Doctoral Thesis

Entropy-stable discontinuous Galerkin finite element methods
with streamline diffusion and shock-capturing for hyperbolic
systems of conservation laws

Author(s):
Hiltebrand, Andreas

Publication Date:
2014

Permanent Link:
https://doi.org/10.3929/ethz-a-010337672

Rights / License:
In Copyright - Non-Commercial Use Permitted
ENTROPY-STABLE DISCONTINUOUS GALERKIN
FINITE ELEMENT METHODS
WITH STREAMLINE DIFFUSION AND SHOCK-CAPTURING
FOR HYPERBOLIC SYSTEMS OF CONSERVATION LAWS

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
ANDREAS EDUARD HILTEBRAND
MSc ETH CSE, ETH Zurich
born on 17.03.1987
citizen of Bachenbülach ZH, Switzerland

accepted on the recommendation of
Prof. Dr. Siddhartha Mishra, ETH Zurich, examiner
Prof. Dr. Rémi Abgrall, University of Zurich, co-examiner
Prof. Dr. Charalambos Makridakis, University of Sussex, co-examiner

2014
Abstract

We propose and analyse a space-time discontinuous Galerkin (DG) finite element method for hyperbolic conservation laws. Entropy stability is obtained by discretising the entropy variables and using entropy-stable numerical fluxes. As solutions of hyperbolic conservation laws can develop discontinuities (shocks) in finite time, we include a streamline diffusion and a shock-capturing term to suppress spurious oscillations in the vicinity of shocks. We show that the approximate solutions converge to an entropy measure-valued solution for systems of conservation laws.

Algorithmically, the discretisation leads to a big nonlinear system of algebraic equations in each time step. A Newton-Krylov method is applied to solve it. However, for efficiency reasons, a preconditioner must be used. We design and apply block Jacobi and block Gauss-Seidel type preconditioners and investigate their performance both analytically (in a simplified setting) and experimentally.

The space-time formulation allows an easy local adaptation of the mesh in space and time. We consider residual-based adaptivity as well as goal-oriented adaptivity using duality estimates.

Due to the implicitness of the scheme, there is no CFL condition that must be satisfied. Therefore, the scheme can also be used in problems with multiple time scales, such as flows near the incompressible limit.

The high resolution and robustness properties of the method are demonstrated in several experiments. We consider, among others, the linear advection equation, Burgers’ equation, the wave equation, and the Euler equations in one or two spatial dimensions. These examples show that also problems with complicated boundaries can be solved, thanks to the finite element formulation.
Zusammenfassung


I would like to thank all the people that have made this thesis possible either directly or indirectly. First of all, I would like to thank Siddhartha Mishra for all the inspiring ideas, for the insights into the theory given to me, and for all the time spend in discussions and in assisting me in any respect.

Furthermore, I would like to mention and thank the following people explicitly: Blanca Ayuso de Dios for the help in preconditioning and for pointing out that superapproximation can be used, Franziska Weber for her patience in seemingly endless discussions on projection errors and for enabling me to understand certain aspects of the theory more thoroughly, and Sandra May for posing apparently simple questions that highlighted fundamental properties of the schemes and allow to embed the scheme into the existing work. Moreover, I would like to thank her, Deep Ray, and Ulrik Fjordholm for critical reading of the thesis, which has improved it a lot.

I would like to thank all the SAMies that have made the stay at the institute a pleasant time in my life. This includes the discussions on various topics, being scientific or non-scientific.

Last but not least, I would like to thank my family. Without their support, which enabled me to concentrate my work, this thesis would just not have been possible.

\(^1\)The research was funded in part by ERC STG. NN 306279 SPARCCLE.
## Contents

- Abstract iii
- Zusammenfassung v
- Acknowledgements vii
- Introduction xiii
- Notation xvii

1. Hyperbolic conservation laws 1
   1.1. Basic notions .................................................... 1
       1.1.1. Integral formulation of conservation laws ............... 1
       1.1.2. Differential formulation of conservation laws ........... 1
       1.1.3. Hyperbolicity ............................................... 2
       1.1.4. Weak solutions ............................................ 2
       1.1.5. Entropy conditions ....................................... 3
       1.1.6. Entropy measure-valued solutions ......................... 4
   1.2. Examples ...................................................... 5
       1.2.1. Scalar conservation laws .................................. 5
       1.2.2. Linear symmetrisable systems ............................. 6
       1.2.3. Nonlinear systems of conservation laws ................. 7

   2.1. Finite volume schemes ......................................... 10
       2.1.1. High-resolution schemes .................................. 10
       2.1.2. Entropy-stable finite volume schemes .................... 11
   2.2. Discontinuous Galerkin schemes ................................ 12

3. The shock-capturing streamline diffusion DG formulation 15
   3.1. The mesh ..................................................... 15
   3.2. Variational formulation ....................................... 15
   3.3. The DG quasilinear form ...................................... 16
       3.3.1. Numerical fluxes .......................................... 16
       3.3.2. Entropy-conservative flux ................................ 18
       3.3.3. Numerical diffusion operators ........................... 18
Contents

3.4. Streamline diffusion operator ........................................ 19
3.5. Shock-capturing operator .................................................. 20
3.6. Entropy stability for nonlinear systems .................................. 21

4. Convergence analysis for systems of conservation laws ........ 31
  4.1. Convergence to measure-valued solution ........................... 31
  4.2. Entropy consistency ..................................................... 38
  4.3. Convergence for scalar conservation laws .......................... 46
    4.3.1. Mathematical Framework ......................................... 46
    4.3.2. Strong convergence of approximate solutions .................. 46
    4.3.3. Convergence analysis for scalar conservation laws .......... 63
  4.4. Convergence for linear symmetrisable systems .................... 65

5. Numerical experiments ..................................................... 67
  5.1. Wave equation .......................................................... 67
    5.1.1. Exact form of equations, entropy formulation ............... 67
    5.1.2. Smooth data: convergence rates ............................... 67
    5.1.3. Discontinuous data ............................................. 69
    5.1.4. Two space dimensions .......................................... 70
  5.2. Euler equations ........................................................ 72
    5.2.1. Entropy-conservative fluxes, numerical diffusion operators 72
    5.2.2. One space dimension ............................................ 72
    5.2.3. Sod shock tube ................................................. 72
    5.2.4. Pressure scaling ................................................ 72
    5.2.5. Lax shock tube ................................................. 73
  5.3. Two space dimensions .................................................. 75
    5.3.1. Vortex advection ............................................... 75
    5.3.2. Radial shock tube .............................................. 77
    5.3.3. NACA 0012 aerofoil ......................................... 77

6. The computational method .................................................. 81
  6.1. Choice of basis functions ............................................. 81
  6.2. Nonlinear and linear solvers ......................................... 82
  6.3. The mesh and mesh generation ....................................... 83
  6.4. Quadrature rules ....................................................... 84
  6.5. Implementation .......................................................... 84

7. Preconditioning .............................................................. 85
  7.1. Block Jacobi preconditioner .......................................... 85
    7.1.1. Description ..................................................... 85
    7.1.2. Fourier analysis ................................................ 85
    7.1.3. Numerical results .............................................. 91
Contents

7.1.4. CFL dependence ................................................. 98
7.2. Block Gauss-Seidel preconditioners .................................. 100
  7.2.1. Analysis of the spectrum ..................................... 101
  7.2.2. Numerical experiments ...................................... 101

8. All-speed flows ....................................................... 107
  8.1. Two-speed advection ............................................ 107
  8.2. 1-D Euler equations ............................................ 111
  8.3. Flow around a cylinder ........................................ 112

9. Adaptivity .............................................................. 117
  9.1. Space-time mesh ................................................ 117
  9.2. Formulation of the method ...................................... 118
  9.3. Implementation aspects ........................................ 120
    9.3.1. Time marching ............................................. 120
    9.3.2. Preconditioning ........................................... 121
    9.3.3. Parallelisation ............................................ 121
  9.4. Residual based adaptivity ...................................... 121
    9.4.1. Linear advection equation ................................ 121
    9.4.2. Sod shock tube ............................................ 122
    9.4.3. Forward facing step flow .................................. 123
    9.4.4. Aerofoil flow ............................................. 126
  9.5. Goal-oriented adaptivity ...................................... 127
    9.5.1. Linear case ............................................... 128
    9.5.2. The nonlinear case ....................................... 128
    9.5.3. Computable estimates and refinement strategy .......... 129
    9.5.4. Linear advection equation ................................ 130
    9.5.5. Burgers’ equation ........................................ 134
    9.5.6. Sod shock tube ............................................ 137
    9.5.7. Lift computation on a NACA 0012 aerofoil .............. 138

10. Conclusions .......................................................... 141
  10.1. Future work ................................................... 142

A. Appendix ............................................................... 145
  A.1. Superapproximation ............................................. 145

. Bibliography ............................................................ 152
Introduction

Conservation is a fundamental phenomenon in physics. Conservation of mass, momentum, energy, etc. are observed everywhere in nature and are the cornerstone for many models that have been successfully used in practice. Adhering to the continuum theory, we can formulate a conservation law as follows [8]:

The amount of a conserved quantity $U$ in a control volume $K$ changes only due to the fluxes $F$ through the boundary $\partial K$ of the control volume $K$.

We assume for the moment that the (concentration of) the conserved quantity $U(x,t) : \Omega \times \mathbb{R}_+ \to \mathbb{R}^m$, $\Omega \subset \mathbb{R}^d$, is smooth. As the conservation holds true for any control volume $K$ in the domain $\Omega$, we can pass from the above integral formulation to a differential formulation

$$U_t + \sum_{k=1}^{d} F^k (U)_{x_k} = 0, \quad (x, t) \in \Omega \times \mathbb{R}_+.$$  (0.1)

Here, $F^k$ is the component of the flux in the $k$-th spatial direction. We assume that the flux $F$ depends only on the conservative variable $U$ and not, say, on $\nabla U$. This is related to the fact that we consider hyperbolic conservation laws in this thesis. Roughly speaking this means that initial disturbances break in each direction into $m$ different waves that travel with finite speed. On the other hand for parabolic equations, such as the heat equation, disturbances immediately affect the solution in the whole domain, that is, they travel with infinite speed.

There are many examples of hyperbolic conservation laws: the Euler equations of gasdynamics, the shallow water equations of oceanography, the magnetohydrodynamic (MHD) equations of plasma physics, and the equations of nonlinear elasticity. We focus mostly on the Euler equations, but we will also consider linear problems such as the advection equation and the wave equation and nonlinear scalar model problems such as the (inviscid) Burgers equation.

It is well-known that solutions of nonlinear hyperbolic conservation laws can develop discontinuities (shocks) in finite time even if the initial data is smooth. This manifests itself in the steepening of waves until they eventually break. Therefore, we can no longer search for classical solutions, but have to rely on weak solutions. This means that the conservation law is only enforced in an integral sense, and unlike the differential formulation, the integral formulation is valid even with discontinuous $U$. However, weak solutions need not be unique and additional admissibility criteria have to be imposed to single out the physically relevant solution. They are often posed in form of entropy conditions, ensuring that the entropy (in the mathematical sense) does not increase.
Scalar conservation laws possess a rich class of entropies, resulting in a unique entropy solution \[33\]. Thus, the theory of scalar conservation laws is well-understood. The situation for systems of conservation laws is completely different. Apart from some partial results in one spatial dimension the field is open. Even though systems are typically endowed with an entropy (related to the physical entropy), it does not suffice to guarantee uniqueness, existence, or stability of solutions.

An approach to overcome this issue is to consider an even weaker notion of solutions: the entropy measure-valued solutions proposed by DiPerna \[9\]. Weak solutions map points in space-time to points in the state space. This is now generalised as follows: measure-valued solutions map points in space-time to probability measures over state space. This measure can be interpreted as the likelihood that a certain state is attained at this point at the given time. Thus, it describes also the oscillatory nature of the solution. In this framework, it might be possible to prove existence of solutions or even well-posedness.

Although the theoretical foundations are far from being complete, numerical methods have successfully been used to approximate hyperbolic conservation laws. In fact, they have even been a driving force for the advances in the theory (e.g. \[16\]). The most widespread class of methods are the finite volume methods \[35\], where the integral formulation of the conservation law is enforced in each volume of a domain partitioning. Monotone schemes enable the carrying over of many properties of the exact solution (of a scalar problem) to the approximate solutions, and thus one obtains convergence for scalar conservation laws. However, these methods are only first-order accurate, and are therefore too diffusive for most practical applications.

Hence, the quest has been to weaken the theoretical requirements on the numerical approximations in favour of higher-order accuracy. Many approaches involve a reconstruction step in space. However, one has to prevent spurious oscillations (at least to a certain degree) and thus, non-oscillatory reconstructions are typically employed. Examples are total variation diminishing (TVD) limiter based schemes \[35\], essentially non-oscillatory (ENO) schemes \[20\], and weighted ENO (WENO) schemes \[46\] . A higher-order temporal accuracy is often gained by using strong stability preserving (SSP) Runge-Kutta (RK) methods \[18\].

An alternative approach for constructing high-order schemes is the discontinuous Galerkin (DG) method (see \[6\] for an overview), where the test and trial functions are discontinuous piecewise polynomials. Here, the high order of accuracy is already built into the finite dimensional spaces and no reconstruction is needed. However, applying an (explicit) time-stepping approach directly to the semidiscrete formulation would introduce instabilities even if one uses SSP RK schemes. The key ingredient to obtain a stable scheme is the use of limiters (after each RK substep) that damp or prevent oscillations \[6\].

Even though these high-resolution schemes have been successfully used in realistic large-scale simulations (see e.g. \[45\]), at the moment there is almost no rigorous analysis of these schemes, especially for systems. Second-order TVD schemes are shown to converge to the entropy solution of scalar conservation laws \[17\]. However, this is not true for ENO and WENO schemes. Further results were shown for linear symmetrisable systems \[14\].
For general systems of conservation laws, a TVD bound is not available even on the continuous level. The entropy stability and a resulting $L^2$ bound are the only generic a priori estimate that one can rely on. Thus, we design entropy-stable schemes that dissipate the total entropy.

In [11], an entropy-stable scheme referred to as the TECNO scheme was designed. It combines entropy-conservative fluxes [47] with numerical diffusion operators based on ENO reconstruction. The approximate solutions generated by the TECNO scheme were shown to converge to an entropy measure-valued solution for systems of conservation laws. However, the scheme has three main drawbacks. Firstly, the proof relies on a conjecture concerning the total variation the ENO reconstruction. Secondly, as the schemes are finite difference schemes, they can be applied only to Cartesian grids. Thirdly, the stability and convergence proof was only carried out in a semidiscrete formulation; it has not yet been shown to carry over to a fully discrete scheme.

In contrast to finite volume methods, entropy stability has gained more attention in finite element methods. In [27], an entropy-stable streamline diffusion space-time finite element method was proposed. Similar streamline diffusion finite element methods for scalar conservation laws were proposed in [30, 31] and extended to a streamline diffusion DG method in [29]. More recently, Barth proposed entropy-stable streamline diffusion and DG methods for systems of conservation laws in [1].

In this thesis, we combine these approaches and construct an entropy-stable streamline diffusion shock-capturing space-time DG scheme. The main advantage of space-time methods is that the formulation immediately allows the use of general unstructured grids. Furthermore, the entropy stability can be obtained for the fully discrete scheme (without any limiters). However, we prove more than only stability: we show that the scheme converges to an entropy measure-valued solution for hyperbolic systems of conservation laws in several space dimensions. We shall discuss many examples to illustrate the robustness and high-resolution properties of the scheme.

The thesis is organised as follows. In Chapter 1, we shortly summarise the main results from the theory of hyperbolic conservation laws that are relevant to this work. Particular focus is given to the different notions of solutions, including the entropy consistency conditions. Additionally, examples of hyperbolic conservation laws are introduced.

Numerical methods to approximate hyperbolic conservation laws are presented in Chapter 2. We compare finite difference, finite volume, and finite element methods and sketch the construction of entropy-stable finite volume schemes.

In Chapter 3, the space-time DG method is proposed. This includes the definition of the three constituting parts: the DG operator, the streamline diffusion operator, and the shock-capturing operator. Furthermore, we prove entropy stability and a resulting weak “BV” estimate.

The convergence proofs are given in Chapter 4. We prove that the approximate solutions converge to an entropy measure-valued solution for systems of conservation laws (under a uniform $L^\infty$ bound). For linear symmetrisable systems, convergence to the weak solution is then a corollary. Moreover, we show convergence to the entropy solution of convex scalar
conservation laws in one spatial dimension.

Several numerical experiments in Chapter 5 demonstrate the robustness and high-resolution property of the scheme. In addition, the effects of the different terms in the formulation are highlighted.

In the remainder of the thesis, the aspects of implementation are treated. Chapter 6 builds the start: there, the computational method is sketched. This includes the description of the Newton-Krylov solver that is used to solve the large nonlinear system for the degrees of freedom.

For efficiency reasons, the linearised system in the Newton method must be preconditioned. In Chapter 7, we construct different preconditioners of block Jacobi and block Gauss-Seidel type. We study their performance analytically and experimentally.

Chapter 8 deals with a particular application of the space-time DG method. As the method is unconditionally stable, we can also apply it to problems with multiple time scales such as flows near the incompressible limit. Standard (explicit) schemes would impose a very restrictive condition on the time step, whereas the space-time DG method is stable under larger time steps. The resulting numerical challenges are investigated in this chapter.

In the last chapter, Chapter 9, the scheme is generalised to space-time meshes that are no longer a Cartesian product of a spatial and a temporal mesh. Hence, we can adapt the mesh locally in space and time. For the refinement strategy, we consider two approaches: residual-based adaptivity and goal-oriented adaptivity. Residual-based adaptivity relies on an ad-hoc criterion, but is nevertheless quite successful in reducing the global error. In goal-oriented adaptivity, one aims at computing a functional of the solution as accurately as possible. Therefore, duality estimates are applied. However, the computation of dual solutions is a numerical challenge. Nevertheless, several examples demonstrate that goal-oriented adaptivity can indeed lead to a high-order convergence of the functional.
We use the following notation throughout the thesis. We consider systems of \( m \) equations. We denote the corresponding vectors or matrices in bold, e.g. \( \mathbf{U} \in \mathbb{R}^m \) or \( \mathbf{A} \in \mathbb{R}^{m \times m} \). The spatial vectors however are not denoted in bold, e.g. \( \mathbf{x} \in \mathbb{R}^d \), where \( d \) denotes the number of spatial dimensions. Components are denoted using subscripts, e.g. \( x = (x_1, x_2, \ldots, x_d) \), or using superscripts. The latter is mostly used when each component is actually a vector, e.g. \( \mathbf{F}^k \), or when we adapt the common notation \( x = x_1, y = x_2 \) and \( z = x_3 \), e.g. \( a = (a^x, a^y) \). For derivatives, we use mostly subscripts, e.g. \( \mathbf{S}_\mathbf{U} \) is the derivative of \( S \) with respect to all components of \( \mathbf{U} \). Scalar products are written using angular brackets, e.g. \( \langle \mathbf{V}, \mathbf{U} \rangle \). For a symmetric positive definite matrix \( \mathbf{A} \in \mathbb{R}^{m \times m} \) and \( \mathbf{a} \in \mathbb{R}^m \), we denote the corresponding norm by \( \| \mathbf{a} \|_\mathbf{A} = \sqrt{\langle \mathbf{a}, \mathbf{Aa} \rangle} \).

For brevity, the arguments in the most common sums are omitted: \( \sum_n \) is the sum over all time steps, i.e. \( \sum_{n=0}^{N-1} \); \( \sum_K \) is the sum over all elements, i.e. \( \sum_{K \in \mathcal{T}} \); and \( \sum_{K'} \) is the sum over all neighbouring elements \( \sum_{K' \in \mathcal{N}(K)} \).

We will call a function \( S(\mathbf{U}) \) convex if \( \mathbf{S}_{\mathbf{UU}} \) is positive semidefinite for all \( \mathbf{U} \). It is called strictly convex if \( \mathbf{S}_{\mathbf{UU}} \) is positive definite for a.e. \( \mathbf{U} \) and lastly it is called uniformly convex if furthermore the smallest eigenvalue of \( \mathbf{S}_{\mathbf{UU}} \) is bounded away from zero.

We denote the Kronecker delta by

\[
\delta_{i,j} = \begin{cases} 
1, & i = j \\
0, & \text{otherwise}
\end{cases} \quad (0.2)
\]

Finally, we denote a generic positive constant as \( C \), possibly different at each occurrence.
1. Hyperbolic conservation laws

In this chapter, we summarize the theory of hyperbolic conservation laws that is relevant to this thesis. Particular focus is given to the different notions of solutions. Moreover, examples of hyperbolic conservation laws are introduced.

1.1. Basic notions

1.1.1. Integral formulation of conservation laws

Many physical phenomena are based on the principle of conservation of quantities such as mass, momentum, energy, etc. Let us consider the concentration of a conserved quantity \( U(x,t) : \Omega \times \mathbb{R}_+ \rightarrow \mathbb{R}^m \) (conservative variable for short) in a control volume \( K \subset \Omega \subset \mathbb{R}^d \). The rate of change of the conserved quantity in the control volume \( K \) is only due to the flux \( F \) at the boundary \( \partial K \). If we assume that the flux \( F \) is only a function of the conservative variable \( U \) (which is true for many physical phenomena), this yields the following mathematical formulation:

\[
\frac{d}{dt} \int_K U \, dx = - \int_{\partial K} \sum_{k=1}^d F_k(U) \nu^k \, d\sigma,
\]

where \( \nu \) is the outward pointing normal, \( F_k \) is the component pointing in the \( k \)-th direction, and \( d \) is the spatial dimension. (1.1) is termed the integral formulation of the conservation law.

The conservation law is supplemented with an initial condition of the form

\[
U(x,0) = U_0(x), \quad x \in \Omega,
\]

where \( U_0 : \Omega \rightarrow \mathbb{R}^m \) is the initial data.

1.1.2. Differential formulation of conservation laws

Under the assumption that the conservative variable \( U \) is smooth, we can apply the divergence theorem to (1.1). This implies

\[
U_t + \sum_{k=1}^d F_k(U)x_k = 0, \quad (x,t) \in \Omega \times \mathbb{R}_+,
\]
1. Hyperbolic conservation laws

where $\Omega \subset \mathbb{R}^d$ is the spatial domain. The initial condition (1.2) is additionally enforced. The equation (1.3) is the differential formulation of the conservation law (1.1). Conservation laws are mostly stated in this form. However, one has to keep in mind that the smoothness assumption might be invalid. We will come back to this point. There are many conservation laws that can be posed in the form (1.3): for instance, the Euler equations of gas dynamics, the magnetohydrodynamic (MHD) equations of plasma physics, the shallow water equations of oceanography, and the equations of nonlinear elasticity [8, 35]. We will discuss some of these examples in more detail in Section 1.2.

1.1.3. Hyperbolicity

The above mentioned examples fall into the category of hyperbolic conservation laws, which is also the focus of this dissertation. The system (1.3) is termed hyperbolic if the directional Jacobian $\sum_{k=1}^d F^k_{\nu} U^k$ is diagonalisable with real eigenvalues for all directions $\nu$. This implies that initial disturbances break into $m$ different waves in each direction, all of them with a finite speed of propagation. This is to be contrasted with parabolic partial differential equations, such as the heat equation, in which a small perturbation influences the solution immediately in the whole domain, i.e. disturbances travel with an infinite speed.

Another distinguishing feature of nonlinear hyperbolic systems of conservation laws is that solutions can develop discontinuities, such as shock waves, in finite time even if the initial data is smooth. Therefore, we no longer search for classical solutions, but have to consider the following more general notion of solutions.

1.1.4. Weak solutions

In this notion of solutions, the conservation law (1.3) is enforced in the weak (distributional) sense: $U \in \left( L^1_{\text{loc}}(\Omega \times \mathbb{R}^+) \right)^m$ is said to be a weak solution of (1.3) if for every compactly supported test function $\varphi \in \left( C^\infty_c(\Omega \times \mathbb{R}^+) \right)^m$ the following integral identity is satisfied:

$$
\int_{\mathbb{R}^+} \int_{\Omega} \left( \langle \dot{U}, \varphi_t \rangle + \sum_{k=1}^d \langle F^k(U), \varphi_{x_k} \rangle \right) dx dt + \int_{\Omega} \langle U_0(x), \varphi(x,0) \rangle dx = 0. \quad (1.4)
$$

Formally, (1.4) is obtained by multiplying (1.3) with the test function $\varphi$, integrating over space and time, and applying integration by parts. The integral form (1.1) and the weak form (1.4) are equivalent [34]. Both of them require less smoothness than the differential formulation (1.3). Mathematically, it is more convenient to work with the weak than with the integral form. Therefore, this formulation is typically used.

Weak solutions are a generalisation of classical solutions, i.e. solutions that satisfy the differential form (1.3). Any classical solution of (1.3) is also a weak solution of (1.3). However, weak solutions are not necessarily unique, in contrast to classical solutions. Hence, additional admissibility criteria or entropy conditions are prescribed to single out the physically relevant solution.
1.1. Basic notions

1.1.5. Entropy conditions

We assume in the following that there exists a uniformly convex entropy function \( S : \mathbb{R}^m \to \mathbb{R} \) and entropy flux functions \( Q^k : \mathbb{R}^m \to \mathbb{R} \) such that the following compatibility conditions are satisfied:

\[
Q^k_U(U) = \langle V, F^k_U(U) \rangle, \quad k = 1, 2, \cdots, d. \tag{1.5}
\]

Here, \( V = S_U(U) \) is called the vector of entropy variables [37]. A weak solution \( U \) of (1.3) is said to be admissible with respect to \((S, Q)\) [8] if it satisfies the entropy inequality

\[
S(U)_t + \sum_{k=1}^d Q^k(U)_{x_k} \leq 0 \tag{1.6}
\]

in the sense of distributions, i.e.

\[
\int_{\mathbb{R}_+} \int_{\Omega} \left( S(U)\varphi_t + \sum_{d=1}^k Q^k(U)\varphi_{x_k} \right) dx dt + \int_{\Omega} S(U_0(x))\varphi(x, 0) dx \geq 0 \tag{1.7}
\]

for every compactly supported test function \( \varphi \in C_c^\infty(\Omega \times \mathbb{R}_+) \) with \( \varphi(x, t) \geq 0 \). It is called an entropy solution if it is admissible with respect to all convex entropies.

For smooth solutions, (1.6) is actually satisfied with equality. To see this, we multiply the conservation law (1.3) with the entropy variables \( V \):

\[
\left\langle V, U_t + \sum_{k=1}^d F^k(U)_{x_k} \right\rangle = 0. \tag{1.8}
\]

The temporal term can be written as

\[
\langle V, U_t \rangle = \langle S_U(U), U_t \rangle = S(U)_t \tag{1.9}
\]

using the definition of the entropy variables. We reformulate the spatial term using the compatibility condition (1.5)

\[
\langle V, F^k(U)_{x_k} \rangle = \langle V, F^k_U(U)_{x_k} \rangle = \langle Q^k_U(U), U_{x_k} \rangle = Q^k(U)_{x_k}, \tag{1.10}
\]

which leads to the differential formulation of entropy conservation,

\[
S(U)_t + \sum_{k=1}^d Q^k(U)_{x_k} = 0. \tag{1.11}
\]

Thus, entropy is dissipated only at discontinuities of \( U \), more precisely at shocks.
1. Hyperbolic conservation laws

Integrating the entropy inequality (1.6) in space results in

\[ \frac{d}{dt} \int_{\Omega} S(U(x,t)) \, dx \leq 0. \]  

(1.12)

Thus, the total entropy does not increase over time. (Physicists define the entropy with a negative sign; for them the total entropy should not decrease.) This bound on the total entropy, together with the uniform convexity of the entropy function, yields an \( L^2 \) stability estimate for the entropy solution \( U \) [8]. This "energy estimate" is currently the only available generic global a priori estimate for systems of conservation laws [8]. Furthermore, the entropy inequality incorporates appropriate small-scale information such as that of vanishing viscosity [8].

The use of the entropy variables symmetrises the system [37]. When convenient, we will work with the following equivalent formulation: Given a strictly convex entropy \( S \), the conservative variables \( U \) and the entropy variables \( V \) are one to one [8]. Consequently, the conservation law (1.3) can be recast in terms of entropy variables as

\[ U(V)_t + \sum_{k=1}^{d} F^k(U(V))_{x_k} = 0, \quad (x,t) \in \Omega \times \mathbb{R}_+. \]  

(1.13)

For brevity of notation, we will write \( F^k(V) = F^k(U(V)) \).

1.1.6. Entropy measure-valued solutions

For scalar conservation laws, every convex function can serve as an entropy. This rich family of entropies implies uniqueness of the entropy solution even for multiple space dimensions [33]. For systems, the question of well-posedness is still open. There are partial results of uniqueness and existence for small data in one space dimension, see [11] for an overview. But in the general multi-dimensional case, no well-posedness results are currently available. In fact, non-uniqueness of entropy solutions for some specific, multi-dimensional systems have been established recently in [5] and references therein. Furthermore, it was recently established in [12] (see also [36]) that state-of-the-art numerical schemes such as the entropy-stable TECNO schemes of [14] may not even converge to an entropy solution as defined above under mesh refinement. Thus, one needs to consider a more general concept of solutions of (1.3) in order to establish existence and stability of solutions and convergence of numerical approximations. A suitable notion of solutions is provided by the framework of entropy measure-valued solutions [12]. As first suggested by DiPerna in [9], we consider the following type of mappings

\[ \mu : \Omega \times \mathbb{R}_+ \rightarrow \text{Prob}(\mathbb{R}^m), \quad (x,t) \mapsto \mu_{x,t}, \]

where \( \text{Prob}(\mathbb{R}^m) \) is the space of probability measures over \( \mathbb{R}^m \) (i.e. non-negative measures with unit mass). We call \( \mu \) a Young measure. A Young measure \( \mu \) is a measure-valued
1.2. Examples

solution of the system (1.3) (based on the equivalent representation (1.13)) if it satisfies,

\[
\int_{\Omega} \int_{\mathbb{R}^+} \left( \langle \langle U, \mu_{x,t} \rangle, \varphi_t \rangle + \sum_{k=1}^{d} \langle \langle F^k, \mu_{x,t} \rangle, \varphi_{x_k} \rangle \right) dx dt + \int_{\Omega} \langle U_0(x), \varphi(x,0) \rangle dx = 0, \quad (1.14)
\]

for all test functions \( \varphi \in (C^\infty_c(\Omega \times [0,\infty)))^m \). Here, we denote

\[\langle g, \mu_{x,t} \rangle = \int_{\mathbb{R}^m} g(\lambda) d\mu_{x,t}(\lambda).\]

Note that we realize the Young measure in terms of the entropy variables \( V \) as (1.14) is based on the formulation (1.13) of the conservation law, and therefore \( U \) and \( F^k \) are functions of the entropy variables \( V \). The corresponding Young measure for the conservative variables \( U \) can be realized as \( U(\mu) \), using the one-to-one mapping \( U(V) \).

Also the concept of entropy conditions carries over to measure-valued solutions: \( \mu \) is an entropy measure-valued solution of (1.3) if it is a measure-valued solution and also satisfies

\[
\int_{\Omega} \int_{\mathbb{R}^+} \left( \langle S, \mu_{x,t} \rangle \varphi_t + \sum_{k=1}^{d} \langle Q^k, \mu_{x,t} \rangle \varphi_{x_k} \right) dx dt + \int_{\Omega} S(U_0(x)) \varphi(x,0) dx \geq 0 \quad (1.15)
\]

for all non-negative test functions \( 0 \leq \varphi \in C^\infty_c(\Omega \times [0,\infty)) \).

The notion of entropy measure-valued solution is a generalisation of the class of entropy solutions: every entropy solution \( U(V(x,t)) \) can be identified with an entropy measure-valued solution by considering the corresponding Young measure \( \mu_{x,t} = \delta_{V(x,t)} \) [9]. However, examples in the recent paper [12] do suggest that entropy measure-valued solutions are not necessarily atomic, i.e. the Young measure attains non-Dirac measures for a set of space-time points with positive measure.

1.2. Examples

Hyperbolic conservation laws are classified based on whether they are linear or nonlinear and whether they are a scalar equation or a system of equations. We present the corresponding classes with representative examples in the following.

1.2.1. Scalar conservation laws

Scalar conservation laws correspond to (1.3) with \( m = 1 \). As already mentioned, every convex function can serve as an entropy and thus, there is a unique entropy solution [33]. We consider the following two examples:
1. Hyperbolic conservation laws

**Linear advection equation**

The linear advection equation is probably the easiest example of a conservation law and is given by

\[ u_t + \sum_{k=1}^{d} a_k u_{x_k} = 0, \quad (x,t) \in \Omega \times \mathbb{R}_+ \tag{1.16} \]

where \( u \) is the conservative variable and \( a = (a_k)_{k=1}^{d} \) is the advection velocity. The solution of the Cauchy-Problem (1.16) with the initial data

\[ u(x,0) = u_0(x), \quad x \in \mathbb{R}^d \tag{1.17} \]

is given by

\[ u(x,t) = u_0(x - at), \quad (x,t) \in \Omega \times \mathbb{R}_+ \tag{1.18} \]

i.e. the initial data is advected in the direction of \( a \). Therefore, only contact discontinuities (due to initial conditions) can occur and the total entropy is conserved. One could also consider a slightly more general version where the advection velocity \( a \) is allowed to depend on space (and time).

**Burgers’ equation**

The easiest example of a nonlinear conservation law is the (inviscid) Burgers equation. It is usually formulated in one spatial dimension and given by

\[ u_t + \frac{1}{2} u^2_x = 0. \tag{1.19} \]

Rewriting it using the chain rule, we obtain

\[ u_t + uu_x = 0. \tag{1.20} \]

This looks similar to the advection equation in one spatial dimension, but there is a crucial difference: instead of using a fixed advection velocity, the unknown quantity itself acts as the advection velocity. This makes the equation nonlinear and allows phenomena such as shocks and rarefactions to develop in the solution.

### 1.2.2. Linear symmetrisable systems

Linear symmetrisable systems are conservation laws of the form

\[ U_t + \sum_{k=1}^{d} A^k U_{x_k} = 0, \quad (x,t) \in \Omega \times \mathbb{R}_+. \tag{1.21} \]

Here, \( A^k \in \mathbb{R}^{m \times m} \) are constant matrices (for simplicity). Assume that there exists \( B \in \mathbb{R}^{m \times m} \) such that
1.2. Examples

(a) $B$ is symmetric, positive definite.

(b) For all $1 \leq k \leq d$, the matrix $BA^k$ is symmetric.

If such a symmetriser $B$ exists, then the linear system (1.21) is called a *symmetrisable* or *Friedrichs system*. It is easy to check that the symmetrisable system (1.21) is equipped with the following entropy formulation:

$$S(U) = \frac{1}{2} \langle BU, U \rangle, \quad Q^k(U) = \frac{1}{2} \langle U, BA^kU \rangle, \quad V = BU. \quad (1.22)$$

Linear scalar conservation laws, such as the linear advection equation, are a (degenerate) subclass of linear symmetrisable systems. A generic example is the following:

**Wave equation**

The wave equation can be written in the form (1.3):

$$h_t + c \sum_{k=1}^{d} (u_k)_{x_k} = 0, \quad (1.23a)$$

$$(u_k)_t + ch_{x_k} = 0, \quad k = 1, \ldots, d. \quad (1.23b)$$

Here, $h$ is the height or pressure variable and $u$ is the velocity vector. (1.23) is already in a symmetric form. Hence, it is a representative of a symmetrisable linear system (with the identity matrix as symmetriser $B$). Note that the wave equation in one spatial dimension can be decoupled into two linear advection equations, but this is no longer true for multiple space dimensions.

1.2.3. Nonlinear systems of conservation laws

Nonlinear systems of conservation laws are the main focus of this thesis. We consider the following two examples:

**Shallow water equations**

The shallow water equations [35] are a simplified model for a liquid flow using a depth-averaged flow velocity:

$$h_t + \sum_{k=1}^{d} (hu)_{x_k} = 0, \quad (1.24a)$$

$$(hu)_t + \sum_{k=1}^{d} (hu_ku_l + \frac{1}{2} gh^2 \delta_{k,l})_{x_k} = 0, \quad l = 1, \ldots, d. \quad (1.24b)$$
1. Hyperbolic conservation laws

with $d = 1$ or $2$. Here, $h$ is the liquid height, $u$ is the (depth-averaged) velocity, and $g$ is the gravity constant. The equation (1.24a) is the continuity equation, while (1.24b) describes the conservation of momentum. The total energy $S(U) = \frac{1}{2} \sum_{k=1}^{d} u_k^2 + \frac{1}{2} gh^2$ together with $Q^k(U) = \frac{1}{2} \sum_{l=1}^{d} u_l^2 u_k + \frac{1}{2} g u_k h^2$, $k = 1, \ldots, d$ serves as an entropy pair for the system. More complex effects such as non-constant bottom topography and the Coriolis force can be built in as additional right hand side terms in (1.24b).

Euler equations

The Euler equations of gasdynamics [35] can be formulated in $d$ space dimensions as

\begin{align}
\rho_t + \sum_{k=1}^{d} (\rho u_k)_{x_k} &= 0, \quad (1.25a) \\
(\rho u_l)_t + \sum_{k=1}^{d} (\rho u_l u_k + p \delta_{k,l})_{x_k} &= 0, \quad l = 1, \ldots, d, \quad (1.25b) \\
(\rho E)_t + \sum_{k=1}^{d} (\rho u_k H)_{x_k} &= 0. \quad (1.25c)
\end{align}

Here, $\rho$ is the density, $u$ is the velocity field and $\rho E$ is the total energy. The system (1.25) describes the conservation of mass, momentum, and energy. It contains the auxiliary quantities pressure $p$, sound speed $c$, and enthalpy $H$, given by

\begin{align*}
p &= (\gamma - 1) \left( \rho E - \frac{1}{2} \rho \sum_{k=1}^{d} u_k^2 \right), \quad c = \sqrt{\frac{\gamma p}{\rho}}, \quad H = \frac{c^2}{\gamma - 1} + \frac{1}{2} \sum_{k=1}^{d} u_k^2,
\end{align*}

where $\gamma$ is the adiabatic exponent. We set $\gamma = 1.4$ in all experiments, which corresponds to a diatomic gas. Furthermore, the specific entropy is $s = \log p - \gamma \log \rho$, and the total entropy for the Euler equation is given by

\begin{align*}
S(U) = -\frac{\rho s}{\gamma - 1}
\end{align*}

with the entropy flux

\begin{align*}
Q^k(U) = -\frac{\rho u_k s}{\gamma - 1}, \quad k = 1, \ldots, d.
\end{align*}
2. Numerical approximation of hyperbolic conservation laws

Given the nonlinearity of the conservation law (1.3), explicit formulas for the solution are available in only a few cases. Hence, one has to approximate the solution numerically. Numerical methods for approximating hyperbolic conservation laws can roughly be split into three categories: finite difference schemes, finite volume schemes, and finite element methods. We note that the classification is not always sharp and that there are also schemes based on very different ideas, such as shock tracking schemes, residual distribution methods, or spectral viscosity methods.

Finite difference schemes approximate the differential form of the conservation law (1.3) by replacing spatial derivatives with finite differences. Advancing in time is usually done by a time stepping method. Finite difference schemes can be formulated quite easily on Cartesian grids. But they have several drawbacks, the most important being that they might show Gibbs phenomena, i.e. the solution is oscillatory if the underlying exact solution forms a shock. Approaches to overcome this problem proceed in similar ways as for finite volume schemes, which we consider next and in more detail.

Finite volume schemes [35] are probably the most common and widespread method to approximate conservation laws. In contrast to finite difference methods, they approximate the integral form of the conservation law (1.1). For this purpose, the domain is split into nonoverlapping control volumes and the degrees of freedom are the mean values (cell averages) of the conservative variable $U$ in each of the control volumes. The true flux at the boundary has to be approximated by a numerical flux depending on the mean values of the cells next to the boundary. Monotone fluxes lead to schemes with preferable theoretical properties with respect to stability and convergence. However, these schemes are restricted to first-order convergence. Therefore, one typically abandons the notion of monotone schemes and looks for total variation diminishing (TVD) or essentially non-oscillatory (ENO) schemes instead. We describe finite volume methods in more detail in Section 2.1.

The last category is finite element methods, which have been very successful in approximating parabolic and elliptic problems. The basic principle is as follows: a weak formulation of the problem is approximated by replacing the infinite-dimensional test and trial spaces by finite-dimensional subspaces. For this purpose, the spatial domain is split into elements, and the restriction of the test and trial functions to each element is a polynomial. For hyperbolic problems, a discontinuous Galerkin (DG) formulation is more suitable [7]. In DG finite element methods, the test and trial functions are allowed to be discontinuous across elements. Hence, and in contrast to continuous finite elements methods, boundary integrals
2. Numerical approximation of hyperbolic conservation laws

at the inner boundary need to be included into the formulation. There is no need for a
reconstruction as the high order of accuracy is already built into the formulation. However,
stability requires the use of limiters, otherwise spurious oscillations at discontinuities can
destroy the numerical approximation. This makes implementation more involved and might
reduce the formal order of accuracy. These methods are described in Section 2.2.

2.1. Finite volume schemes

For simplicity, we consider the one dimensional case here, but the method can be generalized
to Cartesian grids in several space dimensions in an relatively straightforward way. The
spatial domain is partitioned into equidistant intervals (the finite volumes) of size \( \Delta x \) with
midpoint \( x_i, i \in \mathbb{Z} \). The boundary points are denoted by \( x_{i-1/2} \) and \( x_{i+1/2} \). The integral
formulation of the conservation law (1.1) simplifies after integration over time to

\[
\int_{x_{i-1/2}}^{x_{i+1/2}} U(x, t_{n+1}) dx - \int_{x_{i-1/2}}^{x_{i+1/2}} U(x, t_n) dx + \int_{t_n}^{t_{n+1}} F(U(x_{i+1/2})) dt - \int_{t_n}^{t_{n+1}} F(U(x_{i-1/2})) dt = 0.
\]

(2.1)

The unknown quantities are the the mean values of \( U \) inside \([x_{i-1/2}, x_{i+1/2}]\) at time \( t^n \),
denoted by

\[
U^n_i = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} U(x, t^n) dx.
\]

In order to obtain a numerical scheme, the true fluxes at the volume boundaries are replaced
by numerical fluxes \( F_{i+1/2} \) and \( F_{i-1/2} \) that depend on some of the neighbouring \( U^n_i \) values,
giving the numerical scheme

\[
\frac{1}{\Delta t} (U^{n+1}_i - U^n_i) + \frac{1}{\Delta x} (F_{i+1/2} - F_{i-1/2}) = 0.
\]

(2.2)

In the simplest case, the numerical fluxes depend only on the two neighbouring values, i.e.
\( F_{i+1/2} = F(U_i, U_{i+1}) \). The time step \( \Delta t \) needs to satisfy a CFL condition for stability rea-
sons. If monotone numerical fluxes are used, one obtains schemes with favourable properties
in the case of scalar conservation laws, such as \( l_1 \)-contraction, TVD, monotonicity preser-
vation, and convergence, see [34]. However, there is one important drawback: monotone
methods are at most first order accurate. In the next subsection, we consider approaches to
avoid this barrier.

2.1.1. High-resolution schemes

To obtain a higher order of accuracy, the fluxes at the cell boundaries need to be approx-
imated more accurately. Often, the schemes involve a reconstruction step. The aim is to
2.1. Finite volume schemes

compute values $U_{i+1/2+}$ for $x \to x_{i+1/2+}$ and $U_{i+1/2-}$ for $x \to x_{i+1/2-}$ that are more accurate than the mean values $U_i$ and $U_{i+1}$. $U_{i+1/2-}$ and $U_{i+1/2+}$ are then used in the evaluation of the numerical flux $F$, leading to high-order accurate schemes.

The reconstruction process roughly consists of two steps: first a polynomial of degree $p$ is reconstructed in each cell. This is typically done by interpolation, using neighbouring cell values. In a second step, a slope limiter is applied to the polynomial in order to avoid spurious oscillations at discontinuities. Often, the two steps are combined into one single step.

An important class of limiters guarantees that the total variation of the approximate solution decreases over time (TVD) [35]. Near smooth extrema, the schemes reduce to first order because of the TVD property. This can be fixed by the use of total variation bounded (TVB) instead of TVD limiters, but typically they depend on certain tuning parameters.

Furthermore, note that for higher than second order methods, one will need more than just the average values of the two neighbouring cells, i.e. a larger stencil. This will complicate the implementation of boundary conditions because more than just one ghost cell needs to be provided or the scheme needs to be adapted near the boundary.

One way to get higher than second order are the essentially non-oscillatory (ENO) schemes [45]. In this method, the stencil of the reconstruction is successively enlarged either to the left or to the right wherever the oscillation is smaller. Algorithmically, the choice is based on divided differences and a Newton type of interpolation is used.

We will now consider entropy-stable schemes, as these allow proving convergence for high-order schemes.

2.1.2. Entropy-stable finite volume schemes

For simplicity, we keep the time continuous in the following. We consider finite volume schemes of the form

$$\frac{d}{dt} U_i + \frac{1}{\Delta x} (F_{i+1/2} - F_{i-1/2}) = 0. \quad (2.3)$$

Let us first go back to entropy conservation for smooth solutions in the continuous case. Consider again the derivation of (1.11). We imitate the calculations for the discrete case. However, we consider an alternative derivation of the identity (1.10) using the entropy potential $\Psi = \langle V, F \rangle - Q$. We start by remarking that

$$\Psi_V = F + \underbrace{\langle V, F_V \rangle - Q_V}_{0} = F \quad (2.4)$$

due to the compatibility condition (1.5). By adding and subtracting $\Psi_x$, we obtain

$$\langle V, F_x \rangle = (\langle V, F \rangle)_x - \langle V_x, F \rangle = \underbrace{(\langle V, F \rangle - \Psi)_x}_{Q} - \underbrace{(\langle V_x, F \rangle - \Psi_x)}_{0} = Q_x. \quad (2.5)$$

Here, we used
\[ \Psi_x = \langle \Psi, V_x \rangle = \langle V_x, F \rangle \quad \text{(by (2.4))}. \] (2.6)

If a scheme satisfies (2.6) in a discrete manner, i.e. if
\[ \Psi(U_{i+1}) - \Psi(U_i) = \langle V_{i+1} - V_i, F(U_i, U_{i+1}) \rangle, \] (2.7)

we can mimic the continuous proof and obtain an entropy-conservative scheme, see [47, 11] for the details. A flux that satisfies (2.7) is called an entropy-conservative flux. This will be an integral part of the construction of entropy-stable DG schemes.

For scalar conservation laws, we can solve (2.7) for \( F \). The flux is therefore unique. For systems on the other hand, (2.7) specifies only one condition, but \( F \) has \( m \) components. Thus, one has some freedom, and the entropy-conservative fluxes are no longer unique. Fortunately, there are easily computable fluxes at hand for most systems of interest [13, 28].

Finite volume schemes (2.3) with entropy-conservative flux conserve the total entropy. Physically incorrectly, this holds true even if the underlying exact solution develops a shock. Therefore, the approximate solutions start to oscillate wildly in the vicinity of a shock. Hence, additional entropy dissipation, i.e. a diffusion term, is needed.

One can consider different forms of diffusion terms that correspond to Lax-Friedrichs, Rusanov or Roe fluxes, see [11]. However, all of them reduce the order of accuracy to at most first order if they act directly on the mean values in the cells. To get a higher (formal) order, one has to apply a suitable reconstruction without introducing spurious oscillations.

Combining high-order entropy-conservative fluxes with a diffusion based on ENO reconstruction leads to the TECNO schemes, see [11]. They are arbitrarily high-order accurate and convergent to an entropy measure-valued solution for systems of conservation laws (under the assumption that a certain conjecture on ENO reconstruction holds). However, they are only formulated in a semi-discrete manner. The stability and convergence has not yet been shown to carry over to fully discrete schemes, with the exception of stability for the scalar case [15].

2.2. Discontinuous Galerkin schemes

We consider in this section discontinuous Galerkin schemes that employ the method of lines approach, such as Runge-Kutta DG (RKDG) schemes [7]. As is typical for finite element methods, the domain is split into nonoverlapping elements, which we denote by \( K \). DG schemes approximate a weak form of the conservation law (1.3). To discretise, the smooth test and trial functions are replaced by functions that are polynomials of degree \( p \) in each element; we denote the corresponding space by \( V_p \). As the test and trial functions are allowed to be discontinuous, contributions from cell boundaries in the interior of the domain have to be considered in the weak formulation. Replacing these by numerical fluxes \( F \), we
obtain the DG formulation
\[
\sum_K \int_K \left( \langle U_t, W \rangle - \sum_{k=1}^{d} \langle F^k(U), W_{x_k} \rangle \right) dx + \sum_{K,K'} \int_{\partial_{KK'}} F(U_{K,-}, U_{K,+}; \nu_{KK'}) W_{K,-} d\sigma = 0
\] (2.8)
for all \( W \in V_p \). Here, we have employed the notation
\[
\partial_{KK'} = K \cap K', \\
\nu_{KK'} = \text{unit normal for edge } KK' \text{ pointing outwards from element } K, \\
W_{K,\pm}(x,t) = \lim_{h \to 0^+} W(x \pm h \nu, t), \quad \forall x \in \partial_{KK'}. \tag{2.9}
\]
The sums are taken over all elements \( K \) and over all neighbouring elements \( K' \). For the numerical flux, a consistent, conservative, first-order flux, such as the Lax-Friedrichs or the Rusanov flux (see [7]) can be used.

Note that if these numerical fluxes are applied in a finite volume scheme, they reduce the method to first order unless some kind of reconstruction is used. But in a DG scheme the high order is already built into the formulation and there is no need for a reconstruction.

Advancing in time is typically done by the use of a strong stability preserving (SSP) Runge-Kutta method [18]. However, this is not enough to obtain stability; a limiter needs to be applied after every Runge-Kutta substep, see [7]. Here, a total variation diminishing in the means (TVDM) limiter is considered. Additionally, in order not to lose accuracy near smooth extrema, it is extended (also in [7]) to a total variation bounded in the means (TVBM) limiter. As the construction of limiters in multiple dimensions is rather involved and additionally the use of limiters complicates the analysis of the resulting schemes, we avoid using limiters. Instead, we consider a space-time DG formulation with additional streamline diffusion and shock-capturing. This allows us to show entropy stability and convergence for the fully discrete scheme. Furthermore, because of its foundation in finite elements, the scheme can be used on unstructured grids to solve problems on complex domains.
3. The shock-capturing streamline diffusion DG formulation

In this chapter, we introduce the space-time DG method. This chapter is partly based on material that was published [25], but the formulation of the method has evolved since then.

3.1. The mesh

At the $n$-th time level $t^n$, we denote the time step as $\Delta t^n = t^{n+1} - t^n$ and the update time interval as $I^n = [t^n, t^{n+1})$. For simplicity, we assume that the spatial domain $\Omega \subset \mathbb{R}^d$ is bounded and polyhedral and divide into a triangulation $\mathcal{T}$, i.e. a set of open convex polyhedra $K \subset \mathbb{R}^d$. Furthermore, we assume mesh regularity [29] and quasiuniformity. For a generic element (cell) $K$, we denote

$$\Delta x_K = \text{diam}(K), \quad \Delta x_K = \text{width of } K,$$

$$\mathcal{N}(K) = \{K' \in \mathcal{T} : K' \neq K \land \text{meas}_{d-1}(\overline{K} \cap \overline{K'}) > 0\}, \quad \text{(neighbours of } K).$$

The mesh width of the triangulation is $\Delta x(\mathcal{T}) = \max_K \Delta x_K$. A generic space-time element is the prism:

$$K \times I^n.$$

See Figure 3.1 for an illustration of the space-time mesh. We also assume that there exists an (arbitrarily large) constant $C$ such that $(1/C) \Delta x \leq \Delta t^n \leq C \Delta x$ for all time levels $n$.

3.2. Variational formulation

Given a uniformly convex entropy function $S$, the conservative variables $U$ and the entropy variables $V$ are one to one [8]. Consequently, we use the equivalent formulation of the conservation law (1.13) in terms of entropy variables. Following [48, 1], we approximate the conservation law (1.13) by a DG method. On a given triangulation $\mathcal{T}$ with mesh width $\Delta x$, we seek entropy variables

$$V^{\Delta x} \in V_p = (\mathbb{P}_p(\Omega \times [0, T]))^m$$

$$= \left\{ W \in (L^1(\Omega \times [0, T]))^m : W|_{K \times I^n} \text{ is a polynomial of degree } p \text{ in each component} \right\}$$

(3.1)
such that the following quasilinear variational form is satisfied for each $W^\Delta \in V_p$:

$$B(V^\Delta_x, W^\Delta_x) := B_{DG}(V^\Delta_x, W^\Delta_x) + B_{SD}(V^\Delta_x, W^\Delta_x) + B_{SC}(V^\Delta_x, W^\Delta_x) = 0.$$  

We elaborate on each of the three quasilinear forms, which are all nonlinear in the first argument and linear in the second, in the following.

### 3.3. The DG quasilinear form

The form $B_{DG}$ is given by

$$B_{DG}(V^\Delta_x, W^\Delta_x) = -\sum_{n,K} \int_{I^n} \int_K \left( \langle U(V^\Delta_x), W^\Delta_x \rangle + \sum_{k=1}^d \langle F^k(V^\Delta_x), W^\Delta_{x_k} \rangle \right) dx \, dt$$

$$+ \sum_{n,K} \int_K \langle U(V^\Delta_{n+1}, W^\Delta_{n+1}), W^\Delta_{n+1} \rangle dx - \sum_{n,K} \int_K \langle U(V^\Delta_{n-}, W^\Delta_{n+}), W^\Delta_{n+} \rangle dx$$

$$+ \sum_{n,K,K' \in N(K)} \int_{I^n} \int_{\partial KK'} \langle F(V^\Delta_{K,-}, V^\Delta_{K,+}; \nu_{KK'}), W^\Delta_{K,-} \rangle d\sigma(x) dt$$

Here we have employed the notation

$$W_{n,\pm}(x) = W(x, t^n_{\pm}) = \lim_{h \to 0^+} W(x, t \pm h)$$

for all $W \in V_p$, in addition to the notation introduced in (2.9). See Figure 3.1 for an illustration. The temporal numerical flux $U$ and the spatial numerical flux $F$ will be introduced in the following subsections.

$V^\Delta_{0,-}$ is either the initial data or a suitable projection thereof. We remark that the boundary condition is ignored in the above variational form by considering compactly supported (in the spatial domain) solutions and test functions.

The DG form (3.3) can formally be obtained as follows. Start by multiplying the conservation law (1.13) with a test function and integrate over all space-time elements. Perform an integration by parts, which leads to additional boundary terms on both spatial and temporal element boundaries. Replacing the fluxes at the element boundaries by numerical fluxes (depending on both trace values) leads to the form (3.3).

### 3.3.1. Numerical fluxes

Both the temporal and spatial numerical fluxes, need to be specified in order to complete the DG quasilinear form. In order to obtain causality (marching) after each time step, we choose the temporal numerical flux to be the *upwind* flux:

$$U(a, b) = U(a).$$
3.3. The DG quasilinear form

\[ \partial \partial K K K \partial K K K K K \partial x V n, V n, V n, V n, V K, V K, V K, V K, V n, V n, t n, t n+1, I n, t \]

(a) Mesh in one spatial dimension. (b) Generic element in two spatial dimensions.

Figure 3.1.: Space-time mesh.

This ensures that we can use the values at the previous time step in order to compute an update at the time level \( t^n \). A different choice of temporal numerical fluxes will imply that all the degrees of freedom (for all times) are coupled and force us to solve a very large non-linear algebraic system of equations.

The spatial numerical flux is assumed to be conservative, i.e.

\[ \mathbb{F}(a, b; \nu) = -\mathbb{F}(b, a; -\nu) \] (3.6)

for all directions \( \nu \) and all states \( a \) and \( b \), and consistent, i.e.

\[ \mathbb{F}(a, a; \nu) = \sum_{k=1}^{d} \mathbb{F}^k(a)\nu^k, \] (3.7)

for all directions \( \nu \) and all states \( a \). More specifically, we use an entropy-stable numerical flux given by

\[ \mathbb{F}(V^\Delta x K_-, V^\Delta x K_+; \nu_{KK}) = \sum_{k=1}^{d} \mathbb{F}^{k,*}(V^\Delta x K_-, V^\Delta x K_+)^k \nu_{KK} - \frac{1}{2} D(V^\Delta x K_+, V^\Delta x K_-) \] (3.8)

with \( D = D(V^\Delta x K_-, V^\Delta x K_+; \nu_{KK}) \), which consists of the two components described in the following.
3. The shock-capturing streamline diffusion DG formulation

3.3.2. Entropy-conservative flux

The entropy-conservative flux (in the $k$-th direction) is any flux that satisfies the relation

$$\langle b - a, \mathbb{F}^{k,*}(a, b) \rangle = \Psi^k(b) - \Psi^k(a), \quad (3.9)$$

see [47]. Here, $\Psi^k = \langle V, F^k \rangle - Q^k$ is the entropy potential. The existence of such fluxes for any generic conservation law with an entropy framework was shown by Tadmor in [47]. More recently, explicit expressions of entropy-conservative fluxes for specific systems of interest like the shallow water equations [13] and Euler equations [28] have been obtained.

We impose the following additional conditions on the entropy-conservative flux to obtain a consistent (3.7) and conservative (3.6) numerical flux. The entropy-conservative flux itself should be consistent, i.e.

$$\mathbb{F}^{k,*}(a, a) = F^k(a) \quad (3.10)$$

for all $k$ and all states $a$, and symmetric, i.e.

$$\mathbb{F}^{k,*}(a, b) = \mathbb{F}^{k,*}(b, a) \quad (3.11)$$

for all $k$ and all states $a$ and $b$. One could drop the symmetry requirement, but then $\mathbb{F}^{k,*}$ would need to depend on the normal $\nu$ in order to obtain a conservative flux. The path-integral formulation of the entropy-conservative flux in [48] shows the existence of a symmetric entropy-conservative flux for general systems (1.3). Furthermore, the explicit entropy-conservative fluxes given in [13, 28] satisfy these conditions.

3.3.3. Numerical diffusion operators

Following [48, 13, 14], we choose the numerical diffusion operator as

$$D(a, b; \nu) = R_\nu P(a, b; \nu) R_\nu^\top. \quad (3.12)$$

Here, $\Lambda_\nu, R_\nu$ are the eigenvalue and eigenvector matrices of the Jacobian $\partial U(\langle F, \nu \rangle)$ in the normal direction $\nu$. $R_\nu$ is evaluated at an averaged state, e.g. $(a + b)/2$, and scaled such that $R_\nu R_\nu^\top = U V$. Furthermore, $P(a, b; \nu)$ is a diagonal non-negative matrix depending on the eigenvalues $\Lambda_\nu$. Examples of $P$ include

$$P(a, b; \nu) = |\Lambda_\nu((a + b)/2)| \quad (3.13)$$

which leads to a Roe type scheme and

$$P(a, b; \nu) = \max\{\lambda_{\text{max}}(a; \nu), \lambda_{\text{max}}(b; \nu)\} I, \quad (3.14)$$

which leads to a Rusanov type scheme [14], where $\lambda_{\text{max}}(U; \nu)$ is the maximal wave speed in the direction of $\nu$, i.e. $\lambda_{\text{max}}(U; \nu)$ is the spectral radius of $\Lambda_\nu(U)$.
To obtain a conservative flux (i.e., (3.6) is satisfied), we assume that the diffusion matrix is symmetric in the following sense

\[
D(b, a; -\nu) = D(a, b; \nu)
\]  

(3.15)

for all normals \(\nu\) and all states \(a\) and \(b\). Furthermore, we will sometimes assume that \(D\) is uniformly positive definite.

Inserting the numerical fluxes into the form \(B_{DG}\) (3.3), we obtain

\[
B_{DG}(V^{\Delta x}, W^{\Delta x})
= - \sum_{n,K} \int_{I^n} \int_K \left( \langle U(V^{\Delta x}), W_t^{\Delta x} \rangle + \sum_{k=1}^d \langle F_k(V^{\Delta x}), W_{x_k}^{\Delta x} \rangle \right) dx dt
- \sum_{n,K} \int_K \left( \sum_{k=1}^d \langle F_k(V^{\Delta x}), W_{x_k}^{\Delta x} \rangle \right) dx
\]

(3.16)

\[
B_{SD}(V^{\Delta x}, W^{\Delta x})
= \sum_{n,K} \int_{I^n} \int_K \left( \langle U(V^{\Delta x}), W_t^{\Delta x} \rangle + \sum_{k=1}^d \langle F_k(V^{\Delta x}), W_{x_k}^{\Delta x} \rangle \right) d\sigma(x) dt
\]

(3.17)

with intra-element residual:

\[
\text{Res} := U(V^{\Delta x})_t + \sum_{k=1}^d F_k(V^{\Delta x})_{x_k},
\]

(3.18)

and the scaling matrix is chosen as

\[
D^{SD} := C^{SD} \Delta t^n U^{-1}_V(V^{\Delta x}),
\]

(3.19)
3. The shock-capturing streamline diffusion DG formulation

for some positive constant \( C^{SD} \). Note that the intra-element residual is well defined as we are taking first-derivatives of a polynomial function.

Streamline diffusion can be seen as modifying the test function in the DG method by adding a multiple of the linearized form of the hyperbolic operator. For scalar convection problems, streamline diffusion reduces to additional diffusion along the streamlines, which is the origin of its name. Note that for \( W^{\Delta x} = V^{\Delta x} \), we have

\[
U_V(V^{\Delta x}) W^{\Delta x}_t + \sum_{k=1}^{d} F^k_V(V^{\Delta x}) W^{\Delta x}_{x_k} = U_V(V^{\Delta x}) V^{\Delta x}_t + \sum_{k=1}^{d} F^k_V(V^{\Delta x}) V^{\Delta x}_{x_k}
\]

\[
= U(V^{\Delta x})_t + \sum_{k=1}^{d} F^k(V^{\Delta x})_{x_k}
\]

which leads to a weak control of the \( L^2 \) norm of the residual, and thus increased stability.

3.5. Shock-capturing operator

The streamline diffusion operator adds numerical diffusion in the direction of the streamlines. However, we need further numerical diffusion in order to reduce possible oscillations at shocks. We use the following shock-capturing operator (similar to Barth [1]):

\[
B_{SC}(V^{\Delta x}, W^{\Delta x})
= \sum_{n,K} \int_{I^n} \int_K D_{n,K}^{SC}\left( \langle W^{\Delta x}_t, U_V V^{\Delta x} \rangle + \sum_{k=1}^{d} \frac{\Delta x K^2}{(\Delta t^n)^2} \langle W^{\Delta x}_{x_k}, U_V V^{\Delta x}_{x_k} \rangle \right) dx dt,
\]

(3.20a)

with \( \tilde{U}_V = U_V(\tilde{V}_{n,K}) \) for brevity and

\[
\tilde{V}_{n,K} = \frac{1}{\text{meas}(I^n \times K)} \int_{I^n} \int_K V^{\Delta x}(x,t) dx dt.
\]

being the cell average. The scaling factor is

\[
D_{n,K}^{SC} = \frac{\Delta t^n C^{SC} \overline{\text{Res}}_{n,K} + (\Delta t^n)^{\frac{1}{2}} C^{SC} \overline{B \text{Res}}_{n,K}}{\sqrt{\int_{I^n} \int_K \left( \langle V^{\Delta x}_t, U_V V^{\Delta x} \rangle + \sum_{k=1}^{d} \frac{\Delta x K^2}{(\Delta t^n)^2} \langle V^{\Delta x}_{x_k}, U_V V^{\Delta x}_{x_k} \rangle \right) dx dt + \epsilon}},
\]

(3.20b)

with \( \epsilon := |K| (\Delta t^n)^{\frac{1}{2}} \left( \frac{\Delta x}{\text{diam}(I)} \right)^{\theta} \) and \( \theta = 1/2 \) (chosen as 1) and

\[
\overline{\text{Res}}_{n,K} := \sqrt{\int_{I^n} \int_K \langle \text{Res}, U_V^{-1} (V^{\Delta x}) \text{Res} \rangle dx dt}.
\]

(3.20c)
3.6. Entropy stability for nonlinear systems

\[ \text{Res}_{n,K} := \left( \int_K \| U(V^\Delta x_{n,-}) - U(V^\Delta x_{n,+}) \|^2_{U^{-1}V(V^\Delta x_{n,+})} \, dx \right) \]

\[ + \sum_{K'} \int_{I^n} \int_{\partial_{KK'}} \Delta t^n \left\| F(V^\Delta x_{K,n}', V^\Delta x_{K,n}; \nu_{KK'}) - \sum_{k=1}^d F^k(V^\Delta x_{K,n}) \nu^k_{KK'} \right\|^2_{U^{-1}V(V^\Delta x_{K,n})} \, d\sigma dt \right)^{\frac{1}{2}} \]

(3.20d)

Here, \( C^{SC}, \bar{C}^{SC} \) are positive constants.

The idea behind the shock-capturing term is as follows. Near shocks, we need further stabilisation in the form of numerical diffusion. This is not needed in regions where the solution is smooth. More precisely, it should be so small the accuracy of the method does not suffer there. The intra-element residual and the boundary residual are large in the vicinity of shocks and small in smooth regions. Therefore, we add diffusion proportional to the residual. However, in order to be able to apply projection error estimates, integral versions of the intra-element and boundary residual are used and the scaling matrix \( U_V \) is evaluated at the element average. Thus, they are only constant coefficients in each element. Furthermore, we need to normalise the shock-capturing term, which is the reason for the denominator in \( D^{SC}_{n,K} \). The purpose of \( \epsilon \) is to avoid division by zero. The form of \( \epsilon \) and the condition on \( \theta \) will become clear in the entropy stability proof, Theorem 3.6.1.

3.6. Entropy stability for nonlinear systems

The design of the streamline diffusion (SD) shock-capturing (SC) discontinuous Galerkin (DG) scheme (3.2) is motivated by the consideration that it has to be entropy-stable for a generic nonlinear system of conservation laws, equipped with an entropy formulation. We have the following theorem on entropy stability:

**Theorem 3.6.1.** Consider the system of conservation laws (1.3) with a uniformly convex entropy function \( S \) and entropy flux functions \( Q^k \) \((1 \leq k \leq d)\). For simplicity, assume that the exact and approximate solutions have compact support inside the spatial domain \( \Omega \). Let the final time be denoted by \( t^N \). Then, the streamline diffusion shock-capturing discontinuous Galerkin scheme (3.2) approximating (1.3) has the following properties:

(i) The scheme is conservative, i.e. the approximate solutions \( U^\Delta x = U(V^\Delta x) \) satisfy

\[ \int_{\Omega} U^\Delta x(x,t^N_0) \, dx = \int_{\Omega} U^\Delta x(x,t^0_0) \, dx. \]  

(3.21)

(ii) The scheme is entropy-stable, i.e. the approximate solutions satisfy

\[ \int_{\Omega} S(U^\star(t^0_0)) \, dx \leq \int_{\Omega} S(U^\Delta x(x,t^N_0)) \, dx \leq \int_{\Omega} S(U^\Delta x(x,t^0_0)) \, dx, \]  

(3.22)
3. The shock-capturing streamline diffusion DG formulation

with \( \mathbf{U}^* \) being the domain average:

\[
\mathbf{U}^*(t^0_\Omega) = \frac{1}{\text{meas}(\Omega)} \int_{\Omega} \mathbf{U}(\mathbf{V}(x, t^0)) dx.
\]

(iii) Under the assumption that \( \mathbf{U} \), \( \mathbf{V} \), and \( \mathbf{D} \) are uniformly positive definite, we obtain the following weak "BV" estimate:

\[
\sum_{n,K} \int_K \| \mathbf{V}^\n_{n,-} - \mathbf{V}^\n_{n,+} \|^2 dx + \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \langle \mathbf{V}^\Delta_{K,+} - \mathbf{V}^\Delta_{K,-}, \mathbf{D}(\mathbf{V}^\Delta_{K,+} - \mathbf{V}^\Delta_{K,-}) \rangle d\sigma dt
\]

\[
+ \Delta_x \sum_{n,K} \int_{I_n} \int_K \| \mathbf{U}(\mathbf{V}^\Delta)_t + \sum_{k=1}^d \mathbf{F}_k(\mathbf{V}^\Delta)_{x_k} \|^2 dx dt
\]

\[
+ \Delta_x \sum_{n,K} \mathbf{BRes}_{n,K} \left( \int_{I_n} \int_K \| \nabla_{xt} \mathbf{V}^\Delta \|^2 dx dt \right)^{\frac{1}{2}}
\]

\[
+ (\Delta_x)^{\frac{1}{2}} \sum_{n,K} \mathbf{BRes}_{n,K} \left( \int_{I_n} \int_K \| \nabla_{xt} \mathbf{V}^\Delta \|^2 dx dt \right)^{\frac{1}{2}} \leq C
\]  

(3.23)

for some constant \( C \) depending on the initial data, and with the space-time gradient defined by

\[
\nabla_{xt} \mathbf{V}^\Delta = \left( \mathbf{V}^\Delta_t, \mathbf{V}^\Delta_{x_1}, \mathbf{V}^\Delta_{x_2}, \cdots, \mathbf{V}^\Delta_{x_d} \right).
\]

(3.24)

**Proof.** The proof of conservation (property (i)) (3.21) is a straightforward consequence of setting the test function \( \mathbf{W}^\Delta = (\delta_{i,j})_{j=1}^m \) using the conservation property (3.6) in (3.3).

To prove entropy stability, we proceed to show a series of claims.

Claim 1: The streamline diffusion operator (3.17) is positive, i.e.

\[
\mathbf{B}_{SD}(\mathbf{V}^\Delta, \mathbf{V}^\Delta) \geq 0.
\]  

(3.25)

As, \( \mathbf{V}^\Delta \in \mathbf{V}_p \), it is an admissible test function in the quasilinear form \( \mathbf{B}_{SD} \). Setting \( \mathbf{W}^\Delta = \mathbf{V}^\Delta \) in (3.17), we obtain

\[
\mathbf{B}_{SD}(\mathbf{V}^\Delta, \mathbf{V}^\Delta)
\]

\[
= \sum_{n,K} \int_{I_n} \int_K \left\langle \left( \mathbf{U}(\mathbf{V}^\Delta)_t + \sum_{k=1}^d \mathbf{F}_k(\mathbf{V}^\Delta)_{x_k} \right), \mathbf{D}^{SD}_{\text{Res}} \right\rangle dx dt
\]

\[
= \sum_{n,K} \int_{I_n} \int_K \left\langle \mathbf{U}(\mathbf{V}^\Delta)_t + \sum_{k=1}^d \mathbf{F}_k(\mathbf{V}^\Delta)_{x_k}, \mathbf{D}^{SD}_{\text{Res}} \right\rangle dx dt
\]  

22
3.6. Entropy stability for nonlinear systems

\[
\begin{align*}
&= \sum_{n, K} \int_{\Omega_n} \int_{\Omega_K} \langle \text{Res}, D^{SD} \text{Res} \rangle \, dx \, dt \quad \text{(from (3.18))}, \\
&\geq C C^{SD} \Delta x \sum_{n, K} \int_{\Omega_n} \int_{\Omega_K} \| \text{Res} \|^2 \, dx \, dt \\
&= C C^{SD} \Delta x \sum_{n, K} \int_{\Omega_n} \int_{\Omega_K} \| (V^\Delta x)_t \| + \sum_{k=1}^{d} \| F^k (V^\Delta x)_{x_k} \|^2 \, dx \, dt \\
&\geq 0. 
\end{align*}
\] (3.26)

**Claim 2:** The shock-capturing operator (3.20) is positive, i.e.:

\[
\mathcal{B}_{SC}(V^\Delta x, V^\Delta x) \geq 0. 
\] (3.27)

First, we observe that the uniform convexity of the entropy function \( S \) implies that the matrices \( U_V \) and \( U_V^{-1} \) are positive definite. This implies that the term \( D^{SC} \geq 0 \).

We set as test function \( W^\Delta x = V^\Delta x \) in (3.20a) and obtain

\[
\mathcal{B}_{SC}(V^\Delta x, V^\Delta x) \\
= \sum_{n, K} \int_{\Omega_n} \int_{\Omega_K} D^{SC}_{n, K} \left( \langle V^\Delta x_t, U_V V^\Delta x_t \rangle + \sum_{k=1}^{d} \frac{\Delta x^2_k}{(\Delta t^2)^2} \langle V^\Delta x_{x_k}, U_V V^\Delta x_{x_k} \rangle \right) \, dx \, dt \\
\geq \sum_{n, K} \lambda_1 D^{SC}_{n, K} \int_{\Omega_n} \int_{\Omega_K} \| \nabla_{xt} V^\Delta x \|^2 \, dx \, dt \\
\geq 0.
\] (3.28)

Here, \( \lambda_1 \) is the smallest eigenvalue of the positive definite matrix \( U_V \). Under the assumption of (iii), i.e. that \( U_V \) is uniformly positive definite, one obtains \( \lambda_1 > 0 \).

**Claim 3:** Define the spatial part of the DG form \( \mathcal{B}_{DG} \) (3.16) as

\[
\begin{align*}
\mathcal{B}^s_{DG}(V^\Delta x, W^\Delta x) &= - \sum_{n, K} \int_{\Omega_n} \int_{\Omega_K} \sum_{k=1}^{d} \langle F^k(V^\Delta x), W^\Delta x_{x_k} \rangle \, dx \, dt \\
&\quad + \sum_{n, K, K'} \int_{\Omega_n} \int_{\partial_{K K'}} \left( \sum_{k=1}^{d} \langle \mathbb{P}^k (V^\Delta x_{K, -}, V^\Delta x_{K, +}), W^\Delta x_{x_k} \rangle \nu^k_{K K'} \right) \, d\sigma \, dt \\
&\quad - \frac{1}{2} \sum_{n, K, K'} \int_{\Omega_n} \int_{\partial_{K K'}} \langle W^\Delta x_{K, -}, D(V^\Delta x_{K, +} - V^\Delta x_{K, -}) \rangle \, d\sigma \, dt.
\end{align*}
\] (3.29)

We claim that

\[
\mathcal{B}^s_{DG}(V^\Delta x, V^\Delta x) \geq 0. 
\] (3.30)
3. The shock-capturing streamline diffusion DG formulation

From the definition of the entropy potential $\Psi^k$, we obtain that for all $1 \leq k \leq d$,

$$
\Psi^k_{x_k} = (\langle V, F^k \rangle - Q^k)_{x_k} = \langle V_{x_k}, F^k \rangle + \langle V, F^k_{x_k} \rangle - (Q^k)_{x_k} = \langle V_{x_k}, F^k \rangle \quad (\text{from definition of } Q^k).
$$

Therefore,

$$
\sum_{n,K} \int_{I_n} \int_{K} \sum_{k=1}^{d} (F^k(V^{\Delta x}_x), V^{\Delta x}_{x_k}) dx dt = \sum_{n,K} \int_{I_n} \int_{K} \sum_{k=1}^{d} \Psi^k(V^{\Delta x}_{x_k}) dx dt
$$

$$
= \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \sum_{k=1}^{d} \Psi^k(V^{\Delta x}_{K_{k,-}}) \nu^k_{K K'} d\sigma dt.
$$

Using the above identities, we obtain,

$$
B_{DG}^{s}(V^{\Delta x}, V^{\Delta x})
$$

$$
= - \sum_{n,K} \int_{I_n} \int_{K} \sum_{k=1}^{d} (F^k(V^{\Delta x}_x), V^{\Delta x}_{x_k}) dx dt
$$

$$
+ \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \sum_{k=1}^{d} (F^k(V^{\Delta x}_x), V^{\Delta x}_{x_k}) \nu^k_{K K'} d\sigma dt.
$$

$$
- \frac{1}{2} \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \langle V^{\Delta x}_{K_{k,-}}, D(V^{\Delta x}_{K_{k,+}} - V^{\Delta x}_{K_{k,-}}) \rangle d\sigma dt
$$

$$
= \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \sum_{k=1}^{d} \left( \langle F^k(V^{\Delta x}_x), V^{\Delta x}_{x_k} \rangle - \Psi^k(V^{\Delta x}_{x_k}) \right) \nu^k_{K K'} d\sigma dt
$$

$$
- \frac{1}{2} \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \langle V^{\Delta x}_{K_{k,-}}, D(V^{\Delta x}_{K_{k,+}} - V^{\Delta x}_{K_{k,-}}) \rangle d\sigma dt
$$

$$
= \frac{1}{2} \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \sum_{k=1}^{d} \left( \langle F^k(V^{\Delta x}_x), V^{\Delta x}_{x_k} \rangle - \Psi^k(V^{\Delta x}_{x_k}) \right) \nu^k_{K K'} d\sigma dt
$$

$$
+ \frac{1}{2} \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \sum_{k=1}^{d} \left( \langle F^k(V^{\Delta x}_x), V^{\Delta x}_{x_k} \rangle - \Psi^k(V^{\Delta x}_{x_k}) \right) \nu^k_{K K'} d\sigma dt
$$

$$
- \frac{1}{4} \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \langle V^{\Delta x}_{K_{k,-}}, D(V^{\Delta x}_{K_{k,+}} - V^{\Delta x}_{K_{k,-}}) \rangle d\sigma dt
$$

$$
- \frac{1}{4} \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \langle V^{\Delta x}_{K_{k,-}}, D(V^{\Delta x}_{K_{k,+}} - V^{\Delta x}_{K_{k,-}}) \rangle d\sigma dt. \quad (3.31)
$$
3.6. Entropy stability for nonlinear systems

Changing the roles of $K$ and $K'$ in the second and fourth sums and rewriting the arguments of the sums using the fact that the approximate solutions have compact support in $\Omega$, the above expression reduces to

$$
B_{DG}^s(\mathbf{V}^{\Delta x}, \mathbf{V}^{\Delta x})
= -\frac{1}{2} \sum_{n,K,K'} \int_{I^n} \sum_{\partial_{KK'}}^d \left( \langle F_k^{*}(\mathbf{V}^{\Delta x}_{K,-}, \mathbf{V}^{\Delta x}_{K,+}) - \langle F_k^{*}(\mathbf{V}^{\Delta x}_{K,-}, \mathbf{V}^{\Delta x}_{K,+}) \rangle \right) \nu_{KK'}^k d\sigma dt
$$

$$
\text{= 0 from defn entropy-conservative flux (3.9)}
$$

$$
B_{DG}^t(\mathbf{V}^{\Delta x}, \mathbf{V}^{\Delta x})
= -\sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}^{\Delta x}), \mathbf{W}^{\Delta x} \rangle dx dt
$$

$$
\text{+ } \sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}_{n,1,-}^{\Delta x}), \mathbf{W}_{n,1,-}^{\Delta x} \rangle dx - \sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}_{n,-}^{\Delta x}), \mathbf{W}_{n,+}^{\Delta x} \rangle dx.
$$

Claim 4: Define the temporal part of the DG form $B_{DG}$ (3.16) as

$$
B_{DG}^t(\mathbf{V}^{\Delta x}, \mathbf{V}^{\Delta x})
= -\sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}^{\Delta x}), \mathbf{W}^{\Delta x}_t \rangle dx dt
$$

$$
\text{+ } \sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}_{n,1,-}^{\Delta x}), \mathbf{W}_{n,1,-}^{\Delta x} \rangle dx - \sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}_{n,-}^{\Delta x}), \mathbf{W}_{n,+}^{\Delta x} \rangle dx.
$$

We claim that

$$
B_{DG}^t(\mathbf{V}^{\Delta x}, \mathbf{V}^{\Delta x}) \geq \int_{\Omega} S(\mathbf{U}(\mathbf{V}^{\Delta x}(x, t^N)))dx - \int_{\Omega} S(\mathbf{U}(\mathbf{V}^{\Delta x}(x, t^0)))dx.
$$

Setting $\mathbf{W}^{\Delta x} = \mathbf{V}^{\Delta x}$ in (3.33), we obtain

$$
B_{DG}^t(\mathbf{V}^{\Delta x}, \mathbf{V}^{\Delta x})
= -\sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}^{\Delta x}), \mathbf{V}^{\Delta x}_t \rangle dx dt
$$

$$
\text{+ } \sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}_{n,1,-}^{\Delta x}), \mathbf{V}_{n,1,-}^{\Delta x} \rangle dx - \sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}(\mathbf{V}_{n,-}^{\Delta x}), \mathbf{V}_{n,+}^{\Delta x} \rangle dx
$$

$$
\text{= } \sum_{n,K} \int_{I^n} \int_K \langle \mathbf{U}_t(\mathbf{V}^{\Delta x}), \mathbf{V}^{\Delta x} \rangle dx dt \text{ (integrating by parts)}
$$

25
3. The shock-capturing streamline diffusion DG formulation

\[- \sum_{n,K} \int_K \left\langle U(V_{n+1,-}^{\Delta x}), V_{n+1,-}^{\Delta x} \right\rangle dx + \sum_{n,K} \int_K \left\langle U(V_{n+}^{\Delta x}), V_{n+}^{\Delta x} \right\rangle dx \]

\[+ \sum_{n,K} \int_K \left\langle U(V_{n+1,-}^{\Delta x}), V_{n+1,-}^{\Delta x} \right\rangle dx - \sum_{n,K} \int_K \left\langle U(V_{n+}^{\Delta x}), V_{n+}^{\Delta x} \right\rangle dx \]

\[= \sum_{n,K} \int_{I^n} \int_K S(U(V_{n-}^{\Delta x}))_t \, dx \, dt \quad \text{(definition of entropy variables)} \]

\[- \sum_{n,K} \int_K \left\langle \left( U(V_{n-}^{\Delta x}) - U(V_{n+}^{\Delta x}) \right), V_{n+}^{\Delta x} \right\rangle dx \]

\[= \sum_{n,K} \int_K \left( S(U(V_{n+}^{\Delta x})) - S(U(V_{n+}^{\Delta x})) \right) dx \]

\[+ \sum_{n,K} \int_K \left( S(U(V_{n-}^{\Delta x})) - S(U(V_{n+}^{\Delta x})) \right) dx \]

\[- \sum_{n,K} \int_K \left\langle \left( U(V_{n-}^{\Delta x}) - U(V_{n+}^{\Delta x}) \right), V_{n+}^{\Delta x} \right\rangle dx \]

Following [48, 2], we define \( V(\theta) = V_{n+}^{\Delta x} + \theta(V_{n-}^{\Delta x} - V_{n+}^{\Delta x}) = \theta V_{n-}^{\Delta x} + (1 - \theta) V_{n+}^{\Delta x} \) and compute

\[S(U(V_{n-}^{\Delta x})) - S(U(V_{n+}^{\Delta x})) \]

\[= \int_{0}^{1} \frac{d}{d\theta} \left( S(U(V(\theta))) \right) d\theta \]

\[= \int_{0}^{1} \left\langle S(U(V(\theta))), U_V(\theta)V_{\theta}(\theta) \right\rangle \, d\theta \]

\[= \int_{0}^{1} \langle V_{n+}^{\Delta x}, U_V(\theta)V_{\theta}(\theta) \rangle d\theta + \int_{0}^{1} \theta \langle (V_{n-}^{\Delta x} - V_{n+}^{\Delta x}), U_V(\theta)(V_{n-}^{\Delta x} - V_{n+}^{\Delta x}) \rangle d\theta \]

\[= \langle V_{n+}^{\Delta x}, U(V_{n+}^{\Delta x}) - U(V_{n+}^{\Delta x}) \rangle + \int_{0}^{1} \theta \langle (V_{n-}^{\Delta x} - V_{n+}^{\Delta x}), U_V(\theta)(V_{n-}^{\Delta x} - V_{n+}^{\Delta x}) \rangle d\theta. \]

(3.35)

Thus, we obtain

\[B_{DG}^{t}(V_{\Delta x}, V_{\Delta x}) \]

\[= \int_{\Omega} S(U(V_{\Delta x}(x, t^N))) \, dx - \int_{\Omega} S(U(V_{\Delta x}(x, t_0))) \, dx \]

\[+ \sum_{n,K} \int_{I^n} \int_{K} \theta \langle (V_{n-}^{\Delta x} - V_{n+}^{\Delta x}), U_V(\theta)(V_{n-}^{\Delta x} - V_{n+}^{\Delta x}) \rangle d\theta \, dx. \]

(3.35)
3.6. Entropy stability for nonlinear systems

\[
\geq \int_{\Omega} S(U(V^\Delta x(x,t^N)))dx - \int_{\Omega} S(U(V^\Delta x(x,t^0)))dx
+ \sum_{n,K} \frac{\lambda_1}{2} \int_{K} \|V^\Delta x_{n,-} - V^\Delta x_{n,+}\|^2 dx
\geq \int_{\Omega} S(U(V^\Delta x(x,t^N)))dx - \int_{\Omega} S(U(V^\Delta x(x,t^0)))dx.
\]  

(3.36)

Now combining the four claims, we obtain

\[
B_{DG}(V^\Delta x, V^\Delta x) + B_{SD}(V^\Delta x, V^\Delta x) + B_{SC}(V^\Delta x, V^\Delta x) = 0
\Rightarrow B'_{DG}(V^\Delta x, V^\Delta x) + B'_{DG}(V^\Delta x, V^\Delta x) + B_{SD}(V^\Delta x, V^\Delta x) + B_{SC}(V^\Delta x, V^\Delta x) = 0
\Rightarrow \int_{\Omega} S(U(V^\Delta x(x,t^N)))dx - \int_{\Omega} S(U(V^\Delta x(x,t^0)))dx \leq 0
\Rightarrow \int_{\Omega} S(U(V^\Delta x(x,t^N)))dx \leq \int_{\Omega} S(U(V^\Delta x(x,t^0)))dx.
\]

This is the upper bound in the entropy estimate (3.22).

To prove the lower bound on entropy in estimate (3.22), we follow Barth [2] and define the domain average,

\[
U^*(t^N) = \frac{1}{\text{meas}(\Omega)} \int_{\Omega} U(V^\Delta x(x,t^N))dx.
\]

(3.37)

From conservation, we know that \(U^*(t^N) = U^*(t^0)\).

For any given time level, we have

\[
S(U) = S(U^*) + (V(U^*), (U - U^*)) + \int_{0}^{1} (1 - \theta)(S(U - U^*), SUU(U(\theta))(U - U^*))d\theta,
\]

with \(U(\theta) = \theta U + (1 - \theta)U^*\). Integrating the above equation over space and using the definition of domain average (3.37) and the strict convexity of the entropy function, we obtain,

\[
\int_{\Omega} S(U^*)dx \leq \int_{\Omega} S(U)dx
\]

for any time level.

As a consequence of conservation, we obtain,

\[
\int_{\Omega} S(U^*(t^0))dx = \int_{\Omega} S(U^*(t^N))dx \leq \int_{\Omega} S(U(V^\Delta x(x,t^N)))dx,
\]

thus obtaining the lower bound in the entropy estimate (3.22).
3. The shock-capturing streamline diffusion DG formulation

Combining the above claims of positivity and using the above estimate results in the following weak BV estimate:

\[
\sum_{n,K} \int_K \left\| \nabla \Delta x_n(\mathbf{V}^\Delta x_{n,-} - \mathbf{V}^\Delta x_{n,+}) \right\|^2 dx + \sum_{n,K,K'} \int_{I^n} \int_{\partial K \cap K'} \left\langle \nabla \Delta x_{K,+} - \nabla \Delta x_{K,-}, D(\nabla^\Delta x_{K,+} - \nabla^\Delta x_{K,-}) \right\rangle d\sigma dt
+ \Delta x \sum_{n,K} \int_{I^n} \int_K \left\| \mathbf{U}(\mathbf{V}^\Delta x)_t \right\|^2 dx dt
+ \sum_{n,K} \lambda_1 D_{n,K}^{SC} \int_{I^n} \int_K \left\| \nabla_{xt} \mathbf{V} \right\|^2 dx dt \leq C. \tag{3.38}
\]

The weak BV estimate (3.23) follows from the estimate

\[
\Delta x \sum_{n,K} \overline{\text{Res}_{n,K}} \left( \int_{I^n} \int_K \left\| \nabla_{xt} \mathbf{V}^\Delta x \right\|^2 dx dt \right)^{\frac{1}{2}}
+ (\Delta x)^{\frac{1}{2}} \sum_{n,K} \overline{\text{BRes}_{n,K}} \left( \int_{I^n} \int_K \left\| \nabla_{xt} \mathbf{V}^\Delta x \right\|^2 dx dt \right)^{\frac{1}{2}} \leq C \tag{3.39}
\]

that is a consequence of the definitions of \( D_{n,K}^{SC}, \theta \) and the estimate (3.38):

We use the basic estimate

\[
D \leq \max \left\{ E, 2 \frac{D^2}{D + E} \right\} \tag{3.40}
\]

with \( D = \left( \int_{I^n} \int_K \left\| \nabla_{xt} \mathbf{V}^\Delta x \right\|^2 dx dt \right)^{\frac{1}{2}} \) and \( E = |K|^{\frac{1}{2}} (\Delta t^n)^{\frac{1}{2}} \Delta x^{-1+\theta} \). We have

\[
\sum_{n,K} \Delta x \overline{\text{Res}_{n,K}} |K|^{\frac{1}{2}} (\Delta t^n)^{\frac{1}{2}} \Delta x^{-1+\theta}
\leq \Delta x^{\theta} \left( \sum_{n,K} \overline{\text{Res}_{n,K}}^2 \right)^{\frac{1}{2}} \left( \sum_{n,K} |K| \Delta t^n \right)^{\frac{1}{2}}
\leq \Delta x^{\theta} C (\Delta x)^{-\frac{1}{2}} C \quad \text{(by (3.38))}
= C \Delta x^{\theta - \frac{1}{2}}
\leq C \quad \text{(condition on } \theta \text{)}
\]

and we know because of (3.38) that

\[
\sum_{n,K} \frac{\Delta x \overline{\text{Res}_{n,K}}}{D + |K|^{\frac{1}{2}} (\Delta t^n)^{\frac{1}{2}} \Delta x^{-1+\theta}} D^2 \leq C. \tag{3.41}
\]
Therefore, we conclude using (3.40)

\[ \sum_{n,K} \Delta x \overline{\text{Res}}_{n,K} \left( \int_{I^n} \int_{K} \| \nabla_{xt} \mathbf{V} \Delta x \|^2 \, dx \, dt \right)^{\frac{1}{2}} \leq C. \quad (3.43) \]

For the boundary residual, we proceed similarly. We obtain

\[ \sum_{n,K} (\Delta x)^{\frac{3}{2}} \overline{\text{BRes}}_{n,K} |K|^\frac{1}{2} (\Delta t^n)^{\frac{1}{2}} \Delta x^{-1+\theta} \]

\[ \leq \Delta x^{-\frac{1}{2}+\theta} \left( \sum_{n,K} \overline{\text{BRes}}_{n,K}^2 \right)^{\frac{1}{2}} \left( \sum_{n,K} |K| \Delta t^n \right)^{\frac{1}{2}} \]

\[ \leq \Delta x^{-\frac{1}{2}+\theta} CC \quad \text{by (3.38)} \]

\[ = C \Delta x^{\theta-\frac{1}{2}} \]

\[ \leq C \quad \text{(condition on } \theta) \]

and we know because of (3.38)

\[ \sum_{n,K} \frac{(\Delta x)^{\frac{3}{2}} \overline{\text{BRes}}_{n,K}}{D + |K|^\frac{1}{2} (\Delta t^n)^{\frac{1}{2}} \Delta x^{-1+\theta}} D^2 \leq C. \quad (3.45) \]

Thus, we conclude using (3.40)

\[ \sum_{n,K} (\Delta x)^{\frac{3}{2}} \overline{\text{BRes}}_{n,K} \left( \int_{I^n} \int_{K} \| \nabla_{xt} \mathbf{V} \Delta x \|^2 \, dx \, dt \right)^{\frac{1}{2}} \leq C. \quad (3.46) \]

**Remark 3.6.2.** The lowest order version of the streamline diffusion shock-capturing DG scheme (3.2) is given by considering the test functions in the space $V_0$ and reduces to a generalization (with implicit time stepping) of the first-order finite volume scheme proposed by Tadmor in [47]. Thus, the scheme (3.2) can be considered as a finite element fully discrete generalization of well-known entropy-stable finite volume schemes.

**Remark 3.6.3.** The proof of the entropy stability (for the DG form) relies on three main ingredients:

- The discretisation of the entropy variables rather than the conservative variables allows to use $\mathbf{V} \Delta x$ as test function and therefore, to mimic the proof in the continuous case.

- The entropy-stable fluxes then lead to positivity of the spatial part of the DG form.

- Thanks to the space-time formulation, we obtain entropy stability for the fully discrete scheme (on general unstructured meshes).
4. Convergence analysis for systems of conservation laws

Given the problems with well-posedness of entropy solutions in the case of systems of conservation laws in multiple spatial dimensions, one cannot hope to have a scheme that is convergent to the entropy solution. Consistent with this lack of theory for systems of conservation laws, the analysis of numerical schemes for conservation laws has focussed on two special cases:

(i) Scalar conservation laws, i.e. (1.3) with \( m = 1 \).

(ii) Linear symmetrisable systems, i.e. conservation laws of the form (1.21).

A robust numerical scheme should converge to the entropy solutions of scalar conservation laws and the weak solutions of the linear symmetrisable systems. We will show that our numerical scheme (3.2) does indeed converge in both of the above cases (with some restrictions in the scalar case).

However, we would like to prove some results for the general case of a nonlinear system in several space dimensions (in addition to entropy stability). Given the discussion in the introduction and in Chapter 1 about the suitability of entropy measure-valued solutions (1.14), (1.15) as the appropriate solution concept in this context, the main theoretical results of this chapter are to show that approximate solutions of the numerical scheme (3.2) converge to an entropy measure-valued solution for a nonlinear system of conservation laws (1.3). Furthermore, we will show that in the special cases of scalar conservation laws as well as linear symmetrisable systems, convergence to measure-valued solutions automatically implies convergence to entropy solutions. This chapter has been published in [25].

4.1. Convergence to measure-valued solution

We start with the following convergence theorem,

**Theorem 4.1.1.** Let \( U^{\Delta x} = U(V^{\Delta x}) \) be the approximate solutions of the system (1.3) generated by the streamline diffusion shock-capturing DG scheme (3.2). Under the assumptions that the diffusion matrices \( D \) are uniformly positive definite and the approximate solutions satisfy the uniform \( L^\infty \) bound,

\[
\|V^{\Delta x}\|_{L^\infty(\Omega \times \mathbb{R}_+)} \leq C, \tag{4.1}
\]

\( x \)
4. Convergence analysis for systems of conservation laws

the approximate solutions converge to a measure-valued solution \((1.14)\) of the conservation law \((1.3)\).

**Proof.** For simplicity, we will only consider the case \(p \geq 1\). The \(p = 0\) case can be proved using a Lax-Wendroff type argument [17].

To show convergence of the approximate solutions to a measure-valued solution of \((1.3)\), we consider any compactly supported test function \(\varphi \in (C_c^\infty(\Omega \times [0, \infty)))^m\). We denote by

\[
\varphi^{\Delta x} = \Pi^{\Delta x}(\varphi),
\]

the projection into the space \((\mathbb{P}_p)^m\) with \(\Pi^{\Delta x}|_{K \times I^n} : (H^1)^m \rightarrow (\mathbb{P}_p(K \times I^n))^m\) satisfying

\[
\int_{I^n} \int_K \langle (\Pi^{\Delta x}(\varphi))_t, \bar{U}_V W_t \rangle + \sum_{k=1}^d \frac{\Delta x_k^2}{(\Delta t)^2} \langle (\Pi^{\Delta x}(\varphi))_{x_k}, \bar{U}_V W_{x_k} \rangle \, dx dt = \int_{I^n} \int_K \varphi dx dt
\]

for all \(W \in (\mathbb{P}_p(K \times I^n))^m\). Note that the scaling \(U_V\) is a constant matrix as it is evaluated at the (spacetime) cell average \(\bar{V}_{n,K}\). Hence, this projection operator for the infinitely smooth function \(\varphi\) satisfies the following stability and approximation properties (see [29] or other references therein),

\[
\|\nabla_{xt} \varphi^{\Delta x}\|_{L^2(K \times I^n)} \leq C \|\nabla_{xt} \varphi\|_{L^2(K \times I^n)}
\]

\[
\|\varphi - \varphi^{\Delta x}\|_{L^2(K \times I^n)} \leq C \Delta x \|\nabla_{xt} \varphi\|_{L^2(K \times I^n)}
\]

\[
\|\varphi - \varphi^{\Delta x}\|_{L^2(\partial(K \times I^n))} \leq C \Delta x^{\frac{1}{2}} \|\nabla_{xt} \varphi\|_{L^2(K \times I^n)}
\]

(4.4)

It satisfies also the following stronger versions of the approximation properties

\[
\|\varphi - \varphi^{\Delta x}\|_{H^1(K \times I^n)} \leq C \Delta x \|\varphi\|_{H^2(K \times I^n)},
\]

\[
\|\varphi - \varphi^{\Delta x}\|_{L^2(\partial(K \times I^n))} \leq C \Delta x^{\frac{1}{2}} \|\varphi\|_{H^2(K \times I^n)}
\]

(4.5)

which follow from the regularity of solutions of the elliptic equation \((4.3)\) that defines the projection operator and from the fact that the test function \(\varphi\) is infinitely differentiable.

The proof of convergence consists of the following claims:

**Claim 1:** The streamline diffusion operator defined in \((3.17)\) satisfies

\[
B_{SD}(V^{\Delta x}, \varphi^{\Delta x}) \rightarrow 0 \quad \text{as } \Delta x \rightarrow 0.
\]

(4.6)
4.1. Convergence to measure-valued solution

To prove the claim, we calculate the bilinear form (3.17),

\[
| B_{SD}(V_{\Delta x}, \varphi_{\Delta x}) |
\]

\[
= \left| \sum_{n,K} \int_{I_n} \int_{K} \left( \left( U_{V}(V_{\Delta x})\varphi_{\Delta x} + \sum_{k=1}^{d} F_{V}^{k}(V_{\Delta x})\varphi_{x_k} \right) , D^{SD}\text{Res} \right) d^x dt \right|
\]

\[
\leq C \Delta x \| V_{\Delta x} \|_{L^{\infty}(\Omega \times [0,T])} \sum_{n,K} \| \varphi_{\Delta x} \|_{H^1(K \times I_n)^2} \left( \int_{I_n} \int_{K} \| \text{Res} \|^2 d^x dt \right)^{\frac{1}{2}}
\]

\[
\leq C \| V_{\Delta x} \|_{L^{\infty}(\Omega \times [0,T])} \Delta x^{\frac{1}{2}} \left( \sum_{n,K} \| \varphi_{\Delta x} \|_{H^1(K \times I_n)^2} \right)^{\frac{1}{2}} \left( \Delta x \sum_{n,K} \int_{I_n} \int_{K} \| \text{Res} \|^2 d^x dt \right)^{\frac{1}{2}}
\]

\[
\leq C \| V_{\Delta x} \|_{L^{\infty}(\Omega \times [0,T])} \Delta x^{\frac{1}{2}} \| \varphi \|_{H^1(\Omega \times [0,T])} \to 0 \quad \text{as } \Delta x \to 0.
\]

Here, the last estimate follows the stability of the projection operator (4.4) and from the weak BV estimate (3.23).

**Claim 2:** The shock-capturing operator defined in (3.20) satisfies

\[
B_{SC}(V_{\Delta x}, \varphi_{\Delta x}) \to 0 \quad \text{as } \Delta x \to 0.
\]

Define

\[
B_{SC,1}(V_{\Delta x}, \varphi_{\Delta x})
\]

\[
:= \sum_{n,K} \int_{I_n} \int_{K} D_{n,K}^{SC,1} \left( \left( \varphi_{\Delta x}, \bar{U}_{V}V_{\Delta x} \right) + \sum_{k=1}^{d} \frac{\Delta x_k^2}{\Delta t^2 n} \left( \varphi_{x_k}, \bar{U}_{V}V_{x_k} \right) \right) d^x dt,
\]

with

\[
D_{n,K}^{SC,1} = \frac{\Delta x C_{SC}^{Res_{n,K}}}{\sqrt{\int_{I_n} \int_{K} \left( \left( V_{\Delta x}, \bar{U}_{V}V_{\Delta x} \right) + \sum_{k=1}^{d} \frac{\Delta x_k^2}{\Delta t^2 n} \left( V_{x_k}, \bar{U}_{V}V_{x_k} \right) \right) d^x dt + \epsilon}}.
\]

Using the uniform positivity of \( U_{V} \), which is a consequence of the uniform convexity of \( S \) and the \( L^{\infty} \) bound on \( V \), we obtain that

\[
D_{n,K}^{SC,1} \leq C \frac{\Delta x C_{SC}^{Res_{n,K}}}{\left( \int_{K} \int_{I_n} \| \nabla_x V \|^2 d^x dt \right)^{\frac{1}{2}} + \Delta x^{\frac{d}{2} - \frac{1}{2} + \theta}}.
\]
4. Convergence analysis for systems of conservation laws

Therefore,

\[
|\mathcal{B}_{SC,1}(V^{\Delta x}, \varphi^{\Delta x})| \\
\leq C \Delta x \|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \\
\left( \sum_{n,K} \overline{\text{Res}_{n,K}} \|\varphi^{\Delta x}\|_{H^1(K \times I^n)} \left( \frac{\int_K \int_{I^n} \|\nabla_{xt} V\|^2 \, dx \, dt}{\int_K \int_{I^n} \|\nabla_{xt} V\|^2 \, dx \, dt} \right)^{\frac{1}{2}} \frac{1}{2} + \Delta x^{\frac{d}{2} - \frac{1}{2} + \theta} \right)
\]

\[
\leq C(\Delta x)^{\frac{1}{2}} \left( \sum_{n,K} \|\varphi^{\Delta x}\|^2_{H^1(K \times I^n)} \right)^{\frac{1}{2}} \left( \sum_{n,K} \|\text{Res}_{n,K}\| \right)^{\frac{1}{2}} \leq C(\Delta x)^{\frac{1}{2}} \|\varphi\|_{H^1(\Omega \times [0,T])} \rightarrow 0 \text{ as } \Delta x \rightarrow 0.
\]

Here, the last estimate follows the stability of the projection operator and from the weak BV estimate (3.23).

Similarly, define

\[
\mathcal{B}_{SC,2}(V^{\Delta x}, \varphi^{\Delta x}) := \sum_{n,K} \int_I \int_K D_{n,K}^{SC,2} \left( \langle \varphi^{\Delta x}_t, \tilde{U}_V V^{\Delta x} \rangle + \sum_{k=1}^d \Delta x K^2 \langle \varphi^{\Delta x}_{x_k}, \tilde{U}_V V^{\Delta x}_{x_k} \rangle \right) \, dx \, dt,
\]

with

\[
D_{n,K}^{SC,2} = \frac{(\Delta t^n)^{\frac{1}{2}} C^{SC} \overline{\text{BR}_{n,K}}}{\sqrt{\int_I \int_K \left( \langle \nabla^{\Delta x}_t, \tilde{U}_V V^{\Delta x} \rangle + \sum_{k=1}^d \langle \nabla^{\Delta x}_{x_k}, \tilde{U}_V V^{\Delta x}_{x_k} \rangle \right) \, dx \, dt + \epsilon}}.
\]

Using the uniform positivity of $U_V$, we obtain that,

\[
D_{n,K}^{SC,2} \leq C \frac{(\Delta x)^{\frac{1}{2}} C^{SC} \overline{\text{BR}_{n,K}}}{\left( \int_I \int_K \|\nabla_{xt} V\|^2 \, dx \, dt \right)^{\frac{1}{2}} + \Delta x^{\frac{d}{2} - \frac{1}{2} + \theta} \cdot}
\]

Therefore,

\[
|\mathcal{B}_{SC,2}(V^{\Delta x}, \varphi^{\Delta x})| \\
\leq C(\Delta x)^{\frac{1}{2}} \|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \\
\left( \sum_{n,K} \overline{\text{BR}_{n,K}} \|\varphi^{\Delta x}\|^2_{H^2(K \times I^n)} \left( \frac{\int_K \int_{I^n} \|\nabla_{xt} V\|^2 \, dx \, dt}{\int_K \int_{I^n} \|\nabla_{xt} V\|^2 \, dx \, dt} \right)^{\frac{1}{2}} \frac{1}{2} + \Delta x^{\frac{d}{2} - \frac{1}{2} + \theta} \right)
\]

\[
\leq C(\Delta x)^{\frac{1}{2}} \left( \sum_{n,K} \|\varphi^{\Delta x}\|^2_{H^2(K \times I^n)} \right)^{\frac{1}{2}} \left( \sum_{n,K} |\text{BR}_{n,K}|^2 \right)^{\frac{1}{2}} \leq C(\Delta x)^{\frac{1}{2}} \|\varphi\|_{H^2(\Omega \times [0,T])} \rightarrow 0 \text{ as } \Delta x \rightarrow 0.
\]
4.1. Convergence to measure-valued solution

Here, the last estimate follows the stability of the projection operator (4.4) and from the fact that the estimate on the $B\text{Res}$ term is a straightforward consequence of the weak BV estimate (3.23) using the uniform positivity of $D$.

**Claim 3:** Let a part of the DG form (3.16) be defined as

$$B_{\text{DG}}(V^{\Delta x}, \varphi^{\Delta x}) = \sum_{n,K,K'} \int_I \int_{\partial K K'} \left( \sum_{k=1}^{d} \langle F^{k, \ast}(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}), \varphi_{K,-}^{\Delta x} \rangle \nu_{K K'}^k \right) d\sigma dt. \quad (4.8)$$

Then

$$B_{\text{DG}}(V^{\Delta x}, \varphi^{\Delta x}) \to 0 \quad \text{as} \quad \Delta x \to 0. \quad (4.9)$$

Splitting and changing the role of $K$ and $K'$ in the second sum, we perform the following calculation,

$$B_{\text{DG}}(V^{\Delta x}, \varphi^{\Delta x}) = \frac{1}{2} \sum_{n,K,K'} \int_I \int_{\partial K K'} \left( \sum_{k=1}^{d} \langle F^{k, \ast}(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}), (\varphi_{K,-}^{\Delta x} - \varphi) \rangle \nu_{K K'}^k \right) d\sigma dt$$

$$- \frac{1}{2} \sum_{n,K,K'} \int_I \int_{\partial K K'} \left( \sum_{k=1}^{d} \langle F^{k, \ast}(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}), (\varphi_{K,+}^{\Delta x} - \varphi) \rangle \nu_{K K'}^k \right) d\sigma dt.$$

Adding and subtracting $\varphi$ to the above and using the continuity of the smooth test function $\varphi$ across every edge, results in

$$|B_{\text{DG}}(V^{\Delta x}, \varphi^{\Delta x})|$$

$$\leq \frac{1}{2} \left| \sum_{n,K,K'} \int_I \int_{\partial K K'} \left( \sum_{k=1}^{d} \langle F^{k, \ast}(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}), (\varphi_{K,-}^{\Delta x} - \varphi) \rangle \nu_{K K'}^k \right) d\sigma dt \right|$$

$$+ \frac{1}{2} \left| \sum_{n,K,K'} \int_I \int_{\partial K K'} \left( \sum_{k=1}^{d} \langle F^{k, \ast}(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}), (\varphi_{K,+}^{\Delta x} - \varphi) \rangle \nu_{K K'}^k \right) d\sigma dt \right|$$

$$\leq C\|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \sum_{n,K} (\text{meas}(\partial(K \times I^n)))^{1/2} \|\varphi^{\Delta x} - \varphi\|_{L^2(\partial(K \times I^n))},$$

$$\leq C\Delta x^{3/2}\|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \sum_{n,K} (\text{meas}(\partial(K \times I^n)))^{1/2} \|\varphi\|_{H^2(\Omega \times I^n)},$$

$$\leq C\Delta x\|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \|\varphi\|_{H^2(\Omega \times [0,T])} \to 0 \quad \text{as} \quad \Delta x \to 0.$$
4. Convergence analysis for systems of conservation laws

we have

\[ B_{DG}^{\Delta x}(V^{\Delta x}, \varphi^{\Delta x}) \to 0 \quad \text{as} \quad \Delta x \to 0. \]  \quad (4.11)

We proceed as in the previous claim to calculate,

\[ |B_{DG}^{\Delta x}(V^{\Delta x}, \varphi^{\Delta x})| \]

\[ \leq C \|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \sum_{n,K'} \int_{I^n} \int_{\partial_K} \langle (\varphi_{K,-}^{\Delta x} - \varphi), D(V_{K,-}^{\Delta x} - V_{K,-}^{\Delta x}) \rangle \ d\sigma dt \]

\[ + C \|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \sum_{n,K'} \int_{I^n} \int_{\partial_K} \langle (\varphi_{K,+}^{\Delta x} - \varphi), D(V_{K,+}^{\Delta x} - V_{K,+}^{\Delta x}) \rangle \ d\sigma dt \]

\[ \leq C \left( \sum_{n,K,K'} \int_{I^n} \int_{\partial_K} \|V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}\|^2 \ d\sigma dt \right)^{\frac{1}{2}} \left( \sum_{n,K,K'} \|\varphi_{K,+}^{\Delta x} - \varphi\|^2_{L^2(\partial_K \times I^n)} \right)^{\frac{1}{2}} \]

\[ \leq C \Delta x^{\frac{1}{2}} \left( \sum_{n,K,K'} \int_{I^n} \int_{\partial_K} \|V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}\|^2 \ d\sigma dt \right)^{\frac{1}{2}} \|\varphi\|_{H^1(\Omega \times [0,T])} \quad \text{(by (4.4))} \]

\[ \to 0, \quad \text{as} \quad \Delta x \to 0, \quad \text{(by the weak BV estimate (3.23))}. \]

Claim 5: Defining a part of the DG form (3.3) as

\[ B_{DG}^{11}(V^{\Delta x}, \varphi^{\Delta x}) = \sum_{n,K} \int_K \langle U_{n+1,-}^{\Delta x}, \varphi_{n+1,-}^{\Delta x} \rangle \ dx - \sum_{n,K} \int_K \langle U_{n,-}^{\Delta x}, \varphi_{n,+}^{\Delta x} \rangle \ dx \]

we have

\[ B_{DG}^{11}(V^{\Delta x}, \varphi^{\Delta x}) \to - \int_{\Omega} (U_0(x), \varphi(x,0)) \ dx \quad \text{as} \quad \Delta x \to 0. \]  \quad (4.13)

Adding and subtracting \( \varphi_n = \varphi(t^n) \) and using the fact that \( \varphi \) is compactly supported in \([0,T]\),

\[ |B_{DG}^{11}(V^{\Delta x}, \varphi^{\Delta x}) + \sum_{n,K} \int_K \langle U_{0,-}^{\Delta x}, \varphi_0 \rangle \ dx| \]

\[ \leq \sum_{n,K} \int_K \|U_{n,-}^{\Delta x}, \varphi_{n,+}^{\Delta x} - \varphi_n\| \ dx + \sum_{n,K} \int_K \|U_{n+1,-}^{\Delta x}, \varphi_{n+1,-}^{\Delta x} - \varphi_{n+1}\| \ dx \]

\[ \leq C \|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \sum_{n,K} \left( \text{meas}(\partial(K \times I^n)) \right)^{\frac{1}{2}} \left( \|\varphi_{n,-}^{\Delta x} - \varphi_n\|_{L^2(K)} + \|\varphi_{n+1,+}^{\Delta x} - \varphi_{n+1}\|_{L^2(K)} \right) \]

\[ \leq C \Delta x \|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \|\varphi\|_{H^2(\Omega \times [0,T])} \quad \text{(by (4.5))}, \]

\[ \to 0 \quad \text{as} \quad \Delta x \to 0, \]
4.1. Convergence to measure-valued solution

Claim 6: The following bilinear form,
\[
\mathcal{B}_{DG}(\mathbf{V}^{\Delta x}, \varphi - \varphi^{\Delta x})
\]
\[
= - \sum_{n,K} \int_{I_n} \int_{K} \left( \langle \mathbf{U}(\mathbf{V}^{\Delta x}), (\varphi - \varphi^{\Delta x}) \rangle + \sum_{k=1}^{d} \langle \mathbf{F}^k(\mathbf{V}^{\Delta x}), (\varphi - \varphi^{\Delta x}) \rangle \right) dx dt
\] (4.14)

satisfies
\[
\mathcal{B}_{DG}(\mathbf{V}^{\Delta x}, \varphi - \varphi^{\Delta x}) \to 0 \quad \text{as} \quad \Delta x \to 0. \quad (4.15)
\]

We estimate,
\[
|\mathcal{B}_{DG}(\mathbf{V}^{\Delta x}, \varphi - \varphi^{\Delta x})|
\]
\[
\leq \sum_{n,K} \int_{I_n} \int_{K} \left| \langle \mathbf{U}(\mathbf{V}^{\Delta x}), (\varphi - \varphi^{\Delta x}) \rangle + \sum_{k=1}^{d} \langle \mathbf{F}^k(\mathbf{V}^{\Delta x}), (\varphi - \varphi^{\Delta x}) \rangle \right| dx dt
\]
\[
\leq C\|\mathbf{V}^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \|\varphi - \varphi^{\Delta x}\|_{H^1(\Omega \times [0,T])}
\]
\[
\leq C\Delta x \|\mathbf{V}^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \|\varphi\|_{H^2(\Omega \times (0,T))} \to 0 \quad \text{as} \quad \Delta x \to 0.
\]

Here, we have used the stronger approximation result (4.5).

To prove convergence, we observe that
\[
\int_0^T \int_{\Omega} \left( \langle \mathbf{U}(\mathbf{V}^{\Delta x}), \varphi_t \rangle + \sum_{k=1}^{d} \langle \mathbf{F}^k(\mathbf{V}^{\Delta x}), \varphi_{x_k} \rangle \right) dt
\]
\[
= \sum_{n,K} \int_{I_n} \int_{K} \left( \langle \mathbf{U}(\mathbf{V}^{\Delta x}), \varphi_t \rangle + \sum_{k=1}^{d} \langle \mathbf{F}^k(\mathbf{V}^{\Delta x}), \varphi_{x_k} \rangle \right) dt
\]
\[
= - \mathcal{B}_{DG}(\mathbf{V}^{\Delta x}, \varphi - \varphi^{\Delta x}) - \mathcal{B}(\mathbf{V}^{\Delta x}, \varphi^{\Delta x})
\]
\[
+ \mathcal{B}_{DG}^1(\mathbf{V}^{\Delta x}, \varphi^{\Delta x}) + \mathcal{B}_{DG}^2(\mathbf{V}^{\Delta x}, \varphi^{\Delta x}) + \mathcal{B}_{DG}^3(\mathbf{V}^{\Delta x}, \varphi^{\Delta x})
\]
\[
+ \mathcal{B}_{SD}(\mathbf{V}^{\Delta x}, \varphi^{\Delta x}) + \mathcal{B}_{SC}(\mathbf{V}^{\Delta x}, \varphi^{\Delta x})
\]
\[
\to - \int_{\Omega} \langle \mathbf{U}_0(x), \varphi(x,0) \rangle dx \quad \text{as} \quad \Delta x \to 0, \quad \text{(by} \quad \text{(3.2, 4.6, 4.7, 4.9, 4.11, 4.13, 4.15)}).
\]

Hence,
\[
\int_0^T \int_{\Omega} \left( \langle \mathbf{U}(\mathbf{V}^{\Delta x}), \varphi_t \rangle + \sum_{k=1}^{d} \langle \mathbf{F}^k(\mathbf{V}^{\Delta x}), \varphi_{x_k} \rangle \right) dt + \int_{\Omega} \langle \mathbf{U}_0(x), \varphi(x,0) \rangle \to 0 \quad \text{as} \quad \Delta x \to 0.
\]

On account of the uniform \(L^\infty\) bound, the sequence \(\mathbf{V}^{\Delta x}\) converges to a Young measure \(\mu\), in the sense that
\[
f(\mathbf{V}^{\Delta x}) \overset{*}{\rightharpoonup} (f, \mu) \quad \text{in} \quad L^\infty(\Omega \times [0,T]).
\]
4. Convergence analysis for systems of conservation laws

for all \( f \in C(\mathbb{R}^m) \) [9]. Thus,

\[
\lim_{\Delta x \to 0} \int_0^T \int_{\Omega} \left( \langle U(V^{\Delta x}), \varphi_t \rangle + \sum_{k=1}^d \langle F^k(V^{\Delta x}), \varphi_{x_k} \rangle \right) dx dt
\]

\[
= \int_0^T \int_{\Omega} \left( \langle U, \mu_{x,t} \rangle, \varphi_t \rangle + \sum_{k=1}^d \langle (F^k, \mu_{x,t}), \varphi_{x_k} \rangle \right) dx dt.
\]

Hence, combining the above two limits, we see that the Young measure \( \mu \) is a measure-valued solution of (1.3).

4.2. Entropy consistency

The next step in the analysis is to show consistency with the entropy condition (1.15). We do so in the following theorem.

**Theorem 4.2.1.** Let \( V^{\Delta x} \) be the approximate solutions generated by the scheme (3.2). We assume that the diffusion matrices \( D \) are uniformly positive definite and that the approximate solutions are uniformly bounded (4.1). Then, the limit measure-valued solution \( \mu \) satisfies the entropy condition (1.15).

**Proof.** We consider a smooth test function \( \varphi \geq 0 \) and compactly supported in \( \Omega \times [0,T) \) and show consistency with the entropy condition (1.15) in the following series of claims.

**Claim 1:** Let the DG quasilinear form be defined as in (3.3), then

\[
\lim_{\Delta x \to 0} \mathcal{B}(V^{\Delta x}, V^{\Delta x}_t \varphi)
\]

\[
\geq - \int_0^T \int_{\Omega} \left( \langle S, \mu_{x,t} \rangle \varphi_t + \sum_{k=1}^d \langle Q^k, \mu_{x,t} \rangle \varphi_{x_k} \right) dx dt - \int_{\Omega} S(U_0(x)) \varphi(x,0) dx.
\]

(4.16)

To prove the claim we define \( \varphi^n = \varphi(t^n) \) and start with the following computation,

\[
\mathcal{B}^t_{DG}(V^{\Delta x}, V^{\Delta x}_t \varphi)
\]

\[
= - \sum_{n,K} \int_{I^n} \int_K \langle U(V^{\Delta x}), (V^{\Delta x}_t \varphi)_t \rangle dx dt
\]

\[
+ \sum_{n,K} \int_K \langle U(V_{n+1,-}^{\Delta x}), V_{n+1,-}^{\Delta x} \varphi^n + 1 \rangle dx - \sum_{n,K} \int_K \langle U(V_{n,-}^{\Delta x}), V_{n,+}^{\Delta x} \varphi^n \rangle dx
\]

\[
= \sum_{n,K} \int_{I^n} \int_K \langle U_t(V^{\Delta x}), V^{\Delta x}_t \varphi \rangle dx dt \quad \text{(integrating by parts)}
\]

\[
- \sum_{n,K} \int_K \langle U(V_{n+1,-}^{\Delta x}), V_{n+1,-}^{\Delta x} \varphi^n + 1 \rangle dx + \sum_{n,K} \int_K \langle U(V_{n,+}^{\Delta x}), V_{n,+}^{\Delta x} \varphi^n \rangle dx
\]

38
4.2. Entropy consistency

\[ + \sum_{n,K} \int_K \langle U(V_{n+1,-}^\Delta x), V_{n+1,-}^\Delta x, \varphi^{n+1} \rangle \, dx - \sum_{n,K} \int_K \langle U(V_{n,-}^\Delta x), V_{n,+}^\Delta x, \varphi^n \rangle \, dx \]

\[ = \sum_{n,K} \int_{I^n} \int_K S(U(V_n^\Delta x)), \varphi \, dx \, dt \quad \text{(definition of entropy function)} \]

\[ - \sum_{n,K} \int_K \langle (U(V_{n,-}^\Delta x) - U(V_{n,+}^\Delta x)), V_{n,+}^\Delta x, \varphi^n \rangle \, dx \]

\[ = - \sum_{n,K} \int_{I^n} \int_K S(V_n^\Delta x) \varphi \, dx \, dt \quad \text{(integrating by parts)} \]

\[ + \sum_{n,K} \int_K S(V_{n+1,-}^\Delta x) \varphi^{n+1} \, dx - \sum_{n,K} \int_K S(V_{n,+}^\Delta x) \varphi^n \, dx \]

\[ - \sum_{n,K} \int_K \langle (U(V_{n,-}^\Delta x) - U(V_{n,+}^\Delta x)), V_{n,+}^\Delta x, \varphi^n \rangle \, dx \]

\[ = - \sum_{n,K} \int_{I^n} \int_K S(V_n^\Delta x) \varphi \, dx \, dt - \sum_K \int_K S(U(V_{0,-}^\Delta x)) \varphi^0 \, dx \]

\[ + \sum_{n,K} \int_K (S(U(V_{n,-}^\Delta x)) - S(U(V_{n,+}^\Delta x)) - \langle (U(V_{n,-}^\Delta x) - U(V_{n,+}^\Delta x)), V_{n,+}^\Delta x, \varphi^n \rangle) \, dx \]

\[ = - \sum_{n,K} \int_{I^n} \int_K S(V_n^\Delta x) \varphi \, dx \, dt - \sum_K \int_K S(U(V_{0,-}^\Delta x)) \varphi^0 \, dx \]

\[ + \sum_{n,K} \int_K \left( \int_0^1 \theta \langle (V_{n,-}^\Delta x - V_{n,+}^\Delta x), U(V_n(\theta)(V_{n,-}^\Delta x - V_{n,+}^\Delta x)) \rangle \right) \varphi^n \, dx \quad \text{(by (3.35))} \]

\[ \geq - \int_0^T \int_\Omega S(V_n^\Delta x) \varphi \, dx \, dt - \int_\Omega S(U(V_{0,-}^\Delta x)) \varphi^0 \, dx. \quad (4.17) \]

The last step follows from the positivity of \( U \) and positivity of the test function \( \varphi \).

As

\[ \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d \langle F^k(V_n^\Delta x), (V_n^\Delta x, \varphi)_x \rangle \, dx \, dt \]

\[ = - \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d \langle (F^k(V_n^\Delta x))_x, V_n^\Delta x \varphi \rangle \, dx \, dt \]

\[ + \sum_{n,K,K'} \int_{I^n} \int_{\partial K K'} \left( \sum_{k=1}^d \langle F^k(V_{K,-}^\Delta x), V_{K,-}^\Delta x \nu_{KK'}^k \rangle \right) \varphi \, d\sigma \, dt \]
4. Convergence analysis for systems of conservation laws

\[
\begin{align*}
&= - \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d Q^k (V^{\Delta x})_{x_k} \varphi dx dt \quad \text{(definition of entropy flux)} \\
&\quad + \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \left( \sum_{k=1}^d \langle F^k (V^{\Delta x}_{K,-}), V^{\Delta x}_{K,-} \rangle \varphi v^k_{KK'} \right) d\sigma dt \\
&= \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d Q^k (V^{\Delta x}) \varphi_{x_k} dx dt \\
&\quad + \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \sum_{k=1}^d \left( \langle F^k (V^{\Delta x}_{K,-}), V^{\Delta x}_{K,-} \rangle - Q^k (V^{\Delta x}_{K,-}) \right) \varphi v^k_{KK'} d\sigma dt \\
&= \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d Q^k (V^{\Delta x}) \varphi_{x_k} dx dt + \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \sum_{k=1}^d \Psi^k (V^{\Delta x}_{K,-}) \varphi v^k_{KK'} d\sigma dt \quad (4.18)
\end{align*}
\]

and repeating the calculation in (3.31) and (3.32), we obtain,

\[
\begin{align*}
B^s_{DG} (V^{\Delta x}, V^{\Delta x}) \\
&= - \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d \langle F^k (V^{\Delta x}), (V^{\Delta x} \varphi)_{x_k} \rangle dx dt \\
&\quad + \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \left( \sum_{k=1}^d \langle F^{k,*} (V^{\Delta x}_{K,-}, V^{\Delta x}_{K,+}), V^{\Delta x}_{K,-} \rangle \varphi v^k_{KK'} \right) d\sigma dt \\
&\quad - \frac{1}{2} \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \langle V^{\Delta x}_{K,-}, D (V^{\Delta x}_{K,+} - V^{\Delta x}_{K,-}) \rangle d\sigma dt \\
&= - \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d Q^k (V^{\Delta x}) \varphi_{x_k} dx dt \\
&\quad + \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \left( \sum_{k=1}^d \left( \langle F^{k,*} (V^{\Delta x}_{K,-}, V^{\Delta x}_{K,+}), V^{\Delta x}_{K,-} \rangle - \Psi^k (V^{\Delta x}_{K,-}) \right) \varphi v^k_{KK'} \right) d\sigma dt \\
&\quad - \frac{1}{2} \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \langle V^{\Delta x}_{K,-}, D (V^{\Delta x}_{K,+} - V^{\Delta x}_{K,-}) \rangle d\sigma dt \\
&\geq - \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d Q^k (V^{\Delta x}) \varphi_{x_k} dx dt. \quad (4.19)
\end{align*}
\]

Using the definition of the residual (3.18) and the streamline diffusion operator, we have
4.2. Entropy consistency

\[ B_{SD}(V^{\Delta x}, V^{\Delta x} \varphi) \]

\[ = \sum_{n,K} \int_{I^n} \int_{K} \left\langle \left( U(V^{\Delta x})(V^{\Delta x} \varphi)_t + \sum_{k=1}^{d} F^k_{V}(V^{\Delta x})(V^{\Delta x} \varphi)_x \right), D^{SD} \text{Res} \right\rangle dx dt \]

\[ = \sum_{n,K} \int_{I^n} \int_{K} \left\langle \text{Res}, D^{SD} \text{Res} \right\rangle \varphi dx dt \]

\[ + \sum_{n,K} \int_{I^n} \int_{K} \left\langle \left( U(V^{\Delta x}) V^{\Delta x} \varphi_t + \sum_{k=1}^{d} F^k_{V}(V^{\Delta x}) V^{\Delta x} \varphi_{x_k} \right), D^{SD} \text{Res} \right\rangle dx dt \geq T_1. \]

(4.20)

However repeating the calculations in the proof of (4.6), we see that

\[ |T_1| \leq C \| V^{\Delta x} \|_{L^\infty(\Omega \times [0,T])} \Delta x^{1/2} \left( \sum_{n,K} \| \varphi^{\Delta x} \|_{H^1(I^n)}^2 \right)^{1/2} \left( \Delta x \sum_{n,K} \int_{I^n} \| \text{Res} \|^2 dx dt \right)^{1/2} \]

\[ \leq C \| V^{\Delta x} \|_{L^\infty(\Omega \times [0,T])} \Delta x^{3/2} \| \varphi \|_{H^1(\Omega \times [0,T])} \to 0 \quad \text{as} \ \Delta x \to 0. \]

(4.21)

Next, from the definition of the shock-capturing operator and positivity of \( U_V \), we obtain

\[ B_{SC}(V^{\Delta x}, V^{\Delta x} \varphi) \]

\[ = \sum_{n,K} \int_{I^n} \int_{K} D_{n,K}^{SC} \left( \left\langle \left( V^{\Delta x} \varphi \right)_t, \widetilde{U}_V V^{\Delta x} \right\rangle \right. \]

\[ + \sum_{k=1}^{d} \frac{\Delta x K^2}{(\Delta t^h)^2} \left\langle \left( V^{\Delta x} \varphi \right)_{x_k}, \widetilde{U}_V V^{\Delta x}_{x_k} \right\rangle \right) dx dt \]

\[ = \sum_{n,K} \int_{I^n} \int_{K} D_{n,K}^{SC} \left( \left\langle V^{\Delta x}_t, \widetilde{U}_V V^{\Delta x}_t \right\rangle \right. \]

\[ + \sum_{k=1}^{d} \frac{\Delta x K^2}{(\Delta t^h)^2} \left\langle V^{\Delta x}_{x_k}, \widetilde{U}_V V^{\Delta x}_{x_k} \right\rangle \right) \varphi dx dt \]

\[ + \sum_{n,K} \int_{I^n} \int_{K} D_{n,K}^{SC} \left( \left\langle V^{\Delta x} \varphi_t, \widetilde{U}_V V^{\Delta x} \right\rangle \right. \]

\[ + \sum_{k=1}^{d} \frac{\Delta x K^2}{(\Delta t^h)^2} \left\langle V^{\Delta x} \varphi_{x_k}, \widetilde{U}_V V^{\Delta x}_{x_k} \right\rangle \right) dx dt \]

\[ \geq T_2. \]

(4.22)

Repeating the calculation in the proof of (4.7), we obtain

\[ |T_2| \leq C(\Delta x)^{3/2} \| V^{\Delta x} \|_{L^\infty(\Omega \times [0,T])} \| \varphi \|_{H^1(\Omega \times [0,T])} \]

\[ \left( \Delta x \sum_{n,K} \int_{I^n} \int_{K} \| \text{Res} \|^2 dx dt \right)^{1/2} \left( \sum_{n,K} \| B\text{Res}_{n,K} \|^2 \right)^{1/2} \]

\[ \leq C(\Delta x)^{3/2} \| \varphi \|_{H^1(\Omega \times [0,T])} \to 0 \quad \text{as} \ \Delta x \to 0. \]

(4.23)
4. Convergence analysis for systems of conservation laws

From (4.17), (4.19), (4.20) and (4.22), we have

\[ B(V^\Delta x, V^\Delta x \phi) = B_{DG}(V^\Delta x, V^\Delta x \phi) + B_{SD}(V^\Delta x, V^\Delta x \phi) + B_{SC}(V^\Delta x, V^\Delta x \phi) \geq - \left( \int_0^T \int_\Omega (S(V^\Delta x) \phi_t + \sum_{k=1}^d Q^k(V^\Delta x) \phi_{x_k} + \int_\Omega S(U(V^\Delta x_0)) \phi_0^0 dx) \right) + T_1 + T_2. \]

Passing to the limit in the above expression as \( \Delta x \to 0 \), using the commutation of nonlinearities and limit in the sense of measures and using (4.21) and (4.23), results in,

\[ \lim_{\Delta x \to 0} B(V^\Delta x, V^\Delta x \phi) \geq - \int_0^T \int_\Omega \left( \langle S, \mu_{x,t} \rangle \phi_t + \sum_{k=1}^d \langle Q^k, \mu_{x,t} \rangle \phi_{x_k} \right) dxdt - \int_\Omega S(U_0(x)) \phi(x,0) dx, \]

thus proving the claim.

Next, we use the \( H^1 \) projection operator \( \Pi^\Delta x \) from (4.2) and show the following claim,

Claim 2: We have

\[ \lim_{\Delta x \to 0} B(V^\Delta x, V^\Delta x \phi - \Pi^\Delta x(V^\Delta x \phi)) = 0. \]  

(4.24)

We will use the following superapproximation properties of the projection operator (see Section A.1)

\[ \| \nabla x t(V^\Delta x \phi - \Pi^\Delta x(V^\Delta x \phi)) \|_{L^2(K \times I^n)} \leq C \Delta x |V^\Delta x|_{H^1(K \times I^n)} \]  

(4.25)

\[ \| V^\Delta x \phi - \Pi^\Delta x(V^\Delta x \phi) \|_{L^2(K \times I^n)} \leq C \Delta x \| V^\Delta x \phi \|_{L^2(K \times I^n)} \]  

(4.26)

\[ \| V^\Delta x \phi - \Pi^\Delta x(V^\Delta x \phi) \|_{L^2(\partial(K \times I^n))} \leq C \Delta x^{1/2} \| V^\Delta x \phi \|_{L^2(K \times I^n)} \]  

(4.27)

where the constant \( C \) depends on \( \phi \), but not on \( K \times I^n \). We denote

\[ \Pi^\Delta x_{\text{err}}(V^\Delta x \phi) = V^\Delta x \phi - \Pi^\Delta x(V^\Delta x \phi) \]

for the projection error for brevity.

We start with calculations involving the DG quasilinear form (3.3). Let,

\[ B_{DG}^{\text{int}}(V^\Delta x, \Pi^\Delta x_{\text{err}}(V^\Delta x \phi)) \]

\[ = - \sum_{n,K} \int_{I^n} \int_K \left( \langle U(V^\Delta x), (\Pi^\Delta x_{\text{err}}(V^\Delta x \phi))_t \rangle + \sum_{k=1}^d \langle F^k(V^\Delta x), (\Pi^\Delta x_{\text{err}}(V^\Delta x \phi))_{x_k} \rangle \right) dxdt \]
\[\sum_{n,K} \int_{I^n} \int_K \left( \langle U(V^{x})_t, \Pi^{x}_{\text{err}}(V^{x} \varphi) \rangle + \sum_{k=1}^{d} \langle F^k(V^{x})_{x_k}, \Pi^{x}_{\text{err}}(V^{x} \varphi) \rangle \right) dx dt \]

- \[\sum_{n,K} \int_{K} \langle U(V^{x}_{n+1,-}), (\Pi^{x}_{\text{err}}(V^{x} \varphi))^{n+1-} \rangle dx + \sum_{n,K} \int_{K} \langle U(V^{x}_{n+}), (\Pi^{x}_{\text{err}}(V^{x} \varphi))^{n,+} \rangle dx \]

- \[\sum_{n,K,K'} \int_{I^n} \int_{\partial K K'} \left( \sum_{k=1}^{d} \langle F^k(V^{x}(K,-), (\Pi^{x}_{\text{err}}(V^{x} \varphi))_{K,-} \rangle \nu_{K K'}^k \right) d\sigma dt. \]  

Using the definition of the residual (3.18), the weak BV estimate (3.23), and the projection error estimate (4.26), we obtain,

\[|T_3| = \left| \sum_{n,K} \int_{I^n} \int_K \langle \text{Res}, \Pi^{x}_{\text{err}}(V^{x} \varphi) \rangle dx dt \right| \]

\[\leq C \sum_{n,K} \text{Res}_{n,K} \| \Pi^{x}_{\text{err}}(V^{x} \varphi) \|_{L^2(K \times I^n)} \]

\[\leq C \Delta x \sum_{n,K} \text{Res}_{n,K} \| V^{x} \|_{L^2(K \times I^n)} \]  

\[\leq C \Delta x^{\frac{1}{2}} \| V^{x} \|_{L^\infty(\Omega \times [0,T])} \left( \Delta x \sum_{n,K} \int_{I^n} \int_K \| \text{Res} \|^2 dx dt \right)^{\frac{1}{2}} \]

\[\to 0, \quad \text{as} \quad \Delta x \to 0. \]

Using (4.28) and calculating with the full DG quasilinear form, we obtain

\[B_{DG}(V^{x}, (\Pi^{x}_{\text{err}}(V^{x} \varphi))) = T_3 \]

\[- \sum_{n,K} \int_{K} \langle U(V^{x}_{n+1,-}), (\Pi^{x}_{\text{err}}(V^{x} \varphi))^{n+1-} \rangle dx + \sum_{n,K} \int_{K} \langle U(V^{x}_{n+}), (\Pi^{x}_{\text{err}}(V^{x} \varphi))^{n,+} \rangle dx \]

\[- \sum_{n,K,K'} \int_{I^n} \int_{\partial K K'} \left( \sum_{k=1}^{d} \langle F^k(V^{x}(K,-), (\Pi^{x}_{\text{err}}(V^{x} \varphi))_{K,-} \rangle \nu_{K K'}^k \right) d\sigma dt \]

\[+ \sum_{n,K} \int_{K} \langle U(V^{x}_{n+1,-}), (\Pi^{x}_{\text{err}}(V^{x} \varphi))^{n+1-} \rangle dx - \sum_{n,K} \int_{K} \langle U(V^{x}_{n-}), (\Pi^{x}_{\text{err}}(V^{x} \varphi))^{n,+} \rangle dx \]

\[+ \sum_{n,K,K'} \int_{I^n} \int_{\partial K K'} \left( \sum_{k=1}^{d} \langle F^k(V^{x}(K,-), (\Pi^{x}_{\text{err}}(V^{x} \varphi))_{K,-} \rangle \nu_{K K'}^k \right) d\sigma dt \]

\[- \frac{1}{2} \sum_{n,K,K'} \int_{I^n} \int_{\partial K K'} \langle (\Pi^{x}_{\text{err}}(V^{x} \varphi))_{K,-}, D(V^{x}_{K,+} - V^{x}_{K,-}) \rangle d\sigma dt \]

43
= T_3 + \sum_{n,K} \int_K \left( U(V_{n,+}^{\Delta x}) - U(V_{n,-}^{\Delta x}), (\Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi}))_{n,+} \right) dx

+ \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} d \sum_{k=1}^d \left( \mathbb{F}^{k,*}(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}) - \mathbb{F}^{k}(V_{K,-}^{\Delta x}, (\Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi}))_{K,-} \right) \nu_{KK'}^k dx dt

- \frac{1}{2} \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \left( (\Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi}))_{K,-}, \mathbb{D}(V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}) \right)\nu_k d\sigma dt.

(4.30)

Using the definition of the boundary residual in (3.20d), the weak BV estimate (3.23), the estimate (4.29) and the projection error estimate (4.27), the above can be simplified as

\[ | \mathcal{B}_{DG}(V^{\Delta x}, \Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi})) | \leq | T_3 | + \sum_{n,K} \mathbb{B}_{\text{Res},n,K} \| \Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi}) \|_{L^2(\partial(K \times I^n))} \]

\[ \leq | T_3 | + C \Delta x^{\frac{3}{2}} \sum_{n,K} \mathbb{B}_{\text{Res},n,K} \| V^{\Delta x} \|_{L^2(K \times I^n)} \]

\[ \leq | T_3 | + C \Delta x^{\frac{3}{2}} \| V^{\Delta x} \|_{L^\infty(\Omega \times [0,T])} \left( \sum_{n,K} | \mathbb{B}_{\text{Res},n,K} |^2 \right)^{\frac{1}{2}} \]

\[ \to 0, \quad \text{as } \Delta x \to 0. \]

(4.31)

Next, we consider the streamline diffusion quasilinear form (3.17) and use the stability of the projection operator (4.25) and the weak BV estimate (3.23) to obtain,

\[ | \mathcal{B}_{SD}(V^{\Delta x}, \Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi})) | \]

\[ = \left| \sum_{n,K} \int_{I^n} \int_K \left( U(V^{\Delta x})(\Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi}))_t + \sum_{k=1}^d \mathbb{F}^{k}(V^{\Delta x})(\Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi}))_{x_k} \right), \mathbb{D}^{\text{SD} \text{Res},n,K} \right| dx dt \]

\[ \leq C \Delta x \| V^{\Delta x} \|_{L^\infty(\Omega \times [0,T])} \sum_{n,K} \| \nabla_{xt}(\Pi_{\text{err}}^{\Delta x}(V^{\Delta x}_{\varphi})) \|_{L^2(K \times I^n)} \mathbb{R}_{\text{Res},n,K} \]

\[ \leq C \Delta x^2 \| V^{\Delta x} \|_{L^\infty(\Omega \times [0,T])} \sum_{n,K} \| \nabla_{xt} V^{\Delta x} \|_{L^2(K \times I^n)} \mathbb{R}_{\text{Res},n,K} \]

\[ \leq C \Delta x \left( \Delta x \sum_{n,K} \mathbb{R}_{\text{Res},n,K} \left( \int_{I^n} \int_K \| \nabla_{xt} V^{\Delta x} \|^2 dx dt \right)^{\frac{1}{2}} \right) \]

\[ \to 0, \quad \text{as } \Delta x \to 0. \]

(4.32)

Given the orthogonality of the projection operator \( \Pi^{\Delta x} \) (4.3) and the definition of the shock-
4.2. Entropy consistency

capturing operator, it is easy to see that,

\[ B_{\text{SC}}(V^{\Delta x}, \Pi^{\Delta x}(V^{\Delta x} \varphi)) \equiv 0. \]  \(4.33\)

The claim (4.24) is proved by combining (4.31), (4.32), and (4.33).

Using (4.16), (4.24), and the definition of the scheme (3.2), we obtain,

\[
\int_0^T \int_\Omega \left( \langle S, \mu_{x,t} \rangle \varphi_t + \sum_{k=1}^d \langle Q^k, \mu_{x,t} \rangle \varphi_{x_k} \right) \, dx \, dt + \int_\Omega S(U_0(x))\varphi(x,0) \, dx \\
\geq - \lim_{\Delta x \to 0} B(V^{\Delta x}, V^{\Delta x} \varphi) \\
\geq - \lim_{\Delta x \to 0} B(V^{\Delta x}, (V^{\Delta x} \varphi - \Pi^{\Delta x}(V^{\Delta x} \varphi))) \\
- \lim_{\Delta x \to 0} B(V^{\Delta x}, \Pi^{\Delta x}(V^{\Delta x} \varphi)) = 0
\]

\[ \geq 0. \]  \(4.34\)

This proves (1.15).

**Remark 4.2.2.** The main assumption in the convergence to entropy measure-valued solutions is the uniform \(L^\infty\) bound (4.1). The entropy solution of scalar conservation laws satisfies an \(L^\infty\) bound. Therefore, consider the following entropy:

\[ S(U) = - \log(b - U) - \log(U - a). \]  \(4.35\)

provided that the initial data \(U_0(x)\) for the scalar conservation law \(U = U\) in (1.3) satisfies the bound,

\[ a < U_0(x) < b, \quad \forall x \in \Omega, \]

for constants \(a, b \in \mathbb{R}\). See [14] for a proof of how control on the above entropy implies an \(L^\infty\) bound on the approximate solutions. It bounds only \(U^{\Delta x}\) (and not \(V^{\Delta x}\)) though and the singularity of \(S\) makes the resulting scheme difficult to analyse.

However, for systems of conservation laws, it is still an open problem to prove \(L^\infty\) bounds, even at the level of the continuous problem. This bound does exist for some special systems [9] which possess invariant regions. The uniform \(L^\infty\) bound assumption is reasonable as we are interested in characterizing the consistency of the scheme (3.2).

**Remark 4.2.3.** The boundary residual in the shock-capturing term is actually not needed for the stability and convergence proof, so \(C_{\text{SC}} = 0\) is possible. One loses control over the gradient term weighted with the boundary residual in (3.23), but this is not needed in the convergence proof to bound the other terms of the formulation.
4. Convergence analysis for systems of conservation laws

4.3. Convergence for scalar conservation laws

In the following section, we consider the scalar case, i.e. (1.3) with $m = 1$. The work in this section will be published in [22].

4.3.1. Mathematical Framework

We shall use the following theorem to show strong convergence of approximate solutions. It is an adaptation of a result by Panov [42] using so-called H-measures. To this end, we recall that $(\eta, q)$ is a convex entropy/entropy flux pair for (1.3) if $\eta$ is a convex function of $U$ and

$$q_k^k(U) = \eta'(U) F_k^k(U), \quad k = 1, \ldots, d.$$ 

Additionally, we remind that the flux $F$ is genuinely non-linear if for any direction $\nu \in \mathbb{R}^d$ with $|\nu| = 1$, the map $U \mapsto \sum_{k=1}^d F_k^k(U) \cdot \nu_k$ is not affine on any non-trivial interval. We will use the assumption of genuine non-linearity throughout this section.

**Theorem 4.3.1.** Let $F$ be genuinely non-linear and $\{U_\nu\}_{\nu > 0}$ be a family of functions defined in $\Omega \times (0, T)$. If $\{U_\nu\}$ is bounded in $L^\infty(\Omega \times (0, T))$, and

$$\left\{ \partial_t \eta(U_\nu) + \text{div} (q(U_\nu)) \right\}_{\nu > 0}$$

lies in a compact set of $H^{-1}_{\text{loc}}(\Omega \times (0, T))$ for all convex entropy/entropy flux pairs $(\eta, q)$, then there exists a sequence $\{\nu_n\}_{n \in \mathbb{N}}, \nu_n \to 0$ and a function $U \in L^\infty(\Omega \times (0, T))$ such that

$$U_{\nu_n} \to U, \text{ a.e. } (x, t).$$

The following compactness interpolation result (known as Murat’s lemma [38]) is useful in obtaining the $H^{-1}_{\text{loc}}$ compactness needed in Theorem 4.3.1.

**Lemma 4.3.2.** Let $\Omega$ be a bounded open subset of $\mathbb{R}^d$, $d \geq 2$. Suppose that the sequence $\{\mathcal{L}_\nu\}_{\nu > 0}$ of distributions is bounded in $W^{-1,\infty}(\Omega)$. Suppose also that

$$\mathcal{L}_\nu = \mathcal{L}_{1,\nu} + \mathcal{L}_{2,\nu},$$

where $\{\mathcal{L}_{1,\nu}\}_{\nu > 0}$ is in a compact subset of $H^{-1}_{\text{loc}}(\Omega)$ and $\{\mathcal{L}_{2,\nu}\}_{\nu > 0}$ is in a bounded subset of $\mathcal{M}_{\text{loc}}(\Omega)$. Then $\{\mathcal{L}_\nu\}_{\nu > 0}$ is in a compact subset of $H^{-1}_{\text{loc}}(\Omega)$.

4.3.2. Strong convergence of approximate solutions

We start with rewriting the scheme using the following auxiliary theorem.

**Theorem 4.3.3.** Let $U^{\Delta x} = U(V^{\Delta x})$ be the approximate solutions generated by the scheme (3.2) and $(\eta, q)$ be any entropy/entropy flux pair. Consider a smooth test function $\varphi$ compactly supported in $\Omega \times (0, T)$. Then

$$B \left( V^{\Delta x}, \eta'(U^{\Delta x}) \varphi \right) = (B_{DG}^t + B_{DG}^s + B_{SD} + B_{SC}) \left( V^{\Delta x}, \eta'(U^{\Delta x}) \varphi \right),$$

(4.36)
where

\[
\begin{align*}
B_{\text{DG}}^t (V^{\Delta x}, \eta' (U^{\Delta x}) \varphi) &= - \sum_{n,K} \int_{I^n} \int_K \eta(U^{\Delta x}) \varphi_t \, dx \, dt + \frac{1}{2} \sum_{n,K} \int_K \eta''(\theta) (U_{n,-}^{\Delta x} - U_{n,+}^{\Delta x})^2 \varphi^n \, dx \\
&= : \langle L^{\text{int},t}, \varphi \rangle + \langle L^t, \varphi \rangle, \tag{4.37} \\
B_{\text{DG}}^s (V^{\Delta x}, \eta' (U^{\Delta x}) \varphi) &= - \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d q^k (V^{\Delta x}) \varphi_{x_k} \, dx \, dt \\
&+ \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \sum_{k=1}^d (F^k(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}) - D^k(V_{K,-}^{\Delta x})) \eta'(U^{\Delta x}) v_{K,K'}^k \varphi \, d\sigma \, dt \\
&+ \frac{1}{2} \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \sum_{k=1}^d q^k (V_{K,-}^{\Delta x}) v_{K,K'}^k \varphi \, d\sigma \, dt \\
&- \frac{1}{2} \sum_{n,K,K'} \int_{I^n} \int_{\partial_{KK'}} \eta'(U^{\Delta x}) \varphi D(V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}) \, d\sigma \, dt \\
&= : \langle L^{\text{int},s}, \varphi \rangle + \langle L^{s,1}, \varphi \rangle + \langle L^{s,2}, \varphi \rangle + \langle L^{s,3}, \varphi \rangle, \tag{4.38} \\
B_{\text{SD}} (V^{\Delta x}, \eta' (U^{\Delta x}) \varphi) &= + \sum_{n,K} \int_{I^n} \int_K \eta''(U^{\Delta x}) U_V(V^{\Delta x}) \text{Res} (V^{\Delta x}) \varphi \, dx \, dt \\
&+ \sum_{n,K} \int_{I^n} \int_K \left( U_V(V^{\Delta x}) \eta'(U^{\Delta x}) \varphi_t + \sum_{k=1}^d F^k(V^{\Delta x}) \eta'(U^{\Delta x}) \varphi_{x_k} \right) D^{\text{SD}} \text{Res} \, dx \, dt \\
&= : \langle L^{\text{SD},1}, \varphi \rangle + \langle L^{\text{SD},2}, \varphi \rangle, \tag{4.39} \\
B_{\text{SC}} (V^{\Delta x}, \eta' (U^{\Delta x}) \varphi) &= + \sum_{n,K} \int_{I^n} \int_K D^{\text{SC}}_{n,K} \eta''(U^{\Delta x}) U_V(V^{\Delta x}) \left( V^{\Delta x}_l \tilde{V}^{\Delta x}_l + \sum_{k=1}^d \frac{\Delta x^2}{(\Delta t^n)^2} V^{\Delta x}_{x_k} \tilde{U} V^{\Delta x}_{x_k} \right) \varphi \, dx \, dt \\
&+ \sum_{n,K} \int_{I^n} \int_K D^{\text{SC}}_{n,K} \left( \eta'(U^{\Delta x}) \varphi_t \tilde{V}^{\Delta x}_l + \sum_{k=1}^d \frac{\Delta x^2}{(\Delta t^n)^2} \eta'(U^{\Delta x}) \varphi_{x_k} \tilde{U} V^{\Delta x}_{x_k} \right) \, dx \, dt \\
&= : \langle L^{\text{SC},1}, \varphi \rangle + \langle L^{\text{SC},2}, \varphi \rangle. \tag{4.40}
\end{align*}
\]
4. Convergence analysis for systems of conservation laws

Proof. To prove the theorem, we define \( \varphi^n = \varphi(t^n) \) and start with the temporal part,

\[
B_{DG}^t(V^{\Delta x}, \eta'(U^{\Delta x})\varphi) = - \sum_{n,K} \int_{I^n} \int_{K} U(V^{\Delta x})(\eta'(U^{\Delta x})\varphi)_t \, dx \, dt
\]

\[
+ \sum_{n,K} \int_{K} U(V_{n+1,-}^{\Delta x})\eta'(U_{n+1,-}^{\Delta x})\varphi^{n+1} \, dx - \sum_{n,K} \int_{K} U(V_{n,-}^{\Delta x})\eta'(U_{n,+}^{\Delta x})\varphi^n \, dx
\]

\[
= \sum_{n,K} \int_{I^n} \int_{K} U_t(V^{\Delta x})\eta'(U^{\Delta x})\varphi \, dx \, dt \quad \text{(integrating by parts)}
\]

\[
- \sum_{n,K} \int_{K} U(V_{n+1,-}^{\Delta x})\eta'(U_{n+1,-}^{\Delta x})\varphi^{n+1} \, dx + \sum_{n,K} \int_{K} U(V_{n,+}^{\Delta x})\eta'(U_{n,+}^{\Delta x})\varphi^n \, dx
\]

\[
+ \sum_{n,K} \int_{K} U(V_{n+1,-}^{\Delta x})\eta'(U_{n+1,-}^{\Delta x})\varphi^{n+1} \, dx - \sum_{n,K} \int_{K} U(V_{n,-}^{\Delta x})\eta'(U_{n,+}^{\Delta x})\varphi^n \, dx
\]

\[
= \sum_{n,K} \int_{I^n} \int_{K} \eta(U(V^{\Delta x}))_t \varphi \, dx \, dt \quad \text{(definition of entropy function)}
\]

\[
- \sum_{n,K} \int_{K} (U(V_{n,-}^{\Delta x}) - U(V_{n,+}^{\Delta x})) \eta'(U_{n,+}^{\Delta x})\varphi^n \, dx
\]

\[
= - \sum_{n,K} \int_{I^n} \int_{K} \eta(U^{\Delta x})\varphi_t \, dx \, dt \quad \text{(integrating by parts)}
\]

\[
+ \sum_{n,K} \int_{K} \eta(V_{n+1,-}^{\Delta x})\varphi^{n+1} \, dx - \sum_{n,K} \int_{K} \eta(V_{n,+}^{\Delta x})\varphi^n \, dx
\]

\[
- \sum_{n,K} \int_{K} (U(V_{n,-}^{\Delta x}) - U(V_{n,+}^{\Delta x})) \eta'(U_{n,+}^{\Delta x})\varphi^n \, dx
\]

\[
= - \sum_{n,K} \int_{I^n} \int_{K} \eta(U^{\Delta x})\varphi_t \, dx \, dt
\]

\[
+ \sum_{n,K} \int_{K} (\eta(U(V_{n,-}^{\Delta x})) - \eta(U(V_{n,+}^{\Delta x})) - (U(V_{n,-}^{\Delta x}) - U(V_{n,+}^{\Delta x})) \eta'(U_{n,+}^{\Delta x})) \varphi^n \, dx
\]

\[
= - \sum_{n,K} \int_{I^n} \int_{K} \eta(U^{\Delta x})\varphi_t \, dx \, dt + \frac{1}{2} \sum_{n,K} \int_{K} \eta''(\theta)(U_{n,-}^{\Delta x} - U_{n,+}^{\Delta x})^2 \varphi^n \, dx. \quad (4.41)
\]

For the spatial part, we observe first

\[
\sum_{n,K} \int_{I^n} \int_{K} \sum_{k=1}^d F^k(V^{\Delta x})(\eta'(U^{\Delta x})\varphi)_{x_k} \, dx \, dt
\]
4.3. Convergence for scalar conservation laws

\[ \begin{align*}
\sum_{n,K} & \int_{I^n} \int_{K} \sum_{k=1}^{d} \left( F^k(V_{\Delta x}^n) \right)_{x_k} \eta'(U_{\Delta x}^n) \varphi \, dx \, dt \quad \text{(integration by parts)} \\
\sum_{n,K,K'} & \int_{I^n} \int_{\partial_{K,K'}} \left( \sum_{k=1}^{d} F^k(V_{\Delta x}^n) \eta'(U_{\Delta x}^n) \nu_{KK'}^k \right) \varphi \, d\sigma \, dt \\
\sum_{n,K} & \int_{I^n} \int_{K} \sum_{k=1}^{d} q^k(V_{\Delta x}^n) \varphi_x \, dx \, dt \quad \text{(definition of entropy flux)} \\
\sum_{n,K,K'} & \int_{I^n} \int_{\partial_{K,K'}} \sum_{k=1}^{d} \left( F^k(V_{\Delta x}^n) \eta'(U_{\Delta x}^n) - q^k(V_{\Delta x}^n) \right) \varphi \nu_{KK'}^k \, d\sigma \, dt.
\end{align*} \] (4.42)

Repeating the calculation in (3.31) and (3.32), we obtain,

\[ \mathcal{B}_{DG}^s(V_{\Delta x}, \eta'(U_{\Delta x}) \varphi) \]

\[ \begin{align*}
\mathcal{B}_{DG}^s(V_{\Delta x}, \eta'(U_{\Delta x}) \varphi) & = - \sum_{n,K} \int_{I^n} \int_{K} \sum_{k=1}^{d} F^k(V_{\Delta x}^n)(\eta'(U_{\Delta x}^n) \varphi)_{x_k} \, dx \, dt \\
\sum_{n,K,K'} & \int_{I^n} \int_{\partial_{K,K'}} \left( \sum_{k=1}^{d} F^k(V_{\Delta x}^n) \eta'(U_{\Delta x}^n) \nu_{KK'}^k \right) \varphi \, d\sigma \, dt \\
- \frac{1}{2} & \sum_{n,K,K'} \int_{I^n} \int_{\partial_{K,K'}} \eta'(U_{\Delta x}^n) \varphi D(V_{\Delta x}^n + V_{\Delta x}^n - V_{\Delta x}^n) \, d\sigma \, dt \\
= - & \sum_{n,K} \int_{I^n} \int_{K} \sum_{k=1}^{d} q^k(V_{\Delta x}^n) \varphi_x \, dx \, dt \\
\sum_{n,K,K'} & \int_{I^n} \int_{\partial_{K,K'}} \sum_{k=1}^{d} \left( F^k(V_{\Delta x}^n) \eta'(U_{\Delta x}^n) - q^k(V_{\Delta x}^n) \right) \varphi \nu_{KK'}^k \, d\sigma \, dt \\
\sum_{n,K,K'} & \int_{I^n} \int_{\partial_{K,K'}} \left( \sum_{k=1}^{d} F^k(V_{\Delta x}^n, V_{\Delta x}^n) \eta'(U_{\Delta x}^n) \nu_{KK'}^k \right) \varphi \, d\sigma \, dt \\
- \frac{1}{2} & \sum_{n,K,K'} \int_{I^n} \int_{\partial_{K,K'}} \eta'(U_{\Delta x}^n) \varphi D(V_{\Delta x}^n + V_{\Delta x}^n - V_{\Delta x}^n) \, d\sigma \, dt.
\end{align*} \] (4.43)
4. Convergence analysis for systems of conservation laws

Using the definition of the residual (3.18) and the streamline diffusion operator, we have

\[ B_{SD}(V^{\Delta x}, \eta'(U^{\Delta x}) \varphi) \]

\[ = \sum_{n,K} \int_I \int_K \left( U_V(V^{\Delta x})(\eta'(U^{\Delta x})\varphi)_t + \sum_{k=1}^{d} F_{V}^{k}(V^{\Delta x})(\eta'(U^{\Delta x})\varphi)_{x_k} \right) D^SD \text{Res} \, dx \, dt \]

\[ = \sum_{n,K} \int_I \int_K \eta''(U^{\Delta x}) U_V(V^{\Delta x}) \text{Res} D^SD \text{Res} \varphi \, dx \, dt \]

\[ + \sum_{n,K} \int_I \int_K \left( U_V(V^{\Delta x})\eta'(U^{\Delta x})\varphi_t + \sum_{k=1}^{d} F_{V}^{k}(V^{\Delta x})\eta'(U^{\Delta x})\varphi_{x_k} \right) D^SD \text{Res} \, dx \, dt. \]  

(4.44)

Next, from the definition of the shock-capturing operator, we obtain

\[ B_{SC}(V^{\Delta x}, \eta'(U^{\Delta x}) \varphi) \]

\[ = \sum_{n,K} \int_I \int_K D_{n,K}^{SC} \left( \eta'(U^{\Delta x}) \varphi_t U_VV^{\Delta x}_t + \sum_{k=1}^{d} \frac{\Delta x K^2}{(\Delta t^n)^2} \eta'(U^{\Delta x})\varphi_{x_k} \tilde{U}_V V^{\Delta x}_{x_k} \right) dx \, dt \]

\[ = \sum_{n,K} \int_I \int_K D_{n,K}^{SC} \left( \eta''(U^{\Delta x}) U_V(V^{\Delta x})V^{\Delta x}_t \tilde{U}_V V^{\Delta x}_t \right. \]

\[ + \left. \sum_{k=1}^{d} \frac{\Delta x K^2}{(\Delta t^n)^2} \eta''(U^{\Delta x}) U_V(V^{\Delta x}) V^{\Delta x}_{x_k} \tilde{U}_V V^{\Delta x}_{x_k} \right) \varphi \, dx \, dt \]

\[ + \sum_{n,K} \int_I \int_K D_{n,K}^{SC} \left( \eta'(U^{\Delta x}) \varphi_t \tilde{U}_V V^{\Delta x}_t + \sum_{k=1}^{d} \frac{\Delta x K^2}{(\Delta t^n)^2} \eta'(U^{\Delta x}) \varphi_{x_k} \tilde{U}_V V^{\Delta x}_{x_k} \right) dx \, dt. \]  

(4.45)

This completes the proof. \( \square \)

We are now in the position to prove the pre-compactness of \( \{ \partial_t \eta(V^{\Delta x}) + \text{div}(q(V^{\Delta x})) \} \) that allows us later on to show strong convergence.

**Theorem 4.3.4.** Let \( V^{\Delta x} \) be the approximate solutions generated by the scheme (3.2) and \( (\eta, q) \) be any entropy/entropy flux pair. If \( D \) is uniformly positive and \( V^{\Delta x} \) is uniformly bounded in \( L^\infty \), then the sequence

\[ \{ \partial_t \eta(V^{\Delta x}) + \text{div}(q(V^{\Delta x})) \}_{\Delta x > 0} \]

is pre-compact in \( H^{-1}_{\text{loc}}(\Omega \times (0, T)) \).

50
4.3. Convergence for scalar conservation laws

Proof. We consider a test function \( \varphi \) in \( C^\infty_c(\Omega \times (0, T)) \). We define

\[
\langle L_{\Delta x}, \varphi \rangle := \langle \partial_t \eta (V_{\Delta x}) + \text{div} \left( q(V_{\Delta x}) \right), \varphi \rangle 
\]

\[
= - \sum_{n,K} \int_I^n \int_K \eta(U_{\Delta x}) \varphi_t dx dt - \sum_{n,K} \int_I^n \int_K \sum_{k=1}^d q^k(V_{\Delta x}) \varphi_{x_k} dx dt 
\]

\[
= \langle L^{\text{int}, t}, \varphi \rangle + \langle L^{\text{int}, s}, \varphi \rangle 
\]

\[
= B \left( V_{\Delta x}, \eta'(U_{\Delta x}) \varphi - \Pi_{\Delta x} \left( \eta'(U_{\Delta x}) \varphi \right) \right) - \langle L^t, \varphi \rangle - \langle L^{s, 1}, \varphi \rangle - \langle L^{s, 2}, \varphi \rangle - \langle L^{s, 3}, \varphi \rangle - \langle L^{\text{SD}, 1}, \varphi \rangle - \langle L^{\text{SD}, 2}, \varphi \rangle - \langle L^{\text{SC}, 1}, \varphi \rangle - \langle L^{\text{SC}, 2}, \varphi \rangle, \quad \text{(by Theorem 4.3.3)} 
\]

where we have also used the fact that

\[
B \left( V_{\Delta x}, \eta'(U_{\Delta x}) \varphi \right) = B \left( V_{\Delta x}, \eta'(U_{\Delta x}) \varphi - \Pi_{\Delta x} \left( \eta'(U_{\Delta x}) \varphi \right) \right) 
\]

by definition of the scheme. Thus, we use first \( \eta'(U_{\Delta x}) \varphi \) as a test function and later on have to consider the projection error \( \eta'(U_{\Delta x}) \varphi - \Pi_{\Delta x} \left( \eta'(U_{\Delta x}) \varphi \right) \). We will show in a series of claims that the operators considered above are actually pre-compact in \( H^{-1}_{\text{loc}}(\Omega \times (0, T)) \).

Claim 1: First we claim that \( L^t, L^{s, 3}, L^{\text{SD}, 1}, L^{\text{SC}, 1} \) are in a bounded subset of \( M_{\text{loc}}(\Omega \times (0, T)) \).

Note that, in order to prove the claim, we must show that

\[
|\langle L, \varphi \rangle| \leq C \| \varphi \|_{L^\infty(\Omega \times [0, T])} 
\]

for the given operators \( L \).

To begin with, observe that

\[
|\langle L^t, \varphi \rangle| \leq \| \eta'' \|_{L^\infty(\Omega \times [0, T])} \| \varphi \|_{L^\infty(\Omega \times [0, T])} \sum_{n,K} \int_K \left( U_{n, -}^{\Delta x} - U_{n, +}^{\Delta x} \right)^2 dx 
\]

\[
\leq C \| \varphi \|_{L^\infty(\Omega \times [0, T])} \quad \text{(by the weak BV estimate (3.23)).} 
\]

Next, we obtain by splitting and changing the normal direction

\[
\langle L^{s, 3}, \varphi \rangle = \frac{1}{4} \sum_{n,K,K'} \int_I^n \int_{\partial_{KK'}} \left( \eta'(U_{K,+}^{\Delta x}) - \eta'(U_{K,-}^{\Delta x}) \right) \varphi D(V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}) d\sigma dt 
\]

\[
= \frac{1}{4} \sum_{n,K,K'} \int_I^n \int_{\partial_{KK'}} \eta''(U(\theta)) U_V(\theta) \varphi D(V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x})^2 d\sigma dt 
\]
4. Convergence analysis for systems of conservation laws

Hence, we have

\[
\left| \langle \mathcal{L}^{s,3}, \varphi \rangle \right| \leq \| \eta'' \|_{L^\infty(\Omega \times [0,T])} \| V^\Delta x \|_{L^\infty(\Omega \times [0,T])} \| \varphi \|_{L^\infty(\Omega \times [0,T])} \\
\left( \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} D (V^\Delta x_{K,+} - V^\Delta x_{K,-})^2 \, d\sigma dt \right) \\
\leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \left( \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} D (V^\Delta x_{K,+} - V^\Delta x_{K,-})^2 \, d\sigma dt \right) \\
\leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \quad \text{(by the weak BV estimate (3.23)).}
\]

Next, we obtain for the streamline diffusion term

\[
\left| \langle \mathcal{L}^{SD,1}, \varphi \rangle \right| \leq C \| \eta'' \|_{L^\infty(\Omega \times [0,T])} \| \varphi \|_{L^\infty(\Omega \times [0,T])} \left( \Delta x \sum_{n,K} \int_{I_n} \int_{K} |\text{Res}_n|^2 \, dx \, dt \right) \\
\leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \quad \text{(by the weak BV estimate (3.23)).}
\]

Finally, we split \( \langle \mathcal{L}^{SC,1}, \varphi \rangle = \langle \mathcal{L}^{SC,1,1}, \varphi \rangle + \langle \mathcal{L}^{SC,1,2}, \varphi \rangle \), where

\[
\langle \mathcal{L}^{SC,1,1}, \varphi \rangle := \sum_{n,K} \int_{I_n} \int_{K} D^{SC,1}_{n,K} \eta''(U^\Delta x) U_V (V^\Delta x) \left( V^\Delta x_t \tilde{U}_V + \sum_{k=1}^{d} V^\Delta x_{x_k} \tilde{U}_V V^\Delta x_{x_k} \right) \varphi \, dx \, dt,
\]

and

\[
\langle \mathcal{L}^{SC,1,2}, \varphi \rangle := \sum_{n,K} \int_{I_n} \int_{K} D^{SC,2}_{n,K} \eta''(U^\Delta x) U_V (V^\Delta x) \left( V^\Delta x_t \tilde{U}_V + \sum_{k=1}^{d} V^\Delta x_{x_k} \tilde{U}_V V^\Delta x_{x_k} \right) \varphi \, dx \, dt.
\]

It is easy to show that

\[
\left| \langle \mathcal{L}^{SC,1,1}, \varphi \rangle \right| \leq C \| \eta'' \|_{L^\infty} \| V^\Delta x \|_{L^\infty(\Omega \times [0,T])} \| \varphi \|_{L^\infty(\Omega \times [0,T])} \\
\left( \Delta x \sum_{n,K} \text{Res}_{n,K} \left( \int_{I_n} \int_{I_n} \| \nabla_x V \|^2 \, dx \, dt \right)^{1/2} + \epsilon \right) \\
\leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \left( \Delta x \sum_{n,K} \text{Res}_{n,K} \left( \int_{I_n} \int_{I_n} \| \nabla_x V \|^2 \, dx \, dt \right)^{1/2} \right)^{1/2} \\
\leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \quad \text{(by weak BV estimate (3.23)).}
\]
4.3. Convergence for scalar conservation laws

and

\[ |\langle \mathcal{L}_{SC,1,2}, \varphi \rangle| \leq C \| \eta'' \|_{L^\infty} \| V^\Delta x \|_{L^\infty(\Omega \times [0,T])} \| \varphi \|_{L^\infty(\Omega \times [0,T])} \]

\[ \left( (\Delta x)^{\frac{1}{2}} \sum_{n,K} \frac{\text{BRes}_{n,K}}{\| \nabla x t V \|^2 \| x t \|^2 dx dt} \right) \left( \int_K \int_I \| \nabla x t V \|^2 dx dt \right)^{\frac{1}{2}} \]

\[ \leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \left( (\Delta x)^{\frac{1}{2}} \sum_{n,K} \text{BRes}_{n,K} \left( \int_K \int_I \| \nabla x t V \|^2 dx dt \right)^{\frac{1}{2}} \right) \]

\[ \leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \] (by weak BV estimate (3.23)).

Claim 2: We next claim that

\[ \mathcal{L}_{SD,2}, \mathcal{L}_{SC,2} \text{ are in a compact subset of } H^{-1}_{\text{loc}}(\Omega \times [0,T]). \]

Note that, in order to prove the claim, it is sufficient to show that

\[ |\langle L, \varphi \rangle| \leq C(\Delta x)^\alpha \| \varphi \|_{H^1(\Omega \times [0,T])}, \text{ for } \alpha > 0, \]

for the given operator \( L \).

We begin with

\[ |\langle \mathcal{L}_{SD,2}, \varphi \rangle| \leq C \| V^\Delta x \|_{L^\infty(\Omega \times [0,T])} \left( \Delta x \sum_{n,K} \| \varphi \|^2_{H^1(K \times I^n)} \right)^{\frac{1}{2}} \]

\[ \left( \Delta x \sum_{n,K} \int_I \int_K |\text{Res}|^2 dx dt \right)^{\frac{1}{2}} \]

\[ \leq C(\Delta x)^{\frac{1}{2}} \| \varphi \|_{H^1(\Omega \times [0,T])} \] (by weak BV estimate (3.23)).

Repeating the calculation in (4.7), we obtain for the shock-capturing term

\[ |\langle \mathcal{L}_{SC,2}, \varphi \rangle| \leq C(\Delta x)^{\frac{1}{2}} \| V^\Delta x \|_{L^\infty(\Omega \times [0,T])} \| \varphi \|_{H^1(\Omega \times [0,T])} \]

\[ \left( \left( \Delta x \sum_{n,K} \int_I \int_K \| \text{Res} \|^2 dx dt \right)^{\frac{1}{2}} + \left( \sum_{n,K} \| \text{BRes}_{n,K} \|^2 \right)^{\frac{1}{2}} \right) \]

\[ \leq C(\Delta x)^{\frac{1}{2}} \| \varphi \|_{H^1(\Omega \times [0,T])} \] (by weak BV estimate (3.23)).

Claim 3: Next, we claim that

\[ \mathcal{L}^{s,1} + \mathcal{L}^{s,2} \text{ is in a bounded subset of } M_{\text{loc}}(\Omega \times (0,T)). \]
4. Convergence analysis for systems of conservation laws

First, note that by splitting and changing the normal direction

\[ \langle L^{s,2}, \varphi \rangle = \sum_{n,K,K',k=1}^d q^k (V_{K,-}^{\Delta x}) \varphi \nu_{KK'}^k, d\sigma dt \]

\[ = -\frac{1}{2} \sum_{n,K,K',k=1}^d (q^k (V_{K,+}^{\Delta x}) - q^k (V_{K,-}^{\Delta x})) \varphi \nu_{KK'}^k, d\sigma dt \]

\[ = -\frac{1}{2} \sum_{n,K,K',k=1}^d (q^k)'(\theta_1^k) (V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}) \varphi \nu_{KK'}^k, d\sigma dt, \]

where \( \theta_1^k \) lies in the interval \([V_{K,+}^{\Delta x}, V_{K,-}^{\Delta x}]\). Again using the definition of \( q^k \), we have

\[ (q^k)'(V) = \eta'(U(V))(F^k)'(V) \]

and using the consistency relation \( F^k(V) = \mathbb{F}^k,V(V,V) \), we have\(^1\)

\[ (F^k)'(V) = \mathbb{F}^k,V_1(V,V) + \mathbb{F}^k,V_2(V,V). \]

Consequently, we can write

\[ (q^k)'(\theta_1^k) = \eta'(U(\theta_1^k))(F^k)'(\theta_1^k) \]

\[ = \eta'(U(\theta_1^k)) \left( \mathbb{F}^k,V_1 (\theta_1^k, \theta_1^k) + \mathbb{F}^k,V_2 (\theta_1^k, \theta_1^k) \right) \]

For \( L^{s,1} \) on the other hand, the symmetry of the numerical flux, the consistency relation, and the Taylor’s expansion gives

\[ \mathbb{F}^k,V_1 (V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}) - F^k(V_{K,-}^{\Delta x}) \]

\[ = \frac{1}{2} \left( \mathbb{F}^k,V_1 (V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}) + \mathbb{F}^k,V_1 (V_{K,+}^{\Delta x}, V_{K,-}^{\Delta x}) \right) - \frac{1}{2} \left( \mathbb{F}^k,V_2 (V_{K,+}^{\Delta x}, V_{K,-}^{\Delta x}) + \mathbb{F}^k,V_2 (V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}) \right) \]

\[ = \frac{1}{2} (V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}) \left( \mathbb{F}^k,V_1 (\theta_2^k, V_{K,-}^{\Delta x}) + \mathbb{F}^k,V_2 (V_{K,-}^{\Delta x}, \theta_3^k) \right), \]

where \( \theta_2^k, \theta_3^k \) lies in the interval \([V_{K,+}^{\Delta x}, V_{K,-}^{\Delta x}]\).

We now start combining the two parts. Again using Taylor’s expansion once more and

---

\(^1\)Here, \( \mathbb{F}^k,V_1 \) means the partial derivative of \( \mathbb{F}^k,V \) with respect to first variable. Similarly, \( \mathbb{F}^k,V_2 \) denotes the partial derivative with respect to second variable.
4.3. Convergence for scalar conservation laws

the regularity of the numerical flux function, we write

\[ \eta'(U_{\Delta x}^k,_{k,-}) F_{V_1}^{k,*}(\theta_{k}^2, V_{K,-}^\Delta x) - \eta'(\theta_{k}^1) F_{V_1}^{k,*}(\theta_{k}^1, \theta_{k}^1) \]

\[ = \eta'(U_{\Delta x}^k,_{k,-}) \left( F_{V_1}^{k,*}(\theta_{k}^2, V_{K,-}^\Delta x) - F_{V_1}^{k,*}(\theta_{k}^1, \theta_{k}^1) \right) + \left( \eta'(U_{\Delta x}^k,_{k,-}) - \eta'(U_{\Delta x}^k,_{k,-}) \right) F_{V_1}^{k,*}(\theta_{k}^1, \theta_{k}^1) \]

\[ = \eta'(U_{\Delta x}^k,_{k,-}) \left( F_{V_1}^{k,*}(\theta_{k}^2, \theta_{k}^3)(\theta_{k}^2 - \theta_{k}^1) + F_{V_1}^{k,*}(\theta_{k}^3, \theta_{k}^4)(V_{K,-}^\Delta x - \theta_{k}^1) \right) \]

\[ + \eta''(U(\theta_{k}^5)) U_V(\theta_{k}^5)(V_{K,-}^\Delta x - \theta_{k}^1) F_{V_1}^{k,*}(\theta_{k}^1, \theta_{k}^1) \]

\[ =: F_{1}^k. \]

Similarly, we can write

\[ \eta'(U_{\Delta x}^k,_{k,-}) F_{V_2}^{k,*}(V_{K,-}^\Delta x, \theta_{k}^1) - \eta'(\theta_{k}^1) F_{V_2}^{k,*}(\theta_{k}^1, \theta_{k}^1) \]

\[ = \eta'(U_{\Delta x}^k,_{k,-}) \left( F_{V_2}^{k,*}(\theta_{k}^0, \theta_{k}^3)(\theta_{k}^0 - \theta_{k}^1) + F_{V_2}^{k,*}(\theta_{k}^3, \theta_{k}^2)(V_{K,-}^\Delta x - \theta_{k}^1) \right) \]

\[ + \eta''(U(\theta_{k}^5)) U_V(\theta_{k}^5)(V_{K,-}^\Delta x - \theta_{k}^1) F_{V_2}^{k,*}(\theta_{k}^1, \theta_{k}^1) \]

\[ =: F_{2}^k. \]

Now we are in the position to give a proof of the Claim 3. Indeed using all the results above, we have

\[ \langle \mathcal{L}^{s,1} + \mathcal{L}^{s,2}, \varphi \rangle = \frac{1}{2} \sum_{n,K,K'} \int_{I^n} \sum_{\theta_{K,K'}} \sum_{k=1}^d (V_{K,+}^\Delta x - V_{K,-}^\Delta x)(F_{1}^k + F_{2}^k) \nu_{K,K'} \varphi d\sigma dt \]

Finally, it is easy to see that

\[ \| \mathcal{L}^{s,1} + \mathcal{L}^{s,2}, \varphi \| \]

\[ \leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \left( \| \eta'' \|_{L^\infty(\Omega \times [0,T])} \left( \sum_{n,K,K',} \int_{I^n} \int_{\partial_{K,K'}} (V_{K,+}^\Delta x - V_{K,-}^\Delta x)^2 d\sigma dt \right) \right) \]

\[ \leq C \| \varphi \|_{L^\infty(\Omega \times [0,T])} \text{ (by weak BV estimate (3.23)),} \]

where one uses that \( |\theta_{k}^2 - \theta_{k}^1|, |\theta_{k}^0 - \theta_{k}^1|, \) and \( |V_{K,+}^\Delta x - V_{K,-}^\Delta x| \) are bounded by \( |V_{K,+}^\Delta x - V_{K,-}^\Delta x| \).

Next, we use the \( H^1 \) projection operator \( \Pi^\Delta x \) from (4.2) and show the following claim,

Claim 4: We have

\[ \langle \mathcal{L}^{\Pi}, \varphi \rangle \text{ is in a compact subset of } H^1_{\text{loc}}(\Omega \times (0,T)), \]

where we defined

\[ \langle \mathcal{L}^{\Pi}, \varphi \rangle := B \left( V^\Delta x, \eta'(U^\Delta x) \varphi - \Pi^\Delta x(\eta'(U^\Delta x) \varphi) \right). \]
4. Convergence analysis for systems of conservation laws

We continue with defining the projection error as
\[
\Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi) = \eta'(U^\Delta x)\varphi - \Pi_{\Delta x}^\varphi(\eta'(U^\Delta x)\varphi)
\] (4.48)
and splitting
\[
\langle \mathcal{L}^\Pi, \varphi \rangle = \langle \mathcal{L}^{\Pi,\text{DG}}, \varphi \rangle + \langle \mathcal{L}^{\Pi,\text{SD}}, \varphi \rangle + \langle \mathcal{L}^{\Pi,\text{SC}}, \varphi \rangle,
\] (4.49)
where
\[
\langle \mathcal{L}^{\Pi,\text{DG}}, \varphi \rangle := B_{\text{DG}}(V_{\Delta x}, \Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))
\]
\[
\langle \mathcal{L}^{\Pi,\text{SD}}, \varphi \rangle := B_{\text{SD}}(V_{\Delta x}, \Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))
\]
\[
\langle \mathcal{L}^{\Pi,\text{SC}}, \varphi \rangle := B_{\text{SC}}(V_{\Delta x}, \Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi)).
\]

We start with calculations involving the DG quasilinear form (3.3). For the interior terms, we obtain using integration by parts
\[
\mathcal{B}_{\text{DG}}^{\text{int}}(V_{\Delta x}, \Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))
\]
\[
= - \sum_{n,K} \int_{I^n} \int_K U(V_{\Delta x}^{n+1}) (\Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))_{t} dx dt
\]
\[
- \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d F_k(V_{\Delta x}^{n+1}) (\Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))_{x_k} dx dt
\]
\[
= \sum_{n,K} \int_{I^n} \int_K \left( U(V_{\Delta x}^{n+1}) + \sum_{k=1}^d F_k(V_{\Delta x}^{n+1}) \Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi) \right) dx dt
\]
\[
- \sum_{n,K} \int_K U(V_{n+1}^{\Delta x}) (\Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))^{n+1,-} dx
\]
\[
+ \sum_{n,K} \int_K U(V_{n+1}^{\Delta x}) (\Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))^{n,+} dx
\]
\[
- \sum_{n,K,K',} \int_{I^n} \int_{\partial_{K,K'}} \left( \sum_{k=1}^d F_k(V_{K,-}^{\Delta x}), (\Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))_{K,-} \nu_{K,K'}^k \right) d\sigma dt.
\] (4.50)

Using (4.50) and calculating with the full DG quasilinear form, we obtain
\[
\langle \mathcal{L}^{\Pi,\text{DG}}, \varphi \rangle
\]
\[
= B_{\text{DG}}^{\text{int}}(V_{\Delta x}, \Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))
\]
\[
+ \sum_{n,K} \int_K U(V_{n+1}^{\Delta x}) (\Pi_{\Delta x}^{\text{err}}(\eta'(U^\Delta x)\varphi))^{n+1,-} dx
\]
4.3. Convergence for scalar conservation laws

\[ - \sum_{n,K} \int_{K} U(V_{n,K}^{\Delta x}) \left( \Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x})\varphi) \right)^{n,+} dx \]

\[ + \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \sum_{k=1}^{d} \left| \Pi_k^{\Delta x}(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}) \right| \left( \Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x})\varphi) \right)_{K,-} \nu_{K K'}^k d\sigma dt \]

\[ - \frac{1}{2} \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \left( \Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x})\varphi) \right)_{K,-} \Delta D(V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}) d\sigma dt \]

\[ = \sum_{n,K} \int_{I_n} \int_{K} \text{Res} \Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x})\varphi) dx dt \]

\[ + \sum_{n,K} \int_{K} \left( U(V_{n,K}^{\Delta x}) - U(V_{n,K}^{\Delta x}) \right) \left( \Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x})\varphi) \right)^{n,+} dx \]

\[ + \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \sum_{k=1}^{d} \left( \Pi_k^{\Delta x}(V_{K,-}^{\Delta x}, V_{K,+}^{\Delta x}) - F_k(V_{K,-}^{\Delta x}) \right) \left( \Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x})\varphi) \right)_{K,-} \nu_{K K'}^k d\sigma dt \]

\[ - \frac{1}{2} \sum_{n,K,K'} \int_{I_n} \int_{\partial K K'} \left( \Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x})\varphi) \right)_{K,-} \Delta D(V_{K,+}^{\Delta x} - V_{K,-}^{\Delta x}) d\sigma dt \]

\[ =: \langle L^{\Pi,\text{int}}, \varphi \rangle + \langle L^{\Pi,\partial}, \varphi \rangle, \quad (4.51) \]

i.e. by splitting into interior and boundary terms.

Let \( \bar{\varphi} \) be the mean value of \( \varphi \) in each element (i.e. the \( L^2 \) projection of \( \varphi \) onto the set of piecewise constant functions). Note that it therefore satisfies the following estimates

\[
\| \varphi - \bar{\varphi} \|_{L^2(K \times I^n)} \leq C \Delta x \| \varphi \|_{H^1(K \times I^n)},
\]

\[
\| \nabla_{xt} (\varphi - \bar{\varphi}) \|_{L^2(K \times I^n)} \leq C \| \varphi \|_{H^1(K \times I^n)},
\]

\[
\| \varphi - \bar{\varphi} \|_{L^2(\partial(K \times I^n))} \leq C \Delta x^{3/2} \| \varphi \|_{H^1(K \times I^n)}.
\]

We will additionally make use of the following estimates

\[
\| \nabla_{xt} (V^{\Delta x})(\varphi - \bar{\varphi}) \|_{L^2(K \times I^n)}
\]

\[
\leq \| \nabla_{xt} (V^{\Delta x}) \|_{L^\infty(K \times I^n)} \| \varphi - \bar{\varphi} \|_{L^2(K \times I^n)}
\]

\[
\leq C \Delta x^{-1} \| V^{\Delta x} \|_{L^\infty(K \times I^n)} \Delta x \| \varphi \|_{H^1(K \times I^n)} \quad \text{(inverse inequality)}
\]

\[
\leq C \| \varphi \|_{H^1(K \times I^n)},
\]

and

\[
\| \nabla_{xt} (\eta'(U^{\Delta x})(\varphi - \bar{\varphi})) \|_{L^2(K \times I^n)}
\]

\[
\leq \| \nabla_{xt} (\eta'(U^{\Delta x}))(\varphi - \bar{\varphi}) \|_{L^2(K \times I^n)} + \| \eta'(U^{\Delta x}) \nabla_{xt} (\varphi - \bar{\varphi}) \|_{L^2(K \times I^n)}
\]
4. Convergence analysis for systems of conservation laws

\[
\begin{align*}
&\leq \|\eta''(U^{\Delta x})U_V(V^{\Delta x})\nabla_{xt}(V^{\Delta x})(\varphi - \varphi')\|_{L^2(K \times I^n)} \\
&+ \|\eta'(U^{\Delta x})\|_{L^\infty(\Omega \times (0,T))}\|\nabla_{xt}(\varphi - \varphi')\|_{L^2(K \times I^n)} \\
&\leq \|\eta''(U^{\Delta x})U_V(V^{\Delta x})\|_{L^\infty(\Omega \times (0,T))}\|\nabla_{xt}(V^{\Delta x})(\varphi - \varphi')\|_{L^2(K \times I^n)} + C\|\varphi\|_{H^1(K \times I^n)} \\
&\leq C\|\varphi\|_{H^1(K \times I^n)} \quad \text{(by (4.52)).} \\
\end{align*}
\]

We start with the projection error for the interior part and split it in two parts

\[
\langle L_{\Pi,\text{int},1}, \varphi \rangle = \sum_{n,K} \int_{I^n} \int_K \text{Res}(\Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x})\varphi)) \\
= \sum_{n,K} \int_{I^n} \int_K \text{Res}(\eta'(U^{\Delta x})\varphi - \Pi^{\Delta x}(\eta'(U^{\Delta x})\varphi)) \\
= \sum_{n,K} \int_{I^n} \int_K \text{Res}(\eta'(U^{\Delta x})\varphi - \eta'(U^{\Delta x})\bar{\varphi} + \Pi^{\Delta x}(\eta'(U^{\Delta x})\bar{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})\varphi)) \\
+ \sum_{n,K} \int_{I^n} \int_K \text{Res}(\eta'(U^{\Delta x})\bar{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x})\bar{\varphi})) \\
=: \langle L_{\Pi,\text{int},1}, \varphi \rangle + \langle L_{\Pi,\text{int},2}, \varphi \rangle.
\]

We obtain for the first part

\[
\langle L_{\Pi,\text{int},1}, \varphi \rangle \leq C \sum_{n,K} \text{Res}_{n,K} \|\eta'(U^{\Delta x})\varphi - \eta'(U^{\Delta x})\bar{\varphi} + \Pi^{\Delta x}(\eta'(U^{\Delta x})\bar{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})\varphi)\|_{L^2(K \times I^n)}.
\]

As

\[
\|\eta'(U^{\Delta x})\varphi - \eta'(U^{\Delta x})\bar{\varphi} + \Pi^{\Delta x}(\eta'(U^{\Delta x})\bar{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})\varphi)\|_{L^2(K \times I^n)} \\
= \|\eta'(U^{\Delta x})(\varphi - \bar{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})(\varphi - \bar{\varphi}))\|_{L^2(K \times I^n)} \\
\leq C\Delta x\|\nabla_{xt}(\eta'(U^{\Delta x})(\varphi - \bar{\varphi}))\|_{L^2(K \times I^n)} \quad \text{(projection error estimate (4.4))} \\
\leq C\Delta x\|\varphi\|_{H^1(K \times I^n)} \quad \text{(by (4.53)),}
\]

this results in

\[
\langle L_{\Pi,\text{int},1}, \varphi \rangle \leq C\Delta x \sum_{n,K} \text{Res}_{n,K} \|\varphi\|_{H^1(K \times I^n)} \\
\leq C\Delta x^{\frac{1}{2}} \left( \Delta x \sum_{n,K} \text{Res}_{n,K}^2 \right)^{\frac{1}{2}} \|\varphi\|_{H^1(\Omega \times (0,T))} \\
\leq C\Delta x^{\frac{1}{2}} \|\varphi\|_{H^1(\Omega \times (0,T))} \quad \text{(by weak BV estimate (3.23)).}
\]

58
4.3. Convergence for scalar conservation laws

We conclude that $\mathcal{L}_{\Pi,\text{int},1}^\ast$ is in a compact subset of $H_{\text{loc}}^{-1}(\Omega \times (0, T))$.

We obtain for the second part

$$\langle \mathcal{L}_{\Pi,\text{int},2}^\ast, \varphi \rangle \leq C \sum_{n,K} \overline{\text{BRes}}_{n,K} \| \eta'(U^{\Delta x}) \tilde{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x}) \tilde{\varphi}) \|_{L^2(K \times I^n)}.$$ 

We continue with estimating

$$\| \eta'(U^{\Delta x}) \tilde{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x}) \tilde{\varphi}) \|_{L^2(K \times I^n)}$$

$$= \| \tilde{\varphi} \|_{L^\infty(\Omega \times (0, T))} \| \eta'(U^{\Delta x}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})) \|_{L^2(K \times I^n)}$$

$$\leq C \Delta x \| \varphi \|_{L^\infty(\Omega \times (0, T))} \| \nabla_{xt}(\eta'(U^{\Delta x})) \|_{L^2(K \times I^n)} \quad \text{(projection error estimate (4.4))}$$

$$\leq C \Delta x \| \varphi \|_{L^\infty(\Omega \times (0, T))} \| \eta''(U^{\Delta x}) U_V(V^{\Delta x}) \nabla_{xt}(V^{\Delta x}) \|_{L^2(K \times I^n)}$$

$$\leq C \Delta x \| \varphi \|_{L^\infty(\Omega \times (0, T))} \| \eta''(U^{\Delta x}) U_V(V^{\Delta x}) \|_{L^\infty(\Omega \times (0, T))} \| \nabla_{xt}(V^{\Delta x}) \|_{L^2(K \times I^n)}$$

$$\leq C \Delta x \| \varphi \|_{L^\infty(\Omega \times (0, T))} \| \nabla_{xt}(V^{\Delta x}) \|_{L^2(K \times I^n)}. \quad (4.54)$$

This implies

$$\langle \mathcal{L}_{\Pi,\text{int},2}^\ast, \varphi \rangle \leq C \| \varphi \|_{L^\infty(\Omega \times (0, T))} \left( \sum_{n,K} \overline{\text{BRes}}_{n,K} \| \nabla_{xt}(V^{\Delta x}) \|_{L^2(K \times I^n)} \right)$$

$$\leq C \| \varphi \|_{L^\infty(\Omega \times (0, T))} \quad \text{(by weak BV estimate (3.23))}$$

and further, $\mathcal{L}_{\Pi,\text{int},2}^\ast$ is in a bounded subset of $\mathcal{M}_{\text{loc}}(\Omega \times (0, T))$.

For the boundary part, let us first introduce the notation

$$\text{BRes}_{n,K} = \begin{cases} U(V_{n,+}^{\Delta x}) - U(V_{n,-}^{\Delta x}) & \text{on } K \times \{t^n\}, \\
\mathbb{F}(V_{K,-}^{\Delta x} ; V_{K,+}^{\Delta x} ; \nu_{KK'}) - \sum_{k=1}^d F_k(V_{K,-}^{\Delta x}) \nu_{KK'}^k & \text{on } \partial_{KK'} \times I^n. \end{cases}$$

Then, we split analogously

$$\langle \mathcal{L}_{\Pi,\partial}^\ast, \varphi \rangle$$

$$= \sum_{n,K} \left( \int_K (U(V_{n,+}^{\Delta x}) - U(V_{n,-}^{\Delta x})) (\Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x}) \varphi))_{n,+} dx \\
+ \int_{I^n} \int_{\partial_{KK'}} \mathbb{F}(V_{K,-}^{\Delta x} ; V_{K,+}^{\Delta x} ; \nu_{KK'}) - \sum_{k=1}^d F_k(V_{K,-}^{\Delta x}) \nu_{KK'}^k \left( \Pi_{\text{err}}^{\Delta x}(\eta'(U^{\Delta x}) \varphi) \right)_{K,-} d\sigma dt \right)$$

$$= \sum_{n,K} \left( \int_K \text{BRes}_{n,K} (\eta'(U^{\Delta x}) (\varphi - \tilde{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x}) (\varphi - \tilde{\varphi}))_{n,+} dx \\
+ \int_{I^n} \int_{\partial_{KK'}} \text{BRes}_{n,K} (\eta'(U^{\Delta x}) (\varphi - \tilde{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x}) (\varphi - \tilde{\varphi}))_{K,-} d\sigma dt \right)$$

59
4. Convergence analysis for systems of conservation laws

\[ + \sum_{n,K} \left( \int_K \text{BRes}_{n,K} \left( \eta'(U^{\Delta x})\overline{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x})\overline{\varphi}) \right)^{n,+} dx \right. \]

\[ + \int_{I_n} \int_{\partial K,K'} \text{BRes}_{n,K} \left( \eta'(U^{\Delta x})\overline{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x})\overline{\varphi}) \right)_{K,-} d\sigma dt \]

\[ =: \langle \mathcal{L}^{\Pi,\partial,1}, \varphi \rangle + \langle \mathcal{L}^{\Pi,\partial,2}, \varphi \rangle. \]

For the first part, we obtain

\[ |\langle \mathcal{L}^{\Pi,\partial,1}, \varphi \rangle| \leq C \sum_{n,K} \text{BRes}_{n,K} \| \eta'(U^{\Delta x})(\varphi - \overline{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})(\varphi - \overline{\varphi})) \|_{L^2(\partial(K \times I^n))}. \]

We estimate

\[ \| \eta'(U^{\Delta x})(\varphi - \overline{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})(\varphi - \overline{\varphi})) \|_{L^2(\partial(K \times I^n))} \]

\[ \leq C \Delta x^{\frac{1}{2}} \| \nabla_{xt} (\eta'(U^{\Delta x})(\varphi - \overline{\varphi})) \|_{L^2(K \times I^n)} \quad \text{(projection error estimate (4.4))} \]

\[ \leq C \Delta x^{\frac{1}{2}} \| \varphi \|_{H^1(K \times I^n)} \quad \text{(by (4.53))}. \]

This implies

\[ |\langle \mathcal{L}^{\Pi,\partial,1}, \varphi \rangle| \leq C \Delta x^{\frac{1}{2}} \sum_{n,K} \text{BRes}_{n,K} \| \varphi \|_{H^1(K \times I^n)} \]

\[ \leq C \Delta x^{\frac{1}{2}} \left( \sum_{n,K} \text{BRes}_{n,K}^2 \right)^{\frac{1}{2}} \| \varphi \|_{H^1(\Omega \times (0,T))} \]

\[ \leq C \Delta x^{\frac{1}{2}} \| \varphi \|_{H^1(\Omega \times (0,T))} \]

and therefore, the operator \( \mathcal{L}_{\Delta x}^{\Pi,\partial,1} \) is in a compact subset of \( H_{\text{loc}}^{-1}(\Omega \times (0,T)) \).

For the second part, we obtain

\[ |\langle \mathcal{L}^{\Pi,\partial,2}, \varphi \rangle| \leq C \sum_{n,K} \text{BRes}_{n,K} \| \eta'(U^{\Delta x})\overline{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x})\overline{\varphi}) \|_{L^2(\partial(K \times I^n))}. \]

We continue with

\[ \| \eta'(U^{\Delta x})\overline{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x})\overline{\varphi}) \|_{L^2(\partial(K \times I^n))} \]

\[ \leq \| \overline{\varphi} \|_{L^\infty(\Omega \times (0,T))} \| \eta'(U^{\Delta x}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})) \|_{L^2(\partial(K \times I^n))} \]

\[ \leq C \Delta x^{\frac{1}{2}} \| \varphi \|_{L^\infty(\Omega \times (0,T))} \| \nabla_{xt}(\eta'(U^{\Delta x})) \|_{L^2(K \times I^n)} \quad \text{(projection error estimate (4.4))} \]

\[ \leq C \Delta x^{\frac{1}{2}} \| \varphi \|_{L^\infty(\Omega \times (0,T))} \| \nabla_{xt}(V^{\Delta x}) \|_{L^2(K \times I^n)} \quad \text{(as in (4.54))} \]
4.3. Convergence for scalar conservation laws

and

\[ |\langle \mathcal{L}^{\Pi,\partial,2}, \varphi \rangle| \leq C \| \varphi \|_{L^\infty(\Omega \times (0,T))} \left( \Delta x^{3/2} \sum_{n,K} \text{Res}_{n,K} \| \nabla_{xt}(V^{\Delta x}) \|_{L^2(K \times I^n)} \right) \]

\[ \leq C \| \varphi \|_{L^\infty(\Omega \times (0,T))} \quad (\text{by weak BV estimate (3.23))}. \]

Therefore, the operator \( \mathcal{L}^{\Pi,\partial,2} \) is in a bounded subset of \( \mathcal{M}_{\text{loc}}(\Omega \times (0,T)) \).

For the streamline diffusion form, we split in a similar way

\[ \langle \mathcal{L}^{\Pi,\text{SD}}, \varphi \rangle \]

\[ = B_{\text{SD}}(V^{\Delta x}, \Pi^{\Delta x}_{\text{err}}(\eta'(U^{\Delta x})\varphi)) \]

\[ = \sum_{n,K} \int_{I^n} \int_K U_V(V^{\Delta x})(\Pi^{\Delta x}_{\text{err}}(\eta'(U^{\Delta x})\varphi))_t \> D^{\text{SD}} \text{Res} dx dt \]

\[ + \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d F^{k}_V(V^{\Delta x})(\Pi^{\Delta x}_{\text{err}}(\eta'(U^{\Delta x})\varphi))_{x_k} \> D^{\text{SD}} \text{Res} dx dt \]

\[ = \sum_{n,K} \int_{I^n} \int_K \left( U_V(V^{\Delta x}) (\eta'(U^{\Delta x})(\varphi - \bar{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})(\varphi - \bar{\varphi})) \right)_t \]

\[ + \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d \left( F^{k}_V(V^{\Delta x})(\eta'(U^{\Delta x})(\varphi - \bar{\varphi}) - \Pi^{\Delta x}(\eta'(U^{\Delta x})(\varphi - \bar{\varphi})) \right)_{x_k} \> D^{\text{SD}} \text{Res} dx dt \]

\[ + \sum_{n,K} \int_{I^n} \int_K \left( U_V(V^{\Delta x}) (\eta'(U^{\Delta x}) \bar{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x}) \bar{\varphi})) \right)_t \]

\[ + \sum_{n,K} \int_{I^n} \int_K \sum_{k=1}^d \left( F^{k}_V(V^{\Delta x})(\eta'(U^{\Delta x}) \bar{\varphi} - \Pi^{\Delta x}(\eta'(U^{\Delta x}) \bar{\varphi})) \right)_{x_k} \> D^{\text{SD}} \text{Res} dx dt \]

\[ =: \langle \mathcal{L}^{\Pi,\text{SD},1}, \varphi \rangle + \langle \mathcal{L}^{\Pi,\text{SD},2}, \varphi \rangle. \]

We have for the first term

\[ |\langle \mathcal{L}^{\Pi,\text{SD},1}, \varphi \rangle| \leq C \Delta x \sum_{n,K} \text{Res}_{n,K} \left\| \nabla_{xt}(\eta'(U^{\Delta x})(\varphi - \bar{\varphi})) \right\|_{L^2(K \times I^n)} \]

\[ + \left\| \nabla_{xt}(\Pi^{\Delta x}(\eta'(U^{\Delta x})(\varphi - \bar{\varphi})) \right\|_{L^2(K \times I^n)} \].

We further estimate

\[ \| \nabla_{xt}(\eta'(U^{\Delta x})(\varphi - \bar{\varphi})) \|_{L^2(K \times I^n)} \leq C \| \varphi \|_{H^1(K \times I^n)} \quad (\text{by (4.53)}). \]
4. Convergence analysis for systems of conservation laws

and

\[ \| \nabla_x (\Pi^{\Delta x} (\eta'(U^{\Delta x})(\varphi - \bar{\varphi})) ) \|_{L^2(K \times I^n)} \leq \| \nabla_x (\eta'(U^{\Delta x})(\varphi - \bar{\varphi})) \|_{L^2(K \times I^n)} \quad \text{(stability of projection (4.4))} \]

\[ \leq C \| \varphi \|_{H^1(K \times I^n)} \quad \text{(by (4.53))}, \]

which results in

\[ |\langle \mathcal{L}^{\Pi,SD,1}, \varphi \rangle | \leq C \Delta x \sum_{n,K} \text{Re}_n,K \| \varphi \|_{H^1(K \times I^n)} \]

\[ \leq C \Delta x^{1/2} \left( \Delta x \sum_{n,K} \text{Re}_n,K^2 \right)^{1/2} \| \varphi \|_{H^1(\Omega \times (0,T))} \]

\[ \leq C \Delta x^{1/2} \| \varphi \|_{H^1(\Omega \times (0,T))} \quad \text{(by weak BV estimate (3.23))}. \]

We conclude that \( \mathcal{L}^{\Pi,SD,1} \) is in a compact subset of \( H^{-1}_{\text{loc}}(\Omega \times (0,T)) \).

For the second part of the streamline diffusion term, we obtain

\[ |\langle \mathcal{L}^{\Pi,SD,2}, \varphi \rangle | \leq C \Delta x \sum_{n,K} \text{Re}_n,K \| \nabla_x (\eta'(U^{\Delta x})\varphi - \Pi^{\Delta x} (\eta'(U^{\Delta x})\varphi)) \|_{L^2(K \times I^n)}. \]

We continue with

\[ \| \nabla_x (\eta'(U^{\Delta x})\varphi - \Pi^{\Delta x} (\eta'(U^{\Delta x})\varphi)) \|_{L^2(K \times I^n)} \leq \| \varphi \|_{L^\infty(\Omega \times (0,T))} \| \nabla_x (\eta'(U^{\Delta x}) - \Pi^{\Delta x} (\eta'(U^{\Delta x}))) \|_{L^2(K \times I^n)} \]

\[ \leq C \| \varphi \|_{L^\infty(\Omega \times (0,T))} \| \nabla_x (\eta'(U^{\Delta x})) \|_{L^2(K \times I^n)} \quad \text{(projection error estimate (4.4))} \]

\[ \leq C \| \varphi \|_{L^\infty(\Omega \times (0,T))} \| \nabla_x (V^{\Delta x}) \|_{L^2(K \times I^n)} \quad \text{(as in (4.54))} \]

and

\[ |\langle \mathcal{L}^{\Pi,SD,2}, \varphi \rangle | \leq C \| \varphi \|_{L^\infty(\Omega \times (0,T))} \left( \Delta x \sum_{n,K} \text{Re}_n,K \| \nabla_x (V^{\Delta x}) \|_{L^2(K \times I^n)} \right) \]

\[ \leq C \| \varphi \|_{L^\infty(\Omega \times (0,T))} \quad \text{(by weak BV estimate (3.23))}. \]

This implies that \( \mathcal{L}^{\Pi,SD,2} \) is in a bounded subset of \( \mathcal{M}_{\text{loc}}(\Omega \times (0,T)) \).

Given the orthogonality of the projection operator \( \Pi^{\Delta x} \) (4.3) and the definition of the shock-capturing operator, it is easy to see that,

\[ \langle \mathcal{L}^{\Pi,SC}, \varphi \rangle = B_{SC}(V^{\Delta x}, (\Pi^{\Delta x}_{\text{err}} (\eta'(U^{\Delta x})\varphi))) \equiv 0, \quad \text{(4.55)} \]

which trivially implies that \( \mathcal{L}^{\Pi,SC} \) is in a compact subset of \( H^{-1}_{\text{loc}}(\Omega \times (0,T)) \).
4.3. Convergence for scalar conservation laws

Finally, we remark that $L \Delta x$ is in $W^{-1, \infty}(\Omega \times (0, T))$, because

$$|\langle L \Delta x, \phi \rangle| = \left| - \sum_{n,K} \int_{I_n} \int_{K} \eta(U^{\Delta x}) \varphi_t dx dt - \sum_{n,K} \int_{I_n} \int_{K} d \sum_{k=1}^{d} q^{k}(V^{\Delta x}) \varphi_{x_k} dx dt \right| \leq C \|V^{\Delta x}\|_{L^\infty(\Omega \times [0,T])} \|\phi\|_{H^1(\Omega \times [0,T])} \leq C \|\phi\|_{W^{1,\infty}(\Omega \times [0,T])} \tag{4.56}$$

Hence, we can apply Murat’s lemma 4.3.2 and this completes the proof. \qed

4.3.3. Convergence analysis for scalar conservation laws

We start with the following convergence theorem,

**Theorem 4.3.5.** Let $F$ be genuinely non-linear and $U^{\Delta x} = U(V^{\Delta x})$ be the approximate solutions of the scalar conservation law (1.3) generated by the streamline diffusion shock-capturing DG scheme (3.2). If $D$ is uniformly positive and $V^{\Delta x}$ is uniformly bounded in $L^\infty$, then the approximate solutions converge to a weak solution of the conservation law (1.3).

Furthermore, the weak solution is admissible with respect to the entropy $S$.

**Proof.** From Theorem 4.3.4 (pre-compactness of $\{\partial_t \eta (V^{\Delta x}) + \text{div}(q(V^{\Delta x}))\}_{\Delta x > 0}$) together with Theorem 4.3.1, we obtain that $V^{\Delta x}$ converges strongly to a function $V$, as $\Delta x \to 0$. Theorem 4.1.1 (convergence to a measure-valued solution) immediately assures that the limit is a weak solution: In the proof, we obtain

$$\int_0^T \int_{\Omega} \left( U(V^{\Delta x}) \varphi_t + \sum_k F^k(V^{\Delta x}) \varphi_{x_k} \right) dx dt \to - \int_{\Omega} U_0(x) \varphi(x,0) dx \quad \text{as } \Delta x \to 0.\tag{4.57}$$

As the nonlinearities commute with respect to strong convergence, we have that

$$\lim_{\Delta x \to 0} \int_0^T \int_{\Omega} \left( U(V^{\Delta x}) \varphi_t + \sum_k F^k(V^{\Delta x}) \varphi_{x_k} \right) dx dt = \int_0^T \int_{\Omega} \left( U(V) \varphi_t + \sum_k F^k(V) \varphi_{x_k} \right) dx dt$$

Hence, combining the above two limits, we see that $V$ is a weak solution of (1.3).

The consistency with the entropy $S$ can be shown as follows. From Theorem 4.2.1 (admissibility of measure-valued solution), we know that

$$\lim_{\Delta x \to 0} \mathcal{B} \left( V^{\Delta x}, V^{\Delta x} \varphi \right) \geq - \int_0^T \int_{\Omega} \left( S(V) \varphi_t + \sum_{k=1}^{d} Q^k(V) \varphi_{x_k} \right) dx dt - \int_{\Omega} S(U_0(x)) \varphi(x,0) dx, \quad \tag{4.57}$$

63
4. Convergence analysis for systems of conservation laws

and

\[ \lim_{\Delta x \to 0} B \left( V^{\Delta x}, V^{\Delta x} \varphi - \Pi^{\Delta x}(V^{\Delta x} \varphi) \right) = 0. \tag{4.58} \]

Using (4.57), (4.58), and the definition of the scheme (3.2), we obtain,

\[
\int_0^T \int_\Omega \left( S(V) \varphi_t + \sum_{k=1}^d Q^k(V) \varphi_{x_k} \right) dx dt + \int_\Omega S(U_0(x)) \varphi(x, 0) dx \\
\geq - \lim_{\Delta x \to 0} B \left( V^{\Delta x}, V^{\Delta x} \varphi \right) \\
\geq - \lim_{\Delta x \to 0} B \left( V^{\Delta x}, (V^{\Delta x} \varphi - \Pi^{\Delta x}(V^{\Delta x} \varphi)) \right) - \lim_{\Delta x \to 0} B \left( V^{\Delta x}, \Pi^{\Delta x}(V^{\Delta x} \varphi) \right) = 0 \tag{4.59} \]

i.e. \( V \) is consistent with the entropy \( S \).

In one spatial dimension, we obtain the following corollary due to a uniqueness result by Panov [41]:

**Corollary 4.3.6.** Let the assumptions of Theorem 4.3.5 be valid and assume that the flux \( F \) is uniformly convex. For one-dimensional problems \( (d = 1) \), the approximate solutions \( V^{\Delta x} \) converge to the unique entropy solution \( V \) of (1.3) with the given initial data.

Furthermore, the convergence is not only along a subsequence, but the whole sequence converges.

**Proof.** As the entropy \( S \) and the flux \( F \) are uniformly convex by assumption, we obtain that \( V^{\Delta x} \) converges to the unique entropy solution by combining Theorem 4.3.5 with the uniqueness result of Panov [41].

Furthermore, as the limit is unique, one obtains the convergence of the whole sequence by a contradiction argument.

**Remark 4.3.7.** Unfortunately, this does not generalise to multiple spatial dimensions. There, consistency with all entropies is essential. This requires to show that all terms bounded in \( M_{lo} \) in the proof of Theorem 4.3.4 actually have a correct sign. It is straightforward for the streamline diffusion, the shock-capturing, and the temporal terms. For the spatial terms, one has to assume that the numerical is actually an E-flux [40], i.e.

\[
(V^+ - V^-) \left( F(V^-, V^+; \nu) - \sum_{k=1}^d F^k(\theta) \nu^k \right) \leq 0 \tag{4.60} \]

for all \( \theta \) in between \( V^- \) and \( V^+ \) and all directions \( \nu \). This might require using the central flux instead of the entropy-conservative flux and an Enquist-Osher type of diffusion, see [2].
However, the real problem is caused by the projection error. As we can not hope to obtain
the correct sign for these terms, one has to show that the corresponding terms actually vanish
in the limit $\Delta x \to 0$. For the entropy $S$, this is possible thanks to the superapproximation
properties. For more general entropies $\eta$, this is, however, not possible as $\eta'(U\Delta x)$ can be
arbitrarily bad in norms of the form $\|\eta'(U\Delta x)\varphi - \Pi\Delta x(\eta'(U\Delta x)\varphi)\|$. This is actually the reason, why Jaffre et. al. [29] have included a factor $\Delta x^{-\alpha}$ with
$0 < \alpha < 1/2$ in the definition of the shock-capturing operator. See Figure 4.1 for numerical
results with $\alpha > 0$. The underlying problem is the linear advection equation (1.16) and the
initial data is smooth (7.16). Therefore, also the solution is smooth and we observe high order
convergence for moderate values of $\alpha$ on coarse grids. However, as soon as $\Delta x$ becomes small
enough, the factor $\Delta x^{-\alpha}$ destroys the high-resolution property of the scheme as all gradients
in the approximate solutions are diffused out and we observe only a convergence that is very
similar to the one of a scheme with $p = 0$.

Figure 4.1.: Convergence for the linear advection equation using $p = 1$ or $p = 2$ and different
positive values for $\alpha$ together with $p = 0$.

Remark 4.3.8. Note that in [29], the authors additionally require a lower bound on $D_{n,K}^{SC}$
in terms of $\Delta x^{3/2}$ (see equation (2.14) of [29]), which will reduce the convergence rate for
smooth problems. We do not require such a lower bound.

4.4. Convergence for linear symmetrisable systems

Consider the linear symmetrisable system (1.21) and complete the scheme (3.2) with the
entropy-conservative flux:

$$F^{k,*}(a, b) = \frac{1}{2} A^k(a + b).$$

(4.61)
4. Convergence analysis for systems of conservation laws

Following [14], we see that the above flux satisfies (3.9) for the entropy formulation (1.22). Observe that although the system (1.21) is linear, the streamline diffusion shock-capturing DG scheme (3.2) approximating it is nonlinear on account of the nonlinear shock-capturing term (3.20).

We have the following stability and convergence theorem for the scheme (3.2) with flux (4.61) approximating the linear symmetrisable system (1.21).

**Theorem 4.4.1.** Consider the linear symmetrisable system (1.21) with symmetriser $B$. Let $U_{\Delta x} = U(V_{\Delta x})$ be the approximate solutions generated by the streamline diffusion shock-capturing DG scheme (3.2) with numerical flux (3.9). Then, the approximate solutions satisfy the following energy bounds,

$$
\|U_{\Delta x}(\cdot, t^n)\|_{L^2(\Omega)} \leq C\|U_{\Delta x}(\cdot, t^0)\|_{L^2(\Omega)},
$$

(4.62)

for all discrete time levels $t^n$.

Furthermore, the approximate solutions $U_{\Delta x} \rightarrow U$ in $L^2(\Omega \times [0, T])$ and $U$ is the unique weak solution of the system (1.21) with the given initial data.

The proof is just a repetition of Theorem 4.1.1, replacing the $L^\infty$ bound with the $L^2$ bound at appropriate places.
5. Numerical experiments

We test the shock-capturing streamline diffusion DG scheme (3.2) for systems of conservation laws, in both one and two space dimensions. We first consider a linear system.

5.1. Wave equation

5.1.1. Exact form of equations, entropy formulation

The wave equation is given in (1.23). The equation is already in the symmetric form (i.e. the symmetriser can be chosen as the identity matrix) and the entropy is the energy: \( S(U) = \frac{1}{2} (h^2 + \sum_{k=1}^{d} u_k^2) \). Furthermore, the entropy-conservative flux in (3.3) is the average flux of the two neighbouring states. We use the following Rusanov type diffusion operator:

\[
D(a, b; \nu) = \max \{ \lambda_{\text{max}}(a; \nu), \lambda_{\text{max}}(b; \nu) \} U_V = cI
\]  

(5.1)

5.1.2. Smooth data: convergence rates

In the first numerical experiment, we consider the wave equation in one space dimension and set the wave speed to be \( c = 1 \). The domain is \([-1, 1]\) with periodic boundary conditions and the initial data is \( h = \sin(2\pi x) \), \( u = \sin(2\pi x) / 3 \).

As the initial data is smooth, the resulting solutions are smooth and we can compute the exact solution explicitly. The convergence rate for different schemes is provided in Figure 5.1. We compute the \( L^1 \) errors at the final time \( t = 1.5 \) normalized with the \( L^1 \) norm of the solutions. As shown in the figure, we test the scheme (3.2) with different spaces of basis functions, i.e. schemes with piecewise constant, linear, quadratic, and cubic functions. Furthermore, different parts of the quasilinear form in (3.2) are tested. We test with just the DG form (3.3), with the DG and streamline diffusion form as well as with the full scheme (3.2). The results in Figure 5.1 clearly show that the expected order of convergence is attained in this numerical experiment for all the choices of test functions as well as for all constituent parts of the quasilinear form. The best results in this case are obtained with the pure DG form (3.3). Adding streamline diffusion (3.17) results in some smearing and adding the shock-capturing form further smears the solution, even though the design order of accuracy is attained in all cases. The results indicate that the pure DG scheme will work quite well for smooth solutions.
5. Numerical experiments

Figure 5.1.: Convergence for wave equation for smooth initial data.
5.1.3. Discontinuous data

In the next test, we consider the previous configuration with the one dimensional wave equation, but discontinuous initial data,

\[ h(x, 0) = \begin{cases} 1.0, & \text{if } x < 0, \\ 0.0, & \text{if } x > 0. \end{cases} \quad u(x, 0) = \begin{cases} \frac{1}{3}, & \text{if } x < 0, \\ 0.0, & \text{if } x > 0. \end{cases} \]

![Graphs showing convergence for wave equation for discontinuous initial data.](image)

Figure 5.2.: Convergence for wave equation for discontinuous initial data.

In this case, the solutions are discontinuous. The rate of convergence for different schemes is shown in Figure 5.2. As expected, the design order of accuracy is no longer achieved and the maximum rate of convergence is approximately 1. However, increasing the polynomial
order does consistently result in lower errors for the same number of cells. Furthermore, time
snapshots of the solution with different degrees of polynomial basis functions and with dif-
ferent operators are shown in Figure 5.3. The figure clearly shows that there is considerable
improvement in resolution when the polynomial degree is increased from piecewise constant
to piecewise linear. There is a marginal improvement with further increase in the polynomial
degree. Furthermore, the pure DG scheme (3.3) is clearly oscillatory near the discontinuities.
The addition of streamline diffusion (3.17) damps the oscillations to some extent. However,
the best results are obtained with the full scheme (3.2). The shock-capturing term is clearly
needed to reduce oscillations considerably.

Figure 5.3.: Wave equation (h only), discontinuous initial data. Left: SD+SC, different p,
\( N_c = 80 \). Right: Different operators, \( p = 2 \), \( N_c = 80 \).

5.1.4. Two space dimensions

Next we test the wave equations in a two dimensional domain \([-1, 1]^2\) with Dirichlet bound-
ary conditions and with discontinuous initial data:

\[
h = \text{sgn}(\sin(2\pi \langle \alpha, x \rangle)), \quad u = \alpha_1 \text{sgn}(\sin(2\pi \langle \alpha, x \rangle)), \quad v = \alpha_2 \text{sgn}(\sin(2\pi \langle \alpha, x \rangle)),
\]

where \( \alpha = (\cos(\pi/3), \sin(\pi/3)) \). The numerical results with different schemes are provided
in Figure 5.4. The figure clearly shows that all the scheme resolve the solution quite well.
There is clear gain in accuracy when piecewise quadratic basis functions are used instead of
piecewise linears. As in the one dimensional tests, the pure DG and the streamline diffusion
DG schemes are oscillatory near discontinuities. However, these oscillations are significantly
reduced with the addition of the shock-capturing term (3.20).
5.1. Wave equation

(a) no SD/SC, \( p = 1 \)
(b) SD, \( p = 1 \)
(c) SD+SC, \( p = 1 \)
(d) no SD/SC, \( p = 2 \)
(e) SD, \( p = 2 \)
(f) SD+SC, \( p = 2 \)

(g) \( p = 1 \), slice
(h) \( p = 2 \), slice

Figure 5.4.: 2-D wave equation with discontinuous initial data and \( N_c = 840 \). The bottom panel shows cuts in \( \alpha \)-direction through \((0,0)\).
5. Numerical experiments

5.2. Euler equations

Next, we consider the nonlinear Euler equations of gas dynamics. They are given by equation (1.25).

5.2.1. Entropy-conservative fluxes, numerical diffusion operators

Here, we use the entropy-conservative flux for the Euler equations derived in [28], see also [14].

The numerical diffusion operator is of the Rusanov type:

\[
D(a, b; \nu) = \max \{\lambda_{\text{max}}(a; \nu), \lambda_{\text{max}}(b; \nu)\} \frac{(a + b)}{2} \\
= \max \left\{c(a) + \left|\sum_{k=1}^{d} u_k(a) \nu^k\right|, c(b) + \left|\sum_{k=1}^{d} u_k(b) \nu^k\right|\right\} \frac{(a + b)}{2}
\]  

(5.2)

5.2.2. One space dimension

As a first numerical experiment, we consider the Euler equations in one space dimension on the domain \([-5, 5]\) and consider the following initial data.

5.2.3. Sod shock tube

The Sod shock tube is a Riemann problem with left state \(\rho = 1, u = 0, p = 1\) and right state \(\rho = 0.125, u = 0, p = 0.1\). The results for the density and with piecewise quadratic basis functions are shown in Figure 5.5. The results clearly show that the pure DG and the streamline diffusion DG schemes resolve the exact solution (consisting of a rarefaction, a contact and a shock) sharply but with oscillations near both the shock wave as well as near the contact discontinuity. The addition of the shock-capturing terms damps these oscillations considerably with some smearing at the shock wave. However, the contact discontinuity is smeared considerably and the results are quite unsatisfactory.

5.2.4. Pressure scaling

The above results indicate that we need some artificial compression near contacts. To this end, we utilize the fact that the pressure does not jump across contact discontinuities and modify the scaling (3.20b) in the shock-capturing operator (3.20) in the following way:

\[
D_{n,K}^{\text{SC}} = \frac{D_n^p \left(\Delta t^n C_{\text{Res}}_{n,K} + (\Delta t^n)^{\frac{1}{2}} C_{\text{BRes}}_{n,K}\right)}{\sqrt{\int I_n \int K \left(\langle V_{t}^{\Delta x}, \bar{U}V_{t}^{\Delta x}\rangle + \sum_{k=1}^{d} \frac{\Delta x^2}{(\Delta t^n)^2} \langle V_{x_k}^{\Delta x}, \bar{U}V_{x_k}^{\Delta x}\rangle\right) dxdt + \epsilon}}
\]  

(5.3)
5.2. Euler equations

\[
D^p_{n,K} = \Delta x^2_K \frac{1}{\Delta t^n} |K| \int_{I^n} \int_K \sqrt{\sum_{k=1}^{d} p_{x_k}^2} dx dt
\]

Hence, jumps in pressure serve as indicators of contact discontinuities and the shock-capturing operator is switched off near the contacts. Note that introduction of the pressure scaling does not destroy the entropy stability of the scheme as the pressure term \(D^p_{n,K}\) appears as a positive constant. Corresponding results for the Sod shock tube with the new pressure scaling and with piecewise quadratic basis functions are shown in Figure 5.6. The figure clearly shows that the introduction of the new pressure scaling has significantly reduced the smearing at the contact while dampening oscillations near both the shock and the contact. Furthermore, the results for different polynomial degrees, shown in Figure 5.7 demonstrate that there is a significant gain in accuracy when piecewise linears are used in place of piecewise constants. Gains in accuracy with even higher polynomial degrees are more modest.

5.2.5. Lax shock tube

Next, we consider the Lax shock tube, which is the Riemann problem with left state \(\rho = 0.445, u = 0.698, p = 3.528\) and right state \(\rho = 0.5, u = 0, p = 0.571\) on the domain \([-5, 5]\) with Dirichlet boundary conditions. The results are shown in Figure 5.8. The figure shows

---

Figure 5.5.: Sod shock tube, density, different operators, \(p = 2, N_c = 80\)
5. Numerical experiments

Figure 5.6.: Sod shock tube, density, different operators, $p = 2$, $N_c = 80$.

Figure 5.7.: Sod shock tube, density, SD+SC(p), different $p$, $N_c = 80$.  

74
that pure DG as well as the streamline diffusion DG scheme are very oscillatory in the region between the shock and the contact discontinuity. The oscillations are decreased significantly when the shock-capturing term is added but this term smears the contact discontinuity considerably. The sharpness at the contact discontinuity is regained once the pressure scaled shock-capturing term is used, albeit with small oscillations in the intermediate state between the shock and the contact discontinuity. The behaviour of the schemes with increasing polynomial degree in the basis functions is very similar to the Sod shock tube experiment.

![Figure 5.8.](image)

**Figure 5.8.** Lax shock tube, density, $N_c = 80.$

### 5.3. Two space dimensions

#### 5.3.1. Vortex advection

We consider the advection of an Euler vortex (see [14] for the setup) in the domain $[0, 10]^2$, with Dirichlet boundary conditions and a vortex centred at $x_c = 5, y_c = 5$ with $r_c = 1$ as initial condition:

$$
u = 1 - (y - y_c)\phi(r), \quad v = 1 + (x - x_c)\phi(r), \quad \theta = 1 - \frac{\gamma - 1}{2\gamma} \phi(r)^2, \quad s = 0,$$

where $\theta = \frac{p}{\rho}, \ s = \log p - \gamma \log \rho, \ r = \sqrt{(x - x_c)^2 + (y - y_c)^2}, \ \phi(r) = \epsilon e^{\alpha \left(1 - \left(\frac{r}{r_c}\right)^2\right)}, \ \epsilon = \frac{5}{2\pi}$ and $\alpha = 12$.

The exact solution is explicitly known in this case ([14] and references therein) and is a vortex propagating along the diagonal. We compute convergence rates with different schemes and present the results in Figure 5.9. The expected convergence rates for the
5. Numerical experiments

![Graphs showing relative L1-error for different operators and polynomial degrees.](image)

Figure 5.9.: Vortex-Advection, different $p$, different operators.
5.3. Two space dimensions

piecewise constant, linear and quadratic basis functions are obtained. Furthermore, a time snapshot, shown in Figure 5.10, demonstrates that the full scheme (with the pressure scaling) resolves the propagating vortex without excessive smearing.

![Image](image.png)

Figure 5.10.: Vortex-Advection for the Euler equations, density \( p = 2 \), \( N_c = 828 \), \( t = 2 \).

5.3.2. Radial shock tube

We consider the Euler equations in the domain \([-1, 1]^2\) with Dirichlet boundary conditions. The initial data consists of a radial discontinuity at \( r = \sqrt{x^2 + y^2} = 0.4 \):

Outside this circle, the states are \( \rho = 0.125, p = 0.125 \) and inside the circle, the states \( \rho = 1, p = 1 \). Furthermore, \( u = 0, v = 0 \), everywhere in the domain initially.

The results with the scheme (3.2) are shown in Figure 5.11 and demonstrate that the shock-capturing (with pressure scaling) method works very well and the radial discontinuities are resolved sharply. One dimensional slices clearly show that adding shock capturing damps the oscillations generated with a pure DG or streamline diffusion DG scheme.

5.3.3. NACA 0012 aerofoil

As a last experiment, we consider an Euler flow around a NACA 0012 aerofoil. The aerofoil is placed along the \( x \) axis, ranging from \( x = 0 \) (head) to \( x = 1 \) (tail). Slip boundary conditions are used on the aerofoil. An artificial outer boundary is placed on a circle around \((2, 0)\) with radius 4, where the following free-stream values are prescribed: Mach number \( \text{Ma}_\infty \in \{0.5, 0.75, 1.3, 3\} \), pressure \( p_\infty = 8.5419 \), density \( \rho_\infty = 11.4452 \) and an angle of
5. Numerical experiments

Figure 5.11.: Radial shock tube for Euler equations, density, $p = 1$, $N_c = 13440$, $t = 0.2$. 

(c) Slice
5.3. Two space dimensions

attack of $4^\circ$. We compute and display the pressure coefficient
\[ c_p = \frac{(p - p_\infty)}{((1/2)\rho_\infty \| \mathbf{u}_\infty \|^2)}, \]
where $\mathbf{u}_\infty$ is the free-stream flow velocity. At $t = 0$, the flow is initialized by free-stream values. The equations are then solved up to $t = 3.5$; the time by which the steady state is approximately reached.

The results computed with the full scheme (with pressure scaling) using a polynomial degree of 1 are depicted in Figure 5.12. The shocks for the transsonic and supersonic problems are captured almost without oscillations albeit smeared over a few elements. Nevertheless, the flow is well approximated for all Mach numbers under consideration.

![Figure 5.12](image-url)

(a) $Ma_\infty = 0.5$

(b) $Ma_\infty = 0.75$

(c) $Ma_\infty = 1.3$

(d) $Ma_\infty = 3$

Figure 5.12.: NACA 0012 aerofoil flow, pressure coefficient $c_p$, $p = 1$, $N_c = 16704$, $t = 3.5$.  

79
6. The computational method

The approximate solution $V^{\Delta x}$ is only given implicitly by the variational formulation (3.2). In the following, we will sketch how it can be computed, i.e. specify the computational method.

6.1. Choice of basis functions

We start by specifying a suitable basis. The approximate solution $V^{\Delta x}$ is in $V_p$. For the computation we express it as a linear combination of basis functions $\phi_{Kki}^n \in V_p$:

$$V^{\Delta x} = \sum_{K,k,i,n} \hat{v}_{Kki}^n \phi_{Kki}^n,$$

where $0 \leq n \leq N - 1$, $K \in \mathcal{T}$, $1 \leq k \leq m$ and $1 \leq i \leq n_f$ and the coefficients $\hat{v}_{Kki}^n$ are the degrees of freedom. The indices indicate the support of the basis functions: We choose the basis functions $\phi_{Kki}^n$ such that they are nonzero only in one space-time element: $\phi_{Kki}^n$ is nonzero in $K \times I^n$. Additionally, only one component is nonzero—namely the $k$-th component for $\phi_{Kki}^n$—and we use the same scalar basis functions $\phi_{Kki}^n$ for all the components, i.e.

$$(\phi_{Kki})_l = \delta_{kl} \phi_{Kki}^n, \quad 1 \leq l \leq m. \quad (6.2)$$

The scalar basis functions have to span $\mathbb{P}_p(K \times I^n)$. We use monomials, except that they are shifted and scaled:

$$\phi_{Kki}^n|_{K \times I^n} = \left(\frac{t - t^{n+1}}{\Delta t^n}\right)^{p^{t,i}} \prod_{k=1}^d \left(\frac{x_k - \bar{x}_K}{\Delta x_K}\right)^{p^{x,k,i}} \quad (6.3)$$

where $\bar{x}_K$ is the centroid of cell $K$ and $p^{t,i}$, $p^{x,k,i}$, $k = 1, \ldots, d$ is the polynomial degree of the $i$-th scalar basis function in $t$, resp. $x_k$-direction. We use a maximal degree of $p$, hence:

$$p^{t,i} + \sum_{k=1}^d p^{x,k,i} \leq p. \quad (6.4)$$

In general there are $n_f = \binom{1+d+p}{p}$ scalar basis functions. As there are $N_c = |\mathcal{T}|$ cells, $N$ time steps, and $m$ components, the total number of degrees of freedom is $N_c N m n_f = N_c N m \binom{1+d+p}{p}$. 
6. The computational method

6.2. Nonlinear and linear solvers

Since the form $B$ is linear in the test function, fulfilling the variational form (3.2) for all test functions $W^\Delta x \in \mathcal{V}_p$ is equivalent to requiring that (3.2) is satisfied for all basis functions $\phi^n_{K,ki}$. Hence, we define $F^n_{K'lj} = B(V^\Delta x, \phi^n_{K'lj})$. The temporal upwind flux (3.5) allows us to march in time, and thus, we consider each time step separately. At the $n$-th time step, we have to solve the nonlinear system for $\hat{v}^n$

$$F^n_{K'lj}(\hat{v}^n) = 0, \quad K' \in \mathcal{T}, 1 \leq l \leq m, 1 \leq j \leq n_f,$$

where $\hat{v}^n$ is the vector of the degrees of freedom for the $n$-th time slab.

We use a damped Newton method [26] to solve this system: Starting from an initial guess $\hat{v}_0^n$, the approximation is iteratively improved by setting

$$\hat{v}_{i+1}^n = \hat{v}_i^n + \lambda_i \delta \hat{v}_i^n, \quad i = 0, 1, 2, \ldots$$

using the Newton correction $\delta \hat{v}_i^n$. The damping parameter $\lambda_i \leq 1$ is chosen using a line search. The most important step is to solve for the Newton correction $\delta \hat{v}^n$ at the current state $\hat{v}^n$ (for ease of notation we drop the iteration index $i$):

$$J^n(\hat{v}^n)\delta \hat{v}^n = -F^n(\hat{v}^n),$$

where the Jacobian is given by

$$J^n_{K'lj,Kki} = (F^n_{K'lj})_{\hat{v}_i^n K_{ki}}, \quad K, K' \in \mathcal{T}, 1 \leq k, l \leq m, 1 \leq i, j \leq n_f$$

The analytic expression of the Jacobian is used, i.e. the derivatives are computed analytically (by hand) and the corresponding matrix is assembled. The Jacobian has a block structure with blocks of size $mn_f \times mn_f$, which we shall denote by $J^n_{K',K}$. The block $J^n_{K',K}$ corresponds to the coupling of the degrees of freedom (dofs) of cell $K'$ to the dofs of the cell $K$. As only dofs of neighbouring cells are (directly) coupled, only the diagonal blocks $J^n_{K',K}$ and a few offdiagonal blocks are nonzero. The number of nonzero offdiagonal blocks for cell $K'$ is given by the number of neighbours of cell $K'$ and is therefore small. Figure 6.1 illustrates the nonzero structure of $J$ for a two-dimensional Euler case. Even though the sparsity pattern of $J$ might be symmetric, the matrix itself is in general nonsymmetric. Summarizing the facts: the Jacobian is large, sparse and nonsymmetric.

For one dimensional problems or for coarse meshes in two dimensions, a sparse LU decomposition can be used to solve (6.7). However, for larger computations (especially in parallel), this is no longer suitable, and therefore an iterative Krylov solver is used.

We use restarted GMRES (Generalized Minimal RESidual method) [44] as the iterative solver (restarted every 30 iterations). However for Krylov solvers, it is well-known that in order to increase the convergence speed, which in turn improves the efficiency, one needs to use a preconditioner. The next chapter deals with the details of preconditioning.
6.3. The mesh and mesh generation

Figure 6.1.: Nonzero structure of the Jacobian $J$ for a 2-D Euler flow with 207 cells and $p = 2$.

Using a preconditioned Newton-Krylov method is not the only option to compute approximate solutions of (3.2). In [49], a pseudo-timestepping approach with a multigrid acceleration technique is used to solve the constituting nonlinear equations of a slightly different space-time discontinuous Galerkin formulation. A more general fixed point iteration is also conceivable. However, we will not consider these approaches here.

6.3. The mesh and mesh generation

In one space dimension, we have (so far) only used regular meshes. In two spatial dimensions, we have used the distmesh mesh generator [43], which returns a Delauney triangulation of the domain $\Omega$ based on a signed distance function.

The DG method is fully implicit and thus, no CFL condition is required for stability, as manifested in the stability theorem 3.6.1. Nevertheless, for accuracy reasons the time step for the space-time DG method (3.2) is determined using the CFL conditions,

$$\Delta t^n \leq C_{\text{CFL}} \min_{K \in \mathcal{T}, x \in K} \frac{\Delta x_K}{\lambda_{\max}(U \Delta x(x, t^n))},$$

in one space dimension and

$$\Delta t^n \leq C_{\text{CFL}} \min_{K \in \mathcal{T}, x \in K} \frac{|K|}{\lambda_{\max}(U \Delta x(x, t^n))},$$

in two space dimensions. Here $\lambda_{\max}(U) = \max_{\nu} \lambda_{\max}(U; \nu)$ is the maximal wave speed in all directions and the constant $C_{\text{CFL}}$ is typically chosen to be $1/2$. 

83
6. The computational method

6.4. Quadrature rules

There are three types of integration that have to be performed, compare (3.3), (3.17), (3.20): integration over the whole space-time element $K \times I^n$, integration over the spatial elements $K$, and integration over the edges of the elements and the time $\partial_{KK'} \times I^n$. For the tensor product domains $K \times I^n$ and $\partial_{KK'} \times I^n$, we use tensor product integration formulas. Hence, we only need to further specify the quadrature rules for $K$, $\partial_{KK'}$ and $I^n$. For the integration over triangles $K$, Dunavant quadrature rules of order $2p + 2$ are applied [10], while for the integration over $\partial_{KK'}$ and $I^n$ one dimensional Gaussian quadrature rules of order $2p + 2$ are used.

6.5. Implementation

The method was first implemented in Matlab\(^1\). But the more advanced implementation was then done in C++ using the PETSc framework\(^2\). PETSc stands for 'Portable, Extensible Toolkit for Scientific Computation'. It provides parallel vector and matrix types as well as linear and nonlinear solvers for parallel computation using MPI. The use of this framework substantially facilitated the parallel implementation of the method.

The code is written in a generic way by the use of templates, and thus, can be extended easily to accommodate additional conservation laws. Furthermore, the code is written in most parts in a dimension independent way. Therefore, the dimension of the spatial domain can be easily changed by varying a parameter representing the dimension.

---

\(^1\)See http://www.mathworks.com
\(^2\)See http://www.mcs.anl.gov/petsc
7. Preconditioning

As described in the last chapter, the linearised system in each Newton iteration (6.7) is solved using a Krylov solver. It is well-known that in order to increase the computational efficiency the system needs to be preconditioned. In this chapter, we design and investigate various preconditioners. This chapter was published in [24] and uses a slightly different formulation of the method, see [24].

7.1. Block Jacobi preconditioner

7.1.1. Description

Given the block diagonal structure of the Jacobian matrix in (6.7), it is natural to consider block Jacobi preconditioners. The consequent preconditioner $D$ is a block diagonal matrix, specified as

$$D_{K',K} = \delta_{K',K} J_{K',K}^n,$$

so each block corresponds to an element $K$. Here, $\delta$ is the Kronecker delta symbol.

We need to invert the above matrix $D$ in order to apply the preconditioner. But this corresponds only to inverting each block $J_{K,K}^n$ separately. This "local" inversion can be performed by a standard LU decomposition with reasonable effort as the blocks are quite small (of size $mn_f \times mn_f$). Furthermore, the work per block does not grow with the number of cells $N_c$. To illustrate typical block sizes that arise here, we consider the Euler equations in two space dimensions with piecewise quadratic trial functions ($p = 2$). These equations consist of $m = 4$ components and the polynomial degree $p = 2$ leads in two dimensions to $n_f = \left(\frac{1+2+2}{2}\right) = 10$ degrees of freedom per component and cell. Therefore, in this case the block size is $40 \times 40$. This (and even smaller) sizes are representative of typical block sizes in defining the preconditioner and are amenable to a direct solve.

7.1.2. Fourier analysis

Given any diagonalisable matrix $A = XD X^{-1}$, the fundamental estimate for the performance of GMRES was provided by Saad and Schultz [44]:

$$\| r_{m+1} \| \leq \kappa(X) \min_{p \in P_m, p(0) = 1} \max_{\lambda \in \sigma} |p(\lambda)| \| r_0 \|.$$  \hfill (7.1)

Here $\kappa(X) = \|X\| \|X^{-1}\|$ denotes the condition number, $P_m$ the space of polynomials of degree at most $m$, $\sigma$ denotes the spectrum of $A$ and $r_m$ is the residual in the $m$-th step of GMRES.
7. Preconditioning

Given this basic estimate, we aim to compute the eigenvalues of the unpreconditioned matrix $J$ in (6.7) as well as the preconditioned matrix $D^{-1}J$ in order to ascertain if the preconditioner did improve the conditioning. We use the following simplifications in order to make this problem tractable,

- We require periodic boundary conditions as we will use a Fourier transformation of the system. Uniform grids are also assumed for simplification.
- As it is not possible (algebraically) to perform the analysis for a generic linearised system, we will analyse the linear advection equation as a model problem.

**Linear advection in one space dimension.**

We consider the simplest example of a conservation law, the linear advection equation:

$$u_t + au_x = 0$$

(7.2)

in the one-dimensional domain $[-1, 1]$ with periodic boundary conditions. For simplicity we assume $a > 0$ and discretise (7.2) on a regular grid; numbered (here) $0, \ldots, N_c - 1$ and choose a Rusanov diffusion operator in (3.12) resulting in an upwind spatial numerical flux. Neglecting the shock-capturing operator, the resulting Jacobian in (6.7) for this problem has the structure,

$$\frac{1}{\Delta x} J = \begin{pmatrix} A & B \\ B & A \\ \ddots & \ddots \\ B & A \end{pmatrix} = I \otimes A + S \otimes B$$

(7.3)

with fixed matrices $A, B \in \mathbb{R}^{n_f \times n_f}$ (depending on $p$ and the time step). Here, $\otimes$ denotes the Kronecker product and the matrix $S$ is defined as,

$$(S)_{j,k} = \begin{cases} 1, & j = k + 1 \mod N_c \\ 0, & \text{otherwise} \end{cases}$$

(7.4)

Let $F$ be the Fourier matrix, i.e.,

$$F = \frac{1}{\sqrt{N_c}} (\omega^{jk})_{j,k=0}^{N_c-1}, \quad \omega = e^{-2\pi i/N_c}$$

(7.5)

Since the matrix $S$ is a circulant matrix, it can diagonalised by the Fourier matrix as

$$S = F \text{diag}(d) F^H.$$  

(7.6)
7.1. Block Jacobi preconditioner

With $F^H$ denoting the Hermitian of $F$ and

\[
d = \sqrt{N_c} F((S)_{0,0},\ldots,(S)_{0,N_c-1})^T
\]

\[
= \sqrt{N_c} F(0,\ldots,0,1)^T
\]

\[
= (\omega^0(N_c-1),\ldots,\omega^{(N_c-1)(N_c-1)})^T
\]

Therefore,

\[
\frac{1}{\Delta x} J \sim (F \otimes I)^H (I \otimes A + S \otimes B)(F \otimes I)
\]

\[
= (F^H IF) \otimes (IAI) + (F^H SF) \otimes (IBI)
\]

\[
= I \otimes A + \text{diag}(d) \otimes B
\]

\[
\sim \begin{pmatrix}
A + d_0 B \\
A + d_1 B \\
\vdots
\end{pmatrix}
\]

(7.8)

where $\sim$ denotes similarity.

From the above calculation, we have established that $\frac{1}{\Delta x} J$ is similar to a blockdiagonal matrix with blocks $A + d_i B$. Furthermore, observe that the number of cells $N_c$ only enters the expression through the factors $d_i$. However, as $|d_i| = 1$, $\forall i$, the eigenvalues can only lie on curves specified by the eigenvalues of $A + e^{i\phi} B$, parametrized by $\phi$. Hence, refining the mesh (increasing $N_c$) only leads to increase in the denseness of the eigenvalues on these curves. Furthermore, assuming that none of these level curves cross the origin, the closeness of the eigenvalues to zero is almost independent of the mesh size. These findings are validated in figures 7.1 and 7.2. Furthermore as demonstrated in Figure 7.2, the increase of polynomial degree $p$ leads to an increase in the number of degrees of freedom per element (block size) and thus increases the number of curves on which the eigenvalues lie.

Next, we consider the role of the preconditioner,

\[
\frac{1}{\Delta x} D = I \otimes A.
\]

(7.9)

The resulting preconditioned system is

\[
D^{-1} J = (I \otimes A)^{-1} (I \otimes A + S \otimes B)
\]

\[
= I \otimes I + S \otimes A^{-1} B
\]

\[
\sim (F \otimes I)^H (I \otimes I + S \otimes A^{-1} B)(F \otimes I)
\]

\[
= (F^H IF) \otimes (III) + (F^H SF) \otimes (IA^{-1} BI)
\]

\[
= I \otimes I + \text{diag}(d) \otimes A^{-1} B
\]

\[
\sim \begin{pmatrix}
I + d_0 A^{-1} B \\
I + d_1 A^{-1} B \\
\vdots
\end{pmatrix}
\]

(7.10)
Figure 7.1.: Eigenvalues of $\frac{1}{\Delta x} J$ and $D^{-1}J$ for the 1-D linear advection equation with different number of cells for polynomial degree $p = 2$. 

(a) $N_c = 40$

(b) $N_c = 40$, closeup

(c) $N_c = 160$

(d) $N_c = 160$, closeup

7. Preconditioning
7.1. Block Jacobi preconditioner

Let \( \tilde{d} \) denote the vector of eigenvalues of \( A^{-1}B \). Then the eigenvalues of \( I + d_i A^{-1} B \) (and consequently of \( D^{-1}J \)) are given by \( 1 + d_i \tilde{d} \). Since \( |d_i| = 1 \), the eigenvalues lie on circles \( 1 + e^{i\phi} \tilde{d}_j \), parametrized by \( \phi \) (see Figure 7.1 and 7.2). Hence, if the eigenvalues \( \tilde{d}_j \) of \( A^{-1}B \) are small, then the eigenvalues of the preconditioned system are clustered around one (therefore bounded away from the origin). This will clearly improve the spectral contribution to the performance estimate (7.1) of GMRES.

However, it is clear from the estimate (7.1) that it is not enough to consider the distribution of eigenvalues. We also need to study the conditioning of the eigenvector matrix \( X \) in order to ascertain the performance of GMRES. To this end, we denote \( \tilde{V} \) as the matrix corresponding to an eigenbasis of \( A^{-1}B \). (7.10) implies that \( X = F \otimes V \) diagonalises \( D^{-1}J \). Hence, its condition number is given by,

\[
\kappa(X) = \kappa(F \otimes V) \\
= \kappa((F \otimes I)(I \otimes V)) \\
\leq \|F \otimes I\| \|I \otimes V\| \|(F \otimes I)^{-1}\| \|(I \otimes V)^{-1}\| \leq \kappa(V),
\]

(7.11)
as the Fourier matrix \( F \) is unitary. Therefore, the condition number of \( X \) is bounded independently of the number of cells \( N_c \).

Combining the above results with (7.1), we establish that the number of iterations of the Jacobi preconditioned GMRES is independent of mesh size. Furthermore, the smallness of the eigenvalues \( \tilde{d} \) will lead to a clustering of the eigenvalues of the preconditioned matrix, away from the origin (as seen in figures 7.1 and 7.2). Hence, this analysis demonstrates that the Jacobi preconditioner enhances the performance of GMRES, at least for the linear advection equation.

Role of the shock-capturing term

The inherent non-linearity of the shock-capturing term in (3.20) implies that the variational formulation (3.2) is nonlinear, even when the underlying PDE is a linear conservation law such as the advection equation. Consequently, the Jacobian matrix \( J \) in (6.7) is solution dependent. In order to analyse the role played by the shock-capturing term, we consider the linear advection equation (7.2) with periodic initial data (7.16) and display the eigenstructure of the Jacobian at a typical time step in Figure 7.3. The figure compares the eigenstructure of the Jacobian as well as the preconditioned Jacobian, in the absence and in the presence of the shock-capturing term. As seen in the figure, the shock-capturing term has a very minor effect on the eigenvalue distribution in both the non-preconditioned as well as preconditioned matrices. Its main effect is a distortion in the shape of some level curves.
7. Preconditioning

Figure 7.2.: Eigenvalues of $\frac{1}{\Delta x} J$ and $D^{-1}J$ for the 1-D linear advection for different degrees and for $N_c = 160$.

Figure 7.3.: Eigenvalues of $\frac{1}{\Delta x} J$ and $D^{-1}J$ with or without the shock-capturing term (SC).
7.1. Block Jacobi preconditioner

**Linear advection in two space dimensions**

As prototype for multi-dimensional problems, we consider the two-dimensional version of the linear advection equation:

\[ u_t + a^x u_x + a^y u_y = 0 \quad (7.12) \]

in the domain \([-1, 1]^2\) with periodic boundary conditions. We discretise this equation with the space-time DG method (3.2) (neglecting the shock-capturing term) on a uniform rectangular grid with mesh size \(\Delta x\). Following the previous section, we can write the Jacobian as,

\[
\frac{1}{\Delta x^2} J = I \otimes I \otimes A + S^x \otimes I \otimes B + I \otimes S^y \otimes C \quad (7.13)
\]

with fixed matrices \(A, B, C \in \mathbb{R}^{n_f \times n_f \times n_f}\) (depending on polynomial degree \(p\) and the time step) and the matrices \(S^x, S^y\), analogously defined as before. Fourier transforming in both \(x\) and \(y\) directions, one can show that \(\frac{1}{\Delta x^2} J\) is blockdiagonal with blocks of the form

\[
A + e^{i\phi_x} B + e^{i\phi_y} C \quad (7.14)
\]

parametrized by \(\phi^x\) and \(\phi^y\). The Jacobi preconditioned matrix is similar to a blockdiagonal matrix with blocks of the form

\[
I + e^{i\phi_x} A^{-1} B + e^{i\phi_y} A^{-1} C. \quad (7.15)
\]

The eigenstructure in 2-D is more complicated than in 1-D as, in general, \(A^{-1} B\) and \(A^{-1} C\) are not simultaneously diagonalisable.

Hence, we will limit ourselves to numerical results in this case. In Figure 7.4 the eigenvalues of \(\frac{1}{\Delta x^2} J\) and \(D^{-1} J\) are shown for a regular triangular mesh. As in the 1-D case the general structures remain the same, under mesh refinement. The distribution of eigenvalues is denser but is confined to a bounded region of the complex plane. Furthermore, the preconditioner clusters the eigenvalues around 1 indicating that the preconditioner does enhance the performance of GMRES in this case.

**7.1.3. Numerical results**

The Fourier analysis of the previous section indicates that

- The performance of the Jacobi preconditioned GMRES will be independent of the mesh size.

- The performance of the preconditioned GMRES will be significantly better than its unpreconditioned version as the eigenvalues of the preconditioned Jacobian are clustered around 1. Hence, they are bounded away from zero, unlike that of the unpreconditioned version, see Figure 7.1.
7. Preconditioning

Figure 7.4.: Eigenvalues of $\frac{1}{\Delta x^2} J$ and $D^{-1} J$ for the linear advection equation in 2-D for $p = 2$ for different number of cells.
7.1. Block Jacobi preconditioner

The analysis was restricted to the model case of the one-dimensional and two-dimensional advection equations. We will evaluate the performance of the Jacobi preconditioner for this model problem as well as the more complicated Euler equations of gas dynamics.

**Linear advection equation in one space dimension**

We consider the linear advection equation (7.2) in the domain $[-1, 1]$ with periodic boundary conditions. The initial condition is

$$u(x, 0) = \sin(2\pi x). \quad (7.16)$$

The quadratic entropy is used. Hence, the conservative and entropy variables coincide. Furthermore, a Rusanov type numerical diffusion operator in (3.12) is used. In each Newton step the linear system (6.7) is solved up to a relative tolerance of $10^{-4}$. The average number of Krylov iterations per Newton iteration is depicted in Figure 7.5. We compare the unpreconditioned system with the system preconditioned by block Jacobi. As predicted from the analysis, the number of iterations for both the unpreconditioned as well as preconditioned systems is independent of mesh size. However and consistent with the analysis of the previous section, the number of iterations is significantly reduced from around 30 ($p = 1$) or 70 ($p = 2$) in the unpreconditioned case to approximately 5 for the block Jacobi preconditioned system. This demonstrates a significant gain in efficiency due to the block Jacobi preconditioner.

![Figure 7.5](image-url) (a) $p = 1$

![Figure 7.5](image-url) (b) $p = 2$

Figure 7.5.: Number of Krylov iteration per Newton iteration for the one-dimensional linear advection equation.

The approximate solution computed with piecewise linear and piecewise quadratic basis functions (and with the preconditioned system) is plotted in Figure 7.6. The figure shows
7. Preconditioning

that the exact smooth solution is approximated very well with the shock-capturing space-
time DG method. The quality of approximation improves with increasing mesh size or
increasing polynomial degree.

Figure 7.6.: 1-D linear advection, different number of cells.

Linear advection in two space dimensions

We consider the linear advection equation (7.12) in the domain \([-1, 1]^2\) with constant ve-
locities \(a^x = 1\) and \(a^y = 1/2\). The initial condition is a bump at \((-0.2, -0.2)\):

\[
u(x, 0) = \frac{1}{2} e^{-16((x+0.2)^2+(y+0.2)^2)}(1 - e^{-16(x+1)^2})(1 - e^{-16(x-1)^2})(1 - e^{-16(y+1)^2})(1 - e^{-16(y-1)^2}).
\]

(7.17)

At the inflow boundaries, homogeneous Dirichlet boundary conditions are used.

As in the previous case, we consider the quadratic entropy and a Rusanov type numer-
cal diffusion operator. The average number of Krylov iterations per Newton iteration is
displayed in Figure 7.7. The unpreconditioned version works only for \(p = 1\), where slightly
more than 50 steps are needed. For \(p = 2\) the linear solver does not always converge within
the maximum number of 2000 iterations. With the block Jacobi preconditioner, one needs
only about 7 iterations to converge (both for \(p = 1\) and \(p = 2\)). Furthermore, this number
is independent of the mesh size. Thus, this example clearly illustrates the enhancement in
efficiency on account of the preconditioner. The approximate solution, computed with pre-
conditioned space-time DG method (3.2) is shown in Figure 7.8 and shows that the method
is able to approximate the advected solution quite well, both for piecewise linear as well as
piecewise quadratic basis functions.
7.1. Block Jacobi preconditioner

Figure 7.7.: Number of Krylov iteration per Newton iteration for the linear advection equation in two dimensions.

Figure 7.8.: Advection equation, smooth initial data, $N_c = 13440$. 

(a) $p = 1$

(b) $p = 2$
7. Preconditioning

Sod shock tube

We consider now the Euler equations of gas dynamics (1.25). To begin with, we consider the Sod shock tube, which is a one dimensional Riemann problem, see Section 5.2.3. We use the entropy-conservative flux for the Euler equations derived in [28], see also [14], together with a Rusanov type of diffusion. Again, we are interested in the average number of Krylov iterations per Newton iteration, see Figure 7.9. The case without preconditioner is not included as it needs several hundred iterations or fails to converge to the desired tolerance within 2000 iterations. However, with the block Jacobi preconditioner, the specified relative tolerance of $10^{-4}$ is reached within only about 7 ($p = 1$) or 8 ($p = 2$) iterations. Furthermore, the number of iterations is totally independent of the number of mesh points.

The computed density for different mesh resolutions and different polynomial orders is displayed in Figure 7.10. The space-time DG method approximates the solution, consisting of a shock wave, a contact discontinuity and a rarefaction wave, quite well. There is a significant gain in accuracy when the number of mesh points is increased. The subsequent gain in accuracy on increasing the polynomial degree from one to two is rather moderate.

Two dimensional Euler vortex

Next, we consider the advection of an Euler vortex, see Section 5.3.1. The iteration numbers are shown in Figure 7.11. Also in this case, the number of Krylov iterations is quite small for the block Jacobi preconditioned system: about 7 ($p = 1$) or 9 ($p = 2$) iterations per Newton iteration. The number of iterations is also independent of grid size. The computed
7.1. Block Jacobi preconditioner

Figure 7.10.: Sod shock tube, different number of cells.

Figure 7.11.: Number of Krylov iteration per Newton iteration for the vortex advection problem.
7. Preconditioning

![Figure 7.12.: Vortex advection for the Euler equations, density, $N_c = 3312$, $t = 2$.](image)

(a) $p = 1$

(b) $p = 2$

solutions are shown in Figure 7.12 and demonstrate that the shock-capturing DG method approximates the underlying exact solution (vortex advected along the diagonal) quite well. Summarizing the results of the above numerical experiments, we observe that the block Jacobi preconditioner does significantly reduce the number of Krylov iterations per Newton step, compared to the unpreconditioned version. Combined with the observed independence of the number of the iterations with respect to mesh size, the preconditioner increases the efficiency of the shock-capturing space-time DG methods by orders of magnitude.

7.1.4. CFL dependence

The time step for the space-time DG method (3.2) is determined using the CFL number by the formulas (6.9) and (6.10). But one of the attractive features of the shock-capturing space-time DG method is the fact that the method remains entropy-stable as well as convergent to the entropy measure-valued solutions for any finite value of the CFL constant $C_{CFL}$. In particular, large CFL constants (large time steps) are allowed. This possibility of setting large time steps enables the space-time DG method to be efficient in resolving problems involving convergence to steady state (such as in aerodynamic calculations) or problems with multiple time scales (such as in radiation hydrodynamics). See Chapter 8 for further examples. Hence, it would be useful to design preconditioners that can work in the large time step (CFL number) regime.
7.1. Block Jacobi preconditioner

Fourier analysis

We perform the Fourier analysis for the linear advection case, as in the previous section but allow for a dependence on the CFL number. The eigenvalues of the unpreconditioned as well as block Jacobi preconditioned matrices are shown in Figure 7.13). The results show that some of the eigenvalues of $A^{-1}B$ get larger when the time step is increased (for the same mesh size). Consequently, the eigenvalues of the preconditioned matrix are no longer clustered around one and come close to zero. This will lead to a deterioration in the performance of the preconditioner and will increase the number of iterations per Newton step.

![Eigenvalues](a) $CFL = 0.5$, closeup (b) $CFL = 5$, closeup (c) $CFL = 10$, closeup

Figure 7.13.: Eigenvalues of $\frac{1}{\Delta x} J$ and $D^{-1} J$ for the one-dimensional linear advection equation with different CFL values for $N_c = 160$ and $p = 2$.

Numerical results

This lack of robustness of the block Jacobi preconditioner to increasing the CFL number is examined through a numerical experiment for the one-dimensional linear advection equation. The same set-up as in the previous subsection is used and the average number of iterations per Newton step, for both the unpreconditioned and preconditioned systems is shown in Figure 7.14. As seen from the figure and consistent with the Fourier analysis, the number of iterations increases significantly with increasing CFL number. In particular, the block Jacobi preconditioner is quite well-behaved till a moderate CFL number of around 10, but the number of iterations increases significantly with further increase in the CFL number.

The following heuristic explains this observed dependence of the block Jacobi preconditioner with the CFL number. The exact solution of the conservation law has a finite speed of propagation, due to the hyperbolicity of the system. Even though the implicit formulation of the space-time DG method couples all the degrees of freedom (and therefore allows an infinite speed of propagation), one can still expect that most effects are local as they eventually approximate the exact solution. Block Jacobi is acting on the degrees of freedom of each cell separately, i.e. it is a local solver. So for small CFL numbers one expects it
7. Preconditioning

Figure 7.14.: Number of Krylov iteration per Newton iteration for the linear advection equation in dependence of the CFL number ($N_c = 640$).

to be a good preconditioner as the time step is too small for information to propagate significant distances in the domain. However, the higher the CFL number, the further is the information allowed to travel, i.e. it will cross more and more cells in just one time step and the effects become increasingly non-local. Therefore block Jacobi, as it is a local solver, can not resolve these waves and will fail.

7.2. Block Gauss-Seidel preconditioners

The above discussion suggests that a non-local preconditioner might be more robust with respect to increasing the CFL number. In this context, we consider preconditioners of the Gauss-Seidel type. To this end, we introduce an ordering of the cells and thus of the Jacobian $J$ in (6.7) and define the block lower triangular part as

$$L_{K', K} = \mathbb{1}_{K' > K} J^n_{K', K}$$

and the upper triangular part as

$$U_{K', K} = \mathbb{1}_{K' < K} J^n_{K', K}.$$

A simple block Gauss-Seidel preconditioner results from a forward sweep, i.e. using $D + L$ as a preconditioner. Applying the preconditioner to the system $Jx = b$ results for the forward version in

$$x = (D + L)^{-1} b.$$  (7.18)
7.2. Block Gauss-Seidel preconditioners

We can also consider the symmetric version, where an additional backwards sweep using $D + U$ is performed, resulting in,

$$
\begin{align*}
\tilde{x} &= (D + L)^{-1}b, \\
x &= \tilde{x} + (D + U)^{-1}(b - J\tilde{x}).
\end{align*}
$$

(7.19)

The inversion of $D + L$ resp. $D + U$ is performed cell by cell in forward resp. backward direction. For each cell, this corresponds to solving an $mn_f \times mn_f$ system as in the block Jacobi case, but the important difference is that this inversion is performed sequentially.

7.2.1. Analysis of the spectrum

As before, we consider the one-dimensional linear advection equation with periodic boundary conditions. The block GS preconditioned matrix is given by

$$
(D + L)^{-1}J = (D + L)^{-1}(D + L + U) = I + (D + L)^{-1}U.
$$

(7.20)

The matrix $(D+L)^{-1}U$ is strictly block upper triangular (there are only two nonzero blocks) given the structure of $J$ in (7.3). Hence, all eigenvalues are 1 and the block GS is a perfect preconditioner in this case. Furthermore, this fact is independent of the CFL condition. However, this is limited to the advection equation; for more general conservation laws the performance of block GS might be less striking.

7.2.2. Numerical experiments

One-dimensional linear advection

We consider the one-dimensional linear advection equation (7.2) with initial data (7.16). The numerical set-up is exactly as before. The cells are ordered in the advection direction, and we use the forward version of the block GS preconditioner. The iteration numbers are shown in Figure 7.15. Only two iterations are required for the block Gauss-Seidel preconditioned system to converge to desired tolerance. This number is independent of the grid size.

However, the real test of the block Gauss-Seidel preconditioner lies in increasing the CFL number. The corresponding iteration numbers for different preconditioners are shown in Figure 7.16. The number of iterations with respect to the block Gauss-Seidel preconditioner remains at 2 for any value of the CFL number. This is in marked contrast to the significant increase in the number of the iterations for the unpreconditioned and the block Jacobi preconditioned systems. The excellent preconditioning properties of block GS for the advection equation was also observed earlier on in a slightly different context, see [19, 21]. However, it depends on the ordering of the degrees of freedom. There is a single advection direction (a single flow of information) and the cells must be ordered in this downwind direction for the block GS preconditioner to fully unleash its power.
7. Preconditioning

Figure 7.15.: Number of Krylov iteration per Newton iteration for the linear advection equation.

Figure 7.16.: Number of Krylov iteration per Newton iteration for the linear advection equation in dependence of the CFL number ($N_c = 640$).
2-D Euler: Vortex advection

We consider the advection of a vortex for the two-dimensional Euler equations. The numerical set-up is exactly the same as in Section 7.1.3. The cells are ordered in the direction of the vortex advection and we use the symmetric version of block Gauss-Seidel (GS). Figure 7.17 shows a comparison of the Krylov iteration numbers. Using block GS, the iteration numbers drop to about 3 ($p = 1$) or 4 ($p = 2$), which is about one half of the numbers for the block Jacobi preconditioned system. However, the symmetric block GS is twice as expensive (per step) as the block Jacobi. Hence, there is no gain in performance with the block GS over the block Jacobi at this low CFL ($C_{CFL} = 0.5$).

The number of iterations with respect to increasing CFL number are shown in Figure 7.18. Here, both for block Jacobi and block GS, the iteration numbers grow with higher CFL. The increase for block Jacobi is bigger than for block GS. Recalling that the symmetric block GS iteration is about twice as costly as a block Jacobi iteration, the block GS is still better by a factor of about 2 for $CFL = 50$ as it needs only about 28 instead of 126 ($p = 1$) or only about 26 instead of 107 Krylov iterations ($p = 2$). Thus, even if information is no longer flowing in just one direction (the advection direction) but in all directions, the performance gain with block GS over block Jacobi is less striking but still observable.

NACA 0012 aerofoil

As the final numerical experiment, we consider a $Ma_{\infty} = 0.75$ Euler flow around a NACA 0012 aerofoil, see Section 5.3.3. An unstructured mesh (consisting of triangles) is generated.
7. Preconditioning

Figure 7.18.: Number of Krylov iteration per Newton iteration for the vortex advection problem in dependence of the CFL number ($N_c = 3312$).

around the aerofoil. This mesh is finer near the head of the aerofoil than near the tail. As a further modification the shock-capturing operator is modified with a pressure scaling as described in (5.3). Snapshots of the computed pressure coefficient $c_p$ are shown in Figure 7.19.

As the steady state is the object of interest in the current computation, both results are obtained with a very large time step resulting from $C_{CFL} = 500$, in order to accelerate convergence to the steady state. The results show that the transonic flow is resolved quite well with the shock-capturing space-time DG method. The shocks are resolved sharply and without strong oscillations. Furthermore, the smooth regions are also approximated well. The resolution is noticeably better with piecewise quadratic basis functions.

As we are also interested in the performance of the preconditioners developed herein, we plot the average number of Krylov iterations per Newton step in this problem in Figure 7.20. In addition to the block Jacobi and the block GS (forward sweep), a further variant of the block GS, the multi-block GS, is also considered. This preconditioner consists of four sweeps; two forward sweeps, one based on the standard ordering and another on an alternate ordering, coupled with two backward sweeps, performed with the directions reversed. The cells are ordered in the free-stream direction and the alternative ordering is in the orthogonal direction. The results from Figure 7.20 show that the number of iterations is quite high. This is not unexpected, given the very high CFL number that is employed. Nevertheless, all the preconditioners lead to a performance that is very mildly dependent on the mesh size. The block Jacobi preconditioner has the worst performance, in terms of number of iterations as well as their growth with mesh size whereas the block GS and the multi-block GS lead to only a moderate growth. Given the fact that the multi-block GS is more expensive per step, both the block GS preconditioners show a similar performance in this realistic test case.
7.2. Block Gauss-Seidel preconditioners

Figure 7.19.: NACA 0012 aerofoil Euler equations, pressure coefficient $c_p$, $N_c = 2809$.

Figure 7.20.: Number of Krylov iteration per Newton iteration for the Euler NACA0012 aerofoil flow ($CFL = 500$).
8. All-speed flows

Often problems modelled by systems of conservation laws (1.3) consist of multiple time scales, for instance in radiation hydrodynamics or in flows near the incompressible (low Mach number) limit. Typically, resolving all the time scales is not the main interest of the computation. Furthermore, the computational costs associated with such a high resolution would be enormous. Therefore, one concentrates on resolving the time scales (the slow ones) of interest, while ensuring that the lack of resolving the fast time scales does not destroy the stability of the scheme. The DG scheme (3.2) is unconditionally stable. Hence, the time steps can be quite large and are not restricted by any stability condition, as is true for explicit methods. This scheme is therefore well suited to approximate problems involving multiple time scales. In this chapter, we investigate this in more detail. It is published in [23].

In the context of this chapter, we need to pay more attention to the choice of the time step. In general, we use the CFL condition to determine the time step size. More precisely, in one spatial dimension the condition is

$$\Delta t^{n+1} \leq C_{CFL} \min_{K \in T, x \in K} \frac{\Delta x_K}{\lambda_{\max}(U \Delta x(x, t^n))},$$

while in two spatial dimension we use the following condition

$$\Delta t^{n+1} \leq C_{CFL} \min_{K \in T, x \in K} \frac{|K|}{\lambda_{\max}(U \Delta x(x, t^n))},$$

where \(\lambda_{\max}(U) = \max_{\nu} \lambda_{\max}(U; \nu)\) is the maximal wave speed in all directions.

The use of \(\lambda_{\max}(U; \nu)\) in the denominator of (8.1) and (8.2) ensures that all waves speeds are well resolved (for \(C_{CFL} \leq 1\)). To resolve only the time scales of interest, we replace \(\lambda_{\max}\) in the denominator by \(\lambda_{\text{slow}}\), which we define as the maximal wave speed of the slow waves. We call the corresponding constant the reduced CFL number \(CFL_{\text{red}}\) and often chose it to be \(1/2\).

8.1. Two-speed advection.

As a first example to illustrate this strategy, we consider the following linear system of conservation laws:

$$U_t + F(U)_x = 0,$$
8. All-speed flows

with

\[
U = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad F(U) = \frac{1}{2} \left( (a + b)u_1 + (b - a)u_2 \right) \\
\left( b - a \right)u_1 + (a + b)u_2
\]

(8.4)

The system has the wave speeds \(a\) and \(b\), and we assume that \(a\) is much smaller than \(b\). Since the system is symmetric, it is endowed with the entropy framework described in (1.22). Thus, the entropy \(S(U)\) is the quadratic entropy, and the entropy variables \(V\) coincide with the conservative variables \(U\). As we are mainly interested in approximating the wave with the slow speed \(a\), the reduced CFL number is based only on \(\lambda_{\text{max}}^\text{slow}(U) = a\).

We consider the Riemann problem

\[
u_1(x, 0) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases}, \quad u_2(x, 0) = 0
\]

(8.5)
in the domain \((-1, 1)\) with Dirichlet boundary conditions on the left boundary and Neumann boundaries on the right boundary. The slow wave speed is \(a = 1\), while the fast wave speed \(b\) is varied between 1, 10, and 100. The problem is solved from \(t = 0\) to \(T = 0.5\). This implies that the fast waves originating from the initial jump leave the domain before the end of simulation (except for \(b = 1\)).

The entropy-conservative flux is the central flux (4.61) and as the numerical diffusion operator, either a Rusanov (3.14) or a Roe (3.13) type of operator is used. For the Rusanov diffusion operator, the maximal wave speed is chosen as \(a\), in order not to smear out the discontinuities too much. Note that the Roe type of diffusion leads to an upwind flux for this example.

In Figure 8.1, the approximate solution is depicted for piecewise quadratic element using the Roe type of diffusion on different meshes. The fast wave speed is \(b = 100\) and therefore, the effective CFL is 50. Nevertheless, the discontinuity is captured quite nicely, without excessive smearing. There are some spurious oscillations, but they are bounded in amplitude and confined to more or less 2-3 elements around the discontinuity. The results for the simulations with \(b = 1\) and \(b = 10\) or with a Rusanov diffusion operator are similar and thus not shown.

The number of Newton iterations is quite small. For this problem, it is typically around three iterations. The convergence of the linear solver in the Newton iteration is shown in Figure 8.2 for Roe diffusion. Two types of preconditioners are compared for different values of \(b\). We use a block Jacobi preconditioner (BlockJac) and a block Gauss-Seidel preconditioner (BlockGS), where the blocks correspond to elements, cf. Chapter 7. For the block Jacobi preconditioner, the number of iterations increases slightly for a higher value of \(b\). With \(b = 1\), the number of iterations is almost constant at 5 \((p = 1)\) or 6 \((p = 2)\) iterations, while the grid is refined. For higher wave speeds \(b\) (and therefore higher effective CFL numbers), there are pre-asymptotic increases in the number of iterations. But for finer grids, the number of iterations stabilises or is even decreasing. The total number of iterations is quite small (with at most about 17 iterations in one case).
8.1. Two-speed advection.

Figure 8.1.: Approximate solution for the two speed advection problem using polynomial degree $p = 2$ and a fast wave speed of $b = 100$.

Figure 8.2.: Number of Krylov iterations for the two speed advection problem using a Roe type of diffusion with varying fast wave speed $b$ and preconditioners.
8. All-speed flows

The block Gauss-Seidel preconditioner is even more optimal as a preconditioner because it acts as a direct solver and the number of iterations is therefore one in all cases (the Krylov solver would not be needed here). This is a consequence of the fact that the numerical flux reduces to the upwind flux and the cells are ordered in downwind direction.

![Graph](image)

Figure 8.3.: Number of Krylov iterations for the two speed advection problem using the block Jacobi preconditioner with varying fast wave speed $b$ and diffusive fluxes.

In Figure 8.3, we compare the performance of the linear solver for the two different diffusive fluxes: the Roe type (3.13) of diffusion and the Rusanov (3.14) type of diffusion (based on the slow wave speed). As already noted in Figure 8.2, the number of iterations for the Roe type of diffusion is quite small and bounded with respect to grid refinement. For $b = 1$, the Rusanov and Roe diffusion coincide. But for higher values of $b$, the number of iterations is much higher for the Rusanov type of diffusion: about 20 instead of 10 for $b = 10$ and about 130 instead of 13 iterations for $b = 100$ and $p = 1$. For $p = 2$, it is more extreme, and for $b = 100$, the linear solver even fails to converge for fine grids. The linear solver also fails to converge with the combination of the block Gauss-Seidel preconditioner and the Rusanov diffusion for values $b = 10$ or $b = 100$. Thus from these simulations, one can infer that a Roe type of diffusion is preferable over a Rusanov type diffusion operator.

As a last comparison, we investigate how the fast waves are treated by the scheme. For this purpose, we enlarge the domain to $(-1, 10)$ and compute only to the time $t = 0.05$. This implies that the initial jump in the fast wave does not leave the domain. The approximate solutions are shown in Figure 8.4. For coarse grids, the fast wave is smeared over a large part of the domain. It starts getting resolved with further refinement, but its is still highly smeared. This behaviour is expected as the time step is chosen only to resolve the slow, but not the fast waves. As long as the main interest is in the slow waves with the fast waves not dominating the solution, the scheme is effective and can resolve the slow waves, even...
8.2. 1-D Euler equations

In the remainder, we will consider the nonlinear Euler equations of gas dynamics (1.25).

We consider the second problem of [32], which is a one dimensional Euler flow. It is a large amplitude, short wave length density layering set in motion by a periodic train of right-running acoustic pulses. In [32], the Euler equations are nondimensionalised using the reference Mach number $M$, which is $1/51$ in this example. As we use a dimensional form of the equations, the data has to be rescaled compared to [32] and reads

$$
\rho(x, 0) = \tilde{\rho}_0 + \Phi(x)\tilde{\rho}_0 \sin(40\pi x/L) + M\tilde{\rho}_0 \frac{1}{2}(1 + \cos(\pi x/L)),
$$

$$
p(x, 0) = \left(\tilde{p}_0 + M\tilde{p}_0 \frac{1}{2}(1 + \cos(\pi x/L))\right)/M^2,
$$

$$
u(x, 0) = \tilde{u}_0 \frac{1}{2}(1 + \cos(\pi x/L)),
$$

where $x$ is in the domain $[-L, L]$ with $L = 1/M$ and

$$
\tilde{\rho}_0 = 1, \quad \tilde{\rho}_0 = 1/2, \quad \tilde{\rho}_0 = 1, \quad \tilde{p}_0 = 2\gamma, \quad \tilde{u}_0 = 2\sqrt{\gamma}.
$$

The function

$$
\Phi(x) = \begin{cases} 
\frac{1}{2}(1 - \cos(5\pi x/L)), & 0 \leq x \leq 2L/5, \\
0, & \text{otherwise}
\end{cases}
$$

Figure 8.4.: Approximate solution for the two speed advection problem on larger domain using $p = 2$ and $b = 100$. though the effective CFL is quite high.
8. All-speed flows

Figure 8.5.: Density in Klein’s problem II with polynomial degree \( p = 2 \).

is used to smoothly restrict the large amplitude, short wave length density variation to the region \([0, 2L/5]\). The initial density is visualized in Figure 8.5 (a). Periodic boundary conditions are used and the equations are solved from \( t = 0 \) up to \( T = 5.071 \). In this time span, the long wave length acoustic pulse crosses the domain about two and a half times.

The approximate solutions are shown in Figure 8.5 (b). For coarse meshes the density variations are smeared out considerably, but for instance with 1600 spatial cells, they are quite accurately represented. The amplitude of the high frequency variations is of the same size as in the initial condition, but the variations are shifted to the right due to the movement of the acoustic pulses. It is apparent that even though the time step is chosen only according to the slow waves, with an effective CFL of around 18, the approximate solutions are quite accurate.

The number of Krylov iterations per Newton iteration are depicted in Figure 8.6. Again the number of iterations are more or less constant with respect to grid refinement. About 25 \((p = 1)\) or 20 \((p = 2)\) iterations are needed with the block Jacobi preconditioner and the Roe type (3.13) of diffusion. Considerably more iterations are needed or the solver even fails completely with the Rusanov type (3.14) of diffusion, based only on the velocity \( u \). The best combination is again a Roe type of diffusion together with the (symmetric version of the) block Gauss-Seidel preconditioner. For this combination, the average number of Krylov iteration per Newton iteration lies between 2 and 3. Unlike the previous linear case, the preconditioner is no longer a direct solver, but still performs very well. To enable this, the cells need to be sorted in one direction and sweeps in both directions have to be performed, i.e. the symmetric version of Gauss-Seidel is needed (see Chapter 7).

8.3. Flow around a cylinder

We consider an Euler flow around a cylinder with a free-stream Mach number \( M_{a\infty} \). The cylinder has a diameter of 1 and the outer artificial boundary is a circle of radius 10 around
8.3. Flow around a cylinder

The following free-stream variables are imposed:

\[ p_\infty = 101325, \quad \rho_\infty = \frac{p_\infty}{287.05 \cdot 288.15}, \quad u_\infty = \text{Ma}_\infty c_\infty, \quad v_\infty = 0, \quad (8.9) \]

where \( c_\infty = \sqrt{\gamma p_\infty / \rho_\infty} \). At the cylinder, slip boundary conditions are imposed. It is important that the cylinder boundary is accurately resolved. For this purpose, a polynomial mapping of degree \( p \) (for \( p \geq 1 \)) is used to generate curved boundary elements. The equations are solved from \( t = 0 \) up to \( T = 0.02 / \text{Ma}_\infty \).

We use a Roe type diffusive flux. The time step is determined using the reduced CFL that is based only on the velocity \( \lambda_{\text{slow}}(U) = \|u\| \), but not on the sound speed \( c = \sqrt{\gamma p / \rho} \).

In Figure 8.7, the pressure coefficient

\[ c_p = \frac{p - p_\infty}{\frac{1}{2} \rho_\infty \|u_\infty\|^2} \]

is depicted for \( \text{Ma}_\infty = 0.1, \text{Ma}_\infty = 0.01 \), and \( \text{Ma}_\infty = 0.001 \). The resulting \( c_p \) value of the incompressible flow (potential flow) is also shown as a reference. The flow is very well approximated for all the three Mach numbers, not only on the upstream (left) side of the cylinder, but also at the downstream (right) side of the cylinder. For this, a good approximation of the cylinder boundary is crucial.

For the computation, a reduced CFL number of \( \text{CFL}^{\text{red}} = 10 \) was used. This results in an effective CFL of about 65, 590, 6000 for \( \text{Ma}_\infty = 0.1, \text{Ma}_\infty = 0.01 \), and \( \text{Ma}_\infty = 0.001 \), respectively. Despite the high CFL number, the approximate solution is computed accurately.
8. All-speed flows

Figure 8.7.: Pressure coefficient $c_p$ of a flow around a cylinder using polynomial degree $p = 2$ and $N_c = 13282$ spatial elements.
8.3. Flow around a cylinder

The number of Newton iterations is quite low, i.e. below 10. The number of linear solver iterations on the other hand grows quickly with decreasing Mach number. With the symmetric block Gauss-Seidel preconditioner, the number of iterations is about 37, 220, and 850 with $\text{Ma}_\infty$ being 0.1, 0.01, and 0.001, respectively. For $\text{Ma}_\infty = 0.001$, the linear solver even fails on the finest grid, and a multi block GS preconditioner is used instead to compute the above shown results. So the current bottleneck in the computation is the linear solver or its preconditioner. But note that the Mach numbers span three orders of magnitude, while the number of iterations increases approximately by a factor 5 for each order of magnitude.

Figure 8.8.: Pressure coefficient of a flow around a cylinder with $\text{Ma}_\infty = 0.75$ using polynomial degree $p = 2$ and $N_c = 13282$ spatial elements.

As a last example, we consider a transsonic flow with $\text{Ma}_\infty = 0.75$. We use $CFL^{red} = 1$, which results in an effective CFL of roughly 1.6. The approximate solutions for degree 2 and 13282 spatial elements is shown in Figure 8.8. Approximately 4 Newton iterations and, in each Newton iteration, on an average less than 4 Krylov iterations are needed for the symmetric Gauss-Seidel preconditioner. As it can be seen in the figure, the scheme does not only work for low Mach numbers but for also for high Mach numbers. (Note that the flow is not yet in steady state.) See Chapter 7 for more examples of this kind of flows, where the scheme is successfully applied.
9. Adaptivity

Mesh adaptivity can greatly enhance accuracy when compared to a uniform mesh with the same number of cells, especially if singularities such as shocks are present in the solution. As we use a space-time formulation, we consider local mesh refinement in both space and time. We utilise two different approaches towards adaptivity: The first approach is based on residuals of the computed solution, which gives quite promising results with respect to the global error of the solutions. However in most problems, one is not interested in the solution itself, but rather in functionals such as e.g. drag or lift in aerodynamics. Thus, we also consider goal-oriented adaptivity. The error for approximating the functional is estimated using the solution of a dual problem. This leads to a refinement in those parts of the space-time domain that influence the functional error.

9.1. Space-time mesh

We consider a space-time mesh $\tilde{T}$ of $\Omega \times (0, T)$. Each space-time element $\tilde{K} \in \tilde{T}$ is assumed to be a prism of the form

$$\tilde{K} = K \times I$$

(9.1)

with $K \subset \mathbb{R}^d$ being an open convex polyhedra with plane faces and $I$ an interval with $I = [\bar{t}, \bar{t}]$. Furthermore, we use

$$\Delta x_K = \text{diam}(K),$$

$$\Delta t^I = \bar{t} - t,$$

$$\mathcal{N}(K \times I) = \{K' \times I' \in \tilde{T} : K' \neq K \land \text{meas}_{d-1}(\overline{K \cap K'}) > 0 \land \text{meas}_1(\overline{I \cap I'}) > 0\}.$$

The mesh does not need to be conforming, e.g. hanging nodes are allowed. However, we use only a regular refinement of elements, i.e. if a element is refined, then it is refined in both space and time, see Figure 9.1. We therefore assume that there is a constant $C$ such that $(1/C)\Delta x_K \leq \Delta t^I \leq C\Delta x_K$ for all elements $K \times I \in \tilde{T}$. Additionally, we assume that the spatial elements $K$ are shape regular.
9. Adaptivity

![Diagram of space-time elements](image)

(a) One spatial dimension  
(b) Two spatial dimensions

Figure 9.1.: Refinement of space-time elements.

9.2. Formulation of the method

The formulation is a generalisation of the DG method (3.2) described in Chapter 3 for a space-time mesh. We seek

\[ \mathbf{V}^{\Delta x} \in \mathcal{V}_p = \left( \mathbb{P}_p(\Omega \times [0, T]) \right)^m \]

such that the following identity is satisfied

\[ B(\mathbf{V}^{\Delta x}, \mathbf{W}^{\Delta x}) := B_{DG}(\mathbf{V}^{\Delta x}, \mathbf{W}^{\Delta x}) + B_{SD}(\mathbf{V}^{\Delta x}, \mathbf{W}^{\Delta x}) + B_{SC}(\mathbf{V}^{\Delta x}, \mathbf{W}^{\Delta x}) = 0. \]  

for each \( \mathbf{W}^{\Delta x} \in \mathcal{V}_p \).

The DG form is given by

\[ B_{DG}(\mathbf{V}^{\Delta x}, \mathbf{W}^{\Delta x}) = \sum_{K \times I} \int_K \left( \left\langle \mathbf{U}(\mathbf{V}^{\Delta x}), \mathbf{W}_t^{\Delta x} \right\rangle + \sum_{k=1}^d \left\langle \mathbf{F}^k(\mathbf{V}^{\Delta x}), \mathbf{W}_{x_k}^{\Delta x} \right\rangle \right) dx dt + \sum_{K \times [t, \bar{t}]} \int_K \left\langle \mathbf{U}(\mathbf{V}_{t,-}^{\Delta x}, \mathbf{V}_{t,+}^{\Delta x}), \mathbf{W}_{t,-}^{\Delta x} \right\rangle dx - \sum_{K \times [\bar{t}, t]} \int_K \left( \mathbf{U}(\mathbf{V}_{t,-}^{\Delta x}, \mathbf{V}_{t,+}^{\Delta x}), \mathbf{W}_{t,+}^{\Delta x} \right) dx \]  

for each \( \mathbf{W}^{\Delta x} \in \mathcal{V}_p \).
9.2. Formulation of the method

Here we have employed the notation that was introduced in (2.9) and (3.4).

The streamline diffusion term is

\[
\mathcal{B}_{SD}(\mathbf{V}^{\Delta x}, \mathbf{W}^{\Delta x}) = \sum_{K \times I} \int_I \int_K \left( \left( \mathbf{U}_V(\mathbf{V}^{\Delta x}) \mathbf{W}_t^{\Delta x} + \sum_{k=1}^d \mathbf{F}^{k}_V(\mathbf{V}^{\Delta x}) \mathbf{W}^{\Delta x}_{x_k} \right), \mathbf{D}^{SD}_{\text{Res}} \right) dxdt \tag{9.5}
\]

with the scaling matrix

\[
\mathbf{D}^{SD} = C^{SD} \mathbf{t}^I \mathbf{U}^{-1}_V(\mathbf{V}^{\Delta x}). \tag{9.6}
\]

The shock-capturing form is

\[
\mathcal{B}_{SC}(\mathbf{V}^{\Delta x}, \mathbf{W}^{\Delta x}) = \sum_{K \times I} \int_I \int_K D^{SC}_{K \times I} \left( \left( \mathbf{W}_t^{\Delta x}, \mathbf{U}_V \mathbf{V}^{\Delta x}_t \right) + \sum_{k=1}^d \frac{\Delta x_k^2}{(\mathbf{t}^I)^2} \left( \mathbf{W}^{\Delta x}_{x_k}, \mathbf{U}_V \mathbf{V}^{\Delta x}_{x_k} \right) \right) dxdt, \tag{9.7a}
\]

with \( \mathbf{U}_V = \mathbf{U}_V(\mathbf{V}_{K \times I}) \) for brevity and

\[
\mathbf{V}_{K \times I} = \frac{1}{\text{meas}(I \times K)} \int_I \int_K \mathbf{V}^{\Delta x}(x,t) dx dt.
\]

being the cell average and the scaling factor,

\[
D^{SC}_{K \times I} = \frac{\Delta t^I C^{SC} \mathbf{R}^{K \times I} + (\Delta t^I)^2 C^{SC} B^{K \times I} \mathbf{R}^{K \times I}}{\sqrt{\int_I \int_K \left( \left( \mathbf{V}^{\Delta x}_t, \mathbf{U}_V \mathbf{V}^{\Delta x}_t \right) + \sum_{k=1}^d \frac{\Delta x_k^2}{(\mathbf{t}^I)^2} \left( \mathbf{V}^{\Delta x}_{x_k}, \mathbf{U}_V \mathbf{V}^{\Delta x}_{x_k} \right) \right) dxdt + \epsilon}}, \tag{9.7b}
\]

with \( \epsilon = |K|^{\frac{1}{2}} (\Delta t^I)^{-\frac{1}{2}} \left( \frac{\Delta x}{\text{diam}(\Omega)} \right)^\theta \) and \( \theta \geq 1/2 \) (chosen as 1) and

\[
\mathbf{R}^{I \times K} = \sqrt{\int_I \int_K \left( \mathbf{R}_V^{-1}(\mathbf{V}^{\Delta x}) \mathbf{R}_V \right) dxdt}. \tag{9.7c}
\]

\[
B^{K \times I} = \left( \int_K \|\mathbf{U}(\mathbf{V}^{\Delta x}_{l,-}) - \mathbf{U}(\mathbf{V}^{\Delta x}_{l,+}) \|_{\mathbf{U}^{-1}_V(\mathbf{V}^{\Delta x}_{n,+})}^2 \right) dx
\]

\[
+ \sum_{K' \times I'} \int_{I' \cap I' \cap K} \frac{\Delta t^I}{\Delta x_K} \left\| \mathbf{F}(\mathbf{V}^{\Delta x}_{K,-}, \mathbf{V}^{\Delta x}_{K,+}; \nu_{K,K'}) - \sum_{k=1}^d \mathbf{F}^k(\mathbf{V}^{\Delta x}_{K,-}) \nu^k_{K,K'} \right\|^2_{\mathbf{U}^{-1}_V(\mathbf{V}^{\Delta x}_{K,-})} d\sigma dt \right)^{\frac{1}{2}} \tag{9.7d}
\]
9. Adaptivity

9.3. Implementation aspects

9.3.1. Time marching

Consider the mesh in Figure 9.2. Degrees of freedom (dofs) in a given element are coupled to the dofs of another element in two ways. First of all, the dofs of an element depend directly on the dofs of elements that touch at the lower temporal boundary. This not true for the upper temporal boundary because of the upwinding in time. Secondly, a direct dependency occurs as follows: the dofs of the given element can depend on dofs of elements that touch the spatial boundary. For the linear advection equation with \( a > 0 \), the dofs depend only on the dofs in the left (upwind) direction. But in the general case, spatial boundaries introduce dependencies in both directions.

![Figure 9.2: Illustration of the global time marching.](image)

Continue considering the mesh depicted in Figure 9.2. The dofs of all the elements in the time slab \([0, 0.5]\) might depend on each other, but not on other dofs. So we can first solve for these degrees of freedom. Afterwards, we can compute the dofs in the time slab \([0.5, 0.75]\) and so on. Algorithmically, we perform a downwind ordering in time direction of the elements and determine a minimal global time step such that each space-time element is either completely included in or completely excluded from the time slab. In the worst case (with respect to the mesh) one has to solve the problem in just one global time step. But this implies huge memory requirements (especially for the Jacobian) since the amount of requested memory is roughly proportional to the number of space-time element in the current global time step. Thus, one might even want to refine certain cells to obtain a smaller global time step and reduce memory requirements. However, we have not implemented such strategies so far.
9.3.2. Preconditioning

The dependencies described in the last subsection also influence the linearised system in the Newton steps. As the dofs of certain elements might depend very indirectly on dofs in other elements (consider e.g. the cells in the vicinity of the contact discontinuities in Figure 9.3) a block Jacobi preconditioner is not expected to work well on adapted meshes. Here, we use a block Gauss-Seidel type of preconditioner with a downwind ordering of the cells in the temporal direction and a suitable ordering in the spatial directions. For the linear advection equation, the block Gauss-Seidel method acts as a direct solver of the linearised system (in the absence of periodic boundary conditions). However, this has to be considered as a special case, as not all of the dependencies can be respected in general.

9.3.3. Parallelisation

Compared to the non-adaptive version, (distributed memory) parallelisation is substantially trickier. In the non-adaptive version, one can initially distribute the spatial mesh, and afterwards, only needs to exchange the values of certain fixed ghost cells during time marching. For the adaptive code, the mesh looks different in each global time step and the corresponding elements must be distributed in a new way to balance the workload. This complicated memory layout has so far prevented a parallel implementation of the adaptive method.

9.4. Residual based adaptivity

The refinement is based on the following elementwise heuristic indicator,

\[ \eta_{K \times I} = \left( \Delta x_K^{\alpha} \left( \Delta t^I \text{Res}^2_{K \times I} + \text{BRes}^2_{K \times I} \right) \right)^{1/2} \]  

(9.8)

based on the interior and boundary residual together with \( \alpha \geq 0 \). The reason for the scaling with \( \Delta x_K^{\alpha} \) is that the method tends to refine shocks too much. This additional factor favours refinement of bigger cells. A value of \( \alpha = 3 \) has worked well in practice. We refine a fixed ratio of the cells (30 % unless specified), namely those that have the biggest indicator \( \eta_{K \times I} \), and no coarsening is performed.

9.4.1. Linear advection equation

As a first example, we consider the linear advection equation (1.16) with \( a = 1 \) in one spatial dimension on the interval \([-1, 1]\] with periodic boundary conditions. We start with the following discontinuous initial data

\[ u(x, 0) = \begin{cases} 1 & x < 0, \\ 0 & \text{otherwise}. \end{cases} \]  

(9.9)
9. Adaptivity

We start with an initial space-time mesh that consists of only 4 cells and then refine 12 times using the residual based estimator. In Figure 9.3, the resulting mesh and solution is depicted. The mesh refinement nicely follows the contact discontinuities and therefore, they are captured quite sharply.

Figure 9.3.: Mesh with 12550 elements and corresponding solution for the linear advection problem using $p = 2$.

The convergence is visualised in Figure 9.4. Here, adaptive and uniform refinement are compared for different degrees $p$. With uniform refinement, the convergence rate with respect to the number of space-time elements is approximately 0.4 irrespective of the polynomial degree. This corresponds to a convergence rate of 0.8 with respect to $\Delta x$. For the adapted meshes, the convergence rate with respect to the number of space-time cells is roughly 0.8, which is twice as high as compared to uniform refinement. The mesh needs to be reasonably refined, but eventually the adaptivity starts to pay off, for instance for $p = 2$ for $N_c = 12550$ adapted cells and 153600 uniform cells one obtains roughly an error of 0.01. Thus, adaptivity reduces the number of cells needed to achieve the same accuracy by an order of magnitude.

9.4.2. Sod shock tube

As a second example, we consider the Sod shock tube, see Section 5.2.3. Again, we start with a mesh consisting of 4 cells and refine 12 times. The mesh and the approximate solution can be seen in Figure 9.5. The refinement nicely follows the structure of the solution. The mesh is refined the most near the shock, a bit less for the contact discontinuity, and some minor refinement is done in the expansion regions, specifically near the head and tail of the rarefaction wave, where the solution is not smooth. The additional factor $\Delta x_K^p$ in (5.2.3)
9.4. Residual based adaptivity

really improves the mesh quality; without it the refinement would almost exclusively focus on the shock. An even higher value of $\alpha$ might further improve the quality of the mesh and the solution.

In Figure 9.6, the convergence is depicted. With uniform refinement, a convergence rate of almost 0.5 is obtained with respect to the number of space-time cells for $p = 1$ and $p = 2$. This corresponds to almost first order convergence in $\Delta x$. The convergence rate with adaptivity is about 0.8 with respect to $N_c$, which is significantly higher. To get an accuracy of about 0.001 for $p = 2$, 12550 and 366080 cells are needed for adapted and uniform meshes, respectively. Therefore, mesh adaptivity allows to reduce the number of cells by more than an order of magnitude without sacrificing accuracy.

9.4.3. Forward facing step flow

We consider a Mach 3 wind tunnel Euler flow with a forward facing step, studied by Woodward and Collela [50], see also [6]. For the geometry of the wind tunnel, see the mesh in Figure 9.7. Slip boundary conditions are applied at the lower and the top wall. Initially, the tunnel is filled with a gas with density $1.4$, pressure $1$, and velocity $3$ in $x$-direction. Gas with this state is fed at the left boundary. As the flow is supersonic at the right (outflow) boundary, no boundary condition needs to be applied there (formally we use a Neumann boundary).

The flow impinges on the step. This generates a shock that spreads into the domain and that is reflected back and forth between the walls. We simulate the flow up to the time $t = 4$. Note that it has not yet reached the steady state by that time.

We start the computation on a space-time mesh with only 2871 elements and use a polynomial degree $p = 1$. After 6 refinement steps using the residual-based estimator, we obtain a mesh with 2549821 space-time elements. In Figure 9.7, the spatial mesh is shown at dif-

Figure 9.4.: Convergence for the linear advection equation.
9. Adaptivity

Figure 9.5.: Mesh with 12550 elements and corresponding solution for the Sod shock tube problem using $p = 2$.

Figure 9.6.: Convergence for Sod shock tube.
Figure 9.7.: Mesh and corresponding solution for the forward facing step flow problem using $p = 1$. 
9. Adaptivity

Different times. Additionally, the density of the approximate solutions is depicted. It is clearly visible that the mesh refinement concentrates on the shocks. It is able to follow the shock as it evolves in time. Therefore, the shock profiles are resolved nicely. The mesh is also refined towards the corner of the step. This corresponds to the centre of the rarefaction fan, which is an additional singularity of the flow. Actually, it might require even more refinement to eliminate the second Mach stem, which is considered to be a numerical artefact, compare [6]. Nevertheless, the residual-based adaptivity allows to capture the flow using few degrees of freedom.

9.4.4. Aerofoil flow

We consider a two dimensional flow around a NACA 0012 aerofoil, see Section 5.3.3 for the setup. The free-stream Mach number is 0.75, but the flow is supersonic along the upper part of the aerofoil leading to a shock. We start we a mesh consisting only of 3132 elements and refine it 4 times. The resulting mesh is depicted in Figure 9.8. It is actually a space-time mesh, but for the sake of clarity only the spatial mesh at the final time is shown. The mesh is mostly refined towards the head and tail of the aerofoil, and also towards the shock in the flow above the aerofoil. The pressure coefficient $c_p$ is shown in the same figure. The flow and especially the shock is resolved quite nicely. Compare this to Figure 7.19. There, only 47753 space-time elements are used. But even with another uniform refinement (which leads to a mesh with more elements than in the adapted mesh) the shock would not be well resolved comparatively. Once again, adaptivity leads to big improvement.

![Figure 9.8.](image-url)

Figure 9.8.: Closeup of the mesh with 289362 elements and corresponding solution for the aerofoil flow using $p = 2$ at the final time.
Note that as the boundary is curved, it is crucial that it is represented well in the mesh refinement as well as in the formulation itself. Therefore, the new boundary nodes in the refinement are projected to the exact boundary location. Additionally the elements along the aerofoil boundary are curved using a polynomial mapping of degree at least $p$ based on an interpolation of the true boundary.

We repeat the experiment with a supersonic free-stream with a Mach number of 1.3. The corresponding mesh and solution is depicted in Figure 9.9. Here, we used only a polynomial degree $p = 1$, but did 6 refinement steps starting from a mesh with 3132 space-time cells. The bow and tail shock are represented nicely in the mesh and thus also in the solution. The lower tail shock is the weakest of the shocks present in the solution. Therefore, the mesh is refined the least in this region when compared with the other shocks. But nevertheless, also this shock is resolved quite well.

![Mesh and Pressure Coefficient](a) Mesh (b) Pressure coefficient $c_p$

Figure 9.9.: Mesh with 2780781 elements and corresponding solution for the supersonic aerofoil flow using $p = 1$ at the final time.

### 9.5. Goal-oriented adaptivity

Goal-oriented adaptivity aims at computing functionals of the solution as accurately as possible, rather than minimising the global error of the solution. The estimates are duality-based, i.e. not only the primal problem, but also a dual problem needs to be solved. We follow the approach of Barth [2, 3].
9. Adaptivity

9.5.1. Linear case

Let us first consider the linear case. We assume that we have a bilinear form \( B(V, W) \) and a linear functional \( F(V) \) such that the primal (numerical) problem is given as follows: Find \( V^{\Delta x} \in \mathcal{V}_p \) such that

\[
B(V^{\Delta x}, W^{\Delta x}) = F(W^{\Delta x}) \tag{9.10}
\]

for all test functions \( W^{\Delta x} \in \mathcal{V}_p \). In order to bring the scheme (9.3) into the form (9.10) with \( B \) bilinear and \( F \), one may only consider linear symmetrisable systems with the quadratic entropy. Furthermore, the scheme consists only of the DG and streamline diffusion operator; the shock-capturing term must be omitted due to its nonlinearity. The contribution of the initial conditions and of some of the boundary conditions enter into the functional \( F \).

We are interested in the value of a linear functional \( J(V) \). In order to estimate the error in approximating the functional, let us define the following dual (numerical) problem. It is formulated on a suitable infinite-dimensional broken Sobolev space \( \mathcal{V} \) corresponding to the mesh. Find \( \phi \in \mathcal{V} \) such that

\[
B(W, \phi) = J(W) \tag{9.11}
\]

for all \( W \in \mathcal{V} \). It is formally obtained by interchanging the test and trial functions in the bilinear form and replacing the functional \( F \) by the target functional \( J \).

Under the assumption that the dual solution \( \phi \) exists and that the exact solution is contained in \( \mathcal{V} \), we can exactly represent the error as follows

\[
J(V) - J(V^{\Delta x}) = J(V - V^{\Delta x}) \quad \text{(linearity of } J \text{)}
\]

\[
= B(V - V^{\Delta x}, \phi) \quad \text{(dual problem)}
\]

\[
= B(V - V^{\Delta x}, \phi - \pi^{\Delta x} \phi) \quad \text{(orthogonality)}
\]

\[
= B(V, \phi - \pi^{\Delta x} \phi) - B(V^{\Delta x}, \phi - \pi^{\Delta x} \phi) \quad \text{(linearity of } B \text{)}
\]

\[
= J(\phi - \pi^{\Delta x} \phi) - J(V^{\Delta x}, \phi - \pi^{\Delta x} \phi) \quad \text{(primal problem)}
\]

(9.12)

Here, \( \pi^{\Delta x} \) is an arbitrary linear mapping (e.g. a projection) from \( \mathcal{V} \) to \( \mathcal{V}_p \). Note that the right hand side does not depend on the exact solution.

9.5.2. The nonlinear case

We generalise this to the case where \( B \) is quasilinear and \( J \) is nonlinear. For this purpose, we consider the mean value linearisations

\[
B(V, W) = B(V^{\Delta x}, W) + B(V - V^{\Delta x}, W) \quad \text{for all } W \in \mathcal{V} \tag{9.13}
\]

\[
J(V) = J(V^{\Delta x}) + J(V - V^{\Delta x}) \tag{9.14}
\]
Note that $\bar{B}$ and $\bar{J}$ depend on $V$ and $V^{\Delta x}$.

We define the dual problem as: Find $\phi \in \mathcal{V}$ such that

$$
\bar{B}(W, \phi) = \bar{J}(W) \quad (9.15)
$$

To represent the error, we can proceed in an analogous way to the linear case

$$
\mathcal{J}(V) - \mathcal{J}(V^{\Delta x}) = \bar{J}(V - V^{\Delta x}) \quad \text{(mean value of $\mathcal{J}$)}
$$

$$
= \bar{B}(V - V^{\Delta x}, \phi) \quad \text{(dual problem)}
$$

$$
= \bar{B}(V - V^{\Delta x}, \phi - \pi^{\Delta x} \phi) \quad \text{(orthogonality)}
$$

$$
= B(V, \phi - \pi^{\Delta x} \phi) - B(V^{\Delta x}, \phi - \pi^{\Delta x} \phi) \quad \text{(mean value of $B$)}
$$

$$
= \mathcal{F}(\phi - \pi^{\Delta x} \phi) - B(V^{\Delta x}, \phi - \pi^{\Delta x} \phi) \quad \text{(primal problem)}
$$

(9.16)

But even though the final formula looks exactly the same way as in the linear case, there is one important difference. The solution of the dual problem $\phi$ depends via the mean value linearisation on the exact solution $V$.

### 9.5.3. Computable estimates and refinement strategy

The error representation formulas (9.12) and (9.16) can not directly be applied in practice. The reason is that the dual problem is posed in the infinite-dimensional space $\mathcal{V}$ and that for nonlinear problems the mean value linearisation depend on the exact solution. Thus, we approximate the dual solution $\phi$ by $\phi^{\Delta x}$ by replacing $\mathcal{V}$ by $\mathcal{V}_{p+1}$. Note that $\mathcal{V}_{p}$ would not suffice as the right hand side in the error representation formulas would be exactly zero. In order to eliminate the second problem, we replace the mean value linearisation by a linearisation at the approximate solution $V^{\Delta x}$. This implies that we can reuse the evaluation routine of the Jacobian because we have to solve a linear system where the matrix is given by its transpose. More sophisticated strategies can be applied as well, such as those in [3].

In order to perform adaptivity, we have to split up the contributions of each element to the error. We start by splitting up the assembly as follows

$$
\mathcal{B}(V, W) = \sum_{K} \mathcal{B}(V, 1_{K} W), \quad \mathcal{F}(W) = \sum_{K} \mathcal{F}(1_{K} W).
$$

(9.17)

This leads to

$$
|\mathcal{J}(V) - \mathcal{J}(V^{\Delta x})| \approx \left| \mathcal{F}(\phi^{\Delta x} - \pi^{\Delta x} \phi^{\Delta x}) - B(V^{\Delta x}, \phi^{\Delta x} - \pi^{\Delta x} \phi^{\Delta x}) \right|
$$

$$
= \left| \sum_{K} \mathcal{F}_{K}(\phi^{\Delta x} - \pi^{\Delta x} \phi^{\Delta x}) - B_{K}(V^{\Delta x}, \phi^{\Delta x} - \pi^{\Delta x} \phi^{\Delta x}) \right|_{\eta^{K}}
$$

$$
= \left| \sum_{K} \eta^{K} \right|
$$

129
9. Adaptivity

\[ \leq \sum_{\tilde{K}} |\eta^{\tilde{K}}| \] (9.18)

Thus, we use \( |\sum_{K} \eta^{K}| \) as functional error estimate, while \( |\eta^{\tilde{K}}| \) serves as an estimate of contribution of cell \( \tilde{K} \) to the functional error. Hence, we refine those (30% of the) cells that have the largest values of \( |\eta^{\tilde{K}}| \). We will call \( |\sum_{K} \eta^{K}| \) the estimated functional error and \( \sum_{K} |\eta^{K}| \) is called the estimated functional error bound.

9.5.4. Linear advection equation

**Smooth initial data**

As a first example, we consider the linear advection equation as in Section 9.4.1. However, we start with the initial data

\[ u(x, 0) = \sin(2\pi x). \] (9.19)

Thus, the exact solution is smooth. We consider the following target functional

\[ J(u) = \int_{t_a}^{t_b} \int_{x_a}^{x_b} \Psi(t; t_a, t_b) \Psi(x; x_a, x_b) u \, dx \, dt \] (9.20)

with the bump function

\[ \Psi(x; a, b) = \exp \left( 1 - \frac{1}{1 - ((t - \frac{a+b}{2}) \frac{2}{b-a})^2} \right) \] (9.21)

and \([t_a, t_b] = [1.25, 1.5]\) and \([x_a, x_b] = [0, 0.5]\).

We start with a mesh consisting of 4 space-time elements and refine 12 times using the functional error estimate. We use only the DG part of the formulation. Figure 9.10 shows the resulting mesh. The region \([0, 0.5] \times [1.25, 1.5]\) and its domain of dependence is clearly refined. The other regions are almost not refined, which shows that the functional error indicator correctly predicts where to refine. The solution is shown in the same figure. It is well-resolved in the regions where the mesh is refined, but the approximation is quite coarse in the other parts, e.g. in \([0.5, 1] \times [1.125, 1.5]\).

For the linear advection problem, the dual problem is a linear advection equation to be solved backwards in time. The terminal condition are that the dual is zero, but the functional \( J(u) \) acts as a source term. The solution of the dual problem is depicted in Figure 9.10 (c). In most of the domain \( \phi \) is zero. However, in the region \([0, 0.5] \times [1.25, 1.5]\) the value of \( \phi \) increases due to \( J \) and this bump is then advected to the left.

In Figure 9.10, we compare the convergence of the functional \( J \) for uniform refinement, residual-based refinement, and goal-oriented refinement. As the solution is smooth, a uniform refinement works pretty well. Residual-based refinement leads to roughly the same
9.5. Goal-oriented adaptivity

Figure 9.10.: Mesh with 12550 elements and corresponding solution for the linear advection problem using $p = 2$. 
9. Adaptivity

accuracy, and there is a certain advantage for goal-oriented refinement. This reflects the fact that goal-oriented adaptivity does not have to refine the whole domain, but only a certain fixed part of it. However, as the region where refinement is needed is a two dimensional area (in space-time), one can not expect the adaptive method to perform substantially better than with uniform refinement, i.e. the value of the error can be lower, but the convergence rate can not be higher. For all of the three refinement strategies the convergence rate is approximately 2.5, which corresponds to a convergence rate of 5 with respect to the mesh width in a uniform refinement. This agrees with the expected value of \(2p + 1\).

![Figure 9.11: Mesh with 12550 elements and corresponding solution for the linear advection problem using \(p = 2\).](image)

In Figure 9.11, the three refinement strategies are studied in more detail. Depicted is the functional error, the estimated functional error, and the estimated functional error bound (using the absolute values of the indicators). All quantities are normalised with the exact or approximate value of the functional to obtain relative errors. In all three cases, the estimated error agrees quite well with the exact error. For this, it was crucial to compute the functional value \(J\) using a high-order accurate quadrature rule. The error bound is always higher than the estimated error. This shows that indeed there is some cancellation happening in the error estimation. Except on the coarsest meshes, the error bound decreases monotonically, in contrast to the functional error or its estimate, which increase in some of the refinement steps. We think that this corresponds to the fact that the adaptivity is performed with respect to estimated functional error bound, i.e. the cells with large values of the indicator are refined and thus, adaptivity ignores cancellation that might happen.

**Discontinuous initial data**

We consider again the discontinuous initial data (9.9). We use the same target functional, i.e. it is given by (9.20). In Figure 9.12, the mesh, the approximate solution, and the dual solution after 12 refinement steps are depicted. The dual solution looks the same as in the smooth case, as for linear problems there is no dependence of the dual solution on the...
9.5. Goal-oriented adaptivity

Figure 9.12.: Mesh with 12550 elements and corresponding solution for the linear advection problem using $p = 2$. 
9. Adaptivity

approximate solution. The refinement is done again in the region of the dependency of the support of the functional $J$. Furthermore, it is concentrating on the contact discontinuity in this region. In the approximate solution, this contact discontinuity is well-resolved while the other contact discontinuity is almost not resolved at all. However, this does not reduce the accuracy of the functional.

The convergence of uniform, residual-based adaptive, and goal-oriented adaptive refinement is shown in Figure 9.12 (d). The convergence in the case of uniform refinement is the slowest and the rate is only approximately 1.5. In the other two cases, the defect of the convergence rate can be partially removed; it is approximately 2.0. Goal-oriented adaptivity is the most effective strategy for coarse meshes, but residual-based adaptivity is comparable.

![Figure 9.13.: Convergence for the linear advection problem using $p = 2$.](image)

The convergence is analysed in more detail in Figure 9.13. For very coarse meshes, the error is estimated badly, but for reasonably fine meshes for all the three refinement algorithms the error agrees quite well with its estimate. Again, the error bound is consistently higher, which demonstrates the effect of cancellation.

9.5.5. Burgers’ equation

As a second problem, we consider the Burgers equation (1.19) in the domain $[-1, 1]$ with periodic boundary conditions. The initial data is given by

$$u(x, 0) = \begin{cases} 1 & x < 0, \\ 0 & \text{otherwise}, \end{cases}$$

(9.22)

and we solve it up to time $t = 1.5$. The exact solution is given by a rarefaction centred at $-1$ and a shock starting at 0. As target functional, we consider same type of functional as for the advection equation (9.20), but we change the support to $[x_a, x_b] = [0.5, 1]$ times $[t_a, t_b] = [1.25, 1.5]$. Thus, it only touches the rarefaction, but the shock passes through it.

We start with 4 space-time cells and refine 12 times. Figure 9.14 shows the resulting mesh, the approximate solution, and the dual solution. Once more, the functional $J$ acts as
9.5. Goal-oriented adaptivity

(a) Mesh

(b) Solution

(c) Dual solution

(d) Convergence

Figure 9.14.: Mesh with 12550 elements and corresponding solution for the Burgers’ problem using $p = 2$. 
9. Adaptivity

a source term in the dual problem. In the region to the right of the shock, the approximate solution is zero and therefore the dual solution $\phi$ remains constant once the source term is no longer active. In the region to the left of the shock, the approximate solution is one and the dual solution is advected with this velocity. Note however, the integral of $\phi$ increases when going backwards in time. Thus, the shock acts as an additional source term (while the rarefaction would act as a sink). The shock is resolved quite well in the area where the functional $\mathcal{J}$ is nonzero. At earlier times, the mesh is almost not refined towards the shock. This is in fact not needed, because the characteristics enter the shock and therefore, the shock is resolved accurately enough once the mesh refinement starts. One can also argue using the error representation formula (9.16). As only the defect of the dual solution $\phi - \pi \Delta x \phi$ enters the error and the $\phi$ is constant in this region, there is no contribution to the error in this area and the cells do not need to be refined.

Figure 9.14 (d) shows the performance of the different mesh adaptation strategies. After an initial phase with similar performance, the curves separate clearly. With uniform convergence, only a rate of 1.0 is achieved, which corresponds to a rate of 2.0 with respect to the mesh width. For the other two strategies, the convergence rate is roughly doubled. Even though it does not manifest itself in the rate, but there is a distinct advantage of goal-oriented adaptivity over residual-based adaptivity. With the same number of space-time elements, the error is more than an order of magnitude smaller.

The functional error, the estimated error, and the error bound is shown in Figure 9.15. Even though the problem is nonlinear, the error estimate agrees quite well with the actual error in all cases. However, the error bound does not follow the error and stagnates at roughly $10^{-2}$. This is to be contrasted with the linear advection problems, where the error bound was consistently larger than the error, but still converged to zero with roughly the same rate. Thus, in the Burgers’ problem cancellation (due to the shock) is intrinsic. Nevertheless, even though the adaptivity is based on the error bound rather than the error estimate, with goal-oriented adaptivity the right cells are refined and a high-order of accuracy for the functional can be achieved.

Figure 9.15.: Convergence for the Burgers’ problem using $p = 2$. 

136
9.5. Goal-oriented adaptivity

9.5.6. Sod shock tube

As a last one dimensional example, we consider the Sod shock tube, see Section 5.2.3. We choose the following target functional

\[
J(V) = \int_{t_a}^{t_b} \int_{x_a}^{x_b} \Psi(t, t_a, t_b) \Psi(x, x_a, x_b) \rho(V) \, dx \, dt
\]  

(9.23)

with the support \([x_a, x_b] \times [t_a, t_b] = [0, 2.5] \times [1.5, 2]\). The functional is nonlinear as the density \(\rho\) depends in a nonlinear way on the entropy variables \(V\).

![Figure 9.16: Mesh with 12550 elements and corresponding solution for the Sod shock tube using \(p = 2\).](image)

Figure 9.16.: Mesh with 12550 elements and corresponding solution for the Sod shock tube using \(p = 2\).

Also here we start with 4 space-time elements and refine 12 times. The mesh, the solution
9. Adaptivity

and the dual solution are shown in Figure 9.16. The mesh is refined towards the shock, the contact discontinuity, and the beginning and end of the rarefaction. However, this refinement only takes place in region of dependency of the support of $\mathcal{J}$. The other parts of the domain are almost not refined at all. Naturally this also carries over to the approximate solution. The contact discontinuity is captured quite well and the shock is sharply resolved where it is needed.

The convergence is compared with uniform and residual-based refinement in Figure 9.16 (d). For uniform refinement the convergence rate is only 0.5 which corresponds to a value 1.0 with respect to the mesh width. Thus, the convergence is rather slow. However, the rate is roughly doubled for the residual-based and goal-oriented adaptivity. Although the rates for these two strategies are almost identical, goal-oriented adaptivity has an advantage over residual-based adaptivity by almost an order of magnitude.

The error estimation is studied in more detail in Figure 9.17. In contrast with the previous experiments, the error estimate is not completely accurate. The error is slightly larger, but still it follows the estimate approximately. The reason for this discrepancy lies mainly in the nonlinearity of the problem and the approximation of mean-value linearisation by linearisation at the approximate solution. The error bound, on the other hand stagnates, as in the Burgers case. Nevertheless, the goal-oriented adaptivity works and a higher order of convergence of the functional can be achieved.

![Graphs showing convergence](image)

Figure 9.17.: Convergence for the Sod shock tube using $p = 2$.

9.5.7. Lift computation on a NACA 0012 aerofoil

We consider again the $\text{Ma}_\infty = 0.75$ aerofoil flow with an angle of attack $\alpha = 4^\circ$ described in Section 5.3.3. We are interested in the lift at the end time $T = 3.5$, which corresponds to the following functional

$$
\mathcal{J}(\mathbf{V}) = \int_{\Gamma} p(\mathbf{V}(T)) \sum_{k=1}^{2} \nu_{up}^k \nu_k d\sigma,
$$

(9.24)
where $\Gamma$ is the aerofoil surface, $\nu_{up} = (-\sin(\alpha), \cos(\alpha))$ is the upward pointing normal (with respect to the inflow velocity) and $\nu$ is outward pointing normal of the elements.

Figure 9.18.: Convergence for the NACA aerofoil lift computation for $p = 1$.

In Figure 9.18, the convergence of the lift functional is depicted. Uniform refinement is compared to residual-based adaptivity and goal-oriented adaptivity for polynomial degree $p = 1$. Residual-based adaptivity is the worst performing, uniform refinement is slightly better, and goal-oriented performs best.

Figure 9.19.: Convergence for the NACA aerofoil lift computation for $p = 1$.

The convergence is studied in more detail in Figure 9.19. The error estimate is quite accurate for all three refinement strategies. The functional error bound is almost an order of magnitude larger than the error and converges more slowly, indicating that cancellation takes place in the error estimation.

In general, goal-oriented adaptivity is quite challenging numerically, as the dual problem is badly conditioned, requiring many iterations of the linear solver. Furthermore, the non-
9. Adaptivity

linearity of the problem makes the error representation formula (9.16) only approximate, which results in less accurate error estimates and in less efficient adaptivity. Further analysis is needed to make this approach more mature. Nevertheless, the various examples have shown that goal-oriented adaptivity can reduce the number of degrees of freedom needed to obtain a certain functional error considerably, depending on the problem even by orders of magnitude.
10. Conclusions

We have developed and investigated an entropy-stable discontinuous Galerkin scheme to approximate hyperbolic systems of conservation laws. Entropy stability is obtained due to the following three ingredients: the entropy variables rather than the conservative variables are discretised, which allows to use the discretised entropy variables as test functions and therefore, to mimic the proof of entropy stability in the continuous case. The use of entropy-stable fluxes, consisting of an entropy-conservative flux together with a numerical diffusion, leads then to entropy stability. Here, the space-time formulation is essential, as it allows to obtain stability on general unstructured meshes for the fully discrete scheme. However, as (non-linear) hyperbolic conservation laws in general develop discontinuities in finite time, we need further stabilisation mechanisms. We use a streamline diffusion term that permits to control the square of the residual and a shock-capturing term that leads to diffusion proportional to the residual and thus, bounds the gradients of the solution. The combination of these features allows us to show not only stability, but also to show convergence to an entropy measure-valued solution for hyperbolic systems of conservation laws in any number of spatial dimensions (under a uniform $L^\infty$ assumption). Convergence for symmetrisable linear systems is only a corollary of this result. For scalar conservation laws, the approximate solutions are shown to converge to the entropy solution in one spatial dimension for uniformly convex fluxes, and to a weak solution for multiple spatial dimensions.

Since the scheme is implicit, the computation of the approximate solutions needs to be considered carefully. We apply a damped Newton method to solve the large nonlinear systems that result in each time step from the discretisation. The linearised systems are then solved using a Krylov solver. For efficiency reasons, they must be preconditioned. We have designed and investigated preconditioners of block Jacobi and block Gauss-Seidel type. For a fixed (moderate) CFL number, the block Jacobi preconditioner leads to a small number of GMRES iterations that is independent of the mesh width, i.e. it is efficient. For higher CFL numbers, the block Gauss-Seidel preconditioner is more efficient. The unconditional stability allows to approximate problems with multiple time scales such as flows near the incompressible limit.

In the last part of the thesis, we have exploited the space-time formulation of the scheme. Thus, we can adapt the mesh locally in space and time. To this end, we have investigated two types of adaptivity. Residual-based adaptivity relies on an ad hoc criterion using the interior and boundary residual to mark elements for refinement. Nevertheless, the resulting scheme is quite successful in reducing the global error efficiently. The second strategy, goal-oriented adaptivity, aims to compute functional of the solution (such as e.g. drag or lift) as accurately as possible. For the error estimation and the marking of the elements,
10. Conclusions

duality-based estimates are used. The computation of the solution of the dual problem is numerically challenging, but it allows to compute functionals more accurately as shown with numerical experiments.

We have depicted various examples for several different hyperbolic conservations laws. The experiments demonstrate the high-resolution and robustness properties of the method. Furthermore, the finite element formulation enables us to use unstructured meshes, and we can therefore also solve problems on complicated domains.

10.1. Future work

There are many ways along which one could proceed, both analytically and in terms of implementation. From the theoretical point of view, we have assumed the uniform $L^\infty$ bound of the approximate solutions, and this was never violated in the numerical experiments. A proof of this property (even if it is only for the scalar case) would be interesting. Furthermore, we were so far only able to show convergence to the entropy solution of scalar problems in one spatial dimension with uniformly convex fluxes as we could not estimate the projection error, and thus, we could only show consistency with one entropy. It would be interesting to see if the assumptions for the proof can be weakened or to construct counter-examples where the scheme converges to a non-entropic weak solution (which we have not observed so far).

From the implementation point of view, we have observed a growth of the number of iterations of the linear solver with the CFL number. According to our intuition, one should be able to use a multi-grid approach to precondition the linear system. However, we have not been able to construct such a method that works in multiple space dimensions as well. An improvement would be highly desirable as it will reduce the computational workload considerably. In general, the scheme is rather costly compared to a more standard (and therefore explicit) method. Exploiting either the unconditional stability, the parallel efficiency, or the space-time adaptivity makes the DG method more competitive. Nevertheless, the computational method needs to be further improved to enable for example the computation of three dimensional problems in a reasonable time.

Furthermore, one could drop the assumption that the elements are a tensor product of a spatial and a temporal element and consider more general meshes. Using anisotropic refinement, shocks could then be captured even more sharply, resulting in even higher convergence rates. However, it is a challenging task, both from the theoretical and implementational aspects. A related work would be to analyse the error estimation and the adaptivity in a rigorous manner.

There are two further extensions of the project. A first extension is to generalise the entropy-stable space-time DG scheme to convection-diffusion equations. The schemes would be based on interior penalty or on local discontinuous Galerkin methods. This would then allow us to compute viscous flows, as for example compressible Navier-Stokes flows. Another extension of the project could be to design entropy-stable space-time DG methods for non-
10.1. Future work

conservative hyperbolic systems, such as e.g. the multilayer shallow water equations. The method would presumably be based on first order entropy-stable schemes for such problems, developed in a recent paper [4].
A. Appendix

A.1. Superapproximation

In the proof of convergence to an entropy measure-valued solution, we need superapproximation estimates for the projection error. These superapproximation properties were first studied by Nitsche and Schatz in [39]. We start with the definition of the used projection.

We consider any compactly supported test function $\varphi \in (C^\infty_c(\Omega \times [0, \infty)))^m$. We denote by

$$\varphi^\Delta_x = \Pi^\Delta_x(\varphi),$$

the projection into the space $((P^p)^m)$ with $\Pi^\Delta_x|_{K \times I^n} : (H^1)^m \rightarrow ((P^p(K \times I^n))^m$ satisfying

$$\int_{I^n} \int_K \left( (\Pi^\Delta_x(\varphi))_t, \tilde{U}_V W_t \right) + \sum_{k=1}^d \frac{\Delta x^2_K}{(\Delta t)^2} \left( (\Pi^\Delta_x(\varphi))_{x_k}, \tilde{U}_V W_{x_k} \right) dxdt$$

$$= \int_{I^n} \int_K \left( \varphi_t, \tilde{U}_V W_t \right) + \sum_{k=1}^d \frac{\Delta x^2_K}{(\Delta t)^2} \left( \varphi_{x_k}, \tilde{U}_V W_{x_k} \right) dxdt,$$

$$\int_{I^n} \int_K \Pi^\Delta_x(\varphi) dxdt = \int_{I^n} \int_K \varphi dxdt$$

for all $W \in ((P^p(K \times I^n))^m$. Note that the scaling $\tilde{U}_V$ is a constant matrix as it is evaluated at the (spacetime) cell average $\tilde{V}_{n,K}$. Hence, this projection operator for the infinitely smooth function $\varphi$ satisfies the following stability and approximation properties (see [29] or other references therein),

$$\|\nabla_{xt}(\varphi^\Delta_x)\|_{L^2(K \times I^n)} \leq C \|\nabla_{xt}\varphi\|_{L^2(K \times I^n)}$$

$$\|\varphi - \varphi^\Delta_x\|_{L^2(K \times I^n)} \leq \Delta x \|\varphi - \varphi^\Delta_x\|_{L^2(\partial(K \times I^n))} \leq C \Delta x^{\frac{1}{2}} \|\nabla_{xt}\varphi\|_{L^2(K \times I^n)}.$$ (A.3)

We have the following superapproximation properties:

$$\|\nabla_{xt}(V^\Delta_x \varphi - \Pi^\Delta_x(V^\Delta_x \varphi))\|_{L^2(K \times I^n)} \leq C \Delta x |V^\Delta_x|_{H^1(K \times I^n)}$$

$$\|V^\Delta_x \varphi - \Pi^\Delta_x(V^\Delta_x \varphi)\|_{L^2(K \times I^n)} \leq C \Delta x \|V^\Delta_x\|_{L^2(K \times I^n)}$$

$$\|V^\Delta_x \varphi - \Pi^\Delta_x(V^\Delta_x \varphi)\|_{L^2(\partial(I^n \times K))} \leq C \Delta x^{1/2} \|V^\Delta_x\|_{L^2(K \times I^n)}$$

(A.4) (A.5) (A.6)

where the constant $C$ depends on $\varphi$, but not on $K \times I^n$. 

145
A. Appendix

Proof. We denote $h = \text{diam}(K \times I^n)$. Note that $h \leq C \Delta x$ because of the bound $\Delta t^n \leq C \Delta x$. We start with the Ritz-Galerkin projection estimate (see [29] or other references therein)

$$\|\varphi - \Pi_{\Delta x}^p(\varphi)\|_{H^1(K \times I^n)} \leq C h^p \|\varphi\|_{H^{p+1}(K \times I^n)}.$$  \hspace{1cm} (A.7)

Then one can show

$$\|\varphi - \Pi_{\Delta x}^p(\varphi)\|_{L^2(K \times I^n)} \leq C h^{p+1} \|\varphi\|_{H^{p+1}(K \times I^n)}$$  \hspace{1cm} (A.8)

using the Aubin-Nitsche trick. For this the convexity of the elements is essential. For the boundary, we will use the trace inequality

$$\|\varphi\|_{L^2(e)}^2 \leq C l^{-1}_e \|\varphi\|_{L^2(K \times I^n)}^2 + C l_e \|\varphi\|_{H^1(K \times I^n)}^2,$$  \hspace{1cm} (A.9)

where $e$ is an edge of the boundary of $K \times I^n$ and $l_e = \text{diam}(e)$. Note that by definition

$$l_e \leq h,$$  \hspace{1cm} (A.10)

but to use a bound of the form $h \leq C l_e$ one needs a quasiuniform mesh. We will also use inverse inequalities

$$|V_{\Delta x}^p \varphi - \Pi_{\Delta x}^p(V_{\Delta x}^p \varphi)|_{H^1(K \times I^n)} \leq C h^{-2} \|\varphi\|_{H^{m+1}(K \times I^n)}^2 \leq C h^{-2} \sum_{j=0}^p |\phi|_{W^{p+1-j, \infty}(K \times I^n)} \|V_{\Delta x}^p \varphi\|_{H^j(K \times I^n)}^2 \leq C h^{-2} \sum_{j=0}^p |\phi|_{W^{p+1-j, \infty}(K \times I^n)} \|V_{\Delta x}^p \varphi\|_{H^j(K \times I^n)}^2 \leq C h^{-2} \sum_{j=0}^p |\phi|_{W^{p+1-j, \infty}(\Omega \times [0, T])} h^{-j+1} \|V_{\Delta x}^p \varphi\|_{H^j(K \times I^n)} \leq C h^{-2} \sum_{j=0}^p |\phi|_{W^{p+1-j, \infty}(\Omega \times [0, T])} h^{-j+1} \|V_{\Delta x}^p \varphi\|_{H^j(K \times I^n)} \leq C h \|V_{\Delta x}^p \varphi\|_{H^1(K \times I^n)}$$

which proves (A.4).
In the same way, we obtain
\[
\|V^{\Delta x} \varphi - \Pi^{\Delta x}(V^{\Delta x} \varphi)\|_{L^2(K \times I^n)} \leq Ch^{p+1} |V^{\Delta x} \varphi|_{H^{p+1}(K \times I^n)}
\]
\[
\leq Ch^{p+1} \sum_{j=0}^{p+1} \|\varphi\|_{W^{p+1-j, \infty}(K \times I^n)} |V^{\Delta x} \varphi|_{H^j(K \times I^n)} \text{ (using product rule, Hölder)}
\]
\[
\leq Ch^{p+1} \sum_{j=0}^{p} \|\varphi\|_{W^{p+1-j, \infty}(K \times I^n)} h^{-j} \|V^{\Delta x}\|_{L^2(K \times I^n)} \text{ (highest degree vanishes)}
\]
\[
\leq Ch^{p+1} \sum_{j=0}^{p} \|\varphi\|_{W^{p+1-j, \infty}(\Omega \times [0,T])} h^{-j} \|V^{\Delta x}\|_{L^2(K \times I^n)} \text{ (inverse inequality)}
\]
\[
\leq Ch^{p+1} \sum_{j=0}^{p} \|\varphi\|_{W^{p+1-j, \infty}(\Omega \times [0,T])} h^{-j} \|V^{\Delta x}\|_{L^2(K \times I^n)} \text{ (using maximum)}
\]
\[
\leq Ch \|V^{\Delta x}\|_{L^2(K \times I^n)} \text{ (absorbing into const)} \quad (A.13)
\]

which proves (A.5).

For the boundary, we estimate
\[
\|V^{\Delta x} \varphi - \Pi^{\Delta x}(V^{\Delta x} \varphi)\|_{L^2(\partial)}^2 \leq Cl_e^{-1} \|V^{\Delta x} \varphi - \Pi^{\Delta x}(V^{\Delta x} \varphi)\|_{L^2(K \times I^n)}^2 + Cl_e \|V^{\Delta x} \varphi - \Pi^{\Delta x}(V^{\Delta x} \varphi)\|_{H^1(K \times I^n)}^2
\]
\[
\leq Cl_e^{-1} h^2 \|V^{\Delta x}\|_{L^2(K \times I^n)}^2 + Cl_e h^2 \|V^{\Delta x}\|_{H^1(K \times I^n)}^2 \text{ (using (A.13) and (A.12))}
\]
\[
\leq Cl_e^{-1} h^2 \|V^{\Delta x}\|_{L^2(K \times I^n)}^2 + Cl_e h^2 h^{-2} \|V^{\Delta x}\|_{L^2(K \times I^n)}^2 \text{ (inverse inequality)}
\]
\[
\leq Ch \left( \frac{h}{l_e} + \frac{l_e}{h} \right) \|V^{\Delta x}\|_{L^2(K \times I^n)}^2 \quad (A.14)
\]

which proves (A.6) by combining this estimate on all edges of $K \times I^n$. \qed
Bibliography


Bibliography


Bibliography


