter function with similar properties, yet improved shock capturing abilities.

Omitting logarithmic functions, we not only avoid troublesome singularities but also get better control on the reconstruction routine, thus on the total variation. The new limiter reads:

$$\Phi(\theta, \alpha, \beta, \gamma) = \max\left(0, \min\left(\frac{2 + \theta}{3}, \max\left(-\alpha \theta, \min\left(\beta \theta, \frac{2 + \theta}{3}, \gamma\right)\right)\right)\right), \quad (4.19)$$

![Figure 4.1: Shape of the log-limiter (4.13) for different values of $q$. The larger $q$, the less local variation is introduces. Thus numerical diffusion increases, leading to smeared discontinuities. Right: Zoom of the inner region.](image-url)
where the parameters $\alpha, \beta$ and $\gamma$ satisfy the conditions:

$$
\begin{align*}
0 & \leq \alpha \leq 2 \\
1 & \leq \beta \leq 2 \\
1 & \leq \gamma \leq 2
\end{align*}
$$

Similarly to the variation control parameter $q$ (4.14), the parameters $\alpha, \beta$ and $\gamma$ control the shock capturing properties and the total variation of the piecewise-linear limiter $\Phi$. The building block of the generic limiter (4.19) is the quadratic reconstruction, $\phi(\theta) = \frac{1}{3} (2 + \theta)$. The variables $\alpha$ and $\beta$ bound the variation of $\Phi$ for $\theta \in (-\infty, 0]$ and $\theta \in [0, 1]$, respectively. The parameter $\gamma$ represents the upper-bound for the third-order reconstruction for $\theta \in [1, +\infty)$ (see Fig. 3.1). We bound the variables $\beta$ and $\gamma$ according to Harten’s TVD constraints (3.12). So if we set $\alpha = 0$ we recover a TVD limiter

$$
\phi_{\text{TVD}}(\theta) \equiv \Phi(\theta, 0, 2, 2) = \max \left( 0, \min \left( 2\theta, \frac{2 + \theta}{3}, 2 \right) \right),
$$

which essentially reconstructs smooth functions, for which $\theta \in \left[ \frac{2}{3}, 4 \right]$, to third-order accuracy. Still degenerating to first-order accuracy for $\theta \leq 0$. This function obviously is very economic and simple. Its evaluation only consists of two $(\max, \min)$-statements identical to the minmod limiter (3.41). This limiter is essentially identical to the $\kappa = \frac{1}{3}$ Koren-limiter [29]. To our knowledge, originally this limiter formulation goes back to the $\kappa$-scheme of Anderson, Thomas and van Leer [2] (see eq. (3.40)). Yet they do not employ a single limiter formulation, such as eq. (4.21), but instead use minmod limiters to preprocess lateral slopes. Koren on the other hand, uses in his formulation the inverse $r = \frac{\delta_{i+1/2}}{\delta_{i-1/2}}$ of the slope ratio $\theta$ as smoothness measure. Therefore, the cell interface values $\hat{u}_{i+ \frac{1}{2}}^{(\mp)}$ have to be reformulated under consideration of $\phi_{\text{TVD}}(\theta) \delta_{i+1/2} = \psi_{\text{Koren}}(r) \delta_{i-1/2}$.

In the following we will refer to limiter (4.21) as Koren-limiter, in order to emphasize its first appearance. Note, that a similar limiter function has also been proposed by Arora and Roe [3] in the context of fully discretized FV methods. Although they employed a quadratic reconstruction for $\theta \approx 1$, the proposed

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.2.png}
\caption{Left: New limiter $\hat{\phi}$ (4.23) matching for the log-limiter with $q = 1.4$. Right: Zoom of the inner region. Extended monotonicity-region according to [12].}
\end{figure}
4.4 The Relevancy of Extended Monotonicity Regions

The algorithm is only second-order accurate for nonlinear equations (see Section 2.1 for explanation). Similar to classical second-order TVD limiters, the Koren-limiter suffers from clipping of extrema for $\theta \leq 0$.

To be able to capture smooth extrema for $\theta = -1 + O(\Delta x)$ without loss of accuracy, we have to ease the monotonicity requirement, i.e., Harten’s TVD theorem. Setting $\alpha \geq \frac{1}{3}$ expands the limiter function continuously for $\theta < 0$, thus reconstructing smooth data for $\theta \in [-2, -\frac{2}{3\alpha+1}]$ to third-order accuracy.

A further attribute of the proposed limiter is that it does not satisfy the symmetry requirement of classical second-order TVD limiters $\phi(\theta^{-1}) = \phi(\theta^{-1})$, $\forall \theta \in \mathbb{R} \setminus \{0\}$ (see eq. (3.38)). Suppose the data $\tilde{u}^n$ is symmetric in $x$, we want to impose symmetry on the reconstruction, i.e., $\tilde{u}_{i+\frac{1}{2}}^{(+)} = \tilde{u}_{i+\frac{1}{2}}^{(-)}$. This requires that the limiter function $\Phi$ satisfy

$$
1 \leq \Phi(\theta) \leq \theta \Phi(\theta^{-1}) \leq \theta \quad \theta \in [1, \infty)
$$

$$
\theta \leq \theta \Phi(\theta^{-1}) \leq \Phi(\theta) \leq 1 \quad \theta \in (-\infty, 1] \setminus \{0\},
$$

which is due to the inverse evaluation of $\theta$ in (4.12).

In Fig. 4.2 we match the new limiter $\Phi(\theta, 0.5, 2, 1.6)$ with the original logarithmic limiter function $\phi_{\log}(\theta, 1.4)$. The limiter reads

$$
\tilde{\phi}(\theta) = \max\left(0, \min\left(\frac{2 + \theta}{3}, \max\left(-0.5 \theta, \min\left(2 \theta, \frac{2 + \theta}{3}, 1.6\right)\right)\right)\right).
$$

It shares the common characteristics of the logarithmic limiter $\phi_{\log}(\theta, 1.4)$ within the simple framework of a piecewise-linear (max, min)-function. Its rigorous cutoffs for $\theta \to \pm 0$ (Fig. 4.2, right) ensure sharp resolution of discontinuous functions. For $\theta \geq 0$ the limiter (4.23) remains in Harten’s TVD region, thus diminishing variation at discontinuities. Monotonicity restrictions, however, require that the slope $\Phi \delta_{i+1/2} = 0$ if the data are not locally monotone, i.e., in the presence of local extrema or discontinuities. Therefore, total variation bounded methods, such as LDLR, LHR, or ENO / WENO-schemes relax the monotonicity conditions, in order to resolve smooth extrema without loss of accuracy.

### 4.4 The Relevancy of Extended Monotonicity Regions

The relevancy of extending Harten’s TVD region for $\theta < 0$ has been controversially discussed by several authors (see, e.g., [12, 22, 54, 64]). We have clearly shown that classical limiter function, which solely depend on $\theta$ as local smoothness indicator alway degenerate accuracy for non-monotone data (see Section 3.3.2). It is the purpose of this section, to examine the implication of different limiter extensions for $\theta < 0$, on the accurate reconstruction of smooth non-monotone functions.

In Section 3.3 we calculated the truncation error for smooth data with a local extremum (3.52)

$$
L_i(\tilde{u}(t)) = L_{ex}(u(\cdot, t), i) + a \mathcal{E}(\xi, \phi) u''(x_{i-\xi}) \Delta x.
$$
We call the function \( E(\xi, \phi) \), error constant

\[
E(\xi, \phi) = \frac{1}{4}(2 + (2 \xi - 1) \phi(a_1(\xi)) - (1 + 2 \xi) \phi(a_2(\xi)))
\]

(4.25)

with the abbreviations (see eqs. (3.53))

\[
a_1(\xi) = \frac{3 - 2 \xi}{1 - 2 \xi}, \quad \text{and} \quad a_2(\xi) = \frac{2 \xi - 1}{1 + 2 \xi}
\]

The error constant \( E(\xi, \phi) \) gives us the possibility to explicitly quantify the local truncation error, depending on the location \( \xi \) of the extremum, and any limiter \( \phi(\theta(\xi)) \). With this we can clarify the relevancy of different limiter extensions. Note that for \( \xi = \pm \frac{1}{2} \), i.e., when the extremum is precisely located at the right, respectively left cell boundary, limiter function, which are solely based upon \( \theta \) will always recover only first order accuracy (see eq. (3.51), Section 3.3.2).

Note that the sign of the error constant must be considered. Assuming a local minimum \( (u''(x_{i-\xi}) > 0) \) and \( E(\xi, \phi) > 0 \), the numerical approximation \( L_i(\bar{u}(t)) \) converges towards \( L^e(\bar{u}(\cdot), i) \) linearly from above. Accordingly the numerical approximation ”overestimates” the exact result for a given grid size \( \Delta x \). We have to remark that for simplicity we assume \( a > 0 \), i.e., the initial data is upwinded from the left to the right. Analogously, if \( a < 0 \) the initial data is downwinded from the right to the left.

### 4.4.1 Extended Piecewise Linear Limiters

The idea of extending limiter functions for \( \theta < 0 \) is indeed not new. Schulz-Rinne 1993 [54] has extended classical second-order accurate limiter functions, such that smooth data for which \( \theta = -1 + \mathcal{O}(\Delta x) \) holds, are accurately reconstructed. He focused mainly on van Leer’s MCD limiter (3.43), which employs monotonized central differences \( \phi(\theta) = \frac{1}{2} (1 + \theta) \) as a building block. Therefore, Schulz-Rinne required, that the extended limiters must pass continuously through \( \phi(-1) = 0 \).

The piecewise-linear limiter functions suggested in [54] are:

- the extended minmod limiter (see (3.41) for the original)

\[
\phi^{\text{EMM}}(\theta) = \text{sign}(1 + \theta) \cdot \min(1, |\theta|, |1 + \theta|)
\]

(4.26)

- the extended superbee limiter (see (3.42) for the original)

\[
\phi^{\text{ESB}}(\theta) = \begin{cases} 
\text{sign}(1 + \theta) \cdot \min(1, |\theta|, |1 + \theta|) & \text{if } \theta < 0 \\
\max(\min(2 \theta, 1), \min(\theta, 2)) & \text{if } \theta \geq 0 
\end{cases}
\]

(4.27)

- the extended monotonized central differences (see (3.43) for the original)

\[
\phi^{\text{EMCD}}(\theta) = \text{sign}(1 + \theta) \cdot \min(2 \theta, \frac{1}{2}(1 + \theta), 2).
\]

(4.28)
Note that none of Schulz-Rinne’s limiters is inside the extended monotonicity region proposed in [12,64] (see also conditions (3.13)). All limiter functions are significantly extended for $\theta < 0$. Thus, e.g., $\phi_{EMCD}$ employs a linear reconstruction for $\theta \in [-5, -1/5]$.

Schulz-Rinne could prove that the new minmod limiter (4.26) is TVD and that a fully discretized FV-scheme employing $\phi_{EMM}$ satisfies a local maximum principle (see Theorems [54] for details). Consequently the method fulfills Harten’s monotonicity restriction and therefore fails to recover second-order accuracy for non-monotone profiles.

In the following, we will compare in several plots the error constant $\mathcal{E}(\xi, \phi)$ for different limiter functions. To understand these, one has to picture an extremum, which is located at different positions $\xi$ of the computational domain. Since the whole scheme utilizes five cells (see Remark 3.1.1), we plot the error constant also outside the centered cell. This way we can explicitly see, when and how, in terms of the location $\xi$, the error constant evolves.

**Comparison of second-order accurate limiters:**

![Comparison of second-order accurate limiters](Image)

**Figure 4.3:** Left: Plot of error constants for TVD second-order accurate limiter (top) and for extended second-order accurate limiter (bottom). Right: Plot of the corresponding limiters: minmod (solid line), superbee (dashed line) and MCD (dotted line).
In Fig. 4.3 we compare \( E(\xi, \phi) \) for both groups (classical and extended) of limiters. We can observe, that all plotted error constants are perfectly symmetric, with the symmetry axis located at \( \xi = 0.5 \). The error constant for the minmod limiters (original version (3.41) and extended version (4.26)) is always positive. Whereas employing \( \varphi^{MCD} \), \( \varphi^{EMCD} \) and \( \varphi^{SB} \), \( \varphi^{EMCD} \) gives also negative values for \( E(\xi, \phi) \).

Let us assume a maximum \( (u''(x_{i-\xi}) < 0) \) with location \( \xi \in [-1.5, 0] \) or \( \xi \in [-1, 2.5] \). A spatial reconstruction \( L_i(\bar{u}(t)) \) utilizing classical second-order TVD limiters yields

\[
L_i^{MM}(\bar{u}(t)) < L^{EX}(u(\cdot, t), i), \quad \text{and} \quad L_i^{SB}(\bar{u}(t)) > L_i^{MCD}(\bar{u}(t)) \geq L^{EX}(u(\cdot, t), i).
\]

If a minimum \( (u''(x_{i-\xi}) > 0) \) has to be approximated inside \( \xi \in [-1.5, 0] \) or \( \xi \in [-1, 2.5] \), the spatial reconstruction \( L_i(\bar{u}(t)) \) yields

\[
L_i^{MM}(\bar{u}(t)) > L^{EX}(u(\cdot, t), i), \quad \text{and} \quad L_i^{SB}(\bar{u}(t)) < L_i^{MCD}(\bar{u}(t)) \leq L^{EX}(u(\cdot, t), i).
\]

Consequently both limiters, \( \varphi^{MCD} \) and \( \varphi^{SB} \) overestimate a local maximum and underestimate a local minimum. This leads for a given \( \Delta x \) to a more compressive resolution of local extrema. The limiter \( \varphi^{MM} \) always under estimates the exact data, which result in numerical diffusion.

Schulz-Rinne’s expansion of the limiter functions for \( \theta < 0 \) changes the error constant \( E(\xi, \phi) \) for extrema located in \( \xi \in [-0.5, 1.5] \setminus \{0.5\} \). We can clearly observe a decrease of \( |E(\xi, \phi)| \) for \( \varphi^{EMM} \) and \( \varphi^{EMCD} \). The most significant improvement is achieved for the \( \varphi^{EMCD} \) limiter. Its error constant is of machine size precision for \( \xi \in [-1/3, 0] \) and \( \xi \in [1, 1/3] \), hence approximating smooth data to almost second-order accuracy. Remember that this is the only limiter, which utilizes second-order accurate central-differences for \( \theta \in [-5, -1/5] \). For the extended superbee limiter \( \varphi^{ESB} \) the function \( E(\xi, \phi) \) becomes once again negative for \( \xi \in ]0, 1/3[ \). This essentially introduces over and under estimation of local maxima and minima, respectively.

Numerical experiments utilizing the extended limiters confirm our observations. Schulz-Rinne [54] could show, that the convergence rates for \( \varphi^{EMCD} \) were superior compared to result obtained with the original limiter \( \varphi^{MCD} \). The limiter function \( \varphi^{EMCD} \), however, achieved also best results in comparison to \( \varphi^{ESB} \) and \( \varphi^{EMM} \). The new superbee limiter \( \varphi^{ESB} \) resolved sharp discontinuities better, than its original counterpart. However smooth extrema were significantly squared utilizing \( \varphi^{ESB} \). The defect of squaring smooth extrema is often called over-compressiveness (see, e.g., [40, 85]).

**Remark 4.4.1 (Compressive Limiter)** Compressive limiter functions, also not clearly defined, are traditionally associated with the choice of a steeper reconstruction at corners of linear discontinuities. This leads to a squaring of smooth data. For TVD limiters the compressive behavior depends on the gradients around zero, \( \frac{d\phi(0^+)}{d\theta} \) and \( \frac{d\phi(0^-)}{d\theta} \).

The compressive behavior of limiters can be directly linked to Harten’s artificial compression method (see [40]). In [40] one can find several examples of over-compressive limiters and numerical test cases,
which clearly envisage the squaring of smooth structures. We should explicitly emphasize that the observed squaring effects of second-order methods are completely diminished in methods of third-order accuracy. This issue of compressiveness will be addressed through out this work and discussed in details in the coming chapters (Chapter 6).

**Comparison of different versions of \( \Phi(\theta, \alpha, 2, 2) \):**

In Fig. 4.4 we compare \( E(\xi, \Phi(\theta, \alpha, 2, 2)) \) for \( \alpha = 0, 0.5, 2 \). Similar to \( \phi^{EMCD} \), we can observe an improvement of the accuracy with both limiter extensions. Note that the absolute value of \( E(\xi, \Phi) \) is smaller than that of \( \phi^{SB} \) and \( \phi^{ESB} \). This is due to the fact, that \( \Phi(\theta, \alpha, 2, 2) \) employes the reconstruction \( \phi = 2\theta \) in \( 0 \leq \theta \leq 0.4 \). Superbee limiters utilize \( \phi = 2\theta \) for \( 0 \leq \theta \leq 0.5 \), which essentially results in an over-compressive spatial approximation.

### 4.4.2 Extended Smooth Limiters

In Section 3.3.1, we introduced van Leer’s limiter \( \phi^{VL} \) (3.44), which is the smooth counterpart of \( \phi^{MCD} \) (3.43). Similarly, there exist smooth counterparts of the Koren-limiter \( \phi^{TVD} \) (4.21), such as

\[
\phi^{KS}(\theta) = \frac{\theta(2 + \theta)}{2 + \theta(2\theta - 1)},
\]

or

\[
\phi^{GPR-1/3}(\theta) = \frac{\theta(2 + \theta)}{1 + \theta + \theta^2}.
\]

The first limiter \( \phi^{KS} \) has been originally derived by Koren [30] and is based upon Spekreijse’s monotonicity constrains [64]. The second limiter \( \phi^{GPR-1/3} \) has been suggested by Waterson and Deconinck [85] and is based upon a generalized polynomial-ratio (GPR) scheme. Both limiter functions satisfy

\[
\phi^{KS}(1 - \Delta\theta) = \phi^{GPR-1/3}(1 - \Delta\theta) = 1 - \frac{\Delta\theta}{3} + \mathcal{O}(\Delta\theta^2),
\]

**Figure 4.4: Plot of error constants of three different version of \( \Phi(\theta, \alpha, 2, 2) \) eq. (4.19): \( \alpha = 0 \) (full line), \( \alpha = 0.5 \) (broken line) and \( \alpha = 2 \) (dotted line).**
Chapter 4. New Third-Order Accurate Limiters

Figure 4.5: Plot of smooth limiter functions: smooth Koren-limiter $\phi^{KS}$, $\phi^{GPR-1/3}$ and $\hat{\phi}(\theta, 1.4)$ (LogLim) inside Harten’s TVD region (yellow).

Figure 4.6: Plot of modified smooth limiter functions: modified smooth Koren-limiter $\phi^{ModKS}$, $\phi^{ModGPR-1/3}$ and $\hat{\phi}(\theta, 1.4)$ (LogLim) inside Harten’s TVD region (yellow).

with $\Delta \theta \equiv \frac{u''(x_i-\xi)}{u'(x_i-\xi)} \Delta x$. This is essential for third-order accuracy for smooth functions. Note, however, that both limiters deviate from the unlimited quadratic interpolation (3.34) at second-order. Whereas the log-limiter (4.13) deviated at fourth-order from the unlimited quadratic interpolation (see Remark 4.2.1). Both limiters, $\phi^{KS}, \phi^{GPR-1/3}$ are continuously differentiable and become negative for $\theta \in [-1, 0]$. To capture smooth extrema, for which holds $\theta = -1 + O(\Delta x)$ to third-order accuracy, a limiter function must pass continuously through $\phi(-1) = \frac{1}{3}$. Since $\phi^{KS}(-1) = -\frac{1}{5}$ and $\phi^{GPR-1/3}(-1) = -1$, both limiters recover only first-order accuracy for smooth extrema, which lateral slopes have identical absolute values, but different signs.

In Fig. 4.5 we envisage the original limiter functions, respectively their modifications. We also plot the log-limiter $\phi^{log}(\theta, 1.4)$ (4.13) to point-out the main differences. All three functions are continuous, yet only $\hat{\phi}(\theta, 1.4)$ passes continuously through $\phi(-1) = \frac{1}{3}$, which eventually guarantees third-order for $\theta = -1$. Note that both, $\phi^{KS}$ and $\phi^{GPR-1/3}$ are inside Harten’s TVD region for $\theta > 0$.

We suggest a simple modification of the limiters $\phi^{KS}$ and $\phi^{GPR-1/3}$, in order to improve the reconstruction
accuracy of smooth extrema, for which \( \theta = -1 + \mathcal{O}(\Delta x) \) holds. The modified limiters (see Fig. 4.6) read

\[
\phi_{\text{ModKS}}(\theta) = \frac{|\theta|(2 + \theta)}{2 + \theta(2\theta - 1)},
\]

(4.32)

and

\[
\phi_{\text{ModGPR-1/3}}(\theta) = \frac{|\theta|(2 + \theta)}{1 + |\theta| + \theta^2}.
\]

(4.33)

Although these limiters are not continuously differentiable at \( \theta = 0 \), they mimic the quadratic reconstruction \( \phi(\theta) = \frac{1}{3}(2 + \theta) \) for \( \theta < 0 \) better, yielding \( \phi_{\text{ModKS}}(-1) = \frac{1}{5} \) and \( \phi_{\text{ModGPR-1/3}}(-1) = \frac{1}{4} \) (see Fig. 4.6). Consequently only the modified GPR limiter (4.33) approximates smooth extrema with different signs with third-order accuracy.

**Comparison of smooth third-order accurate limiters:**

In Fig. 4.7 and Fig. 4.8 (left plots) we can observe the effects of the proposed modification for \( \phi_{\text{ModKS}} \), respectively \( \phi_{\text{ModGPR-1/3}} \). In addition we always show the error constant \( E(\xi, \phi) \) for all limiters, employing no extensions for \( \theta < 0 \). The modification shows its biggest impact for extrema located in \( \xi \in [-1/6, 1/2] \). In this region the error constant of the modified limiters is always smaller than that of the original limiter functions. Especially for the GPR-1/3 limiters, we can observe, that the limiter version without any extension for \( \theta < 0 \) results in smaller errors than the original limiter function (4.30). Note that limiters utilizing no extension for \( \theta < 0 \) are TVD. In Fig. 4.9 we compare the error constants for different versions of the log-limiter \( \phi^{\text{log}}(\theta, q) \). We can observe an improvement of the accuracy, utilizing the limiter \( \phi^{\text{log}}(\theta, 1) \) instead of \( \phi^{\text{log}}(\theta, 2) \). Decreasing the variation control parameter \( q \) leads to an increased local variation (see Rem. 4.2.2 (Local Variation Control Parameter \( q \))).

**Figure 4.7:** Plot of error constants for different versions of the smooth Koren-limiter: \( \phi^{\text{KS}}(\theta) \) (full line), \( \phi_{\text{ModKS}}(\theta) \) (dashed line), \( \phi^{\text{KS}}(\theta) \) without limiting for \( \theta < 0 \) (dotted line). Reference line \( \phi = \frac{1}{3}(2 + \theta) \).
Figure 4.8: Plot of error constants for different versions of the GPR-1/3 limiter: $\phi^{GPR-1/3}(\theta)$ (full line), $\phi^{ModGPR-1/3}(\theta)$ (dashed line), $\phi^{GPR-1/3}(\theta)$ without limiting for $\theta < 0$ (dotted line).

Figure 4.9: Plot of error constants for different versions of the logarithmic limiter: $\hat{\phi}(\theta, 1)$ (full line), $\hat{\phi}(\theta, 2)$ (dashed line), $\hat{\phi}(\theta, 1)$ without limiting for $\theta < 0$ (dotted line). Reference line $\phi = \frac{1}{3}(2 + \theta)$.

Figure 4.10: Sketch of a smooth initial profile and its discrete representation at different times, i.e., different locations on a uniform grid. For simplicity we draw point values instead of cell mean values.
4.5 Geometrical Interpretation of the TVD Conditions

In the previous section (Section 4.3), we derived a new limiter \( \hat{\phi} \) (4.23), which reconstructs data, for which \( \theta = \pm 1 + O(\Delta x) \) holds, utilizing a quadratic polynomial. Unfortunately in this form, the accuracy of the reconstruction still degenerates to first-order in cells with one vanishing lateral derivative. In the following, we will discuss the problem of resolving smooth extrema with limiters for which \( \theta \approx \pm 0 \) hold.

Let us consider a non-monotone sinusoidal initial function localized at different time-levels on an equidistant grid. Fig. 4.10 illustrates the smooth profile, as well as the discrete set of cell mean values \( \{ \bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1} \} \), i.e., the discrete representation of the profile. The reconstruction of the interface values, based upon the three cells is shown at three different times:

- For \( t_1 \) the data are monotone in the interval \([x_{i-1}, x_i, x_{i+1}]\) and \( \theta = 1 + O(\Delta x) \) holds. Every limiter, by definition, employes a high-order reconstruction in this region.

- For \( t_3 \) the extrema is shifted to the neighboring cell. The slope ratio \( \theta \) becomes negative. Limiters, that satisfy Harten’s TVD constrains recover only the cell mean value itself and consequently clip-off local extrema, for which holds \( \theta = -1 + O(\Delta x) \). The extension of limiters has been discussed in the previous section (Section 4.4).

- For \( t_2 \) one of the lateral difference vanishes. Limiters recover, by definition, only first-order accuracy, independently of the value of the second lateral difference.

It is obvious, that between time-level \( t_1 \) and \( t_3 \), the smoothness measure holds \(-1 < \theta < 1\) and passes continuously through zero for \( t_2 \). The purpose of this section is to derive an indicator function, that allows us to distinguish between smooth extrema and discontinuities for which \( \theta \to \pm 0 \) and \( \theta \to \pm \infty \), without abandon the compact reconstruction scheme. We have to formulate a criterion which prohibits a limiter to pass through zero in the presence of a local extremum with one vanishing lateral derivative.

For this purpose we introduce an asymptotic region based on the indicator function.

**Definition 4.5.1 (Smoothness Indicator)** We introduce the function

\[
\eta(\delta_{i-1/2}, \delta_{i+1/2}) = \frac{\delta_{i-1/2}^2 + \delta_{i+1/2}^2}{(r \Delta x)^2} \tag{4.34}
\]

which is a function of the interface differences \( \delta_{i\pm1/2} \), instead of the ratio \( \theta \). In addition \( \eta \) depends on the grid size \( \Delta x \) and a dimensionless constant \( r > 0 \). The indicator function \( \eta \) spans a circle around \( \theta \approx \pm 0 \) with the radius \( r \) (see Fig. 4.11). We will refer to the constant \( r \) as radius of the asymptotic region.

For \( \eta \leq 1 \) the lateral differences \( \delta_{i\pm1/2} \) are too small for \( \theta \) being a good measure for discontinuities. For \( \eta > 1 \), \( \delta_{i\pm1/2} \) are large enough so that a shock capturing limiter can be applied. We define data, for which \( \eta \leq 1 \) hold as smooth and data, for which \( \eta > 1 \) hold as discontinuous. Consequently data which is inside
the asymptotic region is reconstructed to full third-order accuracy, whereas data outside the asymptotic region is passed to the shock capturing limiter \( \hat{\phi} \) (4.23).

**Lemma 4.5.1** Assume a local extremum, i.e., \( \delta_{i-1/2} = O(\Delta x^2) \) and \( \delta_{i+1/2} = O(\Delta x^2) \). Then the indicator function \( \eta \) yields

\[
\eta(O(\Delta x^2), O(\Delta x^2)) = \frac{2\xi^2 + \frac{1}{2}}{r^2} \left( u''(x_{i-\xi}) \right)^2 \Delta x^2 + O(\Delta x^3).
\]  

(4.35)

**Proof.** We utilize the Taylor expansions of the lateral derivatives \( \delta_{i\pm 1/2} \) around \( x_i - \xi \Delta x \) (see Appendix A) and neglect higher-order terms. In the presence of a local extremum \( u'(x_{i-\xi}) = 0 \) and \( u''(x_{i-\xi}) \neq 0 \), we recover eq. (4.35). ■

Consequently the function \( \eta \) is a measure of the second derivative, i.e., of the curvature of the data for \( u'(x_{i-\xi}) = 0 \). Note that \( \eta = O(\Delta x^2) \) which is smaller than the required linear convergence of \( \eta \) for a third-order reconstruction. In this perspective the quadratic convergence is optimal, since one can not expect more from a three point stencil.

For \( \xi = \frac{1}{2} \), the local extremum is precisely located at the right cell boundaries. The indicator functions eq. (4.34) yields

\[
\eta(0, O(\Delta x^2)) = \left( \frac{u''(x_{i-1/2})}{r} \right)^2 \Delta x^2 + O(\Delta x^3).
\]  

(4.36)

In the presence of a local extremum at the cell interfaces the second derivative \( u''(x_{i-1/2}) \) represents a measure of its curvature. We require for third-order accurate reconstruction \( \eta \leq 1 \). Therefore, local extrema for which holds

\[
u''(x_{i-1/2}) \leq \frac{r}{\Delta x}.
\]  

(4.37)

are accurately resolved. Obviously extrema with large curvature could lead to \( \eta > 1 \) on coarse meshes. The radius of the asymptotic region \( r \) can account for large curvatures on coarse grids. Choosing a large radius leads to a faster, in terms of mesh resolution \( \Delta x \), convergence of \( \eta \) towards zero, hence to third-order accurate interpolation.

In the presence of discontinuities the indicator function measures gradients.

**Lemma 4.5.2** Assume that one of the lateral derivatives is an unbounded function of \( O(1/\Delta x) \). For \( \delta_{i-1/2} = 0 \) and \( \delta_{i+1/2} = O(1) \) the indicator function \( \eta \) yields

\[
\eta(0, O(1)) = \left( u'(x_i) \right)^2 \frac{4}{r^2} + O(\Delta x^2).
\]  

(4.38)

**Proof.** For discontinuous data with one vanishing lateral derivative we have

\[
\delta_{i+1/2} = \bar{u}_{i+1} - \bar{u}_i = \bar{u}_{i+1} - \bar{u}_{i-1}, \quad \text{since} \quad \delta_{i-1/2} = \bar{u}_i - \bar{u}_{i-1} = 0.
\]
The proposed indicator function (4.34) yields
\[
\eta(0, \delta_{i+1/2}) = \left( \frac{\delta_{i+1/2}}{r \Delta x} \right)^2 = \frac{4}{r^2} \frac{4}{r^2} = \left( \frac{\bar{u}_{i+1} - \bar{u}_{i-1}}{2\Delta x} \right)^2 = \left( u'(x_i) \right)^2.
\]
(4.39)

Thus we recover eq. (4.38).

We said, that data for which \( \eta > 1 \) holds is discontinuous and must be fully limited. Jump discontinuities for which holds
\[
u'(x_i) > \frac{r}{2},
\]
(4.40)
will eventually be outside the asymptotic region and have to be fully limited. For a small \( r \) already discontinuities with shallow gradients will be immediately limited. It is important to understand that playing with the parameter \( r \), i.e., increasing or decreasing the size of the asymptotic domain clearly affects the amount of introduced variation for a fixed grid size. Finally our new third-order limiter function reads:
\[
\phi^{O(3)}(\delta_{i-1/2}, \delta_{i+1/2}) = \begin{cases} 
\frac{2+\theta}{3} & \text{if } \eta \leq 1 - \epsilon \\
\hat{\phi}(\theta) & \text{if } \eta \geq 1 + \epsilon \\
\frac{1}{2} \left( 1 - \frac{\eta-1}{\epsilon} \right) \frac{2+\theta}{3} + \left( 1 + \frac{\eta-1}{\epsilon} \right) \hat{\phi}(\theta) & \text{else.}
\end{cases}
\]
(4.41)

The number \( \epsilon \) is a small positive number, which is about the size of the particular machine precision and ensures Lipschitz continuity for \( \phi^{O(3)} \). Note that in the proposed limiter (4.41), the problem of accurately

![Figure 4.11](image_url)

**Figure 4.11:** Sketch of new limiter function including the asymptotic region for smooth extrema with one vanishing interface difference \( \delta_{i\pm1/2} \). Red color indicates third order recovery.
recovering smooth extrema with one vanishing lateral derivative and simultaneously reconstructing sharp gradients is rigorously decoupled. Consequently, in the presence of a discontinuity the new limiter yields $\phi^{O(3)} \to \hat{\phi}(0) = 0$ with $O(\Delta x^2)$. This convergence rate towards zero total variation is significantly higher than that of Artebrant and Schroll’s double logarithmic reconstruction (see (4.10)) or Marquina’s hyperbolic reconstruction (see (4.8)).

Fig. 4.11 points out the characteristic properties of the final limiter (4.41). It is, for simplicity, more a sketch rather than an actual plot of the function $\phi^{O(3)}$. This way we can geometrically clarify the asymptotic region $\eta$, the confidence region and the transition domain. In the first two regions the interface values are reconstructed to full third-order accuracy, since we are "confident" that the input data is smooth. Whereas in the transition region the discretized input data is discontinuous and consequently the data has to be fully limited. The lower limit $(1 - \epsilon)$ and upper limit $(1 + \epsilon)$, i.e., the thickness of the dark circle around the asymptotic region ensure Lipschitz continuity of $\phi^{O(3)}$, and so for the numerical flux function $\mathcal{F}(\tilde{u}_{i+1/2}^(-), \tilde{u}_{i+1/2}^(+)) = F_{i\pm1/2}$ (1.10). Away from the asymptotic domain cuts along constant $\delta_{i-1/2}$ correspond to the limiter function Fig. 4.2.

We want to emphasize that in this formulation the asymptotic region only affects smooth functions with one vanishing lateral derivative and is completely decoupled from the shock capturing limiter $\hat{\phi}(\theta)$. Once $\eta > 1 + \epsilon$ the function $\hat{\phi}(\theta)$ is switched on. This limiter satisfies for $\theta > 0$ Harten’s TVD constraints, consequently not producing variation at all. Note that numerical experiments indicate, that for smooth data the region for $\theta \leq -1$ is very seldom evaluated, only when the smooth extremum is precisely located at the cell center. Most of the time the cells for which $\theta \approx \pm 0$ holds are affected. From the reconstruction point of view, this means that these cells have to be identified as fast as possible, when refining cell size $\Delta x$. Therefore, the choice of the radius $r$ is crucial.

**Remark 4.5.1 (Radius of the Asymptotic Region)** The radius $r$, at least on a coarse grid, controls the compressiveness of the indicator function $\eta$. This is critical since discretized shallow gradient or smooth extrema eventually have identical curvature for a certain grid size $\Delta x$. Because of this we recommend, especially for intermediate jump discontinuities on coarse grids, to set $0 < r < 1$.

We prefer to produce as small variation as possible. Numerical experiments, however, indicate that $r > 1$ is possible and for certain examples favored. Eventually the asymptotic region always converges quadratically to the correct limit. This compromise is the ultimate, independent of the reconstruction Ansatz function, for a local three-point-stencil.
Chapter 5

Relation to Previous Work

The challenge of accurately resolving smooth extrema without introducing spurious oscillations has always been a major concern in the computational fluid dynamics community. In 1987 Harten, Enquist, Osher and Chakravarthy [20] have introduced the idea of essentially non-oscillatory (ENO) schemes. These schemes avoid classical limiters. Instead they utilize adaptive stencils, obtaining information from regions of smoothness if discontinuities are present. For third-order accurate reconstruction, ENO schemes construct three parabolas employing a five-point stencil. The smoothness of the data is measured in terms of high-order derivatives. For a detailed review on ENO scheme see, e.g., [60].

At the same time, different attempts have been proposed to improve the accuracy of limiter functions for smooth extrema. The basic idea is to exchange the smooth limiters, by a mesh size dependent function, which in the limit \( \Delta x \to 0 \) recovers the original limiter. To our knowledge this has been done for the first time by van Albada, van Leer and Robert in 1982 [75] and extended by Venkatakrishna [81–83].

5.1 Smooth Second-Order Accurate Reconstruction

We consider here the van Albada limiter

\[
\phi^{vA}(\theta) = \frac{\theta^2 + \theta}{1 + \theta^2}.
\] (5.1)

This limiter is a continuously differentiable second-order accurate function, for which holds

\[
|\phi^{vA}(1 - \Delta \theta) - \phi^{\text{MCD}}(1 - \Delta \theta)| = \mathcal{O}(\Delta \theta^2), \quad \text{with} \quad \Delta \theta = \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \Delta x.
\] (5.2)

Identically to the smooth van Leer limiter (3.44) (see also eq. (3.45)), \( \phi^{vA} \) converges to Fromm’s scheme (3.30) for \( \theta = 1 + \mathcal{O}(\Delta x) \). A simple modification of this limiter

\[
\phi^{\text{ModvA}}(\theta) = \text{sign}(\theta) \frac{\theta^2 + \theta}{1 + \theta^2}
\] (5.3)
essentially improves the accuracy for $\theta = -1 + O(\Delta x)$. This modification, however, does not avoid the accuracy reduction for smooth extrema for which $\theta \approx 0$ holds (see Section 3.3.2).

To avoid accuracy reduction for $\theta = 0$, one has to reformulate the limiter into a mesh size dependent function. Accordingly the reconstruction of the cell interfaces reads

\[
\begin{align*}
\hat{u}_{i+1/2}^{(-)} &= \bar{u}_i + \frac{1}{2} \Pi \left( \delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x) \right) \\
\hat{u}_{i-1/2}^{(+)} &= \bar{u}_i - \frac{1}{2} \Pi \left( \delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x) \right)
\end{align*}
\]

(5.4)

where $\epsilon$ is actually a small number. Van Albada et al [75] suggested the smooth function

\[
\Pi^{VA} \left( \delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x) \right) = \frac{(\delta_{i+1/2}^2 + \epsilon^2) \delta_{i-1/2} - (\delta_{i-1/2}^2 + \epsilon^2) \delta_{i+1/2}}{3 \delta_{i+1/2}^2 + 3 \delta_{i-1/2}^2 + 6 \epsilon^2},
\]

(5.5)

where

\[
\epsilon^2(\Delta x) = (K \Delta x)^3
\]

(5.6)

is a small number, the so called smallness parameter, with $0 \leq K < 1$. The parameters in front of $\epsilon$ are derived, imposing second-order accuracy conditions on the cell interface reconstruction, i.e., $\hat{u}_{i+1/2}^{(-)} = u(x_i + \Delta x/2) + O(\Delta x^2)$. Accordingly for monotone data the truncation error yields

\[
E_{\Delta x}(i - \xi, t) = |a| \frac{1}{12} u^{(3)}(x_{i-\xi}) (\Delta x)^2 + O(\Delta x^3).
\]

(5.7)

For $K = 0$ we recover the original limiter (5.1). For $K > 0$ we have a scheme, which is essentially not limiter based. It is not possible to extract from (5.5) a function of the form $\phi(\theta)$. Consequently TVD constrains are not applicable. Yet we can still find some estimates on the LV of the reconstruction. The general reconstruction function for 5.4 reads

\[
\begin{align*}
\hat{r}_i(x) &= \bar{u}_i + \frac{1}{\Delta x} \Pi \left( \delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x) \right) (x - x_i).
\end{align*}
\]

(5.8)

Assuming a monotonic profile, the local variation of the reconstruction function 5.8 reads

\[
\begin{align*}
\text{LV}(r_i) &= \int_{x_{i+1/2}}^{x_{i+1/2}} |r'_i(x)| dx \\
&= |\Pi \left( \delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x) \right)|.
\end{align*}
\]

(5.9)

Lemma 5.1.1 Assuming $\delta_{i-1/2} = O(\Delta x)$ and $\delta_{i+1/2} = O(1)$, then the modified van Albada reconstruction (5.5) is locally total variation bounded.

Proof. With $\delta_{i-1/2} = M \Delta x$ and $\delta_{i+1/2} = N$, we obtain from eq. (5.9)

\[
\text{LV}(r_i) = |\Pi^{VA} \left( N, M \Delta x, (K \Delta x)^{3/2} \right)| = |M| \Delta x + O(\Delta x^2)
\]

(5.10)

Hence condition (4.2) holds for all $i$ and the reconstruction is locally total variation bounded.
Note that linear convergence of $\text{LV}(r_i)$ featuring reconstruction (5.8) is eventually given by definition. The smooth function $\Pi(\delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x))$ is constructed such that the enumerator is always one order of $\Delta x$ larger than the denominator. Remark that reconstruction (5.8) is less oscillatory than Marti-
quina’s LHR (4.8).

Lemma 5.1.2 Assume smooth non-monotone data and $\epsilon(\Delta x) = (K \Delta x)^3/2$, then the modified van Al-
bada reconstruction (5.5) is second-order accurate and the local truncation error reads

$$E_{\Delta x}(i - \xi, t) = |a| \frac{1}{12} \left( 3 \left( \frac{u''(x_i - \xi)}{K} \right)^3 + u(3)(x_i - \xi) \right) \Delta x^2 + \mathcal{O}(\Delta x^3).$$

(5.11)

Proof. We utilize the Taylor expansions of the cell interface differences $\delta_{i \pm 1/2}$ around $x_i - \xi \Delta x$ (see Appendix A) and neglect higher-order terms. In the presence of a local extremum $u'(x_i - \xi) = 0$ and $u''(x_i - \xi) \neq 0$ somewhere inside a FV-cell $\xi \in (-1/2, 1/2]$ we recover with $E_{\Delta x}(i - \xi, t)$ (3.19) second-
order accuracy for the linear advection equation. ■

Note that for $\epsilon(\Delta x) = (K \Delta x)$, the truncation error analysis yields

$$E_{\Delta x}(i - \xi, t) = |a| \frac{1}{12} u(3)(x_i - \xi) \Delta x^2 + \mathcal{O}(\Delta x^3),$$

(5.12)

and we recover the local variation $\text{LV}(r_i) = \left| \frac{(K + M N)\Delta x}{2N} \right|$.

It is obvious, that in order to recover smooth extrema accurately already on coarse grids, one would prefer to choose the small number to be of order $\Delta x$. This, however, results eventually in larger variation in the presence of discontinuities. It is also interesting, that for $\epsilon(\Delta x) = (K \Delta x)^n$ with $n \geq 2$, only first-order accuracy for non-monotone functions is recovered. Therefore, the optimal choice for $\epsilon(\Delta x)$ is (5.6). Van Albada suggests to set $K = 0.3$, although a larger value would improve the accuracy of non-monotone data.

In the following, we want to compare the accuracy improvement of the modified reconstruction 5.5 with the original van Albada limiter (5.1), which is essentially TVD.

Consider numerical convergence experiments for the linear advection equation (for simplicity $a = 1$) with initial function $u_0(x) = \sin^4(\pi x)$ in the domain $x \in [0, 1]$. The initial function has a local maximum at $x_0 = 0.5$. The absolute value of the truncation error of the modified van Albada (5.11) with $K = 0.3$ yields,

$$|E_{\Delta x}(i, t)| = 5.7 \cdot 10^5 \Delta x^2.$$

(5.13)

The classical van Albada limiter $\phi^{VA}$ (5.1) clips of smooth extrema and recovers only first-order accuracy. We get the truncation error (see Section 3.3.2, eq. (3.52))

$$E_{\Delta x}(i - \xi, t) = \tilde{E}(\xi, \phi^{VA}) \Delta x,$$

(5.14)
Chapter 5. Relation to Previous Work

with the error function

\[ \tilde{E}(\xi, \phi^A) = \frac{4\pi^2 (1 - 4(-1 + \xi)\xi)}{(5 + 4(\xi - 2)\xi)(1 + 4\xi^2)} \]  

(5.15)

The error constant \( \tilde{E}(\xi, \phi^A) \) is a continuous function of \( \xi \) and has a local maximum at the cell boundary \( \xi = 0.5 \) with magnitude \( 2\pi^2 \). Comparing both errors, we see that the modified scheme (5.5) requires sufficient high resolution, i.e.,

\[ \Delta x \leq \frac{2\pi^2}{5.7} \cdot 10^{-5}, \]  

(5.16)

to accurately recover smooth extrema. Numerical experiment in [82] confirm this observation. Note that here we consider only local truncation errors of the spatial approximation. Time marching schemes, in particular numerical stability, will be discussed in the next chapter.

5.2 Smooth Third-Order Accurate Reconstruction

To obtain smooth, third-order accurate and local variation bounded reconstructions, we can simply modify the continuous differentiable Koren-limiter (4.29) and the GPR-1/3 limiter (4.30). Consequently we recover the following cell interface reconstructions

\[ \hat{u}_{i+1/2}^{(-)} = \bar{u}_i + \frac{1}{2} \Pi (\delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x)) \]
\[ \hat{u}_{i-1/2}^{(+)} = \bar{u}_i - \frac{1}{2} \Pi (\delta_{i-1/2}, \delta_{i+1/2}, \epsilon(\Delta x)) \]  

(5.17)

with the modified limiters

\[ \Pi^{KS}(\delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x)) = \frac{2 \delta_{i+1/2}^2 + \epsilon^2}{2 \delta_{i+1/2}^2 - \delta_{i+1/2} \delta_{i-1/2} + 2 \delta_{i-1/2}^2 + 3 \epsilon^2} \delta_{i+1/2} \]  

(5.18)

and

\[ \Pi^{GPR-1/3}(\delta_{i+1/2}, \delta_{i-1/2}, \epsilon(\Delta x)) = \frac{2 \delta_{i+1/2}^2 + \epsilon^2}{\delta_{i+1/2}^2 + \delta_{i+1/2} \delta_{i-1/2} + \delta_{i-1/2}^2 + 3 \epsilon^2} \delta_{i+1/2} \]  

(5.19)

where \( \epsilon(\Delta x) = (K \Delta x) \) with \( K \geq 0 \). The first reconstruction (5.18) has been originally suggested by Venkatakrishna [81]. The second function is a new reconstruction. Identically to the discussion in the previous section (Section 5.1), we recover for \( K = 0 \) the original smooth limiters. Note that in contrast to van Albada’s modification (5.4), the third-order accurate reconstruction (5.17) utilizes a second evaluation of \( \Pi \). Thus optimizing the use of three data points (compare also reconstruction (3.36) and (3.39), Remark 3.3.1) For \( K > 0 \) we essentially have ENO like third-order accurate schemes. The truncation error analysis for both reconstructions assuming monotone data yields

\[ E_{\Delta x}(i - \xi, t) = |a| \left( C - \frac{u^{(4)}(x_i - \xi)}{12} \right) \Delta x^3 + O(\Delta x^4), \]  

(5.20)
with the abbreviation
\[ C = \frac{u''(x_i - \xi)}{c} \left( \frac{K^2 - (u'(x_i - \xi))^2}{(u''(x_i - \xi))^2} \right) \left( \frac{2u'(x_i - \xi)}{K^2 + (u'(x_i - \xi))^2} \right) \frac{u^{(3)}(x_i - \xi)}{c(K^2 + (u'(x_i - \xi))^2)^2}. \]  
(5.21)

For reconstruction (5.18) \( c = 3 \) and for (5.19) \( c = 6 \). In contrast to the second-order accurate reconstruction (5.5), the truncation error depends on all derivatives of the smooth monotone data. This might eventually lead for certain initial data to large truncation errors.

For non-monotone data, the truncation analysis for both reconstruction yields
\[ E_{\Delta x}(i - \xi, t) = |a| \left( \frac{(u''(x_i - \xi))^3}{cK^2} - \frac{u^{(4)}(x_i - \xi)}{12} \right) \Delta x^3 + O(\Delta x^4). \]  
(5.22)

For reconstruction (5.18) \( c = 3 \) and for (5.19) \( c = 6 \). Consequently we reconstruct local extrema with third-order accuracy. This result is similar to the truncation error of the modified van Albada limiter (5.11), yet the effect of the parameter \( K \) is less profound. Note that both schemes are local variation bounded (compare with (5.10)).

In order to achieve full third-order accuracy for non-monotone data, the choice of the smallness parameter \( \epsilon(\Delta x) \) is crucial. Venkatakrishna [81] originally assumes \( \epsilon(\Delta x) \sim O(\Delta x^{n/2}) \) with \( n > 2 \) and fixes \( n = 2.5 \) in the numerical computations. In fact with this choice the truncation error for non-monotone data yields
\[ E_{\Delta x}(i - \xi, t) = |a| \frac{(u''(x_i - \xi))^3}{3K^{5/2}} \Delta x^{5/2} + O(\Delta x^3), \]

hence degenerating accuracy in the presence of local extrema. Truncation error analysis reveals that only with \( 0 < n \leq 2 \), third-order accuracy for non-monotone data can be achieved. In particular, for \( n = 1 \) the truncation error for monotone and non-monotone data is identical yielding
\[ E_{\Delta x}(i - \xi, t) = -|a| \frac{u^{(4)}(x_i - \xi)}{12} \Delta x^3 + O(\Delta x^4). \]

Consequently third-order accuracy is recovered without possible accuracy degeneration due to higher derivatives of the data, yet more local variation is introduced. It is a fundamental design principle of these modifications, that the smaller the parameter \( \epsilon(\Delta x) \), the less variation gets introduced, and simultaneously more numerical diffusion is produced.

### 5.3 LDLR Smoothness Measure

We could identify also a similar drawback in Artebrant and Schroll’s [5] LDLR scheme. Although LDLR is essentially a limiter free reconstruction, it still needs a local indicator function to distinguish between
smooth and discontinuous data. Artebrant and Schroll [5] suggest a mesh size dependent modification of the function \( a(\delta_{i-1/2}, \delta_{i+1/2}) \) (see Appendix B, eq. (B.2))

\[
\hat{a}(d^-, d^+, \epsilon(\Delta x)) = (1 - \epsilon) \left( 1 + \epsilon - \frac{|d^-|^q |d^+|^q + \epsilon}{|d^-|^{2q} + |d^+|^{2q} + \epsilon} \right) \epsilon(\Delta x) = 0.1 \Delta x^q, \quad (5.23)
\]

with the lateral derivatives \( d^- = \frac{\delta_{i-1/2}}{\Delta x} \) and \( d^+ = \frac{\delta_{i+1/2}}{\Delta x} \). Therefore, the original formulation has been softened similarly to the modification discussed in the previous sections. This way, not only divisions through zero are avoided (see Appendix B, eq. (B.5)), furthermore smooth non-monotone data can be accurately resolved. We want to emphasize that without this modification LDLR cannot accurately approximate smooth extrema with one vanishing lateral derivative. In the presence of an extremum located at the right cell boundary, the smoothness indicator yields

\[
\hat{a}(0, O(\Delta x), 0.1 \Delta x^q) = \left( 0.1 + \frac{(u''(x_{i-1/2}))^{2q}}{0.1} \right) \Delta x^q + O(\Delta x^{2q}). \quad (5.24)
\]

Depending on the parameter \( q \), \( \hat{a}(0, O(\Delta x), 0.1 \Delta x^q) \) converges faster or slower to zero.

When the function \( \hat{a}(0, O(\Delta x), 0.1 \Delta x^q) \) under-runs a certain mesh size independent limit, LDLR switches Lipschitz continuously to a quadratic reconstruction. This conditional threshold allows Artebrant and Schroll to accurately approximate smooth extrema. The larger the variation control parameter \( q \), the faster, in terms of resolution \( \Delta x \), LDLR recovers smooth extrema. Note that for \( q = 1 \) we have only linear convergence. For extrema with large curvature, the error constant \( \mathcal{E}(\hat{a}) \) dominates the convergence of \( \hat{a}(0, O(\Delta x), 0.1 \Delta x^q) \). It simply prohibits the function to reach the conditional threshold of third-order reconstruction. Consequently for \( q = 1 \), LDLR needs sufficient fine resolution to resolve smooth extrema with one vanishing lateral derivative, yet simultaneously produces large variations for discontinuities (see Remark 4.2.2).

All the discussed drawbacks are completely avoided in the proposed reconstruction \( \phi^{O(3)} \) (4.41). The approximation of smooth extrema with one vanishing lateral derivative is completely decoupled from the shock capturing limiter \( \hat{\phi} \). Furthermore the proposed indicator function \( \eta \) (4.34) converges always quadratically, consequently optimizing the use of a three-point stencil.
Chapter 6

Time Integration

Since numerical errors in time and space discretization can be swapped, we need to adjust the time-marching scheme for semi-discrete high-order methods. We consider here only explicit strong stability preserving (SSP) Runge-Kutta time marching methods. SSP schemes are integration methods for ODEs, which preserve the stability properties of first-order Euler time integration (2.1). These methods have been originally named TVD schemes (see, e.g., [59,62]). However, their strong stability properties hold in any norm (not only in the TVD norm), therefore they are more often termed SSP Runge-Kutta schemes [18]. They consist of convex combination of the forward Euler operator, thus they are stable under the identical time step restriction, i.e., Courant number $\nu \leq 1$ (see (1.24)).

Storage is usually an important consideration for large scale scientific computing. Therefore, low-storage SSP Runge-Kutta methods, which only require two memory registers per ODE variable are desirable. A general objective is to find SSP Runge-Kutta methods that are high-order accurate, have low computational cost, respectively storage requirements and are stable for preferably large CFL coefficient. Such schemes are generally termed optimal SSP Runge-Kutta methods (see, e.g., [18] and references therein for a comprehensive review on SSP methods).

Through out this thesis we consider second-order and third-order accurate spatial reconstructions. For second-order accurate semi-discrete schemes, the optimal method is given by:

\[
\begin{align*}
\bar{u}_i^{(1)} & = \bar{u}_i^n + \Delta t L_i(\bar{u}^n) \\
\bar{u}_i^{n+1} & = \frac{1}{2} \bar{u}_i^n + \frac{1}{2} \bar{u}_i^{(1)} + \Delta t L_i(\bar{u}^{(1)}).
\end{align*}
\]  

(6.1)

Third-order accuracy in time is accomplished by integrating the semi-discrete finite volume update (1.9) with

\[
\begin{align*}
\bar{u}_i^{(1)} & = \bar{u}_i^n + \Delta t L_i(\bar{u}^n) \\
\bar{u}_i^{(2)} & = \frac{3}{4} \bar{u}_i^n + \frac{1}{4} \bar{u}_i^{(1)} + \frac{1}{4} \Delta t L_i(\bar{u}^{(1)}) \\
\bar{u}_i^{n+1} & = \frac{1}{3} \bar{u}_i^n + \frac{2}{3} \bar{u}_i^{(2)} + \frac{2}{3} \Delta t L_i(\bar{u}^{(2)}).
\end{align*}
\]  

(6.2)
Both SSP Runge-Kutta time marching schemes have been suggested in [62]. We will in the following refer to the second-order integrator (6.1) as Heun scheme. The third-order SSP Runge-Kutta (SSP33) method (6.2) is the most popular ODE solver used for third- and higher-order semi-discrete FV method. However, Ketchenson [28] has recently suggested several low storage third-order accurate SSP Runge-Kutta method, which are optimal, yet utilize more than three stages. We consider here only the four-stage SSP Runge-Kutta method (SSP43) originally reported in [31]:

\[
\begin{align*}
\bar{u}_i^{(1)} &= \bar{u}_i^n + \Delta t L_i(\bar{u}_i^n) \\
\bar{u}_i^{(2)} &= \bar{u}_i^n + \Delta t L_i(\bar{u}_i^{(1)}) \\
\bar{u}_i^{(3)} &= \frac{2}{3} \bar{u}_i^n + \frac{1}{2} \bar{u}_i^{(2)} + \frac{1}{6} \Delta t L_i(\bar{u}_i^{(2)}) \\
\bar{u}_i^{n+1} &= \bar{u}_i^{(3)} + \frac{1}{2} \Delta t L_i(\bar{u}_i^{(3)}).
\end{align*}
\]

(6.3)

The purpose of this section is to perform a linear stability analysis and discuss the issue of compressive limiters. Moreover we will determine the accuracy for several schemes in the frequency domain introducing the concept of modified equation.

### 6.1 Von Neumann Stability Analysis

The explicit high-order SSP Runge-Kutta schemes consist of multiple combinations of a single Euler’s step

\[
\begin{align*}
u_i^{n+1} &= \frac{\Delta t L u_i^n}{1 + \Delta t L} = M(\Delta t L) u_i^n.
\end{align*}
\]

(6.4)

This is a system of explicit linear equations with the matrix operator \(L\) containing the non-linear spatial approximation for the discrete set of values \(u_i^{n+1} \equiv \{u_i^{n+1}, n = \ldots 0, 1, 2, \ldots \}\), i.e., \(L u_i^n = L_i(u_i^n)\).

For the explicit Euler time-marching scheme the operator \(M\) reads

\[
M(\Delta t L) = I + \Delta t L,
\]

(6.5)

and for the explicit three-stage SSP Runge-Kutta method

\[
M(\Delta t L) = I + \Delta t L + \frac{1}{2}(\Delta t L)^2 + \frac{1}{6}(\Delta t L)^3.
\]

(6.6)

Note \(M\) is a polynomial in the spatial operator \(\Delta t L\). Thus for a fourth-order accurate four-stage Runge-Kutta scheme, we simply add \(\frac{1}{24}(\Delta t L)^4\). Yet no explicit optimal fourth order SSP method which consist of repetitions of forward Euler steps has been found [31]. For the explicit four-stage third-order accurate optimal SSP scheme (6.3), the operator \(M\) reads

\[
M(\Delta t L) = I + \Delta t L + \frac{1}{2}(\Delta t L)^2 + \frac{1}{6}(\Delta t L)^3 + \frac{1}{48}(\Delta t L)^4.
\]

(6.7)
6.1. Von Neumann Stability Analysis

To investigate the linear stability of the algorithm we assume $L$ to be linear. We consider the time evolution of a sinusoidal grid function in an unbounded domain, governed by the scalar linear advection equation (1.4) with $a > 0$

$$u_t + a u_x = 0, \quad u_k(x, 0) = v_k(0) e^{ikx}, \quad x \in (-\infty, \infty). \quad (6.8)$$

The function $v_k(t)$ is the Fourier transform of the exact solution, $k = 2\pi/\lambda$ is the wave number and $i$ represent the imaginary unit. Discretizing eq. (6.8) on a grid with uniform spacing $x_i = i \Delta x$ and $t^n = n \Delta t$ we obtain a system of differential equations (6.4)

$$\frac{du_i}{dt} = L_i(u^n). \quad (6.9)$$

This is the semi-discretization discussed previously (see Chapter 1, eq. (1.9)). The operator $L_i(u^n) = Lu^n$ contains the linear differential spatial approximations. The sinusoidal function

$$u_k(x_i, t) = v_k(t) e^{ikx} \quad (6.10)$$

is a solution of the semi-discretization (6.9) provided that

$$\frac{dv_k}{dt} = L(e^{ikx}) v_k. \quad (6.11)$$

The functions $L(e^{ikx})$, are the spectral functions [84] of the spatial operator $L$. They are the corresponding eigenvalues $\lambda(\tilde{k})$ of the spatial approximation. The parameter $\tilde{k} = k \Delta x$ is the so-called reduced wave number.

We find by analytic integration of (6.11) that the complex amplitude of the solution at time $t$ is given by

$$v_k(t) = v_k(0) e^{-i \frac{\Delta x}{\Delta t} \Psi(\tilde{k}) t}, \quad (6.12)$$

where the function

$$\Psi(\tilde{k}) = \frac{1}{\tilde{k}} \frac{\Delta x}{a} \lambda(\tilde{k}) \quad (6.13)$$

is called the modified wave number [48]. Hence, the sinusoidal solution of semi-discretization reads

$$u_k(x_i, t) = v_k(0) e^{ik(x_i - \frac{a}{\Delta x} \Psi(\tilde{k}) t)}. \quad (6.14)$$

Likewise, the solutions of the exact evolution equation (6.8) are sinusoidal functions

$$u_k(x, t) = v_k(0) e^{ik(x - at)}. \quad (6.15)$$

Such functions considering one wave, i.e., one $k$ at a time, are called sinusoidal trail solution [84]. The difference between the discrete and the continuous solutions, provide information of the (spectral) properties of a particular spatial approximation used for the semi-discretization of eq. (6.8). The real part of the modified wave number $\Psi(\tilde{k})$ is associated with the phase speed, i.e., with the dispersion properties of the spatial discretization, while the imaginary part of the modified wave number $\Psi(\tilde{k})$ is related to its
numerical dissipation.

In order to get information on the full discretization, utilizing a particular time marching scheme, we can calculate the complex amplitude, i.e., the amplification factor of a numerical scheme

\[
\frac{v_k^{n+1}}{v_k^n} = M(\Delta t \lambda(\tilde{k})) \equiv z(\tilde{k}).
\]  

(6.16)

This amplification factor is an approximation of the amplification factor of the exact solution of the advection equation (6.8):

\[
z^{ex}(\tilde{k}) = \frac{u_k(x, \Delta t)}{u_k(x, 0)} = e^{i \nu k \Delta x},
\]  

(6.17)

with the Courant number

\[\nu \equiv \frac{a \Delta t}{\Delta x}. \]  

(6.18)

Numerical stability, in the von Neumann sense, implies that the absolute value of the amplification factor \(z(\tilde{k})\) should not exceed unity. Hence the characteristic equation for the explicit three-stage third-order SSP Runge-Kutta method reads

\[
|z(\tilde{k})| = |1 + \Delta t \lambda(\tilde{k}) + \frac{1}{2} (\Delta t \lambda(\tilde{k}))^2 + \frac{1}{6} (\Delta t \lambda(\tilde{k}))^3| \leq 1.
\]  

(6.19)

Since \(\lambda(\tilde{k})\) are the eigenvalues of the spectral functions of a particular spatial approximations \(L\), we can obtain stability restriction for the whole scheme. Applying the general compact \(\kappa\) reconstruction \(\phi^\kappa\) (3.35) to the advection equation (6.8), the semi-discretization reads

\[
\frac{d}{dt} u_i = \frac{a}{4\Delta x} (3(\kappa - 1)u_i + (\kappa - 1)u_{i-2} + (5 - 3\kappa)u_{i-1} - (1 + \kappa)u_{i+1}).
\]  

(6.20)

This corresponds to the complex spectrum

\[
\Lambda(\nu, \tilde{k}, \kappa) \equiv \Delta t \lambda(\tilde{k}) = -\frac{1}{2} e^{-i\tilde{k}} \left(-1 + e^{i\tilde{k}}\right) \nu (2 - \kappa + \kappa \cos(\tilde{k}) + i \sin(\tilde{k})),
\]  

(6.21)

Thus, depending on \(\kappa \in [-1, 1]\) (see Remark 3.3.1 (Limiter Accuracy)), we recover the complex spectrum for all second-order accurate schemes utilizing a weighted linear combination of \(\{u_{i-2}, u_{i-1}, u_i, u_{i+1}\}\). For \(\kappa = \frac{1}{3}\) we obtain the spectrum for the unique quadratic interpolation (3.34)

\[
\Lambda(\nu, \tilde{k}, 1/3) = -\frac{1}{6} \nu e^{-i\tilde{k}} (e^{i\tilde{k}} - 1)(5 + \cos(\tilde{k}) + 3i \sin(\tilde{k})).
\]  

(6.22)

Inserting eq. (6.22) into the characteristic equation of the third-order SSP Runge-Kutta method (6.19) yields the amplification factor of the Fourier mode. Similarly we obtain the amplitude of the Fourier mode for the scheme utilizing the second-order time marching method (6.1) or the four-stage third-order accurate time marching scheme (6.3), respectively.

We have plotted the absolute value of the amplification \(|z(\tilde{k})|\) for several schemes. In Fig. 6.1 one can see the amplitude of the Fourier mode for the Heun scheme (6.1), utilizing a linear \(\kappa = 0\) (left plot) and a
6.1. Von Neumann Stability Analysis

\[ n = 0.1 \]
\[ n = 1.0 \]
\[ n = 0.5 \]
\[ n = 1.05 \]

0.0 0.5 1.0 1.5 2.0 2.5 3.0
0.0 0.2 0.4 0.6 0.8 1.0 1.2

\[ \text{reduced wavenumber, } kDx \]
\[ \text{amplification factor, } |zHkDxL| \]

\[ n = 0.1 \]
\[ n = 0.5 \]
\[ n = 1.0 \]
\[ n = 1.1 \]

0.0 0.5 1.0 1.5 2.0 2.5 3.0
0.0 0.2 0.4 0.6 0.8 1.0 1.2

\[ \text{reduced wavenumber, } kDx \]
\[ \text{amplification factor, } |zHkDxL| \]

Figure 6.1: Stability region of explicit second-order Runge-Kutta method time integration (6.1) combined with linear \( \kappa = 0 \) (left plot), and quadratic \( \kappa = 1/3 \) (right plot) spatial reconstruction. Time step restriction: |ν| ≤ 1.

All plots indicate that the schemes are less dissipative for very small Courant numbers |ν| ≤ 0.1. However, using the third-order time integration, the dissipation of high frequency waves gets reduced with large Courant numbers |ν| ≥ 1.0 (Fig. 6.2). Furthermore the third-order time marching scheme gets unstable for high frequency waves, but remains stable for small frequencies. Whereas for second-order accurate spatial reconstruction, basically all Fourier modes get unstable. The Fourier modes in the solution have in general the largest amplitude at low wave numbers.

From the numerical evaluation of the wave spectrum, we conclude that the explicit three-stage SSP method is stable for all values of Courant number |ν| ≤ 1.63 and the four-stage SSP scheme remains

Figure 6.2: Stability region of explicit three-stage (6.2) (left plot) four-stage (6.3) (right plot) third-order Runge-Kutta time integration scheme (SSP43 and SSP33), combined with quadratic spatial reconstruction. Time step restriction for SSP33: |ν| ≤ 1.63, for SSP43: |ν| ≤ 2.0.
Chapter 6. Time Integration

stable for all values of Courant number $|\nu| \lesssim 2.0$. Note that using only a second-order time marching scheme plus the third-order spatial reconstruction we would obtain a more restrictive CFL condition $|\nu| \lesssim 0.83$.

Until now we have determined the amplification factor of Fourier modes of different wave-numbers, assuming unlimited linear or quadratic spatial approximation. Piecewise linear limiter functions, however, utilize different schemes, depending on the smoothness, i.e., $\theta$, of the data. Since we consider the time evolution of a sinusoidal grid function, any classical limiter would clip-off its extrema utilizing the reconstruction $\phi(\theta) = \beta \theta$ with $1 \leq \beta \leq 2$. (see Sweby’s respectively Harten’s TVD region). The complex spectrum for $\phi(\theta) = \beta \theta$ reads

$$\tilde{\Lambda}(\nu, \tilde{k}, \beta) = \frac{1}{2} \left( -2 - \beta - e^{-2i\tilde{k} \beta} + 2e^{-i\tilde{k}(1 + \beta)} \right) \nu$$

(6.23)

For $\beta = 1$ we recover the linear-upwind interpolation (3.31) with the complex spectrum

$$\tilde{\Lambda}(\nu, \tilde{k}, 1) = \Lambda(\nu, \tilde{k}, -1) = \frac{1}{2} \left( -3 + 4e^{-i\tilde{k}} - e^{-2i\tilde{k}} \right) \nu.$$  

(6.24)

In Fig. 6.3 we observe that a scheme, which employes $\phi(\theta) = 2\theta$ is absolutely unstable. Independent of the time integrator, the amplification factor always exceeds one, for any CFL number. Employing $\phi(\theta) = \theta$ for the spatial reconstruction, a FV scheme remains stable, for all values of $|\nu| \lesssim 0.5$ for second-order time integration (6.1) and for all values of $|\nu| \lesssim 0.63$ for third-order accurate time integration schemes (6.2). Obviously, a scheme does not utilize $\phi(\theta) = 2\theta$ for all times. Still this analysis confirms that compressive limiter function introduce local instabilities, which eventually give dispersive errors and lead to squaring of smooth structures. These instabilities get essentially compensated by the SSP properties of the time marching scheme, such that the consistent scheme still converges.

To get detailed information on the dispersion and dissipation error of a particular spatial reconstruction, we have calculated the real part respectively the imaginary part of the modified wave number $\Psi(\tilde{k})$ (6.13).

![Figure 6.3: Stability region of explicit second-order Heun (6.1) (left plot) and third-order SSP33 (6.2) (right plot) time integration scheme combined with the spatial reconstruction $\phi(\theta) = \beta \theta$ for $\beta = 1, 2$. Time step restriction for Heun with $\theta = 1$: $|\nu| \leq 0.5$, for SSP33: $|\nu| \leq 0.63$.](image-url)
6.2. Error Analysis in Frequency Domain

To recover the formal order of accuracy of the whole scheme we perform an error analysis in the frequency domain. In contrast to the truncation error analysis in Section 3.3, Chapter 3, we consider here the whole discretization, including a particular time integration. This procedure shows, which model equation is approximated by the scheme to a leading frequency order. The model equation is usually termed modified...
equation. The modified equation is an continuous equivalent PDE, eventually modeled by a particular numerical scheme applied to the advection equation (1.4).

We consider the ratio of the numerical amplification factor and the exact one:

\[ \mathcal{R}(\tilde{k}) = \frac{z(\tilde{k})^{nu}}{z(\tilde{k})^{ex}}. \]  \hfill (6.25)

With a series expansion for \( \Delta x \to 0 \) we recover for \( \phi \) Eq. (3.35) combined with:

- **second-order Heun time integrator (6.1)**

  \[ \mathcal{R}(\nu, \tilde{k}, \kappa)^{\text{Heun}} - 1 = -\frac{1}{12} i \nu (1 - 3\kappa + 2\nu^2) \tilde{k}^3 + \mathcal{O}(\tilde{k}^4) \]  \hfill (6.26)

- **SSP33 Runge-Kutta time integrator (6.2)**

  \[ \mathcal{R}(\nu, \tilde{k}, \kappa)^{\text{SSP33}} - 1 = \frac{1}{12} i (-1 + 3\kappa) \nu \tilde{k}^3 - \frac{1}{24} (\nu (3 - 3\kappa + \nu^3)) \tilde{k}^4 + \mathcal{O}(\tilde{k}^5) \]  \hfill (6.27)

- **SSP43 Runge-Kutta time integrator (6.3)**

  \[ \mathcal{R}(\nu, \tilde{k}, \kappa)^{\text{SSP43}} - 1 = \frac{1}{12} i (-1 + 3\kappa) \nu \tilde{k}^3 - \frac{1}{48} (\nu (6 - 6\kappa + \nu^3)) \tilde{k}^4 + \mathcal{O}(\tilde{k}^5). \]  \hfill (6.28)

Identically to the local truncation error analysis in Section 3.3, Chapter 3, we recover third-order accuracy with \( \kappa = 1/3 \) and SSP33 or SSP43. The ratio of the amplification factors for the third-order schemes approximates 1 to fourth-order \( \mathcal{O}(\tilde{k}^4) \):

\[
\mathcal{R}(\nu, \tilde{k}, 1/3)^{\text{SSP33}} = 1 - \frac{1}{24} (\nu (2 + \nu^3)) \tilde{k}^4 + \mathcal{O}(\tilde{k}^5) \\
\mathcal{R}(\nu, \tilde{k}, 1/3)^{\text{SSP43}} = 1 - \frac{1}{48} (\nu (4 + \nu^3)) \tilde{k}^4 + \mathcal{O}(\tilde{k}^5). \]  \hfill (6.29)

The leading term is a real number which is associated with a diffusive character of the method. Note that the four-stage scheme has a smaller error constant \( \Delta_\nu = \nu^4/48 \) than the three-stage scheme. Comparing this quantity with only a second-order time marching scheme and quadratic spatial reconstruction we obtain

\[ \mathcal{R}(\nu, \tilde{k}, 1/3)^{\text{Heun}} = 1 - \frac{1}{6} i \nu^3 \tilde{k}^3 - \frac{1}{24} \nu (2 - 3\beta + \nu^2) \tilde{k}^4 + \mathcal{O}(\tilde{k}^5). \]  \hfill (6.30)

Similarly to the previous stability analysis, we can determine the accuracy of reconstructions utilizing \( \phi(\theta) = \beta \theta \) with 1 < \( \beta \leq 2 \). The error analysis for the complex spectrum (6.23) yields

\[ \mathcal{R}(\nu, \tilde{k}, \beta)^{\text{Heun}} - 1 = \frac{1}{2} (\beta - 1) \nu \tilde{k}^2 - \frac{1}{6} i \nu (-1 + 3\beta + \nu^2) \tilde{k}^3 + \mathcal{O}(\tilde{k}^4). \]  \hfill (6.31)

For \( \beta = 1 \) we recover linear reconstruction, which gives a dispersive second-order accurate scheme. For 1 < \( \beta \leq 2 \) we have only first order accuracy, yet the scheme is totally unstable since

\[ z^{nu}(\tilde{k}) = \left(1 + \frac{1}{2} (\beta - 1) \nu \tilde{k}^2\right) z^{ex}(\tilde{k}) > z^{ex}(\tilde{k}). \]  \hfill (6.32)
This result approves the previous stability analysis. For the continuous modified equation a reconstruction, which employs $\phi(\theta) = \beta \theta$ with $\theta > 1$ results in
\[
 u_t + a u_x = -\frac{1}{2} a \Delta x (\nu + \beta - 1) u_{xx} + O(\Delta x^2).
\] (6.33)

This equivalent model equation is an ill-posed backward heat equation, with exponentially growing solutions. This result confirms, that compressiveness is essentially local instability.

In Fig. 6.5 we compare numerical results for the linear advection equation (6.20). Two different time integration schemes, namely the second-order Heun (6.1) and the three-stage SSP33 Runge-Kutta method (6.2), are used. The unlimited third-order algorithm produces only small, yet symmetric and local, over and undershoots. In contrast, the second-order scheme with third-order spatial accuracy is dispersive and the oscillations spread as time proceeds and completely destroy the shape of the initial profile.

We are aware that the von Neumann analysis gives only a sufficient condition for stability for linear operators $L$. Reconstructions utilizing limiters are essentially non linear. Thus their behavior in the frequency domain cannot be completely represented by a linear stability analysis. However, the analysis indicates the benefits using a three-stage Runge-Kutta time integration together with third-order spatial reconstruction. We not only obtain a far less restrictive CFL condition than using a second-order time marching scheme, we also improve its dissipative character especially for high frequency waves and its shape preserving properties. Moreover due to the local character of the oscillations, the non linear limiter function will only be limiting few cells around a discontinuity. Whereas in the second-order, dispersive, algorithm the limiter has to be applied on many cells to stabilize the computation.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.5.png}
\caption{Linear advection of a square wave: unlimited quadratic reconstruction with a second-order ($\times$) and third-order ($\bullet$) TVD time-marching scheme, until $t = 2$ (left) and $t = 20$ (right), $\nu = 0.8$ with 200 cells.}
\end{figure}
Chapter 7

Numerical Experiments for Scalar Equations

In this section we conduct a series of convergence and accuracy tests for one-dimensional equations. We put our focus on both, the approximation accuracy for smooth data, as well as the sharp and accurate resolution of discontinuous profiles.

7.1 Advection Equation: Convergence studies

We solve the scalar linear advection equation \( a = 1 \) (1.4)

\[
    u_t = -u_x, \quad u(t = 0, x) = u_0(x),
\]

with different initial conditions on a periodic domain \( x \in [-1, 1] \). First we will consider smooth initial profiles to identify the accuracy of the proposed algorithm. For this purpose we conduct two different convergence studies. All experiments are performed on a uniform FV grid.

Example 1: We solve the advection equation with the initial condition \( u_0(x) = \sin(\pi x) \). The initial profile is advected until \( t = 1 \). We conduct numerical experiments for different radii \( r \) of the asymptotic region (4.34), to clarify its expansion for the accuracy of the proposed scheme. Rather than presenting a table with the convergence rates calculated between two consecutive results, we show in Fig. 7.1 the computed \( L_1 \)- and \( L_{\infty} \)-errors obtained at \( t = 1 \) with Courant number \( \nu = 0.9 \). The discrete norms are given by:

\[
    ||e(\cdot, t^n)||_{L_1} = \Delta x \sum_{i=1}^{N} |\bar{u}(x_i, t^n) - \bar{u}_i^n|
\]

\[
    ||e(\cdot, t^n)||_{L_{\infty}} = \max_{1 \leq i \leq N} |\bar{u}(x_i, t^n) - \bar{u}_i^n|
\]

Here \( N \) is the number of computational cells. In this way we get a clear impression when, in terms of spatial resolution \( \Delta x \) and radius \( r \), the designated error, i.e., the error of the unlimited quadratic
Figure 7.1: Double logarithmic plots of the error vs. number of grid-cells for the advection equation with $u_0(x) = \sin(\pi x)$ at $t = 1$ with $\nu = 0.9$. Solution for different radii $r$ of the asymptotic region (4.34) are shown. Left: $l_1$-norm errors, right: max-norm errors.

Figure 7.2: Solution of the advection equation with a square wave (left) and a sin-wave (right) as initial conditions. Results are calculated at $t = 20$ with $\nu = 0.8$ using 200 cells for different expansion of the asymptotic region.

polynomial reconstruction is reached. We have to keep in mind that no FV-reconstruction based on three cells can be more accurate than the pure quadratic polynomial reconstruction. Calculating the convergence rate between two consecutive results gives in all three cases, $r = 1.0$, $r = 0.1$ and $r = 0.01$, the formal third-order using only 40 cells. However, it is obvious that the designated accuracy is only reached with more cells, especially for a small asymptotic region $r \leq 0.01$.

Note that the curvature of our initial function is $|u''(x)| = \pi^2$. Whereas Artebrant and Schroll [5] use the initial function $u_0(x) = 1.0 + 0.2 \sin(\pi x)$ with smaller curvature $|u''| = 0.2\pi^2$. Consequently, they recover third-order convergence rates with $q = 1.4$ (see Remark 4.2.2) already on 40 grid cells. Using the same initial data, we recover the desired convergence error for $r = 1.0$ with only 10 cells, thus significantly faster. This is due to the quadratic converges of $\eta$ for smooth data (see Lemma 4.5.1).
7.1. Advection Equation: Convergence studies

Fig. 7.2 shows two different solutions of the advection equation: one obtained with a step function and one with a sin-wave as initial function \( u_0(x) \). The picture clarifies the effect of the size of the asymptotic region on discontinuities. Choosing a small radius \( r \leq 0.01 \), the limiter needs a sufficient resolution, around 200 cells, to be able to distinguish between a discretized smooth extremum or a shallow gradient. Note that the two profiles are advected ten times \( (t = 20) \) around the domain \( x \in [-1, 1] \). We can also observe that even for a large asymptotic region, essentially no limiting, the method does not become unstable, unlike an unlimited second-order TVD-method. The over- and under-shoots remain localized and do not grow with time.

**Example 2:** To investigate the long-term evolution properties of several schemes we solve the advection equation (7.1) with the initial condition \( u_0(x) = \sin^4(\pi x) \). The convergence errors are calculated at \( t = 10 \) with Courant number \( \nu = 0.7 \). The key issue of this test case is to examine the importance of adjusting the time marching scheme to the formal order of accuracy.

We perform numerical convergence analyses for the second-order Heun scheme (6.1) and the third-order SSP Runge-Kutta time integrator (6.2) (RK3rd) combined with different spatial reconstructions. The tested schemes are:

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Spatial Reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heun, MCD</td>
<td>Heun ( \mathcal{O}(\Delta t^2) ) with limiter ( \phi(\theta) = \max(0, \min(2 \theta, \frac{1+\theta}{2^2}), 2) )</td>
</tr>
<tr>
<td>Heun, LimO3</td>
<td>Heun ( \mathcal{O}(\Delta t^2) ) with limiter ( \tilde{\phi}(4.23) )</td>
</tr>
<tr>
<td>RK3rd, Koren</td>
<td>RK3rd ( \mathcal{O}(\Delta t^3) ) with Koren-limiter ( \phi_{\text{TVD}}(4.21) )</td>
</tr>
<tr>
<td>RK3rd, LimO3, ( r = 0.1 )</td>
<td>RK3rd ( \mathcal{O}(\Delta t^3) ) with limiter ( \phi^{O(3)}(4.41) )</td>
</tr>
<tr>
<td>RK3rd, Poly. 3rd</td>
<td>RK3rd ( \mathcal{O}(\Delta t^3) ) with quadratic reconstruction ( \phi(\theta) = \frac{2+\theta}{3} )</td>
</tr>
<tr>
<td>Heun, Poly. 2nd</td>
<td>RK3rd ( \mathcal{O}(\Delta t^3) ) with linear reconstruction ( \phi(\theta) = \frac{1+\theta}{2^2} )</td>
</tr>
</tbody>
</table>

The Heun time marching scheme combined with a third-order spatial reconstruction is formally only second-order accurate, exhibiting a dispersive character (6.30) (see also Fig. 6.5). This leads to a phase error when smooth solutions are approximated, especially for long-term evolution.

Fig. 7.3 shows the obtained convergence results with the different algorithms. The two unlimited spatial reconstructions are plotted as reference solutions for a slope corresponding to a second-order (dashed line) and a third-order method (full line), respectively. We employed a quite fine resolved grid since the initial function consists of two sharply peaked but smooth extrema.

The worst results are obtained with the classical second-order TVD method (Heun, MCD). Yet the scheme exhibits almost second-order convergence rate in the \( L_1 \)-norm with only slightly larger errors than its unlimited counterpart. Changing only the spatial reconstruction to the limiter function \( \tilde{\phi}(4.23) \) improves the results. The errors in both norms decrease by an order of about 2, yet the convergence order remains identical.

An improvement of the accuracy is obtained with the third-order Runge-Kutta time integrator combined...
either with the Koren-limiter (4.21) or with \( \hat{\phi} \) (4.23). Although third-order accuracy cannot be recovered completely, extrema clipping without asymptotic region, the convergence rates in the \( L_1 \) - and \( L_\infty \)-norm are about 2.5 and 1.7, respectively. For smooth data both, the Koren-limiter (4.21) and the limiter function \( \hat{\phi} \) (4.23) are of similar accuracy. Differences are not visible in a double logarithmic plot. Both limiter functions yield for smooth extrema with one vanishing lateral derivative

\[
\phi^{\text{TVD}}(0) = \hat{\phi}(0) = 0 \quad \text{and} \quad \frac{d\phi^{\text{TVD}}}{d\theta}(0^+) = \frac{d\hat{\phi}}{d\theta}(0^+) = 2.
\]

Yet the limiter \( \hat{\phi}(\theta) \) (4.23) is slightly more accurate. It employs a quadratic reconstruction for \( \theta \in [-2, -\frac{4}{5}] \) and yields

\[
\frac{d\hat{\phi}}{d\theta}(0^-) = -\frac{1}{2}.
\]

Hence the limiter \( \hat{\phi} \) (4.23) uses also information for the reconstruction from the neighboring cells for \( -2 \leq \theta < 0 \).

Switching on the asymptotic region, hence using limiter \( \phi^{O(3)} \) (4.41) with radius \( r = 0.1 \) improves the convergence results significantly. The initial profile exhibits large curvatures \( |u''| \approx 40 \) at the vicinity of the two extrema. Therefore, the full accuracy of the unlimited quadratic reconstruction in the max-norm is recovered not until about 1000 grid-cells resolution. Utilizing the asymptotic region, however, already improves the accuracy with less computational cells in both norms. Note that no accuracy degeneracy, as reported in [61] for third-order ENO schemes was found with the proposed algorithm (RK3rd, LimO3, \( r = 0.1 \) with asymptotic region).

We can clearly observe that the unlimited second-order scheme (dashed line) needs about 6000 grid-cells

![Double logarithmic plots of the error vs. number of grid-cells for the advection equation with \( u_0(x) = \sin^4(\pi x) \) at \( t = 10 \) with \( \nu = 0.7 \). Comparison of different time-marching schemes combined with different spatial reconstruction. Heun, Poly. 2nd (dashed line): unlimited second-order scheme. RK3rd, Poly. 3rd (dashed line): unlimited third-order scheme. Left: 1-norm errors, right: max-norm errors.](image-url)
to achieve the same accuracy as the RK3rd, Koren-limiter method (or RK3rd LimO3, respectively), which cuts of smooth extrema. Consequently for this example modified second-order algorithms as proposed by [75] (see reconstruction (5.5) in Section 5.1) will also need at least 6000 computational cells and also introduce oscillations. We did not observe any spurious oscillations, neither with RK3rd $\phi^{\text{TVD}}$ (4.21) or with RK3rd $\hat{\phi}$ (4.23).

Although the computational effort increases with the number of stages in the time integrator, it is difficult to achieve the accuracy of the third-order schemes (RK3rd, LimO3, $r = 0.1$ with asymptotic region) with second-order methods (see Fig. 7.3). The analysis of Section 6.1 suggests that we can use larger time steps with the new algorithm. This increases the computational efficiency, even in comparison to second-order methods, without loss of accuracy.

### 7.2 Advection Equation: Initial Condition with Discontinuities

We take a test case, proposed in [23]. It consists of a tight combination of four waves, namely a smooth but narrow Gaussian peak, a square wave, a triangle wave and a half ellipse. The triangle has a smooth transition at its base. We employ periodic boundary conditions on the domain $x \in [-1, 1]$ and advect the initial profile until $t = 20$ utilizing 400 computational cells. This experiment impressively envisages the major accuracy improvement of third-order accurate schemes, in contrast to second-order accurate methods.

In Fig. 7.4 we compare results obtained with classical TVD limiters. We can observe that none of the second-order accurate limiters can resolve the profile appropriately. All four reconstructions produce non-symmetric and smeared out profiles. Especially results obtained with the minmod limiter (3.41) are much to diffusive. It is rather impossible to distinguish the four different waves. The superbee limiter (3.42), although quite accurate on the square wave, tends to "squares out" every smooth structure. This is a well known drawback of the superbee limiter and is associated with its over-compressive behavior (see discussion in Section 4.4 and Section 6.2). Van Leer’s smooth limiter (3.44) gives similar results as the MCD limiter (3.43). The piecewise linear MCD limiter, however, employs $\phi(\theta) = 2 \theta$ for $\theta \in [0, 1/3]$, which eventually results in a more compressive reconstruction of linear discontinuities.

It is obvious that the quality of the reconstruction depends on both, the data and the applied limiter. Superbee sharpens discontinuities, whereas van Leer’s limiter seem to capture smooth structures more accurately. This observation led to the idea of adaptive limiters. Billet and Louedin [7] suggested to apply different limiters, depending on the "local physical fluctuation". In other words they preprocess the left and right lateral difference, such that for empirically fixed bounds, different limiters are used. The final method is essentially van Leer’s $\kappa$-scheme with some additional empirical constants. The resulting scheme, however, still degenerates accuracy at smooth extrema and does not reach the shock-capturing quality of third-order accurate schemes.
Figure 7.4: Advection of discontinuities with $\nu = 0.8$ until $t = 20$ utilizing 400 computational cells. Second-order schemes with different limiters. Top profile: Minmod (3.41) and van Leer’s smooth limiter (3.44). Bottom profile: Superbee (3.42) and MCD limiter (3.43).

In the top plot of Fig. 7.5 we compare smooth essentially third-order accurate reconstructions. We can observe the differences between the log-limiter (4.13), the modified $\Pi^{GPR-1/3}$ (4.33) reconstruction and the smooth ModGPR-1/3 (4.30). Note that only $\Pi^{GPR-1/3}$ does resolve smooth extrema to full third-order accuracy. All schemes utilize the explicit third-order SSP Runge-Kutta time marching scheme (6.2).

Obviously, the best results are obtained with the log-limiter, which is capable of resolving linear discontinuities quite accurately. Both smooth GPR-1/3 reconstructions eventually are not able of distinguishing between the four different waves. We can observe that $\Pi^{GPR-1/3}$ gives better results compared to the original GPR-1/3 limiter. Note that simultaneously it produces the largest variation. Reducing the parameter $K$, we diminish local variation, essentially recovering the original GPR-1/3 limiter (see discussion Section 5.2). Although the log-limiter clips-off smooth extrema, it still approximates the smooth half
ellipse very accurately, especially compared to $\Pi^{GPR-1/3}$ and ModGPR-1/3. This is due to the fact, that for monotone data, eventually the slopes of the half ellipse, the log-limiter recovers the unlimited quadratic reconstruction to fourth-order accuracy (see Remark 4.2.1). Wheres $\Pi^{GPR-1/3}$ and ModGPR-1/3 converges only quadratically to the unlimited quadratic reconstruction (see, e.g., eq. (4.31)).

In the bottom profile of Fig. 7.5 we can observe the relevancy of the shape of the limiter for $\theta < 0$. For this computation, we used the limiter

$$\hat{\phi}(\theta)^{\text{test}} = \max \left(-1, \min \left(\frac{2 + \theta}{3}, \max \left(-2 \theta, \min \left(2 \theta, \frac{2 + \theta}{3}, 2\right)\right)\right)\right)$$

(7.4)

and compared it to $\phi^{\text{TVD}}$ (4.21). We deliberately utilized here the second-order accurate Heun time marching scheme, to emphasize that the formal order of the scheme is critical for reconstruction quality.
As already indicated in Section 6.1, the formal order of the complete algorithm is crucial, not only for its accuracy, but also for its shape preserving, i.e., shock capturing quality. Second-order accurate methods are dispersive, which results in the distinct tendency to square smooth structures, independently if smooth or piecewise linear limiters are employed (see also Fig. 7.4).

It is obvious that $\hat{\phi}(\theta)_{\text{test}}$ violates the monotonicity region of $[12, 64]$, yet we still could not observe any over or undershoots. In comparison to $\phi^{\text{TVD}}$ the amplitude of all four waves is better resolved. The limiter (7.4) reconstructs cells for which hold $\theta \in [-\frac{5}{2}, -\frac{7}{2}]$ to full third-order accuracy. This results in a more accurate approximation of the wave peaks. Note that we have significantly stretched the region of third-order reconstruction for $\theta < 0$ in order to clearly demonstrate its effects. For further calculations we remain in the monotonicity region proposed by [12, 64].

Finally we conducted this experiment also with our new piecewise linear limiter function $\phi^{O(3)}$ (LimO3). In Fig. 7.6 we can see the differences between the proposed algorithm, the third-order ENO method (ENO3), Artebrant and Schroll’s LDLR and Marquina’s LHHR. Note that LHHR, is a further version of the local hyperbolic reconstruction LHR, but LHHR utilizes harmonic mean functions for the preprocessing of the lateral slopes. This preprocessing reduces the local variation to $O(\Delta x)$, yet at the same time introduces more numerical diffusion.

ENO3 is essentially not local since it uses a five-point stencil to choose between the ”smoothest” quadratic reconstruction for the interface values. All schemes use the same explicit third-order SSP Runge-Kutta time marching scheme. Since our method has a less restrictive CFL condition, we run the experiment with $\nu = 0.8$, instead of $\nu = 0.6$ for ENO3.

Especially in contrast to the classical TVD-MUSCL scheme (see Fig. 7.4), the new method gives a good approximation of the exact profile. Typical drawbacks of second-order TVD limiter, such as smearing and squaring of linear waves are completely avoided. The results are perfectly symmetric and corners a very well resolved. The symmetry of the results is due to the fact, that near $\theta = 1$ we recover a smooth function of second degree and the formal third-order accuracy, which results in a rather diffusive than dispersive scheme. Second-order TVD limiters, such as superbee or minmod, are only Lipschitz continuous at $\theta = 1$. This drawback enhances the probability, that the wrong choice of slopes is used for the one-sided approximation. Note, however, that even smooth limiters, such as van Leer’s limiter (3.44) do not produce symmetric profiles (see Fig. 7.4).

All three schemes, i.e., LogLim, LimO3 and LDLR are essentially of the same quality, yet LimO3 has a higher Gaussian peak and produces less overshoots for the half ellipse. Whereas ENO3 is not able to distinguish between the different waves and always produces sinusoidal-like wave patterns and LHHR is quite diffusive. All five methods produce negative values, yet for LimO3 they are of a smaller magnitude. Note, that the asymptotic region scales, in the presence of jump discontinuities, with $O(\Delta x^2)$ (4.39). LogLim is not capable of resolving smooth extrema with oner vanishing lateral derivative. It introduces
more spurious oscillations than LimO3, since it is out of Harten’s TVD region for \( \theta > 0 \).

### 7.3 Nonlinear Scalar Equations

It is our aim to examine the total variation for nonlinear scalar equations. First we solve Burger’s equation

\[
  u_t = - \left( \frac{u^2}{2} \right)_x, \quad u_0(x) = 1 + \frac{1}{2} \sin(\pi(x - 1))
\]

(7.5)

on a periodic domain \( x \in [-1, 1] \) under CFL condition \( \nu = 0.9 \).
Second we calculate the non-convex Buckley–Leverett equation

\[ u_t = -\left(\frac{u^2}{u^2 + a(1-u)^2}\right)_x \, , \quad u_0(x) = \begin{cases} 1 & \text{if } -\frac{1}{2} \leq x \leq 0 \\ 0 & \text{elsewhere} \end{cases} \]  

(7.6)
on the same spatial domain and with Courant number \( \nu = 0.7 \). The constant \( a \) is set to \( \frac{1}{4} \) and the initial profiles are taken from [5]. For the Buckley–Leverett equation we use the local Lax-Friedrichs flux

\[ F_{i+\frac{1}{2}}(\hat{u}^i(\cdot), \hat{u}^{i+1}(\cdot)) = \frac{1}{2} \left( f(\hat{u}^i(\cdot)) + f(\hat{u}^{i+1}(\cdot)) - a_{i+1/2}(\hat{u}^{i+1}(\cdot) - \hat{u}^i(\cdot)) \right) , \]  

(7.7)
with the wave propagation speed \( a_{i+1/2} = \max \left( |\frac{\partial f(u)}{\partial u}| \right) \). See Section 8.5 for a detailed discussion on numerical flux functions.

We calculate for both equations the total variation for a grid function \( \bar{u} \) (see Section 3.2, eq. (3.7))

\[ TV(\bar{u}^n) \equiv \sum_{i=1}^{N} |\bar{u}_{i+1}^n - \bar{u}_i^n| \]  

(7.8)
of the discrete solution.

Fig. 7.7 (left) shows the numerical behavior of the total variation for Burger’s equation and the LimO3 approximation at \( t = 2.0 \). We observe that the calculated total-variation is bounded by \( TV(\bar{u}^0) \) and gradually decreases with time as a jump discontinuity forms at \( t \approx 1.0 \). In contrast to LDLR, the accuracy of the solution with 40 cells is significantly better (see [5]) and the evolving smooth profile for \( t \in [0, 0.8] \) is recovered accurately. Note that for all three resolutions the value of total variation before the shock forms is always larger than of LDLR. Hence the smooth profile is resolved more accurately with few computational cells. This could even be improved with a larger asymptotic region, thus also introducing larger, but bounded spurious oscillations shortly before the jump discontinuity forms. Once
the discontinuity clearly appears the limiter yields $\phi^{O(3)} \rightarrow \hat{\phi}(0) = 0$ and the oscillations completely disappear.

In the case of the Buckley–Leverett equation we analyze a very similar total-variation behavior (see Fig. 7.8). Although we introduce some spurious variation, that eventually diminishes as the mesh is refined. The slight overshots at $x = 0.62$ in the compound wave are similar to the LDLR results (see [5] for details) and are not connected to the asymptotic region. Due to the compressible character of high-order schemes, the shock speed is approximated slower than the characteristic speed at some cells between the rarefaction and the shock. Adding diffusion, i.e., simply decreasing the parameter $\beta$ in $\Phi$ (4.19) damps out the spurious overshoot in the vicinity of the shock.

![Figure 7.8: Total variation of the Buckley–Leverett equation (left) and LimO3 approximation with $\nu = 0.7$ and $r = 0.1$ at $t = 0.4$](image-url)
Chapter 8

Euler Equations of Gas Dynamics

The Euler equations model an inviscid, compressible and non-heat conducting gas. They are derived from the physical principles of conservation of mass (or density), momentum and energy. In their conservation form, the equations are valid whether the flow is smooth or discontinuous. As previously discussed, the solution of hyperbolic conservation laws reduces to a flux estimate at each cell interface. In particular for a given numerical flux function the numerical accuracy improves with the reconstruction order of the cell interface values. Numerical algorithms, more precisely upwind schemes, for hyperbolic conservation laws are generally derived from the scalar advection equation. For systems to assure correct shock speeds the equations are solved always in conservation form. Yet the quantities, which are essentially convected are the characteristic variables. These are the quantitative values associated with every single wave of a linear or linearized system. They determine the wave’s strength.

The third interesting variables are the primitive variables, namely mass, flow velocity and pressure. They could be considered as engineering variables, since they are eventually the quantities measured in an experiment. They are also used for the calculation of the numerical flux function and the approximative Riemann solver.

In the application of high-order schemes to systems of hyperbolic conservation laws all of these variables play essential roles. It is still an open question which of these variables have to be reconstructed, i.e., also limited, in order to achieve best results, in terms of shock capturing. Yet it seems to be a common agreement that methods of higher than second order accuracy should reconstruct the characteristic variables (see e.g. [45, 49, 50]). This assumption is essentially based on the fact, that most of high-order schemes employ ENO / WENO reconstruction techniques for the cell interface values. A simple component wise application of the ENO / WENO reconstruction appears to be inadequate whenever discontinuities are separated by regions of almost constant states [52]. This eventually leads to spurious oscillations around discontinuities on coarse and intermediately refined grids.

In the following will review the various forms of the Euler equations, since we solve the Euler system in most of the numerical test cases for systems. For this purpose we employ Roe’s approximative Riemann
The basic equations for an 2D flow of an inviscid, compressible and non-heat conducting gas read
\[ \partial_t u(x, y, t) + \partial_x f(u(x, y, t)) + \partial_y g(u(x, y, t)) = 0, \] (8.1)
with
\[
\begin{align*}
    u &= \begin{pmatrix}
        \rho \\
        \rho v_x \\
        \rho v_y \\
        E
    \end{pmatrix},
    f(u) &= \begin{pmatrix}
        \rho v_x \\
        \rho v_x^2 + p \\
        \rho v_x v_y \\
        u(E + p)
    \end{pmatrix}, \quad \text{and} \quad
    g(u) &= \begin{pmatrix}
        \rho v_y \\
        \rho v_x v_y \\
        \rho v_y^2 + p \\
        v_y(E + p)
    \end{pmatrix}.
\end{align*}
\] (8.2)

The variables \( \rho, v_x, v_y, p \) and \( E \) denote density, velocity in \( x-, y- \) direction, pressure and total energy, respectively. Assuming a polytropic (perfect) gas, the system is completed by the equation of state
\[ p = (\gamma - 1)\rho \varepsilon, \] (8.3)
where \( \varepsilon \) is the specific internal energy per unit volume and \( \gamma \) is the adiabatic exponent, i.e., the ratio of specific heats. The total energy per unit volume is the defined as
\[ E = \rho \varepsilon + \frac{1}{2} \rho (v_x^2 + v_y^2). \] (8.4)

The 1D Euler equations are obtained by setting \( v_y = 0 \) and \( v_x = v \).
\[ \partial_t u(x, t) + \partial_x f(u(x, t)) = 0, \] (8.5)
with
\[
\begin{align*}
    u &= \begin{pmatrix}
        \rho \\
        m \\
        E
    \end{pmatrix}, \quad \text{and} \quad
    f(u) &= \begin{pmatrix}
        m \\
        \frac{m^2}{\rho} + p \\
        \frac{m}{p}(E + p)
    \end{pmatrix},
\end{align*}
\] (8.6)
where \( m = \rho v \) is the momentum. The total energy, then simply reads
\[ E = \frac{1}{2} \frac{m^2}{\rho} + \frac{1}{\gamma - 1} p. \] (8.7)

8.2 Primitive Variables

We define \( q \) to be the column vector of the primitive variables,
\[ q = (\rho, v, p)^T. \] (8.8)
At smooth regions of $u$, it is possible to rewrite the Euler equations in various comprehensive forms. We can derive from eq. (8.5, 8.6) a nonconservative, i.e., non-divergent form of the Euler equations:

$$
\partial_t q(x,t) + B \partial_x q(x,t) = 0,
$$

(8.9)

where the coefficient matrix $B \in \mathbb{R}^{3 \times 3}$ is defined as

$$
B = \begin{pmatrix}
v & \rho & 0 \\
0 & v & \frac{1}{\rho} \\
0 & \gamma p & v
\end{pmatrix}.
$$

(8.10)

The analytical eigenvalues of the coefficient matrix $B$ are

$$
\lambda_1 = v - c_s, \quad \lambda_2 = v \quad \text{and} \quad \lambda_3 = v + c_s,
$$

(8.11)

where the parameter $c_s$

$$
c_s = \sqrt{\frac{\gamma p}{\rho}}
$$

(8.12)
denotes the speed of sound.

We are in general less interested in smooth solutions of $u$ where the Euler equations can be rewritten in primitive variables. Our main interest is the calculation of shocks utilizing approximative Riemann solvers.

### 8.3 Roe’s Approximative Riemann Solvers

In the following we will discuss a major ingredient in the numerical approximation for systems of hyperbolic conservation laws. For this purpose we return to the conservative form of the Euler equations, which is also valid across shock waves. In 1D the Jacobian matrix $A \in \mathbb{R}^{3 \times 3}$ for the Euler equations is defined as

$$
A = \partial_u f(u) = (\partial_x f(u), \partial_m f(u), \partial_E f(u)).
$$

(8.13)

A straightforward calculation yields the analytical expression of the Jacobian matrix of the flux $f(u)$ (8.6),

$$
A = \begin{pmatrix}
0 & \frac{1}{2}v^2(\gamma - 3) & \frac{1}{2}v^2(\gamma - 1) - v H \\
\frac{1}{2}v^2(\gamma - 3) & v(3 - \gamma) & H - (\gamma - 1)v^2 \\
\frac{1}{2}v^2(\gamma - 1) - v H & H - (\gamma - 1)v^2 & \gamma v
\end{pmatrix},
$$

(8.14)

where $H$ is the total enthalpy

$$
H = \frac{E + p}{\rho} = \frac{v^2}{2} + \frac{\gamma p}{(\gamma - 1)\rho} = \frac{v^2}{2} + \frac{c_s^2}{\gamma - 1}.
$$

(8.15)

The parameter $c_s$ is the speed of sound $c_s = \sqrt{\frac{\gamma p}{\rho}}$ (8.12). The eigenvalues of the matrix $A$ are identical to the eigenvalues of the coefficient matrix $B$ (8.10)

$$
\lambda_1 = v - c_s, \quad \lambda_2 = v \quad \text{and} \quad \lambda_3 = v + c_s.
$$

(8.16)
They have to agree, since both forms, the primitive and the conservative, are equivalent and yield the identical characteristic speeds of the Euler system. The complete set of right-eigenvectors is

\[ R = [r_1, r_2] = \begin{pmatrix} 1 & 1 & 1 \\ u - c_s & u & u + c_s \\ H - uc_s & \frac{1}{2}u^2 & H + uc_s \end{pmatrix} \]  

(8.17)

To define an approximative Riemann solutions for the nonlinear hyperbolic problem (8.5) one has to find a suitable linearization of the problem. The linearization has to be defined locally at each cell interface, i.e., where the Riemann problem is solved,

\[ \partial_t \tilde{u} + A(\tilde{u}_{i+1/2}) \partial_x \tilde{u} = 0. \]  

(8.18)

The \( \tilde{u} \) notation should emphasize, that we have frozen the system locally at the cell interface \( x_i \). The matrix \( A(\tilde{u}_{i+1/2}) \) has to be chosen as an appropriate approximation of the Jacobian \( A = \partial_u f(u) \). The linear Riemann problem is consistent with the original problem, if for all variable \( \bar{u}_i \) hold

\[ A_{i+1/2}(\bar{u}_i, \bar{u}_i) = A(\bar{u}_i, \bar{u}_i). \]  

(8.19)

Conservation of the numerical scheme is maintained, if holds

\[ A_{i+1/2}(\bar{u}_{i+1} - \bar{u}_i) = f(\bar{u}_{i+1}) - f(\bar{u}_i). \]  

(8.20)

for all \( i \). Note that this is essentially the mean value theorem.

Since the hyperbolic character of the initial system is conserved by the linear problem, the matrix \( A_{i+1/2} \) is diagonalizable permitting the decomposition

\[ A(\tilde{u}_{i+1/2}) = R(\tilde{u}_{i+1/2}) \Lambda(\tilde{u}_{i+1/2}) R^{-1}(\tilde{u}_{i+1/2}), \]  

(8.21)

The matrix \( R(\tilde{u}_{i+1/2}) \in \mathbb{R}^{3 \times 3} \) is the set of right-eigenvectors, associated with a chosen average state \( \tilde{u}_{i+1/2} \). The matrix \( \Lambda(\tilde{u}_{i+1/2}) \) is a diagonal matrix with the eigenvalues \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) on the diagonal. The eigenvalues are obtained from the approximated value \( \tilde{u}_{i+1/2} \). The matrix \( R^{-1}(\tilde{u}_{i+1/2}) \) is the complete set of left-eigenvectors.

For the FV scheme we have to compute the numerical flux at the cell boundary \( x_{i+1/2} \). Note that numerical fluxes are functions of two values, i.e., \( F_{i+1/2} = F(\hat{u}^{(-)}_{i+1/2}, \hat{u}^{(+)}_{i+1/2}) \). Consequently the quality of the approximation of the Jacobian \( A(\tilde{u}_{i+1/2}) \) depends, not only upon the sort of average one takes to approximate the value \( \tilde{u}_{i+1/2} \), but also on the quality of the cell reconstructed values \( \hat{u}_{i+1/2}^\pm \)

One simple approximation for the value \( \tilde{u}_{i+1/2} \) is the arithmetic mean

\[ \tilde{u}_{i+1/2} = \frac{1}{2} \left( \hat{u}^{(-)}_{i+1/2} + \hat{u}^{(+)}_{i+1/2} \right). \]  

(8.22)
8.4 Characteristic Variables

A more elaborate average for the Euler equations was proposed by Roe [51] in terms of a parameter vector

\[ \tilde{v}_{i+1/2} = \frac{\sqrt{\rho(-) v(-)} + \sqrt{\rho(+) v(+)} }{\sqrt{\rho(-)} + \sqrt{\rho(+)}} \]

(8.23)

\[ \tilde{H}_{i+1/2} = \frac{\sqrt{\rho(-) H(-)} + \sqrt{\rho(+) H(+)} }{\sqrt{\rho(-)} + \sqrt{\rho(+)}} \]

For simplicity, we have dropped on the right hand side of eqs. (8.23) the underscore \( i + 1/2 \). Yet it should be made clear that both, the Roe averaged velocity \( \tilde{v} \) as well as the Roe averaged total enthalpy \( \tilde{H} \) are obtained from the reconstructed left and right states \( \hat{u}(-) \) and \( \hat{u}(+) \) at the cell boundary \( x_{i+1/2} \).

Substituting eq. (8.23) into the Jacobian (8.14) we obtain the Roe-matrix \( A_{Roe}^{i+1/2} \). Up to know we have just obtained an approximation of the Jacobian \( A^{i+1/2} \) for the locally linearized Euler equations. In the FV scheme we eventually have to approximate the physical flux function at the cell boundaries. Since the system is locally linear, we can employ Godunov’s scheme (see Section 1.3, eq. (1.20, 1.21)). The numerical flux function then reads

\[ F_{i+1/2}^{Roe}(\hat{u}(-), \hat{u}(+)) = \frac{1}{2} \left( f(\hat{u}(-)) + f(\hat{u}(+)) - |A_{Roe}^{i+1/2}| (\hat{u}(+) - \hat{u}(-)) \right) \]

(8.24)

where the eigenvalues \( \lambda_j \), eigenvectors \( r_j \) and the variables \( \alpha_j \) are obtained from the decomposition of the Roe-averaged matrix \( |A_{Roe}^{i+1/2}| \).

\[ |A_{Roe}^{i+1/2}| = R(\tilde{u}_{i+1/2})|\Lambda(\tilde{u}_{i+1/2})|R^{-1}(\tilde{u}_{i+1/2}). \]

(8.25)

This flux is simply another form of the Godunov flux for linear systems. In this form it is also called Roe’s method or wave-propagation form of Godunov’s method [38] since it contains explicit information of the characteristic variable decomposition.

8.4 Characteristic Variables

To be able to calculate the flux function at the cell boundaries, we have to locally freeze the nonlinear conservation law in every time-step. Thus, we obtain in general, a linear system

\[ \partial_t u(x, t) + A \partial_x u(x, t) = 0, \]

(8.26)

with a constant coefficient matrix \( A = f'(u) \). Therefore, the matrices \( R, R^{-1}, \Lambda \) are the related eigenvectors and eigenvalues are all constant. With the decomposition of the approximated Jacobian matrix, we can define a variable transformation for the vector \( w(x, t) : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^p \)

\[ w(x, t) = R^{-1}u(x, t). \]

(8.27)
With this transformation, the linear system becomes diagonal, and we obtain a system of decoupled linear advection equations.

\[ \partial_t w(x, t) + \Lambda \partial_x w(x, t) = 0, \tag{8.28} \]

The new variables \( w(x, t) = (w_1(x, t), \ldots, w_p(x, t))^T \) are the characteristic fields. They are essentially the single waves, which get advected along their characteristics with the velocities \( \lambda_1, \ldots, \lambda_p \) obtained from the Jacobian \( A \).

To apply the high-order reconstruction on the characteristic fields, we have first to transform the conservative variable of the considered cells \([x_{i-1}, x_i, x_{i+1}]\) via

\[ \bar{w}_k = R^{-1}(\tilde{u}_{i+1/2})\bar{u}_k, \quad \text{for } k = i - 1, i, i + 1. \tag{8.29} \]

It is obvious that with larger stencils more cell mean values have to be transformed. Afterwards we apply the proposed reconstruction routine (3.36) on the cell mean values \( \bar{w}_k \), to obtain the cell interface values \( \hat{u}_{i+1/2}^\pm \). This is essentially the characteristic based reconstruction. Finally we obtain the conserved cell interface values \( \hat{u}_{i+1/2}^\pm \) from the inverse transformation

\[ \hat{u}_{i+1/2}^\pm = R(\tilde{u}_{i+1/2})\bar{w}_{i+1/2}^\pm. \tag{8.30} \]

The last step cannot be avoided, since the variables \( \hat{u}_{i+1/2}^\pm \) represent the interface values in the physical conserved space. Numerical flux functions depend explicitly on these values, so that conservation of the scheme is ensured.

Until now we have just calculated the interface values and not yet applied an approximative Riemann solver. We should remark that the characteristic decomposition for the reconstruction is independent of the Riemann solver, one intends to apply. Once the cell interface values \( \hat{u}_{i+1/2}^\pm \) are obtained, the procedure from Section 8.3 can be applied.

It is however a common “believe” that methods of higher than second-order accuracy don’t need to apply complicated Riemann solvers [52]. Believe is indeed the correct transcription, since there exist neither formal proofs nor physical observations which endorse this idea. It appears, however, that in most numerical examples the quality of the solution gets less sensitive to the chosen Riemann solver, the higher the formal accuracy of the scheme is. Obviously characteristic variable decomposition is quite expensive and combined with Roe’s approximative Riemann solver very inefficient. Hence we will discuss, in the following, some approximative Riemann solvers, which are simpler and more efficient than Roe’s Riemann solver (8.24).

### 8.5 Different Flux Formulations

The major disadvantage of linearized Riemann solvers is that, for certain Riemann problems non-physical states with negative density or internal energy could be produced. Einfeldt et. al. [14] have shown for
the Euler equations, that no Godunov-type scheme based on a linearization will preserve positivity of the density during the computation. They have suggested to use schemes, which are positively-conservative, such as the Harten-Lax-van Leer-Einfeldt (HLLE) scheme.

The HLLE approximative Riemann solver was originally reported in [13]. This solver is based on estimates of the maximal and minimal wave speed. The numerical flux function reads

\[ F_{HLLE}^{i+1/2}(\hat{u}^-, \hat{u}^+) = \frac{1}{b^+ - b^-} \left( b^+ f(\hat{u}^-) - b^- f(\hat{u}^+) + b^+ b^- \left( \hat{u}^+ - \hat{u}^- \right) \right). \]  

The HLLE scheme is positively conservative if the absolute value of the largest \( b^+ \) and smallest \( b^- \) wave speeds satisfy certain stability bounds:

\[ b^+ \equiv b^+_{i+1/2} = \max(0, \bar{b}^r_{i+1/2}) \quad \text{and} \quad b^- \equiv b^-_{i+1/2} = \max(0, \bar{b}^l_{i+1/2}), \]  

where \( \bar{b}^l_{i+1/2} \) is the minimal and \( \bar{b}^r_{i+1/2} \) is the maximal eigenvalue of the matrix \( A_{i+1/2} \), respectively. Einfeldt [13] suggested to utilize the eigenvalues obtained from Roe’s linearization (see Roe averages (8.23))

\[ \bar{b}^l_{i+1/2} = \bar{u}_{i+1/2} - \bar{c}_{i+1/2} \quad \text{and} \quad \bar{b}^R_{i+1/2} = \bar{u}_{i+1/2} + \bar{c}_{i+1/2}, \]  

with \( \bar{c}^2_{i+1/2} = (\gamma - 1) \left( \tilde{H}_{i+1/2} - \frac{1}{2} \tilde{u}_{i+1/2}^2 \right) \). This is exactly the way we employ the HLLE flux.

An approximative Riemann solver, which only considers the maximal wave speed is the Local Lax-Friedrichs flux (LLF) (see also Section 7.3, eq. (7.7)). The numerical flux function reads

\[ F_{LLF}^{i+1/2}(\hat{u}^-, \hat{u}^+) = \frac{1}{2} \left( f(\hat{u}^-) + f(\hat{u}^+) - \alpha_{i+1/2} \left( \hat{u}^+ - \hat{u}^- \right) \right), \]  

with the wave propagation speed \( \alpha_{i+1/2} = \max \left( \left| \frac{\partial f(u)}{\partial u} \right| \right) \). The maximum is taken over the relevant range of \( \tilde{u}_{i+1/2} = (\hat{u}^-, \hat{u}^+) \). Usually an simple arithmetic mean is applied to approximate \( \tilde{u}_{i+1/2} \). This is the most dissipative numerical flux. It significantly smears out linear waves, i.e., contact discontinuities for systems.

In the HLLE and LLF scheme, the structure of the Riemann solution is only modeled with reduced amount of wave information. In particular only with information on the speed of the fastest or fastest and slowest wave, respectively. This approach makes both scheme for the Euler equations quite efficient, yet eventually leads to a loss of resolution for waves traveling at intermediate speeds, i.e., contact discontinuities.
Chapter 9

Numerical Experiments for One-Dimensional Systems

In this section we conduct a series of convergence and accuracy tests for the one-dimensional Euler equations and the one-dimensional ideal magnetohydrodynamic (MHD) equations. The order of the scheme is based upon the reconstruction of the interface values and not upon the approximative Riemann solver. Thus the proposed scheme is not tied to any particular flux function. Since our reconstruction is very efficient, we can choose a more costly, in terms of floating point operations, flux function. We use Roe’s approximative solver (see Section 8.3, eq. (8.24)). Most of the calculations are performed with the limiter function applied on conservative variables. In Section 9.2, however, we test the proposed reconstruction algorithm with characteristic based reconstruction.

9.1 Gas Dynamics: 1D Euler Equations

We consider an inviscid, compressible and non-heat conducting gas modeled by the Euler equations in conservative form (8.5, 8.6). The adiabatic exponent is set to be $\frac{5}{3}$ for diatomic gases. We compute initial value problems by means of an initial profile of the primitive variables density $\rho$, velocity $u$ and pressure $p$ defined through the domain.

9.1.1 Nonlinear Traveling Plane Waves

In order to calculate the empirical order of convergence (EOC) for nonlinear systems, we choose a nonlinearly coupled test problem, which develops a shock profile in finite time. We consider a convergence test on periodic boundaries $x \in [0, 1]$ with initial conditions

$$
\begin{pmatrix}
(\rho u p)\nu_0 = \\
1/4 \sin(\pi x) \\
(1 + \frac{1}{2}(\gamma - 1)\frac{\nu_0}{c_0})^\kappa \\
\end{pmatrix}, 
\quad c_0 = \sqrt{\gamma}, \quad \kappa = \frac{2}{\gamma - 1}
$$

(9.1)
Figure 9.1: EOC density profile of a non-linear plane wave for the Euler equations at $t = 0.8$ with $\nu = 0.8$ and $r = 0.01$. Koren-limiter $\phi^{\text{TVD}}(\theta)$ (4.21) with $\nu = 0.8$.

The initially smooth flow will eventually steepen and form a shock wave at $t^* = \frac{8}{\pi(\gamma+1)} \approx 1.06$ [33]. This way the EOC reflects the accuracy of the whole scheme, including a linearized approximated Riemann solver.

In Fig. 9.1 we plot the calculated errors in the density component at $t = 0.8$. We compare the proposed algorithm RK3rd, LimO3 $\phi^{\text{O}(3)}$ (4.41) with RK3rd, Koren-limiter $\phi^{\text{TVD}}$ (4.21). A reference solution is obtained with the third-order unlimited polynomial reconstruction (2.25, 2.26) utilizing 40960 cells.

Due to the rather flat initial density profile we get almost optimal accuracy with already 40 computational cells already with $r = 0.01$. Similar to the previous convergence experiment (see Fig. 7.3) both schemes exhibit nearly identical accuracy on coarse grids. Once the resolution is sufficient refined the asymptotic region is switched on and $\phi^{\text{O}(3)}$ (4.41) recovers full third-order accuracy. Yet the Koren-limiter always smears out smooth structures.

Now we conduct the traveling plane wave problem with two different third-order accurate time-integrators, namely the three-stage SSP33 (6.2) and four-stage SSP43 (6.3) Runge-Kutta scheme. Although both scheme require only two memory registers, SSP34 has to evaluate the numerical flux function four times. This makes the scheme less efficient compared to the SSP33 integrator, which evaluates only three explicit Euler stages.

In Tab. 9.1 we compare the CPU-time in seconds for both time marching schemes utilizing different Courant numbers. Both Runge-Kutta methods have similar performances when used with the maximal admissible (in von Neumann sense) Courant number. Employing identical Courant numbers, the three-stage SPP33 Runge-Kutta is faster.

In Fig. 9.2 we envisage converge results obtained for different Courant numbers. We can clearly observe
9.1. Gas Dynamics: 1D Euler Equations

Figure 9.2: EOC density profile of a non-linear plane wave for the Euler equations at $t = 0.8$ with different Courant numbers and $r = 0.01$.

Table 9.1: CPU-times in seconds for different Runge-Kutta time integration schemes

<table>
<thead>
<tr>
<th>N</th>
<th>SSP33, $\nu = 1.6$</th>
<th>SSP43, $\nu = 2.0$</th>
<th>SSP33, $\nu = 0.5$</th>
<th>SSP34 $\nu = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>160</td>
<td>0.08</td>
<td>0.09</td>
<td>0.23</td>
<td>0.30</td>
</tr>
<tr>
<td>320</td>
<td>0.23</td>
<td>0.23</td>
<td>0.64</td>
<td>0.86</td>
</tr>
<tr>
<td>640</td>
<td>0.70</td>
<td>0.76</td>
<td>2.17</td>
<td>2.83</td>
</tr>
<tr>
<td>1280</td>
<td>2.96</td>
<td>3.00</td>
<td>9.06</td>
<td>11.71</td>
</tr>
<tr>
<td>2560</td>
<td>12.55</td>
<td>13.09</td>
<td>33.27</td>
<td>50.99</td>
</tr>
</tbody>
</table>

the accuracy improvement for a scheme utilizing a small time step. This result confirms the theoretical predictions (see Section 6.1, Fig. 6.2). The differences are, however, not very large, especially compared to results obtained with $\nu = 0.8$ (see Fig. 9.1). Thus for efficiency, we prefer to apply large time steps. This is different in ENO schemes, which have in practice a more restricted CFL condition, due to the nonlinear optimal stencil search employed in the reconstruction.

9.1.2 Riemann Problem: Shock-Blast Interaction

In this rather challenging problem, suggested as benchmark in [87], two blast waves collide and generate multiple strong shocks and rarefactions in the computational domain $x \in [0, 1]$. Classical second-order TVD limiters, such as superbee are either too compressive and produce negative density or they are too diffusive and smear out the contact waves and local extrema. The initial data for the simulation consist of three constant states:

$$(\rho, v, p)_0 = \begin{cases} 
(1, 0, 10^3) & 0 \leq x < 0.1 \\
(1, 0, 10^{-2}) & 0.1 \leq x < 0.9 \\
(1, 0, 10^2) & 0.9 \leq x \leq 0.1 
\end{cases} \quad (9.2)$$
Reflecting boundary conditions are applied at both ends of the domain and the solution is evolved until $t = 0.038$ with $r = 1.0$. Fig. 9.3 and Fig. 9.4 compare the density, total energy and pressure profiles computed on a 800 cell grid with reference- and LDLR-data from [5].

The results obtained with the new scheme LimO3 with Courant number $\nu = 1.5$ approximates the solution more accurately than LDLR. In both the density and the energy distribution we can clearly observe a better resolution of the local extrema in the middle of the domain. The shock is resolved sharply within only two cells without any spurious oscillations and with the correct velocity. The left contact discontinuity is also very accurately approximated.

The Koren-limiter failed on this experiment. The calculation eventually broke down, producing unphysical densities when the two reflected shocks collided together in the middle of the domain. We suspect
that the Koren-limiter is too compressive for this experiment. Similar observations have been made with the second-order accurate, but compressive superbee limiter.

### 9.1.3 Riemann Problem: Shock-Acoustic-Wave Interaction

This test case was proposed in [63]. It simulates an interaction of a supersonic shock wave with a sinusoidal density disturbance involving smooth structures and a sharp shock. A Mach 3 shock wave moves into a smooth acoustic wave, which gets amplified and has higher frequency right behind the shock. We solve the 1D Euler equation on the domain \( x \in [-5, 5] \) with constant extrapolating boundary conditions. The initial conditions are prescribed as

\[
(\rho, v, p)_0 = \begin{cases} 
(3.857143, 2.629369, 10.333333) & x < -4 \\
(1 + 0.2 \sin(5x), 0, 1) & x \geq -4.
\end{cases}
\] (9.3)

In Fig. 9.5 we show computational results for the density component obtained at \( t = 1.8 \) using 400 FV-cells. We compare the solution for the proposed algorithm LimO3 with LDLR results from [5] (Fig. 9.5, right) and with the Koren-limiter (Fig. 9.5, left). Since the solution is dominated by smooth, but narrow structures we could expanded the asymptotic region \( \eta \) (4.34) setting the radius \( r = 1, r = 5 \) and \( r = 10 \). The effect is clearly visible. The local extrema in \( x \in [0.5, 2] \) are considerably better resolved increasing \( r \) and perfectly symmetric. We have to note that we could also run this Riemann problem without any limiting procedure, hence using the full parabolic spatial reconstruction Fig. 9.6 (left). Similar to the observation we have made for the linear advection equation (see Fig. 7.2), the unlimited third-order reconstruction remains stable with minor over- and under-shoots in the vicinity of the shock, consequently allowing us to choose a large asymptotic region. This is in contrast to classical second-order method, which could not simulate this experiment without proper TVD limiters. Yet even compressive TVD limiters, such as van Leer’s smooth harmonic mean limiter \( \phi_{vL} \) (3.44) needs around twice as many computational cells to resolve the smooth structure behind the shock, Fig. 9.6 (right).

![Figure 9.5: Density distribution of Shu–Osher shock-acoustic test case for Euler equations. Comparison at \( t = 1.8 \) of LimO3 \((r = 5)\) and LDLR using 400 cells.](image-url)
Note that if we set \( r \leq 1 \) we obtain qualitatively similar results to LDLR and the Koren-limiter. However in contrast to LDLR we produce no undershoots in the region with steep gradients \( x \in [-2, 0] \). This is identical to the Koren-limiter, which always limits steep gradients to first order. It is an advantage of our reconstruction, that we do not change the TVD property of the limiter \( \phi^{O(3)} \) (4.41) once a discontinuity is detected, i.e., the function \( \hat{\phi} \) is switched on. Enlarging the asymptotic region allows, already on a coarse mesh, the consideration of data with large curvature as smooth structures. This is different in the LDLR scheme. There the total-variation-control-parameter \( q \) has to be decreased. This eventually leads to an improved resolution of the smooth profile in \( x \in [0.4, 2.4] \), but also to more variation in the vicinity of discontinuities, because variation is introduced everywhere.

### 9.2 Characteristic Based Reconstruction

In the previous calculations the reconstruction has been applied on the conservative variables. In the following we will repeat the shock-blast and shock-acoustic test case, but utilize characteristic based reconstruction. We compare the proposed scheme with the superbee limiter and a compressive version of LimO3. The compressive version of \( \hat{\phi} \) (4.23) reads

\[
\hat{\phi}^{\text{comp}}(\theta) = \max \left( 0, \min \left( \frac{2 + \theta}{3}, \max \left( -0.5 \theta, \min \left( 2 \theta, \frac{2 + \theta}{3}, 1.5 \right) \right) \right), \min (2 \theta, 1) \right).
\]  

(9.4)

Similar to the superbee limiter, the function \( \hat{\phi}^{\text{comp}} \) is only Lipschitz continuous near \( \theta = 1 \) and utilizes the reconstruction \( \phi^\theta = 2 \theta \) for \( 0 \leq \theta \leq 0.5 \). The compressive LimO3 scheme, then employs \( \hat{\phi}^{\text{comp}} \) instead \( \hat{\phi} \) outside the asymptotic region. The second-order accurate MUSCL scheme uses the local Lax-Friedrichs flux (8.34), whereas LimO3 employs Roe’s approximative Riemann solver.
9.2. Characteristic Based Reconstruction

Figure 9.7: Density profile of shock-blast interaction problem for Euler equations. Left: Superbee limiter with LLF flux, $\nu = 0.8$. Right: LimO3 ($\tau = 1.0, \nu = 1.5$) utilizing Roe’s flux with $\phi^{\text{comp}}$ (compressive) and with $\hat{\phi}$ (standard). Comparison at $t = 0.038$ using 800 cells.

We envisage in Fig. 9.7 the results for the shock-blast interaction test case. The second-order MUSCL scheme with superbee limiter applied on the conservative variables produces strong oscillations at shocks. Both extrema are significantly overestimated and we can observe squaring effects on the left going rarefaction wave. In contrast, the result obtained with the characteristic based reconstruction is excellent. All shocks are accurately resolved without any spurious oscillations. The overall structure of the reference solution is very well matched. Yet the minimum in the middle of the domain is still overestimated.

In the right plot of Fig. 9.7 we can see the results obtained with LimO3 employing characteristic based reconstruction. Both results match the overall structure of the reference solution very well. As expected the scheme featuring $\hat{\phi}^{\text{comp}}$ is more compressive. We can observe a better resolved left going rarefaction
wave, but also spurious undershoots at the local minimum in the middle of the domain.

The differences between the compressive version of LimO3, \( \hat{\phi}^{\text{comp}} \) (9.4) and the standard version with \( \hat{\phi} \), are better visible in the shock-acoustic test case (see Fig. 9.8). The smooth, but narrow extrema are resolved more accurately with LimO3 featuring \( \hat{\phi}^{\text{comp}} \). Especially compared to van Leer’s MUSCL scheme with superbee limiter. The second-order accurate scheme is dispersive, which leads to a space shift in the transport direction of the narrow extrema. This effect completely vanishes for the third-order accurate LimO3 scheme. Although we employ a compressive limiter \( \hat{\phi}^{\text{comp}} \) no significant squaring effects are visible. These results confirm the previous discussions about compressive limiters. The order of the scheme is as essential as the utilized limiter. Obviously the results obtained with the standard \( \hat{\phi} \) limiter applied on characteristic variables are essentially identical to the previous results where the reconstruction is applied on the conserved fields. (see Fig. 9.5). It appears that for the proposed scheme LimO3, the shock capturing quality does not significantly change with the characteristic based reconstruction. This is for practical consideration quite important, since reconstructing the conserved variables is far more effective.

We should remark that we could not obtain a result with the superbee limiter and Roe’s approximative Riemann solver for the shock-blast experiment. The simulation eventually breaks down, producing unphysical densities, when the two shocks collide in the middle of the domain.

### 9.3 Magnetohydrodynamics (MHD)

We extend the Euler system to the ideal MHD system, which are model equation for the dynamics of plasma, i.e., charge-neutral ionized gas. Plasma dynamics are influenced by magnetic fields induced through the Lorentz-force. Therefore, an additional evolution equation for the magnetic field \( B \) has to be derived. Further terms in the Euler system (8.6) which quantify magnetic force and energy density are also needed. Using the conservative variables density \( \rho \), momentum \( \rho v_n \), magnetic field \( B \), and total energy \( E \), the ideal MHD equations can be written as follows [73]

\[
\begin{align*}
    \partial_t & \begin{pmatrix} 
\rho \\
\rho v_n \\
\rho v_t \\
B_t \\
E 
\end{pmatrix} + \partial_x \begin{pmatrix} 
\rho v_n \\
\rho v_n^2 + p + \frac{1}{2}B_t^2 \\
\rho v_n v_t - B_n B_t \\
v_n B_t - B_n v_t \\
v_n \left( E + p + \frac{1}{2}B_t^2 \right) - B_n B_t v_t 
\end{pmatrix} = 0, \\
(9.5)
\end{align*}
\]

Here the vectorial variables \( v \) and \( B \) are split into their scalar normal components \( v_n, B_n \) and their vectorial transversal components \( v_t, B_t \), respectively. We consider only ideal MHD gases. The total energy \( E \) is related to the pressure \( p \) by:

\[
E = \frac{1}{\gamma - 1} p + \frac{1}{2} \left( \rho \left( v_n^2 + v_t^2 \right) + B_t^2 \right). \\
(9.6)
\]

In all numerical experiments the adiabatic constant is \( \gamma = 5/3 \). All schemes employe the approximated Riemann solver of Wesenberg [86].
The intention of this section is twofold. First we consider a convergence test to prove the formal order of our scheme for a nonlinear rather complicated system. Second we compare the non-uniform converge for one dimensional MHD Riemann problems for different schemes as suggested by [73].

### 9.3.1 Empirical Order of Convergence

We solve the MHD equations on the domain \( x \in [-1, 1] \) with smooth initial data [4],

\[
(\rho, v_n, v_t, B_t, p) = (1.5 + 0.5 \sin(\pi x), 1.5 + 0.25 \cos(\pi x), (1.5 + 0.25 \sin(\pi x), 0), (1.0, 0), 0.25)
\]

and \( B_n = 0.5 \). The errors are calculated for \( t = 0.628 \) using a reference solution obtained on a very fine grid with the pure quadratic reconstruction (2.25, 2.26). We compare the convergence rates of the LimO3 and the LDLR method. The purpose is to envisage the particular difference of the radius \( r \) in the proposed asymptotic domain and the variation-control-parameter \( q \) of LDLR.

Fig. 9.9 shows the calculated empirical convergence error for the density distribution of the LimO3 scheme for various radii of the asymptotic region. In Fig. 9.10 we envisage the empirical convergence error of LDLR. LDLR has been implemented as described in [5]. Since the smooth extrema in the density profile is rather narrow peaked, both methods need around 200 computational cells to reach the desired convergence error.

We can observe, that LimO3 converges faster, in terms of grid size \( \Delta x \), than LDLR. Once the asymptotic region recognizes the smooth extrema the schemes recovers third-order accuracy. The smaller the asymptotic region, the more computational cells are eventually required.

The results obtained for LDLR approve the theoretical consideration of Section 5.3. Although variation
increases with decreasing parameter $q$, the curvature of the local extrema, i.e. $u''$ increases the value of local smoothness measure $\hat{a}(0, \mathcal{O}(\Delta x), \mathcal{O}(\Delta x))$ (5.23). Note, that $\hat{a}(0, \mathcal{O}(\Delta x), \mathcal{O}(\Delta x))$ converges for $q = 1$ only linearly toward zero. For $\hat{a} \to 0$, LDLR recovers eventually quadratic reconstruction.

### 9.3.2 Pseudo-Convergence

The aim of this section is to test the proposed scheme for a Riemann problem, where classical second-order, as well as advanced high-order schemes fail to recover the correct solution on coarse and intermediate grids. We consider an almost coplanar Riemann problem, as proposed in [73], to investigate the convergence properties towards the correct solution. It has been described in details in [73, 74], that schemes of up to a ninth-order accuracy converge initially to a wrong solution. Only with a very fine spatial resolution the methods start to converge to the right solution. As such, the Riemann problem is a hard test case for FV-methods.

The initial condition for the Riemann problem are

\[
(\rho, v_n, \mathbf{v}_t, \mathbf{B}_t, p) = \begin{cases} 
(1, 0 \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, 1) & x < 0 \\
0.2, 0 \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix}, 0.2) & x \geq 0
\end{cases}
\] (9.8)

There exist two solutions to the coplanar Riemann problem for $\alpha = \pi$. The first is a regular solution (r-solution), which contains only shocks and contact discontinuities, but the second solution features non-regular waves, namely compound waves (c-solution). For $\alpha$ different from $\pi$, the solution is regular and should only contain shocks and contact discontinuities. Torrilhon [73] and Torrilhon and Balsara [74] have found that finite volume schemes of different accuracy exhibit always the compound wave for moderate and intermediate resolutions. The methods eventually converge towards the correct physical solution.
only for very fine resolutions.

We investigate the error of the numerical solution obtained by the proposed method LimO3 and compare it to three different schemes: A second-order method (Roe2) with $\nu = 0.9$ (see [73]), a fifth-order WENO (WENO5) scheme with $\nu = 0.4$ (see [74] and references therein), and the LDLR method with $\nu = 0.9$. Roe2, LimO3 and LDLR are conducted with the same approximated Riemann solver and the higher-order methods, LimO3, LDLR and WENO5 use the same explicit three-stage Runge-Kutta time marching scheme. Roe2 utilizes the second-order Heun scheme. All results are obtained at $t = 0.4$ in the domain $x \in [-1, 1.5]$.

Fig. 9.11 envisages the described phenomena. Utilizing coarse or intermediate resolution a the scheme fails the exact physical solution. We can clearly observe that the compound wave remains and that the scheme also fails to recover the slow shock on the right hand side. As the grid gets significantly refined the compound wave eventually disappears and the exact solution is recovered. We should remark, that

![Figure 9.11](image_url)

**Figure 9.11:** Convergence studies of the almost co-planar MHD Riemann problem. Top: Numerical and exact solution with $\alpha = 3.0$. The numerical solution uses 800 computational cells. Bottom: Zoom in the interval $[-0.35, -0.1]$. 
even with $2 \cdot 10^4$ the compound wave pattern is still observable.

The left picture in Fig. 9.12 shows the convergence results of the magnetic field $B_y$ for an almost coplanar Riemann problem consisting only of regular waves (r-solution), i.e., shock and discontinuities. The solution for the Riemann problem consisting of any non-regular waves, c-solutions for compound waves, are shown in the right picture of Fig. 9.12. LimO3, LDLR and WENO5 exhibit eventually the same results.

All three methods start to converge with the same accuracy from 1000 cells upwards to the true solution. As already concluded in [74] very high order, such as WENO5, does not significantly improve the pseudo convergence although they display smaller numerical viscosity than second-order TVD methods (Roe2). Note that LimO3 qualitatively computes the same numerical solution as WENO5, yet using a nearly four times higher Courant number $\nu = 1.5$ and only three-data points for spatial reconstruction. Changing the radius of the asymptotic region did not improve the pseudo convergence.

Figure 9.12: Almost co-planar MHD Riemann problem. $L_1$-errors of the magnetic field $B_y$ for the twist angle $\alpha = 3.0$ with respect to the c-solution (left) and r-solution (right). Lim03 has $\nu = 1.5$ and $r = 0.5$. 
Chapter 10

Finite Volume Methods in Two Dimensions

We compute 2D experiments using a dimension by dimension reconstruction approach on a uniform computation grid, partitioned by finite volumes \( C_{ij}^n = [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}] \times [t^n, t^{n+1}] \). Given the cell average \( \bar{u}_{ij}^n \) at time \( t^n \)

\[
\bar{u}_{ij}^n \equiv \frac{1}{\Delta x \Delta y} \int_{y_{j-1/2}}^{y_{j+1/2}} \int_{x_{i-1/2}}^{x_{i+1/2}} u^n(x, y) \, dx \, dy,
\]

(10.1)

we approximate 2d conservation law (see eq. (8.1))

\[
\partial_t u(x, y, t) + \partial_x f(u(x, y, t)) + \partial_y g(u(x, y, t)) = 0
\]

(10.2)

by the conservative FV scheme

\[
\frac{d}{dt} \bar{u}_{ij} = L_i(\bar{u}_j) + L_j(\bar{u}_i)
\]

(10.3)

\[
= \frac{1}{\Delta x} \left( F_{i-1/2,j} - F_{i+1/2,j} \right) + \frac{1}{\Delta y} \left( G_{i,j-1/2} - G_{i,j+1/2} \right).
\]

The numerical flux functions

\[
F_{i+1/2,j} = F(\bar{u}_{i+1/2,j}^-, \bar{u}_{i+1/2,j}^+) \quad \text{and} \quad G_{i,j+1/2} = G(\bar{u}_{i,j+1/2}^-, \bar{u}_{i,j+1/2}^+)
\]

(10.4)

are approximations of the interface fluxes

\[
F_{i+1/2,j} \equiv \frac{1}{\Delta y} \int_{y_{j-1/2}}^{y_{j+1/2}} f(u(x_{i+1/2}, y, t)) \, dy \quad \text{and} \quad G_{i,j+1/2} \equiv \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} g(u(x, y_{j+1/2}, t)) \, dx.
\]

Integrating \( F_{i+1/2,j} \) we obtain

\[
F_{i+1/2,j} \equiv \frac{1}{\Delta y} \int_{y_{j-1/2}}^{y_{j+1/2}} f(u(x_{i+1/2}, y, t)) \, dy
\]

\[
= f(u(x_{i+1/2}, y_j, t)) + \frac{1}{24} \Delta y^2 f_{yy}(u(x_{i+1/2}, y_j, t)) + O(\Delta y^4).
\]

(10.5)
Consequently for higher than second-order accurate 2d schemes, we have to apply a numerical quadrature rule to get an accurate approximation for the numerical flux function in both directions. For schemes up to fourth-order accuracy one can simply approximate $f_{yy}(u(x_{i+1/2}, y_j, t))$ by a difference scheme. The numerical flux functions can than be approximated by

$$
F_{i+1/2,j} = \frac{11}{12} F_{i+1/2,j} + \frac{1}{24} F_{i+1/2,j-1} + \frac{1}{24} F_{i+1/2,j+1}
$$

(10.6)

$$
G_{i,j+1/2} = \frac{11}{12} G_{i,j+1/2} + \frac{1}{24} G_{i-1,j+1/2} + \frac{1}{24} G_{i+1,j+1/2}
$$

(10.7)

where the fluxes $F_{i+1/2,j-1}$, $F_{i+1/2,j+1}$ and $G_{i-1,j+1/2}$, $G_{i+1,j+1/2}$ are the known 1d numerical fluxes (see, e.g., LLF flux (8.34) or Roe’s flux (8.24)) evaluated at the top and bottom cell in $x$-direction and left and right cell in $y$-direction, respectively. Obviously this flux integral formula doesn’t need additional quadrature points, i.e., additional reconstructions. However it is still slightly more costly, than a simple flux evaluation at the midpoint $F_{i+1/2,j}$. This is the reason, why multidimensional FV schemes of higher than second-order accuracy are rarely used. Instead in practice conservative finite difference schemes are implemented up. The discussion of finite difference methods goes beyond this work and the interested reader is referred to [60].

We recover the interface values $\hat{u}_{i\pm \frac{1}{2},j}^{(\pm)}$ and $\hat{u}_{i,j\pm \frac{1}{2}}^{(\pm)}$ with the proposed third-order reconstruction $\phi^{O(3)}$ applied on the conserved variables. For efficiency reasons we utilize the HLLE approximated Riemann solver (8.31). Since the flux evaluation is done in both directions in every single Euler stage of the third-order SSP Runge-Kutta method, the stability criterion yields:

$$
\nu = 0.5 (\nu_x + \nu_y) \leq 1.6
$$

(10.8)

We compare the following three schemes:

<table>
<thead>
<tr>
<th>TVD-MUSCL</th>
<th>Heun $O(\Delta t^2)$ with van Leer’s harmonic limiter $\phi^{vL}$ (3.44)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RK3rd, LimO3, r = 1.0</td>
<td>RK3rd $O(\Delta t^3)$ with limiter $\phi^{O(3)}$ (4.41)</td>
</tr>
<tr>
<td>RK3rd, Koren</td>
<td>RK3rd $O(\Delta t^3)$ with Koren-limiter $\phi^{TVD}$ (4.21)</td>
</tr>
</tbody>
</table>

### 10.1 Convergence Test

We conduct an accuracy test for the 2d Euler equations proposed in [88]. The initial conditions are:

$$
(\rho, v_x, v_y, p)_0 = (1 + 0.2 \sin(\pi(x + y)), 1, -0.5, 1).
$$

(10.9)

The exact density solution at a later time $t$ is $\rho(t) = 1 + 0.2 \sin(\pi(x + y - 0.5 t))$, while $v_x$, $v_y$ and $p$ remain constant. Hence we simply advect the initial density profile. The computational domain is $(x, y) \in [-1, 1] \times [-1, 1]$ and the convergence errors are computed at $t = 1$.

Similar to the 1d results we obtain third-order accuracy as the mesh is refined. The larger the asymptotic
10.2 Shock-Bubble Experiment

In this experiment we simulate a strong shock wave moving rightwards over a low density gas bubble. The initial conditions and the experiment geometry are sketched in Fig. 10.2. Since the problem is symmetric we compute the upper part $y \geq 0$ only and specify reflecting boundary conditions at $y = 0$ and inflow boundary condition at $x = -0.1$. All results are obtained on a uniform mesh consisting of $1700 \times 500$ FV-cells. For LimO3 and $\phi^{TVD}$ we set the Courant number $\nu = 1.5$, whereas the TVD-MUSCL scheme

![Figure 10.2: Shock-Bubble experimental setup for 2D Euler equation.](image)
Figure 10.3: Emulated Schlieren pictures of shock-bubble simulation with $1700 \times 500$ cells with $\nu = 1.5$. Top: $t = 0.2$. Bottom $t = 0.4$.

Figure 10.4: Zoom of the shock-bubble Schlieren image Fig. 10.3. Second-order TVD-MUSCL scheme with van Leer’s limiter (3.44) with $\nu = 0.9$ at $t = 0.2$. 
10.3 2D Riemann Problem: Four-Shocks

Fig. 10.3 shows the numerical results of LimO3 at different times. The emulated Schlieren pictures display density gradients on a grey scale. In Fig. 10.4 and Fig. 10.5 the results at \( t = 0.4 \) are compared for all three schemes, namely TVD-MUSCL, LimO3 and Koren-limiter. We observe more small scale vortexes in the approximation of LimO3. The contact discontinuities are resolved more accurately with the proposed third-order accurate scheme. Note that due to admissible large time steps, the third-order accurate schemes are computationally as fast as the second-order TVD-MUSCL scheme. Although the accuracy of the solution is already improved with the Koren-limiter (Fig. 10.5, right), considerably more details are still captured with LimO3 (Fig. 10.5, left).

10.3 2D Riemann Problem: Four-Shocks

We consider two 2D Riemann problems originally defined in [55]. Both problems are solved on a square domain \((x, y) \in [0, 1] \times [0, 1]\), with initial constant states in four quadrants. The first Riemann problem features four shock waves with the contact point localized at \((3/4, 3/4)\) as proposed by [57]:

\[
\begin{align*}
(p, v_x, v_y, p)_0 &= \begin{cases} 
(1.5, 0, 0, 1.5) & [0.75, 1] \times [0.75, 1] \\
(0.5323, 1.206, 0, 0.3) & [0, 0.75] \times [0.75, 1] \\
(0.138, 1.206, 1.206, 0.029) & [0, 0.75] \times [0, 0.75] \\
(0.5323, 0, 1.206, 0.029) & [0.75, 1] \times [0, 0.75] 
\end{cases}
\end{align*}
\]  

(10.10)

We compare in Fig. 10.6 and Fig. 10.7 density contour profiles of all three schemes. The numerical solution is obtained at \( t = 0.8 \) with a Courant number of \( \nu = 1.5 \) for LimO3 and Koren-limiter and
Figure 10.6: Density contour plot of Euler-Four-Shocks problem using $1000 \times 1000$ cells at time 0.8. Second-order TVD-MUSCL scheme with van Leer’s smooth limiter (3.44) and $\nu = 0.8$. Number of density contours levels 75 from 0.2 to 1.7.

Figure 10.7: Density contour plots of Euler-Four-Shocks problem using $1000 \times 1000$ cells at time 0.8. Left: LimO3 with $\nu = 1.5$ and $r = 1.0$. Right: Koren-limiter $\phi^{\text{TVD}}$ (4.21) with $\nu = 1.5$. Number of density contours levels 75 from 0.2 to 1.7.

$\nu = 0.8$ for van Leer’s TVD-MUSCL scheme on a uniform grid with $1000 \times 1000$ computational cells. All boundaries are artificial boundaries, i.e., constantly extrapolated. Similar to the previous experiments, the amount and quality of details of the solution increases with the accuracy of the scheme. All three approximations essentially obtain the basic structure of the solution where the four shock waves interact. LimO3, however, resolves the contact discontinuity and the middle and top right plume with more details. The LimO3 and Koren-limiter results are perfectly symmetric. For intermediate to coarse grid resolution
the differences between all three methods where less distinctive. Note that LimO3 needs sufficient high resolution for small scale structures, since the asymptotic region \( \eta(\delta_{i-1/2}, \delta_{i+1/2}) \) (4.34) is mesh size dependent. On coarse grids the solutions obtained with LimO3 and Koren-limiter are of similar quality. Note that the LimO3 simulation is, because of Courant number \( \nu = 1.5 \), as fast as the second-order accurate TVD simulation, yet essentially of third-order accuracy in smooth regions. For comparison with different high-order schemes see [41].

10.4 2D Riemann Problem: Four-Contacts

This Riemann problem consists of two shock waves generated at the origin of the domain. The structure of the solution has the quality of a vortex turning clockwise. The initial conditions are:

\[
(\rho, v_x, v_y, p)_0 = \begin{cases} 
(1, 0.75, -0.5, 1) & [0.5, 1] \times [0.5, 1] \\
(2, 0.75, 0.5, 1) & [0, 0.5] \times [0.5, 1] \\
(1, -0.75, 0.5, 1) & [0, 0.5] \times [0, 0.5] \\
(3, -0.75, -0.5, 1) & [0.5, 1] \times [0, 0.5] 
\end{cases}
\] (10.11)

The generated shock waves propagate outwards following the slip lines, that are spiraled around the origin. We evolve the initial data until \( t = 0.8 \) on a uniform grid with 1000 \( \times \) 1000 cells. Similar to the previous experiments LimO3 (Fig. 10.9 left) exhibits the most detailed self similar solution. We observe a better resolved vortex at the top and bottom of the domain. In contrast to the second-order accurate schemes, LimO3 resolves the contact with more details, recovering small scale structures (compare Fig. 10.9 and Fig. 10.8). The second-order accurate MUSCL scheme is not symmetric and not able to

![Figure 10.8: Density contour plot of Euler-Four-Contacts problem using 1000\times1000 cells at time 0.8. Second-order TVD-MUSCL scheme with van Leer’s smooth limiter (3.44) and \( \nu = 0.8 \). Number of density contours levels 75 from 0.2 to 2.5.](image-url)
accurately resolve the small vortexes (see Fig. 10.8). For comparison with different high-order methods see [58].

10.5 Double Mach Reflection Problem

This is a quite difficult test case, which involves strong shocks as well as multiple Mach stems and a jet along the wall. A detailed discussion of this problem, including results for comparison can be founded in [87].

This problem involves a Mach 10 shock wave in air, which initially makes a $60^\circ$ angle with a reflecting wall. The computational domain is $[0, 4] \times [0, 1]$ and the Mach 10 shock is positioned at a $60^\circ$ angle with the $x$-axis at $(x, y) = (1/6, 0)$. The undisturbed air to the right of the shock has a density $\rho$ of 1.4 and a pressure $p$ of 1. To the left of the Mach 10 shock, the conditions are

$$ (\rho, v_x, v_y, p) = (8.0, 7.1447, -4.125, 116.5). $$  \hfill (10.12)

The region $0 \leq x \leq 1/6$ is always assigned the initial values, i.e., the exact post-shock condition is imposed at the bottom to present an angled wedge. At the top boundary the flow values are set to describe the motion of the shock, i.e., time-dependent conditions determined by the exact solution are set up and driven at the top. The region $1/6 < x \leq 4$ is a reflecting wall and at the left and right boundary we have supersonic inflow / outflow, respectively. The computation has been done on a $2000 \times 500$ grid with the Courant number $\nu = 1.5$.

Figure 10.9: Density contour plots of Euler-Four-Contacts problem using $1000 \times 1000$ cells at time 0.8. Left: LimO3 with $\nu = 1.5$ and $r = 1.0$. Right: Koren-limiter $\phi^{TVD}$ (4.21) with $\nu = 1.5$. Number of density contours levels 75 from 0.2 to 2.5.
In Fig. 10.10 we can see the differences between the second-order accurate MUSCL scheme and the proposed LimO3 algorithm. We can clearly observe that small structures of the complex flow are resolved more accurately. In particular, the rolled up slip line at the bottom, as well as the local extrema at $x = 2.62$, $y = 0.065$ are approximated sharper by LimO3.

Figure 10.10: Density contour plots of Euler double Mach reflection experiment on $2000 \times 500$ cells at time 0.2. Top: second-order MUSCL scheme and van Leer's smooth limiter with $\nu = 0.9$. Bottom: LimO3 with $\nu = 1.5$ and $r = 5.0$. Number of density contours levels 75 from 1.5 to 23.
Chapter 11

Outlook on Systems with Stiff Source Terms

In this chapter we are interested in the numerical solution of hyperbolic conservation laws with stiff relaxation terms. These systems occur in many physical applications, such as non equilibrium Thermodynamics. Our main focus, is the numerical simulation of macroscopic transport equations, which are derived from the kinetic theory of rarefied gas dynamics. The transport equations are balance laws (1.2), which satisfy for in one-dimension the differential form

$$\partial_t u(x, t) + \partial_x f(u(x, t)) = \frac{1}{\epsilon} s(u(x, t)),$$

(11.1)

where $0 < \epsilon \ll 1$ is a dimensionless parameter. The system (11.1) is called relaxation system and the parameter $\epsilon$ is called relaxation parameter. The relaxation parameter is associated with different time scales of the relaxation system. In many physical applications the dissipative / productive processes associated with the source occur on much faster time scales than the fastest wave speed of homogeneous system. Thus the relaxation parameter varies from values of order one to very small values, which in the later case results in balance laws with stiff source terms. In the zero relaxation limit, i.e., $\epsilon \to 0$, the solution $u(x, t)$ formally satisfies a reduced set of conservation equations, called equilibrium system [24], which in fact can be of different nature than the original relaxation system.

To illustrate this theory, we consider a simple model problem [25]

$$h_t + w_x = 0,$$

(11.2)

$$w_t + p_x(h) = \frac{1}{\tau} (f(h) - w),$$

where $0 < \tau \ll 1$ is the relaxation time. We assume for simplicity $p'(h) > 0$, whereby the system (11.2) is hyperbolic with characteristic speeds, i.e., eigenvalues $\pm \sqrt{p'(h)}$. In the asymptotic limit $\tau = 0$ we recover the reduced equation

$$h_t^\tau + f_x(h^\tau) = 0, \quad \text{with} \quad f(h) = w,$$

(11.3)
provided that the characteristic speed \( f'(h^\tau) \) of equation (11.3) satisfies the sub-characteristic condition [38]

\[
-\sqrt{p'(h)} \leq f'(h^\tau) \leq \sqrt{p'(h)}.
\] (11.4)

This stability condition guarantees that the relaxed system has the correct equilibrium velocity. Note that the variable \( h(x, t) \) is the conservative variable in equilibrium, whereas the variable \( w(x, t) \) in the original system (11.2) is a non conservative variable. The flux \( f(h^\tau) \) denotes the flux of the homogenous equilibrium system. It is the leading term in the Chapman-Enskog expansion

\[
w^\tau(x, t) = f(h^\tau(x, t)) + \tau w_1(x, t) + \mathcal{O}(\tau^2),
\] (11.5)

which is simply an expansion in deviations of the uniform local equilibrium. Substituting eq. (11.5) into eq. (11.2) we recover

\[
\begin{align*}
  h^\tau_t + f'(h^\tau) h^\tau_x &= -\tau (w_1)_x + \mathcal{O}(\tau^2) \\
  w_1 &= -f'(h^\tau) h^\tau_t - \tau (w_1)_t - p'(h^\tau) h^\tau_x + \mathcal{O}(\tau^2) \\
  &= (f'(h^\tau) - p'(h^\tau)) h^\tau_x + \mathcal{O}(\tau^2).
\end{align*}
\] (11.6)

If we combine both equations, we yield a second order PDE for \( h^\tau \)

\[
\begin{align*}
  h^\tau_t + f(h^\tau) h^\tau_x &= \tau \left[ (p'(h^\tau) - (f'(h^\tau))^2) h^\tau_x \right] + \mathcal{O}(\tau^2),
\end{align*}
\] (11.7)

with the flux function

\[
f(h^\tau) = w^\tau + \tau \left( p'(h^\tau) - (f'(h^\tau))^2 \right) h^\tau_x.
\] (11.8)

Equation (11.7) is the familiar advection-diffusion equation, which we have already seen in the Fourier analysis of Section 6.2. Similarly to eq. (6.33) we recover an ill-posed diffusion equation if the sub-characteristic condition (11.4) is violated. The advection-diffusion equation (11.7) is the continuous model equation for the long-time behavior \( \mathcal{O}(\tau) \) of the solution. Consequently the asymptotic expansion (11.5) yields a modified equation, which relates the characteristic velocity \( p'(h^\tau) \) of the original relaxation system (11.2) to diffusion and stability.

It is obvious that with the above procedure we get direct information on the flux function (11.8). Jin and Xin [27] utilized this concept for the numerical approximation of conservation laws. They introduced a relaxation approximation on the continuous equation, before they applied a suitable discretization. This led to the development of so called relaxation schemes, which instead of discretizing the original conservation system, discretize an equivalent relaxation system. The major advantage of this concept is, that nonlinear Riemann solvers are avoided, because the nonlinear fluxes are calculated via relaxation. The interested reader is referred to [27] for second-order accurate schemes respectively to [39, 53] for higher-order schemes.

We are rather interested in the development of efficient numerical schemes for stiff relaxation systems. Usually one would like to investigate the physical behavior described by the equilibrium system (11.5)
11.1 Finite-Volume Methods for Stiff Relaxation Systems

by solving the original relaxation system (11.2) utilizing coarse resolution \((\Delta t, \Delta x \gg \epsilon)\). Consequently we like to resolve all processes, including fast reactions, which occur on very short time scales, without necessarily utilizing very small time steps \(\Delta t\). In the following we will focus on so called under-resolved numerical scheme [24].

11.1 Finite-Volume Methods for Stiff Relaxation Systems

We are particularly interested in high-order accurate FV schemes, which are asymptotic preserving and asymptotically accurate. Asymptotic preserving schemes recover in the zero relaxation limit \(\epsilon \to 0\) a consistent and stable discretization of the equilibrium eq. (11.3). Whereas asymptotically accurate scheme preserve the order of approximation accuracy of the FV scheme in the stiff limit.

Identically to the presentation in Chapter 1, we recover a FV method for the balance law (11.1) by integrating over the discrete volume \(C_i\)

\[
\frac{d}{dt} \bar{u}_i = L_i(\bar{u}^n) + \frac{1}{\epsilon} S_i(\bar{u}^n). \quad (11.9)
\]

The value \(S_i(\bar{u}^n)\) denotes the discretization of the \(i\)-th cell average of the source term. It is defined as

\[
S_i(\bar{u}^n) = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} s(u^n(x)) \, dx. \quad (11.10)
\]

Thus for second-order accuracy, we can immediately recover the source term from the point value \(u^n_i\). For higher than second-order accurate schemes, a suitable quadrature formula is required to accurately approximate \(s''(u^n(x_i))\). In contrast to the discussion in Chapter 10 we cannot simply approximate \(s''(u^n(x_i))\) by central differences, since it holds

\[
s''(u^n(x_i)) = s''(\bar{u}^n_i) - \frac{1}{24} u''(u^n(x_i)) \Delta x^2 + O(\Delta x^4) \quad (11.11)
\]

To get a consistent reconstruction one has to also reconstruct the point wise values \(u(x_i), u''(x_i)\), respectively. Hence, for the discretization of the source term, cell averages of neighboring cells must be employed. This coupling makes FV scheme of higher than second-order accuracy quite impractical implementation wise. Pareschi and Russo [47] suggest to employ high-order conservative finite-difference schemes for systems with stiff relaxation. As already indicated in Chapter 10, a complete treatise on finite-difference schemes is beyond the scope of this work. We should, however, make clear that finite-difference methods are based upon point-wise reconstructions and therefore do not require neighboring cells for an accurate source term approximation. This makes finite-difference methods very attractive for numerical approximation of stiff balance laws.
11.2 Implicit-Explicit SSP Runge-Kutta Time Integration

The main difficulty for under-resolver schemes, is the severe stability constraint inherent to stiff ODE initial value problems. To overcome this time step restriction one typically utilizes implicit time integrators for the stiff source and explicit time integration schemes for the convection part. Such schemes are called implicit-explicit (IMEX) SSP Runge-Kutta scheme [9, 46, 47]. They utilize explicit SSP discretization (see Chapter 6) for the non stiff part $L_i(\bar{u}^{n})$. In the following we present some IMEX SSP Runge-Kutta schemes, which were originally introduced by Pareschi and Russo in [46, 47].

For the system of balance laws (11.1), the time marching scheme reads

$$\bar{u}_i^{(k)} = \bar{u}_i^n + \Delta t \sum_{l=1}^{k-1} \tilde{a}_{kl} L_i(\bar{u}^{(l)}) + \Delta t \sum_{l=1}^{k-1} a_{kl} S_i(\bar{u}^{(l)})$$

(11.12)

$$\bar{u}_i^{n+1} = \bar{u}_i^n + \Delta t \sum_{k=1}^{m} \tilde{b}_k L_i(\bar{u}^{(k)}) + \Delta t \sum_{k=1}^{m} b_k S_i(\bar{u}^{(k)})$$

(11.13)

The implicit scheme is diagonally implicit, i.e., for $A = (a_{kl}) \in \mathbb{R}^{m \times m}$ holds $a_{kl} = 0$, for $l > k$. For the time integration to be explicit in $L_i$, the matrix $\tilde{A} = (\tilde{a}_{kl}) \in \mathbb{R}^{m \times m}$ must satisfy $\tilde{a}_{kl} = 0$ for $l \leq k$. An IMEX time marching scheme essentially consists of a successive application of relaxation and convection steps. The relaxation step is always performed at first, which simplifies the solution of the algebraic equation.

The IMEX SSP Runge-Kutta schemes can be represented adequately in a double Butcher array notation. For the explicit SSP scheme we have $\frac{\tilde{c}}{b^T}$ and for the diagonally implicit scheme we have $\frac{c}{b^T}$, with the characteristic coefficients given by $\tilde{c} = \sum_{l=1}^{k-1} \tilde{a}_{kl}$ and $c = \sum_{l=1}^{k} a_{kl}$.

In Tab. 11.1 and Tab. 11.2 we have the tableaus of two second-order accurate IMEX schemes, namely IMEX-SSP2(3,2,2) and IMEX-SSP2(3,3,2). In Tab. 11.3 we present a third-order accurate IMEX scheme, with a four-stage implicit integrator. The first number in the name stands for the order of the explicit SSP Runge-Kutta integrator. The number triple in parenthesis represents in the given order, the number of stages of the implicit and of the explicit scheme and the accuracy order of the whole IMEX time marching scheme, respectively.

It is obvious that the IMEX-SSP2(3,2,2) utilizes the explicit second-order accurate Heun scheme (6.1), and the IMEX-SSP3(4,3,3) employs the explicit third-order accurate SSP33 scheme (6.2). Unfortunately there exists no third-order accurate IMEX scheme which consists of the explicit third-order four-stage SSP43 scheme (6.3) and is diagonally implicit. In fact all possible third-order accurate IMEX schemes, which utilize the explicit SSP43 integrator, have $a_{11} = 0$ in the triangular matrix $A = (a_{kl})$. Then it holds in general [47]

$$\lim_{\epsilon \to 0} s(\bar{u}^{(1)}) \neq 0.$$  (11.14)
11.3 Numerical Examples of Stiff Systems

In this section we will investigate the numerical behavior of the IMEX SSP time marching schemes combined with the proposed third-order accurate limiter $\phi^{O(3)}$ (4.41). We refer to this scheme as IMEX-LimO3, whereas IMEX-WENO stands for the third-order accurate WENO spatial reconstruction combined with IMEX time integration methods.

We apply the IMEX-LimO3 scheme to one simple model problem as well as to two more general relaxation systems. These include the shallow water system, the Broadwell system and Grad’s 13 moment equations. The first two are classical numerical test cases (see, e.g., [10, 24, 47, 52]), whereas the last one
The emphasis here is not on the physical properties of these models. We are rather interested in the numerical behavior of the scheme at zero relaxation limit. We do not perform empirical convergence studies, since these can be found for IMEX-WENO in [47] and for general IMEX Runge-Kutta schemes in [8, 9]. The schemes are implemented on finite-differences utilizing LLF flux (8.34) for the approximation of the non linear flux function. Reference solutions are obtained on very fine grids. All experiments are obtained utilizing an intermediate resolution of 200 grid cells.

### 11.3.1 Shallow Water Model

We consider a standard test case of shallow water flow taken from [24]. The relaxation system reads

\[
\partial_t \begin{pmatrix} h \\ w \end{pmatrix} + \partial_x \left( \begin{pmatrix} w h + \frac{1}{2} h^2 \\ h \end{pmatrix} \right) = \frac{1}{\epsilon} \left( \begin{pmatrix} 0 \\ \frac{1}{2} h^2 - w \end{pmatrix} \right),
\]

\[\alpha = 0.24169426078821, \beta = 0.06042356519705, \eta = 0.12915286960590\]

<table>
<thead>
<tr>
<th></th>
<th>1/2</th>
<th>0/6</th>
<th>2/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0/6</td>
<td>1/6</td>
<td>2/3</td>
</tr>
<tr>
<td>1/2</td>
<td>0/6</td>
<td>1/6</td>
<td>2/3</td>
</tr>
</tbody>
</table>

Table 11.3: Butcher-tableau for the third-order accurate scheme: IMEX-SSP3(4,3,3), explicit (left) and implicit (right).

**Figure 11.1:** Comparison of numerical solution of the shallow water model. Plot of the water height on 200 cells at \( t = 0.5, \epsilon = 10^{-8} \) and \( \nu = 0.8 \). Both scheme utilize IMEX-SSP3(4,3,3) time integrator. WENO results are from [46].
where \( h(x, t) \) is the water height, \( v(x, t) \) is the flow velocity and \( w(x, t) = h v \) is the flux. In the zero relaxation limit, we recover the inviscid Burgers equation

\[
h_t + \left( \frac{1}{2} h^2 \right)_x = \epsilon \left[ (1 + h - h^2)h_x \right]_x. \tag{11.16}
\]

We conduct an experiment with the initial conditions

\[
(h, v)_0 = (1 + 0.2 \sin(8 \pi x), h/2), \tag{11.17}
\]
on the computational domain \( x \in [0, 1] \) with periodic boundary conditions.

In Fig. 11.1 we compare the IMEX-SS3(4,3,3)-LimO3 and IMEX-SS3(4,3,3)-WENO scheme at \( t = 0.5 \) in the stiff regime \( \epsilon = 10^{-8} \). Both schemes are of third-order accuracy, yet WENO utilized a five-point stencil. We can observe that both schemes capture the shocks very accurately with the correct speed. IMEX-LIMO3, however, resolves the shocks with less cells. Since LimO3 utilizes a local reconstruction it is more efficient and its implementation is simpler.

### 11.3.2 Broadwell Model

The Broadwell model is a discrete kinetic model of the nonlinear Boltzmann equation of rarefied gas dynamics. For the derivation consult, e.g., [10, 24] (and references therein). The kinetic model is characterized by the relaxation system

\[
\begin{align*}
\partial_t \begin{pmatrix} \rho \\ m \\ z \end{pmatrix} + \partial_x \begin{pmatrix} m \\ z \\ m \end{pmatrix} &= \frac{1}{\epsilon} \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2} (\rho^2 + m^2 - 2 \rho z) \end{pmatrix},
\end{align*}
\tag{11.18}
\]

![Figure 11.2: Numerical solution of the Broadwell model (11.18) at \( t = 0.25 \), \( \epsilon = 10^{-8} \) with two different time marching schemes. Plot of density (left) and momentum (right) profile on 200 grid cells with \( \nu = 0.9 \) and \( r = 0.1 \).](image-url)
where $\rho$ is the density, $m = \rho v$ is the momentum and $\epsilon$ is the mean free path of the particles. Note that these are the only conservative variables. In local equilibrium, i.e., in zero relaxation limit, the non conservative fluid variable $z$ satisfies

$$z = \frac{1}{2\rho} (\rho^2 + m^2).$$

and the Broadwell system (11.18) reduces to the conservative system

$$\frac{\partial}{\partial t} \left( \begin{array}{c} \rho \\ m \end{array} \right) + \frac{\partial}{\partial x} \left( \begin{array}{c} m \\ \frac{1}{2\rho} (m^2 + \rho^2) \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \end{array} \right).$$

We consider a Riemann problem on $x \in [-1, 1]$ with initial conditions [24]

$$(\rho, v, z)_0 = \begin{cases} (1, 0, 1) & x \leq 0 \\ (0.2, 0, 1) & x > 0 \end{cases}$$

We investigate the second-order accurate IMEX-SSP$2(3,3,2)$ scheme and the third-order accurate IMEX-SSP$3(4,3,3)$ schemes for $\epsilon = 10^{-8}$. We are in particular interested in the shock capturing quality of the methods, as well as in the accuracy of the non equilibrium variable. IMEX-SSP2-LimO3 is formally of second-order, but utilizes $\phi^{O(3)}$ (4.41) for the spatial approximation.

In Fig. 11.2 we plot the density (left figure) and the momentum (right figure) profile. Both under-resolved schemes recover the shock formation in the fluid limit very well. We cannot observe any spurious oscillations. Although IMEX-SSP3-LimO3 is third-order accurate away from shocks, the difference compared with the second-order accurate scheme is rather marginal. In Fig. 11.2 (left plot) we compare the results obtained for the non conserved variable $z$ (11.19). In equilibrium it should hold

$$e \equiv z - \frac{1}{2\rho} (\rho^2 + m^2) = 0.$$
We can clearly observe that the third-order accurate scheme produces stronger deviation from zero. This is associated with the accuracy reduction of not stiffly accurate schemes discussed previously. Similar results can be found in [24].

11.3.3 Grad’s 13 Moments Equation

Grad’s 13 moment equations can be derived in the context of extended thermodynamics for rarefied gases (see, e.g., [65]). The degree of rarefaction of an ideal gas is generally expressed through the Knudsen number

\[ \text{Kn} = \frac{\lambda}{L}, \tag{11.23} \]

which is the ratio between the molecular mean free path \( \lambda \) and a characteristic length scale \( L \). In rarefied gas dynamics the Chapman-Enskog expansion models the nonlinear Boltzmann equation in terms of powers of \( \text{Kn} \). Close to its fluid dynamic limit, i.e., for zeroth-order expansion, we recover the Euler equations. Grad’s equations are a set of PDEs, which include first-order deviations from equilibrium. In one-dimension the governing system reduces to five equations with [72]

\[
\mathbf{u} = \begin{pmatrix} \rho \\ \rho v \\ \rho v^2 + 3p \\ \frac{2}{3}\rho v^2 + \sigma \\ \rho v^3 + 5pv + 2\sigma v + 2q \end{pmatrix}, \tag{11.24}
\]

\[
f(u) = \begin{pmatrix} \rho v \\ \rho v^2 + p + \sigma \\ \rho v^3 + 5pv + 2\sigma v + 2q \\ \frac{2}{3}\rho v^3 + \frac{4}{3}pv + \frac{4}{3}\sigma v + \frac{8}{15}q \\ \rho v^4 + 8pv^2 + 5\sigma v^2 + \frac{32}{5}qv + \frac{2}{p}(5p + 7\sigma) \end{pmatrix}, \tag{11.25}
\]

and the source

\[
s(u) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -\rho \sigma \\ -\rho(\frac{4}{3}q + 2\sigma v) \end{pmatrix}. \tag{11.26}
\]

The variables density \( \rho \), velocity \( v \) and pressure \( p \) are conserved fields, whereas the stress \( \sigma \) and heat flux \( q \) are not conserved fields.

We consider a Riemann problem on \( x \in [-1, 1] \) with the initial conditions [72]

\[
(\rho, v, p, \sigma, q)_0 = \begin{cases} 
(5, 0, 5, 0, 0) & x \leq 0 \\
(1, 0, 1, 0, 0) & x > 0
\end{cases} \tag{11.27}
\]
In Fig. 11.4 we compare the results obtained with a second-order accurate scheme IMEX-SSP2(3,3,2) utilizing van Leer’s smooth limiter (3.44) and IMEX-SSP3(4,3,3)-LimO3 for Kn ≡ ε = 10^{-6}. Both schemes are eventually of very similar quality. The shock structure of the equilibrium system, i.e., the Euler equation, is well recovered. Similarly to previous results we can observe that the non conserved fields, stress σ and heat flux q, are better resolved with the stiffly accurate second-order time marching schemes (compare Fig. 11.5). The results are in very good agreement with the results in [72].

Figure 11.5: Numerical solution of Grad’s 13 moment equation on 200 grid cells with ν = 0.9 and r = 0.1 at t = 0.35. Plot of the stress field (left) and heat flux field (right). Comparison of second-order scheme IMEX-SSP2(3,3,2) utilizing van Leer’s smooth limiter (3.44) and IMEX-SSP3(4,3,3)-LimO3.
Chapter 12

Conclusion and Discussion

In this thesis we have derived and analyzed a new third-order accurate limiter for the numerical solution of hyperbolic conservation laws. Furthermore we have analyzed classical drawbacks, such as accuracy degeneration at local extrema and squaring of smooth data, arising in second-order accurate reconstruction. We could show that squaring effects are associated with local instabilities, resulting from the local simulation of an ill-posed modified equation. In contrast to classical second-order TVD-limiters, the proposed limiter function maintains its formal accuracy at local extrema and recovers perfectly symmetric results.

The new scheme utilizes a local three point stencil. Thus remaining in the standard framework of second-order accurate TVD limiter. This makes the implementation of the proposed algorithm very simple and straightforward, since the spatial reconstruction remains a single function. The proposed limiter is an efficient piecewise linear \((\max, \min)\)-function, which acts as a logical switch depending on the left and right slope and avoids complicated and costly reconstruction techniques. The new limiter utilizes a local piecewise parabolic reconstruction for smooth data and preserves accuracy within the asymptotic region. Similarities to limiter functions obeying Harten’s TVD conditions have been analyzed and clearly revealed the advantages of third-order accuracy. The new algorithm uses, in contrast to second order methods, an explicit third-order SSP Runge-Kutta time marching scheme. Clear advantages in the frequency domain and less restrictive conditions for numerical stability for this time marching method could be shown.

Various numerical experiments indicate the superiority of the proposed limiter over classical second-order TVD limiters. The new scheme also compares favorably with third-order methods such as LDLR, LHHR, ENO / WENO and the Koren-limiter. It has very good shape-preserving properties and does not produce dispersive errors. Numerical simulations with Courant numbers up to \(\nu = 1.6\) make the proposed method just as efficient as classical second-order TVD-MUSCL schemes. Note that TVD methods are computationally more economical than ENO schemes [63]. Costly optimal stencil searches and memory storage are avoided.
We have also applied the new scheme on stiff relaxation systems utilizing IMEX time discretization. The results were satisfactory and of similar quality compared to results obtained with WENO spatial reconstruction. However, it turned out, that there exist no third-order accurate, asymptotic preserving, IMEX time marching schemes, which is stiffly accurate and does not require well prepared initial conditions. We consider this a major drawback, because such schemes in general do not guarantee an accurate approximation of the non conservative variables. A further attribute of third-order accurate IMEX FV schemes is, that, for consistency, the source term has to be approximated to third-order accuracy as well. This makes FV schemes quite impractical for the implementation. Therefore, we suggest to employ second-order, stiffly accurate IMEX time integration combined with the proposed third-order accurate spatial reconstruction.

Most of the research presented in this work is general and not restricted to the model equation calculated here. The new class of limiters, introduced throughout this thesis, can be utilized also in conjunction with other techniques for the computation of the numerical flux function.
Appendix A

Taylor Expansions for Error Analysis

The central idea for the numerical analyses of FV schemes is the cell mean value

\[ \bar{u}_i = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x) \, dx \]

\[ = u(x_i) + \frac{1}{24} u''(x_i) \Delta x^2 + O(\Delta x^4). \quad (A.1) \]

From this we can derive Taylor expansion for the left cell difference \( \delta_{i-1/2} = \bar{u}_{i} - \bar{u}_{i-1} \) and right cell difference \( \delta_{i+1/2} = \bar{u}_{i+1} - \bar{u}_{i} \):

\[ \delta_{i-1/2} = \Delta x u'(x_i) + \frac{1}{2} u''(x_i) \Delta x^2 + \frac{5}{24} u^{(3)}(x_i) \Delta x^3 + O(\Delta x^4) \]

\[ \delta_{i+1/2} = \Delta x u'(x_i) + \frac{1}{2} u''(x_i) \Delta x^2 + \frac{5}{24} u^{(3)}(x_i) \Delta x^3 + O(\Delta x^4). \quad (A.2) \]

Furthermore we get from Taylor expansion around \( x_i - \xi \Delta x \):

\[ u^{(3)}(x_i) = u^{(3)}(x_i - \xi) + O(\Delta x), \quad (A.3) \]

\[ u''(x_i) = \frac{1}{2} \left( 2u''(x_i - \xi) + 2\Delta x \xi u''(x_i - \xi) + O(\Delta x^2) \right), \quad (A.4) \]

\[ u'(x_i) = \frac{1}{6} \left( 6u'(x_i - \xi) + 6\Delta x \xi u''(x_i - \xi) + 3\Delta x^2 \xi^2 u^{(3)}(x_i - \xi) + O(\Delta x^3) \right), \quad (A.5) \]

Inserting eqs. (A.3, A.4, A.5) into eq. (A.2), we recover

\[ \delta_{i-1/2} = u'(x_i - \xi) \Delta x + \frac{\xi - \frac{1}{2}}{2} u''(x_i - \xi) \Delta x^2 \]

\[ + \frac{1}{24} (5 + 12(-1 + \xi)\xi) u^{(3)}(x_i - \xi) \Delta x^3 + O(\Delta x^4) \quad (A.6) \]

\[ \delta_{i+1/2} = u'(x_i - \xi) \Delta x + \frac{\xi + \frac{1}{2}}{2} u''(x_i - \xi) \Delta x^2 \]

\[ + \frac{1}{24} (5 + 12(1 + \xi)\xi) u^{(3)}(x_i - \xi) \Delta x^3 + O(\Delta x^4) \quad (A.7) \]
Another way to derive expression (A.6), is to first expand \( u(x) \) around \( x_i - \xi \Delta x \)

\[
\begin{align*}
  u(x) &= u(x_i - \xi) + (x - x_i + \xi \Delta x) u'(x_i - \xi) \\
  &+ \frac{1}{2}(x - x_i + \xi \Delta x)^2 u''(x_i - \xi) + \frac{1}{6}(x - x_i + \xi \Delta x)^3 u'''(x_i - \xi) + \mathcal{O}(\Delta x^4).
\end{align*}
\] (A.8)

To obtain the centered cell average \( \bar{u}_i \), we simply integrate eq. (A.8) from the left to the right cell boundary, respectively (see cell mean value (A.1)). We recover

\[
\bar{u}_i = u(x_i - \xi) + \xi \Delta x u'(x_i - \xi) + \frac{1}{24}(1 + 12 \xi^2) u''(x_i - \xi) \\
+ \frac{1}{24}\Delta x^3 (\xi + 4 \xi^3) u'''(x_i - \xi) + \mathcal{O}(\Delta x^4).
\] (A.9)

Identically we can derive the right cell averages \( \bar{u}_{i+1} \) and left cell average \( \bar{u}_{i-1} \), respectively. Subtracting the right from the left cell mean value gives us the formula for the right cell difference \( \delta_{i+1/2} \) (A.7).

With the above expansion for the lateral differences (A.6) and (A.7), we can obtain an explicit formula for the local smoothness measure \( \theta^{(\xi)} \). For monotone data, i.e., \( u'(x_i - \xi) \neq 0 \) the smoothness measure yields

\[
\begin{align*}
  \theta_i &= \frac{\delta_{i-1/2}}{\delta_{i+1/2}} \\
  &= \frac{u'(x_i - \xi) \Delta x + (\xi - \frac{1}{2}) u''(x_i - \xi) \Delta x^2 + \mathcal{O}(\Delta x^3)}{u'(x_i - \xi) \Delta x + (\xi + \frac{1}{2}) u''(x_i - \xi) \Delta x^2 + \mathcal{O}(\Delta x^3)} \\
  &= \frac{1 + (\xi - \frac{1}{2}) \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \Delta x + \mathcal{O}(\Delta x^2)}{1 + (\xi + \frac{1}{2}) \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \Delta x + \mathcal{O}(\Delta x^2)} \\
  &= \left(1 + \frac{\xi - \frac{1}{2}}{\xi + \frac{1}{2}} \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \Delta x\right) \left(1 - \frac{\xi + \frac{1}{2}}{\xi - \frac{1}{2}} \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \Delta x\right) + \mathcal{O}(\Delta x^2) \\
  &= 1 - \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \Delta x + \mathcal{O}(\Delta x^2) \\
  &= \theta_{i-1}.
\end{align*}
\] (A.10)

For the Taylor expansion, we assumed that the value \( (\xi + \frac{1}{2}) \frac{u''(x_i - \xi)}{u'(x_i - \xi)} \Delta x \) is bounded and goes to zero for as the grid is refined.
Appendix B

**Derivation of the Logarithmic Limiter**

In order to derive the logarithmic limiter (4.13) we have to formulate the cell interface values of LDLR as function of the smoothness measure \( \theta = \frac{\delta_{i-1/2}}{\delta_{i+1/2}} \) (3.3). The mesh size independent cell interface values of the LDLR are given by:

\[
\hat{u}_{i\pm1/2}^{(\pm)} = \bar{u}_i + c \eta^{(\pm)}(a) + d \eta^{(\pm)}(b),
\]

where \( a, b, c \) and \( d \) are explicit function of the left and right slopes [5]:

\[
a(\delta_{i-1/2}, \delta_{i+1/2}) = 1 - 2 \frac{|\delta_{i-1/2}|^q |\delta_{i+1/2}|^q}{|\delta_{i-1/2}|^{2q} + |\delta_{i+1/2}|^{2q}} = 1 - 2 \frac{|\theta|^q}{1 + |\theta|^{2q}} , \quad \delta_{i+1/2} \neq 0 \quad (B.2)
\]

We define a function depending solely on the slope ratio \( \theta \) (see eq. (4.14))

\[
p = p(\theta) \equiv 1 - a \quad (B.3)
\]

The functions \( b, c \) and \( d \) are algebraic expressions of \( a(\delta_{i-1/2}, \delta_{i+1/2}) \) and the left and right slopes. The logarithm functions come into the reconstruction via the functions [5]:

\[
\eta^{(\pm)}(z) = \begin{cases} 
- \ln(1 - z) + \frac{z}{2} & \text{for } \eta^{(+)}(z) \\
(z - 1) \ln(1 - z) - \frac{z}{2} & \text{for } \eta^{(-)}(z)
\end{cases}
\]

We can reformulate all these functions in terms of \( \theta \) and \( p(\theta) \):

\[
b = \frac{a}{a - 1} = \frac{p - 1}{p} \quad (B.5)
\]

\[
c = \frac{(a - 1)(\delta_{i+1/2}(1 - b) - \delta_{i-1/2})}{b - a} = \frac{p - \theta p^2}{1 - p^2} \delta_{i+1/2} \quad (B.6)
\]

\[
d = \delta_{i-1/2} - c = (\theta - \frac{p - p^2\theta}{1 - p^2})\delta_{i+1/2} \quad (B.7)
\]
Appendix B. Derivation of the Logarithmic Limiter

\[
\eta(a)^{(+)} = -\frac{\ln(p) + 1 - p}{(1 - p)^2} \quad (B.8)
\]

\[
\eta(a)^{(-)} = \frac{(-p) \ln(p) - 1 + p}{(1 - p)^2} \quad (B.9)
\]

To recover the reconstruction (4.11) with a single logarithmic limiter function (4.13), we have to plug the above eqs. (B.5, B.6, B.7, B.8, B.9) into eq. (B.1) and extract once the right \((\delta^{+1/2}_i)\) and the left \((\delta^{-1/2}_i)\) slope.
Bibliography


Curriculum Vitae

PERSONAL


EDUCATION

Ph.D. Program in Applied Mathematics  April 2004 – Present
Swiss Federal Institute of Technology (ETH), Seminar for Applied Mathematics, Zurich, Switzerland.

Diploma in Geophysics  March 2003
Ludwig-Maximilians-University, Institute of Geophysics, Munich, Germany.

Solid Mechanics Research Group of CSIRO Exploration and Mining, Nedlands, Western Australia and School of Mathematical Sciences, Monash University, Melbourne, Australia.

High School Diploma  February 1996
Gymnasium Fürstenried-West, Munich, Germany.