A high order discretization technique for singularly perturbed differential equations

Dissertation approved by the Faculty of Mathematics, Computer Science and Natural Science of RWTH Aachen University to obtain the academic degree *Doktor der Naturwissenschaften* and by Hasselt University to obtain the academic degree *Doctor in Science: Mathematics*

submitted by

Klaus Kaiser M.Sc. RWTH

from

Meschede, Germany

Supervisors:

Univ.-Prof. Dr. rer. nat. Sebastian Noelle, RWTH Aachen University

Prof. Dr. rer. nat. Jochen Schütz, Hasselt University

Referees:

Univ.-Prof. Dr. rer. nat. Sebastian Noelle, RWTH Aachen University

Prof. Dr. rer. nat. Jochen Schütz, Hasselt University

Univ.-Prof. Dr. rer. nat. Claus-Dieter Munz, University of Stuttgart

Date of oral exam:

September 17, 2018

D/2018/2451/64

Fakultät für
Mathematik, Informatik
und Naturwissenschaften

RWTH AACHEN
UNIVERSITY

A high order discretization technique for singularly perturbed differential equations

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University
zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Klaus Kaiser M.Sc. RWTH

aus

Meschede

Berichter:

Univ.-Prof. Dr. rer. nat. Sebastian Noelle

Prof. Dr. rer. nat. Jochen Schütz

Univ.-Prof. Dr. rer. nat. Claus-Dieter Munz

Tag der mündlichen Prüfung:

17.09.2018

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek verfügbar.

Abstract

The compressible Navier-Stokes equations converge towards their incompressible counterpart as the Mach number ε tends to zero. In the case of a weakly compressible flow, i.e. $\varepsilon \ll 1$, the resulting equations can be classified as singularly perturbed differential equations. Unfortunately, these equations set special requirements on numerical methods due to which standard discretization techniques often fail in efficiently computing an accurate approximation. One remedy is to split the equations into a stiff and a non-stiff part and then handle the stiff part implicitly and the non-stiff part explicitly in time. This procedure results in an IMEX method, with the crucial part being the choice of the splitting.

In this thesis the novel RS-IMEX splitting, which uses the $\varepsilon \to 0$ limit to split the equations by a linearization, is coupled with high order IMEX Runge-Kutta schemes. The resulting method is applied to different singularly perturbed differential equations and investigated in its behavior for $\varepsilon \ll 1$. This is done in the following steps:

First, the method is applied to a class of ordinary differential equations and it is proven in which way the resulting discretization suffers from order reduction. For this, it is shown that the convergence behavior depends on ε and that order reduction mainly depends on the implicit part of the discretization. This leads to an improved convergence behavior compared to an established splitting. Numerical computations show the influence of order reduction and a comparison to standard methods is provided.

Second, the isentropic Euler equations are considered to investigate the resulting method in the setting of a weakly compressible flow. For the spatial discretization a discontinuous Galerkin method is used. It is proven that the resulting method is consistent with the $\varepsilon \to 0$ limit of the equations, i.e. the overall algorithm is asymptotically consistent. Then, with the help of numerical computations an investigation of stability and accuracy is provided.

Overall, the method proposed in this thesis is a high order discretization for singularly perturbed differential equations which is consistent the $\varepsilon \to 0$ limit and shows the desired behavior in the low Mach setting.

Zusammenfassung

Die kompressiblen Navier-Stokes-Gleichungen konvergieren gegen das inkompressible Gegenstück, wenn die Mach-Zahl ε gegen Null geht. Für den Fall einer schwach kompressiblen Strömung, d.h. $\varepsilon \ll 1$, können die Gleichungen als singulär gestörte Differentialgleichungen angesehen werden. Diese Gleichungen stellen bestimmte Voraussetzungen an numerische Verfahren, wodurch Standard-Methoden oft nicht in der Lage sind, eine genaue Approximation effizient zu berechnen. Eine Möglichkeit, dieses Problem zu beheben, ist, die Gleichungen in einen steifen und einen nicht steifen Teil zu zerlegen und dann den steifen Teil implizit und den nicht steifen Teil explizit in der Zeit zu diskretisieren. Dieses Verfahren resultiert in eine IMEX-Methode, wobei der entscheidene Teil die Wahl der Zerlegung ist.

In dieser Arbeit wird das neue RS-IMEX Splitting, das den $\varepsilon \to 0$ Limit verwendet, um die Gleichungen mittels einer Linearisierung aufzuteilen, mit IMEX-Runge-Kutta-Verfahren hoher Ordnung gekoppelt. Die resultierende Methode wird auf verschiedene singulär gestörte Differentialgleichungen angewandt und in ihrem Verhalten für $\varepsilon \ll 1$ untersucht. Dies wird in den folgenden Schritten getan:

Zuerst wird die Methode auf eine Klasse von gewöhnlichen Differentialgleichungen angewandt und es wird gezeigt, inwiefern die resultierende Diskretisierung unter Ordnungsreduktion leidet. Hierfür wird gezeigt, dass das Konvergenzverhalten von ε abhängt und dass die Ordnungsreduktion hauptsächlich vom impliziten Teil der Diskretisierung bestimmt wird. Dies führt zu einem verbesserten Konvergenzverhalten verglichen mit einer Standardzerteilung. Numerische Berechnungen zeigen den Einfluss der Ordnungsreduktion und ein Vergleich mit etablierten Methoden wird durchgeführt.

Als zweites wird die Methode auf die isentropen Euler-Gleichungen angewandt und für den Fall schwach kompressibler Strömungen untersucht. Für die räumliche Diskretisierung wird ein unstetiges Galerkin-Verfahren verwendet. Es wird gezeigt, dass die resultierende Methode mit dem $\varepsilon \to 0$ Limit der Gleichungen konsistent ist, sie ist also asymptotisch konsistent. Dann werden mit der Hilfe von numerischen Berechnungen die Stabilität und Genauigkeit untersucht.

In dieser Arbeit wird eine Diskretisierung hoher Ordnung für singulär gestörte Differentialgleichungen vorgeschlagen, die konsistent mit dem $\varepsilon \to 0$ Limit ist und das gewünschte Verhalten im Fall kleiner Mach-Zahlen zeigt.

Samenvatting

In deze thesis worden de Navier-Stokes vergelijkingen voor een zwak-samendrukbare vloeistof beschouwd. Als het Mach getal ε naar nul convergeert, convergeren de Navier-Stokes vergelijkingen naar hun tegenhangers voor onsamendrukbare vloeistoffen. Voor alleen maar zwak samendrukbare vloeistoffen stellen de vergelijkingen dus een singulier gestoord probleem voor. Standaard numerieke methoden zijn minder geschikt voor deze klasse van vergelijkingen omdat ze zich hier vaak zeer inefficiënt gedragen. Een oplossing is dus om de vergelijking in twee delen op te splitsen, namelijk een stijf en een niet-stijf deel. Het stijve deel wordt vervolgens impliciet, en het niet-stijve deel expliciet in de tijd behandeld. Dit levert de IMEX methode; een zeer belangrijk deel is hierbij de keuze van de splitsing.

Hier beschouwen wij de nieuwe RS-IMEX splitsing, die gebruik maakt van de $\varepsilon \to 0$ limiet om de vergelijking via een linearisatie van de flux op te splitsen. De RS-IMEX splitsing wordt aan een IMEX Runge-Kutta methode van hoge orde gekoppeld, de finale methode wordt vervolgens op een aantal singulier gestoorde differentiaalvergelijkingen toegepast. Het gedrag van de methode voor $\varepsilon \ll 1$ wordt geanalyseerd.

Eerst passen wij de methode op een klasse van gewone differentiaalvergelijkingen toe. We tonen aan hoe orde-reductie in het spel komt door te laten zien dat het convergentiegedrag voornamelijk van de impliciete discretisatie afhangt. Dit levert een verbetering ten opzichte van standaardmethoden.

Vervolgens passen wij de methode op de isentrope Euler vergelijking toe. Voor de ruimtelijke discretisatie maken wij gebruik van de discontinue Galerkin methode. We bewijzen dat de methode ook voor $\varepsilon \to 0$ het juiste resultaat levert, de methode is dus asymptotisch consistent. Bovendien wordt stabiliteit en nauwkeurigheid van de methode met behulp van numeriek onderzoek besproken.

De in deze thesis voorgestelde methode is een discretisatie methode van hoge orde voor singulier gestoorde differentiaalvergelijkingen die de $\varepsilon \to 0$ limiet van de vergelijkingen respecteert.

Acknowledgment

During the last years I received unbelievable support of many people. It is nearly impossible to thank all of them and to express how grateful I am for their support. Therefore, it is a pleasure to name at least some of them.

First of all, I would like to thank Sebastian Noelle and Jochen Schütz for their scientific guidance and support during this thesis. Both of them had always an open ear for my problems and time for discussions. It was a privilege to work with them on this topic. Furthermore, I would like to thank Claus-Dieter Munz for reviewing my thesis and also for his helpful comments. In addition, I would like to thank Andrea Beck and Jonas Zeifang for several fruitful discussions which helped to improve this thesis and also resulted in first extensions.

Michael Rom did a great job in proofreading preliminary versions of my thesis. He had tons of comments which improved this work. With Alexander Jaust I had several discussions concerning C++, Netgen, NGSolve, PETSc, LaTeX issues. He always helped me with his broad knowledge in these topics. I'm deeply thankful for the help of both of them.

During the past years I was member of the Institut für Geometrie und Praktische Mathematik (Aachen) and the Computational Mathematics group (Hasselt). The kindness and helpfulness of my colleagues were unbelievable. I'm proud that I found several good friends during this time. Whenever necessary my colleagues, my friends, my family and my girlfriend helped me to clear my mind and to keep on track. I'm deeply grateful for their support and encouragement.

Finally, I would like to thank the Deutsche Forschungsgesellschaft, projects NO 361/3-3 and NO 361/6-1, and Hasselt University, Special Research Fund BOF16BL08, for financial support.

Contents

1. Introduction

In the past decades, several oil spill disasters caused the pollution of large coastal regions and the death of thousands of animals. Two examples are the accidents of the Deepwater Horizon[1] and the Exxon Valdez[2]. In both cases an oil spill was released. Oil spills travel with the movement of the water. Therefore, to predict in which way the oil spill behaves in the deep ocean and which coastal regions are effected, it is necessary to use a proper model of the water flow, see Figure 1.1. The deep ocean contains extremely different scales: the ocean is large compared to the water depth and the water depth is large compared to the height of water waves which are related to the movement of the water. Additionally, standard models also resolve fast gravity waves, which are much faster than the movement of the water. Due to these large scale differences, standard numerical methods can fail for example because of huge computational cost.

Going to a more general setting, the small water waves can be seen as a compressibility effect, which is described by a parameter ε. Therefore, the flow situation can be seen as weakly compressible or nearly incompressible, which means that ε is small. Since we want to resolve the small compressibility effects, an incompressible model cannot be used. On the other hand, the compressible and incompressible models are in direct relation to each other. This can be seen by taking the $\varepsilon \to 0$ limit and obtaining that the model changes its type and transforms from compressible to incompressible [104]. In mathematics, such equations belong to the class of *singularly perturbed differential equations*.

It is often desirable to use a *high order numerical method* which means that the approximation is locally computed from a richer space of ansatz functions. For the oil spill example this means that a low order method needs to divide the complete ocean into small cells to resolve the water movement accurately, while the cells for a high order method are allowed to be much larger to obtain the same accuracy. The goal of this thesis can hence be formulated as the development of a

high order numerical method for singularly perturbed differential equations.

To approximate time dependent flows, it is necessary to discretize temporal derivatives which becomes more difficult as the small parameter ε sets specific requirements on the used method. This is why we need to use an implicit time discretization. Unfortunately, this leads to high computational cost and one would prefer an explicit time discretization method since these are very efficient if ε is much larger than zero. A remedy is to split the equations into two parts and then use for one part an implicit and for the other

[1] https://en.wikipedia.org/wiki/Deepwater_Horizon
[2] https://en.wikipedia.org/wiki/Exxon_Valdez

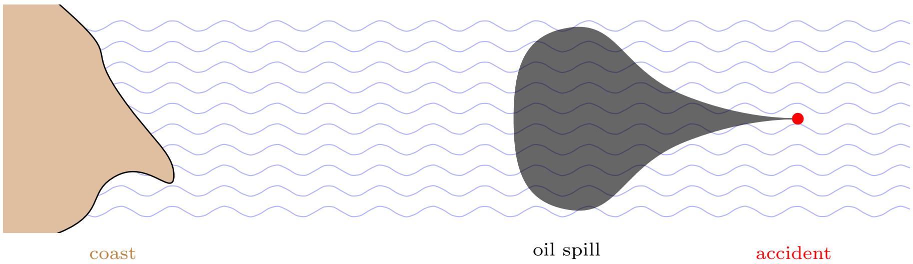


Figure 1.1.: Illustration of an oil spill caused by an accident in the deep ocean and traveling towards a coast. The wave height, which corresponds to compressibility in a more general setting, is extremely small compared to the size of the ocean. Therefore the flow can be seen as weakly compressible.

part an explicit discretization. This results in an IMplicit-EXplicit (IMEX) time discretization [12, 13, 101]. The crucial part is then the choice of the splitting: the splitting is one ingredient which determines whether the resulting discretization respects the special requirements of the equations. In this thesis, we use the newly developed Reference Solution (RS)-IMEX splitting [J3, J4, 186] which relies on the $\varepsilon \to 0$ limit of the model. To obtain a high order discretization we use IMEX Runge-Kutta schemes [12, 101]. Furthermore, we use a discontinuous Galerkin method [38, 40, 42, 43, 44] for the discretization of spatial derivatives.

In principle, the resulting method can directly be applied to the compressible Navier-Stokes equations. The question of existence and uniqueness of solutions for the Navier-Stokes equations is one of the famous millennium problems[3]. It is therefore quite canonical to start the derivation of a new numerical method with smaller test equations which have similar properties as the Navier-Stokes equations but can be analyzed in more detail. Using this strategy, the components of the numerical discretization can be tested and prepared for the full Navier-Stokes equations. The prototype equations we discuss in this thesis are a class of ordinary differential equations and the isentropic Euler equations. The resulting method is then analyzed concerning order reduction, stability and the limiting behavior:

- Order reduction [24, 79, 81] is a problem which occurs if a high order Runge-Kutta scheme is applied to singularly perturbed differential equations and leads to an ε depending range of step sizes where the numerical approximation shows a reduced order of convergence.
- The stability of a numerical method is an extremely important property since it guarantees that the numerical method is able to compute an approximation.
- The equations we consider show a specific behavior as $\varepsilon \to 0$. Ideally, the numerical method is able to resolve this behavior [96], i.e. the numerical method for weakly compressible flows should also change its type as $\varepsilon \to 0$ and transform towards a method for incompressible flows.

In this thesis, we prove in which way the proposed method suffers from order reduction and whether the proposed method is consistent with the $\varepsilon \to 0$ limit. For further investigations and to substantiate our analysis we consider different examples and analyze the performance of the proposed method concerning stability and accuracy. Through this we achieve the following main contributions:

- The RS-IMEX splitting is applied to different singularly perturbed differential equations and coupled with a high order IMEX Runge-Kutta and discontinuous Galerkin method.
- It is shown that the resulting method for ordinary differential equations suffers from order reduction, which depends on the chosen IMEX Runge-Kutta scheme. Furthermore, it is shown that the order reduction is less significant than order reduction obtained by a standard splitting from literature and similar to order reduction obtained by a fully implicit discretization.
- It is shown for the isentropic Euler equations that the resulting method is consistent with the limiting behavior of the equations. Furthermore, it is motivated with numerical computations that the method is stable even for small ε, and accurate for weakly compressible flows.

Overall, the method we propose in this thesis is able to discretize a class of singularly perturbed differential equations, including the isentropic Euler equations, and is consistent with the limiting behavior of the equations. Therefore, the proposed method is a high order discretization for weakly compressible flows which can in principle be extended to the full Navier-Stokes equations. To give an overview of the structure of this thesis we shortly summarize the contents of the following chapters:

- In Chapter 2, we start with an introduction on singularly perturbed differential equations. For this, we introduce a class of ordinary differential equations and the isentropic Euler equations and discuss their behavior as $\varepsilon \to 0$. We also introduce different examples, which are used for the numerical computations later in this thesis.

[3] http://www.claymath.org/millennium-problems/navier–stokes-equation

- In Chapter 3, we introduce the numerical discretization proposed in this thesis, i.e. the IMEX Runge-Kutta method, RS-IMEX splitting and discontinuous Galerkin method. For this, we discuss methods and splittings for weakly compressible flows from literature. We also introduce and discuss asymptotic properties that the resulting method should fulfill.

- In Chapter 4, we investigate the time discretization and observe that the method suffers from order reduction. We are able to show that the order reduction is less significant compared to the same IMEX Runge-Kutta scheme coupled with a standard splitting and similar to a fully implicit discretization. The influence of order reduction is then shown by several numerical computations and a comparison of different IMEX Runge-Kutta schemes is provided.

- In Chapter 5, we consider the numerical method for the isentropic Euler equations to investigate their behavior concerning the asymptotic properties. For this, we prove that the method is consistent with the limiting behavior of the equations, and numerical computations show that for a weakly compressible flow the method is stable and provides the desired order of convergence if ε is small.

- In Chapter 6, the thesis is finalized with a short conclusion and a discussion of further steps in the development of a high order method for singularly perturbed differential equations.

Parts of this thesis are related to previously published works: the results for the RS-IMEX splitting for ordinary differential equations rely on the works [J2, J4] and the results for the isentropic Euler equations rely on the works [J1, J3, P1, C3, J5].

2. Singularly perturbed differential equations

In this chapter, we introduce and analyze singularly perturbed differential equations which can be motivated by weakly compressible flows. We start with a short introduction in the physical setting, including some motivating examples in Section 2.1. Then, we shortly introduce the governing equations used in this thesis, see Section 2.1.1 for the isentropic Euler equations and 2.1.3 for the ordinary differential equations. These equations are singularly perturbed differential equations and therefore we analyze their behavior when the small parameter tends to zero in Section 2.2. Finally, this chapter is closed by introducing several numerical configurations in Section 2.3 for the isentropic Euler and the ordinary differential equations that we use later in this thesis to test the numerical methods.

2.1. Problem settings (low Mach flows)

In the following, a short introduction of compressible flows, see [8, 138] for more details on this topic, is given - the focus in this work is on low Mach flows. In compressible fluid dynamics, flows are often characterized by the Mach number, a dimensionless quantity named after the physicist Ernst Mach[1], describing the relation between the flow velocity $\boldsymbol{u}$ and the local speed of sound c:

$$\mathrm{Ma} := \frac{\|\boldsymbol{u}\|}{c}.$$

The speed of sound describes how fast sound waves move through a fluid and is given by

$$c = \sqrt{\frac{1}{\rho\tau}}, \tag{2.1}$$

where τ is a measure of the fluid's compressibility and ρ its density, see [8]. In more detail, the compressibility describes how the density ρ changes due to a change of the pressure p:

$$\tau = \left(\rho\frac{\partial p}{\partial \rho}\right)^{-1}. \tag{2.2}$$

In Figure 2.1 a characterization of flows depending on the Mach number is given. In this thesis low Mach flows are considered which means that

$$\mathrm{Ma} \ll 1,$$

or in other words that the fluid velocity in these flows is much smaller than the speed of sound. Thus, one observes extremely different velocities in the system, due to waves traveling with a slow velocity and sound waves traveling with the speed of sound. Since the speed of sound depends on the fluid compressibility τ one observes a weakly compressible fluid. This situation can also be described as nearly incompressible. For the specific flow weakly compressible effects or sound waves might be very important. This is why the approximation with an incompressible model (where the speed of sound can be seen as infinite) is not necessarily appropriate.

Naturally, the velocity may change from slow to fast or vice versa during the process one observes. This results in a situation where one would have both low Mach numbers, i.e. a weakly compressible flow, and also large Mach numbers, i.e. a fully compressible flow.

[1] Ernst Mach, 1838 – 1916

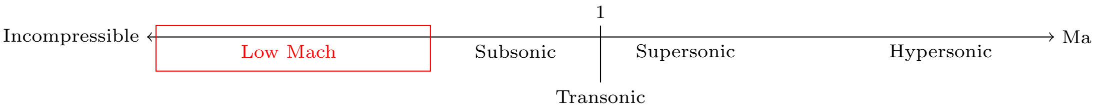


Figure 2.1.: Characterization of flows depending on the Mach number [8, 138], the red rectangle marks flows considered in this thesis. The scale is logarithmic and Ma → 0 corresponds to incompressible flows.

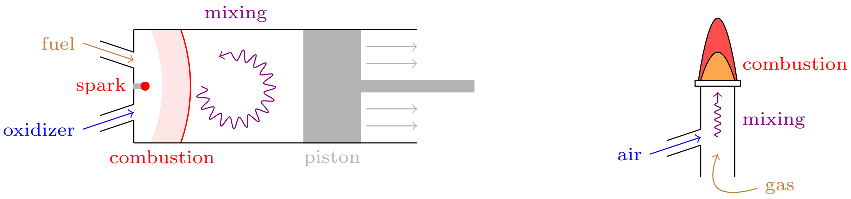


Figure 2.2.: Left: Illustration of combustion in a cylinder, right: Illustration of a Bunsen burner. In both examples the fuel/gas and oxidizer/air are mixed in a chamber. During this mixing particles move very slow compared to the speed of sound and therefore a low Mach flow is obtained.

Using low Mach models in computational fluid dynamics and especially resolving the weakly compressible behavior is a difficult task and a current research topic, see Section 3.2.1 for a discussion on the difficulties and Section 3.1 for recent works in this topic. In the following, two additional, see the introduction for a first example, motivating examplary low Mach flows are presented. All of these examples are an active field of research.

Remark 2.1. *The following physical examples and the example given in the introduction should be understood as a motivation for this thesis. Their full treatment is beyond the scope of this work.*

1) *Combustion:* The combustion in a cylinder of an engine is a relatively slow process, see [145] and the references therein: Fuel and an oxidizer are injected into a cylinder which is closed with a movable piston. The resulting flow is turbulent and therefore both components are mixed for a sufficiently long time. Afterwards, a spark is produced to ignite the gas.

   During this process, the fluid moves relatively slowly compared to the speed of sound. Therefore it is also called slow speed combustion, and consequently a low Mach flow.

   Besides the specific combustion in an engine, combustion in a Bunsen burner can be seen as a low Mach flow as well, see [118]. In Figure 2.2 an illustration of both is given.

2) *Aeroacoustics:* In recent years, aeroacoustics gained increasing interest in industry and science [73, 124]. In this field of research the noise caused by turbulent flow around solid structures is investigated.

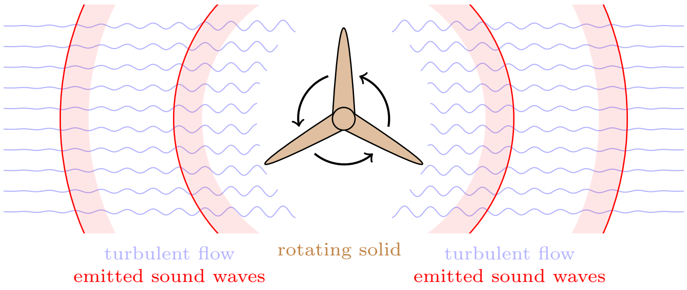


Figure 2.3.: Illustration of a flow around a solid moving structure with emitted sound waves. Due to the airflow around the moving structure a rapid change of pressure can be observed. By this, sound waves are emitted which are faster compared to the airflow and therefore a low Mach flow is given.

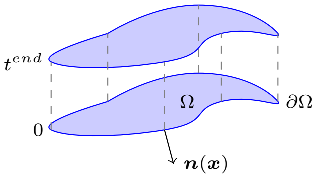


Figure 2.4.: Illustration of the domain Ω_T including the boundary $\partial\Omega$ with a normal-vector $\boldsymbol{n}$.

If a turbulent flow interacts with a solid moving structure, e.g. a fan or a rotor, a rapid change of pressure can be observed on the surface of the structure, see Figure 2.3 for an illustration of this example. This change results in the radiation of sound waves. In general the turbulent flow is much slower than the speed of sound and therefore a low Mach flow is obtained.

This list is by no means complete, there are several additional examples like atmospheric flows [107] or the long time modeling of a supernova [7].

2.1.1. Isentropic Euler equations

In computational fluid dynamics a compressible flow is often modeled by the Euler[2] or Navier[3]-Stokes[4] equations. To construct and test numerical methods in this field, it is often useful to consider simplified equations which show a similar behavior as the original ones but are slightly easier to handle. This is why we consider one special case of the Euler equations – the isentropic Euler equations – in this thesis. The reader is referred to [8, 116, 123, 181] for a more detailed introduction of computational fluid dynamics and the governing equations.

An isentropic flow is present if it is adiabatic, i.e. a process in which there is no exchange of heat with the surroundings, and reversible [8].

Remark 2.2. *For the sake of simplicity, we restrict ourselves to the two-dimensional case, but all results are directly extendable to three dimensions.*

We monitor the flow in a fixed domain Ω, where $\Omega \subset \mathbb{R}^2$ is open but bounded such that $|\Omega| < \infty$ and $\partial\Omega$ denotes the boundary, and over a fixed period of time starting from zero up to $t^{end} \in \mathbb{R}^{>0}$. Consequently, the whole domain of interest, see also Figure 2.4 for an illustration, is given by

$$\Omega_T := (0, t^{end}) \times \Omega \subset \mathbb{R}^{\geq 0} \times \mathbb{R}^2. \tag{2.3}$$

Isentropic flows in two dimensions are described by the scalar density ρ and the vector-valued velocity $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2)^T$. Both depend on time t and space $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)^T$, or more formally

$$\rho : \overline{\Omega_T} \to \mathbb{R},\ (t, \boldsymbol{x}) \mapsto \rho(t, \boldsymbol{x}) \qquad \text{and} \qquad \boldsymbol{u} : \overline{\Omega_T} \to \mathbb{R}^2,\ (t, \boldsymbol{x}) \mapsto \boldsymbol{u}(t, \boldsymbol{x}).$$

Note that we may drop the dependence on $(t, \boldsymbol{x})$ to keep the notation simple. The velocity $\boldsymbol{u}$ is in direct relation to the momentum-density $\rho\boldsymbol{u}$ which is the product of density and velocity. For a fluid it is known that mass and momentum are conserved during the process which means that in a fixed area mass and momentum only change due to the flow through the boundary. Thus, the corresponding densities of an isentropic flow, if they are assumed to be smooth, satisfy the *conservation of mass equation*

$$\partial_t \rho + \nabla_{\boldsymbol{x}} \cdot (\rho\boldsymbol{u}) = 0 \qquad \text{in} \qquad \Omega_T, \tag{2.4}$$

and the *conservation of momentum equation*

$$\partial_t (\rho\boldsymbol{u}) + \nabla_{\boldsymbol{x}} \cdot (\rho\boldsymbol{u} \otimes \boldsymbol{u}) + \nabla_{\boldsymbol{x}} p = 0 \qquad \text{in} \qquad \Omega_T, \tag{2.5}$$

[2] Leonhard Euler, 1707 – 1783
[3] Claude-Louis Navier, 1785 – 1836
[4] Sir George Stokes, 1819 – 1903

where p denotes the pressure. The equations are closed by a law for the pressure which is given for an isentropic flow of an ideal gas by

$$p := p(\rho) := \kappa\rho^{\gamma}, \tag{2.6}$$

where the constants $\kappa > 0$ and $\gamma \geq 1$. Both κ and γ are defined by the considered fluid and γ denotes the ratio of specific heats.

Equations (2.4) and (2.5) use temporal and spatial derivatives. This means that we need to prescribe initial and boundary conditions. Hence,

$$\rho(t = 0, \boldsymbol{x}) = \rho^0(\boldsymbol{x}) \qquad \text{and} \qquad \boldsymbol{u}(t = 0, \boldsymbol{x}) = \boldsymbol{u}^0(\boldsymbol{x}) \qquad \text{for} \qquad \boldsymbol{x} \in \Omega,$$

with given functions ρ^0 and $\boldsymbol{u}^0$. Naturally, the choice of boundary conditions depends on the considered flow. There could be, among others, outflow / inflow, solid wall and far field boundaries. As an example, the solid wall boundary conditions are chosen in such a way that the flow in direction of the boundary is 0, in detail

$$\boldsymbol{u}(t, \boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) = 0 \qquad \text{for} \qquad (t, \boldsymbol{x}) \in [0, t^{end}] \times \partial\Omega, \tag{2.7}$$

where $\boldsymbol{n}(\boldsymbol{x})$ denotes the outer normal vector of the boundary $\partial\Omega$ at $\boldsymbol{x} \in \partial\Omega$. For a detailed discussion on boundary conditions we refer to [8, 181]. The isentropic Euler equations are summarized in the following definition.

Definition 2.3 (Isentropic Euler equations)**.** *Let $\Omega_T \subset \mathbb{R}^{\geq 0} \times \mathbb{R}^2$ be a given domain, then the* isentropic Euler equations *are given by*

$$\partial_t \begin{pmatrix} \rho \\ \rho\boldsymbol{u} \end{pmatrix} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{F} = 0 \qquad \textit{in} \qquad \Omega_T, \qquad \textit{where} \qquad \boldsymbol{F} := \begin{pmatrix} \rho\boldsymbol{u} \\ \rho\boldsymbol{u} \otimes \boldsymbol{u} + p\,\mathrm{Id} \end{pmatrix}$$

and $p = \kappa\rho^{\gamma}$ with $\kappa > 0$ and $\gamma \geq 1$. The equations are equipped with initial conditions

$$\begin{pmatrix} \rho(t = 0, \boldsymbol{x}) \\ \rho(t = 0, \boldsymbol{x})\boldsymbol{u}(t = 0, \boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} \rho^0(\boldsymbol{x}) \\ \rho^0(\boldsymbol{x})\boldsymbol{u}^0(\boldsymbol{x}) \end{pmatrix} \qquad \textit{for} \qquad \boldsymbol{x} \in \Omega,$$

where ρ^0 and $\boldsymbol{u}^0$ are given functions; and suitable boundary conditions.

Remark 2.4. *The unknowns of the isentropic Euler equations are ρ and $\rho\boldsymbol{u}$, but as $\rho\boldsymbol{u}$ can be seen as $\rho \cdot \boldsymbol{u}$, we often use $\boldsymbol{u}$ instead of $\rho\boldsymbol{u}/\rho$.*

The isentropic Euler equations as given in Definition 2.3 form a system of partial differential equations given in conservative form with flux function $\boldsymbol{F}$, which is characterized as hyperbolic, see [74, 114, 121, 124] and the references therein for an introduction on the theory of hyperbolic conservation laws.

Definition 2.5 (Hyperbolic conservation law)**.** *We call a partial differential equation of the form*

$$\partial_t \boldsymbol{w} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{F} = 0$$

hyperbolic if $\nabla_{\boldsymbol{w}} \boldsymbol{F} \cdot \boldsymbol{n}$ has only real eigenvalues and is diagonalizable.

As mentioned in Section 2.1, the speed of sound is an important quantity in the low Mach setting and can be seen as the inverse of the product of compressibility τ and density ρ, see Equation (2.1). For an ideal gas and isentropic flow, we can rewrite the compressibility τ given in Equation (2.2) such that it only depends on the ratio of specific heats and the pressure, $\tau = (\gamma p)^{-1}$, and consequently

$$c = \sqrt{\gamma \frac{p}{\rho}}. \tag{2.8}$$

The importance of the speed of sound can directly be seen by computing the speeds in the system which are given by the eigenvalues $\lambda_{1,2,3}$ of the Jacobian of the flux function $\nabla_{\rho,\rho\boldsymbol{u}}\boldsymbol{F}\cdot\boldsymbol{n}$:

$$\lambda_1 = \boldsymbol{u}\cdot\boldsymbol{n} \qquad \text{and} \qquad \lambda_{2,3} = \boldsymbol{u}\cdot\boldsymbol{n} \pm \sqrt{p'(\rho)}, \tag{2.9}$$

where $\boldsymbol{n}\in\mathbb{R}^2$ is an arbitrary normal direction. For the isentropic flow of an ideal gas, $\sqrt{p'(\rho)}$ can be also computed by $\sqrt{\gamma p/\rho}$ which is the speed of sound c as given in Equation (2.8). Consequently, we obtain

$$\lambda_1 = \boldsymbol{u}\cdot\boldsymbol{n} \qquad \text{and} \qquad \lambda_{2,3} = \boldsymbol{u}\cdot\boldsymbol{n} \pm c.$$

These eigenvalues differ extremely in order of magnitude for a low Mach flow, i.e. if $\|\boldsymbol{u}\| \ll c$. In detail λ_1 describes the slow waves and $\lambda_{2,3}$ the fast waves which travel approximately with the speed of sound.

In the beginning of this section we mentioned that the isentropic Euler equations are chosen to develop and test numerical methods without having one of the physical examples given before in mind. In the following remark, we shortly discuss the relation between the isentropic Euler equations and the flows in the deep ocean.

Remark 2.6. *In Chapter 1, flows in the deep ocean have been mentioned as an example. These types of flows are often modeled by shallow water equations, also called Saint-Venant*[5] *equations, see e.g. [69, 123, 177]. Shallow water equations are identical to isentropic Euler equations with the so-called Froude*[6] *number taking the role of the Mach number and the water height h taking the role of the density ρ. Furthermore, the pressure law is given by $p(h) = \frac{1}{2}h^2$, i.e. $\kappa = \frac{1}{2}$ and $\gamma = 2$. Note that for describing an ocean the bottom topography is needed, which cannot be described with the isentropic Euler equations.*

Non-dimensionalization

All quantities in the isentropic Euler equations as given in Definition 2.3 are physically motivated quantities with corresponding units. It is useful to reformulate the equations in dimensionless quantities such that effects due to different scales and units are eliminated and structural aspects of the equations become clearer.

In the following, we introduce several reference quantities, denoted with $(\cdot)^*$, to non-dimensionalize the isentropic Euler equations as given in Definition 2.3 following the same steps as [5, 78], see also the references therein, for equations in fluid dynamics.

Note that the reference quantities $(\cdot)^*$ are equipped with the corresponding units and that the particular choice depends on the specific flow situation. The resulting dimensionless quantities are denoted by $\overline{(\cdot)}$. In detail, we define

$$\overline{\boldsymbol{x}} := \frac{\boldsymbol{x}}{x^*}, \quad \overline{t} := \frac{t}{t^*}, \quad \overline{\rho} := \frac{\rho(t^*\overline{t}, x^*\overline{\boldsymbol{x}})}{\rho^*}, \quad \overline{\boldsymbol{u}} := \frac{\boldsymbol{u}(t^*\overline{t}, x^*\overline{\boldsymbol{x}})}{u^*} \quad \text{and} \quad \overline{p} := \frac{p(\rho^*\overline{\rho})}{p^*}. \tag{2.10}$$

The change in the temporal and spatial variables directly affects the corresponding derivatives which also have to be changed, in detail

$$\partial_t(\cdot) = \frac{d}{dt}\left(\frac{t}{t^*}\right)\partial_{\overline{t}}(\cdot) = \frac{1}{t^*}\partial_{\overline{t}}(\cdot) \qquad \text{and similarly} \qquad \nabla_{\boldsymbol{x}}(\cdot) = \frac{1}{x^*}\nabla_{\overline{\boldsymbol{x}}}(\cdot).$$

The resulting non-dimensional isentropic Euler equations are summarized in the following lemma. Note that the reference velocity u^* may directly depend on reference time t^* and length x^*. Therefore it is useful to choose some reference quantities in relation to another, see the proof of Lemma 2.7 for more details.

Lemma 2.7. *Assume that non-dimensional variables are given as in* (2.10) *with $u^* = x^*/t^*$, then the*

[5] Adhémar Jean Claude Barré de Saint-Venant, 1797 – 1886

[6] William Froude, 1810 – 1879

isentropic Euler equations in non-dimensional form are given by

$$\partial_{\overline{t}} \begin{pmatrix} \overline{\rho} \\ \overline{\rho \boldsymbol{u}} \end{pmatrix} + \nabla_{\overline{\boldsymbol{x}}} \cdot \overline{\boldsymbol{F}} = 0, \qquad \textit{where} \qquad \overline{\boldsymbol{F}} := \begin{pmatrix} \overline{\rho \boldsymbol{u}} \\ \overline{\rho \boldsymbol{u}} \otimes \overline{\boldsymbol{u}} + \frac{1}{\varepsilon^2} \overline{p}\,\mathrm{Id} \end{pmatrix}$$

$$\textit{and} \qquad \varepsilon = \sqrt{\frac{\rho^*(u^*)^2}{p^*}}.$$

Proof. We start with the conservation of mass equation (2.4) and replace all quantities with their non-dimensional counterpart, i.e.

$$\frac{\rho^*}{t^*}\partial_{\overline{t}}\overline{\rho} + \frac{\rho^* u^*}{x^*}\nabla_{\overline{\boldsymbol{x}}} \cdot (\overline{\rho \boldsymbol{u}}) = 0 \qquad \Leftrightarrow \qquad \partial_{\overline{t}}\overline{\rho} + \frac{u^* t^*}{x^*}\nabla_{\overline{\boldsymbol{x}}} \cdot (\overline{\rho \boldsymbol{u}}) = 0.$$

Naturally, the velocity is in direct relation to the space and time, therefore it is useful to choose $u^* = x^*/t^*$ and the equation simplifies to

$$\partial_{\overline{t}}\overline{\rho} + \nabla_{\overline{\boldsymbol{x}}} \cdot (\overline{\rho \boldsymbol{u}}) = 0.$$

With a similar computation we obtain for the conservation of momentum equation (2.5)

$$\partial_{\overline{t}}(\overline{\rho \boldsymbol{u}}) + \nabla_{\overline{\boldsymbol{x}}} \cdot (\overline{\rho \boldsymbol{u}} \otimes \overline{\boldsymbol{u}}) + \frac{p^*}{\rho^*(u^*)^2}\nabla_{\overline{\boldsymbol{x}}}\overline{p} = 0. \tag{2.11}$$

The lemma is proven by noting that $\varepsilon^2 = \rho^*(u^*)^2/p^*$. □

This non-dimensionalization process introduced the reference parameter ε in front of the pressure gradient, which plays a crucial role for low Mach flows. Namely, it can be identified as the reference quantity for the Mach number to obtain the dimensionless counterpart, i.e.

$$\mathrm{Ma} = \frac{\|\boldsymbol{u}\|}{c} = \frac{u^*\|\overline{\boldsymbol{u}}\|}{c^*\overline{c}} = \varepsilon\overline{\mathrm{Ma}},$$

where $\overline{c}$ denotes the non-dimensionalized speed of sound and c^* the corresponding reference quantity. To obtain this, we consider the speed of sound as given in Equation (2.8) and also compute the corresponding non-dimensional version, hereby following [78],

$$c = \sqrt{\gamma\frac{p}{\rho}} = \sqrt{\gamma\frac{p^*}{\rho^*}\frac{\overline{p}}{\overline{\rho}}} = \sqrt{\gamma\frac{p^*}{\rho^*}}\sqrt{\frac{\overline{p}}{\overline{\rho}}} =: c^*\overline{c},$$

with $c^* = \sqrt{\gamma p^*/\rho^*}$. Then, we can rewrite the parameter ε as

$$\varepsilon^2 = \frac{\rho^*(u^*)^2}{p^*} = \left(\gamma\frac{u^*}{c^*}\right)^2 \qquad \Rightarrow \qquad \varepsilon = \gamma\frac{u^*}{c^*},$$

which corresponds to the reference Mach number. The special role of ε gets more clear if we compute the eigenvalues of the Jacobian of the flux function $\overline{\boldsymbol{F}}$ in an arbitrary normal direction $\boldsymbol{n}$, similarly as done for the dimensional equations in Equation (2.9), $\nabla_{\overline{\rho},\overline{\rho \boldsymbol{u}}}\overline{\boldsymbol{F}} \cdot \boldsymbol{n}$:

$$\lambda_1 = \overline{\boldsymbol{u}} \cdot \boldsymbol{n}, \qquad \lambda_{2,3} = \overline{\boldsymbol{u}} \cdot \boldsymbol{n} \pm \frac{\overline{c}}{\varepsilon}. \tag{2.12}$$

We know that for a low Mach flow, the velocity $\boldsymbol{u}$ is small compared to the speed of sound c. Consequently, if we assume that, due to the non-dimensionalization, $\overline{\boldsymbol{u}}$ and $\overline{c}$ are of the same order of magnitude then we can conclude that a low Mach flow is present if

$$\varepsilon \ll 1.$$

This is what we assume in the following. We conclude this subsection with a short remark on the notation used in the remainder of this thesis.

Remark 2.8. *For the rest of this thesis we use the non-dimensionalized isentropic Euler equations, as given in Lemma 2.7. For the sake of readability we drop the $\overline{(\cdot)}$-notation in the following. Thus, from this point onward t, $\boldsymbol{x}$, ρ, $\boldsymbol{u}$, and p denote dimensionless variables and the term* isentropic Euler equations *identifies the non-dimensionalized equations as given in Lemma 2.7.*

2.1.2. Incompressible Euler equations

A low Mach flow, i.e. a flow with $\varepsilon \ll 1$, can be described as weakly compressible or nearly incompressible. Therefore, the incompressible Euler equations can be seen as an approximation of a nearly incompressible flow if one neglects weakly compressible effects. See [8, 115] for a detailed introduction of incompressible flows.

In the following we consider a given domain Ω_T, see Equation (2.3), and incompressible flow thereon. We assume that both boundary and initial data are such that the incompressible density ρ is constant in space and time. Again, the flow fulfills conservation of mass and momentum if we assume that all quantities are smooth. Due to the constant density ρ, the conservation of mass equation reduces to

$$\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} = 0 \qquad \text{in} \qquad \Omega_T. \tag{2.13}$$

Note that then the boundary conditions of $\boldsymbol{u}$ must satisfy

$$\int_{\partial\Omega} \boldsymbol{u} \cdot \boldsymbol{n} = 0.$$

From this we can rewrite the conservation of momentum equation. In this setting the pressure p^I is given as the so-called mechanic pressure, see [115], and can be seen as an additional variable. This results in

$$\begin{aligned} &\partial_t (\rho \boldsymbol{u}) + \nabla_{\boldsymbol{x}} \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u}) + \nabla_{\boldsymbol{x}} p^I = 0 \qquad \text{in} \qquad \Omega_T \\ \Leftrightarrow\ &\partial_t \boldsymbol{u} + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{u} \otimes \boldsymbol{u}) + \nabla_{\boldsymbol{x}} \frac{p^I}{\rho} = 0 \qquad \text{in} \qquad \Omega_T, \end{aligned}$$

and together with the divergence free constraint for the velocity given in Equation (2.13) in

$$\begin{aligned} \Leftrightarrow\ &\partial_t \boldsymbol{u} + (\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u})\, \boldsymbol{u} + \boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u} + \nabla_{\boldsymbol{x}} \frac{p^I}{\rho} = 0 \qquad \text{in} \qquad \Omega_T \\ \Leftrightarrow\ &\partial_t \boldsymbol{u} + \boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u} + \nabla_{\boldsymbol{x}} \frac{p^I}{\rho} = 0 \qquad \text{in} \qquad \Omega_T. \end{aligned} \tag{2.14}$$

The incompressible Euler equations are summarized in the following definition.

Definition 2.9 (Incompressible Euler equations)**.** *Let $\Omega_T \subset \mathbb{R}^{\geq 0} \times \mathbb{R}^2$ be a given domain, then the* incompressible Euler equations *are given by*

$$\partial_t \begin{pmatrix} 0 \\ \boldsymbol{u} \end{pmatrix} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{F}^I = 0 \qquad \textit{in} \qquad \Omega_T, \qquad \textit{where} \qquad \boldsymbol{F}^I := \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{u} \otimes \boldsymbol{u} + \frac{p^I}{\rho}\, \mathrm{Id} \end{pmatrix}$$

and $\rho > 0$ is a given constant. The equations are equipped with initial conditions

$$\boldsymbol{u}(t = 0, \boldsymbol{x}) = \boldsymbol{u}^0(\boldsymbol{x}) \qquad \textit{for} \qquad \boldsymbol{x} \in \Omega,$$

where $\boldsymbol{u}^0$ is a given function which fulfills

$$\nabla \cdot \boldsymbol{u}^0 = 0,$$

and suitable boundary conditions. Such boundary conditions must fulfill

$$\int_{\partial\Omega} \boldsymbol{u} \cdot \boldsymbol{n} = 0.$$

Remark 2.10. *Note that one can derive a Poisson[7] equation for the pressure by applying the divergence operator on the conservation of momentum equation, i.e.*

$$\begin{aligned} 0 =& \nabla_{\boldsymbol{x}} \cdot \left(\partial_t \boldsymbol{u} + \boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u} + \nabla_{\boldsymbol{x}} \frac{p^I}{\rho} \right) = \partial_t \nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u}) + \nabla_{\boldsymbol{x}} \cdot \left(\nabla_{\boldsymbol{x}} \frac{p^I}{\rho} \right) \\ =& \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u}) + \frac{1}{\rho} \Delta_{\boldsymbol{x}} p^I. \end{aligned}$$

Remark 2.11 (Non-dimensionalization)**.** *As for the isentropic Euler equations, see Lemma 2.7, we non-dimensionalize the incompressible Euler equations. For this, we perform the same steps as in Lemma 2.7 and choose*

$$p^* := \rho^* (u^*)^2.$$

Then, the same equations as given in Definition 2.9 are obtained, where t, $\boldsymbol{x}$, ρ, $\boldsymbol{u}$, and p denote dimensionless quantities. From this point onward the term incompressible Euler equations *identifies the non-dimensional equations.*

2.1.3. Ordinary differential equations

The isentropic Euler equations for a low Mach flow have a special structure due to the small parameter ε. This is similar to singularly perturbed differential equations which are differential equations with a small parameter - in the spirit of the previous section called ε - that cannot directly be set to zero in order to compute an approximation of the solution.

In this subsection, we introduce a class of ordinary differential equations as prototypical examples. Please note that we restrict ourselves to one specific type with two scalar variables. A more detailed introduction to singularly perturbation in the context of ordinary differential equations can be found in [61, 81, 136, 173].

Definition 2.12. *Let $t^{end} \in \mathbb{R}^{>0}$ be given. Then, the* ordinary differential equation (ODE) *we consider in this thesis is defined by*

$$\frac{d}{dt} \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f(y,z) \\ \frac{1}{\varepsilon} g(y,z) \end{pmatrix} \quad \text{in} \quad (0, t^{end}) \quad \text{and} \quad \begin{pmatrix} y(0) \\ z(0) \end{pmatrix} = \begin{pmatrix} y^0 \\ z^0 \end{pmatrix} \tag{2.15}$$

for given $\varepsilon \ll 1$, sufficiently smooth functions $f, g : \mathbb{R}^2 \to \mathbb{R}$ and initial values y^0 and z^0 and

$$y := [0, t^{end}] \to \mathbb{R},\ t \mapsto y(t) \quad \text{and} \quad z := [0, t^{end}] \to \mathbb{R},\ t \mapsto z(t).$$

Furthermore, we assume that $g(y,z)$ fulfills

$$\partial_z g(y,z) \le -1. \tag{2.16}$$

Equations as given in Definition 2.12 are common examples to test numerical methods if a small parameter is present, see exemplarily [24, 81].

Up to this point the functions f and g are arbitrary up to the conditions given in Definition 2.12. The first condition, smoothness, is canonical if one wants to obtain a smooth and unique solution. In addition g needs to fulfill (2.16), see also [24, 81]. The following remark gives more details on (2.16).

Remark 2.13. *We shortly comment on condition* (2.16)*:*

[7] Siméon Denis Poisson, 1781 – 1840

1) In literature, see [24, 81], one uses a more general property which is equivalent to (2.16) *if the function g is scalar.*

2) The upper bound -1 is arbitrary since we can rescale the equation. It is necessary that $\partial_z g(y,z)$ is bounded by a strictly negative constant. Then this condition guarantees the existence of a solution with some special properties. More details on this can be found in Section 2.2.

Next, we introduce two examples of ODEs as given in Definition 2.12.

Michaelis-Menten equation

The Michaelis[8]-Menten[9] equation, see e.g. [68, 137], describes the transformation of a substrate to a product driven by an enzyme (kinetics of an enzyme reaction mechanism). The parameter ε gives the ratio between the enzyme and substrate concentration.

Definition 2.14 (Michaelis-Menten equation)**.** *The* Michaelis-Menten equation *is an ordinary differential equation as given in Definition 2.12 with*

$$f(y,z) = -y + \left(y + \frac{1}{2}\right) z \qquad \textit{and} \qquad g(y,z) = y - (y+1)z.$$

Van der Pol equation

The van der Pol[10] equation arises from a second order differential equation which describes an oscillator with non-linear damping

$$\varepsilon \frac{d^2}{dt^2} y + (y^2 - 1) \frac{d}{dt} y + y = 0, \tag{2.17}$$

where ε describes the damping in the process, i.e. if ε is small there is less damping. The van der Pol equation became a common test example of numerical methods for stiff problems [62], see also [24, 27, 32, 81, 101] and the references therein. We can derive an equation as (2.15) from Equation (2.17) by introducing an additional variable $z = \frac{d}{dt} y$. This results in the van der Pol equation as given in the following definition.

Definition 2.15 (Van der Pol equation)**.** *The* van der Pol equation *is an ordinary differential equation as given in Definition* (2.12) *with*

$$f(y,z) = z \qquad \textit{and} \qquad g(y,z) = (1 - y^2)z - y.$$

2.2. Asymptotic limit

Most equations introduced before, see Lemma 2.7 and Definition 2.12, contain the small parameter ε and for these equations the behavior as $\varepsilon \to 0$ is of special interest. To compute this limit, we start with the ordinary differential equations and introduce the technique of asymptotic expansion. Asymptotic expansions are used to derive the $\varepsilon \to 0$ limit, and conditions to obtain an asymptotic solution are thereby obtained, i.e. a solution which converges towards the solution of the limiting equation as $\varepsilon \to 0$.

2.2.1. Asymptotic expansion for ODEs

We consider the ordinary differential equation as given in Definition 2.12 and take a look at the asymptotic behavior as $\varepsilon \to 0$. This has also been done in [81]. Ideally, a general solution $y(t;\varepsilon)$ and $z(t;\varepsilon)$ exists for all values of $\varepsilon \ll 1$, that converges towards a solution as $\varepsilon \to 0$. In the following, we try to find such a

[8] Leonor Michaelis, 1875 – 1949

[9] Maud Menten, 1879 – 1960

[10] Balthasar van der Pol, 1889 – 1959

solution with the help of an asymptotic expansion as given in the below definition. For a more detailed introduction of asymptotic expansions we refer to [61, 103, 136, 137].

Definition 2.16 (Asymptotic expansion)**.** *In this thesis, the asymptotic expansion of a function $y(t;\varepsilon)$ is defined by a sequence*

$$(y_{(i)}(t))_{i=0}^{\infty} \qquad \textit{such that} \qquad y(t;\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i y_{(i)}(t).$$

Remark 2.17. *The asymptotic sequence $(\varepsilon^i)_{i=0}^{\infty}$ is not a unique choice for an asymptotic expansion. There are several possibilities which depend on the considered problem. In this thesis we restrict ourselves to the presented sequence.*

For simplicity we ignore the choice of initial conditions in the beginning and assume that they are chosen sufficiently. We comment on the initial values later in this section, see Definition 2.20.

We assume that two sequences $(y_{(i)}(t))_{i=0}^{\infty}$ and $(z_{(i)}(t))_{i=0}^{\infty}$ exist, such that y and z are given by

$$y(t;\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i y_{(i)}(t) \qquad \text{and} \qquad z(t;\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i z_{(i)}(t). \tag{2.18}$$

As long as y and z are smooth, these asymptotic expansions can exemplarily be computed with the help of a Taylor[11] series in ε. Then, we can plug Equation (2.18) into Equation (2.15) and collect terms of the same order of ε. For the sake of readability, we drop the dependency on t. This results in

$$\begin{aligned}
0 =& \frac{d}{dt}\begin{pmatrix} \sum\limits_{i=0}^{\infty} \varepsilon^i y_{(i)} \\ \sum\limits_{i=0}^{\infty} \varepsilon^i z_{(i)} \end{pmatrix} - \begin{pmatrix} f\left(\sum\limits_{i=0}^{\infty} \varepsilon^i y_{(i)}, \sum\limits_{i=0}^{\infty} \varepsilon^i z_{(i)}\right) \\ \frac{1}{\varepsilon} g\left(\sum\limits_{i=0}^{\infty} \varepsilon^i y_{(i)}, \sum\limits_{i=0}^{\infty} \varepsilon^i z_{(i)}\right) \end{pmatrix} \\
=& \frac{d}{dt}\begin{pmatrix} \sum\limits_{i=0}^{\infty} \varepsilon^i y_{(i)} \\ \sum\limits_{i=0}^{\infty} \varepsilon^i z_{(i)} \end{pmatrix} - \begin{pmatrix} \sum\limits_{i=0}^{\infty} \varepsilon^i f_{(i)}\left(y_{(0)}, y_{(1)}, \dots, z_{(0)}, z_{(1)}, \dots\right) \\ \sum\limits_{i=0}^{\infty} \varepsilon^{i-1} g_{(i)}\left(y_{(0)}, y_{(1)}, \dots, z_{(0)}, z_{(1)}, \dots\right) \end{pmatrix} \\
=& \begin{pmatrix} \frac{d}{dt} y_{(0)} - f_{(0)}\left(y_{(0)}, y_{(1)}, \dots, z_{(0)}, z_{(1)}, \dots\right) \\ -\varepsilon^{-1} g_{(0)}\left(y_{(0)}, y_{(1)}, \dots, z_{(0)}, z_{(1)}, \dots\right) \end{pmatrix} \\
&+ \sum_{i=1}^{\infty} \begin{pmatrix} \varepsilon^i \left(\frac{d}{dt} y_{(i)} - f_{(i)}\left(y_{(0)}, y_{(1)}, \dots, z_{(0)}, z_{(1)}, \dots\right)\right) \\ \varepsilon^{i-1}\left(\frac{d}{dt} z_{(i-1)} - g_{(i)}\left(y_{(0)}, y_{(1)}, \dots, z_{(0)}, z_{(1)}, \dots\right)\right) \end{pmatrix}
\end{aligned}$$

where $f_{(i)}$ and $g_{(i)}$ for $i = 0, \dots$ are computed with a Taylor expansion, i.e. we get for f

$$f_{(0)}(y_{(0)}, z_{(0)}) = f(y_{(0)}, z_{(0)}) \tag{2.19}$$

$$f_{(1)}(y_{(0)}, y_{(1)}, z_{(0)}, z_{(1)}) = \partial_y f(y_{(0)}, z_{(0)}) y_{(1)} + \partial_z f(y_{(0)}, z_{(0)}) z_{(1)} \tag{2.20}$$

$$\begin{aligned}
f_{(2)}(y_{(0)}, y_{(1)}, y_{(2)}, z_{(0)}, z_{(1)}, z_{(2)}) =& \partial_y f(y_{(0)}, z_{(0)}) y_{(2)} + \partial_z f(y_{(0)}, z_{(0)}) z_{(2)} \\
&+ \frac{1}{2}\left(\partial_{yy} f(y_{(0)}, z_{(0)}) y_{(1)}^2 + \partial_{zz} f(y_{(0)}, z_{(0)}) z_{(1)}^2\right) \\
&+ \partial_{yz} f(y_{(0)}, z_{(0)}) y_{(1)} z_{(1)}
\end{aligned} \tag{2.21}$$

and a similar result for g. Since we assumed that the asymptotic expansion given in Equation (2.18) is valid for all values of $\varepsilon \ll 1$, we can vary in terms of ε and derive several equations which have to be fulfilled. In detail we obtain

$$\frac{d}{dt}\begin{pmatrix} y_{(0)} \\ 0 \end{pmatrix} = \begin{pmatrix} f_{(0)}\left(y_{(0)}, z_{(0)}\right) \\ g_{(0)}\left(y_{(0)}, z_{(0)}\right) \end{pmatrix} \tag{2.22}$$

[11] Brook Taylor, 1685 – 1731

and for $i = 1, \dots$

$$\frac{d}{dt}\begin{pmatrix} y_{(i)} \\ z_{(i-1)} \end{pmatrix} = \begin{pmatrix} f_{(i)}\left(y_{(0)}, \dots, y_{(i)}, z_{(0)}, \dots, z_{(i)}\right) \\ g_{(i)}\left(y_{(0)}, \dots, y_{(i)}, z_{(0)}, \dots, z_{(i)}\right) \end{pmatrix}. \tag{2.23}$$

By starting with Equation (2.22) and continuing with Equation (2.23) we can compute a solution which is valid for all values of ε.

The zeroth order equation

We now consider solely the zeroth order equation (2.22) which can be characterized as a differential algebraic equation [11, 30, 81]. Ideally, we can solve the algebraic equation in $z_{(0)}$ such that we derive a differential equation for $y_{(0)}$.

Lemma 2.18 (Algebraic equation)**.** *Let the differential algebraic equation* (2.22) *with corresponding inital values be given, where g is given as in Definition 2.12 and the initial values are given as an asymptotic expansion and fulfill*

$$g(y^0_{(0)}, z^0_{(0)}) = 0.$$

Then there exists an open set $I \subset \mathbb{R}$ with $y^0_{(0)} \in I$ and a function

$$D := I \to \mathbb{R}, \quad y_{(0)} \mapsto z_{(0)} := D(y_{(0)})$$

such that

$$0 = g(y_{(0)}, D(y_{(0)})).$$

Furthermore the derivative of D is given by

$$D'(y_{(0)}) = -\frac{\partial_y g(y_{(0)}, D(y_{(0)}))}{\partial_z g(y_{(0)}, D(y_{(0)}))}.$$

Proof. Since $\partial_z g$ is bounded by -1, see Definition 2.12, we can apply the implicit function theorem and directly obtain the local existence of the function D with the corresponding derivative. □

From the previous lemma we can conclude that there exists a time interval for which we can represent the solution $z_{(0)}(t)$ by $D(y_{(0)}(t))$. Consequently, the differential algebraic equation reduces to a differential equation for $y_{(0)}$ and from this $z_{(0)}$ can directly be computed,

$$\frac{d}{dt} y_{(0)} = f(y_{(0)}, D(y_{(0)})) \qquad \text{and} \qquad z_{(0)} = D(y_{(0)}). \tag{2.24}$$

Remark 2.19. *Both examples of ordinary differential equations defined before, see Definitions 2.14 and 2.15, fulfill the requirements of Lemma 2.18 if the initial conditions are chosen in a suitable way. The function D is given for Michaelis-Menten (MM) and van der Pol (VDP) by*

$$D_{MM}(y_{(0)}) = \frac{y_{(0)}}{y_{(0)} + 1} \qquad \textit{and} \qquad D_{VDP}(y_{(0)}) = \frac{y_{(0)}}{1 - y^2_{(0)}}.$$

Initial conditions

Up to this point, we have left the choice of initial conditions open and assumed that they are chosen in a suitable way. We assumed that the solution is given as an asymptotic expansion and consequently we do the same for the initial conditions, in detail

$$\begin{pmatrix} y^0 \\ z^0 \end{pmatrix} = \begin{pmatrix} y^0_{(0)} \\ z^0_{(0)} \end{pmatrix} + \varepsilon \begin{pmatrix} y^0_{(1)} \\ z^0_{(1)} \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y^0_{(2)} \\ z^0_{(2)} \end{pmatrix} + \mathcal{O}(\varepsilon^3)$$

We know from Lemma 2.18 that if $y_{(0)}$ is given we can compute $z_{(0)} = D(y_{(0)})$. This should be also valid for the initial conditions to obtain an asymptotic solution. Thus, for given $y^0_{(0)}$ we compute $z^0_{(0)}$ by

$$z^0_{(0)} = D(y^0_{(0)}).$$

Next, y^0 and z^0 must be chosen in such a way that the assumption (2.16) holds, i.e. that $\partial_z g(y,z)$ is bounded by -1. Since $y_{(0)}$ and $z_{(0)}$ are the dominating terms in the asymptotic expansion, a value for $y_{(0)}$ is chosen in such a way that $\partial_z g(y_{(0)}, D(y_{(0)}))$ is bounded by a negative constant smaller than -1. Then there exists an $\varepsilon_0 > 0$ and a time instance $t^{end} > 0$ such that

$$\partial_z g(y(t), z(t)) \leq -1$$

for all $\varepsilon < \varepsilon_0$ and for all $t \in (0, t^{end})$. This can be done since the solution and all functions are sufficiently smooth. The remaining initial conditions $y^0_{(i)}$, $z^0_{(i)}$ for $i = 1, \dots$ can be derived in a similar way. In detail, we choose $y^0_{(i)}$ and then compute $z^0_{(i)}$. For example, for $z^0_{(1)}$ we can obtain

$$\begin{aligned} &\frac{d}{dt} z_{(0)} = \partial_y g(y_{(0)}, D(y_{(0)}))y_{(1)} + \partial_z g(y_{(0)}, D(y_{(0)}))z_{(1)} \\ \Leftrightarrow\ & z_{(1)} = \frac{D'(y_{(0)})f(y_{(0)}, D(y_{(0)})) - \partial_y g(y_{(0)}, D(y_{(0)}))y_{(1)}}{\partial_z g(y_{(0)}, D(y_{(0)}))}, \end{aligned} \tag{2.25}$$

where we used that

$$\frac{d}{dt} z_{(0)} = \frac{d}{dt} D(y_{(0)}) = D'(y_{(0)}) \frac{d}{dt} y_{(0)} = D'(y_{(0)}) f(y_{(0)}, D(y_{(0)})).$$

This procedure can be continued for every $i = 2, \dots$. We can conclude that we have freedom in choosing the initial conditions y^0, but from this we need to compute the initial values z^0 to obtain a solution with the desired asymptotic behavior. Initial conditions which are chosen in such a way are called *well-prepared.*

Definition 2.20 (Well-prepared initial conditions for ODEs)**.** *We call initial conditions for the ordinary differential equation* (2.15) well-prepared *if*

1. *they are given as an asymptotic expansion, i.e.*

$$\begin{pmatrix} y^0 \\ z^0 \end{pmatrix} = \begin{pmatrix} y^0_{(0)} \\ z^0_{(0)} \end{pmatrix} + \varepsilon \begin{pmatrix} y^0_{(1)} \\ z^0_{(1)} \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y^0_{(2)} \\ z^0_{(2)} \end{pmatrix} + \mathcal{O}(\varepsilon^3),$$

2. *they fulfill $\partial_z g(y^0, z^0)) \leq c < -1$, with a constant c*

3. *and the elements of the asymptotic sequence $(y_{(i)}, z_{(i)})$ are valid initial conditions for* (2.22) *and* (2.23)*.*

Initial layer

The question which arises from Definition 2.20 is, what happens if initial conditions are not well-prepared. To see the consequences, we compute the solution of van der Pol's equation with initial conditions

$$y^0 = 2 \qquad \text{and} \qquad z^0 = 1.$$

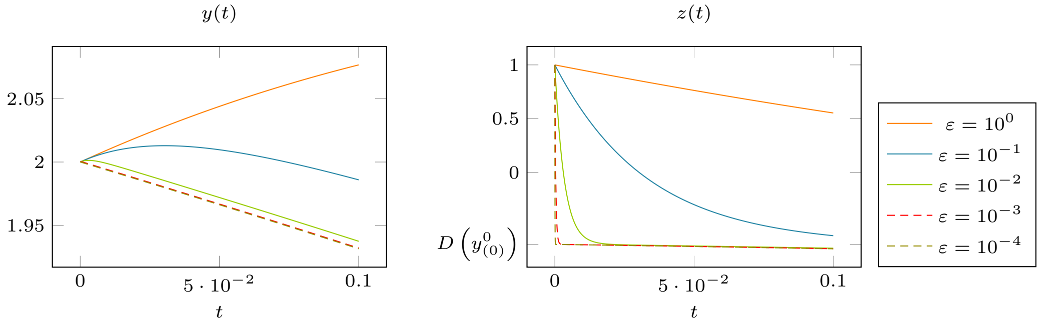


Figure 2.5.: Solution $y(t)$ (left) and $z(t)$ (right) of van der Pol's equation with non well-prepared initial conditions $y^0 = 2$ and $z^0 = 1$ for different values of ε.

We can easily check that these initial conditions are not well-prepared since

$$D_{VDP}(y^0_{(0)}) = D_{VDP}(2) = -\frac{2}{3} \neq 1 =: z^0_{(0)},$$

see Remark 2.19 for the definition of $D_{VDP}(y)$. The solutions for different values of ε are presented in Figure 2.5. First of all, $y(t)$ shows a convergent behavior as $\varepsilon \to 0$. On the other hand, $z(t)$ exhibits a steep layer connecting the initial value of z to the value $D_{VDP}(2)$, with a gradient in $\mathcal{O}(\varepsilon^{-1})$. What we observe is an initial or boundary layer [10, 81, 136, 173]. Such a layer can be computed with the help of a matched asymptotic expansion by re-scaling the time t with $\frac{1}{\varepsilon}$, see [10, 81, 136, 173] for more details, and performing an asymptotic analysis.

Remark 2.21. *Initial layers set special requirements on the numerical method which are not discussed in this thesis, therefore we always assume that well-prepared initial values are given.*

2.2.2. Isentropic Euler equations

We can now apply the theory of asymptotic expansions, see Definition 2.16, to compute the formal $\varepsilon \to 0$ limit of the isentropic Euler equations as given in Lemma 2.7. This is done by following [78]. In the end, we want to see that the compressible isentropic Euler equations converge towards the incompressible Euler equations, see Definition 2.9, as $\varepsilon \to 0$. This convergence is rigorous proven in [104]. Similar results are also obtained in [5, 52, 125, 159, 184], see also the references therein, including the full Euler and Navier-Stokes equations.

Similarly to the ordinary differential equations, we need well-prepared initial conditions. Therefore, well-prepared initial conditions must be valid initial conditions for the incompressible Euler equations since we know from Definition 2.20 that well-prepared initial conditions must coincide with the $\varepsilon \to 0$ limit.

Definition 2.22 (Well-prepared initial conditions)**.** *We call initial conditions for the isentropic Euler equations* well-prepared*, see also [50, 78], if they are given as an asymptotic expansion and fulfill*

$$\rho^0 = \underbrace{const}_{>0} + \mathcal{O}(\varepsilon^2) \qquad \text{and} \qquad \nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}^0 = \mathcal{O}(\varepsilon).$$

Furthermore, the initial conditions are given in such a way that the resulting solutions of the isentropic Euler equations are sufficiently smooth up to the final time instance t^{end}.

Since we consider a partial differential equation we also need well-prepared boundary conditions.

Definition 2.23 (Well-prepared boundary conditions)**.** *We call boundary conditions for the isentropic Euler equations* well-prepared *if they fulfill*

$$\int_{\partial\Omega} \boldsymbol{u} \cdot \boldsymbol{n} \mathrm{d}\sigma = \mathcal{O}(\varepsilon^2) \qquad \text{and} \qquad \rho(\boldsymbol{x}) = const + \mathcal{O}(\varepsilon^2) \quad \text{for} \quad \boldsymbol{x} \in \partial\Omega,$$

where the constant part of ρ is equal to the constant part of ρ^0 in Definition 2.22. We refer to [78] for a more detailed discussion of boundary conditions.

With this we can observe that the isentropic Euler equations are consistent with the incompressible Euler equations as $\varepsilon \to 0$. Please note that we do not show this in all details and mainly motivate the result, for a rigorous proof see [104].

Corollary 2.24. *Let the isentropic Euler equations, as given in Lemma 2.7, with an arbitrary $\varepsilon \ll 1$ be equipped with well-prepared initial and boundary conditions, see Definitions 2.22 and 2.23, and let the solutions ρ and $\boldsymbol{u}$ be given as an asymptotic expansion and sufficiently smooth, then we can observe that the $\varepsilon \to 0$ limit of the isentropic Euler equations is consistent with the incompressible Euler equations, see Definition 2.9.*

Proof. This corollary follows directly from Lemmas 2.25, 2.26, 2.27 and 2.28. From Lemma 2.26 we can conclude that the density $\rho_{(0)}$ is constant in space and time, thus the fluid described by $\rho_{(0)}$ is incompressible. Next, from Lemma 2.27 we can conclude that $\boldsymbol{u}_{(0)}$ is divergence free and finally Lemma 2.28 shows that $\boldsymbol{u}_{(0)}$ and $p_{(2)}$ fulfill equations which are consistent with the incompressible Euler equations given in Definition 2.9. □

Lemma 2.25. *Let the requirements of Corollary 2.24 be given, then the components of the asymptotic expansion fulfill*

$$\partial_t \rho_{(0)} + \nabla_{\boldsymbol{x}} \cdot \rho_{(0)} \boldsymbol{u}_{(0)} = 0, \tag{2.26}$$

$$\partial_t \rho_{(1)} + \nabla_{\boldsymbol{x}} \cdot \left(\rho_{(1)} \boldsymbol{u}_{(0)} + \rho_{(0)} \boldsymbol{u}_{(1)}\right) = 0, \tag{2.27}$$

$$\partial_t \rho_{(0)} \boldsymbol{u}_{(0)} + \nabla_{\boldsymbol{x}} \cdot \left(\rho_{(0)} \boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{(0)} + p_{(2)} \,\mathrm{Id}\right) = 0, \tag{2.28}$$

$$\nabla_{\boldsymbol{x}} \cdot \left(p_{(0)} \,\mathrm{Id}\right) = 0 \tag{2.29}$$

$$\textit{and} \qquad \nabla_{\boldsymbol{x}} \cdot \left(p_{(1)} \,\mathrm{Id}\right) = 0 \tag{2.30}$$

in the domain Ω_T.

Proof. Due to the requirements, the solution is given as an asymptotic expansion, see Definition 2.16, i.e.

$$\begin{aligned} \rho &= \rho_{(0)} + \varepsilon \rho_{(1)} + \varepsilon^2 \rho_{(2)} + \mathcal{O}(\varepsilon^3) \\ \boldsymbol{u} &= \boldsymbol{u}_{(0)} + \varepsilon \boldsymbol{u}_{(1)} + \mathcal{O}(\varepsilon^2) \\ p &= p_{(0)} + \varepsilon p_{(1)} + \varepsilon^2 p_{(2)} + \mathcal{O}(\varepsilon^3). \end{aligned} \tag{2.31}$$

Note that p is given by $p(\rho) = \kappa \rho^\gamma$ and therefore we can compute $p_{(i)}$ for $i = 0, \ldots$ by a Taylor expansion and get

$$\begin{aligned} &p_{(0)} = \kappa \rho_{(0)}^\gamma, \quad p_{(1)} = \kappa \gamma \rho_{(0)}^{\gamma-1} \rho_{(1)} \quad \text{and} \\ &p_{(2)} = \kappa \gamma \rho_{(0)}^{\gamma-1} \rho_{(2)} + \frac{\kappa}{2} \gamma (\gamma - 1) \rho_{(0)}^{\gamma-2} \rho_{(1)}^2. \end{aligned} \tag{2.32}$$

If we insert the asymptotic expansions of all quantities (2.31) in the isentropic Euler equations and rearrange the terms in order of ε, we obtain

$$\partial_t \rho_{(0)} + \nabla_{\boldsymbol{x}} \cdot \rho_{(0)} \boldsymbol{u}_{(0)} + \varepsilon \left(\partial_t \rho_{(1)} + \nabla_{\boldsymbol{x}} \cdot \left(\rho_{(1)} \boldsymbol{u}_{(0)} + \rho_{(0)} \boldsymbol{u}_{(1)}\right)\right) = \mathcal{O}(\varepsilon^2)$$

for the conservation of mass and

$$\partial_t \rho_{(0)} \boldsymbol{u}_{(0)} + \nabla_{\boldsymbol{x}} \cdot \left(\rho_{(0)} \boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{(0)} + p_{(2)} \,\mathrm{Id}\right) + \frac{1}{\varepsilon^2} \nabla_{\boldsymbol{x}} p_{(0)} + \frac{1}{\varepsilon} \nabla_{\boldsymbol{x}} p_{(1)} = \mathcal{O}(\varepsilon)$$

for the conservation of momentum equation. Note that we do not consider the higher order in ε terms since we are only interested in the $\varepsilon \to 0$ limit. If we vary in ε we obtain the Equations (2.26)-(2.30). □

Lemma 2.26. *Let the requirements of Corollary 2.24 be given, then*

$$\rho_{(0)} \equiv const \qquad and \qquad \rho_{(1)} \equiv const,$$

i.e. they are constant in space and time. Furthermore there holds

$$\rho_{(0)} \equiv \rho_{(0)}^0 \qquad and \qquad \rho_{(1)} \equiv 0.$$

Proof. Lemma 2.25 is applicable, therefore we can conclude from Equations (2.29) and (2.30) that $p_{(0)}$ and $p_{(1)}$ are constant in space, and together with Equation (2.32) that $\rho_{(0)}$ and $\rho_{(1)}$ are constant in space.

Next, we consider Equation (2.26) and use that $\partial_t \rho_{(0)} = \frac{d}{dt}\rho_{(0)}$ since $\rho_{(0)}$ is constant in space. We integrate over the whole spatial domain Ω and use the divergence theorem,

$$\begin{aligned} 0 = &\int_\Omega \left(\frac{d}{dt}\rho_{(0)} + \nabla_{\boldsymbol{x}} \cdot \rho_{(0)}\boldsymbol{u}_{(0)} \right) = \frac{d}{dt}\rho_{(0)} \int_\Omega 1\mathrm{dx} + \rho_{(0)} \int_\Omega \left(\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}_{(0)} \right) \mathrm{dx} \\ = &\frac{d}{dt}\rho_{(0)} |\Omega| + \rho_{(0)} \int_{\partial\Omega} \left(\boldsymbol{u}_{(0)} \cdot \boldsymbol{n} \right) \mathrm{d}\sigma. \end{aligned}$$

We assumed that well-prepared boundary conditions, see Definition 2.23, are given. Therefore, we can conclude that the boundary integral equals zero and thus

$$0 = \frac{d}{dt}\rho_{(0)} |\Omega|,$$

which means that $\rho_{(0)}$ is also constant in time. Similarly, we can conclude that $\rho_{(1)}$ is constant in time. Since both quantities are constant in space and time, they are equal to the initial conditions. This concludes the lemma. □

Lemma 2.27. *Let the requirements of Corollary 2.24 be given, then $\boldsymbol{u}_{(0)}$ fulfills*

$$\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}_{(0)} = 0.$$

Proof. This lemma follows directly from the results of Lemma 2.26. Since $\rho_{(0)}$ is constant in space and time, Equation (2.26) reduces to $\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}_{(0)} = 0$. □

Lemma 2.28. *Let the requirements of Corollary 2.24 be given, then $\boldsymbol{u}_{(0)}$ and $p_{(2)}$ fulfill*

$$\partial_t \boldsymbol{u}_{(0)} + \boldsymbol{u}_{(0)} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u}_{(0)} + \frac{1}{\rho_{(0)}} \nabla_{\boldsymbol{x}} p_{(2)} = 0.$$

Proof. We consider Equation (2.28) of Lemma 2.25 and divide by the constant value $\rho_{(0)}$, see Lemma 2.26. Finally, we can rewrite the term $\nabla_{\boldsymbol{x}} \cdot \left(\boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{(0)} \right)$ with the results of Lemma 2.27, and the desired equation is obtained. Note that the pressure $p_{(2)}$ is due to Equation (2.32) and Lemma 2.26 given by

$$p_{(2)} = \kappa\gamma\rho_{(0)}^{\gamma-1}\rho_{(2)}.$$

□

Remark 2.29. *As well-prepared boundary conditions for Corollary 2.24 we assumed that the boundary integral of $\boldsymbol{u}$ in normal direction is in $\mathcal{O}(\varepsilon^2)$. This is fulfilled by periodic and solid wall, see Equation* (2.7)*, boundary conditions.*

2.3. Prototype examples

To test numerical methods for the equations introduced and analyzed before, we need several numerical examples which consist of well-prepared initial and boundary conditions.

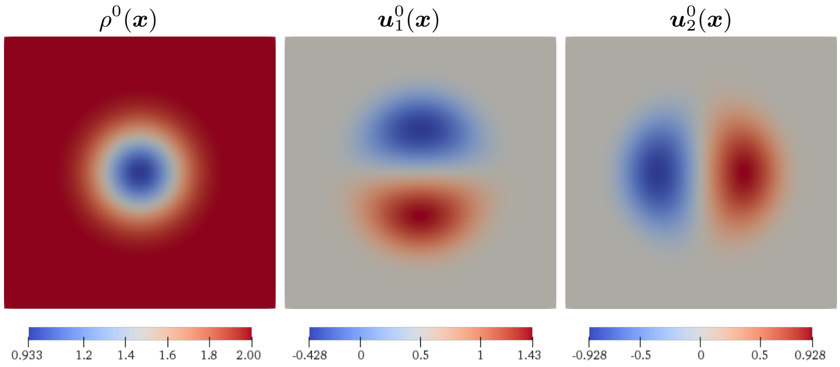


Figure 2.6.: Initial values of ρ, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of the smooth vortex example with $\varepsilon = 1$ as given in Example 2.30.

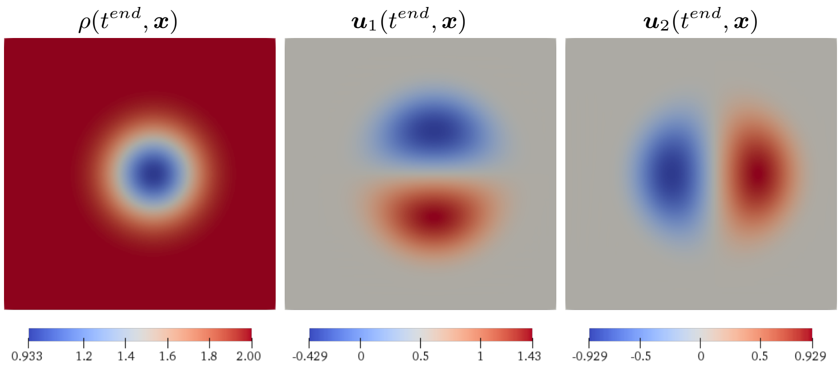


Figure 2.7.: ρ, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of the smooth vortex example with $\varepsilon = 1$ as given in Example 2.30 for $t = t^{end} = 0.1$. In comparison to the initial values given in Figure 2.6 the vortex moved to the right.

2.3.1. Isentropic Euler equations

We consider three different examples for the isentropic Euler equations. The first one is derived to show the convergence behavior of a numerical method and the remaining two are taken from literature to investigate the performance of a numerical method.

Smooth vortex

To compare numerical methods concerning their convergence behavior it is inevitable to consider an example which is sufficiently smooth and where one knows the solution. In general, one is not able to compute an exact solution for the isentropic Euler equations, but if one considers a special type of flow, an analytical solution can be derived.

The following example describes a radial symmetric vortex which is rotating and moving in one direction. A similar example for the low Mach isentropic Euler equations has been derived in [22] from an example in [149, 164]. This vortex is only one time continuously differentiable and therefore not useful to test a high order numerical method.

Example 2.30 (Smooth vortex)**.** *The* smooth vortex *example for the isentropic Euler equations as given in Lemma 2.7 with* $\overline{\Omega_T} := [0, t^{end}] \times [0,1]^2$ *and* $t^{end} = 0.1$ *is given by initial values*

$$\rho^0(\boldsymbol{x}) = 2 + 250,000\varepsilon^2 \begin{cases} \frac{1}{2} e^{2/\Delta r} \Delta\tau - \mathrm{Ei}\left(\frac{2}{\Delta\tau}\right) & \tau < \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases}$$

$$\boldsymbol{u}^0(\boldsymbol{x}) = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix} + 500 \begin{pmatrix} \frac{1}{2} - \boldsymbol{x}_2 \\ \boldsymbol{x}_1 - \frac{1}{2} \end{pmatrix} \cdot \begin{cases} e^{\frac{1}{\Delta\tau}} & \tau < \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases},$$

where

$$\mathrm{Ei}(x) := \int_{-\infty}^{x} \frac{e^s}{s} \mathrm{ds}, \qquad \tau := \left\| \boldsymbol{x} - \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\|_2 \qquad \textit{and} \qquad \Delta\tau := \tau^2 - \frac{1}{4},$$

$\kappa = \frac{1}{2}$, $\gamma = 2$ *and periodic boundary conditions.*

The initial values are plotted in Figure 2.6 for $\varepsilon = 1$ and the corresponding solution at time instance t^{end} is plotted in Figure 2.7. In these figures we can see that the vortex moves with a velocity of $\frac{1}{2}$ in $\boldsymbol{x}_1$-direction. In the following we show the derivation of this example.

First, we seek a stationary solution where the density is radially symmetric and the velocity describes a rotation, i.e. we seek a solution in spherical coordinates (r, φ) of the form

$$\rho(\boldsymbol{x}) = a(r(\boldsymbol{x})) \qquad \text{and} \qquad \boldsymbol{u}(\boldsymbol{x}) = b(r(\boldsymbol{x})) \begin{pmatrix} -\sin(\varphi(\boldsymbol{x}))r(\boldsymbol{x}) \\ \cos(\varphi(\boldsymbol{x}))r(\boldsymbol{x}) \end{pmatrix}, \tag{2.33}$$

where a and b are functions derived in the following and

$$r = \|\boldsymbol{x}\|_2, \qquad \cos(\varphi) = \frac{\boldsymbol{x}_1}{r} \qquad \text{and} \qquad \sin(\varphi) = \frac{\boldsymbol{x}_2}{r}.$$

Then, there holds

$$\begin{aligned} \nabla_{\boldsymbol{x}}\rho = a'(r) \begin{pmatrix} \cos(\varphi) \\ \sin(\varphi) \end{pmatrix}, \qquad \nabla_{\boldsymbol{x}}\boldsymbol{u}_1 = \begin{pmatrix} rb'(r)\cos(\varphi)(-\sin(\varphi)) \\ rb'(r)(-\sin(\varphi)^2) - b(r) \end{pmatrix} \\ \text{and} \qquad \nabla_{\boldsymbol{x}}\boldsymbol{u}_2 = \begin{pmatrix} rb'(r)\cos(\varphi)^2 + b(r) \\ rb'(r)\sin(\varphi)\cos(\varphi) \end{pmatrix}. \end{aligned} \tag{2.34}$$

Since we are interested in a stationary, i.e. a time independent, solution we consider the stationary isentropic Euler equations

$$\nabla_{\boldsymbol{x}} \cdot (\rho\boldsymbol{u}) = 0 \qquad \text{and} \qquad \nabla_{\boldsymbol{x}} \cdot \left(\rho\boldsymbol{u} \otimes \boldsymbol{u} + \frac{1}{\varepsilon^2}\kappa\rho^\gamma \,\mathrm{Id} \right) = 0.$$

For the conservation of mass equation we can directly conclude that

$$\begin{aligned} \nabla_{\boldsymbol{x}} \cdot \rho\boldsymbol{u} &= \nabla_{\boldsymbol{x}}\rho \cdot \boldsymbol{u} + \rho\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} \\ &= \left(ra'(r)b(r) + rb'(r)a(r)\right)\left(-\cos(\varphi)\sin(\varphi) + \sin(\varphi)\cos(\varphi)\right) = 0. \end{aligned}$$

Therefore, a solution of the form (2.33) directly fulfills the conservation of mass equation. Next, we take a look on the conservation of momentum equation. For simplicity we only consider the first one, since the second one can be handled similarly. The equation is given by

$$0 = \boldsymbol{u}_1^2\partial_{\boldsymbol{x}_1}\rho + 2\rho\boldsymbol{u}_1\partial_{\boldsymbol{x}_1}\boldsymbol{u}_1 + \boldsymbol{u}_1\boldsymbol{u}_2\partial_{\boldsymbol{x}_2}\rho + \rho\boldsymbol{u}_1\partial_{\boldsymbol{x}_2}\boldsymbol{u}_2 + \rho\boldsymbol{u}_2\partial_{\boldsymbol{x}_2}\boldsymbol{u}_1 + \frac{\kappa\gamma}{\varepsilon^2}\rho^{\gamma-1}\partial_x\rho$$

After inserting Equation (2.34) and basic calculations, the equation reduces to

$$0 = -a(r)\cos(\varphi)\left(b(r)^2 r + \frac{\kappa\gamma}{\varepsilon^2}a(r)^{\gamma-2}a'(r)\right).$$

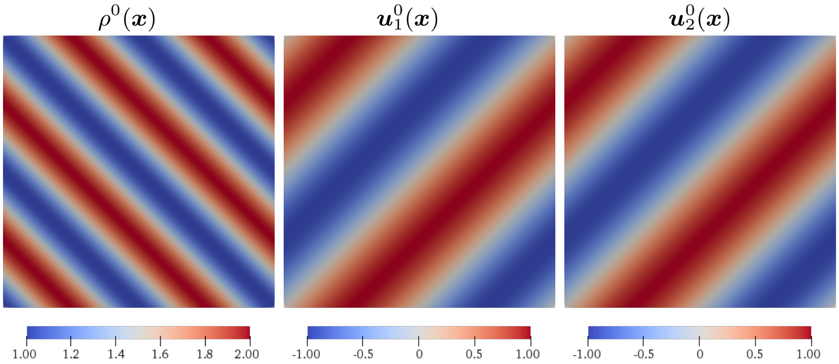


Figure 2.8.: Initial values of ρ, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of the periodic flow example with $\varepsilon = 1$ as given in Example 2.31.

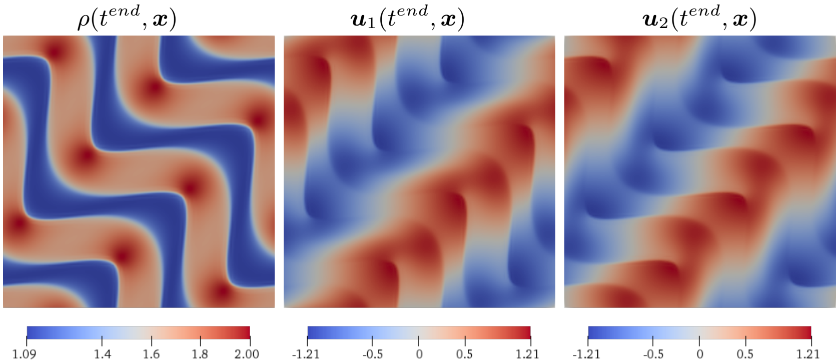


Figure 2.9.: ρ, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of the periodic flow example with $\varepsilon = 1$ as given in Example 2.31 for $t = t^{end} = 0.1$. The corresponding initial values are plotted in Figure 2.8.

This is fulfilled if $a(r) \equiv 0$ or

$$a'(r) = -b(r)^2 r a(r)^{2-\gamma} \frac{\varepsilon^2}{\kappa\gamma}.$$

Consequently, if we choose a function $b(r)$ and constants κ, γ we can compute the function $a(r)$ by solving an ordinary differential equation. Note that we need to choose a sufficiently, ideally infinitely, smooth function for $b(r)$. From this we obtain a stationary solution which is transformed to a non-stationary one by introducing a transport in one direction. This all together results in Example 2.30.

Periodic flow

The periodic flow example is used in [50, 78] and is an example of a flow with periodic boundary conditions where the flow is not only transported in a specific direction as given for the smooth vortex. The initial values for $\varepsilon = 1$ are given in Figure 2.8. After some time the flow shows a rich structure as we can see in Figure 2.9.

Example 2.31 (Periodic flow)**.** *The* periodic flow *example for the isentropic Euler equations as given in Lemma 2.7 with* $\overline{\Omega_T} := [0, t^{end}] \times [0,1]^2$ *and* $t^{end} = 0.1$ *is given by initial values*

$$\rho^0(\boldsymbol{x}) = 1 + \varepsilon^2 \sin(2\pi(\boldsymbol{x}_1 + \boldsymbol{x}_2))^2, \qquad \boldsymbol{u}^0(\boldsymbol{x}) = \begin{pmatrix} \sin(2\pi(\boldsymbol{x}_1 - \boldsymbol{x}_2)) \\ \sin(2\pi(\boldsymbol{x}_1 - \boldsymbol{x}_2)) \end{pmatrix},$$

$\kappa = 1$, $\gamma = 2$ *and periodic boundary conditions.*

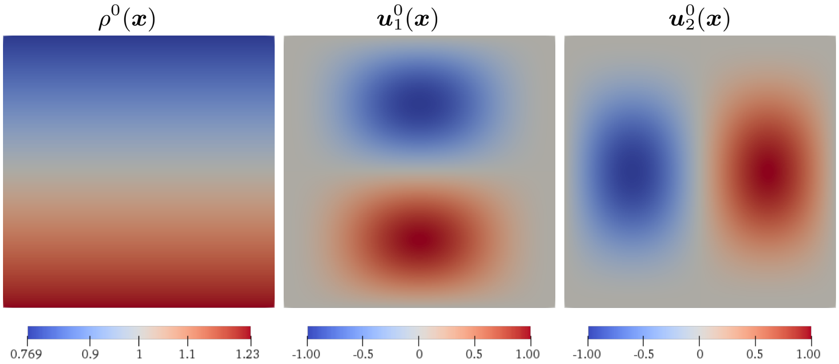


Figure 2.10.: Initial values of ρ, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of the vortex in a Box example with $\varepsilon = 1$ as given in Example 2.32.

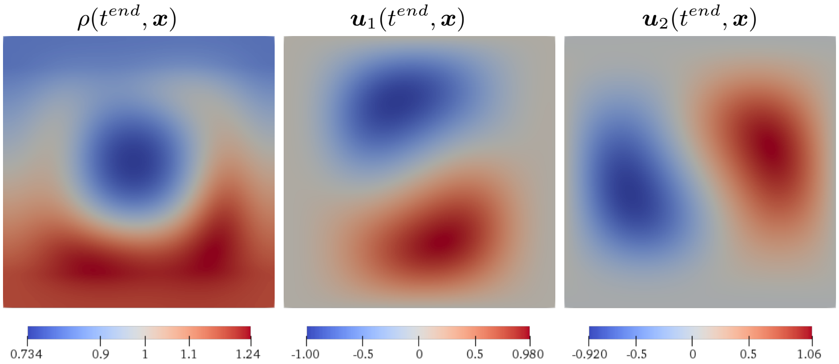


Figure 2.11.: ρ, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of the vortex in a box example with $\varepsilon = 1$ as given in Example 2.32 for $t = t^{end} = 0.1$.

Vortex in a box

The last example for the isentropic Euler equations is the vortex in a box example used in [45, 78]. This example uses a similar velocity field as Example 2.30, but due to the different initial data of ρ, see also Figure 2.10 for the initial values with $\varepsilon = 1$, and solid wall boundary conditions the flow field forces the creation of a vortex which can be seen in Figure 2.11 for the $\varepsilon = 1$ case.

Example 2.32 (Vortex in a box)**.** *The* vortex in a box *example for the isentropic Euler equations as given in Lemma 2.7 with* $\overline{\Omega_T} := [0, t^{end}] \times [0,1]^2$ *and* $t^{end} = 0.1$ *is given by initial values*

$$\rho^0(\boldsymbol{x}) = 1 - \frac{\varepsilon^2}{2} \tanh\left(\boldsymbol{x}_2 - \frac{1}{2}\right), \qquad \boldsymbol{u}^0(\boldsymbol{x}) = \begin{pmatrix} 2\sin(\pi \boldsymbol{x}_1)^2 \sin(\pi \boldsymbol{x}_2)\cos(\pi \boldsymbol{x}_2) \\ -2\sin(\pi \boldsymbol{x}_1)\cos(\pi \boldsymbol{x}_1)\sin(\pi \boldsymbol{x}_2)^2 \end{pmatrix},$$

$\kappa = 1$, $\gamma = 1.4$ *and solid wall boundary conditions.*

2.3.2. Ordinary differential equation

We have shown in Section 2.2.1 how to choose well-prepared initial conditions for equations like Michaelis-Menten, see Definition 2.14, and van der Pol, see Definition 2.15. This is done by choosing initial values y^0 and then computing the corresponding initial values z^0. In the following, we shortly introduce well-prepared initial conditions for both equations.

Please note that we theoretically need to check infinitely many conditions for well-prepared initial data, but we are interested in the case $\varepsilon \ll 1$ and therefore we can assume that terms in $\mathcal{O}(\varepsilon^3)$ are negligibly small.

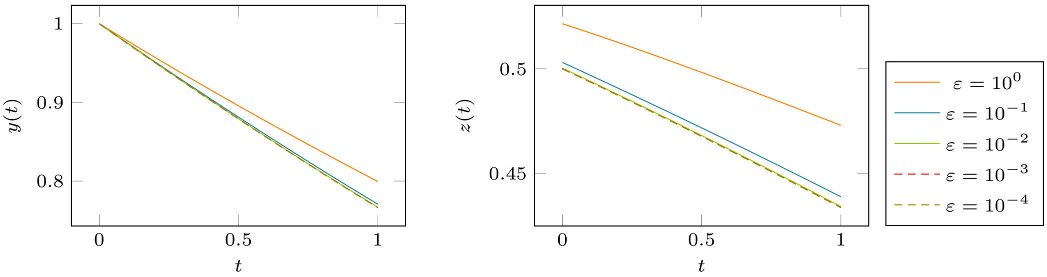


Figure 2.12.: Solution of the Michaelis-Menten equation equipped with well-prepared initial data as given in Example 2.33. The solutions y (left) and z (right) are plotted for different values of ε up to the final time instance $t^{end} = 1$.

Michaelis-Menten equation

We start by choosing $y^0 = 1$ and compute the corresponding initial values z^0.

Example 2.33. *The initial values of the Michaelis-Menten equation, see Definition 2.14, considered in the following are given by*

$$\begin{pmatrix} y^0 \\ z^0 \end{pmatrix} := \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix} + \varepsilon \begin{pmatrix} 0 \\ \frac{1}{32} \end{pmatrix} + \varepsilon^2 \begin{pmatrix} 0 \\ -\frac{5}{512} \end{pmatrix} \qquad \textit{and} \qquad t^{end} = 1.$$

The resulting solution computed with the initial values given in Example 2.33 is shown in Figure 2.12 for different values of ε. We can directly confirm that the solution converges towards a limiting state as $\varepsilon \to 0$ which means that the initial values are indeed well-prepared.

Van der Pol equation

As mentioned before, the van der Pol equation is a standard test equation for methods in this setting. Therefore, we can find well-prepared initial data in literature, see e.g. [24, 81], and it is useful to also consider them in this thesis.

Example 2.34. *The initial values and final time instance of the van der Pol equation, see Definition 2.15, considered in the following are given by*

$$\begin{pmatrix} y^0 \\ z^0 \end{pmatrix} = \begin{pmatrix} 2 \\ -\frac{2}{3} \end{pmatrix} + \varepsilon \begin{pmatrix} 0 \\ \frac{10}{81} \end{pmatrix} + \varepsilon^2 \begin{pmatrix} 0 \\ -\frac{292}{2187} \end{pmatrix} \qquad \textit{and} \qquad t^{end} = 0.55139.$$

Please note that the van der Pol equation describes an oscillating system which means that the solution can form singularities after a finite amount of time. Then it cannot be guaranteed that $\partial_z g$ is bounded by -1 which is needed for the asymptotic behavior of the solution. Furthermore, in [81] an adaptive method is used, i.e. the time step size Δt is adjusted if needed, for computing a numerical solution of the van der Pol equation and observed an extreme drop of the step size of Δt starting from $t = 0.55139$. This is why we choose the final time instance as done in [24, 81].

Again the solution up to the final instance $t^{end} = 0.55139$ is plotted in Figure 2.13 for different values of ε and we can again confirm that the initial values are well-prepared and that the corresponding solution converges towards a limiting state.

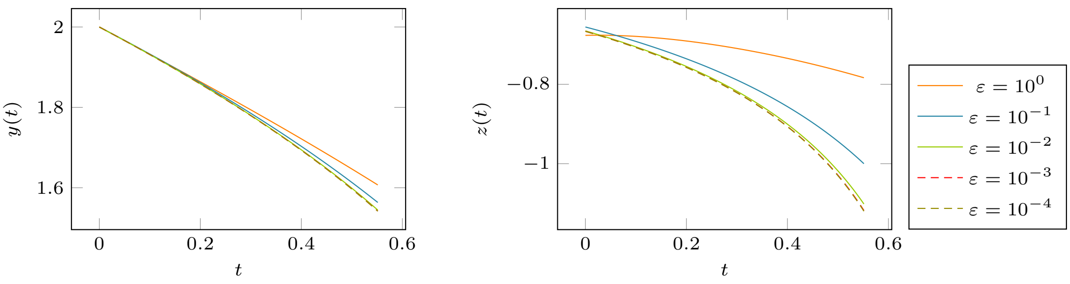


Figure 2.13.: Solution of the van der Pol equation equipped with well-prepared initial data as given in Example 2.34. The solutions y (left) and z (right) are plotted for different values of ε up to the final time instance $t^{end} = 0.55139$.

3. Numerical methods and asymptotic properties

This chapter is devoted to the numerical discretization of singularly perturbed differential equations with a focus on weakly compressible flows. To achieve such a method, we start with a literature overview of numerical methods for weakly compressible flows in Section 3.1 which are not based on a temporal splitting. These methods are discussed in Section 3.3.2. A more detailed overview can be found in [86, 95, 102, 109, 176].

Afterwards, we derive the basic concepts of the numerical discretization in Section 3.2. For this, we formulate and review several numerical properties which are especially relevant for the asymptotic behavior of the numerical method. These properties concern the

- *asymptotic consistency* [95, 96] (the numerical method is consistent with the limiting behavior of the equation),
- *asymptotic stability* [22] (the method is stable for large (with respect to ε) time steps even if $\varepsilon \ll 1$) and
- *asymptotic accuracy* [56, 58] (the numerical method computes a solution with the desired accuracy for large time steps).

Then, we continue Section 3.2 by introducing the method of choice which is an IMEX Runge-Kutta discretization [12, 101] combined with a discontinuous Galerkin spatial discretization [38, 40, 42, 43, 44]. This method is also given in [J1, J5]. The main ingredient of an IMEX Runge-Kutta method is a splitting of the equation into a stiff and a non-stiff contribution. Therefore, in Section 3.3 an introduction to splittings is given including

- an overview on the importance of the properties defined before,
- splittings for weakly compressible flows that can be found in literature
- and the novel splitting technique which we apply to the ordinary differential equation [J2, J4] and the isentropic Euler equations [J1, J3, C3, J5].

The numerical method, which is introduced in this chapter, has been first published in [J4] for ordinary differential equations, in [J1] for isentropic Euler equations and is also used in [J2, P1, J5].

3.1. Literature overview

The main difficulties a numerical method must resolve for low Mach number flows are the asymptotic behavior of the equations and the extreme stiffness of the equation, which affects the accuracy and efficiency of classical numerical methods.

There are several different approaches for numerical discretizations in this setting in literature. To give an overview, we start with Godunov[1] type schemes which can suffer from problems in the low Mach setting. Then, we consider different implicit time integration techniques, followed by methods which use preconditioning to obtain a stationary solution. Finally, we consider approaches which are derived from methods for incompressible equations. Note that methods which use a splitting of the flux function to handle each part of this splitting with a different time discretization are discussed in Section 3.3.2.

[1] Sergei Konstantinovich Godunov, *1929

Godunov type schemes

Godunov type schemes, which are based on solving a Riemann[2] problem, are a classical way to obtain a numerical approximation of a hyperbolic conservation law. Unfortunately, the corresponding Riemann fluxes can suffer from accuracy problems in the low Mach setting, see [51, 76, 129]. This behavior is caused by the creation of spurious waves, see [51] for a detailed investigation. Similar effects can also be observed for classical upwind schemes [77]. Furthermore, standard numerical flux functions, e.g. the Roe[3] flux function, adds too much artificial viscosity to the equation. In detail these inaccuracies can lead to $\mathcal{O}(\varepsilon)$ terms in the pressure while the pressure should, next to constant terms, only contain $\mathcal{O}(\varepsilon^2)$ terms. To overcome this issue, one can find several approaches in literature. In the following we discuss two different approaches. Note that this list is by no means exhaustive.

The first approach is to use a classical Riemann solver but add a preconditioning matrix $\boldsymbol{\Gamma}$ to change the artificial viscosity caused by the numerical method. This idea was introduced in [171]. Exemplarily, the Roe numerical flux function then reads

$$\boldsymbol{h}(\boldsymbol{w}^-, \boldsymbol{w}^+) = \frac{1}{2}\left(\widetilde{\boldsymbol{F}}(\boldsymbol{w}^-) + \widetilde{\boldsymbol{F}}(\boldsymbol{w}^+)\right) + \boldsymbol{\Gamma}\boldsymbol{A}\left(\boldsymbol{w}^- - \boldsymbol{w}^+\right) \cdot \boldsymbol{n},$$

where $\boldsymbol{A}$ is the classical Roe matrix. Similar ideas are used in [17, 20, 126, 139, 150, 155, 175], see also the references therein, for different equations and numerical flux functions.

The second approach is to split the flux function into two parts and then discretize these two parts with different numerical Riemann solvers. In [161], a flux vector splitting based on [2] is used to identify the upwind portion of the flux function and then solve the corresponding terms with an upwind method coupled with an explicit and a semi-implicit time integration method. Another flux splitting technique is given in [153], where one of the splitted flux functions only contains the pressure. Furthermore, [165] uses a flux splitting proposed by [168] to split the flux function into a convective and a pressure part at a cell boundary. Similarly, in [33] a splitting is used and parts of the equation are handled in Lagrange[4] coordinates. The spatial derivatives are then discretized with different numerical methods to obtain correct behavior in the $\varepsilon \ll 1$ regime.

Implicit methods

Explicit time integration methods need extremely small time steps to compute a stable solution if they are applied to weakly compressible flows. This is the case since the time step depends on the Mach number. To overcome this, [16] splits the isentropic Euler equations in convective and pressure terms and then handle the resulting parts with different explicit methods. This results in the ability to use a less restrictive CFL condition than standard explicit methods, but these time steps still depend on the Mach number.

Most fully implicit methods do not suffer from this restriction and can be stable for large time steps. Consequently, this is the canonical standard method in this setting and one can find different implicit time integration methods coupled with finite volume, see [17, 174], high order finite difference, see [53], or discontinuous Galerkin, see [105, 106], methods in literature.

The governing equations are in general non-linear and therefore solving the resulting non-linear system of equations can be very expensive in terms of computational cost. Ideally, the implicit part is linear, which is (often) more efficient to solve. To achieve a linear implicit flux function, one can try to write the flux function $\boldsymbol{F}$, or parts of the equation, in a form

$$\boldsymbol{F}(\boldsymbol{w}) = \boldsymbol{A}(\boldsymbol{w})\boldsymbol{w}, \tag{3.1}$$

where $\boldsymbol{A}$ is a matrix, e.g. the Jacobian of $\boldsymbol{F}$, identify $\boldsymbol{A}(\boldsymbol{w})$ with the explicit time instance and solve the equation where the flux is given by $\boldsymbol{A}(\boldsymbol{w}^n)\boldsymbol{w}^{n+1}$. This results in linear- or semi-implicit methods. In [63,

[2] Bernhard Riemann, 1826 – 1866
[3] Philip Lawrence Roe, *1939
[4] Joseph-Louis de Lagrange, 1736 – 1813

64] this semi-implicit temporal discretization coupled with a discontinuous Galerkin method in the low Mach setting is used. Similar approaches are proposed in [131] for Navier-Stokes equations and in [70] for the Euler-Korteweg equations, both in non-conservative form. In [66] the flux function is linearized around the previous time instance and efficient linear solver techniques are used for the full Navier-Stokes equations.

A completely different approach is to use a discontinuous Galerkin discretization in space *and* time. These methods are called space-time DG, see e.g. [67, 172], and are used for weakly compressible flows in [84, 166]. Note that a space-time discontinuous Galerkin method often needs to solve a large system of non-linear equations since one also uses a polynomial approximation in temporal direction which leads to a much larger discrete system compared to classical one-step temporal methods.

Finally, in [3, 29] relaxation schemes are proposed which derive with a proper transformation a linear system which approximates the original one. This system is larger but linear and therefore often more efficient to solve for an implicit time integration scheme.

Preconditioning methods

In different applications one seeks the stationary solution of an equation, i.e. a solution which fulfills

$$\partial_t \boldsymbol{w} = 0 \qquad \text{and} \qquad \nabla_{\boldsymbol{x}} \cdot \boldsymbol{F}(\boldsymbol{w}) = 0.$$

For this, a common way is to consider the time dependent equations and march in time until one is close to steady state. Therefore, the accuracy in time is insignificant and one is able to use a low order time integration method for a high order spatial discretization. Furthermore, one is even able to precondition the time derivative to reduce the stiffness of the equation, see e.g. [35, 170],

$$\boldsymbol{\Gamma} \partial_t \boldsymbol{w} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{F}(\boldsymbol{w}) = 0,$$

where $\boldsymbol{\Gamma}$ is a given preconditioning matrix. In most cases one needs to iterate to a large final time until a steady state with desired accuracy is reached. Therefore most methods use implicit or linear implicit time stepping. Examples for such schemes with a discontinuous Galerkin method coupled with different time stepping techniques, including explicit, implicit and linear implicit methods, are given in [18, 132, 133, 147]. Furthermore, implicit methods coupled with different spatial discretizations are presented in [35, 110, 119, 143, 174], see also the references therein.

The preconditioning technique can be extended to the time dependent case by introducing an additional time variable τ, then called dual time stepping [93], and using the preconditioning technique to compute a steady state solution with respect to τ, i.e. one considers

$$\boldsymbol{\Gamma} \partial_\tau \boldsymbol{w} + \partial_t \boldsymbol{w} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{F}(\boldsymbol{w}),$$

where $\boldsymbol{\Gamma}$ is a given preconditioning matrix and τ the additional time variable. For example, this has been done in [6, 34, 110, 133]. Please note, that one solves for every time instance a steady state problem, which could result in huge computational cost depending on how fast the resulting steady state solution is obtained. In principle a Newton iteration method to solve a system of non-linear equations can also be seen as dual time stepping.

Extensions of methods for incompressible equations

Weakly compressible flows can be seen as nearly incompressible. Thus, numerical methods which are successful for incompressible equations could be extended to the compressible counterpart.

The equation $\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} = 0$ of incompressible equations, as given in Definition 2.9, reduces to a constraint if one considers a fully explicit time integration scheme, and is not guaranteed to be fulfilled during the time iteration process. Furthermore, an explicit method does not compute an update for the pressure since the equation does not contain any time derivative of the pressure. To overcome this issue, pressure correction methods are introduced in literature, see e.g. [75, 144]. As an example we consider the so called SIMPLE[5] method [144], which performs the following steps, see also [128] and the references therein:

1) The velocity $\boldsymbol{u}$ and the pressure p are rewritten in an initial contribution, $\boldsymbol{u}_{in}$ and p_{in}, and an update, $\delta\boldsymbol{u}$ and δp, i.e.

$$\boldsymbol{u} = \boldsymbol{u}_{in} + \delta\boldsymbol{u} \qquad \text{and} \qquad p = p_{in} + \delta p,$$

where p_{in} is assumed to be given, e.g. computed from the values of the previous time instance.

2) $\boldsymbol{u}_{in}$ is computed with a discretization of the conservation of momentum equation.

3) The divergence equation and the conservation of momentum equation are rewritten in terms of $\delta\boldsymbol{u}$ and δp. Then, the conservation of momentum equation is solved in terms of $\delta\boldsymbol{u}$ and the result is inserted in the conservation of mass equation. This results in an elliptic equation for δp, which is then solved.

4) Finally, from the resulting pressure correction the velocity correction $\delta\boldsymbol{u}$ is computed by the previously derived equation for $\delta\boldsymbol{u}$.

These steps are repeated until the desired accuracy is obtained. This pressure correction procedure can be extended to compressible equations. As an example we consider the method proposed in [128] to show how this is done. In comparison to the incompressible case the divergence equation also depends on the density ρ, i.e. on an additional variable. [128] rewrites the Navier-Stokes equations in primitive variables and then use the equation of state for p to obtain a pressure correction. Thus, the method can be summarized by performing the same steps as before for the SIMPLE method but replacing the third step by

3') The *equation for the pressure* and the conservation of momentum equation are rewritten in terms of $\delta\boldsymbol{u}$ and δp. Then, the conservation of momentum equation is solved in terms of $\delta\boldsymbol{u}$ and the result is inserted in the *equation for the pressure*. This results in an elliptic equation in δp, which is then solved.

Note that this is not the only extension of pressure correction methods for weakly compressible flows, see [124] for an overview and for example [83, 108, 142, 169, 178], see also the references therein, for several different methods for different equations. Some of these methods use an additional technique to reduce the stiffness of the equation. Namely, they use a so-called multiple pressure variable, which means that the pressure variable is decomposed similarly to an asymptotic expansion and then the constant terms drop due to the divergence operator, see [128, 142]. A similar multipressure ansatz is also used in [120, 152].

Additionally, in [152] the equations are solved with a semi-implicit predictor corrector step, i.e. in a first step the global and large scale effects are solved and in a second step an incompressible solver is used which is extended with compressible effects as source term.

In [45] a low order method for the full non-conservative Euler equations is given by using a Helmholtz[6]-Hodge[7] decomposition of the velocity, i.e. the velocity is decomposed in a divergence free and a gradient part, and then solving an incompressible equation with a projection method and an equation for the remaining terms with an implicit Euler method.

The last work we mention, see [98], uses a transformation of the isentropic Euler equations to a kinetic equation, which is then solved with a low order finite difference scheme. It is shown that the resulting solution is an approximation of the incompressible Navier-Stokes equations with a Reynolds[8] number depending on Δt.

[5] Semi-implicit Method for Pressure Linked Equations
[6] Hermann Ludwig Ferdinand von Helmholtz, 1821 – 1894
[7] Sir William Vallance Douglas Hodge, 1903 – 1975
[8] Osborne Reynolds, 1842 – 1912

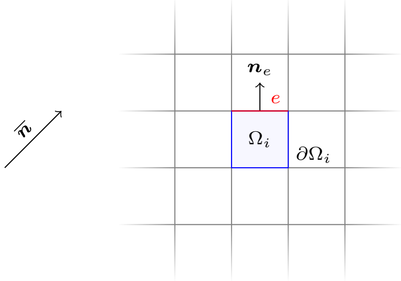


Figure 3.1.: Example of a rectangular cell Ω_i with its neighboring cells, its boundary $\partial\Omega_i$ and the normal vector $\boldsymbol{n}_e$ at one specific cell intersection e. Furthermore a uniform direction vector $\overline{\boldsymbol{n}}$, see Equation (3.12), is given.

3.2. Numerical discretization

In this section we present the basic discretization methods we consider in the rest of this thesis and we introduce the numerical properties a method for singularly perturbed differential equations should fulfill. Therefore, we start by discretizing the domain Ω_T. We assume that the spatial domain Ω, which fulfills $|\Omega| < \infty$, is bounded by a polygon and separated by a triangulation $\mathcal{T}$ into cells Ω_i with $i = 1, \dots, \mathrm{ne}$, where every cell is convex and bounded by a polygon. The triangulation is defined by

$$\mathcal{T} := \{\Omega_i \mid i = 1, \dots, \mathrm{ne}\}, \text{ with } \bigcup_{i=1}^{\mathrm{ne}} \overline{\Omega_i} = \Omega \text{ and } \Omega_i \cap \Omega_j = \emptyset \; \forall i \neq j. \tag{3.2}$$

The boundaries of every cell Ω_i play an important role for the numerical method. Therefore, we define the skeleton of the triangulation $\mathcal{T}$ by

$$\partial\mathcal{T} := \underbrace{\{e \mid e = \partial\Omega_i \cap \partial\Omega_j \quad \text{for} \quad i \neq j\}}_{=:\partial\mathcal{T}^I} \cup \underbrace{\{e \mid e = \partial\Omega_i \cap \partial\Omega\}}_{=:\partial\mathcal{T}^E},$$

i.e. the skeleton is given by all intersections of cells denoted with $\partial\mathcal{T}^I$ and all intersections of cells with the domain boundary denoted with $\partial\mathcal{T}^E$. All cells Ω_i are convex and bounded by polygons, thus normal vectors of an edge $e \in \partial\mathcal{T}$ fulfill

$$\boldsymbol{n}(\boldsymbol{x}) \equiv \boldsymbol{n}_e,$$

i.e. the normal vector of one edge $e \in \mathcal{T}$ is constant. See Figure 3.1 for a rectangular cell with neighboring cells and its normal vector $\boldsymbol{n}_e$ at one edge e.

We also assume that the temporal domain $(0, t^{end})$ is subdivided into cells $\left((t^{n-1}, t^n)\right)_{n=1}^{N}$ with

$$0 =: t^0 < t^1 < \dots < t^{N-1} < t^N := t^{end}. \tag{3.3}$$

For the ease of presentation we assume that the distances between two following time instances are the same, i.e. $t^n - t^{n-1} = \Delta t$ for all $n = 1, \dots N$. Furthermore, again for the ease of presentation, we assume that the spatial grid consists of uniform cells. Then, every edge $e \in \partial\mathcal{T}$ has the same length and we denote this length with Δx, i.e.

$$\Delta x = \max_{e \in \partial\mathcal{T}} \|e\| = \min_{e \in \partial\mathcal{T}} \|e\|.$$

Please note that for an ordinary differential equation as given in Definition 2.12 we only consider the temporal cells.

Assuming that the numerical method computed an approximate solution $\boldsymbol{w}_{\Delta x}^N$ of a smooth solution $\boldsymbol{w}$, then we can compute the L^2-error to measure the accuracy of the numerical approximation. The resulting error is given by

$$\|\boldsymbol{w}(t^{end}) - \boldsymbol{w}^N_{\Delta x}\|_{L^2(\Omega)} = \mathcal{O}(\Delta x^{p^x}) + \mathcal{O}(\Delta t^{p^t}), \tag{3.4}$$

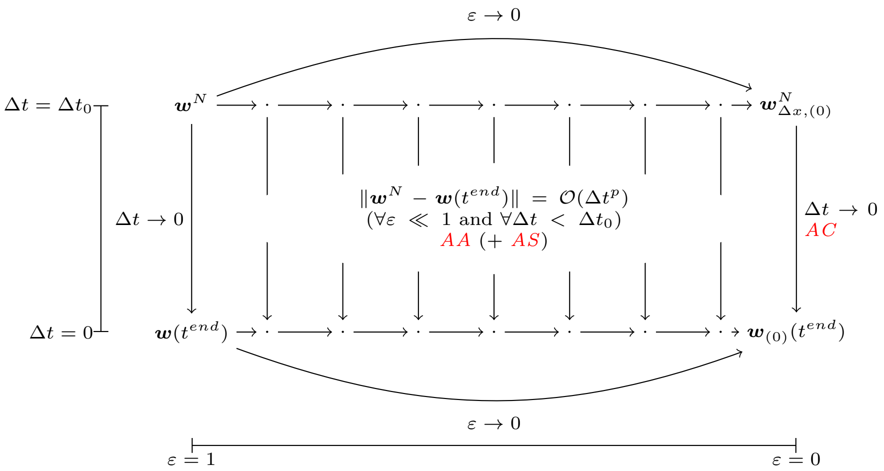
where p^x denotes the order of accuracy of the spatial and p^t the order of accuracy of the temporal discretization method. To achieve a certain degree of accuracy one has two different options. First, one uses a low order method on a very fine grid, e.g. $p^x = p^t$ small and $\Delta x, \Delta t \ll 1$, and second, one uses a high order method on a somehow coarser grid, e.g. p^x and p^t large with Δx, Δt relatively large. In the following we focus on high order temporal and spatial discretization.

3.2.1. Asymptotic properties

To derive a high order method for weakly compressible flows we define different numerical properties the method should fulfill. These different properties are also used in literature, see [22, 56, 58, 95, 96], to investigate numerical methods in the setting of weakly compressible flows, i.e. for $\varepsilon \ll 1$.

- Does the numerical method resolve the behavior of the equations as $\varepsilon \to 0$?
  → *asymptotic consistency (AC), see Definition 3.1.*
- Is the numerical method stable for large values of Δt and for all values $\varepsilon \ll 1$?
  → *asymptotic stability (AS), see Definition 3.3.*
- Does the numerical method compute a solution with the desired accuracy?
  → *asymptotic accuracy (AA), see Definition 3.4.*

These properties are also shown, for the case of a numerical method for ordinary differential equations as given in Definition 2.12, in Figure 3.2. In this figure the behavior of the solution and the convergence behavior of the numerical solution are shown for $\varepsilon \to 0$ and $\Delta t \to 0$. We comment on this figure in more detail later in this section.



$\varepsilon \to 0$
$\Delta t = \Delta t_0$
$\boldsymbol{w}^N$
$\boldsymbol{w}^N_{\Delta x,(0)}$
$\Delta t \to 0$
$\|\boldsymbol{w}^N - \boldsymbol{w}(t^{end})\| = \mathcal{O}(\Delta t^p)$
$(\forall \varepsilon \ll 1 \text{ and } \forall \Delta t < \Delta t_0)$
AA (+ AS)
$\Delta t \to 0$
AC
$\Delta t = 0$
$\boldsymbol{w}(t^{end})$
$\boldsymbol{w}_{(0)}(t^{end})$
$\varepsilon \to 0$
$\varepsilon = 1$
$\varepsilon = 0$

Figure 3.2.: Illustration of the asymptotic consistency, stability and accuracy properties the numerical solution $\boldsymbol{w}^N$ of a numerical method for singularly perturbed ordinary differential equations, see Definition 2.12, should fulfill. ε is fixed for the intermediate values $\cdot$.

In Section 2.2 we have seen that the equation converges towards a corresponding limiting equation as $\varepsilon \to 0$. It is desirable that the numerical method shows a similar behavior and by this resolves the behavior of the equation. This is the *asymptotic consistency* property, see Definition 3.1, which was introduced by Jin [95, 96], and became one of the fundamental properties a numerical method should fulfill for weakly compressible flows and also in general for singularly perturbed differential equations.

Definition 3.1. *We call a numerical method* asymptotically consistent *(AC) if the formal $\varepsilon \to 0$ limit of the numerical method is a consistent discretization of the corresponding limiting equation.*

Remark 3.2. *In literature, see e.g. [95, 96], the* asymptotic consistency *property is extended by stability of the limiting method. If both is fulfilled, viz the limiting method is consistent and stable, then the method is called* asymptotic preserving.

The asymptotic consistency property can also be seen in Figure 3.2 by the connection of the $\varepsilon \to 0$ limiting numerical approximation $\boldsymbol{w}_{(0)}^N$ with the $\varepsilon \to 0$ limiting solution $\boldsymbol{w}_{(0)}(t^{end})$.

For accuracy and efficiency reasons we want to be able to choose a temporal step size Δt which is of the same size as Δx to obtain a stable solution. If we consider an ordinary differential equation a relatively large value of Δt should be selectable. This is why we define the *asymptotic stability* property, which is fulfilled if we can make such a choice.

Definition 3.3. *We call a numerical method* asymptotically stable *(AS), see e.g. [22], if there exists a constant $\Delta t_0 > 0$, which is independent of ε but may depend on Δx, such that the numerical method is stable, i.e. the L^2-norm of the solution is bounded in time, for all $\Delta t < \Delta t_0$ and for all values of $\varepsilon \ll 1$.*

This property cannot directly be seen from Figure 3.2, since the asymptotic accuracy, which is introduced in Definition 3.4, is shown more prominently and can be seen as a special case of asymptotic stability.

Since we are interested in a high order numerical method, we would like to obtain a method which delivers the optimal order of convergence, ideally the order the numerical method would have for the non-stiff case, even for large values of Δt.

Definition 3.4. *We call a numerical method* asymptotically accurate *(AA), see e.g. [56, 58], if the numerical method converges with the same order of accuracy as if it is applied to a non-stiff equation, e.g. the $\varepsilon = 1$ case, starting from a point Δt_0, which is independent of ε but may depend on Δx, and for all values of $\varepsilon \ll 1$. In detail we obtain an error*

$$\|\boldsymbol{w}_{\Delta x}^N - \boldsymbol{w}(t^{end})\| = \mathcal{O}(\Delta t^p) + \mathcal{O}(\Delta x^p) \qquad \textit{for} \qquad \Delta t < \Delta t_0,$$

where p denotes the order of convergence the method would have for the non-stiff case.

This property can be seen in Figure 3.2 by the connections of $\boldsymbol{w}^N$ for different values of ε with the corresponding limiting solution $\boldsymbol{w}(t^{end})$ and the condition that a suitable error behavior can be obtained also for large values of Δt.

Remark 3.5. *Asymptotic accuracy implies asymptotic stability but not vice versa.*

With these properties defined we can derive the numerical method considered in this thesis and test the resulting scheme whether it is a suitable discretization of weakly compressible flows.

3.2.2. IMEX Runge-Kutta

To derive the temporal discretization, we consider an ordinary differential equation of the form

$$\frac{d}{dt}\boldsymbol{w}(t) = \boldsymbol{G}(\boldsymbol{w}(t), t) \quad \text{for} \quad t \in (0, t^{end}) \qquad \text{with} \qquad \boldsymbol{w}(0) = \boldsymbol{w}^0, \tag{3.5}$$

with a given function $\boldsymbol{G}(\boldsymbol{w}, t)$ and initial values $\boldsymbol{w}^0$. Note that we use this more general formulation of an ODE but always keep the examples given in Lemma 2.7 and Definition 2.12 in mind, i.e. the function $\boldsymbol{G}$ could contain the right hand side of an ordinary differential equation or of a partial differential equation.

We are mainly interested in the solution at final time t^{end}. This is why we consider a time iterative method which computes approximations at temporal cell boundaries t^n for $n = 1, \dots, N$. The simplest and most well-known iterative methods are explicit Euler

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n + \Delta t \boldsymbol{G}(\boldsymbol{w}^n, t^n) \tag{3.6}$$

and implicit Euler

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n + \Delta t \boldsymbol{G}(\boldsymbol{w}^{n+1}, t^{n+1}). \tag{3.7}$$

Explicit and implicit schemes suffer from several disadvantages in the setting of singularly perturbed differential equations and especially in the setting of low Mach number flows:

- **Explicit scheme:** For hyperbolic conservation laws it is well known that the temporal discretization size Δt is in direct relation to the spatial discretization size Δx. For a stable method it is necessary that waves only travel over one cell per time step, which means that

  $$\Delta t \sim \varepsilon \Delta x, \tag{3.8}$$

  i.e. for $\varepsilon \ll 1$ a huge number of time steps is needed. This relation is also known as the CFL condition [47], see Remark 3.6 for more details. Thus the resulting method is not *asymptotically stable.*

  On the other hand, performing one step with an explicit method is very cheap in terms of computational cost.

- **Implicit scheme:** Compared to an explicit method one can choose an arbitrary large value of Δt and still obtain a stable method, but because of (3.4) one would like to choose $\Delta t \approx \Delta x$ to achieve the desired accuracy. For this, one needs to solve non-linear (depending on the function $\boldsymbol{G}$) systems of equations, which leads to large computational cost.

  Furthermore, an implicit method adds additional diffusion to the solution, see [111], which affects the accuracy.

Remark 3.6. *The Courant*[9]*-Friedrichs*[10]*-Lewy*[11] *(CFL) condition [47] is needed to obtain a stable numerical method and describes the relation between Δx and Δt by*

$$\Delta t \leq CFL \frac{\Delta x}{\max_i |\lambda_i|}$$

for a proper CFL number. In detail, this relation enforces that waves only travel over one cell during a time step. In the setting of weakly compressible flows one can think about two different CFL numbers: CFL_{conv} which is needed for stability to resolve the slow convective waves:

$$\Delta t \leq CFL_{conv} \frac{\Delta x}{\|\boldsymbol{u}\|_{L^\infty}},$$

and CFL_{acoust} which is needed for stability to resolve the fast acoustic waves

$$\Delta t \leq CFL_{acoust} \frac{\Delta x}{\|\boldsymbol{u}\|_{L^\infty} + \frac{c}{\varepsilon}}.$$

Ideally, one would like to combine the advantages of implicit and explicit methods, i.e. the small computational cost and good accuracy of explicit methods and the good stability of implicit methods, to obtain an optimal scheme in the setting of low Mach number flows. One way to achieve this are IMEX[12] schemes, see e.g. [12, 13, 36, 88, 101, 190]. The basic idea of IMEX schemes is to split the right hand side $\boldsymbol{G}$ of (3.5)

$$\boldsymbol{G}(\boldsymbol{w}, t) = \widetilde{\boldsymbol{G}}(\boldsymbol{w}, t) + \widehat{\boldsymbol{G}}(\boldsymbol{w}, t) \tag{3.9}$$

and handle $\widetilde{\boldsymbol{G}}$ with an implicit and $\widehat{\boldsymbol{G}}$ with an explicit method. The simplest method is the IMEX Euler

[9] Richard Courant, 1888 – 1972

[10] Kurt Otto Friedrichs, 1901 – 1982

[11] Hans Lewy, 1904 – 1988

[12] IMplicit EXplicit

$$\begin{array}{c|c||c|c}
\widetilde{\boldsymbol{c}} & \widetilde{\boldsymbol{A}} & \widehat{\boldsymbol{c}} & \widehat{\boldsymbol{A}} \\
\hline
 & \widetilde{\boldsymbol{b}}^T & & \widehat{\boldsymbol{b}}^T
\end{array}$$

Table 3.1.: Butcher tableaux of an IMEX Runge-Kutta method. If the method has s stages then $\widetilde{\boldsymbol{c}}, \widehat{\boldsymbol{c}}, \widetilde{\boldsymbol{b}}, \widehat{\boldsymbol{b}} \in \mathbb{R}^s$ and $\widetilde{\boldsymbol{A}}, \widehat{\boldsymbol{A}} \in \mathbb{R}^{s \times s}$, where $\widetilde{\boldsymbol{A}}$ is a lower triangular matrix and $\widehat{\boldsymbol{A}}$ a lower triangular matrix with 0 entries on the diagonal.

scheme [48] which is obtained by combining (3.7) and (3.6)

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n + \Delta t \left(\widetilde{\boldsymbol{G}}(\boldsymbol{w}^{n+1}, t^{n+1}) + \widehat{\boldsymbol{G}}(\boldsymbol{w}^n, t^n) \right). \tag{3.10}$$

The choice of the splitting functions $\widetilde{\boldsymbol{G}}$ and $\widehat{\boldsymbol{G}}$ is essential in this setting since asymptotic consistency, stability and accuracy depend on it. We let this choice open for the moment, see Section 3.3.2 for splittings in literature and Section 3.3.3 for the splitting considered in this thesis.

Definition 3.7 (Splitting for an ordinary differential equation)**.** *$\widetilde{\boldsymbol{G}}$ and $\widehat{\boldsymbol{G}}$ form a* splitting *of $\boldsymbol{G}$ if $\widetilde{\boldsymbol{G}}$ and $\widehat{\boldsymbol{G}}$ are consistent with $\boldsymbol{G}$, i.e. $\boldsymbol{G} = \widetilde{\boldsymbol{G}} + \widehat{\boldsymbol{G}}$.*

Definition 3.8 (Splitting for a hyperbolic conservation law)**.** *$\widetilde{\boldsymbol{G}} := -\nabla \cdot \widetilde{\boldsymbol{F}}$ and $\widehat{\boldsymbol{G}} := -\nabla \cdot \widehat{\boldsymbol{F}}$ form a* splitting *of $\boldsymbol{G} := -\nabla \cdot \boldsymbol{F}$ if*

- *$\widetilde{\boldsymbol{G}}$ and $\widehat{\boldsymbol{G}}$ are consistent with $\boldsymbol{G}$, i.e. $\boldsymbol{G} = \widetilde{\boldsymbol{G}} + \widehat{\boldsymbol{G}}$, and*
- *both splitting flux functions $\widetilde{\boldsymbol{F}}$ and $\widehat{\boldsymbol{F}}$ induce a hyperbolic system, see Definition 2.5.*

The IMEX idea can be extended to high order time integration methods, like Runge[13]-Kutta[14] methods [12, 25, 26, 57, 91, 101, 140, 141], linear multistep methods [13, 88, 117], integral deffered correction methods [27, 36] and general linear multistep methods [190]. In the following we focus on IMEX Runge-Kutta methods, which are well studied in literature, see [12, 25, 26, 101, 141], and self-starting.

An IMEX Runge-Kutta method is given by two Runge-Kutta schemes, where both are described by their Butcher tableaux, see Table 3.1. We assume that the matrices $\widetilde{\boldsymbol{A}}$ and $\widehat{\boldsymbol{A}}$ are lower triangular matrices and $\widehat{\boldsymbol{A}}$ has zero diagonal entries. Note that one cannot simply combine two arbitrary high order Runge-Kutta schemes to obtain a high order IMEX Runge-Kutta scheme. The combined method needs to fulfill additional order conditions, see [101, 140].

Definition 3.9 (IMEX Runge-Kutta method)**.** *For an ordinary differential equation* (3.5)*, a given temporal grid* (3.3) *and a given s-stage IMEX Runge-Kutta method do the following for $n = 0, \dots, N-1$:*

1. *Solve for $i = 1, \dots, s$*

$$\begin{aligned}
\boldsymbol{w}^{n,i} = \boldsymbol{w}^n &+ \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \widetilde{\boldsymbol{G}}(\boldsymbol{w}^{n,j}, t^n + \widetilde{\boldsymbol{c}}_j \Delta t) \\
&+ \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \widehat{\boldsymbol{G}}(\boldsymbol{w}^{n,j}, t^n + \widehat{\boldsymbol{c}}_j \Delta t).
\end{aligned}$$

2. *Evaluate*

$$\begin{aligned}
\boldsymbol{w}^{n+1} = \boldsymbol{w}^n &+ \Delta t \sum_{j=1}^{s} \widetilde{\boldsymbol{b}}_j \widetilde{\boldsymbol{G}}(\boldsymbol{w}^{n,j}, t^n + \widetilde{\boldsymbol{c}}_j \Delta t) \\
&+ \Delta t \sum_{j=1}^{s} \widehat{\boldsymbol{b}}_j \widehat{\boldsymbol{G}}(\boldsymbol{w}^{n,j}, t^n + \widehat{\boldsymbol{c}}_j \Delta t).
\end{aligned}$$

[13] Carl Runge, 1856 – 1927

[14] Wilhelm Kutta, 1867 – 1944

Remark 3.10 (Classification of IMEX Runge-Kutta methods)**.** *IMEX Runge-Kutta methods can be classified depending on the structure of $\widetilde{\boldsymbol{A}}$, see [24], in the following classes:*

- ***Type A*** *[140]: If $\widetilde{\boldsymbol{A}}$ is invertible.*
- ***Type CK*** *[101]: If $\widetilde{\boldsymbol{A}}_{1,1} = 0$ and $\widetilde{\boldsymbol{A}}_{2\ldots s,2\ldots s}$ is invertible.*
- ***Type ARS*** *[12]: If of type CK and $\widetilde{\boldsymbol{A}}_{2\ldots s,1} = (0,\ldots,0)^T$.*

Note that this classification is more or less historical. In general one can show that all of these classes are special cases of type CK. For an IMEX Runge-Kutta method of type CK the matrices $\widetilde{\boldsymbol{A}}$ and $\widehat{\boldsymbol{A}}$ can be separated such that

$$\widetilde{\boldsymbol{A}} =: \left(\begin{array}{c|c} 0 & \boldsymbol{0}^T \\ \widetilde{\boldsymbol{\alpha}} & \widetilde{\boldsymbol{B}} \end{array}\right) \qquad \text{and} \qquad \widehat{\boldsymbol{A}} =: \left(\begin{array}{c|c} 0 & \boldsymbol{0}^T \\ \widehat{\boldsymbol{\alpha}} & \widehat{\boldsymbol{B}} \end{array}\right),$$

where $\widetilde{\boldsymbol{B}}$ is invertible.

Next, we introduce the globally stiffly accurate property [26], which is the IMEX equivalent to the stiffly accurate property for implicit Runge-Kutta schemes, see [81], and the first same as last property for explicit Runge-Kutta schemes, see [80]. This property is fulfilled if the last internal stage $\boldsymbol{w}^{n,s}$ is equal to the update step.

Definition 3.11 (Globally stiffly accurate)**.** *An IMEX Runge-Kutta method as given in Definition 3.9 is called* globally stiffly accurate*, see [26], if*

$$\widetilde{\boldsymbol{A}}_{s,1\ldots s} = \widetilde{\boldsymbol{b}}^T \qquad \text{and} \qquad \widehat{\boldsymbol{A}}_{s,1\ldots s} = \widehat{\boldsymbol{b}}^T,$$

i.e. if the update step is given by the last internal stage:

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^{n,s}.$$

To simplify the analysis of non-autonomous ODEs, we consider IMEX Runge-Kutta methods, where both parts are evaluated at the same time instances, i.e. the internal time instances vectors $\widetilde{\boldsymbol{c}}$ and $\widehat{\boldsymbol{c}}$ are the same.

Definition 3.12 (Uniform $\boldsymbol{c}$)**.** *An IMEX Runge-Kutta method as given in Definition 3.9 has a* uniform $\boldsymbol{c}$ *if the internal time instances of the implicit and explicit part are the same, i.e.*

$$\widetilde{\boldsymbol{c}} = \widehat{\boldsymbol{c}} =: \boldsymbol{c}.$$

In the following we use the abbreviation $t^{n,i} := t^n + c_i \Delta t$ for $i = 1,\ldots,s$.

Remark 3.13. *The Butcher tableaux of the IMEX Runge-Kutta methods we use in the rest of this thesis are given in the Appendix, see Tables A.1, A.2, A.3, A.4, A.5, A.6 and A.7.*

3.2.3. Discontinuous Galerkin

There are several possible choices for spatial discretization methods for hyperbolic conservation laws, for example finite difference (FDM), finite volume (FVM), finite element (FEM) or discontinuous Galerkin[15] (DG) methods. For the FDM, FVM and DG methods we refer to [163] and the references therein for an overview and for the FEM method we refer to [4] and the references therein. In Table 3.2, which is adapted from [82], some properties of these methods are shown. The discontinuous Galerkin method shows an enormous flexibility, i.e. the method is able to handle complex geometries, *hp*-adaptivity, is suitable for conservation laws and has also been formulated for elliptic problems [9, 151]. This is why we use and

[15] Boris Grigorjewitsch Galjorkin, 1871 – 1945

	Complex geometries	High order accuracy and *hp* adaptivity	Explicit semi-discrete form	Available for conservation laws	Available for elliptic problems
FDM	No	Yes	Yes	Yes	Yes
FVM	Yes	No	Yes	Yes	Yes
FEM	Yes	Yes	No	Yes	Yes
DG	Yes	Yes	Yes	Yes	Yes

Table 3.2.: Comparison of high order discretization methods adapted from [82].

introduce the DG method in the following. For a more detailed introduction also for different equations we refer to [19, 39, 41, 54, 59, 60, 82] and the references therein. An overview of the development of the discontinuous Galerkin method is given in [39]. Furthermore, a comparison between continuous Galerkin, i.e. FEM, and (hybridized) discontinuous Galerkin is given in [182], see also the references therein.

We consider a spatial triangulation as given in (3.2). On every cell Ω_i with $i = 1, \dots, \mathrm{ne}$ we introduce a polynomial space with polynomials of maximal degree $p \in \mathbb{N}^{>0}$ by

$$\mathbb{P}_i^p := \left\{ v \ : \ v|_{\Omega_i} = \sum_{k,l=0}^{p} v_{k,l} \boldsymbol{x}_1^k \boldsymbol{x}_2^l \quad \text{and} \quad v|_{\Omega \setminus \overline{\Omega_i}} = 0 \right\}.$$

The union of $\mathbb{P}_i^p$ defines a broken polynomial space $V_{\Delta x}$ on the complete domain Ω and is given by

$$V_{\Delta x} := \bigcup_{i=1}^{\mathrm{ne}} \mathbb{P}_i^p.$$

Note that we restrict ourselves to the case where the polynomial degree in every cell is the same, but one could also use a specific polynomial degree for each cell. Since the polynomial space $\mathbb{P}_i^p$ is only defined on one cell, a function $\varphi \in V_{\Delta x}$ could be discontinuous over cell boundaries, i.e. for a point $\boldsymbol{x} \in \partial\mathcal{T}^I$ on the grid skeleton. Therefore, we define the inner $(-)$ and outer $(+)$ value of φ on a given cell boundary $\partial\Omega_i$ by

$$\varphi^{\pm}(\boldsymbol{x}) := \lim_{0<\delta\to 0} \varphi\left(\boldsymbol{x} \pm \delta \boldsymbol{n}(\boldsymbol{x})\right), \tag{3.11}$$

where $\boldsymbol{n}(\boldsymbol{x})$ denotes the outward pointing normal vector at $\partial\Omega_i$. In some cases we might consider a point $\boldsymbol{x} \in e \in \partial\mathcal{T}$ without the mentioning of a specific cell. For this, we assume that a reference direction $\overline{\boldsymbol{n}}$ is given which fulfills

$$\overline{\boldsymbol{n}} \cdot \boldsymbol{n}(\boldsymbol{x}) \neq 0 \qquad \forall \boldsymbol{x} \in \partial\Omega_i \setminus \partial\Omega \quad \text{and} \quad i = 1, \dots, \mathrm{ne}. \tag{3.12}$$

This means that the reference direction is not orthogonal to any normal vector of the grid skeleton. Such a choice is possible since we only consider grids with a finite number of normal vectors on the skeleton. With this reference direction given, we can define an inner $(-)$ and outer $(+)$ value for $\boldsymbol{x} \in e \in \partial\mathcal{T}^I$ similarly to (3.11) by

$$\varphi^{\pm}(\boldsymbol{x}) := \lim_{0<\delta\to 0} \varphi\left(\boldsymbol{x} \pm \delta \boldsymbol{n}(\boldsymbol{x})\right), \tag{3.13}$$

where $\boldsymbol{n}(\boldsymbol{x})$ denotes the normal vector to e in $\boldsymbol{x}$ which fulfills $\boldsymbol{n}(\boldsymbol{x}) \cdot \overline{\boldsymbol{n}} > 0$. For the derivation of the DG method we follow the steps in [38, 40, 42, 43, 44]. A rigorous derivation based on the weak formulation of hyperbolic conservation laws can be found in [134].

We start with the hyperbolic conservation law in two spatial dimensions, e.g. as given in Lemma 2.7, with flux function

$$\boldsymbol{F} : \mathbb{R}^3 \mapsto \left(\mathbb{R}^3\right)^2, \qquad \boldsymbol{w} \to \left(\boldsymbol{F}_1(\boldsymbol{w}), \boldsymbol{F}_2(\boldsymbol{w})\right)$$

and smooth solution $\boldsymbol{w}$, which is split with a splitting as given in Definition 3.8. Then the conservation law is of the form

$$0 = \partial_t \boldsymbol{w} + \nabla_{\boldsymbol{x}} \cdot \widetilde{\boldsymbol{F}}(\boldsymbol{w}) + \nabla_{\boldsymbol{x}} \cdot \widehat{\boldsymbol{F}}(\boldsymbol{w}) \qquad \text{for} \qquad (t, \boldsymbol{x}) \in \Omega_T, \tag{3.14}$$

where $\boldsymbol{w} := \Omega_T \to \mathbb{R}^d$, $\widetilde{\boldsymbol{F}}(\boldsymbol{w}) = \left(\widetilde{\boldsymbol{F}}_1(\boldsymbol{w}), \widetilde{\boldsymbol{F}}_2(\boldsymbol{w})\right)$ and $\widehat{\boldsymbol{F}}(\boldsymbol{w}) = \left(\widehat{\boldsymbol{F}}_1(\boldsymbol{w}), \widehat{\boldsymbol{F}}_2(\boldsymbol{w})\right)$. In the following we seek an approximation $\boldsymbol{w}_{\Delta x}(t)$ of $\boldsymbol{w}$ which fulfills

$$\boldsymbol{w}_{\Delta x} \in C^1((0, t^{end}); V^3_{\Delta x}), \qquad \text{where} \qquad (V_{\Delta x})^3 = V_{\Delta x} \times V_{\Delta x} \times V_{\Delta x}.$$

To find such an approximation we multiply Equation (3.14) by an arbitrary piece-wise smooth function $\varphi \in V_{\Delta x}$ and integrate over the domain Ω. For simplicity we assume that periodic boundary conditions are given. This results in

$$0 = \int_\Omega \partial_t \boldsymbol{w} \varphi \mathrm{dx} + \int_\Omega \left(\nabla_{\boldsymbol{x}} \cdot \widetilde{\boldsymbol{F}}(\boldsymbol{w}) + \nabla_{\boldsymbol{x}} \cdot \widehat{\boldsymbol{F}}(\boldsymbol{w})\right) \varphi \mathrm{dx}.$$

Next, we use integration by parts to get rid of the derivatives in front of the flux functions $\widetilde{\boldsymbol{F}}$ and $\widehat{\boldsymbol{F}}$. We assumed that the function $\boldsymbol{w}$ is continuous over the whole domain, but since we want to replace the function $\boldsymbol{w}$ by its approximation $\boldsymbol{w}_{\Delta x} \in C^1((0, t^{end}); V^3_{\Delta x})$, which might be discontinuous over the cell boundaries, we keep the boundary integrals, i.e.

$$\begin{aligned} 0 = &\sum_{i=1}^{\mathrm{ne}} \int_{\Omega_i} \partial_t \boldsymbol{w} \varphi \mathrm{dx} - \sum_{i=1}^{\mathrm{ne}} \int_{\Omega_i} \left(\widetilde{\boldsymbol{F}}(\boldsymbol{w}) + \widehat{\boldsymbol{F}}(\boldsymbol{w})\right) \nabla_{\boldsymbol{x}} \varphi \mathrm{dx} \\ &+ \sum_{i=1}^{\mathrm{ne}} \int_{\partial\Omega_i} \left(\widetilde{\boldsymbol{F}}(\boldsymbol{w}^-) + \widehat{\boldsymbol{F}}(\boldsymbol{w}^-)\right) \varphi^- \boldsymbol{n} \mathrm{d}\sigma. \end{aligned}$$

Finally, we can replace the exact solution $\boldsymbol{w}$ by its approximation $\boldsymbol{w}_{\Delta x} \in C^1((0, t^{end}); V^3_{\Delta x})$. Due to this, the boundary integrals do not sum up to zero and therefore we introduce numerical flux functions $\widetilde{\boldsymbol{h}}$ and $\widehat{\boldsymbol{h}}$ to stabilize the boundary integral between two neighboring cells. The numerical flux functions we consider in this thesis are given in Definition 3.15. The resulting method is summarized in the following definition.

Definition 3.14 (Semi-discrete discontinuous Galerkin method)**.** *The* semi-discrete discontinuous Galerkin *formulation of Equation* (3.14) *is given by*

$$\begin{aligned} 0 = &\sum_{i=1}^{\mathrm{ne}} \int_{\Omega_i} \partial_t \boldsymbol{w}_{\Delta x} \varphi \mathrm{dx} - \sum_{i=1}^{\mathrm{ne}} \int_{\Omega_i} \left(\widetilde{\boldsymbol{F}}(\boldsymbol{w}_{\Delta x}) + \widehat{\boldsymbol{F}}(\boldsymbol{w}_{\Delta x})\right) \nabla_{\boldsymbol{x}} \varphi \mathrm{dx} \\ &+ \sum_{i=1}^{\mathrm{ne}} \int_{\partial\Omega_i} \left(\widetilde{\boldsymbol{h}}(\boldsymbol{w}^-_{\Delta x}, \boldsymbol{w}^+_{\Delta x}) + \widehat{\boldsymbol{h}}(\boldsymbol{w}^-_{\Delta x}, \boldsymbol{w}^+_{\Delta x})\right) \varphi^- \boldsymbol{n} \mathrm{d}\sigma, \end{aligned} \tag{3.15}$$

where $\varphi \in V_{\Delta x}$ and $\widetilde{\boldsymbol{h}}$ and $\widehat{\boldsymbol{h}}$ are given numerical flux functions. We seek the solution $\boldsymbol{w}_{\Delta x} \in C^1((0, t^{end}); V^3_{\Delta x})$ which fulfills every equation of (3.15) *for all $\varphi \in V_{\Delta x}$.*

This definition lets the choice of numerical flux functions open and one can find different choices in literature, see [112] for a comparison of some of them. In this thesis we choose a flux function which is the local Lax[16]-Friedrichs, also called Rusanov[17] [154], flux function but uses a slightly different stabilization.

Definition 3.15 (Numerical flux function)**.** *The* numerical flux function *considered in this thesis is given by*

$$\widetilde{\boldsymbol{h}}(\boldsymbol{w}^-, \boldsymbol{w}^+) := \frac{1}{2}\left(\widetilde{\boldsymbol{F}}(\boldsymbol{w}^-) + \widetilde{\boldsymbol{F}}(\boldsymbol{w}^+)\right) + \frac{1}{2} \mathrm{Diag}\left\{\varepsilon^{-2}, 1, 1\right\} \left(\boldsymbol{w}^- - \boldsymbol{w}^+\right) \cdot \boldsymbol{n}$$

[16] Peter David Lax, *1926

[17] Viktor Vladimirovich Rusanov, *1919

for the implicit and

$$\widehat{\boldsymbol{h}}(\boldsymbol{w}^-,\boldsymbol{w}^+) := \frac{1}{2}\left(\widehat{\boldsymbol{F}}(\boldsymbol{w}^-) + \widehat{\boldsymbol{F}}(\boldsymbol{w}^+)\right) + \varepsilon\left(\boldsymbol{w}^- - \boldsymbol{w}^+\right)\cdot\boldsymbol{n}$$

for the explicit part. Note that the numerical flux functions depend on the normal vector $\boldsymbol{n}$*, but for simplicity we drop this dependence.*

Remark 3.16. *The numerical flux function for the implicit part introduced in Definition 3.15 can be seen as a preconditioned local Lax-Friedrichs numerical flux function and therefore the basic idea is similar to the preconditioning idea for Roe-type methods, see [171] and Section 3.1.*

This flux function takes a special role in the analytical investigation of the final method, see Chapter 5 for more details.

Remark 3.17. *For simplicity we derived the discontinuous Galerkin method for a given triangulation with periodic boundary conditions. If one considers non-periodic boundary conditions one possible modification of the numerical flux function at domain boundary* $\partial\Omega$*, exemplarily for the implicit flux function, is given by*

$$\widetilde{\boldsymbol{h}}_{\partial\Omega} := \widetilde{\boldsymbol{F}}(\boldsymbol{w}_{\partial\Omega}),$$

where $\boldsymbol{w}_{\partial\Omega}$ *denotes the computed boundary value which fulfills the corresponding boundary conditions.*

Definition 3.14 gives the semi-discrete discontinuous Galerkin method, which contains time derivatives. To obtain a fully-discrete method one introduces basis functions for the space $V_{\Delta x}$, chooses every basis function as test function and derives from this a system of ordinary differential equations. Consequently, we can apply the IMEX Runge-Kutta method as given in Definition 3.9 on these ODEs and obtain a fully-discrete formulation. This all results in an IMEX Runge-Kutta discontinuous Galerkin discretization.

Remark 3.18. *In literature, one finds several different IMEX discontinuous Galerkin methods. Next to the one presented before, there are for example methods which use an IMEX decomposition of the domain, see e.g. [99], or for elliptic equations where the convective and diffusive part are handled with the different IMEX parts, see e.g. [87, 179].*

3.3. Splittings

The main ingredient of an IMEX time discretization is a splitting. To find such a splitting we start with a short review of the importance of asymptotic properties defined in Section 3.2.1. Afterwards, we review splittings for weakly compressible flows from literature to see what has been done before. Finally, we consider the novel splitting analyzed in this thesis.

3.3.1. Importance of asymptotic properties

In the following we consider different splittings and numerical methods which fulfill the properties defined in Section 3.2.1 and which do not. By this, we show the importance of these properties.

Proving that a method is asymptotically consistent follows in general the same steps. For a time iterative method one shows that well-prepared initial data - well-prepared in a discrete sense - are preserved during the iteration process. Furthermore, one shows with the help of an asymptotic expansion that the lowest order terms of this expansion are a consistent discretization of the limiting equation, similarly as we have done for the continuous equations in Section 2.2.

Remark 3.19. *For the ordinary differential equation as given in Definition 2.12 the limiting equation is given in Equation* (2.22) *and well-prepared initial conditions are given in Definition 2.20.*

In the next lemma we consider a numerical method which is not asymptotically consistent. With the help of this example we also comment on why the asymptotically consistent property is needed to be fulfilled in this setting.

Lemma 3.20. *The explicit Euler method as given in Equation* (3.6) *applied to Equation* (2.15)*, i.e.*

$$\begin{pmatrix} y^{n+1} \\ z^{n+1} \end{pmatrix} = \begin{pmatrix} y^{n} \\ z^{n} \end{pmatrix} + \Delta t \begin{pmatrix} f(y^n, z^n) \\ \frac{1}{\varepsilon} g(y^n, z^n) \end{pmatrix}$$

is not asymptotically consistent.

Proof. We show this by computing two steps of the method and showing that then the solution is not given as an asymptotic expansion and the limiting method cannot be computed. For this, we assume that the initial conditions are well-prepared, see Remark 3.19. Then we start by rearranging the terms and replacing all values $y^n, z^n, y^{n+1}, z^{n+1}$ by an asymptotic expansion

$$\begin{pmatrix} y^1_{(0)} - y^0_{(0)} \\ \varepsilon\left(z^1_{(0)} - z^0_{(0)}\right) \end{pmatrix} = \Delta t \begin{pmatrix} f(y^0_{(0)}, z^0_{(0)}) \\ g(y^0_{(0)}, z^0_{(0)}) \end{pmatrix} \tag{3.16}$$

$$+ \varepsilon \Delta t \begin{pmatrix} \mathcal{O}(1) \\ \partial_y g(y^0_{(0)}, z^0_{(0)}) y^0_{(1)} + \partial_z g(y^0_{(0)}, z^0_{(0)}) z^0_{(1)} \end{pmatrix} + \mathcal{O}(\varepsilon^2). \tag{3.17}$$

Separating in terms of ε leads to

$$g(y^0_{(0)}, z^0_{(0)}) = 0 \tag{3.18}$$

and

$$z^1_{(0)} - z^0_{(0)} = \Delta t \left(\partial_y g(y^0_{(0)}, z^0_{(0)}) y^0_{(1)} + \partial_z g(y^0_{(0)}, z^0_{(0)}) z^0_{(1)}\right). \tag{3.19}$$

Equation (3.18) is fulfilled since we assumed that y^0 and z^0 are well-prepared. From (3.19) and the update for $y^1_{(0)}$ we know that these are $\mathcal{O}(\Delta t^2)$ approximations for the exact solution, but in the next step the condition

$$g(y^1_{(0)}, z^1_{(0)}) = 0,$$

must be again fulfilled, which is in general not the case. We can only conclude that

$$\begin{aligned} g(y^1_{(0)}, z^1_{(0)}) =& \partial_y g(y^0_{(0)}, z^0_{(0)})(y^1_{(0)} - y^0_{(0)}) + \partial_z g(y^0_{(0)}, z^0_{(0)})(z^1_{(0)} - z^0_{(0)}) \\ &+ \mathcal{O}((y^1_{(0)} - y^0_{(0)})^2) + \mathcal{O}((y^1_{(0)} - y^0_{(0)})(z^1_{(0)} - z^0_{(0)})) + \mathcal{O}((z^1_{(0)} - z^0_{(0)})^2). \end{aligned}$$

Together with Equation (3.19), the update for $y^1_{(0)}$ in Equation (3.16) and the representation of $z_{(1)}$ given in Equation (2.25) we obtain

$$g(y^1_{(0)}, z^1_{(0)}) = \mathcal{O}(\Delta t^2).$$

Thus there are $\mathcal{O}(\varepsilon^{-1})$ terms remaining in the equation which are compensated by a term in the asymptotic expansion, which means that these values must depend on ε and therefore the asymptotic solution cannot be obtained. □

From the previous proof we obtained that $\mathcal{O}(\varepsilon^{-1})$ terms remain after one step. From this we can also directly see that the explicit discretization of the equation is not asymptotically stable. To obtain an asymptotically consistent method we need that at least some parts of the equation are handled implicitly.

Next, we give a splitting where the resulting method is asymptotically consistent but not asymptotically stable. Please note, that we mostly investigate asymptotic stability with the help of numerical examples since for discontinuous Galerkin schemes coupled with an IMEX time integration it is difficult to prove stability.

Lemma 3.21. *The numerical method*

$$\begin{pmatrix} y^{n+1} \\ z^{n+1} \end{pmatrix} = \begin{pmatrix} y^{n} \\ z^{n} \end{pmatrix} + \Delta t \begin{pmatrix} f(y^n, z^n) \\ \frac{2}{3}\frac{1}{\varepsilon} g(y^n, z^n) \end{pmatrix} + \Delta t \begin{pmatrix} 0 \\ \frac{1}{3}\frac{1}{\varepsilon} g(y^{n+1}, z^{n+1}) \end{pmatrix}, \tag{3.20}$$

is asymptotically consistent but not asymptotically stable.

Proof. We start by proving the AC property. Therefore, we assume that every quantity is given as an asymptotic expansion, which results in

$$\begin{pmatrix} y^{n+1}_{(0)} - y^{n}_{(0)} \\ 0 \end{pmatrix} = \frac{\Delta t}{3} \begin{pmatrix} 3f(y^n_{(0)}, z^n_{(0)}) \\ \frac{2}{\varepsilon} g(y^n_{(0)}, z^n_{(0)}) \end{pmatrix} + \frac{\Delta t}{3\varepsilon} \begin{pmatrix} 0 \\ g(y^{n+1}_{(0)}, z^{n+1}_{(0)}) \end{pmatrix} + \begin{pmatrix} \mathcal{O}(\varepsilon) \\ \mathcal{O}(1) \end{pmatrix}.$$

We consider the $\mathcal{O}(\varepsilon^{-1})$ terms of the second equation and since we assumed well-prepared initial conditions the equation reduces to

$$0 = g(y^{n+1}_{(0)}, z^{n+1}_{(0)}) \qquad \Rightarrow \qquad z^{n+1}_{(0)} = D(y^{n+1}_{(0)}).$$

Thus, the next step is also well-prepared. From this and the $\mathcal{O}(1)$ terms of the first equation one can directly see that this is a consistent discretization of the limiting equation. The method is not asymptotically stable which can be seen by computing the numerical solution for different values of Δt and ε and compare this solution with the exact one. These results are summarized in Figure 3.3, where we obtain that the numerical approximation is not stable for an ε depending range of values of Δt. □

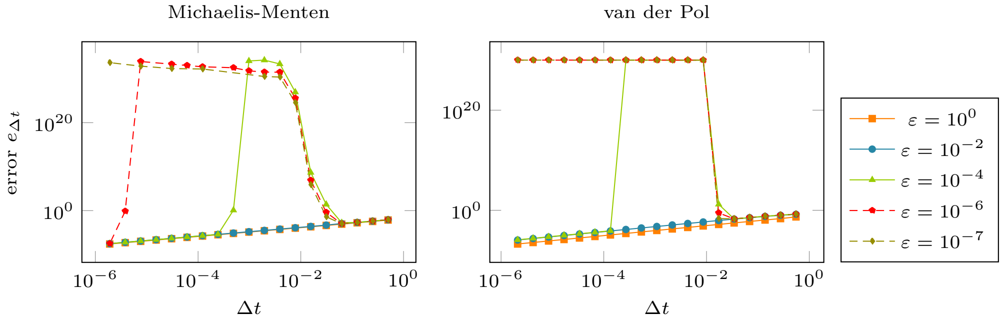


Figure 3.3.: Convergence behavior of the *asymptotically consistent* but not *asymptotically stable* numerical method given in Lemma 3.21, left: Michaelis Menten equation, right: van der Pol equation. Values evaluated as NaN are set to 10^{30}.

The next method we consider is asymptotically stable but not asymptotically accurate. Note that the AA property is a high order property and therefore we give the splitting and a corresponding IMEX Runge-Kutta scheme, whose Butcher tableaux are given in the appendix.

Lemma 3.22. *The numerical method defined by the splitting*

$$\widehat{\boldsymbol{G}} := \begin{pmatrix} f(y,z) \\ 0 \end{pmatrix} \qquad \text{and} \qquad \widetilde{\boldsymbol{G}} := \begin{pmatrix} 0 \\ \frac{1}{\varepsilon} g(y,z) \end{pmatrix}, \tag{3.21}$$

coupled with the third order BPR_353 IMEX Runge-Kutta scheme, see Table A.4, is asymptotically consistent, asymptotically stable but not asymptotically accurate.

Proof. The proof of asymptotic consistency is similar to the proof of Lemma 3.21. In Figure 3.4 we can see that a stable approximation is computed. Furthermore, Figure 3.4 also shows that the method is not asymptotically accurate, i.e. we obtain a drop of the convergence order depending on ε. □

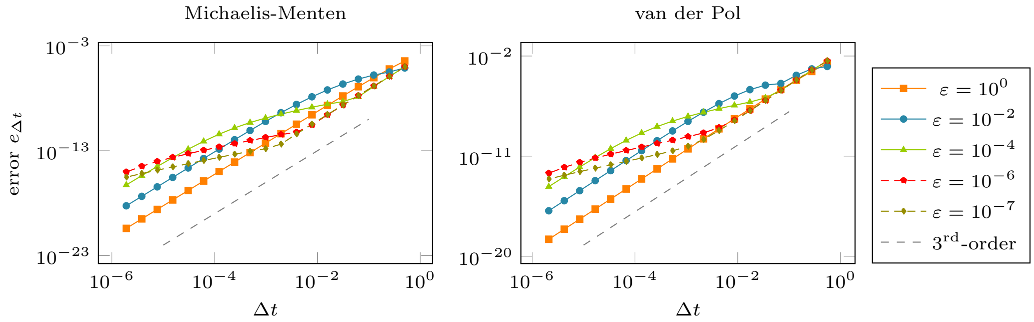


Figure 3.4.: Convergence behavior of the *asymptotically consistent*, *asymptotically stable* but not *asymptotically accurate* numerical method given in Lemma 3.22 with the BPR_353 scheme given in Table A.4. Left: Michaelis Menten equation, right: van der Pol equation.

The splitting defined in Lemma 3.22 takes a special role in the rest of this thesis. It has been investigated in the setting of ordinary differential equations [24, 27] and also extended to kinetic equations [26].

Definition 3.23. *We call the splitting, defined in Lemma 3.22,* standard splitting *in the setting of ordinary differential equations.*

Finally, we also consider a numerical method which fulfills all properties, i.e. it is asymptotically consistent, asymptotically stable and asymptotically accurate. The method is given by considering the same splitting as given in Lemma 3.22 coupled with a different IMEX Runge-Kutta method, whose Butcher tableau is given in the appendix.

Lemma 3.24. *The numerical method defined by the splitting*

$$\widehat{\boldsymbol{G}} := \begin{pmatrix} f(y,z) \\ 0 \end{pmatrix} \qquad \textit{and} \qquad \widetilde{\boldsymbol{G}} := \begin{pmatrix} 0 \\ \frac{1}{\varepsilon} g(y,z) \end{pmatrix}, \tag{3.22}$$

coupled with the third order BHR_553 IMEX Runge-Kutta scheme, see Table A.6, is asymptotically consistent, asymptotically stable and asymptotically accurate.

Proof. We refer to [25] for a proof of the asymptotic accuracy of this method. From this also asymptotic consistency and asymptotic stability follows. This is also shown in Figure 3.5. □

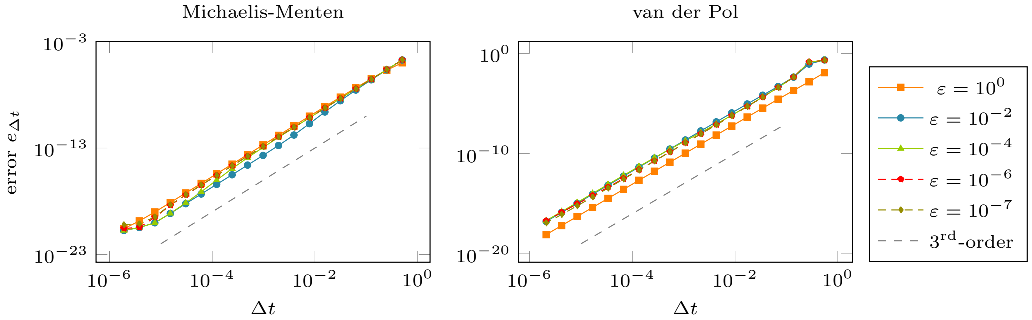


Figure 3.5.: Convergence behavior of the *asymptotically consistent*, *asymptotically stable* and *asymptotically accurate* numerical method given in Lemma 3.24 with the BHR_553 scheme given in Table A.6. Left: Michaelis Menten equation, right: van der Pol equation.

On a first sight, the method given in Lemma 3.24 seems to fulfill all properties defined before, but it needs to solve a non-linear system of equations and therefore huge computational cost are needed. Furthermore, it is not directly extendable to the isentropic Euler equations.

3.3.2. Splittings in literature

There are many different approaches for different equations in literature which use IMEX methods, similar to the ones defined before, to compute low Mach flows. We discuss several of these approaches in the following. A comparison of different splittings coupled with a low order finite volume discretization for the one-dimensional shallow water equations in terms of stability is given in [189].

Using IMEX methods leads to a system of equations which is expensive to solve, especially if the system of equations is non-linear. Therefore, often a time semi-discrete equation, i.e. an equation which results from applying the time discretization but leaving the spatial derivatives continuous, is used to derive an elliptic equation which can be solved more efficiently. Splittings which are discussed in the following and also use this elliptic equation to reduce computational cost are for example given in [28, 46, 49, 50, 78]. Please note, that the elliptic equation directly depends on the chosen time discretization and therefore these methods are often low order methods. The splittings discussed in the following are all derived for one specific equation and are mostly not directly extendable to an arbitrary equation.

For the isentropic Euler equations, and this is also true for the full Euler equations, it is clear that one has to handle the pressure gradient, which is scaled with ε^{-2}, in some sense implicitly. Furthermore, one needs to add some terms from the conservation of mass equation to the implicit part to obtain a hyperbolic implicit system and to fulfill the divergence free constraint in the limit, i.e. to obtain an asymptotically consistent method.

One of the first splittings is given in [1], where the limiting density is used to split the one-dimensional isentropic Euler equations in non-conservative form. In the same work also a splitting of the full Euler equations is given, where the speed of sound of a base flow is used to split the equations in, again, non-conservative form. A few years later in [108] a splitting for the one-dimensional Euler equations is proposed, where the pressure in the conservation of momentum equation and the complete conservation of energy equation are handled implicitly. Starting with this work, there were many splittings developed in the following years.

In [49] a splitting for the isentropic Navier-Stokes equations is introduced, where the pressure is handled implicitly and the velocity is split with the help of a Helmholtz-Hodge decomposition. Next, in [50] for the isentropic Euler equations a splitting is proposed where the pressure gradient is split with a factor which is in $\mathcal{O}(\varepsilon^{-2})$ such that the remaining explicit pressure is non-stiff. Furthermore the conservation of mass equation is handled completely implicit, i.e.

$$\widetilde{\boldsymbol{F}} := \begin{pmatrix} \rho \boldsymbol{u} \\ \frac{1-\alpha\varepsilon^2}{\varepsilon^2} p(\rho)\,\mathrm{Id} \end{pmatrix} \qquad \text{and} \qquad \widehat{\boldsymbol{F}} := \begin{pmatrix} 0 \\ \rho \boldsymbol{u} \otimes \boldsymbol{u} + \alpha p(\rho)\,\mathrm{Id} \end{pmatrix}, \tag{3.23}$$

where α is a splitting coefficient chosen as $\alpha \leq \varepsilon^{-2}$. This results in a non-linear implicit part. In [28, 55] a similar splitting is used, but the pressure gradient is handled completely implicit. These methods are extended to the full Euler equations in [28, 46, 55], see also [89, 90] for similar splittings.

In comparison to [50], in [78] a splitting is introduced where the implicit part is linear. This is obtained by adding a linear term in ρ and handling this implicitly. The coefficient of this term is chosen as the minimum of $p'(\rho)$ at the explicit time instance. This results in an explicit part which is hyperbolic and non-stiff, i.e.

$$\widetilde{\boldsymbol{F}} := \begin{pmatrix} (1-\alpha)\rho \boldsymbol{u} \\ \frac{a(t)}{\varepsilon^2} \rho\,\mathrm{Id} \end{pmatrix} \qquad \text{and} \qquad \widehat{\boldsymbol{F}} := \begin{pmatrix} \alpha \rho \boldsymbol{u} \\ \frac{p(\rho)-a(t)\rho}{\varepsilon^2}\,\mathrm{Id} \end{pmatrix}, \tag{3.24}$$

where α and $a(t)$ are given splitting coefficients with $\alpha \approx \varepsilon^2$ and $a(t) := \min_{\boldsymbol{x}} p'(\rho)$.

In [120] a splitting technique coupled with a low order finite difference method for the shallow water equations is proposed, where the equations are split by separating long and short waves in an intermediate step.

Furthermore, in [146, 187] splittings for one dimensional isentropic Euler equations are given and then discretized with a low order time integration method. In both works, parts of the resulting discretization are handled with a Lagrange projection scheme. Additionally, [146] uses a high order discontinuous Galerkin method for the spatial discretization.

Then, in [135] a splitting for the full Euler equations is proposed, which can be seen as a combination of the ideas in [50] and [78]. In detail [135] handles the pressure completely implicitly and uses the minimum of the pressure variable to split the energy equation. This splitting results in a non-linear implicit part, which includes the calculation of a minimum. Unfortunately, instabilities occurred caused by the splitting and a pressure correction has to be used for stabilization.

Inspired by the stability results in [135], in [21, 22] a splitting for the shallow water equations is published. The basic idea is to split the pressure with a linearization around a reference state, in this setting the lake at rest motivated by the works [71, 72, 148]. This splitting was then extended to other equations, like the Euler equations in atmospheric flow, see [23, 183].

3.3.3. RS-IMEX splitting

As mentioned before, in [135] a splitting for the full Euler equations is derived which is not stable without an additional stabilization. To understand the reason of this, [160] considers an m-dimensional linear hyperbolic equation in one space dimension, i.e.

$$0 = \partial_t \boldsymbol{w} + \boldsymbol{A}\partial_x \boldsymbol{w} = \partial_t \boldsymbol{w} + \left(\widetilde{\boldsymbol{A}} + \widehat{\boldsymbol{A}}\right)\partial_x \boldsymbol{w} \qquad \text{with} \qquad \boldsymbol{A} \in \mathbb{R}^{m\times m} \text{ const}, \tag{3.25}$$

where $\boldsymbol{A}$ has eigenvalues which are in $\mathcal{O}(\varepsilon^{-1})$, to find conditions for a splitting which delivers a stable method if $\varepsilon \ll 1$. For this, a low order discretization is analyzed with the help of a modified equation analysis, see [180] for more details on this. The resulting modified equation is given by

$$\partial_t \boldsymbol{w} + \boldsymbol{A}\partial_x \boldsymbol{w} = \frac{\Delta t}{2}\left(\frac{(\widetilde{\alpha}+\widehat{\alpha})\Delta x}{\Delta t}\operatorname{Id} - \widehat{\boldsymbol{A}}^2 + \widetilde{\boldsymbol{A}}^2 + \widetilde{\boldsymbol{A}}\widehat{\boldsymbol{A}} - \widehat{\boldsymbol{A}}\widetilde{\boldsymbol{A}}\right)\partial_{xx}\boldsymbol{w},$$

where $\widetilde{\alpha}$ and $\widehat{\alpha}$ are the corresponding implicit and explicit stabilization coefficients. Roughly spoken, see again [160] for more details, to obtain a stable discretization one needs a positive diffusion coefficient. In the case of a splitting into a stiff part $\widetilde{\boldsymbol{A}}$ and non-stiff part $\widehat{\boldsymbol{A}}$ one observes that the implicit contribution $(\widetilde{\boldsymbol{A}}^2)$ does not affect stability and that the explicit contribution $(-\widehat{\boldsymbol{A}}^2)$ can be stabilized by choosing an appropriate CFL-condition with $\Delta t = \mathcal{O}(\Delta x)$. What remains is the commutator

$$\widetilde{\boldsymbol{A}}\widehat{\boldsymbol{A}} - \widehat{\boldsymbol{A}}\widetilde{\boldsymbol{A}},$$

which should be small for a stable method, but in general this commutator is in $\mathcal{O}(\varepsilon^{-1})$. However, this analysis is not extendable to non-linear or multi-dimensional problems, but it helps to identify properties a splitting should fulfill also in the non-linear case. This gets more clear if we consider a non-linear conservation law with a splitting, e.g. as given in Equation (3.14). Then we can rewrite the equation by using the chain rule of derivatives,

$$\begin{aligned} 0 =& \partial_t \boldsymbol{w} + \partial_x \widetilde{\boldsymbol{F}}(\boldsymbol{w}) + \partial_x \widehat{\boldsymbol{F}}(\boldsymbol{w}) \\ =& \partial_t \boldsymbol{w} + \nabla_{\boldsymbol{w}} \widetilde{\boldsymbol{F}}(\boldsymbol{w})\partial_x \boldsymbol{w} + \nabla_{\boldsymbol{w}} \widehat{\boldsymbol{F}}(\boldsymbol{w})\partial_x \boldsymbol{w}, \end{aligned}$$

and identify the matrices $\widetilde{\boldsymbol{A}}$ and $\widehat{\boldsymbol{A}}$ of (3.25) with $\nabla_{\boldsymbol{w}}\widetilde{\boldsymbol{F}}$ and $\nabla_{\boldsymbol{w}}\widehat{\boldsymbol{F}}$ respectively. Thus, a splitting should be chosen in such a way that the explicit part is small enough to not affect the stability of the implicit part if $\varepsilon \ll 1$. This was the starting point of the *RS-IMEX* splitting. To understand the basic idea of the RS-IMEX splitting, we consider an ordinary differential equation of the form

$$\frac{d}{dt}\boldsymbol{w} = \boldsymbol{G}(\boldsymbol{w}),$$

where $\boldsymbol{G} = \mathcal{O}(\varepsilon^{-1})$. If we assume that a *reference solution* $\boldsymbol{w}_{ref}$ is given which fulfills

$$\boldsymbol{w}_{ref} - \boldsymbol{w} = \mathcal{O}(\varepsilon),$$

then we can compute a linearization around this reference solution and obtain

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{w}) =& \overbrace{\boldsymbol{G}(\boldsymbol{w}_{ref}) + \nabla_{\boldsymbol{w}}\boldsymbol{G}(\boldsymbol{w}_{ref})(\boldsymbol{w} - \boldsymbol{w}_{ref})}^{=:\widetilde{\boldsymbol{G}}} \\ &+ \underbrace{\boldsymbol{G}(\boldsymbol{w}) - \boldsymbol{G}(\boldsymbol{w}_{ref}) - \nabla_{\boldsymbol{w}}\boldsymbol{G}(\boldsymbol{w}_{ref})(\boldsymbol{w} - \boldsymbol{w}_{ref})}_{=:\widehat{\boldsymbol{G}}}.\end{aligned}$$

As expected, the implicit part is stiff, this can be directly seen by

$$\widetilde{\boldsymbol{G}}(\boldsymbol{w}) = \mathcal{O}(\varepsilon^{-1}) + \mathcal{O}(\varepsilon^{-1}) \cdot \mathcal{O}(\varepsilon),$$

but it is also linear and therefore efficient to solve with a linear equation solver. The hope is, that the explicit part is small,

$$\begin{aligned}\widehat{\boldsymbol{G}}(\boldsymbol{w}) =& \boldsymbol{G}(\boldsymbol{w}) - \boldsymbol{G}(\boldsymbol{w}_{ref}) - \nabla_{\boldsymbol{w}}\boldsymbol{G}(\boldsymbol{w}_{ref})(\boldsymbol{w} - \boldsymbol{w}_{ref}) \\ =& \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{ref})^T \left\{\nabla^2_{\boldsymbol{w}}\boldsymbol{G}(\boldsymbol{\xi})\right\}(\boldsymbol{w} - \boldsymbol{w}_{ref}) \\ =& \mathcal{O}(\varepsilon) \cdot \mathcal{O}(\varepsilon^{-1}) \cdot \mathcal{O}(\varepsilon) \\ =& \mathcal{O}(\varepsilon),\end{aligned}$$

for an intermediate value $\boldsymbol{\xi}$. Thus, the splitting idea can lead to an explicit part which becomes less dominant as $\varepsilon \to 0$ and therefore this can lead to a splitting with the desired stability properties. What remains is the choice of the reference solution $\boldsymbol{w}_{ref}$, but there is a direct canonical choice due to the asymptotic behavior of the equation. We discuss the reference solution for every equation separately.

The RS-IMEX splitting first has been introduced for ordinary differential equations in [J4] and for the isentropic Euler equations in [J3, C3]. In [J2] the order of convergence in the setting of ordinary differential equations is investigated for a special class of IMEX Runge-Kutta methods which we also consider in this thesis. In [C2] the RS-IMEX splitting is applied to a slightly different ordinary differential equation and tested in terms of initial layers. For the isentropic Euler equations the splitting is tested and compared to the splitting given in [78] with a low order finite volume discretization in [J3]. The step to a high order discretization, namely discontinuous Galerkin coupled with IMEX Runge-Kutta schems – again the same method we use in this thesis –, is done in [J1]. Furthermore, in [J5], the same method is compared to fully explicit and implicit discretizations.

Next to these publications the RS-IMEX splitting coupled with a low order finite volume discretization has been tested for the shallow water equations in different configurations: first of all the splitting is tested and compared to splittings from literature in terms of stability in [189]. Additionally, the splitting is analyzed in one space dimensions in [185], two space dimension in [186] and in the case of Coriolis[18] force in [188].

[18] Gaspard Gustave de Coriolis, 1792 – 1843

Ordinary differential equations

For the ordinary differential equation given in Definition 2.12, we know from our asymptotic analysis in Section 2.2.1, that the solution $(y, z)^T$ converges towards a limiting solution $(y_{(0)}, z_{(0)})^T$ if we assume well-prepared initial conditions. This limiting solution can be computed by the limiting equation, see Equation (2.22) or (2.24), and formally satisfies

$$\begin{pmatrix} y \\ z \end{pmatrix} - \begin{pmatrix} y_{(0)} \\ z_{(0)} \end{pmatrix} = \mathcal{O}(\varepsilon).$$

Definition 3.25. *In the setting of ordinary differential equations, see Definition 2.12, we call the limiting solution, i.e. the solution which results from taking the $\varepsilon \to 0$ limit,* reference solution. *We denote this solution with $(\cdot)_{ref}$, in detail*

$$\begin{pmatrix} y_{ref} \\ z_{ref} \end{pmatrix} := \lim_{\varepsilon \to 0} \begin{pmatrix} y \\ z \end{pmatrix}.$$

Note that the reference solution corresponds to the solution of the limiting equation given in Equation (2.22) *or* (2.24).

Note that we always assume well-prepared initial conditions such that the $\varepsilon \to 0$ limit exists. Then, the RS-IMEX splitting for ordinary differential equations, as given in Definition 2.12, can be directly computed.

Definition 3.26. *The RS-IMEX splitting for ordinary differential equations as given in Definition 2.12 in the notation of Equation* (3.9) *with $\boldsymbol{w} = (y, z)^T$ is given by*

$$\widetilde{\boldsymbol{G}} := \begin{pmatrix} f(y_{ref}, z_{ref}) + \partial_y f(y_{ref}, z_{ref})(y - y_{ref}) + \partial_z f(y_{ref}, z_{ref})(z - z_{ref}) \\ \frac{1}{\varepsilon} \left\{ g(y_{ref}, z_{ref}) + \partial_y g(y_{ref}, z_{ref})(y - y_{ref}) + \partial_z g(y_{ref}, z_{ref})(z - z_{ref}) \right\} \end{pmatrix}$$

and

$$\widehat{\boldsymbol{G}} := \begin{pmatrix} f(y, z) \\ \frac{1}{\varepsilon} g(y, z) \end{pmatrix} - \widetilde{\boldsymbol{G}}.$$

Unfortunately, the reference solution cannot be assumed to be given and has to be computed with the help of a numerical method.

Remark 3.27. *For the ordinary differential equation given in Definition 2.12 we obtain the reference solution by solving the limiting Equation* (2.24) *with an explicit method if the function D is given exactly or by solving Equation* (2.22) *with a suitable implicit or IMEX method if the function D is not given.*

Isentropic Euler equations

In the case of the isentropic Euler equations, we follow the same steps as before for ordinary differential equations. In our analysis in Section 2.2 we were able to show that if we assume well-prepared initial and boundary conditions, see Definition 2.20 and Remark 2.29, we can compute the $\varepsilon \to 0$ limit and the limiting solution formally satisfies

$$\begin{pmatrix} \rho \\ \boldsymbol{u} \end{pmatrix} - \begin{pmatrix} \rho_{(0)} \\ \boldsymbol{u}_{(0)} \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\varepsilon^2) \\ \mathcal{O}(\varepsilon) \end{pmatrix}.$$

Basically, the value $\rho_{(0)}$ is constant and given by the initial conditions. Furthermore, $\boldsymbol{u}_{(0)}$ can be computed by solving the incompressible equations, see Definition 2.9. Similarly to Definition 3.25 we can now define $\rho_{(0)}$ and $\boldsymbol{u}_{(0)}$ as the reference solution.

Definition 3.28. *In the setting of isentropic Euler equations, see Lemma 2.7, we call the limiting solution, i.e. the solution which results from taking the $\varepsilon \to 0$ limit,* reference solution*. We denote this solution with $(\cdot)_{ref}$, in detail*

$$\begin{pmatrix} \rho_{ref} \\ \boldsymbol{u}_{ref} \end{pmatrix} := \lim_{\varepsilon \to 0} \begin{pmatrix} \rho \\ \boldsymbol{u} \end{pmatrix}.$$

Note that the reference solution corresponds to $\rho_{(0)}$ defined by initial and boundary values and to $\boldsymbol{u}_{(0)}$ computed by solving the incompressible Euler equations.

With this reference solution defined, we can compute the splitting of flux function $\boldsymbol{F}$ by computing the Taylor expansion up to first order terms of $\boldsymbol{F}$ around the reference solution $(\rho_{ref}, \boldsymbol{u}_{ref})^T$.

Definition 3.29. *The RS-IMEX splitting for the isentropic Euler equations, see Lemma 2.7, is given by*

$$\widetilde{\boldsymbol{F}}(\rho, \rho\boldsymbol{u}) := \boldsymbol{F}(\rho_{ref}, (\rho\boldsymbol{u})_{ref}) + \nabla_{\rho,\rho\boldsymbol{u}} \boldsymbol{F}(\rho_{ref}, (\rho\boldsymbol{u})_{ref}) \begin{pmatrix} \rho - \rho_{ref} \\ \rho\boldsymbol{u} - (\rho\boldsymbol{u})_{ref} \end{pmatrix}$$

and

$$\widehat{\boldsymbol{F}}(\rho, \rho\boldsymbol{u}) := \boldsymbol{F}(\rho, \rho\boldsymbol{u}) - \widetilde{\boldsymbol{F}}(\rho, \rho\boldsymbol{u}).$$

We can directly compute the exact representation of $\widetilde{\boldsymbol{F}}$ and $\widehat{\boldsymbol{F}}$. The closed form is given by

$$\widetilde{\boldsymbol{F}} := \begin{pmatrix} \rho\boldsymbol{u} \\ -\rho\boldsymbol{u}_{ref} \otimes \boldsymbol{u}_{ref} + \rho\boldsymbol{u} \otimes \boldsymbol{u}_{ref} + \boldsymbol{u}_{ref} \otimes \rho\boldsymbol{u} + \frac{1}{\varepsilon^2}\left(p(\rho_{ref}) + p'(\rho_{ref})(\rho - \rho_{ref})\right) \end{pmatrix}$$

for the implicit and

$$\widehat{\boldsymbol{F}} := \begin{pmatrix} 0 \\ \rho(\boldsymbol{u} - \boldsymbol{u}_{ref}) \otimes (\boldsymbol{u} - \boldsymbol{u}_{ref}) + \frac{1}{\varepsilon^2}\left(p(\rho) - p(\rho_{ref}) - p'(\rho_{ref})(\rho - \rho_{ref})\right) \end{pmatrix}$$

for the explicit part. To check whether this is a useful splitting for a hyperbolic conservation law we need to show that both parts are hyperbolic, see Definition 3.8. The implicit part is per definition hyperbolic, but for the explicit part we need to compute the eigenvalues of $\nabla_{\rho,\rho\boldsymbol{u}} \widehat{\boldsymbol{F}} \cdot \boldsymbol{n}$

$$\lambda_1 = 0, \quad \lambda_2 = (\boldsymbol{u} - \boldsymbol{u}_{ref}) \cdot \boldsymbol{n} \quad \text{and} \quad \lambda_3 = 2(\boldsymbol{u} - \boldsymbol{u}_{ref}) \cdot \boldsymbol{n} \tag{3.26}$$

and from this we can conclude that the resulting system is hyperbolic.

Remark 3.30. *For the isentropic Euler equations as given in Lemma 2.7 we obtain the reference solution by solving the limiting equation, i.e. the incompressible Euler equations, see Definition 2.9, with a suitable implicit or IMEX method.*

3.4. Conclusion and summary

In this chapter we introduced the numerical method we consider in this thesis. This is a combination of an IMEX Runge-Kutta method with a discontinuous Galerkin discretization and the RS-IMEX splitting, which we introduced for ordinary differential equations and the isentropic Euler equations. Let us summarize the general idea behind the RS-IMEX splitting in a short remark.

Remark 3.31 (RS-IMEX splitting)**.** *The RS-IMEX splitting is a splitting technique for singularly perturbed differential equations where the asymptotic limit is of special interest. If the asymptotic limit can be computed with an additional equation, then it is used as a reference solution to split the singularly perturbed differential equation with the help of a linearization around this reference solution.*

In principle, this technique can be extended to many equations which have a limit as a small parameter goes to zero. Unfortunately, we cannot guarantee that the resulting splitting fulfills the hyperbolicity condition of Definition 3.8.

The hope of this discretization is that the RS-IMEX splitting has a good performance – if the reference solution is computed efficiently – in this setting, since the implicit part is linear and close to a fully implicit discretization.

Remark 3.32. *There are some splittings in literature which are similar to the RS-IMEX splitting.*

First of all, there is the splitting given in [21, 22] for the shallow water equations, where the lake at rest is used as reference solution to discretize the pressure with a linearization.

Then, in [65] a splitting for kinetic equations at low Knudsen[19] *number is derived, where the collision operator is split with a linearization around an approximation of the corresponding limiting solution.*

An additional similarity to the RS-IMEX splitting can be seen if one considers the IMEX Euler discretization, see Equation (3.10)*, and then uses the explicit state* $\mathbf{w}^n$ *as reference solution. Then the explicit part sums up to zero and the implicit part becomes the Taylor expansion up to second order terms, i.e. a second order method which is similar to linear implicit methods presented in Section 3.1, see also [63, 64] for such a method.*

[19] Martin Hans Christian Knudsen, 1871 – 1949

4. Asymptotic convergence order

Runge-Kutta methods applied to ordinary differential equations, as given in Definition 2.12, suffer from a problem called order reduction [24, 27, 79, 81]. Order reduction is caused by the small parameter ε and can lead, for a p^{th} order method, to a tremendous loss of convergence order. In detail, the following convergence behavior can be obtained for a high order IMEX Runge-Kutta scheme:

– For $\Delta t = \mathcal{O}(\varepsilon^0) = \mathcal{O}(1)$ the order p is obtained.

– For $\mathcal{O}(\varepsilon) < \Delta t < \mathcal{O}(\varepsilon^0) = \mathcal{O}(1)$, which is a relevant case for high order methods, the order of convergence drops below p.

– For $\Delta t < \mathcal{O}(\varepsilon)$ the order p is obtained.

An example of this convergence behavior is given in Figure 4.1 where the error behavior of the ARS_443 scheme (Table A.3) for different values of ε is shown. We can directly obtain the ε depending range of values of Δt where the order of convergence is reduced. In this figure all splittings behave very similarly. As a second example we consider in Figure 4.2 the convergence behavior of the BPR_353 scheme (Table A.4) for different values of ε. In comparison to Figure 4.1 we obtain that the standard splitting shows a more distinctive order reduction than the other two methods. Furthermore, we obtain that the RS-IMEX splitting behaves similarly to the fully implicit discretization.

In this chapter we investigate in which way an IMEX Runge-Kutta scheme coupled with the RS-IMEX splitting suffers from order reduction. Therefore, we first review order reduction in literature for a fully implicit discretization and an IMEX Runge-Kutta scheme coupled with the standard splitting. Afterwards, we prove the main theorem of this chapter which gives a detailed asymptotic order analysis for globally stiffly accurate IMEX Runge-Kutta schemes of type CK with uniform $\boldsymbol{c}$, see Section 3.2.2, coupled with the RS-IMEX splitting. Finally, we consider different IMEX Runge-Kutta schemes to investigate the influence of order reduction on the general convergence behavior in more detail.

Parts of this chapter have been previously published in [J2, J4]. This includes the main theorem, Theorem 4.6, and the corresponding proof, see [J2]. All numerical results have been recomputed with a higher precision but are similar to the computations in [J2, J4].

4.1. Order reduction

In Section 2.2 we assumed that the solution of the ordinary differential equation, see Definition 2.12, is given as an asymptotic expansion. This assumption helps us to derive different differential algebraic equations for the components of the asymptotic expansion, see Equations (2.22) and (2.23). It is quite canonical to do the same for the numerical approximation, i.e. to assume that for $\varepsilon \ll \Delta t$

$$\begin{pmatrix} y^{n+1} \\ z^{n+1} \end{pmatrix} = \begin{pmatrix} y^{n+1}_{(0)} \\ z^{n+1}_{(0)} \end{pmatrix} + \varepsilon \begin{pmatrix} y^{n+1}_{(1)} \\ z^{n+1}_{(1)} \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y^{n+1}_{(2)} \\ z^{n+1}_{(2)} \end{pmatrix} + \mathcal{O}(\varepsilon^3),$$

and then derive the corresponding numerical methods for $y^{n+1}_{(i)}$ and $z^{n+1}_{(i)}$ for $i = 0, 1, \dots$. For this, we can compute the error of every component separately and obtain that the global error is given by

$$\begin{pmatrix} y^{n+1} - y(t^{n+1}) \\ z^{n+1} - z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} y^{n+1}_{(0)} - y_{(0)}(t^{n+1}) \\ z^{n+1}_{(0)} - z_{(0)}(t^{n+1}) \end{pmatrix}$$

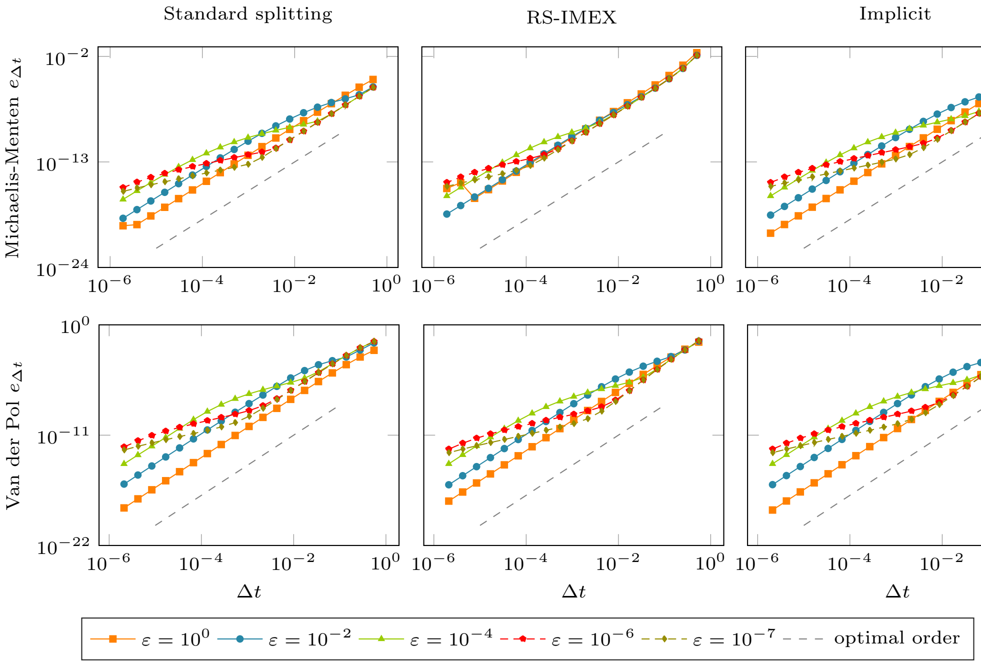


Figure 4.1.: Convergence behavior of the ARS_443 scheme, see Table A.3, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

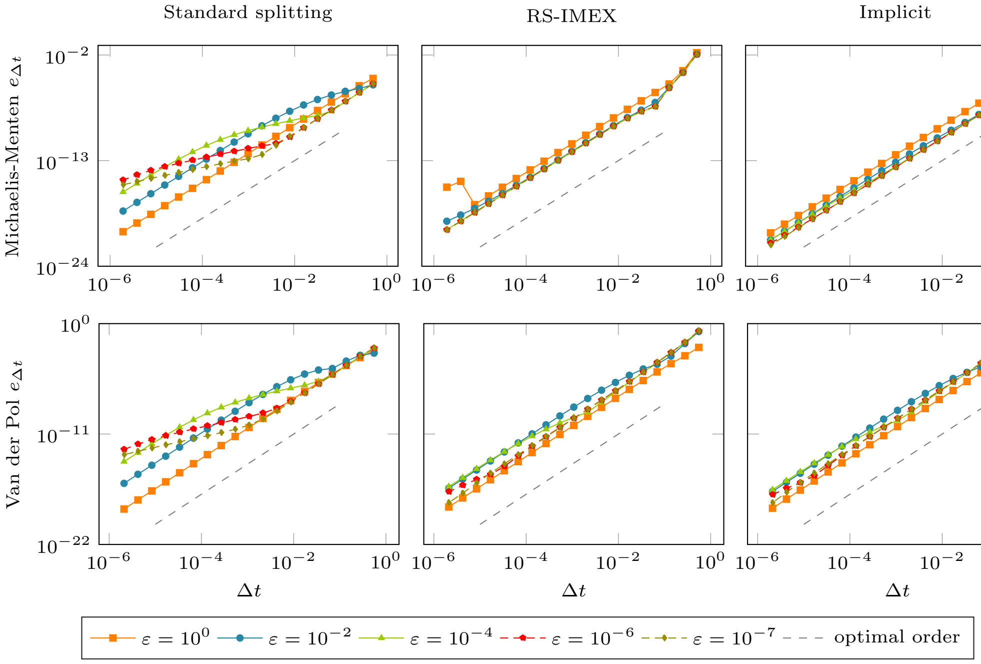


Figure 4.2.: Convergence behavior of the BPR_353 scheme, see Table A.4, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

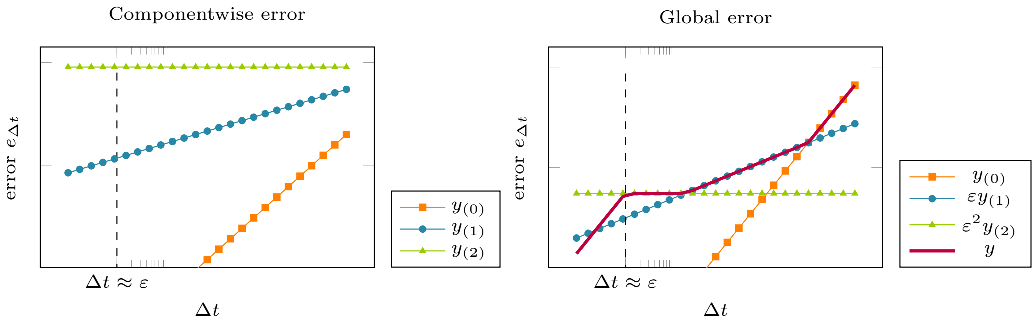


Figure 4.3.: Theoretical example for the convergence behavior of a numerical method for an ordinary differential equation, where the components of the asymptotic expansion are not solved with same accuracy. Left the theoretical error of the components of the asymptotic expansion and right the error scaled with the corresponding power of ε and the global error (solid red line) obtained by the method.

$$
\begin{aligned}
&+\varepsilon \begin{pmatrix} y_{(1)}^{n+1} - y_{(1)}(t^{n+1}) \\ z_{(1)}^{n+1} - z_{(1)}(t^{n+1}) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y_{(2)}^{n+1} - y_{(2)}(t^{n+1}) \\ z_{(2)}^{n+1} - z_{(2)}(t^{n+1}) \end{pmatrix} + \dots \\
= &\begin{pmatrix} \mathcal{O}(\Delta t^{p^y_{(0)}}) \\ \mathcal{O}(\Delta t^{p^z_{(0)}}) \end{pmatrix} + \varepsilon \begin{pmatrix} \mathcal{O}(\Delta t^{p^y_{(1)}}) \\ \mathcal{O}(\Delta t^{p^z_{(1)}}) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} \mathcal{O}(\Delta t^{p^y_{(2)}}) \\ \mathcal{O}(\Delta t^{p^z_{(2)}}) \end{pmatrix} + \dots,
\end{aligned}
$$

where $p_{(i)}^{y/z}$ for $i = 0, \dots$ denote the order of convergence for the i^{th} component of the asymptotic expansion for y and z, respectively. Note that in principle $p_{(i)}^{y/z} < 0$ is also possbile. Ideally, for a p^{th}-order numerical method all components are solved with the desired accuracy such that $p_{(i)}^{y/z} \equiv p$ for all $i = 0, 1, \dots$ or since the asymptotic expansion is only given if $\varepsilon \ll \Delta t$ and the error of the i^{th} component is scaled with ε^i, it might be enough that $p_{(i)}^{y/z} = p - i$ for $i = 0, 1, \dots$. Unfortunately, this is - in general - not valid for an (IMEX) Runge-Kutta time discretization method. Exemplarily, if the components are solved with a convergence behavior as given in Figure 4.3 (left), then the error is scaled with different powers of ε and a convergence behavior as given in Figure 4.3 (right) including a loss of convergence order is obtained. Note that for $\varepsilon \approx \Delta t$ the numerical solution cannot be given as an asymptotic expansion anymore and the classical convergence behavior of the chosen (IMEX) Runge-Kutta scheme can be expected.

Order reduction is investigated in [79, 81] for implicit Runge-Kutta methods. These results are extended in [24, 27] to IMEX Runge-Kutta methods coupled with the standard splitting and in [J2] to IMEX Runge-Kutta methods coupled with the RS-IMEX splitting. In [79, 81] it is shown that order reduction directly depends on the stage-wise structure of Runge-Kutta schemes and on the stage order, which is the minimal internal order, of the method.

Definition 4.1 (Internal order)**.** *The i^{th} stage of an IMEX Runge-Kutta method with classical order of convergence p, see Definition 3.9, has* internal order *$q^i \le p$ if q^i is the maximal value such that there holds*

$$\boldsymbol{w}^{n,i} - \boldsymbol{w}(t^{n,i}) = \mathcal{O}(\Delta t^p) + \mathcal{O}(\Delta t^{q^i+1})$$

for a smooth solution $\boldsymbol{w}$ of a corresponding ordinary differential equation. Similarly, we define $\widetilde{q}^i$ to be the internal order of the resulting Runge-Kutta scheme if only the implicit part is used.

Definition 4.2 (Stage order)**.** *We denote the minimal internal order, see Definition 4.1, of an IMEX Runge-Kutta scheme* stage order *q, i.e.*

$$q := \min_{1 \le i \le s} q^i.$$

Similarly, we define $\widetilde{q}$ to be the stage order of the resulting Runge-Kutta scheme if only the implicit part is used.

Theorem 4.3 (Implicit Runge-Kutta method [79, 81])**.** *Consider an implicit stiffly accurate Runge-Kutta method of classical order p and stage order $\widetilde{q}$ with $\widetilde{q} < p$, where the Butcher matrix $\widetilde{\boldsymbol{A}}$ is invertible, applied to the ordinary differential equation given in Definition 2.12. For $\varepsilon \ll \Delta t$ the numerical error is given by*

$$\begin{pmatrix} y^{n+1} - y(t^{n+1}) \\ z^{n+1} - z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\Delta t^p) \\ \mathcal{O}(\Delta t^p) \end{pmatrix} + \varepsilon \begin{pmatrix} \mathcal{O}(\Delta t^{\widetilde{q}+1}) \\ \mathcal{O}(\Delta t^{\widetilde{q}}) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} \mathcal{O}(\Delta t^{\widetilde{q}}) \\ \mathcal{O}(\Delta t^{\widetilde{q}-1}) \end{pmatrix} + \dots .$$

Proof. The proof of this theorem is given in [79] and also in [81, VI.3.: Thm 3.3 and 3.4]. □

From this theorem we can directly conclude the following:

1. The $\mathcal{O}(1)$ component of the solution is solved with the desired order of accuracy.

2. The $\mathcal{O}(\varepsilon)$ component of the solution is solved with an order of accuracy which depends on the stage order.

3. The $z_{(i)}$ component is solved with an order of accuracy which is one order less than the corresponding order of accuracy of the $y_{(i)}$ component for $i = 1, 2, \dots$.

A fully implicit scheme can be constructed in such a way that a relatively large stage order is given, but due to efficiency reasons one would like to restrict oneself to diagonally implicit schemes, i.e. schemes where the matrix $\widetilde{\boldsymbol{A}}$ is a lower triangular matrix with $\widetilde{\boldsymbol{A}}_{i,i} \neq 0$ for $i = 2, \dots, s$. Choosing this one could only obtain a stage order of at most two. If $\widetilde{\boldsymbol{A}}$ is a lower triangular matrix with $\widetilde{\boldsymbol{A}}_{1,1} = 0$, Theorem 4.3 cannot be applied since $\widetilde{\boldsymbol{A}}$ is not invertible anymore, but the resulting method shows a similar convergence behavior and the theorem can be adjusted if the resulting sub-matrix $\widetilde{\boldsymbol{A}}_{2\dots s, 2\dots s}$ is invertible.

Corollary 4.4. *Consider an implicit stiffly accurate Runge-Kutta scheme of classical order p and stage order $\widetilde{q}$ with $\widetilde{q} < p$, where the Butcher matrix $\widetilde{\boldsymbol{A}}$ is a lower triangular one with $\widetilde{\boldsymbol{A}}_{1,1} = 0$ and $\widetilde{\boldsymbol{A}}_{2\dots s, 2\dots s}$ invertible, applied to the ordinary differential equation given in Definition 2.12. For $\varepsilon \ll \Delta t$ the numerical error is given by*

$$\begin{pmatrix} y^{n+1} - y(t^{n+1}) \\ z^{n+1} - z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\Delta t^p) \\ \mathcal{O}(\Delta t^p) \end{pmatrix} + \varepsilon \begin{pmatrix} \mathcal{O}(\Delta t^{\widetilde{q}+1}) \\ \mathcal{O}(\Delta t^{\widetilde{q}}) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} \mathcal{O}(\Delta t^{\widetilde{q}}) \\ \mathcal{O}(\Delta t^{\widetilde{q}-1}) \end{pmatrix} + \dots .$$

Proof. We refer to [81] and show how the proof of Theorem 4.3 has to be changed to obtain the desired result. For this, we use an asymptotic expansion for every quantity of the equation and the Runge-Kutta method. Then, the lowest order method, i.e. the method for $y_{(0)}^{n+1}$ and $z_{(0)}^{n+1}$, is a stiffly accurate implicit Runge-Kutta scheme applied to the differential algebraic equation (2.22). Since the method is stiffly accurate, this is equivalent to applying the same method to the non-stiff equation (2.24) and therefore the desired accuracy is obtained.

For the error in $y_{(1)}^{n+1}$ and $z_{(1)}^{n+1}$, we do the same steps as in Theorem 3.4 in Section VI.3 in [81], see also [79], with a small modification. The proof works on the difference between numerical and the exact solution. Therefore the first stage, which is equal to the previous time instance $y_{(1)}^{n}$ and $z_{(1)}^{n}$, is directly obtained with the desired accuracy and does not affect the result. The remaining stages contain an implicit matrix which is invertible and the remaining proof works as given in [79, 81]. Finally, the error for $y_{(2)}^{n+1}$ and $z_{(2)}^{n+1}$ is obtained by induction, see again [79, 81]. □

Theorem 4.3 is extended to IMEX Runge-Kutta schemes coupled with the standard splitting, see Definition 3.23, in [24] for general IMEX schemes and in [27] for some special schemes including globally stiffly accurate IMEX Runge-Kutta schemes.

Theorem 4.5 (IMEX Runge-Kutta with standard splitting [24, 27])**.** *Consider a globally stiffly accurate IMEX Runge-Kutta scheme of classical order $p > 1$ and type CK coupled with the standard splitting, see*

Definition 3.23, applied to the ordinary differential equation given in Definition 2.12. Then, if $\varepsilon \ll \Delta t$, the numerical error is given by

$$\begin{pmatrix} y^{n+1} - y(t^{n+1}) \\ z^{n+1} - z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\Delta t^p) \\ \mathcal{O}(\Delta t^p) \end{pmatrix} + \varepsilon \begin{pmatrix} \mathcal{O}(\Delta t) \\ \mathcal{O}(\Delta t) \end{pmatrix} + \dots .$$

Proof. The proof of this theorem can be found in [24, 27]. □

For IMEX Runge-Kutta schemes one always has a stage order of at most one and the errors of the $\mathcal{O}(\varepsilon^2)$ terms are dropped in the theorem in [24, 27]. If we follow the steps in [81], we obtain an error behavior of the form

$$\begin{pmatrix} y^{n+1} - y(t^{n+1}) \\ z^{n+1} - z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\Delta t^p) \\ \mathcal{O}(\Delta t^p) \end{pmatrix} + \varepsilon \begin{pmatrix} \mathcal{O}(\Delta t^q) \\ \mathcal{O}(\Delta t^q) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} \mathcal{O}(\Delta t^{q-1}) \\ \mathcal{O}(\Delta t^{q-1}) \end{pmatrix} + \dots .$$

Compared with Theorem 4.3 and Corollary 4.4 we obtain the following:

1. The $y_{(i)}$ and $z_{(i)}$ component for $i = 1, 2, \dots$ is solved with the same order of accuracy.

2. The $y_{(i)}$ component for $i = 1, 2, \dots$ is solved with one order of accuracy less than given in Corollary 4.4 if we assume that the same stage order is given.

4.2. RS-IMEX splitting

In this section we extend the results given in Theorem 4.3, Corollary 4.4 and Theorem 4.5 to IMEX Runge-Kutta methods coupled with the RS-IMEX splitting. First computations, see Figure 4.1 and 4.2, show that the RS-IMEX splitting leads to a convergence behavior which is more similar to a fully implicit discretization. This is stated in the following theorem.

Theorem 4.6. *Consider a globally stiffly accurate IMEX Runge-Kutta scheme of type CK with uniform* $\mathbf{c}$ *coupled with the RS-IMEX splitting, see Definition 3.26, applied to the ordinary differential equation given in Definition 2.12. Then, if the reference solution is given exactly and $\varepsilon \ll \Delta t$, the numerical error at time instance t^{n+1} with $n\Delta t \leq const$ is given by*

$$\begin{pmatrix} y^{n+1} - y(t^{n+1}) \\ z^{n+1} - z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\Delta t^{r_1}) \\ \mathcal{O}(\Delta t^{r_1}) \end{pmatrix} + \varepsilon \begin{pmatrix} \mathcal{O}(\Delta t^{r_2+1}) \\ \mathcal{O}(\Delta t^{r_2}) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} \mathcal{O}(\Delta t^{r_2}) \\ \mathcal{O}(\Delta t^{r_2-1}) \end{pmatrix} + \dots ,$$

where p denotes the classical order, q the stage order with $q \leq p$, $\widetilde{q}$ the implicit stage order and the constants r_1 and r_2 are given by

$$r_1 := \min\{p, 2(q+1)\} \qquad \textit{and} \qquad r_2 := \min\{r_1 - 1, \widetilde{q}, q+1\}.$$

Proof. For the proof of this theorem we follow the steps of [24, 81] and assume that all quantities can be represented by an asymptotic expansion, similarly as for the continuous equations. With this, we can separate the numerical method in different equations for the components of the asymptotic expansion, this is done in Lemma 4.10, and recalculate the numerical error to prove the desired convergence behavior separately for every component in different theorems, i.e.

$$\begin{pmatrix} y^{n+1} - y(t^{n+1}) \\ z^{n+1} - z(t^{n+1}) \end{pmatrix} = \overbrace{\begin{pmatrix} y_{(0)}^{n+1} - y_{(0)}(t^{n+1}) \\ z_{(0)}^{n+1} - z_{(0)}(t^{n+1}) \end{pmatrix}}^{\text{Thm. 4.14}} + \varepsilon \overbrace{\begin{pmatrix} y_{(1)}^{n+1} - y_{(1)}(t^{n+1}) \\ z_{(1)}^{n+1} - z_{(1)}(t^{n+1}) \end{pmatrix}}^{\text{Thm. 4.23}}$$

$$+\,\underbrace{\varepsilon^2 \begin{pmatrix} y_{(2)}^{n+1} - y_{(2)}(t^{n+1}) \\ z_{(2)}^{n+1} - z_{(2)}(t^{n+1}) \end{pmatrix}}_{\text{Thm. 4.24}} + \dots\ .$$

□

To simplify the resulting analysis we introduce a notation such that we can rewrite the IMEX Runge-Kutta method, see Definition 3.9, in a vector form. We start by rewriting the solutions of the internal stages $y^{n,i}$ and $z^{n,i}$ for $i = 2, \dots, s$ in a vector and we do the same for all time instances and reference solutions. The first stage $y^{n,1} = y^n$ is handled separately. Note that we assume a uniform $\boldsymbol{c}$ and therefore the reference solution used by the implicit and explicit parts are the same at time instance $t^{n,i}$.

Definition 4.7 (Notation: vectors)**.** *We define the following for an s-stage IMEX Runge-Kutta scheme of type CK which is globally stiffly accurate and has a uniform* $\boldsymbol{c}$*:*

1. $\boldsymbol{t} \in \mathbb{R}^{s-1}$ *denotes the vector with all internal time instances, i.e.*

   $$\boldsymbol{t} := (t^{n,2}, \dots, t^{n,s})^T.$$

2. $\boldsymbol{y}^\Delta \in \mathbb{R}^{s-1}$ *and* $\boldsymbol{z}^\Delta \in \mathbb{R}^{s-1}$ *denote the stage vectors which are given by*

   $$\boldsymbol{y}^\Delta := \left(y^{n,2}, \dots, y^{n,s}\right)^T \qquad \textit{and} \qquad \boldsymbol{z}^\Delta := \left(z^{n,2}, \dots, z^{n,s}\right)^T.$$

3. $\boldsymbol{y}_{ref} \in \mathbb{R}^{s-1}$ *and* $\boldsymbol{z}_{ref} \in \mathbb{R}^{s-1}$ *denote the reference solution vectors by*

   $$\boldsymbol{y}_{ref} := \begin{pmatrix} y_{ref}(t^{n,2}) \\ \vdots \\ y_{ref}(t^{n,s}) \end{pmatrix} \qquad \textit{and} \qquad \boldsymbol{z}_{ref} := \begin{pmatrix} z_{ref}(t^{n,2}) \\ \vdots \\ z_{ref}(t^{n,s}) \end{pmatrix}.$$

4. *For the difference between a numerical approximation and the corresponding exact solution we define*

   $$\Delta y^{n,i} = y^{n,i} - y(t^{n,i}) \qquad \textit{and} \qquad \Delta z^{n,i} = z^{n,i} - z(t^{n,i}).$$

   Similarly to $\boldsymbol{y}^\Delta$ *and* $\boldsymbol{z}^\Delta$ *we define* $\Delta\boldsymbol{y}^\Delta$ *and* $\Delta\boldsymbol{z}^\Delta$*.*

5. $\boldsymbol{e} := (1, \dots, 1)^T \in \mathbb{R}^{s-1}$ *denotes the vector filled with ones.*

We can extend this vector notation to function evaluations needed by the IMEX Runge-Kutta scheme.

Definition 4.8 (Notation: function evaluation)**.** *We define* $f(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta) \in \mathbb{R}^{s-1}$ *and* $g(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta) \in \mathbb{R}^{s-1}$*, where* $\boldsymbol{y}^\Delta$ *and* $\boldsymbol{z}^\Delta$ *are given in Definition 4.7, by*

$$f(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta) := \begin{pmatrix} f(y^{n,2}, z^{n,2}) \\ \vdots \\ f(y^{n,s}, z^{n,s}) \end{pmatrix} \qquad \textit{and} \qquad g(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta) := \begin{pmatrix} g(y^{n,2}, z^{n,2}) \\ \vdots \\ g(y^{n,s}, z^{n,s}) \end{pmatrix}.$$

In an analogous way we define $\widetilde{f}(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta)$*,* $\widehat{f}(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta)$*,* $\widetilde{g}(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta)$*,* $\widehat{g}(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta)$ *and* $D(\boldsymbol{y}^\Delta)$*. The function* D *is given by Lemma 2.18. Furthermore, for the evaluation of the function* f *or* g *at the exactly given reference solution* y_{ref} *and* z_{ref} *we define*

$$f(t^{n,i}) := f(y_{ref}(t^{n,i}), z_{ref}(t^{n,i})) \qquad \textit{and} \qquad g(t^{n,i}) := g(y_{ref}(t^{n,i}), z_{ref}(t^{n,i})).$$

In an analogous way we define $\partial_y f(t^{n,i})$, $\partial_z f(t^{n,i})$, $\partial_y g(t^{n,i})$ and $\partial_z g(t^{n,i})$ and compositions of these functions. Furthermore we define

$$f(\boldsymbol{t}) := \Big(f(t^{n,2}), \dots, f(t^{n,s})\Big)^T \qquad \text{and} \qquad g(\boldsymbol{t}) := \Big(g(t^{n,2}), \dots, g(t^{n,s})\Big)^T,$$

and analogously for the corresponding derivatives and compositions of these functions.

With these definitions made, we can rewrite the IMEX Runge-Kutta method in vector notation.

Corollary 4.9. *One step of a globally stiffly accurate IMEX Runge-Kutta method, see Definition 3.9, of type CK with uniform $\boldsymbol{c}$ applied to the ordinary differential equation given in Definition 2.12 with a corresponding splitting can be written as*

$$\begin{aligned}
\boldsymbol{y}^{\Delta} &= y^n \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \widetilde{f}(y^n, z^n) + \widehat{\boldsymbol{\alpha}} \widehat{f}(y^n, z^n) + \widetilde{\boldsymbol{B}} \widetilde{f}(\boldsymbol{y}^{\Delta}, \boldsymbol{z}^{\Delta}) + \widehat{\boldsymbol{B}} \widehat{f}(\boldsymbol{y}^{\Delta}, \boldsymbol{z}^{\Delta})\right) \\
\boldsymbol{z}^{\Delta} &= z^n \boldsymbol{e} + \frac{\Delta t}{\varepsilon} \left(\widetilde{\boldsymbol{\alpha}} \widetilde{g}(y^n, z^n) + \widehat{\boldsymbol{\alpha}} \widehat{g}(y^n, z^n) + \widetilde{\boldsymbol{B}} \widetilde{g}(\boldsymbol{y}^{\Delta}, \boldsymbol{z}^{\Delta}) + \widehat{\boldsymbol{B}} \widehat{g}(\boldsymbol{y}^{\Delta}, \boldsymbol{z}^{\Delta})\right),
\end{aligned}$$

where the matrices $\widetilde{\boldsymbol{B}}$ and $\widehat{\boldsymbol{B}}$ and the vectors $\widetilde{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\alpha}}$ are given as in Remark 3.10. The update is then defined by $y^{n+1} := y^{n,s}$ and $z^{n+1} := z^{n,s}$.

Proof. This corollary follows directly from the previously defined notation, see Definitions 4.7 and 4.8, and the definition of the corresponding IMEX Runge-Kutta method, see Definitions 3.9 and 3.11. □

4.2.1. ε expansion of the IMEX Runge-Kutta method

Following the steps in [24, 81], we assume that all quantities and especially the stage vectors $\boldsymbol{y}^{\Delta}$ and $\boldsymbol{z}^{\Delta}$ are given as an asymptotic expansion, i.e.

$$\boldsymbol{y}^{\Delta} = \boldsymbol{y}^{\Delta}_{(0)} + \varepsilon \boldsymbol{y}^{\Delta}_{(1)} + \varepsilon^2 \boldsymbol{y}^{\Delta}_{(2)} + \dots \quad \text{and} \quad \boldsymbol{z}^{\Delta} = \boldsymbol{z}^{\Delta}_{(0)} + \varepsilon \boldsymbol{z}^{\Delta}_{(1)} + \varepsilon^2 \boldsymbol{z}^{\Delta}_{(2)} + \dots. \tag{4.1}$$

With this we can compute different equations which have to be fulfilled by the components of the asymptotic expansion.

Lemma 4.10. *Under the assumptions of Theorem 4.6, the components of the asymptotic expansion of the stage vectors $\boldsymbol{y}^{\Delta}$ and $\boldsymbol{z}^{\Delta}$ fulfill*

$$\boldsymbol{y}^{\Delta}_{(0)} = y^n_{(0)} \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \widetilde{f}^n_{(0)} + \widehat{\boldsymbol{\alpha}} \widehat{f}^n_{(0)}\right) + \Delta t \left(\widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{f}}^{\Delta}_{(0)} + \widehat{\boldsymbol{B}} \widehat{\boldsymbol{f}}^{\Delta}_{(0)}\right) \tag{4.2}$$

$$0 = \widetilde{\boldsymbol{\alpha}} \widetilde{g}^n_{(0)} + \widehat{\boldsymbol{\alpha}} \widehat{g}^n_{(0)} + \widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{g}}^{\Delta}_{(0)} + \widehat{\boldsymbol{B}} \widehat{\boldsymbol{g}}^{\Delta}_{(0)}, \tag{4.3}$$

where we used the abbreviations $(\cdot)^{n,i}_{(0)} := (\cdot)_{(0)}(y^{n,i}_{(0)}, z^{n,i}_{(0)})$ and $(\boldsymbol{\cdot})^{\Delta}_{(0)} := (\cdot)_{(0)}(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)})$,

$$\boldsymbol{y}^{\Delta}_{(1)} = y^n_{(1)} \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \widetilde{f}^n_{(1)} + \widehat{\boldsymbol{\alpha}} \widehat{f}^n_{(1)}\right) + \Delta t \left(\widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{f}}^{\Delta}_{(1)} + \widehat{\boldsymbol{B}} \widehat{\boldsymbol{f}}^{\Delta}_{(1)}\right) \tag{4.4}$$

$$\boldsymbol{z}^{\Delta}_{(0)} = z^n_{(0)} \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \widetilde{g}^n_{(1)} + \widehat{\boldsymbol{\alpha}} \widehat{g}^n_{(1)}\right) + \Delta t \left(\widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{g}}^{\Delta}_{(1)} + \widehat{\boldsymbol{B}} \widehat{\boldsymbol{g}}^{\Delta}_{(1)}\right), \tag{4.5}$$

where we used the abbreviations $(\cdot)^{n,i}_{(1)} := (\cdot)_{(1)}(y^{n,i}_{(0)}, y^{n,i}_{(1)}, z^{n,i}_{(0)}, z^{n,i}_{(1)})$ and $(\boldsymbol{\cdot})^{\Delta}_{(1)} := (\cdot)_{(1)}(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{y}^{\Delta}_{(1)}, \boldsymbol{z}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(1)})$, and

$$\boldsymbol{y}^{\Delta}_{(2)} = y^n_{(2)} \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \widetilde{f}^n_{(2)} + \widehat{\boldsymbol{\alpha}} \widehat{f}^n_{(2)}\right) + \Delta t \left(\widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{f}}^{\Delta}_{(2)} + \widehat{\boldsymbol{B}} \widehat{\boldsymbol{f}}^{\Delta}_{(2)}\right) \tag{4.6}$$

$$\boldsymbol{z}^{\Delta}_{(1)} = z^n_{(1)} \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \widetilde{g}^n_{(2)} + \widehat{\boldsymbol{\alpha}} \widehat{g}^n_{(2)}\right) + \Delta t \left(\widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{g}}^{\Delta}_{(2)} + \widehat{\boldsymbol{B}} \widehat{\boldsymbol{g}}^{\Delta}_{(2)}\right), \tag{4.7}$$

where we used the abbreviations $(\cdot)^{n,i}_{(2)} := (\cdot)_{(2)}(y^{n,i}_{(0)}, y^{n,i}_{(1)}, y^{n,i}_{(2)}, z^{n,i}_{(0)}, z^{n,i}_{(1)}, z^{n,i}_{(2)})$ and $(\boldsymbol{\cdot})^{\Delta}_{(2)} := (\cdot)_{(2)}(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{y}^{\Delta}_{(1)}, \boldsymbol{y}^{\Delta}_{(2)}, \boldsymbol{z}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(1)}, \boldsymbol{z}^{\Delta}_{(2)})$. The functions $\widetilde{f}_{(i)}$, $\widetilde{g}_{(i)}$, $\widehat{f}_{(i)}$ and $\widehat{g}_{(i)}$ for $i = 0, 1, 2$ are given in Corollaries 4.11, 4.12 and 4.13 for the RS-IMEX splitting.

Proof. We consider the method given in Corollary 4.9, insert the asymptotic expansion for every quantity and use a Taylor expansion to obtain terms in different powers of ε. Separating in ε leads to the given equations with corresponding functions $\widetilde{f}^{n,i}_{(j)}$, $\widetilde{g}^{n,i}_{(j)}$, $\widehat{f}^{n,i}_{(j)}$ and $\widehat{g}^{n,i}_{(j)}$ for $j = 0, 1, 2$. □

The functions $\widetilde{f}^{n,i}_{(j)}$, $\widetilde{g}^{n,i}_{(j)}$, $\widehat{f}^{n,i}_{(j)}$ and $\widehat{g}^{n,i}_{(j)}$ for $j = 0, 1, 2$ directly follow from the proof of Lemma 4.10 and the definition of the RS-IMEX splitting. Note that one can also compute these functions for the standard splitting or the fully implicit scheme to derive the corresponding limiting method.

Corollary 4.11. *The functions $\widetilde{f}^{n,i}_{(0)}$, $\widehat{f}^{n,i}_{(0)}$, $\widetilde{g}^{n,i}_{(0)}$ and $\widehat{g}^{n,i}_{(0)}$ of Lemma 4.10 are given by*

$$\widetilde{f}^{n,i}_{(0)} := f(t^{n,i}) + \partial_y f(t^{n,i})\Delta y^{n,i}_{(0)} + \partial_z f(t^{n,i})\Delta z^{n,i}_{(0)}, \tag{4.8}$$

$$\widehat{f}^{n,i}_{(0)} := f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - f(t^{n,i}) - \partial_y f(t^{n,i})\Delta y^{n,i}_{(0)} - \partial_z f(t^{n,i})\Delta z^{n,i}_{(0)}, \tag{4.9}$$

$$\widetilde{g}^{n,i}_{(0)} := g(t^{n,i}) + \partial_y g(t^{n,i})\Delta y^{n,i}_{(0)} + \partial_z g(t^{n,i})\Delta z^{n,i}_{(0)}, \tag{4.10}$$

$$\widehat{g}^{n,i}_{(0)} := g(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - g(t^{n,i}) - \partial_y g(t^{n,i})\Delta y^{n,i}_{(0)} - \partial_z g(t^{n,i})\Delta z^{n,i}_{(0)}, \tag{4.11}$$

where we used that $\Delta y^{n,i}_{(0)} = y^{n,i}_{(0)} - y_{ref}(t^{n,i})$ and analogously for $\Delta z^{n,i}_{(0)}$ since the reference solution is given exactly.

Corollary 4.12. *The functions $\widetilde{f}^{n,i}_{(1)}$, $\widehat{f}^{n,i}_{(1)}$, $\widetilde{g}^{n,i}_{(1)}$ and $\widehat{g}^{n,i}_{(1)}$ of Lemma 4.10 are given by*

$$\begin{aligned}
\widetilde{f}^{n,i}_{(1)} :=& \partial_y f(t^{n,i}) y^{n,i}_{(1)} + \partial_z f(t^{n,i}) z^{n,i}_{(1)},\\
\widetilde{g}^{n,i}_{(1)} :=& \partial_y g(t^{n,i}) y^{n,i}_{(1)} + \partial_z g(t^{n,i}) z^{n,i}_{(1)},\\
\widehat{f}^{n,i}_{(1)} :=& \Big(\partial_y f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_y f(t^{n,i})\Big) y^{n,i}_{(1)} + \Big(\partial_z f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_z f(t^{n,i})\Big) z^{n,i}_{(1)},\\
\widehat{g}^{n,i}_{(1)} :=& \Big(\partial_y g(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_y g(t^{n,i})\Big) y^{n,i}_{(1)} + \Big(\partial_z g(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_z g(t^{n,i})\Big) z^{n,i}_{(1)}.
\end{aligned}$$

Corollary 4.13. *The functions $\widetilde{f}^{n,i}_{(2)}$, $\widehat{f}^{n,i}_{(2)}$, $\widetilde{g}^{n,i}_{(2)}$ and $\widehat{g}^{n,i}_{(2)}$ of Lemma 4.10 are given by*

$$\begin{aligned}
\widetilde{f}^{n,i}_{(2)} :=& \partial_y f(t^{n,i}) y^{n,i}_{(2)} + \partial_z f(t^{n,i}) z^{n,i}_{(2)},\\
\widetilde{g}^{n,i}_{(2)} :=& \partial_y g(t^{n,i}) y^{n,i}_{(2)} + \partial_z g(t^{n,i}) z^{n,i}_{(2)},\\
\widehat{f}^{n,i}_{(2)} :=& \Big(\partial_y f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_y f(t^{n,i})\Big) y^{n,i}_{(2)} + \Big(\partial_z f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_z f(t^{n,i})\Big) z^{n,i}_{(2)}\\
&+ \frac{1}{2}\partial_{yy} f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) y^{n,i}_{(1)} y^{n,i}_{(1)} + \frac{1}{2}\partial_{zz} f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) z^{n,i}_{(1)} z^{n,i}_{(1)}\\
&+ \partial_{yz} f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) y^{n,i}_{(1)} z^{n,i}_{(1)},\\
\widehat{g}^{n,i}_{(2)} :=& \Big(\partial_y g(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_y g(t^{n,i})\Big) y^{n,i}_{(2)} + \Big(\partial_z g(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_z g(t^{n,i})\Big) z^{n,i}_{(2)}\\
&+ \frac{1}{2}\partial_{yy} g(y^{n,i}_{(0)}, z^{n,i}_{(0)}) y^{n,i}_{(1)} y^{n,i}_{(1)} + \frac{1}{2}\partial_{zz} g(y^{n,i}_{(0)}, z^{n,i}_{(0)}) z^{n,i}_{(1)} z^{n,i}_{(1)}\\
&+ \partial_{yz} g(y^{n,i}_{(0)}, z^{n,i}_{(0)}) y^{n,i}_{(1)} z^{n,i}_{(1)}.
\end{aligned}$$

With the results of Lemma 4.10 and the previous corollaries we get, in the end, methods for the components of the asymptotic expansion. Therefore, we can make a convergence analysis for each component separately to obtain the convergence behavior of the complete method.

4.2.2. The error of $y^{\triangle}_{(0)}$ and $z^{\triangle}_{(0)}$

We start with the zeroth order components of the asymptotic expansion and prove their error behavior.

Theorem 4.14. *Under the assumptions of Theorem 4.6, there holds*

$$\begin{pmatrix} y^{n+1}_{(0)} - y_{(0)}(t^{n+1}) \\ z^{n+1}_{(0)} - z_{(0)}(t^{n+1}) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\Delta t^{r_1}) \\ \mathcal{O}(\Delta t^{r_1}) \end{pmatrix} \qquad \text{with} \qquad r_1 := \min\{p, 2(q+1)\}.$$

Proof. This theorem is proven in several steps, which are given in different lemmas. First, from Lemma 4.16 we can conclude that

$$z_{(0)}^{n+1} - z_{ref}(t^{n+1}) = \mathcal{O}\left((y_{(0)}^{n+1} - y_{ref}(t^{n+1}))\right) + \mathcal{O}(\Delta t^{r_1}),$$

under the assumption that $\boldsymbol{y}_{(0)}^{\triangle} - \boldsymbol{y}_{ref} = \mathcal{O}(\Delta t^{q+1})$ holds, which is valid due to Lemma 4.19, see also Corollary 4.20. Finally, from Lemma 4.19 we obtain that

$$y_{(0)}^{n+1} - y_{ref}(t^{n+1}) = \mathcal{O}(\Delta t^{p+1}) + \mathcal{O}(\Delta t^{r_1}) = \mathcal{O}(\Delta t^{r_1})$$

holds. This all together proves the theorem. □

Ideally, the limiting method corresponds to the RS-IMEX discretization of the limiting equation, see Equation (2.24) coupled with an update step for $z_{(0)}$, i.e.

$$\frac{d}{dt} y_{(0)} = f(y_{(0)}, D(y_{(0)})) \quad \text{and} \quad y_{(0)} = D(y_{(0)}). \tag{4.12}$$

The corresponding RS-IMEX discretization of the limiting equation is given in the following corollary, where $\boldsymbol{v}^{\triangle}$ denotes the approximation of $\boldsymbol{y}_{(0)}$ and $\boldsymbol{\zeta}^{\triangle}$ the approximation of $\boldsymbol{z}_{(0)}$ and we use a similar notation as given in Definitions 4.7 and 4.8.

Corollary 4.15. *A globally stiffly accurate IMEX Runge-Kutta scheme of type CK with uniform* $\boldsymbol{c}$ *coupled with the RS-IMEX splitting applied to Equation* (4.12)*, is given by*

$$\begin{aligned}
\boldsymbol{v}^{\triangle} = v^n \boldsymbol{e} &+ \Delta t \widetilde{\boldsymbol{\alpha}} \left[f(t^n) + \left\{ \partial_y f(t^n) + \partial_z f(t^n) D'(t^n) \right\} (v^n - y_{ref}) \right] \\
&+ \Delta t \widehat{\boldsymbol{\alpha}} \left[f(v^n, D(v^n)) - f(t^n) - \left(\partial_y f(t^n) + \partial_z f(t^n) D'(t^n) \right) (v^n - y_{ref}) \right] \\
&+ \Delta t \widetilde{\boldsymbol{B}} \left[f(\boldsymbol{t}) + \operatorname{Diag} \left\{ \partial_y f(\boldsymbol{t}) + \operatorname{Diag} \left\{ \partial_z f(\boldsymbol{t}) \right\} D'(\boldsymbol{t}) \right\} (\boldsymbol{v}^{\triangle} - \boldsymbol{y}_{ref}) \right] \\
&+ \Delta t \widehat{\boldsymbol{B}} \left[f(\boldsymbol{v}^{\triangle}, D(\boldsymbol{v}^{\triangle})) - f(\boldsymbol{t}) \right] \\
&+ \Delta t \widehat{\boldsymbol{B}} \left[-\operatorname{Diag} \left\{ \partial_y f(\boldsymbol{t}) + \operatorname{Diag} \left\{ \partial_z f(\boldsymbol{t}) \right\} D'(\boldsymbol{t}) \right\} (\boldsymbol{v}^{\triangle} - \boldsymbol{y}_{ref}) \right],
\end{aligned}$$

where the reference solution is the same as the one used in Theorem 4.6 and the notation is the same as given in Definitions 4.7 and 4.8 with $f(t) = f(y_{ref}(t), D(y_{ref}(t)))$*. Furthermore the numerical approximation of* $z_{(0)}$ *is computed by*

$$\boldsymbol{\zeta}^{\triangle} = D(\boldsymbol{v}^{\triangle}).$$

Proof. This discretization can directly be obtained by computing the RS-IMEX splitting for the ordinary differential equation in (4.12) and noting that only one variable, $y_{(0)}$, is present and f fulfills

$$\nabla_y f(y, D(y)) = \partial_y f(y, D(y)) + \partial_z f(y, D(y)) D'(y).$$

Finally an update for $\boldsymbol{\zeta}^{\triangle}$ is added. □

Note that this method would give the desired convergence results since Equation (4.12) is non-stiff and sufficiently smooth. Therefore we can assume that $\boldsymbol{\zeta}^{\triangle}$ and $\boldsymbol{v}^{\triangle}$ are given with the convergence behavior which is prescribed by the chosen IMEX Runge-Kutta method.

In the following we show that the limiting RS-IMEX discretization equals the one given in Corollary 4.15 up to some high order terms. Therefore in a first step we show in which way $\boldsymbol{z}^{\triangle}$ corresponds to $D(\boldsymbol{y}_{(0)})$ and then in which way $\boldsymbol{z}^{\triangle}$ corresponds to $D(\boldsymbol{y}^{\triangle})$.

The error of $\boldsymbol{z}_{(0)}^{\Delta}$

In the following we solely consider the numerical method for $\boldsymbol{z}_{(0)}^{\Delta}$, which is the method given by Equations (4.3), (4.10) and (4.11). The goal is to compare the numerical solution $\boldsymbol{z}_{(0)}^{\Delta}$ with the approximation $\boldsymbol{\zeta}^{\Delta}$ of $z_{(0)}(\boldsymbol{t}) = \boldsymbol{z}_{ref}$ in Equation (4.12) which is given in Corollary 4.15. In a first step we simplify the numerical method for $\boldsymbol{z}_{(0)}^{\Delta}$ under the assumption that $\boldsymbol{y}_{(0)}^{\Delta}$, $y_{(0)}^{n}$ and $z_{(0)}^{n}$ are given with a specific accuracy.

Lemma 4.16. *Under the assumptions of Theorem 4.6, let $\Delta\boldsymbol{y}_{(0)}^{\Delta} = \boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{ref} = \mathcal{O}(\Delta t^{q+1})$, where q is the stage order of the IMEX Runge-Kutta scheme, and let $y_{(0)}^{n}$ and $z_{(0)}^{n}$ be given with accuracy in $\mathcal{O}(\Delta t^{r_1})$. Then the i^{th} component of $\boldsymbol{z}_{(0)}^{\Delta}$ fulfills*

$$z_{(0)}^{n,i} = D(y_{ref}(t^{n,i})) + D'(y_{ref}(t^{n,i}))\Delta y_{(0)}^{n,i} + \mathcal{O}(\Delta t^{r_1}\boldsymbol{e}), \tag{4.13}$$

with $r_1 := \min\{p, 2(q+1)\}$.

Proof. We first consider the terms of the previous time instance and obtain that these terms drop due to the accuracy of $y_{(0)}^{n}$ and $z_{(0)}^{n}$ since

$$\begin{aligned}\widetilde{g}_{(0)}(y_{(0)}^{n}, z_{(0)}^{n}) :=& \underbrace{g(t^n)}_{=0} + \underbrace{\partial_y g(t^n)\Delta y_{(0)}^{n}}_{=\mathcal{O}(\Delta t^{r_1})} + \underbrace{\partial_z g(t^n)\Delta z_{(0)}^{n}}_{=\mathcal{O}(\Delta t^{r_1})}\\ =&\mathcal{O}(\Delta t^{r_1}).\end{aligned}$$

In an analogous way we also obtain $\widehat{g}_{(0)}(y_{(0)}^{n}, z_{(0)}^{n}) = \mathcal{O}(\Delta t^{r_1})$. Then the method we consider reduces to

$$\begin{aligned}0 =&\widetilde{\boldsymbol{B}}\left[\operatorname{Diag}\{\partial_y g(\boldsymbol{t})\}\Delta\boldsymbol{y}_{(0)}^{\Delta} + \operatorname{Diag}\{\partial_z g(\boldsymbol{t})\}\Delta\boldsymbol{z}_{(0)}^{\Delta}\right]\\ &+ \widehat{\boldsymbol{B}}\left[g(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta}) - \operatorname{Diag}\{\partial_y g(\boldsymbol{t})\}\Delta\boldsymbol{y}_{(0)}^{\Delta} - \operatorname{Diag}\{\partial_z g(\boldsymbol{t})\}\Delta\boldsymbol{z}_{(0)}^{\Delta}\right]\\ &+ \mathcal{O}(\Delta t^{r_1}\boldsymbol{e}).\end{aligned}$$

In a next step, we collect all terms which occur in both the explicit and implicit part and multiply by the inverse of $\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}$, thus

$$\begin{aligned}0 =& \operatorname{Diag}\{\partial_y g(\boldsymbol{t})\}\Delta\boldsymbol{y}_{(0)}^{\Delta} + \operatorname{Diag}\{\partial_z g(\boldsymbol{t})\}\Delta\boldsymbol{z}_{(0)}^{\Delta} + \left(\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\right)^{-1}\widehat{\boldsymbol{B}}g(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta})\\ &+ \mathcal{O}(\Delta t^{r_1}\boldsymbol{e}).\end{aligned}$$

Multiplying by the inverse of $\operatorname{Diag}\left\{\partial_z g(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta})\right\}$, rearranging the terms and noting that the derivative of D is given by $-\partial_y g/\partial_z g$, see Lemma 2.18, leads to

$$\begin{aligned}\boldsymbol{z}_{(0)}^{\Delta} =&\boldsymbol{z}_{ref} - \operatorname{Diag}\left\{\frac{\partial_y g}{\partial_z g}(\boldsymbol{t})\right\}\Delta\boldsymbol{y}_{(0)}^{\Delta}\\ &- \operatorname{Diag}\{\partial_z g(\boldsymbol{t})\}^{-1}\left(\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\right)^{-1}\widehat{\boldsymbol{B}}g(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta}) + \mathcal{O}(\Delta t^{r_1}\boldsymbol{e})\\ =&\boldsymbol{z}_{ref} + \operatorname{Diag}\{D'(\boldsymbol{y}_{ref})\}\Delta\boldsymbol{y}_{(0)}^{\Delta}\\ &- \operatorname{Diag}\{\partial_z g(\boldsymbol{t})\}^{-1}\left(\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\right)^{-1}\widehat{\boldsymbol{B}}g(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta}) + \mathcal{O}(\Delta t^{r_1}\boldsymbol{e}).\end{aligned}$$

It remains to show that

$$\operatorname{Diag}\{\partial_z g(\boldsymbol{t})\}^{-1}\left(\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\right)^{-1}\widehat{\boldsymbol{B}}g(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta}) = \mathcal{O}(\Delta t^{r_1}\boldsymbol{e}).$$

This can be directly seen by mathematical induction. For this, we note that $\left(\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\right)^{-1}\widehat{\boldsymbol{B}} =: \boldsymbol{C}$ is a lower triangular matrix with zero entries on the diagonal. Then, the first equation reads

$$z_{(0)}^{n,2} =z_{ref}(t^{n,2}) + D'(y_{ref}(t^{n,2}))\Delta y_{(0)}^{n,2} + \mathcal{O}(\Delta t^{r_1})$$

which is the desired result. Furthermore, we obtain $z_{(0)}^{n,2} - z_{ref}^{n,2} = \mathcal{O}(\Delta t^{q+1})$ due to the assumptions on $\Delta y_{(0)}^{n,2}$. Next, we consider the i^{th} equation and assume that (4.13) holds for $z_{(0)}^{n,j}$ with $j < i$. The i^{th} equation reads

$$\begin{aligned} z_{(0)}^{n,i} =& z_{ref}(t^{n,i}) + D'(y_{ref}(t^{n,i}))\Delta y_{(0)}^{n,i} - \partial_z g(t^{n,i})^{-1} \sum_{j=1}^{i-1} \boldsymbol{C}_{i,j} g(y_{(0)}^{n,j}, z_{(0)}^{n,j}) \\ &+ \mathcal{O}(\Delta t^{r_1}). \end{aligned} \tag{4.14}$$

We expand g, which is evaluated at previous stages, with a Taylor expansion up to second order terms, which are then in $\mathcal{O}(\Delta t^{2(q+1)})$ since we assumed that $y_{(0)}^{n,j}$ is given with a specific accuracy. This leads, for $j < i$, to

$$g(y_{(0)}^{n,j}, z_{(0)}^{n,j}) = \partial_y g(t^{n,j})\Delta y_{(0)}^{n,j} + \partial_z g(t^{n,j})\Delta z_{(0)}^{n,j} + \mathcal{O}(\Delta t^{2(q+1)}). \tag{4.15}$$

We plug (4.13) for the previous stages in and use the definition of D'. All together, we obtain that several terms drop and we get

$$g(y_{(0)}^{n,j}, z_{(0)}^{n,j}) = \mathcal{O}(\Delta t^{r_1}) + \mathcal{O}(\Delta t^{2(q+1)}) = \mathcal{O}(\Delta t^{r_1}).$$

Finally, we plug this result in Equation (4.14) and get

$$\begin{aligned} z_{(0)}^{n,i} =& z_{ref}(t^{n,i}) + D'(y_{ref}(t^{n,i}))\Delta y_{(0)}^{n,i} - \partial_z g(t^{n,i})^{-1} \sum_{j=1}^{i-1} \boldsymbol{C}_{i,j} \mathcal{O}(\Delta t^{r_1}) + \mathcal{O}(\Delta t^{r_1}). \\ =& z_{ref}(t^{n,i}) + D'(y_{ref}(t^{n,i}))\Delta y_{(0)}^{n,i} + \mathcal{O}(\Delta t^{r_1}). \end{aligned}$$

This all together proves this Lemma. □

Corollary 4.17. *In Theorem 4.16 we assumed that $\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{ref} = \mathcal{O}(\Delta t^{q+1})$ holds. To show the result of the theorem for the i^{th}-component of $\boldsymbol{z}_{(0)}^{\Delta}$ we only need that $y_{(0)}^{n,j} - y_{ref}(t^{n,j}) = \mathcal{O}(\Delta t^{q+1})$ for $j < i$ is fulfilled.*

From the results of the previous lemma we can show the relation between $\boldsymbol{z}_{(0)}^{\Delta}$ and $D(\boldsymbol{y}_{(0)}^{\Delta})$.

Lemma 4.18. *Under the assumptions of Lemma 4.16, there holds*

$$\boldsymbol{z}_{(0)}^{\Delta} - D(\boldsymbol{y}_{(0)}^{\Delta}) = \mathcal{O}((\Delta \boldsymbol{y}_{(0)}^{\Delta})^2) + \mathcal{O}(\Delta t^{r_1}).$$

Proof. We compute a Taylor expansion up to second order terms of $D(\boldsymbol{y}_{(0)}^{\Delta})$, i.e.

$$\begin{aligned} D(\boldsymbol{y}_{(0)}^{\Delta}) &= \overbrace{D(\boldsymbol{y}_{ref})}^{=\boldsymbol{z}_{ref}} + \operatorname{Diag}\left\{D'(\boldsymbol{y}_{ref})\right\}\Delta \boldsymbol{y}_{(0)}^{\Delta} + \mathcal{O}((\Delta \boldsymbol{y}_{(0)}^{\Delta})^2) \\ \Leftrightarrow \qquad \boldsymbol{z}_{ref} &= D(\boldsymbol{y}_{(0)}^{\Delta}) - \operatorname{Diag}\left\{D'(\boldsymbol{y}_{ref})\right\}\Delta \boldsymbol{y}_{(0)}^{\Delta} + \mathcal{O}((\Delta \boldsymbol{y}_{(0)}^{\Delta})^2). \end{aligned}$$

We plug this for $\boldsymbol{z}_{ref} = D(\boldsymbol{y}_{ref})$ in Equation (4.13) of Lemma 4.16 and obtain that several terms drop. The resulting equation directly proves this lemma. □

Overall Lemmas 4.16 and 4.18 give an error estimate of $\boldsymbol{z}_{(0)}^{\Delta}$ under the assumption that $\boldsymbol{y}_{(0)}^{\Delta}$ shows a specific behavior. The next step is to show that $\boldsymbol{y}_{(0)}^{\Delta}$ fulfills the assumption made before.

The error of $\boldsymbol{y}_{(0)}^{\Delta}$

We consider the limiting method for $\boldsymbol{y}_{(0)}^{\Delta}$ and show that this method is similar, up to terms in $\mathcal{O}(\Delta t^{r_1})$, to the one given in Corollary 4.15.

Lemma 4.19. *Under the assumptions of Theorem 4.6, the numerical method for $\boldsymbol{y}^{\triangle}_{(0)}$, see Equation* (4.2) *and Corollary 4.11, equals the method given in Corollary 4.15 up to $\mathcal{O}(\Delta t^{r_1})$ terms. Furthermore, the i^{th} component of $\boldsymbol{y}^{\triangle}_{(0)}$ fulfills*

$$y^{n,i}_{(0)} - y_{ref}(t^{n,i}) = \mathcal{O}(\Delta t^{q^i+1}) + \mathcal{O}(\Delta t^{r_1}). \tag{4.16}$$

Proof. For the i^{th} internal stage we need that $y^{n,j}_{(0)}$ for $j < i$ is given with the desired accuracy. For this we again use mathematical induction, i.e. we show the results for the first stage and then for the i^{th} stage under the assumption that it is fulfilled for all previous stages.

We start by only considering the explicit part and use the results of Lemma 4.16 and Corollary 4.17 to replace $\Delta \boldsymbol{z}^{\triangle}_{(0)}$ in Equation (4.9). Next, we use a Taylor expansion in the z-component

$$\begin{aligned} f(\boldsymbol{y}^{\triangle}_{(0)}, \boldsymbol{z}^{\triangle}_{(0)}) =& f(\boldsymbol{y}^{\triangle}_{(0)}, D(\boldsymbol{y}^{\triangle}_{(0)})) + \partial_z f(\boldsymbol{y}^{\triangle}_{(0)}, D(\boldsymbol{y}^{\triangle}_{(0)})) \left(\boldsymbol{z}^{\triangle}_{(0)} - D(\boldsymbol{z}^{\triangle}_{(0)})\right) \\ &+ \mathcal{O}\left(\left(\boldsymbol{z}^{\triangle}_{(0)} - D(\boldsymbol{z}^{\triangle}_{(0)})\right)^2\right), \end{aligned}$$

the results of Lemma 4.18

$$f(\boldsymbol{y}^{\triangle}_{(0)}, \boldsymbol{z}^{\triangle}_{(0)}) = f(\boldsymbol{y}^{\triangle}_{(0)}, D(\boldsymbol{y}^{\triangle}_{(0)})) + \mathcal{O}((\Delta \boldsymbol{y}^{\triangle}_{(0)})^2) + \mathcal{O}(\Delta t^{r_1})$$

and Equation (4.16) to obtain

$$f(\boldsymbol{y}^{\triangle}_{(0)}, \boldsymbol{z}^{\triangle}_{(0)}) = f(\boldsymbol{y}^{\triangle}_{(0)}, D(\boldsymbol{y}^{\triangle}_{(0)})) + \mathcal{O}(\Delta t^{r_1}).$$

Using this plus the results of Lemma 4.18 we obtain that the method for $\boldsymbol{y}^{\triangle}_{(0)}$ is given by

$$\begin{aligned} \boldsymbol{y}^{\triangle}_{(0)} = y^{n}_{(0)} &+ \Delta t \widetilde{\boldsymbol{\alpha}} \left[f(t^n) + \partial_y f(t^n) \Delta y^n_{(0)} + \partial_z f(t^n) \operatorname{Diag}\left\{D'(t^n)\right\} \Delta y^n_{(0)}\right] \\ &+ \Delta t \widehat{\boldsymbol{\alpha}} \left[f(y^n_{(0)}, D(y^n_{(0)})) - f(t^n) - \operatorname{Diag}\left\{\partial_y f(t^n)\right\} \Delta y^n_{(0)} \right.\\ &\qquad \left. - \operatorname{Diag}\left\{\partial_z f(t^n)\right\} \operatorname{Diag}\left\{D'(t^n)\right\} \Delta y^n_{(0)}\right] \\ &+ \Delta t \widetilde{\boldsymbol{B}} \Big[f(\boldsymbol{t}) + \operatorname{Diag}\left\{\partial_y f(\boldsymbol{t})\right\} \Delta \boldsymbol{y}^{\triangle}_{(0)} \\ &\qquad + \operatorname{Diag}\left\{\partial_z f(\boldsymbol{t})\right\} \operatorname{Diag}\left\{D'(\boldsymbol{t})\right\} \Delta \boldsymbol{y}^{\triangle}_{(0)}\Big] \\ &+ \Delta t \widehat{\boldsymbol{B}} \Big[f(\boldsymbol{y}^{\triangle}_{(0)}, D(\boldsymbol{y}^{\triangle}_{(0)})) - f(\boldsymbol{t}) - \operatorname{Diag}\left\{\partial_y f(\boldsymbol{t})\right\} \Delta \boldsymbol{y}^{\triangle}_{(0)} \\ &\qquad - \operatorname{Diag}\left\{\partial_z f(\boldsymbol{t})\right\} \operatorname{Diag}\left\{D'(\boldsymbol{t})\right\} \Delta \boldsymbol{y}^{\triangle}_{(0)}\Big] + \mathcal{O}(\Delta t^{r_1+1}). \end{aligned}$$

This is the same method as given in Corollary 4.15 up to terms in $\mathcal{O}(\Delta t^{r_1+1})$ which sum up to $\mathcal{O}(\Delta t^{r_1})$ during the time iteration. Thus, we can conclude that the i^{th} internal stage fulfills

$$y^{n,i}_{(0)} - y_{ref}(t^{n,i}) = \mathcal{O}(\Delta t^{q^i+1}) + \mathcal{O}(\Delta t^{r_1}).$$

□

In the analysis for the z-component we always assumed that the y-component is given with a specific accuracy. That this assumption is reasonable is shown by the proof of Lemma 4.19 and Corollary 4.17.

Corollary 4.20. $\boldsymbol{y}^{\triangle}_{(0)}$ *fulfills* $\boldsymbol{y}^{\triangle}_{(0)} - \boldsymbol{y}_{ref} = \mathcal{O}(\Delta t^{q+1})$.

We have shown that the limiting method corresponds to the RS-IMEX discretization of the limiting equation up to terms in $\mathcal{O}(\Delta t^{r_1})$ and therefore we obtain the convergence results given in Theorem 4.14.

4.2.3. The error of $\boldsymbol{y}^{\triangle}_{(1)}$ and $\boldsymbol{z}^{\triangle}_{(1)}$

Next, we show the error behavior in $\boldsymbol{y}^{\triangle}_{(1)}$ and $\boldsymbol{z}^{\triangle}_{(1)}$. For this, the basic idea is to show that the numerical methods for $\boldsymbol{y}^{\triangle}_{(1)}$ and $\boldsymbol{z}^{\triangle}_{(1)}$ reduce to a fully implicit discretization of the corresponding equation, where some parts are evaluated exactly, and then follow the same steps as in [81]. The limiting equations for $y_{(1)}$ and $z_{(1)}$, see also Equations (2.23) and (2.20), are given by

$$\frac{d}{dt}\begin{pmatrix} y_{(1)} \\ z_{(0)} \end{pmatrix} = \begin{pmatrix} \partial_y f(y_{(0)}, z_{(0)}) y_{(1)} + \partial_z f(y_{(0)}, z_{(0)}) z_{(1)} \\ \partial_y g(y_{(0)}, z_{(0)}) y_{(1)} + \partial_z g(y_{(0)}, z_{(0)}) z_{(1)} \end{pmatrix}.$$

By applying the RS-IMEX splitting idea on this equation we get the same splitting functions as in Corollary 4.12 and obtain for the continuous case

$$\begin{aligned}\frac{d}{dt}\begin{pmatrix} y_{(1)} \\ z_{(0)} \end{pmatrix} = & \overbrace{\begin{pmatrix} \big(\partial_y f(y_{(0)}, z_{(0)}) - \partial_y f(t^n)\big) y_{(1)} + \big(\partial_z f(y_{(0)}, z_{(0)}) - \partial_z f(t^n)\big) z_{(1)} \\ \big(\partial_y g(y_{(0)}, z_{(0)}) - \partial_y g(t^n)\big) y_{(1)} + \big(\partial_z g(y_{(0)}, z_{(0)}) - \partial_z g(t^n)\big) z_{(1)} \end{pmatrix}}^{\text{explicit}} \\ & + \underbrace{\begin{pmatrix} \partial_y f(t^n) y_{(1)} + \partial_z f(t^n) z_{(1)} \\ \partial_y g(t^n) y_{(1)} + \partial_z g(t^n) z_{(1)} \end{pmatrix}}_{\text{implicit}}.\end{aligned}$$

The reference solution is given, due to Definition 3.25, by $y_{(0)}$ and $z_{(0)}$, thus for the continuous equation the explicit part sums up to zero and only the implicit part remains. Unfortunately, this is not the case for the discretization, due to numerical errors, and we need to show that the explicit part does not affect the accuracy. For this we introduce an additional notation in the following.

Definition 4.21. *A Δ in front of a function denotes the difference between this value and the corresponding exact value, e.g.*

$$\Delta \widetilde{f}^{n,i}_{(1)} := \widetilde{f}^{n,i}_{(1)} - \partial_y f(t^{n,i}) y_{(1)}(t^{n,i}) - \partial_z f(t^{n,i}) z_{(1)}(t^{n,i}).$$

We also used this abbreviation before for $\Delta y^{n,i}_{(0)}$ in Corollary 4.11.

We first show in Lemma 4.22 that if we can rewrite the numerical method for $y_{(1)}$ and $z_{(1)}$ in a specific from, we obtain a convergence result with order $c+1$ and c, respectively, with $c \in \mathbb{Z}$. Then, in Theorem 4.23 we show that we can indeed rewrite the considered method in the form needed by Lemma 4.22 and we derive the choice of c.

Lemma 4.22. *Let the numerical method of Lemma 4.10 be given in such a way that for $c \in \mathbb{N}^{\geq 1}$*

$$\Delta \boldsymbol{y}^{\Delta}_{(1)} = \Delta y^n_{(1)} \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \Delta \widetilde{f}^n_{(1)} + \widetilde{\boldsymbol{B}} \Delta \widetilde{\boldsymbol{f}}^{\Delta}_{(1)} \right) + \mathcal{O}(\Delta t^{c+1}) \tag{4.17}$$

$$\Delta \boldsymbol{z}^{\Delta}_{(0)} = \Delta z^n_{(0)} \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \Delta \widetilde{g}^n_{(1)} + \widetilde{\boldsymbol{B}} \Delta \widetilde{\boldsymbol{g}}^{\Delta}_{(1)} \right) + \mathcal{O}(\Delta t^{c+1}) \tag{4.18}$$

and furthermore

$$\Delta y^{n+1}_{(1)} := \Delta y^{n,s}_{(1)} = \Delta y^n_{(1)} + \Delta t \left(\widetilde{\boldsymbol{A}}_{s,1} \Delta \widetilde{f}^n_{(1)} + \widetilde{\boldsymbol{A}}_{s,2\ldots s} \Delta \widetilde{\boldsymbol{f}}^{\Delta}_{(1)} \right) + \mathcal{O}(\Delta t^{c+2}) \tag{4.19}$$

$$\Delta z^{n+1}_{(0)} := \Delta z^{n,s}_{(0)} = \Delta z^n_{(0)} + \Delta t \left(\widetilde{\boldsymbol{A}}_{s,1} \Delta \widetilde{g}^n_{(1)} + \widetilde{\boldsymbol{A}}_{s,2\ldots s} \Delta \widetilde{\boldsymbol{g}}^{\Delta}_{(1)} \right) + \mathcal{O}(\Delta t^{c+2}), \tag{4.20}$$

where the functions $\widetilde{f}^{n,i}_{(1)}$ and $\widetilde{g}^{n,i}_{(1)}$ are given as in Corollary 4.12. Under the assumptions of Theorem 4.6 and

$$z^m_{(0)} - z_{(0)}(t^m) = \mathcal{O}(\Delta t^{c+1})$$

for $m \leq n+1$, the errors of $y^{n+1}_{(1)}$ and $z^{n+1}_{(1)}$ are given by

$$y^{n+1}_{(1)} - y_{(1)}(t^{n+1}) = \mathcal{O}(\Delta t^{c+1}) \qquad \text{and} \qquad z^{n+1}_{(1)} - z_{(1)}(t^{n+1}) = \mathcal{O}(\Delta t^{c}).$$

Proof. In this proof we follow the same steps as in [81] rewritten to be used with the RS-IMEX splitting.

We first remark that

$$\Delta\widetilde{\boldsymbol{f}}^{\Delta}_{(1)} := \mathrm{Diag}\left\{\partial_y f(\boldsymbol{t})\right\}\Delta\boldsymbol{y}^{\Delta}_{(1)} + \mathrm{Diag}\left\{\partial_z f(\boldsymbol{t})\right\}\Delta\boldsymbol{z}^{\Delta}_{(1)} \tag{4.21}$$

$$\Delta\widetilde{\boldsymbol{g}}^{\Delta}_{(1)} := \mathrm{Diag}\left\{\partial_y g(\boldsymbol{t})\right\}\Delta\boldsymbol{y}^{\Delta}_{(1)} + \mathrm{Diag}\left\{\partial_z g(\boldsymbol{t})\right\}\Delta\boldsymbol{z}^{\Delta}_{(1)} \tag{4.22}$$

since an exact reference solution is assumed to be given. Furthermore we obtain from Equation (4.20) and the requirement that $\Delta z^{n}_{(0)} = \mathcal{O}(\Delta t^{c+1})$ and also $\Delta z^{n+1}_{(0)} = \mathcal{O}(\Delta t^{c+1})$ that

$$\Delta\widetilde{g}^{n,i}_{(1)} = \mathcal{O}(\Delta t^{c}). \tag{4.23}$$

Equation (4.22) can be rewritten in terms of $\Delta\boldsymbol{z}^{\Delta}_{(1)}$ and plugged into Equation (4.21), i.e.

$$\Delta\widetilde{f}^{n,i}_{(1)} - \frac{\partial_z f}{\partial_z g}(t^{n,i})\Delta\widetilde{g}^{n,i}_{(1)} = \left(\partial_y f(t^{n,i}) - \frac{\partial_z f}{\partial_z g}(t^{n,i})\partial_y g(t^{n,i})\right)\Delta y^{n,i}_{(1)}.$$

Next, we can replace $\Delta y^{n,i}_{(1)}$ by Equation (4.17). Then we obtain

$$\Delta\widetilde{f}^{n,i}_{(1)} - \frac{\partial_z f}{\partial_z g}(t^{n,i})\Delta\widetilde{g}^{n,i}_{(1)} = \mathcal{O}(\Delta y^{n}_{(1)}) + \mathcal{O}(\Delta t^{c+1}). \tag{4.24}$$

Note that in Equation (4.17) $\Delta f^{n,i}_{(1)}$ also occurs but with an additional power of Δt. Due to Equations (4.23) and (4.24) we can conclude that $\Delta\widetilde{f}^{n,i}_{(1)} = \mathcal{O}(\Delta y^{n}_{(1)}) + \mathcal{O}(\Delta t^{c+1})$ and therefore the additional $\Delta t\Delta\widetilde{f}^{n,i}_{(1)}$ terms are hidden in the $\mathcal{O}(\Delta y^{n}_{(1)}) + \mathcal{O}(\Delta t^{c+1})$ part of the equation. The following idea is essential in this proof. We define a new variable by

$$\Delta u^{n}_{(1)} := \Delta y^{n}_{(1)} - \frac{\partial_z f}{\partial_z g}(t^{n})\Delta z^{n}_{(0)}. \tag{4.25}$$

and in the next steps we derive the numerical method which is used to compute $\Delta u^{n+1}_{(1)}$ to show from this an error estimate for $\Delta u^{n+1}_{(1)}$. Therefore, considering (4.25) at t^{n+1} and using (4.19) we get

$$\begin{aligned}\Delta u^{n+1}_{(1)} =& \Delta y^{n}_{(1)} + \Delta t\left(\widetilde{\boldsymbol{A}}_{s,1}\Delta\widetilde{f}^{n}_{(1)} + \widetilde{\boldsymbol{A}}_{s,2\dots s}\Delta\widetilde{\boldsymbol{f}}^{\Delta}_{(1)}\right) - \frac{\partial_z f}{\partial_z g}(t^{n+1})\Delta z^{n+1}_{(0)}\\ &+ \mathcal{O}(\Delta t^{c+2}).\end{aligned}$$

We add a zero by $\left(\frac{\partial_z f}{\partial_z g}(t^{n}) - \frac{\partial_z f}{\partial_z g}(t^{n})\right)\Delta z^{n+1}_{(0)}$, rearrange the terms, replace one occurrence of $\Delta z^{n+1}_{(0)}$ by using (4.20) and make some smaller calculations to obtain

$$\begin{aligned}\Delta u^{n+1}_{(1)} =& \Delta y^{n}_{(1)} - \frac{\partial_z f}{\partial_z g}(t^{n})\Delta z^{n}_{(0)}\\ &+ \Delta t\widetilde{\boldsymbol{A}}_{s,1}\left(\Delta\widetilde{f}^{n}_{(1)} - \frac{\partial_z f}{\partial_z g}(t^{n})\Delta\widetilde{g}^{n}_{(1)}\right)\\ &+ \Delta t\widetilde{\boldsymbol{A}}_{s,2\dots s}\left(\Delta\widetilde{\boldsymbol{f}}^{\Delta}_{(1)} - \frac{\partial_z f}{\partial_z g}(t^{n})\Delta\widetilde{\boldsymbol{g}}^{\Delta}_{(1)}\right)\\ &- \left(\frac{\partial_z f}{\partial_z g}(t^{n+1}) - \frac{\partial_z f}{\partial_z g}(t^{n})\right)\Delta z^{n+1}_{(0)} + \mathcal{O}(\Delta t^{c+2}).\end{aligned}$$

Due to the requirements of this theorem we know that $\Delta z^{n+1}_{(0)} = \mathcal{O}(\Delta t^{c+1})$ and therefore we obtain

$$\left(\frac{\partial_z f}{\partial_z g}(t^{n+1}) - \frac{\partial_z f}{\partial_z g}(t^{n})\right)\Delta z^{n+1}_{(0)} = \mathcal{O}(\Delta t^{c+2}).$$

This, together with adding an additional zero by $\widetilde{\boldsymbol{A}}_{s,2\dots s}\,\mathrm{Diag}\left\{\frac{\partial_z f}{\partial_z g}(\boldsymbol{t}) - \frac{\partial_z f}{\partial_z g}(\boldsymbol{t})\right\}\Delta\widetilde{\boldsymbol{g}}^{\Delta}_{(1)}$, observing that

$\Delta\widetilde{\boldsymbol{g}}^{\Delta}_{(1)} = \mathcal{O}(\Delta t^c)$ due to Equation (4.23) and that $\frac{\partial_z f}{\partial_z g}(t^{n,i}) - \frac{\partial_z f}{\partial_z g}(t^n) = \mathcal{O}(\Delta t)$ leads to

$$\begin{aligned}\Delta u^{n+1}_{(1)} =& \Delta u^n_{(1)} + \Delta t \widetilde{\boldsymbol{A}}_{s,1}\left(\Delta \widetilde{f}^n_{(1)} - \frac{\partial_z f}{\partial_z g}(t^n)\Delta\widetilde{g}^n_{(1)}\right)\\ &+ \Delta t \widetilde{\boldsymbol{A}}_{s,2\ldots s}\left(\Delta\widetilde{\boldsymbol{f}}^{\Delta}_{(1)} - \mathrm{Diag}\left\{\frac{\partial_z f}{\partial_z g}(\boldsymbol{t})\right\}\Delta\widetilde{\boldsymbol{g}}^{\Delta}_{(1)}\right) + \mathcal{O}(\Delta t^{c+2}).\end{aligned}$$

By the results of Equation (4.24) we get

$$\Delta u^{n+1}_{(1)} = \Delta u^n_{(1)} + \Delta t\mathcal{O}(\Delta y^n_{(1)}) + \mathcal{O}(\Delta t^{c+2})$$

and due to the definition of $\Delta u^n_{(1)}$, see Equation (4.25), and $\Delta z^n_{(0)} = \mathcal{O}(\Delta t^{c+1})$ we get

$$\Delta u^{n+1}_{(1)} = (1 + C\Delta t)\Delta u^n_{(1)} + \mathcal{O}(\Delta t^{c+2}).$$

From this we can directly conclude that $\Delta u^{n+1}_{(1)} = \mathcal{O}(\Delta t^{c+1})$ and therefore

$$\Delta y^{n+1}_{(1)} = \mathcal{O}(\Delta t^{c+1}).$$

From Equation (4.24) we obtain that $\Delta\widetilde{f}^{n,i}_{(1)} = \mathcal{O}(\Delta t^{c+1})$ and from (4.17) we also get $\Delta\boldsymbol{y}^{\Delta}_{(1)} = \mathcal{O}(\Delta t^{c+1})$. Finally, due to Equation (4.22) and since $\Delta\widetilde{g}^{n,i}_{(1)} = \mathcal{O}(\Delta t^c)$ there holds

$$z^{n+1}_{(1)} - z_{(1)}(t^{n+1}) = \mathcal{O}(\Delta t^c).$$

This all together proves this theorem. □

Next, we can use this lemma to prove the convergence results of $\boldsymbol{y}^{\Delta}_{(1)}$ and $\boldsymbol{z}^{\Delta}_{(1)}$. This is done by showing that the requirements of Lemma 4.22 are fulfilled with a proper choice of the constant c.

Theorem 4.23. *Under the assumptions of Theorem 4.6, there holds*

$$\begin{pmatrix} y^{n+1}_{(1)} - y_{(1)}(t^{n+1}) \\ z^{n+1}_{(1)} - z_{(1)}(t^{n+1}) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\Delta t^{r_2+1}) \\ \mathcal{O}(\Delta t^{r_2}) \end{pmatrix} \qquad \textit{with} \qquad r_2 := \min\{r_1 - 1, \widetilde{q}, q+1\}.$$

Proof. We consider the method given in Equations (4.4) and (4.5), where the functions $\widetilde{f}^{n,i}_{(1)}$ $\widehat{f}^{n,i}_{(1)}$ $\widetilde{g}^{n,i}_{(1)}$ and $\widehat{g}^{n,i}_{(1)}$ are given by Corollary 4.12. The assumptions of Theorem 4.6 are fulfilled, therefore we can apply Lemmas 4.16 and 4.19 to obtain

$$\Delta y^{n,i}_{(0)} = \mathcal{O}(\Delta t^{q^i+1}) + \mathcal{O}(\Delta t^{r_1}) \qquad \text{and} \qquad \Delta z^{n,i}_{(0)} = \mathcal{O}(\Delta t^{q^i+1}) + \mathcal{O}(\Delta t^{r_1})$$

for $i = 2, \ldots, s$. Next, we consider the explicit part $\widehat{f}^{n,i}_{(1)}$ given in Corollary 4.12. If we take $y^{n,i}_{(1)} = y_{(1)}(t^{n,i}) + \Delta y^{n,i}_{(1)}$ and $z^{n,i}_{(1)} = z_{(1)}(t^{n,i}) + \Delta z^{n,i}_{(1)}$, we obtain

$$\begin{aligned}\widehat{f}^{n,i}_{(1)} &= \left(\partial_y f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_y f(t^{n,i})\right)\left(y_{(1)}(t^{n,i}) + \Delta y^{n,i}_{(1)}\right)\\ &\quad + \left(\partial_z f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_z f(t^{n,i})\right)\left(z_{(1)}(t^{n,i}) + \Delta z^{n,i}_{(1)}\right)\\ &= \left(\partial_y f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_y f(t^{n,i})\right) y_{(1)}(t^{n,i}) \qquad (4.26)\\ &\quad + \left(\partial_z f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_z f(t^{n,i})\right) z_{(1)}(t^{n,i})\\ &\quad + \mathcal{O}((\Delta t^{q^i+1} + \Delta t^{r_1})(|\Delta y^{n,i}_{(1)}| + |\Delta z^{n,i}_{(1)}|)).\end{aligned}$$

From Corollary 4.20 we can conclude that $\widehat{f}^{n,i}_{(1)} = \mathcal{O}(\Delta t^{q+1})$ and therefore the i^{th} internal stage of the method reads

$$\begin{aligned}
y^{n,i}_{(1)} =& y^n_{(1)} + \Delta t \left(\widetilde{\boldsymbol{\alpha}}_i \widetilde{f}^{n,1}_{(1)} + \sum_{j=2}^{i} \widetilde{\boldsymbol{A}}_{i,j} \widetilde{f}^{n,j}_{(1)} \right) + \mathcal{O}(\Delta t^{q+2}) \\
z^{n,i}_{(0)} =& z^n_{(0)} + \Delta t \left(\widetilde{\boldsymbol{\alpha}}_i \widetilde{g}^{n,1}_{(1)} + \sum_{j=2}^{i} \widetilde{\boldsymbol{A}}_{i,j} \widetilde{g}^{n,j}_{(1)} \right) + \mathcal{O}(\Delta t^{q+2}).
\end{aligned}$$

The Runge-Kutta method is a standard quadrature rule and therefore we can compute the difference to the exact value and obtain

$$\begin{aligned}
\Delta y^{n,i}_{(1)} =& \Delta y^n_{(1)} + \Delta t \left(\widetilde{\boldsymbol{\alpha}}_i \Delta \widetilde{f}^{n,1}_{(1)} + \sum_{j=2}^{i} \widetilde{\boldsymbol{A}}_{i,j} \Delta \widetilde{f}^{n,j}_{(1)} \right) + \mathcal{O}(\Delta t^{r_2+1}) \\
\Delta z^{n,i}_{(0)} =& \Delta z^n_{(0)} + \Delta t \left(\widetilde{\boldsymbol{\alpha}}_i \Delta \widetilde{g}^{n,1}_{(1)} + \sum_{j=2}^{i} \widetilde{\boldsymbol{A}}_{i,j} \Delta \widetilde{g}^{n,j}_{(1)} \right) + \mathcal{O}(\Delta t^{r_2+1}),
\end{aligned}$$

where we used $\mathcal{O}(\Delta t^{r_2+1}) = \mathcal{O}(\Delta t^{q+2}) + \mathcal{O}(\Delta t^{\widetilde{q}+1}) + \mathcal{O}(\Delta t^{r_1-1})$. For the update step, i.e. for the equations of $y^{n,s}_{(1)}$ and $z^{n,s}_{(1)}$, we want to find an estimate which is one power of Δt more accurate to apply Lemma 4.22 with $c = r_2$. Therefore, we again consider Equation (4.26) and note that the explicit part only depends on the numerical solutions $y^{n,i}_{(0)}$ and $z^{n,i}_{(0)}$ plus terms in $\mathcal{O}(\Delta t^{q^i+1} |\Delta y^{n,i}_{(1)}|)$, $\mathcal{O}(\Delta t^{q^i+1} |\Delta z^{n,i}_{(1)}|)$ and $\mathcal{O}(\Delta t^{r_1})$. These $\mathcal{O}$ terms are more accurate than the remaining part and do not affect the results anymore. $y^{n,i}_{(0)}$ and $z^{n,i}_{(0)}$ stem from the discretization of the limiting equation, see Theorem 4.14, and if we consider only the explicit part and add the numerical discretization of the limiting equation, we obtain that this is an IMEX Runge-Kutta discretization applied to the equations

$$\begin{aligned}
& y' = f(y,z) \\
& 0 = g(y,z) \\
& \delta_1' = (\partial_y f(y,z) - \partial_y f(t)) y_{(1)}(t) + (\partial_z f(y,z) - \partial_z f(t)) z_{(1)}(t) \\
& \delta_2' = (\partial_y g(y,z) - \partial_y g(t)) y_{(1)}(t) + (\partial_z g(y,z) - \partial_z g(t)) z_{(1)}(t),
\end{aligned}$$

where the exact solutions of δ_1 and δ_2 are given by zero. This is why we can conclude that in the update step the integration of the explicit part is given by

$$\Delta t \left(\widehat{\boldsymbol{\alpha}}_s \widehat{f}^{n,1}_{(1)} + \sum_{j=2}^{s-1} \widehat{\boldsymbol{A}}_{s,j} \widehat{f}^{n,s}_{(1)} \right) = \mathcal{O}(\Delta t^{r_1+1}).$$

Therefore we obtain

$$\Delta y^{n+1}_{(1)} = \Delta y^{n,s}_{(1)} = \Delta y^n_{(1)} + \Delta t \left(\widetilde{\boldsymbol{\alpha}}_s \Delta \widetilde{f}^n_{(1)} + \widetilde{\boldsymbol{A}}_{s,2\ldots s} \Delta \widetilde{\boldsymbol{f}}^{\Delta}_{(1)} \right) + \mathcal{O}(\Delta t^{r_1+1}) \tag{4.27}$$

$$\Delta z^{n+1}_{(0)} = \Delta z^{n,s}_{(0)} = \Delta z^n_{(0)} + \Delta t \left(\widetilde{\boldsymbol{\alpha}}_s \Delta \widetilde{g}^n_{(1)} + \widetilde{\boldsymbol{A}}_{s,2\ldots s} \Delta \widetilde{\boldsymbol{g}}^{\Delta}_{(1)} \right) + \mathcal{O}(\Delta t^{r_1+1}). \tag{4.28}$$

Finally, we can apply Lemma 4.22 with $c = r_2 = \min\{r_1 - 1, \widetilde{q}, q+1\}$ and the desired result is obtained. □

4.2.4. The error of $\boldsymbol{y}^{\Delta}_{(2)}$ and $\boldsymbol{z}^{\Delta}_{(2)}$

We continue with the variables $\boldsymbol{y}^{\Delta}_{(2)}$ and $\boldsymbol{z}^{\Delta}_{(2)}$ and show the error result by using the same arguments as given in Theorem 4.23 and in [81].

Theorem 4.24. *Under the assumptions of Theorem 4.6, there holds*

$$\begin{pmatrix} y^{n+1}_{(2)} - y_{(2)}(t^{n+1}) \\ z^{n+1}_{(2)} - z_{(2)}(t^{n+1}) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\Delta t^{r_2}) \\ \mathcal{O}(\Delta t^{r_2-1}) \end{pmatrix} \qquad \text{with} \qquad r_2 := \min\{r_1 - 1, q+1, \widetilde{q}\}.$$

Proof. We consider the method given in Equations (4.6) and (4.7), where the functions $\widetilde{f}_{(2)}$ $\widehat{f}_{(2)}$ $\widetilde{g}_{(2)}$ and $\widehat{g}_{(2)}$ are given by Corollary 4.13, and follow the same steps as done in Theorem 4.23 and [81].

First of all, similarly to Lemma 4.22, we consider all terms as the difference to the exact value. Then, terms that only involve variables $\boldsymbol{y}^{\triangle}_{(i)}$ and $\boldsymbol{z}^{\triangle}_{(i)}$ with $i < 2$ can directly be approximated by a Lipschitz[1] continuity argument, see [81] for more details, with the accuracy of the corresponding variable. Then, the remaining explicit part is handled analogously as done in Theorem 4.23 and we follow the same steps as in Theorem 4.23 and obtain the same result with $c = r_2 - 1$. □

Please note, that in principle Theorem 4.24 can be extended to all components of the asymptotic expansions $y_{(i)}$ and $z_{(i)}$ with $i > 2$, but one always loses an additional order of convergence. With this said, we have proven Theorem 4.6.

4.2.5. Approximate reference solution

One cannot expect that the reference solutions y_{ref} and z_{ref} are given exactly. In general they are computed with a suitable numerical method. Therefore we assume that the reference solution is computed with the same IMEX Runge-Kutta scheme and a proper splitting such that

$$y^{n,i}_{ref} - y_{(0)}(t^{n,i}) = \mathcal{O}(\Delta t^{q^i+1}) + \mathcal{O}(\Delta t^p),$$

where q^i denotes the corresponding internal order and p the overall order of the method. The less accurate reference solution may affect the overall convergence behavior of the method given in Theorem 4.6. In the following we argue why we still obtain the convergence behavior as given in Theorem 4.6, which can also be seen in a comparison in [J4].

We consider the different parts where the reference solution occur and could affect the accuracy. We start with Theorem 4.14, including all related lemmas. For every occurrence of the reference solution in a function, here as an example in f, we can add a zero and get

$$f(y^{n,i}_{ref}, z^{n,i}_{ref}) = f(y_{(0)}(t^{n,i}), z_{(0)}(t^{n,i})) + \left(f(y^{n,i}_{ref}, z^{n,i}_{ref}) - f(y_{(0)}(t^{n,i}), z_{(0)}(t^{n,i}))\right).$$

The first term is the one we assumed to be given in Theorem 4.14, the last term is handled with one part of the IMEX Runge-Kutta method and therefore integrated with the desired accuracy of the considered stage. If the reference solution occurs in a linear part, here as an example in $\partial_y f$, we add zeros and obtain

$$\begin{aligned}
&\partial_y f(y^{n,i}_{ref}, z^{n,i}_{ref}) \left(y^{n,i}_{(0)} - y^{n,i}_{ref}\right) \\
&\qquad =\partial_y f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) \left(y^{n,i}_{(0)} - y_{(0)}(t^{n,i})\right) \\
&\qquad\quad + \left(\partial_y f(y^{n,i}_{ref}, z^{n,i}_{ref}) - \partial_y f(y_{(0)}(t^{n,i}), z_{(0)}(t^{n,i}))\right) \left(y^{n,i}_{(0)} - y^{n,i}_{ref}\right) \\
&\qquad\quad + \partial_y f(y_{(0)}(t^{n,i}), z_{(0)}(t^{n,i})) \left(y_{(0)}(t^{n,i}) - y^{n,i}_{ref}\right).
\end{aligned}$$

The first term is the one we assumed to be given in Theorem 4.14, the last term is integrated and then given with the current internal order and the middle one is in $\mathcal{O}(\Delta t^{2(q+2)})$, thus in $\mathcal{O}(\Delta t^{r_1})$.

In Theorem 4.23, including all related lemmas, the reference solution only occurs in linear terms (linear in the solution) and there only in the derivative. To handle these terms, here as an example $\partial_y f$, we add a zero and obtain

$$\begin{aligned}
\partial_y f(y^{n,i}_{ref}, z^{n,i}_{ref})\Delta y^{n,i}_{(1)} =&\partial_y f(y_{(0)}(t^{n,i}), z_{(0)}(t^{n,i}))\Delta y^{n,i}_{(1)} \\
&+ \left(\partial_y f(y^{n,i}_{ref}, z^{n,i}_{ref}) - \partial_y f(y_{(0)}(t^{n,i}), z_{(0)}(t^{n,i}))\right) \Delta y^{n,i}_{(1)},
\end{aligned}$$

[1] Otto Sigismund Lipschitz, 1832 – 1903

	Ref.	Table	Type	GSA	$\widetilde{\boldsymbol{c}} = \widehat{\boldsymbol{c}}$	s	p	q	$\widetilde{q}$	r_1	r_2
ARS_222	[12]	A.1	ARS	Yes	Yes	3	2	1	1	2	1
DPA_242	[57]	A.2	A	Yes	No	4	2	1	1	–	–
ARS_443	[12]	A.3	ARS	Yes	Yes	5	3	1	1	3	1
BHR_553	[25]	A.6	CK	No	Yes	5	3	1	2	–	–
BPR_353	[26]	A.4	CK	Yes	Yes	5	3	1	2	3	2
SSP_433	[140]	A.5	A	No	No	4	3	1	1	–	–
ARK_4A2	[117]	A.7	CK	Yes	Yes	7	4	1	1	4	1

Table 4.1.: Comparison of IMEX Runge-Kutta schemes considered in this thesis. For all schemes the type, if they are globally stiffly accurate (GSA), have uniform $\boldsymbol{c}$, their number of stages s, the order of convergence p and the (implicit) stage order are given. Furthermore, for those schemes which fulfill the conditions of Theorem 4.6 the values r_1 and r_2 are given.

where the first term is the one we assumed to be given in Theorem 4.23 and the last one is given with a larger accuracy such that it does not affect the results of the theorem.

Overall we can expect that if the reference solution is computed with (parts of) the same IMEX Runge-Kutta method, we get the same convergence behavior as for an exact reference solution. In the following we compute the reference solution with the explicit part of the IMEX Runge-Kutta scheme applied to the ordinary differential equation given in (2.24) plus the update for the $z_{(0)}$ component.

4.3. Numerical experiments

In this section we consider different IMEX Runge-Kutta schemes to investigate in which way order reduction affects the convergence of the numerical solution and how the RS-IMEX splitting behaves in comparison to the standard splitting and a fully implicit discretization.

The numerical results are computed with a quad precision [167] C++ implementation of IMEX Runge-Kutta methods which uses the PETSc library [14, 15] to solve the linear system of equations. To obtain an 'exact' solution, in order to compute an error, we use the implicit part of the BPR_353 scheme on a very fine grid. As error measurement we compute

$$e_{\Delta t} := \sqrt{(y^N - y(t^{end}))^2 + (z^N - z(t^{end}))^2}.$$

Next to this, we also compute the numerical order of convergence which is for two values of Δt, Δt_1 and Δt_2, given by

$$q_{\Delta t} := \frac{\ln(e_{\Delta t_1}/e_{\Delta t_2})}{\ln(\Delta t_1/\Delta t_2)}.$$

Note that there could occur accuracy problems due to cancellation issues for very small values of Δt. In [J2, J4] similar numerical results are published, which are computed with a Matlab [122] implementation of the corresponding IMEX Runge-Kutta schemes.

In the following we first consider IMEX Runge-Kutta schemes which fulfill the conditions of Theorem 4.6 and then we consider schemes which do not have a uniform $\boldsymbol{c}$ and / or are not globally stiffly accurate. All used methods are summarized and classified in Table 4.1, including the stage orders, type, if globally stiffly accurate (GSA) and if have uniform $\boldsymbol{c}$. For those methods which fulfill all conditions of Theorem 4.6 the values r_1 and r_2 are computed.

4.3.1. IMEX Runge-Kutta schemes: globally stiffly accurate and uniform $\boldsymbol{c}$

Examples for IMEX Runge-Kutta schemes which are globally stiffly accurate, of type CK and have a uniform $\boldsymbol{c}$ are the

	Splitting	$\Delta y^N_{(0)}$	$\Delta z^N_{(0)}$	$\Delta y^N_{(1)}$	$\Delta z^N_{(1)}$	$\Delta y^N_{(2)}$	$\Delta z^N_{(2)}$
ARS_222	Standard	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$	$\mathcal{O}(\Delta t^0)$
	RS-IMEX	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$
	Implicit	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$
ARS_443	Standard	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$	$\mathcal{O}(\Delta t^0)$
	RS-IMEX	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$
	Implicit	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$
BPR_353	Standard	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$	$\mathcal{O}(\Delta t^0)$
	RS-IMEX	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$
	Implicit	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^3)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$
ARK_4A2	Standard	$\mathcal{O}(\Delta t^4)$	$\mathcal{O}(\Delta t^4)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$	$\mathcal{O}(\Delta t^0)$
	RS-IMEX	$\mathcal{O}(\Delta t^4)$	$\mathcal{O}(\Delta t^4)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$
	Implicit	$\mathcal{O}(\Delta t^4)$	$\mathcal{O}(\Delta t^4)$	$\mathcal{O}(\Delta t^2)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^1)$	$\mathcal{O}(\Delta t^0)$

Table 4.2.: Comparison of the convergence behavior of the IMEX Runge-Kutta schemes considered in this thesis which fulfill the conditions of Theorem 4.6. For all schemes the expected convergence behavior for the components of the asymptotic expansion at the final time instance are given in the case of the standard (Theorem 4.5), the RS-IMEX splitting (Theorem 4.6) and fully implicit (Theorem 4.4).

- second order ARS_222 scheme, see Table A.1,
- third order ARS_443 scheme, see Table A.3,
- third order BPR_353 scheme, see Table A.4,
- and fourth order ARK_4A2 scheme, see Table A.7.

In Table 4.2 the expected convergence behavior for each of the four schemes is given, which are shown in Corollary 4.4 for the fully implicit method, in Theorem 4.5 for the standard and in Theorem 4.6 for the RS-IMEX splitting.

Lemma 4.10 shows how the numerical method is split if we consider an asymptotic expansion of the numerical solution. We can use this to compute the convergence behavior of the corresponding numerical methods in every component $y_{(i)}$ and $z_{(i)}$ for $i = 0, 1, 2$, i.e. the limiting methods of Lemma 4.10 are implemented to obtain a numerical approximation of each component. These results are summarized in

- Figure 4.8 for the ARS_222,
- Figure 4.4 for the ARS_443,
- Figure 4.5 for the BPR_353
- and Figure 4.9 for the ARK_4A2

scheme. In every figure the dashed line denotes the optimal order of convergence p of the scheme. We first obtain that every component converges with the order of accuracy we expected in Table 4.2, which corresponds to the bounds given by Corollary 4.4, Theorem 4.5 and Theorem 4.6. Then we obtain that the implicit method and the RS-IMEX splitting behave analogously and that the standard splitting computes some components less accurately than the other two. Mainly the $y_{(1)}$ and $y_{(2)}$ component are computed with at least one order of accuracy less than by the RS-IMEX splitting. This becomes most relevant for the BPR_353 scheme since in this case the implicit part has stage order 2 and therefore $y_{(1)}$ is computed with accuracy $\mathcal{O}(\Delta t^3)$ for the implicit method and the RS-IMEX splitting, while the standard splitting computes this component with an accuracy of $\mathcal{O}(\Delta t)$.

Next, we consider the global error $e_{\Delta t}$ to investigate in which way this different convergence behavior affects the overall convergence of the numerical solution. We first obtain in Figure 4.10 that for the

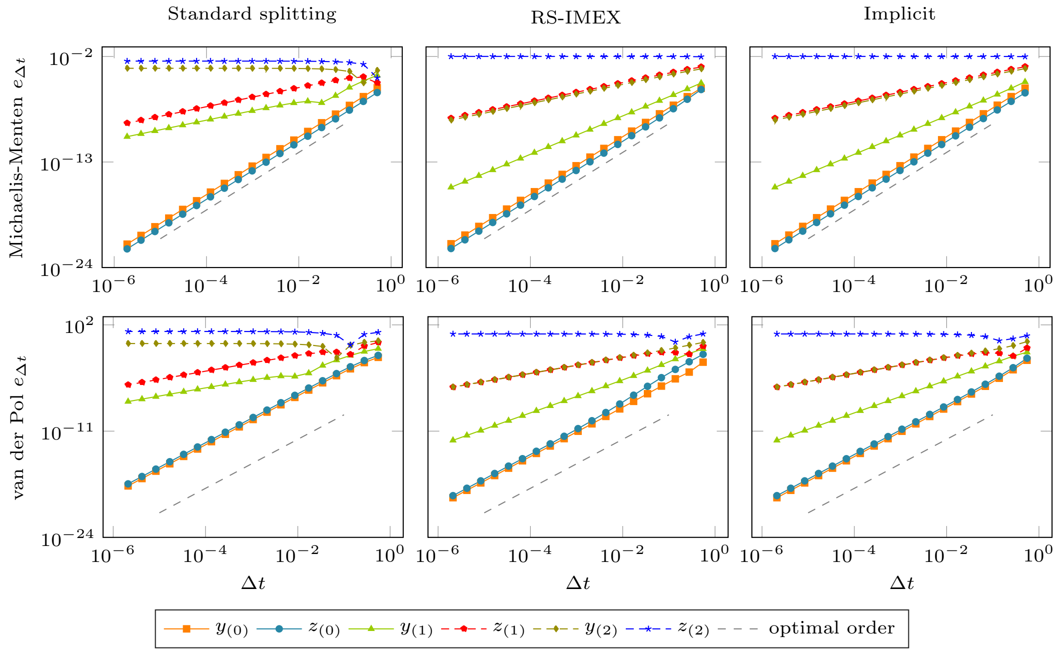


Figure 4.4.: Convergence behavior of the limiting methods given in Lemma 4.10 for the different components of the asymptotic expansion: ARS_443 scheme, see Table A.3, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation.

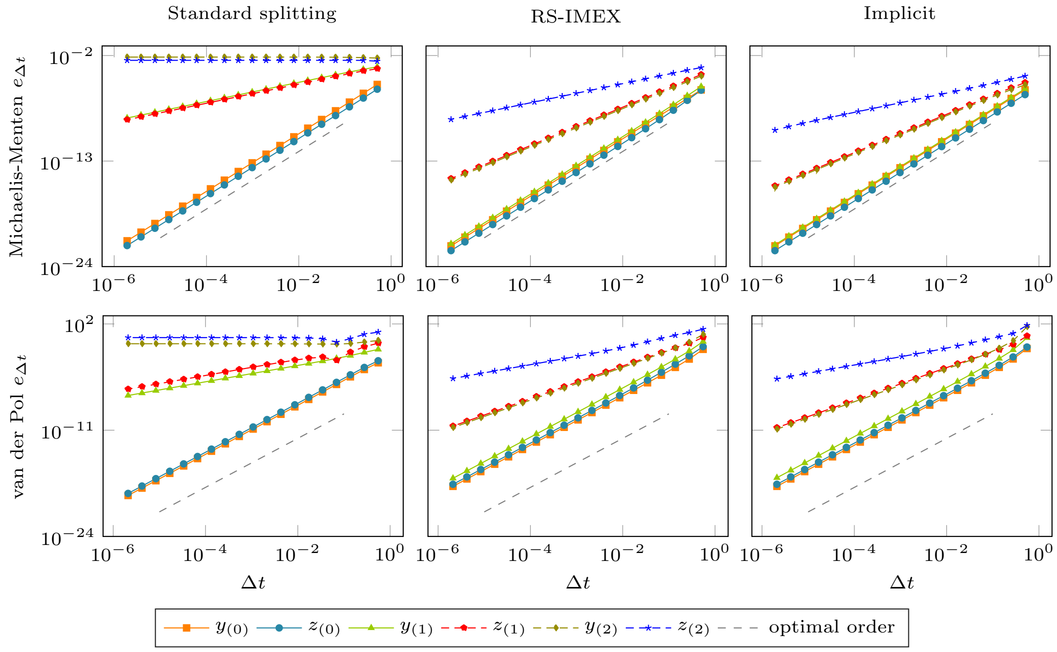


Figure 4.5.: Convergence behavior of the limiting methods given in Lemma 4.10 for the different components of the asymptotic expansion: BPR_353 scheme, see Table A.4, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation.

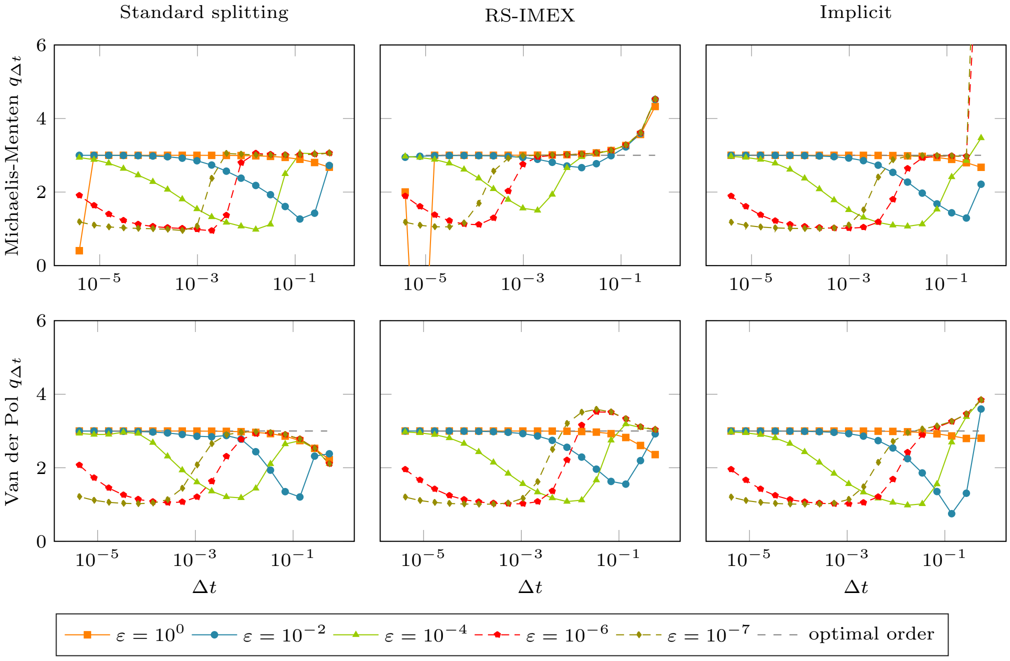


Figure 4.6.: Numerical order of convergence $q_{\Delta t}$ of the ARS_443 scheme, see Table A.4, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

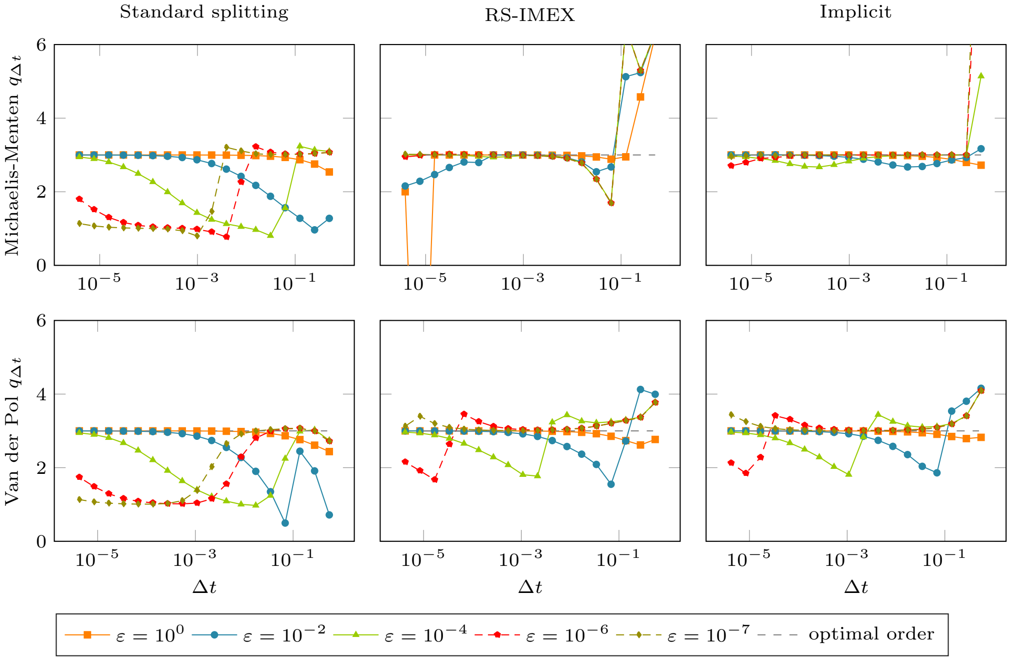


Figure 4.7.: Numerical order of convergence $q_{\Delta t}$ of the BPR_353 scheme, see Table A.4, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

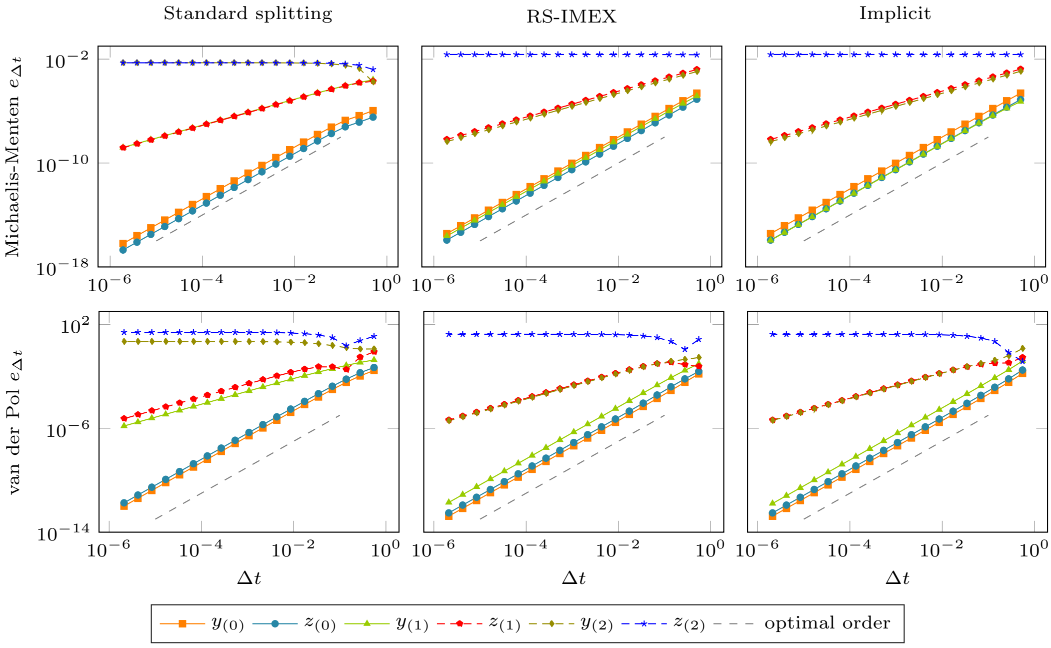


Figure 4.8.: Convergence behavior of the limiting methods given in Lemma 4.10 for the different components of the asymptotic expansion: ARS_222 scheme, see Table A.3, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation.

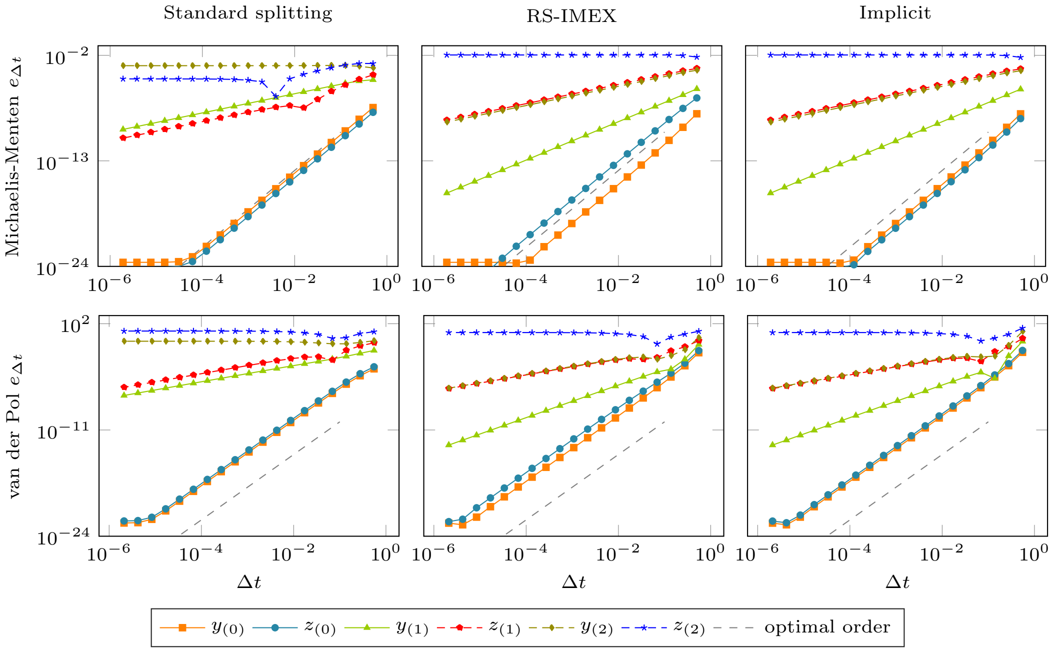


Figure 4.9.: Convergence behavior of the limiting methods given in Lemma 4.10 for the different components of the asymptotic expansion: ARK_4A2 scheme, see Table A.7, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation.

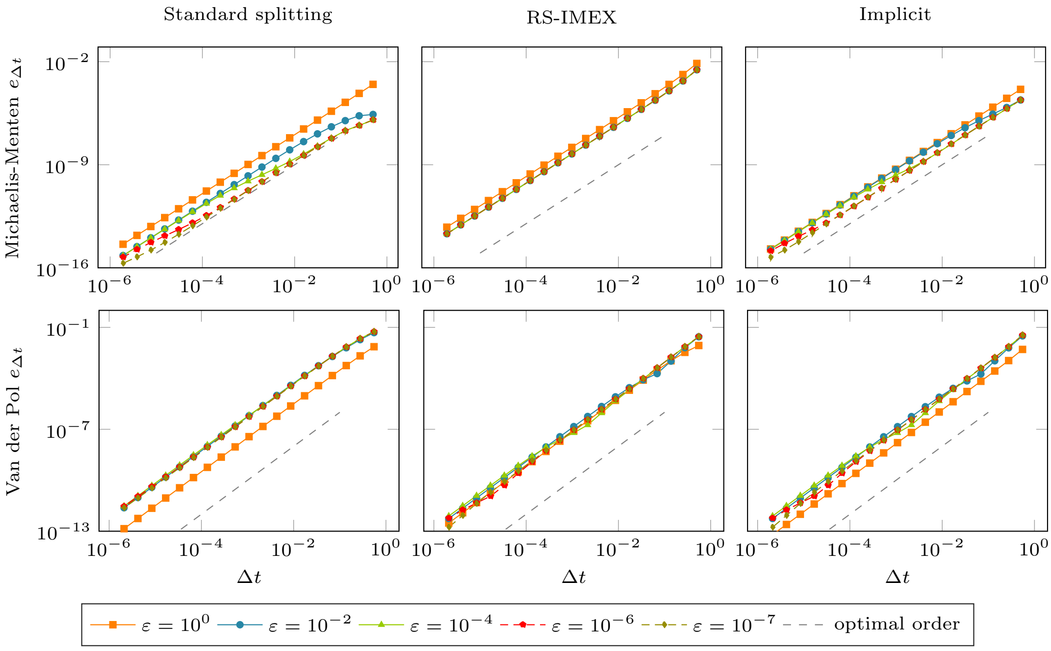


Figure 4.10.: Convergence behavior of the ARS_222 scheme, see Table A.1, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

ARS_222 scheme all methods show the desired convergence behavior and only small order reduction is visible for the RS-IMEX splitting. In Figures 4.1 and 4.11 we obtain for the ARS_443 and ARK_4A2 scheme, respectively, that all splittings show a similar behavior, i.e. order reduction is visible. This can be explained with the results in Figures 4.4 and 4.9, where we can see that the error in the $z_{(1)}$ and $z_{(2)}$ component is the most dominant and converges with the same order of accuracy for all splittings. Finally we obtain in Figure 4.2 that for the BPR_353 scheme the RS-IMEX splitting and the implicit splitting show only small order reduction compared to the standard splitting. This can be explained with the structure of the BPR_353 scheme, where the implicit part has stage order two, see Table 4.1, due to which the $z_{(1)}$ and $z_{(2)}$ component are computed with a larger order of accuracy than for the standard splitting.

The drop of convergence order can, more prominently, be seen in

- Figure 4.12 for the ARS_222,
- Figure 4.6 for the ARS_443,
- Figure 4.7 for the BPR_353
- and Figure 4.13 for the ARK_4A2 scheme.

In these figures the numerical orders of convergence for the different IMEX schemes are plotted. From these figures we also obtain that the BPR_353 scheme, see Figure 4.7, shows an order reduction for the RS-IMEX and fully implicit discretization, but this order reduction is reduced compared to the standard splitting. Similar results are given for the ARS_222 scheme in Figure 4.12 and 4.10.

4.3.2. IMEX Runge-Kutta schemes: not globally stiffly accurate and / or non-uniform c

In Theorem 4.6 we restricted ourselves to IMEX Runge-Kutta schemes which are globally stiffly accurate and have a uniform $\boldsymbol{c}$. It is interesting to see whether one can extend the results also to IMEX schemes which do not fulfill these properties.

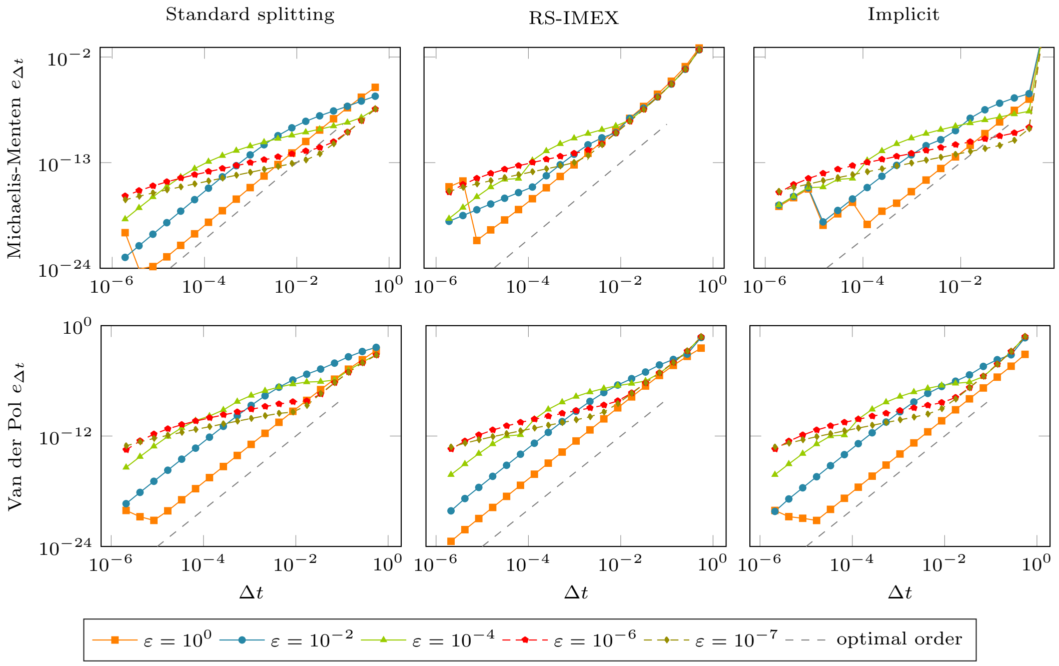


Figure 4.11.: Convergence behavior of the ARK_4A2 scheme, see Table A.7, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

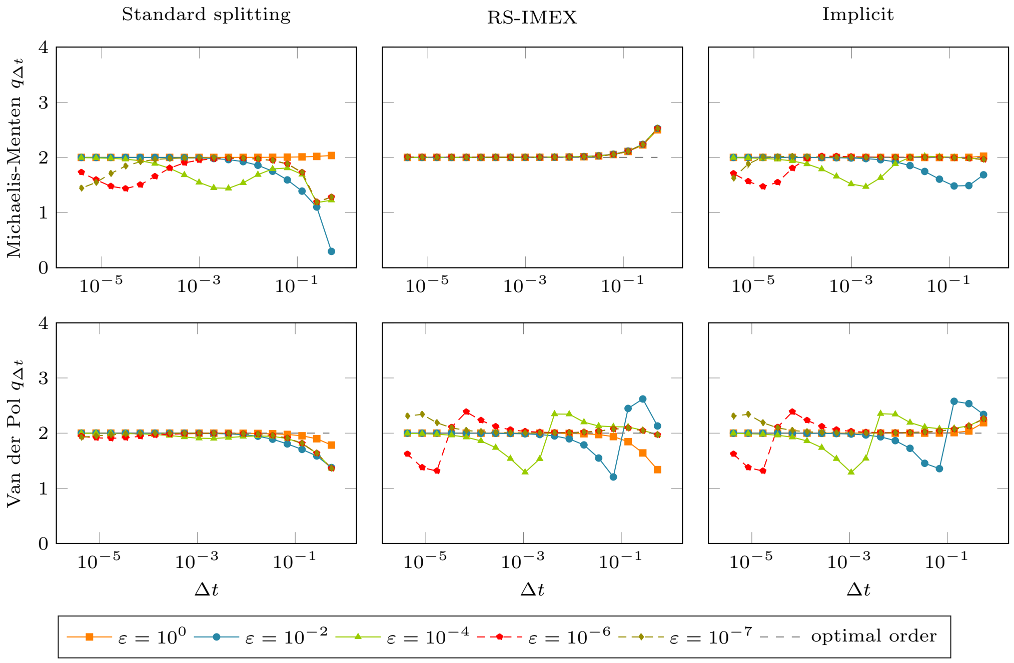


Figure 4.12.: Numerical order of convergence $q_{\Delta t}$ of the ARS_222 scheme, see Table A.1, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

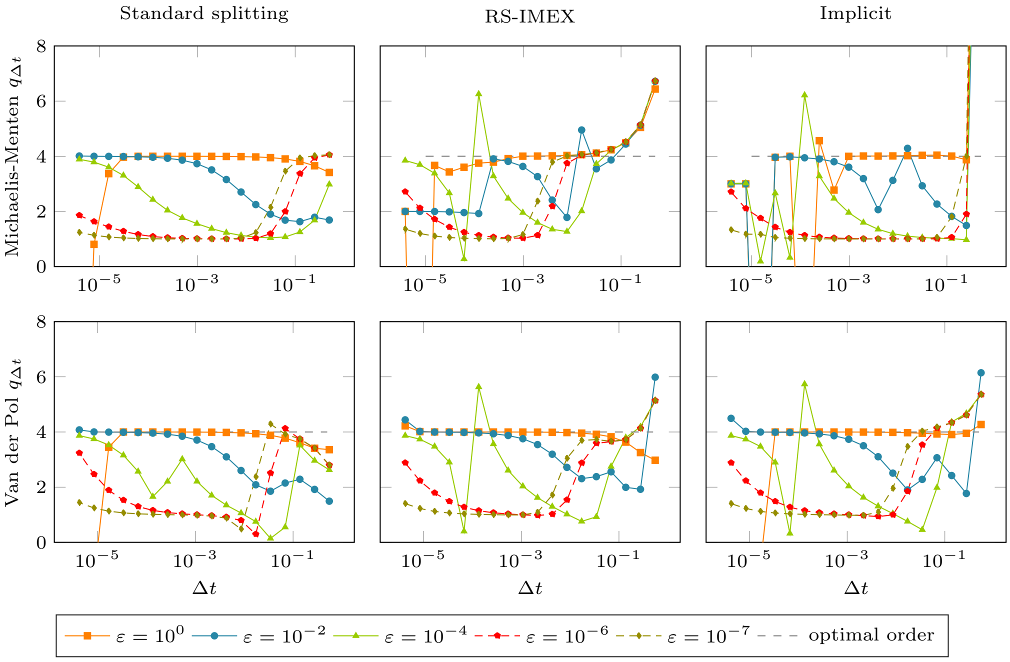


Figure 4.13.: Numerical order of convergence $q_{\Delta t}$ of the ARK_4A2 scheme, see Table A.7, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

Globally stiffly accurate and non-uniform $\boldsymbol{c}$

We consider the DPA_242 scheme, see Table A.2, which is globally stiffly accurate but does not have a uniform $\boldsymbol{c}$, see also Table 4.1. Note that the explicit part of the DPA_242 scheme has order 2 while the implicit part has order 3.

In Figure 4.14 the overall convergence behavior for different values of ε and in Figure 4.15 the numerical order of convergence are given. In all figures and for all methods the order two is plotted as a reference line. We obtain that all methods show an order reduction, but the order reduction of the RS-IMEX splitting is reduced compared to the standard splitting. Note that the $y_{(0)}$ and $z_{(0)}$ components are solved with a larger order of accuracy by the implicit scheme since the implicit part has a larger order of accuracy than the combined IMEX method.

Not globally stiffly accurate

Finally, we consider IMEX Runge-Kutta schemes which are not globally stiffly accurate. Examples are the SSP_433 and the BHR_553 scheme. The second method is designed specifically for the standard splitting such that the overall error is uniformly third order accurate, see [26].

In Figures 4.16 and 4.17 the convergence behaviors of both IMEX Runge-Kutta schemes coupled with the different splittings and for different values of ε are shown. First of all, we obtain that the RS-IMEX splitting performs similarly to the fully implicit scheme if Δt is small enough. For the BHR_553 scheme the desired convergence behavior is observed. For the SSP_433 scheme the RS-IMEX splitting and the fully implicit scheme show a slightly reduced order reduction compared to the standard splitting. This can also be seen if we consider the numerical order of convergence plotted in Figures 4.7 and 4.18. Unfortunately, Figures 4.16 and 4.17 show that for large values of Δt the RS-IMEX splitting becomes unstable for this specific IMEX Runge-Kutta scheme.

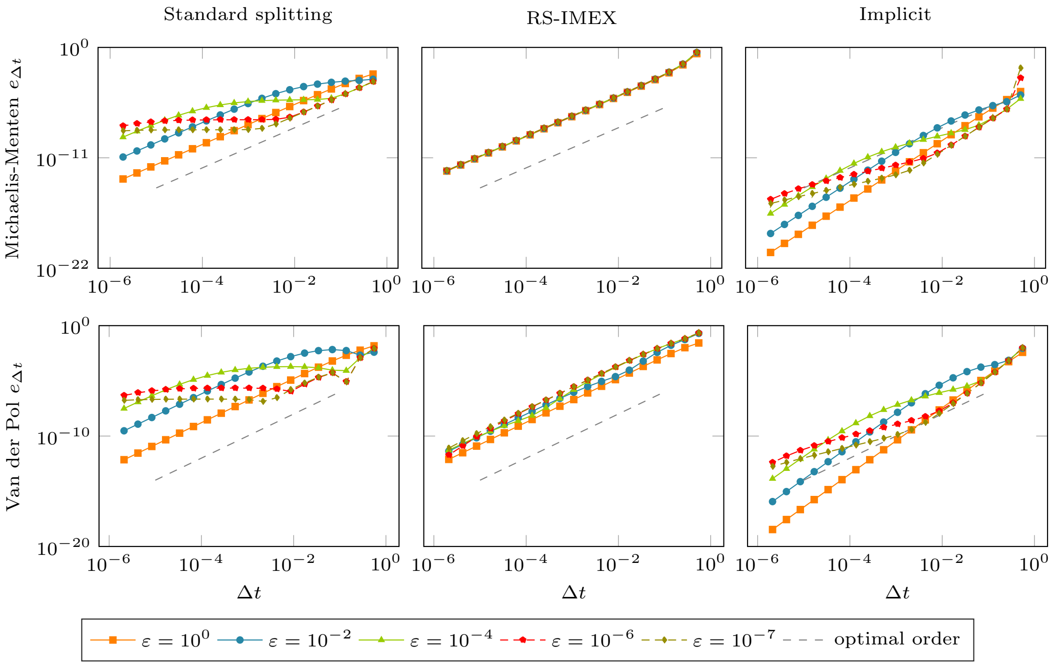


Figure 4.14.: Convergence behavior of the DPA_242 scheme, see Table A.2, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

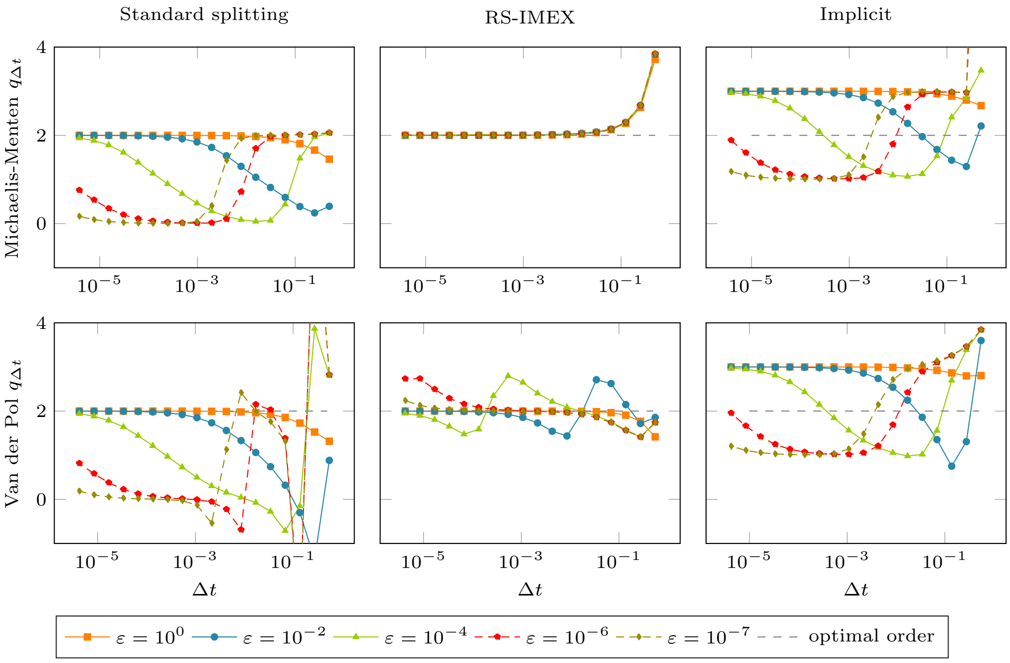


Figure 4.15.: Numerical order of convergence $q_{\Delta t}$ of the DPA_242 scheme, see Table A.2, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

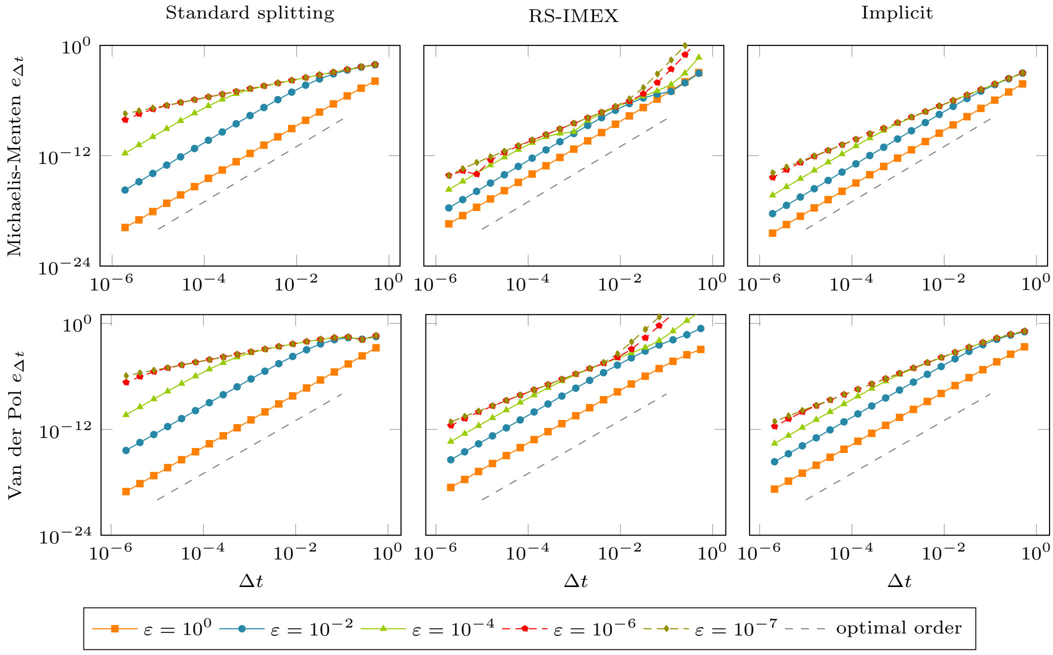


Figure 4.16.: Convergence behavior of the SSP_433 scheme, see Table A.5, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

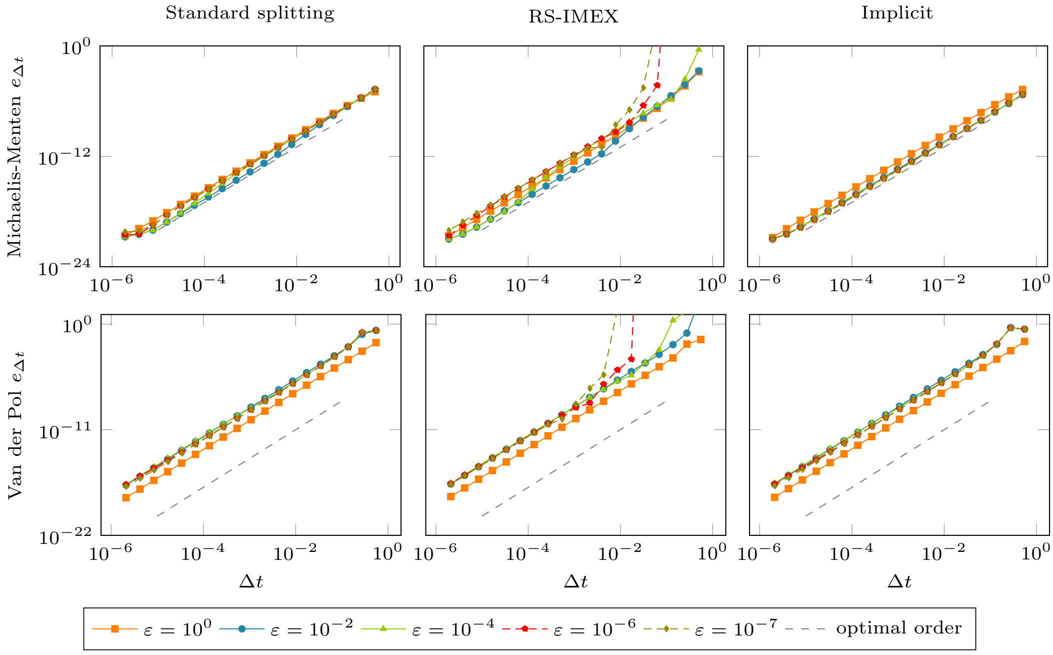


Figure 4.17.: Convergence behavior of the BHR_553 scheme, see Table A.6, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

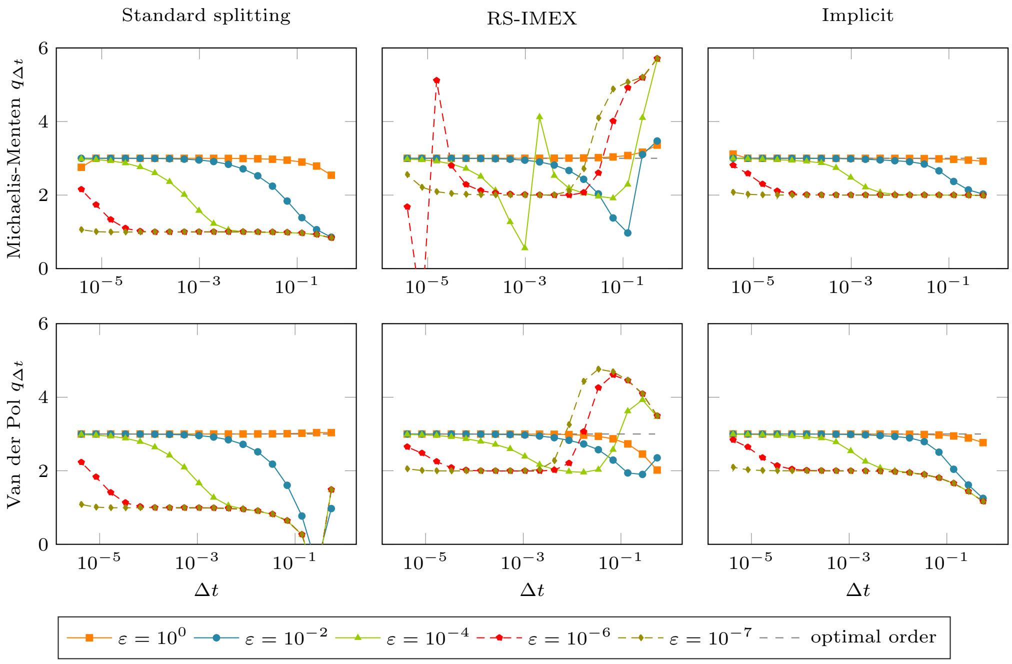


Figure 4.18.: Numerical order of convergence $q_{\Delta t}$ of the SSP_433 scheme, see Table A.5, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

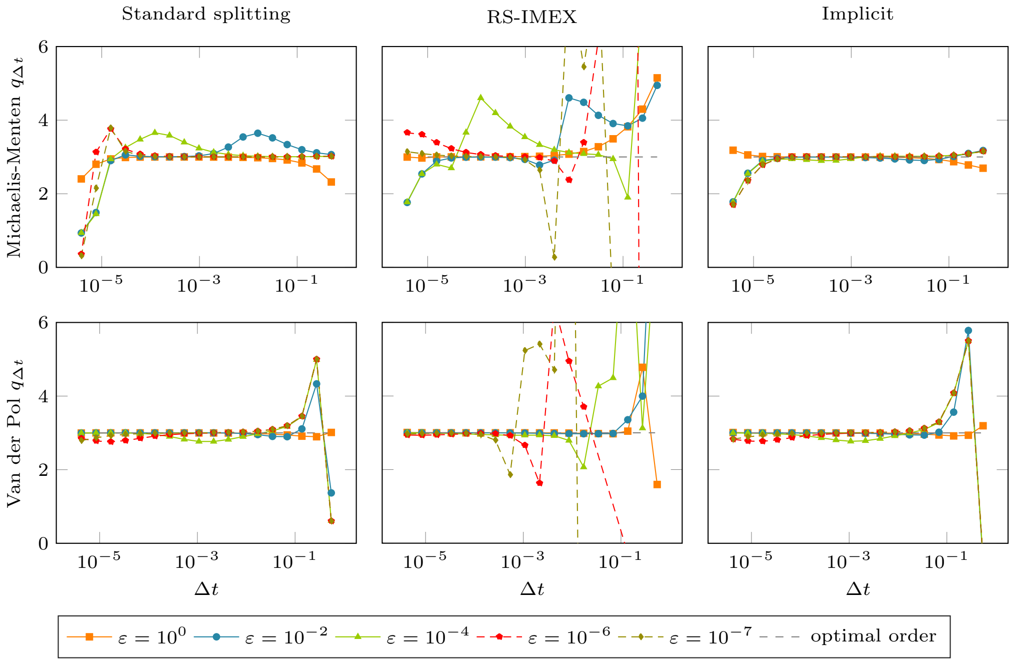


Figure 4.19.: Numerical order of convergence $q_{\Delta t}$ of the BHR_553 scheme, see Table A.6, coupled with the standard splitting (left), RS-IMEX splitting (middle) and fully implicit (right) applied to Michaelis-Menten (top, Definition 2.14) and van der Pol (bottom, Definition 2.15) equation for different values of ε.

4.4. Asymptotic properties

In Section 3.2.1 several properties which are desirable to be fulfilled for a numerical method in the setting of singularly perturbed differential equations are given. In this section we comment whether IMEX schemes coupled with the RS-IMEX splitting fulfill these properties.

In Theorem 4.14 we have shown that the limiting numerical method, i.e. the method to obtain the values $y_{(0)}^{n+1}$ and $z_{(0)}^{n+1}$, is a discretization of the corresponding limiting equation with a specific order of accuracy. As a consequence, we can conclude that the limiting method is a consistent discretization of the limiting equation and therefore that the method is *asymptotically consistent.*

Corollary 4.25. *The numerical method given by the RS-IMEX splitting coupled with a globally stiffly accurate IMEX Runge-Kutta scheme of type CK with uniform* $\mathbf{c}$ *is asymptotically consistent.*

Note that in comparison to the fully implicit method and the standard splitting, both coupled with a (globally) stiffly accurate Runge-Kutta method, the limiting algebraic equation $g(y_{(0)}, z_{(0)}) = 0$ is not fulfilled exactly. Therefore, after one time step the solution fulfills the conditions for well-prepared initial values, see Definition 2.20, only in a discrete sense.

From Theorem 4.6 and the numerical results we can also conclude that the numerical method is *asymptotically stable.* This is the case since the asymptotic expansion holds for every ε small enough and Δt large enough compared to ε but independently of ε.

Corollary 4.26. *The numerical method given by the RS-IMEX splitting coupled with a globally stiffly accurate IMEX Runge-Kutta method of type CK with uniform* $\mathbf{c}$ *is asymptotically stable.*

Theorem 4.6 shows that the *asymptotic accuracy* depends on the chosen IMEX Runge-Kutta scheme. In Section 4.3 we were able to identify several IMEX Runge-Kutta schemes where the resulting numerical method shows a less significant order reduction. These schemes are the BPR_353, DPA_242 and ARS_222 scheme. Note that we have also seen that all schemes show order reduction but in a different magnitude.

Remark 4.27. *The influence of order reduction bases both on the used temporal integration scheme (there is no order reduction for a fully implicit linear multistep method) and the used splitting (the RS-IMEX splitting shows less order reduction compared to the standard splitting). As we only focus on (IMEX) Runge-Kutta methods we investigate the influence of the splitting. In this sense the definition of asymptotic accuracy, see Definition 3.4, is a bit misleading: In general, no splitting shows the convergence behavior as for the non-stiff case if* $\varepsilon \ll \Delta t$*. Therefore, asymptotic accuracy can not be used as a comparison tool.*

Unfortunately, we have also seen, that if the IMEX Runge-Kutta scheme is not globally stiffly accurate, the resulting method is not asymptotically stable since we obtained instabilities for large values of Δt which seem to depend on the value of ε. Furthermore, we have not shown that this method is also asymptotically consistent.

4.5. Conclusion and summary

First of all we have shown in this chapter that proper IMEX Runge-Kutta schemes coupled with the RS-IMEX splitting applied to ordinary differential equations as given in Definition 2.12 do suffer from the phenomenon of order reduction. We have proven that this order reduction is similar to the one obtained by a fully implicit discretization and that it depends on the stage order of the implicit part. Furthermore, we have shown that order reduction depends on the chosen splitting, i.e. we obtained a different convergence behavior for the RS-IMEX splitting than for the standard splitting. From this we were able to identify several IMEX Runge-Kutta schemes which show a less dominant order reduction. Furthermore, we obtained that the globally stiffly accurate property is very important in this setting since it leads to asymptotically stable schemes if the RS-IMEX splitting is used. Note that the importance of the globally stiffly accurate property is also shown in [24, 81] by proving that for a method which is not GSA an additional order reduction in the $y_{(0)}$ and $z_{(0)}$ component is given.

Theorems 4.6 and 4.14 give a the glimpse that the limiting equation can only be solved with an order of accuracy of at most five since the value r_1 is given by

$$r_1 := \min\{p, 2(q+1)\},$$

which is at most five for an IMEX Runge-Kutta method. Fortunately, numerical experiments in [J2] raise the hope that the bounds derived in Theorem 4.14 are not sharp for very high order methods. On the other hand, we have seen that order reduction depends on the stage order of the implicit part and this stage order can be at most two. Therefore the $y_{(1)}$ component can be solved with an order of accuracy of at most three, scaled with ε, and the $z_{(1)}$ component with an order of accuracy of at most two, again scaled with ε.

Finally, we note that order reduction is mainly a problem of methods which use an internal stage structure like Runge-Kutta methods. As an example, linear multistep methods do not suffer from order reduction. This can be seen in [J4] for IMEX BDF schemes coupled with the RS-IMEX splitting.

5. Weakly compressible flows

The method for weakly compressible flows we propose in this thesis, see Sections 3.2 and 3.3.3, is a combination of an IMEX Runge-Kutta scheme coupled with the RS-IMEX splitting and a discontinuous Galerkin discretization:

- In Chapter 4 we have seen that some special IMEX Runge-Kutta schemes coupled with the RS-IMEX splitting show an improved order of convergence compared to a more standard splitting from literature.
- The discontinuous Galerkin method leads to a high order discretization for compressible flows if a relatively large Mach number is given.

Thus, the components itself are able to give a high order discretization strategy in their setting, but it is not clear if the combination of these methods is also useful. Therefore, this chapter is devoted to the performance of IMEX Runge-Kutta schemes coupled with the RS-IMEX splitting and a discontinuous Galerkin spatial discretization for low Mach number flows.

In Section 3.2.1 we have defined different numerical properties [22, 56, 58, 95, 96] which help to decide if a numerical method is suitable in the low Mach context. If all properties are fulfilled a stable high order method can be obtained which is consistent with the asymptotic limit as $\varepsilon \to 0$. Therefore, we check (analytically or numerically) if these properties are fulfilled by the chosen discretization. In a theoretical analysis we are able to assume that the reference solution is given exactly, for numerical computations we need to compute an approximation. Therefore, we derive a proper numerical method for the computation of the reference solution. This method is obtained by computing the $\varepsilon \to 0$ limit of the given discretization and identifying this limit with a fully implicit method. This observation is also motivated by the results of Chapter 4, where we obtained that the RS-IMEX splitting behaves similarly to a fully implicit discretization.

This chapter is organized as follows. We first show in Section 5.1 that the resulting numerical method is asymptotically consistent, i.e. that it is consistent with the $\varepsilon \to 0$ limit of the equations. For this, we start with the semi-discrete case, i.e. the case where the temporal derivatives are discretized with an IMEX Runge-Kutta scheme and the spatial derivatives are left continuous, and then we consider the fully-discrete case. Afterwards, we derive in Section 5.2 the numerical method to compute the reference solution. In Section 5.3, we use an implementation of the numerical scheme and consider the numerical examples given in Section 2.3 to investigate the stability and convergence behavior of the method. The chapter continues with a short discussion on the efficiency and accuracy of the proposed method compared to methods from literature in Section 5.4. Finally, a conclusion and summary is given in Section 5.5.

Parts of this chapter have been previously published in [J1, P1, J5]. This includes the proof of the asymptotic consistency for the fully-discrete method, which has been published in [J1, P1]. A proof of the asymptotic consistency for the semi-discrete and fully-discrete method for the special case of the RS-IMEX splitting is given in [J1]. All numerical results were recomputed, but some of them are similar to the results in [J1, P1, J5].

5.1. Asymptotic consistency

The asymptotic consistency property shows that a numerical method is consistent with the limiting behavior of the equations. In the following we prove that this property is fulfilled by the combination of an

IMEX Runge-Kutta scheme which is globally stiffly accurate with the discontinuous Galerkin method by considering a generalized splitting.

For the special case of the RS-IMEX splitting coupled with a first order finite volume method the asymptotic consistency is proven in [J3], where similarly to [78] the boundary conditions are treated in a special way, which is not extendable to high order discontinuous Galerkin methods. In [J1] the asymptotic consistency of the RS-IMEX splitting coupled with the numerical method proposed in this thesis is shown. In [P1] the results given in [J1] are generalized to a class of splittings. The last work is the one the following section is based on.

We introduce a generalized splitting and show that this splitting coupled with the numerical method is asymptotically consistent. This is done in two steps:

1. We consider the semi-discrete setting and show that an IMEX Runge-Kutta scheme which is globally stiffly accurate coupled with the generalized splitting is asymptotically consistent if we leave the spatial derivatives continuous.
2. We consider the fully-discrete case and show that an IMEX Runge-Kutta scheme which is globally stiffly accurate coupled with the generalized splitting and the discontinuous Galerkin method is asymptotically consistent.

We start with assuming that density and velocity can be represented by an asymptotic expansion such that for $n = 0, \dots, N$ and all internal stages $i = 1, \dots, s$

$$\rho^{n,i} = \rho^{n,i}_{(0)} + \varepsilon\rho^{n,i}_{(1)} + \varepsilon^2\rho^{n,i}_{(2)} + \mathcal{O}(\varepsilon^3) \qquad \text{and} \qquad \boldsymbol{u}^{n,i} = \boldsymbol{u}^{n,i}_{(0)} + \mathcal{O}(\varepsilon) \tag{5.1}$$

for the semi-discrete and

$$\begin{aligned} &\rho^{n,i}_{\Delta x} = \rho^{n,i}_{\Delta x,(0)} + \varepsilon\rho^{n,i}_{\Delta x,(1)} + \varepsilon^2\rho^{n,i}_{\Delta x,(2)} + \mathcal{O}(\varepsilon^3) \qquad \text{and} \\ &\boldsymbol{u}^{n,i}_{\Delta x} = \boldsymbol{u}^{n,i}_{\Delta x,(0)} + \mathcal{O}(\varepsilon) \end{aligned} \tag{5.2}$$

for the fully-discrete case. Next, we can formulate a generalized splitting which we consider in the following.

Definition 5.1 (Generalized splitting)**.** *Let $\mathcal{M}$, $\mathcal{H}$ and $\boldsymbol{\mathcal{K}}$ be given smooth functions, which can be written as an asymptotic expansion if ρ and $\boldsymbol{u}$ are given as an asymptotic expansion, such that*

$$\begin{aligned} \mathcal{M} =& \mathcal{M}_{(0)} + \varepsilon\mathcal{M}_{(1)} + \varepsilon^2\mathcal{M}_{(2)} + \mathcal{O}(\varepsilon^3) \\ \mathcal{H}(\rho) =& \mathcal{H}_{(0)}(\rho_{(0)}) + \varepsilon\mathcal{H}_{(1)}(\rho_{(0)}, \rho_{(1)}) + \varepsilon^2\mathcal{H}_{(2)}(\rho_{(0)}, \rho_{(1)}, \rho_{(2)}) + \mathcal{O}(\varepsilon^3) \\ \boldsymbol{\mathcal{K}}(\rho, \boldsymbol{u}) =& \boldsymbol{\mathcal{K}}_{(0)}(\rho_{(0)}, \boldsymbol{u}_{(0)}) + \mathcal{O}(\varepsilon). \end{aligned}$$

Furthermore, let the functions $\mathcal{M}$ and $\mathcal{H}$ fulfill the following conditions

1. *$\mathcal{M}$ only depends on ε such that $0 \le \mathcal{M}(\varepsilon) < 1$ for all $\varepsilon < 1$ and also $0 \le \mathcal{M}_{(0)} < 1$,*
2. *$\nabla\mathcal{H}_{(0)}(\rho_{(0)}) = \mathcal{H}^*_{(0)}(\rho_{(0)})\nabla\rho_{(0)}$ with $\mathcal{H}^*_{(0)} > 0$ if $\rho_{(0)} > 0$*
3. *and $\nabla\mathcal{H}_{(1)}(\rho_{(0)}, \rho_{(1)}) = \mathcal{H}^*_{(1)}(\rho_{(0)})\nabla\rho_{(1)}$ if $\rho_{(0)} \equiv const > 0$ with $\mathcal{H}^*_{(1)} > 0$.*

Then, we define the generalized splitting by

$$\widetilde{\boldsymbol{F}} := \begin{pmatrix} (1-\mathcal{M})\,\rho\boldsymbol{u} \\ \boldsymbol{\mathcal{K}}(\rho,\boldsymbol{u}) + \frac{1}{\varepsilon^2}\mathcal{H}(\rho)\,\mathrm{Id} \end{pmatrix} \quad \textit{and}$$

$$\widehat{\boldsymbol{F}} := \begin{pmatrix} \mathcal{M}\rho\boldsymbol{u} \\ \rho\boldsymbol{u}\otimes\boldsymbol{u} - \boldsymbol{\mathcal{K}}(\rho,\boldsymbol{u}) + \frac{1}{\varepsilon^2}\left(p(\rho) - \mathcal{H}(\rho)\right)\mathrm{Id} \end{pmatrix}.$$

Before we start to prove the asymptotic consistency, we show that the RS-IMEX splitting fulfills the conditions of Definition 5.1. Therefore we assume that in the following the reference solution is given exactly. With this, we can check if the RS-IMEX splitting fulfills all conditions of the generalized splitting.

Lemma 5.2. *The RS-IMEX splitting fulfills the conditions of a generalized splitting with*

$$\begin{aligned}
\mathcal{M} :=& 0,\\
\boldsymbol{\mathcal{K}} :=& -\rho \boldsymbol{u}_{ref} \otimes \boldsymbol{u}_{ref} + \rho \boldsymbol{u} \otimes \boldsymbol{u}_{ref} + \boldsymbol{u}_{ref} \otimes \rho \boldsymbol{u},\\
\mathcal{H} :=& p(\rho_{ref}) + p'(\rho_{ref})(\rho - \rho_{ref}).
\end{aligned}$$

Proof. The choice of $\mathcal{M}$, $\boldsymbol{\mathcal{K}}$ and $\mathcal{H}$ directly follows from the definition of the RS-IMEX splitting, see Definition 3.29, and the generalized splitting, see Definition 5.1. What remains is to check if conditions 1 to 3 of Definition 5.1 are fulfilled. The first condition is directly fulfilled since $\mathcal{M} = 0$. For the second and third condition we compute the asymptotic expansion of $\mathcal{H}$ under the assumption that ρ is given as an asymptotic expansion, i.e.

$$\begin{aligned}
\mathcal{H} =& p(\rho_{ref}) + p'(\rho_{ref})(\rho_{(0)} + \varepsilon\rho_{(1)} + \varepsilon^2\rho_{(2)} + \mathcal{O}(\varepsilon^3) - \rho_{ref})\\
=& \underbrace{p(\rho_{ref}) + p'(\rho_{ref})(\rho_{(0)} - \rho_{ref})}_{=:\mathcal{H}_{(0)}} + \varepsilon \underbrace{p'(\rho_{ref})\rho_{(1)}}_{=:\mathcal{H}_{(1)}} + \varepsilon^2 \underbrace{p'(\rho_{ref})\rho_{(2)}}_{=:\mathcal{H}_{(2)}} + \mathcal{O}(\varepsilon^3).
\end{aligned}$$

Then $\mathcal{H}^*_{(0)}$ and $\mathcal{H}^*_{(1)}$ are given by

$$\mathcal{H}^*_{(0)} = \mathcal{H}^*_{(1)} = p'(\rho_{ref}) = \gamma\rho_{ref}^{\gamma-1} > 0$$

if $\rho_{ref} > 0$, which is fulfilled since $\rho_{ref} \equiv \rho_{(0)} > 0$ and $\gamma \geq 1$. Thus, all conditions of a generalized splitting are fulfilled. □

Remark 5.3. *In [P1] it is shown that, next to the RS-IMEX splitting, also the splittings given in [78] and in [50], see also Equations* (3.24) *and* (3.23), *fulfill the conditions of Definition 5.1.*

Next, we compute the asymptotic expansion of the splitting functions $\widetilde{\boldsymbol{F}}$ and $\widehat{\boldsymbol{F}}$ of the generalized splitting, see Definition 5.1, by using a Taylor expansion to obtain terms in different powers of ε, i.e.

$$\begin{aligned}
\widetilde{\boldsymbol{F}} =& \frac{1}{\varepsilon^2}\begin{pmatrix} 0 \\ \mathcal{H}_{(0)}\,\mathrm{Id}\end{pmatrix} + \frac{1}{\varepsilon}\begin{pmatrix} 0 \\ \mathcal{H}_{(1)}\,\mathrm{Id}\end{pmatrix} + \begin{pmatrix} \left(1-\mathcal{M}_{(0)}\right)(\rho\boldsymbol{u})_{(0)} \\ \boldsymbol{\mathcal{K}}_{(0)} + \mathcal{H}_{(2)}(\rho)\,\mathrm{Id}\end{pmatrix}\\
&+ \varepsilon\begin{pmatrix} -\mathcal{M}_{(1)}(\rho\boldsymbol{u})_{(0)} + \left(1-\mathcal{M}_{(0)}\right)(\rho\boldsymbol{u})_{(1)} \\ \mathcal{O}(1)\end{pmatrix} + \mathcal{O}(\varepsilon^2)
\end{aligned} \tag{5.3}$$

for the implicit and

$$\begin{aligned}
\widehat{\boldsymbol{F}} =& \frac{1}{\varepsilon^2}\begin{pmatrix} 0 \\ \left(p_{(0)} - \mathcal{H}_{(0)}\right)\mathrm{Id}\end{pmatrix} + \frac{1}{\varepsilon}\begin{pmatrix} 0 \\ \left(p_{(1)} - \mathcal{H}_{(1)}\right)\mathrm{Id}\end{pmatrix}\\
&+ \begin{pmatrix} \mathcal{M}_{(0)}(\rho\boldsymbol{u})_{(0)} \\ (\rho\boldsymbol{u})_{(0)} \otimes \boldsymbol{u}_{(0)} - \boldsymbol{\mathcal{K}}_{(0)} + \left(p_{(2)} - \mathcal{H}_{(2)}\right)\mathrm{Id}\end{pmatrix}\\
&+ \varepsilon\begin{pmatrix} \mathcal{M}_{(1)}(\rho\boldsymbol{u})_{(0)} + \mathcal{M}_{(0)}(\rho\boldsymbol{u})_{(1)} \\ \mathcal{O}(1)\end{pmatrix} + \mathcal{O}(\varepsilon^2)
\end{aligned} \tag{5.4}$$

for the explicit part. We use this for the numerical method to derive ε-independent equations, which are solved by the components of the asymptotic expansion of all variables. This is then used to show the asymptotic consistency by proving that the formal $\varepsilon \to 0$ limit of the method is a consistent discretization of the incompressible Euler equations given in Definition 2.9. The following analysis is given for the semi-discrete and the fully-discrete case separately. Note that the general steps are similar to the steps we performed to obtain that the $\varepsilon \to 0$ limit of the isentropic Euler equations is consistent with the incompressible Euler equations.

5.1.1. Semi-discrete setting

We assume that the spatial derivatives are computed exactly and only a discretization in time is applied. The corresponding method is given in the following corollary.

Corollary 5.4. *A globally stiffly accurate IMEX Runge-Kutta method of type CK coupled with a proper splitting is given by :*

1. *Set* $\boldsymbol{w}^{n,1} = \boldsymbol{w}^n$.
2. *For* $i = 2, \dots, s$*: Seek* $\boldsymbol{w}^{n,i} : \Omega \to \mathbb{R}^3$ *such that*

$$0 = \boldsymbol{w}^{n,i} - \boldsymbol{w}^n + \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \widetilde{\boldsymbol{F}}(\boldsymbol{w}^{n,j}) + \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \widehat{\boldsymbol{F}}(\boldsymbol{w}^{n,j})$$

   holds.
3. *Set* $\boldsymbol{w}^{n+1} = \boldsymbol{w}^{n,s}$.

In the following we prove that the numerical method is asymptotically consistent. Due to the stage-wise structure of an IMEX Runge-Kutta scheme, we prove the asymptotic consistency for one internal stage of the method and then we can follow with the globally stiffly accurate property that the complete method is asymptotically consistent.

Theorem 5.5. *The* i^{th} *internal stage, with* $1 < i \leq s$*, of the IMEX Runge-Kutta method given in Corollary 5.4 coupled with a generalized splitting as given in Definition 5.1 is asymptotically consistent if we assume that* $\boldsymbol{w}^{n,j}$ *for* $j < i$ *is well-prepared in the sense of Definition 2.22. Furthermore* $\boldsymbol{w}^{n,i}$ *is also well-prepared.*

Proof. We assume that all variables are given as an asymptotic expansion. Then we can derive different equations from the discretization of the conservation of mass and momentum equation. These equations are given in Corollaries 5.7 and 5.8.

From Lemmas 5.9 and 5.10 we can conclude that $\rho_{(0)}^{n,i}$ and $\rho_{(1)}^{n,i}$ are constant in space and time. Thus, both values are given by the previous time instances. Next, from Lemma 5.11 we can conclude that $\boldsymbol{u}_{(0)}^{n,i}$ is divergence free. With this we have shown that the i^{th} internal stage fulfills the conditions of well-prepared initial values.

It remains to show that the remaining method is a consistent discretization of the conservation of momentum equation of the incompressible Euler equations. This is done in Lemma 5.12 and therefore we have shown that one stage is asymptotically consistent. □

From Theorem 5.5 we can directly conclude that the complete method is asymptotically consistent.

Corollary 5.6. *The numerical method given in Corollary 5.4 coupled with the generalized splitting given in Definition 5.1 is an asymptotically consistent discretization of the isentropic Euler equations given in Lemma 2.7.*

Proof. We have proven in Theorem 5.5 that the i^{th} internal stage is asymptotically consistent under the assumption that all previous stages are asymptotically consistent. Then, because of the globally stiffly accurate property, we can conclude that the complete method is asymptotically consistent if the initial values are well-prepared. □

ε-expansion of the IMEX Runge-Kutta method

To show that the method is asymptotically consistent we vary in terms of ε to derive different methods for the components of the asymptotic expansion. For this we insert the asymptotic expansion of all variables, see Equation (5.1), and of the flux functions, see Equations (5.3) and (5.4), order the terms in different powers of ε and obtain the different methods. This results in the methods given in Corollaries 5.7 and 5.8.

Corollary 5.7. *If we use an asymptotic expansion in every quantity, we obtain for the discretization of the conservation of mass equation given by the method in Corollary 5.4 coupled with the generalized splitting given in Definition 5.1 the following equations for the i^{th} stage of the method: For the $\mathcal{O}(1)$ terms we obtain*

$$\begin{aligned}0 =& \rho^{n,i}_{(0)} - \rho^{n}_{(0)} + \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \Big((1 - \mathcal{M}_{(0)}) (\rho \boldsymbol{u})^{n,j}_{(0)} \Big) \\ &+ \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \Big(\mathcal{M}_{(0)} (\rho \boldsymbol{u})^{n,j}_{(0)} \Big)\end{aligned}$$

and for the $\mathcal{O}(\varepsilon)$ terms we obtain

$$\begin{aligned}0 =& \Big(\rho^{n,i}_{(1)} - \rho^{n}_{(1)}, \varphi \Big)_{\mathcal{T}} + \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \Big[-\mathcal{M}_{(1)} (\rho \boldsymbol{u})^{n,j}_{(0)} + (1 - \mathcal{M}_{(0)}) (\rho \boldsymbol{u})^{n,j}_{(1)} \Big] \\ &+ \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \Big[\mathcal{M}_{(1)} (\rho \boldsymbol{u})^{n,j}_{(0)} + \mathcal{M}_{(0)} (\rho \boldsymbol{u})^{n,j}_{(1)} \Big] .\end{aligned}$$

Note that $(\rho \boldsymbol{u})^{n,j}_{(0)} = \rho^{n,j}_{(0)} \boldsymbol{u}^{n,j}_{(0)}$ and $(\rho \boldsymbol{u})^{n,j}_{(1)} = \rho^{n,j}_{(1)} \boldsymbol{u}^{n,j}_{(0)} + \rho^{n,j}_{(0)} \boldsymbol{u}^{n,j}_{(1)}$.

Corollary 5.8. *If we use an asymptotic expansion in every quantity, we obtain for the discretization of the conservation of momentum equation given by the method in Corollary 5.4 coupled with the generalized splitting given in Definition 5.1 the following different equations for the i^{th} stage of the method: For the $\mathcal{O}(\varepsilon^{-2})$ terms we obtain*

$$0 = \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \mathcal{H}^{n,j}_{(0)} + \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \Big(p^{n,j}_{(0)} - \mathcal{H}^{n,j}_{(0)} \Big) ,$$

for the $\mathcal{O}(\varepsilon^{-1})$ terms we obtain

$$0 = \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \mathcal{H}^{n,j}_{(1)} + \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \Big(p^{n,j}_{(1)} - \mathcal{H}^{n,j}_{(1)} \Big) ,$$

and for the $\mathcal{O}(1)$ terms we obtain

$$\begin{aligned}0 =& (\rho \boldsymbol{u})^{n,i}_{(0)} - (\rho \boldsymbol{u})^{n}_{(0)} + \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \Big[\boldsymbol{\mathcal{K}}^{n,j}_{(0)} + \mathcal{H}^{n,j}_{(2)} \,\mathrm{Id} \Big] \\ &+ \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \Big((\rho \boldsymbol{u})^{n,j}_{(0)} \otimes \boldsymbol{u}^{n,j}_{(0)} - \boldsymbol{\mathcal{K}}^{n,j}_{(0)} + p^{n,j}_{(2)} \,\mathrm{Id} - \mathcal{H}^{n,j}_{(2)} \,\mathrm{Id} \Big) .\end{aligned}$$

Note that $(\rho \boldsymbol{u})^{n,j}_{(0)} = \rho^{n,j}_{(0)} \boldsymbol{u}^{n,j}_{(0)}$ and that we used the abbreviations

$$\begin{aligned}&\mathcal{H}^{n,j}_{(0)} = \mathcal{H}_{(0)}(\rho^{n,j}_{(0)}), \qquad \mathcal{H}^{n,j}_{(1)} = \mathcal{H}_{(1)}(\rho^{n,j}_{(0)}, \rho^{n,j}_{(1)}), \qquad \mathcal{H}^{n,j}_{(2)} = \mathcal{H}_{(2)}(\rho^{n,j}_{(0)}, \rho^{n,j}_{(1)}, \rho^{n,j}_{(2)}) \\ &\qquad \textit{and} \qquad \boldsymbol{\mathcal{K}}^{n,j}_{(0)} = \boldsymbol{\mathcal{K}}_{(0)}(\rho^{n,j}_{(0)}, (\rho \boldsymbol{u})^{n,j}_{(0)}).\end{aligned}$$

$\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ constant in space and time

Similarly to the continuous equations in Lemma 2.26 we first show that the limiting densities $\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ are constant in space by considering the $\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-1})$ terms of the discretization of the pressure gradient. Then, by considering the conservation of mass discretization we show that $\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ are also constant in time.

Lemma 5.9. *Under the assumptions of Theorem 5.5, $\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ are constant in space.*

Proof. We give the proof for $\rho^{n,i}_{(0)}$ in the following, the one for $\rho^{n,i}_{(1)}$ is analogous. We assumed that all previous time instances are well-prepared, i.e. $\rho_{(0)}^{n,j}$ is constant in space for $j < i$. Thus, for these values the gradient equals zero and we obtain for the $\mathcal{O}(\varepsilon^{-2})$ terms of the conservation of momentum discretization given in Corollary 5.8 that

$$0 = \Delta t \widetilde{\boldsymbol{A}}_{i,i} \nabla_{\boldsymbol{x}} \mathcal{H}_{(0)} \left(\rho_{(0)}^{n,i}\right) = \Delta t \widetilde{\boldsymbol{A}}_{i,i} \underbrace{\mathcal{H}'_{(0)} \left(\rho_{(0)}^{n,i}\right)}_{=\mathcal{H}^*_{(0)}} \nabla_{\boldsymbol{x}} \rho_{(0)}^{n,i},$$

where $\widetilde{\boldsymbol{A}}_{i,i} \neq 0$. From the definition of a generalized splitting we know that $\mathcal{H}^*_{(0)} > 0$ holds and therefore the equation can only hold if $\nabla_{\boldsymbol{x}} \rho_{(0)}^{n,i} \equiv 0$ and consequently $\rho_{(0)}^{n,i}$ is constant in space. □

Lemma 5.10. *Under the assumptions of Theorem 5.5, $\rho_{(0)}^{n,i}$ and $\rho_{(1)}^{n,i}$ are constant in time.*

Proof. Similarly to Lemma 5.9 we prove this for $\rho_{(0)}^{n,i}$ and note that the proof for $\rho_{(1)}^{n,i}$ is analogous. We follow similar steps as for the continuous case, see Corollary 2.24. Therefore, we consider the $\mathcal{O}(1)$ terms of the conservation of mass discretization, see Corollary 5.7, and integrate over the whole domain. Then with Lemma 5.9 and integration by parts for the convective term we obtain

$$\begin{aligned} 0 = &\left(\rho_{(0)}^{n,i} - \rho_{(0)}^{n}\right) \int_{\Omega} 1 \mathrm{dx} + \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \int_{\partial\Omega} (1 - \mathcal{M}_{(0)}) (\rho \boldsymbol{u})_{(0)}^{n,j} \cdot \boldsymbol{n} \mathrm{d}\sigma \\ &+ \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \int_{\partial\Omega} \mathcal{M}_{(0)} (\rho \boldsymbol{u})_{(0)}^{n,j} \cdot \boldsymbol{n} \mathrm{d}\sigma. \end{aligned}$$

The boundary integrals equal zero due to the chosen boundary conditions, see Definition 2.23, or if a periodic boundary is given. Thus

$$0 = \left(\rho_{(0)}^{n,i} - \rho_{(0)}^{n}\right) \int_{\Omega} 1 \mathrm{dx} = \left(\rho_{(0)}^{n,i} - \rho_{(0)}^{n}\right) |\Omega|$$

and therefore $\rho_{(0)}^{n,i}$ is constant in space and time and equals to the corresponding initial values. We can show the same result for $\rho_{(1)}^{n,i}$ by considering the $\mathcal{O}(\varepsilon)$ terms given in Corollary 5.19. □

The previous lemmas prove that the limiting densities $\rho_{(0)}^{n,i}$ and $\rho_{(1)}^{n,i}$ are constant in space and time. Therefore, they are given by the corresponding initial values, i.e.

$$\rho_{(0)}^{n,i} = \rho_{(0)} \qquad \text{and} \qquad \rho_{(1)}^{n,i} = \rho_{(1)}.$$

We use this representation in the following.

Divergence free constraint for $\boldsymbol{u}_{(0)}$

In the semi-discrete setting we left the spatial derivatives continuous and only used the temporal discretization method. Due to this, we are able to show that the divergence free equation for $\boldsymbol{u}_{(0)}$ is fulfilled exactly. This is done in the following lemma.

Lemma 5.11. *Under the assumptions of Theorem 5.5, $\boldsymbol{u}_{(0)}^{n,i}$ fulfills $\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}_{(0)}^{n,i} = 0$.*

Proof. We consider the $\mathcal{O}(1)$ terms of the conservation of mass discretization given in Corollary 5.7 and use the results of Lemma 5.10 to obtain

$$0 = \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \left((1 - \mathcal{M}_{(0)}) \boldsymbol{u}_{(0)}^{n,j}\right) + \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \left(\mathcal{M}_{(0)} \boldsymbol{u}_{(0)}^{n,j}\right).$$

Assuming that all values $\boldsymbol{u}_{(0)}^{n,j}$ for $j < i$ are divergence free we can conclude

$$0 = \nabla_{\boldsymbol{x}} \cdot \left((1 - \mathcal{M}_{(0)}) \boldsymbol{u}_{(0)}^{n,j}\right) \Rightarrow \nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}_{(0)}^{n,j} = 0.$$

The last step can be done since $\mathcal{M}_{(0)} \neq 1$ and does not depend on $\boldsymbol{x}$. □

From Lemmas 5.10, 5.9 and 5.11, we can conclude that the solution of the i^{th} internal stage fulfills the conditions of well-prepared initial values, see Definition 2.22.

Limiting method

It remains to show that the limiting method, which results from the $\mathcal{O}(1)$ terms of the conservation of momentum discretization, together with the divergence free constraint, is a consistent discretization of the incompressible Euler equations given in Definition 2.9.

Lemma 5.12. *Under the assumptions of Theorem 5.5, the $\mathcal{O}(1)$ terms of the conservation of momentum discretization are a consistent discretization of*

$$\partial_t \boldsymbol{u}_{(0)} + \nabla_{\boldsymbol{x}} \cdot \left(\boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{(0)} + \frac{p_{(2)}}{\rho_{(0)}} \operatorname{Id} \right) = 0. \tag{5.5}$$

Proof. We consider the $\mathcal{O}(1)$ terms of the discretization of the conservation of momentum equation as given in Corollary 5.20 and obtain together with Lemmas 5.22 and 5.23

$$\begin{aligned} 0 = \left(\boldsymbol{u}_{(0)}^{n,i} - \boldsymbol{u}_{(0)}^{n}, \varphi \right)_{\mathcal{T}} &+ \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \left(\frac{\boldsymbol{\mathcal{K}}_{(0)}^{n,j}}{\rho_{(0)}} + \frac{\mathcal{H}_{(2)}^{n,j}}{\rho_{(0)}} \operatorname{Id} \right) \\ &+ \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \nabla_{\boldsymbol{x}} \cdot \left(\boldsymbol{u}_{(0)}^{n,j} \otimes \boldsymbol{u}_{(0)}^{n,j} - \frac{\boldsymbol{\mathcal{K}}_{(0)}^{n,j}}{\rho_{(0)}} + \frac{p_{(2)}^{n,j}}{\rho_{(0)}} \operatorname{Id} - \frac{\mathcal{H}_{(2)}^{n,j}}{\rho_{(0)}} \operatorname{Id} \right) \end{aligned}$$

where we used the same abbreviations as in Corollary 5.8. Overall, we obtain a consistent discretization of Equation (5.5) where a splitting technique is used such that the flux function is split into the implicit contribution

$$\frac{\boldsymbol{\mathcal{K}}_{(0)}(\rho_{(0)}, \boldsymbol{u}_{(0)})}{\rho_{(0)}} + \frac{\mathcal{H}_{(2)}(\rho_{(0)}, \rho_{(1)}, \rho_{(2)})}{\rho_{(0)}} \operatorname{Id}$$

and the explicit contribution

$$\boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{(0)} - \frac{\boldsymbol{\mathcal{K}}_{(0)}(\rho_{(0)}, \boldsymbol{u}_{(0)})}{\rho_{(0)}} + \frac{p_{(2)}(\rho_{(0)}, \rho_{(1)}, \rho_{(2)})}{\rho_{(0)}} \operatorname{Id} - \frac{\mathcal{H}_{(2)}(\rho_{(0)}, \rho_{(1)}, \rho_{(2)})}{\rho_{(0)}} \operatorname{Id}. \tag{5.6}$$

□

5.1.2. Fully-discrete setting

We now prove that the discontinuous Galerkin method given in Definition 3.14 with an IMEX Runge-Kutta scheme given in Definition 3.9 coupled with the generalized splitting given in Definition 5.1 is asymptotically consistent. We restrict ourselves to periodic boundary conditions, but we comment on the non-periodic case at the end of this section. We follow the same steps as done for the semi-discrete method up to some differences. The main difference is that, due to the weak formulation the discontinuous Galerkin method is based on, we cannot directly conclude that the limiting densities are constant in space. Furthermore, we cannot show that the divergence free constraint for the limiting velocity is fulfilled exactly, we can show this only in a discrete sense. To make the following analysis better readable, we hide unneeded terms in an additional (non-standard) notation which is given in the following definition.

Definition 5.13 (Notation)**.** *For the discontinuous Galerkin method given in Definition 3.14, we define the following abbreviations:*

- $(\boldsymbol{a}, \boldsymbol{b})_{\mathcal{T}} := \sum_{k=1}^{\text{ne}} \int_{\Omega_k} \boldsymbol{a} \cdot \boldsymbol{b} \mathrm{dx}$,
- $\{\boldsymbol{a}, b\}_{\partial\mathcal{T}} := \sum_{k=1}^{\text{ne}} \int_{\partial\Omega_k} \left(\boldsymbol{a}^- + \boldsymbol{a}^+ \right) b^- \cdot \boldsymbol{n} \mathrm{d}\sigma$

- *and* $[\![\boldsymbol{a}, b]\!]_{\partial\mathcal{T}} := \sum_{k=1}^{\mathrm{ne}} \int_{\partial\Omega_k} \left(\boldsymbol{a}^- - \boldsymbol{a}^+\right) b^- \mathrm{d}\sigma.$

From this, we can rewrite the discontinuous Galerkin method given in Definition 3.14 coupled with a globally stiffly accurate IMEX Runge-Kutta scheme of type CK, see Definition 3.9.

Corollary 5.14. *The discontinuous Galerkin method given in Definition 3.14, coupled with a globally stiffly accurate IMEX Runge-Kutta method of type CK, see Definition 3.9, numerical flux functions as given in Definition 3.15 and periodic boundary conditions is given by the following steps:*

1. *Set* $\boldsymbol{w}_{\Delta x}^{n,1} = \boldsymbol{w}_{\Delta x}^{n}$.
2. *For* $i = 2, \dots, s$*: Seek* $\boldsymbol{w}_{\Delta x}^{n,i} \in V_{\Delta x}^3$ *such that every equation of*

$$
\begin{aligned}
0 = &\left(\boldsymbol{w}_{\Delta x}^{n,i} - \boldsymbol{w}_{\Delta x}^{n}, \varphi\right)_{\mathcal{T}} - \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left(\widetilde{\boldsymbol{F}}(\boldsymbol{w}_{\Delta x}^{n,j}), \nabla\varphi\right)_{\mathcal{T}} \\
&+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left(\left\{ \widetilde{\boldsymbol{F}}(\boldsymbol{w}_{\Delta x}^{n,j}), \varphi \right\}_{\partial\mathcal{T}} + \mathrm{Diag}\left\{ \left(\varepsilon^{-2}, 1, 1\right)^T \right\} \left[\!\!\left[\boldsymbol{w}_{\Delta x}^{n,j}, \varphi \right]\!\!\right]_{\partial\mathcal{T}} \right) \\
&- \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left(\widehat{\boldsymbol{F}}(\boldsymbol{w}_{\Delta x}^{n,j}), \nabla\varphi\right)_{\mathcal{T}} \\
&+ \frac{\Delta t}{2} \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left(\left\{ \widehat{\boldsymbol{F}}(\boldsymbol{w}_{\Delta x}^{n,j}), \varphi \right\}_{\partial\mathcal{T}} + \varepsilon \left[\!\!\left[\boldsymbol{w}_{\Delta x}^{n,j}, \varphi \right]\!\!\right]_{\partial\mathcal{T}} \right)
\end{aligned}
$$

   holds for every $\varphi \in V_{\Delta x}$.
3. *Set* $\boldsymbol{w}_{\Delta x}^{n+1} = \boldsymbol{w}_{\Delta x}^{n,s}$.

One of the main ingredients of the discontinuous Galerkin method is, that the numerical approximation is allowed to be discontinuous over the cell boundaries, but the limiting densities $\rho_{(0)}^{n,i}$ and $\rho_{(1)}^{n,i}$ should be constant and therefore continuous. To overcome this issue we introduced the special numerical stabilization in Definition 3.15. To show that this stabilization leads to continuous limiting densities we need the following lemma.

Lemma 5.15. *Let* $\sigma \in V_{\Delta x}$ *be given in such a way that*

$$
[\![\sigma, \varphi]\!]_{\partial\mathcal{T}} = 0 \tag{5.7}
$$

holds for all $\varphi \in V_{\Delta x}$. *Then* σ *is continuous over* Ω.

Proof. $\sigma \in V_{\Delta x}$ is continuous in the interior of every cell Ω_k for $k = 1, \dots, \mathrm{ne}$. Therefore we only need to prove that σ is continuous over the cell boundaries. We consider (5.7), insert the definition of $[\![\cdot, \cdot]\!]_{\partial\mathcal{T}}$ and since the equation holds for every $\varphi \in V_{\Delta x}$, we can choose $\varphi = \sigma$. This results in

$$
0 = [\![\sigma, \sigma]\!]_{\partial\mathcal{T}} = \sum_{k=1}^{\mathrm{ne}} \int_{\partial\Omega_k} \left(\sigma^- - \sigma^+\right) \sigma^- \mathrm{d}\sigma.
$$

The sum considers every cell intersection twice with switched roles of ±. Therefore, we can rearrange the terms and sum over all edges $e \in \mathcal{T}^I$ and obtain

$$
0 = \sum_{e \in \partial\mathcal{T}^I} \int_{\partial\Omega_k} \left(\sigma^- - \sigma^+\right) \sigma^- \mathrm{d}\sigma + \sum_{e \in \partial\mathcal{T}^I} \int_{\partial\Omega_k} \left(\sigma^+ - \sigma^-\right) \sigma^+ \mathrm{d}\sigma.
$$

Note that we switched from the cell boundary formulation, where ± corresponds to the cell normal vector, see Equation (3.11), to the edge formulation, where ± corresponds to the normal vector in reference direction, see Equation (3.13). Next, we rearrange the terms and obtain

$$
0 = -\sum_{e \in \partial\mathcal{T}^I} \int_{\partial\Omega_k} \left(\sigma^- - \sigma^+\right)^2 \mathrm{d}\sigma.
$$

This equation can only be fulfilled if $\sigma^+ = \sigma^-$ and therefore if σ is continuous over every edge of the inner skeleton $\partial\mathcal{T}^I$. Thus, σ is continuous in Ω. □

In the following, we prove that the numerical method is asymptotically consistent. Due to the stage-wise structure of an IMEX Runge-Kutta scheme, we prove the asymptotic consistency for one internal stage of the method and then we can follow with the globally stiffly accurate property that the complete method is asymptotically consistent. For this, we show that well-prepared initial conditions are preserved during the iteration process. We consider a numerical discretization and therefore the divergence free constraint on $\boldsymbol{u}_{(0)}$ cannot be fulfilled exactly. Therefore we modify the definition of well-prepared initial conditions, see Definition 2.22, for the discrete case.

Definition 5.16 (Well-prepared initial conditions (discrete))**.** *We call the initial conditions for the discretization given in Corollary 5.14 well-prepared if they fulfill the conditions of Definition 2.22 in a discrete sense, i.e.* $\rho^0_{\Delta x} = \underbrace{const}_{>0} + \mathcal{O}(\varepsilon^2)$ *and* $\boldsymbol{u}^0_{\Delta x}$ *stems from a discretization of* $\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}^0 = \mathcal{O}(\varepsilon)$.

Theorem 5.17. *The* i^{th} *internal stage, with* $1 < i \leq s$*, of the method given in Corollary 5.14 coupled with the generalized splitting as given in Definition 5.1 is asymptotically consistent if we assume that* $\boldsymbol{w}^{n,j}_{\Delta x}$ *for* $j < i$ *is well-prepared in the sense of Definition 5.16. Furthermore* $\boldsymbol{w}^{n,i}_{\Delta x}$ *is also well-prepared. Note that for* $1 < i \leq s$ *there holds* $\widetilde{\boldsymbol{A}}_{i,i} \neq 0$.

Proof. We assume that all variables are given as an asymptotic expansion. Then we can derive different equations from the discretization of the conservation of mass and momentum equations. These equations are given in Corollaries 5.19 and 5.20.

Then,

- from Lemma 5.21 we obtain that $\rho^{n,i}_{\Delta x,(0)}$ and $\rho^{n,i}_{\Delta x,(1)}$ are continuous over the whole domain,
- from Lemma 5.22 we obtain that $\rho^{n,i}_{\Delta x,(0)}$ and $\rho^{n,i}_{\Delta x,(1)}$ are constant in space and
- from Lemma 5.23 we obtain that $\rho^{n,i}_{\Delta x,(0)}$ and $\rho^{n,i}_{\Delta x,(1)}$ are constant in time.

Thus, both values are given by the initial conditions.

We cannot show that the divergence free constraint is fulfilled exactly, but from Lemma 5.24 we can conclude that the method for $\boldsymbol{u}_{(0)}$ is consistent to the divergence free equation. Then together with Lemma 5.25 we can show that the limiting method is a consistent discretization of the incompressible equations. Thus, we have shown that one stage of the given method is asymptotically consistent. □

From Theorem 5.17 we can directly conclude that the complete method is asymptotically consistent.

Corollary 5.18. *The IMEX DG method given in Corollary 5.14 coupled with the generalized splitting, see Definition 5.1, is asymptotically consistent.*

Proof. We have proven in Theorem 5.17 that the i^{th} internal stage is asymptotically consistent under the assumption that all previous stages are asymptotically consistent. Then, because of the globally stiffly accurate property, we can conclude that the complete method is asymptotically consistent if the initial values are well-prepared. □

ε-expansion of the IMEX Runge-Kutta discontinuous Galerkin method

Similarly to the semi-discrete case, we consider the numerical method given in Corollary 5.14, insert the asymptotic expansion of every component and then derive different methods for the components of the asymptotic expansion by varying ε.

Corollary 5.19. *We consider the conservation of mass discretization obtained by the method given in Corollary 5.14 coupled with the generalized splitting given in Definition 5.1. If we use an asymptotic expansion in every quantity, we obtain the following different equations for the i^{th} internal stage of the method: For the $\mathcal{O}(\varepsilon^{-2})$ and for the $\mathcal{O}(\varepsilon^{-1})$ terms we obtain*

$$0 = \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left\llbracket \rho^{n,j}_{\Delta x,(0)}, \varphi \right\rrbracket_{\partial \mathcal{T}} \qquad \text{and} \qquad 0 = \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left\llbracket \rho^{n,j}_{\Delta x,(1)}, \varphi \right\rrbracket_{\partial \mathcal{T}},$$

respectively, for the $\mathcal{O}(1)$ terms we obtain

$$\begin{aligned}
0 = &\left(\rho^{n,i}_{\Delta x,(0)} - \rho^{n}_{\Delta x,(0)}, \varphi\right)_{\mathcal{T}} - \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left((1 - \mathcal{M}_{(0)})(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)}, \nabla \varphi\right)_{\mathcal{T}} \\
&+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[\left\{(1 - \mathcal{M}_{(0)})(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)}, \varphi\right\}_{\partial \mathcal{T}} + \left\llbracket \rho^{n,j}_{\Delta x,(2)}, \varphi \right\rrbracket_{\partial \mathcal{T}}\right] \\
&- \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left(\mathcal{M}_{(0)}(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)}, \nabla \varphi\right)_{\mathcal{T}} + \frac{\Delta t}{2} \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left\{\mathcal{M}_{(0)}(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)}, \varphi\right\}_{\partial \mathcal{T}}
\end{aligned}$$

and for the $\mathcal{O}(\varepsilon)$ terms we obtain

$$\begin{aligned}
0 = &\left(\rho^{n,i}_{\Delta x,(1)} - \rho^{n}_{\Delta x,(1)}, \varphi\right)_{\mathcal{T}} \\
&- \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[- \left(\mathcal{M}_{(1)}(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)}, \nabla \varphi\right)_{\mathcal{T}} + \left((1 - \mathcal{M}_{(0)})(\rho \boldsymbol{u})^{n,j}_{\Delta x,(1)}, \nabla \varphi\right)_{\mathcal{T}}\right] \\
&- \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left[\left(\mathcal{M}_{(1)}(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)}, \nabla \varphi\right)_{\mathcal{T}} + \left(\mathcal{M}_{(0)}(\rho \boldsymbol{u})^{n,j}_{\Delta x,(1)}, \nabla \varphi\right)_{\mathcal{T}}\right] \\
&+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[- \left\{\mathcal{M}_{(1)}(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)}, \varphi\right\}_{\partial \mathcal{T}} + \left\{(1 - \mathcal{M}_{(0)})(\rho \boldsymbol{u})^{n,j}_{\Delta x,(1)}, \varphi\right\}_{\partial \mathcal{T}}\right] \\
&+ \frac{\Delta t}{2} \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left[\left\{\mathcal{M}_{(1)}(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)}, \varphi\right\}_{\partial \mathcal{T}} + \left\{\mathcal{M}_{(0)}(\rho \boldsymbol{u})^{n,j}_{\Delta x,(1)}, \varphi\right\}_{\partial \mathcal{T}}\right] \\
&+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left\llbracket \rho^{n,j}_{\Delta x,(3)}, \varphi \right\rrbracket_{\partial \mathcal{T}} + \frac{\Delta t}{2} \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left\llbracket \rho^{n,j}_{\Delta x,(0)}, \varphi \right\rrbracket_{\partial \mathcal{T}}.
\end{aligned}$$

Note that $(\rho \boldsymbol{u})^{n,j}_{\Delta x,(0)} = \rho^{n,j}_{\Delta x,(0)} \boldsymbol{u}^{n,j}_{\Delta x,(0)}$ and $(\rho \boldsymbol{u})^{n,j}_{\Delta x,(1)} = \rho^{n,j}_{\Delta x,(1)} \boldsymbol{u}^{n,j}_{\Delta x,(0)} + \rho^{n,j}_{\Delta x,(0)} \boldsymbol{u}^{n,j}_{\Delta x,(1)}$.

Corollary 5.20. *We consider the conservation of momentum discretization obtained by the method given in Corollary 5.14 coupled with the generalized splitting given in Definition 5.1. If we use an asymptotic expansion in every quantity, we obtain the following different equations for the i^{th} internal stage of the method: For the $\mathcal{O}(\varepsilon^{-2})$ terms we obtain*

$$\begin{aligned}
0 = &\sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[\left(\mathcal{H}^{n,j}_{(0)} \operatorname{Id}, \nabla \varphi\right)_{\mathcal{T}} - \frac{1}{2} \left\{\mathcal{H}^{n,j}_{(0)} \operatorname{Id}, \varphi\right\}_{\partial \mathcal{T}}\right] \\
&+ \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left[\left(\left(p^{n,j}_{\Delta x,(0)} - \mathcal{H}^{n,j}_{(0)}\right) \operatorname{Id}, \nabla \varphi\right)_{\mathcal{T}} - \frac{1}{2} \left\{\left(p^{n,j}_{\Delta x,(0)} - \mathcal{H}^{n,j}_{(0)}\right) \operatorname{Id}, \varphi\right\}_{\partial \mathcal{T}}\right],
\end{aligned}$$

for the $\mathcal{O}(\varepsilon^{-1})$ terms we obtain

$$\begin{aligned}
0 = &\sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[\left(\mathcal{H}^{n,j}_{(1)} \operatorname{Id}, \nabla \varphi\right)_{\mathcal{T}} - \frac{1}{2} \left\{\mathcal{H}^{n,j}_{(1)} \operatorname{Id}, \varphi\right\}_{\partial \mathcal{T}}\right] \\
&+ \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left[\left(p^{n,j}_{\Delta x,(1)} \operatorname{Id} - \mathcal{H}^{n,j}_{(1)} \operatorname{Id}, \nabla \varphi\right)_{\mathcal{T}} - \frac{1}{2} \left\{p^{n,j}_{\Delta x,(1)} \operatorname{Id} - \mathcal{H}^{n,j}_{(1)} \operatorname{Id}, \varphi\right\}_{\partial \mathcal{T}}\right],
\end{aligned}$$

and for the $\mathcal{O}(1)$ terms we obtain

$$\begin{aligned}
0 = &\left((\rho\boldsymbol{u})^{n,i}_{\Delta x,(0)} - (\rho\boldsymbol{u})^{n}_{\Delta x,(0)}, \varphi\right)_{\mathcal{T}} - \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[\left(\boldsymbol{\mathcal{K}}^{n,j}_{(0)} + \mathcal{H}^{n,j}_{(2)} \operatorname{Id}, \nabla\varphi\right)_{\mathcal{T}}\right] \\
&+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[\left\{\boldsymbol{\mathcal{K}}^{n,j}_{(0)} + \mathcal{H}^{n,j}_{(2)} \operatorname{Id}, \varphi\right\}_{\partial\mathcal{T}} + \left[\!\left[(\rho\boldsymbol{u})^{n,j}_{\Delta x,(0)}, \varphi\right]\!\right]_{\partial\mathcal{T}}\right] \\
&- \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left((\rho\boldsymbol{u})^{n,j}_{\Delta x,(0)} \otimes \boldsymbol{u}^{n,j}_{\Delta x,(0)} - \boldsymbol{\mathcal{K}}^{n,j}_{(0)} + p^{n,j}_{\Delta x,(2)} \operatorname{Id} - \mathcal{H}^{n,j}_{(2)} \operatorname{Id}, \nabla\varphi\right)_{\mathcal{T}} \\
&+ \frac{\Delta t}{2} \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left\{(\rho\boldsymbol{u})^{n,j}_{\Delta x,(0)} \otimes \boldsymbol{u}^{n,j}_{\Delta x,(0)} - \boldsymbol{\mathcal{K}}^{n,j}_{(0)} + p^{n,j}_{\Delta x,(2)} \operatorname{Id} - \mathcal{H}^{n,j}_{(2)} \operatorname{Id}, \varphi\right\}_{\partial\mathcal{T}}.
\end{aligned}$$

Note that $(\rho\boldsymbol{u})^{n,j}_{(0)} = \rho^{n,j}_{(0)}\boldsymbol{u}^{n,j}_{(0)}$ and that we used the abbreviations

$$\begin{aligned}
&\mathcal{H}^{n,j}_{(0)} = \mathcal{H}_{(0)}(\rho^{n,j}_{\Delta x,(0)}), \qquad \mathcal{H}^{n,j}_{(1)} = \mathcal{H}_{(1)}(\rho^{n,j}_{\Delta x,(0)}, \rho^{n,j}_{\Delta x,(1)}), \\
&\mathcal{H}^{n,j}_{(2)} = \mathcal{H}_{(2)}(\rho^{n,j}_{\Delta x,(0)}, \rho^{n,j}_{\Delta x,(1)}, \rho^{n,j}_{\Delta x,(2)}) \qquad \textit{and} \\
&\boldsymbol{\mathcal{K}}^{n,j}_{(0)} = \boldsymbol{\mathcal{K}}_{(0)}(\rho^{n,j}_{\Delta x,(0)}, (\rho\boldsymbol{u})^{n,j}_{\Delta x,(0)}).
\end{aligned}$$

$\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ constant in space and time

We show that the limiting densities are constant in space and time. This is done in three steps. First we show in Lemma 5.21 that the limiting densities are continuous over the whole domain, then in a second step we show in Lemma 5.22 that the limiting densities are constant in space and finally in Lemma 5.23 we show that the limiting densities are also constant in time.

Lemma 5.21. *Under the assumptions of Theorem 5.17, $\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ are continuous.*

Proof. In the conservation of mass equation the implicit numerical stabilization in $\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ is the only term in $\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-1})$, respectively, see Corollary 5.19 for the corresponding equations. $\rho^{n,j}_{(0)}$ and $\rho^{n,j}_{(1)}$ are constant in space and time since we assumed that they are well-prepared for $j < i$. Therefore we obtain

$$0 = \frac{\Delta t}{2}\widetilde{\boldsymbol{A}}_{i,i}\left[\!\left[\rho^{n,i}_{\Delta x,(0)}, \varphi\right]\!\right]_{\partial\mathcal{T}} \qquad \text{and} \qquad 0 = \frac{\Delta t}{2}\widetilde{\boldsymbol{A}}_{i,i}\left[\!\left[\rho^{n,i}_{\Delta x,(1)}, \varphi\right]\!\right]_{\partial\mathcal{T}}.$$

Next, since $\widetilde{\boldsymbol{A}}_{i,i} \neq 0$, we can apply Lemma 5.15 and directly obtain that $\rho^{n,i}_{\Delta x,(0)}$ and $\rho^{n,i}_{\Delta x,(1)}$ are continuous over the whole domain. □

Lemma 5.22. *Under the assumptions of Theorem 5.17, $\rho^{n,i}_{\Delta x,(0)}$ and $\rho^{n,i}_{\Delta x,(1)}$ are constant in space.*

Proof. We show that $\rho^{n,i}_{\Delta x,(0)}$ is constant in space, for $\rho^{n,i}_{\Delta x,(1)}$ the prove is analogous. We consider the $\mathcal{O}(\varepsilon^{-2})$ terms of the conservation of momentum equation, see Corollary 5.20, and use integration by parts on both the implicit and explicit part. For this we use the modified notation

$$[\![(\cdot)\boldsymbol{n}_k, \varphi]\!]_{\partial\mathcal{T}} := \sum_{k=1}^{\text{ne}} \int_{\partial\Omega_k} \left((\cdot)^- - (\cdot)^+\right) \cdot \boldsymbol{n}_k \varphi^- \mathrm{d}\sigma$$

and then we get

$$\begin{aligned}
0 = &\sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left(\nabla\mathcal{H}_{(0)}\left(\rho^{n,j}_{\Delta x,(0)}\right), \varphi\right)_{\mathcal{T}} \\
&+ \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left(\nabla p\left(\rho^{n,j}_{\Delta x,(0)}\right) - \nabla\mathcal{H}_{(0)}\left(\rho^{n,j}_{\Delta x,(0)}\right), \varphi\right)_{\mathcal{T}}
\end{aligned}$$

$$+\frac{1}{2}\sum_{j=1}^{i}\widetilde{\boldsymbol{A}}_{i,j}\left[\!\left[\mathcal{H}_{(0)}\left(\rho^{n,j}_{\Delta x,(0)}\right)\boldsymbol{n}_k,\varphi\right]\!\right]_{\partial\mathcal{T}}$$
$$+\frac{1}{2}\sum_{j=1}^{i-1}\widehat{\boldsymbol{A}}_{i,j}\left[\!\left[p\left(\rho^{n,j}_{\Delta x,(0)}\right)\boldsymbol{n}_k-\mathcal{H}_{(0)}\left(\rho^{n,j}_{\Delta x,(0)}\right)\boldsymbol{n}_k,\varphi\right]\!\right]_{\partial\mathcal{T}}.$$

We assumed that all previous time instances are well-prepared, i.e. $\rho^{n,j}_{\Delta x,(0)}$ is constant in space for $j<i$. Thus, for these values the gradient and the jump over all cell-boundaries equal zero. We therefore obtain

$$0=\left(\nabla\mathcal{H}_{(0)}\left(\rho^{n,i}_{\Delta x,(0)}\right),\varphi\right)_{\mathcal{T}}+\frac{1}{2}\left[\!\left[\mathcal{H}_{(0)}\left(\rho^{n,i}_{\Delta x,(0)}\right)\boldsymbol{n}_k,\varphi\right]\!\right]_{\partial\mathcal{T}}.$$

$\rho^{n,i}_{\Delta x,(0)}$ is continuous, see Lemma 5.21, and then the boundary terms drop. Thus

$$0=\left(\nabla\mathcal{H}_{(0)}\left(\rho^{n,i}_{\Delta x,(0)}\right),\varphi\right)_{\mathcal{T}}=\left(\mathcal{H}'_{(0)}\left(\rho^{n,i}_{\Delta x,(0)}\right)\nabla\rho^{n,i}_{\Delta x,(0)},\varphi\right)_{\mathcal{T}}.$$

We can now use a similar argument as in Lemma 5.15 by considering the d^{th} equation with $d=1,2$, choosing $\varphi=\partial_{\boldsymbol{x}_d}\rho^{n,i}_{\Delta x,(0)}$ and using the definition of $(\cdot,\cdot)_{\mathcal{T}}$. Then the equation reads

$$\begin{aligned}0=&\left(\mathcal{H}'_{(0)}\left(\rho^{n,i}_{\Delta x,(0)}\right)\partial_{\boldsymbol{x}_d}\rho^{n,i}_{\Delta x,(0)},\partial_{\boldsymbol{x}_d}\rho^{n,i}_{\Delta x,(0)}\right)_{\mathcal{T}}\\=&\sum_{k=1}^{\text{ne}}\int_{\Omega_k}\mathcal{H}'_{(0)}\left(\rho^{n,i}_{\Delta x,(0)}\right)\left(\partial_{\boldsymbol{x}_d}\rho^{n,i}_{\Delta x,(0)}\right)^2\text{dx}.\end{aligned}$$

Next, $\mathcal{H}'_{(0)}\left(\rho^{n,i}_{\Delta x,(0)}\right)=\mathcal{H}^*_{(0)}$ and from the conditions of a generalized splitting we know that $\mathcal{H}^*_{(0)}>0$ if $\rho^{n,i}_{\Delta x,(0)}>0$. We can assume that $\rho^{n,i}_{\Delta x,(0)}>0$ since otherwise we would obtain a negative density and then an unstable method. Therefore, the equation can only be solved if $\nabla_{\boldsymbol{x}}\rho^{n,i}_{\Delta x,(0)}\equiv 0$ and consequently $\rho^{n,i}_{\Delta x,(0)}$ is constant in space. □

We have shown in Lemma 5.22 that the limiting densities $\rho^{n,i}_{\Delta x,(0)}$ and $\rho^{n,i}_{\Delta x,(1)}$ are constant in space. Therefore, these values only depend on time and we use the notation

$$\rho^{n,i}_{(0)}:=\rho^{n,i}_{\Delta x,(0)}\qquad\text{and}\qquad\rho^{n,i}_{(1)}:=\rho^{n,i}_{\Delta x,(1)}$$

in the following.

Lemma 5.23. *Under the assumptions of Theorem 5.17, $\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ are constant in time.*

Proof. Similarly to Lemma 5.22 we show this for $\rho^{n,i}_{(0)}$ and note that the proof for $\rho^{n,i}_{(1)}$ is analogous. We consider the $\mathcal{O}(1)$ terms of the conservation of mass discretization, see Corollary 5.19. With Lemma 5.22 and by choosing $\varphi\equiv 1$, we obtain

$$\begin{aligned}0=&\left(\rho^{n,i}_{(0)}-\rho^{n}_{(0)},1\right)_{\mathcal{T}}\\&+\frac{\Delta t}{2}\sum_{j=1}^{i}\widetilde{\boldsymbol{A}}_{i,j}\left[\rho^{n,j}_{(0)}\left\{(1-\mathcal{M}_{(0)})\boldsymbol{u}^{n,j}_{\Delta x,(0)},1\right\}_{\partial\mathcal{T}}+\left[\!\left[\rho^{n,j}_{(2)},1\right]\!\right]_{\partial\mathcal{T}}\right]\\&+\frac{\Delta t}{2}\sum_{j=1}^{i-1}\widehat{\boldsymbol{A}}_{i,j}\rho^{n,j}_{(0)}\left\{\mathcal{M}_{(0)}\boldsymbol{u}^{n,j}_{\Delta x,(0)},1\right\}_{\partial\mathcal{T}}.\end{aligned}$$

A periodic domain is given and every cell intersection $e\in\mathcal{T}^I$ is considered twice. Thus,

$$\begin{aligned}\{\boldsymbol{a},1\}_{\partial\mathcal{T}}=&\sum_{k=1}^{\text{ne}}\int_{\partial\Omega_k}\left(\boldsymbol{a}^-+\boldsymbol{a}^+\right)\cdot\boldsymbol{n}\text{d}\sigma\\=&\sum_{e\in\mathcal{T}^I}\int_e\left[\left(\boldsymbol{a}^-+\boldsymbol{a}^+\right)\cdot\boldsymbol{n}+\left(\boldsymbol{a}^++\boldsymbol{a}^-\right)\cdot(-\boldsymbol{n})\right]\text{d}\sigma\end{aligned}$$

$$=0,$$

where $\boldsymbol{a}$ is given by $(1-\mathcal{M}_{(0)})\boldsymbol{u}^{n,j}_{\Delta x,(0)}$ or $\mathcal{M}_{(0)}\boldsymbol{u}^{n,j}_{\Delta x,(0)}$. Similarly, we obtain for the jump

$$\begin{aligned}\left[\!\left[\rho^{n,j}_{(2)},1\right]\!\right]_{\partial\mathcal{T}} =&\sum_{k=1}^{\mathrm{ne}}\int_{\partial\Omega_k}\left(\rho^{n,j,-}_{(2)}-\rho^{n,j,+}_{(2)}\right)\mathrm{d}\sigma\\ =&\sum_{e\in\mathcal{T}^I}\int_e\left[\left(\rho^{n,j,-}_{(2)}-\rho^{n,j,+}_{(2)}+\rho^{n,j,+}_{(2)}-\rho^{n,j,-}_{(2)}\right]\right)\mathrm{d}\sigma\\ =&0.\end{aligned}$$

Thus, all boundary terms sum up to zero and we can conclude, using Lemma 5.22, that

$$0=\left(\rho^{n,i}_{(0)}-\rho^{n}_{(0)},1\right)_{\mathcal{T}}=\left(\rho^{n,i}_{(0)}-\rho^{n}_{(0)}\right)(1,1)_{\mathcal{T}}\qquad\Rightarrow\qquad\rho^{n,i}_{(0)}\equiv\rho^{n}_{(0)}.$$

Consequently, $\rho^{n,i}_{(0)}$ is constant in time. We can show the same result for $\rho^{n,i}_{(1)}$ by considering the $\mathcal{O}(\varepsilon)$ terms given in Corollary 5.19. □

Lemma 5.23 shows that $\rho^{n,i}_{(0)}$ and $\rho^{n,i}_{(1)}$ are constant in time. Then, these values equal the values of the previous time instance. Going back to the initial conditions we can conclude that

$$\rho^{n,i}_{(0)}\equiv\rho_{(0)}\qquad\text{and}\qquad\rho^{n,i}_{(1)}\equiv\rho_{(1)},$$

where $\rho_{(0)}$ and $\rho_{(1)}$ are given by the initial values, and use this representation in the following.

Limiting method

Next, we consider the limiting methods defined by the $\mathcal{O}(1)$ terms of the discretization of the conservation of mass and momentum equation. In comparison to the semi-discrete setting we cannot show that the divergence free constraint on $\boldsymbol{u}^{n,i}_{\Delta x,(0)}$ is given exactly. Therefore, we show that it is fulfilled in a discrete sense, i.e. that $\boldsymbol{u}^{n,i}_{\Delta x,(0)}$ is computed with a consistent discretization of $\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u}_{(0)}=0$. This is shown in the following lemma.

Lemma 5.24. *Under the assumptions of Theorem 5.17, the $\mathcal{O}(1)$ terms of the conservation of mass discretization are a consistent discretization of $\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u}_{(0)}$.*

Proof. We consider the $\mathcal{O}(1)$ terms of the conservation of mass discretization as given in Corollary 5.19 and directly use the results of Lemmas 5.22 and 5.23. Then the equation reads

$$\begin{aligned}0=&-\Delta t\sum_{j=1}^{i}\widetilde{\boldsymbol{A}}_{i,j}\left((1-\mathcal{M}_{(0)})\boldsymbol{u}^{n,j}_{\Delta x,(0)},\nabla\varphi\right)_{\mathcal{T}}-\Delta t\sum_{j=1}^{i-1}\widehat{\boldsymbol{A}}_{i,j}\left(\mathcal{M}_{(0)}\boldsymbol{u}^{n,j}_{\Delta x,(0)},\nabla\varphi\right)_{\mathcal{T}}\\ &+\frac{\Delta t}{2}\sum_{j=1}^{i}\widetilde{\boldsymbol{A}}_{i,j}\left[\left\{(1-\mathcal{M}_{(0)})\boldsymbol{u}^{n,j}_{\Delta x,(0)},\varphi\right\}_{\partial\mathcal{T}}+\rho^{-1}_{(0)}\left[\!\left[\rho^{n,j}_{\Delta x,(2)},\varphi\right]\!\right]_{\partial\mathcal{T}}\right]\\ &+\frac{\Delta t}{2}\sum_{j=1}^{i-1}\widehat{\boldsymbol{A}}_{i,j}\left\{\mathcal{M}_{(0)}\boldsymbol{u}^{n,j}_{\Delta x,(0)},\varphi\right\}_{\partial\mathcal{T}},\end{aligned}$$

which is a consistent discretization of $\nabla_{\boldsymbol{x}}\cdot\boldsymbol{u}_{(0)}=0$. In the end this is an IMEX discretization where the equation is split into the implicit contribution $(1-\mathcal{M}_{(0)})\boldsymbol{u}_{(0)}$ and the explicit contribution $\mathcal{M}_{(0)}\boldsymbol{u}_{(0)}$ with an additional stabilization in $\rho_{(2)}$. Note that $\rho_{(2)}$ corresponds to $p_{(2)}$ due to the asymptotic expansion of the pressure. □

Lemma 5.25. *Under the assumptions of Theorem 5.17, the $\mathcal{O}(1)$ terms of the conservation of momentum discretization are a consistent discretization of*

$$\partial_t \boldsymbol{u}_{(0)} + \nabla_{\boldsymbol{x}} \cdot \left(\boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{(0)} + \frac{p_{(2)}}{\rho_{(0)}} \operatorname{Id} \right) = 0. \tag{5.8}$$

Proof. We consider the $\mathcal{O}(1)$ terms of the conservation of momentum discretization as given in Corollary 5.20. Together with Lemmas 5.22 and 5.23 we obtain

$$\begin{aligned} 0 = {} & \left(\boldsymbol{u}^{n,i}_{\Delta x,(0)} - \boldsymbol{u}^{n}_{\Delta x,(0)}, \varphi \right)_{\mathcal{T}} - \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left(\frac{\boldsymbol{\mathcal{K}}^{n,j}_{(0)}}{\rho_{(0)}} + \frac{\mathcal{H}^{n,j}_{(2)}}{\rho_{(0)}} \operatorname{Id}, \nabla \varphi \right)_{\mathcal{T}} \\ & + \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[\left\{ \frac{\boldsymbol{\mathcal{K}}^{n,j}_{(0)}}{\rho_{(0)}} + \frac{\mathcal{H}^{n,j}_{(2)}}{\rho_{(0)}} \operatorname{Id}, \varphi \right\}_{\partial \mathcal{T}} + \left[\!\left[\boldsymbol{u}^{n,j}_{\Delta x,(0)}, \varphi \right]\!\right]_{\partial \mathcal{T}} \right] \\ & - \Delta t \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left(\boldsymbol{u}^{n,j}_{\Delta x,(0)} \otimes \boldsymbol{u}^{n,j}_{\Delta x,(0)} - \frac{\boldsymbol{\mathcal{K}}^{n,j}_{(0)}}{\rho_{(0)}} + \frac{p^{n,j}_{(2)}}{\rho_{(0)}} \operatorname{Id} - \frac{\mathcal{H}^{n,j}_{(2)}}{\rho_{(0)}} \operatorname{Id}, \nabla \varphi \right)_{\mathcal{T}} \\ & + \frac{\Delta t}{2} \sum_{j=1}^{i-1} \widehat{\boldsymbol{A}}_{i,j} \left\{ \boldsymbol{u}^{n,j}_{\Delta x,(0)} \otimes \boldsymbol{u}^{n,j}_{\Delta x,(0)} - \frac{\boldsymbol{\mathcal{K}}^{n,j}_{(0)}}{\rho_{(0)}} + \frac{p^{n,j}_{(2)}}{\rho_{(0)}} \operatorname{Id} - \frac{\mathcal{H}^{n,j}_{(2)}}{\rho_{(0)}} \operatorname{Id}, \varphi \right\}_{\partial \mathcal{T}}, \end{aligned}$$

where we used the same abbreviations as in Corollary 5.20. Overall we obtain a consistent discretization of Equation (5.8) where a splitting technique is used such that the flux function is split into the implicit contribution

$$\frac{\boldsymbol{\mathcal{K}}_{(0)}(\rho_{(0)}, \boldsymbol{u}_{(0)})}{\rho_{(0)}} + \frac{\mathcal{H}_{(2)}(\rho_{(0)}, \rho_{(1)}, \rho_{(2)})}{\rho_{(0)}} \operatorname{Id}$$

and the explicit contribution

$$\boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{(0)} - \frac{\boldsymbol{\mathcal{K}}_{(0)}(\rho_{(0)}, \boldsymbol{u}_{(0)})}{\rho_{(0)}} + \frac{p_{(2)}(\rho_{(0)}, \rho_{(1)}, \rho_{(2)})}{\rho_{(0)}} \operatorname{Id} - \frac{\mathcal{H}_{(2)}(\rho_{(0)}, \rho_{(1)}, \rho_{(2)})}{\rho_{(0)}} \operatorname{Id}. \tag{5.9}$$

□

Non-periodic boundary conditions

To show the asymptotic consistency of one stage in Theorem 5.17 we assumed that a periodic domain is given, which is a huge restriction. In Definition 2.23 we introduced boundary conditions which are useful to be assumed if one wants to obtain the same limiting behavior for the isentropic Euler equations as given for a periodic domain. Furthermore, in Remark 3.17 we saw how the discontinuous Galerkin method can handle boundary conditions. Therefore, we assume that the boundary values $\rho^{n,i}_{\partial\Omega,\Delta x}$ and $(\rho\boldsymbol{u})^{n,i}_{\partial\Omega,\Delta x}$ of the i^{th} internal stage are prescribed in such a way that the conditions of the following definition are fulfilled.

Definition 5.26. *We assume that the boundary is handled in such a way that*

$$\rho^{n,i}_{\partial\Omega,\Delta x} \equiv \rho^{n,i}_{\Delta x} \qquad \text{and} \qquad \int_{\partial\Omega} (\rho\boldsymbol{u})^{n,i}_{\partial\Omega,\Delta x} \cdot \boldsymbol{n} \mathrm{d}\sigma = \mathcal{O}(\varepsilon^2)$$

holds.

As an example the boundary conditions for $\boldsymbol{u}$ can be enforced for a solid wall boundary, see Equation (2.7), by choosing

$$(\rho\boldsymbol{u})^{n,i}_{\partial\Omega,\Delta x} = (\rho\boldsymbol{u})^{n,i}_{\Delta x} - ((\rho\boldsymbol{u})^{n,i}_{\Delta x} \cdot \boldsymbol{n})\boldsymbol{n},$$

where $\boldsymbol{n}$ is the given normal vector at the boundary.

In the following, we show how the proof of Theorem 5.17 changes if we consider a non-periodic domain. Thus, the numerical method given in Corollary 5.14 is modified in such a way that all present boundary integrals are only considered for inner cell intersections and the terms

$$\sum_{k=1}^{\mathrm{ne}} \int_{\partial\Omega_k \cap \partial\Omega} \widetilde{\boldsymbol{F}}(\rho^{n,j}_{\partial\Omega,\Delta x}, (\rho\boldsymbol{u})^{n,j}_{\partial\Omega,\Delta x})\varphi^- \cdot \boldsymbol{n}\mathrm{d}\sigma$$

are added. This can be done similarly for the explicit part. Since there is no stabilization at the boundary, Lemma 5.21 is not affected and therefore we also obtain that $\rho^{n,i}_{(0)}$ and $\rho^{n,1}_{(1)}$ are continuous over the whole spatial domain. In Lemma 5.22, after integration by parts, we obtain additional boundary terms of the form

$$\sum_{k=1}^{\mathrm{ne}} \int_{\partial\Omega_k \cap \partial\Omega} \left(\mathcal{H}_{(0)}\left(\rho^{n,j,-}_{\Delta x,(0)}\right) - \mathcal{H}_{(0)}\left(\rho^{n,j}_{\partial\Omega,\Delta x,(0)}\right)\right)\varphi^- \boldsymbol{n}\mathrm{d}\sigma.$$

These terms sum up to zero since the boundary function is chosen in such a way that $\rho^{n,j}_{\partial\Omega,\Delta x} \equiv \rho^{n,j}_{\Delta x}$. Similarly, we also obtain that the explicit boundary terms drop. Finally, in Lemma 5.23 we obtain with $\varphi \equiv 1$ additional boundary terms of the form,

$$\sum_{k=1}^{\mathrm{ne}} \int_{\partial\Omega_k \cap \partial\Omega,} (\rho\boldsymbol{u})^{n,j}_{\partial\Omega,\Delta x,(0)} \cdot \boldsymbol{n}\mathrm{d}\sigma.$$

These terms also sum up to zero since we assumed that the boundary conditions fulfill

$$\int_{\partial\Omega} (\rho\boldsymbol{u})^{n,j}_{\partial\Omega,\Delta x,(0)} \cdot \boldsymbol{n}\mathrm{d}\sigma = \mathcal{O}(\varepsilon^2).$$

From this we can conclude that $\rho^{n,i}_{\Delta x,(0)}$ and $\rho^{n,i}_{\Delta x,(1)}$ are constant in space and time. The remaining lemmas are not affected by the different boundary conditions and we can also conclude that the resulting method is asymptotically consistent.

5.2. Incompressible solver

In general, we cannot assume that the reference solution is given exactly. Therefore, we need to compute a numerical approximation of the limiting incompressible equations. Ideally, the reference solution corresponds to the $\mathcal{O}(1)$ terms of the compressible solution. Therefore, we consider the limiting numerical methods given in Lemmas 5.24 and 5.25 and derive a numerical method which does not depend on a reference solution. For this, we first compute the functions $\mathcal{M}_{(0)}$, $\boldsymbol{\mathcal{K}}_{(0)}$ and $\mathcal{H}_{(2)}$ for the RS-IMEX splitting. Due to Lemma 5.2 we can conclude that

$$\mathcal{M}_{(0)} = 0, \qquad \frac{\boldsymbol{\mathcal{K}}_{(0)}}{\rho_{(0)}} = -\boldsymbol{u}_{ref} \otimes \boldsymbol{u}_{ref} + \boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{ref} + \boldsymbol{u}_{ref} \otimes \boldsymbol{u}_{(0)}$$
$$\text{and} \qquad \mathcal{H}_{(2)} = p'(\rho_{ref})\rho_{(2)}.$$

The divergence free equation ($\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} = 0$) is handled with a completely implicit method, see Lemma 5.24 with $\mathcal{M}_{(0)} = 0$, i.e.

$$\begin{aligned} 0 = &-\Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left(\boldsymbol{u}^{n,j}_{\Delta x,(0)}, \nabla\varphi\right)_{\mathcal{T}} \\ &+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[\left\{\boldsymbol{u}^{n,j}_{\Delta x,(0)}, \varphi\right\}_{\partial\mathcal{T}} + \rho^{-1}_{(0)} \left[\!\left[\rho^{n,j}_{\Delta x,(2)}, \varphi\right]\!\right]_{\partial\mathcal{T}}\right]. \end{aligned}$$

Therefore, we consider the explicit part of the limiting conservation of momentum discretization, given in Lemma 5.25, and assume that the reference solution is the same as the computed numerical approximation.

Then, the explicit part sums up to zero and only the implicit part remains, i.e.

$$\begin{aligned}0 = &\left(\boldsymbol{u}^{n,i}_{\Delta x,(0)} - \boldsymbol{u}^{n}_{\Delta x,(0)}, \varphi\right)_{\mathcal{T}} - \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left(\frac{\boldsymbol{\mathcal{K}}^{n,j}_{(0)}}{\rho_{(0)}} + \frac{\mathcal{H}^{n,j}_{(2)}}{\rho_{(0)}} \operatorname{Id}, \nabla\varphi \right)_{\mathcal{T}} \\ &+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left[\left\{ \frac{\boldsymbol{\mathcal{K}}^{n,j}_{(0)}}{\rho_{(0)}} + \frac{\mathcal{H}^{n,j}_{(2)}}{\rho_{(0)}} \operatorname{Id}, \varphi \right\}_{\partial\mathcal{T}} + \left\llbracket \boldsymbol{u}^{n,j}_{\Delta x,(0)}, \varphi \right\rrbracket_{\partial\mathcal{T}} \right].\end{aligned}$$

Thus, the limiting method corresponds to a fully implicit discretization of the limiting equation. Note that the limiting equation works in the variables $p_{(2)}$ which corresponds to $\rho_{(2)}$ by

$$p_{(2)} = p'(\rho_{(0)})\rho_{(2)}.$$

The corresponding method is given in the following corollary.

Corollary 5.27. *The limiting discontinuous Galerkin method given in Lemmas 5.24 and 5.25 corresponds to a fully implicit discretization which is given by:*

1. *Set* $\boldsymbol{w}^{n,1}_{\Delta x} = \boldsymbol{w}^{n}_{\Delta x}$.
2. *For* $i = 2, \ldots, s$*: Seek* $\boldsymbol{w}^{n,i}_{\Delta x} \in V^3_{\Delta x}$ *such that*

$$\begin{aligned}0 = &\operatorname{Diag}\left\{ \left(0,1,1\right)^T \right\} \left(\boldsymbol{w}^{n,i}_{\Delta x} - \boldsymbol{w}^{n}_{\Delta x}, \varphi \right)_{\mathcal{T}} - \Delta t \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left(\boldsymbol{F}^I(\boldsymbol{w}^{n,j}_{\Delta x}), \nabla\varphi \right)_{\mathcal{T}} \\ &+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \left\{ \boldsymbol{F}^I(\boldsymbol{w}^{n,j}_{\Delta x}), \varphi \right\}_{\partial\mathcal{T}} \\ &+ \frac{\Delta t}{2} \sum_{j=1}^{i} \widetilde{\boldsymbol{A}}_{i,j} \operatorname{Diag}\left\{ \left(\left(\rho_{(0)} p'(\rho_{(0)})\right)^{-1}, 1, 1 \right)^T \right\} \left\llbracket \boldsymbol{w}^{n,j}_{\Delta x}, \varphi \right\rrbracket_{\partial\mathcal{T}}\end{aligned}$$

   holds for every $\varphi \in V_{\Delta x}$.
3. *Set* $\boldsymbol{w}^{n+1}_{\Delta x} = \boldsymbol{w}^{n,s}_{\Delta x}$.

Here, $\boldsymbol{w} = \left(p_{(2)}, \boldsymbol{u}_{(0)}\right)^T$ *and* $\boldsymbol{F}^I = \left(\boldsymbol{u}_{(0)}, \boldsymbol{u}_{(0)} \otimes \boldsymbol{u}_{(0)} + p_{(2)} \operatorname{Id}\right)^T$.

In the incompressible equations the pressure $p_{(2)}$ only occurs with spatial gradients. Therefore, it could only be computed in a unique way up to a constant. This constant could be arbitrary large if we use the method given in Corollary 5.27. Then, stability and machine accuracy issues could occur. This is why we perform a pressure correction in every step to force the mean value of $p_{(2)}$ to be zero such that we obtain a unique approximation in every step. This means we compute

$$\overline{p}^{n,i}_{\Delta x,(2)} = p^{n,i}_{\Delta x,(2)} - \frac{1}{|\Omega|} \int_{\Omega} p^{n,i}_{\Delta x,(2)} \mathrm{dx}.$$

5.3. Numerical results

In the following we consider the examples defined in Section 2.3 to obtain in which way the numerical method behaves and which convergence behavior it shows. For this, the numerical methods are implemented in C++, where we use the library Netgen [157] and NGSolve [156] to handle grids, basis functions, quadrature rules and so forth and the libraries PETSc [14, 15] to solve the resulting system of linear equations.

To compute the convergence behavior we consider a set of grids which contain uniformly quadratic cells. The value Δx corresponds to the length of one cell edge and can therefore be computed by $\Delta x = 1/\sqrt{\mathrm{ne}}$ for the used domain $\Omega = [0,1]^2$. An overview of the used grids is given in Table 5.1.

In the following, we apply the numerical method to the examples given in Definitions 2.30, 2.31 and 2.32. To comment on the quality of the numerical approximation we need some kind of error measurement. For

Δx	number of cells
0.25	$4 \times 4 = 16$
0.125	$8 \times 8 = 64$
0.0625	$16 \times 16 = 256$
0.03125	$32 \times 32 = 1024$
0.015625	$64 \times 64 = 4096$
0.0078125	$128 \times 128 = 16384$
0.00390625	$256 \times 256 = 65536$

Table 5.1.: The size of the used grids. Given is the number of cells and the value Δx which is the length of one cell edge and computed by $\Delta x = 1/\sqrt{\text{ne}}$.

number of cells	second order	third order	fourth order
16	64	144	256
64	256	576	1024
256	1024	2304	4096
1024	4096	9216	16384
4096	16384	36864	65536
16384	65536	147456	262144
65536	262144	589824	1048576

Table 5.2.: The number of degrees of freedom per variable for the different grids and different maximum polynomial degrees.

the smooth vortex, an exact solution is available and therefore we can compute the L^2-error between the exact and approximate solution by

$$e_{\Delta x}^2 := \|\boldsymbol{w}_{\Delta x}^N - \boldsymbol{w}(t^{end})\|_{L^2}^2 = \int_\Omega \|\boldsymbol{w}_{\Delta x}^N(\boldsymbol{x}) - \boldsymbol{w}(t^{end}, \boldsymbol{x})\|_2^2 \mathrm{dx},$$

where $\boldsymbol{w}$ denotes the exact solution and $\boldsymbol{w}_{\Delta x}^N$ the numerical approximation at the final time instance t^{end}. For the periodic flow and vortex in a box examples exact solutions are not available. Therefore, we compute a numerical approximation on two different grids with grid size Δx and $\Delta x/2$ and compare them in the L^2-norm

$$e_{\Delta x}^2 := \|\boldsymbol{w}_{\Delta x}^N - \boldsymbol{w}_{\Delta x/2}^{2N}\|_{L^2}^2 = \int_\Omega \|\boldsymbol{w}_{\Delta x}^N(\boldsymbol{x}) - \boldsymbol{w}_{\Delta x/2}^{2N}(\boldsymbol{x})\|_2^2 \mathrm{dx}.$$

Note that this measure only provides a lower bound for the convergence towards the exact solution since

$$\begin{aligned} \|\boldsymbol{w}_{\Delta x}^N - \boldsymbol{w}_{\Delta x/2}^{2N}\|_{L^2} =& \|\boldsymbol{w}_{\Delta x}^N - \boldsymbol{w}(t^{end}) + \boldsymbol{w}(t^{end}) - \boldsymbol{w}_{\Delta x/2}^{2N}\|_{L^2} \\ \leq& \|\boldsymbol{w}_{\Delta x}^N - \boldsymbol{w}(t^{end})\|_{L^2} + \|\boldsymbol{w}(t^{end}) - \boldsymbol{w}_{\Delta x/2}^{2N}\|_{L^2}. \end{aligned}$$

Thus, if the method converges we can see convergence by the behavior of $e_{\Delta x}$ but not vice versa.

In Chapter 4 we investigated different IMEX Runge-Kutta schemes in terms of asymptotic accuracy for the ordinary differential equation given in Definition 2.12. From this we obtained that a globally stiffly accurate Runge-Kutta scheme is desirable in this setting. Therefore, we consider the IMEX DG method with the

- ARS_222 scheme with polynomials of maximal degree one for the second order case,
- BPR_353 and ARS_443 with polynomials of maximal degree two for the third order cases
- and ARK_4A2 with polynomials of maximal degree three for the fourth order case

to solve the compressible equation. An overview on the different number of degrees of freedom for the

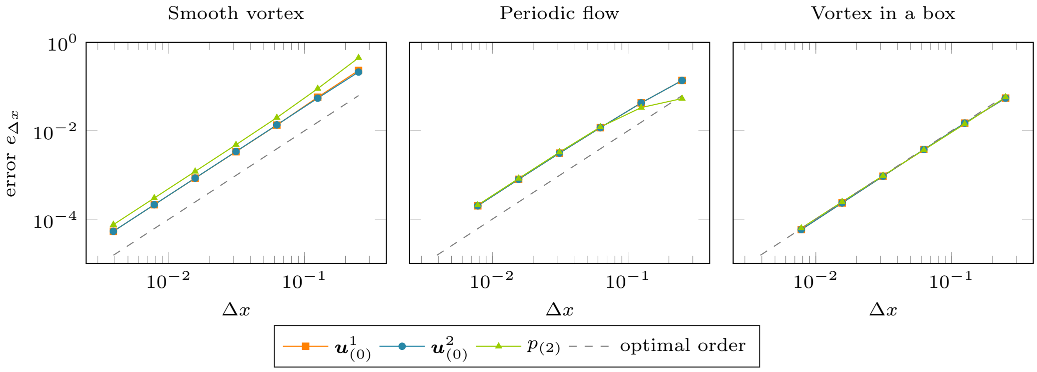


Figure 5.1.: Numerical convergence analysis for the second order discontinuous Galerkin method with the implicit part of the ARS_222 scheme given in Corollary 5.27 for computing the reference solution of the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. The L^2-error in each component is plotted. The dashed line gives the optimal order of convergence.

different orders is given in Figure 5.2

In the following we start with the numerical method for computing the reference solution and investigate the error behavior. Then, we consider the different schemes for the compressible equation and compute their numerical stability behavior to obtain a proper choice of $\Delta t/\Delta x$ to compute the convergence behavior. This is then done at the end of the section.

Note that, beside the smooth vortex example, we do not know anything about the smoothness of the examples and we do not know if there are stiff gradients which can affect the convergence on a coarse grid. Therefore, even if the numerical method is high order accurate there could be a low order convergence behavior for coarse and fine grids. Furthermore, it is not clear for which ε a solution can be represented by an asymptotic expansion, i.e. for which ε we are in the low Mach limit.

5.3.1. Reference solution

To obtain the reference solution for $\boldsymbol{u}$ we use the incompressible solver defined in Corollary 5.27. As reference solution for ρ we use the value $\rho_{(0)}$ which is defined by the initial conditions. The method is fully implicit and therefore computing the reference solution is of high computational cost. This is why we choose a reference solution less accurate than the overall numerical method. In detail, for all numerical examples we compute a reference solution with the same time integration scheme but only with first order, piece-wise linear, polynomials. This procedure results in a reference solution which is at most second order accurate but might be as good as possible to obtain a stable numerical method.

The convergence results for the different time integration schemes are summarized in Figures 5.1, 5.2, 5.3 and 5.4. We obtain that in every case for every example the numerical approximation converges with second order. We do not see any difference between the convergence behavior of the different time integration schemes, therefore we can conclude that the spatial error is dominating and the temporal discretization error only slightly affects the accuracy.

5.3.2. Asymptotic stability

We investigate the asymptotic stability of the proposed numerical method for weakly compressible flows, i.e. we try to find out whether we are able to choose a time step restriction which is independent of ε if $\varepsilon \ll 1$. For this we fix a relatively coarse grid, in this case a grid with $8 \times 8 = 64$ cells, and perform a fixed number, here $N = 2000$, of time steps for different values of $\Delta t/\Delta x$ and ε. Thus in the end this means we vary CFL_{conv} or since Δx is fixed we vary Δt. In every time step we compute the L^2-norm of

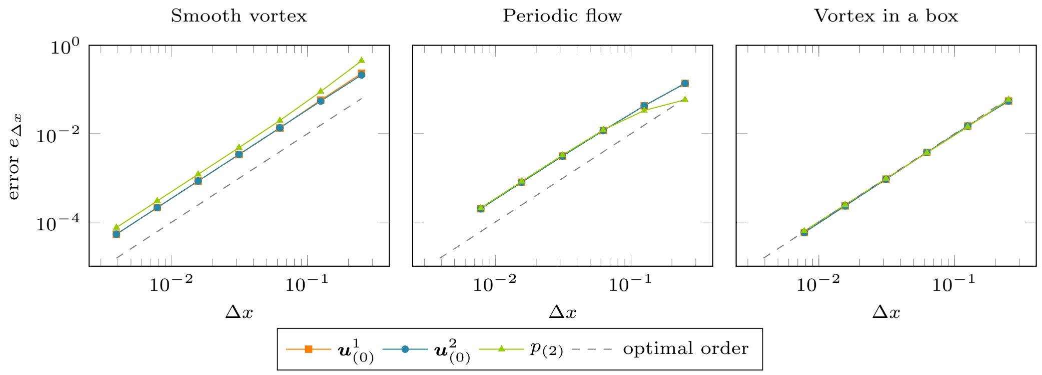


Figure 5.2.: Numerical convergence analysis for the second order discontinuous Galerkin method with the implicit part of the BPR_353 scheme given in Corollary 5.27 for computing the reference solution of the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. The L^2-error in each component is plotted. The dashed line gives the optimal order of convergence.

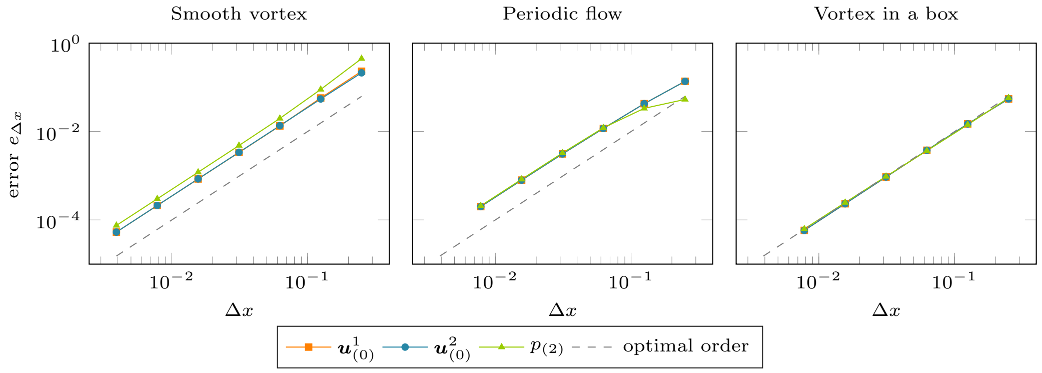


Figure 5.3.: Numerical convergence analysis for the second order discontinuous Galerkin method with the implicit part of the ARS_443 scheme given in Corollary 5.27 for computing the reference solution of the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. The L^2-error in each component is plotted. The dashed line gives the optimal order of convergence.

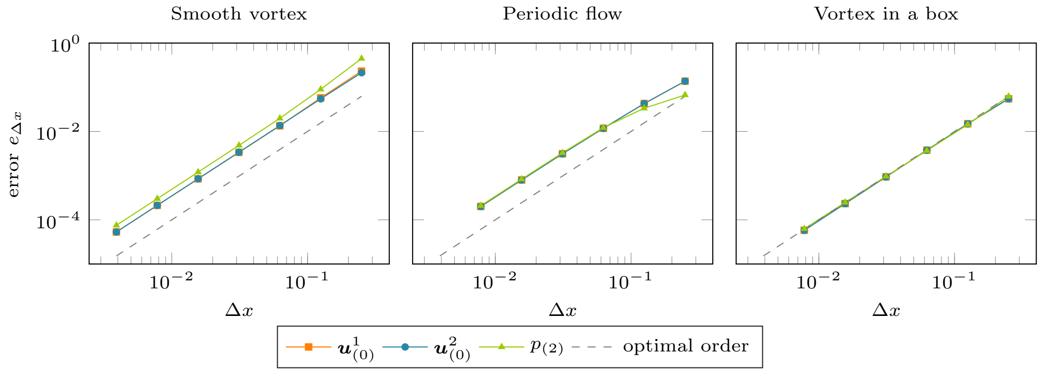


Figure 5.4.: Numerical convergence analysis for the second order discontinuous Galerkin method with the implicit part of the ARK_4A2 scheme given in Corollary 5.27 for computing the reference solution of the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. The L^2-error in each component is plotted. The dashed line gives the optimal order of convergence.

the solution $\|\boldsymbol{w}^{n,i}_{\Delta x}\|_{L^2}$ and check the behavior of this value. If the L^2-norm raises over a specific threshold, here 1000, we stop the computation. Finally, we plot the resulting norms in an $\varepsilon - \Delta t$ diagram. Please note that this procedure does not prove asymptotic stability and only gives a glimpse on the behavior of the scheme. This is in more detail stressed out by the following two comments:

1. Even if the numerical method is unstable, the L^2-norm could be bounded by a small constant, e.g. if a steady state or constant state is reached before instabilities occur.
2. Next to the method given in Corollary 5.14, stability also depends on the equation solver, stopping criteria, implementation details and so forth.

The results, which we discuss in the following in detail, are given

– in Figure 5.5 for the second order method,

– in Figures 5.6 and 5.7 for the third order methods and

– in Figure 5.8 for the fourth order method.

We obtain instabilities for large values of $\Delta t/\Delta x$ and also for large values of ε. It seems like the method gets more stable if ε gets smaller. This observation corresponds to the eigenvalues of the explicit part, which are given in $\mathcal{O}(\varepsilon)$, i.e. for $\varepsilon \ll 1$ the influence of the explicit part is very small. Only for the vortex in a box example it seems like the method is more stable if ε is large, but this could be caused by a steady state or constant solution which is reached.

From the results given in Figures 5.5, 5.6, 5.7 and 5.8 and also experiences from numerical computations we choose different values of $\Delta t/\Delta x$ for the computations which are done in the next section. These choices are summarized in Tables 5.3 and 5.4, where also the corresponding convective CFL number CFL_{conv} is given.

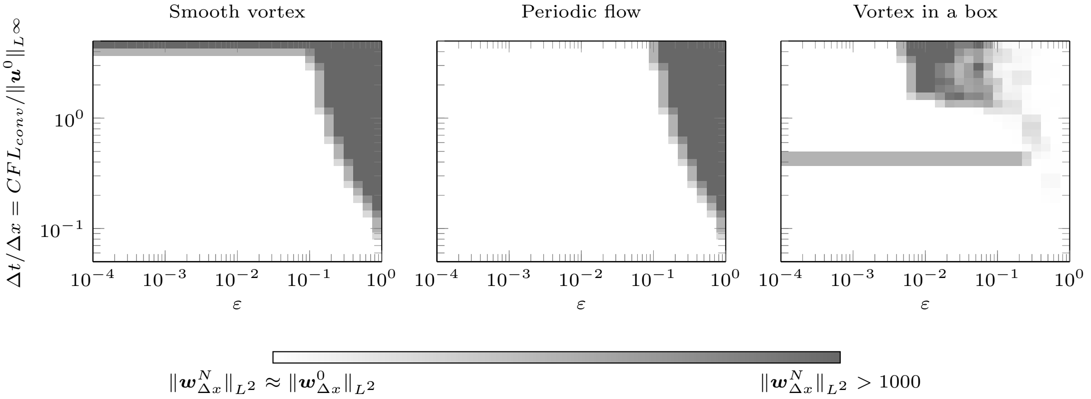


Figure 5.5.: Numerical stability analysis for second order discontinuous Galerkin method with the ARS_222 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right) see Definition 2.32, for the different values of $\Delta t/\Delta x = CFL_{conv}/\|\boldsymbol{u}^0\|_{L^\infty}$ and ε. 2000 steps on an 8×8 grid are performed. An L^2-norm larger than 1000 is set to 1000.

5.3.3. Asymptotic accuracy

Next, we investigate the error behavior of the numerical method proposed in this thesis. The choices of Δt are motivated by the stability results in Section 5.3.2 and are given in Tables 5.3 and 5.4.

The convergence results of

– the second order method are given in Figure 5.9,

Scheme	Example	$\varepsilon = 10^0$		$\varepsilon = 10^{-1}$		$\varepsilon = 10^{-2}$	
		$\frac{\Delta t}{\Delta x}$	CFL_{conv}	$\frac{\Delta t}{\Delta x}$	CFL_{conv}	$\frac{\Delta t}{\Delta x}$	CFL_{conv}
ARS_222	Smooth vortex	0.05	0.074	0.1	0.148	1	1.48
	Periodic flow	0.05	0.05	0.1	0.1	1	1
	Vortex in a box	0.05	0.05	0.1	0.1	0.1	0.1
ARS_443	Smooth vortex	0.05	0.074	1	1.48	1	1.48
	Periodic flow	0.05	0.05	0.1	0.1	0.1	0.1
	Vortex in a box	0.05	0.05	0.1	0.1	0.1	0.1
BPR_353	Smooth vortex	0.01	0.0148	0.01	0.0148	0.1	0.148
	Periodic flow	0.01	0.01	0.01	0.01	0.1	0.1
	Vortex in a box	0.01	0.01	0.1	0.1	0.1	0.1
ARK_4SA	Smooth vortex	0.05	0.074	0.1	0.148	0.1	0.148
	Periodic flow	0.05	0.05	0.05	0.05	0.01	0.01
	Vortex in a box	0.05	0.05	0.05	0.05	0.05	0.05

Table 5.3.: Values of $\Delta t/\Delta x$ and the convective CFL number CFL_{conv}, which is given by $\frac{\Delta t}{\Delta x}\|\boldsymbol{u}^0\|_{L^\infty} = CFL_{conv}$, for different discretization schemes and for $\varepsilon = 1, 10^{-1}, 10^{-2}$. The value $\|\boldsymbol{u}^0\|_{L^\infty}$ is approximately 1.48 for the smooth vortex, 1 for the periodic flow and 1 for the vortex in a box example.

Scheme	Example	$\varepsilon = 10^{-3}$		$\varepsilon = 10^{-4}$	
		$\frac{\Delta t}{\Delta x}$	CFL_{conv}	$\frac{\Delta t}{\Delta x}$	CFL_{conv}
ARS_222	Smooth vortex	1	1.48	1	1.48
	Periodic flow	1	1	1	1
	Vortex in a box	0.1	0.1	0.1	0.1
ARS_443	Smooth vortex	1	1.48	1	1.48
	Periodic flow	0.1	0.1	0.1	0.1
	Vortex in a box	0.1	0.1	0.1	0.1
BPR_353	Smooth vortex	0.1	0.148	0.1	0.148
	Periodic flow	0.1	0.1	0.1	0.1
	Vortex in a box	0.1	0.1	0.1	0.1
ARK_4SA	Smooth vortex	0.1	0.148	0.1	0.148
	Periodic flow	0.05	0.05	0.05	0.05
	Vortex in a box	0.05	0.05	0.05	0.05

Table 5.4.: Values of $\Delta t/\Delta x$ and the convective CFL number CFL_{conv}, which is given by $\frac{\Delta t}{\Delta x}\|\boldsymbol{u}^0\|_{L^\infty} = CFL_{conv}$, for different discretization schemes and for $\varepsilon = 10^{-3}, 10^{-4}$. The value $\|\boldsymbol{u}^0\|_{L^\infty}$ is approximately 1.48 for the smooth vortex, 1 for the periodic flow and 1 for the vortex in a box example.

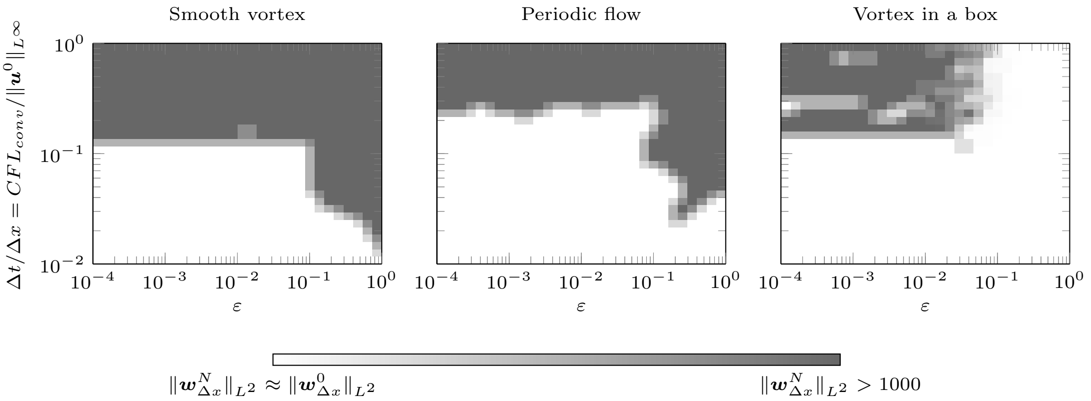


Figure 5.6.: Numerical stability analysis for third order discontinuous Galerkin method with the BPR_353 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right) see Definition 2.32, for the different values of $\Delta t/\Delta x = CFL_{conv}/\|\boldsymbol{u}^0\|_{L^\infty}$ and ε. 2000 steps on an 8×8 grid are performed. An L^2-norm larger than 1000 is set to 1000.

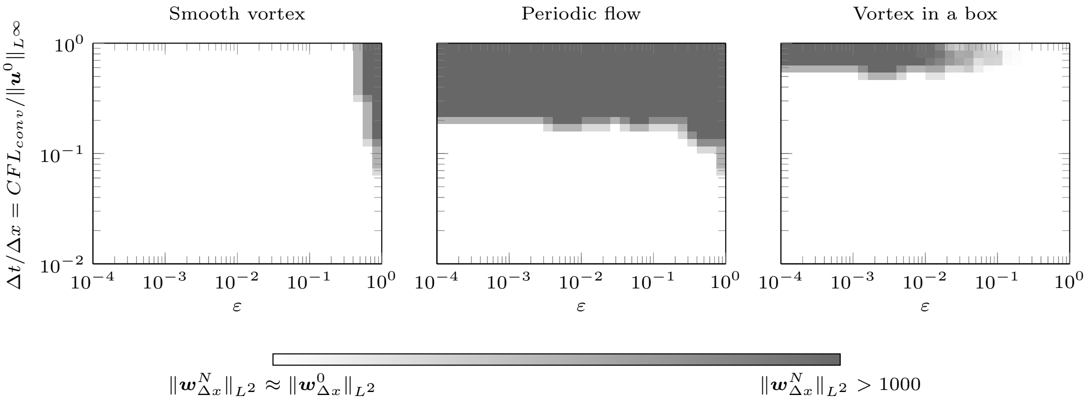


Figure 5.7.: Numerical stability analysis for third order discontinuous Galerkin method with the ARS_443 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right) see Definition 2.32, for the different values of $\Delta t/\Delta x = CFL_{conv}/\|\boldsymbol{u}^0\|_{L^\infty}$ and ε. 2000 steps on an 8×8 grid are performed. An L^2-norm larger than 1000 is set to 1000.

- the third order method with the ARS_443 scheme are given in Figure 5.10,
- the third order method with the BPR_353 scheme are given in Figure 5.11
- and for the fourth order method are given in Figure 5.12.

For the high order methods and very small ε, $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$, we obtain an extreme drop of convergence order. This drop is caused by the extreme stiffness of the equation in combination with the machine accuracy. In more detail we consider the $\varepsilon = 10^{-4}$ case for the fourth order method and observe that the drop of convergence is given if the error $e_{\Delta x}$ reaches approximately 10^{-5}. Furthermore we obtain terms in the numerical method which are given by $\varepsilon^{-2} = 10^8$. Overall we start from a value which is approximately $\mathcal{O}(10^8)$ and converge up to an error of $\mathcal{O}(10^{-5})$. Thus, a relative error of $\mathcal{O}(10^{-13})$ is reached which is close to machine accuracy.

If we ignore the machine accuracy issues, we can obtain that the numerical methods converge with the desired order of accuracy if ε is small. Thus, we see the desired behavior of the method. This observation is also given for large values of ε and the smooth vortex example. For the smooth vortex example only

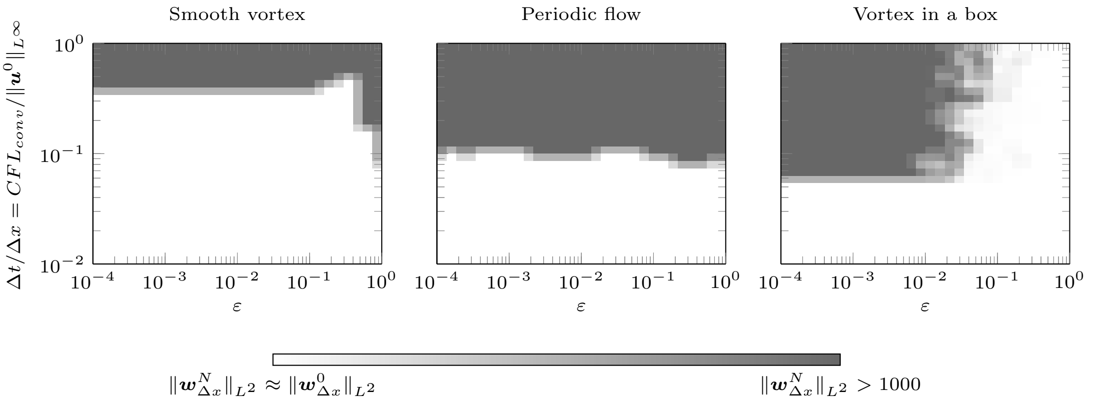


Figure 5.8.: Numerical stability analysis for fourth order discontinuous Galerkin method with the ARK_4A2 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right) see Definition 2.32, for the different values of $\Delta t/\Delta x = CFL_{conv}/\|\boldsymbol{u}^0\|_{L^\infty}$ and ε. 2000 steps on an 8×8 grid are performed. An L^2-norm larger than 1000 is set to 1000.

the BPR_353 scheme, see Figure 5.11, shows a slightly reduced order of convergence for $\varepsilon = 1$. In this case it might be possible that the resulting method is not stable since we observed in Figure 5.6 that the BPR_353 scheme is unstable close to the chosen value of $\Delta t/\Delta x$.

For the periodic flow and vortex in a box examples the results are not clear. We see the optimal order for every example if ε is small, but for large values of ε, i.e. $\varepsilon = 10^0$ and $\varepsilon = 10^{-1}$, we obtain that a convergence with a lower order is given or the desired order is only reached for a small value of Δx. Furthermore for $\varepsilon = 10^{-2}$ we obtain the method converges with the desired order and then the order of convergence drops to a lower one. There could be several explanations for this behavior:

1. This could be caused by steep gradients or a solution which is not smooth enough such that the desired order can only be obtained if the grid resolution is small enough or cannot be obtained. Unfortunately, we do not know the solution of these examples and therefore we cannot say anything about their properties. The assumption on steep gradients or non-smoothness is supported by several observations:

   – For the periodic flow example, the desired order of convergence is in several cases reached after some refinements of the grid for $\varepsilon = 10^{-1}, 10^{-2}$. Thus, steep gradients could be present in $\mathcal{O}(\varepsilon)$, $\mathcal{O}(\varepsilon^2)$ or $\mathcal{O}(\varepsilon^3)$ terms, which are less dominant for very small values of ε.

   – For the periodic flow example this assumption is also supported by Figure 2.9 where steep gradients can be seen for the $\varepsilon = 1$ solution. Note, that the assumption that the solution is given as an asymptotic expansion can only be valid if ε is small enough. Therefore, for $\varepsilon = 1$ it could be that the solution is not in the asymptotic regime and therefore shows a different behavior than for $\varepsilon \ll 1$.

   – In Figure 5.17 the convergence behavior of a fully explicit discretization for $\varepsilon = 1$ and different orders are shown and in every case a similar behavior as for the RS-IMEX scheme is given.

2. Taking a closer look on the convergence behavior in the case of $\varepsilon = 10^{-1}, 10^{-2}$ this looks similar to order reduction phenomena, see Chapter 4 for more details.

Overall, the behavior of the numerical method for the periodic flow and vortex in a box example need further investigation.

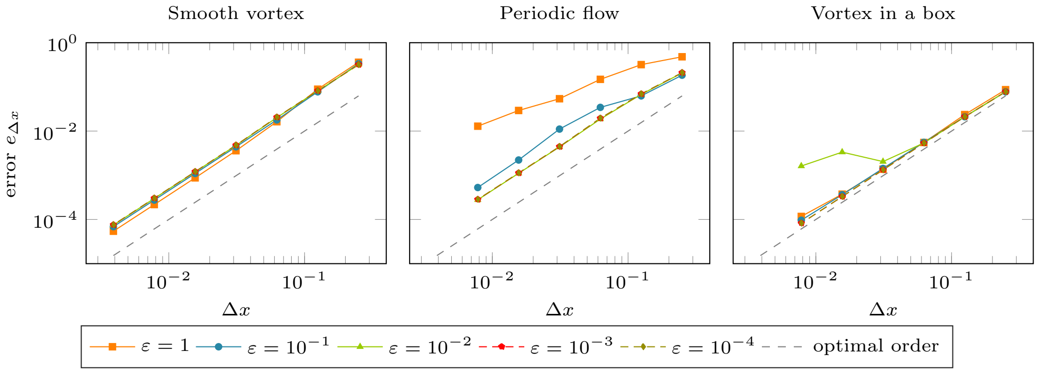


Figure 5.9.: Numerical convergence analysis for second order discontinuous Galerkin method with the ARS_222 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. For the smooth vortex example the L^2-error is plotted, for the other examples the L^2-error between two following numerical approximations are plotted. The dashed line gives the optimal order of convergence.

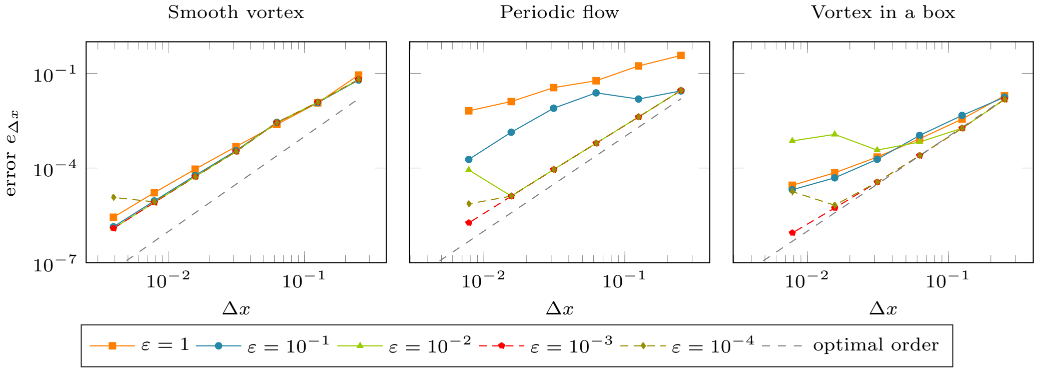


Figure 5.10.: Numerical convergence analysis for Third order discontinuous Galerkin method with the ARS_443 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. For the smooth vortex example the L^2-error is plotted, for the other examples the L^2-error between two following numerical approximations are plotted. The dashed line gives the optimal order of convergence.

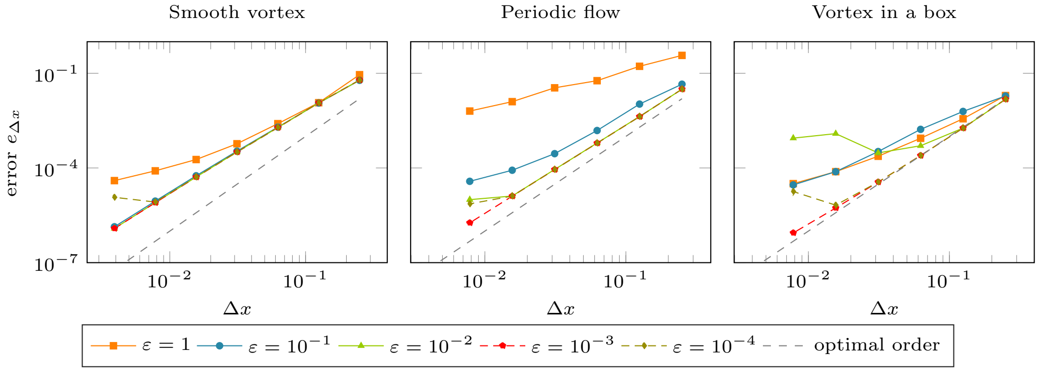


Figure 5.11.: Numerical convergence analysis for third order discontinuous Galerkin method with the BPR_353 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. For the smooth vortex example the L^2-error is plotted, for the other examples the L^2-error between two following numerical approximations are plotted. The dashed line gives the optimal order of convergence.

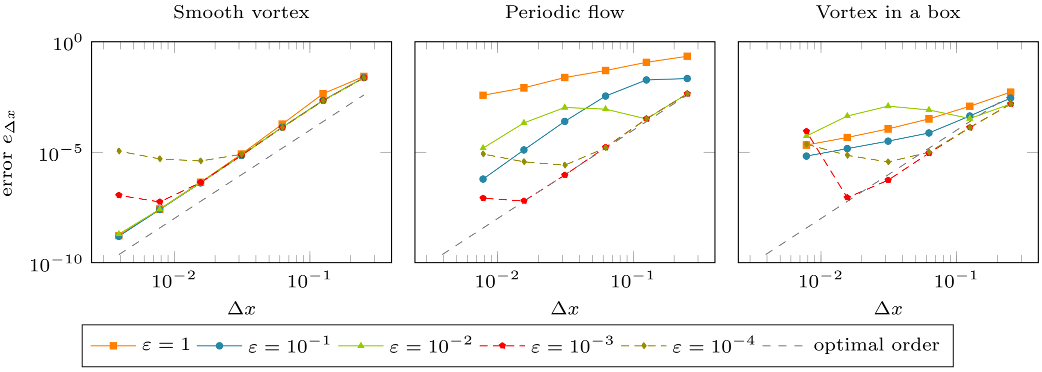


Figure 5.12.: Numerical convergence analysis for fourth order discontinuous Galerkin method with the ARK_4A2 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. For the smooth vortex example the L^2-error is plotted, for the other examples the L^2-error between two following numerical approximations are plotted. The dashed line gives the optimal order of convergence.

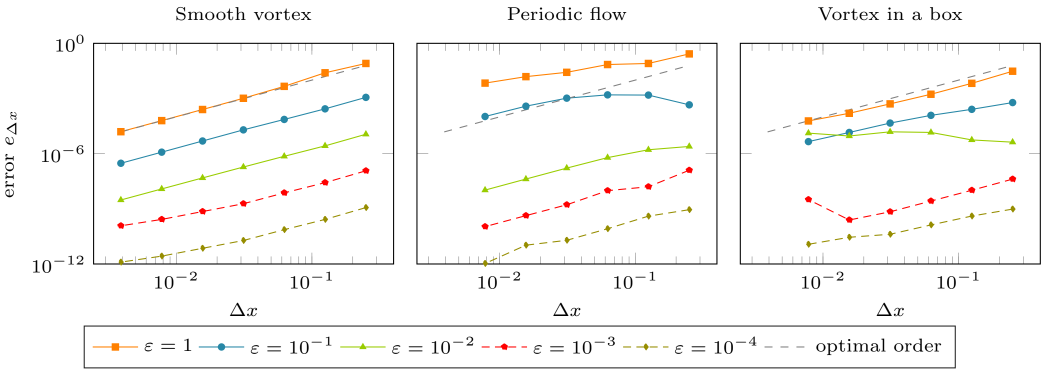


Figure 5.13.: Numerical convergence analysis in ρ for second order discontinuous Galerkin method with the ARS_222 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. For the smooth vortex example the L^2-error in ρ is plotted, for the other examples the L^2-error in ρ between two following numerical approximations are plotted. The dashed line gives the optimal order of convergence.

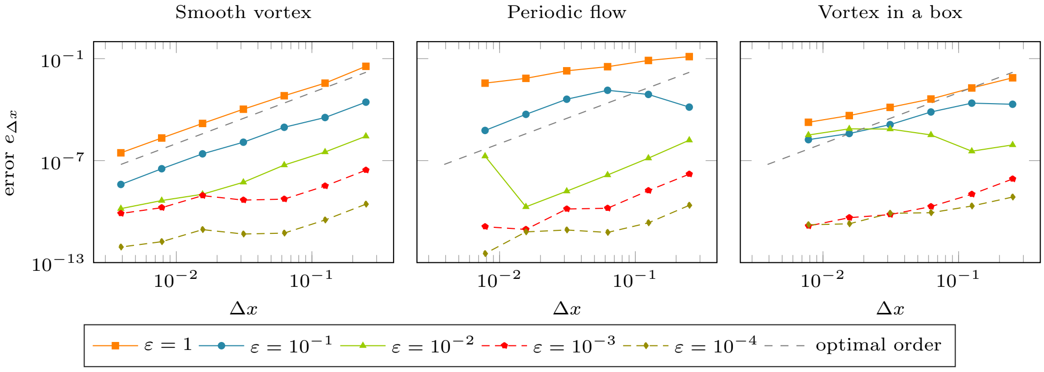


Figure 5.14.: Numerical convergence analysis in ρ for third order discontinuous Galerkin method with the ARS_443 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. For the smooth vortex example the L^2-error in ρ is plotted, for the other examples the L^2-error in ρ between two following numerical approximations are plotted. The dashed line gives the optimal order of convergence.

Asymptotic consistency

In Section 5.1 we have performed an asymptotic consistency analysis to check whether the numerical method is consistent with the limiting behavior of the corresponding equations. To see this property in the numerical examples we compute the L^2-error in ρ. Ideally, we can see that this error behaves like $\mathcal{O}(\varepsilon^2)$ if we consider the different methods in ε. These convergence results of the

- second order method are given in Figure 5.13,
- third order method with the ARS_443 scheme are given in Figure 5.14,
- third order method with the BPR_353 scheme are given in Figure 5.15 and
- fourth order method are given in Figure 5.16.

From these figures we can directly obtain the asymptotic consistency if the numerical method converges with the desired order and if we neglect machine accuracy issues.

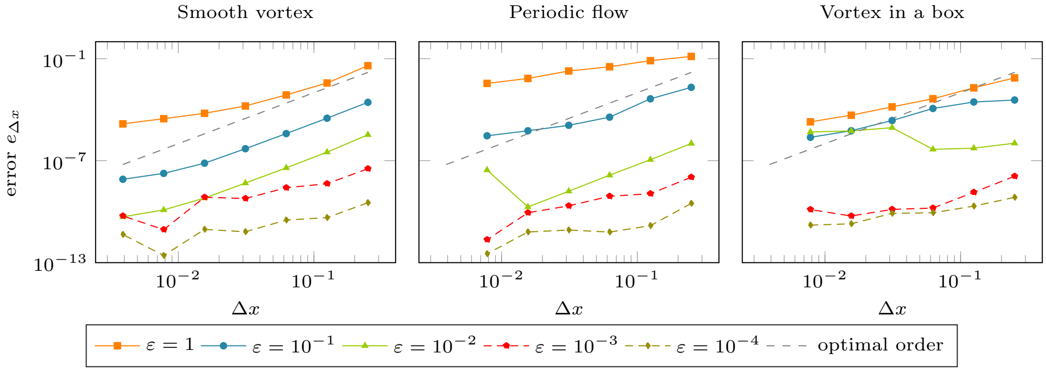


Figure 5.15.: Numerical convergence analysis in ρ for third order discontinuous Galerkin method with the BPR_353 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. For the smooth vortex example the L^2-error in ρ is plotted, for the other examples the L^2-error in ρ between two following numerical approximations are plotted. The dashed line gives the optimal order of convergence.

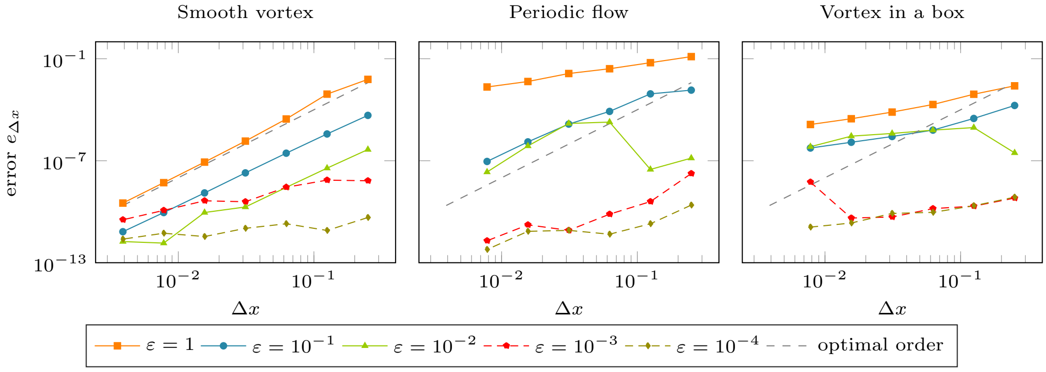


Figure 5.16.: Numerical convergence analysis in ρ for fourth order discontinuous Galerkin method with the ARK_4A2 scheme coupled with the RS-IMEX splitting for the smooth vortex example (left), see Definition 2.30, the periodic flow example (middle), see Definition 2.31, and the vortex in a box example (right), see Definition 2.32. For the smooth vortex example the L^2-error in ρ is plotted, for the other examples the L^2-error in ρ between two following numerical approximations are plotted. The dashed line gives the optimal order of convergence.

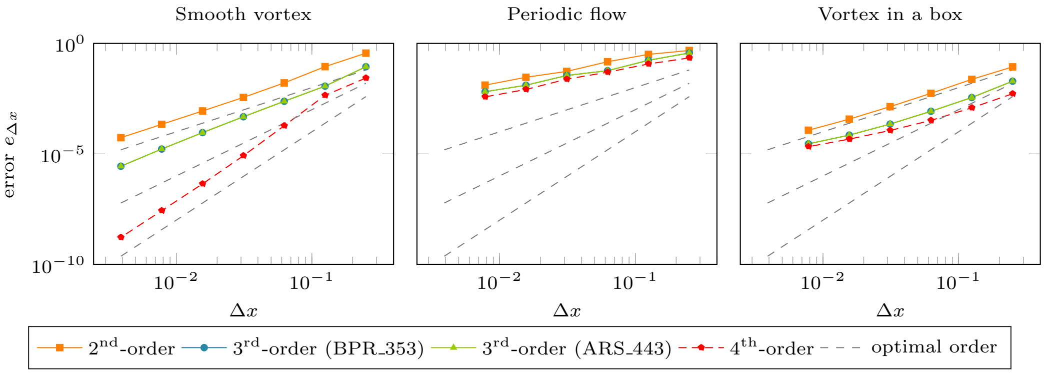


Figure 5.17.: Numerical convergence analysis for an explicit discontinuous Galerkin method coupled with the explicit part of the IMEX Runge-Kutta schemes given in the appendix for the smooth vortex example (left) see Definition 2.30, the periodic flow example (middle) see Definition 2.31 and the vortex in a box example (right) see Definition 2.32. For the smooth vortex example the L^2-error is plotted, for the other examples the L^2-error between two following numerical approximations are plotted. The dashed lines gives the optimal order of convergence.

5.4. Comparison to methods from literature

One main step in the development of a new numerical method is the comparison to standard methods in terms of efficiency and accuracy. In the analysis given in this thesis we ignored the computational cost. If the reference solution is given, we obtain a linear implicit part which should be solved efficiently. Therefore, the computation of the reference solution is one bottleneck in the efficiency of the RS-IMEX splitting. For ordinary differential equations we were able to compute the reference solution with an explicit method, thus an efficient method is obtained. For the isentropic Euler equations we used a fully implicit discretization, which leads to huge computational cost. To reduce this we computed the reference solution less accurately but this might be still an overhead compared to standard numerical methods from literature.

In [J3] a first order finite volume discretization is used to compare the RS-IMEX splitting with the one given in [78]. The reference solution is computed with a splitting of the incompressible Euler equations again with a first order finite volume discretization. The conclusion of this work is that in the low Mach setting the RS-IMEX splitting is able to compute a more accurate numerical approximation and therefore the additional effort is justified. Note that this comparison did not use the numerical method proposed in [78], it only considered the splitting.

The high order discretization proposed in this work is also used in [P1, J5] to compare the RS-IMEX splitting with different splittings or fully explicit/implicit schemes from literature implemented in the Flexi software package [85]. In [J5] the method is compared to an implicit and explicit discretization for several different examples. For a small Mach number the RS-IMEX splitting, where the reference solution is computed with the method given in Section 5.2 with the same order of accuracy as the compressible part, is more efficient than a fully implicit discretization. Furthermore the RS-IMEX splitting is also able to compete with a fully explicit discretization. In [P1] the computation of the reference solution is optimized by two different approaches:

1. The reference solution is computed with a lower order method, which is similar to the way we obtain the reference solution in this work.

2. A completely different reference solution is used which fulfills

$$\rho_{ref} - \rho = \mathcal{O}(\varepsilon^2) \qquad \text{and} \qquad \boldsymbol{u}_{ref} - \boldsymbol{u} = \mathcal{O}(1)$$

but can be computed without solving an additional equation.

The conclusion is that the computational cost of the reference solution can be reduced if the choice of the reference solution is suitable, i.e. choosing the reference solution as $\rho_{ref} = \rho_{(0)}$ and $\boldsymbol{u}_{ref}(t, \cdot) = \frac{1}{\|\Omega\|} \int_\Omega \boldsymbol{u}(t, \boldsymbol{x}) \mathrm{dx}$ evaluated at a previous time instance leads to a more efficient discretization which shows the same convergence behavior as the classical choice. Furthermore, in a comparison to splittings from literature it is shown that the efficiency can compete with the splitting given in [78] while it is more efficient compared to the one given in [50] in the low Mach case.

Finally, in [P1, J5] also the accuracy and consistency in the low Mach regime are compared with the methods from literature. In all cases the proposed method is able to compute an approximation with the same accuracy and to resolve the correct behavior in the asymptotic limit also for more complex examples. This can also be seen as a verification that the method is suitable in the low Mach limit.

5.5. Conclusion and summary

First of all, we have shown that an IMEX Runge-Kutta method coupled with a generalized splitting and a discontinuous Galerkin method is consistent with the asymptotic behavior of the isentropic Euler equations. This was done by proving the asymptotic consistency property in Corollary 5.6 for the semi-discrete method and in Corollary 5.18 for the fully-discrete method.

With the help of numerical examples we have seen that the resulting numerical method seems to be asymptotically stable, i.e. we are able to choose a time step restriction which is independent of ε if $\varepsilon \ll 1$. Furthermore, we have seen that for a smooth test case with known smooth solution the numerical method converges with the desired order of accuracy. For the test cases with unknown solution we have different problems concerning the convergence if ε is relatively large. We assumed that these problems are caused by the unknown behavior of the solution and not by the numerical method. We were not able to prove this assumption, yet we could motivate it.

Note that the results concerning the smooth vortex are also verified by a different implementation of the method proposed in this thesis in [J5]. In the same publication also further more complex examples are computed and the discretization method performed well in the low Mach case. Furthermore, efficiency and performance in the low Mach case for more complex examples are also compared to methods from literature in [P1, J5]. As a conclusion we can say that the numerical method is able to resolve the correct behavior in the low Mach limit and is efficient if one chooses the reference solution in a suitable way.

For very small Mach numbers we obtained that the numerical solution is close to machine accuracy. In [J5] a modification of the RS-IMEX scheme is considered to show that this problem can be resolved by using the reference solution as the $\mathcal{O}(1)$ terms of the compressible solution and then adapt the numerical method to obtain an approximation for the higher order terms in ε. For this, a more accurate reference solution is needed and therefore computational cost become larger. Similar ideas are used in [127, 183] to reduce the stiffness of the equations.

As a conclusion we can state that an IMEX Runge-Kutta scheme coupled with the RS-IMEX splitting and the discontinuous Galerkin method is a suitable discretization for weakly compressible flows, i.e. for the case $\varepsilon \ll 1$. The case of a larger Mach number needs further investigation since then the numerical method is less efficient than a standard fully explicit scheme.

6. Conclusions and outlook

In this work we have investigated the RS-IMEX splitting coupled with high order IMEX Runge-Kutta schemes for different singularly perturbed differential equations. In this final chapter we shortly summarize the results of this thesis and give a detailed outlook on possible extensions for further research.

6.1. Conclusion

In the first part we have introduced the RS-IMEX splitting for ordinary differential equations to perform a comprehensive convergence analysis to prove the order of convergence of a specific class of IMEX Runge-Kutta schemes. From this analysis we were able to conclude the following:

- The resulting numerical method shows an order reduction which depends on the stage order of the implicit part. This result is similar to order reduction proofs for fully implicit Runge-Kutta schemes [81] and the standard splitting coupled with an IMEX Runge-Kutta scheme [24].

- The resulting numerical method shows a similar convergence behavior as a fully implicit discretization and an improved convergence behavior compared to the standard splitting.

In the subsequent numerical computations we were able to see the influence of order reduction on the overall convergence behavior of the numerical method. In addition, we obtained that the globally stiffly accurate property is needed to compute a stable approximation for large values of Δt. With this analysis we were able to show that the numerical discretization is asymptotically consistent, stable and depending on the IMEX Runge-Kutta scheme also more accurate.

The second part of this thesis was devoted to the extension of the RS-IMEX splitting to the isentropic Euler equations. In this step the RS-IMEX splitting is coupled with a high order IMEX Runge-Kutta scheme and a discontinuous Galerkin discretization. For the resulting method we were able to prove the asymptotic consistency, where the special numerical stabilization took a crucial role to show that the limiting densities are continuous in the fully-discrete case. Several numerical computations give the evidence that an asymptotically stable and accurate method for small Mach numbers is given. For large Mach numbers, there were several issues concerning accuracy which could be explained by stability issues for the convective CFL number, the considered examples or order reduction. To thoroughly justify these explanations further investigation is needed. Overall, we can conclude that the proposed method in this thesis is suitable for weakly compressible flows and performs well for small Mach numbers. This result is also verified by a different implementation, tested with the smooth vortex and more complex examples, of the proposed method in [P1, J5].

6.2. Outlook: order reduction for ordinary differential equations

We proved order reduction for the method proposed in this thesis, where we made some restrictions concerning the class of IMEX Runge-Kutta schemes we consider. Furthermore, the results for the ε independent components $y_{(0)}$ and $z_{(0)}$ could be more precise. Therefore, the analysis could be made more complete. Next to this, the results of Theorem 4.6 could be extended to different time integration methods. For example, an extension to general linear methods [31, 81, 92], for which an IMEX extension is available [32, 190], could be interesting.

6.2.1. Extension of Theorem 4.6 and numerical computations

In Theorem 4.6 we performed a detailed convergence analysis to show that IMEX Runge-Kutta methods coupled with the RS-IMEX splitting suffer from a similar order reduction as obtained by a fully implicit discretization. This analysis could be extended in two points:

1. We showed in Theorem 4.14 that the limiting solution is approximated with an accuracy in $\mathcal{O}(\Delta t^{r_1})$, where $r_1 = \min\{p, 2(q+1)\}$ with p the classical order of the method and q the stage order. Numerical experiments in [J4] indicate that an approximation with accuracy in $\mathcal{O}(\Delta t^p)$ is computed. Therefore, the proof of Theorem 4.14 might be extendable to prove this assumption.

2. We restricted ourselves to IMEX Runge-Kutta schemes which are globally stiffly accurate, of type CK and have a uniform $\boldsymbol{c}$. In a first step the results of Theorem 4.6 could be extended to IMEX Runge-Kutta schemes which are globally stiffly accurate, of type CK and do not have a uniform $\boldsymbol{c}$. In a second step also not globally stiffly accurate IMEX Runge-Kutta schemes can be considered. These methods are of special interest since we obtained in the numerical experiments that stability issues for large values of Δt can occur. It would be desirable to prove why these stability problems occur and to prove the behavior of these methods for $\varepsilon \ll 1$.

Finally, the numerical computations can be extended by considering more different IMEX Runge-Kutta schemes to see the influence of order reduction on the overall convergence behavior. Here, especially the case of very high order methods is interesting. Unfortunately, there are only a few IMEX Runge-Kutta schemes available which have a classical order of convergence larger than three. Therefore additional IMEX Runge-Kutta schemes with a high order of accuracy and ideally with a large implicit stage order are needed. One way to obtain such a scheme is the integral deferred correction procedure, see [37] and the references therein and [27, 36] for an IMEX extension. Integral deferred correction methods apply an IMEX Runge-Kutta scheme in a first step to the ordinary differential equation and then in a second step to a differential equation which describes the error between the classical solution and the numerical approximation. By this, the error can be reduced to obtain a very high order. Due to this structure the final method can be written as an IMEX Runge-Kutta scheme, see [27], but this scheme has many stages and is therefore not very efficient.

6.2.2. Different time integration methods

We have shown that the obtained order reduction depends on the stage-wise structure of the time discretization method, i.e. the order of convergence is reduced to the stage order. Unfortunately, for an IMEX Runge-Kutta scheme we can only obtain a stage order of at most one and of at most two for the implicit part. Therefore, we would always obtain order reduction for a high order scheme. The question is if this can be improved by a method which combines the ideas of linear multistep, where no order reduction is obtained, and Runge-Kutta schemes, where order reduction is obtained. Such a method is given by general linear methods [31, 81, 92] which can be seen as a class of time integration schemes where linear multistep and Runge-Kutta methods are sub-classes. In [32, 190] IMEX extensions of general linear methods are given. In [190] also a numerical investigation concerning order reduction for the van der Pol equation is done and no order reduction is obtained.

6.3. Outlook: weakly compressible flows

We have proven that the method proposed in this thesis leads to an asymptotically consistent discretization of the isentropic Euler equations. Numerical computations have shown good convergence results if ε is small. In the following we mention different topics which are useful next steps in the development of a high order method for weakly compressible flows. These topics are the extension to the full Euler equations, the efficiency of the method and adaptation to use explicit schemes for large Mach numbers and the method proposed in this thesis for low Mach numbers.

6.3.1. Full Euler equations

The RS-IMEX splitting can - in principle - be extended to the full Euler equations, which is shown in the following. Similarly to the isentropic Euler equations, the full Euler equations consist of the conservation of mass and momentum. Furthermore, the conservation of energy is added. In non-dimensional form the full Euler equations are given by

$$\partial_t \begin{pmatrix} \rho \\ \rho \boldsymbol{u} \\ E \end{pmatrix} + \nabla_{\boldsymbol{x}} \cdot \begin{pmatrix} \rho \boldsymbol{u} \\ \rho \boldsymbol{u} \otimes \boldsymbol{u} + \frac{1}{\varepsilon^2} p \, \mathrm{Id} \\ \boldsymbol{u}(E+p) \end{pmatrix} = 0, \tag{6.1}$$

where ρ denotes the mass density, $\rho \boldsymbol{u}$ the momentum density and E the energy density. The equations are closed with an equation of state for the pressure which is given by

$$p(\rho, \rho \boldsymbol{u}, E) = (\gamma - 1) \left(E - \frac{\varepsilon^2}{2} \rho \|\boldsymbol{u}\|^2 \right), \tag{6.2}$$

where $\gamma > 1$ is the adiabatic gas constant. In the following, we assume that the velocity $\boldsymbol{u}$ fulfills the boundary conditions

$$\int_{\partial \Omega} \boldsymbol{u} \cdot \boldsymbol{n} \mathrm{d}\sigma = \mathcal{O}(\varepsilon^2). \tag{6.3}$$

We first derive the $\varepsilon \to 0$ limit of the full Euler equations to obtain in which way a reference solution can be computed. Then we compute the RS-IMEX splitting and check if this splitting fulfills the conditions of Definition 3.8.

The incompressible limit as $\varepsilon \to 0$

We follow the same steps as for the isentropic Euler equations to obtain the $\varepsilon \to 0$ limit, see Corollary 2.24. Therefore, we assume that every quantity is given by an asymptotic expansion, i.e.

$$\begin{aligned} &\rho = \rho_{(0)} + \varepsilon \rho_{(1)} + \varepsilon^2 \rho_{(2)} + \mathcal{O}(\varepsilon^3), \\ &\boldsymbol{u} = \boldsymbol{u}_{(0)} + \varepsilon \boldsymbol{u}_{(1)} + \varepsilon^2 \boldsymbol{u}_{(2)} + \mathcal{O}(\varepsilon^3) \qquad \text{and} \\ &E = E_{(0)} + \varepsilon E_{(1)} + \varepsilon^2 E_{(2)} + \mathcal{O}(\varepsilon^3). \end{aligned} \tag{6.4}$$

We first insert these asymptotic expansions in the equation of state of the pressure p, which is given in Equation (6.2), and obtain

$$p(\rho, \rho \boldsymbol{u}, E) = \underbrace{(\gamma - 1) E_{(0)}}_{=: p_{(0)}} + \varepsilon \underbrace{(\gamma - 1) E_{(1)}}_{=: p_{(1)}} + \varepsilon^2 \underbrace{(\gamma - 1) \left(E_{(2)} + \rho_{(0)} \|\boldsymbol{u}_{(0)}\| \right)}_{=: p_{(2)}} + \mathcal{O}(\varepsilon^3).$$

We use this and again the asymptotic expansions given in Equation (6.4) for the conservation of momentum equation, rearrange the terms concerning their power in ε and derive different ε-independent equations which have to be fulfilled. The equations, which correspond to the $\mathcal{O}(\varepsilon^{-2})$ and the $\mathcal{O}(\varepsilon^{-1})$ terms, are given by

$$(\gamma - 1) \nabla_{\boldsymbol{x}} E_{(0)} = 0 \qquad \text{and} \qquad (\gamma - 1) \nabla_{\boldsymbol{x}} E_{(1)} = 0.$$

From this, we can directly conclude that the limiting energy densities $E_{(0)}$ and $E_{(1)}$ are constant in space. Next, we consider the conservation of energy equation, again insert the asymptotic expansion of all variables, and obtain from the $\mathcal{O}(1)$ terms

$$\partial_t E_{(0)} + \gamma \nabla_{\boldsymbol{x}} \cdot \left(\boldsymbol{u}_{(0)} E_{(0)} \right) = 0.$$

We integrate over the whole domain Ω and use integration by parts for the divergence operator. This results in

$$\begin{aligned} 0 &= \int_\Omega \partial_t E_{(0)} \mathrm{dx} + \gamma \int_\Omega \nabla_{\boldsymbol{x}} \cdot \left(\boldsymbol{u}_{(0)} E_{(0)}\right) \mathrm{dx} \\ &= \int_\Omega \partial_t E_{(0)} \mathrm{dx} - \gamma \int_{\partial\Omega} \left(\boldsymbol{u}_{(0)} E_{(0)}\right) \cdot \boldsymbol{n} \mathrm{d}\sigma. \end{aligned}$$

Since $E_{(0)}$ is constant in space and $\boldsymbol{u}_{(0)}$ fulfills the boundary conditions given in Equation (6.3), we obtain that $E_{(0)}$ is constant in time. Similarly, we can conclude that $E_{(1)}$ is constant in time. Then, the $\mathcal{O}(1)$ terms of the conservation of energy equation are given by

$$\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u}_{(0)} = 0.$$

Finally, together with the $\mathcal{O}(1)$ terms of the conservation of mass and momentum equation we obtain

$$\partial_t \begin{pmatrix} \rho_{(0)} \\ (\rho\boldsymbol{u})_{(0)} \\ 0 \end{pmatrix} + \nabla_{\boldsymbol{x}} \cdot \begin{pmatrix} \rho\boldsymbol{u}_{(0)} \\ (\rho\boldsymbol{u})_{(0)} \otimes \boldsymbol{u}_{(0)} + p_{(2)} \, \mathrm{Id} \\ \boldsymbol{u}_{(0)} \end{pmatrix} = 0,$$

which is consistent to the incompressible Euler equations with variable density, see e.g. [8, 115] for more details. For a rigorous proof of the $\varepsilon \to 0$ convergence behavior we refer to [158] and [5] for the case of the Navier-Stokes equations.

RS-IMEX splitting

From the previous analysis we have seen in which way the full Euler equations behave as $\varepsilon \to 0$. Thus, we can obtain a reference solution which fulfills

$$\rho_{ref} - \rho = \mathcal{O}(\varepsilon), \qquad (\rho\boldsymbol{u})_{ref} - \rho\boldsymbol{u} = \mathcal{O}(\varepsilon) \qquad \text{and} \qquad E_{ref} - E = \mathcal{O}(\varepsilon)$$

by choosing

$$\rho_{ref} := \rho_{(0)}, \qquad (\rho\boldsymbol{u})_{ref} := (\rho\boldsymbol{u})_{(0)} \qquad \text{and} \qquad E_{ref} := E_{(0)}.$$

Then, we can compute the RS-IMEX splitting functions similarly as we have done for the isentropic Euler equations. This step is quite straightforward by computing the derivative of the flux function and then computing the corresponding splitting functions, see also Remark 3.31. Next, we compute the eigenvalues of the Jacobian in normal direction of the explicit part to check if the splitting fulfills the conditions of Definition 3.8. These eigenvalues are given by

$$\begin{aligned} &\widehat{\lambda}_1 = 0, \qquad \widehat{\lambda}_2 = \gamma(\boldsymbol{u} - \boldsymbol{u}_{ref}) \cdot \boldsymbol{n} \qquad \text{and} \\ &\widehat{\lambda}_{3,4} = (2 - \frac{1}{2}\gamma)(\boldsymbol{u} - \boldsymbol{u}_{ref}) \cdot \boldsymbol{n} \pm \frac{1}{2}\sqrt{(4 - 4\gamma)\|\boldsymbol{u} - \boldsymbol{u}_{ref}\|^2 + \gamma^2((\boldsymbol{u} - \boldsymbol{u}_{ref}) \cdot \boldsymbol{n})^2}. \end{aligned}$$

First of all, we can directly see that they are non-stiff and fulfill

$$\widehat{\lambda}_i = \mathcal{O}(\varepsilon) \qquad \text{for} \qquad i = 1, 2, 3, 4,$$

which is similar to the explicit eigenvalues of the RS-IMEX splitting for the isentropic Euler equations given in Equation 3.26. On the other hand, we obtain a term in a square root, which depends on the normal vector:

– The first term, $(4 - 4\gamma)\|\boldsymbol{u} - \boldsymbol{u}_{ref}\|^2$, in the square root is negative since $\gamma > 1$.

– The second term, $\gamma^2 \left((\boldsymbol{u} - \boldsymbol{u}_{ref}) \cdot \boldsymbol{n}\right)^2$, in the square root can be small or even zero for an arbitrary normal vector $\boldsymbol{n}$.

Therefore, we can conclude that the eigenvalues become complex if $\boldsymbol{u} \neq \boldsymbol{u}_{ref}$ and the normal vector $\boldsymbol{n}$ is given in such a way that

$$|4 - 4\gamma| \|\boldsymbol{u} - \boldsymbol{u}_{ref}\|^2 > \gamma^2((\boldsymbol{u} - \boldsymbol{u}_{ref}) \cdot \boldsymbol{n})^2.$$

This cannot be prohibited for a general flow situation. Therefore, we obtain a splitting, where the explicit part is not guaranteed to be hyperbolic. This means that the splitting does not fulfill the conditions given in Definition 3.8. It is not clear in which way the numerical method behaves in this setting and consequently we do not know if the RS-IMEX splitting is applicable to the full Euler equations. Solving this issue is currently work in progress and there are two possible remedies:

- A possible solution is to consider a different discretization method for the explicit part. A conservation law with both complex and real eigenvalues is called conservation law of mixed type and such a conservation law is discretized in [97, 162] by adjusting methods for hyperbolic conservation laws.

- Another possible solution is to drop some of the explicit pressure terms. Due to the structure of the RS-IMEX splitting these terms are including the ε^{-2} scaling in $\mathcal{O}(\varepsilon^2)$ and therefore the added error is very small for a low Mach flow. A similar idea is for example used in [23].

6.3.2. Efficiency

The efficiency of a numerical method is very important, it is not useful to construct a method which needs more computational cost than established methods to obtain a similar error. Therefore, it is a canonical step to tune the efficiency of the method proposed in this work. This can be done in several different steps.

It is not clear which accuracy the reference solution needs to have. A first step concerning this is done in Chapter 5 by computing the reference solution with a second order method even for an overall fourth order method. Additionally, one can think about a reference solution which is obtained differently, e.g. by computing the mean value or the minimum. First computations with a reference solution which do not correspond to the asymptotic limit are given in [P1].

To solve the incompressible equation we used a fully implicit method. This leads to large computational cost, but there are several possibilities to reduce the cost:

- One can consider some kind of IMEX splitting of the incompressible equations. First computations with an IMEX splitting for the incompressible equation and a first order scheme are given in [J3].

- One might not need to use a discontinuous Galerkin method to obtain the reference solution. There are several methods which are especially designed for incompressible equations which might be more efficient than the one proposed in this work, e.g. there are pressure correction methods, see [75] and the references therein.

One expensive part of the computations is solving the system of equations resulting from the implicit discretization. For this, one can also think about possibilities to reduce the computational cost:

- For discontinuous Galerkin methods it has been shown that the size of the implicit system of equations can be reduced by adding variables defined on the skeleton of the grid. The resulting method is called hybridized discontinuous Galerkin method and can tremendously reduce the computational cost if a high order discretization is considered [94, 130], see also the references therein. Then, a combination of a discontinuous Galerkin method for the explicit part with a hybridized discontinuous Galerkin method for the implicit part is needed. Such combinations are for example given in [113] for incompressible equations and in [100] for the shallow water equations.

- The linear system of equations is extremely stiff for $\varepsilon \ll 1$ and therefore solving it is expensive. With the reference solution given we have information about the ε-independent and dominating part of the numerical approximation. Therefore, these information could be used for solving the system of equations for example as a preconditioning.

Finally, a comparison to more established methods is needed. Comparisons to different reference solutions and standard methods are done in [P1, J5]. Furthermore, the extension to the isentropic Navier-Stokes equations is directly possible. For the isentropic Navier-Stokes equations, first computations with the splitting given in [78], see also Equation (3.24), and a similar high order discretization are done in [C1], which can be used for a comparison in terms of accuracy.

6.3.3. Adaptive methods

In a general flow situation the Mach number can vary in the domain, i.e. in one part the Mach number is of order one and in another part the Mach number is very small. In such a situation one has one part where a fully explicit method is desirable and a low Mach method needs too much effort. Additionally, one has a part where a fully explicit method needs too much effort and a low Mach method is desirable. Therefore, a combination of both methods would be a canonical choice for this setting. Combining an explicit and an implicit method leads to an IMEX decomposition of the domain [99].

This strategy could be extended to the application of an explicit scheme for parts of the domain where a Mach number of order one is obtained and of an IMEX scheme coupled with the RS-IMEX splitting for the parts of the domain where a low Mach number is obtained. Such a combination is not straightforward since one does not know in which way the two schemes are coupled at the cell boundaries and how to efficiently compute the reference solution. Especially the second point is interesting since it is not useful to compute the reference solution during the whole process on the complete domain. Is there a way to compute the reference solution only on a small part which is needed by the RS-IMEX splitting?

6.4. Summary

Overall, we have shown that the RS-IMEX splitting coupled with a high order IMEX Runge-Kutta scheme can lead to a high order discretization in the setting of singularly perturbed differential equations. We obtained that the resulting method suffers from a less significant order reduction compared to a more standard splitting and provides good convergence results if coupled with a discontinuous Galerkin method for low Mach number flows.

Furthermore, we identified in this chapter several possible extensions of this work and figured out that the development of high order methods for weakly compressible flows is by no means finished. Especially, the extension of the RS-IMEX splitting to the full Euler equations and the comparison with more established methods is currently work in progress.

A. Appendix

A.1. IMEX Runge-Kutta schemes

For the sake of completeness, we summarize the used IMEX Runge-Kutta schemes in this appendix.

0	0	0	0	0	0	0	0
γ	0	γ	0	γ	γ	0	0
1	0	$1-\gamma$	γ	1	δ	1-δ	0
	0	$1-\gamma$	γ		δ	1-δ	0

Table A.1.: The Butcher tableaux of an IMEX Runge-Kutta scheme named ARS_222 [12]. The left and right tableaux give the implicit and explicit part, respectively. The scheme has stage order one, implicit stage order one and overall convergence order two. The constants are given by $\gamma = \frac{2-\sqrt{2}}{2}$ and $\delta = 1 - \frac{1}{2\gamma}$.

1/2	1/2	0	0	0	0	0	0	0	0
2/3	1/6	1/2	0	0	1/3	1/3	0	0	0
1/2	-1/2	1/2	1/2	0	1	1	0	0	0
1	3/2	-3/2	1/2	1/2	1	1/2	0	1/2	0
	3/2	-3/2	1/2	1/2		1/2	0	1/2	0

Table A.2.: The Butcher tableaux of an IMEX Runge-Kutta scheme named DPA_242 [57]. The left and right tableaux give the implicit and explicit part, respectively. The scheme has stage order one, implicit stage order one and overall convergence order two.

0	0	0	0	0	0	0	0	0	0	0	0
1/2	0	1/2	0	0	0	1/2	1/2	0	0	0	0
2/3	0	1/6	1/2	0	0	2/3	11/18	1/18	0	0	0
1/2	0	-1/2	1/2	1/2	0	1/2	5/6	-5/6	1/2	0	0
1	0	3/2	-3/2	1/2	1/2	1	1/4	7/4	3/4	-7/4	0
	0	3/2	-3/2	1/2	1/2		1/4	7/4	3/4	-7/4	0

Table A.3.: The Butcher tableaux of an IMEX Runge-Kutta scheme named ARS_443 [12]. The left and right tableaux give the implicit and explicit part, respectively. The scheme has stage order one, implicit stage order one and overall convergence order three.

						‖						
0	0	0	0	0	0	‖	0	0	0	0	0	0
1	1/2	1/2	0	0	0	‖	1	1	0	0	0	0
2/3	5/18	-1/9	1/2	0	0	‖	2/3	4/9	2/9	0	0	0
1	1/2	0	0	1/2	0	‖	1	1/4	0	3/4	0	0
1	1/4	0	3/4	-1/2	1/2	‖	1	1/4	0	3/4	0	0
	1/4	0	3/4	-1/2	1/2	‖		1/4	0	3/4	0	0

Table A.4.: The Butcher tableaux of an IMEX Runge-Kutta scheme named BPR_353 [26]. The left and right tableaux give the implicit and explicit part, respectively. The scheme has stage order one, implicit stage order two and overall convergence order three.

					‖					
α	α	0	0	0	‖	0	0	0	0	0
0	$-\alpha$	α	0	0	‖	0	0	0	0	0
1	0	$1-\alpha$	α	0	‖	1	0	1	0	0
1/2	β	η	$1/2-\beta-\eta-\alpha$	α	‖	1/2	0	1/4	1/4	0
	0	1/6	1/6	2/3	‖		0	1/6	1/6	2/3

Table A.5.: The Butcher tableaux of an IMEX Runge-Kutta scheme named SSP_433 [140]. The left and right tableaux give the implicit and explicit part, respectively. The scheme has stage order one, implicit stage order one and overall convergence order three. The constants are given by $\alpha = 0.24169426078821$, $\beta = 0.06042356519705$ and $\eta = 0.12915286960590$.

0	0	0	0	0	0
0.871733	0.435867	0.435867	0	0	0
0.871733	0.435867	0	0.435867	0	0
2.34021	-0.0667587	0	1.9711	0.435867	0
1	0.412898	0	0.19734	-0.0461045	0.435867
	0.412898	0	0.19734	-0.0461045	0.435867

0	0	0	0	0	0
0.871733	0.871733	0	0	0	0
0.871733	0.435867	0.435867	0	0	0
2.34021	-0.800998	0	3.14121	0	0
1	0.356753	-0.19734	0.881949	-0.0413622	0
	0.412898	0	0.19734	-0.0461045	0.435867

Table A.6.: The Butcher tableaux of an IMEX Runge-Kutta scheme named BHR_553 [25]. The upper and lower tableaux give the implicit and explicit part, respectively. The scheme has stage order one, implicit stage order two and overall convergence order three.

								‖								
0	0	0	0	0	0	0	0	‖	0	0	0	0	0	0	0	0
1/3	-1/6	1/2	0	0	0	0	0	‖	1/3	1/3	0	0	0	0	0	0
1/3	1/6	-1/3	1/2	0	0	0	0	‖	1/3	1/6	1/6	0	0	0	0	0
1/2	3/8	-3/8	0	1/2	0	0	0	‖	1/2	1/8	0	3/8	0	0	0	0
1/2	1/8	0	3/8	-1/2	1/2	0	0	‖	1/2	1/8	0	3/8	0	0	0	0
1	-1/2	0	3	-3	1	1/2	0	‖	1	1/2	0	-3/2	0	2	0	0
1	1/6	0	0	0	2/3	-1/2	2/3	‖	1	1/6	0	0	0	2/3	1/6	0
	1/6	0	0	0	2/3	-1/2	2/3	‖		1/6	0	0	0	2/3	1/6	0

Table A.7.: The Butcher tableaux of an IMEX Runge-Kutta scheme named ARK_4A2 [117]. The left and right tableaux give the implicit and explicit part, respectively. The scheme has stage order one, implicit stage order one and overall convergence order four.

B. Bibliography

Journal publications

[J1] K. Kaiser and J. Schütz. “A high-order method for weakly compressible flows”. In: *Communications in Computational Physics* 22.4 (2017), pp. 1150–1174.

[J2] K. Kaiser and J. Schütz. “Asymptotic error analysis of an IMEX Runge-Kutta method”. In: *Journal of Computational and Applied Mathematics* 33 (2018), pp. 139–154.

[J3] K. Kaiser, J. Schütz, R. Schöbel, and S. Noelle. “A new stable splitting for the isentropic Euler equations”. In: *Journal of Scientific Computing* 70 (2017), pp. 1390–1407.

[J4] J. Schütz and K. Kaiser. “A new stable splitting for singularly perturbed ODEs”. In: *Applied Numerical Mathematics* 107 (2016), pp. 18–33.

[J5] J. Zeifang, K. Kaiser, A. Beck, J. Schütz, and C.-D. Munz. “Efficient high-order discontinuous Galerkin computations of low Mach number flows”. In: *Communications in Applied Mathematics and Computational Science* 13.2 (2018), pp. 243–270.

Preprints

[P1] K. Kaiser, J. Zeifang, J. Schütz, A. Beck, and C.-D. Munz. “Comparison of different splitting techniques for the isentropic Euler equations”. In: *IGPM Preprint 476 / CMAT Preprint UP-18-01* (2018).

Conference proceedings

[C1] K. Kaiser and J. Schütz. “Asymptotic Preserving Discontinuous Galerkin Method”. In: *Conference Proceedings of the YIC GACM 2015*. 2015.

[C2] K. Kaiser and J. Schütz. “The Influence of the Asymptotic Regime on the RS-IMEX”. In: *Conference proceedings of the ECMI 2016, (in press)*. 2016.

[C3] J. Schütz, K. Kaiser, and S. Noelle. “The RS-IMEX splitting for the isentropic Euler equations”. In: *Conference Proceedings of the YIC GACM 2015*. 2015.

Further References

[1] S. Abarbanel, P. Duth, and D. Gottlieb. “Splitting methods for low Mach number Euler and Navier-Stokes equations”. In: *Computers & Fluids* 17.1 (1989), pp. 1–12.

[2] S. Abarbanel and D. Gottlieb. “Optimal Time Splitting for Two- and Three-Dimensional Navier-Stokes Equations with Mixed Derivatives”. In: *Journal of Computational Physics* 41 (1981), pp. 1–33.

[3] E. Abbate, A. Iollo, and G. Puppo. “An all-speed relaxation scheme for gases and compressible materials”. In: *Journal of Computational Physics* (2017).

[4] R. Abgrall. “High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices”. In: *Journal of Scientific Computing* 73.2-3 (2017), pp. 461–494.

[5] T. Alazard. “Low Mach number limit of the full Navier-Stokes equations”. In: *Archive for Rational Mechanics and Analysis* 180.1 (2006), pp. 1–73.

[6] N. Alkishriwi, M. Meinke, and W. Schröder. “A large-eddy simulation method for low Mach number flows using preconditioning and multigrid”. In: *Computers & fluids* 35.10 (2006), pp. 1126–1136.

[7] A. S. Almgren, J. B. Bell, C. A. Rendleman, and M. Zingale. “Low Mach number modeling of type Ia supernovae. I. Hydrodynamics”. In: *The Astrophysical Journal* 637.2 (2006), p. 922.

[8] J. D. Anderson. *Fundamentals of Aerodynamics*. McGraw Hill New York, 2001.

[9] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. “Unified Analysis of discontinuous Galerkin Methods for elliptic problems”. In: *SIAM Journal on Numerical Analysis* 39 (2002), pp. 1749–1779.

[10] U. M. Ascher, R. M. M. Mattheij, and R. D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. Vol. 13. SIAM, 1988.

[11] U. M. Ascher and L. R. Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*. Vol. 61. Siam, 1998.

[12] U. M. Ascher, S. Ruuth, and R. Spiteri. “Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations”. In: *Applied Numerical Mathematics* 25 (1997), pp. 151–167.

[13] U. M. Ascher, S. Ruuth, and B. Wetton. “Implicit-Explicit Methods for Time-Dependent Partial Differential Equations”. In: *SIAM Journal on Numerical Analysis* 32 (1995), pp. 797–823.

[14] S. Balay, J. Brown, K. Buschelman, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang. *PETSc Web page*.
http://www.mcs.anl.gov/petsc. 2011.

[15] S. Balay, W. D. Gropp, L. C. McInnes, and B. F. Smith. “Efficient Management of Parallelism in Object Oriented Numerical Software Libraries”. In: *Modern Software Tools in Scientific Computing*. Ed. by E. Arge, A. M. Bruaset, and H. P. Langtangen. Birkhäuser Press Boston, 1997, pp. 163–202.

[16] R. Baraille, G. Bourdin, F. Dubois, and A.-Y. Le Roux. “Une version à pas fractionnaires du schéma de Godunov pour l’hydrodynamique”. In: *Comptes Rendus de l’Académie des Sciences* 314 (1992), pp. 147–152.

[17] W. Barsukow, P. V. F. Edelmann, C. Klingenberg, F. Miczek, and F. K. Röpke. “A numerical scheme for the compressible low-Mach number regime of ideal fluid dynamics”. In: *Journal of Scientific Computing* 72.2 (2017), pp. 623–646.

[18] F. Bassi, C. D. Bartolo, R. Hartmann, and A. Nigro. “A discontinuous Galerkin method for inviscid low Mach number flows”. In: *Journal of Computational Physics* 228.11 (2009), pp. 3996–4011.

[19] F. Bassi and S. Rebay. “A High-Order Accurate discontinuous finite-element Method for the Numerical Solution of the Compressible Navier-Stokes Equations”. In: *Journal of Computational Physics* 131 (1997), pp. 267–279.

[20] P. Birken and A. Meister. “Stability of preconditioned finite volume schemes at low Mach numbers”. In: *BIT Numerical Mathematics* 45.3 (2005), pp. 463–480.

[21] G. Bispen. “IMEX Finite Volume Methods for the Shallow Water Equations”. PhD thesis. Johannes Gutenberg-Universität, 2015.

[22] G. Bispen, K. R. Arun, M. Lukáčová-Medvid’ová, and S. Noelle. “IMEX large time step finite volume methods for low Froude number shallow water flows”. In: *Communications in Computational Physics* 16 (2014), pp. 307–347.

[23] G. Bispen, M. Lukáčová-Medvid’ová, and L. Yelash. “Asymptotic preserving IMEX finite volume schemes for low Mach number Euler equations with gravitation”. In: *Journal of Computational Physics* 335 (2017), pp. 222–248.

[24] S. Boscarino. “Error Analysis of IMEX Runge-Kutta Methods Derived from Differential-Algebraic Systems”. In: *SIAM Journal on Numerical Analysis* 45 (2007), pp. 1600–1621.

[25] S. Boscarino. “On an accurate third order implicit-explicit Runge-Kutta method for stiff problems”. In: *Applied Numerical Mathematics* 59 (2009), pp. 1515–1528.

[26] S. Boscarino, L. Pareschi, and G. Russo. “Implicit-explicit Runge–Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit”. In: *SIAM Journal on Scientific Computing* 35.1 (2013), A22–A51.

[27] S. Boscarino, J. Qiu, and G. Russo. “Implicit-Explicit Integral Deferred Correction Methods for Stiff Problems”. In: *arXiv preprint arXiv:1701.04750* (2017).

[28] S. Boscarino, G. Russo, and L. Scandurra. “All Mach Number Second Order Semi-Implicit Scheme for the Euler Equations of Gasdynamics”. In: *arXiv preprint arXiv:1706.00272* (2017).

[29] F. Bouchut, C. Chalons, and S. Guisset. “An entropy satisfying two-speed relaxation system for the barotropic Euler equations. Application to the numerical approximation of low Mach number flows.” In: *HAL Preprint* (2017).

[30] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*. SIAM, 1995.

[31] J. C. Butcher. “General linear methods”. In: *Acta Numerica* 15 (2006), pp. 157–256.

[32] A. Cardone, Z. Jackiewicz, A. Sandu, and H. Zhang. “Construction of highly stable implicit-explicit general linear methods”. In: *Discrete and Continuous Dynamical Systems. Series S* 2015 (2015), pp. 185–194.

[33] C. Chalons, M. Girardin, and S. Kokh. “An all-regime Lagrange-Projection like scheme for the gas dynamics equations on unstructured meshes”. In: *Communications in Computational Physics* 20.1 (2016), pp. 188–233.

[34] K. H. Chent. “On solving the compressible Navier-Stokes equations for unsteady flows at very low Mach numbers”. In: *11th Computational Fluid Dynamics Conference* 3368 (1993).

[35] Y.-H. Choi and C.L. Merkle. “The Application of Preconditioning in Viscous Flows”. In: *Journal of Computational Physics* 105.2 (1993), pp. 207–223.

[36] A. Christlieb, M. Morton, B. Ong, and J-M. Qiu. “Semi-implicit integral deferred correction constructed with additive Runge-Kutta methods”. In: *Communications in Mathematical Sciences* 9 (2011), pp. 879–902.

[37] A. Christlieb, B. Ong, and J.-M. Qiu. “Integral deferred correction methods constructed with high order Runge-Kutta integrators”. In: *Mathematics of Computation* 79.270 (2010), pp. 761–783.

[38] B. Cockburn, S. Hou, and C.-W. Shu. “The Runge–Kutta Local Projection discontinuous Galerkin Finite Element Method for Conservation Laws IV: The Multidimensional Case”. In: *Mathematics of Computation* 54 (1990), pp. 545–581.

[39] B. Cockburn, G. Karniadakis, and C.-W. Shu. “The development of discontinuous Galerkin methods”. In: *discontinuous Galerkin Methods: Theory, Computation and Applications*. Springer, 1999.

[40] 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”. In: *Journal of Computational Physics* 84 (1989), pp. 90–113.

[41] B. Cockburn and C.-W. Shu. “The Local discontinuous Galerkin Method for Time-dependent convection-diffusion systems”. In: *SIAM Journal on Numerical Analysis* 35 (1998), pp. 2440–2463.

[42] B. Cockburn and C.-W. Shu. “The Runge-Kutta discontinuous Galerkin Method for conservation laws V: Multidimensional Systems”. In: *Mathematics of Computation* 141 (1998), pp. 199–224.

[43] B. Cockburn and C.-W. Shu. "The Runge-Kutta Local Projection P^1-discontinuous Galerkin Finite Element Method for Scalar Conservation Laws". In: *RAIRO Mathematical modelling and numerical analysis* 25 (1991), pp. 337–361.

[44] B. Cockburn and C.-W. Shu. "TVB Runge-Kutta Local Projection discontinuous Galerkin Finite Element Method for Conservation Laws II: General Framework". In: *Mathematics of Computation* 52 (1989), pp. 411–435.

[45] P. Colella and K. Pao. "A projection method for low speed flows". In: *Journal of Computational Physics* 149.2 (1999), pp. 245–269.

[46] F. Cordier, P. Degond, and A. Kumbaro. "An Asymptotic-Preserving all-speed scheme for the Euler and Navier-Stokes equations". In: *Journal of Computational Physics* 231 (2012), pp. 5685–5704.

[47] R. Courant, K. Friedrichs, and H. Lewy. "Über die partiellen Differenzengleichungen der mathematischen Physik". In: *Mathematische Annalen* 100.1 (1928), pp. 32–74.

[48] M. Crouzeix. "Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques". In: *Numerische Mathematik* 35.3 (1980), pp. 257–276.

[49] P. Degond, S. Jin, and J.-G. Liu. "Mach-number uniform asymptotic-preserving gauge schemes for compressible flows". In: *Bulletin-Institute of Mathematics Academia Sinica (New Series)* 2.4 (2007), pp. 851–892.

[50] P. Degond and M. Tang. "All speed scheme for the low Mach number limit of the Isentropic Euler equation". In: *Communications in Computational Physics* 10 (2011), pp. 1–31.

[51] S. Dellacherie. "Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number". In: *Journal of Computational Physics* 229.4 (2010), pp. 978–1016.

[52] B. Desjardins, E. Grenier, P.-L. Lions, and N. Masmoudi. "Incompressible limit for solutions of the isentropic Navier-Stokes equations with Dirichlet boundary conditions". In: *Journal de Mathématiques Pures et Appliquées* 78.5 (1999), pp. 461–471.

[53] O. Desjardins, G. Blanquart, G. Balarac, and H. Pitsch. "High order conservative finite difference scheme for variable density low Mach number turbulent flows". In: *Journal of Computational Physics* 227.15 (2008), pp. 7125–7159.

[54] D. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin Methods.* Vol. 69. Springer Science & Business Media, 2011.

[55] G. Dimarco, R. Loubère, and M.-H. Vignal. "Study of a new asymptotic preserving scheme for the Euler system in the low Mach number limit". In: *SIAM Journal on Scientific Computing* 39.5 (2017), A2099–A2128.

[56] G. Dimarco, L. Mieussens, and V. Rispoli. "An asymptotic preserving automatic domain decomposition method for the Vlasov-Poisson-BGK system with applications to plasmas". In: *Journal of Computational Physics* 274 (2014), pp. 122–139.

[57] G. Dimarco and L. Pareschi. "Asymptotic Preserving Implicit-Explicit Runge–Kutta Methods for Nonlinear Kinetic Equations". In: *SIAM Journal on Numerical Analysis* 51.2 (2013), pp. 1064–1087.

[58] G. Dimarco and L. Pareschi. "Exponential Runge-Kutta Methods for Stiff Kinetic Equations". In: *SIAM Journal on Numerical Analysis* 49.5 (2011), pp. 2057–2077.

[59] V. Dolejší and M. Feistauer. "Discontinuous Galerkin Method". In: *Analysis and Applications to Compressible Flow. Springer Series in Computational Mathematics* 48 (2015).

[60] M. Dumbser and C.-D. Munz. "Arbitrary High Order discontinuous Galerkin Schemes". In: *Numerical Methods for Hyperbolic and Kinetic Problems.* 2005, pp. 295–333.

[61] W. Eckhaus. *Matched asymptotic expansions and singular perturbations.* Vol. 6. North-Holland Publishing Company, 1973.

[62] W. H. Enright, T. E. Hull, and B. Lindberg. “Comparing numerical methods for stiff systems of ODEs”. In: *BIT Numerical Mathematics* 15.1 (1975), pp. 10–48.

[63] M. Feistauer, V. Dolejší, and V. Kučera. “On the discontinuous Galerkin method for the simulation of compressible flow with wide range of Mach numbers”. In: *Computing and Visualization in Science* 10.1 (2007), pp. 17–27.

[64] M. Feistauer and V. Kučera. “On a robust discontinuous Galerkin technique for the solution of compressible flow”. In: *Journal of Computational Physics* 224.1 (2007), pp. 208–221.

[65] F. Filbet and S. Jin. “A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources”. In: *Journal of Computational Physics* 229.20 (2010), pp. 7625–7648.

[66] R. Frolov. “An Efficient Algorithm for the Multicomponent Compressible Navier-Stokes Equations in Low and High Mach Number Regimes”. In: *arXiv preprint arXiv:1711.06211* (2017).

[67] G. Gassner, F. Lörcher, and C.-D. Munz. “A discontinuous Galerkin scheme based on a space-time expansion II. Viscous flow equations in multi dimensions”. In: *Journal of Scientific Computing* 34.3 (2008), pp. 260–286.

[68] C. W. Gear, T. J. Kaper, I. G. Kevrekidis, and A. Zagaris. “Projecting to a slow manifold: Singularly perturbed systems and legacy codes”. In: *SIAM Journal on Applied Dynamical Systems* 4.3 (2005), pp. 711–732.

[69] J.-F. Gerbeau and B. Perthame. “Derivation of Viscous Saint-Venant System for Laminar Shallow Water; Numerical Validation”. In: *Discrete and Continuous Dynamical Systems-Series B* 1.1 (2001), pp. 89–102.

[70] J. Giesselmann. “Low Mach asymptotic-preserving scheme for the Euler–Korteweg model”. In: *IMA Journal of Numerical Analysis* 35.2 (2014), pp. 802–833.

[71] F. X. Giraldo and M. Restelli. “High-order semi-implicit time-integrators for a triangular discontinuous Galerkin oceanic shallow water model”. In: *International Journal for Numerical Methods in Fluids* 63.9 (2010), pp. 1077–1102.

[72] F.X. Giraldo, M. Restelli, and M. Läuter. “Semi-implicit formulations of the Navier-Stokes equations: Application to nonhydrostatic atmospheric modeling”. In: *SIAM Journal on Scientific Computing* 32.6 (2010), pp. 3394–3425.

[73] S. Glegg and W. Devenport. *Aeroacoustics of low Mach number flows: fundamentals, analysis, and measurement*. Academic Press, 2017.

[74] E. Godlewski and P.-A. Raviart. *Hyperbolic systems of conservation laws*. Ellipses Paris, 1991.

[75] J.-L. Guermond, P. Minev, and J. Shen. “An overview of projection methods for incompressible flows”. In: *Computer Methods in Applied Mechanics and Engineering* 195.44 (2006), pp. 6011–6045.

[76] H. H. Guillard and A. Murrone. “On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes”. In: *Computers & Fluids* 33.4 (2004), pp. 655–675.

[77] H. Guillard and C. Viozat. “On the behavior of upwind schemes in the low Mach number limit”. In: *Computers & Fluids* 28.1 (1999), pp. 63–86.

[78] J. Haack, S. Jin, and J.-G. Liu. “An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations”. In: *Communications in Computational Physics* 12 (2012), pp. 955–980.

[79] E. Hairer, C. Lubich, and M. Roche. “Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations”. In: *BIT Numerical Mathematics* 28.3 (1988), pp. 678–700.

[80] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations I*. Springer Series in Computational Mathematics, 1987.

[81] E. Hairer and G. Wanner. *Solving ordinary differential equations II.* Springer Series in Computational Mathematics, 1991.

[82] Jan S. Hesthaven and Tim Warburton. *Nodal discontinuous Galerkin Methods: Algorithms, Analysis, and Applications.* Texts in Applied Mathematics 54. Springer Verlag, 2007.

[83] D. R. van der Heul, C. Vuik, and P. Wesseling. "A conservative pressure-correction method for flow at all speeds". In: *Computers & Fluids* 32.8 (2003), pp. 1113–1132.

[84] A. Hiltebrand and S. Mishra. *Efficient computation of all speed flows using an entropy stable shock-capturing space-time discontinuous Galerkin method.* Tech. rep. 2014-17. Switzerland: Seminar for Applied Mathematics, ETH Zürich, 2014.

[85] F. Hindenland, G. Gassner, C. Altmann, A. Beck, M. Staudenmaier, and C.-D. Munz. "Explicit discontinuous Galerkin methods for unsteady problems". In: *Computers & Fluids* 61 (2012), pp. 86–93.

[86] J. Hu, S. Jin, and Q. Li. "Asymptotic-preserving schemes for multiscale hyperbolic and kinetic equations". In: *Handbook of Numerical Analysis* 18 (2017), pp. 103–129.

[87] W. Hundsdorfer and J. Jaffré. "Implicit–explicit time stepping with spatial discontinuous finite elements". In: *Applied Numerical Mathematics* 45.2 (2003), pp. 231–254.

[88] W. Hundsdorfer and S.-J. Ruuth. "IMEX extensions of linear multistep methods with general monotonicity and boundedness properties". In: *Journal of Computational Physics* 225.2 (2007), pp. 2016–2042.

[89] D. Iampietro, F. Daude, P. Galon, and J.-M. Hérard. "A Mach-Sensitive Implicit-Explicit Scheme Adapted to Compressible Multi-scale Flows". In: *Journal of Computational and Applied Mathematics* 340 (2018), pp. 122–150.

[90] D. Iampietro, F. Daude, P. Galon, and J.-M. Hérard. "A Weighted Splitting Approach for Low-Mach Number Flows". In: *International Conference on Finite Volumes for Complex Applications.* Springer. 2017, pp. 3–11.

[91] G. Izzo and Z. Jackiewicz. "Highly stable implicit–explicit Runge–Kutta methods". In: *Applied Numerical Mathematics* 113 (2017), pp. 71–92.

[92] Z. Jackiewicz. *General linear methods for ordinary differential equations.* John Wiley & Sons, 2009.

[93] A. Jameson. *Time dependent calculations using multigrid, with applications to unsteady flows past airfoils and wings.* AIAA Paper 91-1596. 1991.

[94] A. Jaust and J. Schütz. "A temporally adaptive hybridized discontinuous Galerkin method for time-dependent compressible flows". In: *Computers & Fluids* 98 (2014), pp. 177–185.

[95] S. Jin. "Asymptotic Preserving (AP) Schemes for multiscale kinetic and hyperbolic equations: A review". In: *Rivista di Matematica della Universita Parma* 3 (2012), pp. 177–216.

[96] S. Jin. "Efficient Asymptotic-Preserving (AP) schemes for some multiscale kinetic equations". In: *SIAM Journal on Scientific Computing* 21 (1999), pp. 441–454.

[97] S. Jin. "Numerical integrations of systems of conservation laws of mixed type". In: *SIAM Journal on applied Mathematics* 55.6 (1995), pp. 1536–1551.

[98] M. Junk. "Kinetic schemes in the case of low Mach numbers". In: *Journal of Computational Physics* 151.2 (1999), pp. 947–968.

[99] A. Kanevsky, M. H. Carpenter, D. Gottlieb, and J. S. Hesthaven. "Application of implicit-explicit high order Runge-Kutta methods to discontinuous-Galerkin schemes". In: *Journal of Computational Physics* 225.2 (2007), pp. 1753–1781.

[100] S. Kang, F. X. Giraldo, and T. Bui-Thanh. “IMEX HDG-DG: a coupled implicit hybridized discontinuous Galerkin (HDG) and explicit discontinuous Galerkin (DG) approach for shallow water systems”. In: *arXiv preprint arXiv:1711.02751* (2017).

[101] C. A. Kennedy and M. H. Carpenter. “Additive Runge-Kutta schemes for convection-diffusion-reaction equations”. In: *Applied Numerical Mathematics* 44 (2003), pp. 139–181.

[102] I. J. Keshtiban, F. Belblidia, and M. F. Webster. “Compressible flow solvers for low Mach number flows - a review”. In: *International Journal for Numerical Methods in Fluids* 23 (2004), pp. 77–103.

[103] J. Kevorkian and J. D. Cole. *Perturbation Methods in Applied Mathematics.* Springer Berlin / Heidelberg / New York, 1981.

[104] S. Klainerman and A. Majda. “Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids”. In: *Communications on Pure and Applied Mathematics* 34 (1981), pp. 481–524.

[105] B. Klein. “A high-order discontinuous Galerkin solver for incompressible and low-Mach number flows”. PhD thesis. Technischen Universität Darmstadt, 2015.

[106] B. Klein, B. Müller, F. Kummer, and M. Oberlack. “A high-order discontinuous Galerkin solver for low Mach number flows”. In: *International Journal for Numerical Methods in Fluids* 81.8 (2016), pp. 489–520.

[107] R. Klein. “Asymptotics, structure, and integration of sound-proof atmospheric flow equations”. In: *Theoretical and Computational Fluid Dynamics* 23.3 (2009), pp. 161–195.

[108] R. Klein. “Semi-Implicit Extension of a Godunov-Type Scheme Based on Low Mach Number Asymptotics I: One-Dimensional Flow”. In: *Journal of Computational Physics* 121 (1995), pp. 213–237.

[109] R. Klein, N. Botta, T. Schneider, C.-D. Munz, S. Roller, A. Meister, L. Hoffmann, and T. Sonar. “Asymptotic adaptive methods for multi-scale problems in fluid mechanics”. In: *Journal of Engineering Mathematics* 39.1 (2001), pp. 261–343.

[110] T. Kloczko, C. Corre, and A. Beccantini. “Low-cost implicit schemes for all-speed flows on unstructured meshes”. In: *International Journal for Numerical Methods in Fluids* 58.5 (2008), pp. 493–526.

[111] D. Kröner. *Numerical Schemes for Conservation Laws.* Wiley Teubner, 1997.

[112] B. van Leer, J. L. Thomas, P. L. Roe, and R. W. Newsome. “A comparison of numerical flux formulas for the Euler and Navier-Stokes equations”. In: *8th Computational Fluid Dynamics Conference* (1987).

[113] C. Lehrenfeld and J. Schöberl. “High order exactly divergence-free hybrid discontinuous Galerkin methods for unsteady incompressible flows”. In: *Computer Methods in Applied Mechanics and Engineering* 307 (2016), pp. 339–361.

[114] R. J. LeVeque. *Numerical Methods for Conservation Laws.* Birkhäuser Basel, 1990.

[115] P.-L. Lions. *Mathematical topics in fluid mechanics. Vol. 1, Incompressible models.* Oxford Science Publication, Oxford, 1996.

[116] P.-L. Lions. *Mathematical topics in fluid mechanics, Vol. 2, Compressible models.* Oxford Science Publication, Oxford, 1998.

[117] H. Liu and J. Zou. “Some new additive Runge–Kutta methods and their applications”. In: *Journal of Computational and Applied Mathematics* 190.1-2 (2006), pp. 74–98.

[118] G. S. S. Ludford. *Reacting flows: combustion and chemical reactors.* American Mathematical Society, 1986.

[119] H. Luo, J. D. Baum, and R. Löhner. “Extension of Harten-Lax-van Leer scheme for flows at all speeds”. In: *AIAA journal* 43.6 (2005), pp. 1160–1166.

[120] O. Le Maıtre, J. Levin, M. Iskandarani, and O. M. Knio. “A Multiscale Pressure Splitting of the Shallow-Water Equations: I. Formulation and 1D Tests”. In: *Journal of Computational Physics* 166.1 (2001), pp. 116–151.

[121] A. Majda. *Compressible fluid flow and systems of conservation laws in several space variables.* Vol. 53. Springer Science & Business Media, 2012.

[122] MATLAB. *MATLAB and Statistics Toolbox Release 2017a.* Natick, Massachusetts: The MathWorks Inc., 2017.

[123] R. M. M. Mattheij, S. W. Rienstra, and J. H. M. ten Thije Boonkkamp. *Partial differential equations: modeling, analysis, computation.* SIAM, 2005.

[124] A. Meister and J. Struckmeier. *Hyperbolic partial differential equations: theory, numerics and applications.* Springer Science & Business Media, 2012.

[125] G. Métivier and S. Schochet. “The incompressible limit of the non-isentropic Euler equations”. In: *Archive for Rational Mechanics and Analysis* 158.1 (2001), pp. 61–90.

[126] F. Miczek, F. K. Röpke, and P. V. F. Edelmann. “New numerical solver for flows at various Mach numbers”. In: *Astronomy & Astrophysics* 576 (2015), A50.

[127] C.-D. Munz, M. Dumbser, and M. Zucchini. “The multiple pressure variables method for fluid dynamics and aeroacoustics at low Mach numbers”. In: *Numerical Methods for Hyperbolic and Kinetic Problems* (2005), pp. 335–359.

[128] C.-D. Munz, S. Roller, R. Klein, and K. J. Geratz. “The extension of incompressible flow solvers to the weakly compressible regime”. In: *Computers & Fluids* 32.2 (2003), pp. 173–196.

[129] A. Murrone and H. Guillard. “Behavior of upwind scheme in the low Mach number limit: III. Preconditioned dissipation for a five equation two phase model”. In: *Computers & Fluids* 37.10 (2008), pp. 1209–1224.

[130] N. C. Nguyen and J. Peraire. “Hybridizable discontinuous Galerkin methods for partial differential equations in continuum mechanics”. In: *Journal of Computational Physics* 231 (2012), pp. 5955–5988.

[131] F. Nicoud. “Conservative high-order finite-difference schemes for low-Mach number flows”. In: *Journal of Computational Physics* 158.1 (2000), pp. 71–97.

[132] A. Nigro, C. De Bartolo, R. Hartmann, and F. Bassi. “Discontinuous Galerkin solution of preconditioned Euler equations for very low Mach number flows”. In: *International Journal for Numerical Methods in Fluids* 63.4 (2010), pp. 449–467.

[133] A. Nigro, S. Renda, C. De Bartolo, R. Hartmann, and F. Bassi. “A high-order accurate discontinuous Galerkin finite element method for laminar low Mach number flows”. In: *International Journal for Numerical Methods in Fluids* 72.1 (2013), pp. 43–68.

[134] S. Noelle. “On the derivation of the Discontinuous Galerkin method for hyperbolic conservation laws”. In: *IGPM Preprint Nr. 470* (2017).

[135] S. Noelle, G. Bispen, K.R. Arun, M. Lukáčová-Medvid’ová, and C.-D. Munz. “A Weakly Asymptotic Preserving Low Mach Number Scheme for the Euler Equations of Gas Dynamics”. In: *SIAM Journal on Scientific Computing* 36 (2014), B989–B1024.

[136] R. E. O’Malley. *Introduction to singular perturbations.* Academic Press, 1974.

[137] R. E. O’Malley. *Singular perturbation methods for ordinary differential equations.* Vol. 89. Springer Science & Business Media, 2012.

[138] P.H. Oosthuizen and W.E. Carscallen. *Introduction to Compressible Fluid Flow, Second Edition*. Heat Transfer. Taylor & Francis, 2013.

[139] K. Oßwald, A. Siegmund, P. Birken, V. Hannemann, and A. Meister. "L^2Roe: a low dissipation version of Roe's approximate Riemann solver for low Mach numbers". In: *International Journal for Numerical Methods in Fluids* 81.2 (2016), pp. 71–86.

[140] L. Pareschi and G. Russo. "High order asymptotically strong-stability-preserving methods for hyperbolic systems with stiff relaxation". In: *Hyperbolic Problems: Theory, Numerics, Applications*. Springer, 2003, pp. 241–251.

[141] L. Pareschi and G. Russo. "Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation". In: *Journal of Scientific Computing* 25.1-2 (2005), pp. 129–155.

[142] J. H. Park and C.-D. Munz. "Multiple pressure variables methods for fluid flow at all Mach numbers". In: *International Journal for Numerical Methods in Fluids* 49.8 (2005), pp. 905–931.

[143] S. H. Park, J. E. Lee, and J. H. Kwon. "Preconditioned HLLE method for flows at all Mach numbers". In: *AIAA journal* 44.11 (2006), p. 2645.

[144] S. V. Patankar and B. D. Spalding. "A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows". In: *International Journal of Heat and Mass Transfer* 15.10 (1972), pp. 1787–1806.

[145] N. Peters. *Turbulent combustion. Cambridge Monographs on Mechanics*. Cambridge University Press, 2000.

[146] F. Renac. "A robust high-order discontinuous Galerkin method with large time steps for the compressible Euler equations". In: *Communications in Mathematical Sciences* 15.3 (2017), pp. 813–837.

[147] S. M. Renda, R. Hartmann, C. De Bartolo, and M. Wallraff. "A high-order discontinuous Galerkin method for all-speed flows". In: *International Journal for Numerical Methods in Fluids* 77.4 (2015), pp. 224–247.

[148] M. Restelli. "Semi-Lagrangian and semi-implicit discontinuous Galerkin methods for atmospheric modeling applications". In: *PhD thesis Politecnico di Milano* (2007).

[149] M. Ricchiuto and A. Bollermann. "Stabilized residual distribution for shallow water simulations". In: *Journal of Computational Physics* 228(4).4 (2009), pp. 1071–1115.

[150] F. Rieper. "A low-Mach number fix for Roe's approximate Riemann solver". In: *Journal of Computational Physics* 230.13 (2011), pp. 5263–5287.

[151] B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations: theory and implementation*. SIAM, 2008.

[152] S. Roller and C.-D. Munz. "A low Mach number scheme based on multi-scale asymptotics". In: *Computing and Visualization in Science* 3.1-2 (2000), pp. 85–91.

[153] C.-C. Rossow. "A flux-splitting scheme for compressible and incompressible flows". In: *Journal of Computational Physics* 164.1 (2000), pp. 104–122.

[154] V. V. Rusanov. "The calculation of the interaction of non-stationary shock waves and obstacles". In: *USSR Computational Mathematics and Mathematical Physics* 1.2 (1962), pp. 304–320.

[155] M. Sabanca, G. Brenner, and N. Alemdaroğlu. "Improvements to compressible Euler methods for low-Mach number flows". In: *International Journal for Numerical Methods in Fluids* 34.2 (2000), pp. 167–185.

[156] J. Schöberl. "C++11 Implementation of Finite Elements in NGSolve". In: *Institute for Analysis and Scientific Computing, Vienna University of Technology ASC Report* 30 (2014).

[157] J. Schöberl. “NETGEN - An advancing front 2D/3D-mesh generator based on abstract rules”. In: *Computing and Visualization in Science* 1 (1997), pp. 41–52.

[158] S. Schochet. “The compressible Euler equations in a bounded domain: existence of solutions and the incompressible limit”. In: *Communications in Mathematical Physics* 104.1 (1986), pp. 49–75.

[159] S. Schochet. “The mathematical theory of low Mach number flows”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 39.03 (2005), pp. 441–458.

[160] J. Schütz and S. Noelle. “Flux Splitting for stiff equations: A notion on stability”. In: *Journal of Scientific Computing* 64.2 (2015), pp. 522–540.

[161] J. Sesterhenn, B. Müller, and H. Thomann. “Flux-Vector Splitting for Compressible Low Mach Number Flow”. In: *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects.* Springer, 1993, pp. 528–535.

[162] C.-W. Shu. “A numerical method for systems of conservation laws of mixed type admitting hyperbolic flux splitting”. In: *Journal of Computational Physics* 100.2 (1992), pp. 209–436.

[163] C.-W. Shu. “High-order finite difference and finite volume WENO schemes and discontinuous Galerkin methods for CFD”. In: *International Journal of Computational Fluid Dynamics* 17.2 (2003), pp. 107–118.

[164] G. Stecca, A. Siviglia, and E. F. Toro. “A Finite Volume Upwind-Biased Centred Scheme for Hyperbolic Systems of Conservation Laws: Application to Shallow Water Equations”. In: *Communications in Computational Physics* 12.4 (2012), pp. 1183–1214.

[165] D. Sun, C. Yan, F. Qu, and R. Du. “A robust flux splitting method with low dissipation for all-speed flows”. In: *International Journal for Numerical Methods in Fluids* 84.1 (2017), pp. 3–18.

[166] M. Tavelli and M. Dumbser. “A pressure-based semi-implicit space–time discontinuous Galerkin method on staggered unstructured meshes for the solution of the compressible Navier–Stokes equations at all Mach numbers”. In: *Journal of Computational Physics* 341 (2017), pp. 341–376.

[167] *The GCC Quad-Precision Math Library.* Free Software Foundation. 2010.

[168] E. F. Toro and M. E. Vázquez-Cendón. “Flux splitting schemes for the Euler equations”. In: *Computers & Fluids* 70 (2012), pp. 1–12.

[169] Y.-Y. Tsui and T.-C. Wu. “A pressure-based unstructured-grid algorithm using high-resolution schemes for all-speed flows”. In: *Numerical Heat Transfer, Part B: Fundamentals* 53.1 (2008), pp. 75–96.

[170] E. Turkel. “Preconditioned methods for solving the incompressible and low speed compressible equations”. In: *Journal of Computational Physics* 72.2 (1987), pp. 277–298.

[171] E. Turkel, V. N. Vatsa, and R. Radespiel. *Preconditioning Methods for Low-Speed Flows.* Tech. rep. Institute for Computer Applications in Science and Engineering, 1996.

[172] J. J. W. Van der Vegt and H. Van der Ven. “Space–time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows: I. General formulation”. In: *Journal of Computational Physics* 182.2 (2002), pp. 546–585.

[173] F. Verhulst. *Methods and applications of singular perturbations: boundary layers and multiple timescale dynamics.* Vol. 50. Springer Science & Business Media, 2005.

[174] D. Vigneron, J.-M. Vaassen, and J.-A. Essers. “An implicit finite volume method for the solution of 3d low mach number viscous flows using a local preconditioning technique”. In: *Journal of Computational and Applied Mathematics* 215.2 (2008), pp. 610–617.

[175] C. Viozat. “Implicit upwind schemes for low Mach number compressible flows”. PhD thesis. Inria, 1997.

[176] G. Volpe. “Performance of compressible flow codes at low Mach numbers”. In: *AIAA Journal* 31.1 (1993), pp. 49–56.

[177] C. B. Vreugdenhil. *Numerical methods for shallow-water flow.* Vol. 13. Springer Science & Business Media, 2013.

[178] C. Wall, C. D. Pierce, and P. Moin. “A semi-implicit method for resolution of acoustic waves in low Mach number flows”. In: *Journal of Computational Physics* 181.2 (2002), pp. 545–563.

[179] H. Wang, C.-W. Shu, and Q. Zhang. “Stability and Error Estimates of Local Discontinuous Galerkin Methods with Implicit-Explicit Time-Marching for Advection-Diffusion Problems”. In: *SIAM Journal on Numerical Analysis* 53.1 (2015), pp. 206–227.

[180] R. F. Warming and B. J. Hyett. “The Modified Equation Approach to the Stability and Accuracy of Finite-Difference Methods”. In: *Journal of Computational Physics* 14 (1974), pp. 159–179.

[181] P. Wesseling. *Principles of Computational Fluid Dynamics.* Ed. by R. Bank, R. L. Graham, J. Stoer, and R. Varga. Vol. 29. Springer Series in Computational Mechanics. Springer Verlag, 2001.

[182] S. Yakovlev, D. Moxey, R. Kirby, and S. Sherwin. “To CG or to HDG: A Comparative Study in 3D”. In: *Journal of Scientific Computing* 67.1 (2015), pp. 192–220.

[183] L. Yelash, A. Müller, M. Lukáčová-Medvid’ová, F. X. Giraldo, and V. Wirth. “Adaptive discontinuous evolution Galerkin method for dry atmospheric flow”. In: *Journal of Computational Physics* 268 (2014), pp. 106–133.

[184] W.-A. Yong. “A note on the zero Mach number limit of compressible Euler equations”. In: *Proceedings of the American Mathematical Society* 133.10 (2005), pp. 3079–3085.

[185] H. Zakerzadeh. “Asymptotic analysis of the RS-IMEX scheme for the shallow water equations in one space dimension”. In: *IGPM Preprint Nr. 455* (2016).

[186] H. Zakerzadeh. “Asymptotic consistency of the RS-IMEX scheme for the low-Froude shallow water equations: analysis and numerics”. In: *XVI International Conference on Hyperbolic Problems: Theory, Numerics, Applications.* 2016.

[187] H. Zakerzadeh. “On the Mach-uniformity of the Lagrange-Projection scheme”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 51.4 (2017), pp. 1343–1366.

[188] H. Zakerzadeh. “The RS-IMEX scheme for the rotating shallow water equations with the Coriolis force”. In: *International Conference on Finite Volumes for Complex Applications.* Springer. 2017, pp. 199–207.

[189] H. Zakerzadeh and S. Noelle. “A note on the stability of implicit-explicit flux splittings for stiff hyperbolic systems”. In: *IGPM Preprint Nr. 449* (2016).

[190] H. Zhang, A. Sandu, and S. Blaise. “Partitioned and Implicit–Explicit General Linear Methods for Ordinary Differential Equations”. In: *Journal of Scientific Computing* 61.1 (2014), pp. 119–144.