The Finite Volume Particle Method - A Meshfree Method of Second Order for the Numerical Solution of Hyperbolic Conservation Laws

Dissertation
zur Erlangung des Doktorgrades
der Fakultät für Mathematik, Informatik
und Naturwissenschaften
der Universität Hamburg

vorgelegt
im Fachbereich Mathematik

von
Libor Kadrnka
aus Ostrava, Tschechische Republik

Hamburg
2014

Als Dissertation angenommen vom Fachbereich
Mathematik der Universität Hamburg

Auf Grund der Gutachten von
Prof. Dr. Armin Iske
Prof. Dr. Jens Struckmeier

Hamburg, den 28. Januar 2015

Prof. Dr. Michael Hinze
Leiter des Fachbereichs Mathematik

Acknowledgements

I would like to thank my supervisor Prof. Dr. Armin Iske for his support and encouragement throughout my doctor studies. His patience and helpful advice contributed substantially to the finishing of this thesis.

Thanks also to my fellow PhD students for interesting discussions during the years. I would like to thank especially Dr. Christine Kaland, Dr. Claus Goetz and Dr. Martin Rybicki for fruitful discussions and useful advice in various topics. Particular thanks to Dr. Claus Goetz, Dr. Martin Rybicki and Annika Klüpfel for proofreading of this thesis.

My special thanks to my whole family and especially parents for their endless support throughout the years.

Contents

Introduction **1**

1 Preliminaries **3**
1.1 Hyperbolic conservation laws . . . 3
1.2 Finite volume method . . . 9
1.3 Generalized Riemann problem and the ADER method . . . 14
1.4 B-splines . . . 20

2 Finite-volume particle method **23**
2.1 The derivation of finite volume particle method . . . 23
2.2 Properties of FVPM . . . 29
2.3 Adding a particle . . . 36
2.4 Removing a particle . . . 48

3 Polyharmonic spline interpolation and the WENO method **55**
3.1 Polyharmonic spline interpolation . . . 55
3.2 WENO method . . . 63

4 A method of higher order **67**
4.1 Formulation of the method . . . 67
4.2 Local truncation error . . . 78
4.3 Stability analysis . . . 92
4.4 Convergence analysis . . . 98

5 Numerics **103**
5.1 Adding and removing a particle . . . 103
5.2 Higher order scheme . . . 108
5.2.1 Linear advection equation . . . 108
5.2.2 Burgers' equation . . . 110
5.2.3 Linear advection equation with peaks . . . 112
5.2.4 Linearized gas dynamics with a smooth solution . . . 118
5.2.5 Linearized gas dynamics with discontinuous solution . . . 121
5.2.6 Shallow water equations - the dambreak problem . . . 123
5.2.7 Euler equations with smooth solution . . . 126
5.2.8 Euler equations - the Sod problem . . . 129

Conclusions **133**

Bibliography **135**

List of Figures

1.1 An illustration of a triangulation of a non-rectangular domain 11
1.2 Comparison of a classical and a generalized Riemann problem 16
1.3 Example of linear B-splines B_j^1 21

2.1 Examples of functions ψ_i generated by function W in (2.8) 25
2.2 Examples of functions ψ_i generated by function W in (2.9) 26
2.3 Example of discretization errors in the scheme of adding a particle 44

3.1 1-d example of scaling a function ψ_{x_i} into ψ_{hx_i} 58

4.1 An illustration of two neighboring reconstructions 72
4.2 An illustration of two neighboring restricted reconstructions 73
4.3 Functions ψ_i, ψ_i^α and χ_i 74
4.4 An illustration of sets $g(\xi, \Delta x, \Delta t)$ 95

5.1 Example 1, adding a particle 105
5.2 Example 2, adding a particle 105
5.3 Example 3, adding a particle 106
5.4 Example, removing a particle 107
5.5 Linear advection equation. Solutions for $N = 38$ at time $t = 1$ 109
5.6 Burgers' equation. Solutions for $N = 70$ at time $t \doteq 0.85271$ 111
5.7 Linear advection equation with peaks. $HO22$ and $o1$ for $N = 200$ 114
5.8 Linear advection equation with peaks. $HO22$ and $o1$ for $N = 4000$ 115
5.9 Linear advection equation with peaks. $HO32$ and $o1$ for $N = 200$ 115
5.10 Linear advection equation with peaks. $HO32$ and $o1$ for $N = 4000$ 116
5.11 Linear advection equation with peaks. $HO32$ and LS for $N = 200$ 116
5.12 Linear advection equation with peaks. $HO32$ and LS for $N = 4000$ 117
5.13 Lin. gas dynamics with a smooth solution. First component 119
5.14 Lin. gas dynamics with a smooth solution. Second component 120
5.15 Lin. gas dynamics with discontinuous solution. First component 122
5.16 Lin. gas dynamics with discontinuous solution. Second component 122
5.17 Shallow water equations - the dambreak problem. First component 124
5.18 Shallow water equations - the dambreak problem. Second component 125
5.19 Shallow water equations - the dambreak problem. Zoom 125
5.20 Euler equations with smooth solution 128
5.21 Euler equations - the Sod problem 131
5.22 Euler equations - the Sod problem. Zoom 132

Introduction

In the course of technological progress more and more physical problems have become a matter of high interest as an object of investigation. In order to be able to find a solution of these often very complex problems simplified physical models have been introduced. In these models, even if the real problem is simplified, the important phenomena can still be found in the solution. The physical models consist usually of equations, algebraic or differential, and the goal is to find a function solving these equations. Mathematical methods have been developed to find this solution which is often a very challenging task depending on the character and properties of considered equations.

A special part of physics deals with hyperbolic conservation laws, describing e.g., conservation of mass, momentum or energy, leading to time-dependent hyperbolic *partial differential equations* (PDEs). Their fields of application are very diverse, for example computational fluid dynamics (CFD) is a numerical discipline used in car or plane construction as well as in power engineering. The hyperbolic PDEs can also be found in the area of electromagnetism, chemistry, acoustics, crystal growth or in various types of optimization such as shape or material optimization.

Hyperbolic PDEs have several properties that other equations such as elliptic or parabolic do not have. The most important characteristics are that the information propagates with a finite speed and that, independently from the smoothness of underlying data, discontinuities may occur in the solution. Therefore, a special class of methods has been developed over time that aims at solving the equations while at the same time dealing with discontinuities. These methods have to be *conservative*, which mimics the property of the exact solution that is conservative too, i.e., the mass of certain quantities can change only due to the flux through boundary.

One of the first methods developed in this field was the *finite difference method* (FDM), approximating the exact solution at given points in the computational domain, more precisely at nodes of a given computational mesh. However, it turned out that this method is not suitable for computations on complex geometries in higher dimensions with unstructured grid. Another method, the *finite volume method* (FVM), is more appropriate for this task and is one of the methods most widely used to solve hyperbolic conservation laws. This method approximates *local integral means* of the exact solution rather than point values and is conservative due to its construction.

The characteristic feature of the methods discussed above is that they require a mesh for the computation. Especially in higher dimensions, the construction and maintaining of the grid can become very costly or technically difficult. These issues resulted in a development of more flexible methods, so-called *meshfree methods*, that need no mesh for the computation.

The finite volume particle method (FVPM) is a method based on a combination of a meshfree particle method, in this case the so-called *Smoothed Particle Hydrodynamics* (SPH), (see [45]), and of the concept of FVM. The FVPM was first introduced by Hietel, Steiner and Struckmeier [24] and further developed by Junk and Struckmeier [29]. The basic idea is to substitute the *finite volumes* in FVM by volumes associated with *particle basis functions* but to preserve the computational model of *numerical flux* developed in FVM. Instead of a mesh, FVPM relies on an underlying structure given by *particles* - moving or non-moving points in the computational domain - and their interaction provided by the nearest neighbors. In fact, FVPM can be considered to be a generalization of the classical FVM, as stated by Junk [30]. The resulting FVPM is a highly flexible meshfree method suited to use for complicated or time-dependent geometries or for geometries in high dimensions, but still possessing the core feature of FVM - the concept of numerical flux and conservativity.

In the FVM framework, in the course of time, the need for higher accuracy of the method has arised

since the resolution of standard first order methods was not satisfactory for practical computations. The contemporary state-of-the-art method is the ADER (Arbitrary DERivatives) method. It allows for the construction of a numerical flux of arbitrarily high order of accuracy in time and space. It can be considered to be a generalized Godunov method (see [17]) and the fully numerical scheme was developed by Toro and Titarev in [65] by introducing their *Toro-Titarev solver*. The original Godunov method approximates the exact solution with a piecewise constant function defined on finite volumes and the time evolution is provided by solution of local *Riemann problems* on the interfaces between two finite volumes. The generalized Riemann problem generalizes the approximation in the sense that a piecewise polynomial function is reconstructed out of the piecewise constant data. As a result, the time evolution is provided by the solution of *generalized Riemann problems* defined by two polynomials, left and right of an interface. The Toro-Titarev solver was then developed to avoid solving this problem analytically which may become a very difficult task. Nevertheless, this method is still not an ultimate tool for solving hyperbolic PDEs. Computational problems were reported by Castro and Toro in [4] and by Montecinos et al. in [46] for non-linear systems. For a rigorous analysis and a potential remedy we refer to the work of Goetz [18].
In the FVPM framework, most papers deal with a first order meshfree method. To the author's knowledge, it is only in the paper [47] by Nestor, Basa, Lastiwka and Quinlan, that a second order FVPM is introduced, and there is also a method of second order provided by Teleaga in [59]. Nevertheless, there is no rigorous analysis of the convergence of the method. In our work, we are going to bridge this gap and develop a *meshfree method of second order of convergence in time and space*, for which we are able to show a standard result of convergence for a scalar linear equation analytically. For further equations, such as non-linear ones, or systems, we present examples that confirm the second order of convergence numerically.
Moreover, we introduce procedures enabling to *add or remove a particle* from a given general particle distribution which increases computational stability of the method.

The thesis is structured as follows: in the first chapter, *Preliminaries*, we introduce hyperbolic conservation laws and their important properties as well as their numerical treatment via FVM. Also, the concepts of the ADER method and of B-splines are presented.
In chapter 2, we present the derivation and properties of FVPM. We present a known *correction procedure* for geometrical coefficients and extend it to a case of bounded computational domain where boundary terms have to be treated. We formulate also a sufficient condition for the procedure to work properly. Finally, we introduce numerically exact *procedures to add or remove a particle* to the particle distribution for an arbitrary partition of unity used in FVPM. This serves to increase the stability of the method since low or even high density of particles at one place may cause difficulties during the computations. Numerical examples are given in the last chapter.
Polyharmonic spline interpolation and WENO method are presented in chapter 3. Interpolation with polyharmonic splines is a very powerful tool in the field of scattered data approximation and the concept can be also used in the field of PDEs. We adapt already known results to the case of FVPM, which is rather a technical issue, i.e., we analyse *polyharmonic spline interpolation of data given by weighted integral means.* The WENO method is presented since it is a helpful technique to suppress oscillations that may arise when using any type of approximation of discontinuous data.
Main results of this thesis can be found in chapter 4. We consider a one-dimensional conservation law and focus on non-moving particles coupled with linear B-splines. For this setting, we introduce a *method of second order of convergence in time and space.* Making use of the ADER method and Toro-Titarev solver, polyharmonic spline interpolation of the data and the WENO approach presented in the previous chapters leads to the desired scheme. We show the second order of consistency of the scheme for general one-dimensional scalar conservation law and the stability of the scheme for a linear case. Altogether, we prove convergence for a scalar linear PDE. The convergence is verified numerically on other prototype examples in the last chapter.
In chapter 5 several relevant examples are presented to demonstrate the quality of the developed scheme. The analytical results for a scalar linear equation are verified. We show numerically robustness and convergence of second order of the scheme for non-linear scalar equations, linear systems and non-linear systems with smooth data. Non-linear systems with discontinuities produce small non-physical oscillations. However, they do not lead to a blow-up of the whole numerical solution. Possible remedy strategies would be the use of limiters or a modification of the ADER scheme, lying beyond the scope of this thesis.

1 Preliminaries

In the first section of this chapter, the theory of hyperbolic conservation laws is introduced. A finite volume method is presented in order to solve these partial differential equations numerically. The classical formulation of this method leads to a method of first order of accuracy. To increase the order, the generalized concept of classical Riemann problem is shown as well as its utilization in the ADER (Arbitrary DERivatives) method. Finally, the theory of B-splines functions is presented. Altogether, based on this auxiliary theory we will be able to design a meshfree method for numerical solution of hyperbolic problems in chapter 4.

1.1 Hyperbolic conservation laws

We introduce a special type of partial differential equations called *hyperbolic conservation laws*. Conservation laws, or balance laws respectively, are used in physics to model a variety of problems such as fluid dynamics, magneto-hydrodynamics, electromagnetism, motion of elastic materials or traffic flow. Physical meaning of those equations is the conservation of certain quantities, such as mass, momentum, energy or another quantity, whereas the hyperbolicity stands for the mathematical structure of equations. The literature on these equations is vast. To mention just a few, see e.g., [3], [7], [10], [14], [16], [38], [39] or [53].

Let $\mathcal{U} \subset \mathbb{R}^m$ be an open and convex subset. A multidimensional system of conservation laws can be written as a system of partial differential equations in a d-dimensional space for $\mathbf{u} = \mathbf{u}(\mathbf{x}, t) : \mathbb{R}^d \times \mathbb{R}_+ \to \mathcal{U} \subset \mathbb{R}^m$

$$\mathbf{u}_t + \nabla \cdot \boldsymbol{F}(\mathbf{u}) = \mathbf{0} \qquad \text{in } \mathbb{R}^d \ , \ t > 0 \tag{1.1}$$

with initial conditions

$$\mathbf{u}(\ .\ , 0) = \mathbf{u}_0 \qquad \text{in } \mathbb{R}^d \ , \tag{1.2}$$

where $\mathbf{u} = (u_1, \ldots, u_m)^T \in \mathbb{R}^m$ is the vector of conserved quantities, $d \geq 1$, $m \geq 1$ and $\boldsymbol{F} \in \mathbb{R}^{m \times d}$ denotes the flux function of the conservation law, where $\boldsymbol{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_d)$, $\mathbf{f}_j = (f_{1j}, \ldots, f_{mj})^T \in \mathbb{R}^m$, $\mathbf{f}_j : \mathcal{U} \to \mathbb{R}^m$ and $\nabla \cdot \boldsymbol{F}(\mathbf{u}) := \sum_{j=1}^d \partial_{x_j} \mathbf{f}_j(\mathbf{u})$. The initial conditions are given by a function $\mathbf{u}_0 : \mathbb{R}^d \to \mathcal{U}$.
The name "conservation laws" is motivated with its physical meaning, that is, the quantities $\mathbf{u} = (u_1, \ldots, u_m)^T$ are conserved in the following sense:
Consider a bounded domain $\Omega \subset \mathbb{R}^d$ and denote $\mathbf{n} = (n_1, \ldots, n_d)^T$ the outer unit normal to the boundary $\partial\Omega$ of Ω. Then, integrating the equation (1.1) and using the divergence theorem, we obtain

$$\frac{d}{dt} \int_\Omega \mathbf{u} d\mathbf{x} + \sum_{j=1}^d \int_{\partial\Omega} \mathbf{f}_j(\mathbf{u}) n_j d\sigma = \mathbf{0} \ . \tag{1.3}$$

This equation can be then interpreted in the sense that $\int_\Omega \mathbf{u} d\mathbf{x}$ changes in time only due to the flux of $\mathbf{u}$ through the boundary $\partial\Omega$. In other words, the quantity $\mathbf{u}$ is conserved in Ω up to the flux of $\mathbf{u}$ through the boundary $\partial\Omega$.

Remark 1.1 (Boundary conditions)
If we solve the equation (1.1) on an open and bounded domain $\Omega \subset \mathbb{R}^d$*, then additional boundary conditions on the boundary* $\partial\Omega$ *of* Ω *have to be prescribed. We consider then a suitable boundary operator* B*, such that*

$$B(\mathbf{u}) = 0 \qquad \text{in } \partial\Omega \times (0,T) \ .$$

Further remarks on boundary conditions will be given later in this section.

Remark 1.2 (Source term)
If the right hand side of the equation (1.1) is non-zero, we speak about balance laws*, which is a more general case of equation (1.1). Then it has the form*

$$\mathbf{u}_t + \nabla.\ \boldsymbol{F}(\mathbf{u}) = \boldsymbol{S}(\mathbf{u}) \qquad \text{in } \mathbb{R}^d \ , \ t > 0 \ ,$$

where the vector function $\boldsymbol{S}(\mathbf{u})$ *is called a* source term. Shallow water equations *are a typical example of balance laws.*

Let us introduce the following definitions.

Definition 1.3
Let $\mathbb{A}_j(\mathbf{u}) := \frac{D\mathbf{f}_j(\mathbf{u})}{D\mathbf{u}}$ *denote the Jacobi matrices of* $\mathbf{f}_j$*,* $j = 1, \ldots, d$*.*

Definition 1.4 (Cauchy problem)
The problem (1.1) - (1.2) is called Cauchy problem.

Definition 1.5 (Hyperbolicity)
We say that a system of conservation laws (1.1) is (strictly) hyperbolic *if for any* $\mathbf{u} \in \mathcal{U}$ *and any* $\mathbf{n} = (n_1, \ldots, n_d)^T \in \mathbb{R}^d$*,* $\mathbf{n} \neq \mathbf{0}$*, the matrix*

$$\mathbb{A}(\mathbf{u}, \mathbf{n}) = \sum_{j=1}^{d} n_j \mathbb{A}_j(\mathbf{u})$$

has only real (and distinct) eigenvalues $\lambda_1, \ldots, \lambda_m$ *and* m *linearly independent right eigenvectors* $\mathbf{r}_1, \ldots, \mathbf{r}_m$*.*

Example 1.6 (Euler equations, [14], [64])
A typical example of a hyperbolic conservation law are the *Euler equations* used e.g., for modelling of flow in aeronautics, the aviation industry and steam or gas turbine design. They are given for $d = 1, 2, 3$ and $m = d + 2$ by the equation

$$\mathbf{u}_t + \sum_{s=1}^{d} \frac{\partial \mathbf{f}_s(\mathbf{u})}{\partial x_s} = \mathbf{0} \qquad \text{in } \mathbb{R}^d \ , \ t > 0$$

with the vector of unknowns

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \\ \vdots \\ \\ u_m \end{pmatrix} = \begin{pmatrix} \rho \\ \rho v_1 \\ \vdots \\ \rho v_d \\ E \end{pmatrix}$$

and physical fluxes

$$\begin{aligned} \mathbf{f}_s(\mathbf{u}) &= \begin{pmatrix} \rho v_s \\ \rho v_1 v_s + \delta_{1s} p \\ \vdots \\ \rho v_d v_s + \delta_{ds} p \\ (E+p) v_s \end{pmatrix} \\ &= \begin{pmatrix} u_{s+1} \\ u_2 u_{s+1}/u_1 + \delta_{1s}(\gamma-1)\big(u_m - \sum_{i=2}^{m-1} u_i^2/(2u_1)\big) \\ \vdots \\ u_{m-1} u_{s+1}/u_1 + \delta_{m-2,s}(\gamma-1)\big(u_m - \sum_{i=2}^{m-1} u_i^2/(2u_1)\big) \\ u_{s+1}\big(\gamma u_m - (\gamma-1)\sum_{i=2}^{m-1} u_i^2/(2u_1)\big)/u_1 \end{pmatrix} . \end{aligned}$$

One more equation is required to get a closed system of equations. For *ideal gases* it is the *equation of state*

$$p = (\gamma-1)\left(E - \rho|\mathbf{v}|^2/2\right) .$$

For the modelled gas, $\mathbf{v} = (v_1, \dots, v_d)^T$ is the velocity vector with components v_s in the directions x_s, $s = 1, \dots, d$, ρ is the density, p is the pressure and E is the total energy. The parameter $\gamma > 1$ is the *Poisson adiabatic constant.* The variables $\rho, v_1, \dots, v_d, p$ are called *primitive variables* and $u_1 = \rho$,$u_2 = \rho v_1$, …, $u_{m-1} = \rho v_d$, $u_m = E$ are called *conservative variables.*
Obviously $\mathbf{f}_s \in \mathcal{C}^1(\mathcal{U})^m$ with

$$\mathcal{U} = \Big\{ \mathbf{u} \in \mathbb{R}^m \ \Big| \ u_1 = \rho > 0 \ , \ u_s = \rho v_{s-1} \in \mathbb{R} \text{ for } s = 2, \dots, m-1 \ , \\ u_m - \frac{1}{2}\sum_{i=2}^{m-1} \frac{u_i^2}{u_1} = \frac{p}{\gamma-1} > 0 \Big\} .$$

The eigenvalues of the matrix $\mathbb{A}(\mathbf{u}, \mathbf{n})$ for the Euler equations are

$$\begin{aligned} \lambda_1(\mathbf{u},\mathbf{n}) &= \mathbf{v}\cdot\mathbf{n} - a|\mathbf{n}| \ , \\ \lambda_2(\mathbf{u},\mathbf{n}) &= \mathbf{v}\cdot\mathbf{n} \ , \\ &\ \vdots \\ \lambda_{m-1}(\mathbf{u},\mathbf{n}) &= \mathbf{v}\cdot\mathbf{n} \ , \\ \lambda_m(\mathbf{u},\mathbf{n}) &= \mathbf{v}\cdot\mathbf{n} + a|\mathbf{n}| \ , \end{aligned}$$

where $a = \sqrt{\gamma p/\rho}$ is the speed of sound.

Definition 1.7 (Classical solution)
We say that a function $\mathbf{u} : \mathbb{R}^d \times [0,\infty) \to \mathcal{U}$ *is a* classical solution *of the Cauchy problem (1.1) - (1.2), if*

a) $\mathbf{u} \in \mathcal{C}^1(\mathbb{R}^d \times (0,\infty))^m \cap \mathcal{C}(\mathbb{R}^d \times [0,\infty))^m$

b) $\mathbf{u}$ *satisfies (1.1) and (1.2) for all* $(\mathbf{x},t) \in \mathbb{R}^d \times (0,\infty)$ *and* $\mathbf{x} \in \mathbb{R}^d$*, respectively.*

In the linear one-dimensional case, i.e., $\boldsymbol{F}(\mathbf{u}) = \mathbb{A} \cdot \mathbf{u}$, where the constant matrix $\mathbb{A} \in \mathbb{R}^{m\times m}$ has m real eigenvalues and is diagonalizable due to the assumed strict hyperbolicity, the problem can be splitted into m independent scalar one-dimensional problems (1.1) and solved analytically with the *method of characteristics*. If the initial conditions (1.2) are smooth, the existence and uniqueness of the solution is proven (see e.g., [14]).

In the case of non-smooth initial conditions, the concept of the classical solution cannot be used. Moreover, for the nonlinear case, it is well known, that even for smooth data, a shock in the solution can develop in a finite time, which leads to the blow up of classical solution. Therefore, a generalization of classical solutions was introduced, known as *weak solutions*.

Definition 1.8 (Weak solution)
Let $\mathbf{u}_0 \in L^\infty(\mathbb{R}^d,\mathcal{U})^m$*. We say that a function* $\mathbf{u} : \mathbb{R}^d \times [0,\infty) \to \mathcal{U}$ *is a* distributional solution *to the Cauchy problem (1.1) - (1.2), if*

$$\int_{\mathbb{R}^d}\int_0^\infty \left(\mathbf{u}\varphi_t + \boldsymbol{F}(\mathbf{u})\cdot\nabla\varphi\right) dt d\mathbf{x} + \int_{\mathbb{R}^d} \varphi(\mathbf{x},0)\mathbf{u}_0(\mathbf{x}) d\mathbf{x} = 0 \quad \forall \varphi \in \mathcal{C}_0^\infty(\mathbb{R}^d \times [0,\infty)) \ ,$$

where $\mathcal{C}_0^\infty(\mathbb{R}^d \times [0,\infty))$ *denotes the space of all* C^∞*-functions with compact support inside* $\mathbb{R}^d \times [0,\infty)$*.*
Moreover, if $\mathbf{u}$ *is a distributional solution such that the mapping* $t \mapsto \mathbf{u}(\cdot,t)$ *is continuous from* $[0,\infty) \to L^1_{loc}$ *in each component, we say that* $\mathbf{u}$ *is a* weak solution*.*

It can be shown, that a smooth weak solution is a classical solution. Hence, weak solutions are indeed a generalization of classical solutions.
The weak solutions are, in general, not unique. In order to distinguish the physically relevant solution from the non-physical ones, the so-called *entropy weak solutions* can be used, motivated by the physical meaning of the solution.

Definition 1.9 (Entropy and entropy flux)
Let $\mathcal{U} \subset \mathbb{R}^m$ *be a convex set. A convex function* $\eta : \mathcal{U} \to \mathbb{R}$ *(*$\eta \in \mathcal{C}^1(\mathcal{U})$*) is called an* entropy *of system (1.1), if there exist functions* $\mathbb{G} = (G_1,\dots,G_d) : \mathcal{U} \to \mathbb{R}$*, called* entropy fluxes*, such that*

$$\nabla\eta(\mathbf{u}) \cdot \mathbb{A}_s(\mathbf{u}) = \nabla G_s(\mathbf{u}) \ , \quad \mathbf{u} \in \mathcal{U} \ , \ s = 1,\dots,d \ ,$$

whereas we mean $\nabla = \nabla_{\mathbf{u}}$*. The pair* $(\eta,\mathbb{G})$ *is called an* entropy-entropy flux pair*.*

Definition 1.10 (Entropy solution)
We say that a weak solution $\mathbf{u}$ *of (1.1)-(1.2) is an* entropy solution*, if for every entropy* η *of system (1.1) the condition*

$$\frac{\partial}{\partial t}\eta(\mathbf{u}) + \nabla\cdot\mathbb{G}(\mathbf{u}) \leq 0 \tag{1.4}$$

is satisfied in the distributional sense on $\mathbb{R}^d \times (0,\infty)$*, i.e.,*

$$\int_0^\infty \int_{\mathbb{R}^d} \left(\eta(\mathbf{u})\varphi_t + \sum_{j=1}^d G_j(\mathbf{u})\varphi_{x_j} \right) d\mathbf{x}dt \geq 0 \quad \forall \varphi \in \mathcal{C}_0^\infty(\mathbb{R}^d \times (0,\infty)) \ , \ \varphi \geq 0 \ .$$

The existence and uniqueness of weak solutions of the general system (1.1)-(1.2) is still an open problem. Nevertheless, some theoretical results are available.
The theorem 1.11 states the unique solvability for one scalar conservation law. The proof can be found in Dafermos [7].

Theorem 1.11
Let $f_j \in \mathcal{C}^1(\mathbb{R})$, $j = 1,\ldots,d$. *For any* $u_0 \in L^\infty(\mathbb{R}^d)$ *there exists a unique weak entropy solution* u *of the Cauchy problem for a scalar conservation law defined by*

$$\begin{aligned} u_t + \sum_{j=1}^d \frac{\partial f_j(u)}{\partial x_j} &= 0 \ , \quad \mathbf{x} \in \mathbb{R}^d \ , t > 0 \ , \\ u(\mathbf{x},0) &= u_0(\mathbf{x}) \ , \quad \mathbf{x} \in \mathbb{R}^d \ , \end{aligned}$$

and

$$\|u(\cdot,t)\|_{L^\infty(\mathbb{R}^d)} \leq \|u_0\|_{L^\infty(\mathbb{R}^d)} \ .$$

The total variation of a function is defined by

$$TV_{\mathbb{R}}(\mathbf{u}) = \sup\Big\{ \sum_{j=1}^{k-1} |\mathbf{u}(x_{j+1}) - \mathbf{u}(x_j)| \ \Big| \ x_1,\ldots,x_k \in \mathbb{R} \ , \ x_1 < x_2 < \ldots < x_k \ , \ k \in I\!N \Big\} \ .$$

The space of all functions with bounded variation is defined as

$$BV(\mathbb{R},\mathbb{R}^m) := \Big\{ \mathbf{v} \in L^1_{loc}(\mathbb{R},\mathbb{R}^m) \ \Big| \ TV_{\mathbb{R}}(\mathbf{v}) < \infty \Big\} \ .$$

The next theorem states that there exists a solution of the general one-dimensional problem (1.1)-(1.2) for "small data" and under some assumptions on the matrix $\mathbb{A} = D\mathbf{f}(\mathbf{u})/D\mathbf{u}$, where we write $\mathbb{A} := \mathbb{A}_1$ and $\mathbf{f} := \mathbf{f}_1$ for the case $d = 1$. The original proof was provided by Glimm in [16]. Bressan [3] could prove the statement using a different technique. Also the uniqueness in the class of BV functions with small BV-data for one-dimensional case was shown therein. Consider the previous definition of hyperbolicity. For the next theorem we need the following definition:

Definition 1.12 (Genuinely nonlinear, linearly degenerate)
Consider the case $d = 1$. *Let* $\nabla = \nabla_{\mathbf{u}}$.
We say that the k*-th characteristic field is* genuinely nonlinear*, if*

$$\nabla\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \neq 0 \qquad \forall \mathbf{u} \in \mathcal{U}$$

or linearly degenerate*, if*

$$\nabla\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0 \qquad \forall \mathbf{u} \in \mathcal{U} \ .$$

Theorem 1.13 ([14])
Let us assume that $d = 1$*, system (1.1) is strictly hyperbolic and all characteristic fields are either genuinely nonlinear or linearly degenerate in a neighborhood of a constant state* $\overline{\mathbf{u}}$*. Then there exist two positive constants* δ_1 *and* δ_2 *such that for initial data satisfying*

$$\|\mathbf{u}_0 - \overline{\mathbf{u}}\|_{L^\infty(\mathbb{R})^m} \leq \delta_1 \quad , \quad TV_{\mathbb{R}}(\mathbf{u}_0) \leq \delta_2 \ ,$$

the Cauchy problem (1.1)-(1.2) has a global weak entropy solution $\mathbf{u}(x,t)$ *in* $\mathbb{R} \times [0,\infty)$ *satisfying entropy inequality (1.4) in the sense of distributions for any entropy-entropy flux pair and*

$$\begin{aligned} \|\mathbf{u}(\cdot,t) - \overline{\mathbf{u}}\|_{L^\infty(\mathbb{R})^m} &\leq C_0\|\mathbf{u}_0 - \overline{\mathbf{u}}\|_{L^\infty(\mathbb{R})^m} \ , \quad t \in [0,\infty) \ , \\ TV_{\mathbb{R}}(\mathbf{u}(\cdot,t)) &\leq C_0 TV_{\mathbb{R}}(\mathbf{u}_0) \ , \quad t \in [0,\infty) \ , \\ \|\mathbf{u}(\cdot,t_1) - \mathbf{u}(\cdot,t_2)\|_{L^1(\mathbb{R})^m} &\leq C_0|t_1 - t_2| TV_{\mathbb{R}}(\mathbf{u}_0) \ , \quad t_1, t_2 \in [0,\infty) \ , \end{aligned}$$

for some constant $C_0 > 0$*.*

A special case of the Cauchy problem (1.1)-(1.2) in one spatial dimension is the *Riemann problem*, which is the Cauchy problem with initial data given by piecewise constant states. The study of the Riemann problem is important, since solution of this problem is a part of many modern numerical methods, e.g., in the framework of finite volume methods. Because of its simpler structure, it is also possible to find an analytic solution for some Riemann problems.

Definition 1.14
Consider the Cauchy problem for $d = 1$

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = \mathbf{0} \quad , \quad x \in (-\infty,\infty) \ , \ t > 0 \ , \tag{1.5}$$

$$\boldsymbol{u}(x,0) = \begin{cases} \boldsymbol{u}_L & , \quad x < 0 \ , \\ \boldsymbol{u}_R & , \quad x > 0 \ , \end{cases} \tag{1.6}$$

for constant states $\boldsymbol{u}_L$*,* $\boldsymbol{u}_R \in \mathcal{U}$*. This problem is called the* Riemann problem*.*

The solution of Riemann problem is *self-similar* as the next theorem claims. The proof and terminology can be found in [14] and [3]. We just remark that a *piecewise smooth weak solution* is a weak solution, which is piecewise smooth and on the interface of domains of smoothness special conditions for the shock (so-called *Rankine-Hugoniot conditions*, see [14] and the following example) have to be satisfied.

Example 1.15 (Rankine-Hugoniot conditions in 1D)
Consider the Riemann problem (1.5)-(1.6) and a function

$$\boldsymbol{u}(x,t) = \begin{cases} \boldsymbol{u}_L & , \quad x < \lambda t \ , \\ \boldsymbol{u}_R & , \quad x > \lambda t \ , \end{cases}$$

for some $\lambda \in \mathbb{R}$.
The Rankine-Hugoniot conditions for the shock are given by the relation

$$\lambda(\boldsymbol{u}_L - \boldsymbol{u}_R) = \boldsymbol{f}(\boldsymbol{u}_L) - \boldsymbol{f}(\boldsymbol{u}_R) \ .$$

Theorem 1.16
If the Riemann problem (1.5)-(1.6) has a unique piecewise smooth weak solution $\mathbf{u}$*, then* $\mathbf{u}$ *can be written for* $t > 0$ *in the similarity form* $\mathbf{u}(x,t) = \mathbf{D}(x/t)$*, where* $\mathbf{D} : \mathbb{R} \to \mathbb{R}^m$*.*

For the special case of genuinely nonlinear or linearly degenerate eigenvectors of the Jacobi matrix $\mathbb{A}(\mathbf{u})$ there exists an explicit unique solution of the Riemann problem as formulated in the following theorem. Again, the terminology and references for the proof can be found in [14].

Theorem 1.17
Let us assume that for each $\mathbf{u} \in \mathcal{U}$ *all eigenvalues* $\lambda_k(\mathbf{u})$ *of the matrix* $\mathbb{A}(\mathbf{u})$ *are simple and that every characteristic field is either genuinely nonlinear or linearly degenerate.*
Then to any $\mathbf{u}_L \in \mathcal{U}$ *there exists its neighborhood* $B(\mathbf{u}_L) \subset \mathcal{U}$ *such that the following statement holds: for any* $\mathbf{u}_R \in B(\mathbf{u}_L)$ *the Riemann problem (1.5)-(1.6) has a unique solution. This solution consists of at most* $m+1$ *constant states separated by simple waves or entropy shock waves or contact discontinuities. There is exactly one solution of this structure.*

1.2 Finite volume method

Now we are going to introduce a method for numerical solution of hyperbolic conservation laws (1.1)-(1.2), the *finite volume method*. The method is based on the integral formulation of partial differential equations leading to a *conservative* method. This property is very important since it mimics the property of the exact solution to conserve certain quantities. The finite volume mesh, introduced later in this section, offers also more flexibility on the method, e.g., in comparison with the finite difference method. The finite volume method became very popular in the domain of numerical solution of hyperbolic problems and theoretical results could be also revealed. The convergence of FVM to the entropy weak solution in special cases is briefly discussed in [14] and many references to this topic can be also found therein. Convergence theory for finite volume schemes is summarized in LeVeque [39]. Convergence of finite volume schemes in two dimension on unstructured grids is proven in Kröner [36]. See also the papers of Coquel and LeFloch [6] Chainais-Hillairet [5] and Vila [68] for further results.
In this section, we follow the derivation of the method introduced in [14]. First, we define the mesh and in the second step we introduce the numerical scheme itself.

Consider the equations (1.1)-(1.2) in an open and bounded computational domain $\Omega \subset \mathbb{R}^d$ written as

$$\mathbf{u}_t + \sum_{s=1}^{d} \frac{\partial \mathbf{f}_s(\mathbf{u})}{\partial x_s} = \mathbf{0} \qquad \text{in } \Omega \times (0,T) \tag{1.7}$$

with initial conditions

$$\mathbf{u}(\ .\ ,0) = \mathbf{u}_0 \qquad \text{in } \Omega \tag{1.8}$$

and boundary conditions

$$B(\mathbf{u}) = 0 \qquad \text{in } \partial\Omega \times (0,T)\ , \tag{1.9}$$

where B is a suitable boundary operator. We discuss the numerical treatment of boundary conditions later.

Finite volume mesh

Now, for the sake of simplicity, let us consider the case $d = 2$. The 3-dimensional case is treated in [14] and can be extended to an arbitrary dimension.
Denote by Ω_h a polygonal approximation of Ω (i.e., the boundary $\partial\Omega_h$ of Ω_h consists of finite number of closed simple piecewise linear curves). For the sake of simplicity, we assume, that Ω is already a polygon, i.e., $\Omega_h = \Omega$. Otherwise, one has to take the approximation error of Ω by Ω_h into consideration.
The set $\mathcal{D}_h = \{D_i\}_{i\in J}$ with $J \subset \mathbb{Z}^+ = \{0, 1, 2, \ldots\}$ and $h > 0$ is an index set and will be called *finite volume mesh* in Ω_h, if all D_i are closed polygons with mutually disjoint interiors such that

$$\overline{\Omega}_h = \bigcup_{i\in J} D_i \ .$$

The elements $D_i \in \mathcal{D}_h$ are called *finite volumes.* Two finite volumes $D_i, D_j \in \mathcal{D}_h$ are either disjoint or their intersection is formed by a common part of their boundaries ∂D_i and ∂D_j. For the sake of simplicity, we assume the intersection $\Gamma_{ij} = \partial D_i \cap \partial D_j$ to be a single point or a straight line (for the more general case of intersection formed by multiple straight lines see [14]). If the intersection Γ_{ij} is a straight line, we will call these finite volumes *neighbors.*
We denote by $\mathbf{n}_{ij}$ the unit outer normal to ∂D_i on Γ_{ij}, $h_i = \operatorname{diam}(D_i)$, $h = \sup_{i\in J} h_i$. Additionally, let $s(i) = \{j \in J \mid j \neq i \ , \ D_j \text{ is a neighbor of } D_i\}$.
The straight lines of the boundary $\partial\Omega_h$ are denoted by S_j and numbered by negative indices j forming an index set $J_B \subset \mathbb{Z}^- = \{-1, -2, \ldots\}$. Hence, $J \cap J_B = \emptyset$ and $\partial\Omega_h = \bigcup_{j\in J_B} S_j$. For a finite volume D_i adjacent to the boundary $\partial\Omega_h$, i.e., if $S_j \subset \partial\Omega_h \cap \partial D_i$ for some $j \in J_B$, we set

$$\begin{aligned} \gamma(i) &= \{j \in J_B \mid S_j \subset \partial D_i \cap \partial\Omega_h\} \ , \\ \Gamma_{ij} &= S_j \text{ for } j \in \gamma(i) \ . \end{aligned}$$

If D_i is not adjacent to $\partial\Omega_h$, then we put $\gamma(i) = \emptyset$. In order to stay consistent (we assumed that a non-empty intersection of any two neighboring finite volumes is a straight line), we assume also for all $i \in J$ that $\partial\Omega_h \cap D_i$ is either disjoint or a straight line. Again, intersections given by a single point are not considered. Hence, the number of elements of $\gamma(i)$ is at most 1.
By denoting $S(i) := s(i) \cup \gamma(i)$, we have for all $i \in J$

$$\begin{aligned} \partial D_i &= \bigcup_{j\in S(i)} \Gamma_{ij} \ , \\ \partial D_i \cap \partial\Omega_h &= \bigcup_{j\in\gamma(i)} \Gamma_{ij} \ . \end{aligned}$$

Remark 1.18
An example on a mesh is a triangular or quadrilateral mesh, consisting of triangle or quadrilateral finite volumes, respectively. For an illustration on a triangular mesh see the figure 1.1. For more details and also for some terminology of meshes, see [14].

Having the finite volume mesh constructed, we can now derive the finite volume scheme.

Finite volume scheme

Let us assume that $\mathbf{u} : \overline{\Omega} \times [0, T] \to \mathbb{R}^m$ is a classical solution of (1.7), $\mathcal{D}_h = \{D_i\}_{i\in J}$ is a finite volume mesh in a polygonal approximation Ω_h of Ω. Discretize the time interval $[0, T]$ by $0 = t^0 < t^1 < \ldots < t^{N_T} = T$ and denote $\Delta t^n = t^{n+1} - t^n$ the time step between t^n and t^{n+1}. Integrate the equation (1.7) over the set $D_i \times (t^n, t^{n+1})$ and use divergence theorem on D_i. We obtain

$$\int_{D_i} \left(\mathbf{u}(\mathbf{x}, t^{n+1}) - \mathbf{u}(\mathbf{x}, t^n)\right) d\mathbf{x} + \int_{t^n}^{t^{n+1}} \left(\int_{\partial D_i} \sum_{s=1}^{d} \mathbf{f}_s(\mathbf{u})(\mathbf{n}_i)_s d\sigma \right) dt = \mathbf{0} \ ,$$

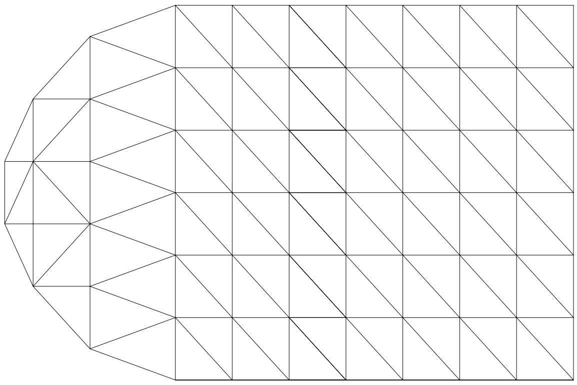


Figure 1.1: *An illustration of a triangulation of a non-rectangular domain.*

which can be rewritten as

$$\int_{D_i} \big(\mathbf{u}(\mathbf{x},t^{n+1})-\mathbf{u}(\mathbf{x},t^n)\big)\,d\mathbf{x}+\int_{t^n}^{t^{n+1}}\left(\sum_{j\in S(i)}\int_{\Gamma_{ij}}\sum_{s=1}^{d}\mathbf{f}_s(\mathbf{u})(\mathbf{n}_{ij})_s d\sigma\right)dt=\mathbf{0}\ .$$

Now we denote the integral averages over the finite volume D_i at time t^n by the value $\mathbf{u}_i^n$

$$\mathbf{u}_i^n=\frac{1}{|D_i|}\int_{D_i}\mathbf{u}(\mathbf{x},t^n)d\mathbf{x}\ ,$$

where $|D_i|$ stands for the area (2-dimensional volume) of the finite volume D_i.
Furthermore, the flux $\sum_{s=1}^{d}\mathbf{f}_s(\mathbf{u})(\mathbf{n}_{ij})_s$ of the quantity $\mathbf{u}$ is approximated with a *numerical flux*

$$\sum_{s=1}^{d}\mathbf{f}_s(\mathbf{u})(\mathbf{n}_{ij})_s\approx\mathbf{g}(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}_{ij})\ ,$$

where the function $\mathbf{g}(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}_{ij})$ depends on the value $\mathbf{u}_i^n$, the neighbors' values $\mathbf{u}_j^n$ and the outer normal vector $\mathbf{n}_{ij}$ between the neighboring finite volumes D_i and D_j. If $j\in J_B$, then there is no neighbor D_j and the value of $\mathbf{u}_j^n$ has to be specified on the basis of boundary conditions, see the corresponding remark below.

Remark 1.19
There are several possibilities to define the numerical flux in the literature. In the upper derivation, we approximate only the function $\sum_{s=1}^{d}\mathbf{f}_s(\mathbf{u})(\mathbf{n}_{ij})_s$ with the numerical flux. Also the spatial or the time integral can be involved in the definition. See e.g., section 1.3.

We end up with the *finite volume scheme*

$$\mathbf{u}_i^{n+1}=\mathbf{u}_i^n-\frac{\Delta t^n}{|D_i|}\sum_{j\in S(i)}\mathbf{g}(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}_{ij})|\Gamma_{ij}|\ ,\quad D_i\in\mathcal{D}_h\ ,\ t^n\in[0,T)\ . \tag{1.10}$$

The initial condition is defined by

$$\mathbf{u}_i^0=\frac{1}{|D_i|}\int_{D_i}\mathbf{u}_0(\mathbf{x})d\mathbf{x}\quad,\quad i\in J\ .$$

The numerical solution of the problem (1.7)-(1.8) via FVM is defined by

$$\mathbf{u}_h(\mathbf{x},t)=\sum_{n=0}^{N_T}\sum_{i\in J}\mathbf{u}_i^n\chi_i(\mathbf{x})\chi_{[t^n,t^{n+1})}(t)\ ,\ \mathbf{x}\in\Omega\ ,\ t\in[0,T]\ , \tag{1.11}$$

where χ_i denotes the characteristic function of interior of D_i for all $i\in J$.

Remark 1.20
In each time step, the numerical solution via FVM is defined as a piecewise constant function with constants $\mathbf{u}_i^n$ *given by the formula (1.10).*

Remark 1.21
The presented scheme is an explicit finite volume method. The construction of an implicit scheme as well as further generalizations can be found in [14].

Remark 1.22 (Boundary conditions)
The numerical flux function $\mathbf{g}(\mathbf{u}_i^n, \mathbf{u}_j^n, \mathbf{n}_{ij})$ *is well-defined for the interior volumes* D_i *for all* $i \in J$. *However, if* D_i *has a common line with* $\partial\Omega_h$, *i.e., it is a border finite volume, some of the values* $\mathbf{u}_j^n$ *have to be defined on the basis of boundary conditions. The treatment of boundary conditions is a very delicate problem and has to be implemented carefully. In principle, the imposition of boundary conditions is a physical problem, but it has to respect its mathematical structure. References to this topic can be found in [14] as well as an implementation of boundary conditions for compressible Euler equations used in practice. Further approaches can be also found in Teleaga [59] and LeVeque [39].*

We assume that the numerical flux $\mathbf{g}(\mathbf{u}, \mathbf{v}, \mathbf{n})$ is defined and continuous on $\mathcal{U} \times \mathcal{U} \times \mathcal{S}_1$, where $\mathcal{U} \subset \mathbb{R}^m$ is the domain of definition of the fluxes $\mathbf{f}_s$ and $\mathcal{S}_1$ is the unit sphere in $\mathbb{R}^d$.
To mimic the properties of the physical flux, the numerical flux has to fulfill certain requirements, given by the next two definitions. These requirements ensure, that constant states and mass will be conserved, see also section 1.1.

Definition 1.23
We say that a numerical flux $\mathbf{g}(\mathbf{u}, \mathbf{v}, \mathbf{n})$ *is* consistent, *if*

$$\mathbf{g}(\mathbf{u}, \mathbf{u}, \mathbf{n}) = \sum_{s=1}^{d} \mathbf{f}_s(\mathbf{u}) n_s \ , \quad \mathbf{u} \in \mathcal{U} \ , \ \mathbf{n} \in \mathcal{S}_1 \ .$$

Definition 1.24
We say that a numerical flux $\mathbf{g}(\mathbf{u}, \mathbf{v}, \mathbf{n})$ *is* conservative, *if*

$$\mathbf{g}(\mathbf{u}, \mathbf{v}, \mathbf{n}) = -\mathbf{g}(\mathbf{v}, \mathbf{u}, -\mathbf{n}) \ , \quad \mathbf{u}, \mathbf{v} \in \mathcal{U} \ , \ \mathbf{n} \in \mathcal{S}_1 \ .$$

The definition of the numerical flux $\mathbf{g}$ completes the numerical scheme. There can be found many numerical fluxes in the literature. We mention *Lax-Friedrichs* and *Steger-Warming* which we will present in the following examples. Other numerical fluxes are e.g., *Godunov*, *Vijayasundaram*, *Van Leer* etc. For further flux functions, also constructed in a very sofisticated way using reconstruction strategies and slope limiters, see e.g., [14], [39] or [64].

Example 1.25 ([14])
Let us define

$$\mathcal{A}(\mathbf{u}, \mathbf{n}) := \sum_{s=1}^{d} \mathbf{f}_s n_s \ .$$

The *Lax-Friedrichs numerical flux* is given by

$$\mathbf{g}(\mathbf{u}, \mathbf{v}, \mathbf{n}) = \frac{1}{2}\left(\mathcal{A}(\mathbf{u}, \mathbf{n}) + \mathcal{A}(\mathbf{v}, \mathbf{n}) - \frac{1}{\lambda}(\mathbf{v} - \mathbf{u})\right) \ , \quad \mathbf{u}, \mathbf{v} \in \mathcal{U} \ , \mathbf{n} \in \mathcal{S}_1 \ .$$

The parameter $\lambda > 0$ is independent of $\mathbf{u}, \mathbf{v}$, but depends in general on edges Γ_{ij}. Consider the scheme with constant time step Δt in two spatial dimensions with an uniform square mesh with edges parallel to x_1- and x_2-axes. Then the parameter λ is given by $\lambda = 2\Delta t/\Delta x$.

Example 1.26 ([14])
If we assume that the system (1.7) is strictly hyperbolic, one can rewrite the matrix $\mathbb{A}(\mathbf{u}) = \sum_{s=1}^{d} \frac{D\mathbf{f}_s(\mathbf{u})}{d\mathbf{u}} n_s$ as

$$\mathbb{A}(\mathbf{u}) = \mathbb{T}\mathbb{\Lambda}\mathbb{T}^{-1}$$

with a regular matrix $\mathbb{T}$ consisting of right eigenvectors of $\mathbb{A}$, where each column represents one eigenvector. The diagonal matrix $\mathbb{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ consists of the eigenvalues λ_i of $\mathbb{A}$. Clearly, $\mathbb{T}$ and $\mathbb{\Lambda}$ are functions of $\mathbf{u}$. Furthermore, let us define the matrices

$$\mathbb{\Lambda}^{\pm} = \mathrm{diag}(\lambda_1^{\pm}, \ldots, \lambda_m^{\pm}) \ ,$$

where

$$\lambda^{+} = \max(0, \lambda) \quad , \quad \lambda^{-} = \min(0, \lambda)$$

and

$$\mathbb{A}^{\pm} = \mathbb{T}\mathbb{\Lambda}^{\pm}\mathbb{T}^{-1} \ .$$

Then the *Steger-Warming numerical flux* is defined as

$$\mathbf{g}(\mathbf{u}, \mathbf{v}, \mathbf{n}) = \mathbb{A}^{+}(\mathbf{u}, \mathbf{n}) \cdot \mathbf{u} + \mathbb{A}^{-}(\mathbf{v}, \mathbf{n}) \cdot \mathbf{v} \ , \quad \mathbf{u}, \mathbf{v} \in \mathcal{U} \ , \mathbf{n} \in \mathcal{S}_1 \ .$$

Another very important component of the finite volume method is the *CFL-condition* (Courant-Friedrichs-Lewy) defining the relation between the time step size Δt^n and mesh size $h = \Delta x$ in order to ensure stability of the method. This condition is only a necessary, not sufficient, condition for the stability, and therefore also for the convergence. It can be interpreted in the sense given by LeVeque [39]: "A numerical method can be convergent only if its numerical domain of dependence contains the true domain of dependence of the PDE, at least in the limit as Δt and Δx go to zero."

Example 1.27 ([14])
For the nonlinear system

$$\begin{aligned} \mathbf{u}_t + \mathbf{f}(\mathbf{u})_x &= \mathbf{0} \quad \text{in } \mathbb{R} \times (0, \infty) \\ \mathbf{u}(x, 0) &= \mathbf{u}_0(x) \ , \ x \in \mathbb{R} \end{aligned}$$

solved by *Lax-Friedrichs scheme* the CFL-condition reads

$$\Delta t^n \leq \mathrm{CFL} \frac{\Delta x}{\sigma(\mathbb{A}(\mathbf{u}_i^n))} \ ,$$

where the suitable number $\mathrm{CFL} \in (0, 1)$, $\mathbb{A}(\mathbf{u}) = \frac{D\mathbf{f}(\mathbf{u})}{D\mathbf{u}}$ is the Jacobi matrix of $\mathbf{f}$ and $\sigma(\mathbb{A})$ denotes the spectral radius of matrix $\mathbb{A}$.

A numerical flux of arbitrary high order of accuracy, the so-called *ADER flux* will be studied in the next section.

1.3 Generalized Riemann problem and the ADER method

In this section, we will present a construction of a finite volume method for a balance law in one spatial dimension as proposed by Toro and Titarev [65] and Toro [64]. This method will be of arbitrary high order of accuracy in time and space. We start with a derivation of the method and motivate with this derivation the introduction of a generalization of the Riemann problem from section 1.1. For the sake of completeness, we will present the method for a one-dimensional balance law (i.e., conservation law with a source term) rather than for a conservation law (compare also to (1.1))

$$\begin{aligned} \boldsymbol{u}_t + \boldsymbol{F}(\boldsymbol{u})_x &= \boldsymbol{S}(\boldsymbol{u}) \quad , \quad x \in \Omega \ , \ t > 0 \ , &(1.12)\\ \boldsymbol{u}(x,0) &= \boldsymbol{u}_0(x) \ , &(1.13) \end{aligned}$$

where $\boldsymbol{u}_0$ is a given initial condition.
A finite volume method for the solution of (1.12)-(1.13) can be constructed as follows (compare with section 1.2). For the sake of simplicity, we will work with uniform discretizations in time and space. Discretize the computational domain Ω with points $x_{i+\frac{1}{2}}$ with $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and define the computational cells $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. The time interval $[0,T]$ is discretized with points t^n and the time step size $\Delta t = t^{n+1} - t^n$. The set $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [t^n, t^{n+1}]$ defines a control volume in the computational and time domain $\Omega \times [0,T]$.
Integrating (1.12) over the control volume yields the exact equation

$$\boldsymbol{u}_i^{n+1} = \boldsymbol{u}_i^n - \frac{\Delta t}{\Delta x}\left(\boldsymbol{F}_{i+\frac{1}{2}} - \boldsymbol{F}_{i-\frac{1}{2}}\right) + \Delta t \boldsymbol{S}_i \ , \tag{1.14}$$

where the cell average

$$\boldsymbol{u}_i^n = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \boldsymbol{u}(x,t^n)dx \tag{1.15}$$

is the spatial integral average of $\boldsymbol{u}(x,t)$ at time $t = t^n$. Further terms are

$$\begin{aligned} \boldsymbol{F}_{i+\frac{1}{2}} &= \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \boldsymbol{F}(\boldsymbol{u}(x_{i+\frac{1}{2}},t))dt \ , &(1.16)\\ \boldsymbol{S}_i &= \frac{1}{\Delta t \Delta x}\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \boldsymbol{S}(\boldsymbol{u}(x,t))dxdt \ , &(1.17) \end{aligned}$$

denoting the time integral mean of the physical flux and the space-time integral mean of the source term, respectively. Suitable approximations of the terms $\boldsymbol{F}_{i+\frac{1}{2}}$ and $\boldsymbol{S}_i$ yield a numerical method. These approximations are then called *numerical flux* and *numerical source*, respectively. The construction of the numerical flux is the core of the numerical method, allowing a higher order of accuracy of the scheme. It should be remarked, that the numerical flux $\boldsymbol{F}_{i+\frac{1}{2}}$ is constructed as a time integral mean of the physical flux, instead of evaluating the flux at a fixed time as done in section 1.2.

The idea of the method dates back to the seminal work of Godunov [17]. In the finite volume framework, the exact solution of the equation (1.12) is approximated with its cell averages (1.15) at the time step t^n, building a piecewise constant approximation (1.11). The cell averages are then evolved to the next time step t^{n+1} solving a local Riemann problem at each cell interface $x_i := \frac{1}{2}(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})$. These evolved cell averages define then the piecewise constant numerical solution at the next time step t^{n+1}. The choice of initial data for the local Riemann problems is therefore crucial for the functionality and the accuracy of the scheme. For a scheme of first order of accuracy the cell average from the left and the cell average from the right of the cell interface x_i are taken as the initial data for the local Riemann problem which is then solved exactly or approximatively. These cell averages can be understood as a first order reconstruction of the exact solution in the corresponding cell. We call this the *Godunov method.*
For a scheme of higher order than one, the exact solution is approximated within each cell utilizing

the neighboring cell average data. This yields a higher order reconstruction. These reconstructions are then used as the initial data for the local Riemann problem. Since they are no longer constant, we have to consider the generalized Riemann problem. The whole scheme can be then considered as a generalized Godunov method. This concept was introduced first by Kolgan in [35] followed by van Leer in [66] and [67].

Definition 1.28
Consider the Cauchy problem

$$\begin{aligned} \boldsymbol{u}_t + \boldsymbol{F}(\boldsymbol{u})_x &= \boldsymbol{S}(\boldsymbol{u}) \quad , \quad x \in (-\infty, \infty) \ , \ t > 0 \ , && (1.18) \\ \boldsymbol{u}(x,0) &= \begin{cases} \boldsymbol{u}_L(x) & , \quad x < 0 \ , \\ \\ \boldsymbol{u}_R(x) & , \quad x > 0 \ , \end{cases} && (1.19) \end{aligned}$$

where $\boldsymbol{u}_L(x)$, $\boldsymbol{u}_R(x)$ are vectors and their components are assumed to be smooth functions of x, with K continuous, non-trivial spatial derivatives away from zero. That is, if the non-negative integers K_L and K_R are minimal numbers such that

$$\begin{aligned} \frac{d^k}{dx^k}\boldsymbol{u}_L(x) &\equiv 0 \qquad \forall k > K_L \ , \ \forall x < 0 \ , \\ \frac{d^k}{dx^k}\boldsymbol{u}_R(x) &\equiv 0 \qquad \forall k > K_R \ , \ \forall x > 0 \ , \end{aligned}$$

and $K := \max\{K_L, K_R\}$.
This problem is then called the generalized Riemann problem *of order K, denoted by GRP_K.*

If we consider the case that the system (1.18) is strictly hyperbolic such that every characteristic field is either genuinely nonlinear or linearly degenerate, then there exists a neighborhood around the origin in which the problem (1.18)-(1.19) has a unique entropy weak solution, provided the jump in the initial condition $|\boldsymbol{u}_L(0) - \boldsymbol{u}_R(0)|$ is sufficiently small. Moreover, at least for small times $t > 0$, the solution of the generalized Riemann problem and the solution of the corresponding classical Riemann problem with initial data $\boldsymbol{u}_L(0)$ and $\boldsymbol{u}_R(0)$ have similar wave structure. For details, see [18], [40] and [57]. An illustration is given in figure 1.2. Details about the wave structure and existence and uniqueness results on the solution of the classical Riemann problem were given in theorem 1.17. The similarity of the wave structures allows to approximate GRP with a sequence of classical Riemann problems as the ADER method does, proposed by Toro and Titarev [65], yielding a robust scheme of higher order. However, it was observed in [4] and [46], that for non-linear systems, numerical difficulties occur if there is a large jump in the initial data of GRP. This phenomenon was described rigorously by Goetz [18]. An analysis of the ADER method concerning consistency and stability can be found in [56] and [62].

Remark 1.29
Note that the definition of generalized Riemann problem is indeed a generalization of the Riemann problem, since GRP_0 defines the classical Riemann problem with piecewise constant initial data.

The *Toro-Titarev solver* works as follows. To define the numerical flux function, the exact solution $\boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + t)$ in (1.16) is approximated by $\boldsymbol{u}_{LR}(t)$ at the interface $x_{i+\frac{1}{2}}$ and then a suitable numerical quadrature is used to achieve the desired accuracy. The function $\boldsymbol{u}_{LR}(t)$ is defined as a truncated Taylor expansion of the exact solution in the time variable. This is a so-called *state expansion.* We only remark that also another approach can be applied - the *direct expansion* of the flux function $\mathbf{F}(\boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + t))$, see e.g., [56].
Using the *Cauchy-Kowalewski procedure*, the time derivatives are expressed in terms of spatial derivatives. These spatial derivatives are defined by the solution of a local generalized Riemann

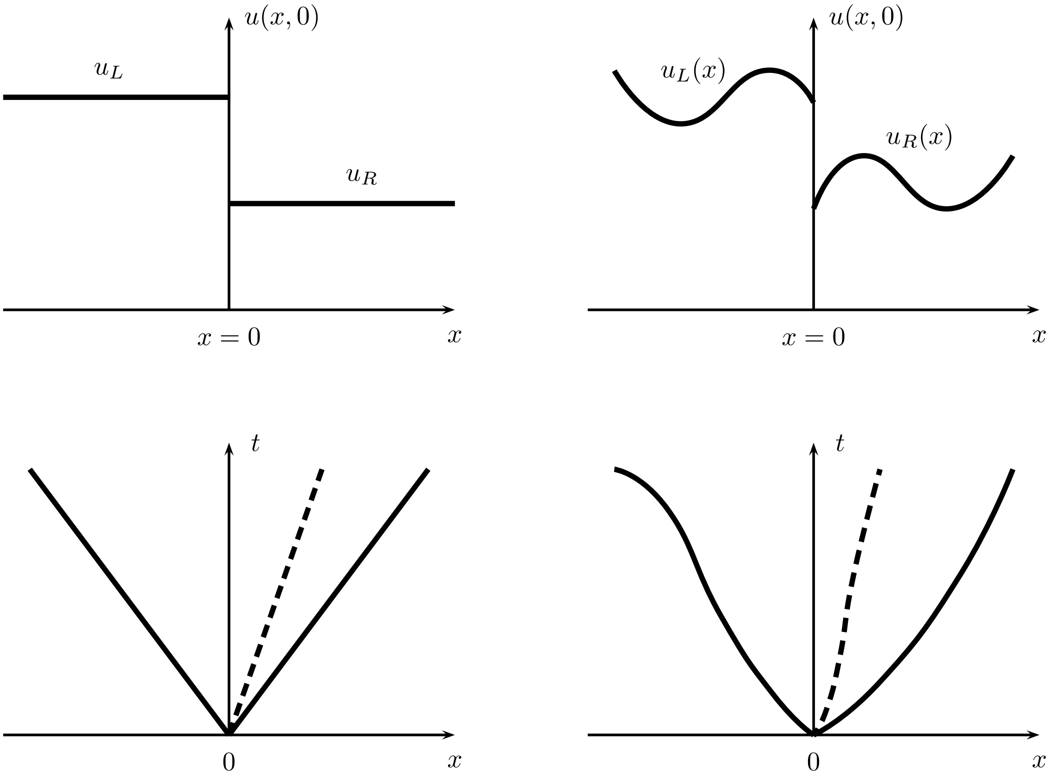


Figure 1.2: *Comparison of a classical and a generalized Riemann problem.*
On the left: classical Riemann problem - above the piece-wise constant initial condition for a single component of $\boldsymbol{u}(x,0)$*, in the bottom the corresponding wave structure of the solution in the* $x-t$ *plane. Characteristics are straight lines here.*
On the right: generalized Riemann problem - above the piece-wise smooth initial condition for a single component of $\boldsymbol{u}(x,0)$*, in the bottom the corresponding wave structure of the solution in the* $x-t$ *plane. Characteristics are curved lines here.*

problem, which is approximated with a sequence of classical Riemann problems with initial data given by the space reconstruction of the exact solution in each computational cell.
For the purposes of the numerical method (1.14) it is not necessary to compute the full solution of a local generalized Riemann problem, which can be a very challenging task. For the method it is enough to determine the solution right at the interface of two neighboring cells. The discretization of numerical source is described later.

Cauchy-Kowalewski procedure

Now we introduce very briefly the Cauchy-Kowalewski procedure. It dates back to the Cauchy-Kowalewski theorem (see e.g., [10]), which states that there exists a unique analytic solution of the initial-value problem (1.12)-(1.13) provided that all involved functions are analytic. Since the technique was used by Lax and Wendroff in [38] too, it is also known as the *Lax-Wendroff procedure.* Using this procedure the time derivatives are expressed in terms of spatial derivatives. The idea and its application are illustrated in the following example.

Example 1.30
Consider the Cauchy problem

$$\begin{aligned} u_t + f(u)_x &= 0 \quad , \quad x \in (-\infty, \infty) \ , \ t > 0 \ , \\ u(x,0) &= u_0(x) \ . \end{aligned} \tag{1.20}$$

We assume f and u_0 to be analytic and so the exact solution u.
As stated above, we first compute the truncated Taylor expansion of $u(x,t)$ to the second order around $(0,0)$

$$\begin{aligned} u(x,t) \quad \approx \quad & u(0,0) + u_t(0,0)t + u_x(0,0)x \\ & + \frac{1}{2}u_{xx}(0,0)x^2 + u_{xt}(0,0)xt + \frac{1}{2}u_{tt}(0,0)t^2 \ . \end{aligned} \tag{1.21}$$

We see that the coefficients of the Taylor expansion are expressed in the terms of the function values $u(0,0)$ and the derivatives $\frac{\partial^k}{\partial x^l t^{k-l}} u(0,0)$. Now we apply the Cauchy-Kowalewski procedure and express the time and mixed derivatives in terms of spatial derivatives of the initial condition only. We use the initial condition and differentiate the equation (1.20), then the coefficients in (1.21) can be expressed in the terms of spatial derivatives of the initial condition:

$$\begin{aligned} u(0,0) &= u_0(0) \ , \\ u_t(0,0) &= -f'(u)u_x \ , \\ u_x(0,0) &= u_0'(0) \ , \\ u_{xx}(0,0) &= u_0''(0) \ , \\ u_{xt}(0,0) &= -\left[f'(u)u_{xx} + f''(u)(u_x)^2\right] \ , \\ u_{tt}(0,0) &= 2f'(u)f''(u)u_x^2 + (f'(u))^2 u_{xx} \ , \end{aligned}$$

where $u = u(0,0) = u_0(0)$, $u_x = u_x(0,0) = u_0'(0)$ and $u_{xx} = u_{xx}(0,0) = u_0''(0)$ on right hand sides. These terms can be finally plugged into the expansion (1.21) and we get an approximation on $u(x,t)$ in spatial derivatives of u_0 only.

Solution of the GRP

Consider again the GRP_K from definition 1.28. The initial states $\boldsymbol{u}_L(x)$ and $\boldsymbol{u}_R(x)$ are assumed to be smooth vector fields away from $x = x_{i+\frac{1}{2}}$. We assume that the requirements of the Cauchy-Kowalewski theorem are satisfied. Then one can use the Cauchy-Kowalewski procedure to construct a solution $\boldsymbol{u}(x,t)$ away from $x = x_{i+\frac{1}{2}}$. We do not require the full solution of the GRP, but only the solution along the t-axis, i.e., along $(x - x_{i+\frac{1}{2}})/t$, is constructed, as a function of time. On the left and right side of the interface we have $K+1$ non-trivial smooth data, with $K+1$ jumps at the interface for each component of the vector $\boldsymbol{u}$, defining the generalized Riemann problem.
To approximate the solution of GRP_K at the interface $x = x_{i+\frac{1}{2}}$, the function $\boldsymbol{u}_{LR}(\tau)$ is introduced, defined as the time power series expansion

$$\boldsymbol{u}_{LR}(\tau) = \boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + 0_+) + \sum_{k=1}^{K} \left[\partial_t^{(k)} \boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + 0_+)\right] \frac{\tau^k}{k!} \ , \tag{1.22}$$

where 0_+ stands for the limit $\lim_{t \to 0_+}$ of functions $\partial_t^{(k)} \boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + t)$, $k = 0, \dots, K$. The *leading term* $\boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + 0_+)$ and the higher-order terms $\partial_t^{(k)} \boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + 0_+)$ will be approximated by the solution of classical Riemann problems as follows.

The leading term accounts for the first-instant interaction of the initial data via the governing

PDE. It is acquired as the solution of the classical Riemann problem

$$
\begin{aligned}
\boldsymbol{u}_t + \boldsymbol{F}(\boldsymbol{u})_x &= \boldsymbol{0} \quad , \quad x \in (-\infty, \infty) \ , \ t > 0 \ , \\
\boldsymbol{u}(x,0) &= \begin{cases} \boldsymbol{u}_L^{(0)} := \lim_{y \to x_{i+\frac{1}{2}-}} \boldsymbol{u}_L(y) & , \quad x < x_{i+\frac{1}{2}} \ , \\ \boldsymbol{u}_R^{(0)} := \lim_{y \to x_{i+\frac{1}{2}+}} \boldsymbol{u}_R(y) & , \quad x > x_{i+\frac{1}{2}} \ . \end{cases}
\end{aligned}
\tag{1.23}
$$

The source term $\boldsymbol{S}(\boldsymbol{u})$ is neglected here. Its influence on the solution of GRP is involved in higher order terms. Denoting the similarity solution of (1.23) by $\boldsymbol{D}^{(0)}\left((x - x_{i+\frac{1}{2}})/t\right)$ (see theorem 1.16) the sought leading term is given by evaluating this solution along the t-axis

$$
\boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + 0_+) = \boldsymbol{D}^{(0)}(0) \ .
$$

This value is called the *Godunov state.*

For higher order terms the Cauchy-Kowalewski method is used to change from time to spatial derivatives, for which evolution equations are constructed and solved.
The time derivatives are expressed as a function of spatial derivatives of $\boldsymbol{u}$ via the Cauchy-Kowalewski procedure, more specifically

$$
\partial_t^{(k)} \boldsymbol{u}(x,t) = \boldsymbol{P}^{(k)} \left(\partial_x^{(0)} \boldsymbol{u}, \partial_x^{(1)} \boldsymbol{u}, \dots, \partial_x^{(k)} \boldsymbol{u} \right) \ .
$$

The source term $\boldsymbol{S}(\boldsymbol{u})$ is also included in the functions $\boldsymbol{P}^{(k)}$. One denotes the derivatives

$$
\boldsymbol{u}_L^{(k)}(y) := \frac{d^k}{dx^k} \boldsymbol{u}_L(y) \quad , \quad \boldsymbol{u}_R^{(k)}(y) := \frac{d^k}{dx^k} \boldsymbol{u}_R(y) \ ,
$$

evaluates $\boldsymbol{u}_L^{(k)}(y)$ and $\boldsymbol{u}_R^{(k)}(y)$ at $y = x_{i+\frac{1}{2}}$ and obtains

$$
\begin{aligned}
\boldsymbol{u}_L^{(k)} &:= \lim_{y \to x_{i+\frac{1}{2}-}} \boldsymbol{u}_L^{(k)}(y) \ , \\
\boldsymbol{u}_R^{(k)} &:= \lim_{y \to x_{i+\frac{1}{2}+}} \boldsymbol{u}_R^{(k)}(y)
\end{aligned}
$$

for $k = 1, \dots, K$.
We differentiate (1.12) k times w.r.t. x and acquire the system of nonlinear inhomogeneous evolution equations

$$
\partial_t(\partial_x^{(k)} \boldsymbol{u}(x,t)) + \boldsymbol{A}(\boldsymbol{u}) \cdot \partial_x(\partial_x^{(k)} \boldsymbol{u}(x,t)) = \boldsymbol{H}^{(k)} \ ,
$$

where the matrix $\boldsymbol{A}(\boldsymbol{u})$ is the Jacobian matrix of the system (1.12) and $\boldsymbol{H}^{(k)}$ is a function yielded by the differentiation and including also the source term $\boldsymbol{S}(\boldsymbol{u})$

$$
\boldsymbol{H}^{(k)} = \boldsymbol{H}^{(k)} \left(\partial_x^{(0)} \boldsymbol{u}, \partial_x^{(1)} \boldsymbol{u}, \dots, \partial_x^{(k)} \boldsymbol{u} \right) \ .
$$

To solve the *GRP*$_K$ efficiently, Toro and Titarev proposed the following simplifications: the source term $\boldsymbol{H}^{(k)}$ is neglected and the resulting homogeneous nonlinear equation is linearized around the Godunov state. The linearization is denoted by

$$
\boldsymbol{A}_{LR}^{(0)} := \boldsymbol{A}(\boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + 0_+))
$$

and one ends up with a sequence of homogeneous, linearized, classical Riemann problems for each $k = 1, \dots, K$

$$
\begin{aligned}
\partial_t(\partial_x^{(k)} \boldsymbol{u}(x,t)) + \boldsymbol{A}_{LR}^{(0)} \cdot \partial_x(\partial_x^{(k)} \boldsymbol{u}(x,t)) &= \boldsymbol{0} \quad , \quad x \in (-\infty, \infty) \ , \ t > 0 \ , \\
\partial_x^{(k)} \boldsymbol{u}(x,0) &= \begin{cases} \boldsymbol{u}_L^{(k)} & , \quad x < x_{i+\frac{1}{2}} \ , \\ \boldsymbol{u}_R^{(k)} & , \quad x > x_{i+\frac{1}{2}} \ . \end{cases}
\end{aligned}
\tag{1.24}
$$

We denote again the similarity solutions of (1.24) by $\boldsymbol{D}^{(k)}\left((x - x_{i+\frac{1}{2}})/t\right)$ yielding the higher order terms

$$\partial_x^{(k)}\boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + 0_+) = \boldsymbol{D}^{(k)}(0) \ .$$

We emphasize, that the Jacobian matrix $\boldsymbol{A}_{LR}^{(0)}$ needs to be evaluated only once and one has to solve K linear PDEs which ensures smaller numerical costs. Numerical experiments show that these simplifications are justified and one acquires a robust and accurate numerical method.

Altogether, plugging the terms $\partial_x^{(k)}\boldsymbol{u}(x_{i+\frac{1}{2}}, t^n + 0_+) = \boldsymbol{D}^{(k)}(0)$ for $k = 0, 1, \ldots, K$ into equation (1.22), one gets the approximation

$$\boldsymbol{u}_{LR}(\tau) = \boldsymbol{D}^{(0)}(0) + \sum_{k=1}^{K} \boldsymbol{D}^{(k)}(0)\frac{\tau^k}{k!} \ .$$

The numerical source (1.17) is treated in a similar way. One assumes, a high-order reconstruction of the exact solution $\boldsymbol{u}(x,t)$ is given for the initial time. Then, one possibility is to compute the volume integral via some numerical quadrature using the appropriate weights and quadrature points in the space domain. For each of these points, a time power series expansion is computed. Further, the time derivatives are substituted with functions of spatial derivatives and the spatial derivatives are evaluated on the initial data. This time-dependent function is then integrated over the time interval. Alternatively, one computes a space-time power serious analogous to the case (1.21) and replaces all time and mixed derivatives by spatial derivatives and integrates over time and space. Both possibilities yield a high-order representation of the numerical source.

For the complete solution, one needs to solve one non-linear Riemann problem to get the leading term and K linear Riemann problems for the higher-order terms. The leading term can be determined with a classical Riemann solver, exact or approximative. There is a variety on Riemann solvers, see e.g., [64], where the exact Riemann solver for the Euler equations for ideal and covolume gas is introduced or many approximative Riemann solvers, let us name the HLL, HLLC, the Riemann solver of Roe or Osher and other ones. For the linear Riemann problem, many well-known systems can be solved analytically, e.g., *linearized gas dynamics* (see [64]).

The reconstruction

The last task remaining to get a fully-discrete numerical scheme of higher order is to introduce a reconstruction procedure to get initial data for GRP_K and for the numerical source.
At any given time t^n one has the data $\{\boldsymbol{u}_i^n\}_i$ available, defining a piece-wise constant approximation of the exact solution. Based on these data, the spatial variation of the exact solution can be reconstructed, usually with polynomials (so one gets a piece-wise polynomial approximation) or as we will see in the chapter 3 with *polyharmonic splines* (where we will handle the more general case of input data $\boldsymbol{u}_i^n$). To this end, usually only the local data set is used to reconstruct the solution at a given cell - one defines a set of neighbors of a given cell i, this set is called *stencil*, denoted by $\mathcal{S}^i$. More specifically, for every two neighboring cells I_i and I_{i+1}, the corresponding reconstructions $\mathbf{R}_i$, based on data $\{\boldsymbol{u}_j^n\}_{j\in\mathcal{S}^i}$, and $\mathbf{R}_{i+1}$, based on data $\{\boldsymbol{u}_j^n\}_{j\in\mathcal{S}^{i+1}}$, are computed, such that

$$\begin{aligned} \mathbf{u}(x,t^n) &\approx \mathbf{R}_i(x) \quad , \quad x \in I_i, \\ \mathbf{u}(x,t^n) &\approx \mathbf{R}_{i+1}(x) \quad , \quad x \in I_{i+1} \end{aligned}$$

with the appropriate approximation order $K+1$. In order to keep the method conservative, the reconstructions have to satisfy

$$\begin{aligned} \frac{1}{\Delta x}\int_{I_i} \mathbf{R}_i &= \mathbf{u}_i^n \ , \\ \frac{1}{\Delta x}\int_{I_{i+1}} \mathbf{R}_{i+1} &= \mathbf{u}_{i+1}^n \ . \end{aligned}$$

If we now take $\mathbf{u}_L(x) := \mathbf{R}_i(x)$ and $\mathbf{u}_R(x) := \mathbf{R}_{i+1}(x)$, the corresponding classical Riemann problems (1.23) and (1.24) can be defined and solved for the initial data

$$\begin{aligned} \mathbf{u}_L^{(k)} &= \lim_{x \to x_{i+\frac{1}{2}-}} \mathbf{R}_i^{(k)}(x) \ , \\ \mathbf{u}_R^{(k)} &= \lim_{x \to x_{i+\frac{1}{2}+}} \mathbf{R}_{i+1}^{(k)}(x) \end{aligned}$$

for $k = 0, \ldots, K$.
The stencil can be chosen as a fixed stencil or a data-dependent stencil. The choice of a fixed stencil leads to a *linear scheme* (i.e., the coefficients of the scheme are constant). According to the Godunov's theorem (see [17]), such a scheme is oscillatory if it is monotone and of accuracy greater than one. That is why variable (adaptive) stencils should be used. Such reconstructions are called *non-linear reconstructions*. These reconstructions can be chosen e.g., on the basis of TVD criterion (see [64]) or using an ENO method (Essentially Non-Oscillatory), see [22]. The state-of-the-art is the WENO method (Weighted Essentially Non-Oscillatory), which will be discussed in detail in chapter 3.

1.4 B-splines

Let us briefly introduce B-spline functions in one dimension and some of their properties. B-splines are of our interest since they build a *partition of unity*, have compact support and are positive on it. Using these functions we can define a partition of unity needed in the finite volume particle method (see chapter 2), allowing us to design a higher order scheme (see chapter 4).
The introduction is based on the textbook by Schaback and Wendland [50], where further theory on B-splines can be found.

Definition and properties

Consider points $\ldots \le x_{-2} \le x_{-1} \le x_0 \le x_1 \le x_2 \le \ldots$, such that $\lim_{j \to \pm\infty} x_j = \pm\infty$.

Remark 1.31
The theory of B-splines allows multiple points x_j (i.e., it is allowed $x_j = x_{j+1} = \ldots = x_{j+k}$ for some $j, k \in \mathbb{Z}$). In the construction of a high order finite volume particle method we will exclude this case and will always require x_j to be distinct.

Definition 1.32
Let $X = \{x_j\}_{j \in \mathbb{Z}}$ and $m \in I\!N_0$. Define

$$\omega_j^m(x) = \begin{cases} \frac{x - x_j}{x_{j+m} - x_j} & , \quad x_j < x_{j+m} \ , \\ 0 & , \quad \textit{otherwise} \ . \end{cases}$$

Then the recursively defined functions

$$B_j^0(x) = \chi_{[x_j, x_{j+1})}(x) = \begin{cases} 1 & , \quad x \in [x_j, x_{j+1}) \ , \\ 0 & , \quad \textit{otherwise} \ , \end{cases}$$

and

$$B_j^m(x) = \omega_j^m(x) B_j^{m-1}(x) + \left(1 - \omega_{j+1}^m(x)\right) B_{j+1}^{m-1}(x) \ , \quad x \in \mathbb{R}$$

will be called B-splines of degree m *corresponding to point set X.*

Example 1.33
B-splines of degree 0 are characteristic functions B_j^0 of intervals $[x_j, x_{j+1})$.

Example 1.34
For $m = 1$ one gets the so-called *hat functions*, i.e., piecesewise linear functions with compact support. They will play an important role in chapter 4. More specifically, they are defined by

$$B_j^1(x) = \begin{cases} \frac{x-x_j}{x_{j+1}-x_j} & , \quad x \in [x_j, x_{j+1}] \ , \\ \frac{x_{j+2}-x}{x_{j+2}-x_{j+1}} & , \quad x \in [x_{j+1}, x_{j+2}] \ , \\ 0 & , \quad \text{otherwise} \ . \end{cases}$$

See also figure 1.3.

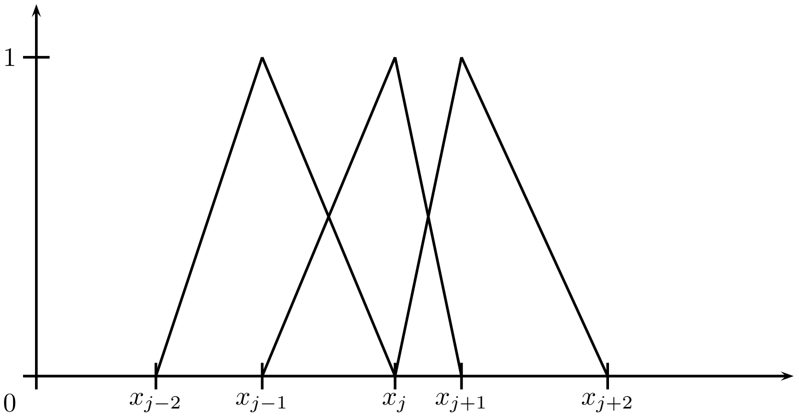


Figure 1.3: *Example of linear B-splines* B_j^1.

We will denote by the symbol $\mathcal{P}_{\widetilde{m}}^d$ the linear space of all d-variate polynomials of order at most $\widetilde{m}$ (i.e., of degree at most $\widetilde{m} - 1$). This notation for general d will be required in further chapters.

Proposition 1.35
The B-spline function B_j^m *consists of piecewise polynomial functions of degree at most* m*, more specifically*

$$B_j^m(x) = \sum_{k=j}^{m+j} b_k^m(x)\chi_{[x_k, x_{k+1})}(x) \ , \quad x \in \mathbb{R} \ ,$$

where $b_k^m \in \mathcal{P}_{m+1}^1$.
Further, B_j^m *is for* $x_j < x_{j+m+1}$ *positive on the interval* (x_j, x_{j+m+1}) *and equal to zero outside of the interval* $[x_j, x_{j+m+1}]$. *More precisely it holds* $B_j^m(x_{j+m+1}) = 0$ *for all* $m \in I\!N_0$ *and* $B_j^m(x_j) = 0$ *if* $x_j < x_{j+m}$.

From the following identity one can deduce two important corollaries, saying that every polynomial can be reproduced by B-splines and B-splines build a partition of unity on $\mathbb{R}$.

Theorem 1.36 (Marsden identity)
Define $\psi_{j,0} \equiv 1$ *and* $\psi_{j,m}(x) = \Pi_{i=1}^m (x_{j+i} - x)$, $m \in I\!N_0$, $j \in \mathbb{Z}$.

Then for every $\xi \in \mathbb{R}$ *it holds*

$$(x-\xi)^m = \sum_{j\in\mathbb{Z}} \psi_{j,m}(\xi) B_j^m(x) \ , \quad x \in \mathbb{R} \ .$$

Corollary 1.37
It holds $\mathcal{P}^1_{m+1} \subseteq span\{B_j^m \ : \ j \in \mathbb{Z}\}$, $m \in I\!N_0$. *Every* $p \in \mathcal{P}^1_{m+1}$ *satisfies for arbitrary* $\xi \in \mathbb{R}$ *the relation*

$$p(x) = \sum_{j\in\mathbb{Z}} \lambda_{j,m}(p) B_j^m(x) \ , \quad x \in \mathbb{R} \ ,$$

where

$$\lambda_{j,m}(p) = \sum_{l=0}^{m} (-1)^l \ \frac{p^{(m-l)}(\xi)}{m!} \psi^{(l)}_{j,m}(\xi) \ .$$

Corollary 1.38
B-splines build a partition of unity, *i.e.,*

$$\sum_{j\in\mathbb{Z}} B_j^m(x) = 1 \ , \quad x \in \mathbb{R} \ .$$

2 Finite-volume particle method

The finite volume particle method (FVPM) is a relatively new method proposed in 1998 in the paper of Hietel, Steiner and Struckmeier [24] for a multidimensional system of conservation laws. This method combines advantages of FVM and meshfree particle methods, based on the *Smoothed Particle Hydrodynamics* method (SPH) [45]. The resulting method is a highly flexible meshfree method utilizing the concept of a numerical flux. Junk and Struckmeier [29] proposed then in 2000 a more stable discretization and could also prove a Lax-Wendroff consistency of the scheme in the scalar case. It says, roughly speaking, that if a numerical solution by FVPM converges, then it converges to a weak solution of the governing PDE. On the other hand, it does not guarantee the convergence. For more details see [29] and for the background of the Lax-Wendroff theorem in the case of finite volume methods see [39], where the theorem was slighly reformulated (with essentially the same assumptions), which is especially useful for the class of TVD (Total Variation Diminishing) methods. The original theorem can be found in [38].

Since then, more analysis of the scheme has been done. The FVPM combines two main features - the numerical flux function from finite volume methods and the meshfree principle. A very motivational text, where both, finite volume and finite volume particle method, are compared, was written by Junk [30]. Different approaches for the correction procedure for geometrical coefficients were proposed in [32],[58] and [59]. The particle motion was studied in [47] or [51] and also [59]. Some theoretical analysis, combining the method with B-splines and interesting physical examples can be found in [31]. Boundary treatment is discussed in [59]. The topic of adaptivity was studied in [51] or [37]. Also other authors deal with this topic, such as e.g., [19], [20], [23], [33], [48] or [60]. In [47], an approach using higher order space discretization and a predictor-corrector method for the time discretization leads to a scheme of second order of accuracy.

The derivation of the FVPM is presented and its properties are discussed. We contribute to the development of the method with a precise formulation of the correction procedure for geometrical coefficients and also prove its correctness. Finally, we develop a new procedure to add and to remove a particle without loss of conservativity of mass and constant states up to the machine precision.

2.1 The derivation of finite volume particle method

Introduction

In this section, we follow [60] in the derivation of the method, which works with a bounded domain, opposed to the original proposal done in [29], where the whole $\mathbb{R}^d$ was considered. The case of moving boundary is described later in this thesis.

Let us consider a system of conservation laws in an open and bounded domain $\Omega \subset \mathbb{R}^d$

$$\mathbf{u}_t + \nabla \cdot \boldsymbol{F}(\mathbf{u}) = \mathbf{0} \qquad \text{in } \Omega \subset \mathbb{R}^d \ , \ t > 0 \tag{2.1}$$

with initial conditions

$$\mathbf{u}(\ .\ ,0) = \mathbf{u}_0 \qquad \text{in } \Omega \subset \mathbb{R}^d \tag{2.2}$$

and with boundary conditions given by

$$B(\mathbf{u}) = 0 \qquad \text{in } \partial\Omega \times (0, \infty)$$

with a given operator $B : \mathcal{U} \to \mathbb{R}$.
Let $n_p \in I\!N$. The functions

$$\begin{array}{rll} \mathbf{x}_i : & [0,\infty) & \to \overline{\Omega}\ ,\ i = 1, \ldots, n_p \\ & t & \mapsto \mathbf{x}_i(t) \end{array} \tag{2.3}$$

are called *particles*. These functions describe the motion of particle points located in the computational domain Ω.

Definition 2.1
Let $W : \mathbb{R} \to \mathbb{R}^+$ denote a compactly supported, non-negative and Lipschitz-continuous function with a non-empty support. We define functions

$$W_i(\mathbf{x}, t) := m_i W(\mathbf{x} - \mathbf{x}_i(t))\ ,\ i = 1, \ldots, n_p \tag{2.4}$$

as a mass packet*, where $m_i > 0$ denotes the* mass of the particle*.*
Further, we define the mass density

$$\sigma(\mathbf{x}, t) := \sum_{i=1}^{n_p} W_i(\mathbf{x}, t)$$

and test functions

$$\psi_i(\mathbf{x}, t) := \frac{W_i(\mathbf{x}, t)}{\sigma(\mathbf{x}, t)} \geq 0\ . \tag{2.5}$$

Usually m_i are set to 1. The test functions are compactly supported and positive on their support. We assume, that

$$\underset{t \in [0,\infty)}{\forall}\ \underset{\mathbf{x} \in \overline{\Omega}}{\forall}\ \underset{i \in \{1,\ldots,n_p\}}{\exists}\ W_i(\mathbf{x}, t) \neq 0\ . \tag{2.6}$$

The last assumption stands for the condition, that the supports of functions ψ_i have to cover the complete computational domain. From the properties of functions ψ_i also follows, that the supports of functions ψ_i overlap.

Observation 2.2

$$\begin{aligned} \sum_{i=1}^{n_p} \psi_i(\mathbf{x}, t) &= 1\ , \\ \sum_{i=1}^{n_p} \nabla_{\mathbf{x}} \psi_i(\mathbf{x}, t) &= 0\ . \end{aligned}$$

We define *volumes*

$$V_i(t) := \int_{\Omega} \psi_i(\mathbf{x}, t) d\mathbf{x}\ ,\ i = 1, \ldots, n_p\ . \tag{2.7}$$

Example 2.3

$$W(x) = \begin{cases} \frac{x+h}{h} & , \quad x \in [-h, 0) \ , \\ \frac{-x+h}{h} & , \quad x \in [0, h) \ , \\ 0 & , \quad \text{otherwise} \ , \end{cases} \tag{2.8}$$

where the parameter $h > 0$ is called the *smoothing length.* Examples on different choice of h are shown in figure 2.1.

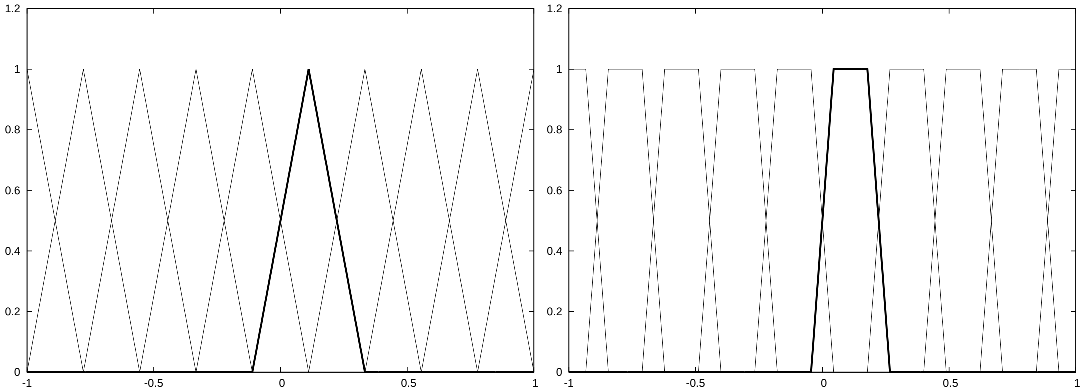


Figure 2.1: *Examples of functions ψ_i generated by function W in (2.8). The shape of all functions is highlighted for one chosen function. On the left functions ψ_i for $h = \Delta x := x_{i+1} - x_i$, on the right for $h = 0.7\Delta x$.*

Example 2.4
Another example on W is

$$W(x) = \begin{cases} (x+h)^2 & , \quad x \in [-h, -\frac{h}{2}) \ , \\ -x^2 + \frac{h^2}{2} & , \quad x \in [-\frac{h}{2}, \frac{h}{2}) \ , \\ (x-h)^2 & , \quad x \in [\frac{h}{2}, h) \ , \\ 0 & , \quad \text{otherwise} \ . \end{cases} \tag{2.9}$$

Examples with this W are shown in figure 2.2.

Derivation

Multiplying the equation (2.1) with the function ψ_i and integrating it over Ω leads to

$$\int\limits_{\Omega} (\mathbf{u}_t + \nabla \cdot \boldsymbol{F}(\mathbf{u}))\, \psi_i \; d\mathbf{x} = 0 \qquad \forall \; i = 1, \ldots, n_p \ ,$$

which yields after using the divergence theorem and interchanging the integral and derivation sign

$$\frac{d}{dt} \int\limits_{\Omega} \psi_i \mathbf{u} \; d\mathbf{x} = \int\limits_{\Omega} (\psi_i)_t \; \mathbf{u} \; d\mathbf{x} + \int\limits_{\Omega} \boldsymbol{F}(\mathbf{u}) \cdot \nabla \psi_i \; d\mathbf{x} - \int\limits_{\partial\Omega} \psi_i \boldsymbol{F}(\mathbf{u}) \cdot \mathbf{n} \; d\sigma \ . \tag{2.10}$$

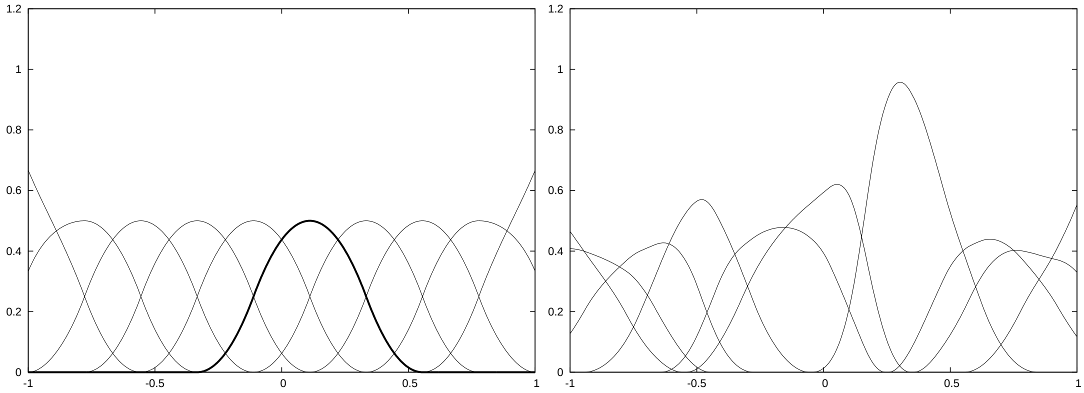


Figure 2.2: *Examples of functions ψ_i generated by function W in (2.9). On the left, the shape of all functions is highlighted for one chosen function, $h = 2\Delta x$. On the right one can see the functions ψ_i for irregular random distribution of particles.*

Define functions

$$\mathbf{\Gamma}_{ij} := \psi_i \frac{\nabla W_j}{\sigma} \ , \quad i, j = 1, \ldots, n_p .$$

Proposition 2.5
For $i = 1, \ldots, n_p$ it holds

$$(\psi_i)_t \quad = \quad \sum_{j=1}^{n_p} \left(\dot{\mathbf{x}}_j \mathbf{\Gamma}_{ij} - \dot{\mathbf{x}}_i \mathbf{\Gamma}_{ji} \right) \ , \tag{2.11}$$

$$\nabla \psi_i \quad = \quad \sum_{j=1}^{n_p} \left(\mathbf{\Gamma}_{ji} - \mathbf{\Gamma}_{ij} \right) \ , \tag{2.12}$$

where $\dot{\mathbf{x}}_i = \dot{\mathbf{x}}_i(t) = \frac{d}{dt}\mathbf{x}_i(t)$ denotes the derivative of $\mathbf{x}_i$ with respect to the time variable t.

Proof.
See [24]. ☐

Remark 2.6
In order to keep the notation simple, we will denote by $\dot{\mathbf{x}}_j \mathbf{\Gamma}_{ij}$ the standard scalar product $\dot{\mathbf{x}}_j^T \mathbf{\Gamma}_{ij}$ of vectors $\dot{\mathbf{x}}_j$ and $\mathbf{\Gamma}_{ij}$.

Denoting the boundary term

$$\mathcal{B}_i := \int_{\partial\Omega} \psi_i \boldsymbol{F}(\mathbf{u}) \cdot \mathbf{n} \; d\sigma$$

and using the proposition 2.5 substitute the corresponding terms $(\psi_i)_t$ and $\nabla \psi_i$ in the equation

(2.10)

$$
\begin{aligned}
\frac{d}{dt}\int\limits_\Omega \psi_i \mathbf{u}\; d\mathbf{x} \;&=\; \sum_{j=1}^{n_p} \int\limits_\Omega \boldsymbol{F}(\mathbf{u}) \cdot (\boldsymbol{\Gamma}_{ji} - \boldsymbol{\Gamma}_{ij}) \;\; d\mathbf{x} \\
&\quad + \sum_{j=1}^{n_p} \left[\int\limits_\Omega \mathbf{u}\, (\dot{\mathbf{x}}_j \boldsymbol{\Gamma}_{ij} - \dot{\mathbf{x}}_i \boldsymbol{\Gamma}_{ji}) \;\; d\mathbf{x} \right] - \mathcal{B}_i \\
&=\; \sum_{j=1}^{n_p} \int\limits_\Omega (\boldsymbol{F}(\mathbf{u}) - \mathbf{u} \otimes \dot{\mathbf{x}}_i) \cdot \boldsymbol{\Gamma}_{ji} \\
&\quad - \sum_{j=1}^{n_p} \int\limits_\Omega (\boldsymbol{F}(\mathbf{u}) - \mathbf{u} \otimes \dot{\mathbf{x}}_j) \cdot \boldsymbol{\Gamma}_{ij} \;\; - \mathcal{B}_i \;.
\end{aligned}
\tag{2.13}
$$

Let us assume, that $\mathbf{u}$ varies only slightly around a constant value $\bar{\mathbf{u}}$ on the intersection of the supports of functions ψ_i and ψ_j as well as $\dot{\mathbf{x}}_i \approx \dot{\mathbf{x}}_j := \bar{\dot{\mathbf{x}}}$, e.g., $\bar{\dot{\mathbf{x}}} = \frac{\dot{\mathbf{x}}_i + \dot{\mathbf{x}}_j}{2}$. Then

$$
\frac{d}{dt}\int\limits_\Omega \psi_i \mathbf{u}\; d\mathbf{x} \;\approx\; -\sum_{j=1}^{n_p} (\boldsymbol{F}(\bar{\mathbf{u}}) - \bar{\mathbf{u}} \otimes \bar{\dot{\mathbf{x}}}) \cdot \int\limits_\Omega (\boldsymbol{\Gamma}_{ij} - \boldsymbol{\Gamma}_{ji})\, d\mathbf{x} \;\; - \mathcal{B}_i \;.
$$

We denote

$$
\boldsymbol{\gamma}_{ij}(t) := \int\limits_\Omega \boldsymbol{\Gamma}_{ij}(\mathbf{x}, t)\;\; d\mathbf{x}\;, \quad i, j = 1, \ldots, n_p
$$

and acquire

$$
\frac{d}{dt}\int\limits_\Omega \psi_i \mathbf{u}\; d\mathbf{x} \;\approx\; -\sum_{j=1}^{n_p} (\boldsymbol{F}(\bar{\mathbf{u}}) - \bar{\mathbf{u}} \otimes \bar{\dot{\mathbf{x}}}) \cdot (\boldsymbol{\gamma}_{ij} - \boldsymbol{\gamma}_{ji}) - \mathcal{B}_i \;. \tag{2.14}
$$

Remark 2.7
Note, that the functions $\boldsymbol{\Gamma}_{ij}$ as functions of $\mathbf{x}$ are located on the intersection of supports of functions ψ_i and ψ_j. Then $\boldsymbol{\gamma}_{ij} = \mathbf{0}$ for non-overlapping particles and for overlapping particles we only have to compute the integral over the intersection $\psi_i \cap \psi_j$ and not over the whole domain Ω. The computation of $\boldsymbol{\Gamma}_{ij}$ is therefore local.

We define new quantities, the coefficients

$$
\mathbf{u}_i(t) := \frac{1}{V_i(t)} \int\limits_\Omega \mathbf{u}(\mathbf{x}, t) \psi_i(\mathbf{x}, t)\;\; d\mathbf{x}\;, \tag{2.15}
$$

which denote the *weighted integral mean* of the function $\mathbf{u}$ with respect to the weight function ψ_i, and

$$
\boldsymbol{\beta}_{ij} := \boldsymbol{\gamma}_{ij} - \boldsymbol{\gamma}_{ji} \quad , \quad \mathbf{n}_{ij} := \frac{\boldsymbol{\beta}_{ij}}{|\boldsymbol{\beta}_{ij}|}\;, \tag{2.16}
$$

where the vector $\mathbf{n}_{ij}$ plays a similar role as the outer unit normal vector between two neighboring cells in FVM. Using this notation, one gets

$$
\frac{d}{dt}(\mathbf{u}_i V_i) \;\approx\; -\sum_{j=1}^{n_p} |\boldsymbol{\beta}_{ij}| \mathbf{g}_{ij} - \mathcal{B}_i\;, \tag{2.17}
$$

where we approximate

$$
(\boldsymbol{F}(\bar{\mathbf{u}}) - \bar{\mathbf{u}} \otimes \bar{\dot{\mathbf{x}}}) \cdot \mathbf{n}_{ij} \approx \mathbf{g}_{ij} = \mathbf{g}(t, \mathbf{x}_i, \mathbf{x}_j, \mathbf{u}_i, \mathbf{u}_j, \mathbf{n}_{ij}) \tag{2.18}
$$

with a function $\mathbf{g}_{ij}$ that is constructed on the basis of a numerical flux function from the framework of FVM. Neglecting the error we arrive at a system of ODEs

$$\frac{d}{dt}(\mathbf{u}_i V_i) \quad = \quad -\sum_{j=1}^{n_p} |\boldsymbol{\beta}_{ij}| \mathbf{g}_{ij} - \mathcal{B}_i \qquad \forall\ i = 1, \ldots, n_p \ . \tag{2.19}$$

We want to choose arbitrarily the quantities $\dot{\mathbf{x}}_i$, $i = 1, \ldots, n_p$, whereas $\boldsymbol{\gamma}_{ij}$, $\boldsymbol{\beta}_{ij}$ can be computed and the discretization of $\mathcal{B}_i$ will be specified later. The only unknowns are therefore $\mathbf{u}_i$ and V_i, for which we have only one equation. To get the second one, one can compute the integral (2.7) using a numerical quadrature or differentiate the equation (2.7) w.r.t. the time t:

$$\frac{d}{dt} V_i(t) = \sum_{i=1}^{n_p} \left(\dot{\mathbf{x}}_j \boldsymbol{\gamma}_{ij} - \dot{\mathbf{x}}_i \boldsymbol{\gamma}_{ji} \right) \ .$$

Finally, one gets the following closed system of ODEs

$$\dot{\mathbf{x}}_i(t) \quad = \quad \mathbf{a}(\mathbf{x}_i, t, \mathbf{u}_i) \ , \tag{2.20}$$

$$\dot{V}_i(t) \quad = \quad \sum_{j=1}^{n_p} \left(\dot{\mathbf{x}}_j \boldsymbol{\gamma}_{ij} - \dot{\mathbf{x}}_i \boldsymbol{\gamma}_{ji} \right) \ , \tag{2.21}$$

$$\frac{d}{dt}(\mathbf{u}_i V_i) \quad = \quad -\sum_{j=1}^{n_p} |\boldsymbol{\beta}_{ij}| \mathbf{g}_{ij} - \mathcal{B}_i \tag{2.22}$$

for all $i = 1, \ldots, n_p$. The vector $\mathbf{a}(\mathbf{x}, t, \mathbf{u})$ is an arbitrarily chosen velocity field. The initial condition is given by the initial values $\mathbf{x}_i(0)$ and

$$\begin{aligned} V_i(0) &= \int_\Omega \psi_i(\mathbf{x}, 0) d\mathbf{x} \ , \\ \mathbf{u}_i(0) &= \frac{1}{V_i(0)} \int_\Omega \mathbf{u}_0(\mathbf{x}) \psi_i(\mathbf{x}, 0) d\mathbf{x} \ . \end{aligned}$$

The semi-discrete scheme (2.20) - (2.22) can be discretized in the time variable e.g., with the explicit Euler method yielding

$$\begin{aligned} \dot{\mathbf{x}}_i^n &= \mathbf{a}(\mathbf{x}_i^n, t^n, \mathbf{u}_i^n) \ , \\ V_i^{n+1} &= V_i^n + \Delta t \sum_{j=1}^{n_p} \left(\dot{\mathbf{x}}_j^n \boldsymbol{\gamma}_{ij}^n - \dot{\mathbf{x}}_i^n \boldsymbol{\gamma}_{ji}^n \right) \ , \\ \mathbf{u}_i^{n+1} V_i^{n+1} &= \mathbf{u}_i^n V_i^n - \Delta t \sum_{j=1}^{n_p} |\boldsymbol{\beta}_{ij}^n| \mathbf{g}_{ij}^n - \mathcal{B}_i \end{aligned}$$

for all $i = 1, \ldots, n_p$. In the above scheme, we use the upper index n to denote the specific quantity at a fixed time t^n. The initial conditions are

$$\begin{aligned} \mathbf{x}_i^0 &= \mathbf{x}_i(0) \ , \\ \mathbf{u}_i^0 &= \frac{1}{V_i^0} \int_\Omega \mathbf{u}_0(\mathbf{x}) \psi_i(\mathbf{x}, 0) d\mathbf{x} \ , \\ V_i^0 &= \int_\Omega \psi_i(\mathbf{x}, 0) d\mathbf{x} \ . \end{aligned}$$

We remark that also another discretization of the scheme in the time variable is possible, e.g., using some multi-step method or a predictor-corrector procedure, as in [47]. Another possibility to discretize is via integrating the equation (2.22) over a time interval as we will do in the chapter 4 to get a higher order scheme.

If we discretize the time interval $[0,T]$, $T>0$ as $0=t^0<t^1<t^2<\ldots<t^{N_T}=T$, then the numerical solution of the problem (2.1)-(2.2) via FVPM is defined by

$$\mathbf{u}_h(\mathbf{x},t)=\sum_{n=0}^{N_T}\sum_{i=1}^{n_p}\mathbf{u}_i^n\psi_i(\mathbf{x},t)\chi_{[t^n,t^{n+1})}(t)\ ,\ \mathbf{x}\in\Omega\ ,\ t\in[0,T]\ . \tag{2.23}$$

Remark 2.8
In each time step, the numerical solution via FVPM is defined as a linear combination of basis functions $\{\psi_i\}_{i=1}^{n_p}$ *with coefficients* $\mathbf{u}_i^n$ *defined in (2.15). These coefficients approximate the weighted integral means of the exact solution*

$$\frac{1}{V_i^n}\int\limits_{\Omega}\mathbf{u}(\mathbf{x},t^n)\psi_i(\mathbf{x},t^n)d\mathbf{x}\approx\mathbf{u}_i^n\ .$$

As we will see later, since the functions $\{\psi_i\}_{i=1}^{n_p}$ *build a partition of unity and under further assumptions on the numerical flux, this numerical scheme is conservative and reproduces constant functions.*

As well as in the FVM, the *CFL-condition* is an important part of the method that provides the stability. For particular choices of CFL-condition we refer to [31] and [59].

2.2 Properties of FVPM

Geometrical coefficients

The coefficients $\boldsymbol{\beta}_{ij}$ from (2.16) are called *geometrical coefficients.* Since they are of particular importance in the FVPM scheme, we will look at them in detail. Their geometrical interpretation in comparison to FVM can be found in [30]. The following proposition lists properties of the geometrical coefficients. Some of them are important for a correct numerical behavior of the method, as we will see later.

Proposition 2.9
The coefficients $\boldsymbol{\beta}_{ij}$ *satisfy the following relations*

$$\boldsymbol{\beta}_{ij} = -\boldsymbol{\beta}_{ji}\ ,\qquad \forall\ i,j=1,\ldots,n_p\ , \tag{2.24}$$

$$\boldsymbol{\beta}_{ij} = \mathbf{0}\ ,\qquad \text{if supp } \psi_i\cap\psi_j=\emptyset\ , \tag{2.25}$$

$$\boldsymbol{\beta}_{ii} = \mathbf{0}\ ,\qquad \forall\ i=1,\ldots,n_p\ , \tag{2.26}$$

(2.27)

$$\sum_{j=1}^{n_p}\boldsymbol{\beta}_{ij} = \begin{cases}\mathbf{0}\ , & \text{if supp } \psi_i\cap\partial\Omega=\emptyset\ ,\\ -\int\limits_{\partial\Omega}\psi_i\boldsymbol{n}d\boldsymbol{\sigma}\ , & \text{if supp } \psi_i\cap\partial\Omega\neq\emptyset\ ,\end{cases} \tag{2.28}$$

(2.29)

$$|\boldsymbol{\beta}_{ij}| = \mathcal{O}(h^{d-1})\ ,\ h\to 0\ , \tag{2.30}$$

$$\boldsymbol{\beta}_{ij} = 2\int\limits_{\Omega}\psi_i\nabla\psi_j d\mathbf{x}-\int\limits_{\partial\Omega}\psi_i\psi_j\mathbf{n}d\sigma\ . \tag{2.31}$$

Proof.
See [59]. ☐

Remark 2.10
The quantity h should be understood as the smoothing length, see (2.8) and (2.9) (for more details see [59] or [60]). In one spatial dimension, it is usually half the length of the support of the functions ψ_i*. For example, in the case of uniform distribution of particle with distance* Δx*, h is chosen*

as some positive multiple of Δx*. In more dimensions, depending on the definition of functions* ψ_i*, it is, e.g., half of the diameter of the support of* ψ_i*. It has a similar meaning as the mesh size* h *in the case of mesh-based methods.*

Proposition 2.11 ([59])
If $\Omega = \mathbb{R}$*, then for every* $\bar{x} \in \mathbb{R}$*, we have*

$$\sum_{x_i(t)\geq\bar{x}} \sum_{x_j(t)\geq\bar{x}} \beta_{ij}(t) = 1 \qquad \forall t \geq 0 \ .$$

Correction procedure

The properties (2.24)-(2.28) ensure the proper behavior of the numerical solution via FVPM. If we can compute the coefficients $\boldsymbol{\beta_{ij}}$ exactly, e.g., using B-splines in one dimension as the partition of unity, the scheme will work properly. Otherwise, we have to use some numerical quadrature to compute the geometrical coefficients. If we use the definition (2.16) rather than the formula (2.31), the equation (2.24) will be satisfied as well as (2.25) and (2.26). However, the relation (2.28) does not have to hold due to the discretization errors of the numerical quadrature applied to compute the geometrical coefficients. Since the usage of a high order accurate numerical quadrature causes high computational costs, other approaches were developed - see [32] and [58], where the geometrical coefficients have been corrected, and [59], where the FVPM scheme has been modified. We will follow the approach proposed in [32] and formulate the algorithm also for the case of bounded domain Ω and prove its correctness. We remark that this approach can be found also in [70].
The principle of the algorithm is the following: for each particle $i \in \{1, \dots, n_p\}$ we compute the error of the sum $\sum_{j=1}^{n_p} \boldsymbol{\beta}_{ij}$ and shift it to the next particle $i+1$. So all the errors are shifted to the last particle. From the lemma 2.12 and theorem 2.15 it then follows, that the error does not accumulate there, see also remark 2.14.
More specifically, we compute the geometrical coefficients out of the definition (2.16) using a numerical quadrature as a first guess. These coefficients will be denoted by $\widetilde{\boldsymbol{\beta}}_{ij}$. As already mentioned, the conditions (2.24)-(2.26) are satisfied for $\widetilde{\boldsymbol{\beta}}_{ij}$.
We denote by

$$\boldsymbol{\Theta}_i := \begin{cases} \boldsymbol{0} & , \text{ if supp } \psi_i \cap \partial\Omega = \emptyset \\ -\int\limits_{\partial\Omega} \psi_i \boldsymbol{n} d\boldsymbol{\sigma} & , \text{ if supp } \psi_i \cap \partial\Omega \neq \emptyset \end{cases} \tag{2.32}$$

the desired value of $\sum\limits_{j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{ij}$, which coincides with the condition (2.28). But due to the numerical integration we have

$$\sum_{j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{ij} = \boldsymbol{\Theta}_i + \mathbf{E}_i \ , \tag{2.33}$$

where $\mathbf{E}_i$ denotes the error of the sum $\sum_{j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{ij}$. The following proposition shows, that these errors do not accumulate.

Lemma 2.12
The sum of errors defined in (2.33) satisfies

$$\sum_{i=1}^{n_p} \boldsymbol{E}_i = \boldsymbol{0} \ . \tag{2.34}$$

Proof.
According to the definition (2.33) it holds

$$\sum_{i=1}^{n_p} \mathbf{E}_i = \sum_{i,j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{ij} - \sum_{i=1}^{n_p} \boldsymbol{\Theta}_i \ .$$

Further, since (2.24) and (2.26) hold, we have

$$\sum_{i,j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{ij} = \sum_{i=1}^{n_p} \sum_{\substack{j=1 \\ j\neq i}}^{n_p} \widetilde{\boldsymbol{\beta}}_{ij} + \sum_{i=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{ii} = \mathbf{0} \ .$$

For the second term we can conclude

$$\sum_{i=1}^{n_p} \boldsymbol{\Theta}_i = \sum_{i=1}^{n_p} \left(- \int\limits_{\partial\Omega} \psi_i \boldsymbol{n} d\boldsymbol{\sigma} \right) = - \int\limits_{\partial\Omega} \sum_{i=1}^{n_p} \psi_i \boldsymbol{n} d\boldsymbol{\sigma} = - \int\limits_{\partial\Omega} \boldsymbol{n} d\boldsymbol{\sigma} = \mathbf{0} \ ,$$

which proves the statement. □

The correction procedure, in which the terms $\widetilde{\boldsymbol{\beta}}_{ij}$ (computed via numerical quadrature) are corrected to $\boldsymbol{\beta}_{ij}$, follows.

Algorithm 2.13

Input: $\widetilde{\boldsymbol{\beta}}_{ij}$, $i, j = 1, \ldots, n_p$. *Output:* $\boldsymbol{\beta}_{ij}$, $i, j = 1, \ldots, n_p$.

```
1. For $i = 1, \ldots, n_p - 1$:
    a) compute $\boldsymbol{\Theta}_i$ according to (2.32) ,
    b) compute $\mathbf{E}_i = \sum_{j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{ij} - \boldsymbol{\Theta}_i$ .
2. For $i = 1, \ldots, n_p - 1$:
    Define
    $\boldsymbol{\beta}_{i,i+1} := \widetilde{\boldsymbol{\beta}}_{i,i+1} - \sum_{k=1}^{i} \mathbf{E}_k$ ,
    $\boldsymbol{\beta}_{ij} := \widetilde{\boldsymbol{\beta}}_{ij}$ , $j = 1, \ldots, n_p;\ j \neq i \pm 1$ ,
    $\boldsymbol{\beta}_{i+1,i} := \widetilde{\boldsymbol{\beta}}_{i+1,i} + \sum_{k=1}^{i} \mathbf{E}_k$ .
3. For $j = 1, \ldots, n_p;\ j \neq n_p - 1$:
    Define $\boldsymbol{\beta}_{n_p,j} := \widetilde{\boldsymbol{\beta}}_{n_p,j}$ .
```

Remark 2.14
In order to express the algorithm in words, consider $(\widetilde{\boldsymbol{\beta}}_{ij})_{i,j}$ *to be a matrix. Then for every* $i = 1, \ldots, n_p - 1$ *we sum the* $\widetilde{\boldsymbol{\beta}}_{ij}$ *and after subtracting the boundary value* $\boldsymbol{\Theta}_i$ *we get the "row" error* $\boldsymbol{E}_i$*. This error is added with a minus sign to the "upper diagonal" term* $\widetilde{\boldsymbol{\beta}}_{i,i+1}$*, so that*

the equation $\sum_{j=1}^{n_p} \boldsymbol{\beta}_{ij} = \mathbf{0}$ *holds. In order to preserve the skew-symmetry condition (2.24), it is necessary to modify also the "lower diagonal" term* $\widetilde{\boldsymbol{\beta}}_{i+1,i}$ *by adding the error (with a plus sign). Then one moves from the "row"* i *to the next "row"* $i+1$ *and repeats the procedure. The errors* $\boldsymbol{E}_i$ *are hereby shifted to the last particle (while the condition (2.28) being fixed), where, due to lemma 2.12, they do not accumulate.*

Theorem 2.15
Under the assumption

$$\mathit{supp}\ \psi_i \cap \mathit{supp}\ \psi_{i+1} \neq \emptyset \qquad \forall\ i = 1, \ldots, n_p - 1\ , \tag{2.35}$$

the coefficients $\boldsymbol{\beta}_{ij}$*,* $i, j = 1, \ldots, n_p$*, defined in the algorithm 2.13, satisfy the conditions (2.24) - (2.28).*

Proof.
Consider $(\boldsymbol{\beta}_{ij})_{ij}$ as a matrix. Then, the only modified coefficient $\widetilde{\boldsymbol{\beta}}_{ij}$ are obviously the upper diagonal ($\widetilde{\boldsymbol{\beta}}_{i,i+1}$, $i = 1, \ldots, n_p - 1$) and the lower diagonal terms ($\widetilde{\boldsymbol{\beta}}_{i+1,i}$, $i = 1, \ldots, n_p - 1$). For the remaining $\widetilde{\boldsymbol{\beta}}_{ij}$ the conditions (2.24) - (2.26) hold.
Because of the assumption (2.35), the modification of the terms $\widetilde{\boldsymbol{\beta}}_{i,i+1}$ and $\widetilde{\boldsymbol{\beta}}_{i+1,i}$ does not violate the condition (2.25).
Due to the 2. step of the algorithm we have for $i = 1, \ldots, n_p - 1$

$$\boldsymbol{\beta}_{i,i+1} \quad = \quad \widetilde{\boldsymbol{\beta}}_{i,i+1} - \sum_{k=1}^{i} \mathbf{E}_k$$

and

$$\boldsymbol{\beta}_{i+1,i} \quad = \quad \widetilde{\boldsymbol{\beta}}_{i+1,i} + \sum_{k=1}^{i} \mathbf{E}_k \overset{(2.24)}{=} -\widetilde{\boldsymbol{\beta}}_{i,i+1} + \sum_{k=1}^{i} \mathbf{E}_k = -\boldsymbol{\beta}_{i,i+1}\ ,$$

which is the condition (2.24) for $\boldsymbol{\beta}_{i,j}$.

Finally, the condition (2.28) holds for $\boldsymbol{\beta}_{i,j}$, since

$$\begin{aligned}
\mathbf{i = 1:} \qquad & \sum_{j=1}^{n_p} \boldsymbol{\beta}_{i,j} &=& \sum_{\substack{j=1\\ j\neq i+1}}^{n_p} \widetilde{\boldsymbol{\beta}}_{i,j} + \left(\widetilde{\boldsymbol{\beta}}_{i,i+1} - \mathbf{E}_i\right) = \sum_{j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{i,j} - \mathbf{E}_i = \boldsymbol{\Theta}_i\ , \\
\mathbf{1 < i < n_p:} \qquad & \sum_{j=1}^{n_p} \boldsymbol{\beta}_{i,j} &=& \sum_{\substack{j=1\\ j\neq i\pm 1}}^{n_p} \widetilde{\boldsymbol{\beta}}_{i,j} + \left(\widetilde{\boldsymbol{\beta}}_{i,i-1} + \sum_{k=1}^{i-1} \mathbf{E}_k\right) + \left(\widetilde{\boldsymbol{\beta}}_{i,i+1} - \sum_{k=1}^{i} \mathbf{E}_k\right) \\
& &=& \sum_{j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{i,j} - \mathbf{E}_i = \boldsymbol{\Theta}_i\ , \\
\mathbf{i = n_p:} \qquad & \sum_{j=1}^{n_p} \boldsymbol{\beta}_{i,j} &=& \sum_{\substack{j=1\\ j\neq i-1}}^{n_p} \widetilde{\boldsymbol{\beta}}_{i,j} + \left(\widetilde{\boldsymbol{\beta}}_{i,i-1} + \sum_{k=1}^{i-1} \mathbf{E}_k\right) = \sum_{j=1}^{n_p} \widetilde{\boldsymbol{\beta}}_{i,j} + \sum_{k=1}^{i-1} \mathbf{E}_k \\
& &=& \boldsymbol{\Theta}_i + \mathbf{E}_i + \sum_{k=1}^{i-1} \mathbf{E}_k = \boldsymbol{\Theta}_i + \sum_{k=1}^{i} \mathbf{E}_k = \boldsymbol{\Theta}_i\ .
\end{aligned}$$

□

Remark 2.16
The condition (2.35) ensures (2.25) to hold and can be achieved in the one dimensional case with a simple ordering of the particles according to their position.

General properties

Proposition 2.17 ([59])
It holds

$$V_i \quad = \quad \mathcal{O}(h^d)\ , h \to 0\ ,$$

where d *is the spatial dimension.*

We extend the result from [59] to particles moving with constant velocity:

Lemma 2.18 (Preserving constant states)
Consider the semi-discrete scheme (2.20)-(2.22). Assume that the velocity field is constant, i.e.,

$$\dot{\mathbf{x}}_i(t) \quad = \quad \mathbf{a}_0 \quad \forall\ t \in [0, \infty)$$

for $\mathbf{a}_0 \in \mathbb{R}^d$*. If* $\mathbf{g}_{ij}$ *is a consistent numerical flux in the case of constant velocity field, i.e., if it holds for every* $\mathbf{c} \in \mathbb{R}^m$

$$(\boldsymbol{F}(\mathbf{c}) - \mathbf{c} \otimes \mathbf{a}_0) \cdot \mathbf{n}_{ij} = \mathbf{g}_{ij} = \mathbf{g}(t, \mathbf{x}_i, \mathbf{x}_j, \mathbf{c}, \mathbf{c}, \mathbf{n}_{ij})\ ,$$

then the scheme (2.20)-(2.22) preserves constant states.

Proof.
We recall the equation (2.22)

$$\dot{\mathbf{u}}_i V_i + \mathbf{u}_i \dot{V}_i \quad = \quad -\sum_{j=1}^{n_p} |\boldsymbol{\beta}_{ij}| \mathbf{g}_{ij} - \mathcal{B}_i\ .$$

Plugging (2.21) into this yields

$$\dot{\mathbf{u}}_i V_i \quad = \quad -\sum_{j=1}^{n_p} |\boldsymbol{\beta}_{ij}| \mathbf{g}_{ij} - \mathcal{B}_i - \mathbf{u}_i \sum_{j=1}^{n_p} \left(\dot{\mathbf{x}}_j \boldsymbol{\gamma}_{ij} - \dot{\mathbf{x}}_i \boldsymbol{\gamma}_{ji} \right)\ .$$

Now consider the constant state $\mathbf{u}_i = \mathbf{c} \in \mathbb{R}^m$ for all $i = 1, \ldots, n_p$. Then

$$\begin{aligned}
\dot{\mathbf{u}}_i V_i \quad &= \quad -\sum_{j=1}^{n_p} |\boldsymbol{\beta}_{ij}| (\boldsymbol{F}(\mathbf{c}) - \mathbf{c} \otimes \mathbf{a}_0) \cdot \mathbf{n}_{ij} - \mathcal{B}_i - \mathbf{c} \sum_{j=1}^{n_p} \left(\mathbf{a}_0 \boldsymbol{\gamma}_{ij} - \mathbf{a}_0 \boldsymbol{\gamma}_{ji} \right) \\
&= \quad -\sum_{j=1}^{n_p} (\boldsymbol{F}(\mathbf{c}) - \mathbf{c} \otimes \mathbf{a}_0) \cdot (\boldsymbol{\gamma}_{ij} - \boldsymbol{\gamma}_{ji}) - \int_{\partial\Omega} \psi_i \boldsymbol{F}(\mathbf{c}) \cdot \mathbf{n}\ d\sigma - \mathbf{c} \otimes \mathbf{a}_0 \cdot \sum_{j=1}^{n_p} \left(\boldsymbol{\gamma}_{ij} - \boldsymbol{\gamma}_{ji} \right) \\
&= \quad -\sum_{j=1}^{n_p} \boldsymbol{F}(\mathbf{c}) \cdot (\boldsymbol{\gamma}_{ij} - \boldsymbol{\gamma}_{ji}) - \int_{\partial\Omega} \psi_i \boldsymbol{F}(\mathbf{c})\ .\ \mathbf{n}\ d\sigma \\
&= \quad -\boldsymbol{F}(\mathbf{c}) \cdot \left(\sum_{j=1}^{n_p} \boldsymbol{\beta}_{ij} + \int_{\partial\Omega} \psi_i \mathbf{n}\ d\sigma \right) \overset{(2.28)}{=} \mathbf{0}
\end{aligned}$$

□

Teleaga showed in [60], that the FVPM is conservative in the sense of classical finite volume method:

Theorem 2.19 (Conservativity, [60])
If the numerical flux function $\boldsymbol{g}$ has the conservative property, i.e.,

$$\boldsymbol{g}(t, \mathbf{x}_i, \mathbf{x}_j, \boldsymbol{u}_i, \boldsymbol{u}_j, \boldsymbol{n}_{ij}) = -\boldsymbol{g}(t, \mathbf{x}_j, \mathbf{x}_i, \boldsymbol{u}_j, \boldsymbol{u}_i, -\boldsymbol{n}_{ij}) \ ,$$

and the coefficients $\boldsymbol{\beta}_{ij}$ satisfy the skew symmetry condition (2.24), then the FVPM (2.20)-(2.22) is conservative in the sense that

$$\frac{d}{dt}\left(\sum_{i=1}^{n_p} V_i \boldsymbol{u}_i\right) = -\int_{\partial\Omega} \boldsymbol{F}(\boldsymbol{u}) \cdot \boldsymbol{n} \ d\sigma \ .$$

Theorem 2.20 (Approximation property, [31] and [33])
Let the barycenter with respect to the test function ψ_i be defined as

$$\boldsymbol{b}_i := \frac{1}{V_i} \int_{\Omega} \boldsymbol{x} \psi_i(\boldsymbol{x}) d\boldsymbol{x}$$

and assume $\boldsymbol{u} \in \mathcal{C}^2(\Omega)$.
Then the discrete quantity $\boldsymbol{u}_i$ satisfies, with respect to $h = diam(supp\ \psi_i)$, the approximation property

$$\boldsymbol{u}_i = \boldsymbol{u}(\boldsymbol{b}_i) + \mathcal{O}(h^2) \ .$$

Theorem 2.21 (Approximation property, [32])
For the reconstruction (2.23) it holds

$$\boldsymbol{u}_h(\boldsymbol{x}, t) = \boldsymbol{u}(\boldsymbol{x}, t) + \mathcal{O}(h) \qquad \forall\ (\boldsymbol{x}, t) \in \Omega \times [0, T] \ .$$

The approximation quality of the reconstruction (2.23) was verified for two-dimensional case. Teleaga showed in [59] under suitable assumptions, that, for fixed $t > 0$, a scalar function $u(\cdot, t) \in H^1(\Omega)$, $\Omega \subset \mathbb{R}^2$ open set, can be approximated with $u_h(\cdot, t)$ from (2.23) with the error

$$\|u_h(\cdot, t) - u(\cdot, t)\|_{L^2(\Omega)} \leq Ch \|\nabla u\|_{L^2(\Omega)} \ ,$$

where the constant $C > 0$ does not depend on h.
Under further assumptions, Junk and Struckmeier [29] proved a Lax-Wendroff consistency result for the semi-discrete form of FVPM, i.e., if the numerical solution converges, then it convergences to a weak solution of the original PDE.
Under suitable assumptions, a L^∞-stability result for the scalar multidimensional case was given by Kaland [31] in the sense that

$$\|u_h(\cdot, t)\|_{L^\infty(\mathbb{R}^d)} \leq \|u_0\|_{L^\infty(\mathbb{R}^d)} \quad \forall t > 0 \ .$$

Considering these assumptions, positivity preserving property, L^1-estimate, weak BV-stability, monotonicity and discrete entropy inequality for the FVPM scheme were also shown. For more details, see [31].

Particle motion

As derived, the FVPM offers the possibility to define the particle motion arbitrarily. It is advantageous for problems with moving boundary, as treated in [31] or [59], since the change of the boundary is implicitly a part of the formulation of the method. However, the original scheme (2.20)-(2.22) has to be reformulated, since the domain $\Omega = \Omega(t)$ depends on time t.
Assume, that the boundary $\Gamma(t)$ of the domain $\Omega(t)$ is moving with a velocity $\boldsymbol{b}(\boldsymbol{x}, t)$, i.e.,

$$\Gamma(t) = \left\{ \boldsymbol{x}(t) \in \mathbb{R}^d \;\middle|\; \boldsymbol{x}(t) = \boldsymbol{x}_0 + \int_0^t \boldsymbol{b}(\boldsymbol{x}(\tau), \tau) d\tau \ , \ \boldsymbol{x}_0 \in \Gamma_0 \right\} \ ,$$

where Γ_0 is the initial boundary.
Considering (2.1) with $\Omega = \Omega(t)$, multiplying with ψ_i and integrating over $\Omega(t)$, one gets the equation (2.10) with new terms containing the velocity $\boldsymbol{b}$

$$\begin{aligned} \frac{d}{dt} \int_{\Omega(t)} \mathbf{u}\psi_i \ d\mathbf{x} &= \int_{\Omega(t)} \frac{d}{dt}(\mathbf{u}\psi_i) \ d\mathbf{x} + \int_{\partial\Omega(t)} \psi_i \mathbf{u} \otimes \boldsymbol{b} \cdot \boldsymbol{n} \ d\sigma \\ = \int_{\Omega(t)} (\psi_i)_t \ \mathbf{u} \ d\mathbf{x} &+ \int_{\Omega(t)} \boldsymbol{F}(\mathbf{u}) \cdot \nabla\psi_i \ d\mathbf{x} - \int_{\partial\Omega(t)} \psi_i \left(\boldsymbol{F}(\mathbf{u}) - \boldsymbol{u} \otimes \boldsymbol{b}\right) \cdot \mathbf{n} \ d\sigma \ . \end{aligned}$$

The derivation of the system of ODEs can be continued exactly as before and one ends up with the system

$$\begin{aligned} \dot{\mathbf{x}}_i(t) &= \mathbf{a}(\mathbf{x}_i, t, \mathbf{u}_i) \ , \\ \dot{V}_i(t) &= \sum_{j=1}^{n_p} \left(\dot{\mathbf{x}}_j \boldsymbol{\gamma}_{ij} - \dot{\mathbf{x}}_i \boldsymbol{\gamma}_{ji}\right) + \int_{\partial\Omega(t)} \psi_i \boldsymbol{b} \cdot \boldsymbol{n} d\sigma \ , \\ \frac{d}{dt}(\mathbf{u}_i V_i) &= -\sum_{j=1}^{n_p} |\boldsymbol{\beta}_{ij}| \mathbf{g}_{ij} - \int_{\partial\Omega(t)} \psi_i \left(\boldsymbol{F}(\mathbf{u}) - \boldsymbol{u} \otimes \boldsymbol{b}\right) \cdot \mathbf{n} \ d\sigma \ . \end{aligned}$$

The terms $V_i(t)$ can be alternatively computed out of the definition (2.7).
Numerical experiments concerning problems with moving boundaries were shown in [59] (model of a flow around an oscillating circle in two dimensions) and [31] (linearized piston problem).

In another approach, the particles are understood as physical particles carrying some physical information. So, e.g., in fluid dynamics we would consider the particles as particles of the fluid having velocity given by the fluid. Such particles are called *(purely) Lagrangian particles.* However, the use of purely Lagrangian particles can cause several numerical difficulties as first observed by Schick [51]. In special situations, for example near a discontinuity, the particles can incline to accumulate, which, due to a CFL condition, makes the time step smaller and the computational costs higher. Big variations of particle density can also cause numerical instabilities. Last but not least, following the Lax-Wendroff consistency result for FVPM, the variations in the velocity of the particles are required to be small. In the example given by Schick, this cannot be achieved with purely Lagrangian particles.
To overcome these difficulties, Schick proposed to introduce “repelling forces” between the particles to avoid the particles to get too close to each other. More specifically, consider the one-dimensional case. Let $u_i^n = u(x_i, t^n)$ denote the Lagrangian velocity of the fluid. Then the velocity of purely Lagrangian particles is defined as

$$\dot{x}_i^n := u_i^n \ .$$

Schick added to it a correction term

$$\dot{x}_i^n := u_i^n + q_i^n \ ,$$

where

$$q_i^n := \sum_j r(x_i^n - x_j^n) n_{ij}$$

with $n_{ij} = \operatorname{sgn}(x_i^n - x_j^n)$ and $r(x)$ being some function approximating the function $f(x) = \frac{1}{x^2}$.

A similar idea was presented later in [47]. The velocity is computed at every time step t^n, the index n will be omitted here. For the multidimensional case the velocity

$$\dot{\boldsymbol{x}}_i := \boldsymbol{u}_i + \boldsymbol{u}'_i$$

was introduced. The correction term $\boldsymbol{u}'_i$ is defined as

$$\boldsymbol{u}'_i = C \frac{\bar{r}_i}{\Delta t} \boldsymbol{R}_i \ ,$$

where

$$\bar{r}_i = \frac{1}{\#N(i)} \sum_{k \in N(i)} r_{ik}$$

is the average particle spacing in the neighborhood $N(i)$ of i. C is a constant, in [47] set to 1/1000 and

$$\boldsymbol{R}_i = \sum_{k \in N(i)} \frac{1}{\left(\frac{r_{ik}}{\bar{r}_i}\right)^2} \boldsymbol{n}_{ik} \ ,$$

where r_{ik} and $\boldsymbol{n}_{ik}$ are the distance and unit vector, respectively, from particle i to particle k. The approach was also demonstrated on relevant numerical examples.

Remark 2.22
In [47], the particle velocity in the numerical flux is computed as the average velocity of particles i and j. More general definition of the velocity used in the numerical flux (depending then on the choice of functions ψ_i) can be found in [31].

2.3 Adding a particle

Similarly to mesh-based methods, it is useful to have a possibility to add a new particle to the given particle distribution during the computation. Having this, we can, e.g., refine the particle distribution in the vicinity of a discontinuity to get a better resolution there. Another reason, on the contrary to the mesh-based methods, is, that using moving particles can cause their poor distribution in the computational domain or even "gaps" in their distribution. In other words, non-overlapping particles arise, i.e., the condition (2.6) is violated. To avoid this situation, a procedure of adding a particle is proposed. If a special choice of partition of unity is used, e.g., B-splines, specific properties of this partition of unity can be used to add a new particle in another, possibly better, way. Anyway, the proposed procedure is general and can be used for every type of particle functions building a partition of unity. In our approach, the supports of the original particle functions do not change.
The addition of a new particle takes place at a fixed time $t = t^n$ and does not depend on the time variable. For the sake of notation simplicity all functions appearing in this chapter will not depend on the time variable t.
Refining of the particle distribution can be seen as a local operation. Then such a procedure of adding a new particle should be defined as a local procedure. Furthermore, a standard requirement is to preserve the mass and constant states. As we will see, the naive direct approach is a global operation causing high computational costs. Our approach "localizes" these operations in order to get a local method preserving mass and constant states at the same time.

The scheme to add a particle

Let us consider a set of particles $\{\boldsymbol{x}_i\}_{i=1}^{n_p}$ in a bounded domain $\overline{\Omega} \subset \mathbb{R}^d$ defined in (2.3) and a corresponding set of particle basis functions $\{\psi_i\}_{i=1}^{n_p}$ defined in (2.5), such that (2.6) holds, with the coefficients $\{\boldsymbol{u}_i\}_{i=1}^{n_p}$ defined in (2.15) for some function $\boldsymbol{u}$.
The linear combination

$$\boldsymbol{u}_h(\boldsymbol{x}) = \sum_{i=1}^{n_p} \boldsymbol{u}_i \psi_i(\boldsymbol{x})$$

builds an approximation of the function $\boldsymbol{u}$ given by FVPM (cf. (2.23)).

The scheme to add a new particle proceeds as follows:

Scheme 2.23

1. Define a new particle set $\{\boldsymbol{x}_i^+\}_{i=1}^{n_p+1}$, such that
$$\boldsymbol{x}_i^+ = \boldsymbol{x}_i \ , \qquad i = 1, \dots, n_p$$
and a new particle $\boldsymbol{x}_{n_p+1}^+ \in \overline{\Omega}$.
2. Define new particle basis functions $\{\psi_i^+\}_{i=1}^{n_p+1}$ due to the definition (2.5) corresponding with the particles $\{\boldsymbol{x}_i^+\}_{i=1}^{n_p+1}$. It is assumed that the position of $\boldsymbol{x}_{n_p+1}^+$ is chosen in such a way, that (2.6) holds for functions $\{\psi_i^+\}_{i=1}^{n_p+1}$. The functions $\{\psi_i^+\}_{i=1}^{n_p+1}$ will build another partition of unity in $\overline{\Omega}$.
3. Define new coefficients $\{\boldsymbol{u}_i^+\}_{i=1}^{n_p+1}$ (the specific choice will be shown later).
4. Define the new function
$$\boldsymbol{u}_h^+(\boldsymbol{x}) := \sum_{i=1}^{n_p+1} \boldsymbol{u}_i^+ \psi_i^+(\boldsymbol{x}) \ , \ \boldsymbol{x} \in \overline{\Omega} \ .$$

Definition 2.24
We say that the scheme 2.23 preserves constant states *if*

$$\boldsymbol{u}_h(\boldsymbol{x}) = \boldsymbol{C} \ , \boldsymbol{C} \in \mathbb{R}^m \quad \Longrightarrow \quad \boldsymbol{u}_h^+(\boldsymbol{x}) = \boldsymbol{C} \ . \tag{2.36}$$

We say that the scheme 2.23 preserves the conservativity property *if*

$$\int_\Omega \boldsymbol{u}_h(\boldsymbol{x}) d\boldsymbol{x} = \int_\Omega \boldsymbol{u}_h^+(\boldsymbol{x}) d\boldsymbol{x} \ . \tag{2.37}$$

Remark 2.25
Remark that

$$\int_\Omega \boldsymbol{u}_h(\boldsymbol{x}) d\boldsymbol{x} = \int_\Omega \boldsymbol{u}_h^+(\boldsymbol{x}) d\boldsymbol{x} \quad \Longleftrightarrow \quad \sum_{i=1}^{n_p} \boldsymbol{u}_i V_i = \sum_{i=1}^{n_p+1} \boldsymbol{u}_i^+ V_i^+ \ .$$

Let us now look at the step 3 of the adding procedure in more detail. The standard direct approach would yield the definition

$$\widetilde{\boldsymbol{u}}_i^+ := \frac{1}{V_i^+} \int_\Omega \boldsymbol{u}_h(\boldsymbol{x}) \psi_i^+(\boldsymbol{x}) d\boldsymbol{x} = \frac{1}{V_i^+} \int_\Omega \sum_{j=1}^{n_p} \boldsymbol{u}_j \psi_j(\boldsymbol{x}) \psi_i^+(\boldsymbol{x}) d\boldsymbol{x} \tag{2.38}$$

for $i = 1, \ldots, n_p+1$. One can show, that the conditions (2.36) and (2.37) hold. The big disadvantage of this approach is its global impact, i.e., after adding a particle we have to compute all coefficients $\widetilde{\boldsymbol{u}}_i^+$, $i = 1, \ldots, n_p + 1$ according to

$$\widetilde{\boldsymbol{u}}_i^+ = \frac{1}{V_i^+} \sum_{j=1}^{n_p} \boldsymbol{u}_j \int_\Omega \psi_j(\boldsymbol{x}) \psi_i^+(\boldsymbol{x}) d\boldsymbol{x} \neq \boldsymbol{u}_i \ ,$$

since $\int_\Omega \psi_j(\boldsymbol{x}) \psi_i^+(\boldsymbol{x}) d\boldsymbol{x} \neq \delta_{ij} \int_\Omega \psi_i^+(\boldsymbol{x}) d\boldsymbol{x} = \delta_{ij} V_i^+$, $i, j = 1, \ldots, n_p$ due to the overlapping of particles.

One can observe, that only the functions neighboring to $\psi_{n_p+1}^+$ change after a particle has been added and for the remaining ones it holds

$$\psi_i^+ = \psi_i \ , \qquad \text{if } \operatorname{supp} \psi_i^+ \cap \operatorname{supp} \psi_{n_p+1}^+ = \emptyset \ , \ i \in \{1, \ldots, n_p\}.$$

We see that the basis built with $\{\psi_i\}_{i=1}^{n_p}$ differs from the new basis $\{\psi_i^+\}_{i=1}^{n_p+1}$ only locally in the surrounding of the added particle $\psi_{n_p+1}^+$. Therefore, steps 1 and 2 are local.

The following definitions and lemmata are auxiliary steps for theorem 2.35, where coefficients $\boldsymbol{u}_i^+$ of our method are defined.

Definition 2.26
The sets of neighbors of the particle $\psi_{n_p+1}^+$ are defined as

$$\begin{aligned} J &:= \{ j \in I\!N \mid \mathit{supp}\ \psi_j \cap \mathit{supp}\ \psi_{n_p+1}^+ \neq \emptyset \} \ , \\ J^+ &:= J \cup \{n_p + 1\} \ . \end{aligned}$$

Condition 2.27
Let $\omega \subset \overline{\Omega}$ be an open set. We define the condition

$$\mathit{supp}\ \psi_{n+1}^+ \subset \omega \ . \tag{2.39}$$

Definition 2.28

$$\begin{aligned} \Delta_j &:= \mathit{supp}\ \psi_j \cap \omega \ , \quad j = 1, \ldots, n_p \ , \\ \Delta_j^+ &:= \mathit{supp}\ \psi_j^+ \cap \omega \ , \quad j = 1, \ldots, n_p + 1 \ , \\ \mathcal{V}_j &:= \int_{\Delta_j} \psi_j \quad , \quad \mathring{\mathcal{V}}_j := \int_{\mathit{supp}\ \psi_j \setminus \Delta_j} \psi_j \ , \quad j = 1, \ldots, n_p \ , \\ \mathcal{V}_j^+ &:= \int_{\Delta_j^+} \psi_j^+ \quad , \quad \mathring{\mathcal{V}}_j^+ := \int_{\mathit{supp}\ \psi_j^+ \setminus \Delta_j^+} \psi_j^+ \ , \quad j = 1, \ldots, n_p + 1 \ . \end{aligned}$$

Lemma 2.29
Let $\omega \subset \overline{\Omega}$ be an open set. Then we have for $j = 1, \dots, n_p$

$$
\begin{aligned}
\Delta_j^+ &= \Delta_j \ , \\
\mathit{supp}\ \psi_j^+ \ \setminus \ \Delta_j^+ &= \mathit{supp}\ \psi_j \ \setminus \ \Delta_j \ , \\
\psi_j^+ &= \psi_j \quad \mathit{on\ supp}\ \psi_j^+ \ \setminus \ \Delta_j^+ \ , \\
V_j &= \mathcal{V}_j + \mathring{\mathcal{V}}_j
\end{aligned}
$$

and for $j = 1, \dots, n_p + 1$

$$
V_j^+ = \mathcal{V}_j^+ + \mathring{\mathcal{V}}_j^+ \ .
$$

Proof.
The claims follow directly from the definitions. □

Lemma 2.30
If $\omega \subset \overline{\Omega}$ satisfies (2.39), then

$$
\begin{aligned}
\mathring{\mathcal{V}}_j^+ &= \mathring{\mathcal{V}}_j \ , \quad j = 1, \dots, n_p \ , \\
\mathring{\mathcal{V}}_{n_p+1}^+ &= 0 \ .
\end{aligned}
$$

Proof.
For $j \in \{1, \dots, n_p\}$:

$$
\mathring{\mathcal{V}}_j^+ = \int\limits_{\mathrm{supp}\ \psi_j^+ \ \setminus \ \Delta_j^+} \psi_j^+ = \int\limits_{\mathrm{supp}\ \psi_j^+ \ \setminus \ \Delta_j^+} \psi_j = \int\limits_{\mathrm{supp}\ \psi_j \ \setminus \ \Delta_j} \psi_j = \mathring{\mathcal{V}}_j
$$

and

$$
\mathring{\mathcal{V}}_{n_p+1}^+ = \int\limits_{\mathrm{supp}\ \psi_{n_p+1}^+ \setminus \Delta_{n_p+1}^+} \psi_{n_p+1}^+ = \int\limits_{\emptyset} \psi_{n_p+1}^+ = 0 \ .
$$

□

Condition 2.31
Let $\omega \subset \overline{\Omega}$ be an open set. We define a set of indices $K \subset \{1, \dots, n_p\}$, such that

$$
\sum_{j \in K} \psi_j(\boldsymbol{x}) = 1 \quad \forall\ \boldsymbol{x} \in \omega \ . \tag{2.40}
$$

Definition 2.32
We define the set

$$
K^+ := K \cup \{n_p + 1\} \ .
$$

Lemma 2.33
Let $\{\psi_j\}_{j=1}^{n_p}$, $\{\psi_j^+\}_{j=1}^{n_p+1}$ *be the above defined partitions of unity in* $\overline{\Omega}$, $\omega \subset \overline{\Omega}$ *an open set and* K *satisfy (2.40). Then:*

a) K^+ *satisfies (2.40), i.e.,* $\sum\limits_{j \in K^+} \psi_j^+(\boldsymbol{x}) = 1 \quad \forall\, \boldsymbol{x} \in \omega$.

b) $i \in \{1, \dots, n_p\}$, $i \notin K \;\Rightarrow\;$ *supp* $\psi_i \cap \omega = \emptyset$, *supp* $\psi_i^+ \cap \omega = \emptyset$.

c) If ω *satisfies (2.39), then* $J \subset K$ *and* $J^+ \subset K^+$.

Proof.

a) K is a set of some indices of functions ψ_i where $\{\psi_i\}_{i=1}^{n_p}$ build a partition of unity in $\overline{\Omega}$ and $\{\psi_i\}_{i\in K}$ build a partition of unity in ω. Then K^+ is a set of some indices of functions ψ_i^+ where $\{\psi_i^+\}_{i=1}^{n_p+1}$ build a partition of unity in $\overline{\Omega}$ and $\{\psi_i^+\}_{i\in K^+}$ build a partition of unity in ω.

b) Let us consider supp $\psi_i \cap \omega \neq \emptyset$. Then there exists $\boldsymbol{x}_0 \in$ supp $\psi_i \cap \omega$, such that $\psi_i(\boldsymbol{x}_0) > 0$ and one can write

$$1 = \sum_{j=1}^{n_p+1} \psi_j(\boldsymbol{x}_0) \overset{\psi_j \geq 0}{\geq} \psi_i(\boldsymbol{x}_0) + \sum_{j \in K} \psi_j(\boldsymbol{x}_0) = \psi_i(\boldsymbol{x}_0) + 1 > 1 \ .$$

Hence, we have contradiction.
The proposition for supp ψ_i^+ follows from supp $\psi_i =$ supp ψ_i^+.

c) Follows directly from b).

$\square$

Lemma 2.34
Let $\{\psi_j\}_{j=1}^{n_p}$ *and* $\{\psi_j^+\}_{j=1}^{n_p+1}$ *be the above defined partitions of unity. Let* ω *be an open set satisfying (2.39). Then it holds*

a) $\sum\limits_{j\in J} V_j = \sum\limits_{i \in J^+} V_i^+ \quad , \quad \sum\limits_{j \notin J} V_j = \sum\limits_{i \notin J^+} V_i^+$.

b) $\sum\limits_{j\in J} \mathcal{V}_j = \sum\limits_{i \in J^+} \mathcal{V}_i^+$.

c) Let K *satisfy (2.40). Then*

$$\sum_{j\in K} V_j = \sum_{i \in K^+} V_i^+ \quad , \quad \sum_{j \notin K} V_j = \sum_{i \notin K^+} V_i^+ .$$

d) Let K *satisfy (2.40). Then* $\sum\limits_{j\in K} \mathcal{V}_j = \sum\limits_{i \in K^+} \mathcal{V}_i^+$.

Proof.

a)

$$1 = \sum_{j=1}^{n_p} \psi_j(\boldsymbol{x}) \qquad \Rightarrow \qquad |\Omega| = \sum_{j=1}^{n_p} \int_{\Omega} \psi_j(\boldsymbol{x}) d\boldsymbol{x} = \sum_{j=1}^{n_p} V_j \ ,$$

$$1 = \sum_{i=1}^{n_p+1} \psi_i^+(\boldsymbol{x}) \qquad \Rightarrow \qquad |\Omega| = \sum_{i=1}^{n_p+1} \int_{\Omega} \psi_i^+(\boldsymbol{x}) d\boldsymbol{x} = \sum_{i=1}^{n_p+1} V_i^+ \ ,$$

i.e.,

$$\sum_{j=1}^{n_p} V_j = \sum_{i=1}^{n_p+1} V_i^+ .$$

Because of

$$\psi_i^+ = \psi_i \ , \qquad \text{if } \ \text{supp } \psi_i^+ \cap \text{supp } \psi_{n+1}^+ = \emptyset \ , \ i \in \{1, \dots, n_p\} \ ,$$

it holds

$$V_i = V_i^+ \ , \qquad i \in \{1, \dots, n_p\} \setminus J \ .$$

Then

$$\sum_{j \notin J} V_j = \sum_{i \notin J^+} V_i^+ \ ,$$

because $\{j \notin J\} = \{i \notin J^+\}$. Finally,

$$\sum_{j \notin J} V_j + \sum_{j \in J} V_j = \sum_{j=1}^{n_p} V_j \quad = \quad \sum_{i=1}^{n_p+1} V_i^+ = \sum_{i \notin J^+} V_i^+ + \sum_{i \in J^+} V_i^+$$

gives

$$\sum_{j \in J} V_j = \sum_{i \in J^+} V_i^+ .$$

b) Due to lemma 2.29 it holds

$$V_j = \mathcal{V}_j + \mathring{\mathcal{V}}_j \ , \quad j \in J \ .$$

Then

$$\sum_{j \in J} V_j = \sum_{j \in J} \mathcal{V}_j + \sum_{j \in J} \mathring{\mathcal{V}}_j$$

$$\Rightarrow \quad \sum_{j \in J} \mathcal{V}_j = \sum_{j \in J} V_j - \sum_{j \in J} \mathring{\mathcal{V}}_j = \sum_{i \in J^+} V_i^+ - \sum_{i \in J^+} \mathring{\mathcal{V}}_i^+ = \sum_{i \in J^+} \mathcal{V}_i^+ \ ,$$

where we used the statements of lemma 2.29 and 2.30.

c) and d) follow immediatelly from a), b) and lemma 2.33c).

□

To keep the notation in the next theorem simple, we set

$$\boldsymbol{u}_{n_p+1} := \boldsymbol{0} \ .$$

Theorem 2.35
Let $\omega \subset \overline{\Omega}$ be an open set and $K \subset \{1,\dots,n_p\}$, such that (2.39) and (2.40) hold.
We define

$$\begin{aligned}
\boldsymbol{u}_i^+ &:= \frac{1}{V_i^+}\left[\int\limits_\omega \sum_{j\in K} \boldsymbol{u}_j\psi_j(\boldsymbol{x})\psi_i^+(\boldsymbol{x})d\boldsymbol{x} + \boldsymbol{u}_i\mathring{\mathcal{V}}_i^+\right],\ i \in K^+, && (2.41)\\
\boldsymbol{u}_i^+ &:= \boldsymbol{u}_i\ ,\ i \in \{1,\dots,n_p+1\}\setminus K^+. && (2.42)
\end{aligned}$$

Then the scheme 2.23 preserves constant states, i.e., the condition (2.36) holds, and the conservativity property (2.37) is satisfied. Especially, we have

$$\sum_{i\in K^+} \boldsymbol{u}_i^+ V_i^+ = \sum_{i\in K} \boldsymbol{u}_i V_i\ .$$

Proof.
We will show in two separate steps that (2.36) and (2.37) hold:

1) (2.36) holds:
A constant function $\boldsymbol{u}_h(\boldsymbol{x})$ can be written as $\boldsymbol{u}_h(\boldsymbol{x}) = \sum_{i=1}^{n_p} \boldsymbol{u}_i\psi_i(\boldsymbol{x})$, where $\boldsymbol{u}_i = \boldsymbol{C}$, $\boldsymbol{C} \in \mathbb{R}^m$ is a constant vector, i.e., $\boldsymbol{u}_h(\boldsymbol{x}) = \boldsymbol{C}\sum_{i=1}^{n_p}\psi_i(\boldsymbol{x}) = \boldsymbol{C}$. Then for $i \in \{1,\dots,n_p+1\}\setminus K^+$

$$\boldsymbol{u}_i^+ = \boldsymbol{C}$$

and for $i \in K^+$

$$\begin{aligned}
\boldsymbol{u}_i^+ &= \frac{1}{V_i^+}\left[\boldsymbol{C}\int\limits_\omega \sum_{j\in K}\psi_j(\boldsymbol{x})\psi_i^+(\boldsymbol{x})d\boldsymbol{x} + \boldsymbol{C}\mathring{\mathcal{V}}_i^+\right]\\
&\overset{(2.40)}{=} \frac{\boldsymbol{C}}{V_i^+}\left[\int\limits_\omega \psi_i^+(\boldsymbol{x})d\boldsymbol{x} + \mathring{\mathcal{V}}_i^+\right]\\
&= \frac{\boldsymbol{C}}{V_i^+}\left[\int\limits_{\Delta_i^+} \psi_i^+(\boldsymbol{x})d\boldsymbol{x} + \mathring{\mathcal{V}}_i^+\right] = \frac{\boldsymbol{C}}{V_i^+}\left[\mathcal{V}_i^+ + \mathring{\mathcal{V}}_i^+\right] = \boldsymbol{C}\ .
\end{aligned}$$

Hence, $\boldsymbol{u}_i^+ = \boldsymbol{C}$, $i = 1,\dots,n_p+1$, and therefore $\boldsymbol{u}_h^+(\boldsymbol{x}) = \sum_{i=1}^{n_p+1}\boldsymbol{u}_i^+\psi_i^+(\boldsymbol{x}) = \sum_{i=1}^{n_p+1}\boldsymbol{C}\psi_i^+(\boldsymbol{x}) = \boldsymbol{C}$ and (2.36) holds.

2) (2.37) holds:

$$\begin{aligned}
\sum_{i\in K^+}\boldsymbol{u}_i^+V_i^+ &= \int\limits_\omega \sum_{j\in K}\boldsymbol{u}_j\psi_j(\boldsymbol{x})\sum_{i\in K^+}\psi_i^+(\boldsymbol{x})d\boldsymbol{x} + \sum_{i\in K^+}\boldsymbol{u}_i\mathring{\mathcal{V}}_i^+\\
&\overset{(2.40)}{=} \int\limits_\omega \sum_{j\in K}\boldsymbol{u}_j\psi_j(\boldsymbol{x})d\boldsymbol{x} + \sum_{i\in K^+}\boldsymbol{u}_i\mathring{\mathcal{V}}_i^+\\
&= \sum_{j\in K}\boldsymbol{u}_j\int\limits_{\Delta_j}\psi_j(\boldsymbol{x})d\boldsymbol{x} + \sum_{i\in K}\boldsymbol{u}_i\mathring{\mathcal{V}}_i = \sum_{j\in K}\boldsymbol{u}_j\mathcal{V}_j + \sum_{i\in K}\boldsymbol{u}_i\mathring{\mathcal{V}}_i\\
&= \sum_{i\in K}\boldsymbol{u}_i(\mathcal{V}_i + \mathring{\mathcal{V}}_i) = \sum_{i\in K}\boldsymbol{u}_iV_i\ .
\end{aligned}$$

According to lemma 2.33, if $i \notin K^+$, then supp $\psi_i^+ \cap \omega = \emptyset$ and further, according to lemma 2.29, it holds

$$V_i^+ = V_i\ ,\quad i \in \{1,\dots,n_p+1\}\setminus K^+.$$

Finally,

$$\begin{aligned}\sum_{i=1}^{n_p+1} \boldsymbol{u}_i^+ V_i^+ &= \sum_{i\in K^+} \boldsymbol{u}_i^+ V_i^+ + \sum_{i\notin K^+} \boldsymbol{u}_i^+ V_i^+ \\ &= \sum_{i\in K} \boldsymbol{u}_i V_i + \sum_{i\notin K} \boldsymbol{u}_i V_i = \sum_{i=1}^{n_p} \boldsymbol{u}_i V_i \ ,\end{aligned}$$

so (2.37) holds.

$\square$

Remark 2.36
The formula for $\boldsymbol{u}_i^+$ in (2.41) is similar to (2.38). But now we split the integrand from (2.38) into two parts: First, we integrate over ω, where (2.40) holds, and (2.36) and (2.37) can be proven there. Second, in $\Omega \setminus \omega$, where the integration would bring troubles, the original value of $\boldsymbol{u}_i$ (the term $\boldsymbol{u}_i\mathring{\mathcal{V}}_i^+$) is kept. Since $\Omega \setminus \omega$ does not involve supp $\psi_{n_p+1}^+$ (the condition (2.39)), the functions ψ_i do not change there ($\psi_i = \psi_i^+$). So, it is reasonable to keep old values in this area.

Theorem 2.35 states how the coefficients $\boldsymbol{u}_i^+$ can be computed to obtain a scheme for which the conditions (2.36) and (2.37) are satisfied and the computation is local. Moreover, there are parameters ω and K that can be chosen freely, allowing us to define different methods, provided that (2.39) and (2.40) are satisfied. Numerical comparison of these methods is done in chapter 5.

Some possible choices of ω and K:

1) Method SUPP
   $\omega :=$ int(supp $\psi_{n_p+1}^+$)
   $K := J$

2) Method JI
   $J_I := \{j \in J \mid \nexists k \notin J : \ \text{supp}\ \psi_k \cap \text{supp}\ \psi_j \neq \emptyset\}$
   (J_I is a set of those neighbors of $\psi_{n_p+1}^+$, whose neighbors are only elements of J, i.e., "inner neighbors")

   $\omega :=$ int($\bigcup_{j\in J_I}$ supp ψ_j)
   $K := J$

3) Method JPLUS
   $J_P := \{j \in I\!N \mid \exists i \in J : \ \text{supp}\ \psi_j \cap \text{supp}\ \psi_i \neq \emptyset\}$
   (J_P is a set of all neighbors of elements of J including J too)

   $\omega :=$ int($\bigcup_{j\in J}$ supp ψ_j)
   $K := J_P$

Numerical implementation

The proposed scheme (2.41)-(2.42) gives us the possibility to add a particle in such a way that constants and mass are preserved. But, if the scheme is implemented directly, it will fail. The reason are the discretization errors arising during the computation of integrals in (2.41). For an illustration, consider the figure 2.3: The new particle is added at the position 0. Functions ψ_i (dashed line), functions ψ_i^+ (solid line), coefficients u_i (crosses) and coefficients u_i^+ (circles) are depicted. In the upper part of the plot, reconstructions of a given function $\sum_i u_i\psi_i$ (dashed line) and $\sum_i u_i^+\psi_i^+$ (solid line), respectively, are depicted. Top figures: on the left the reconstruction using the scheme (2.41)-(2.42) directly, on the right the expected result for a constant function.

Bottom figures: on the left the reconstruction using the scheme (2.41)-(2.42) directly, on the right the expected result for a linear function. The difference is caused by discretization errors.
The remedy is to take these errors into account. We introduce in this section a modification of scheme (2.41)-(2.42), analytically equivalent to the original one, but preserving constant states and conservativity also numerically. In this case, we consider only the *discretization errors*, the *rounding errors* are not taken into account.

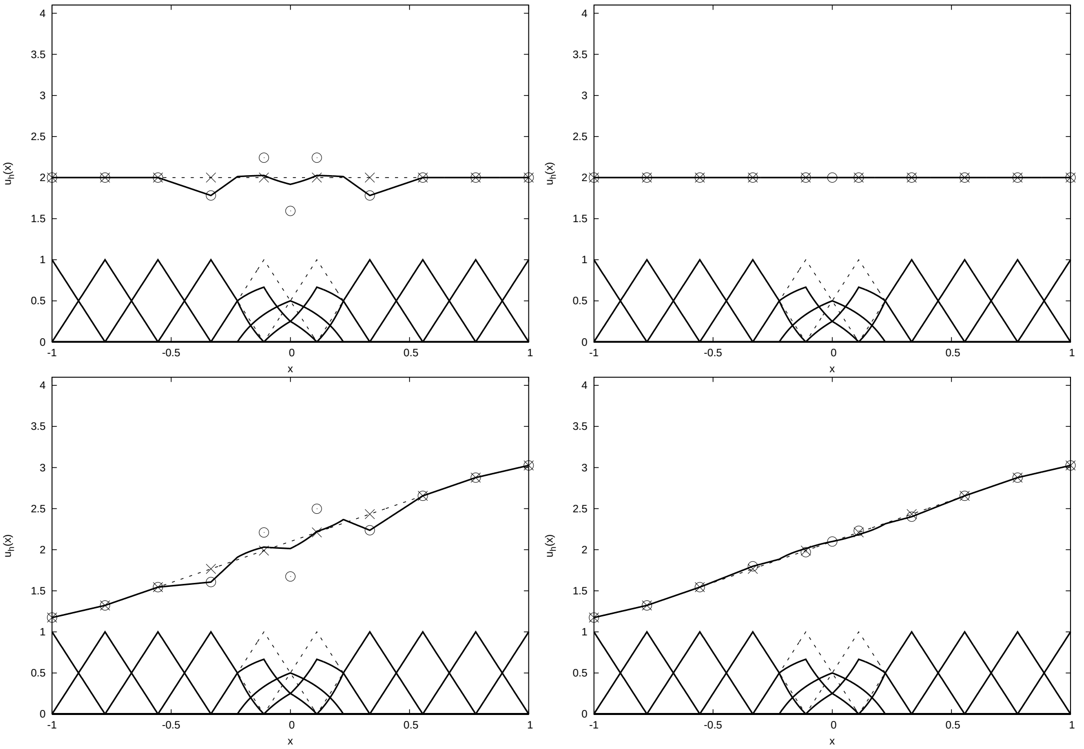


Figure 2.3: *Example of discretization errors in the scheme of adding a particle.*

First important property necessary for the scheme to work properly is the statement of lemma 2.30

$$\overset{\circ}{\mathcal{V}}{}_j^+ \quad = \quad \overset{\circ}{\mathcal{V}}_j \ , \quad j \in K.$$

That is why we compute $\overset{\circ}{\mathcal{V}}{}_j^+$ and $\overset{\circ}{\mathcal{V}}_j$ from the definition and $\mathcal{V}_j^+$ and $\mathcal{V}_j$ due to the lemma 2.29, namely

$$\begin{aligned} \mathcal{V}_j \quad &:= \quad V_j - \overset{\circ}{\mathcal{V}}_j \ , \quad j \in K \ , \\ \mathcal{V}_j^+ \quad &:= \quad V_j^+ - \overset{\circ}{\mathcal{V}}{}_j^+ \ , \quad j \in K \ . \end{aligned}$$

The values V_j^+ are computed due to the definition (2.7) for $j \in K$ and

$$\mathcal{V}_{n_p+1}^+ = V_{n_p+1}^+ := \sum_{j=1}^{n_p} V_j - \sum_{i=1}^{n_p} V_i^+ \ ,$$

in order to conserve the total volume of particles ($\sum_{j=1}^{n_p} V_j = \sum_{i=1}^{n_p+1} V_i^+$), which is necessary for conservativity in case of constant states.
Under the assumptions (2.39) and (2.40) we have proven, that the scheme (2.41)-(2.42) satisfies (2.36) and (2.37). In (2.41) we compute the integrals

$$I_{ji} = \int\limits_{\omega} \psi_j(\boldsymbol{x})\psi_i^+(\boldsymbol{x})d\boldsymbol{x} \ , \quad j \in K, \ i \in K^+ \ ,$$

where the following discretization errors may arise.

Definition 2.37
Define the errors

$$\begin{aligned} E_j &:= \mathcal{V}_j - \sum_{i\in K^+} I_{ji} \ , \quad j \in K, \\ E_i^+ &:= \mathcal{V}_i^+ - \sum_{j\in K} I_{ji} \ , \quad i \in K^+. \end{aligned}$$

If we compute the integrals I_{ji} exactly, the terms E_j and E_i^+ will be 0. In that case, adding and subtracting of these terms is equivalent to adding and subtracting of a zero.

Lemma 2.38
Let $\{\psi_j\}_{j=1}^{n_p}$ and $\{\psi_j^+\}_{j=1}^{n_p+1}$ be the above defined partitions of unity. Let $\omega \subset \overline{\Omega}$ be an open set and $K \subset \{1, \ldots, n_p\}$, such that (2.39) and (2.40) hold. Then

$$\sum_{j\in K} E_j = \sum_{i\in K^+} E_i^+ \ .$$

Proof.
Summing the errors

$$\begin{aligned} E_j &= \mathcal{V}_j - \sum_{i\in K^+} I_{ji} \Bigg/ \sum_{j\in K} \\ E_i^+ &= \mathcal{V}_i^+ - \sum_{j\in K} I_{ji} \Bigg/ \sum_{i\in K^+} \end{aligned}$$

for all j and i, respectively, and subtraction of the sums gives

$$\sum_{j\in K} E_j - \sum_{i\in K^+} E_i^+ \ = \ \sum_{j\in K} \mathcal{V}_j - \sum_{j\in K} \sum_{i\in K^+} I_{ji} - \sum_{i\in K^+} \mathcal{V}_i^+ + \sum_{i\in K^+} \sum_{j\in K} I_{ji} = 0 \ ,$$

due to the lemmata 2.29, 2.30 and 2.34 and the computation of $V_{n_p+1}^+ = \sum_{j=1}^{n_p} V_j - \sum_{i=1}^{n_p} V_i^+$. ☐

The terms in (2.41) can be written as

$$\boldsymbol{u}_i^+ \ = \ \frac{1}{V_i^+} \left[\sum_{j\in K} \boldsymbol{u}_j I_{ji} + \boldsymbol{u}_i \mathring{\mathcal{V}}_i^+ \right] \ , \ i \in K^+.$$

Then, for $i \in \{1, \dots, n_p\}$ we add the term $\frac{1}{V_i^+}\boldsymbol{u}_i E_i^+$ to the right hand side and acquire

$$
\begin{aligned}
\boldsymbol{u}_i^+ &= \frac{1}{V_i^+}\left[\sum_{j\in K}\boldsymbol{u}_j I_{ji} + \boldsymbol{u}_i E_i^+ + \boldsymbol{u}_i \overset{\circ}{\mathcal{V}}_i^+\right] = \frac{1}{V_i^+}\left[\sum_{j\in K}(\boldsymbol{u}_j - \boldsymbol{u}_i) I_{ji} + \boldsymbol{u}_i V_i^+\right] \\
&= \boldsymbol{u}_i + \frac{1}{V_i^+}\sum_{j\in K}(\boldsymbol{u}_j - \boldsymbol{u}_i) I_{ji} \ .
\end{aligned}
$$

The term in (2.41) for $i = n_p + 1$ can be written as

$$
\boldsymbol{u}_{n_p+1}^+ = \frac{1}{V_{n_p+1}^+}\sum_{j\in K}\boldsymbol{u}_j I_{j,n_p+1}.
$$

We add the term $\frac{1}{V_{n_p+1}^+}\sum_{j\in K}\boldsymbol{u}_j\left(E_j - E_j^+\right)$ to this and get

$$
\boldsymbol{u}_{n_p+1}^+ = \frac{1}{V_{n_p+1}^+}\left[\sum_{j\in K}\boldsymbol{u}_j I_{j,n_p+1} + \sum_{j\in K}\boldsymbol{u}_j\left(E_j - E_j^+\right)\right].
$$

Theorem 2.39
Let $\omega \subset \overline{\Omega}$ be an open set and $K \subset \{1, \dots, n_p\}$, such that (2.39) and (2.40) hold.
We define

$$
\begin{aligned}
\boldsymbol{u}_i^+ &:= \boldsymbol{u}_i + \frac{1}{V_i^+}\sum_{j\in K}(\boldsymbol{u}_j - \boldsymbol{u}_i) I_{ji} \ , \quad i \in K^+, i \neq n_p + 1, && (2.43) \\
\boldsymbol{u}_{n_p+1}^+ &:= \frac{1}{V_{n_p+1}^+}\left[\sum_{j\in K}\boldsymbol{u}_j I_{j,n_p+1} + \sum_{j\in K}\boldsymbol{u}_j\left(E_j - E_j^+\right)\right], && (2.44) \\
\boldsymbol{u}_i^+ &:= \boldsymbol{u}_i \ , \quad i \notin K^+. && (2.45)
\end{aligned}
$$

Then the scheme 2.23 preserves constant states, i.e., the condition (2.36) holds, and the conservativity property (2.37) is satisfied. Especially, we have

$$
\sum_{i\in K^+}\boldsymbol{u}_i^+ V_i^+ = \sum_{i\in K}\boldsymbol{u}_i V_i \ .
$$

Proof.
The proof is similar to the proof of the theorem 2.35 and will be performed again in two steps to prove that (2.36) and (2.37) hold.

1) (2.36) holds:
   A constant function $\boldsymbol{u}_h(\boldsymbol{x})$ can be written as $\boldsymbol{u}_h(\boldsymbol{x}) = \sum_{i=1}^{n_p}\boldsymbol{u}_i\psi_i(\boldsymbol{x})$, where $\boldsymbol{u}_i = \boldsymbol{C}$, $\boldsymbol{C} \in \mathbb{R}^m$ is a constant vector, i.e., $\boldsymbol{u}_h(\boldsymbol{x}) = \boldsymbol{C}\sum_{i=1}^{n_p}\psi_i(\boldsymbol{x}) = \boldsymbol{C}$.
   Then, for $i \in \{1, \dots, n_p\}$, we have obviously

$$
\boldsymbol{u}_i^+ = \boldsymbol{C}.
$$

   It remains to show the statement for $i = n_p + 1$.

$$
\begin{aligned}
\boldsymbol{u}_{n_p+1}^+ &= \frac{\boldsymbol{C}}{V_{n_p+1}^+}\left[\sum_{j\in K} I_{j,n_p+1} + \sum_{j\in K} E_j - \sum_{j\in K} E_j^+\right] \\
&= \frac{\boldsymbol{C}}{V_{n_p+1}^+}\left[\sum_{j\in K} I_{j,n_p+1} + \sum_{j\in K^+} E_j^+ - \sum_{j\in K} E_j^+\right] \\
&= \frac{\boldsymbol{C}}{V_{n_p+1}^+}\left[\sum_{j\in K} I_{j,n_p+1} + E_{n_p+1}^+\right] = \frac{\boldsymbol{C}}{V_{n_p+1}^+} V_{n_p+1}^+ = \boldsymbol{C} \ ,
\end{aligned}
$$

because $\mathring{\mathcal{V}}^+_{n_p+1} = 0$, so $E^+_{n_p+1} = V^+_{n_p+1} - \sum\limits_{j\in K} I_{j,n_p+1}$.

Hence, $\boldsymbol{u}_i^+ = \boldsymbol{C}$, $i = 1,\dots,n_p{+}1$, and therefore $\boldsymbol{u}_h^+(\boldsymbol{x}) = \sum_{i=1}^{n_p+1} \boldsymbol{u}_i^+\psi_i^+(\boldsymbol{x}) = \sum_{i=1}^{n_p+1}\boldsymbol{C}\psi_i^+(\boldsymbol{x}) = \boldsymbol{C}$. Finally, (2.36) holds.

2) (2.37) holds:
We use the definitions (2.43) and (2.44) and acquire

$$\begin{aligned}
\sum_{i\in K}\boldsymbol{u}_i^+V_i^+ &= \sum_{i\in K}\boldsymbol{u}_iV_i^+ + \sum_{i\in K}\sum_{j\in K}\boldsymbol{u}_jI_{ji} - \sum_{i\in K}\sum_{j\in K}\boldsymbol{u}_iI_{ji}\ ,\\
\boldsymbol{u}^+_{n_p+1}V^+_{n_p+1} &= \sum_{j\in K}\boldsymbol{u}_jI_{j,n_p+1} + \sum_{j\in K}\boldsymbol{u}_jE_j - \sum_{j\in K}\boldsymbol{u}_kE_k^+\ .
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{i\in K^+}\boldsymbol{u}_i^+V_i^+ &= \sum_{i\in K}\boldsymbol{u}_i^+V_i^+ + \boldsymbol{u}^+_{n_p+1}V^+_{n_p+1}\\
&= \sum_{i\in K}\boldsymbol{u}_iV_i^+ + \sum_{i\in K}\sum_{j\in K}\boldsymbol{u}_jI_{ji} - \sum_{i\in K}\sum_{j\in K}\boldsymbol{u}_iI_{ji}\\
&\quad + \sum_{j\in K}\boldsymbol{u}_jI_{j,n_p+1} + \sum_{j\in K}\boldsymbol{u}_jE_j - \sum_{j\in K}\boldsymbol{u}_kE_k^+\\
&= \sum_{i\in K}\boldsymbol{u}_iV_i^+ + \sum_{j\in K}\boldsymbol{u}_jI_{j,n_p+1}\\
&\quad + \sum_{j\in K}\boldsymbol{u}_j\left(\sum_{i\in K}I_{ji} + E_j\right) - \sum_{i\in K}\boldsymbol{u}_i\left(\sum_{j\in K}I_{ji} + E_i^+\right)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i\in K}\boldsymbol{u}_iV_i^+ + \sum_{j\in K}\boldsymbol{u}_jI_{j,n_p+1}\\
&\quad + \sum_{j\in K}\boldsymbol{u}_j\left(\sum_{i\in K}I_{ji} + \left[\mathcal{V}_j - \sum_{i\in K^+}I_{ji}\right]\right)\\
&\quad - \sum_{i\in K}\boldsymbol{u}_i\left(\sum_{j\in K}I_{ji} + \left[\mathcal{V}_i^+ - \sum_{j\in K}I_{ji}\right]\right)\\
&= \sum_{i\in K}\boldsymbol{u}_iV_i^+ + \sum_{j\in K}\boldsymbol{u}_jI_{j,n_p+1}\\
&\quad + \sum_{j\in K}\boldsymbol{u}_j\left(\mathcal{V}_j - I_{j,n_p+1}\right) - \sum_{i\in K}\boldsymbol{u}_i\mathcal{V}_i^+\\
&= \sum_{i\in K}\boldsymbol{u}_iV_i^+ + \sum_{j\in K}\boldsymbol{u}_j\mathcal{V}_j - \sum_{i\in K}\boldsymbol{u}_i\mathcal{V}_i^+\\
&= \sum_{i\in K}\boldsymbol{u}_i\mathring{\mathcal{V}}_i^+ + \sum_{j\in K}\boldsymbol{u}_j\mathcal{V}_j = \sum_{i\in K}\boldsymbol{u}_i\mathring{\mathcal{V}}_i + \sum_{j\in K}\boldsymbol{u}_j\mathcal{V}_j\\
&= \sum_{j\in K}\boldsymbol{u}_j\mathring{\mathcal{V}}_j + \sum_{j\in K}\boldsymbol{u}_j\mathcal{V}_j = \sum_{j\in K}\boldsymbol{u}_jV_j\ .
\end{aligned}$$

According to the lemma 2.33, if $i \notin K^+$, then supp $\psi_i^+ \cap \omega = \emptyset$ and further, according to the lemma 2.29, it holds

$$V_i^+ = V_i\ , \quad i \in \{1,\dots,n_p+1\}\setminus K^+.$$

Finally,

$$\begin{aligned}\sum_{i=1}^{n_p+1} \boldsymbol{u}_i^+ V_i^+ &= \sum_{i\in K^+} \boldsymbol{u}_i^+ V_i^+ + \sum_{i\notin K^+} \boldsymbol{u}_i^+ V_i^+ \\ &= \sum_{i\in K} \boldsymbol{u}_i V_i + \sum_{i\notin K} \boldsymbol{u}_i V_i = \sum_{i=1}^{n_p} \boldsymbol{u}_i V_i \ ,\end{aligned}$$

hence, (2.37) holds.

□

Remark 2.40
Both schemes, (2.41)-(2.42) and (2.43)-(2.45), will be equivalent, if we can compute integrals exactly. If not, the scheme (2.43)-(2.45) satisfies the properties (2.36) and (2.37) (up to the machine precision), but the method (2.41)-(2.42) does not.

2.4 Removing a particle

In some situations, it is also desired to remove a particle, e.g., to coarsen the distribution of particles or if a particle moves out of the computational domain. Following the concept of section 2.3, we will introduce a method for removing a particle. The structure of the section and of the method itself is very similar to the case of adding a particle, but is presented here due to the technical differences. One further assumption (see condition 2.49) has to be added, since we should not produce “holes” or “gaps” in the particle distribution.

The scheme of removing a particle

We adopt definitions of quantities parallel to the quantities of previous section.
For the sake of simplicity, assume that the index of the particle to be removed is the index n_p.

The linear combination

$$\boldsymbol{u}_h(\boldsymbol{x}) = \sum_{i=1}^{n_p} \boldsymbol{u}_i \psi_i(\boldsymbol{x})$$

builds an approximation of the function $\boldsymbol{u}$.

We introduce the general scheme to remove a particle:

Scheme 2.41

1. Define a new particle set $\{\boldsymbol{x}_i^-\}_{i=1}^{n_p-1}$, such that

$$\boldsymbol{x}_i^- = \boldsymbol{x}_i \ , \qquad i = 1, \ldots, n_p - 1 \ .$$

2. Define new particle basis functions $\{\psi_i^-\}_{i=1}^{n_p-1}$ due to the definition (2.5) corresponding with the particles $\{\boldsymbol{x}_i^-\}_{i=1}^{n_p-1}$. The functions $\{\psi_i^-\}_{i=1}^{n_p-1}$ will build another partition of unity in $\overline{\Omega}$.

3. Define new coefficients $\{\boldsymbol{u}_i^-\}_{i=1}^{n_p-1}$ (the specific choice will be shown later).
4. Define the new function

$$\boldsymbol{u}_h^-(\boldsymbol{x}) := \sum_{i=1}^{n_p-1} \boldsymbol{u}_i^- \psi_i^-(\boldsymbol{x}) \ , \ \boldsymbol{x} \in \overline{\Omega} \ .$$

Definition 2.42
We say that the scheme 2.41 preserves constant states *if*

$$\boldsymbol{u}_h(\boldsymbol{x}) = \boldsymbol{C} \ , \boldsymbol{C} \in \mathbb{R}^m \quad \Longrightarrow \quad \boldsymbol{u}_h^-(\boldsymbol{x}) = \boldsymbol{C} \ . \tag{2.46}$$

We say that the scheme 2.41 preserves the conservativity property *if*

$$\int_\Omega \boldsymbol{u}_h(\boldsymbol{x}) d\boldsymbol{x} = \int_\Omega \boldsymbol{u}_h^-(\boldsymbol{x}) d\boldsymbol{x} \ . \tag{2.47}$$

Definition 2.43
The sets of neighbors of the particle ψ_{n_p} are defined as

$$\begin{aligned} J &:= \{j \in I\!N \mid supp\ \psi_j \cap supp\ \psi_{n_p} \neq \emptyset\}, \\ J^- &:= J \setminus \{n_p\}. \end{aligned}$$

Condition 2.44
Let $\omega \subset \overline{\Omega}$ be an open set. We define the condition:

$$supp\ \psi_{n_p} \subset \omega \ . \tag{2.48}$$

Definition 2.45
Let $\omega \subset \overline{\Omega}$ be an open set. We define

$$\begin{aligned} \Delta_j &:= supp\ \psi_j \cap \omega \ , \quad j = 1, \dots, n_p \ , \\ \Delta_j^- &:= supp\ \psi_j^- \cap \omega \ , \quad j = 1, \dots, n_p - 1 \ , \\ \mathcal{V}_j &:= \int\limits_{\Delta_j} \psi_j \quad , \quad \overset{\circ}{\mathcal{V}}_j := \int\limits_{supp\ \psi_j\ \setminus\ \Delta_j} \psi_j \ , \quad j = 1, \dots, n_p \ , \\ \mathcal{V}_j^- &:= \int\limits_{\Delta_j^-} \psi_j^- \quad , \quad \overset{\circ}{\mathcal{V}}{}_j^- := \int\limits_{supp\ \psi_j^-\ \setminus\ \Delta_j^-} \psi_j^- \ , \quad j = 1, \dots, n_p - 1 \ . \end{aligned}$$

Condition 2.46
Let $\omega \subset \overline{\Omega}$ be an open set. We define a set of indices $K \subset \{1, \dots, n_p\}$ such that

$$\sum_{j \in K} \psi_j(\boldsymbol{x}) = 1 \quad \forall\ \boldsymbol{x} \in \omega \ . \tag{2.49}$$

Remark 2.47
From assumptions (2.48) and (2.49) it follows, that

$$n_p \in K \ .$$

Definition 2.48

$$K^- := K \setminus \{n_p\} \ .$$

Condition 2.49
Let $\omega \subset \overline{\Omega}$ be an open set. We define the condition:

$$\text{the condition (2.6) is satisfied on } \omega \text{ for functions corresponding to the index set } K^-, \text{ i.e., the functions } \{\psi_j\}_{j\in K^-} \text{ overlap in } \omega. \tag{2.50}$$

Lemma 2.50
Let $\{\psi_j\}_{j=1}^{n_p}$ and $\{\psi_j^-\}_{j=1}^{n_p-1}$ be the above defined partitions of unity. Let ω be an open set and $K \subset \{1,\dots,n_p\}$, such that (2.48) and (2.49) hold. Let (2.50) be satisfied. Then we have

$$\sum_{i\in K^-} (\mathcal{V}_i^- - \mathcal{V}_i) = V_{n_p}.$$

Proof.
Similarly as in previous section

$$\sum_{i\in K^-} \mathcal{V}_i^- = \sum_{i\in K} \mathcal{V}_i = \sum_{i\in K^-} \mathcal{V}_i + \mathcal{V}_{n_p} = \sum_{i\in K^-} \mathcal{V}_i + V_{n_p} \ .$$

□

Similar results as in lemmata in the previous section can be achieved. To keep the thesis simple, these technical lemmata are omitted here and only the main result of this section is shown.

Theorem 2.51
Let $\omega \subset \overline{\Omega}$ be an open set and $K \subset \{1,\dots,n_p\}$, such that (2.48) and (2.49) hold. Let (2.50) be satisfied.
We define

$$\boldsymbol{u}_i^- := \frac{1}{V_i^-}\left[\int_\omega \sum_{j\in K} \boldsymbol{u}_j \psi_j(\boldsymbol{x})\psi_i^-(\boldsymbol{x})d\boldsymbol{x} + \boldsymbol{u}_i \mathring{\mathcal{V}}_i^-\right] \ , \ i \in K^-, \tag{2.51}$$

$$\boldsymbol{u}_i^- := \boldsymbol{u}_i \ , \ i \in \{1,\dots,n_p-1\} \setminus K^-. \tag{2.52}$$

Then the scheme 2.41 preserves constant states, i.e., the condition (2.46) holds, and the conservativity property (2.47) is satisfied. Especially, we have

$$\sum_{i\in K^-} \boldsymbol{u}_i^- V_i^- = \sum_{i\in K} \boldsymbol{u}_i V_i \ .$$

Proof.
The proof can be carried out in the same way as the proof of the theorem 2.35. □

Numerical implementation

The discretization errors play again a significant role here. The remedy is achieved in a similar way.
We compute the temporary values

$$\widetilde{V}_j^- \ , \quad j \in K^-$$

due to the formula (2.7). A non-zero computational error $\sum_{j\in K} V_j - \sum_{j\in K^-} \widetilde{V}_j^-$ will occur. In the case of adding a particle there was the term $V_{n_p+1}^+$ where we could "store" this error. In case of removing a particle we split this error among all new terms as follows

$$V_i^- := \widetilde{V}_i^- + \frac{1}{\#(K^-)} \left(\sum_{j\in K} V_j - \sum_{j\in K^-} \widetilde{V}_j^- \right) \ , \quad i \in K^- \ ,$$

in order to conserve the total volume of particles ($\sum_{j=1}^{n_p} V_j = \sum_{i=1}^{n_p-1} V_i^-$), which is necessary for conservativity in case of constant states. Another possibility is to store the error in only one of the terms V_i^-.
We compute $\overset{\circ}{\mathcal{V}}{}_j^-$ and $\overset{\circ}{\mathcal{V}}_j$ from the definition and $\mathcal{V}_j^-$ and $\mathcal{V}_j$ according to

$$\begin{aligned} \mathcal{V}_j &:= V_j - \overset{\circ}{\mathcal{V}}_j \ , \quad j \in K \ , \\ \mathcal{V}_j^- &:= V_j^- - \overset{\circ}{\mathcal{V}}{}_j^- \ , \quad j \in K^- \ . \end{aligned}$$

In this section, under the assumptions (2.48), (2.49) and (2.50) we have proven, that the scheme (2.51)-(2.52) satisfies (2.46) and (2.47). In (2.51) we compute the integrals

$$I_{ji} = \int_\omega \psi_j(\boldsymbol{x})\psi_i^-(\boldsymbol{x})d\boldsymbol{x} \ , \quad j \in K, \ i \in K^- \ ,$$

where substantial discretization errors arise.

Definition 2.52
Define the errors

$$\begin{aligned} E_j &:= \mathcal{V}_j - \sum_{i\in K^-} I_{ji} \ , \quad j \in K, \\ E_i^- &:= \mathcal{V}_i^- - \sum_{j\in K} I_{ji} \ , \quad i \in K^-. \end{aligned}$$

Theorem 2.53
Let $\omega \subset \overline{\Omega}$ be an open set and $K \subset \{1, \ldots, n_p\}$, such that (2.48) and (2.49) hold. Let (2.50) be satisfied.
We define

$$\boldsymbol{u}_i^- = \frac{1}{V_i^-} \left[\sum_{j\in K} \boldsymbol{u}_j I_{ji} + \boldsymbol{u}_{n_p} E_i^- + (\boldsymbol{u}_i - \boldsymbol{u}_{n_p}) E_i + \boldsymbol{u}_i \overset{\circ}{\mathcal{V}}{}_i^- \right] \ , \quad i \in K^- \ , \tag{2.53}$$

$$\boldsymbol{u}_i^- = \boldsymbol{u}_i \ , \quad i \notin K^-. \tag{2.54}$$

Then the scheme 2.41 preserves constant states, i.e., the condition (2.46) holds, and the conservativity property (2.47) is satisfied. Especially, we have

$$\sum_{i\in K^-} \boldsymbol{u}_i^- V_i^- = \sum_{i\in K} \boldsymbol{u}_i V_i \ .$$

Remark 2.54
Other equivalent form of (2.53) is for $i \in K^-$

$$\boldsymbol{u}_i^- \quad = \quad \frac{1}{V_i^-}\left[\sum_{j\in K^-}(\boldsymbol{u}_j - \boldsymbol{u}_{n_p})I_{ji} + \boldsymbol{u}_{n_p}\mathcal{V}_i^- + (\boldsymbol{u}_i - \boldsymbol{u}_{n_p})E_i + \boldsymbol{u}_i\overset{\circ}{\mathcal{V}}{}_i^-\right] \ , \tag{2.55}$$

since

$$\begin{aligned}\sum_{j\in K} \boldsymbol{u}_j I_{ji} + \boldsymbol{u}_{n_p}E_i^- &= \sum_{j\in K^-} \boldsymbol{u}_j I_{ji} + \boldsymbol{u}_{n_p} I_{n_p,i} + \boldsymbol{u}_{n_p}E_i^- \\ = \sum_{j\in K^-} \boldsymbol{u}_j I_{ji} + \boldsymbol{u}_{n_p} I_{n_p,i} + \boldsymbol{u}_{n_p}\left(\mathcal{V}_i^- - \sum_{j\in K} I_{ji}\right) &= \sum_{j\in K^-}(\boldsymbol{u}_j - \boldsymbol{u}_{n_p})I_{ji} + \boldsymbol{u}_{n_p}\mathcal{V}_i^- \ .\end{aligned}$$

We will use this in the proof of theorem 2.53.

Proof of theorem 2.53.
The proof is similar to the proof of the theorem 2.51.

1) (2.46) holds:
   A constant function $\boldsymbol{u}_h(\boldsymbol{x})$ can be written as $\boldsymbol{u}_h(\boldsymbol{x}) = \sum_{i=1}^{n_p} \boldsymbol{u}_i\psi_i(\boldsymbol{x})$, where $\boldsymbol{u}_i = \boldsymbol{C}$, $\boldsymbol{C} \in \mathbb{R}^m$ is a constant vector, i.e., $\boldsymbol{u}_h(\boldsymbol{x}) = \boldsymbol{C}\sum_{i=1}^{n_p}\psi_i(\boldsymbol{x}) = \boldsymbol{C}$.
   Then, for $i \in \{1, \dots, n_p - 1\} \setminus K^-$, we have obviously

   $$\boldsymbol{u}_i^- \quad = \quad \boldsymbol{C}.$$

   And for $i \in K^-$ it follows from (2.55) that

   $$\boldsymbol{u}_i^- \quad = \quad \frac{1}{V_i^-}\left[\boldsymbol{C}\mathcal{V}_i^- + \boldsymbol{C}\overset{\circ}{\mathcal{V}}{}_i^-\right] = \boldsymbol{C} \ ,$$

   hence, (2.46) holds.

2) (2.47) holds:

We multiply (2.55) with V_i^- and sum over K^-. We obtain

$$\begin{aligned}
\sum_{i\in K^-} \boldsymbol{u}_i^- V_i^- &= \sum_{j\in K^-} (\boldsymbol{u}_j - \boldsymbol{u}_{n_p}) \sum_{i\in K^-} I_{ji} \\
&\quad + \boldsymbol{u}_{n_p} \sum_{i\in K^-} \mathcal{V}_i^- + \sum_{i\in K^-} (\boldsymbol{u}_i - \boldsymbol{u}_{n_p}) E_i + \sum_{i\in K^-} \boldsymbol{u}_i \mathring{\mathcal{V}}_i^- \\
&= \sum_{j\in K^-} (\boldsymbol{u}_j - \boldsymbol{u}_{n_p})(\mathcal{V}_j - E_j) \\
&\quad + \boldsymbol{u}_{n_p} \sum_{i\in K^-} \mathcal{V}_i^- + \sum_{i\in K^-} \boldsymbol{u}_i E_i - \boldsymbol{u}_{n_p} \sum_{i\in K^-} E_i + \sum_{i\in K^-} \boldsymbol{u}_i \mathring{\mathcal{V}}_i^- \\
&= \sum_{j\in K^-} \boldsymbol{u}_j \mathcal{V}_j - \sum_{j\in K^-} \boldsymbol{u}_j E_j - \boldsymbol{u}_{n_p} \sum_{j\in K^-} \mathcal{V}_j + \boldsymbol{u}_{n_p} \sum_{j\in K^-} E_j \\
&\quad + \boldsymbol{u}_{n_p} \sum_{i\in K^-} \mathcal{V}_i^- + \sum_{i\in K^-} \boldsymbol{u}_i E_i - \boldsymbol{u}_{n_p} \sum_{i\in K^-} E_i + \sum_{i\in K^-} \boldsymbol{u}_i \mathring{\mathcal{V}}_i^- \\
&= \sum_{j\in K^-} \boldsymbol{u}_j \mathcal{V}_j - \boldsymbol{u}_{n_p} \sum_{j\in K^-} \mathcal{V}_j + \boldsymbol{u}_{n_p} \sum_{i\in K^-} \mathcal{V}_i^- + \sum_{i\in K^-} \boldsymbol{u}_i \mathring{\mathcal{V}}_i^- \\
&= \sum_{j\in K^-} \boldsymbol{u}_j \left(\mathcal{V}_j + \mathring{\mathcal{V}}_j^-\right) + \boldsymbol{u}_{n_p} \sum_{i\in K^-} \left(\mathcal{V}_i^- - \mathcal{V}_i\right) \\
&= \sum_{j\in K^-} \boldsymbol{u}_j V_j + \boldsymbol{u}_{n_p} V_{n_p} \\
&= \sum_{j\in K} \boldsymbol{u}_j V_j \ .
\end{aligned}$$

Finally,

$$\begin{aligned}
\sum_{i=1}^{n_p-1} \boldsymbol{u}_i^- V_i^- &= \sum_{i\in K^-} \boldsymbol{u}_i^- V_i^- + \sum_{i\notin K^-} \boldsymbol{u}_i^- V_i^- \\
&= \sum_{i\in K} \boldsymbol{u}_i V_i + \sum_{i\notin K} \boldsymbol{u}_i V_i = \sum_{i=1}^{n_p} \boldsymbol{u}_i V_i
\end{aligned}$$

and hence, (2.47) holds.

□

Remark 2.55

Both schemes, (2.51)-(2.52) and (2.53)-(2.54), will be equivalent, if we can compute integrals exactly. If not, the scheme (2.53)-(2.54) satisfies the properties (2.46) and (2.47) (up to the machine precision), but the method (2.51)-(2.52) does not.

The choice of ω and K can be made in the same way as in the section 2.3. Numerical examples will be shown in chapter 5.

3 Polyharmonic spline interpolation and the WENO method

To increase the order of accuracy of a scheme, it is usually necessary to construct a high order reconstruction of the solution at each time step. This is true for the finite volume method, for which many results can be found in the literature, as well as for the finite volume particle method. In this chapter we will treat the problem of reconstruction of a function from given data, particularly we will look for an approximation of a function from given weighted integral means of this function (FVPM), which is the generalization of classical integral means appearing in FVM. The obtained knowledge will be utilized in chapter 4 to acquire initial data for local generalized Riemann problems in order to solve a one-dimensional hyperbolic conservation law using FVPM. Therefore, we introduce the scattered data interpolation problem, which will be solved with the polyharmonic spline interpolation and the WENO procedure. The combination of these two methods gives rise to a powerful method to reconstruct a function from given data, proposed by Aboiyar, Georgoulis and Iske [1]. WENO reconstruction by polyharmonic splines is numerically stable, if carefully implemented, and in comparison with the polynomial reconstruction more flexible. Moreover, it reproduces optimal reconstruction with respect to the seminorm in the Beppo-Levi space, and one therefore acquires a natural choice for an oscillation indicator.

3.1 Polyharmonic spline interpolation

The polyharmonic spline interpolation is a popular method in the area of *scattered data interpolation*. Scattered data interpolation problems can be stated in the following way: A data set of an unknown function is given. This data set is typically formed as a set of function values at some given points (*Lagrange interpolation*). One is interested in the interpolation of the data, yielding an approximation of the unknown function. The approach, which we are interested in, is a *kernel-based interpolation* - the polyharmonic spline interpolation. The advantages of this method in combination with the WENO method will be explained throughout this chapter in more detail.
The interpolation can be understood in two ways - the global and the local. For the global interpolation, one increases the amount of interpolated points and expects a convergence at some rate to the sought function. More details can be found in Wendland [69].
We will focus on the local interpolation, defined later in this section, since this approach is suitable to define and analyse a high order FVPM, as done in chapter 4.
The usual function value data set can be also substituted with another data given by a linear functional, e.g., with a data set defined by integral means (*cell averages* in the FVM framework). For the purposes of FVPM, we will consider the case of data given by *weighted integral means*.
In this section, we will follow the work by Iske [26], where the polyharmonic spline interpolation is described for the case of data set given by function values and also the numerical stability is treated, see also [27]. This approach was later extended to the case of data set given by classical integral means of a function, see e.g., [1] or [2], where also the numerical stability for the computation of derivatives was investigated. In this thesis, we will introduce the polyharmonic spline interpolation based on the data set defined by weighted integral means in order to utilize the data

available in the framework of FVPM. Such interpolation is used then in chapter 4 to define a higher order meshfree method for the solution of hyperbolic conservation laws.
We adopt the introduction and the proof techniques given by Iske [26] and adapt them to our problem. The known results for data set given by function values are extended to the case of weighted integral means in lemma 3.2 and theorem 3.3. Moreover, in the theorem 3.3 we introduce the leading error term explicitly, which allows us to analyse the numerical methods in the following chapter 4.

The interpolation problem

In the beginning we define the technical background of the interpolation problem with data given by weighted integral means. We refer to the chapter 2 for the motivation of this paragraph. In that chapter, the basis functions are denoted by ψ_i, whereas we will denote them by $\psi_{\mathbf{x}_i}$ in this chapter since this notation is more suitable for our analysis.

Consider an open and bounded domain $\Omega \subset \mathbb{R}^d$. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \subset \overline{\Omega}$, $d \geq 1$ denote a given scattered point set. For these points construct non-negative and Lipschitz-continuous functions with a compact support

$$\psi_{\mathbf{x}_i} : \mathbb{R}^d \to \mathbb{R} \ , \tag{3.1}$$

such that the family $\{\psi_{\mathbf{x}_i}\}_{\mathbf{x}_i \in X}$ is a partition of unity on $\overline{\Omega} \subset \mathbb{R}^d$, i.e.,

$$\sum_{i=1}^{n} \psi_{\mathbf{x}_i}(\mathbf{x}) = 1 \quad \forall\, \mathbf{x} \in \overline{\Omega} \ .$$

We define the *volumes* by

$$V_{\mathbf{x}_i} := \int\limits_{\operatorname{supp}\, \psi_{\mathbf{x}_i}} \psi_{\mathbf{x}_i}(\mathbf{x}) d\mathbf{x} \ , \quad i = 1, \ldots, n \ .$$

Further, we define a linear functional $\lambda_{\mathbf{x}_i}$ through

$$\lambda_{\mathbf{x}_i}(f) := \frac{1}{V_{\mathbf{x}_i}} \int\limits_{\operatorname{supp}\, \psi_{\mathbf{x}_i}} f(\mathbf{x}) \psi_{\mathbf{x}_i}(\mathbf{x}) d\mathbf{x} \ , \quad i = 1, \ldots, n \tag{3.2}$$

for all suitable functions f, e.g., $f \in L^1_{loc}(\mathbb{R}^d)$.
The symbol λ_X will denote the set of functionals $\lambda_X = \{\lambda_{\mathbf{x}_1}, \ldots, \lambda_{\mathbf{x}_n}\}$.

Now the following interpolation problem can be defined.
For an unknown function $f : \mathbb{R}^d \to \mathbb{R}$, a data vector $f\Big|_{\lambda_X} = (\lambda_{\mathbf{x}_1}(f), \ldots, \lambda_{\mathbf{x}_n}(f))^T \in \mathbb{R}^n$ is given.
The interpolation problem reads: Find an interpolant $s : \mathbb{R}^d \to \mathbb{R}$, s.t. $s\Big|_{\lambda_X} = f\Big|_{\lambda_X}$, i.e.,

$$\lambda_{\mathbf{x}_i}(s) = \lambda_{\mathbf{x}_i}(f) \ , \quad i = 1, \ldots, n \tag{3.3}$$

and $s \in \mathcal{M}$, where $\mathcal{M}$ is a suitable function space.
This problem can be solved in various manners. We will focus on polyharmonic spline interpolation, a kind of kernel-based interpolation.

Polyharmonic spline interpolation

To solve the problem (3.3) with a kernel-based interpolation, we look for an interpolant s of the form

$$s(\mathbf{x}) = \sum_{j=1}^{n} c_j \lambda^{\mathbf{y}}_{\mathbf{x}_j} \phi(\|\mathbf{x} - \mathbf{y}\|) + p(\mathbf{x}) \ , \quad p \in \mathcal{P}^d_m \ , \tag{3.4}$$

where $\phi : [0, \infty) \to \mathbb{R}$ is a fixed *radial basis function*, $\|.\|$ is the Euclidean norm on $\mathbb{R}^d$, and where $\mathcal{P}_m^d$ is the linear space of all d-variate polynomials of order at most m (i.e., of degree at most $m-1$). The dimension of $\mathcal{P}_m^d$ is $q = \dim \mathcal{P}_m^d = \binom{m-1+d}{d}$.
$\lambda_{\mathbf{x}_j}^{\mathbf{y}}$ denotes the action of the linear functional $\lambda_{\mathbf{x}_j}$ with respect to the variable $\mathbf{y}$,

$$\lambda_{\mathbf{x}_j}^{\mathbf{y}} \phi(\|\mathbf{x}-\mathbf{y}\|) := \frac{1}{V_{\mathbf{x}_j}} \int\limits_{\text{supp } \psi_{\mathbf{x}_j}} \phi(\|\mathbf{x}-\mathbf{y}\|)\psi_{\mathbf{x}_j}(\mathbf{y})d\mathbf{y} \ , \quad j = 1, \ldots, n \ .$$

The order m of $p \in \mathcal{P}_m^d$ is given by the order $m \equiv m(\phi)$ of the radial basis function ϕ. For possible choices of radial basis functions ϕ and further details see [1] or [69] and references therein. From now on, we will work only with polyharmonic splines. A good summary on the advantages of polyharmonic splines compared to another radial kernel functions, such as the numerical stability or arbitrary local approximation order, can be found in [27]. We also mention that one of the properties of polyharmonic splines is the reproduction of polynomials.

Polyharmonic spline interpolation is based on the choice of the radial basis function of the form

$$\phi_{d,k}(r) = \begin{cases} r^{2k-d}\log(r) & , \ d \text{ even}, \\ r^{2k-d} & , \ d \text{ odd}, \end{cases}$$

where $2k > d$, $k \in I\!N$. The order of the conditionally positive (negative) definite function $\phi_{d,k}$ is $m = k - \lceil d/2 \rceil + 1$. For more details about conditionally positive definite functions, see [44] and [49].

To find the reconstruction s we have to determine n parameters for the radial basis functions and q parameters for the polynomial part, altogether $n+q$ parameters. The interpolation conditions (3.3) provide n conditions. We get the remaining q conditions by considering linear constraints

$$\sum_{j=1}^{n} c_j \lambda_{\mathbf{x}_j}(p) = 0 \quad \forall p \in \mathcal{P}_m^d \ .$$

The latter constraints have their origin in the theory of conditionally positive definite functions. For details, see e.g., [69].
Under this consideration we have to solve a linear system of size $(n+q) \times (n+q)$

$$\begin{bmatrix} A & P \\ P^T & 0 \end{bmatrix} \cdot \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} f\Big|_{\lambda_X} \\ 0 \end{bmatrix} \ , \tag{3.5}$$

where

$$\begin{aligned} A &= \Big(\lambda_{\mathbf{x}_i}^{\mathbf{x}} \lambda_{\mathbf{x}_j}^{\mathbf{y}} \phi_{d,k}(\|\mathbf{x}-\mathbf{y}\|)\Big)_{1 \le i,j \le n} \in \mathbb{R}^{n \times n} \ , \\ P &= (\lambda_{\mathbf{x}_i}(\mathbf{x}^\alpha))_{1 \le i \le n; |\alpha| < m} \in \mathbb{R}^{n \times q} \ , \\ f\Big|_{\lambda_X} &= (\lambda_{\mathbf{x}_i}(f))_{1 \le i \le n} \in \mathbb{R}^n \ , \end{aligned}$$

for the vectors of unknows $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ and $d = (d_\alpha)_{|\alpha|<m} \in \mathbb{R}^q$. This linear system has always a solution, which is unique, provided that the set of functionals λ_X is $\mathcal{P}_m^d$-*unisolvent*, which is equivalent to requiring that there is no nontrivial polynomial in $\mathcal{P}_m^d$ that vanishes on all functionals from λ_X, i.e., it has to hold for $p \in \mathcal{P}_m^d$

$$\lambda_{\mathbf{x}_j}(p) = 0 \quad \forall j = 1, \ldots, n \quad \Rightarrow \quad p \equiv 0 \ .$$

The unique solvability is proven in [69] which utilizes the theory of conditionally positive definite functions. The proof can be performed in the same way as in the mentioned book by considering the functionals (3.2) instead of function point values.

The local interpolation problem

Consider a fixed point $\mathbf{x}_0 \in \mathbb{R}^d$ and a $\mathcal{P}_m^d$-unisolvent set of functionals λ_X. Let $h > 0$. Furthermore, we denote local volumes

$$V_{h\mathbf{x}_i} := \int\limits_{\text{supp } \psi_{h\mathbf{x}_i}} \psi_{h\mathbf{x}_i}(\mathbf{x}) d\mathbf{x}$$

and local functionals

$$\lambda_{h\mathbf{x}_i}(f) := \frac{1}{V_{h\mathbf{x}_i}} \int\limits_{\text{supp } \psi_{h\mathbf{x}_i}} f(\mathbf{x}) \psi_{h\mathbf{x}_i}(\mathbf{x}) d\mathbf{x} \ ,$$

where we use a shortened notation for the argument $f = f(h\mathbf{x})$ of the functional $\lambda_{h\mathbf{x}_i}$, i.e., $\lambda_{h\mathbf{x}_i}(f)$ stands for $\lambda_{h\mathbf{x}_i}(f) = \lambda_{h\mathbf{x}_i}(f(h\mathbf{x}))$. This notation will be used throughout this section for the sake of notational simplicity.

Symbol $\psi_{h\mathbf{x}_i}$ denotes the scaled function (3.1) with respect to its center $\mathbf{x}_i$ with the scaling parameter $h > 0$, i.e.,

$$\psi_{h\mathbf{x}_i}(h\mathbf{x}) = \psi_{\mathbf{x}_i}(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^d \ .$$

See the figure 3.1 for an example on scaling.

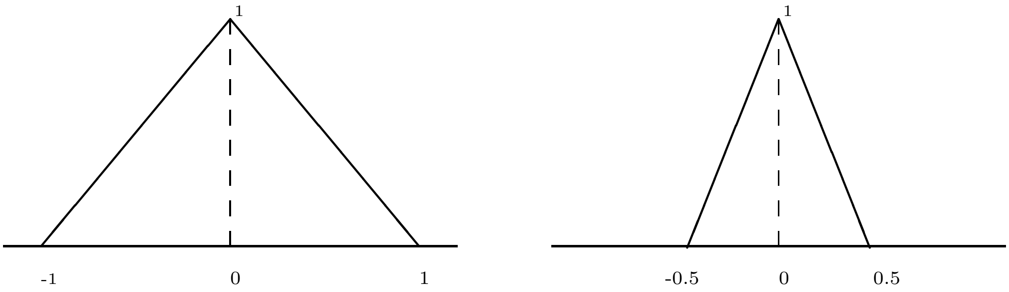


Figure 3.1: *1-d example of scaling a function ψ_{x_i} into ψ_{hx_i}. Function ψ_{x_i} is depicted on the left, function ψ_{hx_i} on the right. Values of h and x_i are chosen to be $h = \frac{1}{2}$, $x_i = 0$.*

The local interpolation problem reads: Find an interpolant $s^h : \mathbb{R}^d \to \mathbb{R}$ in a local neighborhood $U_h(\mathbf{x}_0)$ of $\mathbf{x}_0$, such that

$$\lambda_{h\mathbf{x}_i}\left(s^h(\mathbf{x}_0 + \cdot)\right) = \lambda_{h\mathbf{x}_i}\left(f(\mathbf{x}_0 + \cdot)\right) \ , \quad i = 1, \ldots, n \ . \tag{3.6}$$

For this local interpolation problem the asymptotic bound of the form

$$|s^h(\mathbf{x}_0 + h\mathbf{x}) - f(\mathbf{x}_0 + h\mathbf{x})| = \mathcal{O}(h^p) \ , \ h \to 0$$

is of our interest. The number p is said to be the *approximation order* at x_0.

Since the polyharmonic spline interpolation is shift-invariant, we assume from now on $\mathbf{x}_0 = \mathbf{0}$ without loss of generality. Under this assumption (3.6) becomes

$$\lambda_{h\mathbf{x}_i}(s^h) = \lambda_{h\mathbf{x}_i}(f) \ , \quad i = 1, \ldots, n \ , \tag{3.7}$$

which leads to the next definition.

Approximation order

Definition 3.1
Let s^h denote the polyharmonic spline interpolant, using $\phi_{d,k}$ and satisfying (3.7). We say that the approximation order *of local polyharmonic spline interpolation with respect to the function space $\mathcal{F}$ is p, if for any $f \in \mathcal{F}$ the asymptotic bound*

$$|s^h(h\boldsymbol{x}) - f(h\boldsymbol{x})| = \mathcal{O}(h^p) \ , \ h \to 0$$

holds for any $\boldsymbol{x} \in \mathbb{R}^d$ and any finite $\mathcal{P}_m^d$-unisolvent set of functionals λ_X.

As already explained, the interpolation problem (3.7) for any $h > 0$ and a fixed $\mathcal{P}_m^d$-unisolvent set of functionals λ_X has under the constraints

$$\sum_{j=1}^{n} c_j^h \lambda_{h\mathbf{x}_j}(p) = 0 \quad \forall p \in \mathcal{P}_m^d \tag{3.8}$$

a unique solution s^h of the form

$$s^h(h\mathbf{x}) = \sum_{j=1}^{n} c_j^h \lambda_{h\mathbf{x}_j}^{h\mathbf{y}} \phi_{d,k}(\|h\mathbf{x} - h\mathbf{y}\|) + \sum_{|\alpha|<m} d_\alpha^h (h\mathbf{x})^\alpha \ , \tag{3.9}$$

where the coefficients c_j^h and d_α^h solve the linear system (3.10).
Symbol $\lambda_{h\mathbf{x}_j}^{h\mathbf{y}}$ denotes the action of the linear functional $\lambda_{h\mathbf{x}_j}$ w.r.t. the variable $h\mathbf{y}$,

$$\lambda_{h\mathbf{x}_j}^{h\mathbf{y}} \phi(\|h\mathbf{x} - h\mathbf{y}\|) := \frac{1}{V_{h\mathbf{x}_j}} \int\limits_{\text{supp } \psi_{h\mathbf{x}_j}} \phi(\|h\mathbf{x} - \mathbf{y}\|) \psi_{h\mathbf{x}_j}(\mathbf{y}) d\mathbf{y} \ .$$

The conditions (3.7) and (3.8) can be rewritten as a linear system for coefficients $c^h = (c_1^h, \dots, c_n^h)^T \in \mathbb{R}^n$ and $d^h = (d_\alpha^h)_{|\alpha|<m} \in \mathbb{R}^q$ of the form

$$\begin{bmatrix} \Phi_h & \Pi_h \\ \Pi_h^T & 0 \end{bmatrix} \cdot \begin{bmatrix} c^h \\ d^h \end{bmatrix} = \begin{bmatrix} f\Big|_{\lambda_{hX}} \\ 0 \end{bmatrix} \ , \tag{3.10}$$

where

$$\begin{aligned} \Phi_h &= \left(\lambda_{h\mathbf{x}_i}^{h\mathbf{x}} \lambda_{h\mathbf{x}_j}^{h\mathbf{y}} \phi_{d,k}(\|h\mathbf{x} - h\mathbf{y}\|) \right)_{1 \le i,j \le n} \in \mathbb{R}^{n \times n} \ , \\ \Pi_h &= \left(\lambda_{h\mathbf{x}_i}((h\mathbf{x})^\alpha) \right)_{1 \le i \le n; |\alpha|<m} \in \mathbb{R}^{n \times q} \ , \\ f\Big|_{\lambda_{hX}} &= \left(\lambda_{h\mathbf{x}_i}(f) \right)_{1 \le i \le n} \in \mathbb{R}^n \ . \end{aligned}$$

If we denote

$$A_h = \begin{bmatrix} \Phi_h & \Pi_h \\ \Pi_h^T & 0 \end{bmatrix} \quad , \quad b^h = \begin{bmatrix} c^h \\ d^h \end{bmatrix} \quad \text{and} \quad f_h = \begin{bmatrix} f\Big|_{\lambda_{hX}} \\ 0 \end{bmatrix} \ ,$$

the linear system (3.10) can be rewritten as

$$A_h \cdot b^h = f_h \ .$$

Recall that the *Lagrange representation* of the interpolant s^h in (3.9) given by

$$s^h(h\mathbf{x}) = \sum_{i=1}^{n} L_i^h(h\mathbf{x}) \lambda_{h\mathbf{x}_i}(f) \tag{3.11}$$

with the *Lagrange basis functions* L_i^h satisfying

$$\lambda_{h\mathbf{x}_j}(L_i^h) = \delta_{ij} \ ,$$

where δ_{ij} stands for the *Kronecker delta.*
Moreover, due to the reproduction of polynomials from $\mathcal{P}_m^d$, it holds

$$\sum_{i=1}^{n} L_i^h(h\mathbf{x})\lambda_{h\mathbf{x}_i}(p) = p(h\mathbf{x}) \quad \forall p \in \mathcal{P}_m^d \ .$$

We can construct the Lagrange functions pointwise at any $h\mathbf{x}$ by solving the linear system

$$\begin{bmatrix} \Phi_h & \Pi_h \\ \Pi_h^T & 0 \end{bmatrix} \cdot \begin{bmatrix} L^h(h\mathbf{x}) \\ \mu^h(h\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \varphi_h(h\mathbf{x}) \\ \pi_h(h\mathbf{x}) \end{bmatrix} \ , \tag{3.12}$$

where $L^h(h\mathbf{x}) = (L_i^h(h\mathbf{x}))_{1\leq i\leq n} \in \mathbb{R}^n$ is the vector of point value evaluations of the Lagrange functions at $h\mathbf{x}$ and $\mu^h(h\mathbf{x}) = \left(\mu_\alpha^h(h\mathbf{x})\right)_{|\alpha|<m} \in \mathbb{R}^q$. The right hand side consists of

$$\begin{aligned} \varphi_h(h\mathbf{x}) &= \left(\lambda_{h\mathbf{x}_j}^{h\mathbf{y}} \phi_{d,k}(\|h\mathbf{x} - h\mathbf{y}\|)\right)_{1\leq j\leq n} \in \mathbb{R}^n \ , \\ \pi_h(h\mathbf{x}) &= \left((h\mathbf{x})^\alpha\right)_{|\alpha|<m} \in \mathbb{R}^q \ . \end{aligned}$$

By denoting

$$\nu^h(h\mathbf{x}) = \begin{bmatrix} L^h(h\mathbf{x}) \\ \mu^h(h\mathbf{x}) \end{bmatrix} \quad \text{and} \quad \beta_h(h\mathbf{x}) = \begin{bmatrix} \varphi_h(h\mathbf{x}) \\ \pi_h(h\mathbf{x}) \end{bmatrix} \ ,$$

we can abbreviate the system (3.12) by

$$A_h \cdot \nu^h(h\mathbf{x}) = \beta_h(h\mathbf{x}) \ .$$

Let $\langle \cdot, \cdot \rangle$ denote the inner product of the Euclidean space $\mathbb{R}^p$ for an appropriate $p \in I\!N$. The following computation shows the equivalence of the Lagrange representation (3.11) and the standard representation (3.9) of the interpolant s^h

$$\begin{aligned} s^h(h\mathbf{x}) &= \left\langle L^h(h\mathbf{x}), f\Big|_{\lambda_{hX}} \right\rangle \\ &= \left\langle \nu^h(h\mathbf{x}), f_h \right\rangle \\ &= \left\langle A_h^{-1} \ . \ \beta_h(h\mathbf{x}), f_h \right\rangle \\ &= \left\langle \beta_h(h\mathbf{x}), A_h^{-1} \ . \ f_h \right\rangle \\ &= \left\langle \beta_h(h\mathbf{x}), b^h \right\rangle \ . \end{aligned}$$

The following lemma and theorem are based on the results from [26] but we formulate them for the case of data set given by weighted integral means.

Lemma 3.2
The Lagrange basis functions of polyharmonic spline interpolation are invariant under uniform scalings, i.e., for any $h > 0$, we have

$$L^h(h\boldsymbol{x}) = L^1(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathbb{R}^d \ .$$

Proof.
Let

$$\mathcal{S}_h = \left\{ \sum_{j=1}^{n} c_j \lambda_{h\mathbf{x}_j}^{h\mathbf{y}} \phi(\| \cdot - h\mathbf{y}\|) + p \ : \ p \in \mathcal{P}_m^d \ , \ \sum_{j=1}^{n} c_j \lambda_{h\mathbf{x}_j}(q) = 0 \quad \forall q \in \mathcal{P}_m^d \right\} \ , \ h > 0$$

denote the space of all possible polyharmonic spline interpolants of the form (3.9) satisfying (3.8). We will show that $\mathcal{S}_h$ is a scaled version of $\mathcal{S}_1$, i.e., $\mathcal{S}_h = \{\sigma_h(s) : s \in \mathcal{S}_1\}$, where the dilatation operator is given by $\sigma_h(s) = s(\cdot/h)$. From this follows, due to the uniqueness of the interpolation in either space, $\mathcal{S}_h$ or $\mathcal{S}_1$, that their Lagrange basis functions satisfy $L^h = \sigma_h(L^1)$, which is the statement of this lemma.
We want to show $\mathcal{S}_h = \{\sigma_h(s) : s \in \mathcal{S}_1\}$. To this end, we distinguish two cases - the space dimension d is odd and d is even.
For d odd, substituting $\mathbf{y} = h\mathbf{z}$ we have

$$V_{h\mathbf{x}_j} = \int\limits_{\text{supp }\psi_{h\mathbf{x}_j}} \psi_{h\mathbf{x}_j}(\mathbf{y})d\mathbf{y} = h^d \int\limits_{\text{supp }\psi_{\mathbf{x}_j}} \psi_{h\mathbf{x}_j}(h\mathbf{z})d\mathbf{z} = h^d \int\limits_{\text{supp }\psi_{\mathbf{x}_j}} \psi_{\mathbf{x}_j}(\mathbf{z})d\mathbf{z} = h^d V_{\mathbf{x}_j}$$

and

$$\begin{aligned} \lambda^{h\mathbf{y}}_{h\mathbf{x}_j}\phi_{d,k}(\|h\mathbf{x}-h\mathbf{y}\|) &= \frac{1}{V_{h\mathbf{x}_j}} \int\limits_{\text{supp }\psi_{h\mathbf{x}_j}} \|h\mathbf{x}-\mathbf{y}\|^{2k-d}\psi_{h\mathbf{x}_j}(\mathbf{y})d\mathbf{y} \\ &= \frac{1}{h^d V_{\mathbf{x}_j}}\, h^d \int\limits_{\text{supp }\psi_{\mathbf{x}_j}} \|h\mathbf{x}-h\mathbf{z}\|^{2k-d}\psi_{h\mathbf{x}_j}(h\mathbf{z})d\mathbf{z} \\ &= \frac{h^{2k-d}}{V_{\mathbf{x}_j}} \int\limits_{\text{supp }\psi_{\mathbf{x}_j}} \|\mathbf{x}-\mathbf{z}\|^{2k-d}\psi_{\mathbf{x}_j}(\mathbf{z})d\mathbf{z} \\ &= h^{2k-d}\lambda^{\mathbf{z}}_{\mathbf{x}_j}\phi_{d,k}(\|\mathbf{x}-\mathbf{z}\|)\ . \end{aligned}$$

This gives immediately $\mathcal{S}_h = \{\sigma_h(s) : s \in \mathcal{S}_1\}$, since also

$$0 = \sum_{j=1}^n c_j \lambda_{h\mathbf{x}_j}(q) = \sum_{j=1}^n c_j \lambda_{\mathbf{x}_j}(\widetilde{q})$$

for $q \in \mathcal{P}_m^d$ and $\widetilde{q}(\mathbf{x}) = q(h\mathbf{x})$ (hence, $\widetilde{q} \in \mathcal{P}_m^d$).

Now suppose that d is even. We use again the substitution $\mathbf{y} = h\mathbf{z}$

$$\begin{aligned} \lambda^{h\mathbf{y}}_{h\mathbf{x}_j}\phi_{d,k}(\|h\mathbf{x}-h\mathbf{y}\|) &= \frac{1}{V_{h\mathbf{x}_j}} \int\limits_{\text{supp }\psi_{h\mathbf{x}_j}} \|h\mathbf{x}-\mathbf{y}\|^{2k-d}\log(\|h\mathbf{x}-\mathbf{y}\|)\psi_{h\mathbf{x}_j}(\mathbf{y})d\mathbf{y} \\ &= \frac{h^{2k-d}}{V_{\mathbf{x}_j}} \int\limits_{\text{supp }\psi_{\mathbf{x}_j}} \|\mathbf{x}-\mathbf{z}\|^{2k-d}\log(h\|\mathbf{x}-\mathbf{z}\|)\psi_{\mathbf{x}_j}(\mathbf{z})d\mathbf{z} \\ &= h^{2k-d}\Bigg(\lambda^{\mathbf{z}}_{\mathbf{x}_j}\phi_{d,k}(\|\mathbf{x}-\mathbf{z}\|) \\ &\quad + \log(h)\frac{1}{V_{\mathbf{x}_j}} \int\limits_{\text{supp }\psi_{\mathbf{x}_j}} \|\mathbf{x}-\mathbf{z}\|^{2k-d}\psi_{\mathbf{x}_j}(\mathbf{z})d\mathbf{z}\Bigg)\ . \end{aligned}$$

Therefore, any function $s^h \in \mathcal{S}_h$ has the form

$$s^h(h\mathbf{x}) = h^{2k-d}\left(\sum_{j=1}^n c_j \lambda^{\mathbf{y}}_{\mathbf{x}_j}\phi(\|\mathbf{x}-\mathbf{y}\|) + \log(h)r(\mathbf{x})\right) + p(\mathbf{x})$$

for some $p \in \mathcal{P}_m^d$ and where

$$r(\mathbf{x}) = \sum_{j=1}^n c_j \frac{1}{V_{\mathbf{x}_j}} \int\limits_{\text{supp }\psi_{\mathbf{x}_j}} \|\mathbf{x}-\mathbf{y}\|^{2k-d}\psi_{\mathbf{x}_j}(\mathbf{y})d\mathbf{y}$$

is a polynomial (due to the fact that d is even) of degree at most $m-1$, which can be seen from another form of r

$$\begin{aligned} r(\mathbf{x}) &= \sum_{j=1}^{n} c_j \frac{1}{V_{\mathbf{x}_j}} \sum_{|\alpha|+|\beta|=2k-d} c_{\alpha,\beta}\mathbf{x}^\alpha \int\limits_{\text{supp } \psi_{\mathbf{x}_j}} \mathbf{y}^\beta \psi_{\mathbf{x}_j}(\mathbf{y}) d\mathbf{y} \\ &= \sum_{|\alpha|+|\beta|=2k-d} c_{\alpha,\beta}\mathbf{x}^\alpha \sum_{j=1}^{n} c_j \lambda_{\mathbf{x}_j}^{\mathbf{y}}(\mathbf{y}^\beta) \end{aligned}$$

for some coefficients $c_{\alpha,\beta} \in \mathbb{R}$ with $|\alpha| + |\beta| = 2k - d$. Due to the vanishing moment conditions (3.8) for the coefficients $c_1, \dots, c_n$, the degree of r is at most $2k - d - m = k - d/2 - 1 < m$. Therefore, $s^h \in \sigma_h(\mathcal{S}_1)$, and so $\mathcal{S}_h \subset \sigma_h(\mathcal{S}_1)$.
The inclusion $\mathcal{S}_1 \subset \sigma_h^{-1}(\mathcal{S}_h)$ can be proven accordingly.
From that follows, $\mathcal{S}_h = \sigma_h(\mathcal{S}_1)$, and therefore $L^h(h\mathbf{x}) = L^1(\mathbf{x})$. □

The scale-invariance of the Lagrange basis functions of polyharmonic spline reconstruction spaces has several important corollaries. Firstly, the numerical stability of the reconstruction can be analysed and a preconditioning strategy can be developed. References can be found at the end of this chapter. Secondly, arbitrary local approximation order of the approximation by polyharmonic splines can be investigated and the main result of this section can be stated.

Theorem 3.3
The approximation order of local polyharmonic spline interpolation, using $\phi_{d,k}$, with respect to $C^m(\Omega)$, $\Omega \subset \mathbb{R}^d$, is $m = k - \lceil d/2 \rceil + 1$, i.e.,

$$|s^h(h\boldsymbol{x}) - f(h\boldsymbol{x})| = \mathcal{O}(h^m) \ , \ h \to 0 \ .$$

Moreover, if $f \in C^{m+1}(\Omega)$, $\Omega \subset \mathbb{R}^d$, then

$$s^h(h\boldsymbol{x}) = f(h\boldsymbol{x}) - \sum_{i=1}^{n} L_i^h(h\boldsymbol{x}) \sum_{|\alpha|=m} \frac{1}{\alpha!} D^\alpha f(h\boldsymbol{x})(h\boldsymbol{x}_i - h\boldsymbol{x})^\alpha + \mathcal{O}(h^{m+1}) \ , \ h \to 0 \ .$$

Proof.
Let $h > 0$ and $\mathbf{x} \in \mathbb{R}^d$ be fixed.
We prove the first statement. The m-th order Taylor polynomial of $f \in C^m$ around $h\mathbf{x}$ reads

$$T^m_{f,h\mathbf{x}}(\mathbf{y}) = \sum_{|\alpha|<m} \frac{1}{\alpha!} D^\alpha f(h\mathbf{x})(\mathbf{y} - h\mathbf{x})^\alpha \ .$$

Using the Lagrange representation of s^h and the polynomial reproduction property, we acquire

$$f(h\mathbf{x}) - s^h(h\mathbf{x}) = \sum_{i=1}^{n} L_i^h(h\mathbf{x}) \left[T^m_{f,h\mathbf{x}}(h\mathbf{x}_i) - f(h\mathbf{x}_i)\right] \ .$$

Due to the lemma 3.2, the *Lagrange function*

$$\Lambda(\mathbf{x}) := \sum_{i=1}^{n} |L_i^h(h\mathbf{x})| = \sum_{i=1}^{n} |L_i^1(\mathbf{x})|$$

is uniformly bounded in any local neighborhood of the origin. Since

$$T^m_{f,h\mathbf{x}}(h\mathbf{x}_i) - f(h\mathbf{x}_i) = \mathcal{O}(h^m) \ , \ h \to 0 \quad \forall \ 1 \le i \le n \ ,$$

this implies

$$|s^h(h\mathbf{x}) - f(h\mathbf{x})| = \mathcal{O}(h^m) \ , \ h \to 0 \ .$$

Now, if $f \in C^{m+1}$, then

$$T^m_{f,h\mathbf{x}}(h\mathbf{x}_i) - f(h\mathbf{x}_i) = \sum_{|\alpha|=m} \frac{1}{\alpha!} D^\alpha f(h\mathbf{x})(h\mathbf{x}_i - h\mathbf{x})^\alpha + \mathcal{O}(h^{m+1}) \ , \ h \to 0 \ ,$$

which yields

$$s^h(h\mathbf{x}) = f(h\mathbf{x}) - \sum_{i=1}^n L^h_i(h\mathbf{x}) \sum_{|\alpha|=m} \frac{1}{\alpha!} D^\alpha f(h\mathbf{x})(h\mathbf{x}_i - h\mathbf{x})^\alpha + \mathcal{O}(h^{m+1}) \ , \ h \to 0$$

due to the boundedness of the Lagrange function Λ. □

Optimal reconstruction

Having given the radial basis function $\phi_{d,k}$ with $2k > d$, one can introduce the Beppo-Levi space

$$BL_k(\mathbb{R}^d) := \{v : D^\gamma v \in L^2(\mathbb{R}^d) \ \forall \ |\gamma| = k\} \subset \mathcal{C}(\mathbb{R}^d)$$

equipped with the seminorm $|\cdot|_{BL_k(\mathbb{R}^d)}$ defined by

$$|\cdot|^2_{BL_k(\mathbb{R}^d)} := \sum_{|\gamma|=k} \binom{k}{\gamma} \|D^\gamma v\|^2_{L^2(\mathbb{R}^d)} \ .$$

Based on the work of Duchon in [11], [12] and [13] and presented in [1], [2] and [69], one can show also for data given by the weighted integral means the following: The interpolant $s \in BL_k(\mathbb{R}^d)$, given by (3.4) with a fixed polyharmonic spline kernel $\phi_{d,k}$ satisfying (3.3), is the unique minimiser of the energy $|\cdot|_{BL_k(\mathbb{R}^d)}$ among all interpolants $v \in BL_k(\mathbb{R}^d)$ satisfying $v\Big|_{\lambda_X} = f\Big|_{\lambda_X}$, i.e.,

$$|s|_{BL_k(\mathbb{R}^d)} \leq |v|_{BL_k(\mathbb{R}^d)} \qquad \forall \ v \in BL_k(\mathbb{R}^d) \text{ with } v\Big|_{\lambda_X} = f\Big|_{\lambda_X} \ .$$

In other words, one gets the optimal reconstruction in the Beppo-Levi space $BL_k(\mathbb{R}^d)$. This property allows a natural choice of oscillation indicator, which is the topic of following section.

3.2 WENO method

Let us first consider the framework of the finite volume method (FVM). The following results are straightforward portable to the framework of FVPM.

While solving hyperbolic conservation laws, rapidly changing solutions or solutions with discontinuities may arise. In many methods, these exact solutions are approximated with appropriate functions to achieve higher order of accuracy of the scheme. This may however lead to non-physical oscillations, having origin in the chosen numerical approximation. Therefore new techniques have been developed to avoid these oscillations. We will focus on a WENO (Weighted Essentially Non-Oscillatory) scheme, having its origin in ENO (Essentially Non-Oscillatory) schemes.

The ENO scheme for one-dimensional conservation laws was first proposed by Harten, Engquist, Osher and Chakravarthy in [21]. In the ENO scheme, for each cell of the finite volume discretization, a set of stencils (a set of neighboring cells) is chosen. On each stencil, a reconstruction function based on the data of the stencil is computed. Afterwards, the smoothness of these reconstructions is measured by introducing a suitable oscillation indicator. Finally, the smoothest reconstruction

is selected among all the stencils for the given cell and used in further computations as the best possible approximation.
WENO scheme represents an improvement of the ENO method. First proposals were made by Liu, Osher and Chan in [41] and by Jiang and Shu in [28]. In the ENO scheme, only the smoothest reconstruction is used and the reconstructions built on the remaining stencils are dropped. On the contrary, the WENO reconstruction takes all the reconstructions into account by constructing a convex combination of them. The weights are based on some oscillation indicator, e.g., on that of the ENO scheme. In the resulting scheme, spurious oscillations are avoided.
There are many other works concerning the WENO method, let us mention e.g., [15], [25] or [52].

Let us get back to FVPM. In section 3.1 the interpolation of given data was considered. Let us consider the framework of this previous section but let us look at the interpolation on a stencil. We say, that two particles $\mathbf{x}_i$ and $\mathbf{x}_j$ are *neighbors* if the supports of their corresponding functions $\psi_{\mathbf{x}_i}$ and $\psi_{\mathbf{x}_j}$ overlap.
For a fixed particle $\mathbf{x}_{i_0}$ and its index i_0 we define a *stencil* $\mathcal{S}$ of size n_s as a set of n_s arbitrary indices of neighboring particles $\mathbf{x}_i$, $i \in \{1, \dots, n\}$, such that $i_0 \in \mathcal{S}$.
We also assume that the corresponding linear functionals $\{\lambda_{\mathbf{x}_i}\}_{i\in\mathcal{S}}$ are linearly independent.

Now, the interpolation problem (3.3) on the stencil reads

$$\lambda_{\mathbf{x}_i}(s) = \lambda_{\mathbf{x}_i}(f) \ , \quad i \in \mathcal{S} \ . \tag{3.13}$$

The interpolant has the form

$$s(\mathbf{x}) = \sum_{j\in\mathcal{S}} c_j \lambda^{\mathbf{y}}_{\mathbf{x}_j} \phi(\|\mathbf{x}-\mathbf{y}\|) + p(\mathbf{x}) \ , \quad p \in \mathcal{P}^d_m \ , \tag{3.14}$$

with linear constraints

$$\sum_{j\in\mathcal{S}} c_j \lambda_{\mathbf{x}_j}(p) = 0 \quad \forall \ p \in \mathcal{P}^d_m \ . \tag{3.15}$$

To conserve the unique solvability, it is required that the set of functionals $\{\lambda_{\mathbf{x}_i}\}_{i\in\mathcal{S}}$ is $\mathcal{P}^d_m$-*unisolvent*. Then all results of the previous section remain valid.

For a given index i_0, consider the set of all stencils of given size n_s containing the index i_0

$$\hat{\mathcal{S}} := \{\mathcal{S}_i\}_{i=1}^{N_S} \quad , \quad i_0 \in \mathcal{S}_i \quad , \quad \mathcal{S}_i \text{ stencil} \ ,$$

where N_S stands for the number of such stencils and generally $N_S \neq n_s$.
For each stencil $\mathcal{S}_i \in \hat{\mathcal{S}}$ compute the reconstruction s_i according to (3.14). With weights $\omega_i \geq 0$, given later in this section, define the WENO reconstruction

$$R(\mathbf{x}) := \sum_{i=1}^{N_S} \omega_i s_i(\mathbf{x}) \quad \text{with} \quad \sum_{i=1}^{N_S} \omega_i = 1 \ . \tag{3.16}$$

A suitable choice of the weights ω_i leads to a non-oscillatory reconstruction. The weights ω_i should be small if the reconstruction s_i is highly oscillatory and it should be large if s_i varies slowly. Assuming that the approximation s_i is highly accurate, and therefore the behavior of s_i imitates the behavior of the interpolated function f, these properties of weights ω_i will lead to damped oscillations in the reconstruction $R(\mathbf{x})$.
The weights ω_i of a convex combination are freely to choose. In [41] the choice of weights was based on undivided differences. Another approach can be found in [28]. We will follow the work [1], where a natural choice of oscillation indicator, and therefore also of the weights, was made. Moreover, this choice is in some sense optimal and available directly from the computations.

First of all, we define the *oscillation indicator* $\mathcal{I} : BL_k(\mathbb{R}^d) \to [0, \infty)$ by

$$\mathcal{I}(v) := |v|^2_{BL_k(\mathbb{R}^d)} \quad \text{for } v \in BL_k(\mathbb{R}^d). \tag{3.17}$$

Furthermore, we define the values

$$\widetilde{\omega}_i := \frac{1}{(\epsilon + \mathcal{I}(s_i))^\rho} \;, \tag{3.18}$$

where $\epsilon > 0$ is a fixed parameter used to avoid division by zero and $\rho \in I\!N$ determines the sensitivity of the weights with respect to the oscillation indicator $\mathcal{I}$. In our numerical examples, the parameters are set to $\epsilon = 10^{-6}$ and $\rho = 2$, as in the original paper [1].
Finally, we define the weights

$$\omega_i := \frac{\widetilde{\omega}_i}{\sum_{j=1}^{N_S} \widetilde{\omega}_j} \;. \tag{3.19}$$

From the construction it is obvious that $\omega_i \geq 0$ and $\sum_{i=1}^{N_S} \omega_i = 1$.

As already mentioned, this choice of oscillation indicator is natural, since the indicator $\mathcal{I}$ as a semi-norm in the Beppo-Levi space measures the kth-order variations of interpolants s_i. It is also optimal in that sense, that s_i is the optimal interpolant on the given data on stencil $\mathcal{S}_i$ with respect to the energy $|\cdot|_{BL_k(\mathbb{R}^d)}$ (see previous section).
An additional benefit of such choice is the computation of the weights. It is not necessary to compute the oscillation indicator $\mathcal{I}$ due to the definition (3.17). For the interpolant s from (3.14), we have the formula

$$|s|^2_{BL_k(\mathbb{R}^d)} = c^T A c \;, \tag{3.20}$$

where $c = (c_j)_{j \in \mathcal{S}} \in \mathbb{R}^{n_s}$ is the vector of coefficients in (3.14) (as solution of (3.5)) and $A \in \mathbb{R}^{n_s \times n_s}$ is the matrix appearing in the corresponding linear system (3.5). Both quantities are available during the computation which is a further advantage of this approach. For the use as an oscillation indicator it was introduced by Iske in [27], who followed the work of Madych and Nelson in [42] where the relation (3.20) was introduced. The whole concept is based on the seminal works of Duchon [11], [12] and [13] and Meinguet [43]. The relation (3.20) was introduced by the referenced authors for the case of Lagrangian interpolation but it can be shown also for the case of data given by weighted integral means.

Stability

To obtain numerical stability, polyharmonic splines have to be treated carefully, since matrices arising during the computation may be ill-conditioned. The instabilities arise especially in the situation when the minimal Euclidean distance between barycentres of distinct cells in a stencil is small. This is true for cell averages and can also be shown for other cases. This problem can be overcome by using an appropriate preconditioning, see Iske [26] (for Lagrange interpolation) and Iske, Aboiyar and Georgoulis [2] (for cell averages and also for the derivatives of the interpolant). The preconditioning strategy can be straightforward extended to the case of weighted integral means.

Flexibility

For good approximation quality of the reconstruction it is necessary to select the most suitable stencil. For that, the most possible flexibility in stencil selection is desired. In the classical polynomial WENO reconstruction, the size of a stencil has to be equal to the dimension of the corresponding polynomial space. This restriction reduces the flexibility in the stencil selection. On the contrary, for polyharmonic spline WENO reconstruction it is only required, that the stencil size is at least the dimension of the polynomial space giving an enhanced flexibility to the method. We emphasize that this is true especially in higher dimensions. For more comments and details about the stencil selection see e.g., [2].

4 A method of higher order

In this chapter, we combine various techniques presented in previous chapters to design a high order meshfree method for the numerical solution of hyperbolic conservation laws in one spatial dimension. Methods of higher order of accuracy than one are called *high order scheme.* In our case, we will design a second order scheme which means that the numerical solution converges to the exact solution with order 2 in time and space. We emphasize that in our case we do not speak about a convergence of integral means but about the convergence of the numerical solution to the exact solution in an appropriate function space.
Numerical schemes of higher order have become more and more important in the industrial usage since more phenomena than earlier have been investigated and first order method often do not provide a sufficient resolution of their solution. Schemes of first order are often not able to resolve the fine structure of the solution and information about the exact solution can be lost. See e.g., the results of the example 5.2.3. From the computational point of view, one gets usually better results if a higher order scheme is applied on a coarse mesh, possibly with mesh adaptation, than to use a scheme of first order on a very fine grid.
In the finite volume framework, due to the *Godunov's theorem* [17] non-linear schemes (i.e., schemes with variable coefficients) have to be constructed to achieve higher order of accuracy. Otherwise, non-physical oscillations may arise in the vicinity of large gradients or discontinuities, as reported e.g., in Toro [64]. The ADER method, presented in chapter 1, defines the variable coefficients of the scheme in a very sophisticated way, and in combination with polyharmonic splines and WENO method used in the reconstruction step, embodies a very powerful tool for numerical solution of hyperbolic conservation laws. More details can be found in [1] and [2]. We will follow the principles introduced for FVM and adapt them to FVPM to design a similar but meshfree method.
In the first section, we formulate the scheme using the formulation of FVPM from chapter 2 and the desired higher order of accuracy will be reached by use of the ADER scheme described in chapter 1. For the construction some simplifications are necessary - we focus on the one-dimensional problem, use non-moving particles and make the special choice of partition of unity built by linear B-splines. More details about the FVPM with B-splines can be found in [31]. We will utilize the theory of polyharmonic spline interpolation and the WENO technique from chapter 3 for the reconstruction needed by the ADER method. This combination leads to a meshfree method of second order, for which we will prove second order of consistency for a scalar nonlinear governing PDE in theorem 4.5 and stability for a scalar linear conservation law in theorem 4.20. Hence, we deduce also convergence in theorem 4.26 for a scalar linear conservation law. The convergence for systems and nonlinear PDEs will be verified numerically in chapter 5.

4.1 Formulation of the method

Finite volume particle method

Let us consider the one-dimensional problem on the whole $\mathbb{R}$

$$\begin{aligned} \boldsymbol{u}_t + \boldsymbol{F}(\boldsymbol{u})_x &= \boldsymbol{0} \quad \forall\, x \in \mathbb{R} \quad , \quad \forall\, t > 0 \ , && (4.1) \\ \boldsymbol{u}(x,0) &= \boldsymbol{u}_0(x) \quad \forall\, x \in \mathbb{R} \ , && (4.2) \end{aligned}$$

where $\boldsymbol{u}_0$ is a given initial condition.
Consider the method (2.20) - (2.22) from chapter 2 for the numerical solution of (4.1)-(4.2). Assume non-moving particles, i.e., $\dot{x} = 0$. Under all these considerations the derivation step (2.13) reads

$$\frac{d}{dt} \int\limits_{\mathbb{R}} \psi_i \mathbf{u} \, d\mathbf{x} \quad = \quad \sum_j \int\limits_{\mathbb{R}} \boldsymbol{F}(\mathbf{u}) \cdot \boldsymbol{\Gamma}_{ji} - \sum_{j=1}^{n_p} \int\limits_{\mathbb{R}} \boldsymbol{F}(\mathbf{u}) \cdot \boldsymbol{\Gamma}_{ij} \; . \tag{4.3}$$

The last consideration is a special choice of partition of unity $\sum\limits_i \psi_i(x,t) = 1$ built by functions $\{\psi_i\}_i$. We choose linear B-spline functions from chapter 1, i.e., the functions ψ_i are defined by

$$\psi_i(x,t) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & , \quad x \in [x_{i-1}, x_i] \; , \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & , \quad x \in [x_i, x_{i+1}] \; , \\ 0 & , \quad \text{otherwise} \; , \end{cases}$$

where we assume the points x_i to be pairwise distinct. Notice the index shift in the notation in comparison to chapter 1 (example 1.34) where we were more interested in the properties of B-splines. We remind, that B-splines build automatically by their definition a partition of unity, so that no more construction is necessary. Moreover, in the case of linear B-splines, every particle has at most two neighbors.
Summarized, considering the hyperbolic system (4.1)-(4.2) on the whole real axis, non-moving particles of the scheme (4.1)-(4.2) and choice of linear B-splines as a partition of unity, we can conclude the following relation

$$(\Gamma_{ji} - \Gamma_{ij})\Big|_{\psi_i \cap \psi_j} = \begin{cases} (\psi_i)_x = \frac{-1}{x_{i+1} - x_i} & , \quad j = i+1 \; , \\ (\psi_i)_x = \frac{1}{x_i - x_{i-1}} & , \quad j = i-1 \; , \end{cases}$$

where $\psi_i \cap \psi_j := \operatorname{supp} \psi_i \cap \operatorname{supp} \psi_j$. Hence

$$(\Gamma_{ji} - \Gamma_{ij})\Big|_{\psi_i \cap \psi_j} = const.$$

Using this fact we can rewrite (4.3)

$$\begin{aligned} V_i \frac{d}{dt} \mathbf{u}_i \quad &= \quad \sum_j \int\limits_{\psi_i \cap \psi_j} \boldsymbol{F}(\mathbf{u}) dx \, \frac{1}{|\psi_i \cap \psi_j|} \int\limits_{\psi_i \cap \psi_j} (\Gamma_{ji} - \Gamma_{ij}) dx \\ &= \quad -\sum_j \frac{1}{|\psi_i \cap \psi_j|} \int\limits_{\psi_i \cap \psi_j} \boldsymbol{F}(\mathbf{u}) dx \; \beta_{ij} \; . \end{aligned} \tag{4.4}$$

We approximate

$$\boldsymbol{F}(\mathbf{u}(x,t))\Big|_{\psi_i \cap \psi_j} \approx \boldsymbol{F}(\mathbf{u}(x_i^j, t))$$

with $x_i^j = \frac{1}{2}(x_i + x_j)$ denoting the centre of gravity of $\psi_i \cap \psi_j$ for $j \in \{i-1, i+1\}$.
One gets

$$V_i \frac{d}{dt} \mathbf{u}_i \quad \approx \quad -\sum_j \boldsymbol{F}(\mathbf{u}(x_i^j, t)) \; \beta_{ij} \; .$$

Now we integrate the equation over time interval $[t^n, t^{n+1}]$ and divide by Δt^n

$$\frac{\mathbf{u}_i^{n+1} - \mathbf{u}_i^n}{\Delta t^n} \quad = \quad -\frac{1}{V_i} \sum_j \frac{1}{\Delta t^n} \int_{t^n}^{t^{n+1}} \boldsymbol{F}(\mathbf{u}(x_i^j, t)) dt \; \beta_{ij} \; .$$

The term $\frac{1}{\Delta t^n} \int_{t^n}^{t^{n+1}} \boldsymbol{F}(\mathbf{u}(x_i^j, t)) dt \; \frac{\beta_{ij}}{|\beta_{ij}|}$ is approximated with an appropriate numerical flux

$$\frac{1}{\Delta t^n} \int_{t^n}^{t^{n+1}} \boldsymbol{F}(\mathbf{u}(x_i^j, t)) dt \; \frac{\beta_{ij}}{|\beta_{ij}|} \approx \mathbf{g}_{ij} \; ,$$

more specifically, using an appropriate Gaussian quadrature with weights w_s and quadrature points τ_s, $s = 1, \dots, q_t$ on the time integral

$$\frac{1}{\Delta t^n} \sum_{s=1}^{q_t} w_s \boldsymbol{F}(\mathbf{u}(x_i^j, \tau_s)) \ \frac{\beta_{ij}}{|\beta_{ij}|} \approx \mathbf{g}_{ij} \ .$$

The scheme becomes

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t^n}{V_i} \sum_j |\beta_{ij}| \mathbf{g}_{ij} \ .$$

This derivation allows us to work in the framework of finite volume particle method.
In the case of linear B-splines the coefficients β_{ij} have the simple form

$$\beta_{ij} = \begin{cases} 1 & , \quad j = i+1 \ , \\ -1 & , \quad j = i-1 \ , \\ 0 & , \quad \text{else} \end{cases}$$

and the scheme can be rewritten as

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t^n}{V_i}(\mathbf{g}_{i,i+1} + \mathbf{g}_{i,i-1}) \ .$$

If the numerical flux $\mathbf{g}_{ij}$ is assumed to be conservative, the last equation will read

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t^n}{V_i}(\mathbf{g}_{i,i+1} - \mathbf{g}_{i-1,i}) \ , \tag{4.5}$$

which will motivate the definition of a higher order scheme.

Use of the ADER scheme

The notation used until now was very useful for the derivation of the method (2.20)-(2.22) and to show the conservative form (4.5) of the method. For the purposes of the rest of this chapter we will introduce a slightly different notation concerning the numerical flux: We will write $\mathbf{g}_{i+\frac{1}{2}}$ instead of $\mathbf{g}_{i,i+1}$ which denotes the approximation of the physical flux at point $x_{i+\frac{1}{2}}$.
Formally, we can rewrite the relation (4.5) in the form

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t^n}{V_i}(\mathbf{g}_{i+\frac{1}{2}} - \mathbf{g}_{i-\frac{1}{2}}) \ ,$$

where

$$\mathbf{u}_i^n = \frac{1}{V_i} \int_{\mathbb{R}} \mathbf{u} \psi_i dx$$

and

$$\boldsymbol{g}_{i+\frac{1}{2}} \quad = \quad \frac{1}{\Delta t^n} \sum_{s=1}^{q_t} w_s \boldsymbol{F}(\mathbf{u}(x_{i+\frac{1}{2}}, \tau_s)) \approx \frac{1}{\Delta t^n} \int_{t^n}^{t^{n+1}} \boldsymbol{F}(\boldsymbol{u}(x_{i+\frac{1}{2}}, t)) dt$$

for a suitable numerical quadrature with weights w_s and nodes τ_s and $x_{i+\frac{1}{2}} = \frac{1}{2}(x_i + x_{i+1})$. This is very similar to the finite volume scheme (1.14) and the numerical flux function (1.16).
Further, we follow the construction of the ADER method from chapter 1. We approximate the exact solution with a truncated Taylor expansion in time,

$$\mathbf{u}(x_{i+\frac{1}{2}}, \tau) \approx \mathbf{u}_{i+\frac{1}{2}}^{GRP}(x_{i+\frac{1}{2}}, \tau) = \mathbf{u}(x_{i+\frac{1}{2}}, 0_+) + \tau \mathbf{u}_t(x_{i+\frac{1}{2}}, 0_+)$$

and use the Cauchy-Kowalewski procedure to replace the time derivative with spatial derivative,

$$\mathbf{u}_{i+\frac{1}{2}}^{GRP}(x_{i+\frac{1}{2}}, \tau) \quad = \quad \mathbf{u}(x_{i+\frac{1}{2}}, 0_+) - \tau \boldsymbol{F}'(\mathbf{u}(x_{i+\frac{1}{2}}, 0_+)) \mathbf{u}_x(x_{i+\frac{1}{2}}, 0_+) \ .$$

For data $\mathbf{u}_L^{(j)}, \mathbf{u}_R^{(j)}$, $j = 0, 1$ given later in this section, we approximate the terms $\mathbf{u}(x_{i+\frac{1}{2}}, 0_+)$ and $\mathbf{u}_x(x_{i+\frac{1}{2}}, 0_+)$ with the solution of a generalized Riemann problem, which is approximated with a truncated series of two classical Riemann problems for $\mathbf{u}$ and its derivative $\mathbf{u}_x$. More specifically,

$$\mathbf{u}(x_{i+\frac{1}{2}}, \tau) \quad \approx \quad \mathbf{u}^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, \tau) \approx \mathbf{u}^{(0)}_{i+\frac{1}{2}} - \tau \boldsymbol{F}'(\mathbf{u}^{(0)}_{i+\frac{1}{2}})\mathbf{u}^{(1)}_{i+\frac{1}{2}} \ ,$$

where

$$\begin{aligned} \mathbf{u}(x_{i+\frac{1}{2}}, 0_+) \quad &\approx \quad \mathbf{u}^{(0)}_{i+\frac{1}{2}} = RP_{i+\frac{1}{2}}(\mathbf{u}^{(0)}_L, \mathbf{u}^{(0)}_R) \ , \\ \mathbf{u}_x(x_{i+\frac{1}{2}}, 0_+) \quad &\approx \quad \mathbf{u}^{(1)}_{i+\frac{1}{2}} = LRP_{i+\frac{1}{2}}(\mathbf{u}^{(1)}_L, \mathbf{u}^{(1)}_R) \ . \end{aligned}$$

We denote by

$$\mathbf{u}^{(0)}_{i+\frac{1}{2}} = RP_{i+\frac{1}{2}}(\mathbf{u}^{(0)}_L, \mathbf{u}^{(0)}_R) \tag{4.6}$$

the solution of Riemann problem (1.5)-(1.6) along $(x - x_{i+\frac{1}{2}})/t$ with the governing equation (4.1) and with initial data

$$\mathbf{u}(x, 0) = \begin{cases} \boldsymbol{u}^{(0)}_L & , \quad x < x_{i+\frac{1}{2}} \ , \\ \boldsymbol{u}^{(0)}_R & , \quad x > x_{i+\frac{1}{2}} \ . \end{cases}$$

By the term

$$\mathbf{u}^{(1)}_{i+\frac{1}{2}} = LRP_{i+\frac{1}{2}}(\mathbf{u}^{(1)}_L, \mathbf{u}^{(1)}_R) \tag{4.7}$$

we denote the solution along $(x - x_{i+\frac{1}{2}})/t$ of a linearized Riemann problem for the derivatives $\mathbf{u}_x$

$$\begin{aligned} (\mathbf{u}_x)_t(x, t) + \boldsymbol{F}'(\mathbf{u}^{(0)}_{i+\frac{1}{2}})(\mathbf{u}_x)_x(x, t) \quad &= \quad \mathbf{0} \ , \\ \mathbf{u}_x(x, 0) \quad &= \quad \begin{cases} \boldsymbol{u}^{(1)}_L & , \quad x < x_{i+\frac{1}{2}} \ , \\ \boldsymbol{u}^{(1)}_R & , \quad x > x_{i+\frac{1}{2}} \ , \end{cases} \end{aligned}$$

where we linearize around $\mathbf{u}^{(0)}_{i+\frac{1}{2}}$.

In order to get a complete scheme, it remains to define the initial data $\mathbf{u}_L^{(j)}, \mathbf{u}_R^{(j)}$, $j = 0, 1$. These data are acquired from reconstructions $\mathbf{R}_i$ of the exact solution $\mathbf{u}$ for all $i \in \mathbb{Z}$ by polyharmonic splines based on data $\mathbf{u}_i$, as treated in chapter 3. More specifically, for every $i \in \mathbb{Z}$ one constructs several reconstructions and combine them using a WENO method to get the final non-oscillatory reconstruction $\mathbf{R}_i$. The construction follows.

We will consider *stencils*. A *stencil* is a set of neighboring particles (more precisely, a set of indices of neighboring particles) to a given particle index i involving the particle index i itself too. We define

$$\hat{\mathcal{S}}^i = \{\mathcal{S}^i_l\}_{l=1}^{N_S}$$

the set of all stencils corresponding to the given particle x_i, where $\mathcal{S}^i_l$ is the l-th stencil of the set. The size of a stencil $\mathcal{S}^i_l$ is a given number n_s. The number N_S denotes the number of elements of $\hat{\mathcal{S}}^i$ and it holds $N_S = n_s$ in one dimension.

Consider the linear functional λ_i from (3.2) (we will use the simplified notation λ_i instead of λ_{x_i} in this chapter). For each $l \in \{1, \dots, N_S\}$ find an interpolant $\mathbf{s}^i_l$ on $\mathbf{u}$, such that for every component s^i_l of $\mathbf{s}^i_l$ and u of $\mathbf{u}$ the relation

$$\lambda_j(s^i_l) = \lambda_j(u) \qquad \forall \ j \in \mathcal{S}^i_l \tag{4.8}$$

holds. More precisely, we solve m separate interpolation problems for every component of a vector function $\mathbf{u} = (u_1, \dots, u_m)^T$. The polyharmonic spline function is used as the sought interpolant

in every component.
Having determined the interpolants $\mathbf{s}_l^i$ for each $l = 1, \ldots, N_S$, we compute their oscillation indicators by (3.17) and the corresponding weights $\boldsymbol{\omega}_l^i$ by (3.19) (componentwise for each component of $\mathbf{s}_l^i$). First, we have to define the componentwise product.

Definition 4.1 (Componentwise product)
Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^m$*. The* componentwise product *(also known as Hadamard product, Schur product or elementwise product) of vectors* $\mathbf{v}$ *and* $\mathbf{w}$ *is defined by*

$$\mathbf{v} \circ \mathbf{w} = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \circ \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} := \begin{pmatrix} v_1 w_1 \\ \vdots \\ v_m w_m \end{pmatrix} .$$

Finally, we define the reconstruction $\mathbf{R}_i$ on $\mathbf{u}$ via (3.16), i.e.,

$$\mathbf{R}_i(x) = \sum_{l=1}^{N_S} \boldsymbol{\omega}_l^i \circ \mathbf{s}_l^i(x) \; . \tag{4.9}$$

With this, we have determined for each particle x_i a reconstruction $\mathbf{R}_i$ on the exact solution $\mathbf{u}$, such that

$$\lambda_i(\mathbf{R}_i) = \lambda_i(\mathbf{u}) \; .$$

This motivates the definition of the domain of definition of each reconstruction $\mathbf{R}_i$, which we set to $\operatorname{supp} \psi_i = [x_{i-1}, x_{i+1}]$ (compare with FVM). Consider, for the sake of simplicity, a scalar hyperbolic conservation law, i.e., $m = 1$. This is reasonable since the reconstructions are built for each component of $\mathbf{u}$ separately. In the end, we will formulate the whole method for a general $m \in I\!N$.
Now, consider two neighboring reconstructions R_i and R_{i+1}, cf. figure 4.1. Based on these reconstructions, we want to define a generalized Riemann problem similar to the problem for FVM defined in definition 1.28 (cf. figure 1.2). But, compared to the FVM framework, where characteristic functions χ_i of a given interval are used, our reconstructions in FVPM overlap, since the basis functions ψ_i overlap. That is why the generalized Riemann problem for FVPM has to be defined in a special way.
Consider the figure 4.2. If we take only the function values of R_i and R_{i+1} depicted with full lines in the figure into account, we can immediately define a generalized Riemann problem with these data as initial values. More precisely, consider the center of gravity $x_{i+\frac{1}{2}} := \frac{1}{2}(x_i + x_{i+1})$ of the interval $\operatorname{supp} \psi_i \; \cap \; \operatorname{supp} \psi_{i+1} = [x_i, x_{i+1}]$ and the values of R_i on the interval $[x_{i-1}, x_{i+\frac{1}{2}}]$ and R_{i+1} on the interval $[x_{i+\frac{1}{2}}, x_{i+2}]$. Then the function

$$\widetilde{u}_0(x) = \begin{cases} R_i(x) & , \quad x < x_{i+\frac{1}{2}} \; , \\ R_{i+1}(x) & , \quad x > x_{i+\frac{1}{2}} \end{cases} \tag{4.10}$$

represents a well-defined initial condition for the generalized Riemann problem from definition 1.28.
For $j = 0, 1$ we define the values at the interface $x_{i+\frac{1}{2}}$

$$\begin{aligned} u_L^{(j)} &= \lim_{x \to x_{i+\frac{1}{2}-}} \partial^{(j)} R_i(x) = \partial^{(j)} R_i(x_{i+\frac{1}{2}}) \; , \\ u_R^{(j)} &= \lim_{x \to x_{i+\frac{1}{2}+}} \partial^{(j)} R_{i+1}(x) = \partial^{(j)} R_{i+1}(x_{i+\frac{1}{2}}) \; . \end{aligned}$$

Following chapter 1, the generalized Riemann problem from definition 1.28 with initial conditions defined by (4.10) is then approximated by a truncated series of classical Riemann problems (4.6)

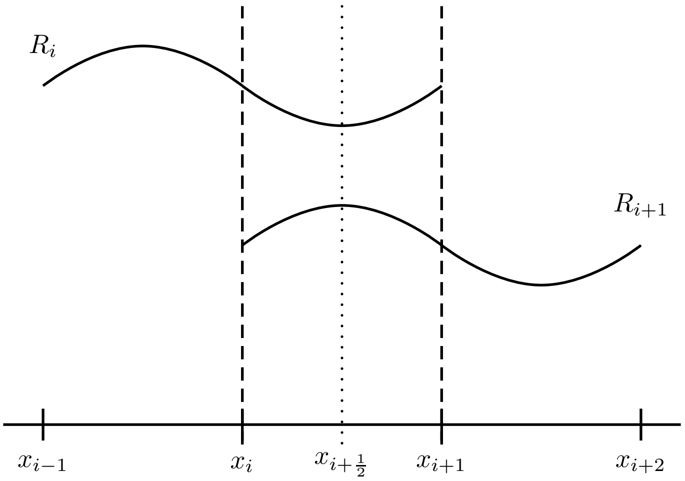


Figure 4.1: *An illustration of two neighboring reconstructions. A scalar case of hyperbolic conservation law, i.e., $m = 1$, is considered. The reconstructions overlap in the interval $[x_i, x_{i+1}]$.*

and (4.7) with initial conditions given by the above defined values $u_L^{(j)}$ and $u_R^{(j)}$. In the following text, we will prove a consistency of definition of the GRP for FVPM and introduce the definition of the high order FVPM.

Remark 4.2
The correct notation would be $u^{(j)}_{L,i+\frac{1}{2}}$, $u^{(j)}_{R,i+\frac{1}{2}}$ and

$$\begin{aligned} &RP_{i+\frac{1}{2}}(u^{(0)}_{L,i+\frac{1}{2}}, u^{(0)}_{R,i+\frac{1}{2}}) \ , \\ &LRP_{i+\frac{1}{2}}(u^{(1)}_{L,i+\frac{1}{2}}, u^{(1)}_{R,i+\frac{1}{2}}) \ , \end{aligned}$$

denoting the dependence of the data on i. For the sake of notational simplicity we omit the index $i + \frac{1}{2}$ of data $u_L^{(j)}$, $u_R^{(j)}$.

Consistency of the definition of the generalized Riemann problem

Consider reconstructions $\mathbf{R}_i$ satisfying

$$\frac{1}{V_i} \int\limits_{\text{supp } \psi_i} \mathbf{R}_i \psi_i dx = \frac{1}{V_i} \int\limits_{\text{supp } \psi_i} \mathbf{u} \psi_i dx \ .$$

Then we defined for FVPM the generalized Riemann problem

$$GRP_{i+\frac{1}{2}}(\mathbf{u}_L(x), \mathbf{u}_R(x)) \tag{4.11}$$

with data

$$\begin{aligned} \mathbf{u}_L(x) &= \mathbf{R}_i(x) \quad , \quad x < x_{i+\frac{1}{2}} \ , \\ \mathbf{u}_R(x) &= \mathbf{R}_{i+1}(x) \quad , \quad x > x_{i+\frac{1}{2}} \ . \end{aligned}$$

This definition is consistent with the definition of GRP in the framework of FVM in the following sense.

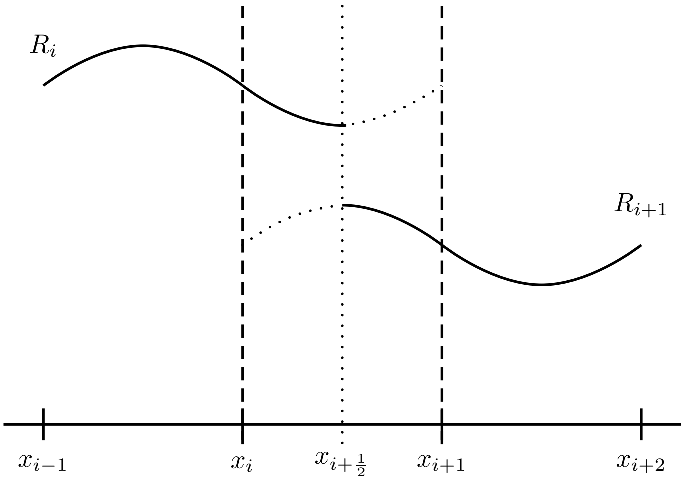


Figure 4.2: *An illustration of two neighboring restricted reconstructions. A scalar case of hyperbolic conservation law, i.e., $m = 1$, is considered. We consider only the values of R_i on $[x_{i-1}, x_{i+\frac{1}{2}}]$ and R_{i+1} on $[x_{i+\frac{1}{2}}, x_{i+2}]$. A generalized Riemann problem with this data can be defined.*

For the sake of simplicity we will consider uniform particle distribution, i.e., $x_{i+1} - x_i = \Delta x$ for all $i \in \mathbb{Z}$. Consider the transformation functions ψ_i^α defined for $\alpha \in [0,1]$ as

$$\psi_i^\alpha(x) = \begin{cases} 0 & , \quad x \in \left(-\infty, x_i - \frac{1+\alpha}{2}\Delta x\right] \cup \left[x_i + \frac{1+\alpha}{2}\Delta x, \infty\right) \ , \\ 1 & , \quad x \in \left[x_i - \frac{1-\alpha}{2}\Delta x, x_i + \frac{1-\alpha}{2}\Delta x\right] \ , \\ \frac{x - x_i + \frac{1+\alpha}{2}\Delta x}{\alpha \Delta x} & , \quad x \in \left[x_i - \frac{1+\alpha}{2}\Delta x, x_i - \frac{1-\alpha}{2}\Delta x\right] \ , \\ \frac{x_i + \frac{1+\alpha}{2}\Delta x - x}{\alpha \Delta x} & , \quad x \in \left[x_i + \frac{1-\alpha}{2}\Delta x, x_i + \frac{1+\alpha}{2}\Delta x\right] \ . \end{cases}$$

The graphs of the functions are shown in figure 4.3.
One can show, that it holds

$$\begin{aligned} \psi_i^1(x) &= \psi_i(x) \ , \\ \psi_i^0(x) &= \chi_i(x) := \chi_{[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]}(x) \ , \\ \|\psi_i^\alpha - \chi_i\|_{L^1(\mathbb{R})} &\overset{\alpha \to 0_+}{\longrightarrow} 0 \ . \end{aligned}$$

For every $\alpha \in [0,1]$ define the interpolation problems

$$\frac{1}{V_i^\alpha} \int\limits_{\mathbb{R}} \mathbf{R}_i^\alpha \psi_i^\alpha dx = \frac{1}{V_i^\alpha} \int\limits_{\mathbb{R}} \mathbf{u} \psi_i^\alpha dx \quad , \quad i \in \mathbb{Z} \ , \tag{4.12}$$

which we solve via polyharmonic splines. Then the problem (4.12) has for every $\alpha \in [0,1]$ a unique solution $\mathbf{R}_i^\alpha$ for every $i \in \mathbb{Z}$. The idea now is, to show that if we change from particle basis functions ψ_i to characteristic functions χ_i via limit transition with respect to α, the limit of reconstructions $\mathbf{R}_i^\alpha$ and $\mathbf{R}_{i+1}^\alpha$ will define a well-posed GRP for FVM. We define then the GRP for FVPM in that sense, i.e., in the sense of FVM.

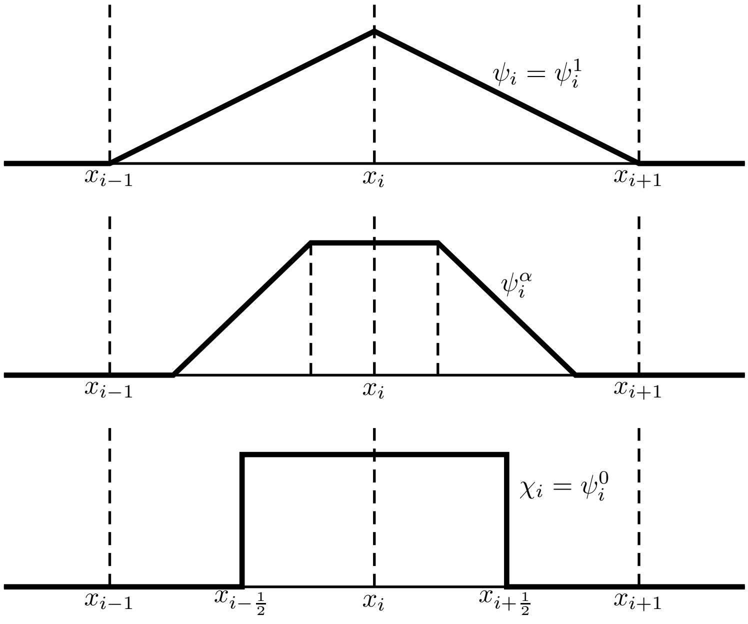


Figure 4.3: *Functions ψ_i, ψ_i^α and χ_i.*

It holds

$$V_i^\alpha = \int\limits_{\text{supp } \psi_i} \psi_i^\alpha(x)dx = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} = \Delta x \quad \forall\ \alpha \in [0,1] \ .$$

Further, one can show that

$$\mathbf{R}_i^\alpha \to \mathbf{R}_i^0 \quad \text{pointwise.}$$

The latter follows from the structure of every single component R_i^α of $\mathbf{R}_i^\alpha$

$$R_i^\alpha(x) = \sum_{l=1}^{N_S} \omega_l^{i,\alpha} s_l^{i,\alpha}(x) \ .$$

The values

$$\omega_l^{i,\alpha} = \omega_l^{i,\alpha}(s_l^{i,\alpha})$$

depends continuously on $s_l^{i,\alpha}$ (see (3.19) for the exact definition). Recall from (3.14) that

$$s_l^{i,\alpha}(x) = \sum_{j \in \mathcal{S}_l^i} c_j^\alpha \lambda_j^{y,\alpha} \phi(\|x-y\|) + \sum_{|\beta|<m} d_\beta^\alpha x^\beta \ ,$$

where the upper index α denotes the dependency on the parameter α. The functional $\lambda_i^{y,\alpha}$ is defined in the same way as in (3.2) with respect to the function ψ_i^α. It can be easily verified that $\lambda_i^{y,\alpha}\phi(\|x-y\|) \to \lambda_i^{y,0}\phi(\|x-y\|)$ for $\alpha \to 0_+$. The coefficients c_j^α and d_β^α are solution of the system (3.5), i.e.,

$$\begin{bmatrix} A^\alpha & P^\alpha \\ (P^\alpha)^T & 0 \end{bmatrix} \cdot \begin{bmatrix} c^\alpha \\ d^\alpha \end{bmatrix} = \begin{bmatrix} u\big|_{\lambda_X^\alpha} \\ 0 \end{bmatrix} ,$$

where

$$\begin{aligned} A^\alpha &= \left(\lambda_i^{x,\alpha}\lambda_j^{y,\alpha}\phi(\|x-y\|)\right)_{i,j} \in \mathbb{R}^{n_s\times n_s} \ , \\ P^\alpha &= \left(\lambda_i^\alpha(x^\beta)\right)_{i,\beta} \in \mathbb{R}^{n_s\times q} \ , \\ u\Big|_{\lambda_X^\alpha} &= \left(\lambda_i^\alpha(u)\right)_i \in \mathbb{R}^{n_s} \ . \end{aligned}$$

Then $c_j^\alpha \to c_j^0$ and $d_\beta^\alpha \to d_\beta^0$. Altogether, using the triangle inequality, one gets the pointwise convergence $\mathbf{R}_i^\alpha \to \mathbf{R}_i^0$.
From that we deduce

$$\begin{tikzcd} \frac{1}{V_i^\alpha}\int\limits_{\mathbb{R}} \mathbf{R}_i^\alpha\psi_i^\alpha dx \arrow[dd] & = & \frac{1}{V_i^\alpha}\int\limits_{\mathbb{R}} \mathbf{u}\psi_i^\alpha dx \arrow[dd] \\ & \alpha \to 0_+ & \\ \frac{1}{\Delta x}\int\limits_{\mathbb{R}} \mathbf{R}_i^0\chi_i dx & = & \frac{1}{\Delta x}\int\limits_{\mathbb{R}} \mathbf{u}\chi_i dx \ . \end{tikzcd}$$

Then a generalized Riemann problem

$$GRP_{i+\frac{1}{2}}(\mathbf{u}_L(x),\mathbf{u}_R(x))$$

on the interface $x_{i+\frac{1}{2}}$ with data

$$\begin{aligned} \mathbf{u}_L(x) = \mathbf{R}_i^0(x) \quad &, \quad x < x_{i+\frac{1}{2}} \ , \\ \mathbf{u}_R(x) = \mathbf{R}_{i+1}^0(x) \quad &, \quad x > x_{i+\frac{1}{2}} \ , \end{aligned}$$

satisfying

$$\frac{1}{\Delta x}\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{R}_i^0 = \frac{1}{\Delta x}\int\limits_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{u} \quad , \quad \frac{1}{\Delta x}\int\limits_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} \mathbf{R}_{i+1}^0 = \frac{1}{\Delta x}\int\limits_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} \mathbf{u}$$

can be defined. This corresponds to the definition in the framework of FVM and the solution of $GRP_{i+\frac{1}{2}}(\mathbf{u}_L(x),\mathbf{u}_R(x))$ defines an approximation on the exact solution $\mathbf{u}$ at the point $x = x_{i+\frac{1}{2}}$ as in the classical ADER method for FVM.

Definition of the method

Let us summarize the definition of the higher order FVPM for the general case $m \geq 1$.

Definition 4.3 (Higher order method)
We define a finite volume particle method for the numerical solution of (4.1)-(4.2) with non-moving particles using linear B-splines as the partition of unity given by the scheme

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t^n}{V_i}(\mathbf{g}_{i+\frac{1}{2}} - \mathbf{g}_{i-\frac{1}{2}}) \quad , \ i \in \mathbb{Z} \tag{4.13}$$

with the numerical flux

$$\mathbf{g}_{i+\frac{1}{2}} = \frac{1}{\Delta t^n}\sum_{s=1}^{q_t} w_s \boldsymbol{F}\left(\mathbf{u}_{i+\frac{1}{2}}^{GRP}(x_{i+\frac{1}{2}},\tau_s)\right) \ . \tag{4.14}$$

The values $\mathbf{u}_{i+\frac{1}{2}}^{GRP}(x_{i+\frac{1}{2}},\tau)$ *are defined by*

$$\mathbf{u}_{i+\frac{1}{2}}^{GRP}(x_{i+\frac{1}{2}},\tau) = \mathbf{u}_{i+\frac{1}{2}}^{(0)} - \tau\boldsymbol{F}'(\mathbf{u}_{i+\frac{1}{2}}^{(0)})\mathbf{u}_{i+\frac{1}{2}}^{(1)} \ , \tag{4.15}$$

where $\mathbf{u}^{(0)}_{i+\frac{1}{2}}$ *and* $\mathbf{u}^{(1)}_{i+\frac{1}{2}}$ *stand for the solutions of classical Riemann problems (4.6) and (4.7), i.e.,*

$$\mathbf{u}^{(0)}_{i+\frac{1}{2}} = RP_{i+\frac{1}{2}}(\mathbf{u}^{(0)}_L, \mathbf{u}^{(0)}_R) \ , \tag{4.16}$$

$$\mathbf{u}^{(1)}_{i+\frac{1}{2}} = LRP_{i+\frac{1}{2}}(\mathbf{u}^{(1)}_L, \mathbf{u}^{(1)}_R) \ . \tag{4.17}$$

Data $\mathbf{u}^{(j)}_L$ *and* $\mathbf{u}^{(j)}_R$ *are defined for* $j = 0, 1$ *by*

$$\mathbf{u}^{(j)}_L = \lim_{x \to x_{i+\frac{1}{2}-}} \partial^{(j)}_x \mathbf{R}_i(x) \ , \tag{4.18}$$

$$\mathbf{u}^{(j)}_R = \lim_{x \to x_{i+\frac{1}{2}+}} \partial^{(j)}_x \mathbf{R}_{i+1}(x) \ , \tag{4.19}$$

where the WENO reconstructions $\mathbf{R}_i(x)$ *are given for* $i \in \mathbb{Z}$ *by (4.9), i.e.,*

$$\mathbf{R}_i(x) = \sum_{l=1}^{N_S} \boldsymbol{\omega}^i_l \circ \mathbf{s}^i_l(x) \ . \tag{4.20}$$

The function $\mathbf{s}^i_l$ *solves for* $l = 1, \ldots, N_S$ *componentwise the interpolation problem (4.8) at a given time level* t^n*, i.e.,*

$$\lambda_j(\mathbf{s}^i_l) = \mathbf{u}^n_j \ , \ j \in \mathcal{S}^i_l$$

by the use of polyharmonic splines.

Remark 4.4
In the practice, in order to get second order of accuracy, we will use the Gaussian quadrature (more specifically Gauss-Legendre quadrature). We will use the number of nodes $q_t = 2$*, nodes* $\tau_s = \pm\sqrt{1/3}$ *and weights* $w_s = 1$ *for the integration over interval* $[-1, 1]$*. We also transform the integral from a general interval* $[a, b]$ *onto* $[-1, 1]$*. The final formula is*

$$\int_a^b f(t)dt = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}\tau + \frac{a+b}{2}\right) d\tau \approx \frac{b-a}{2}\sum_{s=1}^{2} w_s f\left(\frac{b-a}{2}\tau_s + \frac{a+b}{2}\right) \ .$$

In next sections, we are going to analyse the consistency and stability of the above defined scheme for a scalar governing equation. Since we want to prove the order of accuracy to be 2, we will set for polyharmonic splines the parameter $k = 2$. First, we investigate the scheme (4.13)-(4.20) for special choices of parameters: We will simplify the scheme for the case of linear advection equation and $n_s = 2$. This will be later useful to prove stability. Moreover, we will introduce values of WENO weights for $n_s = 3$ in the case of linear advection equation to verify our hypotheses in the section concerning consistency.
Consider for a moment the scheme (4.13)-(4.20) for a scalar linear conservation law $u_t + au_x = 0$, $a > 0$:

The scheme for $n_s = 2$, $k = 2$

We will introduce the whole scheme.
On the stencil $\mathcal{S} = (i, i+1)$ we have to determine the matrices A and P from (3.5). Direct computation yields

$$A = \Delta x^3 \begin{pmatrix} \frac{31}{105} & \frac{841}{420} \\ \frac{841}{420} & \frac{31}{105} \end{pmatrix} \quad , \quad P = \begin{pmatrix} 1 & x_i \\ 1 & x_{i+1} \end{pmatrix} \ .$$

The solution of the system (3.5), i.e.,

$$\begin{bmatrix} A & P \\ P^T & 0 \end{bmatrix} \cdot \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} u\Big|_{\lambda_X} \\ 0 \end{bmatrix} ,$$

where $u\Big|_{\lambda_X} = (u_i, u_{i+1})^T$, is

$$c = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad , \quad d = \frac{1}{\Delta x} \begin{pmatrix} x_{i+1}u_i - x_i u_{i+1} \\ u_{i+1} - u_i \end{pmatrix} .$$

The polyharmonic spline interpolant has then the form

$$s(x) = \sum_{j=1}^{2} c_j \lambda^y_{i+j-1} \phi(|x-y|) + d_1 + d_2 x .$$

We can see that in this case s is a polynomial. Due to (3.20) $|s|_{BL_2} = 0$, and therefore the corresponding WENO weight is $\omega = \frac{1}{n_s} = \frac{1}{2}$.
So we see, we have the interpolant at the point $x_{i+\frac{1}{2}}$ equal to $s^i_2(x_{i+\frac{1}{2}}) = \frac{1}{2}(u^n_i + u^n_{i+1})$ on the stencil $(i, i+1)$ and the value of interpolant s^i_1 is $s^i_1(x_{i+\frac{1}{2}}) = -\frac{1}{2}u^n_{i-1} + \frac{3}{2}u^n_i$ on the stencil $(i-1, i)$. Then

$$R_i(x_{i+\frac{1}{2}}) = \frac{1}{2}s^i_1(x_{i+\frac{1}{2}}) + \frac{1}{2}s^i_2(x_{i+\frac{1}{2}}) = u^n_i + \frac{1}{4}(u^n_{i+1} - u^n_{i-1}) .$$

Analogously one gets

$$R'_i(x_{i+\frac{1}{2}}) = \frac{1}{2\Delta x}(u^n_{i+1} - u^n_{i-1}) .$$

Since both Riemann problems are linear with the characteristic speed $a > 0$, we obtain

$$\begin{aligned} R_i(x_{i+\frac{1}{2}}) &= RP_{i+\frac{1}{2}}\Big(R_i(x_{i+\frac{1}{2}}), R_{i+1}(x_{i+\frac{1}{2}})\Big) , \\ R'_i(x_{i+\frac{1}{2}}) &= LRP_{i+\frac{1}{2}}\Big(R'_i(x_{i+\frac{1}{2}}), R'_{i+1}(x_{i+\frac{1}{2}})\Big) , \end{aligned}$$

and

$$\begin{aligned} g_{i+\frac{1}{2}} &= \frac{1}{\Delta t}\int_0^{\Delta t} a\Big(R_i(x_{i+\frac{1}{2}}) - \tau a R'_i(x_{i+\frac{1}{2}})\Big)\,d\tau \\ &= aR_i(x_{i+\frac{1}{2}}) - \frac{\Delta t}{2}a^2 R'_i(x_{i+\frac{1}{2}}) \\ &= a(u^n_i + \frac{1}{4}\left(u^n_{i+1} - u^n_{i-1}\right)) - \frac{1}{4}\frac{\Delta t}{\Delta x}a^2(u^n_{i+1} - u^n_{i-1}) . \end{aligned}$$

Analogously we get for $g_{i-\frac{1}{2}}$

$$g_{i-\frac{1}{2}} = a\left(u^n_{i-1} + \frac{1}{4}(u^n_i - u^n_{i-2})\right) - \frac{1}{4}\frac{\Delta t}{\Delta x}a^2(u^n_i - u^n_{i-2}) .$$

After plugging $g_{i+\frac{1}{2}}$ and $g_{i-\frac{1}{2}}$ into (4.13) we rearrange and get

$$u^{n+1}_i = \sum_{j=-2}^{1} b_j u^n_{i+j} \tag{4.21}$$

with the coefficients b_j given by (4.38)-(4.41) yielding the explicit scheme.

The WENO weights for $n_s = 3$, $k = 2$

Here we determine only the WENO weights since they are needed in the next section. Consider the stencil $\mathcal{S} = (i-1, i, i+1)$. The matrices A and P read

$$A = \Delta x^3 \begin{pmatrix} \frac{31}{105} & \frac{841}{420} & 10 \\ \\ \frac{841}{420} & \frac{31}{105} & \frac{841}{420} \\ \\ 10 & \frac{841}{420} & \frac{31}{105} \end{pmatrix} \quad , \quad P = \begin{pmatrix} 1 & x_{i-1} \\ 1 & x_i \\ 1 & x_{i+1} \end{pmatrix} .$$

The solution of the corresponding matrix (3.5) with the corresponding right hand side, where $u\Big|_{\lambda_X} = (u_{i-1}, u_i, u_{i+1})^T$, is

$$\begin{aligned} c &= \frac{\mathcal{D}_0}{\Delta x^3} \begin{pmatrix} u_{i-1} - 2u_i + u_{i+1} \\ -2(u_{i-1} - 2u_i + u_{i+1}) \\ u_{i-1} - 2u_i + u_{i+1} \end{pmatrix} , \\ d &= \begin{pmatrix} -\frac{717\Delta x - 604 x_i}{1208\Delta x} u_{i-1} + \frac{1321}{604} u_i - \frac{717\Delta x + 604 x_i}{1208\Delta x} u_{i+1} \\ \\ \frac{1}{2\Delta x}(u_{i+1} - u_{i-1}) \end{pmatrix} , \end{aligned}$$

where the constant $\mathcal{D}_0 = \frac{105}{604}$. The polyharmonic spline interpolant then has the form

$$s(x) = \sum_{j=1}^{3} c_j \lambda^y_{i+j-2} \phi(|x-y|) + d_1 + d_2 x .$$

Then from (3.20) we get

$$|s|^2_{BL_2(\mathbb{R})} = c^T A c = \frac{\mathcal{D}_0}{\Delta x^3}(u_{i-1} - 2u_i + u_{i+1})^2 . \tag{4.22}$$

The WENO weight can be now determined via (3.19).

4.2 Local truncation error

Consider a scalar conservation law in one dimension, i.e.,

$$\begin{aligned} u_t + F(u)_x &= 0 \quad , \quad x \in \mathbb{R} \ , \ t > 0 \ , & (4.23) \\ u(x,0) &= u_0(x) \quad , \quad x \in \mathbb{R} \ . & (4.24) \end{aligned}$$

Assume, that the initial data (4.24) has compact support. Then the solution of the hyperbolic problem (4.23) will have compact support for all times. Let $\Omega \subset \mathbb{R}$ be an open and bounded domain, such that supp $u(\cdot,t) \subset \Omega$ for all times $t \in [0,T]$.
Furthermore, we will consider some smoothness on u over a domain $\overline{\Omega} \times [0,T]$ to get bounds on certain terms.
For the sake of simplicity, let us consider a uniform discretization in time and space

$$x_{i+1} - x_i = \Delta x > 0 \quad \forall\ i \in \mathbb{Z}$$

and

$$t^{n+1} - t^n = \Delta t > 0 \quad \forall\ n \in I\!N_0 \ .$$

Then all volumes have the same size, i.e., $V_i = \Delta x$ for all $i \in \mathbb{Z}$.
Consider the scheme (4.13)-(4.20) applied on (4.23).
The local truncation error is defined by

$$T(x_i, t^n) := \frac{u(x_i, t^n + \Delta t) - u(x_i, t^n)}{\Delta t} + \frac{1}{\Delta x}(g_{i+\frac{1}{2}} - g_{i-\frac{1}{2}}) \ , \tag{4.25}$$

which is the difference of the left and of the right hand side of the formula (4.13) divided by Δt. More specifically, we plug the exact values $u(x_i,t^n)$ instead of u_i^n into the formula (i.e., of the exact solution of (4.23)) and also into the numerical flux functions $g_{i+\frac{1}{2}}$ (see the details below). In order to keep the notation simple, we will still denote the numerical flux with exact values by $g_{i+\frac{1}{2}}$ in this section.

Our goal is to show that the scheme (4.13)-(4.20) is of second order of accuracy in time and space, i.e.,

$$T(x_i,t^n) = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)\ ,\ \Delta t \to 0\ ,\ \Delta x \to 0\ .$$

In the following we introduce useful results to show the local truncation error.

Some results on polyharmonic splines

In order to investigate the local truncation error of scheme (4.13)-(4.20) we will make use of results of chapter 3. To this end, we will adapt the notation. Recall the results and the proof of theorem 3.3 in one dimension for some function f smooth enough:

$$\begin{aligned} f(h\widetilde{x}) - s^h(h\widetilde{x}) &= \sum_k L_k^h(h\widetilde{x}) \left[T^m_{f,h\widetilde{x}}(h\widetilde{x}_k) - f(h\widetilde{x}_k)\right] && (4.26)\\ &= \sum_k L_k^1(\widetilde{x}) \left[T^m_{f,h\widetilde{x}}(h\widetilde{x}_k) - f(h\widetilde{x}_k)\right]\ . && (4.27)\end{aligned}$$

Define $x := h\widetilde{x}$ and $x_j := h\widetilde{x}_j$ and omit the upper index h at the function s^h (in the following analysis we consider always the scaled problem).

Consider for a moment the equivalent notation λ_{x_i} of the functional (3.2) instead of the notation λ_i used otherwise in this chapter. We make the following consideration.

We want to solve the interpolation problems for all $i \in \mathbb{Z}$ and for all $l = 1,\dots,n_s$:

$$\lambda_{h\widetilde{x}_k}\left(s_l^i\right) = \lambda_{h\widetilde{x}_k}(f) \qquad \forall\ k \in \mathcal{S}_l^i\ , \tag{4.28}$$

where

$$\lambda_{h\widetilde{x}_k}(f) := \frac{1}{V_{h\widetilde{x}_k}} \int_{\text{supp}\ \psi_{h\widetilde{x}_k}} f(y)\psi_{h\widetilde{x}_k}(y)dy\ ,$$

$$V_{h\widetilde{x}_k} := \int_{\text{supp}\ \psi_{h\widetilde{x}_k}} \psi_{h\widetilde{x}_k}(y)dy\ ,$$

with the scaling parameter $h > 0$. We choose the special problem for the parameter $h = 1$. In this case

$$\lambda_{\widetilde{x}_k}\left(s_l^i\right) = \lambda_{\widetilde{x}_k}(f) \qquad \forall\ k \in \mathcal{S}_l^i\ .$$

Until now the points $\widetilde{X} = \{\widetilde{x}_k\}_{k=1}^{n_s}$ were some arbitrary points in $\mathbb{R}^d$, s.t. $\lambda_{\widetilde{X}}$ is $\mathcal{P}_m^d$-unisolvent. For our considerations, since we use uniform distributions, let us assume fixed points $\widetilde{X}$, s.t.

$$\widetilde{x}_{k+1} - \widetilde{x}_k = 1\ .$$

The points $\widetilde{X}$ can be considered as *reference points.* Then

$$\Delta x = x_{k+1} - x_k = h\widetilde{x}_{k+1} - h\widetilde{x}_k = h(\widetilde{x}_{k+1} - \widetilde{x}_k) = h\ .$$

With this we have derived the relation between the scaling parameter h in the general local interpolation problem (4.28) and the space discretization of the method (4.13)-(4.20), which will allow us to investigate consistency of the scheme.

The assumption $\widetilde{x}_{k+1} - \widetilde{x}_k = 1$ is without loss of generality with respect to the investigation of convergence of the method. Indeed, for our case with uniform particle distribution, there would be always some constant C, such that

$$\widetilde{x}_{k+1} - \widetilde{x}_k = C \ ,$$

and therefore

$$h = \frac{\Delta x}{\widetilde{x}_{k+1} - \widetilde{x}_k} = \frac{\Delta x}{C} \ .$$

Hence, one would work with Δx scaled with a fixed factor $1/C$.
Let us get back to (4.26)

$$f(h\widetilde{x}) - s^h(h\widetilde{x}) \quad = \quad \sum_k L_k^1(\widetilde{x}) \left[T^m_{f,h\widetilde{x}}(h\widetilde{x}_k) - f(h\widetilde{x}_k) \right] \ .$$

Now we can rewrite it for the "local" variable x as

$$\begin{aligned} f(x) - s(x) &= \sum_k L_k^1(\widetilde{x}) \left[T^m_{f,x}(x_k) - f(x_k) \right] & (4.29) \\ &= \sum_k L_k^1(\widetilde{x}) \left[\sum_{|\alpha|=m} \frac{1}{\alpha!} D^\alpha f(x)(x_k - x)^\alpha + \mathcal{O}(\Delta x^{m+1}) \right] & (4.30) \\ &= \sum_k L_k^1(\widetilde{x}) \sum_{|\alpha|=m} \frac{1}{\alpha!} D^\alpha f(x)(x_k - x)^\alpha + \mathcal{O}(\Delta x^{m+1}) \ , & (4.31) \end{aligned}$$

since L_k^1 are uniformly bounded. We will use this representation of the error later in this section.

The derivative of Lagrange basis functions

We will also need to differentiate the Lagrange basis functions. Recall the relation

$$L_k^1(\widetilde{x}) = L_k^1\left(\frac{x}{h}\right) = L_k^1\left(\frac{x}{\Delta x}\right) \ .$$

Then

$$\frac{d}{dx} L_k^1(\widetilde{x}(x)) = \frac{d}{dx} L_k^1\left(\frac{x}{\Delta x}\right) = \frac{1}{\Delta x}(L_k^1)'(\widetilde{x}) \ .$$

Analysis of local truncation error

The main theorem of this section reads as follows.

Theorem 4.5
Let the WENO weights satisfy $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$, $\Delta x \to 0$, $l = 1, \ldots, n_s$, $i \in \mathbb{Z}$. *Furthermore, assume* $F \in \mathcal{C}^4(\mathcal{U})$, *such that* $F^{(k)}$ *is bounded for* $k = 1, \ldots, 4$. *Assume* $u \in \mathcal{C}^4(\overline{\Omega} \times [0,T])$ *and* $\frac{\Delta t}{\Delta x} = K < +\infty$ *remains constant.*
Then the scheme (4.13)-(4.20) applied on the scalar conservation law in one dimension (4.23) is of second order of consistency in time and space, i.e., it holds for the above defined local truncation error

$$T(x_i, t^n) = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ , \ \Delta t \to 0 \ , \ \Delta x \to 0 \ .$$

Proof.
The local truncation error of the scheme (4.13)-(4.20) is defined in (4.25) by

$$T(x_i,t^n) = \frac{u(x_i,t^n+\Delta t)-u(x_i,t^n)}{\Delta t} + \frac{1}{\Delta x}(g_{i+\frac{1}{2}} - g_{i-\frac{1}{2}}) \ .$$

We want to show

$$T(x_i,t^n) = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ , \ \Delta t \to 0 \ , \ \Delta x \to 0 \ .$$

The proof is divided into six subtasks.

I.

Consider the scheme (4.13)-(4.20). We start with the numerical flux $g_{i+\frac{1}{2}}$ and the reconstructions $R_i(x)$ therein. From the theory of polyharmonic spline interpolation, theorem 3.3 and from the formula (4.31), one gets

$$\begin{aligned} s_l^i(x) &= u(x,t^n) - d_l^i(x) + \mathcal{O}(\Delta x^3) \ , \\ (s_l^i)'(x) &= u_x(x,t^n) - (d_l^i)'(x) + \mathcal{O}(\Delta x^2) \ , \end{aligned}$$

where

$$d_l^i(x) = \sum_{k=1}^{n_s} L_{l,k}^{1,i}(\widetilde{x}) \frac{1}{2!} u_{xx}(x,t^n)(x_i - x)^2$$

is the leading term of the error and $L_{l,k}^{1,i}$ is the k-th Lagrange basis function on the l-th stencil with respect to i, on that we solve the interpolation problem. With 1 is denoted the scaling parameter $h=1$ from the theory of polyharmonic splines. Then

$$\begin{aligned} R_i(x) &= u(x,t^n) + \mathcal{R}_i(x) + \mathcal{O}(\Delta x^3) \ , \\ R_i'(x) &= u_x(x,t^n) + \mathcal{R}_i'(x) + \mathcal{O}(\Delta x^2) \ , \end{aligned}$$

where we used that WENO reconstruction is a convex combination of s_l^i and where we define

$$\begin{aligned} \mathcal{R}_i(x) &:= -\sum_{l=1}^{n_s} \omega_l^i d_l^i(x) \ , \\ \mathcal{R}_i'(x) &:= -\sum_{l=1}^{n_s} \omega_l^i (d_l^i)'(x) \ . \end{aligned}$$

II.
Due to the Max-Min-Principle for the solution of (4.23) (see e.g., [18]) we have

$$\begin{aligned} u_{i+\frac{1}{2}}^{(0)} &= RP_{i+\frac{1}{2}}(u_L^{(0)}, u_R^{(0)}) = u(x_{i+\frac{1}{2}},t^n) + \mathcal{R}_i(x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^3) \ , \\ u_{i+\frac{1}{2}}^{(1)} &= LRP_{i+\frac{1}{2}}(u_L^{(1)}, u_R^{(1)}) = u_x(x_{i+\frac{1}{2}},t^n) + \mathcal{R}_i'(x_{i+\frac{1}{2}}) + \mathcal{O}(\Delta x^2) \ . \end{aligned}$$

For the sake of simplicity, we usually omit to write the arguments of functions u, $\mathcal{R}$ etc. In those cases the argument will be always $(x_{i+\frac{1}{2}}, t^n)$. Then

$$\begin{aligned} u_{i+\frac{1}{2}}^{GRP}(x_{i+\frac{1}{2}}, t^n+\tau) &= u_{i+\frac{1}{2}}^{(0)} - \tau F'(u_{i+\frac{1}{2}}^{(0)}) u_{i+\frac{1}{2}}^{(1)} \\ &= u + \mathcal{R}_i(x) + \mathcal{O}(\Delta x^3) - \tau \ F'(u + \mathcal{R}_i + \mathcal{O}(\Delta x^3)) \ . \\ &\quad . \ \left(u_x + \mathcal{R}_i' + \mathcal{O}(\Delta x^2)\right) \ . \end{aligned}$$

Using Taylor expansion for F' one gets

$$\begin{aligned} u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n+\tau) &= u + \mathcal{R}_i + \mathcal{O}(\Delta x^3) \\ &\quad -\tau \ \left(F'(u) + (\mathcal{R}_i + \mathcal{O}(\Delta x^3))F''(u) + \mathcal{O}(\Delta x^4)\right) \ . \\ &\quad . \ \left(u_x + \mathcal{R}'_i + \mathcal{O}(\Delta x^2)\right) \\ &= u - \tau \ F(u)_x + \mathcal{R}_i - \tau F'(u)\mathcal{R}'_i - \tau \mathcal{R}_i F''(u) u_x \\ &\quad -\tau \mathcal{R}_i \mathcal{R}'_i F''(u) + \mathcal{O}(\Delta x^3) \ , \end{aligned}$$

where we used the fact $\tau = \mathcal{O}(\Delta t)$ and $\tau\mathcal{O}(\Delta x^2) = \mathcal{O}(\Delta x^3)$. The latter is true, since we assumed $\Delta x/\Delta t < +\infty$ to be constant. Further, $\mathcal{R}_i(x_{i+\frac{1}{2}}) = \mathcal{O}(\Delta x^2)$ and $\mathcal{R}'_i(x_{i+\frac{1}{2}}) = \mathcal{O}(\Delta x)$, see lemma 4.6. Then

$$u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n+\tau) \ = \ u - \tau \ F(u)_x + \mathcal{R}_i - \tau F'(u)\mathcal{R}'_i + \mathcal{O}(\Delta x^3) \ .$$

III.

$$F(u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n+\tau)) \ = \ F\left(u - \tau \ F(u)_x + \mathcal{R}_i - \tau F'(u)\mathcal{R}'_i + \mathcal{O}(\Delta x^3)\right) \ .$$

Using Taylor expansion again

$$\begin{aligned} F(u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n+\tau)) &= F(u) + \left(-\tau \ F(u)_x + \mathcal{R}_i - \tau F'(u)\mathcal{R}'_i + \mathcal{O}(\Delta x^3)\right) F'(u) \\ &\quad + \frac{1}{2}\left(-\tau \ F(u)_x + \mathcal{R}_i - \tau F'(u)\mathcal{R}'_i + \mathcal{O}(\Delta x^3)\right)^2 F''(u) + \mathcal{O}(\Delta x^3) \ . \end{aligned}$$

We use $F(u)_x = -u_t$ and $F'(u)u_t = F(u)_t$. After further simplifications

$$\begin{aligned} F(u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n+\tau)) &= F(u) + \tau \ F(u)_t + \mathcal{R}_i F'(u) - \tau \mathcal{R}'_i (F'(u))^2 \\ &\quad + \frac{1}{2}\tau^2 (F(u)_x)^2 F''(u) + \mathcal{O}(\Delta x^3) \ . \end{aligned}$$

IV.
According to Davis, Rabinowitz [9] (see also Davis [8]) the error of Gaussian numerical quadrature is

$$\int_a^b f - \sum_{s=1}^{q_t} w_s f(\tau_s) = \frac{(b-a)^{2q_t+1}(q_t!)^4}{(2q_t+1)[(2q_t)!]^3} f^{(2q_t)}(\xi) \ , \quad \xi \in (a,b) \ ,$$

where $a, b \in \mathbb{R}$, a function $f \in \mathcal{C}^{2q_t}[a,b]$ and where weights w_s and nodes τ_s define the Gaussian quadrature. In other words it holds

$$\sum_{s=1}^{q_t} w_s f(\tau_s) = \int_a^b f + \mathcal{O}((b-a)^{2q_t+1}) \text{ provided that } f^{(2q_t)} \text{ is bounded on } [a,b].$$

Then, application of this result on a Gaussian quadrature with $q_t = 2$, $[a,b] = [0, \Delta t]$ and $f(\cdot) = F(u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n + \cdot))$ on the time integral gives

$$\frac{1}{\Delta t}\sum_{s=1}^{q_t} w_s F(u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n+\tau_s)) \ = \ \frac{1}{\Delta t}\left\{\int_0^{\Delta t} F(u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n+\tau))d\tau + \mathcal{O}(\Delta t^5)\right\}$$

and

$$
\begin{aligned}
&\frac{1}{\Delta t}\sum_{s=1}^{q_t} w_s F(u^{GRP}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}, t^n+\tau_s)) \\
&= \frac{1}{\Delta t}\Bigg\{ \int_0^{\Delta t} \Big[F(u) + \tau\ F(u)_t + \mathcal{R}_i F'(u) - \tau \mathcal{R}_i'(F'(u))^2 + \\
&\qquad + \frac{1}{2}\tau^2 (F(u)_x)^2 F''(u) + \mathcal{O}(\Delta x^3)\Big] d\tau + \mathcal{O}(\Delta t^5)\Bigg\} \\
&= F(u) + \frac{\Delta t}{2}\ F(u)_t + \mathcal{R}_i F'(u) - \frac{\Delta t}{2}\mathcal{R}_i'(F'(u))^2 \\
&\qquad + \frac{\Delta t^2}{6}(F(u)_x)^2 F''(u) + \mathcal{O}(\Delta x^3)\ ,
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
g_{i+\frac{1}{2}} &= F(u) + \frac{\Delta t}{2}\ F(u)_t + \mathcal{R}_i F'(u) - \frac{\Delta t}{2}\mathcal{R}_i'(F'(u))^2 \\
&\quad + \frac{\Delta t^2}{6}(F(u)_x)^2 F''(u) + \mathcal{O}(\Delta x^3)\ .
\end{aligned}
$$

Define

$$
\mathfrak{R}_{i+\frac{1}{2}} := \mathcal{R}_i F'(u) - \frac{\Delta t}{2}\mathcal{R}_i'(F'(u))^2 + \frac{\Delta t^2}{6}(F(u)_x)^2 F''(u)\ ,
$$

then

$$
g_{i+\frac{1}{2}} = F(u) + \frac{\Delta t}{2}\ F(u)_t + \mathfrak{R}_{i+\frac{1}{2}} + \mathcal{O}(\Delta x^3)\ .
$$

Now, we expand the first two terms in a Taylor series at x_i and get

$$
\begin{aligned}
g_{i+\frac{1}{2}} &= F\left(u(x_i,t^n)\right) + (x_{i+\frac{1}{2}} - x_i)F(u)_x(x_i,t^n) + \frac{1}{2}(x_{i+\frac{1}{2}} - x_i)^2 F(u)_{xx}(x_i,t^n) \\
&\quad + \frac{\Delta t}{2}\ \Big(F(u)_t(x_i,t^n) + (x_{i+\frac{1}{2}} - x_i)F(u)_{tx}(x_i,t^n) + \mathcal{O}(\Delta x^2)\Big) \\
&\quad + \mathfrak{R}_{i+\frac{1}{2}} + \mathcal{O}(\Delta x^3) \\
&= F\left(u(x_i,t^n)\right) + \frac{1}{2}\Delta x F(u)_x(x_i,t^n) + \frac{1}{8}\Delta x^2 F(u)_{xx}(x_i,t^n) \\
&\quad + \frac{\Delta t}{2}\ \left(F(u)_t(x_i,t^n) + \frac{1}{2}\Delta x F(u)_{tx}(x_i,t^n)\right) + \mathfrak{R}_{i+\frac{1}{2}} + \mathcal{O}(\Delta x^3)\ .
\end{aligned}
$$

Analogously, the numerical flux $g_{i-\frac{1}{2}}$ can be expanded into

$$
\begin{aligned}
g_{i-\frac{1}{2}} &= F\left(u(x_i,t^n)\right) - \frac{1}{2}\Delta x F(u)_x(x_i,t^n) + \frac{1}{8}\Delta x^2 F(u)_{xx}(x_i,t^n) \\
&\quad + \frac{\Delta t}{2}\ \left(F(u)_t(x_i,t^n) - \frac{1}{2}\Delta x F(u)_{tx}(x_i,t^n)\right) + \mathfrak{R}_{i-\frac{1}{2}} + \mathcal{O}(\Delta x^3)\ .
\end{aligned}
$$

Then

$$
g_{i+\frac{1}{2}} - g_{i-\frac{1}{2}} = \Delta x F(u)_x(x_i,t^n) + \frac{\Delta t}{2}\Delta x F(u)_{tx}(x_i,t^n) + \mathfrak{R}_{i+\frac{1}{2}} - \mathfrak{R}_{i-\frac{1}{2}} + \mathcal{O}(\Delta x^3)\ ,
$$

$$
\begin{aligned}
\frac{1}{\Delta x}\left(g_{i+\frac{1}{2}} - g_{i-\frac{1}{2}}\right) &= F(u)_x(x_i,t^n) + \frac{\Delta t}{2}F(u)_{tx}(x_i,t^n) + \frac{1}{\Delta x}\left(\mathfrak{R}_{i+\frac{1}{2}} - \mathfrak{R}_{i-\frac{1}{2}}\right) \\
&\quad + \mathcal{O}(\Delta x^2)\ .
\end{aligned}
$$

If $\frac{1}{\Delta x}(\mathfrak{R}_{i+\frac{1}{2}} - \mathfrak{R}_{i-\frac{1}{2}}) = \mathcal{O}(\Delta x^2)$, then

$$\frac{1}{\Delta x}\left(g_{i+\frac{1}{2}} - g_{i-\frac{1}{2}}\right) = F(u)_x(x_i,t^n) + \frac{\Delta t}{2} F(u)_{tx}(x_i,t^n) + \mathcal{O}(\Delta x^2)\ .$$

Indeed, $\frac{1}{\Delta x}(\mathfrak{R}_{i+\frac{1}{2}} - \mathfrak{R}_{i-\frac{1}{2}}) = \mathcal{O}(\Delta x^2)$ holds, as stated in theorem 4.12.

V.
We expand the first part of the truncation error

$$\frac{1}{\Delta t}\left(u(x_i,t^n+\Delta t) - u(x_i,t^n)\right) = u_t(x_i,t^n) + \frac{\Delta t}{2} u_{tt}(x_i,t^n) + \mathcal{O}(\Delta t^2)\ .$$

VI.
Putting all together we get the truncation error

$$\begin{aligned}
T(x_i,t^n) &= \frac{1}{\Delta t}\left(u(x_i,t^n+\Delta t) - u(x_i,t^n)\right) + \frac{1}{\Delta x}\left(g_{i+\frac{1}{2}} - g_{i-\frac{1}{2}}\right)\\
&= u_t(x_i,t^n) + F(u)_x(x_i,t^n) + \frac{\Delta t}{2}\left[u_{tt}(x_i,t^n) + F(u)_{tx}(x_i,t^n)\right]\\
&\quad +\mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)\\
&= u_t(x_i,t^n) + F(u)_x(x_i,t^n) + \frac{\Delta t}{2}\Big[u_t(x_i,t^n) + F(u)_x(x_i,t^n)\Big]_t\\
&\quad +\mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)\ .
\end{aligned}$$

Since u is the exact solution of $u_t + F(u)_x = 0$, the corresponding terms vanish and we get

$$T(x_i,t^n) = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)\ ,$$

which finalizes the proof. □

The remainder error

In this technical subsection we are going to prove that for the remainder $\mathfrak{R}_{i+\frac{1}{2}} - \mathfrak{R}_{i-\frac{1}{2}}$ it holds $\mathfrak{R}_{i+\frac{1}{2}} - \mathfrak{R}_{i-\frac{1}{2}} = \mathcal{O}(\Delta x^3)$ as $\Delta x \to 0$.

$$\begin{aligned}
\mathfrak{R}_{i+\frac{1}{2}} &= \mathcal{R}_i F'(u) - \frac{\Delta t}{2}\mathcal{R}_i'(F'(u))^2 + \frac{\Delta t^2}{6}(F(u)_x)^2 F''(u)\ ,\\
\mathfrak{R}_{i-\frac{1}{2}} &= \mathcal{R}_{i-1} F'(u) - \frac{\Delta t}{2}\mathcal{R}_{i-1}'(F'(u))^2 + \frac{\Delta t^2}{6}(F(u)_x)^2 F''(u)\ ,
\end{aligned}$$

where we use the shortened notation for

$$\begin{aligned}
\mathfrak{R}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}) &= \mathcal{R}_i(x_{i+\frac{1}{2}})F'(u)(x_{i+\frac{1}{2}},t^n) - \frac{\Delta t}{2}\mathcal{R}_i'(x_{i+\frac{1}{2}})(F'(u))^2(x_{i+\frac{1}{2}},t^n)\\
&\quad +\frac{\Delta t^2}{6}(F(u)_x)^2(x_{i+\frac{1}{2}},t^n)F''(u)(x_{i+\frac{1}{2}},t^n)\ ,\\
\mathfrak{R}_{i-\frac{1}{2}}(x_{i-\frac{1}{2}}) &= \mathcal{R}_{i-1}(x_{i-\frac{1}{2}})F'(u)(x_{i-\frac{1}{2}},t^n) - \frac{\Delta t}{2}\mathcal{R}_{i-1}'(x_{i-\frac{1}{2}})(F'(u))^2(x_{i-\frac{1}{2}},t^n)\\
&\quad +\frac{\Delta t^2}{6}(F(u)_x)^2(x_{i-\frac{1}{2}},t^n)F''(u)(x_{i-\frac{1}{2}},t^n)\ .
\end{aligned}$$

Then

$$\begin{aligned}
&\mathfrak{R}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}) - \mathfrak{R}_{i-\frac{1}{2}}(x_{i-\frac{1}{2}})\\
&= \mathcal{R}_i(x_{i+\frac{1}{2}})F'(u)(x_{i+\frac{1}{2}},t^n) - \mathcal{R}_{i-1}(x_{i-\frac{1}{2}})F'(u)(x_{i-\frac{1}{2}},t^n)\\
&-\frac{\Delta t}{2}\left\{\mathcal{R}_i'(x_{i+\frac{1}{2}})(F'(u))^2(x_{i+\frac{1}{2}},t^n) - \mathcal{R}_{i-1}'(x_{i-\frac{1}{2}})(F'(u))^2(x_{i-\frac{1}{2}},t^n)\right\}\\
&+\frac{\Delta t^2}{6}\left\{(F(u)_x)^2(x_{i+\frac{1}{2}},t^n)F''(u)(x_{i+\frac{1}{2}},t^n) - (F(u)_x)^2(x_{i-\frac{1}{2}},t^n)F''(u)(x_{i-\frac{1}{2}},t^n)\right\}\ .
\end{aligned}$$

We estimate the three difference terms above. First, we introduce three auxiliary lemmata. In the next three lemmata, we estimate the difference terms. In the end, we formulate the estimate on $\mathfrak{R}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}) - \mathfrak{R}_{i-\frac{1}{2}}(x_{i-\frac{1}{2}})$. We will then also discuss the arising assumption $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$, $\Delta x \to 0$.

Lemma 4.6
Let $u \in \mathcal{C}^3(\overline{\Omega} \times [0,T])$. Then for $\Delta x \to 0$

$$\begin{aligned} \mathcal{R}_i(x_{i+\frac{1}{2}}) &= \mathcal{O}(\Delta x^2) \ , \\ \mathcal{R}_i'(x_{i+\frac{1}{2}}) &= \mathcal{O}(\Delta x) \ . \end{aligned}$$

Proof.
The proof follows immediatelly from the definitions of $\mathcal{R}_i$ and $\mathcal{R}_i'$ as a Taylor expansion remainder for polyharmonic splines, from boundedness of Lagrange functions and their derivatives. □

Lemma 4.7
Under the assumptions $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$, $\Delta x \to 0$, $l = 1, \ldots, n_s$, $i \in \mathbb{Z}$ and $u \in \mathcal{C}^3(\overline{\Omega} \times [0,T])$ it holds

$$\mathcal{R}_i(x_{i+\frac{1}{2}}) - \mathcal{R}_{i-1}(x_{i-\frac{1}{2}}) = \mathcal{O}(\Delta x^3) \ , \ \Delta x \to 0 \ .$$

Proof.
From the definitions

$$\begin{aligned} \mathcal{R}_i(x_{i+\frac{1}{2}}) &= -\sum_{l=1}^{n_s} \omega_l^i d_l^i(x_{i+\frac{1}{2}}) \ , \\ \mathcal{R}_{i-1}(x_{i-\frac{1}{2}}) &= -\sum_{l=1}^{n_s} \omega_l^{i-1} d_l^{i-1}(x_{i-\frac{1}{2}}) \ , \end{aligned}$$

where

$$\begin{aligned} d_l^i(x_{i+\frac{1}{2}}) &= \sum_{k=1}^{n_s} L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}}) \frac{1}{2!} u_{xx}(x_{i+\frac{1}{2}}, t^n)(x_i - x_{i+\frac{1}{2}})^2 \ , \\ d_l^{i-1}(x_{i-\frac{1}{2}}) &= \sum_{k=1}^{n_s} L_{l,k}^{1,i-1}(\widetilde{x}_{i-\frac{1}{2}}) \frac{1}{2!} u_{xx}(x_{i-\frac{1}{2}}, t^n)(x_{i-1} - x_{i-\frac{1}{2}})^2 \ , \end{aligned}$$

$L_{l,k}^{1,i}$ is the k-th Lagrange basis function on the l-th stencil with respect to i, on which we solve the interpolation problem. With 1 is meant the scaling parameter $h = 1$.

Then

$$\mathcal{R}_i(x_{i+\frac{1}{2}}) - \mathcal{R}_{i-1}(x_{i-\frac{1}{2}}) = -\sum_{l=1}^{n_s} \left[\omega_l^i (d_l^i(x_{i+\frac{1}{2}}) - d_l^{i-1}(x_{i-\frac{1}{2}})) + d_l^{i-1}(x_{i-\frac{1}{2}})(\omega_l^i - \omega_l^{i-1}) \right] \ .$$

Since $d_l^{i-1}(x_{i-\frac{1}{2}}) = \mathcal{O}(\Delta x^2)$ and from the assumption $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$, the second term in the sum is of order $\mathcal{O}(\Delta x^3)$.

For the first term we have

$$\begin{aligned} d_l^i(x_{i+\frac{1}{2}}) - d_l^{i-1}(x_{i-\frac{1}{2}}) &= \sum_{k=1}^{n_s} \Big\{ L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}}) \frac{1}{2!} u_{xx}(x_{i+\frac{1}{2}}, t^n)(x_i - x_{i+\frac{1}{2}})^2 \\ &\quad - L_{l,k}^{1,i-1}(\widetilde{x}_{i-\frac{1}{2}}) \frac{1}{2!} u_{xx}(x_{i-\frac{1}{2}}, t^n)(x_{i-1} - x_{i-\frac{1}{2}})^2 \Big\} \ . \end{aligned}$$

Due to the shift-invariance of the polyharmonic splines interpolation it holds for the Lagrange basis functions

$$L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}}) = L_{l,k}^{1,i-1}(\widetilde{x}_{i-\frac{1}{2}}) \ .$$

Further, we have $x_i - x_{i+\frac{1}{2}} = x_{i-1} - x_{i-\frac{1}{2}} = -\Delta x$, i.e.,

$$d_l^i(x_{i+\frac{1}{2}}) - d_l^{i-1}(x_{i-\frac{1}{2}}) = \sum_{k=1}^{n_s} L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}}) \frac{1}{2} \Delta x^2 (u_{xx}(x_{i+\frac{1}{2}}, t^n) - u_{xx}(x_{i-\frac{1}{2}}, t^n))$$

and

$$\left| d_l^i(x_{i+\frac{1}{2}}) - d_l^{i-1}(x_{i-\frac{1}{2}}) \right| \leq \frac{1}{2} \Delta x^2 \ \max_i \left| u_{xx}(x_{i+\frac{1}{2}}, t^n) - u_{xx}(x_{i-\frac{1}{2}}, t^n) \right| \ \sum_{k=1}^{n_s} \left| L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}}) \right| \ .$$

Since $\sum_{k=1}^{n_s} \left| L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}}) \right|$ is uniformly bounded and because of the assumptions on u, that u_{xx} is Lipschitz continuous in x on $\overline{\Omega}$, we have

$$\left| d_l^i(x_{i+\frac{1}{2}}) - d_l^{i-1}(x_{i-\frac{1}{2}}) \right| \leq C \Delta x^3$$

for some constant C independent of Δx.
Because the WENO weights satisfy $0 \leq \omega_l^i \leq 1$, it holds

$$\mathcal{R}_i(x_{i+\frac{1}{2}}) - \mathcal{R}_{i-1}(x_{i-\frac{1}{2}}) = \mathcal{O}(\Delta x^3) \ , \ \Delta x \to 0 \ .$$

□

Lemma 4.8
Under the assumptions $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$, $\Delta x \to 0$, $l = 1, \ldots, n_s$, $i \in \mathbb{Z}$ *and* $u \in \mathcal{C}^3(\overline{\Omega} \times [0, T])$,

$$\mathcal{R}_i'(x_{i+\frac{1}{2}}) - \mathcal{R}_{i-1}'(x_{i-\frac{1}{2}}) = \mathcal{O}(\Delta x^2) \ , \ \Delta x \to 0 \ .$$

Proof.
Consider the definitions

$$\begin{aligned} \mathcal{R}_i'(x_{i+\frac{1}{2}}) &= -\sum_{l=1}^{n_s} \omega_l^i (d_l^i)'(x_{i+\frac{1}{2}}) \ , \\ \mathcal{R}_{i-1}'(x_{i-\frac{1}{2}}) &= -\sum_{l=1}^{n_s} \omega_l^{i-1} (d_l^{i-1})'(x_{i-\frac{1}{2}}) \ . \end{aligned}$$

We have

$$\begin{aligned}
(d_l^i)'(x_{i+\frac{1}{2}}) &= \sum_{k=1}^{n_s}\Big\{\frac{1}{\Delta x}(L_{l,k}^{1,i})'(\widetilde{x}_{i+\frac{1}{2}})\frac{1}{2!}u_{xx}(x_{i+\frac{1}{2}},t^n)(x_i-x_{i+\frac{1}{2}})^2 \\
&\quad +L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}})\frac{1}{2!}\Big[u_{xxx}(x_{i+\frac{1}{2}},t^n)(x_i-x_{i+\frac{1}{2}})^2 \\
&\quad +u_{xx}(x_{i+\frac{1}{2}},t^n)2(x_i-x_{i+\frac{1}{2}})(-1)\Big]\Big\}\ , \\
(d_l^{i-1})'(x_{i-\frac{1}{2}}) &= \sum_{k=1}^{n_s}\Big\{\frac{1}{\Delta x}(L_{l,k}^{1,i-1})'(\widetilde{x}_{i-\frac{1}{2}})\frac{1}{2!}u_{xx}(x_{i-\frac{1}{2}},t^n)(x_{i-1}-x_{i-\frac{1}{2}})^2 \\
&\quad +L_{l,k}^{1,i-1}(\widetilde{x}_{i-\frac{1}{2}})\frac{1}{2!}\Big[u_{xxx}(x_{i-\frac{1}{2}},t^n)(x_{i-1}-x_{i-\frac{1}{2}})^2 \\
&\quad +u_{xx}(x_{i-\frac{1}{2}},t^n)2(x_{i-1}-x_{i-\frac{1}{2}})(-1)\Big]\Big\}\ .
\end{aligned}$$

The latter follows from

$$L_{l,k}^{1,i}(\widetilde{x}) = L_{l,k}^{1,i}\left(\frac{x}{\Delta x}\right)\ ,$$

see the beginning of this chapter, and therefore

$$\frac{d}{dx}L_{l,k}^{1,i}(\widetilde{x}(x))\Big|_{x=x_{i+\frac{1}{2}}} = \frac{1}{\Delta x}(L_{l,k}^{1,i})'(\widetilde{x}_{i+\frac{1}{2}})\ .$$

It holds

$$\mathcal{R}_i'(x_{i+\frac{1}{2}})-\mathcal{R}_{i-1}'(x_{i-\frac{1}{2}}) = -\sum_{l=1}^{n_s}\Big[\omega_l^i\Big((d_l^i)'(x_{i+\frac{1}{2}})-(d_l^{i-1})'(x_{i-\frac{1}{2}})\Big)+(d_l^{i-1})'(x_{i-\frac{1}{2}})\left(\omega_l^i-\omega_l^{i-1}\right)\Big]\ .$$

Since $(d_l^{i-1})'(x_{i-\frac{1}{2}}) = \mathcal{O}(\Delta x)$ and from the assumption $\omega_l^i-\omega_l^{i-1} = \mathcal{O}(\Delta x)$, the second term in the sum is of order $\mathcal{O}(\Delta x^2)$.
For the first term we have

$$\begin{aligned}
&(d_l^i)'(x_{i+\frac{1}{2}})-(d_l^{i-1})'(x_{i-\frac{1}{2}}) \\
&\quad = \sum_{k=1}^{n_s}\frac{1}{\Delta x}\frac{1}{2}(L_{l,k}^{1,i})'(\widetilde{x}_{i+\frac{1}{2}})\Delta x^2\Big[u_{xx}(x_{i+\frac{1}{2}},t^n)-u_{xx}(x_{i-\frac{1}{2}},t^n)\Big] \\
&\qquad +\sum_{k=1}^{n_s}\frac{1}{2}\Delta x^2 L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}})\Big[u_{xxx}(x_{i+\frac{1}{2}},t^n)-u_{xxx}(x_{i-\frac{1}{2}},t^n)\Big] \\
&\qquad +\sum_{k=1}^{n_s}(-1)\Delta x L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}})\Big[u_{xx}(x_{i+\frac{1}{2}},t^n)-u_{xx}(x_{i-\frac{1}{2}},t^n)\Big]\ ,
\end{aligned}$$

where we used the following consideration: From the shift-invariance of the polyharmonic splines interpolation, the Lagrange basis functions satisfy

$$(L_{l,k}^{1,i})'(\widetilde{x}_{i+\frac{1}{2}}) = (L_{l,k}^{1,i-1})'(\widetilde{x}_{i-\frac{1}{2}})\ ,$$

where we used $x_i-x_{i+\frac{1}{2}} = x_{i-1}-x_{i-\frac{1}{2}} = -\Delta x$.

$$\begin{aligned}
&\Big|(d_l^i)'(x_{i+\frac{1}{2}})-(d_l^{i-1})'(x_{i-\frac{1}{2}})\Big| \\
&\quad \le \max_i\Big|u_{xx}(x_{i+\frac{1}{2}},t^n)-u_{xx}(x_{i-\frac{1}{2}},t^n)\Big|\frac{1}{2}\Delta x\sum_{k=1}^{n_s}\Big|(L_{l,k}^{1,i})'(\widetilde{x}_{i+\frac{1}{2}})\Big| \\
&\quad +\max_i\Big|u_{xxx}(x_{i+\frac{1}{2}},t^n)-u_{xxx}(x_{i-\frac{1}{2}},t^n)\Big|\frac{1}{2}\Delta x^2\sum_{k=1}^{n_s}\Big|L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}})\Big| \\
&\qquad +\max_i\Big|u_{xx}(x_{i+\frac{1}{2}},t^n)-u_{xx}(x_{i-\frac{1}{2}},t^n)\Big|\Delta x\sum_{k=1}^{n_s}\Big|L_{l,k}^{1,i}(\widetilde{x}_{i+\frac{1}{2}})\Big|\ .
\end{aligned}$$

The term $\sum_{k=1}^{n_s}\left|(L_{l,k}^{1,i})'\right|$ is uniformly bounded since the polyharmonic spline interpolation $s_l^i(x) = \sum_{k=1}^{n_s} L_{l,k}^{1,i}(x)\lambda_k(u)$ (the Lagrange representation) is at least $\mathcal{C}^1(\overline{\Omega})$. Because of the assumptions on u, u_{xx} is Lipschitz continuous and u_{xxx} is bounded on $\overline{\Omega}$. Hence, we have

$$\left|(d_l^i)'(x_{i+\frac{1}{2}}) - (d_l^{i-1})'(x_{i-\frac{1}{2}})\right| \quad \leq \quad C\Delta x^2$$

for some constant C independent of Δx.
Because the WENO weights satisfy $0 \leq \omega_l^i \leq 1$, it holds

$$\mathcal{R}_i'(x_{i+\frac{1}{2}}) - \mathcal{R}_{i-1}'(x_{i-\frac{1}{2}}) = \mathcal{O}(\Delta x^2)\ ,\ \Delta x \to 0\ .$$

☐

Now, we estimate the three difference terms of $\mathfrak{R}_{i+\frac{1}{2}} - \mathfrak{R}_{i-\frac{1}{2}}$ with the help of the previous three lemmata.

Lemma 4.9
Let F' be Lipschitz-continuous and bounded on $\mathcal{U} \subset \mathbb{R}$, $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$, $\Delta x \to 0$, $l = 1, \ldots, n_s$, $i \in \mathbb{Z}$ and $u \in \mathcal{C}^3(\overline{\Omega} \times [0,T])$. Then

$$\mathcal{R}_i(x_{i+\frac{1}{2}})F'(u)(x_{i+\frac{1}{2}}, t^n) - \mathcal{R}_{i-1}(x_{i-\frac{1}{2}})F'(u)(x_{i-\frac{1}{2}}, t^n) = \mathcal{O}(\Delta x^3)\ ,\ \Delta x \to 0\ .$$

Proof.

$$\begin{aligned}
&\left|\mathcal{R}_i(x_{i+\frac{1}{2}})F'(u)(x_{i+\frac{1}{2}}, t^n) - \mathcal{R}_{i-1}(x_{i-\frac{1}{2}})F'(u)(x_{i-\frac{1}{2}}, t^n)\right| \\
&\leq \left|\mathcal{R}_i(x_{i+\frac{1}{2}})\right|\left|F'(u)(x_{i+\frac{1}{2}}, t^n) - F'(u)(x_{i-\frac{1}{2}}, t^n)\right| \\
&\quad + \left|F'(u)(x_{i-\frac{1}{2}}, t^n)\right|\left|\mathcal{R}_i(x_{i+\frac{1}{2}}) - \mathcal{R}_{i-1}(x_{i-\frac{1}{2}})\right| \\
&\leq C\Delta x^3\ ,
\end{aligned}$$

which follows from lemmata 4.6, 4.7 and from the assumptions on F and u. ☐

Lemma 4.10
Let F' be Lipschitz-continuous and bounded on $\mathcal{U} \subset \mathbb{R}$, $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$, $\Delta x \to 0$, $l = 1, \ldots, n_s$, $i \in \mathbb{Z}$ and $u \in \mathcal{C}^3(\overline{\Omega} \times [0,T])$. Then

$$\mathcal{R}_i'(x_{i+\frac{1}{2}})(F'(u))^2(x_{i+\frac{1}{2}}, t^n) - \mathcal{R}_{i-1}'(x_{i-\frac{1}{2}})(F'(u))^2(x_{i-\frac{1}{2}}, t^n) = \mathcal{O}(\Delta x^2)\ ,\ \Delta x \to 0\ .$$

Proof.

$$\begin{aligned}
&\left|\mathcal{R}_i'(x_{i+\frac{1}{2}})(F'(u))^2(x_{i+\frac{1}{2}}, t^n) - \mathcal{R}_{i-1}'(x_{i-\frac{1}{2}})(F'(u))^2(x_{i-\frac{1}{2}}, t^n)\right| \\
&\leq \left|\mathcal{R}_i'(x_{i+\frac{1}{2}})\right|\left|(F'(u))^2(x_{i+\frac{1}{2}}, t^n) - (F'(u))^2(x_{i-\frac{1}{2}}, t^n)\right| \\
&\quad + \left|(F'(u))^2(x_{i-\frac{1}{2}}, t^n)\right|\left|\mathcal{R}_i'(x_{i+\frac{1}{2}}) - \mathcal{R}_{i-1}'(x_{i-\frac{1}{2}})\right| \\
&\leq C\Delta x^2\ ,
\end{aligned}$$

which follows from lemmata 4.6, 4.8, from the formula $a^2 - b^2 = (a+b)(a-b)$ and from the assumptions on F and u. $\square$

Lemma 4.11
Let F'' be Lipschitz-continuous and bounded on $\mathcal{U} \subset \mathbb{R}$ and $u \in \mathcal{C}^2(\overline{\Omega} \times [0,T])$. Then

$$(F(u)_x)^2(x_{i+\frac{1}{2}}, t^n)F''(u)(x_{i+\frac{1}{2}}, t^n) - (F(u)_x)^2(x_{i-\frac{1}{2}}, t^n)F''(u)(x_{i-\frac{1}{2}}, t^n) = \mathcal{O}(\Delta x) \ , \qquad \Delta x \to 0 \ .$$

Proof.
Since u is the exact solution of the conservation law, we have $F(u)_x = -u_t$.

$$\begin{aligned}
&\left|u_t^2(x_{i+\frac{1}{2}}, t^n)F''(u)(x_{i+\frac{1}{2}}, t^n) - u_t^2(x_{i-\frac{1}{2}}, t^n)F''(u)(x_{i-\frac{1}{2}}, t^n)\right| \\
&\quad \le \left|u_t^2(x_{i+\frac{1}{2}}, t^n)\right|\left|F''(u)(x_{i+\frac{1}{2}}, t^n) - F''(u)(x_{i-\frac{1}{2}}, t^n)\right| \\
&\quad + \left|F''(u)(x_{i-\frac{1}{2}}, t^n)\right|\left|u_t^2(x_{i+\frac{1}{2}}, t^n) - u_t^2(x_{i-\frac{1}{2}}, t^n)\right| \le C\Delta x
\end{aligned}$$

which is a result of the formula $a^2 - b^2 = (a+b)(a-b)$ and the assumptions on F and u. $\square$

Now we can state the theorem concerning the remainder error.

Theorem 4.12
Let $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$, $\Delta x \to 0$, $l = 1, \dots, n_s$, $i \in \mathbb{Z}$ and assume $F : \mathcal{U} \to \mathbb{R}$ s.t. F' is bounded and F'' is bounded and Lipschitz-continuous. Assume $u \in \mathcal{C}^3(\overline{\Omega} \times [0,T])$. If $\frac{\Delta t}{\Delta x} = K < +\infty$ remains constant, then

$$\mathfrak{R}_{i+\frac{1}{2}} - \mathfrak{R}_{i-\frac{1}{2}} = \mathcal{O}(\Delta x^3) \ , \ \Delta x \to 0 \ .$$

Proof.
Since

$$\begin{aligned}
&\mathfrak{R}_{i+\frac{1}{2}}(x_{i+\frac{1}{2}}) - \mathfrak{R}_{i-\frac{1}{2}}(x_{i-\frac{1}{2}}) \\
&= \mathcal{R}_i(x_{i+\frac{1}{2}})F'(u)(x_{i+\frac{1}{2}}, t^n) - \mathcal{R}_{i-1}(x_{i-\frac{1}{2}})F'(u)(x_{i-\frac{1}{2}}, t^n) \\
&- \frac{\Delta t}{2}\left\{\mathcal{R}_i'(x_{i+\frac{1}{2}})(F'(u))^2(x_{i+\frac{1}{2}}, t^n) - \mathcal{R}_{i-1}'(x_{i-\frac{1}{2}})(F'(u))^2(x_{i-\frac{1}{2}}, t^n)\right\} \\
&+ \frac{\Delta t^2}{6}\left\{(F(u)_x)^2(x_{i+\frac{1}{2}}, t^n)F''(u)(x_{i+\frac{1}{2}}, t^n) - (F(u)_x)^2(x_{i-\frac{1}{2}}, t^n)F''(u)(x_{i-\frac{1}{2}}, t^n)\right\} \ ,
\end{aligned}$$

the statement of the theorem is a direct result of lemmata 4.9, 4.10, 4.11 and the assumption that $\frac{\Delta t}{\Delta x} = K$ remains constant. $\square$

Let us discuss the assumption

$$\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x) \ , \Delta x \to 0 \qquad (4.32)$$

on the WENO weights. If the data and the solution of (4.23) are smooth, it is reasonable to assume that ω_l^i and ω_l^{i-1} do not vary "a lot" which is the meaning of the condition. Indeed, in the case of parameters $n_s = 2$, $k = 2$ and a linear governing PDE the scheme (4.13)-(4.20) becomes a linear scheme (a scheme with constant coefficients) and all $\omega_l^i = 1/n_s$. Then $\omega_l^i - \omega_l^{i-1} = 0$ and (4.32) holds. The following lemma states the validity of condition (4.32) in the case $n_s = 3$, $k = 2$.

Lemma 4.13
Consider the scheme (4.13)-(4.20) for parameters $n_s = 3$, $k = 2$ and for the scalar linear equation

$$u_t + au_x = 0 \ , \ a > 0.$$

Let $u \in \mathcal{C}^3(\overline{\Omega} \times [0,T])$ and assume $u_{xx} \neq 0$ on $\overline{\Omega} \times [0,T]$.
Then

$$\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x) \ , \Delta x \to 0 \quad \forall l = 1, \ldots, n_s \ , \ i \in \mathbb{Z} \ .$$

Proof.
We want to avoid the complicated notation of $\widetilde{\omega}_l^i$ in (3.18). Instead, we will use the notation $\widetilde{\omega}_j$, since in the case $n_s = 3$, $k = 2$ we can describe ω_l^i in a simpler way. So let

$$\widetilde{\omega}_j = \frac{1}{\left(\varepsilon + |s_{(j)}|^2_{BL^2(\mathbb{R})}\right)^2} \ ,$$

where $s_{(j)}$ is the interpolant (3.14) based on data u_{j-1}, u_j and u_{j+1}, i.e., on the stencil $(j-1, j, j+1)$ with the center j. Then

$$\omega_l^i = \frac{\widetilde{\omega}_{i-2+l}}{\widetilde{\omega}_{i-1} + \widetilde{\omega}_i + \widetilde{\omega}_{i+1}} \ , \tag{4.33}$$

$$\omega_l^{i-1} = \frac{\widetilde{\omega}_{i-3+l}}{\widetilde{\omega}_{i-2} + \widetilde{\omega}_{i-1} + \widetilde{\omega}_i} \ . \tag{4.34}$$

Using this special notation we can directly see that some terms on right hand side are the same for both ω_l^i and ω_l^{i-1}.
We want to prove $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x)$ for $l = 1, 2, 3$. Due to the symmetry of the terms in (4.33)-(4.34) it is enough to prove it for one fixed $l \in \{1, 2, 3\}$, the other cases follow analogously. Let us consider $l = 1$.
Recall equation (3.20), i.e.,

$$|s_{(j)}|^2_{BL^2(\mathbb{R})} = c_{(j)}^T A_{(j)} c_{(j)}$$

for the corresponding coefficients vector $c_{(j)}$ and matrix $A_{(j)}$.
The direct computation for $n_s = 3$, $k = 2$ gives in the case of linear governing PDE due to the identity (4.22)

$$c_{(j)}^T A_{(j)} c_{(j)} = \frac{\mathcal{D}_0}{\Delta x^3} \left(u_{j-1} - 2u_j + u_{j+1}\right)^2 \quad , \quad \mathcal{D}_0 = \frac{105}{604} \ . \tag{4.35}$$

For the consistency analysis we assume, that instead of data u_j we look for an interpolation with data $u(x_j, t^n)$, where u is the exact solution of (4.23). Then from (4.35) we deduce

$$c_{(j)}^T A_{(j)} c_{(j)} = \frac{\mathcal{D}_0}{\Delta x} \left(u_{xx}(x_j, t^n) + \mathcal{O}(\Delta x^2)\right)^2 \ ,$$

so that

$$\widetilde{\omega}_j = \frac{1}{\left(\varepsilon + \frac{\mathcal{D}_0}{\Delta x}[u_{xx}(x_j, t^n) + \mathcal{O}(\Delta x^2)]^2\right)^2} = \frac{\Delta x^2}{\left(\varepsilon \Delta x + \mathcal{D}_0[u_{xx}(x_j, t^n) + \mathcal{O}(\Delta x^2)]^2\right)^2} \ .$$

Define

$$\alpha_j = \left(\varepsilon\Delta x + \mathcal{D}_0[u_{xx}(x_j,t^n) + \mathcal{O}(\Delta x^2)]^2\right)^2 \ .$$

Then

$$\widetilde{\omega}_j = \frac{\Delta x^2}{\alpha_j}$$

and (4.33)-(4.34) for $l = 1$ reads

$$\begin{aligned} \omega_1^i &= \frac{\alpha_i\alpha_{i+1}}{\alpha_{i-1}\alpha_i + \alpha_i\alpha_{i+1} + \alpha_{i-1}\alpha_{i+1}} \ , \\ \omega_1^{i-1} &= \frac{\alpha_{i-1}\alpha_i}{\alpha_{i-2}\alpha_{i-1} + \alpha_{i-1}\alpha_i + \alpha_{i-2}\alpha_i} \ . \end{aligned}$$

Furthermore,

$$\omega_1^i - \omega_1^{i-1} = \frac{\alpha_i^2(\alpha_{i-1}^2 - \alpha_{i+1}\alpha_{i-2}) + \alpha_{i-1}\alpha_i\alpha_{i+1}(\alpha_{i-1} - \alpha_{i-2})}{(\alpha_{i-1}\alpha_i + \alpha_i\alpha_{i+1} + \alpha_{i-1}\alpha_{i+1})(\alpha_{i-2}\alpha_{i-1} + \alpha_{i-1}\alpha_i + \alpha_{i-2}\alpha_i)} \ .$$

Since $u \in \mathcal{C}^3$, the terms α_j can be rewritten in the form

$$\alpha_j = \mathcal{D}_0^2[u_{xx}]^4(x_j,t^n) + \mathcal{O}(\Delta x) \ .$$

Then, from the assumption $u_{xx} \neq 0$ it follows for the denominator

$$(\alpha_{i-1}\alpha_i + \alpha_i\alpha_{i+1} + \alpha_{i-1}\alpha_{i+1})(\alpha_{i-2}\alpha_{i-1} + \alpha_{i-1}\alpha_i + \alpha_{i-2}\alpha_i) = \mathcal{O}(1) \ .$$

To complete the proof it remains to show that

$$\alpha_i^2(\alpha_{i-1}^2 - \alpha_{i+1}\alpha_{i-2}) + \alpha_{i-1}\alpha_i\alpha_{i+1}(\alpha_{i-1} - \alpha_{i-2}) = \mathcal{O}(\Delta x)$$

holds for the numerator. Indeed,

$$\begin{aligned} &|\alpha_i^2(\alpha_{i-1}^2 - \alpha_{i+1}\alpha_{i-2}) + \alpha_{i-1}\alpha_i\alpha_{i+1}(\alpha_{i-1} - \alpha_{i-2})| \\ &\leq |\alpha_i^2||\alpha_{i-1}^2 - \alpha_{i+1}\alpha_{i-2}| + |\alpha_{i-1}\alpha_i\alpha_{i+1}||\alpha_{i-1} - \alpha_{i-2}| \ . \end{aligned}$$

Since $u \in \mathcal{C}^3$, the following terms are bounded

$$\begin{aligned} |\alpha_i^2| &\leq C_1 \ , \\ |\alpha_{i-1}\alpha_i\alpha_{i+1}| &\leq C_2 \end{aligned}$$

and since u_{xx} is Lipschitz-continuous at x

$$|\alpha_{i-1} - \alpha_{i-2}| \leq \mathcal{O}(\Delta x) \ ,$$

$$\begin{aligned} |\alpha_{i-1}^2 - \alpha_{i+1}\alpha_{i-2}| &\leq |\alpha_{i-1}^2 - \alpha_{i-1}\alpha_{i+1} + \alpha_{i-1}\alpha_{i+1} - \alpha_{i+1}\alpha_{i-2}| \\ &\leq |\alpha_{i-1}||\alpha_{i-1} - \alpha_{i+1}| + |\alpha_{i+1}||\alpha_{i-1} - \alpha_{i-2}| \leq \mathcal{O}(\Delta x) \ . \end{aligned}$$

Altogether

$$\alpha_i^2(\alpha_{i-1}^2 - \alpha_{i+1}\alpha_{i-2}) + \alpha_{i-1}\alpha_i\alpha_{i+1}(\alpha_{i-1} - \alpha_{i-2}) = \mathcal{O}(\Delta x) \ ,$$

i.e.,

$$\omega_1^i - \omega_1^{i-1} = \mathcal{O}(\Delta x) \ , \ \Delta x \to 0 \ ,$$

which completes the proof. □

Remark 4.14
In the case of $u_{xx} \equiv 0$ *the claim of lemma 4.13 remains valid. In that case it holds*

$$\omega_l^i = \frac{1}{n_s} = \frac{1}{3} \quad \forall l = 1,\ldots,n_s \ , \forall i \in \mathbb{Z} \ ,$$

and therefore $\omega_l^i - \omega_l^{i-1} = \mathcal{O}(\Delta x) \ , \ \Delta x \to 0.$

4.3 Stability analysis

Consider the linear advection equation

$$u_t + au_x = 0\ ,\ \ a > 0\ . \tag{4.36}$$

The scheme (4.13)-(4.20) applied on (4.36) can be in the case $n_s = 2$, $k = 2$ through direct computation (see (4.21)) rewritten for all $i \in \mathbb{Z}$ into the form

$$u_i^{n+1} = \sum_{j=-2}^{1} b_j u_{i+j}^n \tag{4.37}$$

with

$$\begin{aligned} b_{-2} &= \left(-\frac{1}{4}\right)\nu(1-\nu)\ , & (4.38)\\ b_{-1} &= \frac{1}{4}\nu(5-\nu)\ , & (4.39)\\ b_0 &= 1-\frac{1}{4}\nu(3+\nu)\ , & (4.40)\\ b_1 &= \left(-\frac{1}{4}\right)\nu(1-\nu) & (4.41) \end{aligned}$$

and

$$\nu = a\frac{\Delta t}{\Delta x}\ .$$

Remark 4.15
In the case $a < 0$ the resulting scheme has similar form and can be treated in the same way as it follows for $a > 0$. For the sake of simplicity, the case $a < 0$ is not considered in this thesis.

We are going to investigate the stability of the scheme. Roughly speaking, stability means, that the global error of numerical solution defined by the scheme does not grow "catastrophically" and can be bounded. We will define the term of *stability* and introduce a stability result in the sense of L^2-norm, also known as *von Neumann stability.* To this end, a theoretical procedure for investigating a scheme on L^2-stability will be presented, based on books by Strikwerda [55] and LeVeque [39], where it is presented for finite difference and finite volume methods. We will see that it is also possible to apply the principle on a finite volume particle method.

For $v = \{v_i\}_{i=-\infty}^{\infty}$ consider the norm

$$\|v\|_{2,\Delta x}^2 = \sum_{i=-\infty}^{\infty} \Delta x |v_i|^2\ ,$$

which approximates and is equal to the norm

$$\|f\|_{L^2(\mathbb{R})}^2 = \int_{-\infty}^{\infty} |f|^2$$

in the case of finite differences and finite volumes, respectively.

Definition 4.16
We say that a numerical scheme

$$u_i^{n+1} = \sum_{j=-L}^{R} b_j u_{i+j}^n\ , \quad L, R \in I\!N_0 \tag{4.42}$$

is stable *in the norm* $\|\cdot\|$, *if*

$$\|u^{n+1}\| \leq \|u^0\| \qquad \forall\ n \in I\!N_0\ ,$$

where

$$u^n := \{u_i^n\}_i\ .$$

Stability theory

We will use the idea of *von Neumann stability*, i.e., stability in L^2-norm, described in LeVeque [39] and in more detail in Strikwerda [55].
For the discrete function u^n we assume that there exists its Fourier transform

$$\hat{u}^n(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{i=-\infty}^{\infty} e^{-I\xi i\Delta x} u_i^n \Delta x$$

for $\xi \in \mathbb{R}$, where I is the complex unit (i.e., $I^2 = -1$). This is true if, for instance, $u^n \in l_1$.

Remark 4.17
We emphasize that the symbol i stands for an integer and the symbol I for the complex unit in this section.

The Fourier inversion formula is given by

$$u_i^n = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{I\xi i\Delta x} \hat{u}^n(\xi) d\xi\ . \tag{4.43}$$

Application of the difference method (4.42) on u_i^n and manipulations with the terms typically gives the expression

$$u_i^{n+1} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{u}^n(\xi) g(\xi, \Delta x, \Delta t) e^{I\xi i\Delta x} d\xi\ .$$

This compared to

$$u_i^{n+1} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{I\xi i\Delta x} \hat{u}^{n+1}(\xi) d\xi$$

given by (4.43) yields

$$\hat{u}^{n+1}(\xi) = g(\xi, \Delta x, \Delta t) \hat{u}^n(\xi)\ . \tag{4.44}$$

We used also the fact, that the Fourier transform is unique.
Further, the idea utilizes the *Parseval's relation*

$$\|u^n\|_{2,\Delta x} = \|\hat{u}^n\|_2\ .$$

So, in order to show that u^n is bounded, it is sufficient to show that the Fourier transform $\hat{u}^n$ remains bounded. For this reason we apply (4.44) iteratively and get

$$\hat{u}_i^{n+1}(\xi) = g^{n+1}(\xi, \Delta x, \Delta t) \hat{u}^0(\xi)\ .$$

If the condition

$$|g(\xi, \Delta x, \Delta t)| \leq 1 \tag{4.45}$$

holds, then

$$\|\hat{u}^{n+1}\|_2 \leq \|\hat{u}^0\|_2$$

and finally from the Parseval's relation we also obtain

$$\|u^{n+1}\|_{2,\Delta x} \leq \|u^0\|_{2,\Delta x} \ .$$

So, in order to investigate stability of the scheme (4.42) it is sufficient to verify the condition (4.45). However, if condition (4.45) is violated, the scheme will not be stable.
Alternatively, $g(\xi, \Delta x, \Delta t)$ can be determined by plugging $u_i^n = e^{I\xi i \Delta x}$ into the formula (4.42). It is a usual shortcut of the von Neumann analysis.

Stability of the scheme (4.37)

Due to the discussion above we will consider the case

$$u_{i+j}^n = e^{(i+j)\xi I \Delta x} \ , \tag{4.46}$$

where I is the imaginary unit.
We plug (4.46) into (4.37) and get

$$u_i^{n+1} = \left(\sum_{j=-2}^{1} b_j e^{j\xi I \Delta x} \right) e^{i\xi I \Delta x} = \left(\sum_{j=-2}^{1} b_j e^{j\xi I \Delta x} \right) u_i^n \ .$$

We define

$$g(\xi, \Delta x, \Delta t) = \sum_{j=-2}^{1} b_j e^{j\xi I \Delta x} \ , \tag{4.47}$$

which leads to

$$u_i^{n+1} = g(\xi, \Delta x, \Delta t) u_i^n \ .$$

If we can show

$$|g(\xi, \Delta x, \Delta t)| \leq 1,$$

then

$$\|u^{n+1}\|_{2,\Delta x} \leq \|u^n\|_{2,\Delta x} \ ,$$

which ensures the stability, since the repeated application of this inequality gives

$$\|u^{n+1}\|_{2,\Delta x} \leq \|u^0\|_{2,\Delta x} \ .$$

An illustration of sets $\{g(\xi, \Delta x, \Delta t) \mid \xi \in \mathbb{R}\}_\nu \subset \mathbb{C}$ for $\nu \in \{0, 1, \dots, 10\}$, where $\nu = a\frac{\Delta t}{\Delta x}$ is given in figure 4.4 promising stability for $0 \leq \nu \leq 1$. Now we are going to prove it analytically.

Lemma 4.18
Consider the scheme (4.37). Then, for $g = g(\xi, \Delta x, \Delta t)$ defined in (4.47), the following identity holds:

$$\begin{aligned}
|g|^2 \quad = \quad & \left(1 - \frac{1}{4}\nu(3+\nu)\right)^2 + \frac{1}{16}\nu^2(1-\nu)^2 + (-3+\nu)^2\frac{1}{4}\nu^2 \\
& + \frac{1}{2}\nu(1-\nu)\left(1 - \frac{1}{4}\nu(3+\nu)\right) \\
& + \left[2\nu(1 - \frac{1}{4}\nu(3+\nu)) + \frac{1}{2}\nu^2(1-\nu)(-3+\nu) + \frac{1}{2}\nu^2(1-\nu)\right] cos(\xi\Delta x) \\
& + \left[\nu^2 - \frac{1}{4}\nu^2(-3+\nu)^2 - \nu(1-\nu)(1 - \frac{1}{4}\nu(3+\nu))\right] cos^2(\xi\Delta x) \\
& + \frac{1}{2}\nu^2(1-\nu)^2\cos^3(\xi\Delta x) \ .
\end{aligned}$$

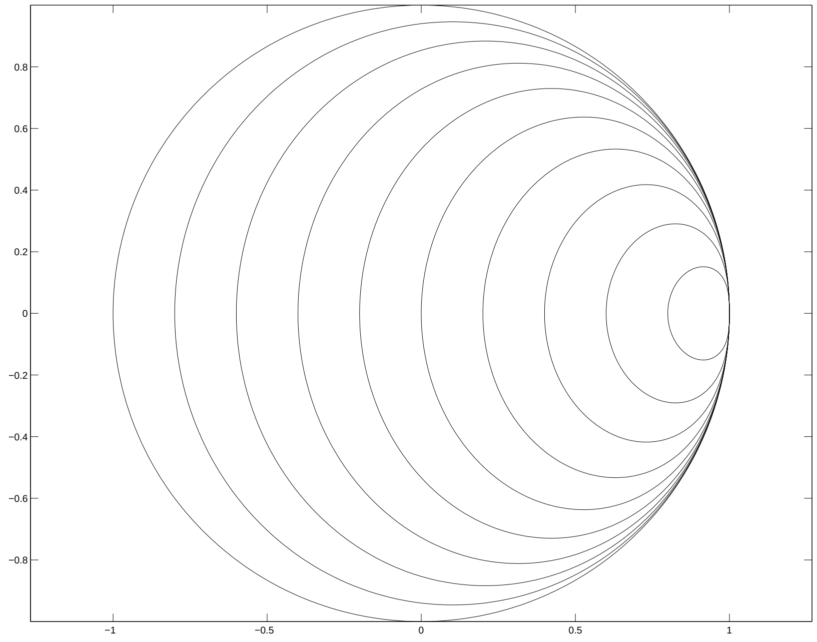


Figure 4.4: *An illustration of sets* $g(\xi, \Delta x, \Delta t)$*. Sets* $\{g(\xi, \Delta x, \Delta t) \mid \xi \in \mathbb{R}\}_\nu \subset \mathbb{C}$ *are depicted for* $\nu \in \{0, 1, \ldots, 10\}$*, where* $\nu = a\frac{\Delta t}{\Delta x}$*. For* $\nu = 1$ *one gets the unit circle (corresponding to the case* $|g(\xi, \Delta x, \Delta t)| = 1$*), for* $\nu = 0$ *the set is equal to the point* 1*.*

Proof.
Denote $g = a + Ib$. Then, $|g|^2 = a^2 + b^2$. From the identity $e^{\alpha I \xi \Delta x} = \cos(\alpha\xi\Delta x) + I\sin(\alpha\xi\Delta x)$ for $\alpha \in \mathbb{R}$ and using $\cos(-y) = \cos(y)$, $\sin(-y) = -\sin(y)$, $\cos(2y) = 2\cos^2(y) - 1$, $\sin(2y) = 2\sin(y)\cos(y)$ one gets

$$\begin{aligned} a &= 1 - \frac{1}{4}\nu(3+\nu) - \frac{1}{4}\nu(1-\nu)(2\cos^2(\xi\Delta x) - 1) + \nu\cos(\xi\Delta x) \ , \\ b &= \frac{1}{2}\nu\sin(\xi\Delta x)\left[(1-\nu)\cos(\xi\Delta x) - 3 + \nu\right] \ . \end{aligned}$$

Some simple algebraic manipulations, such as using $\sin^2(y) = 1 - \cos^2(y)$ and transforming the result to the form $\sum_{k=0}^{3} a_k \cos^k(\xi\Delta x)$, lead to the statement of the lemma. □

The function $|g|^2(\xi, \Delta x, \Delta t)$ of variables ξ, Δx and Δt seems still quite complicated to analyse. But we should notice, that the values of the functions $|\cos^k(\xi\Delta x)|$ are bounded by 1 for all $\xi \in \mathbb{R}$, $\Delta x > 0$. So, if we look for a maximum of $|g|^2$, we should look for which values of $\cos^k(\xi\Delta x)$ there is the maximum. To this end, we use the following technique introduced by Knobloch in [34], who used it to investigate stability of another scheme.
Define a polynomial

$$\varphi_\nu(c) = \sum_{k=0}^{3} a_k^\nu c^k$$

with the coefficients depending on parameter ν

$$\begin{aligned}
a_0^\nu &= \left(1-\frac{1}{4}\nu(3+\nu)\right)^2+\frac{1}{16}\nu^2(1-\nu)^2+(-3+\nu)^2\frac{1}{4}\nu^2 \\
&\quad +\frac{1}{2}\nu(1-\nu)\left(1-\frac{1}{4}\nu(3+\nu)\right) , \\
a_1^\nu &= 2\nu(1-\frac{1}{4}\nu(3+\nu))+\frac{1}{2}\nu^2(1-\nu)(-3+\nu)+\frac{1}{2}\nu^2(1-\nu) , \\
a_2^\nu &= \nu^2-\frac{1}{4}\nu^2(-3+\nu)^2-\nu(1-\nu)(1-\frac{1}{4}\nu(3+\nu)) , \\
a_3^\nu &= \frac{1}{2}\nu^2(1-\nu)^2
\end{aligned}$$

taken as coefficients corresponding to $\cos^k(\xi\Delta x)$ from lemma 4.18. Then

$$|g(\xi,\Delta x,\Delta t)|^2=\sum_{k=0}^{3}a_k^\nu\cos^k(\xi\Delta x)=\varphi_\nu(\cos(\xi\Delta x))$$

and we will investigate the polynomial $\varphi_\nu(c)=\sum_{k=0}^3 a_k^\nu c^k$ as a function of the variable c to find the maximum of $|g|$. Since $|g|^2\geq 0$, we also have $\varphi_\nu\geq 0$ on $[-1,1]$.

Lemma 4.19
Let $0\leq\nu\leq 1$. Then

$$\max_{c\in[-1,1]}\varphi_\nu(c)\leq 1 .$$

Proof.
In the cases $\nu=0$ and $\nu=1$ we have $\varphi_\nu(c)\equiv 1$ for all c.
Consider $\nu\in(0,1)$. The function φ_ν can have extrema at the boundary points -1, 1 and all points, where $\varphi_\nu'=0$.

$$\begin{aligned}
\varphi_\nu(-1) &= (1-2\nu)^2\in[0,1) , \\
\varphi_\nu(1) &= 1 .
\end{aligned}$$

Further,

$$\varphi_\nu'(c)=3a_3^\nu c^2+2a_2^\nu c+a_1^\nu=0 \quad\Leftrightarrow\quad c=c_\pm=\frac{-a_2^\nu\pm\sqrt{(a_2^\nu)^2-3a_1^\nu a_3^\nu}}{3a_3^\nu} .$$

For the argument of the square root we have

$$\begin{aligned}
(a_2^\nu)^2-3a_1^\nu a_3^\nu &= \nu^2(\underbrace{-1+2\nu-2\nu^2+\nu^3}_{<0})^2 \\
\Rightarrow\sqrt{(a_2^\nu)^2-3a_1^\nu a_3^\nu} &= \nu(1-2\nu+2\nu^2-\nu^3) .
\end{aligned}$$

This leads to

$$c_\pm=\frac{-a_2^\nu\pm\nu(1-2\nu+2\nu^2-\nu^3)}{3a_3^\nu}$$

and the points

$$\begin{aligned}
c_+ &= \left(-\frac{1}{3}\right)\frac{-4+5\nu-2\nu^2+\nu^3}{\nu(1-\nu)^2} , \\
c_- &= 1
\end{aligned}$$

are solutions of the equation $\varphi'_\nu(c) = 0$.
Now we want to show that $c_+ \notin [-1,1]$. To this end, define the right hand side of the formula for c_+ as a function of $\nu \in (0,1)$

$$\rho(\nu) \quad = \quad \left(-\frac{1}{3}\right) \frac{-4+5\nu-2\nu^2+\nu^3}{\nu(1-\nu)^2} \ .$$

It holds

$$\lim_{\nu\to 0_+} \rho(\nu) = +\infty \quad , \quad \lim_{\nu\to 1_-} \rho(\nu) = +\infty$$

and

$$\rho'(\nu) = 0 \quad \Leftrightarrow \quad \left(-\frac{1}{3}\right) \frac{-4(-1+2\nu)}{\nu^2(1-\nu^2)} = 0 \quad \Leftrightarrow \quad \nu = \frac{1}{2} \ ,$$

i.e.,

$$\rho(\nu) \geq \rho\left(\frac{1}{2}\right) = 5 \ .$$

Hence, we have for all $\nu \in (0,1)$

$$c_+ = \rho(\nu) \geq 5 \text{ , i.e., } c_+ \notin [-1,1] \ .$$

Consider the function $\varphi_\nu(c)$. Since $c_\pm \geq 1$, we have $\varphi'_\nu(c) > 0$ for $c \in (-1,1)$. Together with $\varphi_\nu(-1) \in [0,1)$ and $\varphi_\nu(1) = 1$ it follows

$$\max_{c\in[-1,1]} \varphi_\nu(c) \leq 1 \ .$$

□

Theorem 4.20
Consider the scheme (4.37) and the corresponding function $g(\xi,\Delta x,\Delta t)$ *defined in (4.47). Let* $0 \leq \nu \leq 1$. *Then*

$$|g(\xi,\Delta x,\Delta t)| \leq 1$$

and

$$\|u^{n+1}\|_{2,\Delta x} \leq \|u^0\|_{2,\Delta x} \ .$$

In other words, the scheme (4.37) is stable in $\|\cdot\|_{2,\Delta x}$*-norm.*

Proof.
The result follows directly from lemma 4.19 and from the discussion above. □

The theorem 4.20 states that the scheme (4.37) is stable in $\|\cdot\|_{2,\Delta x}$-norm if $0 \leq \nu \leq 1$. This is an optimal condition since it corresponds to the CFL-condition in the case of a linear hyperbolic PDE. In the practical use, due to numerical reasons (e.g., round-off errors that we have not considered), the CFL-condition is usually, also for linear hyperbolic PDEs, relaxed to $0 \leq \nu \leq CFL$ for $CFL \in (0,1)$ a suitable number, e.g., $CFL = 0.95$.

4.4 Convergence analysis

In the previous text of this chapter we have studied the scheme (4.13)-(4.20) in the terms of consistency for the scalar conservation law (4.23) and stability for the linear advection equation (4.36), where the latter is a special case of (4.23). In this section, we will make use of these results and combine them to get convergence results of our scheme for the equation (4.36). We will prove the expected convergence rate 2 of the scheme at a fixed time $T > 0$.

Definition 4.21
Let $u_h(x,t^n) = \sum_i u_i^n \psi_i(x)$ denote the numerical solution of equation (4.36) provided by the FVPM (2.23) with non-moving particles at time $t = t^n = n\Delta t$. Let $u(x,t)$ denote the exact solution of (4.36). Assume $\frac{\Delta t}{\Delta x} = K < +\infty$ remains constant. We say, that the method is convergent *at time $T = N_T \Delta t$ in the norm $\|.\|$, if*

$$\lim_{\Delta t \to 0,\ N_T\Delta t = T} \Big\| u_h(x,T) - u(x,T) \Big\| = \lim_{\Delta t \to 0,\ N_T\Delta t = T} \Big\| \sum_i u_i^{N_T} \psi_i(x) - u(x,T) \Big\| = 0 \ .$$

Moreover, if

$$\Big\| u_h(x,T) - u(x,T) \Big\| = \mathcal{O}(\Delta t^s) + \mathcal{O}(\Delta x^r) \ , \ \Delta t \to 0 \ , \ \Delta x \to 0 \ ,$$

the method is said to be convergent of order s in time and r in space.

Due to the triangle inequality we obtain for the error of the approximation at time t^n

$$\begin{aligned} \Big\| u_h(x,t^n) - u(x,t^n) \Big\| &= \Big\| \sum_i u_i \psi_i(x) - u(x,t^n) \Big\| \\ &\le \Big\| \sum_i u_i \psi_i(x) - \sum_i u(x_i,t^n)\psi_i(x) \Big\| + \Big\| \sum_i u(x_i,t^n)\psi_i(x) - u(x,t^n) \Big\| \ . \end{aligned} \tag{4.48}$$

The second term can be estimated for $u \in \mathcal{C}^2(\overline{\Omega} \times [0,T])$ by

$$\Big\| \sum_i u(x_i,t^n)\psi_i(x) - u(x,t^n) \Big\| \le C\Delta x^2$$

in the L^1-, L^2- or L^∞-norm on a bounded domain, since it is an approximation of u by a piecewise linear function. For the estimate see e.g., [50].

A general approach to investigate convergence of a numerical scheme for the solution of hyperbolic PDEs was presented in LeVeque [39]. We will follow this approach to show

$$\Big\| \sum_i u_i \psi_i(x) - \sum_i u(x_i,t^n)\psi_i(x) \Big\| = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ ,$$

which together with the previous considerations will lead to a proof of convergence of the scheme (4.13)-(4.20) of order 2. We begin with the referenced approach.

Convergence theory

A numerical scheme for the solution of hyperbolic conservation laws can be investigated on convergence in the following way. One is interested in the global error of the scheme, which is, however,

difficult to determine directly. A way how to circumvent this difficulty, is to divide the investigation into two parts. One computes the local error of every single time step, i.e., the *consistency* of the scheme is studied. Then, provided that the scheme is *stable*, i.e., the growth of local errors can be bounded, a bound on the global error can be found in terms of local errors and the convergence order can be proven.
More specifically, a general explicit numerical method can be written as

$$u^{n+1} = \mathcal{N}(u^n) \ , \tag{4.49}$$

where $\mathcal{N}$ represents the numerical operator mapping the approximate solution at one time step to the approximate solution at the next step. Values $u^n = \{u_i^n\}_i$ represent some discrete values of the scheme, e.g., point values, integral means or weighted integral means, depending on which method is investigated. Let $u_{ex}^n = \{u(x_i, t^n)\}_i$ be the values of exact solution at (x_i, t^n). The *local truncation error* $\tau^n = \{T(x_i, t^n)\}_i$ is defined by the difference of left and right hand side of the equation (4.49) divided by Δt, where we use exact values u_{ex}^n and u_{ex}^{n+1} instead of values u^n and u^{n+1}, respectively,

$$\tau^n := \frac{1}{\Delta t}\left(u_{ex}^{n+1} - \mathcal{N}(u_{ex}^n)\right) \ ,$$

compare also with (4.25).

Remark 4.22
In section 4.2, we have shown that $\tau^n = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)$ *as* $\Delta t \to 0$, $\Delta x \to 0$ *for the scheme (4.13)-(4.20) and linear advection equation (4.36).*

Further, we define the *global error of coefficients* at time t^n by

$$E^n := u^n - u_{ex}^n \ .$$

The numerical method (4.49) applied on u^n yields

$$u^{n+1} = \mathcal{N}(u^n) = \mathcal{N}(u_{ex}^n + E^n) \ ,$$

which gives

$$\begin{aligned} E^{n+1} &= u^{n+1} - u_{ex}^{n+1} \\ &= \mathcal{N}(u_{ex}^n + E^n) - u_{ex}^{n+1} \\ &= \mathcal{N}(u_{ex}^n + E^n) - \mathcal{N}(u_{ex}^n) + \mathcal{N}(u_{ex}^n) - u_{ex}^{n+1} \\ &= [\mathcal{N}(u_{ex}^n + E^n) - \mathcal{N}(u_{ex}^n)] - \Delta t \tau^n \ . \end{aligned}$$

Stability theory allows to bound the first term $[\mathcal{N}(u_{ex}^n + E^n) - \mathcal{N}(u_{ex}^n)]$ and the consistency analysis gives a bound on the one-step error $\Delta t \tau^n$ yielding then a convergence result.
Assuming $\mathcal{N}$ is *contractive* in the norm $\|\cdot\|$, i.e.,

$$\|\mathcal{N}(P) - \mathcal{N}(Q)\| \le \|P - Q\| \qquad \text{for all suitable } P, Q \ , \tag{4.50}$$

we have

$$\begin{aligned} \|E^{n+1}\| &\le \|\mathcal{N}(u_{ex}^n + E^n) - \mathcal{N}(u_{ex}^n)\| + \Delta t\|\tau^n\| \\ &\le \|E^n\| + \Delta t\|\tau^n\| \ . \end{aligned}$$

In the case of a linear operator $\mathcal{N}$, the contractive condition (4.50) reduces to

$$\|\mathcal{N}\| \le 1 \ , \tag{4.51}$$

since for a linear operator we have $\|\mathcal{N}(P) - \mathcal{N}(Q)\| \le \|\mathcal{N}\|\|P - Q\| \le \|P - Q\|$.
A recursive application gives

$$\|E^{N_T}\| \le \|E^0\| + \Delta t \sum_{n=0}^{N_T - 1} \|\tau^n\| \ .$$

Assume a uniform bound on $\|\tau^n\|$, i.e.,

$$\|\tau\| := \max_{0 \le n \le N_T - 1} \|\tau^n\| \ .$$

Then

$$\|E^{N_T}\| \quad \le \quad \|E^0\| + \Delta t N_T \|\tau\| = \|E^0\| + T\|\tau\| \ . \tag{4.52}$$

Remark 4.23
Weaker requirements on $\mathcal{N}$ can be considered:

$$\|\mathcal{N}(P) - \mathcal{N}(Q)\| \le (1 + \alpha \Delta t)\|P - Q\| \qquad \text{for all suitable } P, Q$$

with a constant α independent of Δt.

Consider the inequality (4.52). If the error in the initial data satisfies $\|E^0\| \to 0$ and $\|\tau\| \to 0$, then the numerical method (4.49) is convergent. Moreover, if $\|E^0\| = \mathcal{O}(\Delta t^s) + \mathcal{O}(\Delta x^r)$ and $\|\tau\| = \mathcal{O}(\Delta t^s) + \mathcal{O}(\Delta x^r)$, then the numerical method is convergent of order s in time and r in space.

Convergence of the scheme (4.37)

Consider the linear advection equation (4.36) and the scheme (4.13)-(4.20) for $n_s = k = 2$, i.e., the scheme (4.37). Consider now the L^2-norm for the convergence. Nevertheless, we will be able to show convergence also in L^1-norm based on the convergence in L^2-norm.
Get back to the inequality (4.48). We have shown that the second term on right hand side can be bounded by $\mathcal{O}(\Delta x^2)$. For the first term one can show an estimate on L^2-norm via direct computation

$$\begin{aligned} \left\| \sum_i \left(u_i^n - u(x_i, t^n)\right) \psi_i(x) \right\|_2^2 \quad &\le \quad \frac{4}{3} \Delta x \sum_i |u_i^n - u(x_i, t^n)|^2 \\ &= \quad \frac{4}{3} \left\| u^n - u_{ex}^n \right\|_{2,\Delta x}^2 = \frac{4}{3} \left\| E^n \right\|_{2,\Delta x}^2 \ . \end{aligned} \tag{4.53}$$

For $\mathcal{N}_1(u_i^n) = u_i^n - \frac{\Delta t}{\Delta x}(g_{i+\frac{1}{2}} - g_{i-\frac{1}{2}})$ analysed in section 4.2 we have a bound on local truncation error

$$\|\tau\|_{2,\Delta x} = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ , \ \Delta t \to 0 \ , \ \Delta x \to 0 \ .$$

This bound follows from theorem 4.5 and from the assumption that the initial data has compact support (and so the solution of hyperbolic PDE for all times).
Due to theorem 4.20 we have the stability result

$$\|\mathcal{N}_2\|_{2,\Delta x} \le 1$$

for $\mathcal{N}_2(u_i^n) = g(\xi, \Delta x, \Delta t) u_i^n$ analysed in section 4.3.
The numerical methods $\mathcal{N}_1$ and $\mathcal{N}_2$ coincide in the case of linear advection equation (4.36) and parameters $n_s = k = 2$.

Lemma 4.24
Assume $u_0 \in \mathbb{C}^2(\overline{\Omega})$. Then it holds for the scheme (4.13)-(4.20)

$$\|E^0\|_{2,\Delta x} = \mathcal{O}(\Delta x^2) \ , \ \Delta x \to 0 \ .$$

Proof.

$$\|E^0\|_{2,\Delta x}^2 = \Delta x \sum_{i=1}^{n_p} |u_i^0 - u_0(x_i)|^2 \le \Delta x \; n_p C^2 \Delta x^4$$

for a constant $C > 0$ independent of Δx. This follows from

$$u_i^0 = u_0(x_i) + \mathcal{O}(\Delta x^2) \ ,$$

due to theorem 2.20 (where $b_i = x_i$). Then

$$\|E^0\|_{2,\Delta x}^2 \le |\Omega| C^2 \Delta x^4 \ .$$

□

Altogether, we plug the bounds on $\|\tau\|_{2,\Delta x}$ and $\|E^0\|_{2,\Delta x}$ into (4.52) and obtain the following result.

Lemma 4.25
Let $u \in \mathcal{C}^4(\overline{\Omega} \times [0,T])$ be the exact solution of the linear advection equation (4.36). Consider the scheme (4.13)-(4.20) with parameters $n_s = k = 2$ (i.e., the scheme (4.37)) for numerical solution of (4.36). Let $0 \le \nu \le 1$. Then the global error of coefficients is of order 2 in time and space in the norm $\| \cdot \|_{2,\Delta x}$, i.e.,

$$\|E^{N_T}\|_{2,\Delta x} \quad = \quad \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ , \ \Delta t \to 0 \ , \ \Delta x \to 0$$

for a fixed time $T = N_T \Delta t$.

Now we can state the main theorem:

Theorem 4.26
Let $u \in \mathcal{C}^4(\overline{\Omega} \times [0,T])$ be the exact solution of the linear advection equation (4.36). Consider the scheme (4.13)-(4.20) with parameters $n_s = k = 2$ (i.e., the scheme (4.37)) for numerical solution of (4.36). Let $0 \le \nu \le 1$. Then the scheme (4.37) is convergent of order 2 in time and space in L^1- and L^2-norm, i.e.,

$$\begin{aligned}
\Big\| u_h(x,T) - u(x,T) \Big\|_1 \quad &= \quad \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ , \ \Delta t \to 0 \ , \ \Delta x \to 0 \ , \\
\Big\| u_h(x,T) - u(x,T) \Big\|_2 \quad &= \quad \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ , \ \Delta t \to 0 \ , \ \Delta x \to 0
\end{aligned}$$

for a fixed time $T = N_T \Delta t$.

Proof.
The result for the L^2-norm follows from the previous considerations - consider the inequality (4.48) and the estimate on the second term

$$\begin{aligned}
\Big\| u_h(x,T) - u(x,T) \Big\|_2 &= \Big\| \sum_i u_i \psi_i(x) - u(x,T) \Big\|_2 \\
&\le \Big\| \sum_i u_i \psi_i(x) - \sum_i u(x_i,T) \psi_i(x) \Big\|_2 + \mathcal{O}(\Delta x^2).
\end{aligned}$$

The first term can be estimated as

$$\left\| \sum_i u_i \psi_i(x) - \sum_i u(x_i,T)\psi_i(x) \right\|_2 \quad \leq \quad \sqrt{\frac{4}{3}} \left\| E^{N_T} \right\|_{2,\Delta x} = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ ,$$

where we used the relation (4.53) and lemma 4.25. Altogether, one gets the desired estimate.

For the L^1-norm we have due to the Hölder's inequality

$$\|f\|_{L^1(\Omega)} \leq |\Omega|^{\frac{1}{2}} \|f\|_{L^2(\Omega)}$$

for any function $f \in L^2(\Omega)$.
We apply the above inequality with $f := u_h(\cdot,T) - u(\cdot,T)$ and obtain

$$\|u_h(x,T) - u(x,T)\|_{L^1(\Omega)} \leq |\Omega|^{\frac{1}{2}} \|u_h(x,T) - u(x,T)\|_{L^2(\Omega)} \leq |\Omega|^{\frac{1}{2}} C(\Delta t^2 + \Delta x^2) \ ,$$

i.e., it is also

$$\|u_h(x,T) - u(x,T)\|_1 = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \ .$$

$\square$

For the case of data smooth enough, we have hereby proven the convergence of order 2 of the numerical solution obtained via the scheme (4.37) to the exact solution of a linear advection equation in the L^1- and L^2-norm.

5 Numerics

After the purely theoretical part, we will present numerical results that confirm results from previous chapters. First, we apply the algorithms, proposed in chapter 2, to add and remove a particle on some examples to demonstrate its functionality concerning conservative approximation. Also relevant examples with poor or too dense particle distribution are presented.

The second section of this chapter is more extensive. We apply the high order meshfree scheme developed in chapter 4 on several relevant examples. We confirm numerically the theoretical result of convergence of second order in the case of scalar linear equation in the L^1- and L^2-norm. Moreover, we show the convergence to the exact solution also in the L^∞-norm. Furthermore, we present examples on a wider class of hyperbolic conservation laws, such as non-linear scalar equations and also linear and non-linear systems, for which the numerical scheme converges with order two to the exact solution. We present also examples with a discontinuous solution. For such solutions, it is not possible to measure the order of convergence. However, we observe better resolution of the discontinuities in comparison to a first order method. One can conclude that the method is robust and of a good approximation quality for scalar hyperbolic conservation laws as well as for linear systems. For non-linear systems with smooth exact solutions, the method works still very well. In the case of discontinuities in the solution one obtains non-physical oscillations in their vicinity. Further techniques have possibly to be utilized to suppress them, such as the principle of *limiters* (an overview on limiters and particular definitions can be found in Toro [64]) or a modification of the ADER method, based on the analysis done by Goetz [18]. In the latter, it is shown that the Toro-Titarev solver does not act properly for non-linear systems with discontinuous data. A possible remedy could be the *LeFloch-Raviart expansion* presented therein. The inclusion of one or both techniques lies however beyond the scope of this thesis.

We emphasize, that the presented convergence is not only the convergence of *weighted integral means* to their exact values, but convergence of the numerical solution function to the exact solution function in a given function space.

5.1 Adding and removing a particle

Adding a particle

In this section, numerical results are shown for the theory developed in chapter 2, namely to add and remove a particle. We provide computations for the methods SUPP, JI and JPLUS in order to compare them (for the description of the methods, see the end of the section 2.3). We will use the numerically optimized schemes (2.43)-(2.45) and (2.53)-(2.54) exclusively. For the sake of simplicity, we will present one-dimensional results, but we remark that the presented method can be used also for vector-valued functions in arbitrary dimensions. The methods SUPP, JI and JPLUS yield comparable results in the presented examples.

As proven in theorems 2.39 and 2.53, the schemes preserve constant states and are conservative up to the machine precision, which also numerical computations show for all three methods, namely SUPP, JI and JPLUS. The numerical results for constant functions are not presented here. We only remark that they confirm the theoretical results.

We will rather investigate the more interesting class of non-constant functions and compare the

Method	$\int u_h^+ - \int u_h$	$\|u_h^+ - u_h\|_{L^1}$	$\|u_h^+ - u_h\|_{L^2}$	$\|u_h^+ - u_h\|_{L^\infty}$
SUPP	0.0000E+00	1.8912E-02	2.3039E-02	4.8887E-02
JI	0.0000E+00	2.0671E-02	2.3411E-02	4.5856E-02
JPLUS	0.0000E+00	2.7201E-02	2.8884E-02	6.1403E-02

Table 5.1: *Example 1. Conservativity and errors.*

Method	$\int u_h^+ - \int u_h$	$\|u_h^+ - u_h\|_{L^1}$	$\|u_h^+ - u_h\|_{L^2}$	$\|u_h^+ - u_h\|_{L^\infty}$
SUPP	0.0000E+00	1.5390E-02	1.8478E-02	3.3207E-02
JI	0.0000E+00	1.6158E-02	1.7997E-02	3.0921E-02
JPLUS	-1.7764E-15	2.2505E-02	2.2643E-02	4.3492E-02

Table 5.2: *Example 2. Conservativity and errors.*

methods SUPP, JI and JPLUS to each other. Also different partitions of unity will be chosen.

In all figures, the structure of the visualization is the same. The new particle is added at the position 0. Functions ψ_i of the original partition of unity are represented with a dashed blue line having values between 0 and 1, “new” functions ψ_i^+ with a solid red line, original coefficients u_i with blue crosses and coefficients u_i^+ with red circles. In the upper part of the plot, a reconstruction of a given function $\sum_i u_i\psi_i$ is shown with a dashed blue line and $\sum_i u_i^+\psi_i^+$ with a solid red line respectively. In the figure one can immediately see how the partition of unity changes by adding a new particle.

In the table, the conservativity and norms of the difference of the reconstructions $u_h = \sum_i u_i\psi_i$ and $u_h^+ = \sum_i u_i^+\psi_i^+$ are shown for different methods. We achieve conservativity up to the machine precision.

Different partitions of unity are used. They are based on the function

$$W_i(x) = \begin{cases} 1 + \frac{x-x_i}{H} & , \quad x \in [x_i - H, x_i] \ , \\ 1 - \frac{x-x_i}{H} & , \quad x \in [x_i, x_i + H] \ , \\ 0 & , \quad \text{otherwise} \ , \end{cases}$$

where W_i denotes the function (2.4) and $H \in \mathbb{R}_+$ is a parameter given in each example. The basis functions are then defined as $\psi_i = \frac{W_i}{\sum_j W_j}$.

The symbol Δx will denote the distance $\Delta x = x_{i+1} - x_i$ in the case of uniformly distributed particles.

Example 1 (2 neighbors)

The given function is $u(x) = x^2 + x + 2$, $x \in [-1, 1]$.

The particles are distributed uniformly, $x_i = -1 + 2\frac{i-1}{9}$, $i = 1, \ldots, 10$. We choose $H = \Delta x$, so that the functions ψ_i are equal to B-spline functions. See figure 5.1 and table 5.1 for results. We can see that the method is conservative. The resulting approximation changes due to the change of underlying structure of particles.

Example 2 (6 neighbors)

The given function is $u(x) = x^2 + x + 2$, $x \in [-1, 1]$.

The particles are distributed uniformly, $x_i = -1 + 2\frac{i-1}{19}$, $i = 1, \ldots, 20$. We choose $H = 2\Delta x$, so we have another particle basis functions. For results see figure 5.2 and table 5.2. The method is again conservative up to the machine precision.

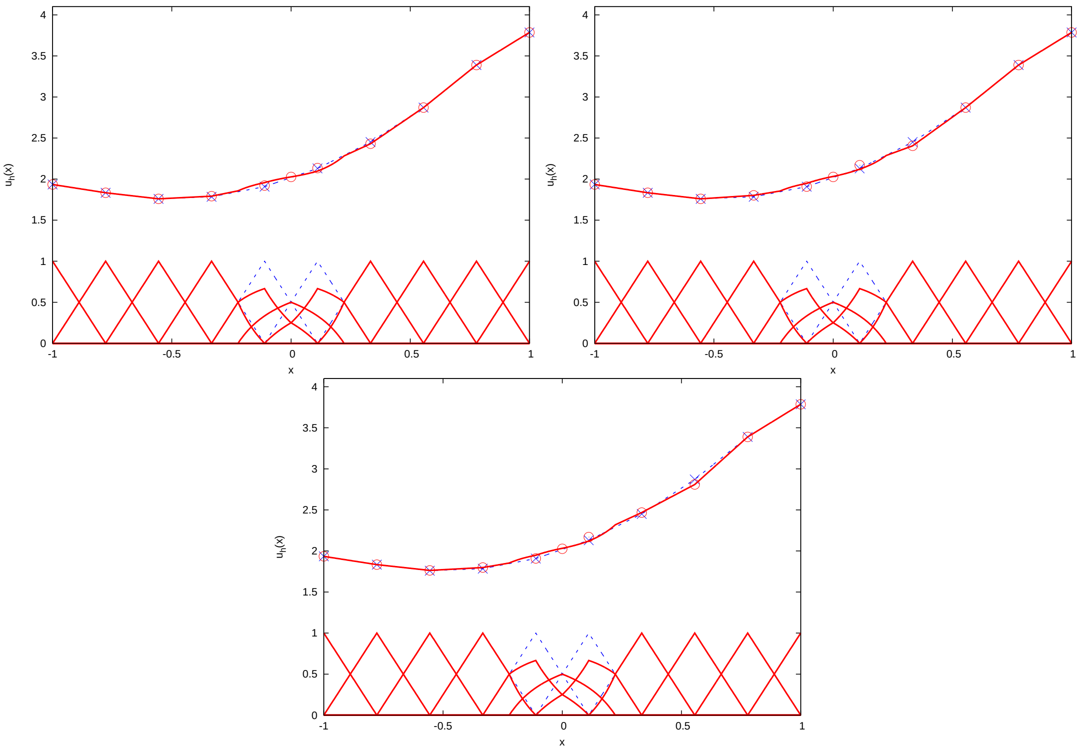


Figure 5.1: *Example 1, adding a particle. From left to right and down the methods SUPP, JI and JPLUS.*

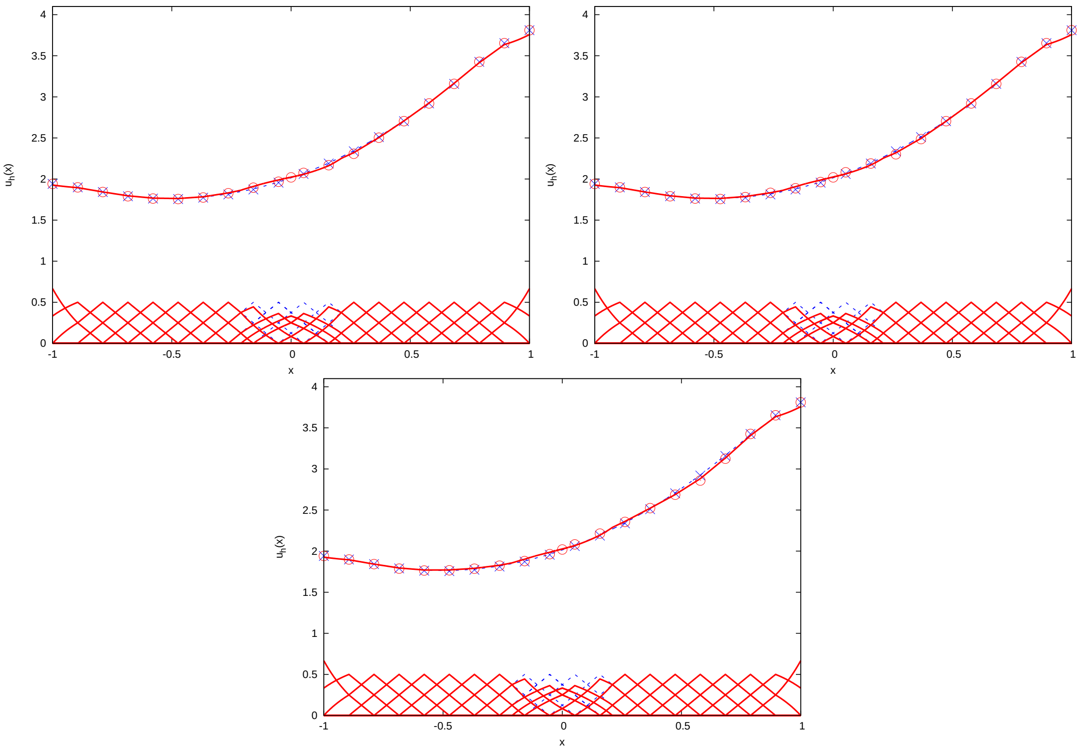


Figure 5.2: *Example 2, adding a particle. From left to right and down the methods SUPP, JI and JPLUS.*

Example 3 (the "gap")

The given function is $u(x) = x^2 + x + 2$, $x \in [-1, 1]$.
We deal with the situation, in which we have a "gap" in the particle distribution. The particles

Method	$\int u_h^+ - \int u_h$	$\|u_h^+ - u_h\|_{L^1}$	$\|u_h^+ - u_h\|_{L^2}$	$\|u_h^+ - u_h\|_{L^\infty}$
SUPP	-8.8818E-16	2.8652E-02	4.8770E-02	1.2978E-01
JI	-8.8818E-16	3.5942E-02	4.6899E-02	1.2771E-01
JPLUS	-8.8818E-16	4.2708E-02	4.9874E-02	1.2771E-01

Table 5.3: *Example 3. Conservativity and errors.*

are distributed as follows: $x_i = -1 + (i-1)\Delta\widetilde{x}$, $i = 1, \ldots, 5$ and $x_i = -1 + (i-1)\Delta\widetilde{x} + \frac{8}{45}$, $i = 6, \ldots, 10$, where $\Delta\widetilde{x} = \frac{82}{405}$. In the vicinity of the point $x = 0$, we have the interval in which ψ_5 and ψ_6 overlap. Its length is $\frac{10}{405} = \frac{10}{82}\Delta\widetilde{x} \doteq 0.12\Delta\widetilde{x}$. In order to improve the poor particle distribution here, a new particle is added. We have $H = \Delta\widetilde{x}$. See figure 5.3 and table 5.3. The example confirms that the algorithm suites for the case of poor particle distribution.

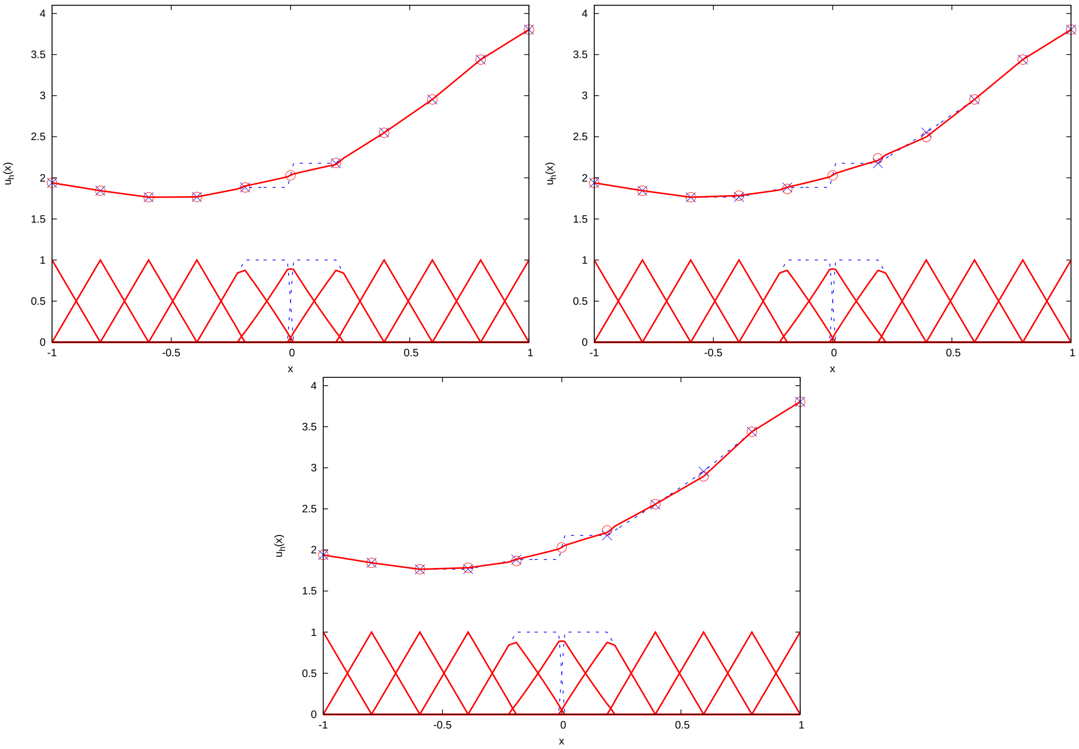


Figure 5.3: *Example 3, adding a particle. From left to right and down the methods SUPP, JI and JPLUS.*

Removing a particle

For completeness, we present also an example on removing a particle.

Example

The given function is $u(x) = x^2 + x + 2$, $x \in [-1, 1]$.
In this example, the particle distribution is too dense. The particles are distributed as follows: $x_i = -1 + (i-1)\Delta\widetilde{x}$, $i = 1, \ldots, 5$, $x_6 = 0$ and $x_i = -1 + (i-2)\Delta\widetilde{x}$, $i = 7, \ldots, 11$, where $\Delta\widetilde{x} = \frac{2}{9}$. One can see, that in the vicinity of the point $x = 0$ the particle distribution density is higher. The particle $x_6 = 0$ will be removed. We have $H = \Delta\widetilde{x}$. See figure 5.4 and table 5.4. The algorithm works as expected and we get a conservative approximation.

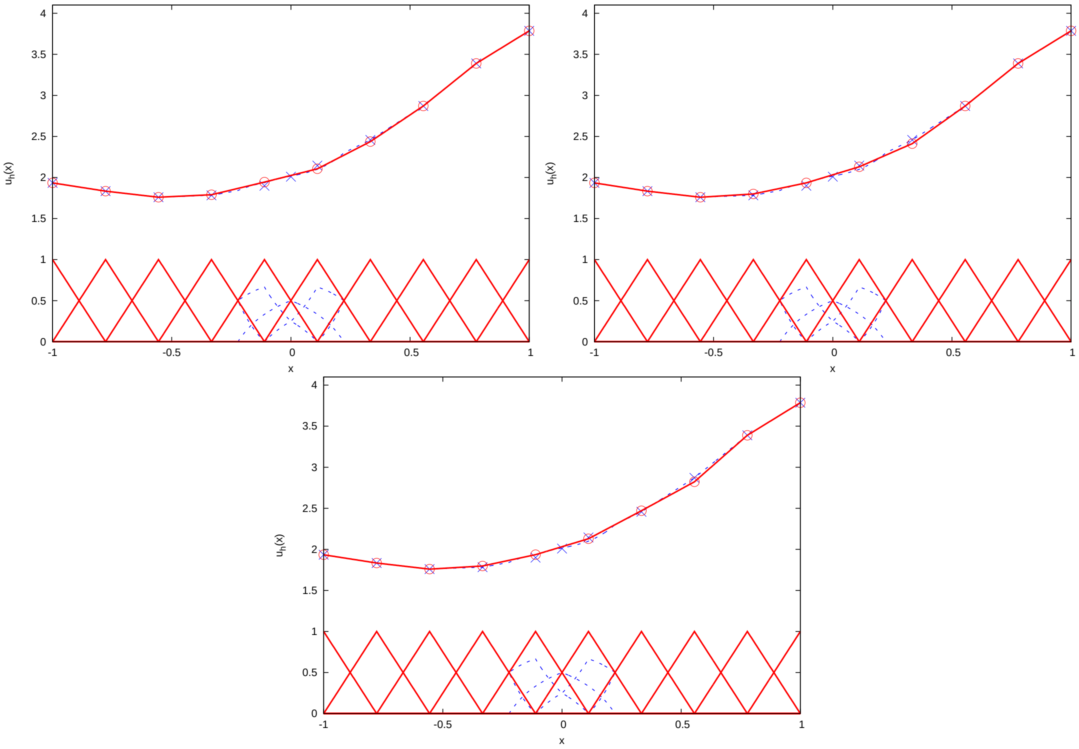


Figure 5.4: *Example, removing a particle. From left to right and down the methods SUPP, JI and JPLUS.*

Method	$\int u_h^+ - \int u_h$	$\|u_h^+ - u_h\|_{L^1}$	$\|u_h^+ - u_h\|_{L^2}$	$\|u_h^+ - u_h\|_{L^\infty}$
SUPP	8.8818E-16	1.1776E-02	1.3698E-02	2.9739E-02
JI	0.00000E+00	2.0068E-02	2.3260E-02	4.5083E-02
JPLUS	0.00000E+00	2.1865E-02	2.3989E-02	4.8213E-02

Table 5.4: *Example on removing a particle. Conservativity and errors.*

5.2 Higher order scheme

In this section we are going to verify numerically the second order of convergence of the scheme (4.13)-(4.20) in the L^1-, L^2- and L^∞-norm.

In cases where discontinuities occur in the exact solution, the second order of convergence cannot be determined by reason of the lack of smoothness. Nevertheless, one observes a better resolution of the numerical solution in the vicinity of discontinuities, compared to a method of first order of accuracy.

In the following examples, we compute the numerical solution with two methods - the scheme (4.13)-(4.20) with parameters $n_s = 3$, $k = 2$ (denoted by $HO32$) and a first order scheme (denoted by $o1$). The only exception is the example 5.2.3 of *linear advection equation with peaks* where we compare some more methods. We use the same CFL-number for all computations. The CFL-condition is in our case defined as

$$\Delta t = CFL \frac{\Delta x}{S_{max}} ,$$

where $CFL \in (0,1)$ and S_{max} is an estimate on the maximum wave speed of a given problem, determined on the basis of the knowledge of the exact solution.

The first order method used here is a standard FVPM as defined in chapter 2, with the same structure as our second order scheme, i.e., with linear B-splines defining the partition of unity. As the numerical flux $\mathbf{g}_{ij}^n$ from (2.22) we apply the Godunov approach, i.e., we solve local Riemann problems defined by the coefficients $\mathbf{u}_i^n$ of (2.23). It is also possible to use an arbitrary numerical flux of first order, e.g., Lax-Friedrichs, Steger-Warming or Vijayasundaram numerical flux. The only difference between the first and second order method used here for comparison, is the applied numerical flux - in fact we compare the ADER-FVPM of first and of second order of accuracy.

The Riemann problems, that arise in the method (4.13)-(4.20), are solved exactly in each example. For the shallow water equations and Euler equations we use exact Riemann solvers introduced by Toro in [63] and [64], respectively.

In each example, we apply the numerical scheme for a sequence of particle distributions for a parameter $N = N_0,\ 2N_0,\ 4N_0, \ldots$, where the number of particles is $N + 1$. Then we compare the numerical solution with the exact solution and compute an error in the appropriate norm denoted by $E_p(N) = \|\mathbf{u}_h - \mathbf{u}_{exact}\|_p$ for $p \in \{1, 2, \infty\}$. In the smooth cases, the knowledge of these errors allows us to compute the numerical order of convergence via the formula

$$k_p = \frac{\log(E_p(N)/E_p(2N))}{\log(2)} .$$

The errors $E_p(N)$ are computed at 100000, in some cases at 200000 points. We remark that the determined values of k_p attain the expected value 2, i.e., the second order of accuracy is shown for the presented examples. In the non-smooth cases, k_p is not computed and only the errors $E_p(N)$ are shown for completeness. The exact solution is always depicted on a grid of 100000 or 200000 points connected with a line.

5.2.1 Linear advection equation

Consider a simple linear advection equation

$$\begin{aligned} u_t + u_x &= 0 \quad , \quad x \in [-0.5, 0.5] \quad , \quad t \in [0,1] , \\ u(x,0) &= \sin(2\pi x) \end{aligned}$$

with periodic boundary conditions. The exact solution is determined by the method of characteristics.

It holds for the exact solution, that the sinus wave is shifted to the right and due to the periodic boundary conditions it arises from the left. At the final time the exact solution is equal to the initial condition. We solved this equation with $CFL = 0.95$. The numerical order of convergence was proven to be 2 in all three considered norms, see table 5.5 and compare to table 5.6 (first order method $o1$). This confirms the theoretical results proven in chapter 4. See also figure 5.5.

N	$E_1(N)$	k_1	$E_2(N)$	k_2	$E_\infty(N)$	k_∞
38	1.3185E-02	—	1.5993E-02	—	3.9305E-02	—
76	3.5368E-03	1.90	4.2355E-03	1.92	9.9899E-03	1.98
152	9.0043E-04	1.97	1.0438E-03	2.02	2.3318E-03	2.10
304	2.2538E-04	2.00	2.6039E-04	2.00	6.7206E-04	1.79
608	5.6355E-05	2.00	6.4714E-05	2.01	1.7453E-04	1.95

Table 5.5: *Linear advection equation. Errors and convergence order for the high order method* $HO32$ *at time* $t = 1$.

N	$E_1(N)$	k_1	$E_2(N)$	k_2	$E_\infty(N)$	k_∞
38	1.9134E-02	—	2.1263E-02	—	3.1137E-02	—
76	8.9284E-03	1.10	9.9185E-03	1.10	1.4303E-02	1.12
152	4.3003E-03	1.05	4.7766E-03	1.05	6.8253E-03	1.07
304	2.1087E-03	1.03	2.3421E-03	1.03	3.3300E-03	1.04
608	1.0439E-03	1.01	1.1595E-03	1.01	1.6442E-03	1.02

Table 5.6: *Linear advection equation. Errors and convergence order for the first order method* $o1$ *at time* $t = 1$.

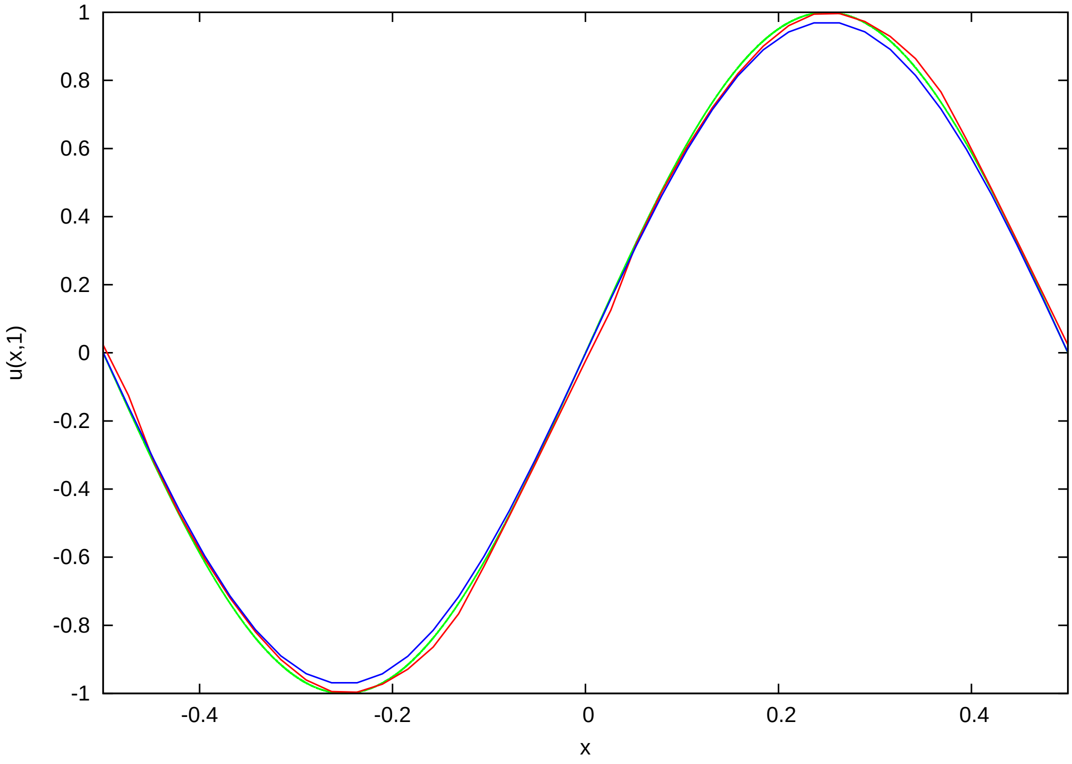


Figure 5.5: *Linear advection equation. Solutions for* $N = 38$ *at time* $t = 1$. *The exact solution (green), the high order solution* $HO32$ *(red) and first order solution* $o1$ *(blue) are depicted.*

5.2.2 Burgers' equation

In this example we will test our method on a non-linear scalar equation, the so-called Burgers' equation

$$\begin{aligned} u_t + \left(\frac{1}{2}u^2\right)_x &= 0 \quad , \quad x \in [0, 2\pi] \quad , \quad t \in \left[0, \frac{19}{70}\pi\right] \doteq [0, 0.85271] \ , \\ u(x,0) &= \frac{1}{2} + \sin(x) \end{aligned}$$

with periodic boundary conditions. The exact solution is determined by the method of characteristics. We solved this equation with $CFL = 0.95$.

The final time $T = \frac{19}{70}\pi \doteq 0.85271 < T_{disc}$ ensures that the exact solution is still smooth, so we are able to determine the numerical convergence order. The initial sinus function graph is deformed in time and at the time $T_{disc} = 1$ a discontinuity occurs in the exact solution.

We verify numerically that the method $HO32$ is of second order of convergence even for non-linear equations (see table 5.7 and compare to table 5.8). Hence, our method is applicable also on non-linear cases of hyperbolic PDEs. For solutions at final time T see figure 5.6.

N	$E_1(N)$	k_1	$E_2(N)$	k_2	$E_\infty(N)$	k_∞
70	1.4503E-02	—	1.6042E-02	—	4.3619E-02	—
140	3.7217E-03	1.96	4.4309E-03	1.86	1.6881E-02	1.37
280	9.4923E-04	1.97	1.0905E-03	2.02	4.1130E-03	2.04
560	2.3292E-04	2.03	2.5577E-04	2.09	1.0214E-03	2.01
1120	5.8143E-05	2.00	6.2516E-05	2.03	2.2689E-04	2.17

Table 5.7: *Burgers' equation. Errors and convergence order for the high order method $HO32$ at time $t = T$.*

N	$E_1(N)$	k_1	$E_2(N)$	k_2	$E_\infty(N)$	k_∞
70	1.0605E-01	—	7.9519E-02	—	1.8952E-01	—
140	5.5180E-02	0.94	4.8524E-02	0.71	1.4902E-01	0.35
280	2.9711E-02	0.89	2.8205E-02	0.78	8.6879E-02	0.78
560	1.5350E-02	0.95	1.5080E-02	0.90	4.7696E-02	0.87
1120	7.8671E-03	0.96	7.8477E-03	0.94	2.4997E-02	0.93

Table 5.8: *Burgers' equation. Errors and convergence order for the first order method $o1$ at time $t = T$.*

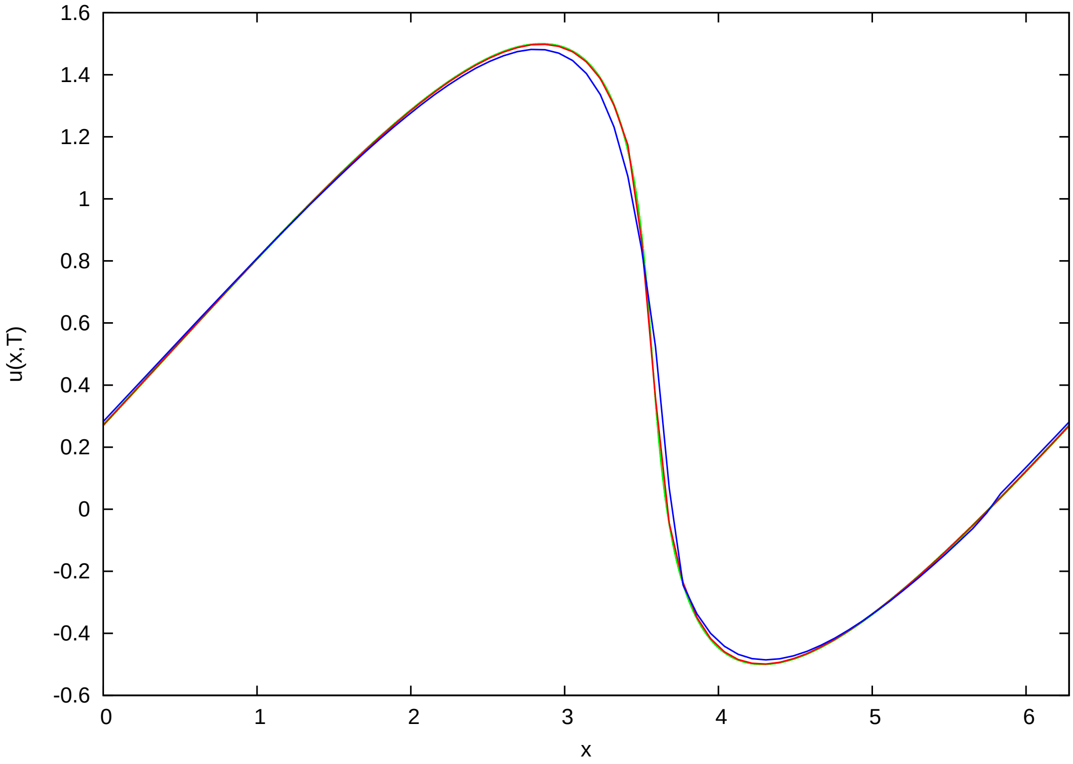


Figure 5.6: *Burgers' equation. Solutions for* $N = 70$ *at time* $t \doteq 0.85271$*. The exact solution (green), the high order solution* $HO32$ *(red) and first order solution* $o1$ *(blue) are depicted.*

5.2.3 Linear advection equation with peaks

Let us consider again the linear advection equation, but with different initial condition (initial condition from [61] and references therein)

$$\begin{aligned} u_t + u_x &= 0 \quad , \quad x \in [-1,1] \quad , \quad t \in [0,2] \ , \\ u(x,0) &= \begin{cases} \exp(-\log(2)(x+0.7)^2/0.0009) & , \quad x \in [-0.8,-0.6] \ , \\ 1 & , \quad x \in [-0.4,-0.2] \ , \\ 1-|10x-1| & , \quad x \in [0.0,0.2] \ , \\ \left[1-100(x-0.5)^2\right]^{1/2} & , \quad x \in [0.4,0.6] \ , \\ 0 & , \quad \text{otherwise} \end{cases} \end{aligned}$$

with periodic boundary conditions. The exact solution is determined by the method of characteristics. Since there are discontinuities in the solution, we solve this equation with $CFL = 0.8$. The final exact solution at time 2 is equal to the initial condition. Since the exact solution is not smooth enough, we do not determine the convergence order. However, we are interested in another phenomenon of our method, namely in the resolution of discontinuities and in the approximation of sharp peaks of the solution.

For this example, we have to analyse the reconstruction step (4.20) more carefully. As discussed in the chapter 4, there are several choices of parameters n_s and k. If we choose $n_s = k = 2$, we get purely polynomial reconstruction of the exact solution with constant WENO weights. We will denote it by $HO22$. For $n_s = 3$ and $k = 2$ we obtain a non-trivial polyharmonic splines reconstruction, denoted by $HO32$. The WENO coefficients are in this case variable (i.e., non-constant). These two methods will be then compared as well as the first order method $o1$.

Let us consider the case $n_s = 3$ and $k = 2$. This setting means, one has 3 stencils of length 3. On each of these stencils of length 3 we have 2 degrees of freedom to determine a polynomial. Then there are two possibilities. One defines another reconstruction including this polynomial as a part of the reconstruction (the *polyharmonic splines approach*, i.e., the method $HO32$). The second possibility is to use the *least squares fitting* to approximate the data given on a stencil by the polynomial. However, the second approach is no more an interpolation on the data.

In more detail, consider a stencil of length 3, the corresponding linear functionals λ_1, λ_2 and λ_3 from (3.2) on this stencil and corresponding data to be interpolated: u_1, u_2 and u_3. We are looking for a solution p to the interpolation problem

$$\lambda_i(p) = u_i \quad , \quad i = 1,2,3 \ .$$

If we do not use a polynomial of degree 2, but of degree 1, we cannot interpolate. Then for a polynomial of the form $p(x) = d_0 + d_1 x$ we solve the least square problem

$$\begin{bmatrix} \lambda_1(p) & 1 \\ \lambda_2(p) & 1 \\ \lambda_3(p) & 1 \end{bmatrix} \cdot \begin{bmatrix} d_0 \\ d_1 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \ .$$

We implemented the method using QR-decomposition in order to achieve a higher numerical stability. If we take constant WENO coefficients, we obtain similar results as for the method $HO22$. For a comparison with the method $HO32$, we have to define some non-constant WENO weights. We define these weights as in (3.18) and (3.19). However, the oscillation indicator has to be chosen in another way. We follow Jiang and Shu and their proposal on oscillation indicators on pages 8 and 9 in [28]. Their proposal is related to a quadratic polynomial on stencils of length 3, but can be applied also on a linear one, as we do. Since some terms vanish for a linear polynomial, we end up with oscillation indicators

$$\begin{aligned} \mathcal{I}(s_1^j) &= \frac{1}{4}(u_{j-2} - 4u_{j-1} + 3u_j)^2 \ , \\ \mathcal{I}(s_2^j) &= \frac{1}{4}(u_{j-1} - u_{j+1})^2 \ , \\ \mathcal{I}(s_3^j) &= \frac{1}{4}(3u_j - 4u_{j+1} + u_{j+2})^2 \ , \end{aligned}$$

for stencils $(j-2, j-1, j)$, $(j-1, j, j+1)$ and $(j, j+1, j+2)$. In the following, we will name this method LS and compare it to $HO32$. The errors are shown in tables 5.9 - 5.12 and diverse comparisons of solutions in figures 5.7 - 5.12. We can see that the higher order methods $HO22$, $HO32$ and LS yield significantly better results than the first order method $o1$. There is a lot of information lost in the first order method $o1$ if a small amount of particles is used. For large N the first order method $o1$ yields a better solution but the resolution of discontinuities and peaks is still very poor.

$HO32$ and $HO22$ have very similar solutions with good resolutions of discontinuities and peaks. This holds except for the area of discontinuities where the truly polynomial reconstruction by $HO22$ with constant WENO weights causes a formation of oscillations. On the contrary, the method $HO32$ remains stable due to non-trivial polyharmonic splines and the WENO reconstruction.

If we now compare $HO32$ with LS, we deduce that for smaller number of particles, $HO32$ yields better results than LS. The reason for this is, that the peaks and discontinuities are strongly smeared out. For large N the methods $HO32$ and LS seem to yield comparable results. Nevertheless, based on this example, we would recommend to use the method $HO32$ rather than LS if discontinuities or peaks are present in the solution.

N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
200	6.8338E-02	1.1139E-01	5.5199E-01
400	3.4464E-02	8.0949E-02	5.6467E-01
800	1.8610E-02	6.2309E-02	5.7764E-01
1600	1.0723E-02	4.8928E-02	5.9055E-01
3200	6.4434E-03	3.8831E-02	6.0290E-01

Table 5.9: *Linear advection equation with peaks. Errors for the method* $HO22$ *at time* $t = 2$.

N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
200	4.9498E-02	1.0235E-01	5.5406E-01
400	2.8030E-02	7.6803E-02	5.6894E-01
800	1.7081E-02	5.9546E-02	5.8966E-01
1600	9.7031E-03	4.6671E-02	6.2372E-01
3200	6.0091E-03	3.9064E-02	6.7396E-01

Table 5.10: *Linear advection equation with peaks. Errors for the method* $HO32$ *at time* $t = 2$.

N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
200	2.8547E-01	2.7568E-01	6.2751E-01
400	1.9399E-01	2.1188E-01	5.0569E-01
800	1.2706E-01	1.6177E-01	5.0311E-01
1600	8.1075E-02	1.2391E-01	5.0218E-01
3200	5.1154E-02	9.6760E-02	5.0150E-01

Table 5.11: *Linear advection equation with peaks. Errors for the first order method* $o1$ *at time* $t = 2$.

N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
200	8.7075E-02	1.3103E-01	5.0237E-01
400	4.4786E-02	9.1427E-02	5.0874E-01
800	2.2433E-02	6.7441E-02	5.0991E-01
1600	1.1539E-02	5.1022E-02	5.1051E-01
3200	6.2025E-03	3.8989E-02	5.1100E-01

Table 5.12: *Linear advection equation with peaks. Errors for the method LS at time* $t = 2$*.*

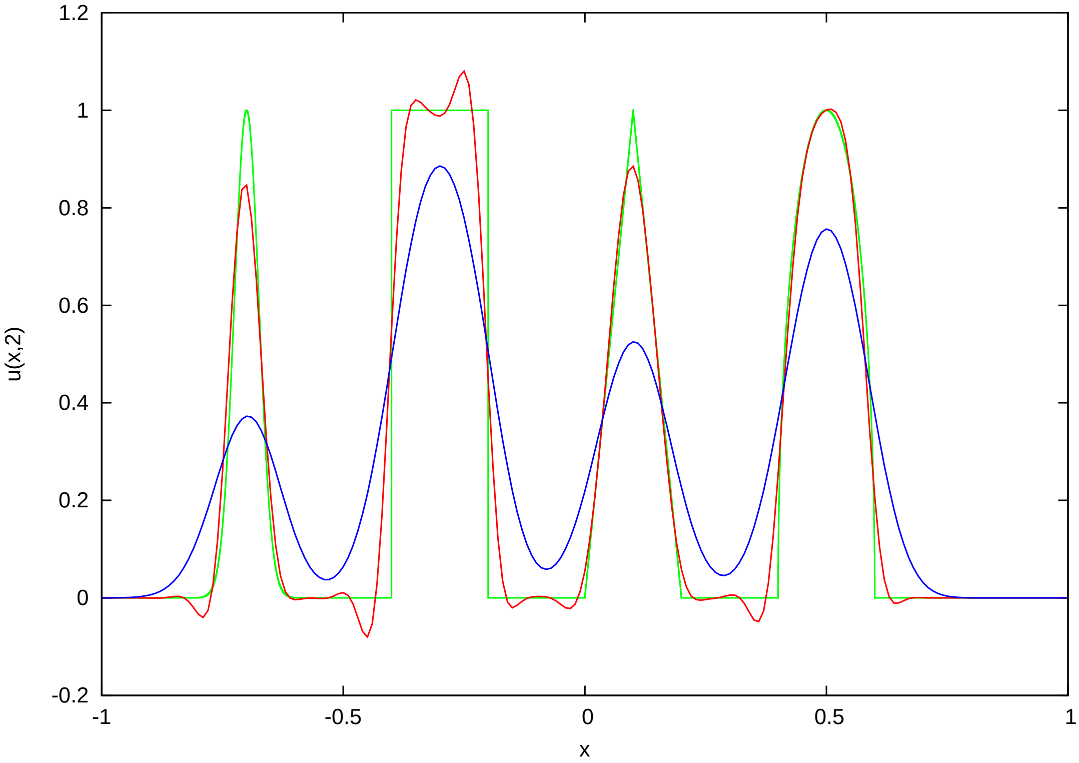


Figure 5.7: *Linear advection equation with peaks. HO22 and o1 for* $N = 200$*. The exact solution (green), the high order solution HO22 (red) and first order solution o1 (blue) are depicted at time* $t = 2$*.*

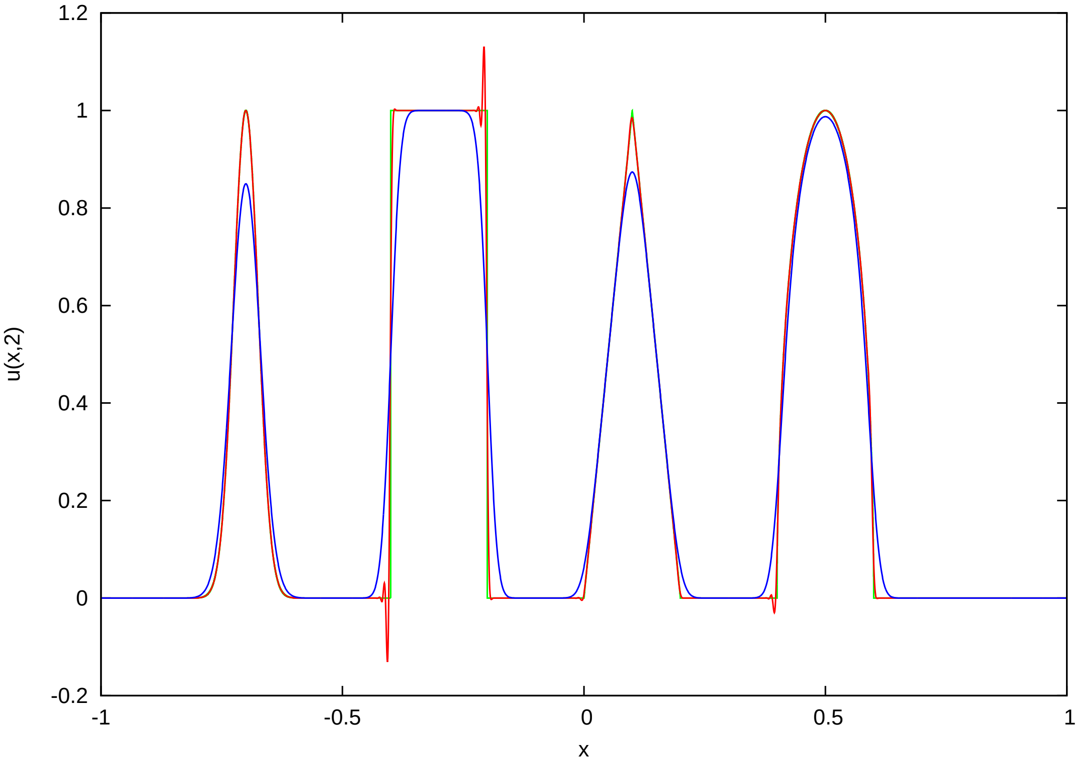


Figure 5.8: *Linear advection equation with peaks. HO22 and o1 for* $N = 4000$*. The exact solution (green), the high order solution HO22 (red) and first order solution o1 (blue) are depicted at time* $t = 2$*.*

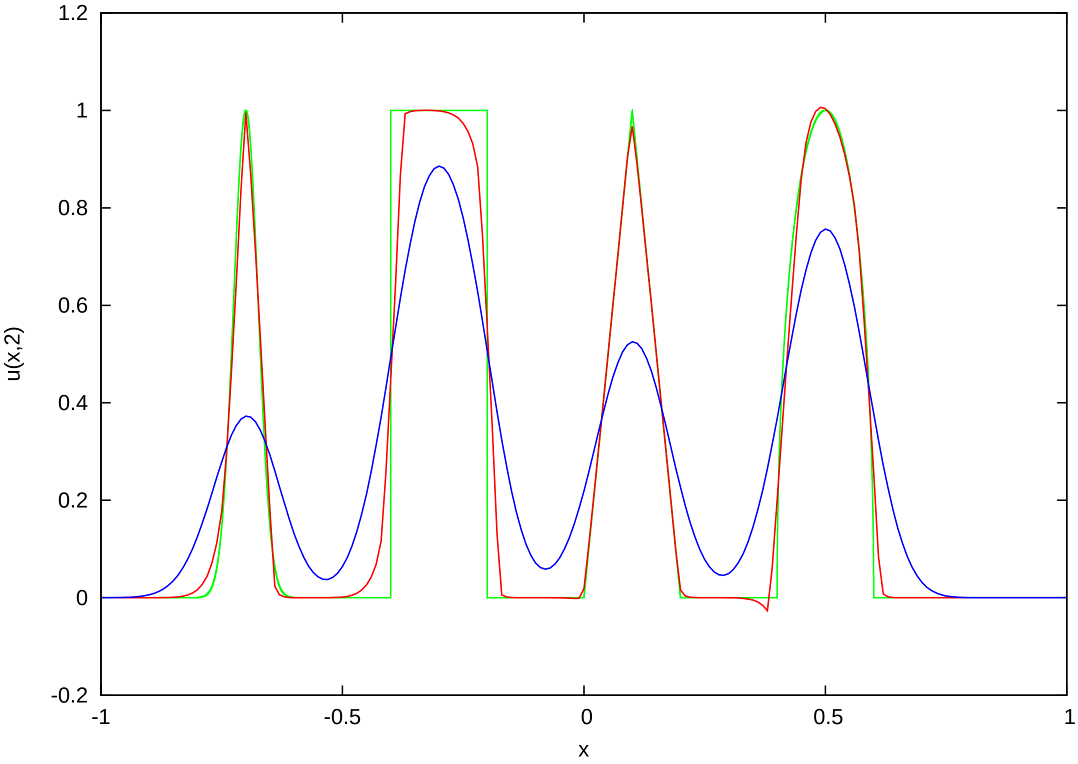


Figure 5.9: *Linear advection equation with peaks. HO32 and o1 for* $N = 200$*. The exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted at time* $t = 2$*.*

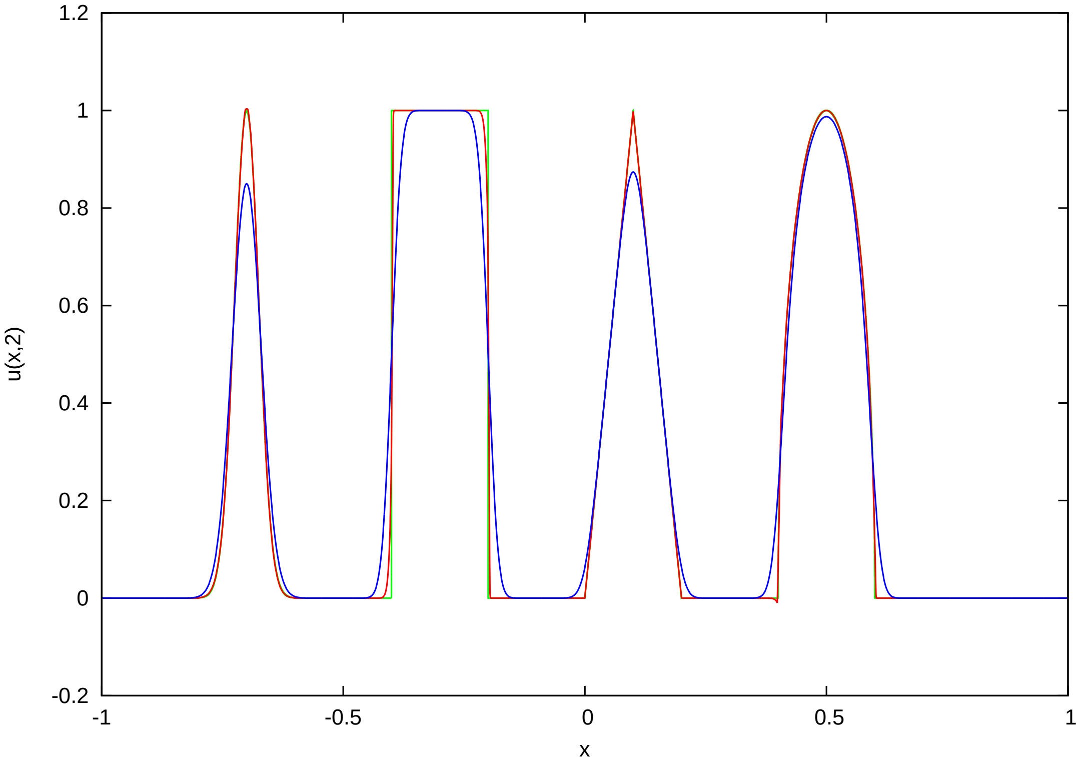


Figure 5.10: *Linear advection equation with peaks. HO32 and o1 for $N = 4000$. The exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted at time $t = 2$.*

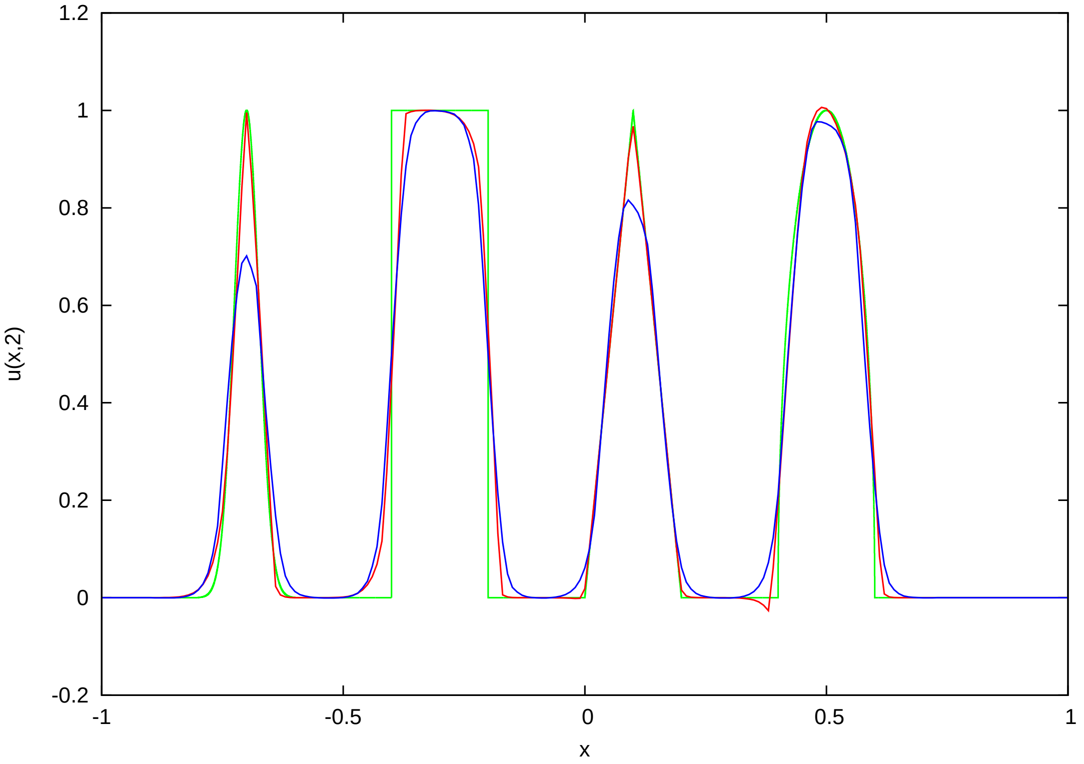


Figure 5.11: *Linear advection equation with peaks. HO32 and LS for $N = 200$. The exact solution (green), the high order solution HO32 (red) and high order solution LS (blue) are depicted at time $t = 2$.*

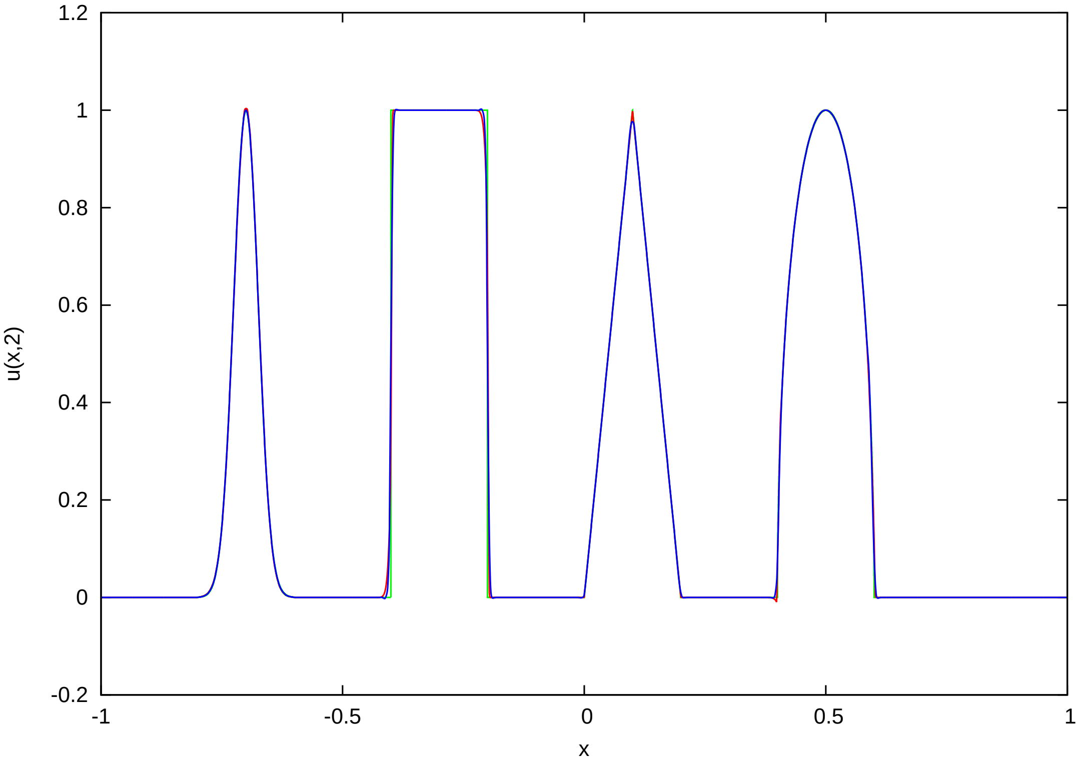


Figure 5.12: *Linear advection equation with peaks. HO32 and LS for* $N = 4000$*. The exact solution (green), the high order solution HO32 (red) and high order solution LS (blue) are depicted at time* $t = 2$*.*

5.2.4 Linearized gas dynamics with a smooth solution

Consider the problem of *linearized gas dynamics* presented in [64]

$$\mathbf{u}_t + \mathbb{A} \cdot \mathbf{u}_x \quad = \quad \mathbf{0} \quad , \quad x \in [-0.5, 0.5] \quad , \quad t \in [0,1] \ ,$$

where

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \rho \\ u \end{pmatrix} \quad , \quad \mathbb{A} = \begin{pmatrix} 0 & \rho_0 \\ a^2/\rho_0 & 0 \end{pmatrix}$$

with parameters a (sound speed) and ρ_0 (reference density). The unknowns are density ρ and velocity u. The exact solution is determined by the method of characteristics (see e.g., [64]). We set $a = 1$ and $\rho_0 = 1$.
As the initial condition, we choose the smooth function

$$\mathbf{u}(x,0) \quad = \quad \begin{pmatrix} 2 + \sin^4(2\pi x) \\ 0 \end{pmatrix}$$

in order to determine the order of convergence (compare also to the initial condition in the example *Euler equations with smooth solution*). We solve this equation with $CFL = 0.95$. The results for the higher order method $HO32$ and the first order method $o1$ can be found in tables 5.13 and 5.14 and the components of solutions in figures 5.13 and 5.14. We can see that our method is also applicable on systems of hyperbolic PDEs. We obtain second order of convergence in all three considered norms. The relatively big error of the high order scheme in the second component (see figure 5.14) is caused by the fact, that the second component of the exact solution is non-constant (consisting of “waves”) for $0 < t < 1$ and is equal to the constant 1 at the time $t = 1$. The method $HO32$ is not able to resolve this transition from non-constant solution to a constant one fine enough. The same behavior as for the high order method can be observed also for the first order method but with a much smaller magnitude. To avoid this behavior one can e.g., use smaller time steps to approximate the constant solution better. From table 5.13 we can see, that these computational “oscillations” will decrease quadratically, if the number of particles grows, i.e., this behavior has no influence on the convergence itself and on the convergence order.

	N	$E_1(N)$	k_1	$E_2(N)$	k_2	$E_\infty(N)$	k_∞
u_1	95	2.3643E-03	—	2.8817E-03	—	7.7941E-03	—
	190	4.6739E-04	2.34	6.7324E-04	2.10	2.1843E-03	1.84
	380	9.1872E-05	2.35	1.1754E-04	2.52	3.6661E-04	2.57
	760	1.9079E-05	2.27	2.3983E-05	2.29	5.9156E-05	2.63
	1520	4.7545E-06	2.00	5.9514E-06	2.01	1.4588E-05	2.02
u_2	95	1.6104E-02	—	1.8669E-02	—	3.2828E-02	—
	190	4.5389E-03	1.83	5.2142E-03	1.84	9.3228E-03	1.82
	380	1.1835E-03	1.94	1.3779E-03	1.92	2.5177E-03	1.89
	760	2.9900E-04	1.98	3.4916E-04	1.98	6.0508E-04	2.06
	1520	7.4857E-05	2.00	8.7466E-05	2.00	1.5196E-04	1.99

Table 5.13: *Linearized gas dynamics with a smooth solution. Errors and convergence order for the high order method* $HO32$ *in components* $u_1 = \rho$ *and* $u_2 = u$ *at time* $t = 1$.

	N	$E_1(N)$	k_1	$E_2(N)$	k_2	$E_\infty(N)$	k_∞
u_1	95	1.7581E-02	—	2.1033E-02	—	4.2642E-02	—
	190	8.7022E-03	1.01	1.0467E-02	1.01	2.1104E-02	1.01
	380	4.3250E-03	1.01	5.2160E-03	1.00	1.0473E-02	1.01
	760	2.1554E-03	1.00	2.6029E-03	1.00	5.2166E-03	1.01
	1520	1.0759E-03	1.00	1.3001E-03	1.00	2.6029E-03	1.00
u_2	95	9.4974E-04	—	1.1233E-03	—	1.9703E-03	—
	190	2.5788E-04	1.88	3.0308E-04	1.89	5.2719E-04	1.90
	380	6.7082E-05	1.94	7.8612E-05	1.95	1.3637E-04	1.95
	760	1.7099E-05	1.97	2.0011E-05	1.97	3.4666E-05	1.98
	1520	4.3156E-06	1.99	5.0472E-06	1.99	8.7384E-06	1.99

Table 5.14: *Linearized gas dynamics with a smooth solution. Errors and convergence order for the first order method o1 in components* $u_1 = \rho$ *and* $u_2 = u$ *at time* $t = 1$.

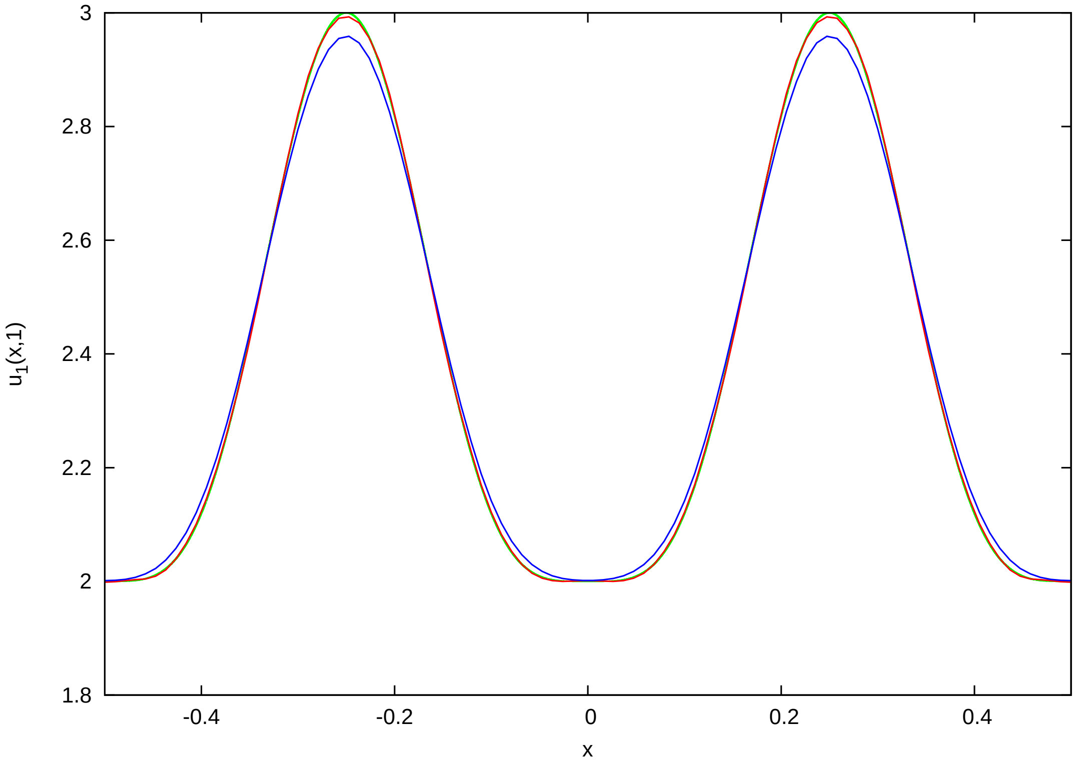


Figure 5.13: *Lin. gas dynamics with a smooth solution, first component. First component of the exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted for* $N = 95$ *at time* $t = 1$.

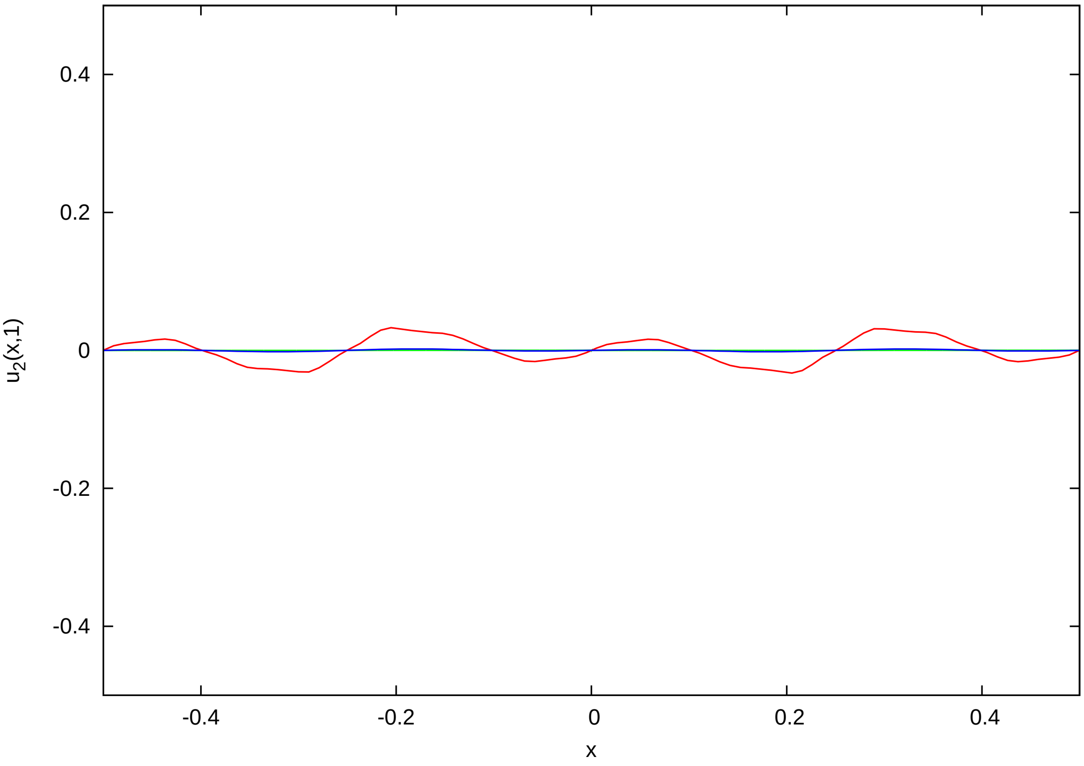


Figure 5.14: *Lin. gas dynamics with a smooth solution, second component. Second component of the exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted for $N = 95$ at time $t = 1$.*

5.2.5 Linearized gas dynamics with discontinuous solution

Consider the same governing equation as in the previous example on *linearized gas dynamics* with the same parameter values $a = 1$, $\rho_0 = 1$. The initial condition is set to be

$$\mathbf{u}(x,0) = \begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} & , \quad x < 0 \ , \\ \\ \begin{pmatrix} 0.5 \\ 0 \end{pmatrix} & , \quad x > 0 \ . \end{cases}$$

The exact solution is obtained with the method of characteristics and can be found in [64]. We solve this equation due to the discontinuities in the solution with $CFL = 0.8$ over the time interval $[0, 0.288]$ with fixed boundary conditions. The errors of solutions can be found in tables 5.15 and 5.16 and graphs in figures 5.15 and 5.16. We can see that in the case of linear hyperbolic system the constant parts of solutions are reproduced correctly by the method $HO32$ (compare to the following non-linear cases). Moreover, we acquire also better approximation on discontinuities in comparison to a classical first order method $o1$.

	N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
u_1	50	8.1157E-03	2.4019E-02	1.3042E-01
	100	4.7541E-03	1.8484E-02	1.3657E-01
	200	2.8778E-03	1.4671E-02	1.3605E-01
	400	1.7349E-03	1.1290E-02	1.3779E-01
	800	1.1191E-03	8.6132E-03	1.3968E-01
u_2	50	8.0230E-03	2.4015E-02	1.3043E-01
	100	4.7519E-03	1.8484E-02	1.3657E-01
	200	2.8778E-03	1.4671E-02	1.3605E-01
	400	1.7355E-03	1.1290E-02	1.3779E-01
	800	1.1400E-03	8.6139E-03	1.3968E-01

Table 5.15: *Linearized gas dynamics with discontinuous solution. Errors for the high order method* $HO32$ *in components* $u_1 = \rho$ *and* $u_2 = u$ *at time* $t = 0.288$.

	N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
u_1	50	1.4352E-02	3.2555E-02	1.2990E-01
	100	9.8694E-03	2.6956E-02	1.2869E-01
	200	6.8750E-03	2.2468E-02	1.2772E-01
	400	4.8245E-03	1.8809E-02	1.2698E-01
	800	3.3983E-03	1.5780E-02	1.2636E-01
u_2	50	1.4352E-02	3.2555E-02	1.2990E-01
	100	9.8694E-03	2.6956E-02	1.2869E-01
	200	6.8750E-03	2.2468E-02	1.2772E-01
	400	4.8245E-03	1.8809E-02	1.2698E-01
	800	3.3983E-03	1.5780E-02	1.2636E-01

Table 5.16: *Linearized gas dynamics with discontinuous solution. Errors for the first order method* $o1$ *in components* $u_1 = \rho$ *and* $u_2 = u$ *at time* $t = 0.288$.

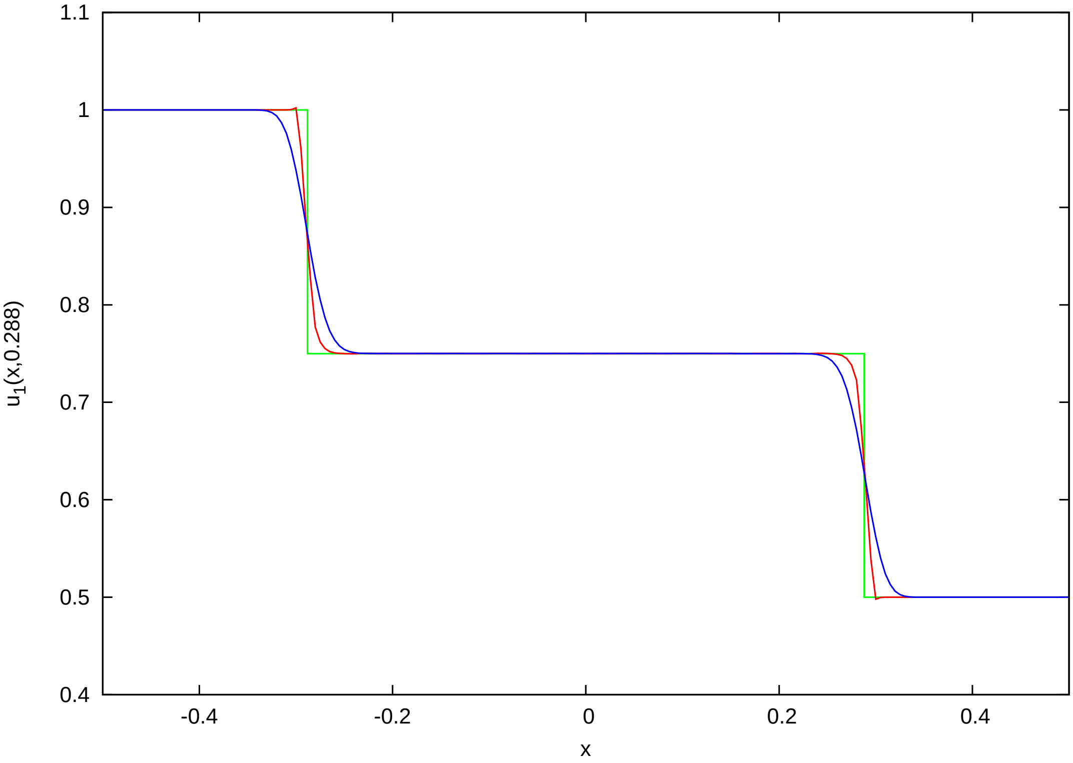


Figure 5.15: *Lin. gas dynamics with discontinuous solution, first component. First component of the exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted for* $N = 200$ *at time* $t = 0.288$*.*

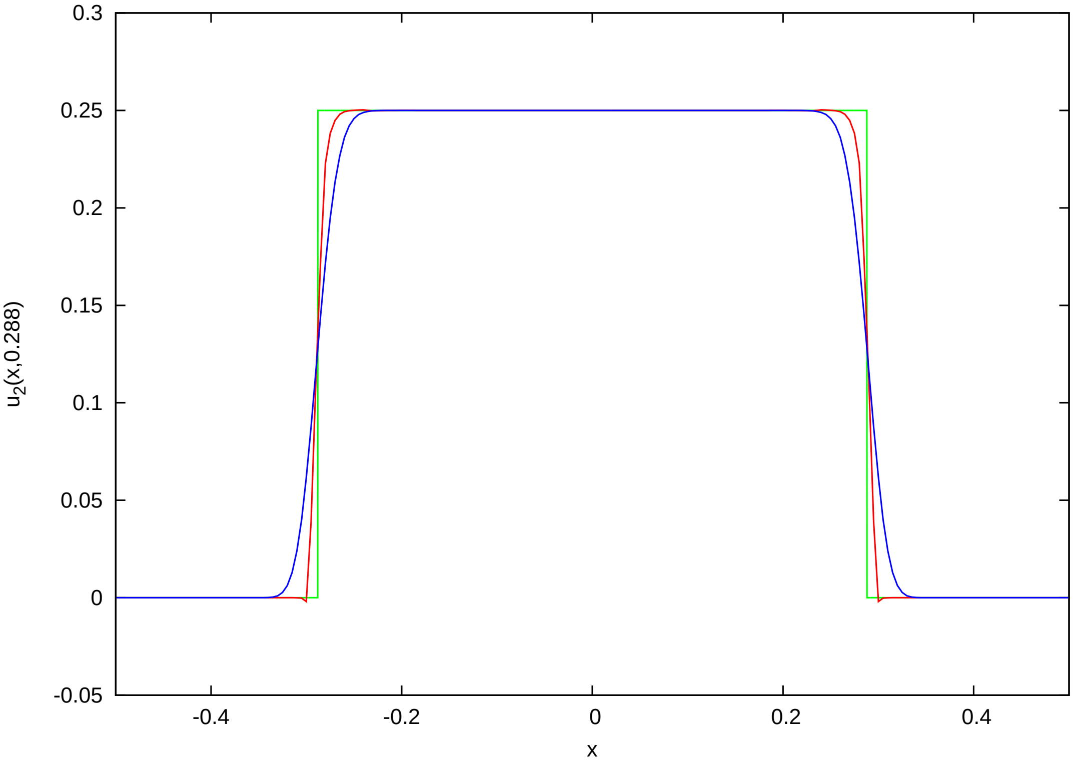


Figure 5.16: *Lin. gas dynamics with discontinuous solution, second component. Second component of the exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted for* $N = 200$ *at time* $t = 0.288$*.*

5.2.6 Shallow water equations - the dambreak problem

For more details about shallow water equations see e.g., [39] and [63].
In this example we consider the *shallow water equations* and the so-called *dambreak problem* given by the initial condition. This problem can be considered as a model for a real dambreak, where water is considered to be the investigated fluid. The problem reads

$$\mathbf{u}_t + \mathbf{F}(\mathbf{u})_x = \mathbf{0} \quad , \quad x \in [-2, 2] \quad , \quad t \in \left[0, \frac{32}{130}\right] \doteq [0, 0.24615]$$

with

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} h \\ hu \end{pmatrix} \quad , \quad \mathbf{F}(\mathbf{u}) = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \end{pmatrix} = \begin{pmatrix} u_2 \\ \frac{u_2^2}{u_1} + \frac{1}{2}gu_1^2 \end{pmatrix} ,$$

where $g = 9.81$ stands for acceleration due to gravity. The unknowns are water depth h and fluid velocity u. The initial condition is

$$\mathbf{u}(x,0) = \begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} & , \quad x < 0 \ , \\ \begin{pmatrix} 2 \\ 0 \end{pmatrix} & , \quad x > 0 \ . \end{cases}$$

We solve this equation due to the discontinuous initial condition with $CFL = 0.8$ with fixed boundary conditions. The exact solution is computed via the exact Riemann solver proposed by Toro in [63]. The resulting errors are given in tables 5.17 and 5.18, graphs in figures 5.17 - 5.19. We can see that we acquire smaller errors for the high order method $HO32$ rather than for the first order method $o1$. The discontinuity and the rarefaction wave are resolved better and are smeared out significantly less than for the first order method. In the parts where the solution remains constant, both methods are comparable and the constant states are conserved. Nevertheless, due to the non-linear character of the governing equation, we obtain non-physical oscillations in the middle part of the solution given by the high order method. Apparently, the WENO approach is not able to damp enough the oscillations in the case of non-linear hyperbolic systems (compare with the previous linear case of linearized gas dynamics) and possibly further techniques, e.g., limiters or a modification of the ADER method, have to be applied. On the other hand, the oscillations remain small in magnitude and do not spread out to further parts of solution, so that the character of solution remains correct with a good resolution of the discontinuity and rarefaction wave.

	N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
u_1	50	4.6763E-02	6.1898E-02	2.3905E-01
	100	2.2775E-02	3.8628E-02	2.5605E-01
	200	1.2791E-02	2.7460E-02	2.3443E-01
	400	7.4648E-03	2.0110E-02	2.3409E-01
	800	4.4334E-03	1.4134E-02	2.3307E-01
u_2	50	1.7408E-01	2.4887E-01	1.0703E+00
	100	8.4215E-02	1.5759E-01	1.1255E+00
	200	4.6565E-02	1.1593E-01	1.0477E+00
	400	2.5408E-02	8.4095E-02	1.0679E+00
	800	1.4909E-02	5.8593E-02	1.0568E+00

Table 5.17: *Shallow water equations - the dambreak problem. Errors for the high order method* $HO32$ *in components* $u_1 = h$ *and* $u_2 = hu$ *at time* $t \doteq 0.24615$.

	N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
u_1	50	1.1775E-01	1.0661E-01	2.3313E-01
	100	7.1722E-02	7.7310E-02	2.3305E-01
	200	4.2422E-02	5.4564E-02	2.3206E-01
	400	2.4579E-02	3.7782E-02	2.2873E-01
	800	1.4076E-02	2.5890E-02	2.2988E-01
u_2	50	4.4175E-01	4.1595E-01	1.0228E+00
	100	2.6753E-01	3.0335E-01	1.0240E+00
	200	1.5776E-01	2.1562E-01	1.0275E+00
	400	9.1239E-02	1.5067E-01	1.0457E+00
	800	5.2166E-02	1.0446E-01	1.0669E+00

Table 5.18: *Shallow water equations - the dambreak problem. Errors for the first order method* o1 *in components* $u_1 = h$ *and* $u_2 = hu$ *at time* $t \doteq 0.24615$.

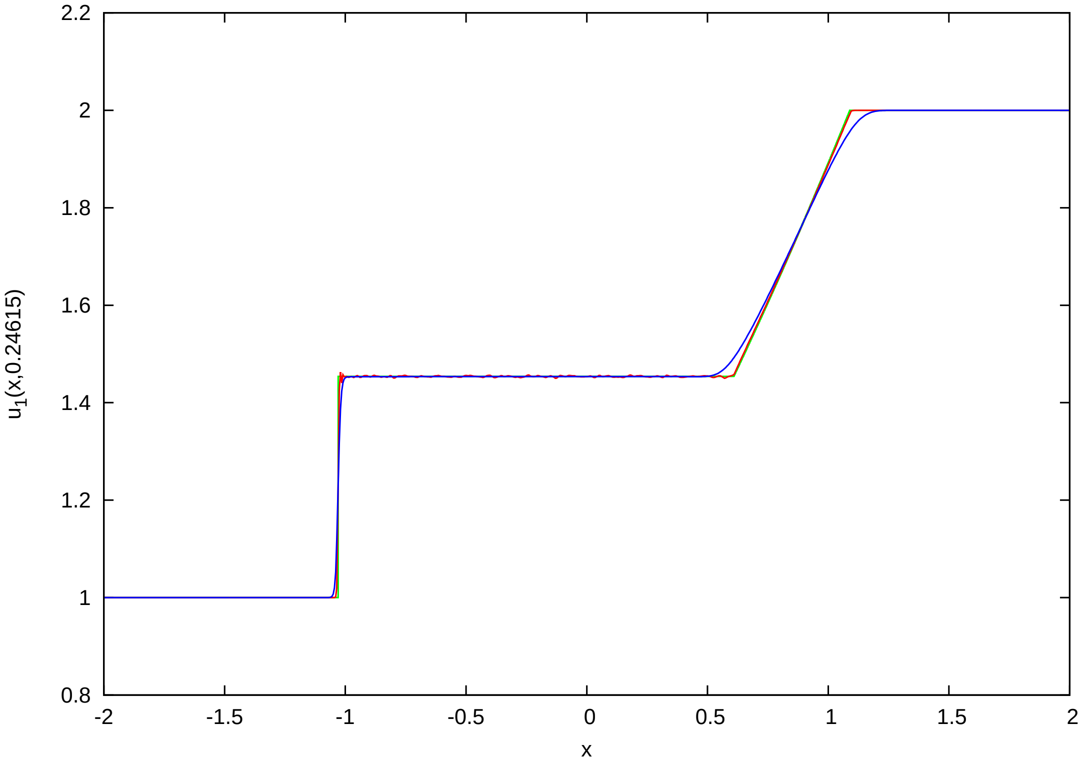


Figure 5.17: *Shallow water equations - the dambreak problem, first component. First component of the exact solution (green), the high order solution HO32 (red) and first order solution* o1 *(blue) are depicted for* $N = 800$ *at time* $t \doteq 0.24615$.

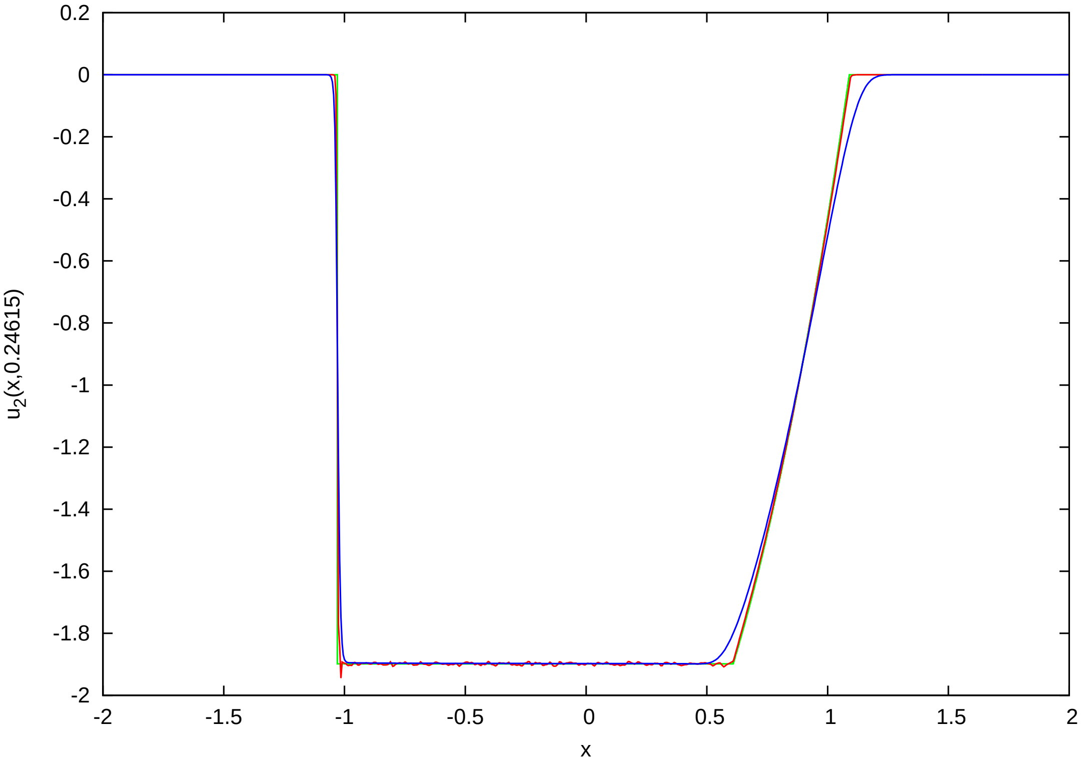
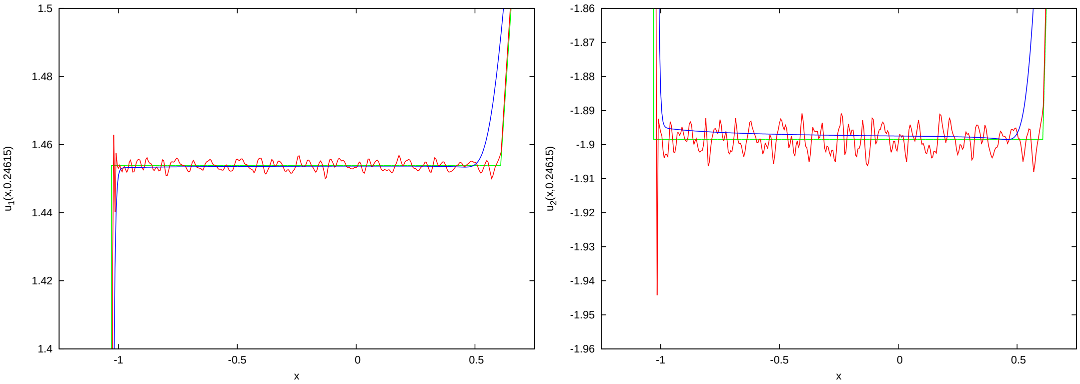


Figure 5.18: *Shallow water equations - the dambreak problem, second component. Second component of the exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted for* $N = 800$ *at time* $t \doteq 0.24615$.

Figure 5.19: *Shallow water equations - the dambreak problem. Zoom. The middle part of the solution with the same scale for h and hu is depicted. On the left the first and on the right the second component of solutions are depicted for* $N = 800$ *at time* $t \doteq 0.24615$. *The exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted.*



5.2.7 Euler equations with smooth solution

Consider the one-dimensional Euler equations from example 1.6

$$\mathbf{u}_t + \mathbf{F}(\mathbf{u})_x = \mathbf{0} \quad , \quad x \in [-0.5, 0.5] \quad , \quad t \in [0, 0.5]$$

with the vector of unknowns

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix}$$

and the physical flux

$$\mathbf{F}(\mathbf{u}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(E+p) \end{pmatrix} = \begin{pmatrix} u_2 \\ \frac{1}{2}(3-\gamma)\frac{u_2^2}{u_1} + (\gamma-1)u_3 \\ \gamma \frac{u_2}{u_1} u_3 - \frac{1}{2}(\gamma-1)\frac{u_2^3}{u_1^2} \end{pmatrix} .$$

The adiabatic exponent γ is chosen to be $\gamma = 1.4$. We follow the test problem provided by Titarev [61], and choose the initial values of ρ, u and p to be

$$\rho(x,0) = 2 + \sin^4(2\pi x) \quad , \quad u(x,0) = 1 \quad , \quad p(x,0) = 1 \ .$$

The exact solution is then given by

$$\rho(x,t) = 2 + \sin^4(2\pi(x-t)) \quad , \quad u(x,t) = 1 \quad , \quad p(x,t) = 1 \ .$$

We should remark here that it is usual to use the quantities ρ, u and p (also called *primitive* or *physical variables*) rather than the *conservative variables* ρ, ρu and E to describe the fluid dynamics given by Euler equations. We will follow this practice and analyse the results in the language of primitive variables ρ, u and p, even if the whole computation is done for conservative variables. For relevant relations see example 1.6.
We solve this equation with $CFL = 0.8$ with periodic boundary conditions. Since the solution is smooth, we determine also the convergence order. For the high order method $HO32$, the errors and experimental convergence orders for $u_1 = \rho$ can be found in table 5.19. The errors for u and p are in the same table. The figure 5.20 shows the results in comparison to the exact solution and to the first order method $o1$, notice the different scales. We can see that we get very good results for the quantity ρ in comparison to the first order method $o1$ and we can also deduce the second order of accuracy. However, for the quantities u and p we have to deal with “noise” in the solution. It is caused by the fact, that they are not the conservative variables, so that their constant value 1 is not preserved necessarily. Nevertheless, this noise is relatively small and gets smaller with the number of particles getting bigger. This noise behaves quite randomly and there is no relation between the noise in the solutions for different values of N, e.g., $N \in \{40, 80, \ldots\}$. That is why, we do not determine the order of convergence in the variables u and p.
Results for first order method are given in table 5.20.
Altogether we can conclude that the method $HO32$ achieves the second order of convergence also for non-linear hyperbolic systems.

	N	$E_1(N)$	k_1	$E_2(N)$	k_2	$E_\infty(N)$	k_∞
ρ	40	1.4405E-02	—	1.8785E-02	—	6.5954E-02	—
	80	5.0079E-03	1.52	6.5202E-03	1.53	2.1523E-02	1.62
	160	1.3767E-03	1.86	1.7504E-03	1.90	6.2411E-03	1.79
	320	3.5974E-04	1.94	4.5911E-04	1.93	1.7093E-03	1.87
	640	9.8186E-05	1.87	1.2858E-04	1.84	5.6486E-04	1.60
u	40	1.3291E-09	—	2.1595E-09	—	9.6460E-09	—
	80	9.9971E-09	—	2.1536E-08	—	1.1160E-07	—
	160	5.9818E-06	—	1.0447E-05	—	4.9602E-05	—
	320	2.5609E-06	—	4.5866E-06	—	2.7454E-05	—
	640	9.0074E-06	—	1.2536E-05	—	7.6327E-05	—
p	40	2.4764E-09	—	3.9297E-09	—	1.4939E-08	—
	80	1.7015E-08	—	3.6162E-08	—	1.8727E-07	—
	160	1.0575E-05	—	1.8144E-05	—	8.3781E-05	—
	320	4.8868E-06	—	8.8729E-06	—	4.7302E-05	—
	640	1.7048E-05	—	2.3750E-05	—	9.7393E-05	—

Table 5.19: *Euler equations with smooth solution. Errors and convergence order for the high order method* $HO32$ *in* ρ, u *and* p *at time* $t = 0.5$.

	N	$E_1(N)$	k_1	$E_2(N)$	k_2	$E_\infty(N)$	k_∞
ρ	40	1.6026E-01	—	1.8078E-01	—	3.4194E-01	—
	80	9.6227E-02	0.74	1.1070E-01	0.71	2.1721E-01	0.65
	160	5.3761E-02	0.84	6.3063E-02	0.81	1.2550E-01	0.79
	320	2.8607E-02	0.91	3.4016E-02	0.89	6.7966E-02	0.88
	640	1.4785E-02	0.95	1.7721E-02	0.94	3.5440E-02	0.94
u	40	2.6512E-16	—	3.7293E-16	—	1.1102E-15	—
	80	3.9293E-16	—	4.8993E-16	—	1.4433E-15	—
	160	3.4388E-16	—	5.0513E-16	—	3.9968E-15	—
	320	4.0785E-16	—	5.6237E-16	—	3.2196E-15	—
	640	7.0836E-14	—	8.6359E-14	—	1.5632E-13	—
p	40	4.9656E-13	—	4.9656E-13	—	4.9827E-13	—
	80	4.9654E-13	—	4.9654E-13	—	4.9827E-13	—
	160	4.9636E-13	—	4.9636E-13	—	4.9827E-13	—
	320	4.9661E-13	—	4.9661E-13	—	5.0004E-13	—
	640	6.7762E-13	—	6.9118E-13	—	9.2104E-13	—

Table 5.20: *Euler equations with smooth solution. Errors and convergence order for the first order method* $o1$ *in* ρ, u *and* p *at time* $t = 0.5$.

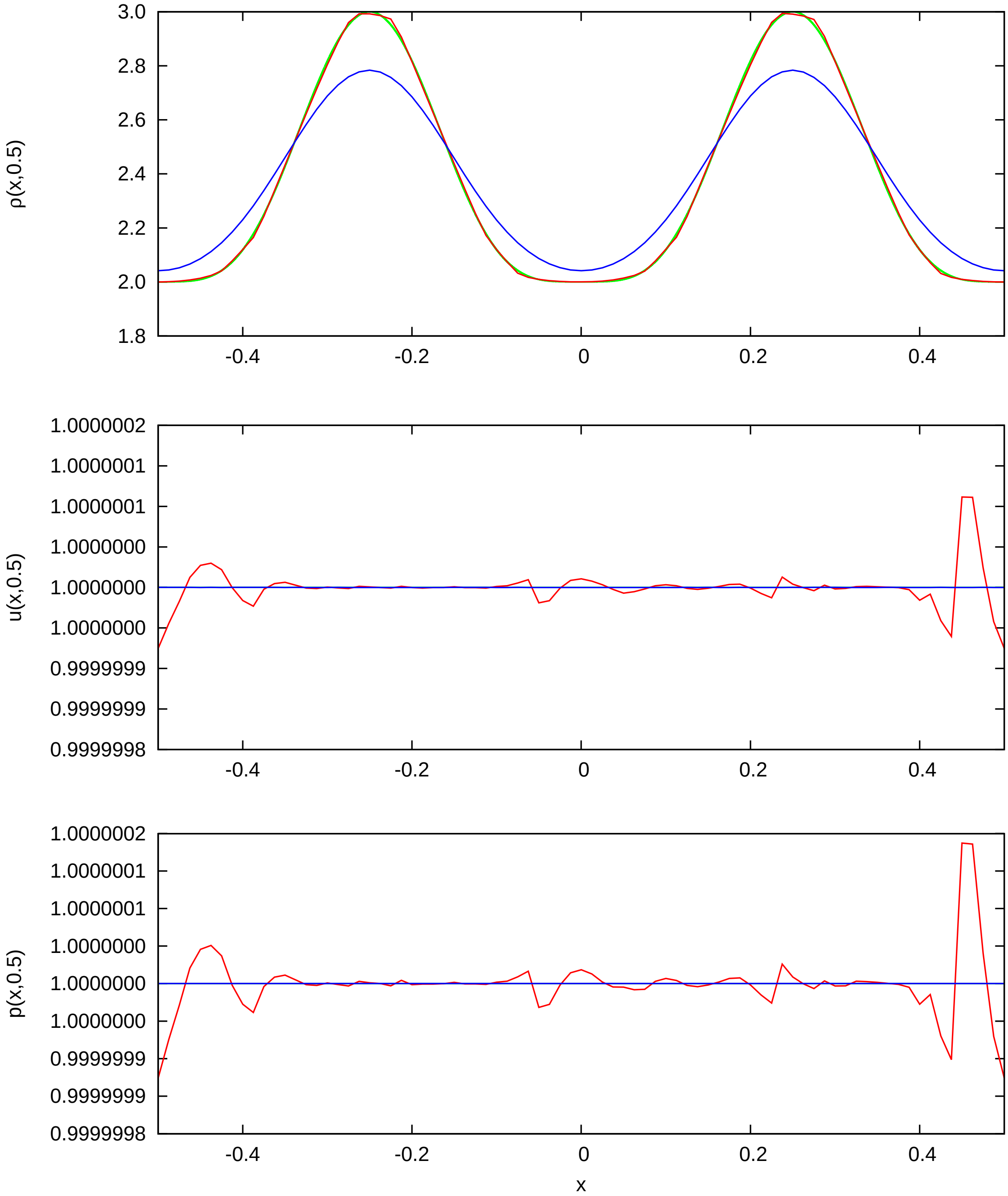


Figure 5.20: *Euler equations with smooth solution. Up to down graphs of* ρ*,* u *and* p *of solutions for* $N = 80$ *at time* $t = 0.5$ *with different scales. The exact solution (green), the high order solution* $HO32$ *(red) and first order solution* $o1$ *(blue) are depicted.*

5.2.8 Euler equations - the Sod problem

The last example is the so-called *Sod problem* ([54]), also known as *shock tube problem.* We consider again the Euler equations from the previous example, i.e.,

$$\mathbf{u}_t + \mathbf{F}(\mathbf{u})_x = \mathbf{0} \quad , \quad x \in [-0.5, 0.5] \quad , \quad t \in \left[0, \frac{13}{55}\right] \doteq [0, 0.23636]$$

with the vector of unknowns

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix}$$

and the physical flux

$$\mathbf{F}(\mathbf{u}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(E+p) \end{pmatrix} = \begin{pmatrix} u_2 \\ \frac{1}{2}(3-\gamma)\frac{u_2^2}{u_1} + (\gamma-1)u_3 \\ \gamma\frac{u_2}{u_1}u_3 - \frac{1}{2}(\gamma-1)\frac{u_2^3}{u_1^2} \end{pmatrix} .$$

The initial condition is

$$\begin{pmatrix} \rho \\ u \\ p \end{pmatrix}(x,0) = \begin{cases} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} & , \quad x < 0 \ , \\ \\ \begin{pmatrix} 0.125 \\ 0 \\ 0.1 \end{pmatrix} & , \quad x > 0 \ . \end{cases}$$

We set again γ to be $\gamma = 1.4$. We solve this equation with $CFL = 0.8$ with fixed boundary conditions. The exact solution is computed via the exact Riemann solver proposed by Toro in [64]. The resulting errors and graphs are given in tables 5.21 and 5.22 and figures 5.21 and 5.22, respectively. We obtain similar results as in the example on the shallow water equations. The constant states are preserved in the parts of solution where the solution is constant from the initial time. The shocks, contact discontinuity and rarefaction waves are approximated better with the high order method $HO32$ than with the first order method $o1$. This is a very good result especially for the contact discontinuity since the classical first order methods smear out the solution usually very strongly. In the middle parts of the solution we get again non-physical oscillations due to the non-linearity of the problem. Again, further techniques to suppress it can possibly be applied. However, the character of the solution is conserved and the oscillations do not spread out from the middle part of the solution.

	N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
ρ	100	5.4789E-03	1.1253E-02	8.6315E-02
	200	2.8617E-03	7.5084E-03	8.3835E-02
	400	1.5470E-03	5.1216E-03	8.2140E-02
	800	9.1642E-04	3.6779E-03	9.2145E-02
	1600	6.0233E-04	2.9021E-03	1.0157E-01
u	100	1.1347E-02	3.7863E-02	5.6965E-01
	200	6.0217E-03	2.9609E-02	5.6320E-01
	400	3.1242E-03	2.0477E-02	5.5908E-01
	800	1.6639E-03	1.2704E-02	5.8740E-01
	1600	1.1128E-03	1.0205E-02	5.5970E-01
p	100	5.0506E-03	1.0519E-02	1.1374E-01
	200	2.5918E-03	6.9785E-03	1.0224E-01
	400	1.3554E-03	4.4870E-03	1.0493E-01
	800	7.5058E-04	2.8963E-03	1.1720E-01
	1600	4.7378E-04	2.1023E-03	1.0554E-01

Table 5.21: *Euler equations - the Sod problem. Errors for the high order method* $HO32$ *in* ρ, u *and* p *at time* $t \doteq 0.23636$.

	N	$E_1(N)$	$E_2(N)$	$E_\infty(N)$
ρ	100	1.6467E-02	2.4948E-02	9.1735E-02
	200	1.0461E-02	1.8342E-02	8.9560E-02
	400	6.6143E-03	1.3794E-02	8.7690E-02
	800	4.1736E-03	1.0659E-02	8.6144E-02
	1600	2.6362E-03	8.4519E-03	8.4890E-02
u	100	2.4711E-02	6.0842E-02	6.1111E-01
	200	1.4109E-02	4.2447E-02	6.1957E-01
	400	7.9847E-03	2.9331E-02	6.1283E-01
	800	4.4845E-03	2.0214E-02	6.1846E-01
	1600	2.5083E-03	1.4231E-02	6.1476E-01
p	100	1.3702E-02	2.4644E-02	1.1473E-01
	200	8.2008E-03	1.6453E-02	1.1485E-01
	400	4.8433E-03	1.0837E-02	1.1427E-01
	800	2.8240E-03	7.0842E-03	1.1640E-01
	1600	1.6268E-03	4.6328E-03	1.1484E-01

Table 5.22: *Euler equations - the Sod problem. Errors for the first order method* $o1$ *in* ρ, u *and* p *at time* $t \doteq 0.23636$.

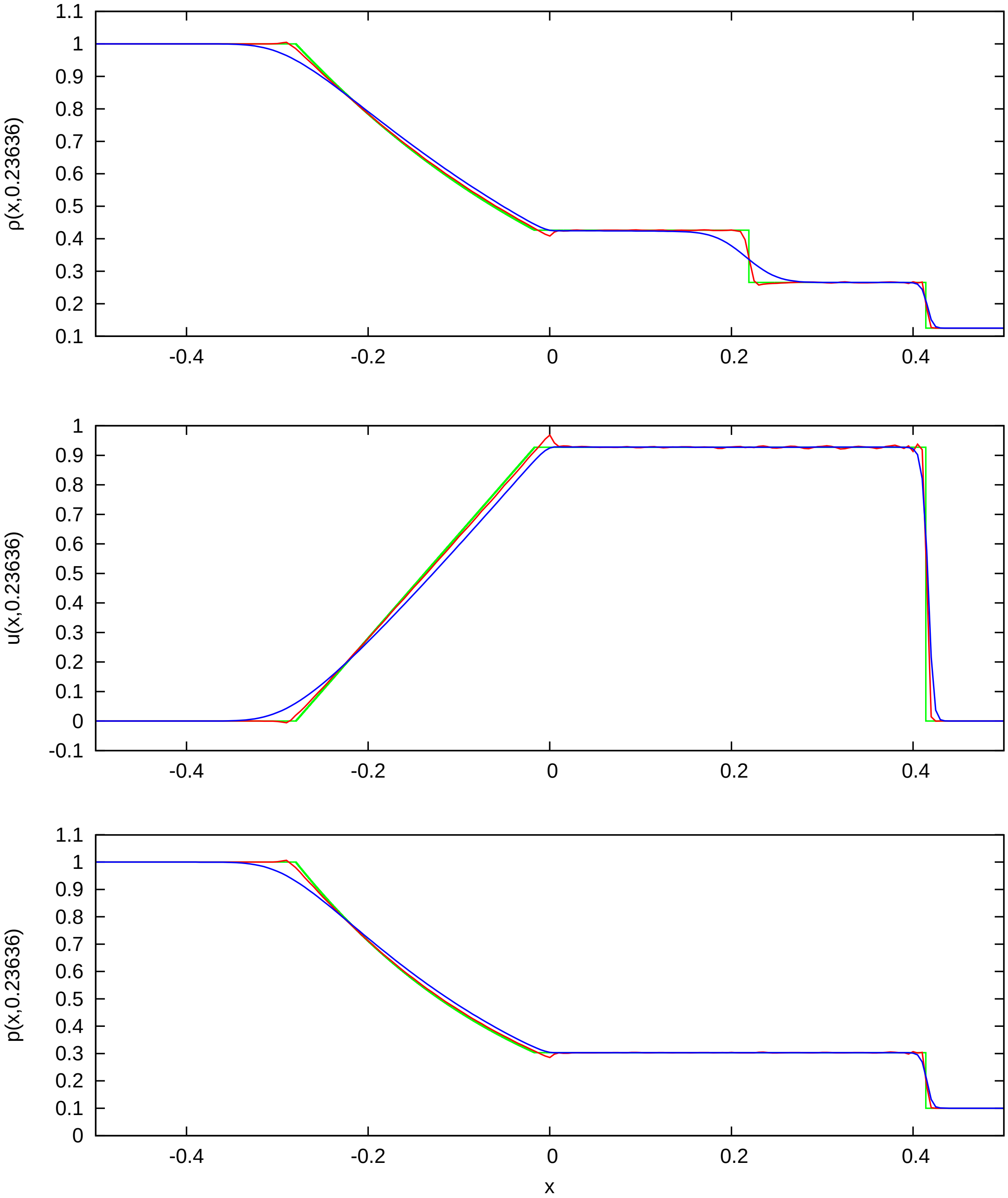


Figure 5.21: *Euler equations - the Sod problem. Up to down graphs of* ρ, u *and* p *of solutions for* $N = 200$ *at time* $t \doteq 0.23636$. *The exact solution (green), the high order solution HO32 (red) and first order solution o1 (blue) are depicted.*

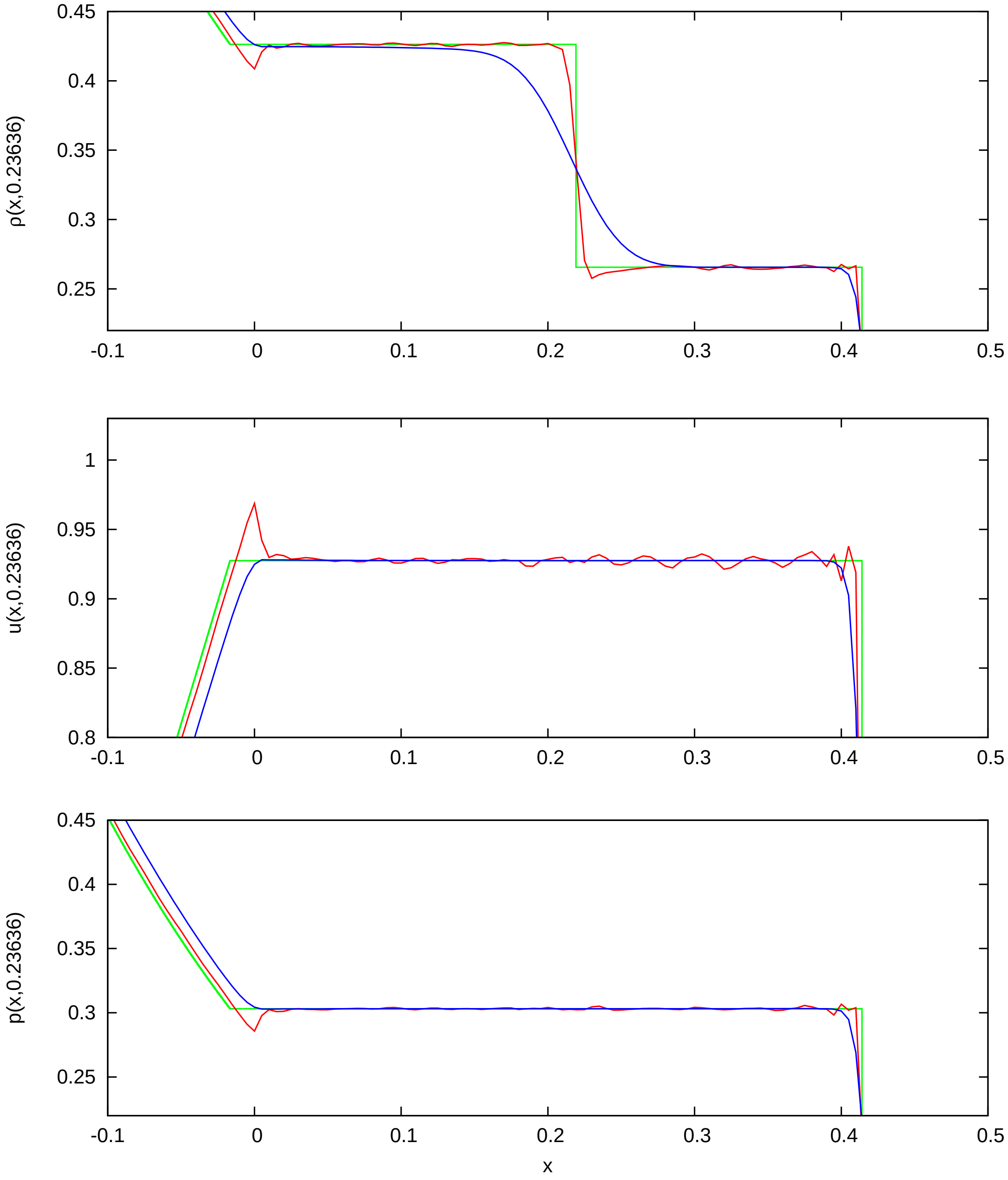


Figure 5.22: *Euler equations - the Sod problem. Zoom. The middle part of the solution with the same scale for* ρ*,* u *and* p *is considered. Up to down graphs of* ρ*,* u *and* p *of solutions for* $N = 200$ *at time* $t \doteq 0.23636$*. The exact solution (green), the high order solution* $HO32$ *(red) and first order solution* $o1$ *(blue) are depicted.*

Conclusions

We have dealt with hyperbolic conservation laws and their numerical treatment. The wide-spread finite volume method (FVM) is a very suitable tool to compute a numerical solution of these partial differential equations (PDEs). However, due to computational experience it is useful to develop new methods, that combine advantages of FVM and are not mesh-based, in order to increase the flexibility of the method. The proposals of Hietel, Steiner and Struckmeier [24] and Junk and Struckmeier [29] deal with this assignment and their finite volume particle method (FVPM) seems to be a good step to accomplish this goal. As FVM can attain higher order of accuracy, it is also desirable to construct FVPM of higher order.

Although FVPM were analyzed by various authors, providing computational stability of the method remains a challenging task we have tried to fulfill. It may seem to be of merely technical meaning, but it has a deeper mathematical background. We have shown, for the general formulation of FVPM, how to define the correction procedure for geometrical coefficients for bounded domains. Furthermore, we have introduced a scheme that enables us to add a new particle to an existing particle distribution, in order to preserve overlapping of the particles. A similar procedure was developed to remove an existing particle, in the case of high density of particles. We have shown, that both methods preserve constant states and are conservative, up to machine precision. Having defined these procedures, one can apply them on arbitrary FVPM to ensure stability in the above mentioned sense.

Polyharmonic spline interpolation (see [12], [26]) is a technique used to interpolate given data. The results for data given in the form of point values or classical integral means are already known. In this work, we have done the rigorous analysis for the case of weighted integral means. It was combined with the WENO approach (see [15], [25], [52]) to construct a FVPM of higher order. We have considered the ADER method (see [64], [65]), developed for the mesh-based FVM and we have shown, that it is possible to adapt principles of the ADER method also on a meshfree scheme, at least in one spatial dimension. For the proposed scheme, we have proven the second order of convergence for a scalar linear PDE. Further cases, such as non-linear PDEs or systems, were successfully tested numerically. We have observed, that the method is robust and attains second order of accuracy in areas where the solution is smooth; in non-smooth areas, the scheme yields at least a better resolution of shocks and rarefaction waves in comparison to a first order method. Even if the method seems to work well, there are difficulties concerning non-linear systems. This behavior needs more attention and has to be investigated in future. A remedy of the occurring oscillations could be the use of limiters or an analysis of the ADER scheme. As a matter of fact, one should take into account, that even the classical FVM ADER method with Toro-Titarev solver does not work properly in this case. Based on [18], one could try to modify the Toro-Titarev solver in the case of FVPM and make use of the LeFloch-Raviart expansion to circumvent the formation of oscillations.

The proposed method can be considered to be the first step in the construction of a method of arbitrary high order of convergence in arbitrary spatial dimensions. As known from the FVM framework, the idea of the ADER method allows even to achieve arbitrary high order of discretization in time and space. An even bigger challenge is to find a suitable and numerically efficient, but more general partition of unity given in FVPM. But having found this, the combination of the latter two steps may lead to a method of arbitrary high order of convergence. Also, another generalization of the method, namely a formulation of the method in more spatial dimensions, is desirable, since many practical computations take place in higher dimensions. Introducing a high order method combined with moving particles is also a matter of particular interest. Providing an outlook for further investigation, these assignments may become an object of research in the future.

Bibliography

[1] T. Aboiyar, E. H. Georgoulis, and A. Iske. High order WENO finite volume schemes using polyharmonic spline reconstruction. 2006.

[2] T. Aboiyar, E. H. Georgoulis, and A. Iske. Adaptive ADER Methods Using Kernel-Based Polyharmonic Spline WENO Reconstruction. *SIAM Journal on Scientific Computing*, 32(6):3251–3277, 2010.

[3] A. Bressan. *Hyperbolic systems of conservation laws: the one dimensional Cauchy problem*, volume 20. Oxford University Press, 2000.

[4] C. E. Castro and E. F. Toro. Alternative solvers for the derivative riemann problem for hyperbolic balance laws. *Isaac Newton Institute for Mathematical Sciences*, 2006.

[5] C. Chainais-Hillairet. Finite volume schemes for a nonlinear hyperbolic equation. Convergence towards the entropy solution and error estimate. *Mathematical Modelling and Numerical Analysis*, 33(1):129–156, 1999.

[6] F. Coquel and P. Le Floch. Convergence of finite difference schemes for conservation laws in several space dimensions: A general theory. *SIAM Journal on Numerical Analysis*, 30(3):675–700, 1993.

[7] C. Dafermos. *Hyperbolic Conservation Laws in Continuum Physics.* Springer, Heidelberg, 2000.

[8] P. J. Davis. *Interpolation and approximation.* Blaisdell New York, 1963.

[9] P. J. Davis and P. Rabinowitz. *Numerical integration.* Blaisdell Publishing Company London, 1967.

[10] E. DiBenedetto. *Partial Differential Equations.* Birkhäuser Boston, 2010.

[11] J. Duchon. Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *R.A.I.R.O. Analyse Numeriques*, 10(R3):5–12, 1976.

[12] J. Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *In: Constructive Theory of Functions of Several Variables, W. Schempp and K. Zeller (eds.), Springer*, pages 85–100, 1977.

[13] J. Duchon. Sur l'erreur d'interpolation des fonctions de plusieurs variables par les D^m-splines. *R.A.I.R.O. Analyse Numeriques*, 12(4):325–334, 1978.

[14] M. Feistauer, J. Felcman, and I. Straškraba. *Mathematical and computational methods for compressible flow.* Oxford University Press, 2003.

[15] O. Friedrich. Weighted essentially non-oscillatory schemes for the interpolation of mean values on unstructured grids. *Journal of Computational Physics*, 144(1):194–212, 1998.

[16] J. Glimm. Solutions in the large for nonlinear hyperbolic systems of equations. *Communications on Pure and Applied Mathematics*, 18(4):697–715, 1965.

[17] S. K. Godunov. A Finite Difference Method for the Computation of Discontinuous Solutions of the Equations of Fluid Dynamics. *Mat. Sb.*, (47):357–393, 1959.

[18] C. Goetz. Approximate Solutions of Generalized Riemann Problems for Hyperbolic Conservation Laws and Their Application to High Order Finite Volume Schemes. PhD Thesis, Universität Hamburg, 2013.

[19] M. Griebel and M. A. Schweitzer. A particle-partition of unity method for the solution of elliptic, parabolic, and hyperbolic PDEs. *SIAM Journal on Scientific Computing*, 22(3):853–890, 2000.

[20] M. Griebel and M. A. Schweitzer. *Meshfree methods for partial differential equations.* Number 1 in Lecture notes in computational science and engineering. Springer, 2003.

[21] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes, III. *Journal of Computational Physics*, 71(2):231–303, 1987.

[22] A. Harten and S. Osher. Uniformly High-Order Accurate Nonoscillatory Schemes I. *SIAM J. Numer. Anal.*, (24(2)):279–309, 1987.

[23] D. Hietel, M. Junk, R. Keck, and D. Teleaga. The finite-volume-particle method for conservation laws. In *Proc. of the GAMM Workshop Discrete Modelling and Discrete Algorithms in Continuum Mechanics*, pages 132–141, 2000.

[24] D. Hietel, K. Steiner, and J. Struckmeier. A Finite-Volume Particle Method for Compressible Flows. *Mathematical Models and Methods in Applied Sciences*, 10:1363–1382, 2000.

[25] C. Hu and C. Shu. Weighted essentially non-oscillatory schemes on triangular meshes. *Journal of Computational Physics*, 150(1):97–127, 1999.

[26] A. Iske. On the approximation order and numerical stability of local Lagrange interpolation by polyharmonic splines. *International Series of Numerical Mathematics*, pages 153–165, 2003.

[27] A. Iske. On the Construction of Kernel-Based Adaptive Particle Methods in Numerical Flow Simulation. In *Recent Developments in the Numerics of Nonlinear Hyperbolic Conservation Laws*, pages 197–221. Springer, 2013.

[28] G. S. Jiang and C. W. Shu. Efficient Implementation of Weighted ENO Schemes. 126:202–228, 1996.

[29] J. Junk, M. , Struckmeier. Consistency analysis of mesh-free methods for conservation laws. *GAMM-Mitteilungen*, 2:96–126, 2001.

[30] M. Junk. Do Finite Volume Methods Need a Mesh? *International Workshop on Meshfree Methods for PDE*, 2001. Bonn.

[31] C. Kaland. The Finite Volume Particle Method – Recent Developments and Applications. PhD Thesis, Universität Hamburg, 2013.

[32] R. Keck. The Finite Volume Particle Method: A Meshless Projection Method for Incompressible Flow. *PhD Thesis*, 2002. Univ. Kaiserslautern, Shaker Verlag.

[33] R. Keck and D. Hietel. A projection technique for incompressible flow in the meshless finite volume particle method. *Advances in Computational Mathematics*, 23(1-2):143–169, 2005.

[34] P. Knobloch. *Author's notices from the lecture "Approximate and Numerical Methods", Charles University, Prag, Czech Republic.* 2007.

[35] V. P. Kolgan. Application of the principle of minimizing the derivative to the construction of finite-difference schemes for computing discontinuous solutions of gas dynamics (in Russian). *Uch. Zap. TsaGI, Russia*, pages 68–77, 1972 (Reprinted and translated in: J. Comput. Phys., 230 (2011), pp. 2384 – 2390).

[36] D. Kröner. *Numerical schemes for conservation laws*, volume 22. Wiley Chichester, 1997.

[37] B. P. Lamichhane. The applications of finite volume particle method for moving boundary. 2000. Master Thesis, Universität Kaiserslautern.

[38] P. Lax and B. Wendroff. Systems of Conservation Laws. (13):217–237, 1960.

[39] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems.* 2002.

[40] T. Li and W. Yu. Boundary value problems for quasilinear hyperbolic systems. 1985.

[41] X. D. Liu, T. Osher, and S. Chan. Weighted Essentially Non-oscillatory Schemes. 115:200–212, 1994.

[42] W. R. Madych and S. A. Nelson. Multivariate interpolation: a variational theory. *unpublished preprint*, 1983.

[43] J. Meinguet. An intrinsic approach to multivariate spline interpolation at arbitrary points. *B.N. Sahney (Ed.), Polynomial and Spline Approximation, Reidel, Dordrecht*, pages 163–190, 1979.

[44] C. A. Micchelli. *Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions.* Springer, 1984.

[45] J. J. Monaghan. Smoothed particle hydrodynamics. *Annual review of astronomy and astrophysics*, 30:543–574, 1992.

[46] G. Montecinos, C. E. Castro, M. Dumbser, and E. F. Toro. Comparison of solvers for the generalized riemann problem for hyperbolic systems with source terms. *Journal of Computational Physics*, 231(19):6472–6494, 2012.

[47] R. M. Nestor, M. Basa, M. Lastiwka, and N. J. Quinlan. Extension of the Finite Volume Particle Method to Viscous Flow. *J. Comput. Phys.*, 228(5):1733–1749, March 2009.

[48] R. M. Quinlan, N. J. , Nestor. Fast exact evaluation of particle interaction vectors in the finite volume particle method. In *Meshfree Methods for Partial Differential Equations V*, pages 219–234. Springer, 2011.

[49] R. Schaback and H. Wendland. Characterization and construction of radial basis functions. In *Multivariate Approximation and Applications*, pages 1–24. Cambridge University Press, 2001.

[50] R. Schaback and H. Wendland. *Numerische Mathematik.* Springer Berlin, 2005.

[51] C. Schick. Adaptivity for particle methods in fluid dynamics. *Diploma Thesis*, 2000. Department of Mathematics, University of Kaiserslautern.

[52] C. W. Shu. High order weighted essentially non-oscillatory schemes for convection dominated problems. 51:82–126, 2009.

[53] J. Smoller. *Shock waves and reaction-diffusion equations.* Springer-Verlag New York Inc., 1983.

[54] G. A. Sod. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of Computational Physics*, 27(1):1–31, 1978.

[55] J. C. Strikwerda. *Finite difference schemes and partial differential equations.* Chapman and Hall New York, 1989.

[56] Y. Takakura and E. F. Toro. *Arbitrarily accurate non-oscillatory schemes for nonlinear scalar conservation laws with source terms.* 32nd AIAA Fluid Dynamics Conference and Exhibit 2002, St. Louis, Missouri, 2002.

[57] L. Tatsien and W. Libin. *Global Propagation of Regular Nonlinear Hyperbolic Waves.* Brikhäuser, Boston, 2009.

[58] D. Teleaga. Numerical Studies of a Finite-Volume Particle Method for Conservation Laws. *Master Thesis*, 2000. Dep. of Math., Univ. Kaiserslautern.

[59] D. Teleaga. A Finite-Volume Particle Method for Conservation Laws. *PhD Thesis*, 2005. Dep. of Math., Univ. Hamburg.

[60] J. Teleaga, D. , Struckmeier. A finite-volume particle method for conservation laws on moving domains. *Int. J. Numer. Meth. Fluids*, 58:945–967, 2008.

[61] V. A. Titarev. Derivative Riemann problem and ADER schemes. PhD Thesis, University of Trento, 2005.

[62] V. A. Titarev and E. F. Toro. Analysis of ADER and ADER-WAF schemes. *IMA journal of numerical analysis*, 27(3):616–630, 2007.

[63] E. F. Toro. *Shock-Capturing Methods for Free-Surface Shallow Flows.* Wiley and Sons Ltd, Chichester, 2001.

[64] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction.* Springer, 2009.

[65] E. F. Toro and V. A. Titarev. Derivative Riemann solvers for systems of conservation laws and ADER methods. *Journal of Computational Physics*, 212(1):150–165, 2006.

[66] B. van Leer. Towards the ultimate conservative difference scheme I. The quest of monotonicity. In *Lecture Notes in Physics*, pages 163–168, 1973.

[67] B. van Leer. Towards the ultimate conservative difference scheme. II. Monotonicity and conservation combined in a second-order scheme. *Journal of Computational Physics*, (14):361–370, 1974.

[68] J.-P. Vila. Convergence and error estimates in finite volume schemes for general multidimensional scalar conservation laws. I. Explicite monotone schemes. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 28(3):267–295, 1994.

[69] H. Wendland. *Scattered Data Approximation.* Cambridge University Press, 2005.

[70] Z. Yang. Efficient Calculation of Geometric Parameters in the Finite Volume Particle Method. Master Thesis, Universität Kaiserslautern, 2001.

Abstract

We study numerical methods for the solution of hyperbolic conservation laws with particular emphasis on meshfree methods. The concept of the Finite Volume Particle Method (FVPM) and properties of the scheme are presented. We contribute to the development of the method with new results concerning stability and order of accuracy. To provide computational stability of a general FVPM we propose algorithms to add and to remove a particle to a given particle distribution. Furthermore, we focus on one-dimensional scalar problems and design and analyse a FVPM of second order of accuracy. To this end, a kernel-based high order spatial reconstruction scheme is combined with the ADER approach for the flux evaluation. Polyharmonic splines are used as kernel functions in the reconstruction step. We analyse the local approximation order of polyharmonic splines for the case of data given by weighted integral means, as needed in FVPM. To suppress oscillations in the reconstruction, we use the WENO technique. We generalize the ADER method and the Toro-Titarev solver in order to apply them on a meshless scheme and provide hereby the solution of a corresponding generalized Riemann problem with initial data given by the WENO approximation by polyharmonic splines. The resulting scheme yields a prototype of highly flexible high order meshfree method. Numerical examples are given to show the second order of convergence and robustness of the method also for non-linear equations as well as for systems of conservation laws.

Zusammenfassung

Wir studieren numerische, insbesondere gitterfreie, Methoden zur Lösung hyperbolischer Erhaltungsgleichungen. Das Konzept der Finite Volumen Partikel Methode (FVPM) und ihre Eigenschaften werden präsentiert. Wir tragen zu der Entwicklung der Methode mit neuen Resultaten bezüglich der Stabilität und Genauigkeitsordnung bei. Um die Stabilität einer allgemeinen FVPM zu gewährleisten, entwerfen wir Algorithmen zum Hinzufügen und Entfernen eines Partikels bezüglich einer gegebenen Partikelverteilung. Darüber hinaus betrachten wir eindimensionale skalare Probleme und befassen uns mit der Konstruktion und Analyse einer FVPM zweiter Ordnung. Zu diesem Zweck kombinieren wir kernbasierte Rekonstruktion höherer Ordnung im Raum mit der ADER-Methode für die Flussauswertung. Als Kern-Funktionen in dem Rekonstruktionsschritt benutzen wir polyharmonische Splines. Für den in FVPM auftretenden Fall der Daten, die durch gewichtete Integraldurschnitte gegeben sind, analysieren wir die lokale Approximationsordnung der polyharmonischen Splines. Mögliche Oszillationen werden mittels des WENO-Verfahrens gedämpft. Wir verallgemeinern die ADER-Methode und den Toro-Titarev-Löser, um sie an gitterfreie Schemata anzuwenden, und lösen hiermit das entsprechende verallgemeinerte Riemann Problem mit Anfangsdaten, welche durch die WENO-Approximation mit polyharmonischen Splines gegeben werden. Das resultierende Schema stellt den Prototyp einer hochflexiblen gitterfreien Methode höherer Ordnung dar. Schließlich werden numerische Beispiele präsentiert, die die Konvergenz zweiter Ordnung und Robustheit des Schemas auch für nicht-lineare Gleichungen sowie für Systeme hyperbolischer Erhaltungsgleichungen zeigen.

Lebenslauf

Name: Libor Kadrnka
E-Mail: libor.kadrnka@gmail.com

25.01.1985	geboren in Ostrava, Tschechische Republik.
05/2004	Erhalt der Allgemeinen Hochschulreife (Abitur) am Gymnasium, Ostrava-Hrabůvka, Tschechische Republik.
09/2004 - 09/2007	Bachelor-Studium Mathematik an der Karls Universität in Prag, Tschechische Republik.
09/2007	Bachelor-Staatsexamen Mathematik mit der Note 1,0.
09/2007 - 09/2010	Master-Studium Numerische und Berechnungsmathematik an der Karls Universität in Prag, Tschechische Republik.
10/2008 - 09/2009	Teilnahme am Erasmus-Austauschprogramm: Studium der Mathematik an der Universität Hamburg.
09/2010	Master-Staatsexamen Numerische und Berechnungsmathematik mit der Note 1,0.
10/2010 - 09/2014	Wissenschaftlicher Mitarbeiter am Fachbereich Mathematik der Universität Hamburg.