Second-Order Implicit Methods for Conservation Laws with Applications in Water Supply Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

Vom Fachbereich Mathematik der Technischen Universität Darmstadt
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation

Tag der Einreichung:	20.10.2017
Tag der mündlichen Prüfung:	12.12.2017

Referent:	Prof. Dr. Jens Lang
1. Korreferent:	Prof. Dr. Gerd Steinebach
2. Korreferent:	Prof. Dr. Winnifried Wollner

von

Lisa Sabine Wagner, M. Sc.

aus Bayreuth

Darmstadt, D17
2018

Abstract

In this thesis, we develop and analyse numerical methods for the simulation of water transport processes in networks. In this context, the possibility of combining such a method with adjoint-based optimization algorithms is of special importance. These algorithms are used in a simulation-based assistance system which computes energy-optimized operation plans of drinking water supply networks.

In the first part, we develop and analyse suitable numerical methods to solve the so-called *water hammer equations* which describe the flow of water through pressurized pipes. From a mathematical point of view, the challenges are the hyperbolic character of this one-dimensional system on the one hand, and a possibly stiff source term modelling the friction effects on the other hand. For the time integration, we use so-called *strong stability preserving* (SSP) *singly-diagonal implicit Runge-Kutta* (SDIRK) methods. Such methods are advantageous with respect to their numerical implementation and further, they preserve the nonlinear stability which is an important property in the context of hyperbolic partial differential equations. Concerning hyperbolic equations, there are two important characteristic features which numerical methods need to display – being conservative and handling discontinuities and shocks. For this reason, we use *Finite Volume* and *Discontinuous Galerkin* methods for the spatial discretization.

For the fully discrete schemes, which are combinations of the schemes mentioned above, we derive important properties: *well-balancedness* with respect to the water hammer equations and a *discrete maximum principle*. As a result, the numerical methods are able to exactly approximate the stationary state of the water hammer equations, which can be used to prove asymptotic stability. Further, the numerical solution which is computed by the methods lies in a certain range, which depends on the initial condition. All theoretical results are additionally verified by numerical tests.

The results presented here were achieved within a project that aims to develop a simulation-based assistance system for drinking water supply. We therefore describe the structure of the entire system in the second part of the thesis. In particular, we take a closer look at the incorporated optimization module and the model equations for all network components. The assistance system is capable of successfully reducing the energy consumption of the whole network, which we demonstrate by two examples based on real data provided by our project partners.

Zusammenfassung

Diese Dissertation beschäftigt sich mit der Entwicklung und Untersuchung numerischer Verfahren zur Simulation des Wassertransports in Netzwerken. Dabei spielt insbesondere die Kombinierbarkeit mit adjungiert-basierten Optimierungsalgorithmen eine Rolle, welche in einem simulations-basierten Assistenzsystem zur Berechnung energie-optimierter Betriebsfahrpläne für Trinkwasserversorgungsnetze Anwendung finden.

Im ersten Teil dieser Arbeit entwickeln und untersuchen wir geeignete numerische Verfahren zur Lösung der *Water-Hammer Gleichungen*, die den Wasserfluss in einem druckbehafteten Rohr beschreiben. Die Schwierigkeiten hierbei sind einerseits der hyperbolische Charakter dieses eindimensionalen Systems und andererseits ein möglicherweise steifer Quellterm für die Modellierung von Reibungseffekten. Für die Zeitintegration benutzen wir sogenannte *strong stability preserving* (SSP) *singly-diagonal implicit Runge-Kutta* (SDIRK) Verfahren, welche sowohl Vorteile bezüglich der numerischen Implementierung haben, als auch die nichtlineare Stabilität erhalten. Diese stellt insbesondere im Kontext von hyperbolischen Gleichungen eine wichtige Eigenschaft dar. Da bei hyperbolischen partiellen Differentialgleichungen die Erhaltungseigenschaft von Bedeutung ist und insbesondere Unstetigkeiten auftreten können, verwenden wir *Finite-Volumen* und *Discontinuous Galerkin* Verfahren zur Ortsdiskretisierung.

Für die Volldiskretisierung in Form einer Kombination der oben genannten Verfahren leiten wir wichtige Eigenschaften her: *Well-balancedness* bezüglich der Water-Hammer Gleichungen und ein *diskretes Maximumsprinzip*. Damit erfüllt das numerische Verfahren den stationären Zustand der Water-Hammer Gleichungen exakt, womit auch die asymptotische Stabilität nachgewiesen werden kann. Ferner liegt die numerische Lösung innerhalb eines durch die Anfangsbedingung festgelegten Intervalls. Alle theoretischen Ergebnisse werden mithilfe numerischer Tests verifiziert.

Diese Arbeit entstand im Rahmen eines Projekts, das sich mit der Entwicklung eines simulations-basierten Assistenzsystems in der Trinkwasserversorgung beschäftigt. Im zweiten Teil der Dissertation beschreiben wir daher die Gesamtstruktur des Systems und gehen insbesondere auf das eingebaute Optimierungsmodul und die Modellgleichungen für alle vorhandenen Netzwerkkomponenten ein. Das Assistenzsystem ist in der Lage, den Energieverbrauch des gesamten Netzwerks erfolgreich zu reduzieren, was wir anhand zweier Beispiele basierend auf realen Daten illustrieren.

Acknowledgement

First, I would like to thank my supervisor Prof. Dr. Jens Lang. During the three years he guided me through important research aspects and at the same time, he gave me enough space to realize my own research ideas. I am very thankful for his support.

Next, I want to say thank you to all members and former members of the research group *numerical analysis and scientific computing* of the TU Darmstadt for the nice working atmosphere. Especially, I want to thank our secretary Elke Dehnert for her permanent efforts during the entire three years.

Special thanks also goes to my family, namely my parents Gislinde and Erwin and my sister Carina who supported me in my whole life and in all my qualifications. I also want to mention my partner Tobias who already supported me during the mathematics studies.

Last but not least, I want to thank my project partners for the helpful comments and discussions and a fruitful collaboration between the whole project phase. Especially, I want to thank Oliver Kolb for his permanent support of my research ideas and for his helpful comments.
Further, this thesis was founded by the Federal Ministry of Education and Research (BMBF) founding measure *Future-oriented Technologies and Concepts for an Energy-efficient and Resource-saving Water Management* (ERWAS) especially within the project *Energy-Management system Water Supply* (EWave).
This work was partially supported by the GSC CE 233 and the special research area (SFB) Transregio TRR154 *Mathematische Modellierung, Simulation und Optimierung am Beispiel von Gasnetzwerken.*

Contents

1 Introduction

Supply with drinking water plays an important role in our daily lives. Water supply companies operate numerous plants distributed across Germany to sustain the population with drinking water. Water is dispensed through sophisticated water supply networks. To ensure steady supply for all consumers, complex operation plans are being developed. However, this process involves a great deal of human interaction and expertise. Over the last decades, automated water management systems have become more and more important, especially for the reduction of energy consumption [54].

For the simulation and optimization of complete water supply networks, several non-trivial steps need to be undertaken. As a first step, all single components of the network including storage tanks, pumps, filter stations and many more need to be modelled mathematically. Due to the variety of components, different kinds of equations like ordinary, partial and algebraic differential equations are involved and thus, the coupling needs to be done appropriately. There are different software packages for the simulation of networks, for instance EPANET [66], STANET [76] and KANET [38]. However, only the steady state equations are used in these codes. Note that the optimization of those networks is a much more complex task due to the complexity and size of the system. Available software packages usually have problems computing the correct solution [3, 16].

In order to close this gap, the project *EWave* (**E**nergiemanagement **Wa**sser**Ve**rsorgung) which is part of the BMBF group project ERWAS [19], was initiated. This thesis was established within this project, whereby the overall goal was the development of a water management system that computes energy-optimized operation plans for water supply networks. Note that we use time-dependent equations for the simulation and optimization of water supply networks which especially enables us to incorporate real-time constraints as well. In particular, we developed an automated energy management system which acts and reacts in an online modus and which was eventually implemented and tested on an existing complex water work of our industrial partner. Here, automated needs to be understood in the sense that the interaction of the user is drastically reduced. Even though within ERWAS, also other projects like EnWasser, H2Opt, and ENERWA [19] deal with energy issues in drinking water supply. The main difference besides the available online modus is that EWave can deal with much more complex networks.

We describe the general structure of the automatic energy-optimized management system in Chapter 6. The system is built of different modules which were first developed separately and eventually coupled. The core of these modules, namely the simulation or optimization module, consists of different software packages, for instance ANACONDA (**A**daptive **N**umerical **A**lgorithms for **C**ontrol **O**ptimization on **N**etworks **DA**rmstadt) [42] stemming from a previous project, or TWaveSim [77] which was established within EWave. At this point, we describe the optimization module and the mathematical modelling of typical network components used in water supply networks in more detail. In particular, we describe the simulation and optimization tool ANACONDA, which functioned as a basis of the optimization module and was continuously extended. Finally, we present some results of the EWave system in the online modus and show the efficiency of the system with respect to energy-optimized operation plans.

Most components of the water supply network are connected via pipes which we therefore consider as components of special importance. In addition to this and in contrary to the other components, water flow through pressurized systems [1] is typically described by hyperbolic partial differential equations (PDE). The main mathematical contribution of this thesis is the development and analysis of suitable numerical methods for the computation of water flow through pressurized pipes. Such methods constitute the basis for an optimization algorithm and their efficient solution plays a crucial role for the overall performance of the water management system. Here, we use a coupled one-dimensional system of balance laws, the *water hammer equations* (WHE) [2]. These equations pose a special difficulty in the sense that they involve a possibly stiff source term that models the friction within the pipe. Therefore, in Chapter 2, we describe the structure of the source term in a detailed way and also analyse the complete system mathematically.

The design of numerical methods suitable for hyperbolic equations has a long history [10, 11, 12, 24, 32, 51, 70]. A crucial requirement of these methods is that they must be able to resolve discontinuities since hyperbolic equations naturally admit shocks or discontinuities even for smooth data. In [42, 43], the so-called IBOX (**I**mplicit BOX) scheme which is also implemented in ANACONDA has been developed to solve the water hammer equations. Even though the scheme is stable and easy to implement, it is not optimal for some aspects. It is a lower-order scheme and introduces a relatively large amount of numerical diffusion. To overcome these drawbacks, we want to develop higher-order methods, where the focus is on higher-order time integration. For this purpose, we use so-called *singly-diagonally implicit* Runge-Kutta (SDIRK) methods [8, 31, 59] that allow us to keep computation times moderate, cf., Section 3.1. Another reason for using such methods is that they can handle the stiff source term present in this system. For the accurate approximation of e.g. shocks, we use so-called SSP (**S**trong **S**tability **P**reserving) methods [22, 26, 35, 40]. Such methods allow us to maintain the nonlinear stability for hyperbolic PDE.

As mentioned before, we aim to use higher-order time integration in a combination with lower-order spatial discretization. Therefore, we use the Method of Lines approach (MOL) to combine a SSP SDIRK method with well-known spatial discretizations including lower- and first-order finite volume methods (FV) and discontinuous Galerkin schemes (DG) with linear ansatz and test functions, cf., Section 3.2. Both discretizations are well suited for the solution of hyperbolic problems, especially since they can be formulated in *conservative form*. Even though some results can be extended to higher-order schemes in a natural way, we note that lower- or first-order discretizations are sufficient in our setting, namely the computation of networks, because we want keep the spatial mesh as coarse as possible. Finally, we state and analyse the fully discrete schemes which are a combination of the introduced temporal and spatial discretizations in Section 3.3. To maintain the total variation diminishing (TVD) property for first-order spatial discretizations, limiters can be used [81, 85] to avoid a *Gibbs phenomenon* [25]. We discuss suitable slope and flux limiters which are applicable for both the DG and FV method in Section 3.4.

The main results are stated in Chapter 4 where we establish some important properties. First, we show that the numerical methods admit a unique solution where we utilize a one-sided Lipschitz condition for the semi-discretization. This enables us to directly show the *well-balancedness* [57] with respect to the WHE. This property ensures that the numerical solutions approximate the steady states of the analytical system exactly. Further, this can be used to show that the considered methods are asymptotically stable. Second, we show that the numerical methods fulfill a *discrete maximum-principle* [53, 88, 89]. This kind of stability property ensures that the numerical solution stays in a certain range determined by the initial conditions of the system. Note that this is also true for the system case if the range is chosen appropriately.

For verification of all theoretical properties, we show some numerical results for the developed methods in Chapter 5. All methods were implemented using MATLAB [37]. In order to verify the maximum-principle, we show results for linear and nonlinear scalar conservation laws, i.e. the transport equation and the Buckley-Leverett equation. We also use the IBOX scheme introduced before for comparison. Further, we apply all methods to the WHE with and without source term. In both cases we verify the maximum-principle (type) property. Further, we show examples for the exact approximation of the stationary state.

2 Modelling of Water Flow in pressurized Pipes

Before dealing with water supply networks (c.f. Chapter 6), we turn to the network component that we focus on in the following, i.e. pressurized pipes. While other network components distribute and control the flow through the network, the actual transport of water happens in pipes. In the following, we motivate and analyse a suitable model for the flow of water through pressurized pipes.

2.1 The water hammer equations

Water is a nearly incompressible and Newtonian fluid. Its fluid dynamical behaviour is therefore described by the *Navier Stokes equations* [82]. We assume that the temperature and therefore also the density of the fluid is constant. In this case, the flow is described by the conservation of mass and momentum.

From a modelling point of view, these equations describe far more complex effects than those appearing in the flow through pipes. We consider a straight pipe with constant diameter d and length L with $L >> d$. Further, $\mathcal{A}$ denotes the cross sectional area, $\mathbf{g_r}$ is the gravitational acceleration constant and $\mathbf{a}$ the speed of sound in water. The first two are considered to be constant, while $\mathbf{a}$ depends on the pipe's wall thickness, its diameter, the modulus of elasticity and Poisson's ratio of the wall material [55]. Integration over the cross-section of the pipe and assuming the transversal velocities to be zero results in the one-dimensional continuity equation

$$\partial_t \mathrm{h}(x,t) + \frac{\mathbf{a}^2}{\mathbf{g_r}\mathcal{A}} \partial_x \mathrm{Q}(x,t) = 0, \tag{2.1}$$

where t denotes time and x the position along the pipe. The unknown variables h and Q are the *piezometric head* and the *flow rate*, respectively. Note that for the piezometric head (or *pressure head*), we have $\mathrm{h}(x,t) = z_0 + \frac{p(x,t)}{\varrho_0 \mathbf{g_r}}$, where p denotes the pressure in the fluid, ϱ_0 is the density and z_0 the so-called *geodetic head* of the pipe. The flow rate is related to the flow velocity v via $\mathrm{Q} = \mathcal{A}v$. Note that the one-dimensional continuity equation in this form can be derived from the equation for the calculation of water hammer by Allievi [2].

A one-dimensional equation for the conservation of momentum can be derived in a similar fashion from the Navier Stokes equations. Following [55, Chap. 7], we have

$$\partial_t \mathrm{Q}(x,t) + \mathbf{g_r}\mathcal{A}\partial_x \mathrm{h}(x,t) = -\lambda(\mathrm{Q}(x,t)) \tfrac{\mathrm{Q}(x,t)|\mathrm{Q}(x,t)|}{2d\mathcal{A}} =: -\mathsf{g}(\mathrm{Q}(x,t)). \tag{2.2}$$

The expression $\mathsf{g}(\mathrm{Q})$ is called *source term* and models the influence of the wall roughness. The *friction coefficient* λ depends on the wall shear stress τ_0, [55, Chap. 7, p. 362]. Even though τ_0 is not known a priori, several models for λ which only depend on Q are available. Regarding the flow of water through pipes, we distinguish between laminar and

turbulent flow, resulting in different formulas for λ [5, 55]. We introduce the *Reynolds number*

$$Re(\mathrm{Q}) = \frac{|v|d}{\nu} = \frac{4|\mathrm{Q}|}{\pi d\nu}, \tag{2.3}$$

where ν denotes the kinematic viscosity of water. The dimensionless Reynolds number can be interpreted as the ratio between viscous and kinematic effects. The flow is considered laminar if $Re(\mathrm{Q}) \leq 2300$ and turbulent if $Re(\mathrm{Q}) > 4000$. Otherwise, we speak of laminar-turbulent transition. In the laminar case, we have

$$\lambda(\mathrm{Q}) = \frac{64}{Re}, \tag{2.4}$$

which results in $\mathtt{g}(\mathrm{Q})$ being a linear function in Q. For turbulent flow, the friction coefficient is given by the implicit formula of Colebrook and White

$$\frac{1}{\sqrt{\lambda(\mathrm{Q})}} = -2\log_{10}\left(\frac{2.51}{Re(\mathrm{Q})\sqrt{\lambda(\mathrm{Q})}} + \frac{k}{3.71d}\right), \tag{2.5}$$

with k being the pipe roughness. For practical computations, we utilize the *Swamee and Jain approximation*

$$\lambda = \frac{0.25}{\left(\ln\left[\frac{k}{3.7d} + \frac{5.74}{Re^{9/10}}\right]\right)^2} \tag{2.6}$$

for the solution of (2.5). Note that cubic interpolation was employed in [42] for the transition zone.

For simplicity, we assume that all relevant material properties, like e.g. the wall thickness of the pipe or the viscosity of water, are constant. More precisely, we set $\mathbf{a} = 1450\,m/s$, i.e. the speed of sound in water, $\nu = 1.31 \cdot 10^{-6}\,m^2/s$, i.e. the kinematic viscosity of water at 10°C and $\mathbf{g_r} = 9.81\ m/s^2$, i.e. the gravitational acceleration constant, in the following.

Equations (2.1) and (2.2) form the so-called *water hammer equations* (WHE). For further reference, see also [5, 55].

The friction coefficient

An important property of the friction coefficient is its monotone dependence on Q. Taking the derivative of λ in (2.4) with respect to Q, we have

$$\lambda'(\mathrm{Q}) = -\frac{16\pi d\nu\,\mathrm{sgn}(\mathrm{Q})}{\mathrm{Q}^2} \tag{2.7}$$

for the laminar case. It can easily be seen that the derivative with respect to Q is negative if Q is positive and vice versa. For the derivative of λ in (2.5) we have

$$\lambda'(\mathrm{Q}) = \frac{\lambda(\mathrm{Q})}{-\frac{1}{2}\mathrm{Q} - \frac{\mathrm{Q}^2 \ln(10)}{2.51\pi d\nu} \operatorname{sgn}(\mathrm{Q}) \left(\frac{2.51\pi d\nu}{4|\mathrm{Q}|\sqrt{\lambda(\mathrm{Q})}} + \frac{k}{3.71d} \right)} \tag{2.8}$$

for turbulent flow, which also results in a negative derivative. Therefore, we get the already mentioned monotone dependence on Q in the friction coefficient. For the Swamee and Jain approximation (2.6), we have

$$\lambda'(\mathrm{Q}) = -\frac{\frac{9\mathrm{Q}}{20|\mathrm{Q}|^{29/10}}}{\left(\frac{k}{3.7d} + \frac{5.74}{Re^{9/10}} \right)^2 \ln\left[\frac{k}{3.7d} + \frac{5.74}{Re^{9/10}} \right]} < 0. \tag{2.9}$$

2.2 Mathematical description of the water hammer equations

We move on to a more precise mathematical description of the water hammer equations. Let us assume that $-\infty < x_l < x_r < \infty$ and $\Omega = [x_l, x_r]$. The length of the pipe is $L = |\Omega| = x_r - x_l$. Further, let $T > 0$ be the time horizon.

We introduce the variable $\mathrm{H} = \mathbf{g_r}\mathcal{A}\mathrm{h}$. The water hammer equations are then given as

$$\begin{aligned} \partial_t \mathrm{H} + \mathbf{a}^2 \partial_x \mathrm{Q} &= 0 \\ \partial_t \mathrm{Q} + \partial_x \mathrm{H} &= -\mathsf{g}(\mathrm{Q}) \end{aligned} \qquad (x,t) \in \Omega \times [0,T], \tag{2.10}$$

where the source term g has to be chosen depending on the Reynolds number. The system is complemented with the initial conditions $\mathrm{H}(x,0) = \mathrm{H}_0$ and $\mathrm{Q}(x,0) = \mathrm{Q}_0$ to form a Cauchy problem. For simplicity, we mainly use periodic boundary conditions in space, i.e. $\mathrm{H}(x_l,\cdot) = \mathrm{H}(x_r,\cdot)$ and $\mathrm{Q}(x_l,\cdot) = \mathrm{Q}(x_r,\cdot)$. Of course, other boundary conditions can be treated as well.

The water hammer equations form a system of hyperbolic balance laws. In particular, a system of the form

$$\partial_t \mathbf{w} + \partial_x \mathsf{f}(\mathbf{w}) = \mathsf{G}(\mathbf{w}) \tag{2.11}$$

is called *hyperbolic* if the Jacobian of the flux function f has only real eigenvalues and linear independent eigenvectors. If it has no multiple eigenvalues, we speak of a *strictly hyperbolic* system. In general, hyperbolic equations do not admit *classical solutions*. This is why the concept of *weak solutions* is introduced. In many cases, the existence of weak solutions can be shown, although in most cases, they are not unique. To exclude nonphysical solutions, so-called *entropy conditions* are usually utilized. For further details on the theory of hyperbolic equations, we refer the reader to [24].

We introduce the variable $\mathbf{w} = (\mathrm{H}, \mathrm{Q})^\top$, the flux function $\mathsf{f}(\mathbf{w}) = (\mathbf{a}^2\mathrm{Q}, \mathrm{H})^\top$ and the vectorial source term $\mathsf{G} = (0, -\mathsf{g}(\mathrm{Q}))^\top$, where $\mathsf{g}(\mathrm{Q}) = \lambda(\mathrm{Q})\frac{\mathrm{Q}|\mathrm{Q}|}{2d\mathcal{A}}$. The water hammer equations then take the general form (2.11).

2.2.1 Eigenvalues of the flux function

We compute the Jacobian matrix of the flux function,

$$\nabla_{\mathbf{w}}\mathsf{f}(\mathbf{w}) = \begin{pmatrix} 0 & \mathbf{a}^2 \\ 1 & 0 \end{pmatrix}. \tag{2.12}$$

The eigenvalues of this Jacobian are important quantities in the theory of hyperbolic equations. We have $\nu_{1/2} = \pm\mathbf{a}$ and thus, we have a strictly hyperbolic system.

Further, we compute the Jacobian matrix of the vectorial source term

$$\nabla_{\mathbf{w}}\mathsf{G}(\mathbf{w}) = \begin{pmatrix} 0 & 0 \\ 0 & -\frac{|\mathrm{Q}|}{2d\mathcal{A}}(\lambda'(\mathrm{Q})\mathrm{Q} + 2\lambda(\mathrm{Q})) \end{pmatrix}. \tag{2.13}$$

Consequently, for the eigenvalues we have

$$\mu_1 = -\frac{|\mathrm{Q}|}{2d\mathcal{A}}(\lambda'(\mathrm{Q})\mathrm{Q} + 2\lambda(\mathrm{Q})) \qquad \text{and} \qquad \mu_2 = 0. \tag{2.14}$$

An important property of the source term $\mathsf{G}(\mathbf{w})$ is its dissipativity, i.e. $\mu_1, \mu_2 \leq 0$. Since μ_1 is a function of $\lambda'(\mathrm{Q})$, we distinguish two cases. Utilizing (2.4) and (2.7), we get the following for the laminar flow

$$\mu_1 = -\frac{|\mathrm{Q}|}{2d\mathcal{A}}\left(-\frac{16\pi d\nu\,\mathrm{sgn}(\mathrm{Q})}{\mathrm{Q}^2}\mathrm{Q} + 2\frac{64\pi d\nu}{4|\mathrm{Q}|}\right) = -\frac{8\pi\nu}{\mathcal{A}} < 0. \tag{2.15}$$

Using (2.5) and (2.8), we get the following for the turbulent flow

$$\mu_1 = -\frac{|\mathrm{Q}|}{2d\mathcal{A}}\left(\frac{\lambda(\mathrm{Q})\mathrm{Q}}{-\frac{1}{2}\mathrm{Q} - \frac{\mathrm{Q}^2\ln(10)}{2.51\pi d\nu}\,\mathrm{sgn}(\mathrm{Q})\left(\frac{2.51\pi d\nu}{4|\mathrm{Q}|\sqrt{\lambda(\mathrm{Q})}} + \frac{k}{3.71d}\right)} + 2\lambda(\mathrm{Q})\right). \tag{2.16}$$

Additionally using (2.6) and (2.9), we get

$$\mu_1 = -\frac{|\mathrm{Q}|}{2d\mathcal{A}\ln\left[\frac{k}{3.7d} + \frac{5.74}{Re^{9/10}}\right]}\left(\frac{0.5}{\ln\left[\frac{k}{3.7d} + \frac{5.74}{Re^{9/10}}\right]} - \frac{\frac{9\mathrm{Q}^2}{10|Q|^{29/10}}}{\left(\frac{k}{3.7d} + \frac{5.74}{Re^{9/10}}\right)^2}\right). \tag{2.17}$$

In Domschke[17] source terms of these forms were computed and analysed using real data. Independent of the sign of Q, they found $\mu_1 \leq 0$ in the turbulent case.

2.2.2 The stationary state

The stationary state $(\bar{\mathrm{H}}, \bar{\mathrm{Q}})$ of the water hammer equations, i.e. a solution of the WHE (2.10) which does not depend on t, is of special importance for applications.

If we assume that the time derivatives $\partial_t \mathrm{H}$ and $\partial_t \mathrm{Q}$ in (2.10) are zero, we get the system of stationary equations

$$\begin{aligned} \mathbf{a}^2 \partial_x \mathrm{Q} &= 0 \\ \partial_x \mathrm{H} &= -\lambda(\mathrm{Q}) \tfrac{\mathrm{Q}|\mathrm{Q}|}{2d\mathcal{A}}. \end{aligned} \tag{2.18}$$

It can easily be seen that the first component of the solution is given by the stationary flow rate

$$\mathrm{Q}(x) \equiv \bar{\mathrm{Q}}, \tag{2.19}$$

where $\bar{\mathrm{Q}}$ depends on the given boundary conditions. Inserting the stationary flow rate into the second equation of the system, we observe that H decreases linearly in the sense that

$$\bar{\mathrm{H}}(x_1) - \bar{\mathrm{H}}(x_0) = -\mathsf{g}(\bar{\mathrm{Q}})(x_1 - x_0), \tag{2.20}$$

for all $x_l \leq x_0 \leq x_1 \leq x_r$, for the solution of the pressure head. Altogether the stationary state of the WHE is determined by (2.20) and (2.19). Note that here we cannot use periodic boundary conditions because of the structure of the stationary state. In particular, this means that the stationary state is not periodic if $\mathrm{Q} \neq 0$. Therefore, we need to use so-called *inflow and outflow boundary conditions*. Considering the eigenvalues for the system, we can, for instance, specify a boundary condition for the pressure head at x_l and specify a boundary condition for the flow rate at x_r. This means

$$\mathrm{Q}(x_r) = \bar{\mathrm{Q}} \tag{2.21}$$

$$\mathrm{H}(x_l) = h_0, \tag{2.22}$$

where h_0 is a given function. For the other remaining boundaries we cannot specify a boundary condition analytically but we may have to specify a boundary condition for the numerical methods [51, Section 3.11]. We refer to this later on when we derive the well-balancedness property.

2.3 The linear system

We turn to the linear system to derive some further properties. Setting the source term $\mathsf{g} \equiv 0$, (2.10) transforms into

$$\partial_t \mathbf{w} + \mathbf{A} \partial_x \mathbf{w} = 0, \tag{2.23}$$

where the matrix $\mathbf{A}$ is defined as

$$\mathbf{A} = \begin{pmatrix} 0 & \mathbf{a}^2 \\ 1 & 0 \end{pmatrix}. \tag{2.24}$$

We introduce the diagonal matrix

$$\Lambda = \begin{pmatrix} \mathbf{a} & 0 \\ 0 & -\mathbf{a} \end{pmatrix} \tag{2.25}$$

containing the eigenvalues of $\mathbf{A}$. We have $\mathbf{A} = \mathbf{R}\Lambda\mathbf{R}^{-1}$, where

$$\mathbf{R} = \begin{pmatrix} \mathbf{a} & -\mathbf{a} \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{R}^{-1} = \begin{pmatrix} \frac{1}{2\mathbf{a}} & \frac{1}{2} \\ -\frac{1}{2\mathbf{a}} & \frac{1}{2} \end{pmatrix}. \tag{2.26}$$

The matrix $\mathbf{R}$ and $\mathbf{R}^{-1}$ contain the right and left eigenvectors of $\mathbf{A}$, respectively, i.e.

$$\mathbf{A}r_k = \nu_k r_k \quad \text{and} \quad l_k^\top \mathbf{A} = \nu_k l_k^\top. \tag{2.27}$$

This means that l_k is an eigenvector of $\mathbf{A}^\top$.

The eigenvectors r_k form a basis of $\mathbb{R}^2$, and after normalization we have

$$l_j^\top r_k = \delta_{jk}, \tag{2.28}$$

for $1 \leq j, k \leq 2$. δ_{jk} denotes the Kronecker symbol.

We introduce the characteristic variables v_1, v_2 through

$$v = (v_1,\, v_2)^\top = \mathbf{R}^{-1}(\mathrm{H}, \mathrm{Q})^\top = \begin{pmatrix} \frac{1}{2\mathbf{a}}\mathrm{H} + \frac{1}{2}\mathrm{Q} \\ -\frac{1}{2\mathbf{a}}\mathrm{H} + \frac{1}{2}\mathrm{Q} \end{pmatrix}. \tag{2.29}$$

Now we transform (2.23) into the characteristic system by multiplying it with $\mathbf{R}^{-1}$, resulting in

$$\partial_t v + \Lambda \partial_x v = 0. \tag{2.30}$$

Due to the diagonal structure of Λ, the characteristic system consists of two independent scalar equations

$$\begin{aligned} \partial_t v_1 + \mathbf{a}\partial_x v_1 &= 0 \\ \partial_t v_2 - \mathbf{a}\partial_x v_2 &= 0. \end{aligned} \tag{2.31}$$

2.3.1 Analytical solution

For the decoupled system (2.31) we can compute an analytical solution of the form

$$v_k(x,t) = l_k^\top \mathbf{w}_0(x - \nu_k t), \tag{2.32}$$

where $\mathbf{w}_0 = (\mathrm{H}_0, \mathrm{Q}_0)^\top$ is a given initial profile for (2.23). Transforming back to the original variables, we obtain

$$\begin{aligned} \mathrm{H}(x,t) &= \mathbf{a}(v_1 - v_2) \qquad (2.33)\\ &= \mathbf{a}(\tfrac{1}{2\mathbf{a}}\mathrm{H}_0(x-\nu_1 t) + \tfrac{1}{2}\mathrm{Q}_0(x-\nu_1 t) + \tfrac{1}{2\mathbf{a}}\mathrm{H}_0(x-\nu_2 t) - \tfrac{1}{2}\mathrm{Q}_0(x-\nu_2 t)) \\ \mathrm{Q}(x,t) &= (v_1 + v_2) \qquad (2.34)\\ &= \tfrac{1}{2\mathbf{a}}\mathrm{H}_0(x-\nu_1 t) + \tfrac{1}{2}\mathrm{Q}_0(x-\nu_1 t) - \tfrac{1}{2\mathbf{a}}\mathrm{H}_0(x-\nu_2 t) + \tfrac{1}{2}\mathrm{Q}_0(x-\nu_2 t). \end{aligned}$$

2.3.2 The source term

If we include the source term, (2.10) can be written as

$$\partial_t \mathbf{w} + \mathbf{A}\partial_x \mathbf{w} = \mathsf{G}(\mathbf{w}), \tag{2.35}$$

where the matrix $\mathbf{A}$ is defined in (2.24) and $\mathsf{G}(\mathbf{w}) = (0, -\mathsf{g}(\mathrm{Q}))^T$ with $\mathsf{g}(\mathrm{Q}) = \lambda(\mathrm{Q})\frac{\mathrm{Q}|\mathrm{Q}|}{2d\mathcal{A}}$. We can now also transform system (2.35) into the characteristic system by multiplying with $\mathbf{R}^{-1}$, which results in

$$\partial_t v + \Lambda \partial_x v = \tilde{\mathsf{G}}(v), \tag{2.36}$$

where Λ is defined in (2.25) and $\tilde{\mathsf{G}} = \mathsf{G}(\mathbf{R}v)$.

Due to the diagonal structure of Λ, the characteristic system now consists of two coupled scalar equations

$$\begin{aligned} \partial_t \tilde{\mathrm{H}} + \mathbf{a}\partial_x \tilde{\mathrm{H}} &= -1/2\mathsf{g}(\tilde{\mathrm{H}} + \tilde{\mathrm{Q}}) \\ \partial_t \tilde{\mathrm{Q}} - \mathbf{a}\partial_x \tilde{\mathrm{Q}} &= -1/2\mathsf{g}(\tilde{\mathrm{H}} + \tilde{\mathrm{Q}}). \end{aligned} \tag{2.37}$$

Note that the scalar equations are coupled by the source term, not by the flux functions.

3 Numerical Methods for Conservation Laws

In this section, we introduce numerical methods which are suited for solving one-dimensional hyperbolic systems of partial differential equations with a possibly stiff source term such as the water hammer equations (2.10). The numerical solution of partial differential equations is a complex task in general and has been investigated intensively in the last decades. To treat hyperbolic equations, specialized numerical methods are necessary. For an overview, we refer to [10, 11, 12, 51] and the references therein.

For reasons of clarity and comprehensibility, we discuss the numerical solution of the system

$$\partial_t \mathsf{u} + \partial_x \mathsf{f}(\mathsf{u}) = \mathsf{g}(\mathsf{u}), \tag{3.1}$$

supplemented with the initial condition $\mathsf{u}(\cdot, 0) = \mathsf{u}_0$. Further, we assume periodic boundary conditions in space to simplify the notation. Of course, the assumption of other boundary conditions is also possible. Note that all methods presented here can be extended to the system case that means u contains more than one component.

The temporal and spatial variables are usually treated separately, but also *space-time* methods exist (see e.g. [83] for discontinuous Galerkin methods), which treat both variables simultaneously. Here, we only treat the former, whereby we distinguish between *Rothe methods* and the *Method of Lines* (MOL). Even though both methods are generally applicable, we use the latter to derive the methods presented here.

The MOL approach enables us to treat the temporal and spatial discretization separately. We start by introducing suitable time discretization schemes of higher-order. Subsequently, we shortly discuss some of the most common spatial discretization techniques, namely the *Finite Volume method* and the *Discontinuous Galerkin method*. The methods we use for the balance laws are then generated by a suitable combination. Next, we shortly discuss the *implicit box scheme* [42, 43] which we use for comparison in our numerical tests. For higher-order spatial discretizations, we describe slope and flux limiters in the last section of this chapter.

3.1 Temporal discretization

We discuss time integration methods suitable for the solution of hyperbolic equations. We introduce a finite dimensional operator $\mathbf{F} : \mathbb{R}^l \to \mathbb{R}^l$, $l \in \mathbb{N}$. However, we keep in mind that $\mathbf{F}$ typically results from a spatial semi-discretization of the hyperbolic equation. Let $T > 0$ and $\mathsf{u} \in \mathbb{R}^l$ be the solution of

$$\partial_t \mathsf{u} = \mathbf{F}(\mathsf{u}), \quad \mathsf{u}(0) = \mathsf{u}_0 \text{ for all } t \in [0, T]. \tag{3.2}$$

For a simpler notation, we assume u to be the discrete pendant of a scalar quantity. An extension to the discretization of vectorial quantities as for the water hammer equation requires only minor changes. However, the definition of $\mathbf{F}$ has to be handled with care in this case.

Equation (3.2) is a system of ordinary differential equations. The existence of solutions is fairly well understood from a theoretical point of view, c.f. [13]. Therefore, we assume $\mathbf{F}$ to be locally Lipschitz continuous with Lipschitz constant L. We use equidistant time steps of size Δt and let $\mathsf{u}^n \approx \mathsf{u}(t^n)$ with $t^n = n\Delta t$ for $n = 0, ..., M$, where u is the analytical solution of the equation.

We use single step methods for the discretization, more precisely *Runge-Kutta* (RK) methods. Those methods have a long history and were originally developed by the Germans mathematicians C. Runge (1895) [67] and M. W. Kutta (1901) [48]. Further investigations were made by K. Heun (1900) and J. C. Butcher (1965) [8].

3.1.1 Singly-diagonally implicit Runge-Kutta methods

In the context of hyperbolic equations, some Runge-Kutta methods proved to be more convenient than others. We distinguish several types of methods. The presence of a stiff source term suggests using an implicit method. In contrary to explicit ones, they exhibit superior stability properties. Concerning implicit methods, we distinguish between *fully implicit* RK methods (IRK), *singly implicit* RK methods (SIRK), *diagonally implicit* RK methods (DIRK) and *singly-diagonally implicit* RK methods (SDIRK) [15, Chap. 3]. The difference becomes clear immediately if we introduce the *Butcher tableau* [8, Chap. 2, Section 23]

$$\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}, \tag{3.3}$$

where $A \in \mathbb{R}^{m\times m}$, $b, c \in \mathbb{R}^m$ and $m \in \mathbb{N}$ denotes the number of stages. The non-zero entries of A determine the type of RK method at hand. A DIRK method would for instance have the form

$$A = \begin{pmatrix} * & & 0 \\ \vdots & \ddots & \\ * & \dots & * \end{pmatrix}.$$

If all entries on the diagonal are equal, we then call it an SDIRK method. If they are zero, this results in an explicit method.

The introduced Butcher tableau is a short notation for the so-called *Butcher form* of the RK method

$$\begin{aligned} \mathsf{u}^{(i)} &= \mathsf{u}^n + \Delta t \sum_{j=1}^{m} \mathbf{a}_{ij} \mathbf{F}(\mathsf{u}^{(j)}), \quad 1 \le i \le m, \\ \mathsf{u}^{n+1} &= \mathsf{u}^n + \Delta t \sum_{j=1}^{m} b_j \mathbf{F}(\mathsf{u}^{(j)}), \end{aligned} \tag{3.4}$$

which computes a new iterate u^{n+1} for the solution of (3.2) from given u^n. The terms $\mathsf{u}^{(i)}$ are called intermediate stages and Δt denotes the step size. For an SDIRK method written in this form, we have $a_{ij} = 0$ for $j > i$, and $a_{ii} = \gamma \neq 0$ for $i = \{1, \dots, m\}$.

The special form of DIRK methods allows for a sub-sequential computation of the intermediate stages. This reduces the computation time compared to full IRK methods. Regarding SDIRK methods, we can achieve an additional reduction of computation costs if we use the approximation of the Jacobian matrix only once computed by the simplified Newton's method.

3.1.2 Shu-Osher formulation

For the convenient treatment of stability properties, we introduce the *Shu-Osher* formulation. Any Runge-Kutta method in Butcher form (3.4) can be transformed to this formulation, which was developed by Shu and Osher [71]. It was initially developed for explicit Runge-Kutta methods firstly and they used it to prove that explicit methods can be rewritten as a convex combination of explicit Euler steps. For implicit Runge-Kutta methods, a generalization of the Shu-Osher formulation exists, i.e. the *modified Shu-Osher form* [21, 35]

$$\mathsf{u}^{(i)} = v_i \mathsf{u}^n + \sum_{j=1}^{m} \alpha_{ij} \left(\mathsf{u}^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} \mathbf{F}(\mathsf{u}^{(j)}) \right), \quad 1 \le i \le m \tag{3.5a}$$

$$\mathsf{u}^{n+1} = \mathsf{u}^{(m+1)} = v_{m+1} \mathsf{u}^n + \sum_{j=1}^{m} \alpha_{m+1,j} \left(\mathsf{u}^{(j)} + \Delta t \frac{\beta_{m+1,j}}{\alpha_{m+1,j}} \mathbf{F}(\mathsf{u}^{(j)}) \right) \tag{3.5b}$$

with coefficients $v \in \mathbb{R}^{m+1}, \alpha, \beta \in \mathbb{R}^{(m+1)\times m}$. Analogous to the standard formulation, the consistency requirement

$$v_i + \sum_{j=1}^{m} \alpha_{ij} = 1, \quad 1 \le i \le m+1 \tag{3.6}$$

ensures that the initial value problem (IVP)

$$\begin{aligned} \partial_t \mathsf{u}(t) &= 0 \\ \text{with} \quad \mathsf{u}(t_0) &= \mathsf{u}_0 \end{aligned} \tag{3.7}$$

is solved exactly. If $v_i, \alpha_{ij}, \beta_{ij} \ge 0$ for all i, j, each stage $\mathsf{u}^{(i)}$ is formally a convex combination of explicit Euler steps. Note that for (S)DIRK methods, the summation in (3.5a) is replaced by $\sum_{j=1}^{i}$.

For the next step, we need to exclude so-called defective methods. This means that the Runge-Kutta methods have to be zero-well defined [26, Chap. 3, Section 3.1] or equivalently irreducible [21].

Definition 1. *We call a RK method zero-well defined if the stage equations have a unique solution and the solution of the IVP* (3.7) *is computed exactly.*

Any irreducible RK method of form (3.4) can be uniquely represented by its Butcher coefficients. Even though any RK method in the modified Shu-Osher form can be transformed back into (3.4), there are multiple sets of coefficients $(\alpha)_{ij}$, $(\beta)_{ij}$ and $(v)_i$ which represent the same RK method. To overcome this problem, another formulation, the

so-called *canonical Shu-Osher form* was developed in [26, Chap. 3, Section 3.3]. Before introducing this form, we establish a compact notation.

Vector notation.

For many computations, it is useful to write Runge-Kutta methods in a more compact form. Let us introduce the extended matrices

$$\begin{aligned}(\tilde{\alpha})_{i,j} &= \alpha_{i,j} \quad \text{for} \quad j \le m \quad \text{and} \quad (\tilde{\alpha})_{i,m+1} = 0,\\ (\tilde{\beta})_{i,j} &= \beta_{i,j} \quad \text{for} \quad j \le m \quad \text{and} \quad (\tilde{\beta})_{i,m+1} = 0,\end{aligned}$$

where we keep the notation α and β if no confusion is possible.

Let us define the vectors

$$\begin{aligned}\mathsf{y}_i &= \mathsf{u}^{(i)} \in \mathbb{R}^l\\ \mathsf{f}_i &= F(\mathsf{u}^{(i)}) \in \mathbb{R}^l, \quad i = 1, ..., m,\end{aligned}$$

and $\mathsf{f}_{m+1} = 0$ and l is the number of the spatial steps. We have $\mathsf{y} = (\mathsf{y}_i)_{i=1,...,m+1} \in \mathbb{R}^{l(m+1)}$ and $\mathsf{f} = (\mathsf{f}_i)_{i=1,...,m+1} \in \mathbb{R}^{l(m+1)}$.

Further, let us introduce the Kronecker product

$$C \otimes D = \begin{pmatrix} c_{11}D & \dots & c_{1n}D \\ \vdots & \ddots & \vdots \\ c_{m1}D & \dots & c_{mn}D \end{pmatrix}$$

for $C \in \mathbb{R}^{m\times n}$ and $D \in \mathbb{R}^{p\times q}$. We have $C \otimes D \in \mathbb{R}^{mp\times nq}$. With these definitions, the modified Shu-Osher form (3.5a)-(3.5b) can be written as

$$\begin{aligned}\mathsf{y} &= (v \otimes \mathbf{I}_l)\mathsf{u}^n + (\alpha \otimes \mathbf{I}_l)\mathsf{y} + \Delta t(\beta \otimes \mathbf{I}_l)\mathsf{f}\\ \mathsf{u}^{n+1} &= \mathsf{y}_{m+1},\end{aligned} \tag{3.8}$$

where $\mathbf{I}_l \in \mathbb{R}^{l\times l}$ denotes the identity matrix and $\alpha, \beta \in \mathbb{R}^{(m+1)\times(m+1)}$ as defined above.

We illustrate this notation with a short example.

Example 1. *If we consider a two-stage fully implicit RK method in modified Shu-Osher form, we have e.g.* $v = (v_1, v_2, v_3)^\top$,

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & \alpha_{32} & 0 \end{pmatrix} \qquad \textit{and} \qquad \beta = \begin{pmatrix} \beta_{11} & \beta_{12} & 0 \\ \beta_{21} & \beta_{22} & 0 \\ \beta_{31} & \beta_{32} & 0 \end{pmatrix}.$$

This results in

$$(v \otimes \mathbf{I}_l)\mathsf{u}^n = \begin{pmatrix} v_1\mathsf{u}^n \\ v_2\mathsf{u}^n \\ v_3\mathsf{u}^n \end{pmatrix}, \quad (\alpha \otimes \mathbf{I}_l)\mathsf{y} = \begin{pmatrix} \alpha_{11}\mathbf{I}_l & \alpha_{12}\mathbf{I}_l \\ \alpha_{21}\mathbf{I}_l & \alpha_{22}\mathbf{I}_l \\ \alpha_{31}\mathbf{I}_l & \alpha_{32}\mathbf{I}_l \end{pmatrix} \mathsf{y} = \begin{pmatrix} \alpha_{11}\mathsf{y}_1 & \alpha_{12}\mathsf{y}_2 \\ \alpha_{21}\mathsf{y}_1 & \alpha_{22}\mathsf{y}_2 \\ \alpha_{31}\mathsf{y}_1 & \alpha_{32}\mathsf{y}_2 \end{pmatrix}$$

and

$$\Delta t(\beta \otimes \mathbf{I}_l)\mathsf{f} = \Delta t \begin{pmatrix} \beta_{11}\mathbf{I}_l & \beta_{12}\mathbf{I}_l \\ \beta_{21}\mathbf{I}_l & \beta_{22}\mathbf{I}_l \\ \beta_{31}\mathbf{I}_l & \beta_{32}\mathbf{I}_l \end{pmatrix} \mathsf{f} = \Delta t \begin{pmatrix} \beta_{11}\mathsf{f}_1 & \beta_{12}\mathsf{f}_2 \\ \beta_{21}\mathsf{f}_1 & \beta_{22}\mathsf{f}_2 \\ \beta_{31}\mathsf{f}_1 & \beta_{32}\mathsf{f}_2 \end{pmatrix}.$$

Formulation (3.8) *is therefore a compact way to formulate RK methods.*

The canonical Shu-Osher form.

Any irreducible RK method of the form (3.4) is uniquely determined by its Butcher coefficients. Utilizing the compact notation, (3.4) can be written as

$$\mathsf{y} = (e \otimes \mathbf{I}_l)\mathsf{u}^n + \Delta t \left(\begin{pmatrix} A & 0 \\ b^\top & 0 \end{pmatrix} \otimes \mathbf{I}_l \right) \mathsf{f}, \tag{3.9}$$

where $e = (1, 1, ..., 1)^\top \in \mathbb{R}^{m+1}$. Using this and solving (3.8) for y, we have

$$\begin{aligned} \mathsf{y} &= (\mathbf{I}_{l(m+1)} - (\alpha \otimes \mathbf{I}_l))^{-1}(v \otimes \mathbf{I}_l)\mathsf{u}^n + (\mathbf{I}_{l(m+1)} - (\alpha \otimes \mathbf{I}_l))^{-1}(\Delta t\beta \otimes \mathbf{I}_l)\mathsf{f} \\ &= (e \otimes \mathbf{I}_l)\mathsf{u}^n + \Delta t(\mathbf{I}_{l(m+1)} - (\alpha \otimes \mathbf{I}))^{-1}(\beta \otimes \mathbf{I}_l)\mathsf{f}, \end{aligned} \tag{3.10}$$

where we used $(\mathbf{I}_{l(m+1)} - (\alpha \otimes \mathbf{I}_l))^{-1}(v \otimes \mathbf{I}_l) = (e \otimes \mathbf{I}_l)$, implied by the consistency requirement (3.6). If we choose $\alpha = 0$ and $\beta = \beta_0$ for some suitable β_0, what results from (3.10) is

$$\mathsf{y} = (e \otimes \mathbf{I}_l)\mathsf{u}^n + \Delta t(\beta_0 \otimes \mathbf{I}_l)\mathsf{f} \tag{3.11}$$

and with (3.9) it implies

$$\beta_0 = \begin{pmatrix} A & 0 \\ b^\top & 0 \end{pmatrix}. \tag{3.12}$$

A comparison of equations (3.11) and (3.10) returns

$$(\beta_0 \otimes \mathbf{I}_l) = (\mathbf{I}_{l(m+1)} - (\alpha \otimes \mathbf{I}_l))^{-1}(\beta \otimes \mathbf{I}_l). \tag{3.13}$$

We use this connection to infer a smart choice for α and β in (3.10). Remember that both matrices chosen in different ways result in the same RK method. This is done in a way that $\beta_{ij} \neq 0$ and the ratio $r = \alpha_{ij}/\beta_{ij}$ is constant for every i, j. We denote the matrices meeting this requirement by α_r and β_r. Note that due to (3.5a)-(3.5b) this choice implies that the resulting method is a convex combination of explicit Euler steps with constant step size r. Using $\alpha_r = r\beta_r$ in (3.13) and assuming that the method is zero-well-defined, we can solve for β_r if $\mathbf{I}_{m+1} + r\beta_0$ is invertible. The thereby determined coefficients are then given by

$$\beta_r = \beta_0(\mathbf{I}_{m+1} + r\beta_0)^{-1} \tag{3.14}$$

$$\alpha_r = r\beta_r = r\beta_0(\mathbf{I}_{m+1} + r\beta_0)^{-1} \tag{3.15}$$

$$v_r = (\mathbf{I}_{m+1} - \alpha_r)e = (\mathbf{I}_{m+1} + r\beta_0)^{-1}e. \tag{3.16}$$

The *canonical Shu-Osher form* is then given as

$$\mathsf{y} = (v_r \otimes \mathbf{I}_l)\mathsf{u}^n + (\alpha_r \otimes \mathbf{I}_l)\left(\mathsf{y} + \frac{\Delta t}{r}\mathsf{f}\right). \tag{3.17}$$

Note that (3.9) together with (3.12) corresponds to (3.17) with $r = 0$ in (3.14)-(3.17). The ratio r is related to the so-called SSP coefficient, which we introduce in the next section.

3.1.3 SSP property

When designing numerical methods for hyperbolic equations, special stability aspects are important. A typical feature of hyperbolic equations is the formation of shocks even for smooth initial profiles. This requires a careful design of numerical methods, since oscillations can occur if standard approaches are used. We want to design methods which will guarantee nonlinear stability in suitable norms, for instance the maximum norm or the bounded/total variation semi-norm. Some other convex functionals representing various non-oscillatory properties can also be used. This leads us to the so-called SSP (**S**trong **S**tability **P**reserving) methods.

The theory of SSP was mainly developed by two groups. Prominent representatives of the first group are the mathematicians Shu, Osher, Gottlieb and Tadmor, who are experts in the field of hyperbolic partial differential equations. In [68, 71], Shu and Osher developed SSP time discretizations which they called *total variation diminishing*. In [27], Gottlieb and Shu studied SSP RK methods with respect to their optimality. In Shu et al. [28], the term *SSP* was used for the first time. Note that the expressions *SSP* and *TVD* are used simultaneously in different publications. Another group, more focused on the theory of ordinary differential equations, made investigations of positivity in Bolley et al. [4] and of contractivity and monotonicity in Spijker [75], all concerning linear ODE systems. They showed that these properties cannot hold for unconditional RK methods of an order higher than one. It was shown that the SSP property has a connection to the radius of monotonicity for special methods. The main developments towards the absolute monotonicity of RK methods and its connection to contractivity for nonlinear equations was made in Kraaijevanger [46]. Finally, the equivalence of both theories, that is the SSP theory and the theory of absolute monotonicity, was detected in [20, 21, 34, 35]. The SSP property is related to the time integration, which is the reason why we use the MOL approach at this point.

Note that for suitable spatial discretizations (i.e. suitable right-hand side $\mathbf{F}$ in (3.2)), the explicit Euler method generates a stable scheme under various norms provided the step size Δt is small enough. Using implicit methods, the region of stable step sizes can be enlarged. This is well-known concerning for instance the implicit Euler method. We want to employ a higher-order time integration, e.g. in the form of SDIRK methods, which maintain the SSP property of the explicit Euler method for larger step sizes. Note that the SSP property and theory is well-known for higher-order explicit RK methods [26].

Applying the explicit Euler method to (3.2), we have

$$\mathsf{u}^{n+1} = \mathsf{u}^n + \Delta t \mathbf{F}(\mathsf{u}^n). \tag{3.18}$$

We say that $\mathbf{F}$ satisfies the *explicit Euler condition* if there exists a time step Δt_{EE} such that

$$\|\mathsf{u}^n + \Delta t \mathbf{F}(\mathsf{u}^n))\| \leq \|\mathsf{u}^n\|, \tag{3.19}$$

for all $\Delta t \leq \Delta t_{EE}$. Here, $\|\cdot\|$ can for example denote the supremum norm $\|\xi\|_\infty = \sup_i |\xi_i|$, the total variation semi-norm $\|\xi\|_{TV} = \sum_i |\xi_{i+1} - \xi_i|$, or other suitable convex functionals satisfying $\|\lambda v + (1-\lambda)w\| \leq \lambda\|v\| + (1-\lambda)\|w\|$ for $0 \leq \lambda \leq 1$.

Utilizing the modified Shu-Osher form, we introduce the *SSP-coefficient*

$$C(\alpha, \beta) = \begin{cases} \min_{i,j} \frac{\alpha_{ij}}{\beta_{ij}}, & \alpha_{ij}, \beta_{ij}, v_i \text{ are non-negative,} \\ 0, & \text{otherwise} \end{cases}. \tag{3.20}$$

We consider $C(\alpha, \beta)$ to be infinite if $\beta_{ij} = 0$. The following theorem shows the SSP property of zero-well-defined implicit RK methods provided the explicit Euler condition is fulfilled.

Theorem 1 (Gottlieb et al. [26])**.** *Suppose that* $\mathbf{F}$ *satisfies the explicit Euler condition* (3.19)*, and let* u^n *denote the solution at step* n *given by applying a zero-well-defined Runge-Kutta method* (3.5a)-(3.5b) *to the initial value problem* (3.2)*. Then* u^n *satisfies the strong stability bound*

$$\|\mathsf{u}^{n+1}\| \leq \|\mathsf{u}^n\|,$$

provided that the time step satisfies

$$0 \leq \Delta t \leq C(\alpha, \beta)\Delta t_{EE}, \tag{3.21}$$

where $C(\alpha, \beta)$ *is the SSP coefficient.*

Proof. A similar proof can also be found in [26, Chap. 3, Section 3.1]. The idea of this proof is to show that the stages fulfill the SSP property. This implies that the approximate solution u^{n+1} fulfills the SSP property. If $C(\alpha, \beta) = 0$, then $\Delta t = 0$, giving us $u^{n+1} = u^n$ and the statement is trivial. Suppose $C(\alpha, \beta) > 0$. Taking the norm on both sides of (3.5a)-(3.5b) and using the consistency condition (3.6) as well as convexity of the norm and the explicit Euler condition (3.19), we obtain the bound

$$\begin{aligned} \|u^{(i)}\| &= \left\| \left(1 - \sum_{j=1}^m \alpha_{ij}\right) u^n + \sum_{j=1}^m \alpha_{ij} \left(u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} F(u^{(j)})\right) \right\| \\ &\leq \left(1 - \sum_{j=1}^m \alpha_{ij}\right) \|u^n\| + \sum_{j=1}^m \alpha_{ij} \left\| u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} F(u^{(j)}) \right\| \\ &\leq \left(1 - \sum_{j=1}^m \alpha_{ij}\right) \|u^n\| + \sum_{j=1}^m \alpha_{ij} \|u^{(j)}\|. \end{aligned} \tag{3.22}$$

Now let q be the index of the Runge-Kutta stage with largest norm, i.e., choose $q \in 1, 2, ..., m+1$ such that $\|u^{(i)}\| \leq \|u^{(q)}\|$ for all $1 \leq i \leq m+1$. Then taking $i = q$ in (3.22) yields

$$\|u^{(q)}\| \leq \left(1 - \sum_{j=1}^{m} \alpha_{qj}\right) \|u^n\| + \sum_{j=1}^{m} \alpha_{qj} \|u^{(j)}\| \leq \left(1 - \sum_{j=1}^{m} \alpha_{qj}\right) \|u^n\| + \sum_{j=1}^{m} \alpha_{qj} \|u^{(q)}\|.$$

If we assume $1 - \sum_{j=1}^{m} \alpha_{qj} > 0$, we can solve the above equation for $\|u^{(q)}\|$, and get

$$\|u^{(q)}\| \leq \|u^n\|,$$

which implies $\|u^{(i)}\| \leq \|u^n\|$ for all $i = 1, .., m+1$. If we suppose $1 - \sum_{j=1}^{m} \alpha_{qj} = 0$, we obtain the following from (3.22)

$$\|u^{(q)}\| \leq \sum_{j=1}^{m} \alpha_{qj} \|u^{(j)}\|.$$

Since q is chosen such that $\|u^{(i)}\| \leq \|u^{(q)}\|$ for all $1 \leq i \leq m+1$, this implies that $\|u^{(j)}\| = \|u^{(q)}\|$ for every j such that $\alpha_{qj} \neq 0$. Let $J = \{j : \alpha_{qj} \neq 0\}$. If there exists any $j^* \in J$ such that $1 - \sum_{j=1}^{m} \alpha_{qj} \neq 0$, we can take $q = j^*$ and apply the argument above. If not, then it follows that the stages with indices in J depend only on each other, and not on u^n. In this case, the method is not zero-well-defined and the assumption is violated. □

The SSP coefficient $C(\alpha_r, \beta_r)$ is related to the radius of absolute monotonicity of the RK method in [46]. Using the canonical Shu-Osher form, we introduce the ratio r as long as (3.14)-(3.16) holds. With the following theorem of Gottlieb et. al [26, Chap. 3, Theorem 3.2], we can establish a connection between r and the SSP coefficient.

Theorem 2 (Gottlieb at el. [26])**.** *Consider a Runge-Kutta method with Butcher coefficient array β_0. For the SSP coefficient of the method, we have*

$$C = \max\{r \geq 0| \quad (\mathbf{I}_{m+1} + r\beta_0)^{-1} \text{ exists}, \alpha_r \geq 0 \text{ and } v_r \geq 0\}, \tag{3.23}$$

where the inequalities have to be understood component-wise.

Proof. See [26, Chap. 3, Section 3.3]. □

3.1.4 Optimal SDIRK methods

In this section, we discuss SDIRK methods which satisfy the SSP property with an optimal SSP coefficient (3.20). An m-stage method of order p is called optimal if it has the greatest possible SSP coefficient C defined by (3.20). Such methods have been investigated by Ferracina et al. [22] and Kocsis et al. [41].

First, we observe that for general problems, we need to require $A \geq 0$ and $b > 0$ component-wise for the Butcher coefficients (3.4) to guarantee that the SSP coefficient is greater than zero [26]. In general the order of an m-stage SDIRK method cannot exceed

$m+1$ [59]. The positivity requirement of the Butcher coefficient, however, introduces additional order barriers. Note that Ketcheson et al. [40] showed that any SDIRK method with positive SSP coefficient has order $p \leq 4$. For general irreducible implicit SSP RK methods, this bound can be relaxed to $p \leq 6$. It was further shown by Gottlieb et. al [26, Chap. 5, Section 5.1.3] that SDIRK methods which satisfy the SSP property suffer from the same order barriers as explicit methods.

Examples for optimal SSP SDIRK methods are the implicit Euler method ($m = 1, p = 1$) and the implicit midpoint rule ($m = 1, p = 2$). While the implicit Euler method has an infinite SSP coefficient, we have $C = 2$ for the implicit midpoint rule. Numerical tests in [22] suggest that for $p = 2$ and $m \geq 3$, optimal SDIRK methods have SSP coefficient $C = 2m$, and we obtain for the Butcher coefficients

$$A_{ij} = \begin{cases} \frac{1}{2m}, & i = j \\ \frac{1}{m}, & j < i \\ 0, & \text{otherwise} \end{cases}$$

and $b_i = \frac{1}{m}$. In the case $p = 3$ and $m \geq 3$, the SSP coefficient is $C = m - 1 + \sqrt{m^2 - 1}$ and

$$A_{ij} = \begin{cases} \frac{1}{2}(1 - \sqrt{\frac{m-1}{m+1}}), & i = j \\ \frac{1}{\sqrt{m^2-1}}, & j < i, \\ 0, & \text{otherwise} \end{cases}$$

and $b_i = \frac{1}{m}$. For $p = 4$ and $m = 3$, an analytical representation is available [22].

In [41], the authors derived the upper bound $2m$ for the optimal SSP coefficient of second-order DIRK methods. Within this class, the SDIRK methods presented above are optimal.

Example 2 (SDIRK(2,2) in modified Shu-Osher form)**.** *We discuss an example of an optimal second-order SSP SDIRK method. The SDIRK(2,2) method with the Butcher tableau*

$$\begin{array}{c|cc} 1/4 & 1/4 & 0 \\ 3/4 & 1/2 & 1/4 \\ \hline & 1/2 & 1/2 \end{array} \tag{3.24}$$

can be written in the compact form (3.8) *with* $v = (\frac{1}{2}, 0, 0)^\top$ *and*

$$\alpha = \begin{pmatrix} 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} 1/8 & 0 & 0 \\ 1/8 & 1/8 & 0 \\ 0 & 1/4 & 0 \end{pmatrix}. \tag{3.25}$$

This SSP SDIRK method is optimal with the SSP coefficient $C = 2m = 4$ *[22]. In fact, the method is already in canonical Shu-Osher form and we have* $r = 4$*. It can be proven that the SSP coefficient fulfills Theorem 2.*

3.2 Spatial discretization

In the following section, we introduce suitable spatial semi-discretizations of the WHE (2.10). More precisely, we discretize the term

$$\mathbf{F}(\mathbf{w}) = -\partial_x \mathsf{f}(\mathbf{w}) + \mathsf{G}(\mathbf{w}).$$

Note that the result will serve as right-hand side $\mathbf{F}(\mathsf{u})$ in (3.2).

Let us comment on the most common discretization techniques. The most simple ansatz is a finite difference approximation [29, Chap. 2]. However, those methods are not conservative in general, which is a crucial criterion when dealing with hyperbolic equations. Another common technique for the discretization of differential equations is the finite element method [6]. Those methods usually approximate a solution with continuous functions, which is not appropriate for hyperbolic equations since shocks can occur.

A discontinuous version of the finite element method, i.e. the discontinuous Galerkin method [10, 61], will be discussed in Section 3.2.2. They are more flexible and can thus handle shocks. Also, conservative methods can be formulated. One of the most common methods in the context of hyperbolic equations are finite volume methods [51]. Those methods are specially designed for conservation laws and will be discussed in Section 3.2.1.

For all considered methods, we define a mesh on $\Omega = [x_l, x_r]$ first, where we use a uniform mesh size for ease of presentation. Let $\Omega_h = \cup_{\mathbf{I}_j}$ with $\mathbf{I}_j = (x_{j-1/2}, x_{j+1/2})$ and $\Delta x = |x_{j+1/2} - x_{j-1/2}|$ for all $j \in 1, 2, ..., N$. Both components H and Q are discretized on the same mesh and we imply periodic boundary conditions for simplicity, i.e. $\mathrm{H}(x_{1/2}) = \mathrm{H}(x_l) = \mathrm{H}(x_{N+1/2}) = \mathrm{H}(x_r)$ and $\mathrm{Q}(x_{1/2}) = \mathrm{Q}(x_l) = \mathrm{Q}(x_{N+1/2}) = \mathrm{Q}(x_r)$. Of course, other boundary conditions can also be used, that is inflow and outflow boundary conditions in order to represent the stationary state of the WHE.

For notational reasons, we discuss the spatial discretization for a scalar variable u as in Chapter 3.1. All concepts can be generalized to vectorial unknowns and can therefore be transferred to the WHE (2.10).

3.2.1 Finite Volume discretization

In some sense, the finite volume method is closely related to a finite difference method. However, its formulation is based on the integral form of the differential equation, resulting in a scheme which is well-suited for hyperbolic equations.

Let us introduce the main idea. The domain Ω is divided into the so-called *control cells* $\mathbf{I}_j$ and conservation is assumed on every such cell. We introduce the average cell value

$$\bar{\mathsf{u}}_j(t) = \frac{1}{\Delta x} \int_{\mathbf{I}_j} \mathsf{u}(x,t) dx. \tag{3.26}$$

Integration of (3.1) over each cell and using the Gauss divergence theorem yields

$$\partial_t \bar{\mathsf{u}}_j(t) = -\frac{1}{\Delta x}(\mathsf{f}(\mathsf{u}_{j+1/2}) - \mathsf{f}(\mathsf{u}_{j-1/2})) + \frac{1}{\Delta x}\int_{\mathbf{I}_j} \mathsf{G}(\mathsf{u}(x,t))dx, \tag{3.27}$$

where we denote $\mathsf{u}_{j\pm1/2} := \mathsf{u}(x_{j\pm1/2}, t)$. A finite volume discretization usually refers to a temporal integration of (3.27). In most cases explicit time integration is used, e.g. the centered method, the Upwind method, the Lax-Friedrichs method, or the Lax-Wendroff method [51]. Since we use the MOL approach and the temporal discretization was already introduced in Section 3.1, we merely discuss the spatial semi-discretization here.

We denote

$$\mathbf{F}_{j\pm1/2} := \mathsf{f}(\mathsf{u}_{j\pm1/2}) \quad \text{and} \quad \mathsf{G}_j := \frac{1}{\Delta x}\int_{\mathbf{I}_j} \mathsf{G}(\mathsf{u}(x,t))dx. \tag{3.28}$$

The appropriate approximation of the fluxes $\mathbf{F}_{j\pm1/2}$ in terms of the average values $\bar{\mathsf{u}}_j$ is the main difficulty in the design of finite volume methods.

Remark 1. *If we assume* $\mathsf{G}(\mathsf{u}(x,t)) = 0$ *in* (3.27)*, the time derivative of the cell average is balanced by the fluxes through the boundary of the cell* $\mathbf{I}_j$*. This form is called conservation form and every scheme which can be written in this form is called conservative. This important property is crucial when developing numerical methods for conservation laws and is necessary to guarantee the physical correctness of the scheme.*

From equations (3.26) and (3.28), we obtain the semi-discrete form

$$\partial_t \bar{\mathsf{u}}_j = -\frac{1}{\Delta x}\left(\mathbf{F}_{j+1/2} - \mathbf{F}_{j-1/2}\right) + \mathsf{G}_j. \tag{3.29}$$

The choice of suitable flux functions approximating the fluxes $\mathbf{F}_{j-1/2}$ is discussed in Section 3.2.3.

3.2.2 Discontinuous Galerkin discretization

The first discontinuous Galerkin (DG) method was introduced by Reed and Hill in 1973 in order to deal with first-order PDEs describing the steady transport of neutrons [62]. An extension to time-dependent hyperbolic problems was made by Chavent and Cockburn [9]. Extensions with respect to the numerical analysis of the method dealing with hyperbolic problems can also be found in the literature [10, 12, 61, 70].

As already mentioned, the DG method can be interpreted as finite element method allowing discontinuities. The DG method is based on the weak formulation of a partial differential equation. We define the space

$$V_{per}^l = \{v \in (L^\infty(\Omega))^l : \text{ periodic boundary conditions on } \partial\Omega\},$$

where the value l describes the number of components for the considered system of equations. A weak formulation of (3.1) is to find $\mathsf{u}(\cdot\,,t) \in V_{per}^l$, such that

$$\int_\Omega \partial_t \mathsf{u} v dx - \int_\Omega \mathsf{f}(\mathsf{u}) \partial_x v dx + \int_{\partial\Omega} \mathsf{f}(\mathsf{u}) v dx - \int_\Omega \mathsf{G}(\mathsf{u}) v dx = 0 \tag{3.30}$$

for all $v \in V_{per}^l$ and $t \in [0,T]$, where we multiply by a test function v, integrate over the domain Ω and use integration by parts.

To derive a semi-discrete formulation, we introduce the finite dimensional space $V_h^{per} = \{v \in V_h^k : \text{ periodic boundary conditions on } \partial\Omega\} \subset V_{per}^l$, where

$$V_h^k = \{v \in L^2(\Omega) \mid v|_{\mathbf{I}_j} \in P^k(\mathbf{I}_j), j = 1, ..., N\}$$

is a broken polynomial space. The space $P^k(\mathbf{I}_j)$ contains the polynomials of degree at most k on $\mathbf{I}_j$. Note that piecewise polynomials lie in L^∞. As an approximation, we search for $\mathsf{u}_h \in V_h^{per}$, such that

$$\begin{aligned} \int_{\mathbf{I}_j} \partial_t \mathsf{u}_h v_h dx - \int_{\mathbf{I}_j} \mathsf{f}(\mathsf{u}_h) \partial_x v_h dx = & \mathsf{f}(\mathsf{u}_{h,j-1/2}^+) v_{h,j-1/2}^+ \\ & - \mathsf{f}(\mathsf{u}_{h,j+1/2}^-) v_{h,j+1/2}^- + \int_{\mathbf{I}_j} \mathsf{G}(\mathsf{u}_h) v_h dx \end{aligned} \tag{3.31}$$

for all $j = 1, \ldots, N$, $v_h \in V_h^{per}$ and $t \in [0,T]$. Here, we use periodic boundary conditions to simplify the notation. Other boundary conditions could be incorporated as well.

Note that u_h and v_h are generally discontinuous at the element boundaries. By $\mathsf{u}_{h,j\pm1/2}^\pm$ we denote the left or the right limit on the point $x_{j\pm1/2}$, respectively.

The boundary terms $\mathsf{f}(\mathsf{u}_{h,j\pm1/2}^\pm) v_{h,j\pm1/2}^\pm$ need to be approximated in a suitable manner. Here, we can utilize similar ideas as in the finite volume case. We discuss the so-called *numerical fluxes* in Section 3.2.3.

Formulation as system of linear equations.

In contrast to the finite volume method, the explicit unknowns are not obvious in the discontinuous Galerkin case. Utilizing polynomial ansatz functions, we derive an explicit form of the finite dimensional approximation u_h. Note that in the case of the water hammer equation, the following has to be understood component-wise.

Let $\varphi_i^n(x)$ be a basis of V_h^{per} for $i = 1, \ldots, D$ and $n = 1, \ldots, N$. We use the expansion

$$\mathsf{u}_h = \sum_{n=1}^{N} \sum_{i=1}^{D} \mathsf{u}_i^n(t) \varphi_i^n(x).$$

Here, D denotes the degrees of freedom on each mesh element $\mathbf{I}_j$, whereas N again denotes the number of mesh elements.

Plugging this expansion into the integral Equation (3.31), we obtain an algebraic equation for the unknowns $\mathsf{u}_i^n(t)$. Once the basis functions are chosen, the spatial integrals can be computed exactly, except for the integral for the source term g in (2.10). There

are several possibilities to define the basis functions $\varphi_i^n(x)$. We distinguish between *nodal* and *modal* basis functions [61]. Nodal basis functions are usually suited for easy numerical integration whereas modal basis functions are usually orthogonal polynomials. A nodal basis is for instance given by Lagrange polynomials. We use the Legendre polynomials, which form a modal basis. These are defined on the reference element $[-1, 1]$ by the recursion formula

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \quad \text{for } n = 1, 2, ..., N,$$

with $P_0 = 1$ and $P_1(x) = x$. They fulfill the orthogonality property

$$\int_{-1}^{1} P_n(x)P_m(x)dx = \frac{1}{2n+1}\delta_{mn}$$

with the Kronecker delta δ_{mn}. The Legendre polynomials up to order 3 are depicted in Figure 3.1.

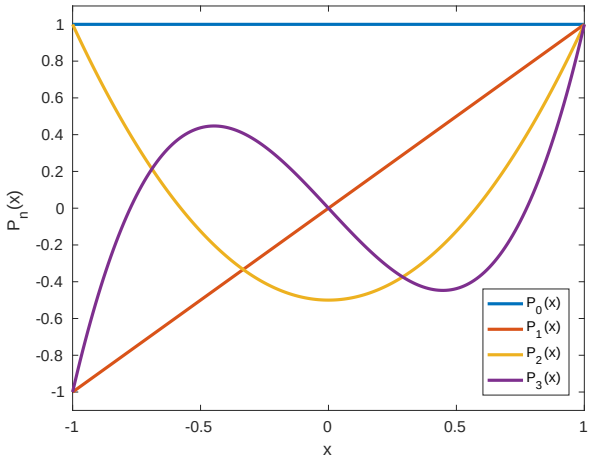

Figure 3.1: Legendre polynomials of degree $n \leq 3$

Finally, this yields an algebraic equation of the form

$$M\partial_t U + DU + BU = \mathsf{g}(U)$$

where $U = (\mathsf{u}_1^1(t), \mathsf{u}_D^1(t), ..., \mathsf{u}_1^N(t), \mathsf{u}_D^N(t))^T$. The entries of the *mass matrix* M are given by

$$m(\phi_i, \phi_j) = \int_{\Omega_h} \phi_i \phi_j \, dx,$$

and for the convection term we obtain

$$(DU)_j = \int_{\Omega_h} \mathsf{f}(\mathsf{u}_h)\phi_j \, dx.$$

The mass matrix is positive definite, has block diagonal structure and is invertible. For linear flux functions, the convective term can also be represented by a matrix with block

diagonal structure. The term B contains terms describing the numerical flux, which will be explained and computed at a later point.

3.2.3 Numerical fluxes

The finite volume and the discontinuous Galerkin method share the characteristic that boundary terms appear in both formulations. Those terms have to be approximated numerically depending on the unknowns, i.e. the cell averages in the finite volume case and point values for the DG method. We introduce the so-called *numerical flux function* $\hat{\mathbf{F}}$ as an approximation for the values on the element boundaries. It is reasonable to define $\hat{\mathbf{F}}$ as a function of the neighbouring cell averages or point values, respectively. In particular, we discuss numerical flux functions of the form

$$\hat{\mathbf{F}}(\mathsf{u}_{j-1}, \mathsf{u}_j) = \mathbf{F}_{j-1/2} \quad \text{or} \quad \hat{\mathbf{F}}(\mathsf{u}^-_{h,j-1/2}, \mathsf{u}^+_{h,j-1/2}) = \mathsf{f}(\mathsf{u}^+_{h,j-1/2}).$$

Remark 2 (High resolution methods)**.** *Note that in order to construct methods of higher-order, the flux function also needs to depend on other degrees of freedom. In the DG case, this can be realized using polynomials of higher-order. For the finite volume scheme, such flux functions can be constructed using a greater stencil. Another possibility is to use so-called* **E**ssentially **n**on-**o**scillatory *(ENO) or* **W**eighted **E**ssentially **n**on-**o**scillatory *(WENO) methods [71].*

In order to construct reasonable schemes which actually converge, the numerical flux needs to satisfy several important properties. Considering linear initial value problems it is well-known that besides stability, also consistency is a crucial requirement for building a convergent scheme, e.g. Lax's equivalence theorem [50].

Definition 2 (Consistent numerical flux)**.** *A numerical flux function $\hat{\mathbf{F}}$ is called consistent if*

$$\hat{\mathbf{F}}(u, u) = \mathsf{f}(u), \tag{3.32}$$

where f denotes the continuous flux function.

Stability is usually ensured in the sense that the numerical flux needs to be Lipschitz continuous. Note that it is particularly important to ensure that the right-hand side in (3.2) is Lipschitz continuous, which guarantees the existence of a solution. In the case of higher-order discretizations, limiters usually need to be employed in order to guarantee stability, c.f. Section 3.4. Note that for the time discretization, the SSP property takes care of the stability.

We emphasize the connection between finite volume and discontinuous Galerkin schemes. Taking $v_h = 1$ in (3.31) and assuming $\mathsf{g} = 0$, we infer

$$\int_{\mathbf{I}_j} \partial_t \mathsf{u}_h dx = \hat{\mathbf{F}}(\mathsf{u}^-_{h,j-1/2}, \mathsf{u}^+_{h,j-1/2}) - \hat{\mathbf{F}}(\mathsf{u}^-_{h,j+1/2}, \mathsf{u}^+_{h,j+1/2}), \tag{3.33}$$

which corresponds to the finite volume formulation. In particular, we see that the scheme is conservative.

Another important property of the numerical flux is its monotonicity. We note that a lower-order DG method with piecewise constant ansatz functions in $P^0(\Omega_h)$ provides a monotone finite volume scheme, see (3.33).

Definition 3 (Monotone numerical flux)**.** *A numerical flux* $\hat{\mathbf{F}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *is called monotone if it is non-decreasing in its first and non-increasing in its second argument. More precisely,*

$$\partial_{\mathsf{u}^-} \hat{\mathbf{F}}(\mathsf{u}^-, \mathsf{u}^+) \geq 0, \quad \text{and} \quad \partial_{\mathsf{u}^+} \hat{\mathbf{F}}(\mathsf{u}^-, \mathsf{u}^+) \leq 0.$$

We introduce some well-known flux functions which satisfy the previous properties. Note that these flux functions are applicable for finite volume and discontinuous Galerkin schemes to the same extend. Further, they can be extended to the system case, even though we present them in a form suitable for scalar conservation laws.

The *Godunov flux* is given as

$$\hat{\mathbf{F}}(\mathsf{u}^-, \mathsf{u}^+) = \begin{cases} \min_{x \in [\mathsf{u}^-, \mathsf{u}^+]} \mathsf{f}(x), & \mathsf{u}^- \leq \mathsf{u}^+, \\ \max_{x \in [\mathsf{u}^-, \mathsf{u}^+]} \mathsf{f}(x), & \text{otherwise} \end{cases}.$$

The Godunov flux shows good performance in the case of nonlinear conservation laws. It is monotone and coincides with the *upwind flux* in the linear case.

A flux function which proved to be robust in practical computations is the *generalized Lax-Friedrichs flux*

$$\hat{\mathbf{F}}(\mathsf{u}^-, \mathsf{u}^+) = \frac{1}{2} \left(\mathsf{f}(\mathsf{u}^-) + \mathsf{f}(\mathsf{u}^+) - \omega(\mathsf{u}^+ - \mathsf{u}^-) \right), \tag{3.34}$$

with the stabilization parameter $\omega > 0$. Again, it coincides with the upwind flux for the choice $\omega = |\mathsf{f}'|$ which is constant in the linear case. In the nonlinear case, we use

$$\omega = \sup_{\mathsf{u} \in \mathcal{U}} |\mathsf{f}'(\mathsf{u})|,$$

where $\mathcal{U}$ is the set of admissible states. Using this flux function results in a monotone scheme. However, the scheme suffers from strong numerical diffusion.

As a remedy to this problem, we can use the *local Lax-Friedrichs flux* with

$$\omega = \sup_{x \in [\mathsf{u}^-, \mathsf{u}^+]} |\mathsf{f}'(x)|,$$

again being a monotone flux. For systems, $|\mathsf{f}'|$ has to be substituted with the eigenvalue with largest absolute value of $\nabla \mathsf{f}$, i.e. $\omega = a$ in case of the water hammer equations.

Remark 3 (Linear case)**.** *Regarding scalar conservation laws with linear flux function, the generalized Lax-Friedrichs flux reduces to the upwind flux for* $\omega = 1$ *and the* centered flux *for* $\omega = 0$*. While the upwind flux is a monotone flux, the centered flux is not.*

3.3 Fully discrete schemes

Let us discuss a fully discrete scheme generated by combining the temporal and spatial discretizations discussed so far. The time interval $[0, T]$ is divided using an equidistant mesh of size Δt. We refer to the points $t^n = n\Delta t$ for $n = 0, \dots, M$. The discrete variable u^n is used as an approximation for $\mathsf{u}(\cdot, t^n)$.

For a fully discrete method, u^n will be a vector containing the spatial degrees of freedom. For spatial discretization of the interval $\mathbf{I}$, we use equally-sized control cells $\mathbf{I}_j$ of size Δx. We use u_j^n as an approximation for $\mathsf{u}(\cdot, t^n)_{|\mathbf{I}_j}$. Note that depending on the spatial discretization, several degrees of freedom are necessary to represent u_j^n.

3.3.1 SDIRK Finite Volume schemes

To derive a fully discrete scheme for the solution of the water hammer equations, we introduce the discrete variables

$$\mathrm{H} = (\mathrm{H}_j)_{j=1,\dots,N}$$
$$\mathrm{Q} = (\mathrm{Q}_j)_{j=1,\dots,N}$$

where the index j corresponds to the control cell $\mathbf{I}_j$. Further, we introduce the system variable

$$\mathrm{W} = (\mathrm{H}_1, \mathrm{Q}_1, \dots, \mathrm{H}_N, \mathrm{Q}_N)^\top.$$

The periodic boundary conditions are incorporated by setting $\mathrm{H}_1 = \mathrm{H}_N$ and $\mathrm{Q}_1 = \mathrm{Q}_N$. However, we can also set other boundary conditions like fixed values or so-called inflow and outflow boundary conditions. We refer to this at a later point.

We apply a finite volume semi-discretization with local Lax-Friedrichs flux resulting in a system of ordinary differential equations of the form

$$\partial_t \mathrm{W} = -\tfrac{1}{\Delta x}\Big[\mathbf{I}_{low,N} \otimes \mathrm{A}^+ + \mathbf{I}_N \otimes |\mathrm{A}| + \mathbf{I}_{up,N} \otimes \mathrm{A}^-\Big]\mathrm{W} + G(\mathrm{W}) \tag{3.35}$$

where

$$\mathrm{A}^+ = \begin{pmatrix} -\frac{\mathbf{a}}{2} & -\frac{\mathbf{a}^2}{2} \\ -\frac{1}{2} & -\frac{\mathbf{a}}{2} \end{pmatrix}, \qquad \mathrm{A}^- = \begin{pmatrix} \frac{\mathbf{a}}{2} & -\frac{\mathbf{a}^2}{2} \\ -\frac{1}{2} & \frac{\mathbf{a}}{2} \end{pmatrix}, \text{ and } |\mathrm{A}| = \mathrm{A}^+ - \mathrm{A}^- = \begin{pmatrix} -\mathbf{a} & 0 \\ 0 & -\mathbf{a} \end{pmatrix}$$

and the matrices $\mathbf{I}_{low,N}, \mathbf{I}_{up,N}$ are $N \times N$ matrices that contain entries 1 on the first lower sub-diagonal or the first upper sub-diagonal, respectively, i.e.,

$$I_{\mathrm{low},N} = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix} \qquad I_{\mathrm{up},N} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \dots & \dots & 0 \end{pmatrix}.$$

$\mathbf{I}_N$ denotes the identity matrix and the operator $G(\mathrm{W})$ is defined as $G(\mathrm{W}) = (0, -\mathsf{g}(\mathrm{Q}_1), \ldots, 0, -\mathsf{g}(\mathrm{Q}_N))^\top$.

Let us introduce the characteristic variables

$$\tilde{\mathrm{W}} = (\tilde{\mathrm{H}}_1, \tilde{\mathrm{Q}}_1, \ldots, \tilde{\mathrm{H}}_N, \tilde{\mathrm{Q}}_N)^\top$$

where $\tilde{\mathrm{H}}_j = \frac{1}{2\mathbf{a}}\mathrm{H}_j + \frac{1}{2}\mathrm{Q}_j$ and $\tilde{\mathrm{Q}}_j = -\frac{1}{2\mathbf{a}}\mathrm{H}_j + \frac{1}{2}\mathrm{Q}_j$. A multiplication with $\mathbf{I}_N \otimes R^{-1}$ results in a system of similar structure for the characteristic variables, i.e.,

$$\partial_t \tilde{\mathrm{W}} = -\frac{1}{\Delta x}\tilde{L}_h \tilde{\mathrm{W}} + \tilde{G}(\tilde{\mathrm{W}}). \tag{3.36}$$

We have

$$\tilde{L}_h = \mathbf{I}_{low,N} \otimes \Lambda^+ + \mathbf{I}_N \otimes |\Lambda| + \mathbf{I}_{up,N} \otimes \Lambda^-, \tag{3.37}$$

where Λ^+, $|\Lambda|$ and Λ^- are defined by

$$\Lambda^+ = \begin{pmatrix} \mathbf{a} & 0 \\ 0 & 0 \end{pmatrix}, \qquad \Lambda^- = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{a} \end{pmatrix}, \text{ and } |\Lambda| = \Lambda^+ - \Lambda^- = \begin{pmatrix} \mathbf{a} & 0 \\ 0 & -\mathbf{a} \end{pmatrix},$$

respectively. The source term is given by

$$\begin{aligned} \tilde{G}(\tilde{\mathrm{W}}) = (&-\tfrac{1}{2}\mathsf{g}(\tilde{\mathrm{H}}_1 + \tilde{\mathrm{Q}}_1), -\tfrac{1}{2}\mathsf{g}(\tilde{H}_1 + \tilde{Q}_1), -\tfrac{1}{2}\mathsf{g}(\tilde{H}_2 + \tilde{Q}_2), -\tfrac{1}{2}\mathsf{g}(\tilde{H}_2 + \tilde{Q}_2), \ldots, \\ &-\tfrac{1}{2}\mathsf{g}(\tilde{H}_{N-1} + \tilde{Q}_{N-1}), -\tfrac{1}{2}\mathsf{g}(\tilde{H}_{N-1} + \tilde{Q}_{N-1}), -\tfrac{1}{2}\mathsf{g}(\tilde{H}_N + \tilde{Q}_N), -\tfrac{1}{2}\mathsf{g}(\tilde{H}_N + \tilde{Q}_N))^T. \end{aligned} \tag{3.38}$$

Note that for the DG discretization using zero-order polynomials, we are working with the same system for the discrete variables. We apply an SDIRK method to (3.35) resulting in

$$\mathrm{W}^{n+1} = v_{m+1}\mathrm{W}^n + \sum_{j=1}^{m} \alpha_{m+1,j} \left(\mathrm{W}^{(j)} + \Delta t \frac{\beta_{m+1,j}}{\alpha_{m+1,j}} \left(-\frac{1}{\Delta x} L_h \ \mathrm{W}^{(j)} + G(\mathrm{W}^{(j)}) \right) \right), \tag{3.39}$$

where the stages can be computed as in (3.5a)-(3.5b). We can perform an analogous computation for the characteristic variables. To get a suitable first-order method, there is the possibility to combine an SDIRK method with a finite volume method using flux limiting which we introduce in Section 3.4.

3.3.2 SDIRK Discontinuous Galerkin schemes

Let us discuss the discretization using a DG method with first-order Legendre polynomials. We introduce the local basis functions for each element $\mathbf{I}_j$:

$$\varphi_j^0(x) \equiv 1, \quad \varphi_j^1(x) = \frac{2}{\Delta x}(x - x_j).$$

The corresponding degrees of freedom for the water hammer equations are denoted by $\mathrm{H}_j^0, \mathrm{H}_j^1$ and $\mathrm{Q}_j^0, \mathrm{Q}_j^1$, respectively. The periodic boundary conditions are implied in the sense that we use the numerical flux function term $\hat{\mathbf{F}}(\mathrm{H}_N^0 + \mathrm{H}_N^1, \mathrm{H}_1^0 - \mathrm{H}_1^1)$ in the first

and last cell. Note again that we can also apply other boundary conditions. Again, we introduce the system variable

$$\mathrm{W} = (\mathrm{H}_1^0, \mathrm{Q}_1^0, \mathrm{H}_1^1, \mathrm{Q}_1^1, \dots, \mathrm{H}_N^0, \mathrm{Q}_N^0, \mathrm{H}_N^1, \mathrm{Q}_N^1)^\top.$$

The semi-discretization again using the local Lax-Friedrichs flux is then given as

$$\partial_t M_h \mathrm{W} = ((\mathbf{I}_{low,N} \otimes C) + (\mathbf{I}_N \otimes B) + (\mathbf{I}_{up,N} \otimes D))\,\mathrm{W} + G(\mathrm{W}), \tag{3.40}$$

where the diagonal mass matrix is given as

$$M_h = \Delta x\, \mathrm{diag}(1, 1, 1/3, 1/3, \dots, 1, 1, 1/3, 1/3) \tag{3.41}$$

and

$$C = \begin{pmatrix} \frac{\mathbf{a}}{2} & \frac{\mathbf{a}^2}{2} & \frac{\mathbf{a}}{2} & \frac{\mathbf{a}^2}{2} \\ -\frac{1}{2} & \frac{\mathbf{a}}{2} & -\frac{1}{2} & \frac{\mathbf{a}}{2} \\ -\frac{\mathbf{a}}{2} & -\frac{\mathbf{a}^2}{2} & -\frac{\mathbf{a}}{2} & -\frac{\mathbf{a}^2}{2} \\ -\frac{1}{2} & -\frac{\mathbf{a}}{2} & -\frac{1}{2} & -\frac{\mathbf{a}}{2} \end{pmatrix},\ B = \begin{pmatrix} -\mathbf{a} & 0 & 0 & -\mathbf{a}^2 \\ 0 & -\mathbf{a} & 1 & 0 \\ 0 & \mathbf{a}^2 & -\mathbf{a} & 0 \\ 1 & 0 & 0 & -\mathbf{a} \end{pmatrix},\ D = \begin{pmatrix} \frac{\mathbf{a}}{2} & -\frac{\mathbf{a}^2}{2} & -\frac{\mathbf{a}}{2} & \frac{\mathbf{a}^2}{2} \\ \frac{1}{2} & \frac{\mathbf{a}}{2} & -\frac{1}{2} & -\frac{\mathbf{a}}{2} \\ \frac{\mathbf{a}}{2} & -\frac{\mathbf{a}^2}{2} & -\frac{\mathbf{a}}{2} & \frac{\mathbf{a}^2}{2} \\ -\frac{1}{2} & \frac{\mathbf{a}}{2} & \frac{1}{2} & -\frac{\mathbf{a}}{2} \end{pmatrix}.$$

For the integral of the source term, we use the trapezoidal rule for integration and have

$$G(\mathrm{W}) = \left(0, -\tfrac{\Delta x}{2}(\mathrm{g}[\mathrm{Q}_j^0 - \mathrm{Q}_j^1] + \mathrm{g}[\mathrm{Q}_j^0 + \mathrm{Q}_j^1]), 0, \tfrac{\Delta x}{2}(\mathrm{g}[\mathrm{Q}_j^0 - \mathrm{Q}_j^1] - \mathrm{g}[\mathrm{Q}_j^0 + \mathrm{Q}_j^1])\right)_j^\top.$$

Once again, we can apply an SDIRK scheme and obtain

$$\mathrm{W}^{n+1} = v_{m+1}\mathrm{W}^n + \sum_{j=1}^{m} \alpha_{m+1,j} \left(\mathrm{W}^{(j)} + \Delta t \frac{\beta_{m+1,j}}{\alpha_{m+1,j}} \left(M_h^{-1} \left(K_h \mathrm{W}^{(j)} + G(\mathrm{W}^{(j)})\right)\right)\right),$$

where $\mathrm{W}^{(j)}$ can be computed by (3.5a)-(3.5b). Analogous for the characteristic variables, this yields

$$M_h \partial_t \tilde{\mathrm{W}} = \tilde{K}_h \tilde{\mathrm{W}} + \tilde{G}(\tilde{\mathrm{W}}) \tag{3.42}$$

with

$$\tilde{K}_h = \mathbf{I}_{low,N} \otimes \tilde{C} + \mathbf{I}_N \otimes \tilde{B} + \mathbf{I}_{up,N} \otimes \tilde{D}$$

and

$$\tilde{C} = \begin{pmatrix} a & 0 & a & 0 \\ 0 & 0 & 0 & 0 \\ -a & 0 & -a & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},\ \tilde{B} = \begin{pmatrix} -a & 0 & -a & 0 \\ 0 & -a & 0 & a \\ a & 0 & -a & 0 \\ 0 & -a & 0 & -a \end{pmatrix},\ \tilde{D} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & a & 0 & -a \\ 0 & 0 & 0 & 0 \\ 0 & a & 0 & -a \end{pmatrix}.$$

The source term is defined as

$$\tilde{G}(\tilde{\mathrm{W}}) = \tfrac{\Delta x}{4}\Big(-(\mathrm{g}[M_j] + \mathrm{g}[m_j]), \mathrm{g}[M_j] - \mathrm{g}[m_j], -(\mathrm{g}[M_j] + \mathrm{g}[m_j]), \mathrm{g}[M_j] - \mathrm{g}[m_j]\Big)_j^\top$$

with $M_j = \tilde{\mathrm{H}}_j^0 + \tilde{\mathrm{Q}}_j^0 - (\tilde{\mathrm{H}}_j^1 + \tilde{\mathrm{Q}}_j^1)$ and $m_j = \tilde{\mathrm{H}}_j^0 + \tilde{\mathrm{Q}}_j^0 + \tilde{\mathrm{H}}_j^1 + \tilde{\mathrm{Q}}_j^1$ for all $j = 1, ..., N$.
The application of the SDIRK scheme is as described above.

3.3.3 The implicit box scheme

The schemes we discussed so far can be written in the form

$$\mathsf{u}_j^{n+1} = F(\mathsf{u}_{j-1}^n, \mathsf{u}_j^n, \mathsf{u}_{j+1}^n, \mathsf{u}_{j-1}^{n+1}, \mathsf{u}_j^{n+1}, \mathsf{u}_{j+1}^{n+1}). \tag{3.43}$$

Note that in general, upper indices refer to time and lower indices refer to space.

For later reference, we also introduce the implicit box scheme (IBOX) [43] which we use for comparison. When applied to (3.1), the scheme can be stated as

$$\frac{\mathsf{u}_{j-1}^{n+1} + \mathsf{u}_j^{n+1}}{2} = \frac{\mathsf{u}_{j-1}^n + \mathsf{u}_j^n}{2} - \frac{\Delta t}{\Delta x}(\mathsf{f}(\mathsf{u}_j^{n+1}) - \mathsf{f}(\mathsf{u}_{j-1}^{n+1})) + \Delta t \frac{\mathsf{g}(\mathsf{u}_{j-1}^{n+1}) + \mathsf{g}(\mathsf{u}_j^{n+1})}{2}. \tag{3.44}$$

Note that this scheme can be derived using the implicit Euler method and a Petrov-Galerkin finite element method of first-order. Stability and convergence towards the entropy solution of this scheme can be proven if the lower bound

$$\frac{\Delta t}{\Delta x} \geq \frac{1}{2\lambda_{\min}} \tag{3.45}$$

is satisfied with $f' \geq \lambda_{\min} > 0$. Here, $\lambda_{\min}$ is the smallest eigenvalue of the Jacobian and g is dissipative. If we have $f' \leq -\lambda_{\min} < 0$, the statements can be proven analogously and we need to shift the indices for the proof.

3.3.4 The CFL condition

Another important property which is necessary but not-sufficient for the convergence of the scheme to the exact solution is the so-called *Courant Friedrich Levy* (CFL) condition

$$\left| \frac{\Delta t}{\Delta x} \partial_{\mathsf{u}} \mathbf{F}(\mathsf{u}) \right| \leq c \tag{3.46}$$

for all u and a constant $c \in \mathbb{R}$. Descriptively, this condition ensures that the numerical method can only be convergent if the numerical domain of dependence contains the domain of dependence of the PDE [51, Chapt. 4.4]. This should be true at least in the limit with $\Delta t, \Delta x \to 0$. This condition is necessary for schemes with explicit time integration. Since the SSP coefficient of SDIRK methods depends on the time step of the explicit Euler method, the SDIRK methods suffer from the CFL condition indirectly. Note that the CFL condition does not need to be fulfilled by the IBOX scheme.

3.4 Limiter

As mentioned before, hyperbolic equations admit shocks, i.e. discontinuities, even for smooth initial data. Especially for higher-order schemes, the *Gibbs phenomenon* appears in computations. This causes non-physical behaviour for instance a negative density of the numerical solution of the compressible Euler equations.

We have already introduced the concept of nonlinear stability in suitable norms. The specific norm measuring oscillations is the *total variation* (TV) [32] semi-norm. We distinguish between so-called *total variation diminishing* (TVD) [32], *total variation bounded* (TVB) [11, 12], *essentially non-oscillatory* (ENO) [69], and *weighted essentially non-oscillatory* (WENO) [71] schemes.

These methods utilize so-called *limiters* to guarantee stability in the corresponding norm. In general, we distinguish two types of limiters, namely *slope limiters* and *flux limiters*.

3.4.1 Scalar limiters

We discuss the concept of limiters for the scalar case. An extension of the same ideas to the system case will be presented in the next section. We first discuss the finite volume setting before the concepts are conferred to the discontinuous Galerkin setting.

A norm used to measure oscillations is the *total variation* semi-norm

$$TV(\mathsf{u}) = \sum_{i=-\infty}^{\infty} |\mathsf{u}_i - \mathsf{u}_{i-1}|.$$

Note that for periodic functions, this can be defined on bounded domains. Also note that in the case of DG methods, the TV norm is defined for the mean values on the elements.

Next, we introduce the concept of so-called TVD finite volume methods [51].

Definition 4 (TVD FV methods [51])**.** *A time stepping method* (3.43) *is called* ***t****otal* ***v****ariation* ***d****iminishing (TVD) if the iterates* u^{n+1} *satisfy*

$$TV(\mathsf{u}^{n+1}) \leq TV(\mathsf{u}^n) \quad \textit{for all data } \mathsf{u}^n. \tag{3.47}$$

Further, we introduce a property that guarantees that monotonicity is preserved using a time stepping scheme.

Definition 5 (Monotonicity-preserving [51])**.** *A time stepping method is called* monotonicity preserving *if*

$$\mathsf{u}_i^n \geq \mathsf{u}_{i+1}^n \quad \textit{for all } i$$

implies

$$\mathsf{u}_i^{n+1} \geq \mathsf{u}_{i+1}^{n+1} \quad \textit{for all } i.$$

Note that any TVD scheme is also monotonicity-preserving.

As mentioned before, the concept of TVD methods does not apply to DG methods in a straightforward manner. In particular, the TVD property does not hold for the discrete point values

$$u^-_{j+1/2} := \begin{cases} u_{h|_{I_j}}(x_{j+1/2}), & j \neq 0 \\ u_{h|_{I_N}}(x_{N+1/2}), & j = 0 \end{cases} \quad u^+_{j+1/2} := \begin{cases} u_{h|_{I_{j+1}}}(x_{j+1/2}), & j < N \\ u_{h|_{I_1}}(x_{1/2}), & j = N \end{cases}, \tag{3.48}$$

which are usually the unknowns in a DG formulation. However, if we compute the average over the elements of the numerical solution, we can gain a slightly different condition.

Definition 6 (TVDM DG methods [61])**.** *A numerical method is called **t**otal **v**ariation **d**iminishing in the **m**eans (TVDM) if*

$$\mathrm{TVM}(\mathsf{u}_h^{n+1}) := \sum_{j=1}^{D} |\bar{\mathsf{u}}_{j+1}^{n+1} - \bar{\mathsf{u}}_j^{n+1}| \leq \mathrm{TVM}(\mathsf{u}_h^n). \tag{3.49}$$

Here, $\bar{\mathsf{u}}_j^{n+1}$ denotes the mean value of the discrete function over $\mathbf{I}_j$.

An important tool to show the TVD property of a method is Harten's Lemma.

Lemma 1 (Harten's Lemma [32, 51, 61])**.** *Consider a method of the form*

$$\mathsf{u}_j^{n+1} = \mathsf{u}_j^n + \frac{\Delta t}{\Delta x}\left[C_{j+1/2}(\mathsf{u}_{j+1}^n - \mathsf{u}_j^n) - D_{j-1/2}(\mathsf{u}_j^n - \mathsf{u}_{j-1}^n)\right]$$

for all $j = 1, .., N$, periodic boundary conditions $u_0 := u_N$ and $u_{N+1} := u_1$, and with given real numbers $C_{j+1/2}$ and $D_{j-1/2}$ for all $j = 1, .., N$. Further, assume that

$$C_{j+1/2} \geq 0, \quad D_{j-1/2} \geq 0, \quad \lambda(C_{j+1/2} + D_{j+1/2}) \leq 1 \tag{3.50}$$

with $D_{N+1/2} := D_{1/2}$ by periodicity. Then the scheme fulfills the TVD property (3.47) *for FV or* (3.49) *for DG schemes, respectively.*

Proof. Proof can be found in [61, Chap. 3, p. 111]. □

Before we discuss some explicit limiters, we note that requiring a method to be TVD introduces some severe limitations, as the following theorem shows.

Theorem 3 (Osher and Chakravarthy [60])**.** *Near smooth extrema TVD methods degenerate to first-order accuracy.*

Proof. Proof can be found in [60]. □

To generate methods which do not have this defect, the TVD requirement has to be relaxed.

Definition 7 (TVB methods [51])**.** *A method is called **t**otal **v**ariation **b**ounded (TVB) if the iterates satisfy*

$$TV(\bar{\mathsf{u}}^{n+1}) \leq (1 + C\Delta t)TV(\bar{\mathsf{u}}^n) \tag{3.51}$$

for all data $\bar{\mathsf{u}}^n$ and for a constant $C > 0$.

Slope limiters

We discuss the concept of slope limiters for the DG method. Slope limiters can also be defined for finite volume schemes which we omit here. In general, limiters are designed to compute solutions which do not inherit spurious oscillations. To understand this, we introduce the decomposition

$$\mathsf{u}_j(x) := \bar{\mathsf{u}}_j + \tilde{\mathsf{u}}_j(x) \tag{3.52}$$

on each element, where $\bar{\mathsf{u}}_j$ denotes the mean value on the interval $\mathbf{I}_j$. If u is a piecewise linear function, the variation $\tilde{\mathsf{u}}_j$ can be interpreted as the slope on $\mathbf{I}_j$. A slope limiter tries to avoid oscillations by limiting the slope in a suitable manner. To do so, we substitute u with

$$\mathsf{u}_j(x) = \bar{\mathsf{u}}_j + (x - x_j)\sigma_j, \tag{3.53}$$

i.e. the slope is set to σ_j. This amounts to a linear reconstruction of u.

The limiter σ can be chosen in such a way that the resulting method fulfills the TVD property. An example for such a limiter is the *MinMod* limiter

$$\sigma_j = \xi\left(\frac{\mathsf{u}^-_{j+1/2} - \mathsf{u}^+_{j-1/2}}{\Delta x}, \frac{\bar{\mathsf{u}}_{j+1} - \bar{\mathsf{u}}_j}{\Delta x}, \frac{\bar{\mathsf{u}}_j - \bar{\mathsf{u}}_{j-1}}{\Delta x}\right) \tag{3.54}$$

with

$$\xi(a_1, a_2, a_3) = \begin{cases} s\min(|a_1|, |a_2|, |a_3|), & s = \operatorname{sign}(a_1) = \operatorname{sign}(a_2) = \operatorname{sign}(a_3) \\ 0, & \text{otherwise} \end{cases}. \tag{3.55}$$

Together with the suitable numerical flux, a numerical scheme containing this limiter fulfills the TVD and also the TVDM property. The limiter belongs to the class of so-called *TVD limiters*.

Note that for general limiters, we can utilize Harten's famous Lemma. For a combination of the DG method with the explicit Euler time stepping, we have the following theorem.

Theorem 4 (TVDM scheme [61, Chap. 3, p. 111])**.** *Let $\hat{\mathbf{F}}$ be a monotone, Lipschitz continuous numerical flux. Assume that there is $\theta > 0$ s.t.*

$$-1 \leq \frac{\tilde{u}^+_{j+1/2} - \tilde{u}^+_{j-1/2}}{\bar{u}_{j+1} - \bar{u}_j} \leq \theta \quad -1 \leq \frac{\tilde{u}^-_{j+1/2} - \tilde{u}^-_{j-1/2}}{\bar{u}_j - \bar{u}_{j-1}} \leq \theta \tag{3.56}$$

for all $j = 1, ..., N$. Further, assume that the CFL condition

$$\Delta t \leq \frac{1}{(1+\theta)(L_1 + L_2)}\Delta x, \tag{3.57}$$

where L_1 and L_2 are the Lipschitz constants of the numerical flux w.r.t. the first and second argument. Then the scheme is TVD(M).

Proof. First of all, we need to check the assumptions of Harten's Lemma. In both schemes, we obtain the Equation (3.33) and can reformulate it as

$$\begin{aligned} u_j^{n+1} =& u_j^n - \lambda \left(\hat{F}(u^-_{j+1/2}, u^+_{j+1/2}) - \hat{F}(u^-_{j+1/2}, u^+_{j-1/2}) \right) \\ & - \lambda \left(\hat{F}(u^-_{j+1/2}, u^+_{j-1/2}) - \hat{F}(u^-_{j-1/2}, u^+_{j-1/2}) \right) \\ =& \bar{u}_j + \lambda \left(C_{j+1/2}(\bar{u}_{j+1} - \bar{u}_j) - D_{j-1/2}(\bar{u}_j - \bar{u}_{j-1}) \right) \end{aligned}$$

with

$$\begin{aligned} C_{j+1/2} &= \frac{\hat{F}(u^-_{j+1/2}, u^+_{j+1/2}) - \hat{F}(u^-_{j+1/2}, u^+_{j-1/2})}{u^+_{j-1/2} - u^+_{j+1/2}} \left(1 + \frac{\tilde{u}^+_{j+1/2} - \tilde{u}^+_{j-1/2}}{\bar{u}_{j+1} - \bar{u}_j} \right) \\ D_{j-1/2} &= \frac{\hat{F}(u^-_{j+1/2}, u^+_{j-1/2}) - \hat{F}(u^-_{j-1/2}, u^+_{j-1/2})}{u^-_{j+1/2} - u^-_{j-1/2}} \left(1 + \frac{\tilde{u}^-_{j+1/2} - \tilde{u}^-_{j-1/2}}{\bar{u}_j - \bar{u}_{j-1}} \right), \end{aligned}$$

where we used the decomposition of u_h on each mesh element (3.52). This yields that the values $C_{j+1/2}$ and $D_{j-1/2}$ are positive due to the monotonicity of the numerical flux and the assumption (3.56). Next, we need to check the third condition of (3.50) of Harten's Lemma. With the Lipschitz constants L_2 and L_1, we can compute an upper bound for the values $C_{j+1/2}$ and $D_{j+1/2}$ and get

$$C_{j+1/2} \leq L_2(1+\theta) \quad \text{and} \quad D_{j+1/2} \leq L_1(1+\theta).$$

Together with the CFL condition (3.57), we get

$$\lambda(C_{j+1/2} + D_{j+1/2}) \leq \frac{\Delta t}{\Delta x}(L_2 + L_1)(1+\theta) \leq 1.$$

□

As mentioned before, the problem arises that TVD methods degenerate to first-order accuracy near critical points. It is however possible to define limiters which do not degenerate and still are TVBM. We consider the TVB correction

$$\xi_{TVB}(a_1, a_2, a_3) = \begin{cases} a_1, & |a_1| \leq Mh^2 \\ \xi(a_1, a_2, a_3), & \text{otherwise} \end{cases} \tag{3.58}$$

for (3.54), where $M = \frac{2}{3}M_2$ and M_2 is an approximate value of the second derivative near smooth critical points of the initial function $\mathsf{u}_0(x)$, i.e. $M_2 = \max_{x\in\Omega, \mathsf{u}_0'(x)=0} |\partial_{xx} u_0(x)|$.

Lemma 2 (TVBM schemes)**.** *The statements of Theorem 4 hold with replacing Equation* (3.55) *by* (3.58) *and the word TVDM by TVBM.*

Flux limiter

Another possibility to reduce spurious oscillations is applying *flux limiters*. We discuss those limiters for finite volume schemes. For these schemes the values on the control cell boundaries are important, see (3.29). The idea of flux limiters is to substitute these values with the expression

$$\mathsf{u}_{j+1/2} = \bar{\mathsf{u}}_j + \tfrac{1}{2}\psi(\theta_j)(\bar{\mathsf{u}}_{j+1} - \bar{\mathsf{u}}_j) \text{ where } \quad \theta_j = \frac{\bar{\mathsf{u}}_j - \bar{\mathsf{u}}_{j-1}}{\bar{\mathsf{u}}_{j+1} - \bar{\mathsf{u}}_j} \tag{3.59}$$

for all $j = 1, ..., N$. The function ψ is called *limiter function*. This function is chosen such that the numerical solution admits better accuracy compared to the case without limiting. Note that we have to distinguish the cases $\mathsf{f}' \geq 0$ and $\mathsf{f}' < 0$ for the continuous flux function. For the latter, we have to use

$$\mathsf{u}_{j+1/2} = \bar{\mathsf{u}}_{j+1} + \tfrac{1}{2}\psi(\frac{1}{\theta_{j+1}})(\bar{\mathsf{u}}_j - \bar{\mathsf{u}}_{j+1}) \text{ where } \theta_{j+1} = \frac{\bar{\mathsf{u}}_{j+1} - \bar{\mathsf{u}}_j}{\bar{\mathsf{u}}_{j+2} - \bar{\mathsf{u}}_{j+1}}, \tag{3.60}$$

which corresponds to a reflection around the value $x_{j+1/2}$ for all $j = 1, ..., N-1$.

Considering that $f' \geq 0$, a sufficient condition for a flux limiter to be TVD is

$$0 \leq \psi(\theta) \leq 2 \text{ and } \quad 0 \leq \frac{1}{\theta}\psi(\theta) \leq \upsilon \tag{3.61}$$

for a positive constant υ, see Figure 3.2 and [36, Chap. 3, Section 1]. If $f' \leq 0$, then the sufficient conditions read

$$0 \leq \psi(\tfrac{1}{\theta}) \leq 2 \text{ and } \quad 0 \leq \theta\psi(\tfrac{1}{\theta}) \leq \upsilon \tag{3.62}$$

for a positive constant υ. Note that for TVD limiters, which are restricted by one from above, the first inequality in (3.61) or (3.62) reads $0 \leq \psi(\theta) \leq 1$ or $0 \leq \psi(\frac{1}{\theta}) \leq 1$. A limiter which fulfills this condition is, e.g., the *Koren limiter* [45]

$$\psi(\theta) = \max(0, \min(2, \min(2/3 + 1/3\theta, 2\theta))). \tag{3.63}$$

It satisfies condition (3.61) with $\upsilon = 2$ and is based on a third-order upwind scheme [36, Example 1.1, p. 217].

Example 3. *We list some TVD limiters which are visualized in Figure 3.2:*

- *MinMod [65]:* $\psi_{MM}(\theta) = \max(0, \min(1, \theta))$ *with* $\lim_{\theta\to\infty} \psi_{MM}(\theta) = 1$
- *van Leer [85]:* $\psi_{VL}(\theta) = \frac{\theta+|\theta|}{\theta^2+1}$ *with* $\lim_{\theta\to\infty} \psi_{VL}(\theta) = 2$
- *Super-bee [65]:* $\psi_{SB}(\theta) = \max(0, \min(2\theta, 1), \min(\theta, 2))$ *with* $\lim_{\theta\to\infty} \psi_{SB}(\theta) = 2$
- *van Albada [84]:* $\psi_{AL}(\theta) = \frac{\theta^2+\theta}{\theta^2+1}$ *with* $\lim_{\theta\to\infty} \psi_{AL}(\theta) = 1$

Note that all these limiters lie in the TVD region *defined in [81]. The upper and lower bound of this region is formed by the Super-bee and the MinMod limiter, respectively.*

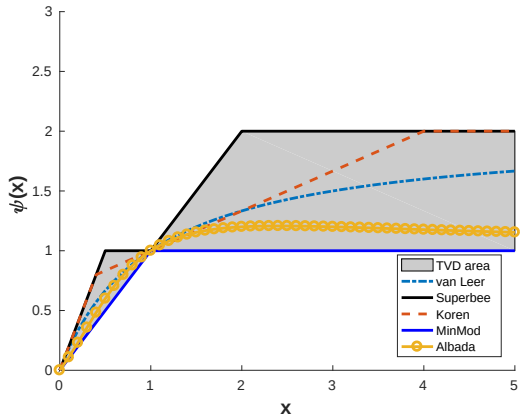

Figure 3.2: Different flux limiters for the finite volume scheme in the TVD region of Sweby [81]

Lemma 3. *Let us assume that* $\mathsf{f}' \geq 0$. *It follows that a finite volume scheme with flux limiting is given as*

$$\partial_t \bar{\mathsf{u}}_j = -\frac{1}{\Delta x}\left(\mathsf{f}(\bar{\mathsf{u}}_j + \tfrac{1}{2}\psi(\theta_j)(\bar{\mathsf{u}}_{j+1} - \bar{\mathsf{u}}_j)) - \mathsf{f}(\bar{\mathsf{u}}_{j-1} + \tfrac{1}{2}\psi(\theta_{j-1})(\bar{\mathsf{u}}_j - \bar{\mathsf{u}}_{j-1}))\right),$$

which corresponds to a limited version of (3.29) *with* $\mathsf{G} = 0$. *Again using Harten's Lemma and Equation* (3.61), *we can show that this scheme in combination with the explicit Euler method is TVD.*

Proof. Applying the explicit Euler method in time, we get

$$\bar{\mathsf{u}}_j^{n+1} = \underbrace{\bar{\mathsf{u}}_j^n + \frac{\Delta t}{\Delta x}\left(\mathsf{f}(\bar{\mathsf{u}}_j^n + \tfrac{1}{2}\psi(\theta_j)(\bar{\mathsf{u}}_{j+1}^n - \bar{\mathsf{u}}_j^n)) - \mathsf{f}(\bar{\mathsf{u}}_{j-1}^n + \tfrac{1}{2}\psi(\theta_{j-1})(\bar{\mathsf{u}}_j^n - \bar{\mathsf{u}}_{j-1}^n))\right)}_{G_\lambda(\bar{\mathsf{u}}_{j-2}^n, \bar{\mathsf{u}}_{j-1}^n, \bar{\mathsf{u}}_j^n, \bar{\mathsf{u}}_{j+1}^n)}. \quad (3.64)$$

Using the mean value Theorem, we can rewrite the right-hand side and get

$$\bar{\mathsf{u}}_j^{n+1} = \bar{\mathsf{u}}_j^n + \frac{\Delta t}{\Delta x}\mathsf{f}'(r_j^n)\left(1 - \tfrac{1}{2}\psi(\theta_j) + \frac{1}{2\theta_j}\psi(\theta_j)\right)(\bar{\mathsf{u}}_{j-1}^n - \bar{\mathsf{u}}_j^n) \quad (3.65)$$

with $r_j^n \in [\mathsf{u}_{j-1/2}^n, \mathsf{u}_{j+1/2}^n]$ and we denote $\iota_j(\bar{\mathsf{u}}_j^n) := \frac{1}{\Delta x}\mathsf{f}'(r_j^n)\left(1 - \frac{1}{2}\psi(\theta_j) + \frac{1}{2\theta_j}\psi(\theta_j)\right)$. Scheme (3.65) is a special case of the more general space periodic scheme

$$\bar{\mathsf{u}}_j^{n+1} = \bar{\mathsf{u}}_j^n + \Delta t \iota_j(\bar{\mathsf{u}}_j^n)(\bar{\mathsf{u}}_{j-1}^n - \bar{\mathsf{u}}_j^n) - \kappa_j(\bar{\mathsf{u}}_j^n)(\bar{\mathsf{u}}_j^n - \bar{\mathsf{u}}_{j+1}^n).$$

For this scheme, we can apply Lemma 1 (Harten's Lemma) with the sufficient conditions $\iota_j(w) \geq 0$, $\kappa_j(w) \geq 0$ and $\Delta t(\iota_{j+1}(\mathsf{u}) + \kappa_j(\mathsf{u})) \leq 1$ for all j. Together with the conditions (3.61), we find that the scheme (3.64) is TVD, if

$$\frac{\Delta t f'(w)}{\Delta x} \leq \frac{1}{1+\upsilon},$$

see [36, III, Section 1]. □

3.4.2 System case

We discuss an extension of the limiter concept to systems of conservation laws. To this end, we consider system (2.10) with $g = 0$. The most simple method is to apply the limiter in a component-wise manner. If we use this in combination with the local Lax-Friedrich flux (3.34), this approach is easy to implement. However, for this conservative setting, there is no TVD or TVB theory available even for linear flux functions. Numerically, we also observe so-called *wriggles* when using this ansatz.

A solution to this is the *characteristic field decomposition*, c.f. Section 2.3. The local projection can be applied to the characteristic variables $\tilde{\mathrm{H}}$ and $\tilde{\mathrm{Q}}$ separately, see (2.37). As in Chapter 2.3, we set $\mathrm{W} = (\mathrm{H}, \mathrm{Q})^\top$. With the matrix R defined in (2.26) we have

$$(\tilde{\mathrm{H}}_j, \tilde{\mathrm{Q}}_j)^\top = \tilde{\mathrm{W}}_j = R^{-1}\mathrm{W}_j.$$

Remark 4 (Characteristic field decomposition for nonlinear systems)**.** *In the nonlinear case, the characteristic field decomposition is not so easy to understand since the Jacobian Matrix* $\mathbf{A}$ *of the quasi-linear form*

$$\partial_t \mathbf{w} + \partial_x \mathsf{f}(\mathbf{w}) =: \partial_t \mathbf{w} + \mathbf{A}\partial_x \mathbf{w} = 0$$

is dependent on the state $\mathbf{w}$*. Considering a Riemann problem, this results in getting two states and two different Jacobian matrices in general. To define a suitable approximation of the Jacobian matrix, Roe defined the so-called Roe matrices [64].*

For this decoupled system, we can apply the finite volume scheme with flux limiting in a component-wise manner. For the first component we have

$$\tilde{\mathrm{H}}_{j+1/2} = \tilde{\mathrm{H}}_j + \tfrac{1}{2}\psi(\theta_j)(\tilde{\mathrm{H}}_{j+1} - \tilde{\mathrm{H}}_j) \text{ where } \quad \theta_j = \frac{\tilde{\mathrm{H}}_j - \tilde{\mathrm{H}}_{j-1}}{\tilde{\mathrm{H}}_{j+1} - \tilde{\mathrm{H}}_j} \tag{3.66}$$

or

$$\tilde{\mathrm{H}}_{j+1/2} = \tilde{\mathrm{H}}_{j+1} + \tfrac{1}{2}\psi(\frac{1}{\theta_{j+1}})(\tilde{\mathrm{H}}_j - \tilde{\mathrm{H}}_{j+1}) \text{ where } \theta_{j+1} = \frac{\tilde{\mathrm{H}}_{j+1} - \tilde{\mathrm{H}}_j}{\tilde{\mathrm{H}}_{j+2} - \tilde{\mathrm{H}}_{j+1}}, \tag{3.67}$$

respectively, depending on the sign of the first eigenvalue of the system matrix. The scheme for the second component is defined analogously. We can now evaluate the flux functions

$$F_{j+1/2} = \mathsf{f}(\tilde{\mathrm{H}}_{j+1/2}) \quad \text{ and } \quad F_{j-1/2} = \mathsf{f}(\tilde{\mathrm{H}}_{j-1/2}).$$

The scheme for the original or conservative variables is generated by component-wise multiplication with R. To show that the scheme is TVD, we proceed as in the scalar case for the single components. Let us discuss the DG method. We use piecewise linear reconstruction (3.53) and TVB (3.58) or TVD correction (3.55), respectively. We denote the average values

$$(\bar{\tilde{\mathrm{H}}}_j, \bar{\tilde{\mathrm{Q}}}_j)^\top = \bar{\tilde{\mathrm{W}}}_j = R^{-1}\bar{\mathrm{W}}_j$$

where $\bar{\mathrm{W}}_j = (\bar{\mathrm{H}}_j, \bar{\mathrm{Q}}_j)^\top$. We apply scheme (3.48) in a component-wise manner with the local projection limiting

$$\sigma_j = \xi \left(\frac{\bar{\tilde{\mathrm{H}}}_{j+1/2} - \bar{\tilde{\mathrm{H}}}_{j-1/2}}{\Delta x}, \frac{\bar{\tilde{\mathrm{H}}}_{j+1} - \bar{\tilde{\mathrm{H}}}_j}{\Delta x}, \frac{\bar{\tilde{\mathrm{H}}}_j - \bar{\tilde{\mathrm{H}}}_{j-1}}{\Delta x} \right).$$

The corresponding piecewise linear reconstruction (3.53) reads

$$\tilde{\mathrm{H}}_j(x, \cdot\,) = \bar{\tilde{\mathrm{H}}}_j + (x - x_j)\sigma_j. \tag{3.68}$$

The second component can be treated analogously. The local Lax-Friedrichs flux for the characteristic variables reads

$$\hat{\mathbf{F}}(\tilde{\mathrm{H}}^-_{j+1/2}, \tilde{\mathrm{H}}^+_{j+1/2}) = \frac{1}{2}[(\tilde{\mathrm{f}}^+_{j+1/2}) + (\tilde{\mathrm{f}}^-_{j+1/2}) - \omega_{j+1/2}((\tilde{\mathrm{H}}^+_{j+1/2}) - (\tilde{\mathrm{H}}^-_{j+1/2}))], \tag{3.69}$$

where $\omega_{j+1/2}$ is the maximal eigenvalue of the system matrix. If we utilize the results from the scalar case, we gain a TVBM scheme after transforming back to the original variables. Note that this approach works for the linear system without source term only.

Remark 5. *As already mentioned before, we can derive a scheme from a combination of a SDIRK method and the finite volume method with generalized Lax-Friedrichs flux* (3.34) *and the flux limiter* (3.59) *or* (3.60)*, respectively. We call this scheme* SDIRKFLUX.

4 Well-Balancedness and the discrete Maximum-Principle

We introduced higher-order time discretizations suitable for the solution of hyperbolic balance laws. These methods are built as a combination of an SSP SDIRK time stepping scheme and a spatial discretization by finite volumes or a discontinuous Galerkin method. The main advantage of higher-order SSP SDIRK methods is that larger time steps are possible compared to the explicit Euler method.

In the following, we show that the schemes built on SSP SDIRK methods are well-balanced and satisfy a discrete maximum-principle. So far, the SSP property is the most restrictive aspect with regard to the step size. To show the well-balancedness of the schemes, it is crucial to ensure that no further restriction on the step size occurs.

4.1 Existence and uniqueness

In this section, we show the existence and uniqueness of discrete solutions. Due to the fact that we use the MOL approach, the proof is based on the theory for ordinary differential equations. It is well-known that an equation of the form

$$\partial_t \mathbf{w} = \mathbf{F}(\mathbf{w}) \tag{4.1}$$

admits a unique solution if $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitz continuous function. When applying an implicit RK method to (4.1), the scheme admits a unique solution under a suitable time step restriction related to the Lipschitz constant L of $\mathbf{F}$ [79, Chapter 7/8].

For our further considerations, we introduce the *one-sided Lipschitz condition.*

Definition 8 (One-sided Lipschitz condition (OSLIP))**.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ induced by a scalar product $\langle\cdot,\cdot\rangle$ and let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a function. If there exists a constant $\nu \in \mathbb{R}$ such that*

$$\langle \mathbf{F}(\mathbf{w}) - \mathbf{F}(\tilde{\mathbf{w}}), \mathbf{w} - \tilde{\mathbf{w}} \rangle \leq \nu \|\mathbf{w} - \tilde{\mathbf{w}}\|^2, \tag{4.2}$$

for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^n$, then $\mathbf{F}$ fulfills the one-sided Lipschitz condition. The constant ν is called one-sided Lipschitz constant *(OSLC).*

Note that in contrast to the usual Lipschitz constant, the OSLC can be negative. Next, we introduce a constant corresponding to the Butcher matrix A of the RK method.

Definition 9 ([31, Chap. IV.14, p. 229])**.** *Consider the inner product $\langle u, v\rangle_D = u^T D v$ where $D = diag(d_1, ..., d_s)$ with $d_i > 0$. Then, we denote by $\alpha_D(A^{-1})$ the largest number α such that*

$$\langle u, A^{-1}u\rangle_D \geq \alpha_D(A^{-1})\langle u, u\rangle_D \tag{4.3}$$

for all $\mathsf{u} \in \mathbb{R}^m$. Further, we set

$$\alpha_0 = \sup_{D>0} \alpha_D(A^{-1}).$$

Note that α_0 is in some sense the optimal coercivity constant for the mapping A^{-1}. For SDIRK methods, we can compute the constant α_0 explicitly.

Lemma 4. *Let A be the Butcher-matrix of a SDIRK methods, c.f. Section 3.1.1, with diagonal entries $\gamma > 0$. Then*

$$\alpha_0(A^{-1}) = \frac{1}{\gamma}.$$

Proof. Similar to the proof in [31, Chap. IV.14, p. 236], we define the diagonal matrix D with entries $1, \varepsilon^2, \varepsilon^4, ..., \varepsilon^{2s-s}$, where s denotes the number of stages for the SDIRK method. With D, we obtain

$$D^{-1/2}A^{-1}D^{-1/2} + (D^{-1/2}A^{-1}D^{-1/2})^T = \mathrm{diag}(1/\gamma, ..., 1/\gamma) + \mathcal{O}(\varepsilon),$$

such that $\alpha_0(A^{-1}) \geq \frac{1}{\gamma} + \mathcal{O}(\varepsilon)$. We also obtain an upper bound by putting $\mathsf{u} = e_i$ with e_i being the i-th unit vector in (4.3) and assume that the diagonal entries of A^{-1} are $\frac{1}{\gamma}$. We therefore have an upper bound $\alpha_0(A^{-1}) \leq \frac{1}{\gamma}$. The lower bound for $\varepsilon \to 0$ and the defined upper bound lead to the statement above. □

We have the following existence Theorem for SDIRK methods.

Theorem 5 (Existence and Uniqueness [31])**.** *Let $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and satisfy condition* (4.2) *with a constant $\nu \in \mathbb{R}$. Let A be the invertible Butcher matrix of the SDIRK method and a time step $\Delta t > 0$ such that*

$$\Delta t \nu < \frac{1}{\gamma} \tag{4.4}$$

with the constant γ from Lemma 4. Then the SDIRK scheme applied to (4.1) *has a unique solution.*

Proof. Proof may be found in [31, Chapter IV.14]. □

Let us comment on this result. As discussed above, the SSP property can only be satisfied under certain time step restrictions. However, operators $\mathbf{F}$ stemming from a semi-discretization of, e.g., the water hammer equations can have a large Lipschitz constant and thus, we would obtain an additional restriction when using Theorem 5 with $\nu = L$. Therefore, we compute the one-sided Lipschitz constant of the spatial discretizations introduced in Section 3.2.

Note that the constant γ can be computed explicitly for the two-stage SDIRK method of Example 2.

Example 4 (Computation of upper bound (4.4))**.** *The inverse of the Butcher matrix in Example 2 is given by*

$$A^{-1} = \begin{pmatrix} 4 & 0 \\ -8 & 4 \end{pmatrix}.$$

From [31, Theorem 14.6], we have $\alpha_0(A^{-1}) = \frac{1}{\gamma}$ with $\gamma = \frac{1}{4}$ being the diagonal entry of the Butcher matrix. Consequently, we have $\Delta t \nu < 4$ in Theorem 5.

The one-sided Lipschitz constant (OSLC)

We compute the one-sided Lipschitz condition for operators $\mathbf{F}$ stemming from a semi-discretization with a FV or DG method. To show that no additional time step restriction is induced by Theorem 5, we show that $\nu \leq 0$ for specific cases.

We introduce an important tool to estimate the OSLC, i.e. the *logarithmic matrix norm*.

Definition 10 (Logarithmic matrix norm [79, Chap. 7, p. 196])**.** *Let $\|\cdot\|$ be an arbitrary vector norm in $\mathbb{R}^n$ and $|||\cdot|||$ the corresponding matrix norm. For $A \in \mathbb{R}^{n\times n}$, the limit*

$$\mu[A] = \lim_{\epsilon \to +0} \frac{|||I + \epsilon A||| - 1}{\epsilon} \tag{4.5}$$

is called logarithmic matrix norm of A.

Note that the limit (4.5) exists for every norm $\|\cdot\|$ and every $A \in \mathbb{R}^{n\times n}$, c.f. [79, Chap. 7]. Further, $\mu[A]$ is not actually a norm, since it can become negative.

In fact, if there exists a norm $\|\cdot\|$ such that $\mu[D\mathbf{F}] < 0$, then $\mathbf{F}$ satisfies condition (4.2) with $\nu \leq 0$.

Theorem 6 ([79, Chap. 7, Theorem 7.2.6])**.** *Let $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and let $\|\cdot\|$ be the norm induced by the scalar product $\langle\cdot,\cdot\rangle$. Then the OSLIP* (4.2) *holds for any ν that satisfies*

$$\mu[D\mathbf{F}] \leq \nu.$$

Proof. Proof may be found in [79, Chap. 7]. □

For special choices of the norm $\|\cdot\|$, we can compute the logarithmic matrix norm explicitly.

Lemma 5 ([79, Chap. 7, Theorem 7.2.4])**.** *Let $A \in \mathbb{R}^{n\times n}$ be an arbitrary matrix and*

$$\|A\|_\infty = \max_{i=1,..,n} \sum_{j=1}^{n} |a_{ij}|, \quad \|A\|_1 = \max_{j=1,..,n} \sum_{i=1}^{n} |a_{ij}|,$$
$$\text{and} \qquad \|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}.$$

For the corresponding logarithmic matrix norms, we get

$$\mu_\infty[A] = \max_{i=1,..,n} (a_{ii} + \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|), \quad \mu_1[A] = \max_{j=1,..,n} (a_{jj} + \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|),$$
$$\text{and} \qquad \mu_2[A] = \lambda_{\max}(\tfrac{1}{2}(A + A^T)).$$

Proof. A proof can also be found in [79, Chap. 7]. □

Note that $\|\cdot\|_2$ is the only norm used here which is induced by a scalar product. Since μ_1 and μ_∞ are much simpler to compute, we exploit the following relation:

Lemma 6 ([80]). *Let $A \in \mathbb{R}^{n\times n}$ be an arbitrary matrix and $\mu_2[A] < \infty$. Then*

$$\mu_2(A) \leq \tfrac{1}{2}\mu_\infty(A + A^T) \leq \tfrac{1}{2}(\mu_1(A) + \mu_\infty(A)). \tag{4.6}$$

In the following subsections, we compute the corresponding logarithmic norms for the mapping F resulting from different spatial discretizations.

The finite volume case

For the finite volume discretization of the water hammer equations, we have from (3.35)

$$\partial_t \mathrm{W} = -\tfrac{1}{\Delta x} L_h \mathrm{W} + G(\mathrm{W}) = F(\mathrm{W}). \tag{4.7}$$

We compute the Jacobian DF of the right-hand side F and have

$$\mu_\infty[DF] = \max\left\{ \frac{\mathbf{a}}{2\Delta x} + \max_{1,..,N} g'(\mathrm{Q}_j), \frac{\mathbf{a}^2}{\Delta x} \right\},$$

which grows for $\Delta x \to 0$. Also, $\mathbf{a}$ is usually a large constant, c.f. Section 2.1. Since we cannot choose $\nu = 0$ here, we utilize a transformation to characteristic variables, c.f. Section 2.3. Again, computing the gradient of $\tilde{L}_h$ and $\tilde{G}$ in (3.36), we get

$$\mu_1[DF] = \mu_\infty[DF] = \max_{j=1,..,N} -\tfrac{1}{2}\mathbf{g}'(\tilde{\mathrm{H}}_j + \tilde{\mathrm{Q}}_j).$$

Using Lemma 6, we have

$$\mu_2[DF] \leq \tfrac{1}{2}\left(\mu_\infty[DF] + \mu_1[DF]\right) = \max_{j=1,..,N} -\tfrac{1}{2}\mathbf{g}'(\tilde{\mathrm{H}}_j + \tilde{\mathrm{Q}}_j) \leq 0.$$

Consequently, we showed that the solution of an SDIRK scheme applied to the characteristic system has a solution for arbitrary step sizes, c.f. Theorem 5. By multiplication with $\mathbf{I}_N \otimes R$, we show the same result for the original system.

If we set $\nu = 0$, the inequality resulting from (4.2) is

$$\langle \tilde{F}(\tilde{W}_1) - \tilde{F}(\tilde{W}_2), \tilde{W}_1 - \tilde{W}_2 \rangle \leq 0.$$

This means that our system is dissipative.

A first-order DG scheme

For the DG discretization of the water hammer equations, we have from (3.40)

$$\partial_t M_h \mathrm{W} = K_h \mathrm{W} + G(\mathrm{W}). \tag{4.8}$$

We note that the inverse of the mass matrix (3.41) is of the form

$$M_h^{-1} = \frac{1}{\Delta x}\,\mathrm{diag}(1, 1, 3, 3, \ldots, 1, 1, 3, 3).$$

Consequently, we have a $1/\Delta x$ term appearing when computing the logarithmic matrix norms. Therefore, Lemma 6 is not applicable since the appearing terms are not bounded for $\Delta x \to 0$.

Let us introduce the norm [49]

$$\|x\|_{M_h^{1/2}} = \langle M_h^{1/2}x, M_h^{1/2}x\rangle = \langle x, x\rangle_{M_h^{1/2}} \text{ for all } x \in \mathbb{R}^N,$$

where $M_h^{1/2}$ is a symmetric positive definite matrix, i.e. $SM_h^{1/2}S^{-1} = \sqrt{\Lambda}$ with Λ a diagonal matrix with positive entries. The corresponding scalar product induces the logarithmic norm

$$\mu_{M_h^{1/2}}[A] = \max_{x\neq 0} \frac{\langle Ax, x\rangle_{M_h^{1/2}}}{\|x\|_{M_h^{1/2}}} \tag{4.9}$$

for some matrix A. Once again, we use the characteristic system and compute the Jacobian of

$$M_h^{-1}\Big(\tilde{K}_h \mathrm{W} + \tilde{G}(\mathrm{W})\Big)$$

and get

$$J = M_h^{-1}\underbrace{\Big(\mathbf{I}_{low,N}\otimes\tilde{C} + \mathbf{I}_N\otimes\tilde{B} + \mathbf{I}_{up,N}\otimes\tilde{D}\Big)}_{\tilde{\tilde{K}}_h} + M_h^{-1}(D\tilde{G}).$$

We have

$$D\tilde{G} = \frac{\Delta x}{4}\,\mathrm{diag}(\tilde{E}_j)$$

with

$$\tilde{E}_j = \begin{pmatrix} -(g'(M_j)+g'(m_j)) & -(g'(M_j)+g'(m_j)) & (g'(M_j)-g'(m_j)) & (g'(M_j)-g'(m_j)) \\ (g'(M_j)-g'(m_j)) & (g'(M_j)-g'(m_j)) & -(g'(M_j)+g'(m_j)) & -(g'(M_j)+g'(m_j)) \\ -(g'(M_j)+g'(m_j)) & -(g'(M_j)+g'(m_j)) & (g'(M_j)-g'(m_j)) & (g'(M_j)-g'(m_j)) \\ (g'(M_j)-g'(m_j)) & (g'(M_j)-g'(m_j)) & -(g'(M_j)+g'(m_j)) & -(g'(M_j)+g'(m_j)) \end{pmatrix}$$

for $j = 1, .., N$, where we have $M_j = \tilde{\mathrm{H}}_j^0 + \tilde{\mathrm{Q}}_j^0 - (\tilde{\mathrm{H}}_j^1 + \tilde{\mathrm{Q}}_j^1)$ and $m_j = \tilde{\mathrm{H}}_j^0 + \tilde{\mathrm{Q}}_j^0 + \tilde{\mathrm{H}}_j^1 + \tilde{\mathrm{Q}}_j^1$. Plugging $M_h^{-1}\tilde{K}_h$ into the specially defined μ-Norm (4.9), we get

$$\begin{aligned}\mu_{M_h^{1/2}}[M_h^{-1}(\tilde{K}_h + D\tilde{G})] &= \max_{x\neq 0}\frac{(M_h^{1/2}x)^T M_h^{1/2}M_h^{-1}(\tilde{K}_h + D\tilde{G})x}{\|x\|_{M_h^{1/2}}} \\ &= \max_{x\neq 0}\frac{x^T M_h^{T/2}M_h^{1/2}M_h^{-1}(\tilde{K}_h + D\tilde{G})x}{\|x\|_{M_h^{1/2}}} = \max_{x\neq 0}\frac{x^T(\tilde{K}_h + D\tilde{G})x}{\|x\|_{M_h^{1/2}}},\end{aligned}$$

where $M_h^{T/2} = M_h^{1/2}$ because of the diagonal structure of the mass matrix. It remains to be shown that $\tilde{K}_h + D\tilde{G}$ is negative semi-definite. Clearly, $\tilde{K}_h$ is negative semi-definite since it contains blocks of negative semi-definite matrices $\tilde{C} + \tilde{B} + \tilde{D}$ with eigenvalues 0

and $-2\mathbf{a}$. The matrix $D\tilde{G}$ again has block structure and the blocks are negative semi-definite. This can be shown by computing the logarithmic norms μ_1 and μ_∞, which are $-\max_j \mathsf{g}'(m_j)$ with m_j defined above. Using Lemma 6 results in

$$\mu_2(M_h^{-1}(D\tilde{G}) \leq 1/2(-2\max_j \mathsf{g}'(m_j)) \leq -\max_j \mathsf{g}'(m_j).$$

Using the fact that the sum of two negative semi-definite matrices is still negative semi-definite, the Jacobian of the right-hand side is negative semi-definite as well. Consequently, we showed that the solution of an SDIRK scheme applied to the characteristic system has a solution for arbitrary step sizes, c.f. Theorem 5. By multiplication with $\mathbf{I}_N \otimes R$, we can show the same result for the original system. By setting $\nu = 0$, we obtain from (4.2) the inequality

$$\langle \tilde{F}(\tilde{W}_1) - \tilde{F}(\tilde{W}_2), \tilde{W}_1 - \tilde{W}_2 \rangle \leq 0.$$

Again, this means that our system is dissipative.

4.2 Well-balancedness of different space discretizations according to WHE

We established the existence and uniqueness of solutions to the proposed methods. Another important property for such schemes is their *well-balancedness*.

Definition 11 (Well-balancedness)**.** *We call a method well-balanced with respect to, e.g., the water hammer equations if the numerical solution satisfies any stationary state of the corresponding partial differential equations exactly, i.e.*

$$\mathbf{w}_h(t^n) = \mathbf{w}_h(t^{n-1})$$

if $\mathbf{w}_h(t^{n-1})$ *is the stationary state.*

Therefore, to show the well-balancedness of the schemes, we compute the discrete stationary state, and show that it is a constant solution of the discrete scheme. Since the solution is unique, the method will maintain the stationary state and is thus well-balanced with respect to the water hammer equations (2.10).

The methods of interest here are combinations of SSP SDIRK methods with several spatial discretizations. Since the RK methods we use fulfill the consistency requirement (3.6), the solution will be constant in time if

$$\partial_t \mathrm{W} = 0$$

for the semi-discrete scheme. This means that, to show well-balancedness, we merely need to show that the right-hand side of the semi-discrete schemes (4.7) and (4.8) become zero, if the stationary state $(\bar{\mathrm{H}}, \bar{\mathrm{Q}})$ is inserted.

Note that we cannot imply periodic boundary conditions here, since the stationary state is not periodic if $\bar{\mathrm{Q}} \neq 0$. Instead, we use *inflow and outflow boundary conditions* with

suitable boundary data. These computations will be demonstrated in the following sections.

4.2.1 The finite volume case

Concerning the discrete stationary state, recall that $\mathrm{Q}_j = \bar{\mathrm{Q}}$ (2.19) and

$$\mathrm{H}_{j+1} - \mathrm{H}_j = -\Delta x \mathbf{g}(\bar{\mathrm{Q}}),$$

c.f. (2.20) for all j. Inserting the stationary state in (3.35), we have the components of $L_h\mathrm{W} + G(\mathrm{W})$ given as

$$-\frac{1}{\Delta x}\left[\frac{1}{2}\begin{pmatrix}\mathbf{a}^2(\bar{\mathrm{Q}}+\bar{\mathrm{Q}})\\ \mathrm{H}_{j+1}+\mathrm{H}_j\end{pmatrix} - \frac{\mathbf{a}}{2}\begin{pmatrix}-\Delta x\mathbf{g}(\bar{\mathrm{Q}})\\ \bar{\mathrm{Q}}-\bar{\mathrm{Q}}\end{pmatrix} - \frac{1}{2}\begin{pmatrix}\mathbf{a}^2(\bar{\mathrm{Q}}+\bar{\mathrm{Q}})\\ \mathrm{H}_j+\mathrm{H}_{j-1}\end{pmatrix} + \frac{\mathbf{a}}{2}\begin{pmatrix}-\Delta x\mathbf{g}(\bar{\mathrm{Q}})\\ \bar{\mathrm{Q}}-\bar{\mathrm{Q}}\end{pmatrix}\right] + \begin{pmatrix}0\\ -\mathbf{g}(\mathrm{Q}_j)\end{pmatrix}$$

for all j. As already mentioned above, we use in- and outflow boundaries, see also Section 2.2.2. In order to do this, we need to introduce so-called *ghosts cells* on the left and right boundary of the domain. For lower-order finite volume schemes we need two ghosts cells, I_0 and I_{N+1} with the values Q_0, H_0, Q_{N+1} and H_{N+1}. For the values H_0 and Q_{N+1} we set fixed constants. Using first-order extrapolation yields the remaining values

$$Q_0 = 2Q_1 - Q_2 \text{ and } H_{N+1} = 2H_N - H_{N-1}.$$

Remark 6. *Note that first-order extrapolation are not easy to handle since they can lead to stability problems [51, Chapter 3.11 and 7], e.g. oscillations can appear.*

We thus obtain

$$\partial_t \mathrm{W} = 0 \tag{4.10}$$

if we insert the stationary state into the right-hand side. Due to the consistency requirement (3.6), the SDIRK method computes a constant solution.

Note that for the zero-order DG method the result is equivalent.

Remark 7. *In the case of the finite volume method with flux limiting, as introduced in Section 3.4 and Remark 5, we also have that the scheme in combination with the SDIRK scheme is well-balanced. Together with the existence and uniqueness of a solution for the finite volume case, which we showed earlier in this section, and by inserting the stationary state, we get $\partial_t \mathrm{W} = 0$, which verifies the statement. Note that we also use inflow and outflow boundaries similar as mentioned above.*

4.2.2 First-order DG method

For the DG case, we again utilize (2.19) resulting in $\mathrm{Q}_j^0 = \bar{\mathrm{Q}}$, $\mathrm{Q}_j^1 = 0$, which means that the cell-average values are constant and the slopes of the linear function Q_j^1 are zero for all j. Again using (2.20) for the pressure head H, the slopes are determined by $-\mathtt{g}(\bar{\mathrm{Q}})$. This can be expressed as

$$\begin{aligned} \mathrm{H}_j^0 + \mathrm{H}_j^1 - (\mathrm{H}_j^0 - \mathrm{H}_j^1) &= -\Delta x \mathtt{g}(\bar{\mathrm{Q}}) \\ \mathrm{H}_j^0 - \mathrm{H}_j^1 - (\mathrm{H}_{j-1}^0 + \mathrm{H}_{j-1}^1) &= 0 \end{aligned}$$

for all j or equivalently as

$$\mathrm{H}_j^1 = -\tfrac{\Delta x}{2}\mathtt{g}(\bar{\mathrm{Q}}) \tag{4.11}$$

$$\mathrm{H}_j^0 - \mathrm{H}_{j-1}^0 = -\Delta x \mathtt{g}(\bar{\mathrm{Q}}). \tag{4.12}$$

Using this, we have for the right-hand side of (3.40)

$$\begin{aligned} &\tfrac{\mathtt{a}^2}{2}(\bar{\mathrm{Q}} + 0 - 2\cdot 0 - \bar{\mathrm{Q}} + 0) + \tfrac{\mathtt{a}}{2}(\mathrm{H}_{n-1}^0 + \mathrm{H}_{n-1}^1 - 2\mathrm{H}_n^0 + \mathrm{H}_{n+1}^0 - \mathrm{H}_{n+1}^1) \\ &\tfrac{1}{2}(\mathrm{H}_{n-1}^0 + \mathrm{H}_{n-1}^1 - 2\mathrm{H}_n^1 - \mathrm{H}_{n+1}^0 + \mathrm{H}_{n+1}^1) + \tfrac{\mathtt{a}}{2}(\bar{\mathrm{Q}} + 0 - 2\bar{\mathrm{Q}} + \bar{\mathrm{Q}} - 0) \\ &- \tfrac{\Delta x}{2}(\mathtt{g}(\bar{\mathrm{Q}} - 0) + \mathtt{g}(\bar{\mathrm{Q}} + 0)) \end{aligned} \tag{4.13}$$

for the first component and

$$\begin{aligned} &-\tfrac{\mathtt{a}^2}{2}(\bar{\mathrm{Q}} + 0 - 2\bar{\mathrm{Q}} + \bar{\mathrm{Q}} - 0) + \tfrac{\mathtt{a}}{2}(\mathrm{H}_{n+1}^0 - \mathrm{H}_{n+1}^1 - 2\mathrm{H}_n^1 - \mathrm{H}_{n-1}^0 - \mathrm{H}_{n-1}^1) \\ &-\tfrac{1}{2}(\mathrm{H}_{n-1}^0 + \mathrm{H}_{n-1}^1 - 2\mathrm{H}_n^0 + \mathrm{H}_{n+1}^0 - \mathrm{H}_{n+1}^1) + \tfrac{\mathtt{a}}{2}(\bar{\mathrm{Q}} - 0 - 2\cdot 0 - \bar{\mathrm{Q}} - 0) \\ &+ \tfrac{\Delta x}{2}(\mathtt{g}(\bar{\mathrm{Q}} - 0) - \mathtt{g}(\bar{\mathrm{Q}} + 0)) \end{aligned} \tag{4.14}$$

for the second component, where we again imply inflow and outflow boundary conditions. Here, we again need to use one ghosts cell at each boundary. We call the ghosts cell I_0 and I_{N+1} with the cell values Q_0^0, H_0^0, Q_0^1, H_0^1, Q_{N+1}^0, H_{N+1}^0, Q_{N+1}^1 and H_{N+1}^1. Again, we set fixed constants for H_0^0, H_0^1, Q_{N+1}^0 and Q_{N+1}^1. It reads

$$\begin{aligned} H_0^1 = -\tfrac{\Delta x}{2}\mathtt{g}(Q_1^0) \text{ and } H_0^0 = 1 + \tfrac{\Delta x}{2}\mathtt{g}(Q_1^0) \\ Q_{N+1}^1 = 0 \text{ and } Q_{N+1}^0 = \bar{Q}. \end{aligned}$$

Remark 8. *In analogous to the finite volume case, the first-order extrapolation of the boundary DG values can also lead to stability problems, e.g. oscillations can appear.*

The remaining values are determined with linear extrapolation,

$$\begin{aligned} H_{N+1}^0 = 2H_N^0 - H_{N-1}^0 \text{ and } H_{N+1}^1 = 2H_N^1 - H_{N-1}^1, \\ Q_0^1 = 2Q_1^1 - Q_2^1 \text{ and } Q_0^0 = 2Q_1^0 - Q_2^0. \end{aligned}$$

Using equations (4.11) and (4.12) in (4.13) and (4.14) for all inner points and for the boundary points of the introduced boundary conditions, we get

$$\partial_t \mathrm{W} = 0$$

if we insert the stationary state into the right-hand side.

Note that we can employ similar arguments for higher-order DG schemes to show that the method is well-balanced.

4.3 Satisfaction of the maximum-principle for the fully discrete scheme

In this section we discuss the discrete maximum principle. Note that entropy solutions of scalar conservation laws fulfill a strict maximum principle [88]. This means that any solution u lies in the interval $[m, M]$, where

$$m = \min_x \mathsf{u}_0(x) \quad \text{and} \quad M = \max_x \mathsf{u}_0(x) \tag{4.15}$$

are the maximum and the minimum of the initial state u_0 and $m, M \in \mathbb{R}$.

We want to show that the discrete solution of the introduced methods also satisfies a (discrete) maximum principle, which means that the numerical solution stays in the range $[m, M]$. To proof this point, we provide a theorem which allows us to reduce the statement to an explicit Euler step. We study scalar balance laws first and subsequently extend the results to linear systems with and without source term.

4.3.1 Connection to explicit Euler

Let us return to the semi-discrete ODE system (3.2) discussed in Section 3.1. The function $\mathbf{F}(\mathsf{u})$ represents the spatial discretization and is assumed to be Lipschitz continuous. We show that if a maximum principle is satisfied for the explicit Euler method, this automatically carries over to SSP SDIRK methods. It is important to notice that the SSP property of the explicit Euler condition (3.19) is only fulfilled for the maximum norm $\|\cdot\|_\infty$ here.

Theorem 7. *Assume that the explicit Euler scheme for all time step sizes $\Delta t \leq \Delta t_{EE}$ applied to* (3.2) *satisfies a discrete maximum principle, i.e. $\mathsf{u}^{n+1} \in [m, M]$ component-wise if $\mathsf{u}^n \in [m, M]$ component-wise for some constants m, M. Then the iterates of an SSP SDIRK method for all time step sizes $\Delta t \leq C_{SSP}\Delta t_{EE}$, where C_{SSP} is the SSP coefficient, also fulfills a discrete maximum principle when applied to* (3.2).

Proof. Any SSP SDIRK method can be written in the modified Shu-Osher form

$$\mathsf{u}^{(i)} = v_i \mathsf{u}^n + \sum_{j=1}^{i} \alpha_{ij} \left(\mathsf{u}^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} \mathbf{F}(\mathsf{u}^{(j)}) \right), \quad 1 \leq i \leq m \tag{4.16a}$$

$$\mathsf{u}^{n+1} = \mathsf{u}^{(m+1)} = v_{m+1} \mathsf{u}^n + \sum_{j=1}^{m} \alpha_{m+1,j} \left(\mathsf{u}^{(j)} + \Delta t \frac{\beta_{m+1,j}}{\alpha_{m+1,j}} \mathbf{F}(\mathsf{u}^{(j)}) \right). \tag{4.16b}$$

From (4.16b) and the consistency condition $v_{m+1} + \sum_{j=1}^{m} \alpha_{m+1,j} = 1$, we see that u^{n+1} is a convex combination of u^n and the explicit Euler steps applied to $\mathsf{u}^{(j)}$ for all j. Consequently, $\mathsf{u}^{n+1} \in [m, M]$ if the terms in brackets in (4.16b) lie in the range $[m, M]$. Due to the stepsize restriction $\Delta t \leq C_{SSP}\Delta t_{EE}$ and the range condition from the explicit Euler step, it suffices to show that $\mathsf{u}^{(j)} \in [m, M]$ for all j. We show that $\mathsf{u}^{(1)} \in [m, M]$. From the modified Shu-Osher form, we have that

$$\mathsf{u}^{(1)} = v_1 \mathsf{u}^n + \alpha_{11} \left(\mathsf{u}^{(1)} + \Delta t \frac{\beta_{11}}{\alpha_{11}} F\left(\mathsf{u}^{(1)}\right) \right). \tag{4.17}$$

The SSP property for the explicit Euler step yields

$$\begin{aligned}\left\|\mathsf{u}^{(1)}\right\|_\infty &= \left\|v_1\mathsf{u}^n + \alpha_{11}\left(\mathsf{u}^{(1)} + \Delta t\frac{\beta_{11}}{\alpha_{11}}F\left(\mathsf{u}^{(1)}\right)\right)\right\|_\infty \\ &\leq v_1\left\|\mathsf{u}^n\right\|_\infty + \alpha_{11}\left\|\mathsf{u}^{(1)} + \Delta t\frac{\beta_{11}}{\alpha_{11}}F\left(\mathsf{u}^{(1)}\right)\right\|_\infty \\ &\underbrace{\leq}_{\text{SSP Cond.}} v_1\left\|\mathsf{u}^n\right\|_\infty + \alpha_{11}\left\|\mathsf{u}^{(1)}\right\|_\infty.\end{aligned}$$

Note that we cannot use any convex functional here, but we need to use the maximum norm. From the consistency requirement $v_1 + \alpha_{11} = 1$, it follows that $\left\|\mathsf{u}^{(1)}\right\|_\infty \leq \left\|\mathsf{u}^n\right\|_\infty$. Let us now assume that $m, M > 0$ and $|m| < |M|$. All other cases can be shown completely analogously. From the previous estimate, we already have that $\mathsf{u}^{(1)} \in [-M, M]$. It remains to show that $\mathsf{u}^{(1)} > m$. From the maximum principle for the explicit Euler step, we deduce

$$\min_i \mathsf{u}_i^{(1)} \leq \min_i \left(\mathsf{u}_i^{(1)} + \Delta t\frac{\beta_{11}}{\alpha_{11}}F\left(\mathsf{u}^{(1)}\right)_i\right). \tag{4.18}$$

Let $\mathsf{u}_q^{(1)}$ be the minimal component of the vector $\mathsf{u}^{(1)}$, i.e. $\operatorname{argmin}_i \mathsf{u}_i^{(1)} =: \mathsf{u}_q^{(1)}$. The minimum exists, since we consider numerical methods with a finite number of sampling points. We can then make the following approximation

$$\begin{aligned}\mathsf{u}_q^{(1)} &= v_1\mathsf{u}_q^n + \alpha_{11}\left(\mathsf{u}_q^{(1)} + \Delta t\frac{\beta_{11}}{\alpha_{11}}F\left(\mathsf{u}^{(1)}\right)\right) \\ &\geq v_1\mathsf{u}_q^n + \alpha_{11}\min_i\left(\mathsf{u}_i^{(1)} + \Delta t\frac{\beta_{11}}{\alpha_{11}}F\left(\mathsf{u}^{(1)}\right)_i\right) \\ &\underbrace{\geq}_{\text{Cond}(4.18)} v_1\mathsf{u}_q^n + \alpha_{11}\mathsf{u}_q^{(1)}.\end{aligned}$$

This implies that $\mathsf{u}_q^{(1)} \geq \mathsf{u}_q^n \geq \min_i \mathsf{u}_i^n =: m$ and therefore $\mathsf{u}^{(1)} \geq m$ component-wise. Note that from the maximum principle for the explicit Euler step we have that

$$\max_i \mathsf{u}_i^{(1)} \geq \max_i \left(\mathsf{u}_i^{(1)} + \Delta t\frac{\beta_{11}}{\alpha_{11}}F\left(\mathsf{u}^{(1)}\right)_i\right), \tag{4.19}$$

which can be utilized in a similar fashion if $|m| > |M|$. For the remaining stages $\mathsf{u}^{(i)}$, we have

$$\mathsf{u}^{(i)} = v_i\mathsf{u}^n + \sum_{j=1}^{i-1}\alpha_{ij}\left(\mathsf{u}^{(j)} + \Delta t\frac{\beta_{ij}}{\alpha_{ij}}F(\mathsf{u}^{(j)})\right) + \alpha_{ii}\left(\mathsf{u}^{(i)} + \Delta t\frac{\beta_{ii}}{\alpha_{ii}}F(\mathsf{u}^{(i)})\right).$$

We now use the exact same argument since we already know that $v_i\mathsf{u}^n + \sum_{j=1}^{i-1}\alpha_{ij}\left(\mathsf{u}^{(j)} + \Delta t\frac{\beta_{ij}}{\alpha_{ij}}F(\mathsf{u}^{(j)})\right)$ lie in the range $[m, M]$.

□

This theorem enables us to prove that our methods satisfy a discrete maximum principle by studying explicit Euler time stepping. For this reason, we discuss the combination of the introduced spatial discretizations with the explicit Euler method in the following sections.

Concerning scalar conservation laws, this has been shown for explicit SSP Runge-Kutta methods in [89]. Note that since the source term g satisfies $\mathsf{g}(0) = 0$ and $\mathsf{g}' \leq 0$, we treat the range $[-m, M]$ where $m, M \geq 0$.

4.3.2 The scalar case

We begin with the discussion of the numerical solution of the scalar balance law

$$\partial_t \mathsf{u} + \partial_x \mathsf{f}(\mathsf{u}) = \mathsf{g}(\mathsf{u}) \tag{4.20}$$

with the initial condition $\mathsf{u}(x, 0) = \mathsf{u}_0(x)$. For the source term, we assume $\mathsf{g}(0) = 0$ and $\mathsf{g}' \leq 0$. Without loss of generality, we assume that u_0 lies in the range $[-m, M]$ component-wise for some constants $m, M \geq 0$, where $-m = \min\{0, \min_x \mathsf{u}_0(x)\}$ and M defined as in (4.15). We consider the combination of the explicit Euler method with several spatial discretizations.

Finite Volume schemes with explicit Euler method

As introduced in Section 3.2.1, a three-point finite volume scheme for the scalar balance law (4.20) combined with the explicit Euler method has the form

$$\mathsf{u}_j^{n+1} = \mathsf{u}_j^n - \tfrac{\Delta t}{\Delta x}\left[\hat{\mathbf{F}}(\mathsf{u}_j^n, \mathsf{u}_{j+1}^n) - \hat{\mathbf{F}}(\mathsf{u}_{j-1}^n, \mathsf{u}_j^n)\right] + \Delta t \mathsf{g}(\mathsf{u}_j^n) =: H_\lambda(\mathsf{u}_{j+1}^n, \mathsf{u}_j^n, \mathsf{u}_{j-1}^n), \tag{4.21}$$

with $\lambda = \frac{\Delta t}{\Delta x}$ and $\hat{\mathbf{F}}$ being the monotone local Lax-Friedrichs flux (3.34) in this case.

Note that the consistent local Lax-Friedrichs flux is continuous in both arguments, non-decreasing in its first argument, and non-increasing in its second argument, c.f. Definition 3. We assume that the step sizes Δt and Δx satisfy the CFL condition

$$\lambda\omega + \Delta t L \leq 1, \tag{4.22}$$

where L is the Lipschitz constant of the source term g and $\omega = \max_u |f'(u)|$. With this equation, we can show that the function $H_\lambda(\mathsf{u}_{j+1}^n, \mathsf{u}_j^n, \mathsf{u}_{j-1}^n)$ is increasing in all three arguments.

Recall that

$$H_\lambda(\mathsf{u}_{j+1}^n, \mathsf{u}_j^n, \mathsf{u}_{j-1}^n) = \mathsf{u}_j^n - \lambda\left[\tfrac{1}{2}\left(\mathsf{f}(\mathsf{u}_{j+1}^n) - \mathsf{f}(\mathsf{u}_{j-1}^n)\right) - \tfrac{\omega}{2}\left(\mathsf{u}_{j+1}^n - 2\mathsf{u}_j^n + \mathsf{u}_{j-1}^n\right)\right] + \Delta t \mathsf{g}(\mathsf{u}_j^n),$$

and thus, we have for its derivatives

$$\begin{aligned}\frac{\partial H_\lambda}{\partial \mathsf{u}_{j-1}^n} &= \frac{\lambda}{2}\left(\mathsf{f}'(\mathsf{u}_{j-1}^n) + \omega\right) \geq 0,\\ \frac{\partial H_\lambda}{\partial \mathsf{u}_{j+1}^n} &= \frac{\lambda}{2}\left(\omega - \mathsf{f}'(\mathsf{u}_{j+1}^n)\right) \geq 0,\\ \frac{\partial H_\lambda}{\partial \mathsf{u}_{j}^n} &= 1 - \lambda\omega + \Delta t \mathsf{g}'(\mathsf{u}_j^n) \geq 0,\end{aligned}$$

where we used the definition of ω for the first two assertions. To show the third inequality, we observe

$$0 \leq \lambda\omega - \Delta t \mathsf{g}'(\mathsf{u}_j^n) \leq \lambda\omega + \Delta t L \leq 1$$

from the CFL condition (4.22).

Together with the CFL condition (4.22), the assumptions on the source term $g(0) = 0$ and $g'(\cdot) \leq 0$, and the monotonicity of H_λ, the scheme fulfills the strict maximum principle, since

$$-m \leq H_\lambda(-m,-m,-m) \leq H_\lambda(\mathsf{u}_{j+1}^n, \mathsf{u}_j^n, \mathsf{u}_{j-1}^n) = \mathsf{u}_j^{n+1} \leq H_\lambda(M,M,M) \leq M.$$

Higher-order DG schemes with explicit Euler method

Let us turn to the spatial discretization using a DG method with higher-order polynomial ansatz functions. We consider piecewise linear polynomials as a first step and establish a more general result afterwards. The maximum principle is valid for the mean values on each control cell. When applying a DG method in combination with the explicit Euler method to (4.20), then the mean values fulfill the following relation

$$\bar{\mathsf{u}}_j^{n+1} = K_\lambda(\bar{\mathsf{u}}_j^n, \mathsf{u}_{j+1/2}^{-,n}, \mathsf{u}_{j+1/2}^{+,n}, \mathsf{u}_{j-1/2}^{-,n}, \mathsf{u}_{j-1/2}^{+,n}),$$

where

$$K_\lambda := \bar{\mathsf{u}}_j^n - \lambda\left[\hat{\mathbf{F}}(\mathsf{u}_{j+1/2}^{-,n}, \mathsf{u}_{j+1/2}^{+,n}) - \hat{\mathbf{F}}(\mathsf{u}_{j-1/2}^{-,n}, \mathsf{u}_{j-1/2}^{+,n})\right] + \Delta t \mathsf{g}(\bar{\mathsf{u}}_j^n). \tag{4.23}$$

Unfortunately, the term K_λ does not have the same structure as H_λ from the lower-order finite volume case. In fact, the assumption that all arguments of K_λ lie in the desired range is no sufficient condition for the fact that $\bar{\mathsf{u}}_j^{n+1}$ lies in the same range [89]. However, we can utilize that the discrete solution is a piecewise linear polynomial. Thus, the mean value $\bar{\mathsf{u}}_j^n$ can be exactly computed using the two-point Legendre Gauss-Lobatto quadrature rule for the interval $\mathbf{I}_j$,

$$\bar{\mathsf{u}}_j^n = \frac{1}{\Delta x}\int_{\mathbf{I}_j} \mathsf{u}_h^n(x)\,dx = \hat{w}_1 \mathsf{u}_h^n(\hat{x}_j^1) + \hat{w}_2 \mathsf{u}_h^n(\hat{x}_j^2) = \hat{w}_1 \mathsf{u}_{j-1/2}^{+,n} + \hat{w}_2 \mathsf{u}_{j+1/2}^{-,n}, \tag{4.24}$$

where the quadrature points are $x_{j-1/2}$ and $x_{j+1/2}$ and we have for the quadrature weights $\hat{w}_1 = \hat{w}_2 = \frac{1}{2}$. Using this decomposition, we show the following Theorem.

Theorem 8. *Assume that* $\bar{\mathsf{u}}_j^n$ *and* $\mathsf{u}_{j+1/2}^{-,n}, \mathsf{u}_{j+1/2}^{+,n}, \mathsf{u}_{j-1/2}^{-,n}, \mathsf{u}_{j-1/2}^{+,n}$ *lie in the range* $[-m, M]$ *with* $m, M \geq 0$ *and that relation* (4.24) *holds true. Further, assume that the CFL condition*

$$\lambda\omega + \Delta t L \leq 1/2 \tag{4.25}$$

is valid. Then, we have $\bar{\mathsf{u}}_j^{n+1} \in [-m, M]$ *in* (4.23).

Proof. Inserting equation (4.24) into equation (4.23), we get

$$\begin{aligned}\bar{\mathsf{u}}_j^{n+1} =& \frac{1}{2}\left(u_{j-1/2}^{+,n} + u_{j+1/2}^{-,n}\right) - \lambda\left[\hat{F}(u_{j+1/2}^{-,n}, u_{j+1/2}^{+,n}) - \hat{F}(u_{j-1/2}^{-,n}, u_{j-1/2}^{+,n})\right] \\ &+ \Delta t g\left(\frac{1}{2}(u_{j-1/2}^{+,n} + u_{j+1/2}^{-,n})\right).\end{aligned}$$

By adding and subtracting $\hat{F}(u_{j-1/2}^{+,n}, u_{j+1/2}^{-,n})$, we have

$$\begin{aligned}\bar{\mathsf{u}}_j^{n+1} &= \frac{1}{2}\left(u_{j-1/2}^{+,n} - 2\lambda(\hat{F}(u_{j+1/2}^{-,n}, u_{j+1/2}^{+,n}) - \hat{F}(u_{j-1/2}^{-,n}, u_{j+1/2}^{-,n}))\right) \\ &+ \frac{1}{2}\left(u_{j+1/2}^{-,n} - 2\lambda(\hat{F}(u_{j-1/2}^{+,n}, u_{j+1/2}^{-,n}) - \hat{F}(u_{j-1/2}^{-,n}, u_{j-1/2}^{+,n}))\right) + \Delta t g\left(\frac{1}{2}\left(u_{j-1/2}^{+,n} + u_{j+1/2}^{-,n}\right)\right) \\ &= \frac{1}{2}\left(H_{\lambda/2}(u_{j-1/2}^{-,n}, u_{j-1/2}^{+,n}, u_{j+1/2}^{-,n}) + H_{\lambda/2}(u_{j-1/2}^{+,n}, u_{j+1/2}^{-,n}, u_{j+1/2}^{+,n})\right) \\ &\quad + \Delta t g\left(\frac{1}{2}\left(u_{j-1/2}^{+,n} + u_{j+1/2}^{-,n}\right)\right),\end{aligned} \tag{4.26}$$

where $H_{\lambda/2}$ is defined as in (4.21) with $g = 0$. Utilizing the same arguments as in the finite volume case, the assertion follows. □

For ansatz functions using higher-order polynomials we can proceed in a similar fashion. The average value $\bar{\mathsf{u}}_j^{n+1}$ can again be exactly computed using the quadrature rule

$$\bar{u}_j^n = \frac{1}{\Delta x}\int_{I_j} u_h(x)dx = \sum_{k=1}^{N} \hat{\omega}_k u_h(x_j^k) = \sum_{k=2}^{N-1} \hat{\omega}_k u_h(x_j^k) + \hat{\omega}_1 u_{j-1/2}^{+,n} + \hat{\omega}_N u_{j+1/2}^{-,n} \tag{4.27}$$

with the quadrature points $x_j^k \in S_j = \{x_{j-1/2} = x_j^1, x_j^2, ..., x_j^{N-1}, x_j^N = x_{j+1/2}\}$ and the quadrature weights $\hat{\omega}_k$ with $\sum_k \hat{\omega}_k = 1$.

Theorem 9 ([89])**.** *Let the assumptions of Theorem 8 hold true where we assume relation* (4.27) *instead of* (4.24)*. Assume additionally that the values* $u_h(\hat{x}_j^b)$ *for all quadrature points* $\hat{x}_j^b$ *lie in the range* $[-m, M]$ *with* $m, M \geq 0$ *and the modified CFL condition*

$$\lambda\omega + \hat{\omega}_1 \Delta t L \leq \hat{\omega}_1.$$

Then we have $\bar{\mathsf{u}}_j^{n+1} \in [-m, M]$ *in* (4.23).

Proof. The proof is analogous to the proof of Theorem 8, see [89]. □

To apply Theorem 8, we need to verify that the polynomial u_h evaluated at the quadrature points lies in the range $[-m, M]$ with $m, M \geq 0$. Especially for higher-order polynomials this is not automatically the case. Therefore we introduce a modification which ensures this condition and maintains the higher-order accuracy of the approximation.

Let $\tilde{\mathsf{u}}_h$ denote the modified polynomial which is generated using the TVD MinMod limiter (3.53)-(3.55). Then we have the following.

Lemma 7. *Assume for the linear polynomial u_h that $\bar{\mathsf{u}}_j^n$ lies in the range $[m, M]$ with $m, M \in \mathbb{R}$ for all j. Then $\tilde{\mathsf{u}}_h$ is a linear polynomial and $\tilde{\mathsf{u}}_h(x) \in [m, M]$ for all $x \in \mathbf{I}_j$, $j = 1, \ldots, N$.*

Proof. The proof can be found in Appendix A. □

Besides the well-known order-reduction at local extrema [10, Section 3.13], the limiter maintains the second-order accuracy.

Lemma 8. *Assuming $\bar{u}_j^n \in [m, M]$ with $m, M \in \mathbb{R}$, the Equations* (3.54) *and* (3.55) *give a second-order accurate limiter in regions where the numerical solution is monotone.*

Proof. Since the original scheme is second-order accurate, we need to show that the difference of the two polynomials

$$u_h(x) = \bar{\mathsf{u}}_j^n + \frac{u_h(x_{j+1/2}^{-,n}) - u_h(x_{j-1/2}^{+,n})}{\Delta x}(x - x_M), \text{ and}$$
$$\tilde{u}_h(x) = \bar{\mathsf{u}}_j^n + \xi \left(\frac{u_h(x_{j+1/2}^{-,n}) - u_h(x_{j-1/2}^{+,n})}{\Delta x}, \frac{\bar{\mathsf{u}}_{j+1}^n - \bar{\mathsf{u}}_j^n}{\Delta x}, \frac{\bar{\mathsf{u}}_j^n - \bar{\mathsf{u}}_{j-1}^n}{\Delta x} \right) (x - x_M),$$

is of the order $O(\Delta x^2)$. By assumption and the definition of ξ, see (3.54) and (3.55), the case that $\xi = 0$ can not occur since the numerical solution is monotone. Second, for the case

$$\xi = \frac{u_h(x_{j+1/2}^{-,n}) - u_h(x_{j-1/2}^{+,n})}{\Delta x},$$

we have $u_h - \tilde{u}_h = 0$. As discussed in Section 3.4, the limiter becomes only active in the case of an over- or undershoot. Here, we consider the case

$$\xi = \frac{\bar{\mathsf{u}}_{j+1}^n - \bar{\mathsf{u}}_j^n}{\Delta x}.$$

We assume to have an overshoot and no undershoot, which means $M_j > M$ and $m_j \geq m$, where $M_j = u_h(x^{-,n}_{j+1/2})$ and $m_j = u_h(x^{+,n}_{j-1/2})$ denotes the maximal and minimal value in the cell I_j. All other cases can be treated analogously. We obtain

$$\begin{aligned}\tilde{u}_h(x) - u_h(x) &= \frac{x - x_M}{\Delta x}(\bar{\mathsf{u}}^n_{j+1} - \bar{\mathsf{u}}^n_j - u_h(x^{-,n}_{j+1/2}) + u_h(x^{+,n}_{j-1/2})) \\ &= \frac{x - x_M}{\Delta x}(\bar{\mathsf{u}}^n_{j+1} - \bar{\mathsf{u}}^n_j - M_j + m_j) \\ &\underset{\bar{\mathsf{u}}^n_{j+1},\bar{\mathsf{u}}^n_j \in [m,M], M_j > M, m \leq m_j}{\leq} \underbrace{\frac{x - x_m}{\Delta x}}_{\leq \frac{1}{2}} (\underbrace{M - M_j}_{O(\Delta x^2)} + \underbrace{m_j - m}_{O(\Delta x^2)}) \leq O(\Delta x^2),\end{aligned}$$

where the last two differences are of order $O(\Delta x^2)$ since the scheme is second-order accurate. □

With the modified linear polynomials, we obtain the modified scheme

$$\bar{\mathsf{u}}^{n+1}_j = \bar{\mathsf{u}}^n_j - \lambda \left[\hat{F}(\tilde{u}^{-,n}_{j+1/2}, \tilde{u}^{+,n}_{j+1/2}) - \hat{F}(\tilde{u}^{-,n}_{j-1/2}, \tilde{u}^{+,n}_{j-1/2})\right] + \Delta t g(\bar{\mathsf{u}}^n_j), \tag{4.28}$$

where $\tilde{u}^{-,n}_{j+1/2} = \tilde{u}_{h_{|I_j}}(x_{j+1/2})$ and $\tilde{u}^{+,n}_{j-1/2} = \tilde{u}_{h_{|I_j}}(x_{j-1/2})$.

To extend the previous results to higher-order polynomials, we again need to make sure that the values of the modified polynomial lie in the range $[m, M]$ for all quadrature points. This can be achieved using the linear scaling limiter introduced by Liu and Osher [53], which reads

$$\tilde{\xi} = \min\left\{\left|\frac{M - \bar{\mathsf{u}}^n_j}{M_j - \bar{\mathsf{u}}^n_j}\right|, \left|\frac{m - \bar{\mathsf{u}}^n_j}{m_j - \bar{\mathsf{u}}^n_j}\right|, 1\right\}$$

with $M_j = \max_{x \in I_j} u_{h_{|I_j}}(x)$ and $m_j = \min_{x \in I_j} u_{h_{|I_j}}(x)$. Here, the modified polynomial is defined as $\tilde{u}_{h_{|I_j}}(x) = \tilde{\xi}(u_{h_{|I_j}}(x) - \bar{\mathsf{u}}^n_j) + \bar{\mathsf{u}}^n_j$. Note that the satisfaction of the maximum principle and maintenance of higher-order accuracy has been shown in [88].

For all introduced schemes that satisfy a discrete maximum principle, we have the following stability result.

Theorem 10. *Assuming inflow and outflow boundary conditions. If the numerical solution of* (4.28) *is positive, we have*

$$\sum_j |\bar{\mathsf{u}}^{n+1}_j| \leq \sum_j |\bar{\mathsf{u}}^n_j|.$$

Proof. Taking the sum of (4.28) over j, we get $\sum_j \bar{\mathsf{u}}^{n+1}_j = \sum_j (\bar{\mathsf{u}}^n_j + \Delta t g(\bar{\mathsf{u}}^n_j))$. Since $\bar{\mathsf{u}}^{n+1}_j, \bar{\mathsf{u}}^n_j \in [0, M]$, we have

$$\sum_j |\bar{\mathsf{u}}^{n+1}_j| = \sum_j \bar{\mathsf{u}}^{n+1}_j = \sum_j (\bar{\mathsf{u}}^n_j + \Delta t g(\bar{\mathsf{u}}^n_j)) \leq \sum_j |\bar{\mathsf{u}}^n_j|$$

because of $g'(\cdot) \leq 0$ and $g(0) = 0$. □

Let us further remark that a maximum principle cannot be proven for the TVB limiter.

FV schemes with flux limiting with explicit Euler method

Finally, let us consider the maximum principle for finite volume schemes with flux limiters. Consider for the continuous flux function $f' \geq 0$. Applying the flux limiter method with (3.59) to the balance law (4.20), utilizing the mean value theorem, we get the formula

$$\begin{aligned} \mathsf{u}_j^{n+1} &= \mathsf{u}_j^n + \tfrac{\Delta t}{\Delta x} f'(r_j^n)[1 - \tfrac{1}{2}\psi(\theta_{j-1}) + \tfrac{1}{2\theta_j}\psi(\theta_j)](\mathsf{u}_{j-1}^n - \mathsf{u}_j^n) \\ &\quad + \tfrac{\Delta t}{2}\left(\mathsf{g}(\mathsf{u}_j^n + \tfrac{1}{2}\psi(\theta_j)(\mathsf{u}_{j+1}^n - \mathsf{u}_j^n)) + \mathsf{g}(\mathsf{u}_{j-1}^n + \tfrac{1}{2}\psi(\theta_{j-1})(\mathsf{u}_j^n - \mathsf{u}_{j-1}^n))\right). \end{aligned} \tag{4.29}$$

Again utilizing the mean value theorem and Equation (3.59) for the value θ_j, we get

$$\begin{aligned} \mathsf{u}_j^{n+1} = \mathsf{u}_j^n &+ \tfrac{\Delta t}{\Delta x} f'(r_j^n)\left[1 - \tfrac{1}{2}\psi(\theta_{j-1}) + \tfrac{1}{2\theta_j}\psi(\theta_j)\right](\mathsf{u}_{j-1}^n - \mathsf{u}_j^n) \\ &+ \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)\left[\left(1 + \tfrac{1}{2\theta_j}\psi(\theta_j) + \tfrac{1}{2}\psi(\theta_{j-1})\right)\mathsf{u}_j^n + \left(1 - \tfrac{1}{2}\psi(\theta_{j-1}) - \tfrac{1}{2\theta_j}\psi(\theta_j)\right)\mathsf{u}_{j-1}^n\right]. \end{aligned} \tag{4.30}$$

Further, utilizing the monotonicity of the flux function, we show the following theorem.

Theorem 11. *Consider Scheme* (4.29) *and assume that the limiter function ψ lies in the TVD region. Under the assumption that u_{j-1}^n, u_j^n and u_{j+1}^n lie in the range of $[-m, M]$ with $m, M > 0$ and without loss of generality $|M| > |m|$, and the CFL condition*

$$2\tfrac{\Delta t}{\Delta x}\omega + \tfrac{3}{2}\Delta t L \leq \tfrac{|m|}{|M|}, \tag{4.31}$$

where L is the Lipschitz constant of the source term g, we have $\mathsf{u}_j^{n+1} \in [-m, M]$.

Proof. Resorting the terms in (4.30), we obtain

$$\begin{aligned} \mathsf{u}_j^{n+1} &= \mathsf{u}_j^n\left(1 - \tfrac{\Delta t}{\Delta x} f'(r_j^n)\left[1 - \tfrac{1}{2}\psi(\theta_{j-1}) + \tfrac{1}{2\theta_j}\psi(\theta_j)\right] + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)\left[1 + \tfrac{1}{2}\psi(\theta_{j-1}) + \tfrac{1}{2\theta_j}\psi(\theta_j)\right]\right) \\ &\quad + \mathsf{u}_{j-1}^n\left(\tfrac{\Delta t}{\Delta x} f'(r_j^n)\left[1 - \tfrac{1}{2}\psi(\theta_{j-1}) + \tfrac{1}{2\theta_j}\psi(\theta_j)\right] + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)\left[1 - \tfrac{1}{2}\psi(\theta_{j-1}) - \tfrac{1}{2\theta_j}\psi(\theta_j)\right]\right) \\ &= A\mathsf{u}_j^n + B\mathsf{u}_{j-1}^n. \end{aligned}$$

From Equation (3.61) it follows that $0 \leq \psi(\theta_j) \leq 2$ and $0 \leq \frac{1}{\theta_j}\psi(\theta_j) \leq 2$ and therefore, we get

$$\begin{aligned} 0 \leq a =: 1 - \tfrac{1}{2}\psi(\theta_{j-1}) + \tfrac{1}{2\theta_j}\psi(\theta_j) \leq 2 \\ 1 \leq b =: 1 + \tfrac{1}{2\theta_j}\psi(\theta_j) + \tfrac{1}{2}\psi(\theta_{j-1}) \leq 3 \\ -1 \leq c =: 1 - \tfrac{1}{2}\psi(\theta_{j-1}) - \tfrac{1}{2\theta_j}\psi(\theta_j) \leq 1. \end{aligned}$$

with

$$A = 1 - \tfrac{\Delta t}{\Delta x} f'(r_j^n)a + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)b$$
$$B = \tfrac{\Delta t}{\Delta x} f'(r_j^n)a + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)c.$$

With the three inequalities, the CFL condition (4.31) with $m, M > 0$ and $|M| > |m|$, the condition $2\frac{\Delta t}{\Delta x}\omega + 3\frac{\Delta t}{2}L \leq 1$, and the assumption of the source term $\mathsf{g}'(\cdot) \leq 0$, we obtain

$$0 \leq A \quad \text{and} \quad 0 \leq A + B = 1 + \Delta t \mathsf{g}'(s_j^n) \leq 1.$$

Unfortunately, the constant B can become negative. If we assume $B \geq 0$ which is especially the case if $-1 \leq c \leq 0$, we obtain that u_j^{n+1} is in some sense similar to a convex combination of u_j^n and u_{j-1}^n and it follows easily that $\mathsf{u}_j^{n+1} \in [-m, M]$. Now, we have to consider the case, where

$$B = \tfrac{\Delta t}{\Delta x} f'(r_j^n)a + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)c \leq 0.$$

This is especially the case if $0 \leq c \leq 1$. And for this case, we also have to show that the new value u_j^{n+1} stays in the range $[-m, M]$. First, it yields that

$$\begin{aligned}|A| + |B| =& |1 - \tfrac{\Delta t}{\Delta x} f'(r_j^n)a + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)b| + |\tfrac{\Delta t}{\Delta x} f'(r_j^n)a + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)c| \\ &\leq |1 - \tfrac{\Delta t}{\Delta x} f'(r_j^n)a + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)c| + |\tfrac{\Delta t}{\Delta x} f'(r_j^n)a + \tfrac{\Delta t}{2}\mathsf{g}'(s_j^n)c| \leq 1,\end{aligned}$$

where we use the assumptions $0 \leq c \leq 1$, $a \geq 0$, $b \geq c$ and $\mathsf{g}'(\cdot) \leq 0$. Second, we have to take into consideration that the old values u_j^n and u_{j-1}^n can have different signs. To show that $\mathsf{u}_j^{n+1} = A\mathsf{u}_j^n + B\mathsf{u}_{j-1}^n \in [-m, M]$, we have to consider four different cases. Under the assumption $A \geq 0$, $B \leq 0$, we can have $\mathsf{u}_j^n, \mathsf{u}_{j-1}^n \geq 0$, $\mathsf{u}_j^n, \mathsf{u}_{j-1}^n \leq 0$, $\mathsf{u}_j^n \leq 0$ and $\mathsf{u}_{j-1}^n \geq 0$ or $\mathsf{u}_j^n \geq 0$ and $\mathsf{u}_{j-1}^n \leq 0$. We consider two of these four cases. The other two cases can be shown with similar arguments.
For the first case, we assume $A \geq 0$, $B \leq 0$ and $\mathsf{u}_j^n, \mathsf{u}_{j-1}^n \geq 0$. Then, we get

$$\begin{aligned}\mathsf{u}_j^{n+1} &= A\mathsf{u}_j^n + B\mathsf{u}_{j-1}^n \underbrace{\leq}_{B\leq 0, A\geq 0} A\mathsf{u}_j^n \underbrace{\leq}_{f', a, b\geq 0, g'\leq 0} \mathsf{u}_j^n \\ \mathsf{u}_j^{n+1} &= A\mathsf{u}_j^n + B\mathsf{u}_{j-1}^n \underbrace{=}_{B\leq 0} A\mathsf{u}_j^n - |B|\mathsf{u}_{j-1}^n \geq -|B|\mathsf{u}_{j-1}^n \underbrace{\geq}_{\text{CFL Cond.(4.31)}} -m.\end{aligned}$$

All together, we get $\mathsf{u}_j^{n+1} = A\mathsf{u}_j^n + B\mathsf{u}_{j-1}^n \in [-m, M]$. For the second case, we assume $A \geq 0$, $B \leq 0$ and $\mathsf{u}_j^n \geq 0$ and $\mathsf{u}_{j-1}^n \leq 0$. Then, we can approximate

$$\begin{aligned}|\mathsf{u}_j^{n+1}| = |A\mathsf{u}_j^n + B\mathsf{u}_{j-1}^n| &\underbrace{\leq}_{B\leq 0, A\geq 0, \Delta\text{-Inequal.}} |A||\mathsf{u}_j^n| + |B||\mathsf{u}_{j-1}^n| \underbrace{\leq}_{|A|+|B|\leq 1} \max\{|\mathsf{u}_j^n|, |\mathsf{u}_{j-1}^n|\} \\ \mathsf{u}_j^{n+1} = A\mathsf{u}_j^n + B\mathsf{u}_{j-1}^n &\underbrace{\geq}_{B, \mathsf{u}_{j-1}^n\leq 0, A, \mathsf{u}_j^n\geq 0,} 0.\end{aligned}$$

All together, we again obtain $\mathsf{u}_j^{n+1} = A\mathsf{u}_j^n + B\mathsf{u}_{j-1}^n \in [-m, M]$. □

Remark 9. *Note that, if we assume $m, M > 0$ and $|m| > |M|$ in Theorem 11, the CFL condition reads $2\frac{\Delta t}{\Delta x}\omega + \frac{3}{2}\Delta t L \leq \frac{|M|}{|m|}$.*

4.3.3 Linear system without source term

We discuss the maximum principle for the system (2.10) where we assume $\mathbf{g} \equiv 0$ and $\mathbf{a} \geq 0$. We want to show that the numerical solution stays in a certain range for both components which is a type of a maximum principle and differs from that formulated for the scalar case, see Section 4.3.2. For brevity and simplicity, we call this property *discrete maximum principle*.

To show this, we utilize the transformation into characteristic variables $\tilde{\mathrm{H}}$ and $\tilde{\mathrm{Q}}$, c.f. (2.31). Of course, the range can be different for the single components, but we need to define the limits of the range with

$$m = \min\{\min_x \tilde{\mathrm{H}}_0(x), \min_x \tilde{\mathrm{Q}}_0(x)\}, \quad M = \max\{\max_x \tilde{\mathrm{H}}_0(x), \max_x \tilde{\mathrm{Q}}_0(x)\}, \tag{4.32}$$

where $m, M \in \mathbb{R}$. Because of the absence of the source term, we can define the range in a more general way.

Finite Volume methods

In the lower-order FV setting, we get for the characteristic variables

$$\tilde{\mathrm{H}}_j^{n+1} = \tilde{\mathrm{H}}_j^n - \lambda\left[\hat{\mathbf{F}}(\tilde{\mathrm{H}}_j^n, \tilde{\mathrm{H}}_{j+1}^n) - \hat{\mathbf{F}}(\tilde{\mathrm{H}}_{j-1}^n, \tilde{\mathrm{H}}_j^n)\right] =: H_{1,\lambda}(\tilde{\mathrm{H}}_{j-1}^n, \tilde{\mathrm{H}}_j^n, \tilde{\mathrm{H}}_{j+1}^n) \tag{4.33a}$$

$$\tilde{\mathrm{Q}}_j^{n+1} = \tilde{\mathrm{Q}}_j^n - \lambda\left[\hat{\mathbf{F}}(\tilde{\mathrm{Q}}_j^n, \tilde{\mathrm{Q}}_{j+1}^n) - \hat{\mathbf{F}}(\tilde{\mathrm{Q}}_{j-1}^n, \tilde{\mathrm{Q}}_j^n)\right] =: H_{2,\lambda}(\tilde{\mathrm{Q}}_{j-1}^n, \tilde{\mathrm{Q}}_j^n, \tilde{\mathrm{Q}}_{j+1}^n) \tag{4.33b}$$

with

$$H_{1,\lambda}(\tilde{\mathrm{H}}_{j-1}^n, \tilde{\mathrm{H}}_j^n, \tilde{\mathrm{H}}_{j+1}^n) = \tilde{\mathrm{H}}_j^n - \lambda\left[\frac{\mathbf{a}}{2}(\tilde{\mathrm{H}}_{j+1}^n - \tilde{\mathrm{H}}_{j-1}^n) - \frac{\mathbf{a}}{2}(\tilde{\mathrm{H}}_{j+1}^n - 2\tilde{\mathrm{H}}_j^n + \tilde{\mathrm{H}}_{j-1}^n)\right]$$

$$H_{2,\lambda}(\tilde{\mathrm{Q}}_{j-1}^n, \tilde{\mathrm{Q}}_j^n, \tilde{\mathrm{Q}}_{j+1}^n) = \tilde{\mathrm{Q}}_j^n - \lambda\left[\frac{\mathbf{a}}{2}(\tilde{\mathrm{Q}}_{j-1}^n - \tilde{\mathrm{Q}}_{j+1}^n) - \frac{\mathbf{a}}{2}(\tilde{\mathrm{Q}}_{j+1}^n - 2\tilde{\mathrm{Q}}_j^n + \tilde{\mathrm{Q}}_{j-1}^n)\right].$$

With the CFL condition $\lambda\omega \leq 1$, the functions $H_{1,\lambda}$ and $H_{2,\lambda}$ are monotone increasing in all arguments, c.f. Appendix A.2. From the consistency of the numerical flux function $\hat{\mathbf{F}}$, we have

$$m = H_{r,\lambda}(m, m, m) \leq H_{r,\lambda}(\tilde{\mathrm{W}}_{r,j-1}^n, \tilde{\mathrm{W}}_{r,j}^n, \tilde{\mathrm{W}}_{r,j+1}^n) \leq H_{r,\lambda}(M, M, M) = M$$

with $r = 1, 2$ and $\tilde{\mathrm{W}}_{1,j}^n = \tilde{\mathrm{H}}$ and $\tilde{\mathrm{W}}_{2,j}^n = \tilde{\mathrm{Q}}$ for all j, n.

Higher-order DG methods

Since the characteristic variables form a decoupled system, we may apply Theorem 8 with $g \equiv 0$ in a component-wise manner. For the components, we get

$$\tilde{\mathrm{H}}_j^{n+1} = \bar{\tilde{\mathrm{H}}}_j^n - \lambda \left[\hat{\mathbf{F}}(\tilde{\mathrm{H}}_{j+1/2}^{-,n}, \tilde{\mathrm{H}}_{j+1/2}^{+,n}) - \hat{\mathbf{F}}(\tilde{\mathrm{H}}_{j-1/2}^{-,n}, \tilde{\mathrm{H}}_{j-1/2}^{+,n}) \right] \tag{4.34a}$$

$$\tilde{\mathrm{Q}}_j^{n+1} = \bar{\tilde{\mathrm{Q}}}_j^n - \lambda \left[\hat{\mathbf{F}}(\tilde{\mathrm{Q}}_{j+1/2}^{-,n}, \tilde{\mathrm{Q}}_{j+1/2}^{+,n}) - \hat{\mathbf{F}}(\tilde{\mathrm{Q}}_{j-1/2}^{-,n}, \tilde{\mathrm{Q}}_{j-1/2}^{+,n}) \right] \tag{4.34b}$$

where the right-hand sides are denoted by $K_{1,\lambda}(\bar{\tilde{\mathrm{H}}}_j^n, \tilde{\mathrm{H}}_{j+1/2}^{-,n}, \tilde{\mathrm{H}}_{j+1/2}^{+,n}, \tilde{\mathrm{H}}_{j-1/2}^{-,n}, \tilde{\mathrm{H}}_{j-1/2}^{+,n})$ and $K_{2,\lambda}(\bar{\tilde{\mathrm{Q}}}_j^n, \tilde{\mathrm{Q}}_{j+1/2}^{-,n}, \tilde{\mathrm{Q}}_{j+1/2}^{+,n}, \tilde{\mathrm{Q}}_{j-1/2}^{-,n}, \tilde{\mathrm{Q}}_{j-1/2}^{+,n})$. In the case of higher-order DG schemes, we can apply Theorem 9 on a component-wise basis and with $g \equiv 0$. Note that we can either use the TVD limiter (3.55) in the first-order case or the linear scaling limiter [88] for higher-order.

Finite Volume methods with flux limiting

For the finite volume scheme with flux limiting, we can proceed in a similar fashion as for the scalar case. We therefore need to use Theorem 11 component-wise with $\mathsf{g} \equiv 0$ and the CFL condition $2\frac{\Delta t}{\Delta x}\omega \leq 1$. Note that in this case, we have for characteristic variables

$$\tilde{\mathrm{H}}_j^{n+1} = \tilde{\mathrm{H}}_j^n - \lambda \left(f\left(\tilde{\mathrm{H}}_j^n + \tfrac{1}{2}\psi(\theta_j)\left(\tilde{\mathrm{H}}_{j+1}^n - \tilde{\mathrm{H}}_j^n \right) \right) - f\left(\tilde{\mathrm{H}}_{j-1}^n + \tfrac{1}{2}\psi(\theta_{j-1})\left(\tilde{\mathrm{H}}_j^n - \tilde{\mathrm{H}}_{j-1}^n \right) \right) \right) \tag{4.35a}$$

$$\tilde{\mathrm{Q}}_j^{n+1} = \tilde{\mathrm{Q}}_j^n - \lambda \left(f\left(\tilde{\mathrm{Q}}_{j+1}^n + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta_{j+1}} \right)\left(\tilde{\mathrm{Q}}_j^n - \tilde{\mathrm{Q}}_{j+1}^n \right) \right) - f\left(\tilde{\mathrm{Q}}_j^n + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta_j} \right)\left(\tilde{\mathrm{Q}}_{j-1}^n - \tilde{\mathrm{Q}}_j^n \right) \right) \right) \tag{4.35b}$$

with (3.59) for the Equation (4.35a) and (3.60) for Equation (4.35b). Particular, applying the mean value Theorem to the differences of f and using the Equation (3.59) or (3.60), we obtain that the new values $\tilde{\mathrm{H}}_j^{n+1}$, $\tilde{\mathrm{Q}}_j^{n+1}$ can be written as a convex combination of $\tilde{\mathrm{H}}_j^n$ and $\tilde{\mathrm{H}}_{j-1}^n$ or $\tilde{\mathrm{Q}}_j^n$ and $\tilde{\mathrm{Q}}_{j-1}^n$. Therefore, it follows that $\tilde{\mathrm{H}}_j^{n+1} \in [\tilde{\mathrm{H}}_j^n, \tilde{\mathrm{H}}_{j-1}^n]$ and $\tilde{\mathrm{Q}}_j^{n+1} \in [\tilde{\mathrm{Q}}_j^n, \tilde{\mathrm{Q}}_{j-1}^n]$ and this implies $\tilde{\mathrm{H}}_j^{n+1}, \tilde{\mathrm{Q}}_j^{n+1} \in [m, M]$.

It remains to be shown that the discrete solution in the conservative setting also stays in a certain range $[m, M]$. However, using the transformation with the matrix $\mathbf{R}$, the range given for the characteristic variables transfers to a certain range for the conservative variables. Since we stay within this range in the characteristic setting, this translates directly to the conservative variables. Note, however, that if we apply the numerical scheme directly to the conservative setting, we have no such conjecture.

4.3.4 Linear system with source term

Let us now turn to the full water hammer equation including the source term. Again, we assume $\mathbf{a} \geq 0$ and $\mathbf{g}(0) = 0$ and $-\mathbf{g}'(\cdot) \leq 0$ for the source term. Utilizing the transformation to characteristic variables, the explicit Euler method takes the form

$$\tilde{\mathrm{W}}^{n+1} = \tilde{\mathrm{W}}^n + \frac{\Delta t}{\Delta x}\tilde{L}_h(\tilde{\mathrm{W}}^n) + \Delta t\tilde{G}(\tilde{\mathrm{W}}^n),$$

where $\tilde{L}_h$ and $\tilde{G}$ are suitable operators. Because of the coupled system and the properties of the source term, we need a symmetric interval, as can be seen below in detail. Therefore, we assume that $\tilde{\mathrm{W}}^n \in [-M, M]$ component-wise for $M = \max(|\tilde{m}|, |\tilde{M}|)$ with $\tilde{m}$ and $\tilde{M}$ defined in (4.32). We distinguish between several spatial discretizations.

Finite Volume methods

For lower-order FV or a DG scheme with constant ansatz and test functions, we get

$$\tilde{\mathrm{H}}_j^{n+1} = \underbrace{\tilde{\mathrm{H}}_j^n - \lambda\left[\frac{\mathbf{a}}{2}(\tilde{\mathrm{H}}_{j+1}^n - \tilde{\mathrm{H}}_{j-1}^n) - \frac{\mathbf{a}}{2}(\tilde{\mathrm{H}}_{j+1}^n - 2\tilde{\mathrm{H}}_j^n + \tilde{\mathrm{H}}_{j-1}^n)\right] - \frac{\Delta t}{2}\mathbf{g}(\tilde{\mathrm{H}}_j^n + \tilde{\mathrm{Q}}_j^n)}_{G_{1,\lambda}(\tilde{\mathrm{H}}_{j-1}^n, \tilde{\mathrm{H}}_j^n, \tilde{\mathrm{H}}_{j+1}^n, \tilde{\mathrm{Q}}_j^n)} \tag{4.36a}$$

$$\tilde{\mathrm{Q}}_j^{n+1} = \underbrace{\tilde{\mathrm{Q}}_j^n - \lambda\left[\frac{\mathbf{a}}{2}(\tilde{\mathrm{Q}}_{j-1}^n - \tilde{\mathrm{Q}}_{j+1}^n) - \frac{\mathbf{a}}{2}(\tilde{\mathrm{Q}}_{j+1}^n - 2\tilde{\mathrm{Q}}_j^n + \tilde{\mathrm{Q}}_{j-1}^n)\right] - \frac{\Delta t}{2}\mathbf{g}(\tilde{\mathrm{H}}_j^n + \tilde{\mathrm{Q}}_j^n)}_{G_{2,\lambda}(\tilde{\mathrm{Q}}_{j-1}^n, \tilde{\mathrm{Q}}_j^n, \tilde{\mathrm{Q}}_{j+1}^n, \tilde{\mathrm{H}}_j^n))}. \tag{4.36b}$$

We show that the functions $G_{1,\lambda}(\tilde{\mathrm{H}}_{j-1}^n, \tilde{\mathrm{H}}_j^n, \tilde{\mathrm{H}}_{j+1}^n, \tilde{\mathrm{Q}}_j^n)$ and $G_{2,\lambda}(\tilde{\mathrm{Q}}_{j-1}^n, \tilde{\mathrm{Q}}_j^n, \tilde{\mathrm{Q}}_{j+1}^n, \tilde{\mathrm{H}}_j^n))$ are monotonically increasing in the first three arguments and monotonically decreasing in the last argument, c.f. Appendix A.2. Note that we require the CFL condition

$$\lambda\omega + \Delta t\frac{L}{2} \leq 1. \tag{4.37}$$

Inserting the range limits $-M$ and M, we get

$$G_{r,\lambda}(-M, -M, -M, M) = -M - \frac{\Delta t}{2}g(-M + M) = -M$$
$$G_{r,\lambda}(M, M, M, -M) = M - \frac{\Delta t}{2}g(M - M) = M$$

with $r = 1, 2$. Here we use that $g(0) = 0$. Together with the monotonicity property and the values for the range limits, we get

$$-M \leq G_{1,\lambda}(\tilde{\mathrm{H}}_{j-1}^n, \tilde{\mathrm{H}}_j^n, \tilde{\mathrm{H}}_{j+1}^n, \tilde{\mathrm{Q}}_j^n), G_{2,\lambda}(\tilde{\mathrm{Q}}_{j-1}^n, \tilde{\mathrm{Q}}_j^n, \tilde{\mathrm{Q}}_{j+1}^n, \tilde{\mathrm{H}}_j^n)) \leq M.$$

Higher-order DG methods

For the DG(1) method, we use (4.23) with the cell-averages (4.24) like in the scalar case. For the characteristic variables, we have

$$\begin{aligned}\tilde{\mathrm{H}}_j^{n+1} &= D_{1,\lambda}(\tilde{\mathrm{H}}_{j-1/2}^{-,n}, \tilde{\mathrm{H}}_{j-1/2}^{+,n}, \tilde{\mathrm{H}}_{j+1/2}^{-,n}, \tilde{\mathrm{Q}}_{j-1/2}^{+,n}, \tilde{\mathrm{Q}}_{j+1/2}^{-,n}) := \frac{1}{2}\left(\tilde{\mathrm{H}}_{j-1/2}^{+,n} + \tilde{\mathrm{H}}_{j+1/2}^{-,n}\right) \qquad (4.38a)\\ &\quad - \lambda\left[\mathbf{a}(\tilde{\mathrm{H}}_{j+1/2}^{-,n} - \tilde{\mathrm{H}}_{j-1/2}^{-,n})\right] - \frac{\Delta t}{2}\mathrm{g}\left(\frac{1}{2}\left(\tilde{\mathrm{H}}_{j-1/2}^{+,n} + \tilde{\mathrm{H}}_{j+1/2}^{-,n} + \tilde{\mathrm{Q}}_{j-1/2}^{+,n} + \tilde{\mathrm{Q}}_{j+1/2}^{-,n}\right)\right),\\ \tilde{\mathrm{Q}}_j^{n+1} &= D_{2,\lambda}(\tilde{\mathrm{Q}}_{j-1/2}^{+,n}, \tilde{\mathrm{Q}}_{j+1/2}^{-,n}, \tilde{\mathrm{Q}}_{j+1/2}^{+,n}, \tilde{\mathrm{H}}_{j-1/2}^{+,n}, \tilde{\mathrm{H}}_{j+1/2}^{-,n}) := \frac{1}{2}\left(\tilde{\mathrm{Q}}_{j-1/2}^{+,n} + \tilde{\mathrm{Q}}_{j+1/2}^{-,n}\right) \qquad (4.38b)\\ &\quad - \lambda\left[\mathbf{a}(\tilde{\mathrm{Q}}_{j-1/2}^{+,n} - \tilde{\mathrm{Q}}_{j+1/2}^{+,n})\right] - \frac{\Delta t}{2}\mathrm{g}\left(\frac{1}{2}\left(\tilde{\mathrm{H}}_{j-1/2}^{+,n} + \tilde{\mathrm{H}}_{j+1/2}^{-,n} + \tilde{\mathrm{Q}}_{j-1/2}^{+,n} + \tilde{\mathrm{Q}}_{j+1/2}^{-,n}\right)\right).\end{aligned}$$

To show that the two functions $D_{1,\lambda}$ and $D_{2,\lambda}$ are monotonically increasing in the first three arguments and monotonically decreasing in the last two arguments, c.f. Appendix A.2, we require the CFL condition

$$\lambda\omega + \frac{\Delta t}{4}L \leq 1/2. \tag{4.39}$$

Inserting the range limits M and $-M$, we get

$$\begin{aligned}D_{1,\lambda}(M, M, M, -M, -M) = D_{2,\lambda}(M, M, M, -M, -M) = M - \Delta t g(0) = M,\\ D_{1,\lambda}(-M, -M, -M, M, M) = D_{2,\lambda}(-M, -M, -M, M, M) = -M - \Delta t g(0) = -M.\end{aligned}$$

And consequently

$$\begin{aligned}-M \leq D_{1,\lambda}(\tilde{\mathrm{H}}_{j-1/2}^{-,n}, \tilde{\mathrm{H}}_{j-1/2}^{+,n}, \tilde{\mathrm{H}}_{j+1/2}^{-,n}, \tilde{\mathrm{Q}}_{j-1/2}^{+,n}, \tilde{\mathrm{Q}}_{j+1/2}^{-,n}) \leq M,\\ -M \leq D_{2,\lambda}(\tilde{\mathrm{Q}}_{j-1/2}^{+,n}, \tilde{\mathrm{Q}}_{j+1/2}^{-,n}, \tilde{\mathrm{Q}}_{j+1/2}^{+,n}, \tilde{\mathrm{H}}_{j-1/2}^{+,n}, \tilde{\mathrm{H}}_{j+1/2}^{-,n}) \leq M.\end{aligned}$$

For the case of higher-order polynomials ($k \geq 1$), we get the following formulation:

$$\begin{aligned}\tilde{\mathrm{H}}_j^{n+1} &= \sum_{b=2}^{N-1} \hat{\omega}_b h_j(\hat{x}_j^b) + \hat{\omega}_N H_{\lambda/\hat{\omega}_N}(\tilde{\mathrm{H}}_{j-1/2}^{+,n}, \tilde{\mathrm{H}}_{j+1/2}^{-,n}, \tilde{\mathrm{H}}_{j+1/2}^{+,n}) \qquad (4.40a)\\ &\quad + \hat{\omega}_1 H_{\lambda/\hat{\omega}_1}(\tilde{\mathrm{H}}_{j-1/2}^{-,n}, \tilde{\mathrm{H}}_{j-1/2}^{+,n}, \tilde{\mathrm{H}}_{j+1/2}^{-,n}) - \frac{\Delta t}{2}\mathrm{g}\left(\sum_{b=1}^{N}\left(\hat{\omega}_b q_j(\hat{x}_j^b) + \hat{\omega}_b h_j(\hat{x}_j^b)\right)\right),\\ \tilde{\mathrm{Q}}_j^{n+1} &= \sum_{b=2}^{N-1} \hat{\omega}_b q_j(\hat{x}_j^b) + \hat{\omega}_N H_{\lambda/\hat{\omega}_N}(\tilde{\mathrm{Q}}_{j-1/2}^{+,n}, \tilde{\mathrm{Q}}_{j+1/2}^{-,n}, \tilde{\mathrm{Q}}_{j+1/2}^{+,n}) \qquad (4.40b)\\ &\quad + \hat{\omega}_1 H_{\lambda/\hat{\omega}_1}(\tilde{\mathrm{Q}}_{j-1/2}^{-,n}, \tilde{\mathrm{Q}}_{j-1/2}^{+,n}, \tilde{\mathrm{Q}}_{j+1/2}^{-,n}) - \frac{\Delta t}{2}\mathrm{g}\left(\sum_{b=1}^{N}\left(\hat{\omega}_b q_j(\hat{x}_j^b) + \hat{\omega}_b h_j(\hat{x}_j^b)\right)\right),\end{aligned}$$

where we denote the right-hand sides with $B_{1,\lambda}$ and $B_{2,\lambda}$. Under the CFL condition

$$\lambda\omega + \hat{\omega}_1 \Delta t \frac{L}{2} \leq \hat{\omega}_1, \tag{4.41}$$

we can prove that the right-hand side in (4.40a) is monotonically increasing in the arguments $H^{\pm,n}_{j\pm1/2}$ and $h_j(\hat{x}^b_j)$ and monotonically decreasing in the arguments $q_j(\hat{x}^b_j)$ for all $j = 1, ..., N$, c.f. Appendix A.2. For Equation (4.40b), the proof is analogous.

Flux limiting FV scheme

For the FV scheme with flux limiting (3.59) and (3.60), we get the two equations

$$\begin{aligned}
\tilde{\mathrm{H}}^{n+1}_j = \tilde{\mathrm{H}}^n_j &- \lambda\left(\mathsf{f}\left(\tilde{\mathrm{H}}^n_j + \tfrac{1}{2}\psi(\theta^H_j)(\tilde{\mathrm{H}}^n_{j+1} - \tilde{\mathrm{H}}^n_j)\right) - \mathsf{f}\left(\tilde{\mathrm{H}}^n_{j-1} + \tfrac{1}{2}\psi(\theta^H_{j-1})(\tilde{\mathrm{H}}^n_j - \tilde{\mathrm{H}}^n_{j-1})\right)\right) \\
&- \tfrac{\Delta t}{2}\Big[\mathsf{g}\left(\tilde{\mathrm{H}}^n_j + \tfrac{1}{2}\psi(\theta^H_j)(\tilde{\mathrm{H}}^n_{j+1} - \tilde{\mathrm{H}}^n_j) + \tilde{\mathrm{Q}}^n_j + \tfrac{1}{2}\psi(\theta^Q_j)(\tilde{\mathrm{Q}}^n_{j+1} - \tilde{\mathrm{Q}}^n_j)\right) \qquad (4.42a) \\
&+ \mathsf{g}\left(\tilde{\mathrm{H}}^n_{j-1} + \tfrac{1}{2}\psi(\theta^H_{j-1})(\tilde{\mathrm{H}}^n_j - \tilde{\mathrm{H}}^n_{j-1}) + \tilde{\mathrm{Q}}^n_{j-1} + \tfrac{1}{2}\psi(\theta^Q_{j-1})(\tilde{\mathrm{Q}}^n_j - \tilde{\mathrm{Q}}^n_{j-1})\right)\Big], \\
\tilde{\mathrm{Q}}^{n+1}_j = \tilde{\mathrm{Q}}^n_j &- \lambda\left(\mathsf{f}\left(\tilde{\mathrm{Q}}^n_{j+1} + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^Q_{j+1}}\right)(\tilde{\mathrm{Q}}^n_j - \tilde{\mathrm{Q}}^n_{j+1})\right) - \mathsf{f}\left(\tilde{\mathrm{Q}}^n_j + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^Q_j}\right)(\tilde{\mathrm{Q}}^n_{j-1} - \tilde{\mathrm{Q}}^n_j)\right)\right) \\
&- \tfrac{\Delta t}{2}\Big[\mathsf{g}\left(\tilde{\mathrm{Q}}^n_{j+1} + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^Q_{j+1}}\right)(\tilde{\mathrm{Q}}^n_j - \tilde{\mathrm{Q}}^n_{j+1}) + \tilde{\mathrm{H}}^n_{j+1} + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^H_{j+1}}\right)(\tilde{\mathrm{H}}^n_j - \tilde{\mathrm{H}}^n_{j+1})\right) \\
&+ \mathsf{g}\left(\tilde{\mathrm{Q}}^n_j + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^Q_j}\right)(\tilde{\mathrm{Q}}^n_{j-1} - \tilde{\mathrm{Q}}^n_j) + \tilde{\mathrm{H}}^n_j + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^H_j}\right)(\tilde{\mathrm{H}}^n_{j-1} - \tilde{\mathrm{H}}^n_j)\right)\Big]. \qquad (4.42b)
\end{aligned}$$

With the mean value Theorem applied to the differences of the flux functions, we obtain

$$\begin{aligned}
\tilde{\mathrm{H}}^{n+1}_j = \tilde{\mathrm{H}}^n_j &+ \lambda\mathsf{f}'(v_j)\left[1 - \tfrac{1}{2}\psi(\theta^H_{j-1}) + \tfrac{1}{2\theta^H_j}\psi(\theta^H_j)\right](\tilde{\mathrm{H}}^n_{j-1} - \tilde{\mathrm{H}}^n_j) \qquad (4.43a) \\
&- \tfrac{\Delta t}{2}\Big[\mathsf{g}\left(\tilde{\mathrm{H}}^n_j + \tfrac{1}{2}\psi(\theta^H_j)(\tilde{\mathrm{H}}^n_{j+1} - \tilde{\mathrm{H}}^n_j) + \tilde{\mathrm{Q}}^n_j + \tfrac{1}{2}\psi(\theta^Q_j)(\tilde{\mathrm{Q}}^n_{j+1} - \tilde{\mathrm{Q}}^n_j)\right) \\
&+ \mathsf{g}\left(\tilde{\mathrm{H}}^n_{j-1} + \tfrac{1}{2}\psi(\theta^H_{j-1})(\tilde{\mathrm{H}}^n_j - \tilde{\mathrm{H}}^n_{j-1}) + \tilde{\mathrm{Q}}^n_{j-1} + \tfrac{1}{2}\psi(\theta^Q_{j-1})(\tilde{\mathrm{Q}}^n_j - \tilde{\mathrm{Q}}^n_{j-1})\right)\Big], \\
\tilde{\mathrm{Q}}^{n+1}_j = \tilde{\mathrm{Q}}^n_j &- \lambda\mathsf{f}'(w_j)\left[1 - \tfrac{1}{2}\psi(\tfrac{1}{\theta^Q_{j+1}}) + \tfrac{1}{2}\theta^Q_j\psi(\tfrac{1}{\theta^Q_j})\right](\tilde{\mathrm{Q}}^n_{j+1} - \tilde{\mathrm{Q}}^n_j) \qquad (4.43b) \\
&- \tfrac{\Delta t}{2}\Big[\mathsf{g}\left(\tilde{\mathrm{Q}}^n_{j+1} + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^Q_{j+1}}\right)(\tilde{\mathrm{Q}}^n_j - \tilde{\mathrm{Q}}^n_{j+1}) + \tilde{\mathrm{H}}^n_{j+1} + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^H_{j+1}}\right)(\tilde{\mathrm{H}}^n_j - \tilde{\mathrm{H}}^n_{j+1})\right) \\
&+ \mathsf{g}\left(\tilde{\mathrm{Q}}^n_j + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^Q_j}\right)(\tilde{\mathrm{Q}}^n_{j-1} - \tilde{\mathrm{Q}}^n_j) + \tilde{\mathrm{H}}^n_j + \tfrac{1}{2}\psi\left(\tfrac{1}{\theta^H_j}\right)(\tilde{\mathrm{H}}^n_{j-1} - \tilde{\mathrm{H}}^n_j)\right)\Big],
\end{aligned}$$

where v_j and w_j some intermediate points between the arguments of each flux differences and $\mathsf{f}'(v_j) \geq 0$ and $\mathsf{f}'(w_j) \leq 0$. For notation purposes, we denote the right-hand side of Equation (4.42a) by

$$Z_{1,\lambda}(\tilde{\mathrm{H}}^n_{j-2}, \tilde{\mathrm{H}}^n_{j-1}, \tilde{\mathrm{H}}^n_j, \tilde{\mathrm{H}}^n_{j+1}, \tilde{\mathrm{Q}}^n_{j-2}, \tilde{\mathrm{Q}}^n_{j-1}, \tilde{\mathrm{Q}}^n_j, \tilde{\mathrm{Q}}^n_{j+1})$$

and the right-hand side of Equation (4.42b) by

$$Z_{2,\lambda}(\tilde{\mathrm{Q}}^n_{j-1}, \tilde{\mathrm{Q}}^n_j, \tilde{\mathrm{Q}}^n_{j+1}, \tilde{\mathrm{Q}}^n_{j+2}, \tilde{\mathrm{H}}^n_{j-1}, \tilde{\mathrm{H}}^n_j, \tilde{\mathrm{H}}^n_{j+1}, \tilde{\mathrm{H}}^n_{j+2}).$$

Let us comment on a discrete maximum principle for Scheme (4.42a)-(4.42b). Let us assume that the limiter function lies in the TVD region and that $\tilde{\mathrm{H}}^n_{j-2}$, $\tilde{\mathrm{H}}^n_{j-1}$, $\tilde{\mathrm{H}}^n_j$, $\tilde{\mathrm{H}}^n_{j+1}$, $\tilde{\mathrm{H}}^n_{j+2}$, $\tilde{\mathrm{Q}}^n_{j-2}$, $\tilde{\mathrm{Q}}^n_{j-1}$, $\tilde{\mathrm{Q}}^n_j$, $\tilde{\mathrm{Q}}^n_{j+1}$, and $\tilde{\mathrm{Q}}^n_{j+2}$ lie in the range $[-M, M]$. We want to show that $\tilde{\mathrm{H}}^{n+1}_j, \tilde{\mathrm{Q}}^{n+1}_j \in [-M, M]$.

We prove the statement for Equation (4.42a) only. The proof for the second component (4.42b) is the same. As a first step, we rewrite Equation (4.42a) as

$$\tilde{\mathrm{H}}_j^{n+1} = \tilde{\mathrm{H}}_j^n - \lambda \left[\mathsf{f}(H1) - f(H2)\right] - \tfrac{\Delta t}{2} \left[\mathsf{g}(H1 + Q1) + \mathsf{g}(H2 + Q2)\right],$$

where $H1 = \tilde{\mathrm{H}}_j^n + \frac{1}{2}\psi(\theta_j^H)(\tilde{\mathrm{H}}_{j+1}^n - \tilde{\mathrm{H}}_j^n)$, $H2 = \tilde{\mathrm{H}}_{j-1}^n + \frac{1}{2}\psi(\theta_{j-1}^H)(\tilde{\mathrm{H}}_j^n - \tilde{\mathrm{H}}_{j-1}^n)$, $Q1 = \tilde{\mathrm{Q}}_j^n + \frac{1}{2}\psi(\theta_j^Q)(\tilde{\mathrm{Q}}_{j+1}^n - \tilde{\mathrm{Q}}_j^n)$ and $Q2 = \tilde{\mathrm{Q}}_{j-1}^n + \frac{1}{2}\psi(\theta_{j-1}^Q)(\tilde{\mathrm{Q}}_j^n - \tilde{\mathrm{Q}}_{j-1}^n)$. The coupling in Scheme (4.42a)-(4.42b) allows for the argument in the function g to have the wrong sign which means we can not simply use the dissipativity of the source term. However, we observe that this can only happen if

$$|\tilde{\mathrm{H}}_j^n - \lambda \left[\mathsf{f}(H1) - f(H2)\right]| \leq M - \varepsilon \tag{4.44}$$

for some $\varepsilon > 0$. From Equation (4.43a), we observe that the term on the right-hand side can be seen as a convex combination of $\tilde{\mathrm{H}}_j^n$ and $\tilde{\mathrm{H}}_{j-1}^n$. To prove this, we assume $\tilde{\mathrm{H}}_j^n - \lambda \left[\mathsf{f}(H1) - f(H2)\right] = M$ which implies $\tilde{\mathrm{H}}_j^n = H1 = H2 = M$. This implies that

$$H1 + Q1 \geq 0 \quad \text{and} \quad H2 + Q2 \geq 0,$$

and we can again utilize the dissipativity of the source term. We can argue in the same way for the lower bound $-M$. Consequently, if e.g. $H1 + Q1 \leq 0$, we have that Equation (4.44) holds and we choose Δt small enough such that

$$|\tfrac{\Delta t}{2} \left[\mathsf{g}(H1 + Q1) + \mathsf{g}(H2 + Q2)\right]| \leq \varepsilon.$$

Remark 10. *Unfortunately, this does not prevent that Δt has to be chosen arbitrary small and for any fixed CFL condition, we may construct a counter example, where the discrete maximum principle does not hold. However, applying Theorem 7, we see that the SDIRKFV with flux limiting also fulfills some kind of discrete maximum principle in our numerical examples in Chapter 5.*

Summary

We introduced a combination of SDIRK time integration schemes with several finite volume and DG schemes with local Lax-Friedrichs flux. We established the existence and uniqueness of solutions of (3.2) without requiring an additional step size restriction. We showed that the resulting schemes are well-balanced with respect to the water hammer equations. Further, we showed that all presented schemes fulfill some kind of discrete maximum-principle. For this purpose, we utilized a connection between the SSP SDIRK time integration and the explicit Euler method. Note that we utilized the characteristic decomposition for the system case, which enabled us to reuse the results from the scalar case.

5 Numerical Results

In this chapter, we verify the results obtained so far numerically. An important point is the SSP property, i.e. the TV stability. This is strongly related to the satisfaction of the discrete maximum-principle.

For our numerical tests, we utilize a combination of an SDIRK method with the different spatial discretizations presented in Chapter 3. For the time integration, we use an optimal SSP SDIRK method with Butcher tableau

$$\begin{array}{c|cc} 1/4 & 1/4 & 0 \\ 3/4 & 1/2 & 1/4 \\ \hline & 1/2 & 1/2 \end{array}, \tag{5.1}$$

see also Example 2 in Section 3.1.4. To compare the different combinations of this SDIRK method and the spatial discretizations, we try to keep the spatial mesh size constant for all methods whenever possible. As has been already mentioned, we use the MOL approach with finite volume or discontinuous Galerkin methods for the spatial discretization. We denote the method with the lower-order finite volume method combined with the SDIRK scheme by *SDIRKFV*. Further, *SDIRKFLUX* denotes the combination with the finite volume method with flux limiter which was introduced in Section 3.4, see also Remark 5. Here, we use the introduced Koren limiter, see (3.63). Combinations with the DG method with linear ansatz and test functions with TVB or TVD limiter are denoted by *SDIRKDGTVB* and *SDIRKDGTVD* and without limiter by *SDIRKDG*, see Section 3.3.2. For comparison, we also show results which were obtained using the IBOX scheme, c.f. Section 3.3.3, and we denote this scheme by *IBOX*.

To verify the desired properties, we treat examples for linear and nonlinear scalar conservation laws first. In a second part, we also show the performance of the method when applied to the water hammer equations.

5.1 Scalar balance law

In this section, we mainly treat two examples, i.e., the transport equation as an example of a linear conservation law and the Buckley-Leverett equation [7], which serves as an example for a nonlinear equation. Let us fix the geometrical setting. We set $\Omega = [0, 1]$ for the spatial variable and $0 \leq t \leq T = 1$ for the time variable. We use periodic boundary conditions in space and an equidistant grid spacing for simplicity. Let us denote the spatial mesh size by $\Delta x = \frac{1}{N_x}$ where N_x is the number of control cells. Analogously, we set $\Delta t = \frac{T}{N_t}$ for the temporal step size where N_t denotes the number of time steps.

For the TVD stability, we use the discrete TV norm for the space periodic setting

$$|\mathsf{u}|_{TV} = \sum_{j=1}^{N_x} |\mathsf{u}_j - \mathsf{u}_{j-1}|, \tag{5.2}$$

where $\mathsf{u}_0 = \mathsf{u}_{N_x}$. Note that for the first-order DG case, TV stability holds for the mean values and therefore, u has to be substituted with $\bar{\mathsf{u}}$. Further, we introduce the ratio

$$\mu(\Delta t) = \max_{1 \leq n \leq N_t} \frac{|\mathsf{u}^n|_{TV}}{|\mathsf{u}^{n-1}|_{TV}} \tag{5.3}$$

which equals one if the TV norm is exactly preserved by the time stepping scheme. We have $\mu(\Delta t) \leq 1$ if the method is TVD(M).

Transport equation

We consider the linear conservation law

$$\partial_t \mathsf{u} + \mathbf{a} \partial_x \mathsf{u} = 0 \tag{5.4}$$

with initial condition $\mathsf{u}(x, 0) = \mathsf{u}^0(x)$. The analytical solution of this equation is given by

$$\mathsf{u}(x, t) = \mathsf{u}^0(x - \mathbf{a}t). \tag{5.5}$$

For simplicity, we set $\mathbf{a} = 1$. Further, we use the non-smooth box profile

$$\mathsf{u}^0(x) = \begin{cases} 1, & 0 \leq x < 0.3 \\ 2, & 0.3 \leq x \leq 0.7 \\ 1, & 0.7 < x \leq 1 \end{cases} \tag{5.6}$$

as initial condition. Due to the analytical solution, we expect that this box is transported with velocity $\mathbf{a} = 1$ and due to the periodic boundary conditions, we expect $\mathsf{u}(x, T) = \mathsf{u}^0(x)$.

The results for $\mathsf{u}(x, T)$ with $T = 1$, $N_x = 200$, $N_t = 600$ and $N_t = 300$ for the IBOX scheme are shown in Figure 5.1. We have the CFL condition $\frac{\Delta t}{\Delta x}\mathbf{a} = 0.3333$ and thus, the lower bound of the IBOX scheme (3.45) is fulfilled. We observe that all methods lie in the range $[1, 2]$, i.e. they satisfy the discrete maximum-principle, except for the DG without slope limiter which shows slight over- and undershoots at the discontinuities. Note that even though the DG without slope limiter does not satisfy the maximum-principle, it resolves the discontinuities rather sharply and the transport velocity is correct. All methods show different amounts of numerical diffusion where the schemes with limiters perform best.

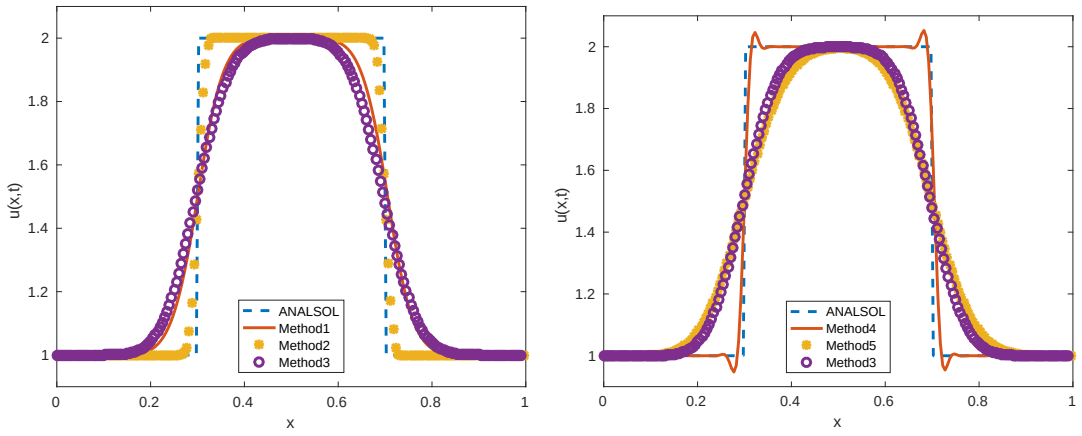

Figure 5.1: Results for the linear transport equation with initial box profile at time $t = 1$. We use $N_x = 200$ for the spatial resolution and $N_t = 600$ time steps (for the IBOX, we use $N_t = 300$). Left: Method1 to Method3 represent the SDIRKDG with TVD limiter, the SDIRKFLUX and the IBOX method. Right: Method4 and Method5 denote the SDIRKDG without limiter and the SDIRKFV. The IBOX and the analytical solution ANASOL are depicted in both.

We show the values of μ in Table 5.1. From this, we obtain the TV stability for all schemes except for the SDIRKDG. Note also that we only show the values of the SDIRKDGTVB scheme whereas the values of the SDIRKDGTVD are almost identical.

Table 5.1: TV norm for the results depicted in Figure 5.1 at different times t and the value of μ in the last column.

Method	$\lvert \mathsf{u}_{\lvert t=\Delta t}\rvert_{TV}$	$\lvert \mathsf{u}_{\lvert t=0.5}\rvert_{TV}$	$\lvert \mathsf{u}_{\lvert t=1}\rvert_{TV}$	$\mu(\Delta t)$
SDIRKFLUX	2.0000	2.0000	2.0000	1.0000
SDIRKDGTVB	2.0000	1.9999	1.9998	1.0000
SDIRKFV	2.0000	1.9998	1.9906	1.0000
SDIRKDG	2.2083	2.4168	2.4427	1.1041
IBOX	2.0000	1.9999	1.9990	1.0000

To verify the theoretical results made in Section 4.3.2, we show the numerical solution of the slightly modified transport equation

$$\partial_t \mathsf{u} + \mathbf{a}\partial_x \mathsf{u} = \mathsf{g}(\mathsf{u}), \tag{5.7}$$

where $\mathsf{g}(\mathsf{u}) = -\lambda \mathsf{u}|\mathsf{u}|$ with constant $\lambda = 0.1$. We again use $\mathsf{u}(x, 0) = \mathsf{u}^0(x)$ as initial condition, where $\mathsf{u}^0(x)$ is defined in (5.6).

In Figure 5.2, we show the numerical results for $t = 0.5$ and $t = T = 1$, respectively. As before, we observe numerical diffusion for all methods. The DG method without limiter again admits over- and undershoots at the discontinuities. Note that due to the additional diffusion term g, the range for the maximum principle will be $[-1, 2]$, which is also observed in Section 5.2. Again, this is fulfilled for all methods except for the DG method without slope limiter. For $t \to \infty$, the solution will converge to zero in all cases.

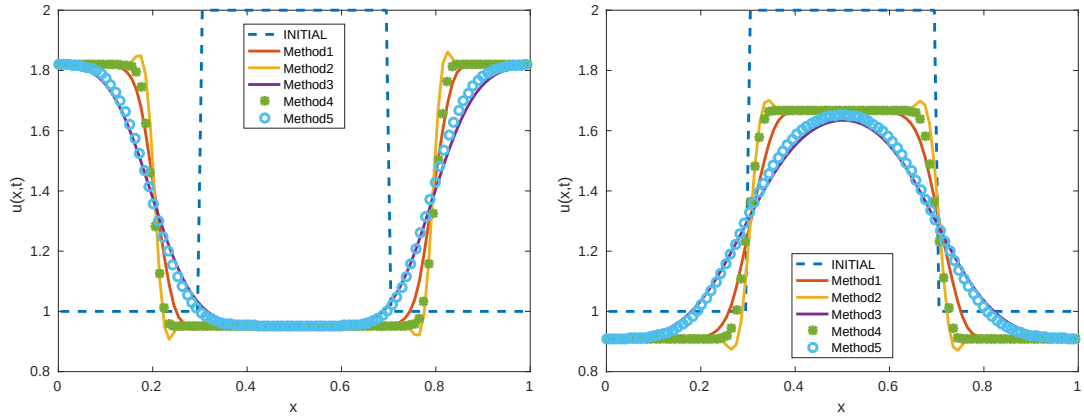

Figure 5.2: Results for the modified transport equation with initial box profile at time $t = 0.5$(left) and $t = 1$(right). $N_x = 100$ is used for the spatial resolution and $N_t = 350$ is the number of time steps (for the IBOX we use $N_t = 130$). The initial profile INITIAL is depicted in both pictures. Method1 to Method5 denote the SDIRKDG with TVD limiter, the SDIRKDG without limiter, the SDIRKFV, the SDIRKFLUX and the IBOX scheme.

Buckley Leverett equation

As a second example, we compute the numerical solution for the Buckley-Leverett equation

$$\partial_t \mathsf{u} + \partial_x \mathsf{f}(\mathsf{u}) = 0 \tag{5.8}$$

with the nonlinear flux function $\mathsf{f}(\mathsf{u}) = 3\mathsf{u}^2/(3\mathsf{u}^2 + (1-\mathsf{u})^2)$. This equation is an example for a hyperbolic equation that builds shocks even for smooth initial data. We use the sinusoidal initial condition $\mathsf{u}(x,0) = \mathsf{u}^0(x)$, where

$$\mathsf{u}^0(x) = 0.4 + 0.5\sin(\pi x) \tag{5.9}$$

for $x \in \Omega$. We show the results for all schemes except for the DG method without slope limiting in Figure 5.3 for $t = 0.15$ and $t = 1$, respectively. For the spatial resolution, we use again $N_x = 100$. In all cases except for the IBOX, we use $N_t = 130$ time steps. Note that for the implicit box scheme the lower bound

$$2\mathbf{a}\frac{\Delta t}{\Delta x} \geq 1 \tag{5.10}$$

has to be satisfied, see also (3.45). Therefore, we use $N_t = 65$ time steps and $N_x = 730$ grid points. Due to the fine spatial discretization, the IBOX resolves the shock very sharply. The reference solution is computed using the finite volume method with flux limiter for $N_x = 400$ and $N_t = 1200$.

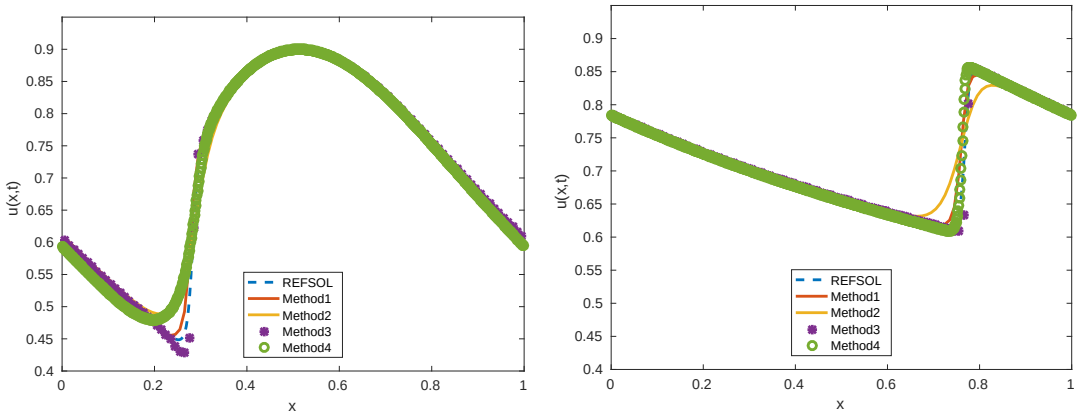

Figure 5.3: Results for the Buckley-Leverett equation with sinusoidal initial profile at time $t = 0.15$(left) and $t = 1$(right). $N_x = 100$ is used for the spatial resolution and $N_t = 130$ is the number of time steps (for the IBOX we use $N_t = 65$ and $N_x = 730$). The reference solution REFSOL, which is the SDIRKFLUX with $N_x = 400$ and $N_t = 1200$, is depicted in both pictures. Method1 to Method4 denotes the SDIRKDG with TVD limiter, the SDIRKFV, the SDIRKFLUX and the IBOX scheme.

From Table 5.2, we observe that all schemes are TVD(M). Note that SDIRKFLUX gives the best values whereas the values of the other schemes results in $\mu < 1$. Additionally,

Table 5.2: TV norm for the results depicted in Figure 5.3 at different times t and the value of μ in the last column.

Method	$\lvert \mathsf{u}_{\lvert t=\Delta t} \rvert_{TV}$	$\lvert \mathsf{u}_{\lvert t=0.5} \rvert_{TV}$	$\lvert \mathsf{u}_{\lvert t=1} \rvert_{TV}$	$\mu(\Delta t)$
SDIRKFLUX	0.9840	0.7086	0.4921	1.0000
SDIRKDGTVB	0.9641	0.6721	0.4534	0.9992
SDIRKFV	0.9809	0.6225	0.3957	0.9982
IBOX	0.9422	0.7051	0.4945	0.9931
REFSOL	0.9948	0.6921	0.4767	0.9997

the discrete maximum principle is clearly fulfilled for the range $[0.4, 0.9]$ stemming from the initial profile (5.9).

We showed that the proposed schemes lie in the class of SSP schemes for linear and nonlinear scalar test examples. The discrete maximum principle is fulfilled and shocks can be resolved satisfactorily.

5.2 The water hammer equations

Let us now turn to the numerical solution of the water hammer equations

$$\begin{aligned} \partial_t \mathrm{H} + \mathbf{a}^2 \partial_x \mathrm{Q} &= 0, \\ \partial_t \mathrm{Q} + \partial_x \mathrm{H} &= -\mathsf{g}(\mathrm{Q}), \end{aligned} \tag{5.11}$$

where $\mathsf{g}(\mathsf{u}) = \frac{\lambda(\mathsf{u})\mathsf{u}|\mathsf{u}|}{2dA}$, c.f. Chapter 2. Here, we use the laminar case for λ, see Equations (2.4) with (2.3) and (2.7). For convenience, we use $(x,t) \in \Omega \times [0,T]$ with $\Omega = [0,1]$ and $T = 1$, which corresponds to a pipe with length $L = 1$. Using the initial conditions

$$\begin{pmatrix} \mathrm{H}(x,0) \\ \mathrm{Q}(x,0) \end{pmatrix} = \begin{pmatrix} \mathrm{H}^0(x) \\ \mathrm{Q}^0(x) \end{pmatrix},$$

the analytical solution for the case $\mathsf{g} = 0$ is given as

$$\begin{aligned} \mathrm{H}(x,t) &= \frac{1}{2}\mathrm{H}^0(x - \mathbf{a}t) + \frac{\mathbf{a}}{2}\mathrm{Q}^0(x - \mathbf{a}t) - \frac{\mathbf{a}}{2}\mathrm{Q}^0(x + \mathbf{a}t) + \frac{1}{2}\mathrm{H}^0(x + \mathbf{a}t), \\ \mathrm{Q}(x,t) &= \frac{1}{2\mathbf{a}}\mathrm{H}^0(x - \mathbf{a}t) + \frac{1}{2}\mathrm{Q}^0(x - \mathbf{a}t) + \frac{1}{2}\mathrm{Q}^0(x + \mathbf{a}t) - \frac{1}{2\mathbf{a}}\mathrm{H}^0(x + \mathbf{a}t), \end{aligned} \tag{5.12}$$

c.f. Section 2.3. If no further conditions are stated, we use periodic boundary conditions in space and (5.12) needs to be understood accordingly. For simplicity, we use $d = 0.2821\,m$, $A = 1\,m^2$, $a = 1.45\,m/s$ and $\nu = 0.13\,m^2/s$ for the examples in this Section.

5.2.1 The linear system without source term

If no source term is present, we can utilize the transformation to the characteristic variables $\tilde{\mathrm{H}}$ and $\tilde{\mathrm{Q}}$. In this case, we have to solve the decoupled system

$$\partial_t \tilde{\mathrm{H}} + \mathbf{a}\partial_x \tilde{\mathrm{H}} = 0, \tag{5.13}$$

$$\partial_t \tilde{\mathrm{Q}} - \mathbf{a}\partial_x \tilde{\mathrm{Q}} = 0, \tag{5.14}$$

for $(x,t) \in \Omega \times [0,T]$ and for periodic boundary conditions. We use the initial condition $\tilde{\mathrm{H}}(x,0) = \tilde{\mathrm{Q}}(x,0) = \mathsf{u}^0(x)$, where $\mathsf{u}^0(x)$ is once again the box profile (5.6). This results in an initial profile

$$\mathrm{H}(x,0) = \mathbf{a}(\mathsf{u}^0(x) - \mathsf{u}^0(x)) = 0, \quad \text{and} \quad \mathrm{Q}(x,0) = \mathsf{u}^0(x) + \mathsf{u}^0(x) = 2\mathsf{u}^0(x) \tag{5.15}$$

for the conservative variables. Note that in this case, the analytical solution reduces to

$$\begin{aligned} \mathrm{H}(x,t) &= \mathbf{a}\Big(\mathsf{u}^0(x - \mathbf{a}t) - \mathsf{u}^0(x + \mathbf{a}t)\Big), \\ \mathrm{Q}(x,t) &= \mathsf{u}^0(x - \mathbf{a}t) + \mathsf{u}^0(x + \mathbf{a}t), \end{aligned}$$

see also equations (2.33) and (2.34). We computed the solution for the characteristic variables, which correspond to the linear transport with transport velocity $\pm\mathbf{a}$ as discussed in Section 5.1. The transformation to the conservative variables is shown in Figure 5.4. Additionally, we applied all available schemes to the coupled system (5.11) with $\mathsf{g} = 0$ and the corresponding initial conditions. The results are depicted in Figure 5.5. Note

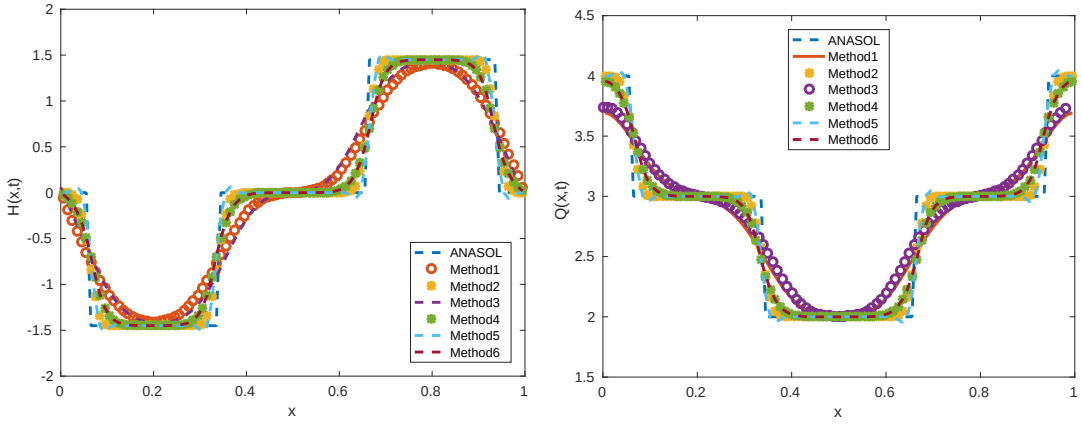

Figure 5.4: Results for the system (5.11) computed using the characteristic variables and the transformation with spatial resolution $N_x = 100$ and $N_t = 350$ (for the IBOX we have $N_t = 160$) at time $t = 0.25$. Method1 to Method6 denote the SDIRKFV, the SDIRKFLUX, the IBOX, the SDIRKDG with TVB limiter, the SDIRKDG without limiter and the SDIRKDG with TVD limiter. ANASOL represents the analytical solution at $t = 0.25$.

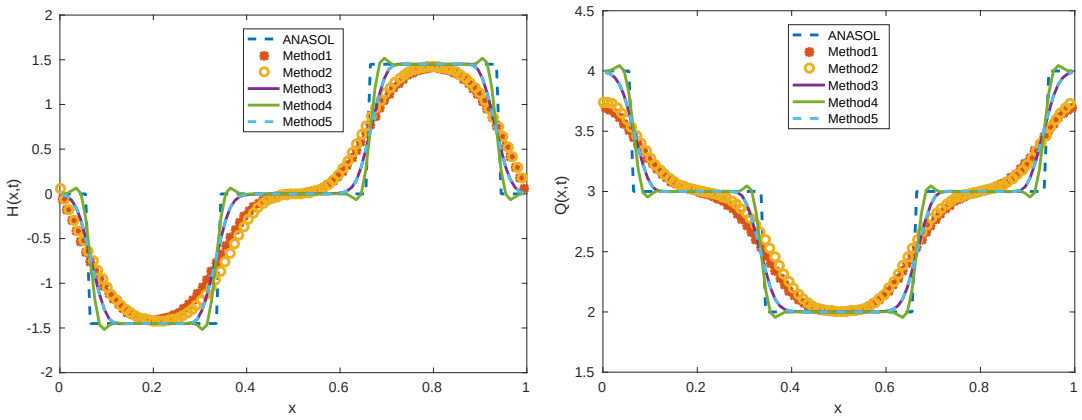

Figure 5.5: Results for the system (5.11) computed using the conservative variables with spatial resolution $N_x = 100$ and $N_t = 350$ (for the IBOX we use $N_t = 160$) at time $t = 0.25$. Method1 to Method5 denote the SDIRKFV, the IBOX, the SDIRKDG with TVB limiter, the SDIRKDG without limiter and the SDIRKDG with TVD limiter. ANASOL represents the analytical solution at $t = 0.25$.

that the schemes in both figures show partially strong numerical diffusion whereas the lower-order schemes like SDIRKFV or IBOX show more numerical diffusion than the higher-order schemes like SDIRKFLUX or SDIRKDG. This can be resolved with higher spatial resolution. The graphs clearly show that all methods except the finite volume with flux limiting can be applied to the coupled system and we can expect the same results as for the transformed characteristic case. As discussed in Section 3.4.2, the flux limiting method is not directly applicable to the coupled system and the detour over the characteristic variables has to be taken. As for the scalar case, we observe the typical numerical diffusion while the TV stability is maintained except for the SDIRKDG. Similar

as in the case of the linear transport equations, the value of μ for the SDIRKDGTVB scheme amounts to one. All values of the ratio μ are listed in Table 5.3.

Table 5.3: TV norm for the results depicted in Figure 5.4 at different times t and the value of μ in the last column. For the second component $\tilde{\mathrm{Q}}$, the results are the same.

Method	$\lvert\tilde{\mathrm{H}}_{\lvert t=\Delta t}\rvert_{TV}$	$\lvert\tilde{\mathrm{H}}_{\lvert t=0.5}\rvert_{TV}$	$\lvert\tilde{\mathrm{H}}_{\lvert t=1}\rvert_{TV}$	$\mu(\Delta t)$
SDIRKFLUX	2.0000	2.0000	2.0000	1.0000
SDIRKDGTVB	2.0000	1.9999	1.9987	1.0000
SDIRKFV	2.0000	1.9623	1.7814	1.0000
SDIRKDG	2.0495	2.3882	2.4165	1.0430
SDIRKDGTVD	2.0000	1.9999	1.9987	1.0000
IBOX	2.0000	1.9775	1.8327	1.0000

The interval limits of the range for the discrete maximum-principle are defined as

$$m = \min\{\min_x \tilde{\mathrm{H}}_0(x), \min_x \tilde{\mathrm{Q}}_0(x)\} = 1 \quad M = \max\{\max_x \tilde{\mathrm{H}}_0(x), \max_x \tilde{\mathrm{Q}}_0(x)\} = 2,$$

see assumptions in Section 4.3.3, where we use the box profile (5.6) as initial profile, as has been already mentioned above. Therefore, the characteristic variables $\tilde{\mathrm{H}}$ and $\tilde{\mathrm{Q}}$ need to lie in the range $[1, 2]$ if the discrete maximum principle is fulfilled. It can be shown numerically that it is the case for the introduced methods except for the SDIRKDG.

5.2.2 The full system

We show the results for the full water hammer equations using the numerical methods of Chapter 3. We utilize the transformation to the characteristic variables, i.e. we solve the system

$$\begin{aligned} \partial_t \tilde{\mathsf{H}} + a\partial_x \tilde{\mathsf{H}} &= -\tfrac{1}{2}\mathsf{g}(\tilde{\mathsf{H}} + \tilde{\mathsf{Q}}), \\ \partial_t \tilde{\mathsf{Q}} - a\partial_x \tilde{\mathsf{Q}} &= -\tfrac{1}{2}\mathsf{g}(\tilde{\mathsf{H}} + \tilde{\mathsf{Q}}), \end{aligned} \tag{5.16}$$

with $x \in \Omega \times [0,T]$. The source term is defined as $\mathsf{g}(\mathsf{u}) = \frac{\lambda(\mathsf{u})\mathsf{u}|\mathsf{u}|}{2dA}$, c.f. Chapter 2. We distinguish between the laminar and the turbulent case for the choice of the friction coefficient λ. Note that in both cases, the small constant ν in the definition of λ ensures that the source term is several magnitudes smaller than the other terms in (5.16). For that reason the constants for the laminar version of λ are chosen such that $\mathsf{g}(\mathsf{u}) = C\mathsf{u}$ with $C = 8\pi\nu 1000 = 0.0327$ with the kinematic viscosity of water ν introduced in Chapter 2. Note that we only show results for the laminar case in this section, since the turbulent case is not in the focus of interest considering our applications later on. For the initial conditions, we use again $\tilde{\mathsf{H}}(x,0) = \tilde{\mathsf{Q}}(x,0) = \mathsf{u}^0(x)$ with $\mathsf{u}^0(x)$ from (5.6). We show the numerical solution for the characteristic variables at $t = 0.25$ in Figure 5.6. Due

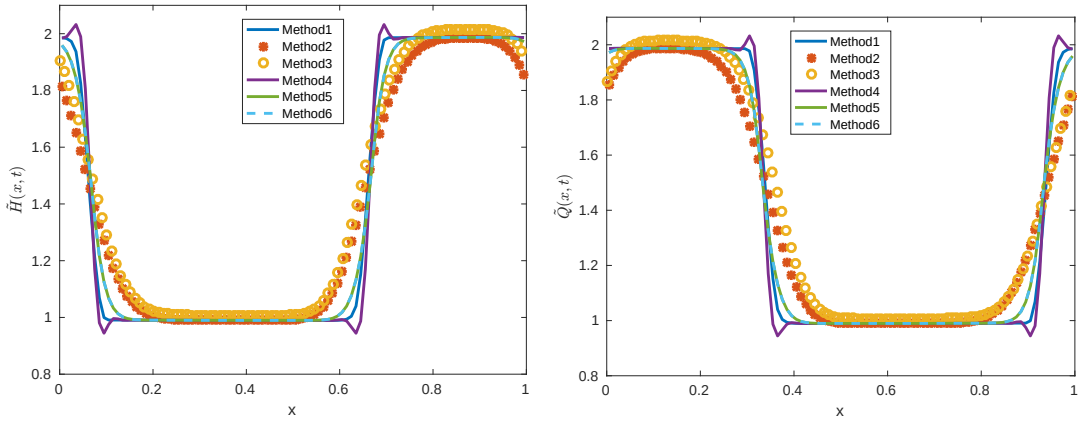

Figure 5.6: Results for the system (5.16) computed using the characteristic variables with spatial resolution $N_x = 100$ and $N_t = 350$ (for the IBOX we have $N_t = 160$) at time $t = 0.25$. Method1 to Method6 denote the SDIRKFLUX, the SDIRKFV, the IBOX, the SDIRKDG without limiter, the SDIRKDG with TVD limiter and the SDIRKDG with TVB limiter.

to the source term, we observe that the box profile is clinched. For all schemes except for the scheme without limiter, we have a TV stable scheme and all methods lie in the range $[-2,2]$. The interval limits are computed by

$$\tilde{m} = \min\{\min_x \tilde{\mathsf{H}}_0(x), \min_x \tilde{\mathsf{Q}}_0(x)\} = 1 \quad \tilde{M} = \max\{\max_x \tilde{\mathsf{H}}_0(x), \max_x \tilde{\mathsf{Q}}_0(x)\} = 2$$

and $M = \max(|\tilde{m}|, |\tilde{M}|) = 2$, see assumptions made in Section 4.3.3 and 4.3.4. The range is then as defined above. For $t \to \infty$, the solution converges to zero in all cases. Again, we depict the values of μ in Table 5.4. We observe that all schemes except the SDIRKDG and IBOX schemes have the ratio $\mu \leq 1$. Now we can again compute the results for the

Table 5.4: TV norm for the results depicted in Figure 5.6 at different times t and the value of μ in the last column. For the second component $\tilde{\mathrm{Q}}$, the results are the same.

Method	$\lvert\tilde{\mathrm{H}}_{\lvert t=\Delta t}\rvert_{TV}$	$\lvert\tilde{\mathrm{H}}_{\lvert t=0.5}\rvert_{TV}$	$\lvert\tilde{\mathrm{H}}_{\lvert t=1}\rvert_{TV}$	$\mu(\Delta t)$
SDIRKFLUX	1.9998	1.9865	1.9688	0.9999
SDIRKDGTVB	1.9998	1.9846	1.9671	0.9999
SDIRKFV	1.9998	1.9460	1.7535	0.9999
SDIRKDG	2.0493	2.3715	2.3776	1.0428
SDIRKDGTVD	1.9998	1.9846	1.9671	0.9999
IBOX	2.0004	2.0096	1.8930	1.0002

conservative variables with the transformation of the characteristic variables. For the conservative variables, we get the transformed initial profile

$$\mathrm{H}_0 = \mathbf{a}(\tilde{\mathrm{H}}_0 - \tilde{\mathrm{Q}}_0) = 0 \tag{5.17}$$

$$\mathrm{Q}_0 = \tilde{\mathrm{H}}_0 + \tilde{\mathrm{Q}}_0, \tag{5.18}$$

where Q_0 leads to a box profile with the bounds 2 and 4. Again, we want to define the limits of the range and get

$$\tilde{m} = \min\{\min_x \mathrm{H}_0(x), \min_x \mathrm{Q}_0(x)\} = 2 \quad \tilde{M} = \max\{\max_x \tilde{\mathrm{H}}_0(x), \max_x \tilde{\mathrm{Q}}_0(x)\} = 4.$$

This yields $M = \max(|\tilde{m}|, |\tilde{M}|) = 4$ and results in the range $[-4, 4]$ for the conservative variables. Here, we observe that the expanded range is indeed necessary, since the initial profile of Q has an influence on H due to the coupling. The results for the conservative variables are shown in Figure 5.7. If we compare the results in Figure 5.7 to the results

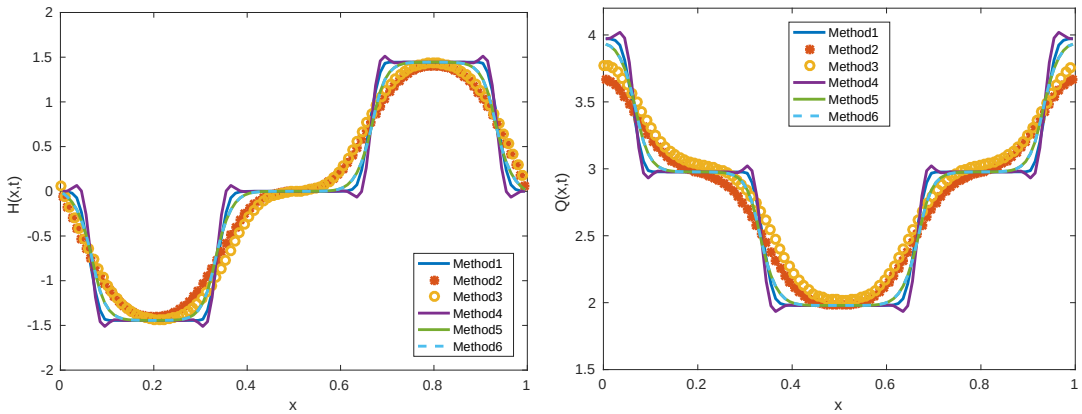

Figure 5.7: Results for the system (5.16) computed using the characteristic variables after transformation to the conservative variables with spatial resolution $N_x = 100$ and $N_t = 350$ (for the IBOX we have $N_t = 160$) at time $t = 0.25$. Method1 to Method6 denote the SDIRKFLUX, the SDIRKFV, the IBOX, the SDIRKDG without limiter, the SDIRKDG with TVD limiter and the SDIRKDG with TVB limiter.

shown in Figure 5.4, we see that the results do not differ even with the presence of the

source term. This is because the effect of the source term is much smaller than the effect of the transport term in the WHE. Again, we observe that the lower-order schemes show more numerical diffusion than the higher-order schemes and the SDIRKDG without slope limiter shows under- and overshoots at the discontinuities.

To numerically account for the fact that all schemes are well-balanced, we insert the stationary state (2.19) and (2.20) into our schemes as initial condition. We get the results showed in Figure 5.8 for $T = 1$. We see that all schemes stay in this stationary state and therefore, they are well-balanced. If we use a long enough time horizon, we can observe

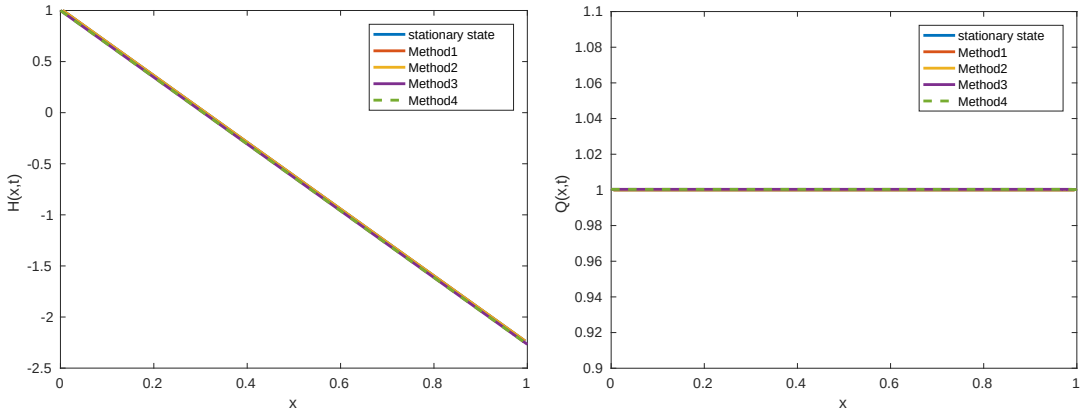

Figure 5.8: Stationary state of WHE for Method1 to Method4 with resolution $N_x = 100$ and $N_t = 350$ at time $T = 1$. Method1 to Method4 represent the SDIRKFV, the SDIRKFLUX, the SDIRKDG without limiter and the SDIRKDG with TVD limiter.

that all schemes are also asymptotically stable, which is another important property of numerical schemes.

In this chapter, we showed numerical results using our introduced numerical methods. For the scalar case, we considered the transport equation and the Buckley-Leverett equation to verify the TV stability of the schemes. For the system case, we showed numerical results of the WHEs with and without source term. For all schemes we verified a discrete maximum principle, where we used the characteristic variables in the WHEs system case. Finally, we showed the well-balancedness of all methods numerically.

6 EWave

The project EWave[1] is part of the cooperation project ERWAS[2] funded by the BMBF[3] and pursues the development of an efficient energy management system for water supply networks. The overall goal is to develop a system which provides energy-optimized operation plans for water supply facilities. The result is an innovative and cognitive energy management system. To obtain satisfactory results, we need to handle the processes of water extraction, water preparation and water distribution. The resulting system should moreover enable companies to handle the balancing act between the increasing requirements of energy efficiency, quality of drinking water and security of supply. Several partners are involved in this project. These cooperation partners are

- Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Prof. Dr. Alexander Martin, Dr. Antonio Morsi, Dr. Björn Geißler
- Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Prof. Dr. Günther Leugering, Maximilian Walther
- Hochschule Bonn-Rhein-Sieg (HBRS), Prof. Dr. Gerd Steinebach, Tim Jax, Patrick Haussmann, David Dreistadt
- Techniche Universität Darmstadt (TU DA), Prof. Dr. Jens Lang, Lisa Wagner
- Universiät Mannheim (U MA), Prof. Dr. Oliver Kolb
- RWW Rheinisch-Westfälische Wasserwerksgesellschaft mbH, Dr. Michael Plath, Stefan Fischer, Ronald Roepke, Martin Launer
- Siemens AG, Process Industries and Drives, Vertical Sales, Dr. Andreas Pirsing
- Siemens AG, Corporate Technology Research in Digitalization and Automation, Modeling and Simulation Technologies, Dr. Annelie Sohr, PhD Moritz Allmaras, Tim Schenk
- Bilfinger GreyLogix aqua GmbH, Olaf Kremsier.

6.1 Description of the project

Let us describe the project in more detail. The EWave assistance system should be able to provide an energy-optimized operation procedure based on the operation plans of the water work and the connected water supply network. In Figure 6.1, we see a prototypical example of the water work Holsterhausen, see also Section 6.5.1 for a detailed description. To provide an energy-optimized operation procedure, we need to take into consideration the interaction of all active facilities, e.g. given boundary conditions on the one hand and also the complexity of the energy market and energy consumption of the network

[1] **E**nergiemanagement **Wa**sser**ve**rsorgung
[2] Zukunftsfähige Technologien und Konzepte für eine **e**nergieeffiziente und **r**essourcenschonende **Was**serwirtschaft
[3] Bundesministerium für Bildung und Forschung

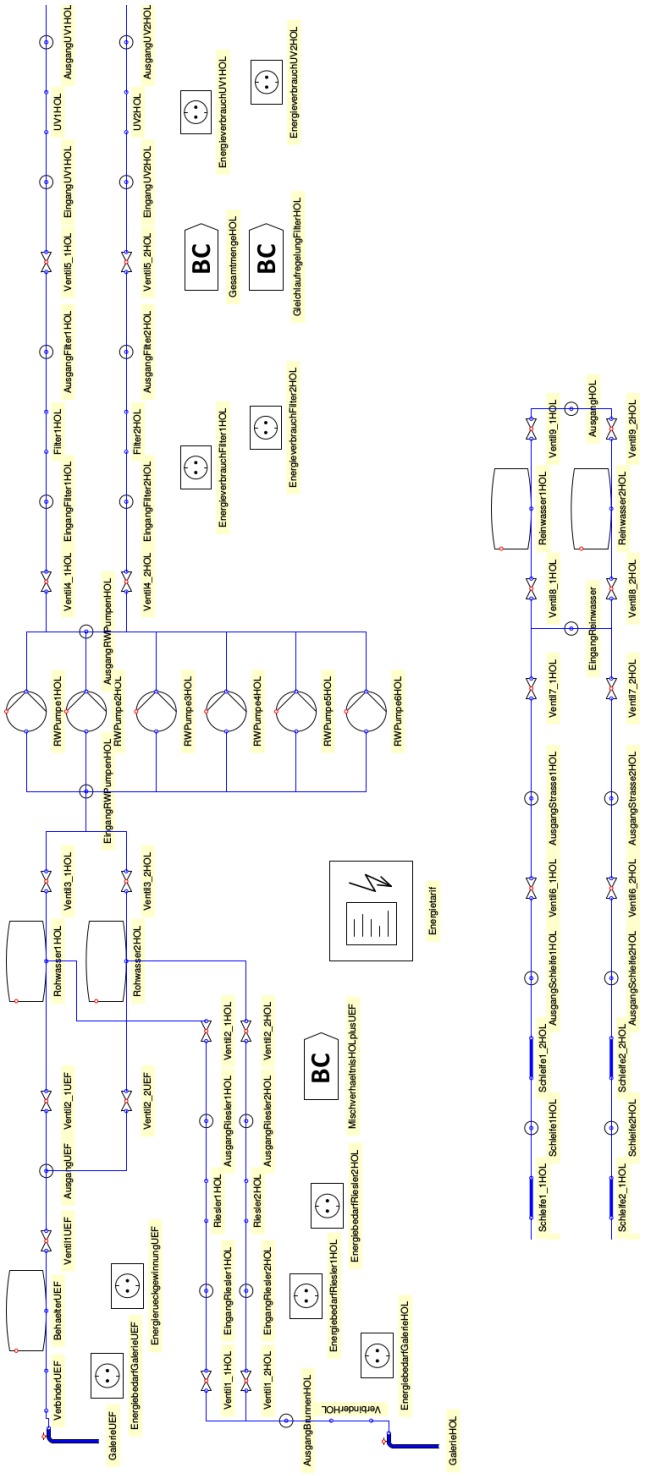

Figure 6.1: Part of the water supply network of RWW with 5 tanks and 6 pumps. In the water work Holsterhausen water is extracted from two wells *GalerieHOL* and *GalerieUEF*. After that, it is transported through different tanks and pumps and then flows through two filter streets where it is prepared and finally distributed. To manage this process there are a lot of network components like tanks, pumps, valves and connections necessary which are marked with different symbols, here. There also are other symbols which describe boundary conditions of the network or energy consumptions of single components. The picture was provided by Annelie Sohr from Siemens AG.

on the other hand. In particular, the current electricity rate is taken into account by the running system, which should also lower the cost for the water supply companies.

For the optimization process performed by the EWave system, several technical and operational restrictions apply. Some of those restrictions like for example water quality or security of supply must strictly be adhered to while others may be relaxed if necessary. The output, i.e. the optimized operation plans, are given in terms of runtime or switching time points of network pumps, respectively. The distribution of the required production output over the available water works is controlled as well. To achieve the above mentioned aims, EWave is divided into different major tasks which are administrated by different project partners.

In the first major project task, water supply and distribution requirements of the simulation-based assistance system, which, as already mentioned above, also has an integrated optimization tool for energy efficiency, were developed and determined. Further, so-called *KPIs* (**K**ey **P**erformance **I**ndicators) were developed to evaluate the energy efficiency of the water companies. To obtain practically relevant results, all developments are being discussed with other water supply companies in the form of workshops. Further, different measure stations were established to deliver the right data to the EWave system.

The second major project task deals with the preparation and evaluation of structure, process and external data as well as with the methods for modelling development, for model operations and for evaluation of results with respect to the energy efficiency in EWave. With real data from RWW, applicable methods for the automatized creation of abstract water supply networks as well as for the half-automatized model calibration and for the computation of the actual network state are derived. As far as model operation is concerned, a forecasting of the demand for drinking water at each pressure zone based on measurements and statistical analysis is necessary, see [33]. With the coupling of the water and energy models, the evaluation of the energy efficiency of the operation control can be considered. Therefore, cost functionals of the different network components are derived from energy acquisition data.

This work and in particular this chapter is integrated in the third major project task. Here, we deal with the development of dynamical simulation and optimization models for transport processes in pipes of water supply networks. For this purpose, we carry out a modular assessment of the technical facilities of water supply. Based on an integrated water and energy model collection, we use adjoint-based optimization methods to gain an energetic observation of single network components as well as of entire networks in a water supply system. We pay special attention to the practically required real-time optimization including switching operations.

In the fourth major project task, concepts for integrated decisions and operation support based on simulation and mathematical optimization of concrete use cases were developed. For this purpose, a decision-based optimization tool was integrated into the simulation-based system EWave. What is important in this respect is coupling the decision-driven optimization and the simulation in order to obtain a physically correct system. This system is then capable of modeling switching dynamics, transport processes and energy couplings with other networks and also of optimizing operation plans with respect to integrated efficiency.

The final major project task deals with the pilot application of the assistance system EWave at RWW. With this application, we test the practical use of the prototypically implemented simulation-based offline- and online assistance system on the one hand and verify its usage based on the previously defined test scenarios on the other hand. Altogether, the EWave system can be used as a planning tool for conceptual analyses. Further, EWave is designed as an assistance system which enables an energy and cost optimized operation plan of water supply networks with respect to the simulation and mathematical optimization approaches.

6.2 Description of sequential control – EWave system

In this section, we describe the EWave assistance system [72]. The main steps here are to establish the mathematical models for the single components, and the description of the network model as directed finite graph. Consequently, all network components will be modeled either as a edge or a node of the graph.

In a first step of the assistance system called *EWave Engineering*, all edges, nodes and energy components of the network are loaded. The initial state of the system is computed from measurements and from the drinking water demand prognosis, which are provided by the water work. This is already part of the *EWave Initializing* step, where the simulation (SIMT) and optimization tool are called in sequence. The simulation tool determines the initial state of the network. In this step, an aggregation of the network is used to reduce the computational costs. As an example, we show the original pipe network of RWW with 15114 pipes on the left and the aggregated pipe network with 98 pipes on the right in Figure 6.2. Note that the aggregation of the network is done beforehand in an half-automated step. Therefore, the EWave system deals with a condensed network model, in which important structures are identified.

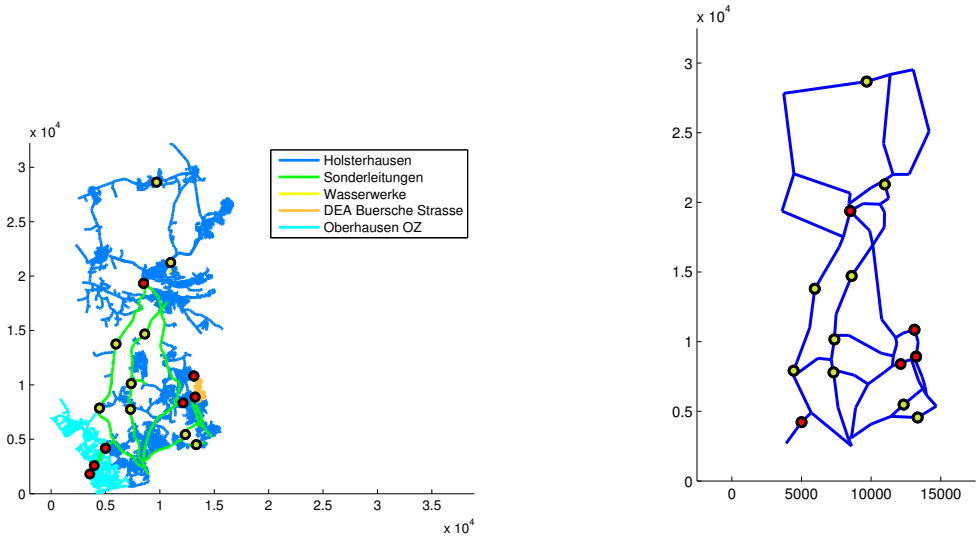

Figure 6.2: Left: The whole water network of RWW with 15114 pipes. Right: The aggregated water network of RWW with 98 pipes. The unit of the x- and y-axis are meters. Both pictures were provided by Gerd Steinebach and Tim Jax from HBRS.

The next step is to call the optimization tool which consists of two subsequent calls, namely the discrete (DOPT) and the continuous optimization tool (COPT), see Figure 6.3. Utilizing the complete state information, DOPT computes all discrete decisions

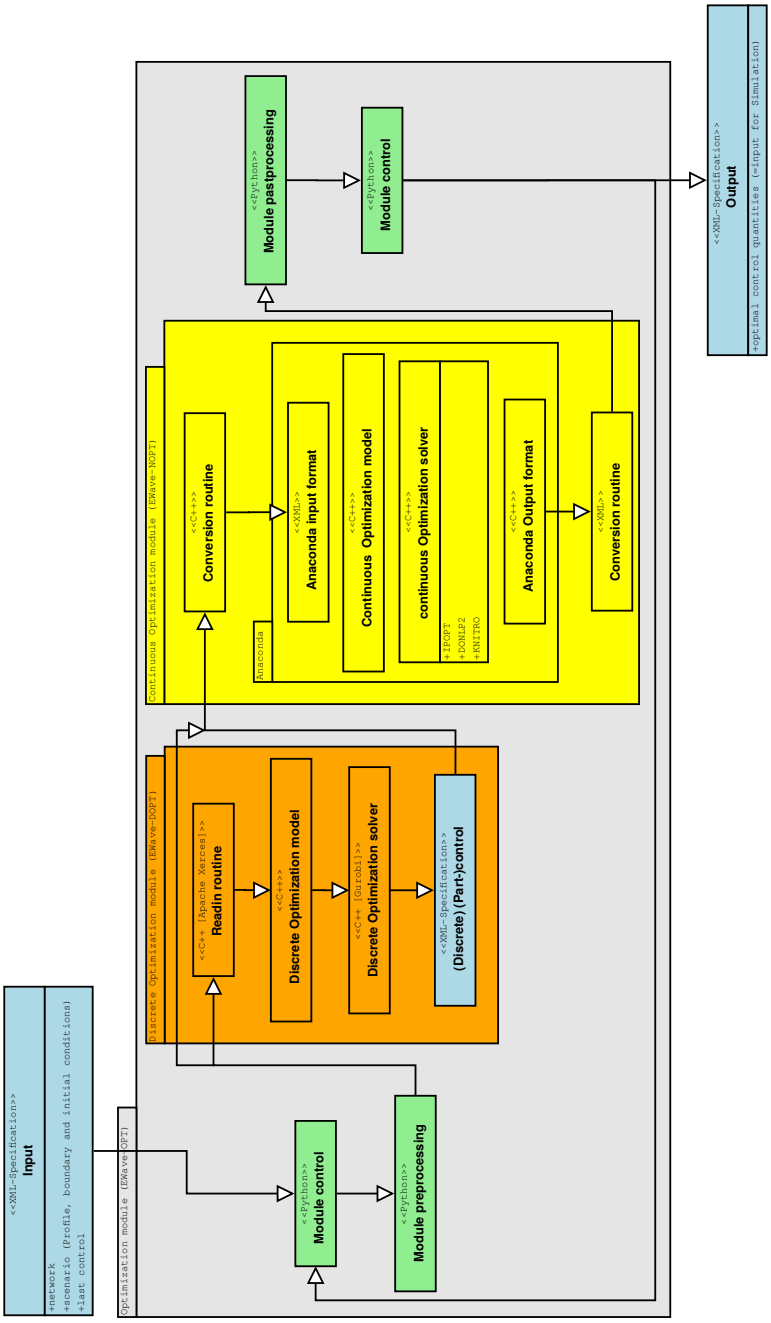

Figure 6.3: Optimization module of EWave divided into two parts – DOPT and COPT. With the input parameters computed by SIMT, DOPT computes the discrete control with the help of a discrete optimization solver. After this step, the continuous optimization module COPT is called. With a chosen continuous optimization solver and with the discrete control as input data, COPT computes the optimal solution and the optimal control quantities, respectively. The picture was provided by Antonio Morsi from FAU.

such as the switching points of controllable pumps for a given time horizon, e.g. 12 or 24 hours. Note that even though DOPT only controls the fixed speed pumps, it also determines the switching points of the continuous controllable pumps. For the mathematical description of the network components, DOPT uses so called quasi-stationary equations which are in a way simplified compared to the model equations of SIMT and COPT. The latter are described in detail in the next Section. More about the mathematical models of the DOPT module can be found in [58]. Once the discrete decisions are computed, COPT is called. Here, the input parameters are again the complete initial state and also the initial control provided by DOPT. By solving a continuous nonlinear optimization problem, all controls are eventually determined.

The continuous optimization process, which uses the software tool ANACONDA[4], is described in more detail in Section 6.4. Before that, let us comment on the possible appearance of infeasible solutions. Since different software tools are coupled in this cycle, it can for instance occur that the initial state is not feasible for the optimization problem. This can, to some extend, be overcome by an automatic relaxation of certain constraints. Therefore, ANACONDA is first called to do a feasibility check. After that, the possibly violated constraints are relaxed for the first thirty minutes. Then, ANACONDA is called a second time to compute the optimal solution. If the solution computed by COPT is infeasible, EWave uses the optimal results from DOPT to give an overall result. More about the optimization tool can be found in [23]. Finally, all results are transferred to the master display where the control room operators see the switching operations of the pumps or the degree of opening of controlled valves in the message window of EWave. Figure 6.4 shows which external sources and computing modules are necessary for the complete process. For example, the simulation module can compute the initial states with the measure data as input parameters. Further, the optimization module computes the optimal solution with respect to the given constraints and input data of the simulation module.

[4] **A**daptive **N**umerical **A**lgorithms for **C**ontrol **O**ptimization on **N**etworks **DA**rmstadt [42]

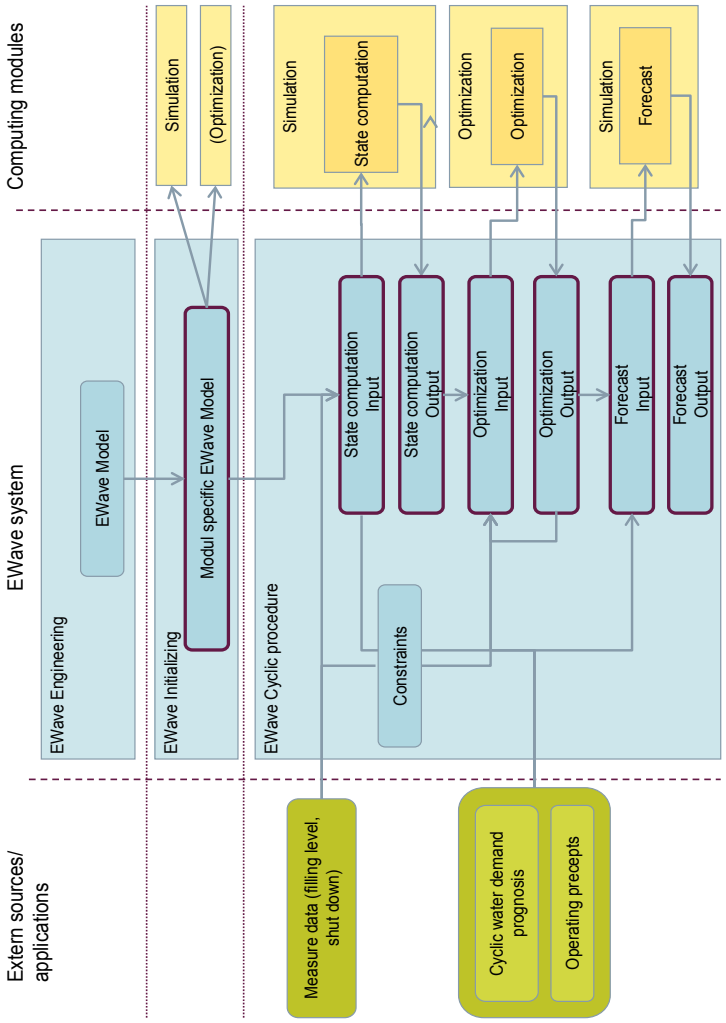

Figure 6.4: The overall EWave process with measurement data, SIMT, DOPT and COPT as single steps. After the EWave Engineering and the EWave Initializing step, the EWave Cyclic procedure is called. For this procedure, different extern sources (for instance measure data and cyclic water demand prognosis) and computing modules are necessary. For example, the simulation module (SIMT) is responsible for the state computation and the forecast computation after the optimization (DOPT and COPT). The picture was provided by Annelie Sohr from Siemens AG.

6.3 Modelling of water supply networks

Before we describe the simulation and optimization tool ANACONDA in more detail, we need suitable models for the description of water supply networks. We have already discussed the modelling of water flow through single pipes in the first part of the thesis. It remains to be clarified how those are coupled with different network components and what the modelling of the other network components look like [42, 44].

6.3.1 The network

We model the network as a finite directed graph G = (V,E) with vertices V and edges E, c.f. Figure 6.5. For example, the sets $\delta^-_{v_2}$ and $\delta^+_{v_2}$ describe the ingoing and outgoing edges at the node v_2. The arcs, i.e., the directed edges, are used to model network

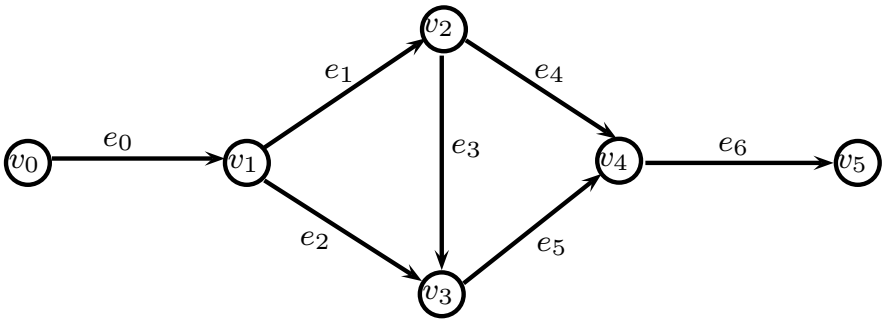

Figure 6.5: Network with edges $E = \{e_0, e_1, ..., e_6\}$ and vertices $V = \{v_0, v_1, ..., v_5\}$, e.g., $\delta^-_{v_2} = \{e_1\}$ and $\delta^+_{v_2} = \{e_3, e_4\}$

components like pumps, valves and pipes. All other components like, e.g., tanks, suppliers and consumers, are modeled as vertices which are called nodes. We distinguish between boundary and coupling nodes. Most components are modeled using ordinary differential equations or in particular algebraic equations. The only exception is the modelling of pipes which is described by hyperbolic partial differential equations, see Chapter 2. For the connection of the different components, the use of suitable coupling conditions is therefore an important ingredient.

6.3.2 Network components

Let us describe the single network components in more detail. For convenience, all components are described using the same state variables, namely the flow rate q and the piezometric head $h = z_0 + \frac{p}{g\rho_0}$. We denote the elevation of a node by z_0, the gravitational acceleration by $g = 9.81\frac{m}{s^2}$, the pressure by p and the density of water by $\rho_0 = 1000\frac{kg}{m^3}$. This section gives a short summary of common network components and their mathematical modelling, see also [44].

Inner nodes and coupling conditions

Coupling nodes, as the name already describes, are responsible for coupling two or more edges. As nodes they do not have any physical expansion. The most reasonable coupling conditions are therefore the conservation of mass and the continuity of the piezometric heads of all in- and outgoing edges. Utilizing the notation depicted in Figure 6.5, we denote the index set of in- and outgoing arcs for a vertex $v \in V$ by δ_v^- and δ_v^+, respectively. With each connection $e \in E$, we associate an disjoint interval $[x_e^a, x_e^b]$. Note that every connection appears exactly once as an ingoing and once as an outgoing arc, i.e. there exist unique v_1, v_2 such that $e \in \delta_{v_1}^-$ and $e \in \delta_{v_2}^+$.

The conservation of mass at each node v can then be expressed as

$$q(v,t) = \underbrace{\sum_{j \in \delta_v^+} q(x_j^b, t)}_{=:q_{out}(v,t)} - \underbrace{\sum_{i \in \delta_v^-} q(x_i^a, t)}_{=:q_{in}(v,t)} = 0. \tag{6.1}$$

The continuity of the piezometric heads for all in- and outgoing edges then reads

$$\begin{aligned} h(x_i^a, t) =: h_{in}(v,t) = h(v,t) \quad & \text{for all } i \in \delta_v^- \\ h(x_j^b, t) =: h_{out}(v,t) = h(v,t) \quad & \text{for all } j \in \delta_v^+. \end{aligned} \tag{6.2}$$

These conditions hold for all inner nodes which just serve as a connection of different components.

Consumers and suppliers as boundary nodes

Boundary nodes in the network are so-called consumers and suppliers. Consequently, all boundary conditions have to be specified at these nodes. Here, the demand of consumers and the capability of the water works can be incorporated. Depending on the connected edge, either the flow rate or the piezometric head are given in terms of a fixed, but time depending, profile. In particular, the same equations as for inner nodes can be used with $q(v,t)$ and $h(v,t)$ given by the user.

Storage Tanks

However, nodes exist which are capable of storing water. Those nodes are modelling components called *storage tanks*. For the modelling, not only the incoming flow rate and the piezometric head from the neighbouring edge are important, but also the so-called *inner pressure head*

$$h_{inner}(t) = z_0 + s(t) \tag{6.3}$$

with z_0 being the constant head due to the elevation of the storage tank and $s(t)$ being the time-dependent filling level of the tank. The change of $h_{inner}(t)$ is modeled by the ordinary differential equation

$$\partial_t h_{inner}(t) = \frac{1}{A}\partial_t V(t) = \frac{1}{A}\Big(\sum_{j \in \delta_v^+} q(x_j^b, t) - \sum_{i \in \delta_v^-} q(x_i^a, t) + r(t)\Big), \tag{6.4}$$

where $V(t)$ is the volume of the contained water and A the cross-sectional area of the storage tank which we assume to be constant.
As above, $q(x_i^a, t)$ and $q(x_i^b, t)$ denote the flow rates at the neighbouring edges. The variable $r(t)$ denotes some manual in- or outflow, which could also model some supplier and consumer. For example, $r(t)$ could denote the ground water inflow if the storage tank describes a water well. In this case, we have

$$r(t) = A\frac{h_G(t) - h_{inner}(t)}{T}, \tag{6.5}$$

where $h_G(t)$ describes the piezometric pressure head of the ground water and T describes the regeneration time of the well.
Besides Equation (6.4), we need an additional coupling condition for each connected edge. Thus, we assume a relationship between the storage inflow $q_i = q(x_i^b, t)$ or storage outflow $q_i = -q(x_i^a, t)$ and the corresponding pressure head $h_i = h(x_i^{a/b}, t)$ of the form

$$q_i = C\,\mathrm{sgn}(h_i - h_{inner})\sqrt{|h_i(t) - h_{inner}(t)|}, \tag{6.6}$$

where C denotes the discharge coefficient and h_i the pressure head of the connected edge. Using an adjoint-based optimization approach, it is not advantageous to use model equations with square roots, since we have to compute derivatives. Therefore, we use equivalent model equations with respect to the state variables of the form

$$\begin{aligned} q_i(t)|q_i(t)| &= C^2(h_i(t) - h_{inner}(t)) \\ h_i(t) - h_{inner}(t) &= \tilde{\zeta} q_i(t)|q_i(t)| \end{aligned} \tag{6.7}$$

with $\tilde{\zeta} = 1/C^2$ instead of Equation (6.6). Altogether, we have the state Equations (6.4) and (6.7) for the storage tanks.

Water flow through pipes

To model the water flow through pipes, we use the water hammer equations

$$\partial_t h(x,t) + \frac{\mathbf{a}^2}{\mathbf{g_r}\mathcal{A}}\partial_x q(x,t) = 0 \tag{6.8}$$

$$\partial_t q(x,t) + \mathbf{g_r}\mathcal{A}\partial_x h(x,t) = -\lambda(q(x,t))\tfrac{q(x,t)|q(x,t)|}{2d\mathcal{A}}, \tag{6.9}$$

which were already introduced and analysed in Chapter 2. If we omit the time derivatives in the above equations, we get the (quasi-)stationary model

$$\begin{aligned} q(t) &= q_{in}(t) = q_{out}(t) \\ h_{in}(t) - h_{out}(t) &= \lambda(q(t))\tfrac{L}{2\mathbf{g_r}d\mathcal{A}^2} q(t)|q(t)|, \end{aligned}$$

where $L = x_r - x_l$ is the length of the pipe, which corresponds to Equations (2.19) and (2.20). Note that pipes or in general edges have to be connected with nodes at both ends. With the help of these nodes, we can construct a geographical allocation of the pipe.

Connections

Connections are fictive elements which correspond to short pieces of pipes. Their only purpose is the connection of two nodes v and w. We therefore assume that no pressure loss occurs, which can be expressed as

$$h(v,t) := h_{in}(t) = h(w,t) := h_{out}(t) \tag{6.10}$$

for each connection (v,w). As usual, also mass (flow) conservation, i.e. $q(v,t) := q_{in}(t) = q(w,t) := q_{out}(t)$, is assumed. Note that these connections do not describe any physical process.

Note that in some situations, it is advantageous to have special connections with a characteristic curve for the pressure head. To model a filter or tricklers[5] in water works, connections are used which have a pre-defined pressure loss ΔH. Thereby, ΔH can depend on the flow rate q such that

$$\Delta H(q) = h_{in}(t) - h_{out}(t) = \alpha_0 + \alpha_1 q + \alpha_r q^r. \tag{6.11}$$

Instead of the specification of the pressure loss, we can also define a characteristic curve $H(q) = \alpha_0 + \alpha_1 q + \alpha_r q^r$ with

$$h_{in}(t) = H(q) \quad \text{or} \tag{6.12}$$
$$h_{out}(t) = H(q) \tag{6.13}$$

at in- or outgoing directions. Note that in these cases Equation (6.10) is substituted by Equations (6.11), (6.12), or (6.13).

Valves

One of the most important components for control of water flow in the water network are valves. Different types of valves exist, which control the amount or direction of water flow using different techniques. For all valves we have

$$q_{in}(t) = q_{out}(t), \tag{6.14}$$

while we denote the flow rate by q in the model equations. In the following, we list a few important valve types used in our test cases.

Gate valves: Gate valves are controlled by the time-dependent control variable

$$u(t) := \frac{A_s(t)}{A}, \tag{6.15}$$

where $u(t) \in [0,1]$ is the opening degree, A denotes the cross-sectional area and A_s denotes the opened area. If the gate valve is not closed, i.e., $q, A_s \neq 0$, we have

$$v_s = \frac{q}{A_s}. \tag{6.16}$$

[5] in German *Rieseler*

for the flow velocity through the valve. In the (partially) opened state the pressure loss ΔH satisfies

$$\Delta H = h_{in}(t) - h_{out}(t) = \frac{\zeta}{2g} v_s(t)|v_s(t)|, \tag{6.17}$$

where ζ denotes the pressure loss coefficient. Utilizing Equation (6.16), we get

$$\left(\frac{A_s}{A}\right)^2 h_s = \frac{\zeta}{2gA^2} q|q|. \tag{6.18}$$

Combining all, we get the model equation for the gate valve

$$u(t)^2(h_{in}(t) - h_{out}(t)) = \frac{\zeta}{2gA^2} q(t)|q(t)| = \tilde{\zeta} q(t)|q(t)|. \tag{6.19}$$

Check valves: Check valves allow the flow only in one direction, which is controlled by a pressure difference. Consequently, if the pressure on the outflow side is greater than the pressure on the inflow side, i.e.,

$$h_{in}(t) - h_{out}(t) < 0,$$

the flow has to be zero, i.e., $q(t) = 0$. In the case

$$h_{in}(t) - h_{out}(t) \geq 0,$$

there is a non-negative flow, which satisfies the relation

$$h_{in}(t) - h_{out}(t) = \tilde{\zeta} q(t)|q(t)|.$$

As above, the pressure loss coefficient $\tilde{\zeta}$ can depend on the flow velocity. Both cases can be summarized with the equation

$$(h_{in}(t) - h_{out}(t))_+ = \tilde{\zeta} q(t)|q(t)| \tag{6.20}$$

with $(x)_+ := \max(x, 0)$.

Control valves: Control valves, as the name suggests, enable us to control the state variable, i.e. the flow rate or the pressure head. More precisely, the degree of opening $u(t)$ is chosen in such a way that the required flow rate or pressure head is met. There are three possibilities to control the degree of opening of the control valve. We either can take the flow rate q or the inlet pressure head $h_{in}(t)$ or outlet pressure head $h_{out}(t)$, as control quantity. Each control quantity represents the value, we want to steer and $q_0(t)$ or $h_0(t)$ are the values we want to reach. We call the latter quantity the target quantity.

We formulate a first approach. We assume that the change of opening $u'(t)$ is proportional to the difference of the control quantity to the target quantity h_0. Further, we need to consider that the degree of opening is limited. Therefore, we set

$$u = \max\{CV_{\min}, \min\{u, CV_{\max}\}\}$$

with the minimal degree of opening $CV_{\min} = \epsilon > 0$, e.g. $\epsilon = 10^{-6}$ and the maximal degree of opening $CV_{\max} = 1$. Thus, we introduce contributing factors which are meant to prevent $u(t)$ from leaving the allowable domain. For the control quantity $h_{out}(t)$, the equation reads

$$u'(t) = \frac{h_0 - h_{out}(t)}{\alpha}(f^+(1 - u(t)) + f^-(u(t) - u_{min})); \quad u(0) = u_0 \tag{6.21}$$

with the factors

$$f^+ = \frac{\operatorname{sgn}(h_0 - h_{out}(t)) + 1}{2}, \quad f^- = \frac{1 - \operatorname{sgn}(h_0 - h_{out}(t))}{2}.$$

For example, if the control target $h_0 > h_{out}$, we get $f^+ = 1$ and $f^- = 0$. Consequently, the right-hand side of Equation (6.21) is positive and its limit is $u(t) \to 1$ with $t \to 0$. Thus, the bound $u(t) \leq 1$ is satisfied. In the stationary state, the control guarantees the observance of the control target $h_{out} = h_0$, provided that the control target can be achieved with the degree of opening $u \in [CV_{\min}, CV_{\max}]$. For the other control targets h_{in} and q, we have to adapt Equation (6.21). In ANACONDA, we implement the change of the degree of opening, see Equation (6.21), slightly different. Shortly explained, we use a linear differential equation for the change of the degree of opening, where we determine its right-hand side in such a way that the restrictions of u are satisfied during our computations.

Besides that, we additionally get the equation describing the pressure loss by

$$\Delta H(q, u) = h_{in}(t) - h_{out}(t) = \frac{u|u|(h_{out} - h_{in})}{\tilde{\zeta}(q, u)|q| + q}, \tag{6.22}$$

where $\tilde{\zeta}(q, u)$ is the function describing the pressure loss coefficient. Note that in (6.22), we sometimes have to use smoothed functions for terms like $|u|$ to omit divisions by zero.

Altogether, we have to distinguish between gate valves, check valves and control valves. If a gate or control valve is located behind a pump, the valve must not be closed if the pump is switched off. We therefore need to check if the bound $u(t) \geq CV_{\min}$ is satisfied. Numerical tests show that $CV_{\min} = \epsilon = 10^{-6}$, as mentioned above, is a good choice for the lower bound. Another option is to not define the pressure loss coefficient $\tilde{\zeta}$ by a fixed value ζ_0, but by a characteristic curve

$$\tilde{\zeta}(q, u) = \zeta_0 + \alpha_1 q + \alpha_r q^{r_q} + \beta_1 u + \beta_r u^{r_u}. \tag{6.23}$$

It also makes sense that the value $\tilde{\zeta}$ is restricted by $10^{-3} \leq \tilde{\zeta} \leq 10^3$, for example.

Pumps

Let us discuss different kinds of pumps, which are the main controlling facilities for the flow in the network. There are both pumps with fixed speed, which can only be switched on and off, and pumps with variable speed, which admit a continuous control. In both cases, we have again

$$q_{in}(t) = q_{out}(t) = q. \tag{6.24}$$

There are different approaches for characteristic curves of pumps in the literature. In [55, p. 389], the pressure increase H of a pump with fixed speed is modeled by

$$H(q) := h_{out} - h_{in} = \alpha_0 + \alpha_1 q + \alpha_2 q^2,$$

where $\alpha_0 > 0$ and $\alpha_1, \alpha_2 \leq 0$. We use higher-order models, like in [30, p. 44-45], where the lower-order terms are omitted, i.e.,

$$H(q) := h_{out} - h_{in} = \alpha_0 + \alpha_r q^r \tag{6.25}$$

for some $r \in \mathbb{R}^+$. In [30], the following approach for pumps with variable speed n is chosen,

$$H(q, n) := h_{out} - h_{in} = n^2\alpha_0 + n\alpha_1 q + \alpha_2 q^2. \tag{6.26}$$

Another possibility is that the coefficients α_0, α_1, α_r and r linearly depend on the speed. The simplest way to construct this is by declaring a minimal and maximal speed n_0 and n_1 and by also declaring a coefficient set with respect to this speed. Consider, for example, the pump with current speed n. The current coefficient α_0 is computed with

$$\alpha_0 = \alpha_{0,n_0} + \frac{n(t) - n_0}{n_1 - n_0} \left(\alpha_{0,n_1} - \alpha_{0,n_0}\right). \tag{6.27}$$

To define the operating status of a pump, a fixed value is defined for pumps with fixed speed. This value lies in the interval $[0, 1]$ with 0 if the pump is off and with 1 if the pump is on. Pumps with variable speed can attain the value 0, 1 or n (which means that the pump operates with speed n). Besides the direct specification of the speed, we can also formulate it automatically subject to the other system quantities. In practice, we use h_{is} of the pump or the pressure head of network components which lie behind the considered pump. In this work, we only consider the pressure head at the exit of the pump or in a storage tank. The control follows from the fact that the speed $n(t)$ is increased if the target pressure h_{target} is undershot, or the speed is decreased if the target pressure is exceeded. Consequently, we get the ordinary differential equation

$$n'(t) = \frac{h_{target} - h_{is}(t)}{\alpha}(n_1 - n_0) \tag{6.28}$$

for the change of speed, where the constant α depends on the characteristic curves of the pump. Additionally, we apply a projection, such that the so-computed speed always lies in the interval $[n_0, n_1]$ and the induced flow rate of the pump cannot become negative. With the parameter α, we can formulate the delay time of the control.

Power consumption:
One aim of the project EWave is to define and optimize the energy consumption of the water network. The energy of the network is mostly consumed by pumps and connections which have pressure characteristic curves, e.g. UV-irradiation. We therefore need to model this consumption and make the following assumptions:

- in principle every network element can consume energy in the form of electrical power P_{el},
- the power P_{el} can depend on the flow q through the element,
- energy consumption has no influence on the hydraulics of the network and can thus be computed after the simulation of the network,
- energy recovery is denoted by a negative value for P_{el},

The computation of the electrical power P_{el} in a network element follows either a characteristic curve

$$P_{el} = \beta(t)(\alpha_0 + \alpha_1 q + \alpha_r q^r) \tag{6.29}$$

or a q-dependent step function

$$P_{el} = \beta(t) \sum_{i=1}^{n} p_i \chi_{[q_i, q_{i+1})}(q) \quad \text{with} \quad \chi_{[q_i, q_{i+1})}(q) = \begin{cases} 1, & q \in [q_i, q_{i+1}), \\ 0, & \text{otherwise.} \end{cases} \tag{6.30}$$

For the definition of the step function, we need value pairs (q_i, p_i) with $i = 1, ..., n$ and $q_1 < q_2 < ... < q_n$. We assume that $q_{n+1} = \infty$. The factor $\beta(t)$ is necessary to turn the power consumption of a network component on ($\beta(t) = 1$) or off ($\beta(t) = 0$). We can also consider electrical power, which is proportional to a variable energy contract. Therefore, we set $\beta(t) = C(t)$, where $C(t)$ describes the time-dependent function for the energy contract. For example, if a pump is turned off after several hours, the value of β should be zero. With the choice of $\alpha_0 = 1$, $\alpha_1 = \alpha_r = 0$ and with the value $\beta(t)$, we can assign electrical power to the network element. If we have computed the characteristic curve indicating the power of a pump, we can compute its efficiency with

$$\eta = \frac{\epsilon q \Delta p}{P_{el}}, \tag{6.31}$$

where the numerator contains the small constant ϵ, the hydraulic power q and the pressure difference Δp.

6.4 Network Simulation and Optimization Tool – ANACONDA

In this section, we describe the software tool **ANACONDA** which is integrated in the EWave module COPT. The software package, which was developed in [42], mainly consists of two parts, namely the simulation and optimization tool.

6.4.1 Simulation Tool

ANACONDA is capable of simulating a network, consisting of the components described in the previous section. Therefore, the numerical approximation of the ODEs, PDEs and (D)AEs has to be handled. For the simulation of the network, we need initial and boundary conditions. Note that the initial control states of all controllable elements in the considered time horizon have to be given as well. The solution is computed on a finite time horizon $[t_{beg}, t_{end}]$. Appropriate discretizations for the given model equations are available. For the discretization in space and time, which is for instance needed for the WHE (6.8) - (6.9), ANACONDA gives two possibilities:

- the method of Lines (MOL) ansatz with a combination of FV or WENO methods in space and explicit RK, DIRK or SDIRK methods in time
- the fully discrete implicit box scheme IBOX (see [43]) or other fully discrete schemes, like a Lax-Friedrichs scheme.

For the different discretization approaches, we need to consider different restrictions for temporal and spatial steps. First, we need to consider the CFL condition for the MOL ansatz with explicit time schemes. For the (S)DIRK methods, we need suitable SSP conditions, see Section 3.1.3 and 3.1.4, and for the implicit box scheme, we need to make sure that a lower bound for the ratio of space and time step is fulfilled, see Equation (3.45) or [43]. As a default for the discretization of the ODEs, the implicit Euler method with equidistant time steps $t_{beg} = t_0 < t_1 < \ldots < t_N = t_{end}$ is used. The final result is a fully discretized coupled system of (nonlinear) algebraic equations

$$E(y, u) \stackrel{!}{=} 0,$$

which depends on the state variables $y^T = (y(t_0)^T, y(t_1)^T, \ldots, y(t_N)^T)$, like pressure heads and flow rates, and on the control variables $u^T = (u(t_0)^T, u(t_1)^T, \ldots, u(t_N)^T)$, like the speed of pumps or the degree of opening of control valves. The resulting system is solved using Newton's method for given initial conditions $y(t_0) = y_0$ and control variables for all time steps. Boundary and coupling conditions are included in $E(y, u)$. Since we use one-step methods, the operator $E(y, u)$ is of the special form

$$E(y,u) = \begin{pmatrix} y(t_0) - y_0 \\ F(t_0, t_1, y(t_0), y(t_1), u(t_0), u(t_1)) \\ \ldots \\ F(t_{N-1}, t_N, y(t_{N-1}), y(t_N), u(t_{N-1}), u(t_N)) \end{pmatrix} = 0. \tag{6.32}$$

This enables us to solve these equations block-wise in an iterative manner, which corresponds to a time-stepping. In every time step, an equation of the form

$$F(t_{j-1}, t_j, y(t_{j-1}), y(t_j), u(t_{j-1}), u(t_j)) = 0$$

has to be solved. While for explicit schemes the solution just involves an evaluation depending on the previous time step, we need to use Newton's method for implicit equations. Note that Newton's method is known to be locally superlinear convergent for suitable initial values. However, global convergence can not be guaranteed in general.

To make this method more applicable to the framework of a water supply network, we utilize two modifications: the damped Newton method and the simplified Newton method. With the damped Newton method, we want to realize that the method converges for considerably more initial values than the non-modified method. A step size control allows us to improve the domain of convergence. With the simplified Newton method, the Jacobian matrix is not computed after every time step and causes considerably less computational effort. More about the used Newton methods can be found in Kolb[42].

One problem associated with the simulation of water supply networks is that the underlying linear systems of equations might not have a unique solution or are ill-conditioned. This can be the case if valves are closed in such a way that one part of the network is cut off. Then, the matrix of the underlying system becomes singular. To overcome this problem, we use a physically reasonable regularization of the underlying matrices, see Kolb[42] for more details.

6.4.2 Optimization Tool

Due to the 2-stage approach (DOPT - COPT) within the EWave system, our main task is to solve continuous nonlinear optimization problems. For this purpose, we use derivative-based optimization techniques. For given or fixed discrete control decisions, we can formulate the optimal control problem of the form

$$\begin{aligned} &\min_u f(y(u), u) \\ &\text{s.t. } E(y(u), u) = 0 \\ &\qquad u_{\min} \leq u \leq u_{\max} \end{aligned} \tag{6.33}$$

with state vector y, control vector u, objective function f and the equality constraints $E(y, u)$ which corresponds to the solution of the network equations. Before solving an optimization problem, ANACONDA uses the integrated simulation tool to compute the state vector y by solving the system $E(y, u) = 0$ for y, see Section 6.4.1. In the optimization tool, we consider the functions f and E solely depending on the control u. Note that the procedure for a more general optimization problem is described in [54, p. 22 -25]. Within the optimization tool of ANACONDA, we use nonlinear optimization solvers like IPOPT [86] or DONLP2 [73, 74]. Mainly, the objective function of a optimization problem of the form (6.33) models energy costs. In our test examples the objective function consists of pump costs or costs for connections with characteristic curves like for example UV-filters. Additionally, there are fictive costs like penalty terms for, e.g., low filling

levels in storage tanks. For the objective function, which contains all this, we consider cost terms of the form

$$\int_{t_{beg}}^{t_{end}} f_k(t, y(t), u(t))\, dt \tag{6.34}$$

and usually approximate these integrals with

$$\sum_{j=0}^{N} w_j f_k(t_j, y(t_j), u(t_j)),$$

for weights w_j and the time steps t_j. Altogether, we get the entire costs

$$f(y, u) = \sum_{k=1}^{M} \sum_{j=0}^{N} w_j f_k(t_j, y(t_j), u(t_j)), \tag{6.35}$$

where M is the number of objective functions and N is the number of time steps.

Description of different constraints

We use different types of constraints in ANACONDA. For the description of the optimization task later on, we need to consider technical bounds for network elements and thus, pressure or flow rate are limited by a minimum and maximum value at the element. Mathematically, we can write the corresponding constraint in the form

$$z_c(t, y(t), u(t)) \geq 0 \tag{6.36}$$

for all $t \in [t_{beg}, t_{end}]$. The discretized version of this inequality reads

$$Z_{c,j} := z_c(t_j, y(t_j), u(t_j)) \geq 0 \tag{6.37}$$

for all $j \in \{0, ..., N\}$. We evaluate the constraint in this form at every time step. Another possibility is to evaluate the constraint at certain so-called *check points*, e.g., every fourth time step. We call such a constraint **multi-valued constraint** because of the multiple evaluations over time. Within the optimization task of the pilot test network, which will be described in Section 6.5, we need special multi-valued constraints. We call them *linear sum constraints*. A characteristic of these constraints is that more state variables of different components are involved. One example is the sum of different flow rates of valves to make a synchronization control or to guarantee a special mixing ratio of water coming from different wells. All other constraints used in the pilot network can be seen as multi-valued constraints. For completeness, we give a short overview of other possible constraints implemented in ANACONDA.

Another type of constraint is so-called **single-valued constraint**, which has the form

$$Z_c(y, u) := \min_{j \in \{0,...,N\}} z_c(t_j, y(t_j), u(t_j)) \geq 0 \tag{6.38}$$

evaluated at all time steps or for special check points. In comparison to Equation (6.37), we can here reduce the effective or total number of constraints and also reduce the

computational effort. Note that because of this reduction the optimization tool gets less information out of the constraints. We can also formulate

$$Z_c(y,u) := \min_{j\in\{0,\dots,N\}} z_c(t_j, y(t_j), u(t_j)) \leq 0. \tag{6.39}$$

This type of constraint is not fulfilled in every time step, but at least once. With such a formula, we describe for example the *breathing* of a tank. For instance when having a lower bound m where the filling level of the tank becomes zero, this bound is reached at least once during a chosen time horizon. This means that $\min h \leq 0$. In contrast, the tank is fully filled at least once if the filling level of the tank reaches the upper bound M, i.e., $\min(-h) \leq -M$ with $M > 0$. Here, it is also possible to evaluate Equation (6.39) at some check points.

A last type of constraint is the so-called **terminal constraint** and can be written as

$$Z_c(y,u) = z_c(t_N, y(t_N), u(t_N)) \geq 0. \tag{6.40}$$

Terminal constraints are used to avoid finite horizon effects. Within EWave, we handle this kind of constraints indirectly. For example, the lower bounds of the filling level of the tanks are set to their initial values at the end of the time horizon.

Adjoint calculus

In ANACONDA, we apply a first-discretized adjoint approach to compute reduced gradients of the objective function or any of the (in)equality constraints in (6.33), called $f(y(u),u)$. These quantities are sufficient to give sensitivity information. We now want to compute the reduced gradient $\frac{d}{du}f(y(u),u)$. To achieve this, we need to derive the *adjoint equations* which can be computed using the Lagrange function $L(y,u)$. Finally, we solve the linear system of adjoint equations

$$(\partial_y E(y(u),u))^T \xi = -(\partial_y f(y(u),u))^T, \tag{6.41}$$

where the matrix on the left-hand side is independent of f. Therefore, the matrix $\partial_y E(y(u),u)$ and decompositions of it only need to be computed once. After computation of the solution for (6.41), we get the reduced gradients from

$$\frac{d}{du} f(y(u),u) = \partial_u f(y(u),u) + \xi^T \partial_u E(y(u),u), \tag{6.42}$$

where the matrix $\partial_u E(y(u),u)$ is also independent of f and we therefore again only need one evaluation. Note that the system has to be solved for all (in)equalities in (6.33) and therefore, the evaluation can be very costly. Here, we consider time-dependent problems.

This means that the system of discretized model equations $E(y,u)$ has a special structure (6.32). For the matrix in (6.41), we get

$$(\partial_y E(y,u))^\top = \begin{pmatrix} I & & & & \\ A_1 & B_1 & & & \\ & A_2 & B_2 & & \\ & & \ddots & \ddots & \\ & & & A_N & B_N \end{pmatrix}^\top$$

with

$$\begin{cases} A_j = \partial_{y_j} F(t_{j-1}, t_j, y(t_{j-1}), y(t_j), u(t_{j-1}), u(t_j)) \\ B_j = \partial_{y_{j-1}} F(t_{j-1}, t_j, y(t_{j-1}), y(t_j), u(t_{j-1}), u(t_j)) \end{cases} \quad \forall j = 1, ..., N,$$

and ∂_{y_j} denoting the derivative with respect to $y(t_j)$. For the objective function $f(y,u)$, the set of adjoint equations (6.41) becomes

$$\begin{pmatrix} I & A_1^T & & & \\ & B_1^T & A_2^T & & \\ & & B_2^T & \ddots & \\ & & & \ddots & A_N^T \\ & & & & B_N^T \end{pmatrix} \begin{pmatrix} \xi(t_0) \\ \xi(t_1) \\ \vdots \\ \vdots \\ \xi(t_N) \end{pmatrix} = - \begin{pmatrix} \partial_{y_0} f(y,u)^T \\ \partial_{y_1} f(y,u)^T \\ \vdots \\ \vdots \\ \partial_{y_N} f(y,u)^T \end{pmatrix}.$$

The partial derivatives $\partial_{y_j} f(y,u)^T$ for all $j = 0, ..., N$ belong to the block-wise partitioning of the state y according to the time steps. The entire system above can be solved block-wise backwards in time, and therefore the size of the systems to be solved reduces.

6.5 The pilot test-network of RWW

One of the main goals of the project EWave was the simulation and optimization of a complex network. Here, we will roughly describe such a network provided by RWW. For the introduced network, we show the results of the EWave system in form of the state variables at important components.

6.5.1 Network description

Let us describe the pilot test-network of RWW in more detail. The main parts of the network are the water work Dorsten-Holsterhausen and the water distribution network.

A major part of the water work Holsterhausen has been shown in Figure 6.1. We show a more schematic description in Figure 6.6. Inside the water work several non-trivial

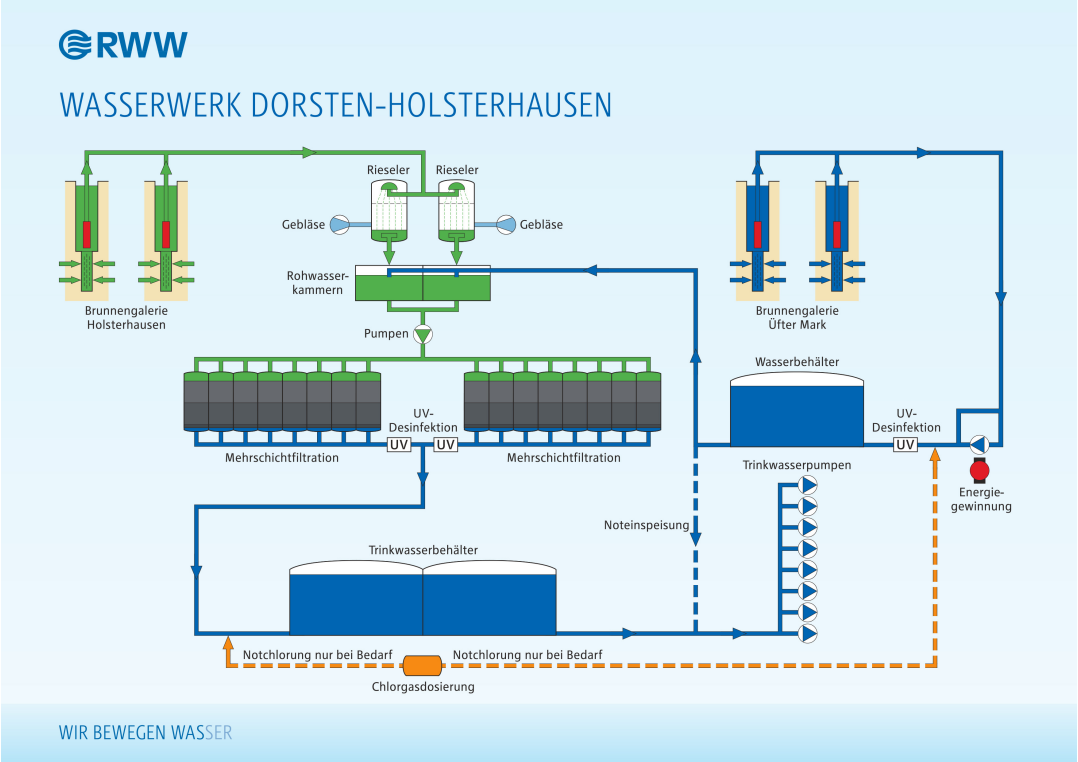

Figure 6.6: Flow diagram of the water work Dorsten-Holsterhausen. This picture was provided by Michael Plath and Stefan Fischer from RWW.

processes take place. The first step is to extract raw water from the available wells. For this water extraction, two well galleries *Galerie HOL* and *Galerie UEF* which are named *Brunnengalerie Holsterhausen* in Figure 6.6 are available in our test-network. They consist of either 42 or 11 vertical water wells. Note that due to the geodetical elevation difference, the water stemming from the Galerie UEF has high potential energy, which is transformed into electrical energy via an energy recovery turbine at the entrance of the water work. Further, the water stemming from the different impoundments comes with different iron concentration. Therefore, the raw water from Galerie HOL is aerated with two oxidators and both water contributions are mixed at a special ratio in the

raw water chambers[6]. The proportion of the raw water coming from Galerie UEF, lies between 20% and 40%. This mixing process also ensures that homogeneous raw water with the right amount of nitrat is created. In a next step, the water is led through two filter streets consisting of eight multimedia filter[7], where six raw water pumps are used. Note that the filters have to be cleaned on a regular basis. This is done either if the minimum preparation amount or the maximal pressure difference is reached. In these cases, the filters are irrigated with a defined irrigation program. After the transportation through the filters, the water is transmitted through the *UV reactors* and is from now on called drinking water. The water can be stored and deacidified in the drinking water tanks and further distributed from there via eight drinking water pumps, c.f. Figure 6.7. Note that the drinking water pumps consist of two variable-speed controlled pumps

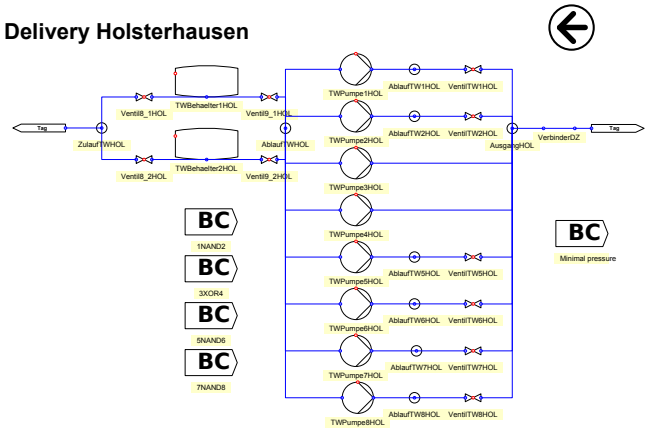

Figure 6.7: Water delivery out of the water work Holsterhausen with two tanks and 8 pumps. After every fixed speed pump a gate valve is located. Here, the water is transported with a minimal pressure out of the waterwork. The picture was provided by Annelie Sohr from Siemens AG.

TWPumpe3/4HOL and six fixed speed pumps.

The water now enters the second part of the pilot test-network, which is the water distribution network as shown in Figure 6.2. Besides a sophisticated pipe network, some additional parts, which are located at certain cities and communities, are necessary to guarantee a consistent supply. These are, for example, the pressure boosting station Buersche Straße and the tank complex Gladbeck, as depicted in Figure 6.8. Besides that, a further tank complex called Tackenberg exists, but is not shown on a picture. The water support from Gladbeck is ensured by four fixed pumps. Concerning the water support at Buersche Straße, the pressure through the pressure boosting Buersche Straße is continuously increased via two pumps. The tank complex Tackenberg also supports the water distribution via four fixed speed pumps. From 10 pm to 6 am, the tank complex is filled with 80 % from Holsterhausen and 20 % from water work Styrum-Ost.

Note that we apply here the network aggregation as described in Section 6.2, to get a reduced network. Altogether, the network under consideration consists of 7 tanks, 2

[6]in German: Rohwasserkammern
[7]in German: Mehrschichtfilter

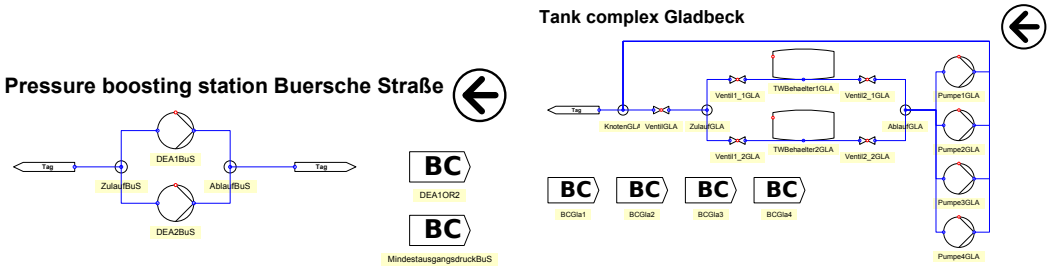

Figure 6.8: Left: Pressure boosting Buersche Straße with two pumps. One Constraint here is the minimal outlet pressure after the pumps. Right: Tank complex Gladbeck with two tanks and four pumps. Four boundary conditions are determined here. Both pictures were provided by Annelie Sohr from Siemens AG.

controllable suppliers, 52 suppliers, 102 pipes, 15 short pipes, 27 gate valves, 1 check valve, 5 control valves and 22 pumps.

6.5.2 Numerical results

We show the numerical results for the EWave system applied to the previously described pilot test-network with respect to two test examples. In particular, we describe and analyse the results of COPT computing the continuous control quantities with the software package ANACONDA. Note that the network contains 11 continuously controllable network quantities, which enter the control vector u of the optimization problem (6.33). Since ANACONDA needs discrete control quantities as input, we use the control quantities computed by DOPT, see Section 6.2, or the historical control quantities used by RWW. As a benchmark we use the results of DOPT. In both examples the input parameters of DOPT, e.g. initial filling level of tanks and initial switching points of pumps, are computed by SIMT. Further, we use the results of SIMT, see Section 6.2, and of DOPT as initial values and initial control of COPT, respectively. To solve the discretized problems in both examples, ANACONDA uses the solver IPOPT [39, 86], as mentioned before.

Let us discuss the different constraints. All control quantities are limited by box constraints, e.g., the controllable flow rate of the two wells *GalerieHOL* and *GalerieUEF*. After time discretization, where we use four time points per hour, we will end up with $11 \cdot 4 \cdot T$ constraints, where T is the time horizon in hours. We will show examples for 12 and 24 hours. For all controllable components, where the pressure head and the flow rate are not the controllable variables, we have to apply state constraints in many cases. This is the case for, e.g., the flow rate of all discretely and continuously controllable pumps. To guarantee that the filling level of the tanks lies in a certain range, we apply constraints for the inner pressure head. The pressure zone Holsterhausen contains 52 fictive suppliers where we also apply box constraints for their pressure heads.

Other constraints are employed in order to control whole regions in the network. For example, the minimal pressure in the pressure boost Buersche Straße and the pressure zone Holsterhausen are ensured using the inequalities

$$h(\text{Referenz\_Gladb\_Sp103}) \geq 99, \tag{6.43}$$

where Referenz_Gladb_Sp103 is a fictive tank and

$$h(\text{AblaufBuS}) \geq 132.4, \tag{6.44}$$

where AblaufBuS is a coupling node. As discussed above, a special mixing ratio of water coming from GalerieUEF and GalerieHOL has to be guaranteed to ensure the water quality. Therefore, we use the linear sum constraint

$$\tfrac{2}{3}q(\text{Ve2\_}1HOL) - q(\text{Ve2\_}1UEF) + \tfrac{2}{3}q(\text{Ve2\_}2HOL) - q(\text{Ve2\_}2UEF) \geq 0, \tag{6.45}$$

which is abbreviated with *MischverhaeltnisHOLPLUSUEF*.
Note that for example $q(\text{Ve2\_}1HOL)$ denotes the flow rate at the valve Ve2_1HOL. Further, we have to ensure that the same amount of water flows through the identically constructed filter streets, thus we set

$$q(\text{Ve5\_}1HOL) = q(\text{Ve5\_}2HOL). \tag{6.46}$$

This constraint is called *GleichlaufregelungFilterHOL*. A last constraint ensures that the amount of water, which comes from the water work Holsterhausen, does not exceed a given bound, and has the form

$$-q(\text{Ve5\_}1HOL) - q(\text{Ve5\_}2HOL) \geq -1.222220 \tag{6.47}$$

which is abbreviated with *GesamtMengeHOL*. For the following examples, we use a time step size of $\Delta t = 900$s. The spatial mesh size is chosen such that the mesh is as coarse as possible, but the numerical accuracy is still high enough. Although preliminary implementations of SDIRK methods in combination with finite volume methods have been included, ANACONDA uses the implicit box scheme, see Section 3.3.3, for the computation of the WHE (6.8) and (6.9). The objective function for these scenarios will be to minimize the energy consumption of the whole network.

12h example

To illustrate the accuracy and efficiency of COPT, i.e. the optimization done by ANACONDA, we show results for a 12h real-life example. To show the numerical results without optimization and compare it to the optimized numerical results later on, we compute the simulation done by ANACONDA with the discrete optimized control of DOPT as input parameters and call it SIM_ANA. Mainly, we compare the three quantities COPT, DOPT and SIM_ANA here.

The energy consumers in this example are the pumps RWPumpe1/2/3/5HOL, TWPumpe2/3/4/6/8HOL, DEA1/2BuS and some connections with pressure characteristic curves modelling for instance UV filters UV1/2HOL or the tricklers Riesler1/2HOL. We show the results for several characteristic components in the network together with the corresponding upper and lower bounds. Note that for all control variables the box constraints are fulfilled after the optimization.

The corresponding inner pressure heads of the 7 tanks are shown in Figure 6.9 and Figure 6.10. We observe that our simulation data fit to the DOPT data. This was

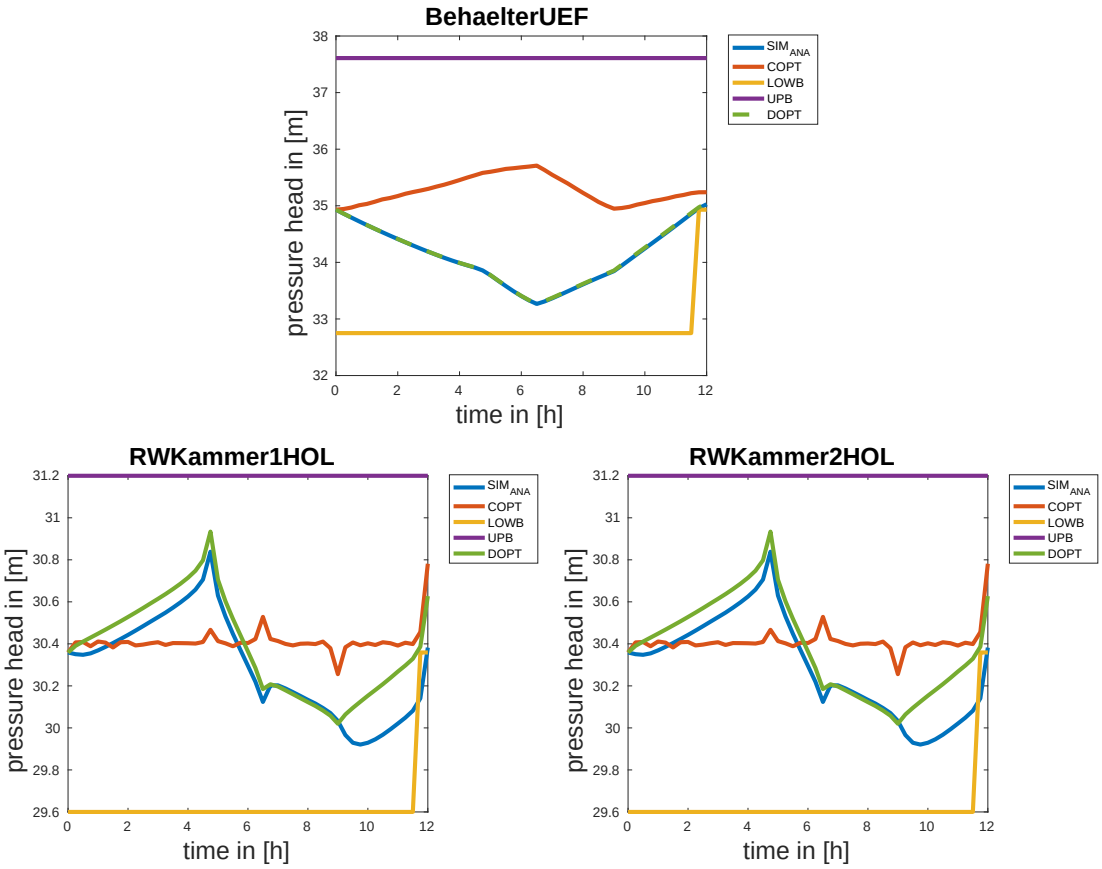

Figure 6.9: Pressure heads of the tanks with the corresponding lower and upper bounds. Bounds for BehaelterUEF: LOWB $= 32.75\,m$ and UPB $= 37.61\,m$. Bounds for RWKammer1/2HOL: LOWB $= 29.6\,m$ and UPB $= 31.2\,m$. Note that in the last time step, LOWB is set to the corresponding initial value.

expected because our simulation data are based on the control decisions made by DOPT.

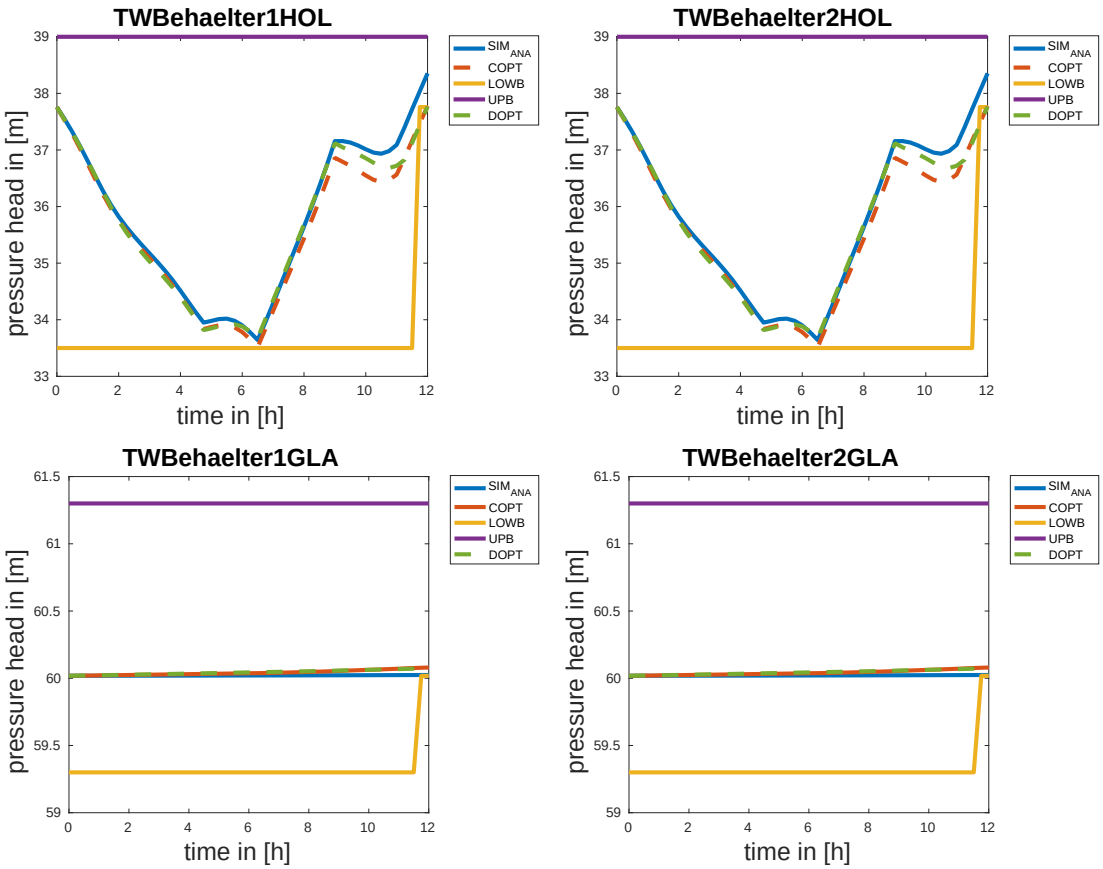

Figure 6.10: Pressure heads of the tanks with the corresponding lower and upper bounds. Bounds for TWBehaelter1/2HOL: LOWB $= 33.5\,m$ and UPB $= 39\,m$. Bounds for TWBehaelter1/2GLA: LOWB $= 59.3\,m$ and UPB $= 61.3\,m$. Note that in the last time step, LOWB is set to the corresponding initial value.

In the figures showing the RWKammer1/2HOL, we see that the simulation data violates the lower bound even though we use the optimal results of DOPT as initial values for COPT which do not violate the bound. If we look at the results of DOPT, we see that DOPT does not violate the lower bound. This effect appears because DOPT and COPT uses different model equations, as mentioned before in this Chapter, and therefore, different numerical results can appear.

The extraction of water out of the tanks TWBehaelter1/2GLA seems to be consistent because of the non-changing pressure heads, see Figure 6.10, and there are no significant differences between DOPT, COPT and SIM_ANA. As already mentioned above, the lower bound LOWB restricting the pressure head of the tanks is set to the corresponding initial value of the pressure head at final time step $(42300, 43200]$. With this requirement, we can avoid that the tanks tick over in every scenario.

Further, we show the flow rates when the pumps are switched on in Figure 6.11. We first look at the pumps called *RWPumpe1/2/3/4/5/6HOL*, which are located behind RWKammer1/2HOL. Only the first three pumps are switched on in this scenario. The pump RWPumpe5HOL is switched off immediately. Note that these pumps are all fixed

speed pumps. Here, we also have upper and lower bounds which need to be fulfilled. The

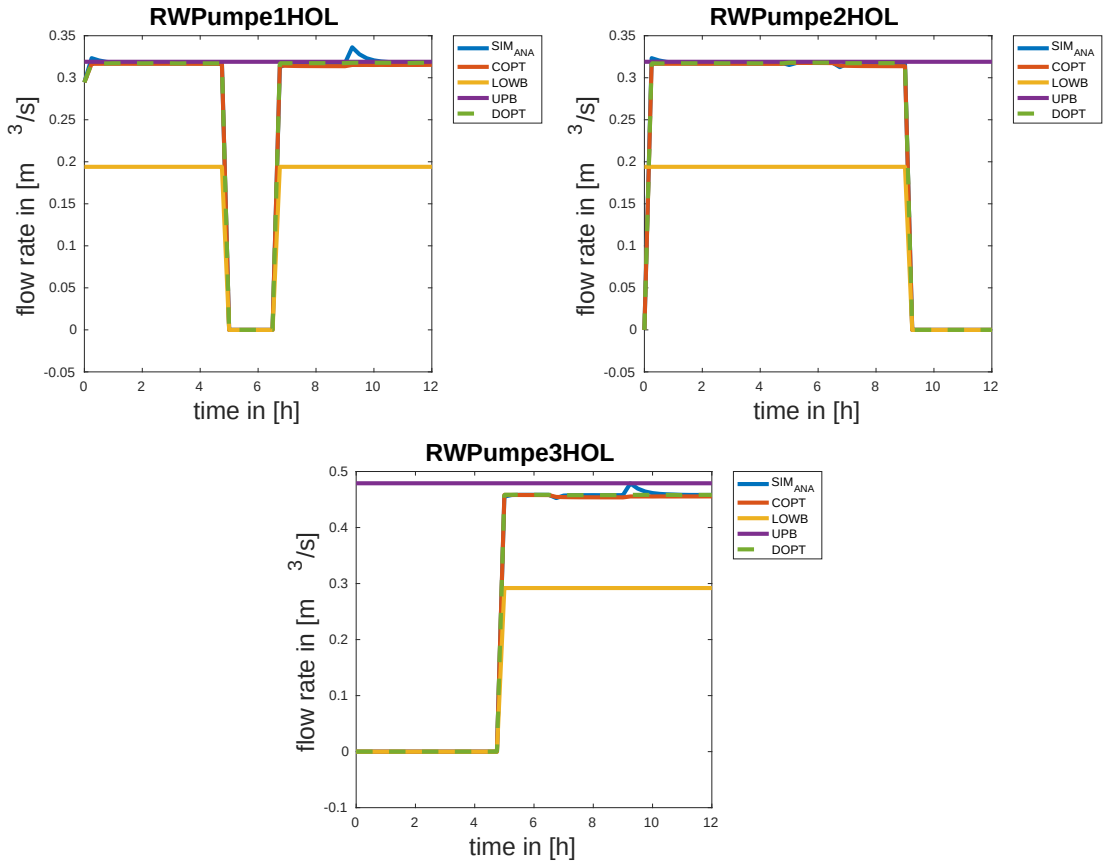

Figure 6.11: Flow rates of the RW pumps with the corresponding lower and upper bounds. Bounds for RWPumpe1HOL and RWPumpe2HOL: LOWB $= 0.194\,m^3/s$ and UPB $= 0.319\,m^3/s$. Bounds for RWPumpe3HOL: LOWB $= 0.292\,m^3/s$ and UPB $= 0.479\,m^3/s$.

lower bounds are set to $-0.0001\,m^3/s$ when the pumps are switched off. This is true for all pumps. Before the optimization step, the flow rate of RWPumpe1HOL violates the upper bound UPB twice. After the optimization, the flow rate lies in the range $[0.194, 0.319]$ or $[-0.0001, 0.319]$, respectively. Next, we look at the drinking water pumps called *TWPumpe1/2/3/4/5/6/7/8HOL* which are located behind the two drinking water tanks called *TWBehaelter1/2HOL*. The corresponding flow rates for the pumps that are switched on are shown in Figure 6.12. Here, the flow rates of TWPumpe3/6HOL are not shown since they are only switched on in the initial state. The pump TWPumpe2HOL is only switched on in the last three time steps. Here, we can observe that the flow rate computed by SIM_ANA clearly lies under the lower bound LOWB whereas the flow rates computed by COPT and DOPT do not lie under the given bound. The different flow rates of TWPumpe8HOL all lie in the bounds. Considering the different flow rates of TWPumpe4HOL, we can observe that SIM_ANA violates the upper bound UPB twice. Note that only the pump TWPumpe4HOL is a continuously controllable pump. It is controlled with respect to the pressure head of the fictive tank *Referenz_Gladb_Sp103*.

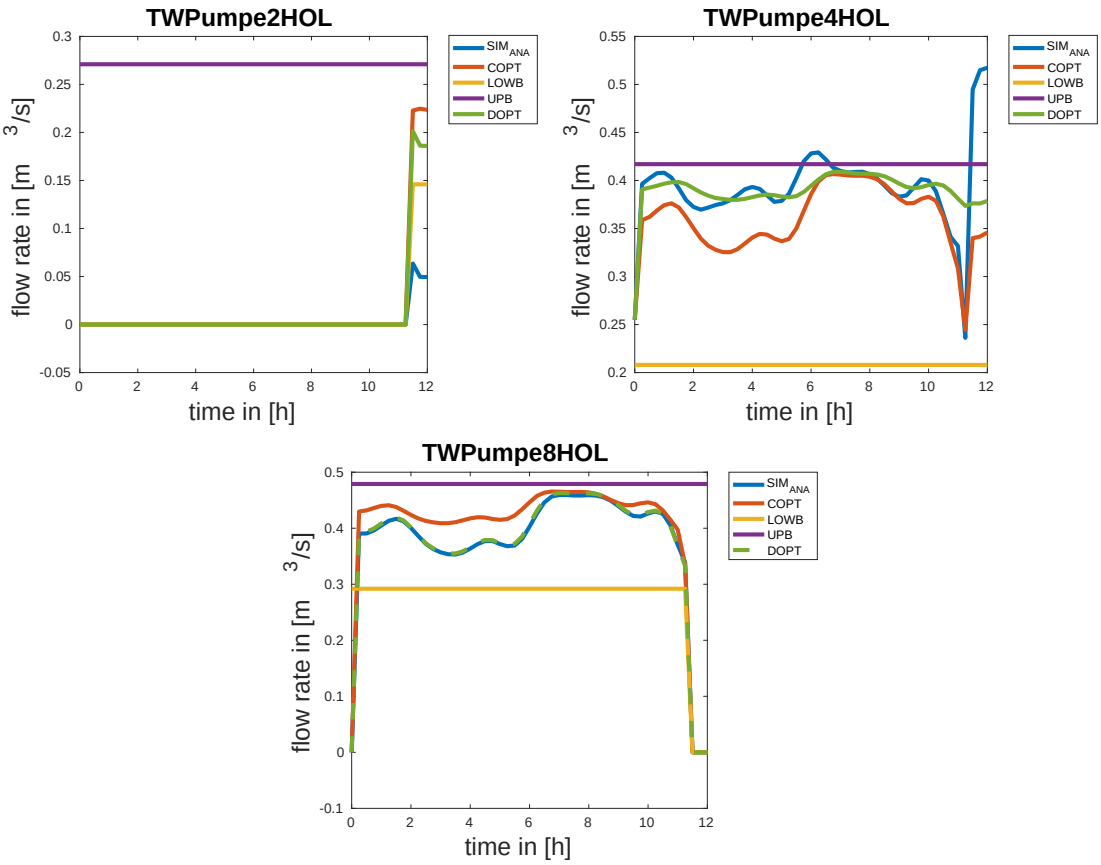

Figure 6.12: Flow rates of the TW pumps with the corresponding lower and upper bounds. Bounds for TWPumpe2HOL: LOWB $= 0.146\,m^3/s$ and UPB $= 0.271\,m^3/s$. Bounds for TWPumpe4HOL: LOWB $= 0.208\,m^3/s$ and UPB $= 0.417\,m^3/s$. Bounds for TWPumpe8HOL: LOWB $= 0.292\,m^3/s$ and UPB $= 0.479\,m^3/s$.

In Figure 6.13, we see the controllable wells *GalerieUEF* and *GalerieHOL*, whose flow rates lie in the required bounds before and after the continuous optimization. Besides that, we can observe that the water extraction out of GalerieUEF computed by COPT is more consistent than that of DOPT and SIM_ANA. Altogether, we can observe that more water is extracted from GalerieHOL in COPT in comparison to SIM_ANA and DOPT.

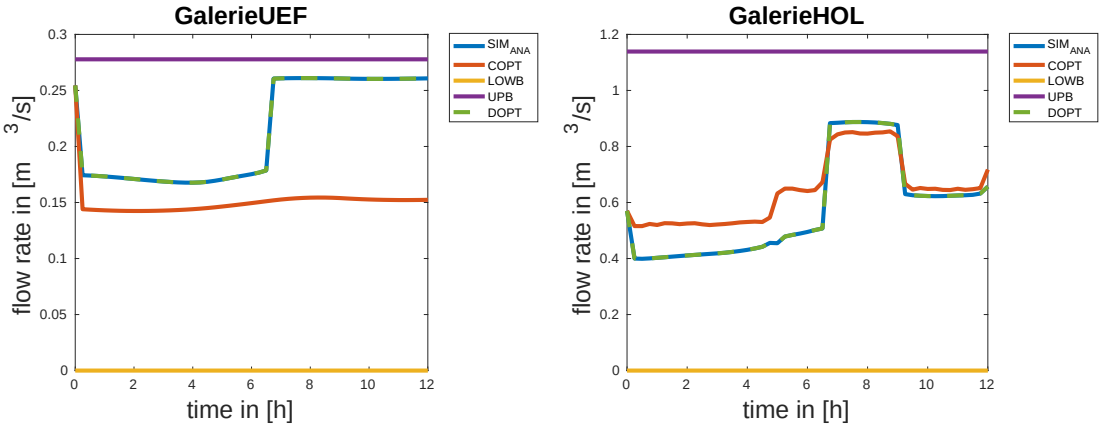

Figure 6.13: Flow rates of the wells with the corresponding lower and upper bounds. Bounds for GalerieUEF: LOWB $= 0\,m^3/s$ and UPB $= 0.277778\,m^3/s$. Bounds for GalerieHOL: LOWB $= 0\,m^3/s$ and UPB $= 1.138889\,m^3/s$

Further, we need to fulfill constraints which guarantee that the pressure in the network does not fall below a given minimum pressure head. This restriction is for example given at the Buersche Strasse, see Figure 6.14. Note that neither the simulation data nor the

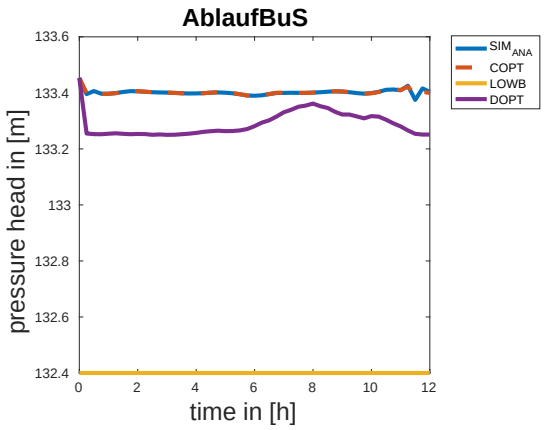

Figure 6.14: Pressure head at the Buersche Strasse. The lower bound LOWB $= 132.4\,m$ do not have to be violated.

optimization data violate that bound. Figure 6.15 shows the results with respect to constraints (6.47) and (6.45). Note that we only plot the results of COPT and SIM_ANA here. Whereas the graphs of COPT and SIM_ANA are almost identical with respect to the constraint (6.47), the graphs describing the *MischverhaeltnisHOLPLUSUEF* are very different. Considering the results corresponding to the constraint *GleichlaufregelungFilterHOL* which we show not in a plot here, we observe that the bound is slightly violated at the first time step in the simulation and optimization data. The violation lies in the range of $10^{-6}\,m^3/s$ which is an acceptable value and fulfill the tolerance of the Newton solver. Concerning the three constraints, the graphs developed by the simulation and optimization data lie within the required bounds.

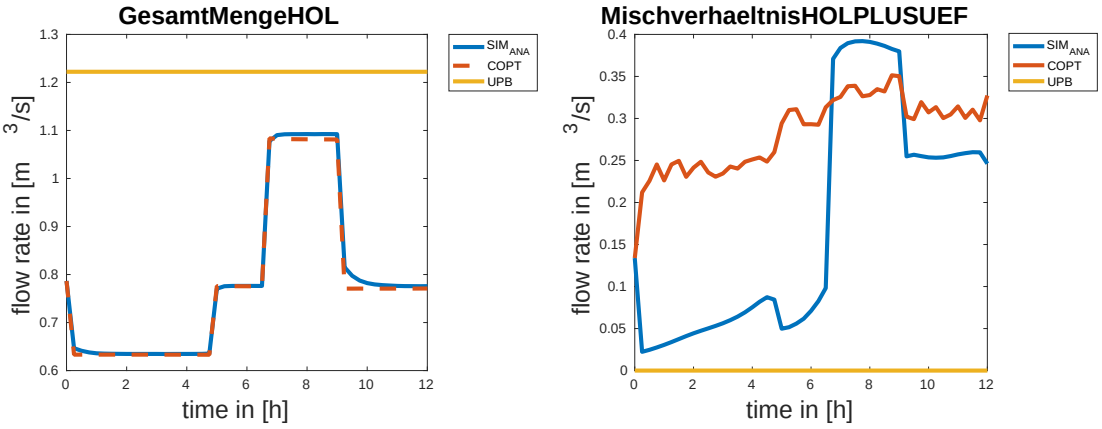

Figure 6.15: Two special constraints which guarantee the water quantity and quality of the water work. Left: Results considering Constraint (6.47). Right: Results considering Constraint (6.45).

The simulation result with the DOPT control reaches an overall energy consumption of 73925 Megajoule. Note that the rate of energy consumption stems from an infeasible solution. Altogether, the optimized energy consumption amounts to 74134 Megajoule. We can observe that our optimized value for the cost functional is slightly higher than that of DOPT, but we have found a solution which is local optimal and feasible and thus, accepted as optimal. Our solver needs 254.729 seconds for the optimization task. For the nonlinear optimization, we use the solver IPOPT [86], which computed a feasible solution with a maximal violation of $2.11869 \cdot 10^{-12}$ of all constraint values.

24h example

To show the efficiency and the potential of energy optimization of the complete EWave system, we consider a 24h example with $T = 86400$ seconds. To give a better comparison between the optimized values of the EWave system (COPT), we show the results which are obtained by the simulation of the real operation plans provided by RWW which are developed using historical data. We denote the simulated real data of RWW, computed by ANACONDA, by *SIM_REAL*. For completeness, we also show the results of DOPT.

The energy consumers in this example are the pumps RWPumpe1/2/3/5HOL, TW-Pumpe2/3/4/8HOL and DEA1/2BuS and some connections with characteristic curves, for instance UV1/2HOL and Riesler1/2HOL. Here, we again show results concerning important network components.

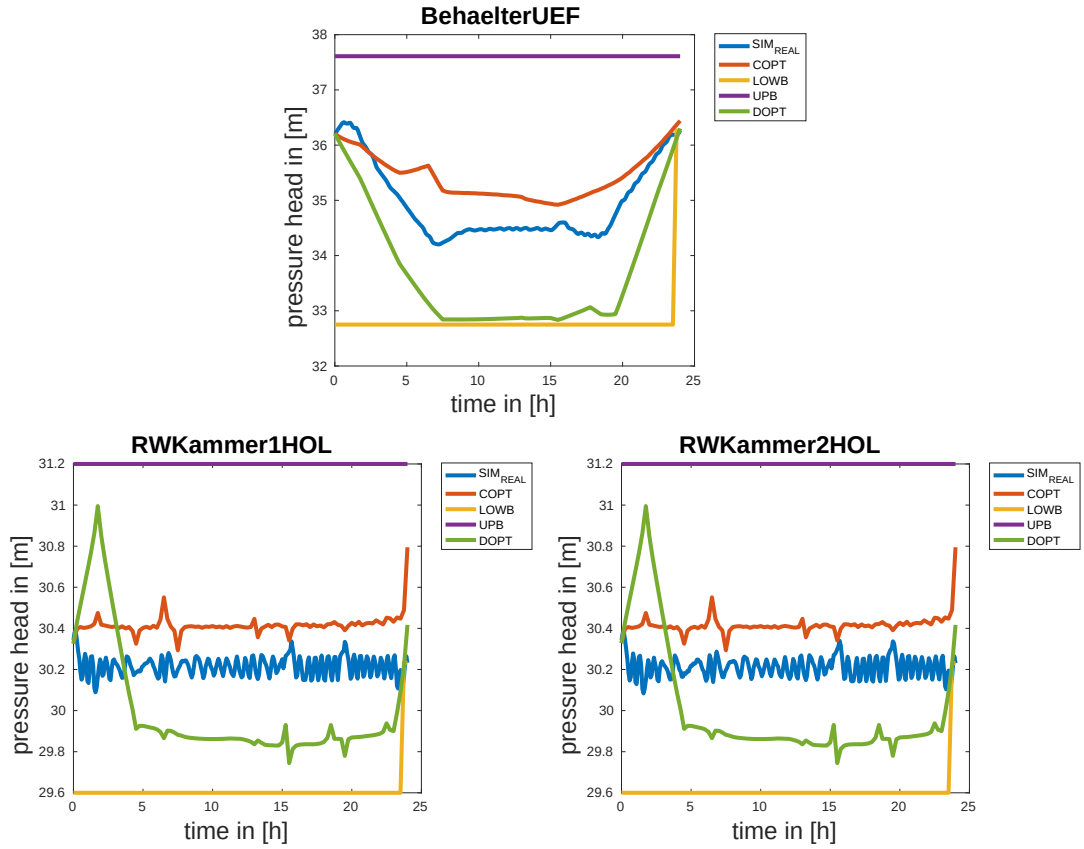

Figure 6.16: Pressure heads of the tanks with the corresponding lower and upper bounds. Bounds for BehaelterUEF: LOWB $= 32.75\,m$ and UPB $= 37.61\,m$. Bounds for RWKammer1/2HOL: LOWB $= 29.6\,m$ and UPB $= 31.2\,m$. Note that in the last time step, LOWB is set to the corresponding initial value.

The results of the evolution of the pressure heads at the tanks before and after the continuous optimization are depicted in Figure 6.16 and Figure 6.17. Again, for all constraints corresponding to the pressure heads of the tanks, the lower bound is set to the initial value at the final time step $(85500, 86400]$. In comparison, we can observe that the evolutions of the pressure heads of the real data differ from that of the optimized data. Considering the graphs of the pressure heads of the first three tanks, we can see that the

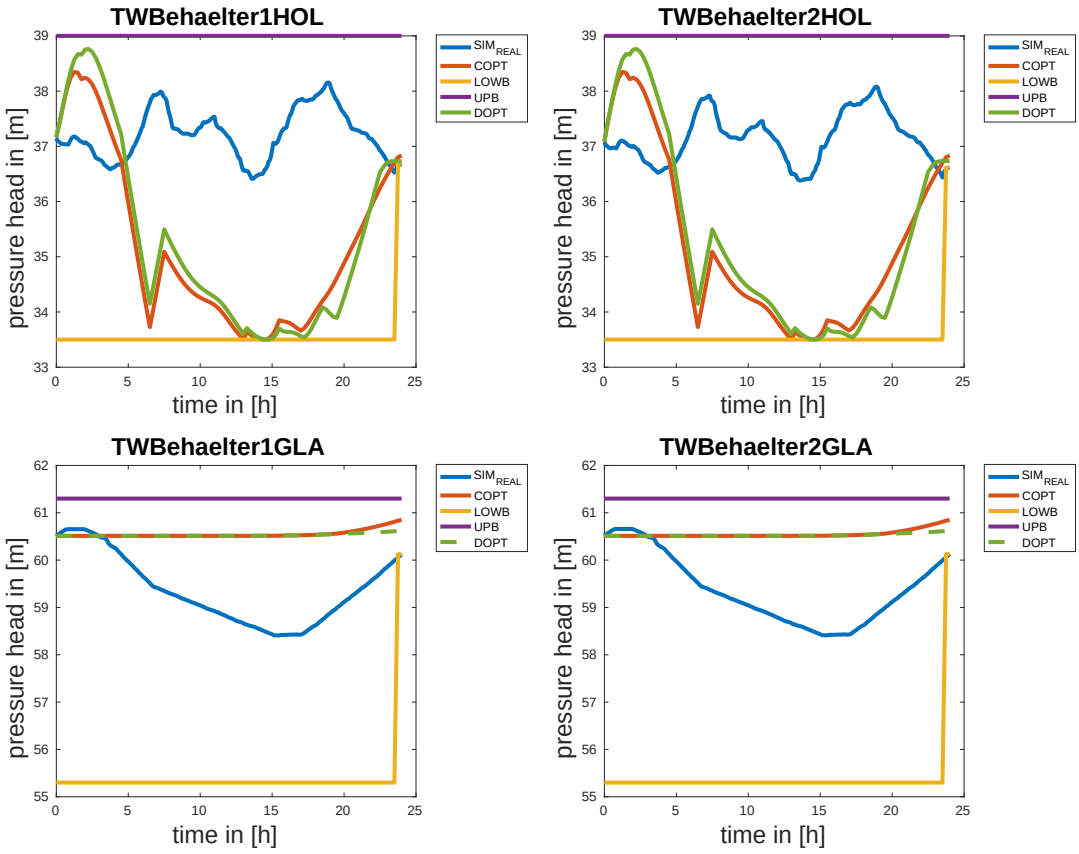

Figure 6.17: Pressure heads of the tanks with the corresponding lower and upper bounds. Bounds for TWBehaelter1/2HOL: LOWB $= 33.5\,m$ and UPB $= 39\,m$. Bounds for TWBehaelter1/2GLA: LOWB $= 55.3\,m$ and UPB $= 61.3\,m$. Note that in the last time step, LOWB is set to the corresponding initial value.

graph of COPT lies almost always above the two other graphs. This means that after the optimization there is more water in the tanks. Concerning the pressure heads computed by COPT and DOPT corresponding to TWBehaelter1/2HOL, we see that they reach the lower bound after 13 hours, whereas the pressure heads computed by SIM_REAL stay mostly near the upper bound UPB. As in the 12h example, the evolution of the pressure heads developed by DOPT and COPT of TWBehaelter1/2GLA are almost constant. The comparison to the results developed by SIM_REAL shows that more water is stored in the tanks.

The results for the running raw water pumps are shown in Figure 6.18. All other raw water pumps are not used and therefore, switched off in this scenario. Note that we only present the graphs of COPT and DOPT in the following figures to show their behaviour with respect to the required bounds. In the first picture of Figure 6.18, we see that the flow rate computed by COPT slightly violates the upper bound. This is acceptable because we set the value $0.05\,m$ for the pressure head and $10^{-3}\,m^3/s$ for the flow rate for the maximal violation of the bounds in agreement with our practical partners. Especially, these violation values were accepted by our industrial partner RWW to be exact enough. We therefore define a solution to be feasible and locally optimal if the bounds are fulfilled

even with these violations. Considering the results of DOPT and COPT in Figure 6.18, we can observe that the flow rates of RWPumpe2/3/5HOL lie in the corresponding bounds.

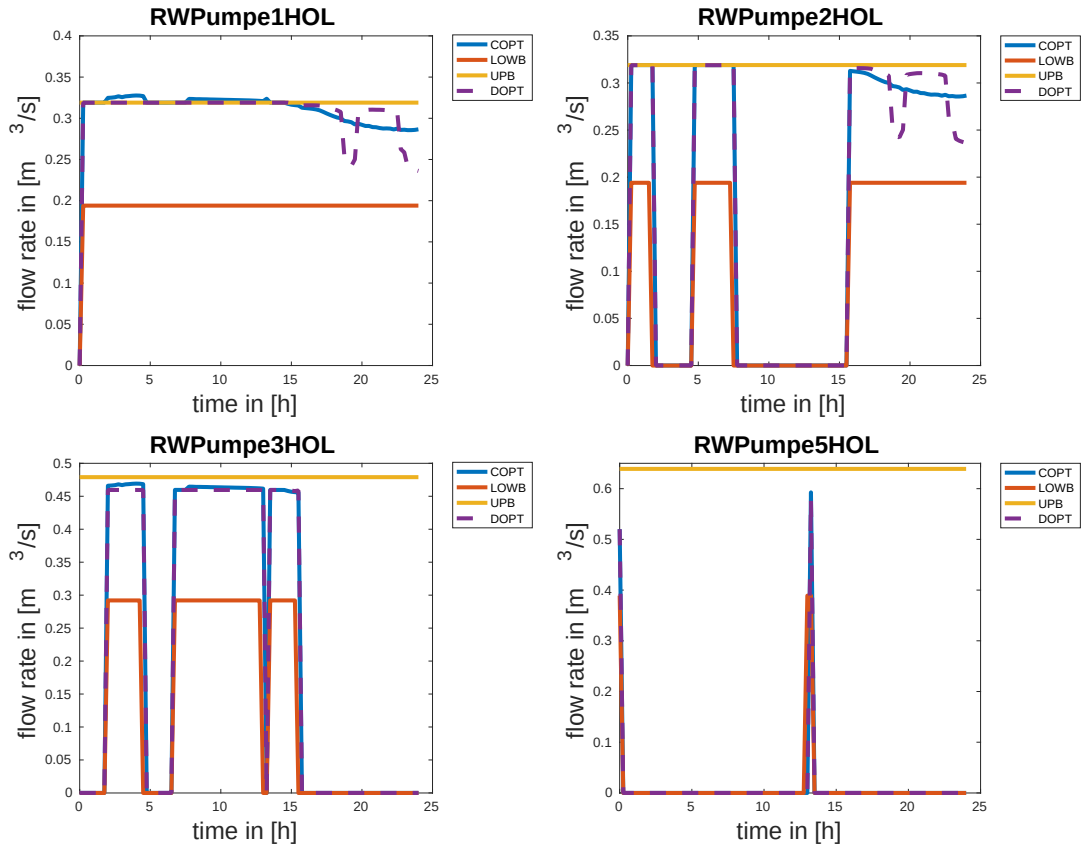

Figure 6.18: Flow rates of the RW pumps with the corresponding lower and upper bounds. Bounds for RWPumpe1HOL: LOWB $= 0.194\,m^3/s$ and UPB $= 0.319\,m^3/s$. Bounds for RWPumpe3HOL: LOWB $= 0.292\,m^3/s$ and UPB $= 0.479\,m^3/s$. Bounds for RWPumpe4HOL: LOWB $= 0.389\,m^3/s$ and UPB $= 0.639\,m^3/s$

The drinking water pumps that are used are depicted in Figure 6.19. Whereas the TWPumpe4HOL is used over the whole time horizon, the two other pumps are switched off for some time. Note that the time for which they are switched off is sometimes very short but thus, it is acceptable because of the minimal run- or downtime is 1800 seconds. Notice also that the flow rate of TWPumpe2HOL as well as the flow rate of TWPumpe4HOL, both computed by COPT, violate the upper bound in an accepted range.

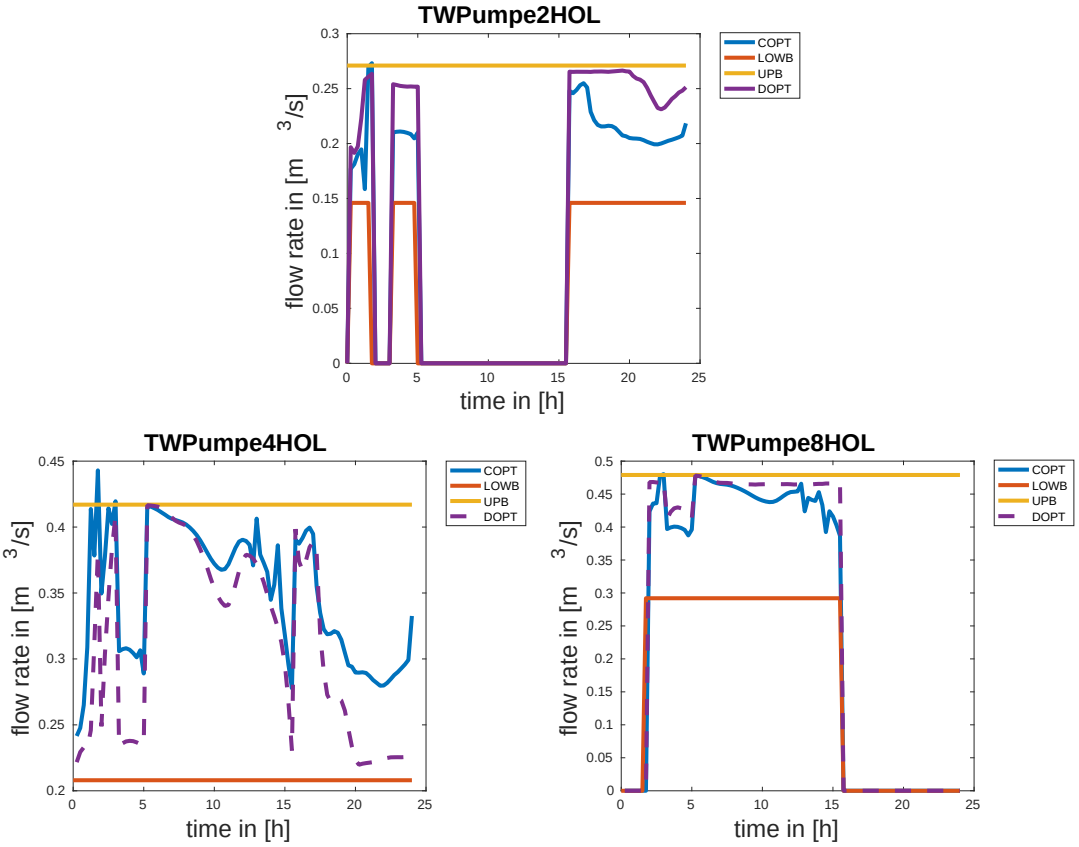

Figure 6.19: Flow rates of the TW pumps with the corresponding lower and upper bounds. Bounds for TWPumpe2HOL: LOWB = $0.146\,m^3/s$ and UPB = $0.271\,m^3/s$. Bounds for TWPumpe4HOL: LOWB = $0.208\,m^3/s$ and UPB = $0.417\,m^3/s$. Bounds for TWPumpe8HOL: LOWB = $0.292\,m^3/s$ and UPB = $0.479\,m^3/s$.

We also compare the switching plans of the raw water pumps, see Figure 6.20, of the drinking water pumps, see Figure 6.21 as well as of the pumps located at Buersche Strasse, see Figure 6.22. The shown graphs are computed by DOPT or REAL which again denotes the simulation of the real historical data of RWW called before SIM_REAL. We observe that overall, the entire switching plans of the single pump stations are different. While DOPT uses the pumps RWPumpe1/2/3/5HOL, REAL uses the same combination of pumps except of RWPumpe2HOL. In the results of DOPT RWPumpe3HOL is switched on three times while the results of REAL show that RWPumpe3HOL is switched on four times. Another observation is that DOPT uses RWPumpe5HOL only in two short intervals whereas REAL uses this pump more often.

Regarding the switching plans of the drinking water pumps, we can see that DOPT uses TWPumpe2/4/8HOL, whereas RWW utilizes TWPumpe1/4/5HOL. Note that in the results of DOPT, the variable speed pump TWPumpe4HOL is switched on over the whole time, whereas the fixed speed pumps TWPumpe2HOL and TWPumpe8HOL are switched off for some time intervals. Note also that the switching plan of DOPT use TWPumpe8HOL instead of TWPumpe5HOL. Altogether, we can observe that both

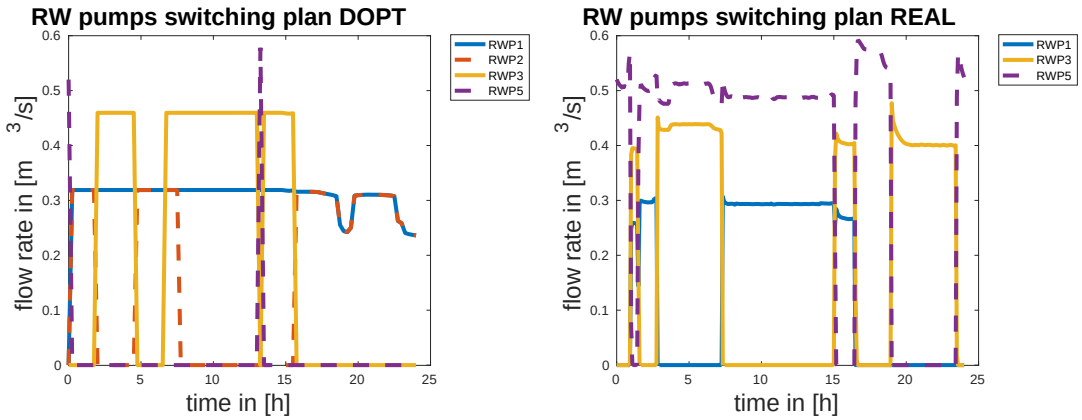

Figure 6.20: Switching plans of the raw water pumps of DOPT and REAL. Whereas DOPT uses the pumps RWPumpe1/2/3/5HOL, REAL uses the same combination of pumps except of RWPumpe2HOL.

switching plans contain the variable speed pump TWPumpe4HOL over the whole 24 hours.

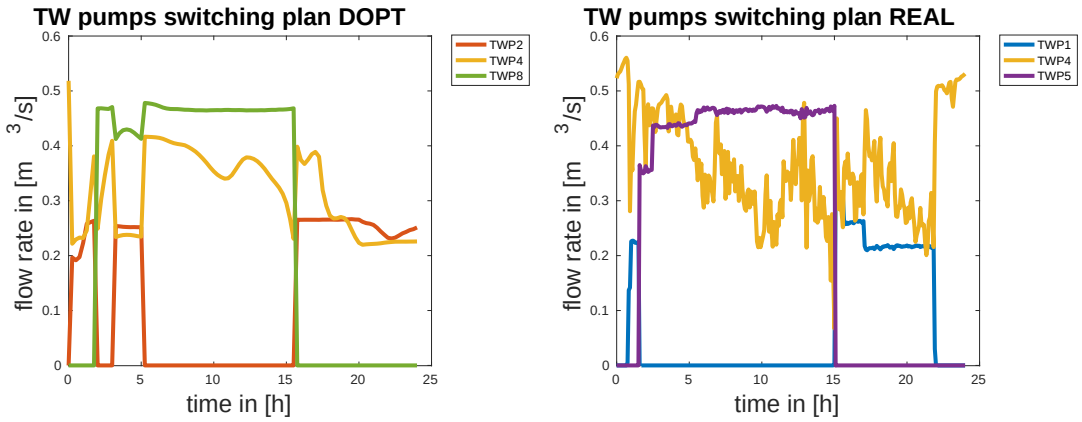

Figure 6.21: Switching plans of the drinking water pumps of DOPT and REAL. While EWave (especially DOPT) uses the pumps TWPumpe2/4/8HOL, REAL uses the pumps TWPumpe1/4/5HOL.

In Figure 6.22, the comparison of the pumps at Buersche Strasse shows that REAL uses both pumps. DOPT, on the other hand, only needs DEA1BuS to guarantee the corresponding pressure head and flow rate values. Therefore, EWave developed an optimized switching plan, which also is in some sense more steady and smooth than that of REAL.

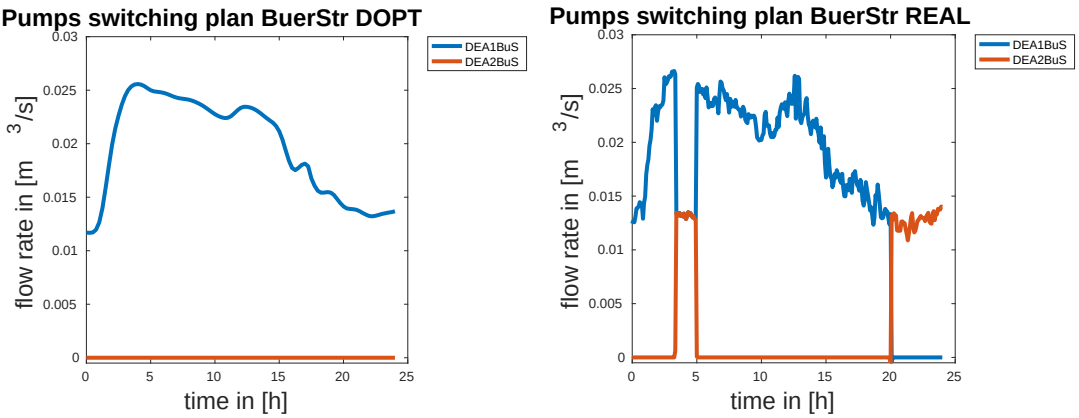

Figure 6.22: Switching plans of pumps of DOPT and REAL at Buersche Strasse. While DOPT only uses DEA1BuS, REAL uses both pumps.

Note that the simulated flow rates in the operation plans of RWW fulfill the pre-defined bounds most of the time, which is expected because we set the bounds according to the instructions of RWW. But in contrast to this, considering the minimum pressure constraint at the pressure boost zone Buersche Strasse, RWW violates this constraint, whereas EWave does not, see also Figure 6.23. Note that the oscillations which appear in SIM_REAL are the measured variations from the historical operation plans of RWW.

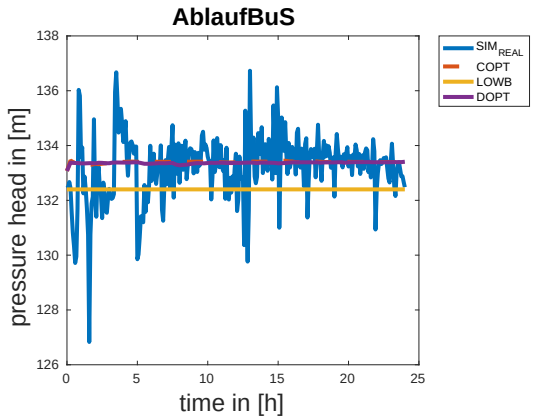

Figure 6.23: Evolution of the pressure head considering AblaufBuS at Buersche Strasse. Note that the minimum pressure constraint is violated by the data of RWW.

The initial value of the cost functional in IPOPT is 131643 Megajoule, which corresponds to the optimal value of the cost functional in DOPT. After the continuous optimization, we get a slightly higher value of 134311 Megajoule. The difference of the optimal value

computed by DOPT and COPT is due to the different modelling in both approaches. For the optimization task our solver needs 9973.75 seconds. For the nonlinear optimization, we use the solver IPOPT [39, 86], which computed a feasible solution with a maximal violation of 0.0259716 of all constraint values.

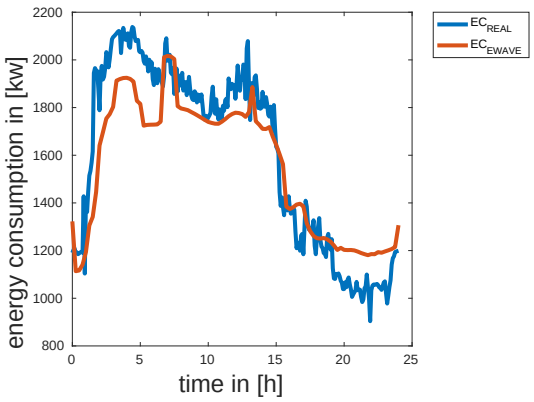

Figure 6.24: EC_REAL represents the energy consumption caused by operation plans of RWW. EC_EWAVE represents the energy consumption computed by EWave.

If we compare the energy consumption caused by operation plans of RWW and that caused by the operation plans of EWave, we get the results in Figure 6.24. In sum, we get an energy consumption of 37673 kWh for the RWW operation plans. For the operation plans computed by EWave, the energy consumption is 37308 kWh. We observe an improvement of approximatively 1%.

The discussed examples show that COPT can achieve optimal or locally optimal solutions in given time horizons. Unfortunately, in most cases of the 24h-examples, COPT can not compute optimal solutions because of the non-compliance of the given time horizon. It can be assumed that this results from the very difficult and complex interaction of the different mathematical models used in the different modules in the EWave system. Nevertheless, the complete EWave system shows its efficiency and potential of energy optimization in the pilot phase at RWW.

Summary

In this chapter, we have introduced the project EWave with all its complex tasks. We took a closer look at the optimization module of the EWave system, which consists of a discrete optimization and a continuous optimization tool. We also introduced the mathematical modelling of the network components for the water supply network. After that, we took a short look into the software, which build the continuous optimization tool COPT. In the last two sections, we showed the structure of the considered pilot test-network and presented numerical results with respect to that network. In the first example, we discussed the results before and after the continuous optimization of COPT in a 12h time horizon and focused on the accuracy and efficiency of the used software package ANACONDA. In the second example, we discussed the results of the EWave system based on a 24h time horizon and focused on the efficiency and potential of optimization of the complete assistance system. For this purpose, we compared the simulated operation plans used by RWW to the operation plans computed by EWave. We observed an improvement of the energy consumption using the results of the EWave system. In contrary to the real data, the pressure constraints were satisfied strictly, see Figure 6.23.

7 Conclusion

In this thesis, we developed suitable numerical methods for the simulation of water flow through pressurized pipes, which is an important aspect of the simulation-based and energy-optimized EWave assistance system.

In the first part, we introduced the water hammer equations that describe the flow of water through pressurized pipes. We discussed suitable numerical methods for the solution of such hyperbolic systems. In particular, we used a combination of so-called SSP SDIRK methods for the time integration and finite volume or discontinuous Galerkin methods for the spatial discretization. For these methods, we established well-balancedness with respect to the water hammer equations and a discrete maximum principle. The main step to achieve this was to prove the existence and uniqueness of a solution for the discrete schemes. This in turn enabled us to use the stationary state as initial condition and to show that the methods maintain the stationary state over time, which corresponds to the well-balancedness with respect to the water hammer equations. To show a discrete maximum principle, we ensured that the solution of the discrete schemes lies in a certain range whose limits are determined by the minimum and maximum value of the initial conditions. For the case of a one-dimensional system, an important aspect is that we use the characteristic decomposition to obtain the variables which satisfy such a principle. Care has to be taken concerning the choice of such a range if the diffusive source term is present.

In the second part, we described the EWave assistance system in more detail. The main results were the simulation and optimization with respect to the energy consumption of a real-life water supply network. After describing the structure and internal processes of EWave, we took a closer look at the optimization module, in particular the continuous optimization tool which is based on the software ANACONDA. We gave a fundamental description of the mathematical modelling of the network components used and showed numerical results for the pilot test-network. We illustrated the capability of EWave to significantly reduce the energy consumption of the network. In a 12h example, we presented the optimal operation plans computed by the system, which we compared to the operation plans stemming from the discrete optimization step. In the second example, we compared the simulated states and control variables which were computed on the basis of the operation plans of RWW to the optimized states and control variables of EWave. As a result, we observed an improvement of the energy consumption when our system is used.

We believe that the SSP SDIRK method in combination with finite volume or discontinuous Galerkin schemes is very promising for its applications in a complete water supply network. In this thesis, we established the well-balancedness and a discrete maximum-principle on a single pipe. Beyond this achievements, we believe that these properties can also be maintained in an entire network if suitable coupling conditions are used. Regarding the project EWave, we see possibilities for developing an assistance system that can be used for different types of water networks. To achieve this goal, the whole process of EWave, however, needs to be automated. This affects the network calibration, the network aggregation and the finding of suitable constraints in the optimization tool and remains a task for future projects.

A Appendix

A.1 Proof of Lemma 7

Lemma 7. *Assume for the linear polynomial* u_h *that* $\bar{\mathsf{u}}_j^n$ *lies in the range* $[m, M]$ *with* $m, M \in \mathbb{R}$ *for all* j*. Then* $\tilde{\mathsf{u}}_h$ *is a linear polynomial and* $\tilde{\mathsf{u}}_h(x) \in [m, M]$ *for all* $x \in \mathbf{I}_j$*,* $j = 1, \dots, N$*.*

Proof. We need to consider several cases. First, we note that in many possible cases including local minima and maxima, the MinMod limiter reduces the modified polynomial to $\tilde{\mathsf{u}}_h = \bar{\mathsf{u}}_j^n$ since the slopes do not have equal signs. We only treat the case where all slopes are positive, depicted in Figure A.1, while the other case can be treated analogously. Because of the linearity of the polynomials, the minimal and maximal values on

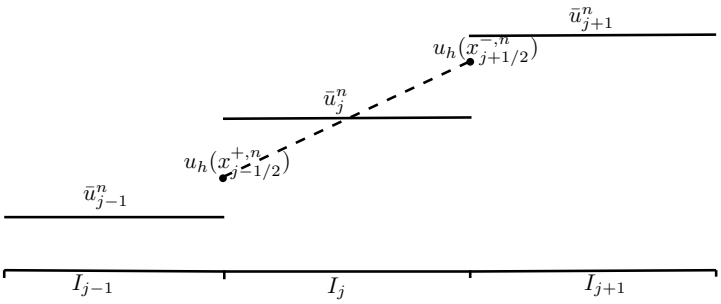

Figure A.1: All slopes have positive signs

the interval I_j are attained in $x_{\min} = x_{j-1/2}^+$ and $x_{\max} = x_{j+1/2}^-$. Recall that

$$\tilde{u}_{h_{|I_j}}(x) = \bar{u}_j^n + (x - x_j)\xi \left(\frac{u_h(x_{j+1/2}^{-,n}) - u_h(x_{j-1/2}^{+,n})}{\Delta x}, \frac{\bar{\mathsf{u}}_{j+1}^n - \bar{\mathsf{u}}_j^n}{\Delta x}, \frac{\bar{\mathsf{u}}_j^n - \bar{\mathsf{u}}_{j-1}^n}{\Delta x} \right)$$

and

$$m \leq \bar{\mathsf{u}}_{j-1}^n \leq \bar{\mathsf{u}}_j^n \leq \bar{\mathsf{u}}_{j+1}^n \leq M.$$

Based on Figure A.1, it is clear that if the limiter becomes active, the new polynomial is in the range $[m, M]$, since we have

$$(x - x_j)\xi \leq 0.5\Delta x \xi \leq 0.5 \min(\bar{\mathsf{u}}_{j+1}^n - \bar{\mathsf{u}}_j^n, \bar{\mathsf{u}}_j^n - \bar{\mathsf{u}}_{j-1}^n)$$

for the slope. Next, we need to ensure that the limiter actually becomes active if any over- or undershoot occurs. We denote $m_j = \min_x u_{h_{|I_j}}(x) = u_h(x_{j-1/2}^{+,n})$ and $M_j = \max_x u_{h_{|I_j}}(x) = u_h(x_{j+1/2}^{-,n})$. If $m_j < m$ and/or $M_j > M$, we either have $M_j - m_j >$

$\bar{\mathsf{u}}_{j+1}^n - \bar{\mathsf{u}}_j^n$ or $M_j - m_j > \bar{\mathsf{u}}_j^n - \bar{\mathsf{u}}_{j-1}^n$, respectively. For instance, if over- and undershoot occur simultaneously and additionally $M_j - M > m - m_j$, we get

$$\begin{aligned} M_j - m_j &\underset{M_j>M}{>} M - m_j \underset{m_j<m}{>} M - m > 1/2(M-m) = M - 1/2(M+m) \\ &\underset{\bar{\mathsf{u}}_{j+1}^n \in [m,M]}{>} \bar{\mathsf{u}}_{j+1}^n - 1/2(M+m) \underset{M_j>M}{>} \bar{\mathsf{u}}_{j+1}^n - 1/2(M_j+m) \\ &\underset{M_j - M > m - m_j}{>} \bar{\mathsf{u}}_{j+1}^n - \underbrace{1/2(M_j+m_j)}_{\bar{\mathsf{u}}_j^n}. \end{aligned}$$

All other cases can be treated analogously.

□

A.2 Monotonicity of the flux functions

For the functions $H_{1,\lambda}$ and $H_{2,\lambda}$ of (4.33a)-(4.33b), we have the derivatives:

$$\begin{aligned} \frac{\partial H_{1,\lambda}}{\partial \tilde{H}_{j-1}^n} &= \frac{\partial H_{2,\lambda}}{\partial \tilde{Q}_{j+1}^n} = \lambda a \geq 0 \\ \frac{\partial H_{1,\lambda}}{\partial \tilde{H}_{j+1}^n} &= \frac{\partial H_{2,\lambda}}{\partial \tilde{Q}_{j-1}^n} = 0 \geq 0 \\ \frac{\partial H_{1,\lambda}}{\partial \tilde{H}_{j}^n} &= \frac{\partial H_{2,\lambda}}{\partial \tilde{Q}_{j}^n} = 1 - \lambda a \geq 0 \text{ because of the CFL condition } \lambda a \leq 1. \end{aligned}$$

Computations for the derivatives of (4.36a)-(4.36b) with the CFL condition (4.37):

$$\begin{aligned} \frac{\partial G_{1,\lambda}}{\partial \tilde{H}_{j-1}^n} &= \frac{\partial G_{2,\lambda}}{\partial \tilde{Q}_{j+1}^n} = \frac{\lambda}{2}(a+\omega) \geq 0 \\ \frac{\partial G_{1,\lambda}}{\partial \tilde{H}_{j+1}^n} &= \frac{\partial G_{2,\lambda}}{\partial \tilde{Q}_{j-1}^n} = 0 \geq 0 \\ \frac{\partial G_{1,\lambda}}{\partial \tilde{Q}_{j}^n} &= \frac{\partial G_{2,\lambda}}{\partial \tilde{H}_{j}^n} = -\frac{\Delta t}{2} g'\left(\tilde{H}_j^n + \tilde{Q}_j^n\right) \leq 0 \text{ because } -g'(\cdot) \leq 0 \\ \frac{\partial G_{1,\lambda}}{\partial \tilde{H}_{j}^n} &= \frac{\partial G_{2,\lambda}}{\partial \tilde{Q}_{j}^n} = 1 - \left(\omega\lambda + \frac{\Delta t}{2} g'\left(\tilde{H}_j^n + \tilde{Q}_j^n\right)\right) \overset{!}{\geq} 0 \\ &\to 1 - \omega\lambda \geq \frac{\Delta t}{2} g'\left(\tilde{H}_j^n + \tilde{Q}_j^n\right) = \frac{\Delta t}{2}\left|g'\left(\tilde{H}_j^n + \tilde{Q}_j^n\right)\right| \leq \frac{\Delta t L}{2} \text{ because } g'(\cdot) \geq 0 \\ &\to 1 \geq \Delta t \left(\frac{1}{\Delta x}\omega + \frac{L}{2}\right) \\ &\to \Delta t \leq \frac{1}{\frac{1}{\Delta x}\omega + \frac{L}{2}}. \end{aligned}$$

Computations for the derivatives of (4.38a)-(4.38b) with CFL condition (4.39):

$$\frac{\partial D_{1,\lambda}}{\partial \tilde{H}^{-,n}_{j-1/2}} = \frac{\partial D_{2,\lambda}}{\partial \tilde{Q}^{+,n}_{j+1/2}} = \lambda a \geq 0$$
$$\frac{\partial D_{1,\lambda}}{\partial \tilde{H}^{+,n}_{j-1/2}} = \frac{\partial D_{2,\lambda}}{\partial \tilde{Q}^{-,n}_{j+1/2}} = \frac{1}{2} - \frac{\Delta t}{4} g'\left(\frac{1}{2}\left(\tilde{H}^{+,n}_{j-1/2} + \tilde{H}^{-,n}_{j+1/2} + \tilde{Q}^{+,n}_{j-1/2} + \tilde{Q}^{-,n}_{j+1/2}\right)\right) \geq 0$$
$$\frac{\partial D_{1,\lambda}}{\partial \tilde{H}^{-,n}_{j+1/2}} = \frac{\partial D_{2,\lambda}}{\partial \tilde{Q}^{+,n}_{j-1/2}} = \frac{1}{2} - \lambda a - \frac{\Delta t}{4} g'\left(\frac{1}{2}\left(\tilde{H}^{+,n}_{j-1/2} + \tilde{H}^{-,n}_{j+1/2} + \tilde{Q}^{+,n}_{j-1/2} + \tilde{Q}^{-,n}_{j+1/2}\right)\right) \geq 0$$
$$\frac{\partial D_{1,\lambda}}{\partial \tilde{Q}^{+,n}_{j-1/2}} = \frac{\partial D_{1,\lambda}}{\partial \tilde{Q}^{-,n}_{j+1/2}} = \frac{\partial D_{2,\lambda}}{\partial \tilde{H}^{+,n}_{j-1/2}} = \frac{\partial D_{2,\lambda}}{\partial \tilde{H}^{-,n}_{j+1/2}}$$
$$= -\frac{\Delta t}{4} g'\left(\frac{1}{2}\left(\tilde{H}^{+,n}_{j-1/2} + \tilde{H}^{-,n}_{j+1/2} + \tilde{Q}^{+,n}_{j-1/2} + \tilde{Q}^{-,n}_{j+1/2}\right)\right) \leq 0$$

To show that the second derivative is non-negative, we use the inequality $\Delta t \leq \frac{2}{L}$. For the third derivative, we need the inequality $\Delta t \leq \frac{1}{\frac{2a}{\Delta x} + \frac{L}{2}}$.

Computations for the derivatives of (4.40a)-(4.40b) with the CFL conditions (4.41):

$$\frac{\partial B_{1,\lambda}}{\partial \tilde{H}^{-,n}_{j-1/2}} = \frac{\partial B_{2,\lambda}}{\partial \tilde{Q}^{+,n}_{j+1/2}} = \lambda \mathbf{a} \geq 0$$
$$\frac{\partial B_{1,\lambda}}{\partial \tilde{H}^{+,n}_{j-1/2}} = \frac{\partial B_{2,\lambda}}{\partial \tilde{Q}^{-,n}_{j+1/2}} = \hat{\omega}_N - \frac{\Delta t}{2} \mathrm{g}'\left(\sum_{b=1}^{N}\left(\hat{\omega}_b q_j(\hat{x}^b_j) + \hat{\omega}_b h_j(\hat{x}^b_j)\right)\right)\hat{\omega}_N \geq 0 \rightarrow \Delta t \leq \frac{2}{L}$$
$$\frac{\partial B_{1,\lambda}}{\partial \tilde{H}^{-,n}_{j+1/2}} = \frac{\partial B_{2,\lambda}}{\partial \tilde{Q}^{+,n}_{j-1/2}} = \hat{\omega}_1 - \lambda \mathbf{a} - \frac{\Delta t}{2} \mathrm{g}'\left(\sum_{b=1}^{N}\left(\hat{\omega}_b q_j(\hat{x}^b_j) + \hat{\omega}_b h_j(\hat{x}^b_j)\right)\right)\hat{\omega}_1 \geq 0 \rightarrow \Delta t \leq \frac{1}{\frac{\hat{\omega}_1 \mathbf{a}}{\Delta x} + \frac{L}{2}}$$
$$\frac{\partial B_{1,\lambda}}{\partial h_j(\hat{x}^b_j)} = \frac{\partial B_{2,\lambda}}{\partial q_j(\hat{x}^b_j)} = \hat{\omega}_b - \frac{\Delta t}{2} \mathrm{g}'\left(\sum_{b=1}^{N}\left(\hat{\omega}_b q_j(\hat{x}^b_j) + \hat{\omega}_b h_j(\hat{x}^b_j)\right)\right)\hat{\omega}_b \geq 0 \rightarrow \Delta t \leq \frac{2}{L}$$
$$\frac{\partial B_{1,\lambda}}{\partial q_j(\hat{x}^b_j)} = \frac{\partial B_{2,\lambda}}{\partial h_j(\hat{x}^b_j)} = -\frac{\Delta t}{2} \mathrm{g}'\left(\sum_{b=1}^{N}\left(\hat{\omega}_b q_j(\hat{x}^b_j) + \hat{\omega}_b h_j(\hat{x}^b_j)\right)\right)\hat{\omega}_b \leq 0.$$

We know that $\hat{\omega}_1 = \hat{\omega}_N$ and $h_j(H_{j-1/2}) = H^{+,n}_{j-1/2}$, $h_j(H_{j+1/2}) = H^{-,n}_{j+1/2}$, $q_j(Q_{j-1/2}) = Q^{+,n}_{j-1/2}$, $q_j(Q_{j+1/2}) = Q^{-,n}_{j+1/2}$.

Bibliography

[1] J. Abreu, E. Cabrera, J. Izquierdo, and J. García-Serra. Flow modeling in pressurized systems revisited. *Journal of Hydraul. Eng. 125*, 11:1154–1169, 1999.

[2] L. Allievi, R. Dubs, and V. Bataillard. *Allgemeine Theorie über die veränderliche Bewegung des Wassers in Leitungen*. Springer, Berlin, 1909.

[3] R. Álvarez, N. B. Gorev, I. F. Kodzhespirova, Y. Kovalenko, S. Negrete, A. Ramos, and J. Rivera. Pseudotransient continuation method in extended period simulation of water distribution systems. *Journal of Hydraul. Eng. 134*, 10:1473–1479, 2008.

[4] C. Bolley and M. Crouzeix. Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques. *R.A.I.R.O. Anal. Numér.*, 12:237–245, 1978.

[5] G. Bollrich. *Technische Hydromechanik 1.* HUSS-Medien GmbH, 2007.

[6] D. Braess. *Finite Elemente – Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer Verlag, Bochum, 1996.

[7] S. E. Buckley and M. C. Leverett. Mechanism of fluid displacements in sands. *Transactions of the AIME*, 142:107–116, 1942.

[8] J. C. Butcher. *Numerical methods for ordinary differential equations*. Whiley, 2003.

[9] G. Chavent and B. Cockburn. The local projection P^0, P^1-discontinuous Galerkin finite element method for scalar conservation laws. *M^2AN Mathematical Modelling and numerical Analysis*, 23:565–592, 1989.

[10] B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor. *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*. Springer, 1997. Lecture Notes in Mathematics.

[11] B. Cockburn, S. Y. Lin, and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws iii: One-dimensional systems. *Journal of Computational Physics*, 84:90–113, 1989.

[12] B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws ii: General framework. *Mathematics of Computation*, 52:411–435, 1989.

[13] L. Collatz. *Differentialgleichungen*. Teubner Studienbücher Mathematik, B.G. Teubner Stuttgart, 1990.

[14] G. Dahlquist. *Stability and error bounds in the numerical integration of ordinary differential equations*. Trans. of Royal Inst. of Techn., N. 130, Stockholm, 1959.

[15] K. Dekker and J.G. Verwer. *Stability of Runge-Kutta methods for stiff nonlinear differential equations*. North-Holland, 1984.

[16] J. Deuerlein, A. R. Simpson, and E. Gross. The never ending story of modeling control-devices in hydraulic systems analysis. *In Proceedings of water distribution system analysis ASCE*, pages 1–12, 2008.

[17] P. Domschke. *Adjoint-Based Control of Model and Discretization Errors for Gas Transport in Networked Pipelines*. PhD thesis, TU Darmstadt, 2011.

[18] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations*. Studentliteratur, 1996.

[19] ERWAS. Zukunftsfähige technologien und konzepte für eine energieeffiziente und ressourcenschonende wasserwirtschaft – erwas. `http://www.fona.de/mediathek/pdf/SammelmappeERWAS_barrierefrei_web.pdf`, April 2017.

[20] L. Ferracina and M. N. Spijker. Stepsize restrictions for total-variation diminishing property in the general Runge-Kutta methods. *SIAM Journal of Numerical Analysis*, 42:1073–1093, 2004.

[21] L. Ferracina and M.N. Spijker. An extension and analysis of the Shu-Osher representation of the Runge-Kutta methods. *Mathematics of Computation*, 74:201–219, 2004.

[22] L. Ferracina and M.N. Spijker. Strong stability of singly-diagonally-implicit Runge-Kutta methods. *Applied Numerical Mathematics*, 58:1675–1686, 2008.

[23] B. Geißler, A. Morsi, and M. Walther. Spezifikation EWave-Optimierungsmodul. Intern Document, November 2015.

[24] E. Godlewski and P.-A. Raviart. *Numerical Approximation of hyperbolic systems of conservation laws*, volume 118. Applied Mathematical Sciences, Springer, 2002.

[25] D. Gottlieb and C.-W. Shu. On the Gibbs Phenomen and its resolution. *SIAM Review 4*, 39:644–668, 1997.

[26] S. Gottlieb, D. Ketcheson, and C. W. Shu. *Stong stability Preserving Time Discetizations*. World Scientific Press, 2010.

[27] S. Gottlieb and C.-W. Shu. Total-variation diminishing Runge-Kutta schemes. *Mathematics of Computation*, 67:73–85, 1998.

[28] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability preserving high-order time discretization methods. *SIAM:Review*, 43:89–112, 2001.

[29] C. Großmann and H.-G. Roos. *Numerische Behandlung partieller Differentialgleichungen*. Teubner Studienbücher Mathematik Verlag, Wiesbaden, 2005.

[30] C. Hähnlein. *Numerische Modellierung zur Betriebsoptimierung von Wasserverteilnetzen*. PhD thesis, TU Darmstadt, 2008.

[31] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*. Springer Verlag, 1991.

[32] A. Harten. High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49:357–393, 1983.

[33] P. Hausmann and G. Steinebach. Dokumentation – TWaveProg – Statistische Trinkwasserbedarfsprognose für das Wasserwerk Dorsten-Holsterhausen. Intern Document, Januar 2016.

[34] I. Higueras. On strong stability preserving time discetization methods. *Journal of Scientific Computing*, 21:193–223, 2004.

[35] I. Higueras. Representations of Runge-Kutta methods and strong stability preserving methods. *SIAM Journal on Numerical Analysis*, 43:924–948, 2005.

[36] W. Hundsdorfer and J.G. Verwer. *Numerical Solution of time-dependent advection-diffusion-reaction equations*, volume 33. Springer Series in Computational Mathematics, Springer, Heidelberg, 2003.

[37] The MathWorks Inc. Matlab R2016b, 2016.

[38] KANET. `http://kanet.iwg.uni-karlsruhe.de`.

[39] Y. Kawajir, F. Margot, C. Laird, S. Vigerske, and A. Wächter. Introduction to IPOPT: A tutorial for downloading, installing and using IPOPT. `https://www.coin-or.org/Ipopt/documentation/`, April 2015.

[40] D.I. Ketcheson, C.B. Macdonald, and S. Gottlieb. Optimal implicit strong stability preserving Runge-Kutta methods. *Applied Numerical Mathematics*, 59:373–392, 2009.

[41] T.A. Kocsis and A. Nèmeth. Optimal second order diagonally implicit SSP Runge-Kutta methods. *arXiv*, 2014.

[42] O. Kolb. *Simulation and Optimization of Gas and Water Supply Networks*. PhD thesis, TU Darmstadt, 2011.

[43] O. Kolb, J. Lang, and P. Bales. An implicit box scheme for subsonic compressible flow with dissipative source term. *Numerical Algoritms*, 53:293–307, 2010.

[44] O. Kolb and G. Steinebach. Modellkatalog und Datenstrukturen Wasserversorgung. Intern Document, 2016.

[45] B. Koren. *A robust upwind discretization for advection, diffusion and source terms. Numerical Methods for Advection-Diffusion Problems*, volume 45. Notes on numerical fluid mechanics, Vieweg, Braunschweig, 1993.

[46] J. F. B. M. Kraaijevanger. Contractivity of Runge-Kutta methods. *BIT*, 31:482–528, 1992.

[47] E. J. Kubatko, B. A. Yeager, and D. I. Ketcheson. Optimal strong-stability-preserving Runge-Kutta time discretizations for discontinuous Galerkin methods. *Journal of Scientific Computing*, 60:313–344, 2014.

[48] M. W. Kutta. *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*. PhD thesis, Universität München, 1901.

[49] J. Lang. Gleichmäßige Beschränktheit der Jacobi-Matrix bei mittels FEM semidiskretisierten linearen parabolischen Differentialgleichungen. Intern Document.

[50] P. D. Lax and R. D. Richtmyer. Survey of the Stability of linear finite Difference Equations. *Communications on pure and applied Mathematics*, IX:267–293, 1956.

[51] R. J. Leveque. *Finite volume methods for hyperbolic problems*. Cambridge University Press, 2002.

[52] B. Q. Li. *Discontinuous Finite Elements in Fluid Dynamics and Heat Transfer*. Springer, 2006.

[53] X.-D. Liu and S. Osher. Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I. *SIAM J. Numer. Anal.*, 33:760–779, 1996.

[54] A. Martin, K. Klamroth, J. Lang, G. Leugering, A. Morsi, M. Oberlack, M. Ostrowski, R. Rosen, and Editors. *Mathematical optimization of water networks*. International Series of Numerical Mathematics, Birkhäuser, Springer Basel, 2012.

[55] H. Martin and R. Pohl. *Technische Hydromechanik 4*. Verlag Bauwesen, 2000.

[56] G. Matthies and F. Schieweck. Higher order variational time discretizations for nonlinear systems of ordinary differential equations. *Preprint No. 23/2011, Fakultät für Mathematik, Otto-von-Guericke Universität Magdeburg*, pages 1–30, 2011.

[57] A. Meister and S. Ortleb. On unconditionally positive implicit time integration for DG scheme applied to shallow water flows. *International J. for Numer. Meth. in Fluids*, 2014.

[58] A. Morsi and B. Geißler. EWave-DOPT: Mathematical Model. Intern Document, August 2016.

[59] S. P. Nórsett. Semi explicit Runge-Kutta methods. Technical report, Report Dept. Math.No. 6/74, Univ. Trondheim, 1974.

[60] S. Osher and S. Chakravarthy. High resolution schemes and the entropy condition. *SIAM J. Numer. Anal.*, 21:955–984, 1984.

[61] D.A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of Mathèmatiques et Applications. Springer, 2012.

[62] W. H. Reed and T. R. Hill. Triangular mesh methods for neutron transport equation. *Technical Report La-UR-73-0479, Los Alamos Scientific Laboratory*, 1973.

[63] T. Richter, A. Springer, and B. Vexler. Efficient numerical realization of discontinuous Galerkin methods for temporal discretization of parabolic problems. *Numerische Mathematik 124(1)*, pages 151–182, 2013.

[64] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43:357–372, 1981.

[65] P. L. Roe. Characteristic-based schemes for the Euler equations. *Annu. Rev. Fluid Mech.*, 18:337–365, 1986.

[66] L.A. Rossmann. *EPANET 2 users manual*. U.S. Enviromental Protection Agency, Cincinnati, OH, 2000.

[67] C. Runge. Über die numerische Auflösung von Differentialgleichungen. *Springer*, 46:167–178, 1895.

[68] C.-W. Shu. Total-variation diminishing time discretizations. *SIAM Journal on Scientific and Statistical Computing*, 9:1073–1084, 1988.

[69] C.-W. Shu. *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, volume 1697. Advanced Numerical Approximation of Nonlinear Hyperbolic Equations.Series Lecture Notes in Mathematics, Springer, 2006.

[70] C.-W. Shu. Discontinuous Galerkin methods: general approach and stability. Lecture notes, 2009.

[71] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, 77:439–471, 1988.

[72] A. Sohr, R. Rosen, and T. Schenk. Energiemanagementsystem EWave für Trinkwasserversorgungssysteme. Technical report, wwt Modernisierungsreport, 2014/2015.

[73] P. Spellucci. A new technique for inconsistent qp problems in the sqp method. *Mathematical Methods of Operations Research*, 47(3):355–400, 1998.

[74] P. Spellucci. A SQP method for general nonlinear programs using only equality constrained subproblems. *Mathematical Programming*, 82(3):413–448, 1998.

[75] M. N. Spijker. Contractivity in the numerical solution of initial value problems. *Numerische Mathematik*, 42:271–290, 1983.

[76] STANET. `https://www.stafu.de`.

[77] G. Steinebach. TWaveSim – Ein Prozesssimulator für die Trinkwasserversorgung. Intern Document, March 2016.

[78] K. Strehmel and R. Weiner. *Linear implizite Runge-Kutta Methoden und ihre Anwendung*. Teubner-Verlag Stuttgart-Leipzig, 1992.

[79] K. Strehmel, R. Weiner, and H. Podhaisky. *Numerik gewöhnlicher Differentialgleichungen*. Springer Spektrum, 2012.

[80] T. Ström. On logarithmic norms. *SIAM J. Numer. Anal.*, 12, No. 5:741–753, 1975.

[81] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21(5):995–1011, 1984.

[82] R. Temam. *Navier-Stokes equations. Theory and numerical Analysis*. AMS Chelsea Publishing. American Mathematical Society, Providence Rhode Island, 2001.

[83] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25. Springer Series in Computational Mathematics, Springer, Berlin. Second edition, 2006.

[84] G. D. van Albada, B. van Leer, and W. W. Roberts Jr. A comparative study of the computational methods in cosmic gas dynamics. *Astronomy and Astrophysics*, 108:76–84, 1982.

[85] B. van Leer. Towards the ultimate conservation difference scheme. *Journal of Computational Physics*, 32:1–136, 1974.

[86] A. Wächter and L.T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[87] L. Wagner, J. Lang, and O. Kolb. Second order implicit schemes for scalar conservation laws. In *Lecture Notes in Comp. Science and Eng.*, volume 112, pages 33–41, 2016.

[88] X. Zhang and C.-W. Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *Journal of Computational Physics*, 2010.

[89] X. Zhang and C.-W. Shu. Maximum-principle-satisfying high-order schemes for conservation laws: survey and new developments. *Proceedings of the Royal Society A*, 2011.

[90] S. Zhao and G. W. Wei. A unified discontinuous Galerkin framework for time integration. *Mathematical Methods in the Applied Sciences*, 37:1042–1071, 2014.

Curriculum Vitae

Lisa Sabine Wagner
Born: September 6, 1988 in Bayreuth, Germany

Academic positions

05/'14 – 12/'17	Member, Numerical analysis and scientific computing, TU Darmstadt
05/'14 – 10/'17	Research assistant, BMBF Project EWave, TU Darmstadt
04/'16 – 12/'17	Associate member, GSC CE, TU Darmstadt

Education

04/2014	M.Sc. in Wirtschaftsmathematik, Universität Bayreuth
11/2011	B.Sc. in Wirtschaftsmathematik, Universität Bayreuth

Publications

L. Wagner, J. Lang, and O. Kolb. *Second order implicit schemes for scalar conservation laws*, Lecture Notes in Comp. Science and Eng., Vol. 112 (2016), pp. 33–41.