HIGH-ORDER ACCURATE ENTROPY STABLE NUMERICAL
SCHEMES FOR HYPERBOLIC CONSERVATION LAWS

A dissertation submitted to
ETH ZÜRICH

for the degree of
Doctor of Sciences

presented by
ULRIK SKRE FJORDHOLM
M.Sc. Computational Science and Engineering, University of Oslo
born 1 December 1985
citizen of Norway

accepted on the recommendation of
Prof. Dr. Siddhartha Mishra, ETH Zürich, examiner
Prof. Dr. Eitan Tadmor, University of Maryland, co-examiner
Prof. Dr. Chi-Wang Shu, Brown University, co-examiner

2013
Abstract

The role of entropy for the stability of hyperbolic conservation laws is well-understood, and to some extent, also for first- and second-order accurate numerical schemes. The stability of higher-order accurate schemes, however, is to a large degree an open problem. In this thesis we adopt the framework of entropy stability as a design principle. We combine entropy conservative fluxes with appropriate diffusion operators, and to obtain high order accuracy, we consider diffusion operators using the ENO reconstruction procedure. We show that the ENO procedure satisfies the so-called sign property, which ensures entropy stability of our (arbitrarily) high-order accurate finite difference scheme. Moreover, we pose and argue for a conjecture on the total variation of the ENO reconstruction, which would imply a weak total variation bound for our scheme. For hyperbolic systems, the reconstruction is performed using a novel, computationally efficient characteristic-wise decomposition. The scheme is easily generalized to multi-dimensional systems on Cartesian meshes.

To study the convergence properties of our scheme for scalar equations, we consider the framework of compensated compactness. Under the assumption that our conjecture on the total variation bound of ENO holds, we prove convergence to the entropy solution. For multi-dimensional hyperbolic systems we are able to prove convergence to an entropy measure-valued solution.

The robustness, accuracy and computational efficiency of the scheme are demonstrated in a series of numerical experiments.
Zusammenfassung


Die Robustheit, Genauigkeit und Effizienz der Verfahren werden in verschiedenen numerischen Experimenten demonstriert.
Acknowledgements

Of all the people who have helped me through the past 3.5 years, I wish to mention three in particular. First, I am very grateful to Sidharth Mishra for all that he has taught me and for always having time for a discussion, math-related or not.
Second, I thank Eitan Tadmor for generously sharing from his immense experience, for his keen interest and enthusiasm, and for his warm hospitality in Maryland.
Finally, thank you, Ingeborg, for your immense patience with me. Your support has meant more to me than I am able to convey here.
# Contents

Abstract iii  
Zusammenfassung v  
Acknowledgements vii  
Introduction xi  
Notation xiv  

Chapter 1. Hyperbolic conservation laws  
1.1. Basic notions and examples 1  
1.2. Entropy pairs and the entropy condition 5  
1.3. Existence and uniqueness for scalar conservation laws 9  
1.4. Compensated compactness 10  
1.5. Existence and uniqueness for systems of conservation laws 13  
1.6. Measure-valued solutions 13  
1.7. Summary and outlook 14  

Chapter 2. Entropy stable methods of first order  
2.1. Finite volume and finite difference methods 18  
2.2. Entropy conservative and entropy stable methods 18  
2.3. Entropy stable methods for scalar equations 21  
2.4. Entropy conservative schemes for systems 23  
2.5. Entropy stable schemes for systems 25  
2.6. Summary 27  

Chapter 3. Convergence theory for finite volume methods  
3.1. Strong convergence of finite volume schemes 30  
3.2. Multi-dimensional measure-valued solutions of systems 34  
3.3. Convergence of entropy stable methods 38  

Chapter 4. Higher-order accurate entropy stable methods  
4.1. High-order entropy conservative methods 39  
4.2. High-order entropy stable methods for scalar equations: the ELW scheme 40
4.3. Reconstruction based entropy stable schemes: scalar equations 42  
4.4. Reconstruction based entropy stable schemes: systems of equations 44  
4.5. Summary and outlook 47  

Chapter 5. ENO stability properties 49  
5.1. The ENO reconstruction method 50  
5.2. The ENO sign property 52  
5.3. Upper bound on jumps 57  
5.4. The ENO interpolation method 60  
5.5. A conjecture on the total variation of ENO reconstruction 64  

Chapter 6. The TECNO scheme 67  
6.1. The TECNO scheme for scalar conservation laws 67  
6.2. The TECNO scheme for systems of equations 69  
6.3. Numerical experiments: scalar conservation laws 70  
6.4. Systems of conservation laws 73  

Chapter 7. Conclusions 83  
Future work 83  

Appendix A. Young measures 85  

Bibliography 87
Introduction

A large class of phenomena in fields such as meteorology, oceanography, aerodynamics and plasma physics can be modelled by conservation laws. If \( u(x, t) \) denotes the concentration of some quantity, such as gas in a container or cars on a highway, then the net change of the total amount of \( u \) in a domain \( D \) can be expressed in terms of the integral equation

\[
\frac{d}{dt} \int_D u(x, t) \, dx + \int_{\partial D} f(x, t) \cdot n(x) \, d\sigma(x) = \int_D g(x, t) \, dx.
\]

Here, \( f(x, t) \) is a function specifying the net flux through a point \( x \) at time \( t \), and \( g(x, t) \) specifies the production (or destruction) of the quantity at \( x, t \). If \( g \equiv 0 \), then there is no net overall change of the quantity \( u \), and the equation above states that \( u \) is a conserved variable – it can only change due to flux through the boundary of \( D \). In many cases, the flux function \( f \) can be written as a function of \( u \). Applying the divergence theorem to the equation, we obtain the conservation law

\[
u_t + \text{div}(f(u)) = 0.
\]

Supplying the partial differential equation with appropriate initial data,

\[
u(x, 0) = u_0(x),
\]

we obtain a Cauchy problem for \( u \).

An important subclass of such equations are hyperbolic conservation laws. Roughly speaking, a conservation law is hyperbolic if information travels at a finite speed. Thus, contrary to parabolic partial differential equations, local changes in the solutions of hyperbolic conservation laws have only local consequences.

Another fundamental difference from parabolic PDE is the appearance of discontinuities even for smooth initial data. Once discontinuities are present, the differential equation is no longer meaningful in the classical sense. By interpreting the equation in a weaker sense, we allow for discontinuous solutions, but also allowing for the existence of multiple solutions to the same Cauchy problem. Thus, to single out a solution we must impose some additional selection criterion. For hyperbolic conservation laws this comes in the form of an entropy
condition, which asserts that there must be no production of entropy\(^1\). Along with some additional conditions on the structure of \( f \), the entropy condition implies uniqueness of solutions to the conservation law. This unique solution is called the entropy solution \([GR91]\).

Since the start of serious research on hyperbolic conservation laws in the 1940s, theory and numerics have gone hand in hand. Indeed, several of the fundamental existence results for conservation laws were obtained through numerical approximation \([Gli65, Lax71, BCP00]\). Early on it was shown that numerical schemes must in certain respects be consistent, not only with the conservation law itself, but also with respect to the entropy condition, in order to attain convergence to the entropy solution \([Lax71]\). Such schemes were later termed entropy stable. The study of entropy stable schemes is the main topic of this thesis.

By the start of the 1980s, the convergence theory for first-order accurate numerical schemes for scalar conservation laws was well-developed \([CM80]\). Through the 1980s, the focus shifted towards developing schemes with a higher (than first) order of accuracy. Among the many different approaches are the ENO and WENO methods, which combine stable first-order schemes with higher-order reconstruction methods, resulting in an overall high-order accurate scheme \([HEOC87, SO88, LOC94, JS96]\). Although these schemes have enjoyed widespread popularity in the industrial and research communities, there is a lack of rigorous stability analysis of high-order accurate schemes \([Shu99, Section 4.3]\).

The main purpose of this thesis is to address this lack of stability results for high-order accurate schemes. We construct high-order accurate entropy stable finite difference schemes for both scalar conservation laws and systems of equations. The entropy dissipation rate of these schemes gives a weak bound on the gradient of the solution; this allows us to prove strong convergence to the entropy solution for scalar equations and convergence to an entropy measure-valued solution for systems of equations.

The following is an outline of the thesis. The introductory Chapter 1 gives a brief introduction to the theory of hyperbolic conservation laws, with emphasis on the role of entropy conditions. We give an introduction to the method of compensated compactness, and discuss the concept of measure-valued solutions as introduced by DiPerna \([DiP85]\).

In Chapter 2 we discuss the theory of first-order accurate, entropy stable finite difference methods for both scalar conservation laws and systems of hyperbolic conservation laws. We demonstrate the construction of first-order accurate entropy conservative and entropy stable methods, which will serve as building blocks for the higher-order accurate methods in subsequent chapters.

Chapter 3 deals with general convergence theory for numerical approximations to hyperbolic conservation laws. By a careful analysis of previous work, we derive sufficient conditions, as weak as possible, for the (strong) convergence of numerical methods to the entropy solution. This is then contrasted with less restrictive conditions for convergence to entropy measure-valued solutions. We apply the theory discussed here to prove convergence

---

\(^1\) According to the second law of thermodynamics, entropy should never decrease, but due to a difference in sign, we require (mathematical) entropy never to increase.
of the schemes from Chapter 2 to the entropy solution for one-dimensional convex scalar conservation laws.

In Chapter 4 we discuss higher-order extensions of the entropy stable schemes from Chapter 3. As a first example we introduce a high-order accurate, entropy stable scheme which converges for scalar equations, but which exhibits oscillations around shocks. To ameliorate this, we explore finite difference schemes using nonoscillatory reconstruction procedures and derive a sufficient condition for entropy stability for both scalar equations and systems. As we show in Chapter 5, the ENO reconstruction procedure satisfies precisely this condition.

Chapter 6 is the main contribution of the thesis. We apply the theory developed in Chapter 4 to obtain nonoscillatory, high-order accurate, entropy stable schemes, which we refer to as TECNO schemes. Using the convergence theory from Chapter 3, we prove that the scheme converges to the entropy solution for scalar equations, and converges to a measure-valued solution for systems of equations. Finally, we compare our scheme to existing schemes in a series of numerical experiments.
Notation

We introduce here notation which is used throughout the thesis. The canonical inner product on $\mathbb{R}^n$ is $x \cdot y = x^T y = \sum_{k=1}^{n} x_k y_k$, with the associated norm $|x| = \sqrt{x \cdot x}$. The subset $B_r(x)$ of $\mathbb{R}^n$ denotes the open ball with center $x \in \mathbb{R}^n$ and radius $r > 0$, and the unit hypersphere in $\mathbb{R}^n$ is denoted $S^{n-1} := \{ x \in \mathbb{R}^n : |x| = 1 \}$. We write for $\mathbb{R}^+$: $[0, \infty)$ the nonnegative real axis. The identity function on a space $X$ is denoted $\text{id}_X(x) := x$. The subscript is dropped when the domain is clear from the context. We denote the indicator function of a set $U$ by $1_U(x) = \begin{cases} 1 & \text{if } x \in U, \\ 0 & \text{if } x \notin U. \end{cases}$ If $A \subset \mathbb{R}^n$ is Lebesgue measurable then we denote its Lebesgue measure by $\text{meas}(A)$.

We denote by $C^k(\mathbb{R}^n, \mathbb{R}^m)$ the set of $k$ times continuously differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. If $m = 1$ then we write $C^k(\mathbb{R}^n) := C^k(\mathbb{R}^n, \mathbb{R})$. The space of smooth, compactly supported functions on $U \subset \mathbb{R}^n$ is denoted $\mathcal{D}(U) := C^\infty_c(U)$. The dual space $\mathcal{D}'(U)$ is the space of distributions on $U$. The Jacobian of a function $f \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ is denoted by a prime, $f'(u) = \frac{df}{du}(u)$. Partial derivatives are denoted by subscripts, $u_x = \frac{\partial u}{\partial x}$. If $\{u_k\}_{k \in \mathbb{N}} \subset L^\infty(\mathbb{R}^n)$ is a weak-$\ast$ convergent sequence then we denote its limit by $\overline{u} := \text{w-}\ast \lim u_k$.

For symmetric matrices $A, B \in \mathbb{R}^{n \times n}$, we mean by $A \preceq B$ that the matrix $B - A$ is positive semi-definite (positive definite if there is strict inequality). A function $\eta \in C^2(\mathbb{R}^n)$ is convex if $\eta''(u) \geq 0$ for all $u$, strictly convex if $\eta''(u) > 0$ for a.e. $u$ and uniformly convex if $\eta''(u) \geq m$ for all $u$ for an $m > 0$.

For a grid function $\{u_i : i \in \mathbb{Z}\}$ we denote its jump and average by $[u]_{i+i/2} := u_{i+1} - u_i$ and $\overline{u}_{i+i/2} = \frac{1}{2}(u_i + u_{i+1})$. 
CHAPTER 1

Hyperbolic conservation laws

In this chapter we summarize the important existence, uniqueness and regularity results related to hyperbolic conservation laws. Because of the appearance of discontinuities in solutions of such equations, the differential equation is not meaningful in the classical sense. We will look at two different, weaker solution concepts: weak solutions and measure-valued solutions. To obtain uniqueness we must introduce entropy conditions that single out a “physically relevant” solution.

In Section 1.4 we give a short review of compensated compactness, a method for showing strong convergence of a sequence of approximate solutions. The framework introduced in this section will be used in Chapter 3 to show convergence of general finite volume methods, and in Chapter 4 to show convergence of the high-order accurate, entropy stable methods introduced there. In Section 1.6 we study the much weaker concept of measure-valued solutions, and argue why this should be viewed as a natural solution concept for general nonlinear evolution equations. We end the chapter with an overview of the rest of the thesis.

Most of the theorems in this chapter will be stated without proofs, with references where appropriate.

1.1. Basic notions and examples

We consider systems of conservation laws,

\[
\begin{align*}
  u_t + \text{div}(f(u)) &= 0, && (x, t) \in \mathbb{R}^d \times \mathbb{R}_+, \\
  u(\cdot, 0) &= u_0,
\end{align*}
\]

(1.1)

where \( u : \mathbb{R}^d \times \mathbb{R}_+ \rightarrow U \) is the vector of conserved variables, taking values in some connected, nonempty set \( U \subset \mathbb{R}^N \). The function \( f = (f^1, \ldots, f^d) : U \rightarrow \mathbb{R}^{N \times d} \) is the flux function, which is assumed to be at least \( C^2 \) on \( U \). The conservation law is hyperbolic if for all \( u \in U \) and \( \xi \in S^{d-1} \), the Jacobian \( \xi \cdot f'(u) = \sum_{n=1}^d \xi^a \frac{d}{du}(u) \) has \( N \) real eigenvalues \( \lambda_1(u) \leq \cdots \leq \lambda_N(u) \) with linearly independent eigenvectors \( r_1(u), \ldots, r_N(u) \) (the dependence on \( \xi \) is suppressed). It is strictly hyperbolic if the eigenvalues are distinct, \( \lambda_1(u) < \cdots < \lambda_N(u) \). An eigenpair \((\lambda_k, r_k)\) of \( \xi \cdot f' \) is referred to as a wave family. The \( k \)-th wave family \((\lambda_k, r_k)\) is genuinely nonlinear if \( \lambda'_k(u) \cdot r_k(u) \neq 0 \) for all \( u \in U \), and is linearly degenerate if \( \lambda'_k(u) \cdot r_k(u) = 0 \) for all \( u \in U \). If \( \lambda'_k(u) \cdot r_k(u) \) vanishes only in parts of \( U \) then the conservation law is termed nonconvex; such conservation laws will not be considered here. We refer to [GR91] for a thorough introduction to the theory of hyperbolic conservation laws.
It is easy to see that every scalar \((N = 1)\) conservation law is hyperbolic; it is genuinely nonlinear iff \(f''(u) \neq 0\) for all \(u \in U\), and is linearly degenerate iff \(f\) is affine. A linear system of equations with flux \(f(u) = Au\) is hyperbolic if \(A\) is diagonalizable with real eigenvalues. All wave families of linear systems are linearly degenerate.

1.1.1. Examples of hyperbolic conservation laws. We give here some examples of hyperbolic conservation laws. The simplest example of a scalar conservation law is the linear advection equation,

\[(1.2) \quad u_t + au_x = 0,\]

whose unique solution is \(u(x, t) = u_0(x - at)\). A common example of a nonlinear scalar conservation law is Burgers’ equation,

\[(1.3) \quad u_t + \left(\frac{u^2}{2}\right)_x = 0.\]

The wave equation in divergence form is the linear system of conservation laws

\[(1.4) \quad \begin{pmatrix} u^1_x \\ u^2_x \end{pmatrix} + \begin{pmatrix} cu^2_x \\ cu^1_x \end{pmatrix} = 0,\]

where \(c \in \mathbb{R}\) is a constant. The system can be written in the form \(u_t + Au_x = 0\) for a diagonalizable matrix \(A = R\Lambda R^{-1} \in \mathbb{R}^{2 \times 2}\). Defining the characteristic variable \(\hat{u} := R^{-1}u\), we have the characteristic decomposition \(\hat{u}_t + \Lambda \hat{u}_x = 0\). Each equation in this system is a linear advection equation, which has a unique solution, and hence the system has a unique solution.

A simple but important example of a one-dimensional, nonlinear hyperbolic system of conservation laws \((d = 1, N > 1)\) is the shallow water equations, which model a body of water under the assumption that the horizontal length scales are much larger than the vertical length scale. The velocity \(w\) of the water can then be approximated by a depth-averaged velocity, \(w(x, z) \approx w(x)\). Denoting by \(h(x)\) the height (or depth) of the water column at \(x\) and by \(g\) the gravitational constant, one obtains the system

\[(1.5) \quad \begin{pmatrix} h \\ hw \end{pmatrix}_t + \left(\frac{hw}{2gh^2 + hw^2}\right)_x = 0.\]

To model a fully three-dimensional body of water, a second spatial dimension may be added [LeV02]. Real-world effects such as a nontrivial bottom topography, the Coriolis force, wind etc. may be incorporated by adding appropriate terms.

Another important example is the Euler equations of gas dynamics. Denoting by \(\rho\) the gas density, \(w\) its velocity, \(E\) total energy, and \(p\) the pressure, this system of equations is given by

\[(1.6) \quad \begin{pmatrix} \rho \\ \rho w \\ E \end{pmatrix}_t + \left(\begin{pmatrix} \rho w^2 + p \\ (E + p)w \end{pmatrix}\right)_x = 0.\]
The pressure \( p \) is given by an equation of state, which for ideal gases is

\[
E = \frac{p}{\gamma - 1} + \frac{1}{2} \rho w^2,
\]

where \( \gamma > 1 \) is the adiabatic constant.

### 1.1.2. The formation of shocks

We give an example that illustrates the main difficulty that arises when dealing with hyperbolic conservation laws.

Consider the scalar, one-dimensional conservation law (1.1) \((d = N = 1)\), and let the initial data \( u_0 \in C^1(\mathbb{R}) \) and flux function \( f \in C^2(\mathbb{R}) \) be given. We will show the local existence of a classical solution of the Cauchy problem (1.1) using the method of characteristics. Assume for the moment that a classical solution \( u \) of (1.1) is given. For \((x, t) \in \mathbb{R} \times \mathbb{R}_+\), let the backwards characteristic \( \xi(s) \) be defined by

\[
\xi'(s) = f'(u(\xi(s), s)) \quad 0 < s < t
\]

\[
\xi(t) = x.
\]

Using (1.1) we see that\( \frac{d}{ds}u(\xi(s), s) = 0 \), and hence \( u(x, t) = u_0(\xi(0)) \). The characteristic \( \xi \) can therefore be solved for explicitly as \( \xi(s) = \xi(0) + sf'(u(\xi(s), s)) \), so it follows that

\[
u(x, t) = u_0(\xi(0)) = u_0(\xi(s) + sf'(u(\xi(s))) = u_0(\xi(s) + sf'(u(x, t))).
\]

Define \( \Phi((x, t), u) := u - u_0(\xi(s) + sf'(u(x, t))) \). Since

\[
\Phi((x_0, 0), u_0(x_0)) = 0 \quad \text{and} \quad \frac{\partial \Phi}{\partial u}((x_0, 0), u_0(x_0)) = 1
\]

for every \( x_0 \in \mathbb{R} \), the implicit function theorem gives the existence of an open set \( U \subset \mathbb{R}^2 \) containing \((x_0, 0)\) and a \( C^1 \) function \( u : U \rightarrow \mathbb{R} \) such that

\[
u(x, t) = u_0(\xi(s) + sf'(u(x, t))) \quad \forall (x, t) \in U.
\]

This function is a solution of (1.1); indeed, suppressing function arguments for the moment,

\[
u_t + f(u)_x = \frac{-f'(u)u'_0}{1 + tu'_0f''(u)} + \frac{f''(u)u'_0}{1 + tu'_0f''(u)} = 0 \quad \forall (x, t) \in U
\]

for small \( t \). Given any bounded interval \([a, b] \subset \mathbb{R} \) we can repeat this process on all points \( x_0 \) in a dense, countable subset of \([a, b] \) and find that there is a \( T > 0 \) such that a classical solution of (1.1) exists in \([a, b] \times [0, T] \).

Next, we demonstrate that in all but the most trivial cases, we cannot hope for the \textit{global} existence of a classical solution. Define the characteristic \( \xi(x, t) = x + tf'(u_0(x)) \). From the discussion above we know that

\[
u_s(\xi(x, t), t) = \frac{u'_0(x)}{1 + tu'_0(x)f''(u_0(x))} = \frac{u'_0(x)}{1 + t \frac{d}{dx}f'(u_0(x))}.
\]
Hence, if the function \( x \mapsto \frac{d}{dx} f'(u_0(x)) \) is ever negative, then

\[
\lim_{t \uparrow t_0} u_\xi(x, t) = \infty, \quad t_0 := -\frac{1}{\min_{x \in \mathbb{R}} \left( \frac{d}{dx} f'(u_0(x)) \right)},
\]

and \( u \) is no longer differentiable.

### 1.1.3. Weak solutions.

It is clear that we cannot expect solutions to be defined in the classical sense. Instead, we interpret the differential equation in the sense of distributions.

**Definition 1.1.** A function \( u \in L^\infty(\mathbb{R}^d \times [0, \infty), \mathbb{R}^N) \) is a weak solution of (1.1) if it satisfies (1.1) in \( D'(\mathbb{R}^d \times [0, \infty)) \), that is, if

\[
(1.7) \quad \int_{\mathbb{R}^d} u \varphi_t + f(u) \text{div}(\varphi) \, dx + \int_{\mathbb{R}^d} u_0(x) \varphi(x, 0) \, dx = 0 \quad \forall \varphi \in D(\mathbb{R}^d \times [0, \infty)).
\]

Since the \( x \)-axis \( \mathbb{R}^d \times \{0\} \) as a subset of \( \mathbb{R}^d \times [0, \infty) \) has measure 0, the expression \( u(x, 0) = u_0(x) \) is not necessarily meaningful. However, it may be shown that weak solutions admit the initial data in the following weak sense:

\[
\lim_{T \to 0} \frac{1}{T} \int_0^T \int_{\mathbb{R}^d} (u(x, t) - u_0(x)) \psi(x) \, dx \, dt = 0 \quad \forall \psi \in D(\mathbb{R}^d).
\]

The following result provides an easy way of checking whether a given function is a weak solution.

**Theorem 1.2.** Let \( u : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^N \) be piecewise differentiable with jump discontinuities along \( d \)-dimensional smooth submanifolds of \( \mathbb{R}^d \times \mathbb{R}^+ \) with normals \( (n, -\sigma) \in S^{d-1} \times \mathbb{R} \). Then the following are equivalent:

1. \( u \) is a weak solution of (1.1).
2. \( u \) satisfies (1.1) at its points of differentiability, and along each surface of discontinuity it satisfies the Rankine-Hugoniot jump condition

\[
(1.8) \quad n \cdot \|f(u)\| - \sigma \|[u]\| = 0.
\]

where

\[
|[u]| = u^+(x, t) - u^-(x, t) \quad \text{and} \quad u^\pm(x, t) = \lim_{h \downarrow 0} u(x \pm nh, t).
\]

We call the type of discontinuities as in the previous theorem shocks. In the particular case \( d = 1 \), the types of discontinuities allowed in Theorem 1.2 are curves \( (\gamma(t), t) \) in the \( x-t \) plane, and (1.8) reads

\[
\|[f(u)]\| - \sigma \|[u]\| = 0,
\]

where \( \sigma(t) = \gamma'(t) \).
Example 1.3. Using the Rankine-Hugoniot condition it is not hard to come up with weak solutions. As an example, consider Burgers’ equation (1.3). We let the initial data be given by

\[ u_0(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x. \end{cases} \]

A Cauchy problem with piecewise constant data separated by a single jump discontinuity, like the above, is called a Riemann problem. A continuous solution of this Riemann problem is given by

\[ u_1(x, t) = \begin{cases} 0 & \text{if } x < 0 \\ x/t & \text{if } 0 \leq x < t \\ 1 & \text{if } t \leq x. \end{cases} \]

However, using the Rankine-Hugoniot conditions (1.8), we can easily come up with more weak solutions, for instance

\[ u_2(x, t) = \begin{cases} 0 & \text{if } x < t/2 \\ 1 & \text{if } t/2 \leq x \end{cases} \quad \text{and} \quad u_3(x, t) = \begin{cases} 0 & \text{if } x < t/4 \\ 1/2 & \text{if } t/4 \leq x < 3t/4 \\ 1 & \text{if } 3t/4 \leq x. \end{cases} \]

\( u_2 \) and \( u_3 \) trivially satisfy the conservation law where it is constant, and along their lines of discontinuity \( x = t/2 \) and \( x = t/4, x = 3t/4 \), respectively, the Rankine-Hugoniot condition (1.8) holds, with shock speeds \( \sigma = 1/2 \) and \( \sigma = 1/4, \sigma = 3/4 \), respectively.

1.2. Entropy pairs and the entropy condition

The example in the previous section demonstrates that the distributional interpretation of the differential equation is too weak to obtain uniqueness. It is clear that we must enforce additional constraints to single out a unique solution. These constraints come in the form of entropy conditions.

Definition 1.4. A pair of functions \((\eta, q)\) with \( \eta \in C^2(\mathbb{R}^N) \) and \( q = (q^1, \ldots, q^d) \in C^2(\mathbb{R}^N, \mathbb{R}^d) \) is called an entropy pair for (1.1) provided \( q'(u)^T = \eta'(u) \cdot f'(u) \) for all \( u \in U \), i.e. \( q^k'(u)^T = \eta'(u) \cdot (f^k)'(u) \) for \( k = 1, \ldots, d \). An entropy pair \((\eta, q)\) is convex if \( \eta \) is convex.

Taking the inner product of (1.1) with \( \eta'(u) \) and using the chain rule, we see that smooth solutions \( u \) satisfy the additional conservation law

\[ 0 = \eta'(u) \cdot (u_t + f'(u) \text{ grad}(u)) = \eta(u)_t + \text{div}(q(u)). \]

In particular, integrating (1.9) over \( \mathbb{R}^d \times [0, T] \) and assuming that \( \lim_{|x| \to \infty} q(u(x, t)) = 0 \) fast enough (say, if \( u \) has compact support), we see that

\[ \int_{\mathbb{R}^d} \eta(u(x, T)) \, dx = \int_{\mathbb{R}^d} \eta(u_0(x)) \, dx. \]

Hence, for smooth solutions, the total amount of entropy is constant in time.
The derivation of (1.9) of course only holds when \( u \) is differentiable. However, we may generalize it to the case of a piecewise differentiable function with jump discontinuities, using the same argument as in the proof of the Rankine-Hugoniot condition. If \( u \) has a discontinuity along a surface with normal \( (n, -\sigma) \in S^{d-1} \times \mathbb{R} \), then the right-hand side of (1.9) is a Dirac \( \delta \)-mass with weight

\[
\eta(u)_t + \text{div}(q(u)) = \frac{1}{\sqrt{1 + \sigma^2}} (n \cdot [q(u)] - \sigma [\eta(u)]).
\]

In the one-dimensional case \( d = 1 \), the entropy production takes the form

\[
\frac{1}{\sqrt{1 + \sigma^2}} ([q(u)] - \sigma [\eta(u)]), \quad \sigma := \gamma'(t)
\]

whenever \( u \) has a discontinuity along a curve \((\gamma(t), t)\).

Example 1.5. Returning to the example from the previous section, we see that \( u_1 \) satisfies the entropy identity (1.9) everywhere for any entropy pair, as it is absolutely continuous. For \( u_2 \) and \( u_3 \), we use the convex entropy pair \( \eta(u) = \frac{u}{2}, q(u) = \frac{u^3}{3} \) for concreteness. Then

\[
\eta(u_2)_t + q(u_2)_x \big|_{x=1/2} = \frac{1}{\sqrt{\frac{5}{3}}} \left( \frac{1}{3} - \frac{1}{2} \right) = \frac{1}{6 \sqrt{5}} > 0,
\]

\[
\eta(u_3)_t + q(u_3)_x \big|_{x=1/4} = \eta(u_3)_t + q(u_3)_x \big|_{x=3/4} = \frac{1}{96} > 0.
\]

Thus, \( u_2 \) and \( u_3 \) produce entropy across their discontinuities. Indeed, we shall soon show that \( u_2 \) and \( u_3 \) produce entropy with respect to any convex entropy pair.

The stability condition that we impose is that entropy never increases. In the example above, this would exclude the solutions \( u_2 \) and \( u_3 \), but not \( u_1 \).

Definition 1.6. Let \((\eta, q)\) be a convex entropy pair for (1.1). A function \( u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+) \) is admissible with respect to \((\eta, q)\) if

\[
\eta(u)_t + \text{div}(q(u)) \leq 0
\]

in \( \mathcal{D}(\mathbb{R}^d \times \mathbb{R}_+) \), that is, if

\[
\int_{\mathbb{R}^d \times \mathbb{R}_+} \eta(u) \varphi_t + q(u) \text{div}(\varphi) \, dxdt + \int_{\mathbb{R}^d} \eta(u_0(x)) \varphi(x, 0) \, dx \geq 0
\]

for all \( 0 \leq \varphi \in \mathcal{D}(\mathbb{R}^d \times \mathbb{R}_+) \). A weak solution of (1.1) is an entropy solution if it is admissible with respect to all convex entropy pairs.

Remark. Formally, we may integrate (1.10) in space and time to find that if \( |q(u)| \to 0 \) as \( |x| \to \infty \), then the total amount of entropy of an entropy solution decreases in time:

\[
\int_{\mathbb{R}^d} \eta(u(x,t)) \, dx \leq \int_{\mathbb{R}^d} \eta(u_0(x)) \, dx.
\]

If, say, \( \eta(u) = |u|^p \) \( (p \gg 1) \) is an entropy for (1.1), then (1.11) provides an a priori \( L^p \)-stability bound on \( u(\cdot, t) \) for all \( t > 0 \).
As for weak solutions, we can derive a necessary and sufficient condition for (1.10) when the function \( u \) has isolated jump discontinuities.

**Proposition 1.7** (Hopf [Hop69]). Let \( u : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^N \) be piecewise differentiable with jump discontinuities along surfaces with normals \((n, -\sigma) \in S^{d-1} \times \mathbb{R}\), and let \((\eta, q)\) be a convex entropy pair. Then the following are equivalent:

(i) \( u \) is admissible with respect to \((\eta, q)\).

(ii) \( u \) satisfies (1.9) at its points of differentiability, and along each surface of discontinuity it satisfies

\[
n \cdot \llbracket q(u) \rrbracket - \sigma \llbracket \eta(u) \rrbracket \leq 0. \tag{1.12}
\]

In the scalar, one-dimensional case, Proposition 1.7 admits a geometrical interpretation, as follows.

**Theorem 1.8** (Oleinik [Ole57]). Consider the scalar, one-dimensional conservation law \((d = N = 1)\), and let \( u \) be as in Proposition 1.7. Then the following are equivalent:

(i) (1.12) is satisfied for all convex entropy pairs \((\eta, q)\).

(ii)

\[
\frac{f(u^+) - f(u^-)}{u^+ - u^-} \leq \frac{f(u) - f(u^-)}{u - u^-} \quad \text{for all } u \text{ between } u^- \text{ and } u^+. \tag{1.13}
\]

**Proof.** We apply the fundamental theorem of calculus to the quantities \( \llbracket \eta(u) \rrbracket \) and \( \llbracket q(u) \rrbracket \) and integrate by parts:

\[
\llbracket \eta(u) \rrbracket = \int_{u^-}^{u^+} \eta'(u) \, du = \eta'(u)(u - u^-)|_{u=-}^{u=+} - \int_{u^-}^{u^+} \eta''(u)(u - u^-) \, du
\]

\[
= \eta'(u^+)\llbracket u \rrbracket - \int_{u^-}^{u^+} \eta''(u)(u - u^-) \, du
\]

and

\[
\llbracket q(u) \rrbracket = \int_{u^-}^{u^+} \eta'(u)f'(u) \, du = \eta'(u)(f(u) - f(u^-))|_{u=-}^{u=+} - \int_{u^-}^{u^+} \eta''(u)(f(u) - f(u^-)) \, du
\]

\[
= \eta'(u^+)\llbracket f(u) \rrbracket - \int_{u^-}^{u^+} \eta''(u)(f(u) - f(u^-)) \, du.
\]

Hence,

\[
\llbracket q(u) \rrbracket - \sigma \llbracket \eta(u) \rrbracket = \eta'(u^+)\llbracket f(u) \rrbracket - \eta''(u)(\sigma(u - u^-) - (f(u) - f(u^-))) \, du.
\]

The first term vanishes because of the Rankine-Hugoniot condition (1.8). If the remaining integral is to be nonpositive for all convex entropies \( \eta \), then we must have

\[
\sigma(u - u^-) - (f(u) - f(u^-)) \leq 0
\]

for all \( u \) between \( u^- \) and \( u^+ \), which again using the Rankine-Hugoniot condition is precisely (1.13). \qed
(1.13) is the Oleinik E-condition. It says that if \( u^- < u^+ \) then \( f(u) \) should lie above the straight line connecting \((u^-, f(u^-))\) and \((u^+, f(u^+))\), while if \( u^- > u^+ \) then it should lie below. In the special case where \( f \) is strictly convex, we see that a shock is admissible if and only if \( u^- > u^+ \), since a convex function lies below all of its chords. This explains the results in Example 1.5, where \( u^- = 0 < u^+ = 1 \), and so any weak solution containing a shock is nonadmissible.

1.2.1. Parametrized entropy pairs. A parametrized entropy pair is a pair of functions \( \eta(u, k), q(u, k) \) for \( u, k \in \mathbb{R}^N \) such that \((\eta(\cdot, k), q(\cdot, k))\) is an entropy pair for all \( k \) and \( \eta(u, k), q(u, k) = q(k, u) \). In the scalar case \((N = 1)\), it is easily seen that
\[
(1.14) \quad \eta(u, k) = |u - k|, \quad q(u, k) = \text{sgn}(u - k)(f(u) - f(k))
\]
is a parametrized entropy pair. This is the famous Kruzkov entropy pair. Kruzkov used this, along with his celebrated doubling of variables argument, to show \( L^1 \)-stability of the entropy solution of (1.1) with respect to the initial data.

It is straightforward to generalize the doubling of variables argument to systems of conservation laws \((N > 1)\); see [Tad03, Theorem 2.2]. However, a general construction of families of entropy pairs is only available for certain symmetric systems [Tad87b] and 2 \( \times \) 2 systems [Lax71, Section 3].

1.2.2. Entropy variables. The function \( u \mapsto \eta'(u) \) has appeared several times in the previous sections. This function is called the entropy variable and is denoted by \( \varphi(u) := \eta'(u) \). If \( \eta \) is strongly convex, \( \eta'' > 0 \), then \( u \mapsto \varphi(u) \) is invertible, and we denote its inverse by \( u(\varphi) \). The mapping \( u \mapsto \varphi \) then induces a change of variables, and we can pose the conservation law in entropy variables:
\[
u(\varphi)_t + f(u(\varphi))_x = 0.
\]
This symmetrizes the system, in the sense that the matrix \( u'(\varphi) = (\eta''(u(\varphi)))^{-1} \) is symmetric positive definite, and \( \frac{d}{du} f(u(\varphi)) \) is symmetric. To see the latter, let the entropy potential be the scalar function defined by \( \psi(\varphi) = \varphi \cdot f(u(\varphi)) - q(u(\varphi)) \). It is easy to verify that \( \psi'(\varphi) = f(u(\varphi)) \), and hence \( \frac{d}{du} f(u(\varphi)) = \psi''(\varphi) \) is symmetric. This symmetrization was first observed by Godunov [God61], and was later developed by Mock [Moc80].

1.2.3. Examples of entropies. For a general scalar conservation law \((N = 1)\), all convex functions \( \eta \) give rise to an entropy pair \((\eta, q)\) by defining
\[
q(u) = \int^u \eta'(s)f'(s) \, ds.
\]
Thus, scalar conservation laws are endowed with a rich family of entropies, an indispensable tool in the stability analysis of these equations. This extends to certain linear systems of conservation laws. For the wave equation (1.4), we can let e.g. \( \eta(u) = (u_1)^2 + (u_2)^2 \), with corresponding entropy flux \( q(u) = u^T Au = 2cu_1u_2 \).
The selection of entropies for nonlinear general systems of conservation laws is much more limited. For the shallow water equations, the only available entropy is given by the total energy of the solution,

\[ \eta(u) = \frac{1}{2} (hw^2 + gh^2), \quad q(u) = \frac{1}{2} hw^2 + gwh. \]

The corresponding entropy variable \( v := \eta'(u) \) and entropy potential \( \psi(v) := v \cdot f(u) - q(u) \) are given by

\[ v = \left( gh - \frac{w^2}{2} \right), \quad \psi(v) = \frac{1}{2} gwh. \]

For the Euler equations (1.6), the relevant entropy pair is given by the entropy of the solution,

\[ \eta(u) = -\rho s, \quad q(u) = -\rho ws, \]

where the thermodynamic entropy is defined by \( s := \log(p) - \gamma \log(\rho). \) The entropy variable and entropy potential are

\[ v = \begin{pmatrix} \frac{\gamma - s}{\gamma - 1} - \frac{\rho w}{2p} \\ \frac{\rho w}{p} \\ -\rho \end{pmatrix}, \quad \psi(v) = \rho w. \]

Harten [Har83] generalized this to an entire family of entropy pairs; however, only the above entropy pair simultaneously symmetrizes the Navier-Stokes equations when heat conduction is taken into account [HFM86].

### 1.3. Existence and uniqueness for scalar conservation laws

It was shown by Kruzkov in 1970 that for scalar conservation laws, there exists a unique entropy solution in the class of functions of bounded variation.

**Theorem 1.9 (Kruzkov [Kru70]).** Consider the scalar conservation law (1.1). For each \( u_0 \in L^{\infty}(\mathbb{R}^d) \), there exists a unique entropy solution \( u \) of the corresponding Cauchy problem. This function satisfies

\[ \|u(\cdot, t)\|_{L^{\infty}(\mathbb{R}^d)} \leq \|u_0\|_{L^{\infty}(\mathbb{R}^d)} \quad \text{for all } t > 0, \]

and it attains the initial data in the following \( L^1 \) sense: there is a set \( \mathcal{T} \subset [0, \infty) \) of measure zero such that for all \( x_0 \in \mathbb{R} \) and \( r > 0 \),

\[ \lim_{t \downarrow 0, t \notin \mathcal{T}} \int_{B_r(x_0)} |u(x, t) - u_0(x)| \, dx = 0. \]

If \( u, \tilde{u} : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R} \) are entropy solutions of (1.1) with initial datum \( u_0, \tilde{u}_0 \in L^{\infty}(\mathbb{R}^d) \cap L^1(\mathbb{R}^d) \), respectively, then

\[ \int_{\mathbb{R}^d} |u(x, t) - \tilde{u}(x, t)| \, dx \leq \int_{\mathbb{R}^d} |u_0(x) - \tilde{u}_0(x)| \, dx \quad \text{for all } t > 0. \]
In particular, if \( u_0 \in BV(\mathbb{R}^d) \), then

\[
TV(u(\cdot, t)) \leq TV(u_0) \quad \text{for all } t > 0.
\]

The \( L^1 \)-stability with respect to initial data stated in the previous theorem may be generalized to arbitrary generalized entropy pairs; see [Tad03, Theorem 2.2].

If we restrict ourselves to uniformly convex flux functions, then it is enough to impose admissibility with respect to a single entropy pair, as seen in the following theorem.

**Theorem 1.10** (Panov [Pan94]). Consider the scalar, one-dimensional conservation law (1.1). Assume that \( f \) is uniformly convex and \( u_0 \in L^\infty(\mathbb{R}) \), and let \( u \in L^\infty(\mathbb{R} \times \mathbb{R}_+) \) be a weak solution that is admissible with respect to a strictly convex entropy pair \((\eta, q)\). Then \( u \) is the entropy solution.

### 1.4. Compensated compactness

Kruzkov originally proved the existence part of Theorem 1.9 by showing that the solutions of a parabolic regularization of (1.1) converge strongly to an entropy solution as the diffusion coefficient goes to zero. The proof relies on a priori stability bounds on the regularized equation, most importantly a local bound on the total variation of the solution.

More generally, to show the convergence of a sequence of approximate solutions \( \{u^\varepsilon\}_{\varepsilon > 0} \), a common approach is to show that the sequence has uniformly bounded total variation, and hence, by Helly’s theorem, has a strongly convergent subsequence. However, this approach is not always practicable; most importantly for us, one cannot hope for uniform total variation bounds for high-order accurate numerical methods.

The compensated compactness method, due to Murat and Tartar [Mur78, Tar79], seeks to overcome this difficulty by imposing a much weaker bound on the gradients of the approximating sequence. More precisely, it is asserted that the entropy residual \( \eta(u^\varepsilon)_t + q(u^\varepsilon)_x \) lies in a compact subset of \( H^{-1}_{loc}(\mathbb{R} \times \mathbb{R}^+) \). Using Murat’s div-curl lemma, this gives enough control on oscillations to be able to pass to the (strong) limit \( \varepsilon \to 0 \).

In this section we give a brief introduction to the compensated compactness method, as applied to scalar one-dimensional conservation laws. The approach that we present relies on Young measures, and a short summary of Young measures is given in Appendix A.

The following result, due to Murat [Mur78], is the main result in the theory of compensated compactness.

**Lemma 1.11** (Div-curl lemma). Let \( \Omega \subset \mathbb{R}^2 \) be open. Let \( D^\varepsilon, E^\varepsilon \in L^2(\Omega, \mathbb{R}^2) \) be such that \( D^\varepsilon \to D, E^\varepsilon \to E \) weakly in \( L^2(\Omega, \mathbb{R}^2) \) as \( \varepsilon \to 0 \), and assume that

\[
\{\text{div}(D^\varepsilon)\}_{\varepsilon > 0} \quad \text{and} \quad \{\text{curl}(E^\varepsilon)\}_{\varepsilon > 0} \quad \text{are precompact in } H^{-1}(\Omega).
\]

Then

\[
(1.19) \quad \{\text{div}(D^\varepsilon)\}_{\varepsilon > 0} \quad \text{and} \quad \{\text{curl}(E^\varepsilon)\}_{\varepsilon > 0} \quad \text{are precompact in } H^{-1}(\Omega).
\]

**Then**

\[
(1.20) \quad D^\varepsilon \cdot E^\varepsilon \to D \cdot E \quad \text{in } \mathcal{D}'(\Omega).
\]

(Here, we have denoted \( \text{div}(D) = \frac{\partial D_1}{\partial y_1} + \frac{\partial D_2}{\partial y_2} \), \( \text{curl}(D) = \frac{\partial D_2}{\partial y_1} - \frac{\partial D_1}{\partial y_2} \) and \( D \cdot E = D_1 E_1 + D_2 E_2 \).
The div-curl lemma is applied to the functions
\begin{equation}
D^\varepsilon := \begin{pmatrix} f(u^\varepsilon) \\ \nu^\varepsilon \end{pmatrix}, \quad E^\varepsilon := \begin{pmatrix} -\xi^\varepsilon \\ q(u^\varepsilon) \end{pmatrix}
\end{equation}
for some entropy pair \((\eta, q)\). We then have
\[
\text{div}(D^\varepsilon) = u^\varepsilon_t + f(u^\varepsilon)_x, \quad \text{curl}(E^\varepsilon) = \eta(u^\varepsilon)_t + q(u^\varepsilon)_x.
\]
By asserting that these quantities lie in a “nice” set, we are providing enough control on oscillations in \(u^\varepsilon\) to conclude that the weak convergence \(u^\varepsilon \rightharpoonup u\) is in fact strong.

The main result of this section, which we pose next, gives the connection between bounds on the entropy residual \(\eta(u^\varepsilon)_t + q(u^\varepsilon)_x\) and the strong convergence of the approximating sequence. The result was originally shown by Tartar [Tar79], who asserted precompactness of these sequences for all entropy pairs \((\eta, q)\). We pose two versions of the theorem, both of which only require precompactness for two entropy pairs: if the flux function \(f\) is strictly convex, then we may choose any strictly convex entropy pair (Theorem 1.12); for general flux functions we have to select a specific entropy pair (Theorem 1.13).

**Theorem 1.12.** Let \(\Omega \subset \mathbb{R}^2\) be open and bounded. Let \(\{u^\varepsilon\}_{\varepsilon>0}\) be a bounded sequence in \(L^\infty(\Omega)\) such that
\begin{equation}
\{u^\varepsilon_t + f(u^\varepsilon)_x\}_{\varepsilon>0} \quad \text{and} \quad \{\eta(u^\varepsilon)_t + q(u^\varepsilon)_x\}_{\varepsilon>0}
\end{equation}
are precompact in \(H^{-1}(\Omega)\),
for a strictly convex entropy pair \((\eta, q)\), and assume that \(f\) is strictly convex. Then \(\{u^\varepsilon\}_{\varepsilon>0}\) has a subsequence that converges pointwise a.e. and in \(L^p(\Omega)\) for all \(p \in [1, \infty)\) to a function \(u \in L^\infty(\Omega)\).

**Proof.** Let \(D^\varepsilon, E^\varepsilon\) be defined by (1.21). By the Fundamental Theorem of Young measures, there is a subsequence (still indexed by \(\varepsilon\)) such that
\[
D^\varepsilon \rightharpoonup \bar{D} := \begin{pmatrix} f \\ \nu \end{pmatrix}, \quad E^\varepsilon \rightharpoonup \bar{E} := \begin{pmatrix} -\bar{\eta} \\ \bar{q} \end{pmatrix}
\]
weakly-\(\ast\) in \(L^\infty(\Omega, \mathbb{R}^2)\),
and since \(D^\varepsilon, E^\varepsilon\) are \(L^\infty\)-bounded on a bounded domain, the weak convergence is also in \(L^2(\Omega, \mathbb{R}^2)\). The compactness criterion (1.19) follows from (1.22), so by the div-curl lemma, we have \(D^\varepsilon \cdot E^\varepsilon \rightharpoonup \bar{D} \cdot \bar{E} \) in \(\mathcal{D}'(\Omega)\). But then also \(D^\varepsilon \cdot E^\varepsilon \rightharpoonup \bar{D} \cdot \bar{E} \) weakly-\(\ast\) in \(L^\infty\), which using the Young measure \(\nu\) can be written as
\begin{equation}
\langle v_y, \lambda \rangle \langle v_y, q \rangle - \langle v_y, \eta \rangle \langle v_y, f \rangle = \langle v_y, \lambda q - \eta f \rangle \quad \text{a.e.} \quad y \in \Omega,
\end{equation}
or more compactly, \(\bar{u} \bar{q} - \bar{\eta} \bar{f} = \lambda \bar{q} - \eta \bar{f}\). This is the so-called Murat-Tartar commutator relation. Using this relation, we will show that the function
\[
H(a, b) := (b - a)(q(b) - q(a)) - (\eta(b) - \eta(a))(f(b) - f(a))
\]
satisfies the conditions of Lemma A.3, and consequently, by Lemma A.2, the convergence \(u^\varepsilon \rightharpoonup u\) is strong.
Clearly, \( H(a, a) = 0 \). We have
\[
\frac{\partial H}{\partial b}(a, b) = q(b) - q(a) + (b - a)q'(b) - (\eta(b) - \eta(a))f'(b) - \eta'(b)(f(b) - f(a))
\]
and
\[
\frac{\partial^2 H}{\partial b^2}(a, b) = \eta''(b)((b - a)f'(b) - (f(b) - f(a)) + f''(b)(b - a)\eta'(b) - (\eta(b) - \eta(a)))
\]
\[
= \eta''(b) \int_a^b f'(b) - f'(s) \, ds + f''(b) \int_a^b \eta'(b) - \eta'(s) \, ds.
\]
Hence, \( \frac{\partial H}{\partial b}(a, a) = 0 \) for all \( a \in \mathbb{R} \), and if \( b \geq a \) then \( \frac{\partial^2 H}{\partial b^2}(a, b) \geq 0 \), with equality only for \( b \) in a set of measure zero, since \( f \) and \( \eta \) are strictly convex. Hence, \( H(a, b) > 0 \) for all \( a < b \), and as \( H(b, a) = H(a, b) \), we have \( H(a, b) > 0 \) for all \( a \neq b \). Last, we have for a.e. \( y \in \Omega \)
\[
\int_{\mathbb{R}^2} H \, d\nu_y \times \nu_y = \int_{\mathbb{R}^2} (b - a)(q(b) - q(a)) - (\eta(b) - \eta(a))(f(b) - f(a)) \, d\nu_y(b)d\nu_y(a)
\]
\[
= 2\left( \langle \nu_y, \text{id} q \rangle - \langle \nu_y, \text{id} \rangle \langle \nu_y, q \rangle + \langle \nu_y, \eta f \rangle - \langle \nu_y, \eta \rangle \langle \nu_y, f \rangle \right) = 0,
\]
where we have used Fubini’s theorem, the fact that \( \nu_y(\mathbb{R}) = 1 \), and (1.23). Hence, \( H \) satisfies (A.1). \hfill \Box

**Theorem 1.13** (Chen-Lu [CL89]). Let \( \Omega \subset \mathbb{R}^2 \) be open and bounded. Let \( \{u^\varepsilon\}_{\varepsilon > 0} \) be a bounded sequence in \( L^\infty(\Omega) \) such that (1.22) is satisfied with the entropy pair \( \eta = f \), \( q(u) = \int_a^bf'(s)^2 \, ds \). Assume that
\begin{equation}
(1.24) \quad \text{meas} \{ s \in \mathbb{R} : f''(s) = 0 \} = 0.
\end{equation}
Then \( \{u^\varepsilon\}_{\varepsilon > 0} \) has a subsequence that converges pointwise a.e. and in \( L^p(\Omega) \) for all \( p \in [1, \infty) \) to a function \( u \in L^\infty(\Omega) \).

**Proof of Theorem 1.13.** The proof follows that of Theorem 1.12 closely. The Murat-Tartar commutator relation (1.23) holds as before, but in this case the function \( H \) is
\[
H(a, b) := (b - a)(q(b) - q(a)) - (f(b) - f(a))^2.
\]
Clearly, \( H(a, a) = 0 \). By the Cauchy-Schwarz inequality we have for every \( a < b \)
\[
(f(b) - f(a))^2 = \left( \int_a^b f'(s) \, ds \right)^2 \leq (b - a) \int_a^b f'(s)^2 \, ds = (b - a)(q(b) - q(a)),
\]
so \( H(a, b) \geq 0 \). But there is equality in the Cauchy-Schwarz inequality if and only if \( f' \) is constant on \((a, b)\), which by (1.24) it cannot be. Hence, \( H(a, b) > 0 \) for all \( a < b \), and since \( H(a, b) = H(b, a) \), the first part of (A.1) follows. The second part follows by the same argument as in Theorem 1.12, so in conclusion, \( \nu \) is a function, i.e. \( \nu_y = \delta_{u(y)} \) for a.e. \( y \in \Omega \) for a \( u \in L^\infty(\Omega) \). \hfill \Box

**Definition 1.14.** Functions satisfying (1.24) are called **genuinely nonlinear**.
1.5. Existence and uniqueness for systems of conservation laws

For general $N \times N$ systems of hyperbolic conservation laws there is no global well-posedness result like Kruzkov’s Theorem 1.9. However, some partial results exist. Lax showed in his seminal paper [Lax57] existence and uniqueness of the entropy solution of the one-dimensional Riemann problem under the assumption that the two states are sufficiently close. Glimm [Gli65] proved existence of a weak solution to the Cauchy problem for all $u_0 \in D$, the $L^1$-closure of the set of all piecewise constant functions with “sufficiently small” total variation. This was accomplished by showing that his random choice method is stable with respect to what is now called the Glimm functional. More recently, Bressan et al. [BLY99, BCP00] showed that the front tracking method is stable with respect to the Glimm functional, and that it converges to the unique entropy solution for all $u_0 \in D$. We sum this up as follows.

**Theorem 1.15.** Let $d = 1$ and assume that the system (1.1) is strictly hyperbolic and that all wave families are either genuinely nonlinear or nonlinearly degenerate. Then for all $u_0 \in L^1$ with sufficiently small total variation, there exists a unique entropy solution to the Cauchy problem (1.1).

1.6. Measure-valued solutions

Summarizing the remarks in the previous section, there is a lack of global, “large data” existence and uniqueness results for systems of hyperbolic conservation laws. We describe here the alternative approach of measure-valued solutions, initiated by DiPerna in his seminal paper [DiP85]. Unlike weak solutions, which are (almost everywhere defined) functions, a measure-valued solution is a Young measure, a mapping which to each point assigns a probability distribution, describing the likely value at that point. This should be seen as a natural solution concept for evolution equations: for instance, one rarely, if ever, know the precise initial conditions, as all measuring processes have some inherent uncertainty.

When a sequence of functions converges pointwise to a function $u$, then the Young measure $\nu$ associated with the sequence is a unit mass concentrated at $u$, $\nu(x,t) = \delta_{u(x,t)}$. In this case $(\nu(x,t), g(\lambda)) = g((\nu(x,t), \lambda)) = g(u(x,t))$ for all $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$ and $g \in C(\mathbb{R}^N)$. Hence, if $u$ were a weak solution of (1.1) then

$$\langle \nu, \lambda \rangle_t + \text{div}(\langle \nu, f \rangle) = 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^d \times \mathbb{R}_+).$$

This expression is well-defined also when $\nu$ is not everywhere a point mass (although in this case the relation $\langle \nu, f(\lambda) \rangle = f(\langle \nu, \lambda \rangle)$ is not satisfied). We take this as the definition of a measure-valued solution of (1.1).

**Definition 1.16** (DiPerna [DiP85]). A Young measure $\nu$ defined on $\mathbb{R}^d \times \mathbb{R}_+$ is a measure-valued solution of (1.1) if

$$\langle \nu, \lambda \rangle_t + \text{div}(\langle \nu, f \rangle) = 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^d \times \mathbb{R}_+).$$
If \((\eta, q)\) is a convex entropy pair then a measure-valued solution \(\nu\) of (1.1) is admissible with respect to \((\eta, q)\) if

\[
\langle \nu, \eta \rangle_t + \text{div}(\langle \nu, q \rangle) \leq 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^d \times \mathbb{R}_+).
\]

It is called an entropy measure-valued solution of (1.1) if it is admissible with respect to all convex entropy pairs \((\eta, q)\).

To see that the concept of (entropy) measure-valued solutions is weaker than that of weak solutions, it is enough to note that definitions (1.25), (1.26) are linear in \(\nu\); hence, if \(\nu^0\) and \(\nu^1\) are two (entropy) measure-valued solutions, then so is

\[
\nu^\theta := (1 - \theta)\nu^0 + \theta\nu^1, \quad \theta \in (0, 1).
\]

The definition (1.7) of weak solutions, on the other hand, is in general nonlinear in \(u\), and so an analogously defined function \(u^\theta\) will not be a weak solution. See Schochet [Sch89] for further examples of measure-valued solutions.

It was shown in [DiP85] that for scalar conservation laws, an entropy measure-valued solution that has initial Dirac mass \(u_0(x)\) coincides with the corresponding entropy solution at all later times – what is known as weak-strong uniqueness. This was proved by showing that if \(\nu\) is an entropy measure-valued solution and \(u\) is the entropy solution of the Cauchy problem (1.1), then for all parametrized entropy pairs \((\eta, q)\) (cf. Section 1.2.1) the following stability bound holds:

\[
\langle \nu_y, \eta(u(y), \lambda) \rangle_t + \text{div}(\langle \nu_y, q(u(y), \lambda) \rangle) \leq 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^d \times \mathbb{R}_+).
\]

Letting \((\eta, q)\) be the Kruzkov entropy pair (1.14), then the above implies that the function

\[
V(t) := \int_{\mathbb{R}^d} \langle \nu(x,t), |\lambda - u(x,t)| \rangle \, dx
\]

is nonincreasing, analogous to the stability bound (1.18) for entropy solutions. In particular, if the initial data \(u_0\) is assumed in a sufficiently strong sense, then \(V(t) = 0\) for all times, and hence \(\nu(x,t) = \delta_{u(x,t)}\) for almost all \((x, t)\).

More recently, Brenier, De Lellis and Székelyhidi showed weak-strong uniqueness for measure-valued solutions of the incompressible Euler equations [BLS11]. They also proved that if there is a Lipschitz continuous solution of the hyperbolic system (1.1), then any entropy measure-valued solution of (1.1) coincides with the classical solution.

### 1.7. Summary and outlook

Based on the discussion in this chapter, we may draw the following conclusions. The theory of existence, uniqueness and regularity for scalar conservation laws is mature and well-developed. The availability of a large class of entropies enables a full stability analysis, hence uniqueness of entropy solutions. Convergence frameworks such as functions of total variation and compensated compactness allows for passing to the limit in approximate solutions, thus implying existence of entropy solutions.

For hyperbolic systems of equations, much less is known. No existence or uniqueness results for “large data” are available for general systems. The weaker concept of measure-valued solutions makes it easier to show existence of solutions, with the possible disadvantage of allowing for pathological solutions.
In view of these remarks, it seems that the most that one can hope for from a numerical approximation of (1.1) is that it
- is consistent with respect to entropy conditions,
- is high-order accurate,
- converges strongly to the entropy solution for scalar conservation laws,
- converges to a measure-valued solution for systems of conservation laws.

The remainder of this thesis is devoted to developing numerical schemes satisfying these properties.
As has been demonstrated in the previous chapter, entropy conditions play a crucial role in the stability analysis of hyperbolic conservation laws. In this chapter we consider finite volume methods which satisfy a discrete version of the entropy admissibility criterion (1.10) for convex entropy pairs. Such schemes will be called entropy stable methods. Entropy stability was studied first in Lax’ important 1971 paper [Lax71]. There, he established that the Lax-Friedrichs method is entropy stable, and as a corollary, that, should the method converge pointwise, then the limit is the unique entropy solution. This result was generalized by Harten, Hyman and Lax in 1976, who showed that all monotone schemes for scalar conservation laws are entropy stable [HHLK76]. Osher developed in his 1984 paper [Osh84] the theory of E-schemes, of which monotone methods are a subset. E-schemes are designed precisely to be entropy stable with respect to all convex entropies, and as they are total variation diminishing, they converge to the entropy solution.

As was shown in [Osh84], all E-schemes are at most first-order accurate. The work towards obtaining higher-order accurate entropy stable schemes was initiated by Tadmor [Tad84], [Tad87a]. He found a convenient relation that guarantees that a method is entropy conservative, meaning that it satisfies a discrete version of the entropy equality (1.9). Such schemes are automatically second-order accurate, and methods of higher (than second) order have been developed more recently in [LMR03]. However, the lack of diffusion in entropy conservative methods will lead to oscillations in the vicinity of shocks. By adding numerical diffusion we get rid of such oscillations, but generally at the cost of accuracy. In Chapter 4 we will design higher-order numerical diffusion operators. In the present chapter we restrict ourselves to first-order schemes.

In Section 2.2 we review the theory of entropy conservative and entropy stable schemes. In Sections 2.3–2.5 we demonstrate how to construct, and give concrete examples of, entropy stable schemes for scalar equations and for systems. The interested reader is referred to Tadmor’s review paper [Tad03] for further details.

For simplicity we only consider the one-dimensional hyperbolic conservation law (1.1), $d = 1$, although everything can easily be generalized to multi-dimensional equations.
2.1. Finite volume and finite difference methods

In its simplest, one-dimensional form, a finite volume (finite difference) method for (1.1) is a method of the form

\[ \frac{d}{dt} u_i + \frac{F_{i+1/2} - F_{i-1/2}}{\Delta x} = 0, \]

here written in the semi-discrete form. The **numerical flux function** \( F_{i+1/2} \) is a function of \( 2m \) \((m \in \mathbb{N})\) variables, \( F_{i+1/2} = F(u_{i-m+1}, \ldots, u_{i+m}) \), and we require that it is **consistent**, i.e., that \( F(u, \ldots, u) = f(u) \) for all \( u \in U \). The spatial domain \( x \in \mathbb{R} \) is partitioned into cells \( I_i = [x_{i-1/2}, x_{i+1/2}] \) with uniform grid size \( x_{i+1/2} - x_{i-1/2} \equiv \Delta x \) and we approximate \( u_i(t) \approx u_i^0(x_i) \) (finite volume) or \( u_i(t) \approx u_i^0(x_i) \) (finite difference). The initial data is sampled as \( u_i^0 : = \frac{1}{\Delta x} \int_{I_i} u_0(x) dx \) (finite volume) or \( u_i^0 : = u_0(x_i) \) (finite difference).

In a practical implementation the temporal dimension must also be discretized, but for clarity of exposition we will leave it continuous. In numerical experiments we integrate in time using strong stability preserving third-order Runge-Kutta methods from [S088, GS98]. These methods consist of subsequent applications and convex combinations of forward Euler steps,

\[ u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} \left( F_{i+1/2}^n - F_{i-1/2}^n \right). \]

Standard stability analysis dictates that \( \Delta t \) should be chosen such that the CFL number \( \frac{\Delta t}{\Delta x} \max_{i \in \mathbb{Z}} |f'(u_i)| \) is less than 1; see [S088, LeV02].

2.2. Entropy conservative and entropy stable methods

Recall that an **entropy pair** \((\eta, q)\) for (1.1) is a pair of functions \( \eta, q : \mathbb{R}^N \to \mathbb{R} \) such that \( q'(u) = \eta'(u)^T f'(u) \). In the remainder we will assume that \( \eta \) is strictly convex.

**Definition 2.1.** Let \((\eta, q)\) be a convex entropy pair. We say that the finite volume method (finite difference method) (2.1) is **entropy conservative** if computed solutions satisfy the discrete entropy equality

\[ \frac{d}{dt} \eta(u_i) + \frac{Q_{i+1/2} - Q_{i-1/2}}{\Delta x} = 0, \]

for some \( 2m \)-point numerical flux function \( Q_{i+1/2} = Q(u_{i-m+1}, \ldots, u_{i+m}) \) such that \( Q(u, \ldots, u) = q(u) \). We say that it is **entropy stable** if computed solutions satisfy the discrete entropy inequality

\[ \frac{d}{dt} \eta(u_i) + \frac{Q_{i+1/2} - Q_{i-1/2}}{\Delta x} \leq 0. \]

Clearly, (2.2) is a discrete version of the entropy equality (1.9), whereas (2.3) is a discrete version of the entropy stability criterion (1.10). Summing (2.3) over \( i \in \mathbb{Z} \) and integrating
over \( t \in [0, T] \), we see that the total amount of entropy of an entropy stable scheme decreases in time:

\[
\sum_i \eta(u_i(T)) \Delta x \leq \sum_i \eta(u_i(0)) \Delta x,
\]

which is a discrete version of (1.11).

Presently we find sufficient conditions for entropy conservation and entropy stability. Recall from Section 1.2.2 that the entropy potential is defined as \( \eta(v) := v \cdot f(u(v)) - q(u(v)) \), where \( \omega(u) = \eta'(u) \) is the entropy variable, and that it satisfies \( \omega'(v) = f(u(v)) \). As was seen in Section 1.2, the entropy identity (1.9) follows by multiplying the conservation law by entropy variables:

\[
0 = v \cdot (u_t + f(u)_x) = \eta(u)_t + q(u)_x.
\]

The use of the chain rule in the computation \( v \cdot f(u)_x = q(u)_x \) may be stated in terms of the entropy potential, as follows:

\[
v \cdot f(u)_x = (v \cdot f(u))_x - v_x \cdot f(u) = (v \cdot f(u) - \psi(v))_x - (v_x \cdot f(u) - v_x \cdot \psi'(v)) = q(u)_x,
\]

where we in the second line have added and subtracted \( \psi(v)_x = v_x \cdot \psi'(v) \). We mimic this use of the chain rule for the scheme (2.1). Denote \( v_i = v(u_i), \psi_i = \psi(u_i) \), etc., and let \( u_{i+1/2} = \frac{1}{2}(u_i + u_{i+1}) \) and \( \|u\|_{i+1/2} = u_{i+1} - u_i \). Then, multiplying the flux part of (2.1) by \( v_i \), we obtain

\[
v_i \cdot \frac{F_{i+1/2} - F_{i-1/2}}{\Delta x} = \frac{v_{i+1/2} \cdot F_{i+1/2} - v_{i-1/2} \cdot F_{i-1/2}}{\Delta x} = \frac{\|v\|_{i+1/2} \cdot F_{i+1/2} + \|v\|_{i-1/2} \cdot F_{i-1/2}}{2\Delta x}
\]

\[
= \frac{(v_{i+1/2} \cdot F_{i+1/2} - \bar{\psi}_{i+1/2}) - (v_{i-1/2} \cdot F_{i-1/2} - \bar{\psi}_{i-1/2})}{\Delta x}
\]

\[
- \frac{(\|v\|_{i+1/2} \cdot F_{i+1/2} - \|\psi\|_{i+1/2}) + (\|v\|_{i-1/2} \cdot F_{i-1/2} - \|\psi\|_{i-1/2})}{2\Delta x}.
\]

Hence, if we define \( Q_{i+1/2} = Q(u_i, u_{i+1}) := v_{i+1/2} \cdot F_{i+1/2} - \bar{\psi}_{i+1/2} \), then (2.1) multiplied by \( v_i \) is the same as

\[
dt \eta(u_i) + \frac{Q_{i+1/2} - Q_{i-1/2}}{\Delta x} = \frac{r_{i+1/2} + r_{i-1/2}}{2\Delta x},
\]

where \( r_{i+1/2} := \|v\|_{i+1/2} \cdot F_{i+1/2} - \|\psi\|_{i+1/2} \) is the local entropy production.

If we denote \( v(s) := v_i + s\|v\|_{i+1/2} \), then

\[
r_{i+1/2} = \|v\|_{i+1/2} \cdot F_{i+1/2} - \int_0^1 \frac{d}{ds} \psi(v(s)) \, ds
\]

\[
= \int_0^1 \|v\|_{i+1/2} \cdot F_{i+1/2} - \|v\|_{i+1/2} \cdot \psi'(v(s)) \, ds
\]

\[
= \int_0^1 \|v\|_{i+1/2} \cdot (F_{i+1/2} - f(u(v(s)))) \, ds.
\]
Hence, if the above integrand is everywhere nonpositive, then (2.5) implies that the method is entropy stable. Schemes that satisfy this are the so-called \( E \)-schemes.

**Definition 2.2** (Osher [Osh84]). An \( E \)-scheme is a scheme (2.1) whose numerical flux \( F \) satisfies

\[
(v_{i+1} - v_i) \cdot (F_{i+1/2} - f(u(v(s)))) \leq 0 \quad \forall \ s \in [0, 1]
\]

(2.6)

(where \( v(s) := v_i + s(v_{i+1} - v_i) \) for all \( v_i, v_{i+1} \) and all convex entropy pairs \( (\eta, \varphi) \).)

In the scalar case \( (N = 1) \), (2.6) is equivalent to

\[
\text{sgn}(u_{i+1} - u_i)(F_{i+1/2} - f(u)) \leq 0 \quad \forall u \text{ between } u_i \text{ and } u_{i+1},
\]

since \( \|v\|_{i+1/2} = \frac{1}{\eta''(\xi)} \|u\|_{i+1/2} \) for some \( \xi \) between \( u_i \) and \( u_{i+1} \). The latter condition is independent of the entropy pair, and hence holds for all entropy pairs. Osher showed in [Osh84] that \( E \)-schemes for scalar conservation laws converge strongly to the unique entropy solution. However, it was also shown that \( E \)-schemes are at most first order accurate. Hence, should we simultaneously wish for entropy stability and high-order accuracy, we cannot demand entropy stability with respect to all convex entropy pairs.

Going back to (2.5), it is obvious that if \( r_{i+1/2} = 0 \) for all \( i \) then the scheme is entropy conservative. This is precisely Tadmor’s entropy conservation condition.

**Theorem 2.3** (Tadmor [Tad87a]). If \( F \) satisfies

\[
\|v\|_{i+1/2} \cdot F_{i+1/2} = \|\psi\|_{i+1/2}
\]

(2.7)

for all \( i \), then the scheme (2.1) is entropy conservative with numerical entropy flux

\[
Q_{i+1/2} = \bar{v}_{i+1/2} \cdot F_{i+1/2} - \overline{\psi}_{i+1/2}.
\]

Moreover, in the scalar case \( (N = 1) \) it is second-order accurate in smooth regions of \( u \).

**Proof.** Entropy conservation follows directly from (2.5) and the definition of \( r_{i+1/2} \). Accuracy in this case means that

\[
\frac{F_{i+1/2} - F_{i-1/2}}{\Delta x} = f(u(x))_{x=x_i} + O(\Delta x^2).
\]

Indeed, by the trapezoidal rule,

\[
F_{i+1/2} = \frac{F_{i+1/2} \|v\|_{i+1/2}}{\|v\|_{i+1/2}} = \frac{\|\psi\|_{i+1/2}}{\|v\|_{i+1/2}} = \frac{1}{\|v\|_{i+1/2}} \int_{v_i}^{v_{i+1}} \psi'(v) \, dv
\]

\[
= \frac{1}{\|v\|_{i+1/2}} \int_{v_i}^{v_{i+1}} f(u(v)) \, dv = \frac{f(u_i) + f(u_{i+1})}{2} - \frac{\|v\|_{i+1/2}^2}{12} \psi''(\xi_{i+1/2})
\]

for a \( \xi_{i+1/2} \) between \( v_i \) and \( v_{i+1} \). Hence,

\[
\frac{F_{i+1/2} - F_{i-1/2}}{\Delta x} = \frac{f(u_{i+1}) - f(u_{i-1})}{2\Delta x} + O(\Delta x^2) = f(u(x))_{x=x_i} + O(\Delta x^2).
\]

\( \square \)
From (2.5) it is obvious that if \( r_i^{+1/2} \leq 0 \) for all \( i \), then the method is entropy stable. Rather than stating this as a theorem, we state the result in terms of the numerical diffusion of the scheme. Let \( F_i^{+1/2} \) be any numerical flux, and let \( \tilde{F}_i^{+1/2} \) be a flux satisfying (2.7). By inspecting \( F_i^{+1/2} - \tilde{F}_i^{+1/2} \), it is not hard to show that there is a matrix \( D_i^{+1/2} \) such that

\[
F_i^{+1/2} = \tilde{F}_i^{+1/2} - D_i^{+1/2} \| v \|_{i+1/2}.
\]

Inserting into \( r_i^{+1/2} \), we obtain the following result.

**Theorem 2.4** (Tadmor [Tad87a]). Assume that \( D_i^{+1/2} \geq 0 \). Then the scheme with flux (2.8) is entropy stable, with the numerical entropy flux

\[
Q_i^{+1/2} = \tilde{\nu}_{i+1/2} \cdot \tilde{F}_{i+1/2} - \nu_{i+1/2} \cdot \tilde{F}_{i+1/2} - D_i^{+1/2} \| v \|_{i+1/2}.
\]

More precisely, it satisfies the entropy dissipation estimate

\[
\frac{d}{dt} \eta(u_i) + \frac{Q_i^{+1/2} - Q_i^{+1/2}}{\Delta x} = -\frac{\| v \|_{i+1/2} \cdot D_i^{+1/2} \| v \|_{i+1/2} + \| v \|_{i-1/2} \cdot D_i^{+1/2} \| v \|_{i-1/2}}{2\Delta x}.
\]

**Proof.** We have

\[
r_i^{+1/2} = \| v \|_{i+1/2} \cdot (\tilde{F}_{i+1/2} - D_i^{+1/2} \| v \|_{i+1/2}) - \| \psi \|_{i+1/2}
\]

\[
= -\| v \|_{i+1/2} \cdot D_i^{+1/2} \| v \|_{i+1/2} \leq 0
\]

because \( \tilde{F} \) satisfies (2.7); hence, entropy stability follows from (2.5). \( \square \)

This results suggests a way of designing entropy stable schemes: first, find an entropy conservative numerical flux \( \tilde{F} \) through the relation (2.7), and then add numerical diffusion in the form (2.8). This will be a recurring theme throughout this thesis. From here on we shall denote all entropy conservative numerical flux functions by \( \tilde{F} \), and entropy stable numerical fluxes by \( F \).

### 2.3. Entropy stable methods for scalar equations

In the scalar, one-dimensional case \( (N = d = 1) \) the construction of an entropy conservative numerical flux is trivial using the entropy conservation criterion (2.7): given a convex entropy pair \( (\eta, q) \), define \( \tilde{F}_i^{+1/2} \) by

\[
\tilde{F}_i^{+1/2} = \begin{cases} 
\| \psi \|_{i+1/2} / \| v \|_{i+1/2} & \text{if } u_i \neq u_{i+1} \\
q(u_i) / \| v \|_{i+1/2} & \text{if } u_i = u_{i+1}.
\end{cases}
\]

Thus, there is a unique three-point entropy conservative scheme for any entropy pair [Tad87a].

**Example 2.5.** Consider the linear advection (1.2). For simplicity we choose the square entropy \( \eta(u) = \frac{u^2}{2} \). Then

\[
v(u) = u, \quad q(u) = \frac{u^2}{2}, \quad \psi(u) = \frac{u^2}{2}.
\]

We obtain the entropy conservative flux

\[
\tilde{F}_i^{+1/2} = \frac{1}{2} \frac{\| u \|_{i+1/2}}{\| v \|_{i+1/2}} = \bar{u}_{i+1/2}.
\]

\( \blacksquare \)
Example 2.6. Consider Burgers’ equation (1.3). Again we can take the square entropy \( \eta(u) = \frac{u^2}{2} \). We then easily find that
\[
v(u) = u, \quad q(u) = \frac{u^3}{3}, \quad \psi(u) = \frac{u^3}{6}.
\]
Thus, a flux for Burgers’ equation that is entropy conservative with respect to \( \eta \) must satisfy
\[
\|u\|_{i+1/2} \tilde{F}_{i+1/2} = \frac{u^2_1}{6},
\]
which is satisfied if and only if
\[
\tilde{F}_{i+1/2} = \frac{u_j^2 + u_i u_{i+1} + u_{i+1}^2}{6}.
\]
Incidentally, this flux is also entropy conservative with respect to the entropy \( \eta(u) = \frac{(u-k)^2}{2} \) for any fixed \( k \in \mathbb{R} \).

If \( D_{i+1/2} \) is any nonnegative number, then the flux (2.8) is entropy stable with respect to \((\eta, q)\). By the mean value theorem there is a \( \zeta_{i+1/2} \) between \( u_i \) and \( u_{i+1} \) such that \( \|v\|_{i+1/2} = v'(\zeta_{i+1/2})\|u\|_{i+1/2} = \eta''(\zeta_{i+1/2})\|u\|_{i+1/2} \). Since \( \eta \) is convex, we have \( P_{i+1/2} = \eta''(\zeta_{i+1/2})D_{i+1/2} \geq 0 \), and so (2.8) admits the equivalent formulation
\[
F_{i+1/2} = \tilde{F}_{i+1/2} - P_{i+1/2}\|u\|_{i+1/2}.
\]
We may now choose \( P_{i+1/2} \) to be any nonnegative number. For instance, we may choose the Lax-Friedrichs diffusion
\[
P_{i+1/2} := \frac{1}{2} \max_{j \in \mathbb{Z}} (|f'(u_j)|).
\]
or the Local Lax-Friedrichs diffusion
\[
P_{i+1/2} := \frac{1}{2} \max (|f'(u_i)|, |f'(u_{i+1})|).
\]
It is clear that the construction of a first-order accurate entropy stable method for scalar conservation laws is more or less trivial. What remains, however, is to choose for which entropy pair the scheme should be entropy stable. As shown in (2.4), a scheme that is entropy stable with respect to \((\eta, q)\) has nonincreasing total entropy \( \sum_i \eta(u_i(t))\Delta x \). Thus, setting \( \eta(u) = |u|^r \) for an \( r \in [1, \infty) \) gives an \( L^r \)-stable scheme,
\[
\sum_i |u_i(t)|^r \Delta x \leq \sum_i |u_i(0)|^r \Delta x.
\]
By a specific choice of \( \eta \), this extends to the case \( r = \infty \).

Proposition 2.7. Suppose that \( u_0 \in L^\infty(\mathbb{R}^d) \) has compact support. Let \( a, b \in \mathbb{R} \) such that \( a < \inf_x u_0(x) \leq \sup_x u_0(x) < b \) and let \( (\eta, q) \) be the uniformly convex entropy pair with
\[
\eta(u) := \frac{1}{u - a} + \frac{1}{b - u}.
\]
Then any scheme that is entropy stable with respect to \((\eta, q)\) satisfies \( a < u_i(t) < b \) for all \( i \in \mathbb{Z}, t > 0 \) and \( \Delta x > 0 \).
Hence, the numerical flux \( \eta_i(t) \) is simply the general solution \( \eta_i(t) \) property in many applications. Unfortunately, the singularities of \( \eta_i \) for all \( i \in \mathbb{Z} \), where \( E^{\Delta x}(t) := \sum_j \eta(u_{j,x}(t)) \Delta x \).

If \( a^{\Delta x} \) and \( b^{\Delta x} \) are the unique numbers such that \( a < a^{\Delta x} < b^{\Delta x} < b \) and \( \eta(a^{\Delta x}) = \eta(b^{\Delta x}) = \frac{E^{\Delta x}(0)}{\Delta x} \), then necessarily \( a^{\Delta x} \leq u_{i,x}^{\Delta x}(t) \leq b^{\Delta x} \) for all \( i \in \mathbb{Z}, t \geq 0 \) and \( \Delta x > 0 \). In particular, \( a < u_{i,x}^{\Delta x}(t) < b \) for all \( i \in \mathbb{Z}, t \geq 0 \) and \( \Delta x > 0 \).

Choosing the entropy (2.11) thus automatically gives an \( L^\infty \)-bound, an important property in many applications. Unfortunately, the singularities of \( \eta \) at \( u = a \) and \( u = b \) makes the analysis of the resulting scheme difficult.

### 2.4. Entropy conservative schemes for systems

For systems of conservation laws \((N > 1)\), (2.7) is a scalar equation for \( N \) unknowns \( \tilde{F}_{i+1/2}, \ldots, \tilde{F}_{i+N/2} \), and so in general there is no unique solution. Tadmor [Tad87a] derived one general solution \( \tilde{F}_{i+1/2} \) as follows: given \( u_i, u_{i+1} \), let \( v(s) = v_i + s \|v\|_{i+1/2} \). Then

\[
\|v\|_{i+1/2} = \int_0^1 \frac{d}{ds} \psi(v(s)) \, ds = \|v\|_{i+1/2} \cdot \int_0^1 \psi'(v(s)) \, ds = \|v\|_{i+1/2} \cdot \int_0^1 f(u(v(s))) \, ds.
\]

Hence, the numerical flux

\[
\tilde{F}_{i+1/2} := \int_0^1 f(u(v(s))) \, ds
\]

is automatically entropy conservative. As an example we can consider the linear wave equation (1.4) with the entropy pair \( \eta(u) = (u_1)^2 + (u_2)^2 \), \( q(u) = 2cu_1u_2 \). Then the entropy variable is simply \( v(u) = u \), which inserted into (2.12) gives

\[
\tilde{F}_{i+1/2} = \int_0^1 A(u_i + s \|u\|_{i+1/2}) \, ds = \frac{1}{2} (Au_i + Au_{i+1}),
\]

which when inserted into the scheme (2.1) gives the central scheme

\[
\frac{d}{dt} u_i + \frac{f(u_{i+1}) - f(u_{i-1})}{2\Delta x} = 0.
\]

Unfortunately, the integral (2.12) may be very hard to evaluate for even the simplest nonlinear systems. Instead, we shall take a more direct approach, relying on explicit case-by-case computations. As an example we consider the shallow water equations (1.5), with the energy (1.15) as an entropy. Inserting the entropy variable and entropy potential (1.16) into (2.7), we find that an entropy conservative flux \( \tilde{F}_{i+1/2} = (\tilde{F}^1_{i+1/2}, \tilde{F}^2_{i+1/2}) \) must satisfy

\[
\tilde{F}^1[gh] - \frac{1}{2} \tilde{F}^1_{i+1/2} \|w^2\| + \tilde{F}^2[2w] = \frac{1}{2} g \|h^2w\|
\]

(where we have suppressed subindices for the moment). Using the identity

\[
[ab] = \overline{a}[b] + [a] \overline{b}
\]
and grouping together jumps in the same variable, we obtain
\[ g[h]\left(\tilde{F}^1 - \overline{h}\overline{w}\right) + w\left[\tilde{F}^2 - \tilde{F}^1\overline{w} - \frac{1}{2}gh^2\right] = 0. \]

Hence, if
\[
(2.15) \quad \tilde{F}_{i+1/2} = \left(\begin{array}{c} \tilde{F}^1_{i+1/2} \\ \tilde{F}^2_{i+1/2} \end{array}\right) = \left(\begin{array}{c} \overline{h}_{i+1/2}\overline{w}_{i+1/2} \\ \overline{h}_{i+1/2}\overline{w}_{i+1/2} + \frac{1}{2}gh^2_{i+1/2} \end{array}\right)
\]

then \( \tilde{F}_{i+1/2} \) satisfies (2.7) and hence is entropy conservative. However, this is not the only solution. For instance, with a different choice of \( \tilde{F}^1 \) we can obtain another entropy conservative flux,
\[
(2.16) \quad \tilde{F}_{i+1/2} = \left(\begin{array}{c} \tilde{F}^1_{i+1/2} \\ \tilde{F}^2_{i+1/2} \end{array}\right) = \left(\begin{array}{c} \overline{h}_{i+1/2}\overline{w}_{i+1/2} \\ \overline{h}_{i+1/2}\overline{w}_{i+1/2} + \frac{1}{2}gh_{i+1} \end{array}\right).
\]

**Lemma 2.8** ([Fjo09, FMT09]). The scheme (2.1) with \( \tilde{F} \) given by either (2.15) or (2.16) is second-order accurate and entropy conservative with respect to the energy (1.15).

For the Euler equations (1.6), the process is more involved since the entropy is a logarithmic function of the conserved variables. Ismail and Roe [IR09] derived an entropy conservative numerical flux based on the logarithmic average
\[
(a^\ln)_{i+1/2} := \frac{[a]_{i+1/2}}{[\log a]_{i+1/2}}.
\]

Define the parameter vectors \( z \) as
\[
(2.17) \quad z := \left(\begin{array}{c} z^1 \\ z^2 \\ z^3 \end{array}\right) := \sqrt{\frac{\rho}{p}} \left(\begin{array}{c} 1 \\ u \\ p \end{array}\right).
\]

Then the entropy conservative flux of [IR09] is
\[
\tilde{F}_{i+1/2} = \left(\begin{array}{c} \tilde{F}^1_{i+1/2} \\ \tilde{F}^2_{i+1/2} \\ \tilde{F}^3_{i+1/2} \end{array}\right)^T
\]

with
\[
(2.18) \quad \tilde{F}^1_{i+1/2} := \frac{z^3}{z^1_{i+1/2}} (z^3)^{\ln}_{i+1/2} \overline{z}^3_{i+1/2} \\
\tilde{F}^2_{i+1/2} := \frac{z^3_{i+1/2}}{\overline{z}^1_{i+1/2}} + \frac{z^2}{z^3_{i+1/2}} \tilde{F}^1_{i+1/2} \\
\tilde{F}^3_{i+1/2} := \frac{1}{2} \frac{z^2}{z^1_{i+1/2}} \left( \frac{\gamma + 1}{\gamma - 1} \frac{z^3}{z^1_{i+1/2}} \right) + \frac{\tilde{F}^2_{i+1/2}}{2}.
\]
2.5. Entropy stable schemes for systems

Given an entropy conservative flux \( \tilde{F}_{i+1/2} \), we add diffusion in the form (2.8) to obtain entropy stability. Any positive matrix \( D_{i+1/2} \) will do; we give here two choices, based on existing finite volume schemes. Recall that the Rusanov flux is defined as

\[
F_{i+1/2} := \frac{f(u_i) + f(u_{i+1})}{2} - \frac{c_{i+1/2}}{2} \|u\|_{i+1/2}
\]

for \( c_{i+1/2} = \max_{k=1,\ldots,N} (|\lambda_k(u_i)|, |\lambda_k(u_{i+1})|) \), where \( \lambda_k(u) \) is the \( k \)-th eigenvalue of \( f'(u) \). The Roe flux is defined as

\[
F_{i+1/2} := \frac{f(u_i) + f(u_{i+1})}{2} - \frac{1}{2} R_{i+1/2} |\Lambda_{i+1/2}| R_{i+1/2}^{-1} \|u\|_{i+1/2},
\]

where \( R_{i+1/2} = (r_1(u_{i+1/2}), \ldots, r_N(u_{i+1/2})) \) and \( \Lambda_{i+1/2} = \text{diag}(\lambda_1(u_{i+1/2}), \ldots, \lambda_N(u_{i+1/2})) \) are the matrices of eigenvectors and eigenvalues, respectively, of \( f'(u_{i+1/2}) \), and \( |\Lambda| := \text{diag}(|\lambda_1|, \ldots, |\lambda_N|) \). Here, \( u_{i+1/2} \) is an appropriately chosen intermediate state [Roe81]. Both of the diffusion operators \( \frac{c_{i+1/2}}{2} \|u\|_{i+1/2} \) and \( \frac{1}{2} R_{i+1/2} |\Lambda_{i+1/2}| R_{i+1/2}^{-1} \|u\|_{i+1/2} \) are of the generic form

\[
(2.19)
\]

\[ R_{i+1/2} A_{i+1/2} R_{i+1/2}^{-1} \|u\|_{i+1/2} \]

for a nonnegative diagonal matrix \( A_{i+1/2} \). We wish to manipulate this expression to arrive at an expression of the form \( D_{i+1/2} \|v\|_{i+1/2} \). Heuristically, we may write

\[
(2.20) \quad \|u\|_{i+1/2} \approx u_0(v_{i+1/2}) \|v_{i+1/2}\|,
\]

where \( u_0 = \frac{da}{dt} \). The matrix \( u_0(v) \) is symmetric positive definite whenever \( \eta \) is strictly convex, but something more can be said of it.

**Lemma 2.9** (Barth [Bar98]). Let \( B = R A R^{-1} \) be a diagonalizable matrix and let \( S \in \mathbb{R}^{N \times N} \) be a symmetric positive definite matrix such that \( BS \) is symmetric. Then the columns of \( R \) may be scaled such that \( S = RR^T \). In particular, \( BS = R R^T \).

Since the s.p.d. matrix \( u_0(v) \) right-symmetrizes \( f'(u) = R(u) \Lambda(u) R(u)^{-1} \) (where \( u = u(v) \); see Section 1.2.2), we find that the columns of the eigenvector matrix \( R(u) \) may be scaled such that \( u_0(v) = R(u) R(u)^T \) for all \( u = u(v) \). We take \( R_{i+1/2} \) to be that matrix \( R(u(v_{i+1/2})) \), and find that (2.19) can be written as

\[
R_{i+1/2} A_{i+1/2} R_{i+1/2}^{-1} \|u\|_{i+1/2} \approx R_{i+1/2} A_{i+1/2} R_{i+1/2}^{-1} u_0(v_{i+1/2}) \|v\|_{i+1/2}
\]

\[ = R_{i+1/2} A_{i+1/2} R_{i+1/2}^{-1} \|v\|_{i+1/2},
\]

As \( A_{i+1/2} \) is a nonnegative diagonal matrix, the diffusion matrix \( D_{i+1/2} := R_{i+1/2} A_{i+1/2} R_{i+1/2}^T \) is clearly also nonnegative. Hence, the flux

\[
(2.21) \quad F_{i+1/2} = \tilde{F}_{i+1/2} - R_{i+1/2} A_{i+1/2} R_{i+1/2}^T \|v\|_{i+1/2}
\]

\(^1\)The mean value theorem does not necessarily hold for vector-valued functions. Nonetheless, an intermediate state \( v_{i+1/2} \) satisfying (2.20) with equality may be found in many cases. In general, the difference between the two sides in (2.20) is \( O(\|v\|^2) \), and will be ignored.
is entropy stable. It remains to select the eigenvalue matrix $A_{i+1/2}$. We make three choices, inspired by the Lax-Friedrichs, Local Lax-Friedrichs and Roe schemes, respectively. Below, we denote the identity matrix in $\mathbb{R}^{N \times N}$ by $\text{Id}$.

**ELF:** Let $A_{i+1/2} = \frac{1}{2}c_i \text{Id}$, with $c := \max_{j \in \mathbb{Z}, k=1, \ldots, N}(|\lambda_k(u_j)|)$. This gives the diffusion matrix

$$D_{i+1/2} = \frac{1}{2}c_i R_{i+1/2} R^T_{i+1/2},$$

**ELLF:** Let $A_{i+1/2} = \frac{1}{2}c_i R_{i+1/2} R^T_{i+1/2}$, with $c_i := \max_{k=1, \ldots, N}(|\lambda_k(u_i)|, |\lambda_k(u_{i+1})|)$. This gives the diffusion matrix

$$D_{i+1/2} = \frac{1}{2}c_i R_{i+1/2} R^T_{i+1/2},$$

**ERoe:** Let $A_{i+1/2} = \frac{1}{2}i \Lambda_{i+1/2}$, with $\Lambda_i := \text{diag}(\lambda_1(u_{i+1/2}), \ldots, \lambda_N(u_{i+1/2}))$ and $u_{i+1/2} = \frac{u_i + u_{i+1}}{2}$ (or any other intermediate value). This gives the diffusion matrix

$$D_{i+1/2} = \frac{1}{2}R_{i+1/2} |\Lambda_{i+1/2}| R^T_{i+1/2}.$$  

One might also consider so-called polynomial viscosity matrices; see [DPRV99].

For concreteness we give the explicit form of an entropy stable diffusion operator for the shallow water equations (1.5). In this case, Lemma 2.9 takes the following form.

**Lemma 2.10 ([Fjo09]).** Denote

$$u_v := \frac{du}{dv} = \frac{1}{g} \begin{bmatrix} 1 & w \\ w^2 + gh & 1 \end{bmatrix}.$$  

(i) Define the following scaled version of the eigenvector matrix $R$ of $f'(u)$:

$$R = \frac{1}{\sqrt{2g}} \begin{bmatrix} 1 & w & 1 \\ w - \sqrt{gh} & w + \sqrt{gh} & 1 \end{bmatrix}. $$

Then $RR^T = u_v$.

(ii) For $u_i, u_{i+1} \in \mathbb{R}_+ \times \mathbb{R}$ we have $[u]_{i+1/2} = (u_v)_{i+1/2} [v]_{i+1/2}$, where

$$(u_v)_{i+1/2} := u_v(v(u_{i+1/2})) = \frac{1}{g} \begin{bmatrix} 1 & \bar{w}_{i+1/2} \\ \bar{w}_{i+1/2} & \bar{w}_{i+1/2}^2 + gh_{i+1/2} \end{bmatrix}$$

is $u_v$ evaluated at $u_{i+1/2} = (\bar{h}_{i+1/2}, \bar{h}_{i+1/2} \bar{w}_{i+1/2}).$

**Proof.** The results follow by insertion.  

Therefore, the entropy stable scheme for the shallow water equations will be the scheme with flux (2.21), with $\bar{F}_{i+1/2}$ being any of the entropy conservative fluxes (2.15) and (2.16), $D_{i+1/2}$ being any of the three choices (2.22), (2.23) and (2.24), and $R_{i+1/2}$ being given by (2.25) with $h = \bar{h}_{i+1/2}, w = \bar{w}_{i+1/2}$. 
2.6. Summary

We may summarize this chapter as follows. Entropy stability gives an a priori bound on the total entropy production of a scheme. Our strategy of constructing entropy stable schemes is to combine an entropy conservative flux with appropriate diffusion operators, expressed in terms of the entropy variable. Using Tadmor’s entropy conservation condition, it is easy to construct entropy conservative schemes for scalar equations, and with a bit more work also for systems of equations.

Entropy stability is obtained by adding an appropriate diffusion term. We propose three such terms, inspired by the Lax-Friedrichs, Local Lax-Friedrichs and Roe schemes, respectively.
CHAPTER 3

Convergence theory for finite volume methods

In his pioneering 1983 paper [DiP83], DiPerna demonstrated how to use compensated compactness to show convergence of finite volume methods to the entropy solution. In particular, DiPerna used Murat’s Lemma (Lemma 3.1) to show that under certain assumptions, the sequences of measures

\[ \{u_t + f(u)_x\}_{\Delta x > 0} \quad \text{and} \quad \{\eta(u)_t + q(u)_x\}_{\Delta x > 0}, \]

\( u \) being the approximate solution generated by the scheme, are precompact in \( H^{-1}(\Omega) \). Combined with Theorems 1.12 or 1.13, this proves strong (pointwise and \( L^p \)) convergence of the finite volume approximation.

At the time of the publication of [DiP83], it was already well-known that all monotone methods converge to the entropy solution; cf. Crandall and Majda [CM80]. The novelty in DiPerna’s compensated compactness technique, however, lies in its generality. By avoiding stability requirements such as monotonicity or requiring that the method is total variation diminishing or an E-scheme, it is applicable to a much broader class of numerical methods. The downside is that it is only applicable to scalar, one-dimensional conservation laws (and to certain systems of two equations [DiP83]). The alternative approach of H-measures is applicable to multi-dimensional scalar conservation laws; see Panov [Pan10].

As DiPerna’s paper is only concerned with first-order methods, certain assumptions are not posed as weakly as they might be. The main purpose of the present chapter is to revisit and refine his proof of convergence, sharpening its hypotheses to the absolute necessity. This will enable us to prove convergence of high-order accurate methods for scalar, one-dimensional conservation laws in Chapter 4. In Section 3.2 we investigate necessary conditions for the (weak-\( \ast \)) convergence of finite volume methods to measure-valued solutions. As this is a weaker solution concept, it should not come as a surprise that these conditions are weaker than the conditions for convergence to the entropy solution. Moreover, since we do not use compensated compactness, we are able to handle multi-dimensional equations on unbounded domains.

Finally, Section 3.3 demonstrates how the convergence framework in Section 3.1 can be applied to prove convergence of the first-order accurate entropy stable methods described in Section 2.3.
3.1. Strong convergence of finite volume schemes

We consider the scalar, one-dimensional conservation law (1.1). If \( u_i(t) \) is the computed solution, we define the function \( u(x,t) = u_i(t) \) for \( x \in I_i \). Similarly, for a flux function \( F_i(t) \), we denote \( F(x,t) = F_i(t) \) for \( x \in I_i \), so that its spatial distributional derivative is

\[
F_x = \sum_i (F_{i+1/2} - F_{i-1/2}) \delta_{x_{i-1/2}},
\]

where \( \delta_{x_{i-1/2}} \) denotes the Dirac unit mass at \( x_{i-1/2} \).

We will show that the sets \( \{ u_i + f(u)_x \}_{\Delta x > 0} \) and \( \{ \eta(u)_t + q(u)_x \}_{\Delta x > 0} \) are precompact in \( H^{-1} \), where \( u = u^{\Delta x} \) is computed using the finite volume methods (2.1) with grid size \( \Delta x \). (For each \( \Delta x > 0 \) we shall suppress the dependence of the computed solution \( u_i(t) \) on \( \Delta x \).) To this end, we must relate the terms \( f(u)_x \) and \( q(u)_x \) to their discretized counterparts

\[
\frac{F_{i+1/2} - F_{i-1/2}}{\Delta x} \quad \text{and} \quad \frac{Q_{i+1/2} - Q_{i-1/2}}{\Delta x},
\]

respectively. (The terms \( u_i \) and \( \eta(u)_t \) are not discretized, thus easing the analysis.) As these are nonlinear functions of \( u \), we will require a bound on the spatial variation of \( u \) to pass to the limit \( \Delta x \to 0 \).

The term \( q(u)_x \) is not discretized explicitly, but many schemes have an inherent discretization of this term. We shall refer to any function of \( 2m \) variables \( Q = Q(u_1, \ldots, u_{2m}) \) satisfying the consistency relation \( Q(u_i, \ldots, u_{i+m}) = q(u) \) for all \( u \in \mathbb{R} \) as a numerical entropy flux; we denote \( Q_{i+1/2} = Q(u_{i-m+1}, \ldots, u_{i+m}) \). For instance, for a 3-point scheme \( (m = 1) \) we can set

\[
Q(u_i, u_{i+1}) = \bar{v}_{i+1/2} F_{i+1/2} - \bar{\psi}_{i+1/2},
\]

(see Section 1.2.2 and Chapter 2). Clearly, any numerical flux function \( F \) is a numerical entropy flux for the entropy pair \( (\text{id}, f) \), where \( \text{id}(u) = u \) denotes the identity function.

Our goal will be to show precompactness of the sequence of measures \( \{ \eta(u)_t + q(u)_x \}_{\Delta x > 0} \) by studying the measures

\[
\frac{d}{dt} \eta(u)_t + \frac{Q_{i+1/2} - Q_{i-1/2}}{\Delta x}.
\]

Here and in the remainder, we assume that \( u \) is a scalar function.

Let \( \Omega \subset \mathbb{R} \times (0, T) \) be an open, bounded set such that \( \text{supp}(u) \subset \Omega \) for all \( \Delta x > 0 \). We will view \( \eta(u)_t + q(u)_x \) as both a functional on functions defined on \( \Omega \), as well as a measure on \( \Omega \). For \( 0 < t_1 < t_2 \) and Borel-measurable \( A \subset \mathbb{R} \), denote by \( \eta_t + q_x \) the measure

\[
(\eta_t + q_x)(A \times [t_1, t_2]) = \int_{t_1}^{t_2} \left( \int_A \eta(u(x,t)) \, dx + \sum_i (q(u_i(t)) - q(u_{i-1}(t))) \delta_{x_{i-1/2}}(A) \right) \, dt
\]

and by \( \eta_t + Q_x \)

\[
(\eta_t + Q_x)(A \times [t_1, t_2]) = \int_{t_1}^{t_2} \left( \int_A \eta(u(x,t)) \, dx + \sum_i (Q_{i+1/2}(t) - Q_{i-1/2}(t)) \delta_{x_{i-1/2}}(A) \right) \, dt.
\]
We may also view \( \eta_i + q_x \) as the functional

\[
(\eta_i + q_x)(\varphi) := \int_{\Omega} \varphi \, d(\eta_i + q_x)
\]

for \( \varphi \) in, e.g., \( W^{-1,p}(\Omega) \). We need the following lemma, a version of which is due to Murat [Mur81].

**Lemma 3.1** (Murat’s Lemma). Let \( \Omega \subset \mathbb{R}^n \) be open and bounded and let \( \{\mu_j\}_{j \in \mathbb{N}} \) be a bounded sequence in \( W^{-1,p}(\Omega) \) for some \( 2 < p \leq \infty \). Assume that \( \mu_j = \chi_j + \pi_j \), where \( \chi_j \) is precompact in \( H^{-1}(\Omega) \) and \( \{\pi_j\} \) is bounded in \( M(\Omega) \). Then \( \{\mu_j\} \) is precompact in \( H^{-1}(\Omega) \).

Here, \( M(\Omega) \) is the set of Radon measures on \( \Omega \); see Appendix A. The following auxiliary lemma provides the means to prove \( H^{-1} \)-precompactness of the sequence of entropy residuals.

**Lemma 3.2.** Let \( \Omega \subset \mathbb{R} \times [0,T] \) be bounded, let \( (\eta, q) \) be an entropy pair and let \( Q \) be a numerical entropy flux satisfying the Lipschitz condition

\[
|q(u_0) - Q(u_{m+1}, \ldots, u_m)| \leq C (|u_{m+2} - u_{m+1}| + \cdots + |u_m - u_{m-1}|) \quad \forall \ u_{m+1}, \ldots, u_m.
\]

Assume that a sequence of functions \( u = u^{\Delta x} (\Delta x > 0) \) satisfies

\[
\exists \ M > 0 : \|u^{\Delta x}\|_{L^\infty(\Omega)} \leq M \quad \forall \ \Delta x > 0
\]

\[
\supp(u^{\Delta x}) \subset \Omega \quad \forall \ \Delta x > 0
\]

\[
\int_0^T \sum_i |u^{\Delta x}_{i+1} - u^{\Delta x}_i| \Delta x \, dt \to 0 \quad \text{as} \ \Delta x \to 0 \quad \text{for some} \ P \in [2, \infty)
\]

\[
|\eta_i + Q_x| (\Omega) \leq C \quad \forall \ \Delta x > 0.
\]

Then \( \eta_i + q_x \), viewed as a sequence of functionals parametrized by \( \Delta x > 0 \), is precompact in \( H^{-1}(\Omega) \).

**Proof of Lemma 3.2.** The proof consists of applying Murat’s lemma to \( \eta_i + q_x \). This functional is bounded in \( W^{-1,\infty}(\Omega) \), since

\[
|(\eta_i + q_x)(\varphi)| = \left| \int_{\Omega} \eta(u)\varphi_i + q(u)\varphi_i \, dx \right| \leq (\|\eta(u)\|_{L^\infty(\Omega)} + \|q(u)\|_{L^\infty(\Omega)})\|\varphi\|_{W^{1,1}(\Omega)},
\]

which is bounded, by (3.2) and the continuity of \( \eta \) and \( q \).

We decompose

\[
(\eta_i + q_x)(\varphi) = \int_{\Omega} \varphi \, d(\eta_i + q_x) + \int_{\Omega} \varphi \, d((\eta_i + q_x) - (\eta_i + Q_x)).
\]

For the first term we have

\[
|\mathcal{E}^1(\varphi)| \leq \|\varphi\|_{L^\infty(\Omega)} |\eta_i + Q_x| (\Omega) \leq C \|\varphi\|_{L^\infty(\Omega)}
\]
3. CONVERGENCE THEORY FOR FINITE VOLUME METHODS

by (3.5). Hence, $\mathcal{E}^1$ is bounded in $\mathcal{M}(\Omega)$. It remains to show that $\mathcal{E}^2$ is precompact in $H^{-1}(\Omega)$. Indeed,

$$|\mathcal{E}^2(\varphi)| = \left| \int_\Omega \varphi \, d(q(u)_x - Q)_x \right|$$

$$= \left| \int_0^T \sum_i (\varphi(x_{i+1/2}, t) - \varphi(x_{i-1/2}, t)) \, (q(u_i) - Q_{i+1/2}) \, dt \right|$$ (summation by parts)

$$\leq \|\varphi\|_{L^2(\Omega)} \left( \int_0^T \sum_i |q(u_i) - Q(u_{i-m+1}, \ldots, u_{i+m})|^2 \, \Delta x \, dt \right)^{1/2}$$ (Hölder inequality)

$$\leq C\|\varphi\|_{L^2(\Omega)} \left( \int_0^T \sum_i |u_{i+1} - u_i|^2 \, \Delta x \, dt \right)^{1/2}$$ (by (3.1))

$$\leq C\|\varphi\|_{L^2(\Omega)} |\Omega|^{(p/2)'} \left( \int_0^T \sum_i |u_{i+1} - u_i|^p \, \Delta x \, dt \right)^{1/p}$$ (Hölder inequality using $P/2 \geq 1$)

$$\rightarrow 0$$ (by (3.3), (3.4)).

In conclusion, Murat’s lemma implies that $\eta_t + q_x$ is precompact in $H^{-1}(\Omega)$. \qed

Remarks.

- In practice, assumption (3.4) will be proved by showing

$$\int_0^T \sum_i |u_{i+1} - u_i|^p \, dt \leq C,$$

which clearly implies (3.4).

- The connection between (3.4), (3.5) and (3.6) is as follows. We may estimate the left-hand side of (3.5) using the entropy residual (2.5):

$$|\eta_t + Q_x(\Omega)| \leq \int_0^T \sum_i \|r_{i+1/2}\| \, dt = \int_0^T \sum_i \|v\|_{i+1/2} F_{i+1/2} - \|\psi\|_{i+1/2} \, dt$$

$$\leq \sum_i \|v\|_{i+1/2} F_{i+1/2} - \|f(u)\|_{i+1/2} + \|v\|_{i+1/2} F_{i+1/2} - \|f(u)\|_{i+1/2} \, dt$$

$$\leq C \sum_i |u_{i+1} - u_i|^2 \, dt,$$

by the Lipschitz continuity of $F_{i+1/2}$ and the definition of $q$ and $v$. Thus, should the sufficient condition (3.6) with $P = 2$ hold, then both (3.4) and (3.5) are satisfied. This is the bound asserted by DiPerna [DiP83, p. 59]. For $P > 2$, however, one cannot conclude (3.5) on the basis of (3.4) or (3.6).

- For a $p$-th order accurate numerical scheme, the entropy residual will be of the order of $|u_{i+1} - u_i|^{p+1}$ (see Chapter 4 and [Lax71]). Thus, it is essential to consider
the general case $P > 2$, as one cannot hope for (3.6) with $P = 2$ for a high-order accurate scheme.

We can now state the main theorem of this section, where we make the connection between the abstract convergence Lemma 3.2 and entropy stable methods.

**Theorem 3.3.** Let $\Omega \subset \mathbb{R} \times [0, T]$ be bounded and $(\eta, q)$ be an entropy pair such that either

(i) both $\eta$ and $f$ are strictly convex, or
(ii) $\eta = f$ and $f$ is genuinely nonlinear.

Let $u$ be computed by the finite difference scheme (2.1), and assume that there is a numerical entropy flux $Q_{i+1/2}$ such that

$$
\frac{d}{dt} \eta(u_i) + \frac{Q_{i+1/2} - Q_{i-1/2}}{\Delta x} \leq 0.
$$

Suppose that

(A1) $\exists M > 0 : \|u\|_{L^\infty(\Omega)} \leq M \quad \forall \Delta x > 0$

(A2) $\text{supp}(u) \subset \Omega \quad \forall \Delta x > 0$

(A3) $\int_0^T \sum_i |u_{i+1} - u_i|^P \Delta x dt \to 0 \quad \text{as } \Delta x \to 0 \text{ for some } P \in [2, \infty),$

and that $F$ and $Q$ satisfy the Lipschitz conditions

(A4)$_F$ $|f(u_0) - F(u_{m+1}, \ldots, u_m)| \leq C \left( |u_{m+2} - u_{m+1}| + \cdots + |u_m - u_{m-1}| \right)$

(A4)$_Q$ $|q(u_0) - Q(u_{m+1}, \ldots, u_m)| \leq C \left( |u_{m+2} - u_{m+1}| + \cdots + |u_m - u_{m-1}| \right)$

for all $u_{m+1}, \ldots, u_m \in [-M, M]$. Then there is a subsequence of $\{u_{\Delta x}\}_{\Delta x > 0}$ which converges a.e. in $\Omega$ and in $L^p(\Omega)$ for all $1 \leq p < \infty$ to a weak solution of (1.1) which is admissible with respect to $(\eta, q)$. In particular, in case (i), the whole sequence converges and the limit is the unique entropy solution.

**Proof.** We apply Lemma 3.2 to the entropy pairs $(id, f)$ and $(\eta, q)$ to show that the sequences $\{u_t + f(u)_{x}\}_{\Delta x > 0}$ and $\{\eta_t + q_x\}_{\Delta x > 0}$ are precompact in $H^{-1}(\Omega)$. Convergence then follows from Theorem 1.12 (case (ii)) or Theorem 1.13 (case (iii)).

Assumptions (A1), (A2), (A3) are precisely (3.2), (3.3), (3.4). Assumptions (A4)$_F$, (A4)$_Q$ are the same as (3.1) for the numerical entropy fluxes $F$ and $Q$, respectively. (3.5) trivially holds for $(id, f)$ since

$$
|u_t + F_{x}(\Omega)| = \int_0^T \sum_i \left| \frac{d}{dt} u_i + \frac{F_{i+1/2} - F_{i-1/2}}{\Delta x} \right| \Delta x dt = 0
$$
(by (2.1)), and for \( (\eta, q) \) since

\[
|\eta_l + Q_i(\Omega) - Q_{i+1/2} - Q_{i-1/2}| \Delta x \Delta t = - \int_0^T \sum_i \left( \frac{d}{dt} \eta(u_i) + \frac{Q_{i+1/2} - Q_{i-1/2}}{\Delta x} \right) \Delta x \Delta t
\]

(by (3.7))

\[
= \sum_i \eta(u_i(0)) \Delta x - \sum_i \eta(u_i(T)) \Delta x
\]

\[\leq 2 \text{diam}(\Omega) \|\eta\|_{L^\infty([-M,M])} \leq C
\]

since \( u \) is \( L^\infty \)-bounded on a compact domain and \( \eta \) is continuous. Hence, Lemma 3.2 implies the compactness property (1.22), so in case (i) and (ii) we may apply Theorem 1.12 and 1.13, respectively, implying the existence of a strongly convergent subsequence of \( \{u\}_{\Delta x > 0} \). By the Lax-Wendroff theorem, the limit is a weak solution. Admissibility with respect to \( (\eta, q) \) follows similarly.

In particular, in case (i), the limit is the unique entropy solution, by Panov’s Theorem 1.10. By the uniqueness of the limit, every convergent subsequence converges to the same limit, and hence the whole sequence converges to the entropy solution. \( \square \)

### 3.2. Multi-dimensional measure-valued solutions of systems

The compensated compactness method, as utilized in the previous section, is applicable only to one-dimensional, scalar conservation laws on bounded domains. What it lacks in generality, it makes up for in strength: we are able to show strong convergence to a distributional solution under the relatively weak assumptions \((A1)-(A4_0)\). In this section we take the opposite approach: by weakening our assumptions further, we will be able to prove the weaker result of (weak) convergence to a measure-valued solution for general systems in any number of space dimensions on unbounded domains.

We will restrict ourselves to two-dimensional conservation laws \((d = 2)\). The generalization to higher dimensions is straightforward, and is left out for notational simplicity. Thus, consider the two-dimensional conservation law (1.1); we denote the components of the spatial variable by \((x, y) \in \mathbb{R}^2\). The spatial domain is partitioned into cells \( I_{i,j} := [x_{i-1/2}, x_{i+1/2}) \times [y_{j-1/2}, y_{j+1/2}) \), and the solution \( u(x, y, t) \) is approximated by cell averages \( u_{i,j}(t) \approx \frac{1}{\Delta x \Delta y} \int_{I_{i,j}} u(x, y, t) \, dx \, dy \) or by point values \( u_{i,j}(t) \approx u(x_i, y_j, t) \). A prototypical finite volume (finite difference) scheme is of the form

\[
\frac{d}{dt} u_{i,j} + \frac{F^x_{i+1/2,j} - F^x_{i-1/2,j}}{\Delta x} + \frac{F^y_{i,j+1/2} - F^y_{i,j-1/2}}{\Delta y} = 0.
\]

(3.8)

The numerical flux function \((F^x_{i+1/2,j}, F^y_{i,j+1/2})\) is assumed to be consistent with (1.1), i.e., \((F^x(u, \ldots, u), F^y(u, \ldots, u)) = f(u)\). We will denote \( \Delta x u_{i,j} := u_{i+1,j} - u_{i,j} \) and \( \Delta y u_{i,j} := u_{i,j+1} - u_{i,j} \).
Lemma 3.4. Let \( N \geq 1 \), let \( \Omega \subset \mathbb{R}^2 \times [0, \infty) \), and let \( u = u^{\Delta x, \Delta y} \) be computed from (3.8). Assume that

(B1) \[ \exists M > 0 : \|u^{\Delta x, \Delta y}\|_{L^\infty(\Omega)} \leq M \quad \forall \Delta x, \Delta y \]

(B3) \[ \exists P \in [2, \infty) : \int_0^T \sum_{i,j} \left( |\Delta x u_{i,j}^{\Delta x, \Delta y}|^P + |\Delta y u_{i,j}^{\Delta x, \Delta y}|^P \right) \Delta x \Delta y dt \to 0 \quad \text{as} \quad \Delta x, \Delta y \to 0. \]

and that for all \( u_{m+1}, \ldots, u_m \),

(B4) \[
\begin{cases}
|f^x(u_0) - F^x(u_{m+1}, \ldots, u_m)| \leq C (|u_{m+2} - u_{m+1}| + \cdots + |u_m - u_{m-1}|) \\
|f^y(u_0) - F^y(u_{m+1}, \ldots, u_m)| \leq C (|u_{m+2} - u_{m+1}| + \cdots + |u_m - u_{m-1}|)
\end{cases}
\]

Then the Young measure associated with \( u_{\Delta x, \Delta y} \) is a measure-valued solution to (1.1).

Proof. (B1) gives the existence of the Young measure \( \nu \); cf. Theorem A.1. Denote \( h := \max(\Delta x, \Delta y) \). For all \( \varphi \in C_0^2(\Omega) \), we have

\[
\int_{\Omega} \langle \nu, \lambda \rangle \varphi + \langle \nu, f \rangle \cdot \nabla \varphi \, dxdt = \lim_{h \to 0} \int_0^T \sum_{i,j} \left( -\left( \frac{d}{dt} u_{i,j} \right) \varphi + f(u_{i,j}) \cdot \nabla \varphi \right) \, dxdt
\]

\[
= \lim_{h \to 0} \int_0^T \sum_{i,j} \left( \left( \frac{F^x_{i+1/2,j} - F^x_{i-1/2,j}}{\Delta x} + \frac{F^y_{i,j+1/2} - F^y_{i,j-1/2}}{\Delta y} \right) \varphi + f(u_{i,j}) \cdot \nabla \varphi \right) \, dxdt
\]

\[
= \lim_{h \to 0} \int_0^T \sum_{i,j} \left( \left( \frac{F^x_{i+1/2,j} - F^x_{i-1/2,j}}{\Delta x} + \frac{F^y_{i,j+1/2} - F^y_{i,j-1/2}}{\Delta y} \right) \varphi_{i,j} \Delta x \Delta y + f(u_{i,j}) \cdot \int_{\Omega} \nabla \varphi \, dx \right) dt
\]

\[
= \lim_{h \to 0} \int_0^T \sum_{i,j} \left( f^x(u_{i,j}) - F^x_{i+1/2,j} \right) \nabla \varphi \Delta x \Delta y + \left( f^y(u_{i,j}) - F^y_{i,j+1/2} \right) \nabla \varphi \Delta x \Delta y
\]

\[
+ f(u_{i,j}) \cdot \int_{\Omega} \left( \nabla \varphi - \nabla \varphi \right) \, dx \right) dt
\]

where we have denoted

\[
\varphi_{i,j} := \frac{1}{\Delta x \Delta y} \int_{\Omega_{i,j}} \varphi \, dxdy \quad \text{and} \quad \nabla \varphi := (\nabla \varphi^x, \nabla \varphi^y) = \left( \frac{\varphi_{i+1,j} - \varphi_{i,j}}{\Delta x}, \frac{\varphi_{i,j+1} - \varphi_{i,j}}{\Delta y} \right).
\]
The first term vanishes as $h \to 0$, since
\[
\int_0^T \sum_{i,j} \left( |f^x(u_{i,j}) - F^x_{i+1/2,j}| \left| \nabla \phi^x \right| + |f^y(u_{i,j}) - F^y_{i,j+1/2}| \left| \nabla \phi^y \right| \right) \Delta x \Delta y dt \\
\leq C \int_0^T \sum_{i,j} \left( \left| \Delta_x u_{i,j} \right| \left| \nabla \phi^x \right| + \left| \Delta_y u_{i,j} \right| \left| \nabla \phi^y \right| \right) \Delta x \Delta y dt \\
\leq C \left( \int_0^T \sum_{i,j} \left( \left| \Delta_x u_{i,j} \right|^P + \left| \Delta_y u_{i,j} \right|^P \right) \Delta x \Delta y dt \right)^{1/P} \left( \int_\Omega \left| \frac{\partial \phi}{\partial x} \right|^P + \left| \frac{\partial \phi}{\partial y} \right|^P \right)^{1/P} \\
\to 0
\]
by (B4), Hölder’s inequality and (B3). The second term also goes to zero because
\[
\sum_{i,j} \left( \nabla \phi^x, \nabla \phi^y \right) \mathbb{1}_{i,j} \to \nabla \phi
\]
pointwise.

Completely analogously, we can show convergence to an entropy measure-valued solution, provided the approximate solution satisfies a discrete entropy inequality (3.9)
\[
\frac{d}{dt} \eta(u_{i,j}) + \frac{Q^x_{i+1/2,j} - Q^x_{i-1/2,j}}{\Delta x} + \frac{Q^y_{i,j+1/2} - Q^y_{i,j-1/2}}{\Delta y} \leq 0
\]
for a numerical entropy flux $(Q^x_{i+1/2,j}, Q^y_{i,j+1/2})$ consistent with a convex entropy pair $(\eta, q)$, i.e.,
\[
(Q^x(u, \ldots, u), Q^y(u, \ldots, u)) = q(u).
\]

**Lemma 3.5.** Assume that $u$ satisfies (2.3) for a numerical entropy flux $Q$. Assume that (B1), (B3) are satisfied, and that for all $u_{-m+1}, \ldots, u_m$,

(B4)
\[
\begin{cases}
|q^x(u_0) - Q^x(u_{-m+1}, \ldots, u_m)| \leq C (|u_{-m+2} - u_{-m+1}| + \cdots + |u_m - u_{m-1}|) \\
|q^y(u_0) - Q^y(u_{-m+1}, \ldots, u_m)| \leq C (|u_{-m+2} - u_{-m+1}| + \cdots + |u_m - u_{m-1}|)
\end{cases}
\]

Then the Young measure $\nu$ associated with $\{u^{\Delta x, \Delta y}\}_{\Delta x, \Delta y > 0}$ is admissible with respect to $(\eta, q)$, i.e., it satisfies (1.26).

**Remark.** We may allow for a nonzero term on the right-hand side which vanishes weakly in $L^2(\Omega)$ as $\Delta x, \Delta y \to 0$. However, for the sake of simplicity we do not pose this general result.
3.2. MULTI-DIMENSIONAL MEASURE-VALUED SOLUTIONS OF SYSTEMS

Proof of Lemma 3.5. The proof follows that of Lemma 3.4 closely. For all \( \varphi \in C^1_c(\Omega) \), we have

\[
\int_\Omega \langle \nu, \eta \rangle \varphi_t + \langle \nu, q \rangle \cdot \nabla \varphi \, dx \, dt = \lim_{h \to 0} \int_0^T \sum_{i,j} \int_{I_i,j} \left( -\left( \frac{d}{dt} \eta(u_{i,j}) \right) \varphi + q(u_{i,j}) \cdot \nabla \varphi \right) \, dx \, dt
\]

\[
\geq \lim_{h \to 0} \int_0^T \sum_{i,j} \int_{I_i,j} \left( \frac{Q^x_{i+1/2,j} - Q^x_{i-1/2,j}}{\Delta x} + \frac{Q^y_{i,j+1/2} - Q^y_{i,j-1/2}}{\Delta y} \right) \varphi_{i,j} \Delta x \Delta y + q(u_{i,j}) \cdot \int_{I_i,j} \nabla \varphi \, dx \right) \, dt
\]

\[
= \lim_{h \to 0} \int_0^T \sum_{i,j} \left( \frac{Q^x_{i+1/2,j} - Q^x_{i-1/2,j}}{\Delta x} + \frac{Q^y_{i,j+1/2} - Q^y_{i,j-1/2}}{\Delta y} \right) \varphi_{i,j} \Delta x \Delta y + (q'(u_{i,j}) - Q^y_{i,j+1/2}) \nabla \varphi^y \Delta x \Delta y
\]

\[
+ q(u_{i,j}) \cdot \int_{I_i,j} (\nabla \varphi - \nabla \varphi^r) \, dx \right) \, dt
\]

with the same notation as in Lemma 3.4. As before, the first term vanishes as \( h \to 0 \), since

\[
\int_0^T \sum_{i,j} \left( |q^x(u_{i,j}) - Q^x_{i+1/2,j}| \left| \nabla \varphi^x \right| + \left| q^y(u_{i,j}) - Q^y_{i,j+1/2} \right| \left| \nabla \varphi^y \right| \right) \Delta x \Delta y \, dt
\]

\[
\leq C \int_0^T \sum_{i,j} \left( \left| \Delta x u_{i,j} \right| \left| \nabla \varphi^x \right| + \left| \Delta y u_{i,j} \right| \left| \nabla \varphi^y \right| \right) \Delta x \Delta y \, dt
\]

\[
\leq C \left( \int_0^T \sum_{i,j} \left( \left| \Delta x u_{i,j} \right|^p + \left| \Delta y u_{i,j} \right|^p \right) \Delta x \Delta y \, dt \right)^{1/p} \left( \int_\Omega \left| \frac{\partial \varphi}{\partial x} \right|^p + \left| \frac{\partial \varphi}{\partial y} \right|^p \right)^{1/p}
\]

\[
\to 0
\]

The second term vanishes for the same reason as in Lemma 3.4. \( \square \)
3.3. Convergence of entropy stable methods

As an illustration of the framework presented in Section 3.1, we shall prove convergence of the first-order accurate, entropy stable schemes for scalar, one-dimensional conservation laws developed in Section 2.3. The approach will be the same for higher-order accurate schemes in Chapter 6, although the technical details will be somewhat more involved.

Theorem 3.6. Let \( f \) be strictly convex and let \((\eta, q)\) be any entropy pair with \( \eta''(u) \geq \eta_0 \) \( \forall u \) for some \( \eta_0 > 0 \). Let \( u_0 \in L^\infty(\mathbb{R}) \) have compact support. Assume that the flux (2.8) is entropy stable with respect to \((\eta, q)\), and that \( c \leq D_{i+1/2} \leq c \) for all \( i \in \mathbb{Z} \) and \( t \in [0, T] \), for some \( c > 0 \). Assume that there is an \( M > 0 \) such that the solution computed by the scheme satisfies \( \|u^\Delta x\|_{L^\infty(\mathbb{R} \times [0, T])} \leq M \) for all \( \Delta x > 0 \). Assume that \( \|u\|_{L^p(\mathbb{R} \times [0, T])} \) for all \( 1 \leq p < \infty \) converges pointwise a.e. in \( \mathbb{R} \times [0, T] \) to the entropy solution of (1.1).

Proof. We show that conditions (A1) – (A4_\text{Q}) of Theorem 3.3 are satisfied. The assumptions (A1) on uniform boundedness and (A2) on compact support hold by hypothesis. Integrating the entropy dissipation estimate (2.9) over \( i \in \mathbb{Z} \), \( t \in [0, T] \) gives

\[
\sum_i \eta(u_i(T)) \Delta x - \sum_i \eta(u_i(0)) \Delta x = - \int_0^T \sum_i D_{i+1/2} \|v(t)\|_{i+1/2}^2 dt.
\]

Hence, we obtain the following bound on the spatial variation,

\[
\int_0^T \sum_i D_{i+1/2} \|v(t)\|_{i+1/2}^2 dt \leq C.
\]

Since \( \|u\| \leq \|u'(v)\|_{L^\infty(\mathbb{R})} \|v\| \), we have

\[
C \geq \int_0^T \sum_i D_{i+1/2} \|v(t)\|_{i+1/2}^2 dt \geq \frac{\eta_0^2}{c} \int_0^T \sum_i \|u(t)\|_{i+1/2}^2 dt,
\]

so the weak total variation bound (A3) is satisfied with \( P = 2 \). The Lipschitz conditions (A4_F) and (A4_Q) hold by the \( L^\infty \) bound on \( u \) and the boundedness of \( D_{i+1/2} \). The result now follows from Theorem 3.3 (i).

Remark. The only real, non-technical assumption in Theorem 3.6 is the \( L^\infty \)-bound on \( u^\Delta x \). Although choosing the entropy (2.11) would have provided such a bound, this particular entropy is not continuous, and hence Theorem 3.3 is not applicable.
CHAPTER 4

Higher-order accurate entropy stable methods

By Theorem 2.4, an entropy stable numerical flux can be written in the generic form

\[ F_{i+1/2} = \tilde{F}_{i+1/2} - D_{i+1/2} \|v\|_{i+1/2}, \]

where \( \tilde{F} \) is an entropy conservative flux function, \( v \) is the entropy variable and \( D_{i+1/2} \) is any positive matrix. In this chapter we investigate whether higher-order accurate schemes can admit such a representation. More precisely, we will design such schemes by replacing \( \tilde{F}_{i+1/2} \) by a high-order accurate entropy conservative flux, and deriving sufficient conditions on \( D_{i+1/2} \) for the scheme to be entropy stable and high-order accurate.

4.1. High-order entropy conservative methods

The entropy conservative schemes from Chapter 2, being 3-point schemes, are limited to second-order accuracy. However, following the procedure of LeFloch, Mercier and Rohde [LMR03], we can use these fluxes as building blocks to obtain \( 2k \)-th order accurate entropy conservative fluxes for any \( k \in \mathbb{N} \). These consist of linear combinations of second-order accurate entropy conservative fluxes \( \tilde{F} \), and have the form

\[ \tilde{F}^{2k}_{i+1/2} = \sum_{r=1}^{k} \alpha_r^k \sum_{s=0}^{r-1} \tilde{F}(u_{i-s}, u_{i-s-s+1}). \] (4.1)

**Theorem 4.1** ([LMR03], Theorem 4.4). For \( k \in \mathbb{N} \), assume that \( \alpha_1^k, \ldots, \alpha_k^k \) solve the \( k \) linear equations

\[ 2 \sum_{r=1}^{k} r \alpha_r^k = 1, \quad \sum_{s=0}^{2s-1} \alpha_r^k = 0 \quad (s = 2, \ldots, k), \]

and define \( \tilde{F}^{2k} \) by (4.1). Then the finite difference scheme with flux \( \tilde{F}^{2k} \) is

- \( 2k \)-th order accurate, in the sense that for sufficiently smooth solutions \( u \) we have

\[ \frac{\tilde{F}^{2k}(u_{i-k+1}, \ldots, u_{i+k}) - \tilde{F}^{2k}(u_{i-k}, \ldots, u_{i+k-1})}{\Delta x} = f(u) \bigg|_{u=u_i} + O(\Delta x^{2k}). \]

- entropy conservative – it satisfies the discrete entropy identity

\[ \frac{d}{dt} \eta(u_i(t)) + \frac{Q^{2k}_{i+1/2} - Q^{2k}_{i-1/2}}{\Delta x} = 0 \]
4.2. High-order entropy stable methods for scalar equations: the ELW scheme

Given the construction of arbitrarily high-order accurate entropy conservative fluxes in the previous section, we may now form the numerical flux

$$F_{i+1/2} = \tilde{F}^{2k}_{i+1/2} - D_{i+1/2} \|v\|_{i+1/2},$$

If $D \equiv 0$ then this gives a $2k$-th order accurate scheme, but without any numerical diffusion, computed solutions will have oscillations around shocks. Instead, we may require that $D_{i+1/2} = O(\|v\|^{p-1}_{i+1/2})$. This gives a scheme with truncation error $O(\Delta x^{2k} + \|v\|^p) = O(\Delta x^p)$, provided $2k \geq p$, and so is $p$-th order accurate, but still has at least some numerical diffusion. For instance, we can let $D_{i+1/2} = c_{i+1/2} \|v\|_{i+1/2}^{p-1}$, with $c_{i+1/2}$ being determined by the local wave speeds, say,

$$c_{i+1/2} = \frac{1}{2} \max(\{f'(u_i), f'(u_{i+1})\}).$$

We call the scheme with flux

$$F_{i+1/2} = \tilde{F}^{2k}_{i+1/2} - c_{i+1/2} \|v\|_{i+1/2}^{p-1} \|v\|_{i+1/2},$$

with $k$ such that $2k \geq p$, the $ELW_p$ (Entropy stable Lax-Wendroff) scheme. Applying Theorem 2.4, we obtain the following.

**Proposition 4.2.** The $ELW_p$ scheme is formally $p$-th order accurate and entropy stable; it satisfies the entropy dissipation estimate

$$\frac{d}{dt} \eta(u_i) + \frac{Q_{i+1/2} - Q_{i-1/2}}{\Delta x} = -\frac{c_{i+1/2} \|v\|_{i+1/2}^{2k+1}}{2\Delta x} + \frac{c_{i-1/2} \|v\|_{i-1/2}^{2k+1}}{2\Delta x} \leq 0.$$

From the entropy dissipation estimate (4.7) we can deduce the following convergence result.
Proposition 4.3. Assume that $f$ is strictly convex and that $(\eta, q)$ is uniformly convex. Let $u_0 \in L^\infty(\mathbb{R})$ have compact support. Assume that $c_{i+1/2} \geq \frac{1}{c}$ for all $i \in \mathbb{Z}$ for some $c > 0$, and that there is an $M > 0$ such that the solution $u^{\Delta x}$ computed by the ELW$_p$ scheme satisfies $\|u^{\Delta x}\|_{L^\infty(\mathbb{R} \times [0, T])} \leq M$ for all $\Delta x > 0$. Then $u^{\Delta x}$ converges pointwise a.e. in $\mathbb{R} \times [0, T]$ and in $L^p(\mathbb{R} \times [0, T])$ for all $1 \leq p < \infty$ to the entropy solution of (1.1).

Proof. As in Theorem 3.6, we show that conditions (A1)–(A4$_Q$) of Theorem 3.3 (i) are satisfied. Conditions (A1), (A2) hold by hypothesis. For (A3), we integrate (4.7) over $i \in \mathbb{Z}$, $t \in [0, T]$ to find that

$$\int_0^T \sum_i c_{i+1/2} \|v\|_{i+1/2}^{2k+1} dt \geq \frac{\eta_0}{c} \int_0^T \|u\|_{i+1/2}^{2k+1} dt$$

where $\eta_0 > 0$ is such that $\eta''(u) \geq \frac{1}{\eta_0}$ for all $u$. Hence, (A3) holds with $P = 2k + 1$. The Lipschitz conditions (A4$_F$), (A4$_Q$) follow from the Lipschitz continuity of $\tilde{F}$ and the $L^\infty$-bound on $u$. □

4.2.1. Numerical experiment: Linear advection. Consider the one-dimensional linear advection equation (1.2). We use initial data $u_0(x) = \sin(\pi x)$, compute on the periodic domain $x \in [-1, 1]$ with 50 gridpoints up to $t = 1$, and we use a CFL number of 0.9. The computed solution is shown in Figure 4.1. There is very little loss of accuracy, and it is difficult to discern the computed solution from the exact solution.

![Exact vs ELW3](image)

**Figure 4.1.** Linear advection equation with sinusoidal initial data. Right: closeup around $x = -0.5$.

Next, we compute the experimental order of accuracy for the third- and fourth-order ELW schemes. As shown in Table 4.2, the expected order of accuracy is reached immediately on this smooth test problem.
4.2.2. Numerical experiment: Burgers’ equation. Consider Burgers’ equation (1.3) with initial data $u_0(x) = (1 + \sin(\pi x))/2$. This Cauchy problem develops a discontinuity at $t = \frac{2}{n} \approx 0.636$. We compute in the domain $x \in [-1, 1]$ up to $t = 2$ using periodic boundary conditions. The solution computed with the fourth-order ELW$_4$ scheme on meshes with 50, 100 and 200 grid points is shown in Figure 4.3(a). For comparison, we plot the solution computed with the fourth-order entropy conservative scheme (4.3) in Figure 4.3(b). Note that this is the entropy conservative flux used in the ELW$_4$ scheme (4.6). The ELW diffusion operator clearly has a smoothening effect on the underlying entropy conservative scheme. The ELW scheme still has oscillations in the wake of the shock, although it is reasonably accurate in the rest of the domain. As the grid size is lowered, the magnitude of the oscillations away from the shock decreases.

4.3. Reconstruction based entropy stable schemes: scalar equations

Although the ELW scheme is entropy stable, high-order accurate and displays the correct order of convergence in numerical examples, it exhibits large oscillations around shocks. This motivates the use of higher-order nonoscillatory reconstruction instead of the rather naive choice of diffusion in (4.6). In this section we derive necessary conditions on the reconstruction procedure to obtain an entropy stable scheme for scalar equations. The procedure is generalized to systems in the next section.

For our purposes, a $p$-th order reconstruction procedure is a map $\mathcal{R}_i : \{v_j\}_{j \in \mathbb{Z}} \mapsto v_i(x)$ that takes a set of point values $\{v_j\}_{j \in \mathbb{Z}}$, and for each cell $I_i$ produces a $(p - 1)$-th order polynomial $v_i(x)$ which (a) interpolates the point value $v_i$, and (b) approximates the underlying function to order $O(\Delta x^p)$. (These notions are made more precise in Chapter 5; see also [Shu99]). In particular, we are interested in the cell interface values $v_i^\pm := v_i(x_{i+1/2})$. Note that since the entropy conservative fluxes from Section 4.1 are finite difference schemes, we are interpolating over point values, not cell averages. The initial data is sampled correspondingly.
4.3. RECONSTRUCTION BASED ENTROPY STABLE SCHEMES: SCALAR EQUATIONS

Figure 4.3. Burgers’ equation. The exact solution is shown in red.

Schemes of the form (4.5) will generally be of order $O(\Delta x)$ whenever $D_{i+1/2} = O(1)$. Instead of the diffusion operator $D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2}$ we shall consider diffusion terms of the form $D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2}$, where $\langle \langle v \rangle \rangle_{i+1/2} := v^+_i - v^-_i$ and $v^\pm_i$ are the cell interface values of a $p$-th order reconstructed function $v_i(x) := R_i((v_j)_{j \in \mathbb{Z}})$. This gives a numerical flux of the form

\begin{equation}
F^p_{i+1/2} = \tilde{F}^{2k}_{i+1/2} - D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2},
\end{equation}

where $D_{i+1/2} \geq 0$ is yet to be determined. Since $D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2} = O(\Delta x^p)$, the truncation error of this method is of order $O(\Delta x^{2k} + \Delta x^{p-1})$; if $D_{i+1/2}$ is continuous with respect to $u$, then the truncation error is $O(\Delta x^{2k} + \Delta x^p)$. Hence, if $k$ is chosen such that $2k \geq p$ then the scheme is (formally) $p$-th order accurate. Note that the reconstructed variables are only used in the diffusion operator, not in the entropy conservative flux $\tilde{F}^{2k}$.

To see if the scheme is entropy stable we write

\begin{equation}
D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2} = D_{i+1/2} \frac{\langle \langle v \rangle \rangle_{i+1/2}}{\| v \|_{i+1/2}} \| v \|_{i+1/2}.
\end{equation}

Thus, if $\frac{\langle \langle v \rangle \rangle_{i+1/2}}{\| v \|_{i+1/2}} \geq 0$, or equivalently,

\begin{equation}
\langle \langle v \rangle \rangle_{i+1/2} = B_{i+1/2} \| v \|_{i+1/2}
\end{equation}

where $B_{i+1/2} \geq 0$, the scheme is entropy stable.
for a $B_{i+1/2} \geq 0$, then the diffusion term $D_{i+1/2} \langle v \rangle_{i+1/2}$ can be written as a nonnegative term times $\|v\|_{i+1/2}$, and hence the scheme is entropy stable. We will say that the reconstruction procedure $\mathcal{R}$ satisfies the sign property if the reconstructed values $v_i^+, v_{i+1}^-$ satisfy (4.9).

**Example 4.4.** We investigate which second-order reconstruction methods satisfy (4.9). Let $\varphi$ be a slope limiter with the symmetry property $\varphi(\theta^{-1}) = \varphi(\theta)\theta^{-1}$ for all $\theta \neq 0$ (see [LeV02]). Define the quotients

$$
\theta_i^- = \frac{\|v\|_{i+1/2}}{\|v\|_{i-1/2}}, \quad \theta_i^+ = \frac{\|v\|_{i-1/2}}{\|v\|_{i+1/2}}.
$$

We denote the “slope” in grid cell $I_i$ by

$$
\sigma_i = \frac{1}{\Delta x} \varphi(\theta_i^-) \|v\|_{i-1/2} = \frac{1}{\Delta x} \varphi(\theta_i^+) \|v\|_{i+1/2},
$$

the second equality follows from the symmetry of $\varphi$. Hence, the reconstructed values at the left and right cell interfaces of grid cell $I_i$ are given by $v_i^- = v_i - \frac{1}{2} \varphi(\theta_i^-) \|v\|_{i-1/2}$ and, respectively, $v_i^+ = v_i + \frac{1}{2} \varphi(\theta_i^+) \|v\|_{i+1/2}$. We obtain

$$
\langle v \rangle_{i+1/2} = v_{i+1} - v_i - \frac{1}{2} \left( \varphi(\theta_i^+) + \varphi(\theta_{i+1}^-) \right) \|v\|_{i+1/2} = \left( 1 - \frac{1}{2} \left( \varphi(\theta_i^+) + \varphi(\theta_{i+1}^-) \right) \right) \|v\|_{i+1/2}.
$$

Hence, since $\theta_i^+$ and $\theta_{i+1}^-$ may be chosen independently of each other, the sign property (4.9) is satisfied if and only if $\varphi(\theta) \leq 1$ for all $\theta \in \mathbb{R}$. It is easily seen that the minmod limiter, given by

$$
\varphi_{\text{min}}(\theta) = \begin{cases} 
0 & \text{if } \theta < 0 \\
\theta & \text{if } 0 \leq \theta \leq 1 \\
1 & \text{otherwise}
\end{cases}
$$

satisfies $\varphi(\theta) \leq 1$. In fact, the minmod limiter is the only symmetric TVD limiter that satisfies the sign property. 

### 4.4. Reconstruction based entropy stable schemes: systems of equations

In this section we generalize the method described in the previous section to systems of conservation laws. The generalization of the construction of the higher-order entropy conservative flux $\bar{F}^{2k}$ is trivial, the only difference being that the expression (4.1) is vector-valued instead of scalar. The generalization of the diffusion term $D_{i+1/2} \langle v \rangle_{i+1/2}$ is less obvious. As in Section 2.5, we consider diffusion matrices which are of the form

$$
D_{i+1/2} = R_{i+1/2}A_{i+1/2}R_{i+1/2}^T, \quad R_{i+1/2} \text{ invertible, } A_{i+1/2} \geq 0 \text{ diagonal.}
$$

Like in the previous section, we define our numerical flux as

$$
F^p_{i+1/2} = \bar{F}^{2k}_{i+1/2} - D_{i+1/2} \langle v \rangle_{i+1/2},
$$
where $D_{i+1/2}$ is now a nonnegative matrix and $\langle \langle v \rangle \rangle_{i+1/2} = v_{i+1}(x_{i+1/2}) - v_i(x_{i+1/2})$ for some reconstructed function $v_i(x)$, yet to be determined. The following preliminary lemma generalizes the computations leading to the sign condition (4.9) to systems of equations.

**Lemma 4.5.** For each $i \in \mathbb{Z}$, let $D_{i+1/2}$ be given by (4.11). Let $v_i(x)$ be a polynomial reconstruction of the entropy variables in the cell $I_i$ such that for each $i$, there exists a diagonal matrix $B_{i+1/2} \geq 0$ such that

\begin{equation}
\langle \langle v \rangle \rangle_{i+1/2} = (R^T_{i+1/2})^{-1} B_{i+1/2} R^T_{i+1/2} \|v\|_{i+1/2}
\end{equation}

Then the scheme with numerical flux (4.12) is entropy stable with numerical entropy flux

\[ Q^{p}_{i+1/2} = \bar{Q}^{2k}_{i+1/2} - \frac{1}{2} \bar{v}_{i+1/2} \cdot D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2} \]

and $\bar{Q}^{2k}$ is defined in (4.2). More precisely, it satisfies the entropy dissipation estimate

\begin{equation}
\frac{d}{dt} \eta(u_i) + \frac{Q^{p}_{i+1/2} - Q^{p}_{i-1/2}}{\Delta x} = - \frac{\|v\|_{i+1/2} \cdot D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2} + \|v\|_{i-1/2} \cdot D_{i-1/2} \langle \langle v \rangle \rangle_{i-1/2}}{2\Delta x} \leq 0.
\end{equation}

**Proof.** Multiplying the finite difference scheme (2.1) by $v^T_i$ and imitating the proof of Theorem 2.4, we obtain

\[ \frac{d}{dt} \eta(u_i) = - \frac{1}{\Delta x} \left( \bar{Q}^{2k}_{i+1/2} - \bar{Q}^{2k}_{i-1/2} \right) + \frac{1}{\Delta x} \left( v^T_i D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2} - v^T_i D_{i-1/2} \langle \langle v \rangle \rangle_{i-1/2} \right) \]

\[ = - \frac{1}{\Delta x} \left( Q^{p}_{i+1/2} - Q^{p}_{i-1/2} \right) - \frac{1}{2\Delta x} \left( \|v\|^T_{i+1/2} D_{i+1/2} \langle \langle v \rangle \rangle_{i+1/2} + \|v\|^T_{i-1/2} D_{i-1/2} \langle \langle v \rangle \rangle_{i-1/2} \right). \]

Suppressing vector and matrix indices $i + 1/2$ for the moment, we have by (4.11) and (4.13)

\[ \|v\|^T D \langle \langle v \rangle \rangle = R^T \bar{R}^T R^T B R^T \langle \langle v \rangle \rangle = \|v\|^T \bar{R} \bar{R}^T \|v\| = \left( R^T \|v\| \right)^T A B \left( R^T \|v\| \right) \geq 0 \]

(since $B_{i+1/2} \geq 0$), and (4.14) follows. \hfill \Box

**Remark.** If the reconstructed variables satisfies (4.13), then the numerical flux (4.12) admits the equivalent representation

\[ F^p_{i+1/2} = F^{2k}_{i+1/2} - \frac{1}{2} R_{i+1/2} A_{i+1/2} B_{i+1/2} R^T_{i+1/2} \|v\|_{i+1/2}. \]

This reveals the role of $B_{i+1/2}$ as limiting the amount of numerical diffusion: in smooth parts of the flow, we have $B_{i+1/2} \approx 0$, and we are left with the entropy conservative flux. \hfill \blacksquare

**4.4.1. Reconstruction along scaled entropy variables.** Lemma 4.5 provides sufficient conditions on the reconstruction for the scheme to be entropy stable. It remains to find a reconstruction procedure that satisfies the crucial condition (4.13). Assume for the moment that $v_i, v_{i+1}$ and $v^+_i, \bar{v}^+_i$ are given. Define the *scaled entropy variables*

\begin{equation}
\langle \langle \bar{w} \rangle \rangle_{i+1/2} = B_{i+1/2} \langle \langle w \rangle \rangle_{i+1/2}, \quad B_{i+1/2} \geq 0 \text{ diagonal,}
\end{equation}

The condition (4.13) now reads

\begin{equation}
\langle \langle \bar{w} \rangle \rangle_{i+1/2} = B_{i+1/2} \langle \langle w \rangle \rangle_{i+1/2}, \quad B_{i+1/2} \geq 0 \text{ diagonal,}
\end{equation}
where \( \langle \tilde{w} \rangle_{i+1/2} = \tilde{w}_{i+1} - \tilde{w}_i^+ = R^T_{i+1/2} \langle v \rangle_{i+1/2} \) and \( \langle w \rangle_{i+1/2} = w_{i+1} - w_i^+ = R^T_{i+1/2} \| v \|_{i+1/2} \). Thus, each component of \( \tilde{w}_i \) must satisfy the property (4.9).

Let \( \mathcal{R} \) be a reconstruction procedure that satisfies the sign property (4.9). As the values we are reconstructing, the scaled entropy variables, are centred at cell interfaces, we must modify the reconstruction method somewhat. For a fixed \( l \in \{1, \ldots, N\} \), denote the \( l \)-th component of \( w \) and \( \tilde{w} \) by \( z \) and \( \tilde{z} \), respectively. For each \( i \), define the point value \( \alpha_i^l = z_i^- \), and inductively

\[
\begin{align*}
\alpha_{j+1}^l &= \alpha_j^l + \langle z \rangle_{j+1/2} \quad (j = i, i + 1, \ldots) \\
\alpha_{j-1}^l &= \alpha_j^l - \langle z \rangle_{j-1/2} \quad (j = i, i - 1, \ldots)
\end{align*}
\]

(recall that \( \langle z \rangle_{j+1/2} = z_{j+1}^- - z_j^+ \), the \( l \)-th component of \( \langle w \rangle_{j+1/2} \)). Similarly, we define \( \beta_i^l = z_i^+ \) and

\[
\begin{align*}
\beta_{j+1}^l &= \beta_j^l + \langle z \rangle_{j+1/2} \quad (j = i, i + 1, \ldots) \\
\beta_{j-1}^l &= \beta_j^l - \langle z \rangle_{j-1/2} \quad (j = i, i - 1, \ldots).
\end{align*}
\]

Then \( \alpha \) and \( \beta \) retain the cell interface jumps of \( z \),

\[
\| \alpha_i^l \|_{j+1/2} = \| \beta_i^l \|_{j+1/2} = \langle z \rangle_{j+1/2} \quad \text{for all } j.
\]

As \( \beta \) and \( \alpha \) have the same jump at a cell interface, we have

\[
\alpha_{j+1}^l = \beta_j^l \quad \text{for all } j.
\]

Let \( \Phi_j^l(x) := \mathcal{R}_j(\alpha_k^l) \) and \( \Psi_j^l(x) := \mathcal{R}_j(\beta_k^l) \) be the reconstructions of \( \alpha_i^l \) and \( \beta_i^l \) in cell \( I_j \). We obtain left and right reconstructed values:

\[
\tilde{z}_i^- := \Phi_j^l(x_i^-) \quad \text{and} \quad \tilde{z}_i^+ := \Psi_j^l(x_i^+) \quad \text{for all } i.
\]

This process is repeated in each grid cell \( I_i \) and for each component of \( w_i^\pm \). Finally, we define the reconstructed values \( w_i^\pm \) by

\[
v_i^\pm := \left( R^T_{i+1/2} \right)^{-1} \tilde{w}_i^\pm,
\]

cf. (4.15).

**Lemma 4.6.** If the reconstruction procedure \( \mathcal{R} \) satisfies the sign property (4.9), then the reconstructed values (4.19) satisfy (4.13).

**Proof.** By construction it suffices to show that the reconstructed scaled entropy variables \( \tilde{w}_i^\pm \) satisfy (4.16). Again, denote the \( l \)-th component of \( w_i \) and \( \tilde{w}_i \) by \( z_i^- \) and \( \tilde{z}_i \), respectively. Because of (4.18), the piecewise polynomial function \( \Psi^l \) is precisely equal to \( \Phi^{i+1} \). Hence,

\[
\langle \tilde{z} \rangle_{i+1/2} = \tilde{z}_{i+1}^- - \tilde{z}_i^+ = \Phi^{i+1}_j(x_{i+1/2}) - \Psi^l_j(x_{i+1/2}) = \Phi^l_{i+1}(x_{i+1/2}) - \Psi^l_j(x_{i+1/2}),
\]

which by assumption has the same sign as \( \beta_i^l - \beta_j^l \), which by definition equals \( \langle z \rangle_{i+1/2} \). \( \square \)
Remarks. We make a few remarks regarding the computational complexity of the reconstruction procedure detailed above.

• Although the definition (4.19) of the reconstructed entropy variables depends on the inverse of $R^T_{i+1/2}$, it is not necessary to compute this inverse. Indeed, inserting (4.19) into the diffusion operator $D_{i+1/2} \langle v \rangle_{i+1/2}$ of our scheme, we get

$$D_{i+1/2} \langle v \rangle_{i+1/2} = R_{i+1/2} A_{i+1/2} (R^T_{i+1/2})^{-1} \langle \tilde{w} \rangle_{i+1/2} = R_{i+1/2} A_{i+1/2} \langle \tilde{w} \rangle_{i+1/2}.$$

• Most reconstruction procedures $R$ encountered in practice, such as the TVD, ENO and WENO procedures, satisfy the additional property of being linear with respect to constants:

$$(4.20) \qquad R_i (\{c + v_j\}_{j \in \mathbb{Z}}) = R_i (\{v_j\}_{j \in \mathbb{Z}}) \quad \forall \ c \in \mathbb{R}.$$ Let $K \in \mathbb{Z}$ be a fixed index in the mesh. Since

$$\beta_j^i = \beta_j^K + (\beta_i^i - \beta_i^K) \quad \text{for all } j,$$

it follows from the linearity property (4.20) that

$$\Psi_j^i (x) = \Psi_j^K (x) + (\beta_i^i - \beta_i^K) \quad \text{for all } j \in \mathbb{Z} \text{ and } x \in I_j.$$

In fact, we will only be interested in the cell interface jumps, which by the discussion above can be written as

$$\tilde{z}_{i+1}^- - \tilde{z}_i^+ = \Psi^K_{i+1/2} (x_{i+1/2}) - \Psi^K_i (x_{i+1/2}).$$

Hence, it suffices to reconstruct on only one mesh $\{\beta_j^K\}_{j \in \mathbb{Z}}$ for all grid cells, as opposed to reconstructing on different meshes for each cell. This greatly increases the efficiency of the reconstruction procedure, and should be contrasted to the more standard approach of reconstructing locally over meshes $\{R_{i+1/2}^{-1} u_j\}_{j \text{ close to } i}$; see [Shu99].

• Finally, most reconstruction procedures depend only on the (un)divided differences $[\beta_j^K, \ldots, \beta_{j+p}^K]$ of the data points. By (4.17), we have $[\beta_j^K, \beta_{j+1}^K] = [z_j^+, z_{j+1}^+]$, so the data points $\{z_j^+\}_{j \in \mathbb{Z}}$ may be inserted directly in the first step of the computation of (un)divided differences. Hence, the values $\{\beta_j^K\}_{j \in \mathbb{Z}}$ do not even have to be computed.

4.5. Summary and outlook

We have presented high-order accurate entropy conservative and entropy stable schemes for both scalar conservation laws and systems of conservation laws. The main idea is to combine high-order accurate entropy conservative fluxes, described in Section 4.1, with entropy stable diffusion operators, described in subsequent sections.

The ELW scheme, presented in Section 4.2, is high-order accurate, entropy stable and converges for convex scalar conservation laws. However, the scheme has unacceptably large oscillations near shocks. Instead of tuning the diffusion coefficient to particular problems, we
considered nonoscillatory reconstruction procedures. The main conclusion from Section 4.3 was that if the reconstruction procedure satisfies the sign property — namely, that the jump $v_{i+1}^r - v_i^r$ in reconstructed values at the cell interface has the same sign as the jump $v_{i+1} - v_i$ in point values — then the resulting flux (4.8) is high-order accurate and entropy stable. This was generalized to systems of equations in Section 4.4, where we described a reconstruction procedure along the components of the so-called scaled entropy variables. This procedure gives an entropy stable scheme, provided the underlying, scalar reconstruction procedure $\mathcal{R}$ satisfies the sign property.

We found that the only symmetric second-order TVD limiter satisfying the sign property is the minmod limiter. However, the question of the existence of higher-order accurate reconstruction procedures satisfying the sign property was left open. In Chapter 5 we will see that the ENO procedure satisfies the sign property. This will allow us to construct non-oscillatory, arbitrarily high-order accurate finite difference schemes in Chapter 6.
CHAPTER 5

ENO stability properties

The ENO, or Essentially Non-Oscillatory, reconstruction procedure was introduced in 1987 by Harten et al. [HEOC87] in the context of accurate simulations for piecewise smooth solutions of nonlinear conservation laws. The ENO procedure generates a piecewise polynomial approximation of a function from its cell averages. The essence of the ENO procedure is its ability to accurately recover discontinuous functions. The starting point is a collection of cell averages \( \{\bar{u}_i\}_{i \in \mathbb{Z}} \) over consecutive intervals \( I_i = [x_{i-1/2}, x_{i+1/2}) \),

\[
\bar{u}_i := \frac{1}{|I_i|} \int_{I_i} u(x) \, dx,
\]

from which one can form the piecewise constant approximation of the underlying function \( u(x) \),

\[
\tilde{u}(x) := \sum_i \bar{u}_i \mathbb{1}_{I_i}(x).
\]

But the averaging operator \( u \mapsto \bar{u} \) is limited to first order accuracy, whether \( u \) is smooth or not. The purpose of the ENO procedure is to reconstruct a higher order approximation of \( u(x) \) from its given cell averages,

\[
\{\bar{u}_i\}_{i \in \mathbb{Z}} \mapsto \tilde{u}(x)
\]

for a piecewise polynomial function \( \tilde{u} \) approximating the underlying function \( u \) to order \( O(h^p) \). In the context of finite volume methods for hyperbolic conservation laws, this reconstructed function is inserted in the numerical flux function to obtain a scheme of the generic form

\[
\frac{d}{dt} u_i + \frac{1}{\Delta x} (F_{i+1/2} - F_{i-1/2}) = 0, \quad F_{i+1/2} := F(\tilde{u}_i^+, \tilde{u}_{i+1}^-)
\]

where \( \tilde{u}_i^\pm = \tilde{u}(x_{i \pm 1/2}) \).

Since its introduction, the ENO procedure and its extensions, [SO89, Har89, Shu90], have been used with a considerable success in computational fluid dynamics; we refer to the review article of Shu [Shu99] and the references therein. Moreover, ENO and its various extensions, in particular, with subcell resolution scheme (ENO-SR), [Har89], have been applied to problems in data compression and image processing in [Har89, CZ02, CDM03, ACDD05, ACD+08] and references therein.

There are only a few rigorous results about the global accuracy of the ENO procedure. In [ACDD05], the authors proved the second-order accuracy of ENO-SR reconstruction of
piecewise-smooth $C^2$ data. Multi-dimensional global accuracy results for the so-called ENO-EA method were obtained in [ACD*08]. Despite the extensive literature on the construction and implementation of ENO and its variants for the last 25 years, we are not aware of any global, mesh independent stability properties.

In the present chapter we introduce and prove the sign-property of the ENO procedure, which says that the jump $\tilde{u}_{i+1}^+ - \tilde{u}_i^-$ of reconstructed values $\tilde{u}$ at the cell interfaces $x_{i+1/2}$ always has the same sign as the jump $\bar{u}_{i+1} - \bar{u}_i$ in cell averages. Moreover, we provide an upper bound to the jump in reconstructed values relative to the jump in cell averages. The importance of stability bounds on the values $\tilde{u}_i^+, \tilde{u}_{i+1}^-$ will be seen in the finite volume method (5.2); it is precisely these values that are used in the finite volume method. That being said, we emphasize that the results in this chapter are purely results on reconstruction and interpolation theory, and no further assumptions or connections to numerical methods for differential equations will be made.

The results presented in this chapter is joint work with Eitan Tadmor and have been published in [FMT12b].

5.1. The ENO reconstruction method

The ENO procedure is a reconstruction procedure, taking cell averages $\bar{u}_i$ and constructing a piecewise $(p-1)$-th order polynomial function $\tilde{u}(x)$,

\[
\tilde{u}(x) = \sum_i \tilde{u}_i \mathbb{1}_{I_i}(x) \quad \mapsto \quad \tilde{u}(x) := \sum_i \tilde{u}_i(x) \mathbb{1}_{I_i}(x),
\]

which addresses the following requirements:

**Accurate:** It approximates $u(x)$ up to order $p$ in the sense that

\[
(5.3) \quad \tilde{u}(x) = u(x) + O(h^p), \quad h = \max_i |I_i|
\]

whenever $u$ is sufficiently smooth in a neighborhood of $x$. 

**Conservative:** It is conservative in sense of conserving the original cell averages,

\[
(5.4) \quad \frac{1}{|I_i|} \int_{I_i} \tilde{u}_i(x) dx = \bar{u}_i.
\]

**Non-oscillatory:** It avoids “Gibbs phenomena”.

The accuracy requirement suggests applying a standard polynomial interpolation. To take into account the conservation requirement, this interpolation is performed on the primitive of $\tilde{u}$; differentiating the resulting interpolant, it will be seen that the conservation requirement is fulfilled. Last, the rather vague criterion of being non-oscillatory is addressed by performing the interpolation over a variable, data-dependent stencil; it is the latter that is the heart of the ENO method.

Let $U(x) := \int_{-\infty}^x u(s) \, ds$ denote the primitive of $u(x)$. Although the underlying function $u(x)$ might be unavailable, its primitive $U(x)$ is computable at the cell interfaces $x_{j+1/2}$ through
the relation

\begin{equation}
U_{j+1/2} = \int_{-\infty}^{x_{j+1/2}} u(s) ds = \sum_{k=-\infty}^{j} \int_{x_{k-1/2}}^{x_{k+1/2}} u(s) ds = \sum_{k=-\infty}^{j} |I_k| \bar{u}_k.
\end{equation}

If \( \tilde{U}(x) \) is any function that interpolates \( \{U_{j+1/2}\}_{j \in \mathbb{Z}} \), so that \( \tilde{U}(x_{j+1/2}) = U_{j+1/2} \ \forall \ j \), then (5.5) implies that its derivative \( \tilde{\bar{u}}(x) := \tilde{U}'(x) \) satisfies (5.4). If \( F \) moreover satisfies \( \tilde{U}(x) = U(x) + O(h^{p+1}) \) then \( \tilde{\bar{u}} \) satisfies the accuracy requirement (5.3). This leads us to defining \( \tilde{U}_i(x) \) as the \( p \)-th order polynomial interpolating \( p + 1 \) point values \( U_{r_i}, \ldots, U_{r_i+p-1/2}, U_{r_i+p+1/2}, \ldots, U_{r_i+p} \). To address the non-oscillatory requirement, the (half-integer) left stencil index \( r \) must be determined in a data-dependent manner.

When the underlying data is sufficiently smooth, the accuracy requirement can be met by interpolating the primitive \( U \) on any set of \( p + 1 \) point-values

\( \{U_r, \ldots, U_{i-1/2}, U_{i+1/2}, \ldots, U_{r+p}\} \).

To satisfy the conservation property (5.4), the stencil of interpolation must include \( U_{i-1/2} \) and \( U_{i+1/2} \). There are \( p \) such stencils, ranging from the leftmost stencil corresponding to an index \( r = i - p + 1/2 \) to the rightmost stencil corresponding to an index \( r = i - 1/2 \). Since we are interested in the approximation of piecewise smooth functions, we must carefully select the stencil in order to avoid spurious oscillations. The main idea behind the ENO procedure is the use of a stencil with a data-dependent index, \( r = r_i \), which is adapted to the smoothness of the data. The choice of ENO stencil is accomplished in an iterative manner, based on divided differences of the data.

**Algorithm 5.1** (ENO reconstruction – stencil selection).

Let point values of the primitive \( U_{i-p+1/2}, \ldots, U_{i+p-1/2} \) be given. Then the left stencil indices \( r_i^1, \ldots, r_i^p \) are determined as follows.

\[
\begin{align*}
  r_i^1 &\leftarrow i - 1/2 \\
  \text{for } k = 1 \rightarrow p - 1 \text{ do} \\
    \text{if } |U[x_{r_i^{k-1}}, \ldots, x_{r_i^{k+1}}]| < |U[x_{r_i^{k}}, \ldots, x_{r_i^{k+1}}]| \text{ then} \\
    & r_i^{k+1} \leftarrow r_i^k - 1 \\
    \text{else} \\
    & r_i^{k+1} \leftarrow r_i^k \\
    \text{end if} \\
  \text{end for}
\end{align*}
\]

The divided differences \( U[x_j, \ldots, x_{j+k}] \) are a good measure of the \( k \)-th order of smoothness of \( U(x) \). Thus, the ENO procedure is based on data-dependent stencils which are chosen in the direction of smoothness, in the sense of preferring the smallest divided differences.
The Newton representation of the $p$-th degree interpolant $\bar{U}_i(x)$, based on point-values $U(x_{i,r}), \ldots, U(x_{i+1,r+p})$, is given by

\begin{equation}
\bar{U}_i(x) = \sum_{k=0}^{p} U[x_{i,r}, \ldots, x_{i+k}] \prod_{m=0}^{k-1} (x - x_{i-1+m}). \label{ENO_interpolant}
\end{equation}

(We set $r_{i}^{-1} = r_{i}^0 = r_{i}^1 = i - \frac{1}{2}$.) Observe that in (5.6), we have summed the contributions of stencils in their “order of appearance” in the stencil selection procedure, rather than the usual sum of stencils from left to right. Differentiating $\bar{U}_i$, we obtain the ENO reconstruction:

**Definition 5.2.** For each cell $I_i$, let the stencil index $r_{i}^p$ be computed according to Algorithm 5.1; let $\bar{U}_i(x)$ be the interpolant of $V$ over the stencil $\{U_{j}^{r_{i}^{p}+p}\}_{j=r_{i}^p}$, and define

\begin{equation}
\tilde{u}_i(x) := \bar{U}_i(x) = \sum_{k=1}^{p} U[x_{i,r}, \ldots, x_{i+k}] \prod_{m=0}^{k-1} (x - x_{i-1+m}). \label{ENO_reconstruction}
\end{equation}

Then the composite function $\tilde{u}(x) = \sum_{i} \tilde{u}_i(x) \mathbb{1}_{I_i}(x)$ is called the \textit{p-th order ENO reconstruction} of $\{\tilde{u}_i\}_{i \in \mathbb{Z}}$.

### 5.2. The ENO sign property

We can now state the main result of this chapter.

**Theorem 5.3** (The sign property). Fix an integer $p > 1$, and let $\tilde{u}_i(x) = \sum_{i} \tilde{u}_i(x) \mathbb{1}_{I_i}(x)$ be the $p$-th order ENO reconstruction of $\{\tilde{u}_i\}_{i \in \mathbb{Z}}$. Let $\tilde{u}_i^+ := \tilde{u}_i(x_{i+1/2})$ and $\tilde{u}_i^- := \tilde{u}_i(x_{i-1/2})$ denote the reconstructed point-values to the left and right of the cell interface $x_{i+1/2}$. Then the following sign property holds at all interfaces:

\begin{equation}
\begin{cases}
\text{if } \tilde{u}_{i+1} - \tilde{u}_i > 0 \text{ then } \tilde{u}_{i+1}^- - \tilde{u}_i^+ \geq 0; \\
\text{if } \tilde{u}_{i+1} - \tilde{u}_i < 0 \text{ then } \tilde{u}_{i+1}^- - \tilde{u}_i^+ \leq 0; \\
\text{if } \tilde{u}_{i+1} - \tilde{u}_i = 0 \text{ then } \tilde{u}_{i+1}^- - \tilde{u}_i^+ = 0.
\end{cases} \label{ENO_sign_property}
\end{equation}

Moreover, there is a constant $C_p$, depending only on $p$ and on the mesh-ratio of neighboring grid cells, $\max_{|j-i| \leq p} \left(\left|J_{j+1}\right|/\left|J_i\right|\right)$, such that

\begin{equation}
0 \leq \frac{\tilde{u}_{i+1}^- - \tilde{u}_i^+}{\tilde{u}_{i+1} - \tilde{u}_i} \leq C_p. \label{ENO_bound}
\end{equation}

The sign property tells us that at each cell interface, the jump of the reconstructed ENO pointvalues has the same sign as the jump in the underlying cell averages. The sign property is illustrated in Figure 5.1, which shows a third-, fourth- and fifth-order ENO reconstruction of randomly chosen cell averages. Even though the reconstructed polynomial may have large variations within each cell, its jumps at cell interfaces always have the same sign as the jumps of the cell averages. Moreover, the relative size of these jumps is uniformly bounded. We
remark that the inequality on the left-hand side of (5.9) is a direct consequence of the sign property (5.8).

It should be emphasized that the main Theorem 5.3 is valid for any order of ENO reconstruction and for any mesh size. It is valid for non-uniform meshes and does in no way depend on the underlying function $u$.

The proof of both the sign property and the related upper-bound (5.9) depends on the judicious choice of ENO stencils in Algorithm 5.1, and it may fail for other choices of ENO-based algorithms. In particular, the popular WENO methods, which are based on upwind or central weighted ENO stencils [LOC94, JS96] fail to satisfy the sign property, as can easily be confirmed numerically.

The aim of this section is to prove the sign property (5.8). To this end we derive a novel expression of the interface jump, $\tilde{u}_{i}^{+} - \tilde{u}_{i}^{-}$, as a sum of terms which involve $(p + 1)$-th divided differences of $V$, and we show that each summand in this expression has the same sign as $\bar{u}_{i+1} - \bar{u}_{i}$. For the remainder of this chapter we will assume for notational convenience that $i = 0$.

We denote the $j$-th divided difference of the primitive $U$ as

$$D_{[r, r+j]} := U[x_{r}, \ldots, x_{r+j}], \quad D_{[r, r+j]} = \frac{D_{[r+1, r+j]} - D_{[r, r+j-1]}}{x_{r+j} - x_{r}}, \quad r = \ldots, -3/2, -1/2.$$  

Thus, for example, by (5.5) we have $D_{[-1/2, 3/2]} = U[x_{-1/2}, x_{1/2}, x_{3/2}] = (\bar{u}_{1} - \bar{u}_{0})/(x_{3/2} - x_{-1/2})$. If $D_{[-1/2, 3/2]} = 0$, or in other words, if $\bar{u}_{0} = \bar{u}_{1}$, then it is easy to see that the ENO procedure will end up with identical stencils for $I_{0}$ and $I_{1}$, which in turn yields $\tilde{u}_{0}^{+} = \tilde{u}_{1}^{-}$. We may therefore assume that $D_{[-1/2, 3/2]} \neq 0$, and the sign property will be proved by showing that

$$\text{(5.10) Sign property:} \quad \begin{cases} \text{if} & D_{[-1/2, 3/2]} > 0 \quad \text{then} \quad \tilde{u}_{1}^{-} - \tilde{u}_{0}^{+} \geq 0; \\ \text{if} & D_{[-1/2, 3/2]} < 0 \quad \text{then} \quad \tilde{u}_{1}^{-} - \tilde{u}_{0}^{+} \leq 0. \end{cases}$$
From (5.7), we find that the value of $\tilde{u}_0^+$ is

$$(5.11) \quad \tilde{u}_0^+ = \tilde{u}_0(x_{1/2}) = \sum_{k=1}^{p} D_{r_0^k, r_0^k + k} \prod_{l=0}^{k-1} \prod_{m=0}^{k-1+l} (x_{1/2} - x_{r_0^{k-1} + m}) = \sum_{k=1}^{p} D_{r_0^k, r_0^k + k} \prod_{m=0}^{k-1} \prod_{m \neq l} (x_{1/2} - x_{r_0^{k-1} + m})$$

The last equality follows from the fact that all but the one term corresponding to $l = 1/2 - r_0^{k-1}$ drop out.

To simplify notation, we use $\prod_{j \in J, \alpha_j \neq \alpha_j}$ to denote a product which skips any of its zero factors. Thus, for example, a simple shift of indices in (5.11) yields

$$\tilde{u}_0^+ = \sum_{k=1}^{p} D_{r_0^k, r_0^k + k} \prod_{m=0}^{k-2} (x_{1/2} - x_{r_0^{k-1} + m}).$$

In an similar fashion, we handle the ENO reconstruction at cell $I_1$. Let $r_1^1, \ldots, r_1^p$ be the signature of that cell. Note that $r_0^k \leq r_1^k$, since the ENO reconstruction at stage $k$ in cell $I_1$ cannot select a stencil further to the left than the one used in cell $I_0$. If $r_0^k = r_1^k$, then the two interpolation stencils are the same, and so $\tilde{u}_1^- - \tilde{u}_0^+ = 0$. Hence, we only need to consider the case $r_0^k < r_1^k$ for all $k$. The reconstructed value of $\tilde{u}_1(x) = \tilde{U}_1'(x)$ at $x = x_{1/2}$ is given by

$$\tilde{u}_1^- = \tilde{u}_1(x_{1/2}) = \sum_{k=1}^{p} D_{r_1^k, r_1^k + k} \prod_{m=0}^{k-2} (x_{1/2} - x_{r_1^{k-1} + m})$$

The jump in the values reconstructed at $x = x_{1/2}$ is then given by

$$(5.12) \quad \tilde{u}_1^- - \tilde{u}_0^+ = \sum_{k=1}^{p} \left[ D_{r_1^k, r_1^k + k} \prod_{m=0}^{k-2} (x_{1/2} - x_{r_1^{k-1} + m}) - D_{r_0^k, r_0^k + k} \prod_{m=0}^{k-2} (x_{1/2} - x_{r_0^{k-1} + m}) \right].$$

The following lemma provides a much needed simplification for the rather intimidating expression (5.12), in terms of a key identity, which is interesting in its own right.

**Lemma 5.4.** The jump of the reconstructed point-values in (5.12) is given by

$$(5.13) \quad \tilde{u}_1^- - \tilde{u}_0^+ = \sum_{r=1}^{r_0^{p-1}} D_{r, r+p+1} (x_{r+p+1} - x_r) \prod_{m=0}^{p-1} (x_{1/2} - x_{r+m+1}).$$

**Proof.** We proceed in two steps. In the first step, we consider the special case when the two stencils that are used by the ENO reconstruction in cells $I_0$ and $I_1$ are only one grid cell apart, say, $r_0^p = r_1^p - 1 = r$ and in this case, Lemma 5.4 claims that $\tilde{u}_1^- - \tilde{u}_0^+$ equals

$$(5.14) \quad \tilde{u}_1(x_{1/2}) - \tilde{u}_0(x_{1/2}) = D_{r, r+p+1} (x_{r+p+1} - x_r) \prod_{m=0}^{p-1} (x_{1/2} - x_{r+m+1}).$$
Indeed, the interpolant of \( U(x_{r+1}), \ldots, U(x_{r+p}), U(x_r) \), assembled in the specified order from left to right and then adding \( U(x_r) \) at the end, is given by

\[
\tilde{U}_0(x) = \sum_{k=0}^{p-1} D_{[r+1,r+1+k]} \prod_{m=0}^{k-1} (x - x_{r+1+m}) + D_{[r,r+p]} \prod_{m=0}^{p-1} (x - x_{r+1+m}).
\]

Similarly, the interpolant of \( U(x_{r+1}), \ldots, U(x_{r+p+1}) \), assembled in the specified order from left to right, is given by

\[
\tilde{U}_1(x) = \sum_{k=0}^{p-1} D_{[r+1,r+k+1]} \prod_{m=0}^{k-1} (x - x_{r+1+m}) + D_{[r+1,r+p+1]} \prod_{m=0}^{p-1} (x - x_{r+1+m})
\]

(cf. (5.6)). Thus, their difference amounts to

\[
\tilde{U}_1(x) - \tilde{U}_0(x) = \left( D_{[r+1,r+p+1]} - D_{[r,r+p]} \right) \prod_{m=0}^{p-1} (x - x_{r+1+m}) = D_{[r,r+p+1]} \left( x_{r+p+1} - x_r \right) \prod_{m=0}^{p-1} (x - x_{r+1+m}),
\]

which reflects the fact that \( \tilde{U}_0 \) and \( \tilde{U}_1 \) coincide at the \( p \) points \( x_{r+1}, \ldots, x_{r+p} \). Differentiation yields

\[
\tilde{u}_1(x) - \tilde{u}_0(x) = D_{[r,r+p+1]} \left( x_{r+p+1} - x_r \right) \sum_{l=0}^{p-1} \prod_{m\neq l}^{p-1} (x - x_{r+1+m}).
\]

At \( x = x_{l/2} \), all product terms on the right vanish except for \( l = -r - l/2 \), since \( x_{l/2} \) belongs to both stencils. We end up with

\[
\tilde{u}_1(x_{l/2}) - \tilde{u}_0(x_{l/2}) = D_{[r,r+p+1]} \left( x_{r+p+1} - x_r \right) \prod_{m=0}^{p-1} (x_{l/2} - x_{r+m+1}).
\]

This shows that (5.14) holds, verifying Lemma 5.4 in the case that the stencils associated with \( I_0 \) and \( I_1 \) are separated by just one cell.

In step two, we extend this result for arbitrary stencils, where \( I_0 \) and \( I_1 \) are associated with arbitrary offsets, \( r_0^p \leq r_1^p \). Denote by \( \tilde{u}^{[r]} \) the interpolant at points \( x_r, \ldots, x_{r+p} \), so that \( \tilde{u}_0 = \tilde{u}^{[r_0]} \) and \( \tilde{u}_1 = \tilde{u}^{[r_1]} \). Using the representation from the first step for the difference of one-cell shifted stencils, \( (\tilde{u}^{[r+1]} - \tilde{u}^{[r]})(x_{l/2}) \), we can write the jump at the cell interface as a
telescoping sum,

\[(\tilde{u}_1 - \tilde{u}_0)(x_{l/2}) = \left(\tilde{u}^{[p]}_0 - \tilde{u}^{[p]}_1\right)(x_{l/2})\]

\[= \sum_{r=r_0^p}^{r_0^p-1} (\tilde{u}^{(r+1)} - \tilde{u}^{(r)})(x_{l/2})\]

\[= \sum_{r=r_0^p}^{r_0^p-1} D_{[r,r+p+1]}(x_{r+p+1} - x_r) \prod_{m=0}^{p-1} (x_{l/2} - x_{r+m+1}),\]

and (5.13) follows.

Next, we apply Lemma 5.4 to conclude the proof of the sign property. To this end, we show that each non-zero summand in (5.13) has the same sign as \(\tilde{u}_1 - \tilde{u}_0\). Since

\[(5.15) \quad \text{sgn} \left( (x_{r+p+1} - x_r) \prod_{m=0}^{p-1} (x_{l/2} - x_{r+m+1}) \right) = (-1)^{r+p-\frac{1}{2}}, \quad r = -p + 1/2, \ldots, -1/2,\]

then in view of (5.10), it remains to prove the following essential lemma.

**Lemma 5.5.** Let \(\{r_0^k\}_{k=1}^p\) and \(\{r_1^k\}_{k=1}^p\) be the signatures of the ENO stencils associated with cells \(I_0\) and, respectively, \(I_1\). Then the following holds:

\[(5.16a) \quad \text{if } D_{[-\frac{1}{2},\frac{1}{2}]} > 0 \text{ then } (-1)^{r+p-\frac{1}{2}}D_{[r,r+p+1]} \geq 0, \quad r = r_0^p, \ldots, r_1^p - 1,\]

\[(5.16b) \quad \text{if } D_{[-\frac{1}{2},\frac{1}{2}]} < 0 \text{ then } (-1)^{r+p-\frac{1}{2}}D_{[r,r+p+1]} \leq 0, \quad r = r_0^p, \ldots, r_1^p - 1.\]

Since \(r\) runs over half-integers, (5.16) imply that each non-zero term in the sum (5.13) has the same sign as \(D_{[-\frac{1}{2},\frac{1}{2}]}\), and Theorem 5.3 follows from the sign property, (5.10).

**Proof.** We consider the case (5.16a) where \(D_{[-\frac{1}{2},\frac{1}{2}]} > 0\); the case (5.16b) can be argued similarly. The result clearly holds for \(p = 1\), where \(r_0^p = r_1^p - 1 = -1/2\). The general case follows by induction on \(p\). Assuming that (5.16) holds for some \(p \geq 1\), namely, that \((-1)^{r+p-\frac{1}{2}}D_{[r,r+p+1]} \geq 0\) for \(r = r_0^p, \ldots, r_1^p - 1\), we will verify that it holds for \(p + 1\). Indeed,

\[(5.17) \quad (-1)^{r+p+\frac{1}{2}}D_{[r,r+p+2]} = (-1)^{r+p+\frac{1}{2}} \frac{D_{[r+1,r+p+2]} - D_{[r,r+p+1]}}{x_{r+p+2} - x_r} = \frac{|D_{[r+1,r+p+2]}| + |D_{[r,r+p+1]}|}{x_{r+p+2} - x_r} \geq 0\]

for \(r = r_0^p, \ldots, r_1^p - 2\), by the induction hypothesis. Thus, it remains to examine \(D_{[r,r+p+2]}\) when \(r = r_0^{p+1} < r_0^p\) and \(r = r_1^{p+1} \geq r_1^p\).
5.3. UPpER BOUND ON JUMPS

In this section, we will prove (5.9), which establishes an upper bound on the size of the jump in reconstructed values in terms of the jump in the underlying cell averages. We need the following lemma.

**Lemma 5.6.** Let $r_p^0, r_1^p$ be the (half-integer) offsets of the ENO stencils associated with cell $I_0$ and $I_1$, respectively. Then

\[
D_{[r, r^p + p + 1]} \frac{(-1)^{p+1}}{D_{[l, l^p + 1]}} \leq C_{r, p}, \quad r = r_0^p, \ldots, r_1^p - 1,
\]

where the constants $C_{r, p}$ are defined recursively, starting with $C_{r, 1} = 1$, and

\[
C_{r, k+1} = \frac{2}{x_{r+k+2} - x_r} \max(C_{r, k}, C_{r+1, k}) \quad \forall r.
\]

The quantity on the left in (5.18a) was shown to be bounded from below by zero in Lemma 5.5; here we prove an upper bound. The constants $C_{r, p}$ only depend on the grid sizes $|I|.

**Proof.** The result clearly holds for $p = 1$. We prove the general induction step passing from $p \mapsto p + 1$. Using the recursion relation

\[
D_{[r, r^p + p + 2]} = \frac{D_{[r, r^p + p + 1]} - D_{[r, r + p + 1]}}{\Delta x}, \quad \Delta x := x_{r+p+2} - x_r,
\]
we have
\begin{align}
0 \leq \frac{D_{[r,r+p+2]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p+1/2} = 1 \left( \frac{D_{[r+1,r+1+p+1]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p+1/2} + \frac{D_{[r,r+p+1]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p-1/2} \right) \\
\leq \frac{C_{r+1,p} + C_{r,p}}{\Delta x} \leq C_{r,p+1}, \quad r = r_{0}, \ldots, r_{1} - 2.
\end{align}

(5.19)

We turn to the remaining cases.

(a) As in Lemma 5.5, if \(r_{0} + 1 = r_{0} - 1\) then the induction step is already covered in (5.19), so the only remaining case is \(r := r_{0} + 1 = r_{0} - 1\), corresponding to Lemma 5.5(a). In this case, \(|D_{[r,r+p+2]}| \leq |D_{[r+1,r+p+2]}|\), so
\begin{align}
\frac{D_{[r,r+p+2]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p+1/2} = 1 \left( \frac{D_{[r+1,r+1+p+1]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p+1/2} + \frac{D_{[r,r+p+1]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p-1/2} \right) \\
\leq \frac{2C_{r+1,p}}{\Delta x} \leq C_{r,p+1}.
\end{align}

(b) If \(r_{1} + 1 = r_{1} - 1\) then the induction step is already covered in (5.19), so the only remaining case is \(r := r_{1} + 1 = r_{1} - 1\), corresponding to Lemma 5.5(b). In this case, \(|D_{[r+1,r+p+2]}| \leq |D_{[r,r+p+2]}|\), so
\begin{align}
\frac{D_{[r,r+p+2]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p+1/2} = 1 \left( \frac{D_{[r+1,r+1+p+1]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p+1/2} + \frac{D_{[r,r+p+1]}(x) - p_{+}}{D_{[-1/2,1/2]}(x)} (-1)^{r+p-1/2} \right) \\
\leq \frac{2C_{r,p}}{\Delta x} \leq C_{r,p+1}.
\end{align}

\(\square\)

Using the explicit form (5.13) of the jump \(\hat{u}_{1} - \hat{u}_{0}\), we get the following explicit expression of the upper-bound asserted in (5.9).

**Theorem 5.7.** Let \(\hat{u}_{1}^{+}\) and \(\hat{u}_{0}^{-}\) be the point-values reconstructed by the ENO Algorithm 5.1 at the cell interface \(x = x_{1/2}^{+}\) and \(x = x_{1/2}^{-}\), respectively. Then
\begin{equation}
\frac{\hat{u}_{1}^{+} - \hat{u}_{0}^{-}}{\hat{u}_{1} - \hat{u}_{0}} \leq C_{p} := \frac{1}{x_{1/2}^{+} - x_{1/2}^{-}} \sum_{r = -p + 1/2}^{-1/2} C_{r,p} \left( x_{r} - x_{r+1}^{-} \right) \left( x_{r} - x_{r+m+1}^{+} \right).
\end{equation}
The number 10 perturbation ensures that cells cannot return to Figure 5.1, we see that although the jumps at the cell interfaces can get large,

\( \left| \tilde{u}_{i+1/2} - \tilde{u}_{i-1/2} \right| \leqslant \frac{1}{\Delta x_{i+1/2}} \sum_{r=0}^{r_p} C_{r,p} \left| (x_{i+r+1} - x_r) \prod_{m=0}^{p-1} (x_{i/2} - x_{r+m+1/2}) \right| \)

When the mesh is uniform, \( |I_i| \equiv h \), the expression for the upper bound \( C \) can be calculated explicitly. The recursion relation (5.18b) yields \( C_{r,p} = \frac{2^p}{2^{p-1}h^{p-1}(p+1)!} \), and the coefficient of the \( (p + 1) \)-th order divided differences in (5.13) is

\[ \left| (x_{i+r+1} - x_r) \prod_{m=0}^{p-1} (x_{i/2} - x_{r+m+1/2}) \right| = h^p(p + 1)(-r - 1/2)!(p + r - 1/2)!. \]

Thus, we arrive at the following bound on the jump in reconstructed values.

**Corollary 5.8.** Assume that the mesh is uniform with mesh size \( |I_i| \equiv \text{const} \). Let \( \tilde{u}_{i+1/2} \) and \( \tilde{u}_{i-1/2} \) be the pointvalues reconstructed at the cell interface \( x = x_{i+1/2} \) and \( x = x_{i-1/2} \), respectively, by the \( p \)-th order ENO Algorithm. Then

\[ \frac{\tilde{u}_{i+1/2} - \tilde{u}_{i-1/2}}{\tilde{u}_{i+1/2} - \tilde{u}_{i-1/2}} \leqslant C_p = 2^{p-1} \frac{1}{p!} \sum_{k=0}^{p-1} k!(p - k - 1)!. \]

Table 5.2 shows the upper bound (5.20) on \( (\tilde{u}_{i+1/2} - \tilde{u}_{i-1/2})/(\tilde{u}_{i+1/2} - \tilde{u}_{i-1/2}) \) for some values of \( p \). Returning to Figure 5.1, we see that although the jumps at the cell interfaces can get large, they cannot exceed \( C_p \) times the size of the jump in cell averages, regardless of the values in neighbouring cell averages.

It can be shown that the bound \( C_p \) given in Theorem 5.7 is sharp. Indeed, the worst-case scenarios for orders of accuracy \( p = 2, 3, 4, 5 \) are shown in Figure 5.3. The mesh in this figure is \( x_{i+1/2} = i \), and the cell averages are chosen as

\[ \tilde{u}_i = \begin{cases} 0 & \text{if } i \text{ is odd} \\ 1 & \text{if } i \text{ is even and } i \leqslant 4 \\ 1 - 10^{-10} & \text{if } i \text{ is even and } i > 4. \end{cases} \]

The number \( 10^{-10} \) is chosen at random; any small perturbation will give the same effect. This perturbation ensures that cells \( I_i \) for \( i \leqslant 4 \) interpolate over a stencil to the left of the cell.
Table 5.2. Upper bounds on relative jumps for uniform meshes.

### Upper bounds $C_p$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$C_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>$10/3 = 3.333\ldots$</td>
</tr>
<tr>
<td>4</td>
<td>$16/3 = 5.333\ldots$</td>
</tr>
<tr>
<td>5</td>
<td>$128/15 = 8.533\ldots$</td>
</tr>
<tr>
<td>6</td>
<td>$208/15 = 13.866\ldots$</td>
</tr>
</tbody>
</table>

![Figure 5.3. Worst case cell interface jumps for $p = 2, 3, 4, 5$.](image)

The ENO interpolation method

The ENO Algorithm can be formulated as a nonlinear interpolation procedure. The starting point is a given collection of point-values $u_i = u(x_i)$ at given grid-points $\{x_i\}_{i \in \mathbb{Z}}$. The purpose of the ENO procedure in this context is to recover a highly accurate approximation
of \( u(x) \) from its point-values \( u_i = u(x_i) \),

\[
\text{ENO: } \{u_i\}_{i \in \mathbb{Z}} \mapsto \tilde{u}(x) := \sum_k \tilde{u}_k(x) I_{I_k}(x),
\]

where now \( I_i = [x_{i-1/2}, x_{i+1/2}) \) and \( x_{i+1/2} := \frac{x_i + x_{i+1}}{2} \). Here, \( \tilde{u}_k(x) \) are polynomials of degree \( p - 1 \) which interpolate the given data,

\[
\tilde{u}_i(x_i) = u_i.
\]

Moreover, the ENO interpolant \( \tilde{u}(x) \) is essentially non-oscillatory in the sense of recovering \( u(x) \) to order \( O(h^p) \). In particular, since \( u(x) \) may experience jump discontinuities, we wish to recover the point-values, \( \tilde{u}_i^+ := \tilde{u}_i(x_{i+1/2}) \) and \( \tilde{u}_i^- := \tilde{u}_i(x_{i+1/2}) \), with high-order accuracy,

\[
\left| \tilde{u}_i^+ - u(x_{i+1/2}^-) \right| + \left| \tilde{u}_i^- - u(x_{i+1/2}^+) \right| = O(h^p).
\]

The ENO interpolant \( \tilde{u}(x) \) is based on the usual divided differences \( \{u[x_i, \ldots, x_{i+j}]\}_{i \in \mathbb{Z}} \), defined by

\[
u[x_i] = u_i, \quad u[x_i, \ldots, x_{i+j}] := \frac{u[x_{i+1}, \ldots, x_{i+j}] - u[x_i, \ldots, x_{i+j-1}]}{x_{i+j} - x_i}.
\]

We denote by \( \{\ell_{i,k}^p\}_{k=1} \) the left index of the ENO stencil associated with grid point \( x_i \). In this case of ENO interpolation, these indices are non-positive integers, corresponding to the integer indices of the prescribed gridpoints \( x_j \). These offsets are selected according to the following ENO selection procedure.

**Algorithm 5.9 (ENO interpolation).**

Let point values \( u_{i-p+1}, \ldots, u_{i+p-1} \) be given. Then the left stencil indices \( \ell_{i,1}^1, \ldots, \ell_{i,p}^p \) are determined as follows.

1. \( \ell_{i,1}^1 \leftarrow i \)
2. for \( k = 1 \rightarrow p - 1 \) do
   1. if \( |u[x_{\ell_{i,k}^1}, \ldots, x_{\ell_{i,k+1}^k}]| < |u[x_{\ell_{i,k}^1}, \ldots, x_{\ell_{i,k+1}^k}]| \) then
      1. \( \ell_{i,k+1}^k \leftarrow \ell_{i,k}^k - 1 \)
   2. else
      1. \( \ell_{i,k+1}^k \leftarrow \ell_{i,k}^k \)
   end if
end for

To simplify the notation we let \( d_{i,j+1}^k \) abbreviate the divided differences \( u[x_i, \ldots, x_{i+j}] \). We set \( \tilde{u}_i(x) \) as the interpolant of \( v \) over the stencil \( \{u_k\}_{k=\ell_{i,k}^p}^{\ell_{i,p-1}^p} \):

\[
\tilde{u}_i(x) = \sum_{k=0}^{p-1} d_{1,i,k}^{\ell_{i,k}^k} \prod_{m=0}^{k-1} \left( x - x_{\ell_{i,m}^{k+1}} \right).
\]

We refer to the composite function \( \tilde{u}(x) = \sum_i \tilde{u}_i(x) I_{I_i}(x) \) as the \( p \)-th order ENO interpolation of \( \{u_i\}_{i \in \mathbb{Z}} \).
In the following theorem we state the main stability result for this version of the ENO interpolation procedure, analogous to the sign property of the ENO reconstruction procedure from cell averages.

**Theorem 5.10 (The sign property for ENO interpolation).** Fix an integer \( p > 1 \). Given the point-values \( \{u_i\} \), let \( \tilde{u}(x) \) be the \( p \)-th order ENO interpolant of these point-values, and define \( \tilde{u}^+_i := \tilde{u}(x_{i+1/2}) \) and \( \tilde{u}^-_i := \tilde{u}(x_{i+1/2}) \) denote the reconstructed point-values to the left and right of the cell interface \( x_{i+1/2} \). Then the following sign property holds at all interfaces:

\[
\begin{aligned}
\begin{cases}
\text{if } u_{i+1} - u_i > 0 & \text{then } \tilde{u}^-_{i+1} - \tilde{u}^+_i \geq 0; \\
\text{if } u_{i+1} - u_i < 0 & \text{then } \tilde{u}^-_{i+1} - \tilde{u}^+_i \leq 0; \\
\text{if } u_{i+1} - u_i = 0 & \text{then } \tilde{u}^-_{i+1} - \tilde{u}^+_i = 0.
\end{cases}
\end{aligned}
\]

(5.21)

Moreover, there is a constant \( c_p \), depending only on \( p \) and on the mesh-ratio \( \max_{|i-j| < p} (|x_{j+1} - x_j|/|x_j - x_{j-1}|) \) of neighboring grid cells, such that

\[
0 \leq \frac{\tilde{u}^-_{i+1} - \tilde{u}^+_i}{u_{i+1} - u_i} \leq c_p.
\]

(5.22)

The proof of Theorem 5.10 is very similar to the proof of Theorem 5.3. We therefore only sketch the arguments without details.

**5.4.1. The sign property for the ENO interpolant.** We focus on the jump across the interface at \( x = x_{1/2} \). Let \( \{\ell_j^k\}_{k=1}^p \) and \( \{\ell_{1j}^p\}_{k=1}^p \) be the ENO stencils associated with gridpoints \( x_0 \) and \( x_1 \), respectively. Recall that in this case of ENO interpolation, these offsets are integers, \(-p + 1 \leq \ell^k_0 \leq 0, -p + 2 \leq \ell^k_1 \leq 1\). Our first key step is to compute the jump at the interface point \( x_{1/2} \) in the case where the ENO procedure are separated by just one point, namely, \( \ell^p_0 = \ell^p_1 - 1 \). As before, we denote the reconstructed values at the interface \( x = x_{1/2} \) by \( \tilde{u}^+_0 = \tilde{u}_0(x_{1/2}) \) and \( \tilde{u}^-_1 = \tilde{u}_1(x_{1/2}) \).

**Lemma 5.11.** If \( \ell := \ell^p_0 = \ell^p_1 - 1 \) then

\[
\tilde{u}^-_1 - \tilde{u}^+_0 = \frac{\ell^p_1 - \ell^p_0}{\ell^p_0} \prod_{m=1}^{p-1} (x_{j+1/2} - x_{\ell+1/2}).
\]

By assembling a telescoping sum of several such stencils we obtain

**Corollary 5.12.** For general \( \ell^p_0 \leq \ell^p_1 \), we have

\[
\tilde{u}^-_1 - \tilde{u}^+_0 = \sum_{\ell = \ell^p_0}^{\ell^p_1 - 1} \prod_{m=1}^{p-1} (x_{j+1/2} - x_{\ell+1/2}).
\]

(5.23)

Since \( u_1 - u_0 = (x_1 - x_0) d_{[0,1]} \), we wish to show that the jump in reconstructed values at the cell interface has the same sign as \( d_{[0,1]} \). To this end we show that each summand in
(5.23) has the same sign as $d_{\{0,1\}}$. Indeed, since

$$
\text{sgn} \left( (x_{\ell+p} - x_\ell) \prod_{m=1}^{p-1} (x_{\ell/2} - x_{\ell+m}) \right) = (-1)^{\ell+p+1}, \quad -p + 1 \leq \ell \leq 0,
$$

it suffices to prove the following:

**Lemma 5.13.** If $\ell_0^p, \ell_1^p$ are selected according to the ENO stencil selection procedure, then

$$
\begin{align*}
\text{if } d_{\{0,1\}} & \geq 0 & (\ell, \ell+1) \geq 0; \\
\text{if } d_{\{0,1\}} & < 0 & (\ell, \ell+1) \leq 0,
\end{align*}
$$

$$\ell = \ell_0^p, \ldots, \ell_1^p - 1.$$

**Remark.** The sign property for ENO interpolation does not depend on the specific value of the cell interface point $x_{\ell/2}$. Indeed, this value only appears in (5.24), which holds for any value $x_{\ell/2} \in (x_0, x_1)$. Hence, we can pose the following, stronger version of (5.21):

$$
\forall x_{\ell+1/2} \in [x_\ell, x_{\ell+1}] \begin{cases}
\text{if } u_{i+1} - u_i > 0 & \bar{u}_{i+1}(x_{\ell+1/2}) - \bar{u}_i(x_{\ell+1/2}) \geq 0; \\
\text{if } u_{i+1} - u_i < 0 & \bar{u}_{i+1}(x_{\ell+1/2}) - \bar{u}_i(x_{\ell+1/2}) \leq 0; \\
\text{if } u_{i+1} - u_i = 0 & \bar{u}_{i+1}(x_{\ell+1/2}) - \bar{u}_i(x_{\ell+1/2}) = 0.
\end{cases}
$$

Note that this does not easily translate to ENO reconstruction, because the cell interface $x_{\ell+1/2}$ appears explicitly in the definition of divided differences, and hence there is a nonlinear dependence on $x_{\ell+1/2}$.

**5.4.2. Upper bounds on the relative jumps for ENO interpolant.** Next, we show the corresponding upper bound on $\bar{u}_1^* - \bar{u}_0^*$ for ENO reconstruction with point values.

**Lemma 5.14.** If $\ell_0^p, \ell_1^p$ are selected according to the ENO stencil selection procedure, then

$$
0 \leq \frac{d_{\{\ell, \ell+1\}}}{d_{\{0,1\}}} (-1)^{\ell+p+1} \leq c_{\ell,p} \quad \ell = \ell_0^p, \ldots, \ell_1^p - 1,
$$

where $c_{\ell,p}$ are defined recursively, starting with $c_{\ell,1} = 1$, and

$$
c_{\ell,k+1} = \frac{2}{x_{\ell+k} - x_\ell} \max(c_{\ell,k}, c_{\ell+1,k}).
$$

When the mesh is uniform we can compute $c_{\ell,p}$ explicitly. If $x_{i+1} - x_i \equiv h$ then it is straightforward to show that $c_{\ell,p} \equiv (2/h)^{p-1} 1/p!$. Moreover, the coefficient of the $(p+1)$-th order divided differences in (5.23) is

$$
\left( x_{\ell+p} - x_\ell \right) \prod_{m=1}^{p-1} (x_{\ell/2} - x_{\ell+m}) = h^p \prod_{m=1}^{p-1} \left( (1/2 - \ell - m) \right).
$$

Thus, we arrive at the following bound on the jump in reconstructed values.
Table 5.4. Upper bounds on relative jumps for uniform meshes.

<table>
<thead>
<tr>
<th>$p$</th>
<th>Upper bound $c_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>10.375</td>
</tr>
<tr>
<td>6</td>
<td>18.25</td>
</tr>
</tbody>
</table>

**Theorem 5.15.** Assume that the mesh is uniform with mesh size $x_{i+1} - x_i \equiv h$, and let $\tilde{u}_{1/2}^\pm$ be the $p$-th order ENO interpolation to the left and right of $x = x_{1/2}$. Then

$$\frac{\tilde{u}_i^- - \tilde{u}_i^+}{u_1 - u_0} \leq c_p := 2^{p-1} \frac{1}{(p-1)!} \sum_{\ell=-p+1}^{0} \left| \prod_{m=1}^{p-1} \left( \frac{1}{2} - \ell - m \right) \right| .$$

Table 5.4 shows the upper bound on $(\tilde{u}_i^- - \tilde{u}_i^+)/ (u_1 - u_0)$ for $p \leq 6$. As for the ENO reconstruction procedure in Section 5.3, it may be shown that these bounds are sharp.

#### 5.5. A conjecture on the total variation of ENO reconstruction

We have established optimal local lower and upper bounds on the jumps of reconstructed values in terms of the jumps in cell averages, $|\tilde{u}_i^- - \tilde{u}_i^+| \leq C|\bar{u}_{i+1} - \bar{u}_i|$ for a $C > 0$. It is obvious that the converse cannot hold, since the jump in reconstructed values may be zero when the jump in cell averages is not. However, a global estimate of this form might hold, as stated in the following conjecture.

**Conjecture 5.16.** Fix an integer $p > 1$ and let $\tilde{u}_i(x)$ be the $p$-th order ENO reconstruction of $\{\bar{u}_j\}_{j \in \mathbb{Z}}$ in cell $I_i$. Assume that $\{\bar{u}_i\}_{i \in \mathbb{Z}}$ has compact support. Then there is a $C > 0$, depending only on $p$ and on the mesh $\{x_{i+1/2}\}_{i \in \mathbb{Z}}$, such that

$$\sum_{i \in \mathbb{Z}} |\tilde{u}_{i+1} - \tilde{u}_i|^{p+1} \leq C\|\bar{u}\|_p^{p-1} \sum_{i \in \mathbb{Z}} (\tilde{u}_{i+1} - \tilde{u}_i)(\tilde{u}_{i+1} - \bar{u}_i).$$

By the sign property, each summand on the right-hand side of (5.26) is nonnegative. The conjecture has been verified numerically for a large array of stencil data, and for uniform meshes the optimal constant seems to be $C = 2^{p-1}$. However, the global nature of the estimate makes a proof seem very difficult, and local estimates such as those in the previous sections are not sufficient. For instance, if

$$\tilde{u}_i = \begin{cases} 
0 & i \leq 0 \\
\frac{i}{n} & 1 \leq i < n \\
1 & n \leq i,
\end{cases}$$

...
for an $n \in \mathbb{N}$, then $\bar{u}_{i+1}^+ - \bar{u}_i^- = 0$ in all but two cells, but $\bar{u}_{i+1} - \bar{u}_i = 1/n$ for $i = 0, \ldots, n-1$. Still,

$$\sum_i |\bar{u}_{i+1} - \bar{u}_i|^{p+1} = \frac{1}{n^p} \leq \frac{1}{n^2} = \|\bar{u}\|^{p-1}_{\ell^\infty} \sum_i (\bar{u}_{i+1}^- - \bar{u}_i^+)(\bar{u}_{i+1} - \bar{u}_i).$$

The cell averages and reconstructed polynomials for $p = 2$, $n = 10$ are shown in Figure 5.5(a). This example extends to the case where $\bar{u}_i$ for $i = 1, \ldots, n$ is the cell average of a polynomial of order less than $p$: the left-hand side is $O(1/n^p)$, while the right-hand side has a fixed (independent of $n$) number of nonzero summands, each of which is $O(1/n^2)$.

The above examples indicate that the worst-case scenario for the estimate (5.26) is not when $u$ is smooth. Indeed, from numerical experiments it seems that the worst-case scenario (that is, when equality is attained in (5.26)) is approached as $n \to \infty$ with the oscillatory data

$$\bar{u}_i = \begin{cases} 
0 & i \leq 0 \text{ or } i \geq n \\
1 & i = 0 \text{ or } i = 1 \text{ (mod 4)} \\
-1 & i = 2 \text{ or } i = 3 \text{ (mod 4)}.
\end{cases}$$

The reconstructed polynomials when $p = 2$, $n = 20$ is plotted in Figure 5.5(b). In this particular example, the ratio between the left- and the right-hand sides of (5.26) is about 1.82, and as $n$ increases, the ratio approaches $2^{p-1} = 2$. No other “worst cases” have been found. This indicates an upper bound of $C = 2^{p-1}$.

Figure 5.5. ENO reconstructed values and cell averages.
The TECNO scheme

In this chapter we conclude the construction of the high-order accurate, entropy stable finite difference schemes that were introduced in Chapter 4. The two essential components of that construction was (i) high-order entropy conservative fluxes, as discussed in Section 4.1, and (ii) a reconstruction procedure satisfying the sign property (4.9). As shown in the previous chapter, the ENO procedure satisfies precisely this property.

The results presented in this chapter is joint work with Eitan Tadmor and have been published in [FMT12a].

We proceed with the precise definition of our scheme and state its stability properties, first for scalar equations and then for systems of equations.

6.1. The TECNO scheme for scalar conservation laws

Consider the scalar, one-dimensional conservation law (1.1), endowed with a convex entropy pair \((\eta, q)\). As before, we denote the entropy variable by \(3 = \eta'(u)\).

**Definition 6.1.** For a \(p \in \mathbb{N}\), let \(k \in \mathbb{N}\) be such that \(2k \geq p\). Let \(\tilde{F}\) be a two-point entropy conservative flux, and define \(\tilde{F}^{2k}\) by (4.1). Let \(D_{i+1/2} \geq 0\) and let \(v_i^\pm := v_i(x_i \pm 1/2)\) be the cell interface values of the \(p\)-th order ENO interpolation \(v_i(x)\) of the point values \(\{v_j\}_{j \in \mathbb{Z}}\). Then the TECNO\(_p\) scheme is the finite difference scheme (2.1) with flux (4.8).

**Theorem 6.2.** The TECNO\(_p\) scheme for scalar conservation laws is (formally) \(p\)-th order accurate and entropy stable, with the numerical entropy flux

\[
Q_{i+1/2}^p = \tilde{Q}_{i+1/2}^{2k} - \tilde{v}_{i+1/2} D_{i+1/2} \frac{\langle v \rangle_{i+1/2}}{\Delta x}.
\]

More precisely, it satisfies the entropy dissipation estimate

\[
\frac{d}{dt} \eta(u_i) + \frac{Q_{i+1/2}^p - Q_{i+1/2}^p}{\Delta x} = -\frac{D_{i+1/2} \|v\|_{i+1/2} \langle v \rangle_{i+1/2} + D_{i-1/2} \|v\|_{i-1/2} \langle v \rangle_{i-1/2}}{2\Delta x} \leq 0.
\]

If \(\tilde{F}\) and \(\tilde{Q}\) are Lipschitz continuous (i.e., satisfy (A4\(_F\)), (A4\(_Q\)) with \(p = 1\)) and \(D_{i+1/2} \leq c\) for a \(c > 0\), then \(F^p\) and \(\tilde{Q}^p\) are also Lipschitz continuous (i.e., satisfy (A4\(_F\)), (A4\(_Q\))).

**Proof.** Accuracy follows from the fact that \(\tilde{F}_{i+1/2}^{2k}\) is accurate to order \(2k \geq p\), and that \(\langle v \rangle_{i+1/2} = O(\Delta x^p)\) in smooth regions. Entropy stability follows analogously to Theorem 2.4; the term \(\|v\|_{i+1/2} \langle v \rangle_{i+1/2}\) is nonnegative due to the ENO sign property, Theorem 5.10.
If $\tilde{F}$ is Lipschitz continuous then clearly $\tilde{F}^{2k}$ is. The Lipschitz continuity of $F^p$ follows from the upper bound on $D$, the continuity of $v$ with respect to $u$ and the upper bound (5.22) of the ENO interpolant, $\langle v \rangle_{i+1/2} \leq c_p \|v\|_{i+1/2}$. The proof for $Q^p$ is identical. □

**Remark.** The generalization of the TECNO scheme to multi-dimensional conservation laws ($d > 1$) on Cartesian meshes is straightforward: perform the one-dimensional reconstruction in each direction and then sum up, as described in Section 3.2. Note, however, that the convergence Theorem 6.3 only holds for the one-dimensional scheme, as it relies on compensated compactness. It remains to be investigated whether the proof of convergence can be generalized to Panov’s convergence framework for multi-dimensional scalar conservation laws [Pan10].

From (6.1) we can deduce that

$$\int_0^T \sum_i \|v\|_{i+1/2} \langle v \rangle_{i+1/2} \leq C,$$

provided $D_{i+1/2} \geq \frac{1}{c} > 0$ for a $c > 0$. To be able to deduce a bound on the spatial variation of the form (A3), we would need a bound of the form

$$\sum_i \|v\|_{i+1/2}^p \leq C \sum_i \|v\|_{i+1/2} \langle v \rangle_{i+1/2}$$

for some $P \geq 2$. This is precisely Conjecture 5.16. Assuming for the moment that the conjecture holds, we can prove convergence in the following sense.

**Theorem 6.3.** Let $f$ be strictly convex and let $(\eta, q)$ be uniformly convex. Let $u_0 \in L^\infty(\mathbb{R})$ have compact support. Assume that $\frac{1}{c} \leq D_{i+1/2} \leq c$ for all $i \in \mathbb{Z}$ and $t \in [0, T]$, for some $c > 0$, and that there is an $M > 0$ such that the solution computed by the scheme satisfies $\|u^{\Delta t}\|_{L^\infty(\mathbb{R} \times [0, T])} \leq M$ for all $\Delta x > 0$. Then, assuming that Conjecture 5.16 is true, the computed solution $\{u^{\Delta t}\}_{\Delta x > 0}$ converges pointwise a.e. and in $L^p(\mathbb{R} \times [0, T])$ for all $1 \leq p < \infty$ to the entropy solution of (1.1).

**Proof.** As in Theorem 3.6, we show that conditions (A1)–(A4Q) of Theorem 3.3 (i) are satisfied. Conditions (A1), (A2) hold by hypothesis. From (6.1) it follows that

$$\sum_i \eta(u_i(T)) \Delta x - \sum_i \eta(u_i(0)) \Delta x = -\int_0^T \sum_i D_{i+1/2} \|v\|_{i+1/2} \langle v \rangle_{i+1/2} dt,$$

so $0 \leq \int_0^T \sum_i D_{i+1/2} \|v\|_{i+1/2} \langle v \rangle_{i+1/2} dt \leq C$ for a $C > 0$. Hence,

\[
C \geq \int_0^T \sum_i D_{i+1/2} \|v\|_{i+1/2} \langle v \rangle_{i+1/2} dt \geq \frac{1}{c} \int_0^T \|v\|_{i+1/2} \langle v \rangle_{i+1/2} dt \geq \frac{1}{cM^{p-1}} \int_0^T \sum_i \|v\|_{i+1/2}^{p+1} dt,
\]

6. THE TECNO SCHEME
by the lower bound on $D$, the $L^\infty$-bound on $u$ and Conjecture 5.16. Uniform convexity of $\eta$ implies that $\|v_i\|_{i+1/2} \geq \eta_0 \|u_i\|_{i+1/2}$ for some $\eta_0 > 0$, so (A3) holds with $P = p + 1$. The Lipschitz conditions (A4$_F$), (A4$_Q$) follow from Theorem 6.2. 

6.2. The TECNO scheme for systems of equations

Using the reconstruction procedure described in Section 4.4, we can construct a TECNO scheme for systems of equations. Thus, consider the system of equations (1.1) endowed with an entropy pair $(\eta, q)$.

Definition 6.4. For a $p \in \mathbb{N}$, let $k \in \mathbb{N}$ be such that $2k \geq p$. Let $\tilde{F}$ be a two-point entropy conservative flux; let $D_{i+1/2}$ be of the form (4.11) and let the reconstructed entropy variables $v_i^+$ defined as in Section 4.4. Then the TECNO$_p$ scheme is the finite difference scheme (2.1) with flux (4.12).

Theorem 6.5. The TECNO$_p$ scheme for systems of conservation laws is (formally) $p$-th order accurate and entropy stable, with the numerical entropy flux

$$Q^p_{i+1/2} = \tilde{Q}_{i+1/2}^k - \tilde{v}_{i+1/2} \cdot D_{i+1/2} \langle v \rangle_{i+1/2}.$$

More precisely, it satisfies the entropy dissipation estimate

$$\frac{d}{dt} \eta(u_i) + \frac{Q^p_{i+1/2} - Q^p_{i-1/2}}{\Delta x} = - \frac{\|v\|_{i+1/2} \cdot D_{i+1/2} \langle v \rangle_{i+1/2} + \|v\|_{i-1/2} \cdot D_{i-1/2} \langle v \rangle_{i-1/2}}{2\Delta x} \leq 0.$$

If $\tilde{F}$ and $\tilde{Q}$ are Lipschitz continuous (i.e., satisfy (A4$_F$), (A4$_Q$) with $p = 1$) and $D_{i+1/2} \leq c$ for a $c > 0$, then $F^p$ and $\tilde{Q}^p$ are also Lipschitz continuous (i.e., satisfy (A4$_F$), (A4$_Q$)).

Proof. Entropy stability follows from Lemma 4.5. The rest is completely analogous to the scalar case. 

We end this section with a proof of the fact that the TECNO scheme converges weakly-* to a measure-valued solution. This result can be proven for the $d$-dimensional scheme, but for the sake of simplicity we prove only the one-dimensional version.

Theorem 6.6. Let $(\eta, q)$ be nonnegative and uniformly convex. Assume that $\frac{1}{c} \leq A_{i+1/2} \leq c$ for all $i$ for some $c > 0$, and that the solution $u^{\Delta x}$ computed with the TECNO$_p$ scheme is uniformly $L^\infty$-bounded, i.e. $\|u^{\Delta x}\|_{L^\infty(\mathbb{R} \times [0,T])} \leq M$ for all $\Delta x > 0$. Then, assuming Conjecture 5.16 is true, the Young measure associated with the sequence $u^{\Delta x}$ is an entropy measure-valued solution.

Proof. We verify conditions (B1), (B3), (B4$_F$), (B4$_Q$) of Lemmas 3.4 and 3.5. The $L^\infty$-bound (B1) follows by hypothesis, and the Lipschitz conditions (B4$_F$) and (B4$_Q$) follow from Theorem 6.5. We verify (B3) only for the one-dimensional ($d = 1$) case for the sake of clarity. From (6.2) it follows that

$$\sum_i \eta(u_i(T)) \Delta x - \sum_i \eta(u_i(0)) \Delta x = - \int_0^T \sum_i \|v\|_{i+1/2} \cdot D_{i+1/2} \langle v \rangle_{i+1/2} dt,$$
so \( 0 \leq \int_0^T \sum_i |v_i|_{i+1/2} \cdot D_{i+1/2} \langle v \rangle_{i+1/2} \ dt \leq C \) for a \( C > 0 \). By construction of the reconstructed values, there is for each \( i \in \mathbb{Z} \) and \( t \in [0, T] \) a diagonal matrix \( B_{i+1/2} \geq 0 \) such that (4.13) is satisfied. Hence, suppressing indices for the moment,

\[
[v]^T D \langle v \rangle = [v]^T R A B R^T [v] = (R^T [v])^T A B (R^T [v])
\]

\[
= \langle w \rangle^T A B \langle w \rangle \geq \frac{1}{C} \langle w \rangle^T B \langle w \rangle
\]

where \( \langle w \rangle_{i+1/2} = w_i^- - w_i^+ \) and \( w_i^\pm := R_{i+1/2} u_i \) are the scaled entropy variables defined in (4.15). The diagonal entries of \( B_{i+1/2} \) are precisely the ratio (5.22) coming from the ENO reconstruction of the components of \( w \). Hence, by Conjecture 5.16, we have (again suppressing indices)

\[
\sum_i \langle w \rangle^T B \langle w \rangle = \sum_i N B^n \langle w^n \rangle^2 \geq C \sum_i N \frac{|\langle w^n \rangle|^{p+1}}{|w^n|_{L^\infty}^{p-1}} \geq \frac{C}{|w|_{L^\infty}^{p-1}} \sum_i |\langle w \rangle|^{p+1},
\]

the last equality following from the equivalence of norms on \( \mathbb{R}^N \). Since \( \eta \) is uniformly convex and bounded on the compact set \( \{u : |u| \leq M\} \), there is an \( \eta_0 > 0 \) such that \( \frac{1}{\eta_0} \leq \eta''(u) \leq \eta_0 \) for all \( u \). The \( L^\infty \)-bound on \( u \) implies that \( |w|_{L^\infty} \leq \sqrt{\eta_0} M \). By the definition of \( w \), we have

\[
|\langle w \rangle|_{i+1/2}^2 = |v|_{i+1/2}^T R_{i+1/2} R_{i+1/2}^T \langle v \rangle_{i+1/2} = |v|_{i+1/2}^T u_0(v_{i+1/2}) |v|_{i+1/2} \geq \frac{1}{\eta_0} |\langle v \rangle|_{i+1/2}^2
\]

since \( u_0(v) = (\eta''(u(v)))^{-1} \geq \frac{1}{\eta_0} \). By the mean value theorem for vector-valued functions, we have \( |\langle v \rangle|_{i+1/2} \geq \frac{1}{\eta_0} |\langle u \rangle|_{i+1/2} |. \) In conclusion,

\[
\sum_i |\langle u \rangle|_{i+1/2}^{p+1} \leq C \sum_i |\langle v \rangle|_{i+1/2}^T D_{i+1/2} \langle v \rangle_{i+1/2}
\]

for a \( C \) depending only on \( p, \eta_0, c, \) and \( M \), so (B3) holds with \( P = p + 1 \). \( \square \)

6.3. Numerical experiments: scalar conservation laws

In this section and the next we compare the TECNO method with standard ENO methods in a series of numerical experiments. We demonstrate that our methods perform as well or better through qualitative comparisons of computed solutions, and quantitative comparisons of convergence rates.

We consider the TECNO method with Lax-Friedrichs-type ELF \( p \) diffusion operator, (2.22), and compare it with the ENO-LF finite difference scheme of [SO88]. We choose the square entropy \( \eta(u) = u^2 \).

6.3.1. Linear advection. Consider the one-dimensional linear advection equation (1.2). Since all discontinuities for this equation are contact discontinuities, entropy should be preserved over time, i.e.

\[
\int_{\mathbb{R}} \eta(u(x, t)) \ dx = \int_{\mathbb{R}} \eta(u_0(x)) \ dx.
\]
6.3. NUMERICAL EXPERIMENTS: SCALAR CONSrrvation LAwS

![Image of graphs showing numerical experiments]

Figure 6.1. Linear advection equation with sinusoidal initial data. Right: closeup around $x = -0.5$.

Figure 6.2. Relative entropy over time

We use the same initial data as in Section 4.2.1. Figure 6.1 shows the solution computed with the TECNO$_3$ scheme on a mesh of 50 grid points. The solution is clearly computed to a high degree of accuracy. Other schemes perform qualitatively very similar and so are not displayed here.

As previously noted, total entropy should be preserved over time. However, the error in the time stepping method will introduce numerical diffusion in the schemes. In Figure 6.2 we display the relative change in total entropy,

$$ E(t) - E(0) $$

$$ E(t) := \int_{-1}^{1} \eta(u(x,t)) \, dx. $$

The relative change is monotonously decreasing, meaning that there is no entropy production. Moreover, the entropy decay decreases with higher order of accuracy and smaller grid sizes.
Finally, we compute with both the TECNO\textsubscript{p} and ENO-LF\textsubscript{p} schemes on a series of meshes. The error and approximate orders of convergence are displayed in Table 6.3. Apart from a slight deterioration in the \(L^\infty\)-order of convergence for the fourth-order schemes, the schemes display the expected order of convergence.

<table>
<thead>
<tr>
<th>(n)</th>
<th>(L^1) error</th>
<th>(L^\infty) error</th>
<th>(L^1) error</th>
<th>(L^\infty) error</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>7.32e-4</td>
<td>–</td>
<td>8.97e-4</td>
<td>–</td>
</tr>
<tr>
<td>100</td>
<td>9.18e-5</td>
<td>3.00</td>
<td>9.32e-5</td>
<td>3.00</td>
</tr>
<tr>
<td>200</td>
<td>1.15e-5</td>
<td>3.01</td>
<td>1.15e-5</td>
<td>3.02</td>
</tr>
<tr>
<td>400</td>
<td>1.44e-6</td>
<td>3.01</td>
<td>1.42e-6</td>
<td>3.01</td>
</tr>
<tr>
<td>600</td>
<td>4.26e-7</td>
<td>3.00</td>
<td>4.20e-7</td>
<td>3.01</td>
</tr>
<tr>
<td>800</td>
<td>1.80e-7</td>
<td>3.01</td>
<td>1.77e-7</td>
<td>3.01</td>
</tr>
<tr>
<td>1000</td>
<td>9.21e-8</td>
<td>3.02</td>
<td>9.04e-8</td>
<td>3.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(n)</th>
<th>(L^1) error</th>
<th>(L^\infty) error</th>
<th>(L^1) error</th>
<th>(L^\infty) error</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>2.42e-4</td>
<td>–</td>
<td>2.52e-4</td>
<td>–</td>
</tr>
<tr>
<td>100</td>
<td>1.55e-5</td>
<td>3.55</td>
<td>1.56e-5</td>
<td>4.01</td>
</tr>
<tr>
<td>200</td>
<td>6.11e-7</td>
<td>3.32</td>
<td>6.17e-8</td>
<td>3.97</td>
</tr>
<tr>
<td>400</td>
<td>1.22e-8</td>
<td>3.26</td>
<td>1.25e-8</td>
<td>3.93</td>
</tr>
<tr>
<td>600</td>
<td>3.90e-9</td>
<td>3.24</td>
<td>4.04e-9</td>
<td>3.94</td>
</tr>
<tr>
<td>800</td>
<td>1.61e-9</td>
<td>3.22</td>
<td>1.68e-9</td>
<td>3.92</td>
</tr>
<tr>
<td>1000</td>
<td>7.58e-11</td>
<td>–</td>
<td>5.96e-11</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(n)</th>
<th>(L^1) error</th>
<th>(L^\infty) error</th>
<th>(L^1) error</th>
<th>(L^\infty) error</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>2.39e-4</td>
<td>–</td>
<td>2.40e-4</td>
<td>–</td>
</tr>
<tr>
<td>100</td>
<td>7.56e-6</td>
<td>4.98</td>
<td>7.58e-6</td>
<td>4.98</td>
</tr>
<tr>
<td>200</td>
<td>2.37e-7</td>
<td>5.00</td>
<td>2.37e-7</td>
<td>5.00</td>
</tr>
<tr>
<td>400</td>
<td>7.40e-9</td>
<td>5.00</td>
<td>7.42e-9</td>
<td>5.00</td>
</tr>
<tr>
<td>600</td>
<td>9.75e-10</td>
<td>5.00</td>
<td>9.77e-10</td>
<td>5.00</td>
</tr>
<tr>
<td>800</td>
<td>2.31e-10</td>
<td>5.00</td>
<td>2.32e-10</td>
<td>5.00</td>
</tr>
<tr>
<td>1000</td>
<td>7.58e-11</td>
<td>5.00</td>
<td>7.60e-11</td>
<td>5.00</td>
</tr>
</tbody>
</table>

Table 6.3. Experimental order of accuracy with TECNO and ENO-LF. \(n\) denotes the number of gridpoints.

6.3.2. Burgers’ equation. We repeat the numerical experiment from Section 4.2.2. The solution computed with the TECNO\textsubscript{3} scheme on meshes with 50, 100 and 200 grid points is
shown in Figure 6.4. The scheme approximates the solution well, apart from small oscillations in the wake of the shock. Contrast this to the ELW scheme in Figure 4.3, which has large oscillations after the shock.

### 6.4. Systems of conservation laws

#### 6.4.1. The wave equation

Consider the wave equation (1.4) with the entropy conservative flux (2.13). We consider the TECNO scheme with the ELF diffusion operator (2.22), which for \( p = 1 \) gives the numerical flux

\[
F_{i+1/2} = \frac{1}{2} (Au_i + Au_{i+1}) - \frac{1}{2} c (u_{i+1} - u_i).
\]

The TECNO\(_p\) scheme is compared with the ENO-LF\(_p\) scheme using flux splitting and reconstruction along characteristic directions. We use initial data

\[
\begin{align*}
    u^1(x, 0) &= \begin{cases} 
    x - 1 & \text{if } x < 0 \\
    1 - x & \text{if } x > 0
    \end{cases} \\
    u^2(x, 0) &= 0
\end{align*}
\]

on the domain \( x \in [-1, 1] \) with periodic boundary conditions. The solution features an initial jump discontinuity at \( x = 0 \) which breaks into two linear (contact) discontinuities. We use a mesh of 50 points and a CFL number of 0.4. Computed solutions at time \( t = 1.5 \) are displayed in Figure 6.5. The two methods resolve the flow with a comparable level of accuracy.

#### 6.4.2. Shallow water equations

Consider the one-dimensional shallow water equations (1.5). For the TECNO schemes we use the two-point entropy conservative flux (2.15) together with the ELLF diffusion operator (2.23). We compare the TECNO scheme with the ENO-LLF scheme using flux splitting in characteristic variables [Shu99]. The gravitational constant is set to \( g = 1 \).

**A dambreak problem.** We consider a dambreak problem for the shallow water equations with initial data

\[
\begin{align*}
    h(x, 0) &= \begin{cases} 
    1.5 & \text{if } |x| < 0.2 \\
    1 & \text{if } |x| > 0.2
    \end{cases} \\
    hw(x, 0) &= 0
\end{align*}
\]
for \( x \in [-1, 1] \) with periodic boundary conditions. The exact solution consists of two shocks separated by two rarefactions. We display computed heights in Figure 6.6. For a CFL number of 0.5, Figure 6.6(a) reveals that the TECNO schemes are comparable to the ENO-LLF schemes of corresponding order. If we increase the CFL number to 0.9 then there are large oscillations in the ENO-LLF scheme (Figure 6.6(b)), while the TECNO scheme still gives reasonable results.

### 6.4.3. Euler equations.

Consider the Euler equations (1.6). We use the TECNO scheme with entropy conservative flux given by (2.18) and the ELLF diffusion matrix (2.23). The eigenvalues and eigenvectors of the Jacobian are computed at the arithmetic average of the left and right states. The method is compared to the ENO-LLF scheme. In all experiments we set \( \gamma = 1.4 \).

#### Sod shock tube.

The Sod shock tube experiment is the Riemann problem

\[
\begin{align*}
\mathbf{u}(x, 0) &= \begin{cases} 
\mathbf{u}_L & \text{if } x < 0 \\
\mathbf{u}_R & \text{if } x > 0
\end{cases}
\end{align*}
\]

with

\[
\begin{pmatrix}
\rho_L \\
w_L \\
p_L
\end{pmatrix} = \begin{pmatrix} 1 \\
0 \\
1
\end{pmatrix}, \quad \begin{pmatrix}
\rho_R \\
w_R \\
p_R
\end{pmatrix} = \begin{pmatrix} 0.125 \\
0 \\
0.1
\end{pmatrix}
\]

in the computational domain \( x \in [-5, 5] \). The initial discontinuity breaks into a left-going rarefaction wave, a right-going contact discontinuity and a right-going shock wave. The computed density at time \( t = 2 \) is shown in Figure 6.7. There is little or no difference between the two schemes. The results demonstrate that increasing \( p \) or the number of gridpoints increases the accuracy of the computed solution.
Lax shock tube. We consider the Euler equations in the computational domain $[-5, 5]$ with Riemann initial data (6.3) given by

$$
\begin{pmatrix}
\rho_L \\
w_L \\
p_L
\end{pmatrix} = 
\begin{pmatrix}
0.445 \\
0.698 \\
3.528
\end{pmatrix},
\begin{pmatrix}
\rho_R \\
w_R \\
p_R
\end{pmatrix} = 
\begin{pmatrix}
0.5 \\
0 \\
0.571
\end{pmatrix}.
$$

The computed density at $t = 1.3$ is shown in Figure 6.8. The results for ENO and TECNO schemes are very similar in this experiment. There are slight oscillations behind the shock for the TECNO schemes.

Shock-Entropy wave interaction. This numerical example was proposed by Shu and Osher in [SO89] and is a good test of a scheme’s ability to resolve a complex solution with both strong and weak shocks and highly oscillatory but smooth waves. The computational domain is $x \in [-5, 5]$ and we use initial data

$$
u(x, 0) =
\begin{cases}
u_L & \text{if } x < -4 \\
u_R & \text{otherwise},
\end{cases}
$$
6. THE TECNO SCHEME

Figure 6.7. Density at \( t = 2 \) computed with ENO-LLF and TECNO for the Sod shock tube.

\[
\begin{bmatrix}
\rho_L \\
\rho L w_L \\
\rho_L w_L \\
E_L \\
\end{bmatrix} = 
\begin{bmatrix}
3.857143 \\
2.629369 \\
10.33333 \\
\end{bmatrix}, \quad \begin{bmatrix}
\rho_R \\
\rho R w_R \\
\rho_R w_R \\
E_R \\
\end{bmatrix} = 
\begin{bmatrix}
1 + \frac{1}{5} \sin(5x) \\
0 \\
1 \\
\end{bmatrix}.
\]

The approximate solutions are computed on a mesh of 200 grid points, corresponding to about 7 grid points for each period of the entropy waves. The solution computed by the ENO-LFF and TECNO schemes are displayed in Figure 6.9. There are very minor differences between the ENO and TECNO schemes of the same order. The test also illustrates that the higher order schemes perform better than the low order schemes.

6.4.4. Numerical experiments for two dimensional Euler equations. We test the TECNO schemes for the two dimensional Euler equations

\[
\begin{bmatrix}
\rho \\
\rho w_x \\
\rho w_y \\
E \\
\end{bmatrix}_t + \begin{bmatrix}
\rho w_x^2 + p \\
\rho w_x w_y \\
\rho w_y^2 + p \\
(E + p)w_x \\
\end{bmatrix}_x + \begin{bmatrix}
\rho w_x^2 + p \\
\rho w_x w_y \\
\rho w_y^2 + p \\
(E + p)w_y \\
\end{bmatrix}_y = 0.
\] (6.4)
The density \( \rho \), velocity field \( (w^x, w^y) \), pressure \( p \) and total energy \( E \) are related by the equation of state \( E = \frac{p}{\gamma - 1} + \frac{\rho((w^x)^2 + (w^y)^2)}{2} \).

The reader is referred to [FMT12b] for the two-dimensional entropy conservative flux for (6.4). We use Roe-type diffusion matrices \( \Lambda^x \) and \( \Lambda^y \) of the form (2.24) in the TECNO diffusion operator.

**Long-time vortex advection.** We start by testing the TECNO schemes on a smooth test case for the two-dimensional Euler equations, taken from Shu [Shu97]. The initial data is set in terms of velocity \( u \) and \( v \), temperature \( \theta = \frac{p}{\rho} \) and thermodynamic entropy \( s = \log p - \gamma \log \rho \):

\[
\begin{align*}
    w^x &= 1 - (y - y_c)\varphi(r), \\
    w^y &= 1 + (x - x_c)\varphi(r), \\
    \theta &= 1 - \frac{\gamma - 1}{2\gamma}\varphi(r)^2, \\
    s &= 0,
\end{align*}
\]

where \((x_c, y_c)\) is the initial center of the vortex, \( r := \sqrt{(x - x_c)^2 + (y - y_c)^2} \) and

\[
\varphi(r) := e^{\alpha(1 - \tau^2)}, \quad \tau := \frac{r}{r_c}.
\]
We set the parameters $\gamma = 7/5$, $\epsilon = \frac{5}{2\pi}$, $\alpha = 1/2$, $r_c = 1$ and $(x_c, y_c) = (5, 5)$. The exact solution to this initial value problem is simply

$$u(x, y, t) = u(x - t, y - t, 0).$$

In other words, the initial vortex, centered at $(x_c, y_c)$ is advected diagonally with a velocity of 1 in the $x$- and $y$-directions.

The computational domain is set to $[0, 10] \times [0, 10]$ and we use periodic boundary conditions to simulate the flow over the entire plane. We compute up to $t = 100$, during which time the vortex will have traversed through the domain 10 times and should end up exactly where it started. Figure 6.10 shows the computed density at the final time step on a mesh of $200 \times 200$ points. Clearly, there is a significant gain in accuracy with increased order of accuracy, and the 3rd and 4th order TECNO schemes deviate only by a few percent from the exact solution. This experiment illustrates the robust performance of high-order TECNO schemes in resolving smooth solutions.

**Vortex-shock interaction.** This problem consists of a single left-moving shock which interacts with a right-moving vortex, and has been taken from [Shu97]. The initial shock has the values

$$u(x, 0) = \begin{cases} u_L & \text{if } x < 0.5 \\ u_R & \text{otherwise,} \end{cases}$$
with \((\rho_L, p_L, u_L, v_L) = (1, 1, \sqrt{\gamma}, 0)\) and

\[
\rho_R = \rho_L \left( \frac{1 + \beta p_R}{\beta + p_R} \right), \quad \rho_R = 1.3
\]

\[
w^1_R = \sqrt{\gamma} + \sqrt{2} \left( \frac{1 - p_R}{\sqrt{\gamma - 1} + p(\gamma + 1)} \right), \quad v_R = 0,
\]

where \(\beta := \frac{\gamma + 1}{\gamma - 1}\). The left state \(u_L\) is then perturbed slightly by adding a vortex. The exact values are specified by the perturbation in velocity, temperature and entropy:

\[
\tilde{w}^1 = \frac{y - y_c}{r_c} \varphi(r), \quad \tilde{w}^2 = -\frac{x - x_c}{r_c} \varphi(r), \quad \tilde{\theta} = -\frac{1}{4\sigma \gamma} \varphi(r)^2, \quad \tilde{s} = 0.
\]

Here, \(r\) and \(\varphi\) are exactly as in the previous experiments. We set the free parameters to be \(\epsilon = 0.3\), \((x_c, y_c) = (0.25, 0.5)\), \(r_c = 0.05\) and \(\alpha = 0.204\). With these parameters, the jump in pressure across the shock wave is about twice as big as the magnitude of the vortex.

We compute on the domain \([0, 1] \times [0, 1]\) up to time \(t = 0.35\). The domain is partitioned into \(200 \times 200\) grid points. The computed densities are plotted in Figure 6.11. The results show that the TECNO schemes resolve both the shock as well as the smooth vortex well. There is a gain in resolution as higher order accurate schemes are employed. The results are comparable to those obtained with standard ENO and WENO schemes in [Shu97].
Cloud-shock interaction: The initial data for this test case is set to be

\[
\rho = \begin{cases} 
3.86859 & \text{if } x < 0.05 \\
10 & \text{if } r < 0.15 \\
1 & \text{otherwise}
\end{cases}
\]

\[
p = \begin{cases} 
167.345 & \text{if } x < 0.05 \\
10 & \text{otherwise}
\end{cases}
\]

\[
w^x = \begin{cases} 
11.2536 & \text{if } x < 0.05 \\
1 & \text{otherwise}
\end{cases}
\]

\[w^y \equiv 0,\]

where \( r := \sqrt{x^2 + y^2} \). The computational domain is \([0, 1] \times [0, 1]\) with Neumann type non-reflecting boundary conditions. The exact solution in this case consists of the interaction of a right moving shock with a high density bubble, resulting in a complicated pattern that includes both bow and tail shocks as well as smooth regions in the center of the domain. The computed densities on a mesh of 200×200 points at time \( t = 0.06 \) are shown in Figure 6.12. For the sake of comparison, a reference solution computed with the TECNO\(_3\) scheme on a mesh of 1400×1400 points is also shown. The results illustrate that the TECNO schemes are stable and resolve the complex solution quite well. There is a clear gain in accuracy with the TECNO\(_3\) and TECNO\(_4\) schemes compared to the TECNO\(_2\) scheme.
Figure 6.12. TECNO schemes on the cloud-shock interaction problem. The density at time $t = 0.06$ is plotted on a mesh of $200 \times 200$ points. A reference solution is also plotted for comparison.
CHAPTER 7

Conclusions

We have developed high-order accurate finite difference schemes for both scalar conservation laws and systems of hyperbolic conservation laws. The entropy stability of the scheme provides a bound on the gradient which allows us to prove convergence to the entropy solution for scalar equations, and convergence to an entropy measure-valued solution for systems of equations. The scheme is non-oscillatory and robust with respect to different equations and initial data. It is free of any tuning parameters and is straightforward to implement. Finally, it is computationally cheap, both for one-dimensional scalar problems and for multi-dimensional systems of equations. We have seen that the scheme performs well on a series of test problems compared to existing finite difference schemes.

Future work

We conclude with the following suggestions for future work:

The convergence proof of the TECNO scheme relies on Conjecture 5.16 about the total variation bound of the ENO procedure. A proof (or a counterexample) of this conjecture would be interesting also in its own right.

The high-order accurate entropy conservative fluxes which the TECNO schemes are built on rely on a Cartesian mesh. To make the scheme more applicable to real-world problems, a generalization to unstructured meshes would be necessary. Entropy conservative schemes on unstructured meshes would presumably take the form of finite element methods; see [Tad03, Section 9]. These would then be coupled with appropriately chosen diffusion operators to guarantee the stability and convergence of the method.

The schemes considered in this thesis are semi-discrete. It would be interesting to investigate explicit time integration methods such that the resulting fully discrete scheme is entropy stable. The nonlinearity of the spatial discretization makes this a formidable challenge. For linear spatial discretizations, it has been shown that the third-order explicit Runge-Kutta scheme is entropy stable [Tad02], but preliminary investigations indicate that this does not carry over to the fourth- or fifth-order scheme.
APPENDIX A

Young measures

In this chapter we review the theory of measure-valued functions, so-called Young measures. We refer the reader to e.g. [MNRR96] for a more thorough introduction.

We denote by $C_0(\mathbb{R}^n)$ the set of continuous functions $g : \mathbb{R}^n \to \mathbb{R}$ such that $|g(\lambda)| \to 0$ as $|\lambda| \to \infty$, and equip it with the supremum norm. The dual space of $C_0(\mathbb{R}^n)$ is denoted $M(\mathbb{R}^n)$, and it may be shown that this space is isometrically isomorphic to the set of finite Radon measures on $\mathbb{R}^n$ (a Radon measure is a Borel measure that is inner regular and locally finite). We will not distinguish between these equivalent definitions of $M$ and will frequently write for $g \in C_0(\mathbb{R}^n), \mu \in M(\mathbb{R}^n)$

$$\langle \mu, g \rangle = \int_{\mathbb{R}^n} g(\lambda) \, d\mu(\lambda).$$

A frequently occurring subset of $M(\mathbb{R}^n)$ is the set of probability measures, $\text{Prob}(\mathbb{R}^n) := \{\mu \in M(\mathbb{R}^n) : \mu \geq 0, \mu(\mathbb{R}^n) = 1\}$. (By $\mu \geq 0$ we mean $\mu(B) \geq 0$ for all Borel sets $B$, or equivalently, $\langle \mu, g \rangle \geq 0$ for all nonnegative $g \in C(\mathbb{R}^n)$.)

Let $\Omega \subset \mathbb{R}^m$ be open. A Young measure is a weak-$^*$ measurable function $\nu : \Omega \to \text{Prob}(\mathbb{R}^n)$; we denote $\nu$ as being parameterized by its argument, $\nu_y := \nu(y)$. Weak-$^*$ measurability means that the function $y \mapsto \langle \nu_y, g \rangle$ is measurable for all $g \in C(\mathbb{R}^n)$. Given a Young measure $\nu$, we denote $\overline{g}(y) := \langle \nu_y, g \rangle$ and $\overline{\mu}(y) := \langle \nu_y, \text{id} \rangle = \int_{\mathbb{R}^n} \lambda \, d\nu_y(\lambda)$. (We frequently abuse notation and write $\langle \nu_y, \lambda \rangle = \int_{\mathbb{R}^n} \lambda \, d\nu_y(\lambda)$.) The most obvious example of a Young measure is a function. If $u : \Omega \to \mathbb{R}^n$ is any measurable function, then $\nu_y := \delta_{u(y)}$ defines a Young measure which satisfies $\overline{g} = g(\overline{\mu})$ for all $g \in C(\mathbb{R}^n)$.

The full power of Young measures is seen in the following Fundamental Theorem of Young measures: given any sequence of functions, it is enough to assume a uniform $L^\infty$ bound to guarantee (weak) convergence towards a Young measure.

**Theorem A.1** (Tartar [Tar79]). Let $\Omega \subset \mathbb{R}^n$ be open and let $u^k : \Omega \to \mathbb{R}^n$ be a bounded sequence in $L^\infty(\Omega, \mathbb{R}^n)$. Let $K \subset \mathbb{R}^n$ be the bounded set $K = \bigcup_{k \in \mathbb{N}} \text{range}(u^k)$. Then there is a subsequence, still denoted by $u^k$, and a Young measure $\nu$ with $\text{supp}(\nu_y) \subset K$ such that

$$g(u^k) \rightharpoonup \overline{g} \quad \text{in} \quad L^\infty(\Omega)$$

for all $g \in C(\mathbb{R}^n)$, where $\overline{g}(y) = \langle \nu_y, g \rangle = \int_{\mathbb{R}^n} g(\lambda) \, d\nu_y(\lambda)$.
Theorem A.1 has been generalized to unbounded sequences of functions that decay “sufficiently fast” at infinity; see Ball [Bal89].

**Lemma A.2** (DiPerna [DiP85], Corollary 2.1). Let \( \Omega \subset \mathbb{R}^m \) be open and let \( \{u_k\}_{k \in \mathbb{N}} \) be a bounded sequence in \( L^\infty(\Omega, \mathbb{R}^n) \), converging weakly-* to some \( u \in L^\infty(\Omega, \mathbb{R}^n) \). Let \( \nu \) be the Young measure associated with \( \{u_k\}_k \). Then the following are equivalent:

1. \( \nu_y = \delta_{u(y)} \) for a.e. \( y \in \Omega \).
2. \( u_k \to u \) in \( L^p_{\text{loc}}(\Omega, \mathbb{R}^n) \) for every \( 1 \leq p < \infty \), and for every compact \( K \subset \mathbb{R}^m \) there is a subsequence such that \( u_k \to u \) a.e. in \( K \cap \Omega \).

**Proof.** By extending \( u_k \) and \( u \) by zero we may assume that \( \Omega = \mathbb{R}^m \). If \( \nu_y = \delta_{u(y)} \) for a.e. \( y \) then \( \overline{g}(y) = g(u(y)) \) for a.e. \( y \). Hence, \( g(u_k) \xrightarrow{\ast} g(u) \) in \( L^\infty \), so in particular,

\[
\int_K (u - u_k)^2 \varphi \, dy = \int_K (u^2 + u_k^2 - 2u_k u) \varphi \, dy \to 0
\]

for every \( \varphi \in L^1(K) \) and compact \( K \subset \mathbb{R}^m \). Setting \( \varphi = 1 \), we find that \( u_k \to u \) in \( L^2_{\text{loc}} \). Convergence in \( L^p_{\text{loc}} \) for all \( 1 \leq p < \infty \) follows from this and the \( L^\infty \) bound on \( u_k \). Last, pointwise convergence follows from a result due to Riesz [Bar66, pp. 71-73].

Conversely, assume that \( u_k \to u \) pointwise a.e. Then \( g(u_k(y)) \to g(u(y)) = \overline{g}(y) \) for a.e. \( y \in \Omega \), or equivalently, \( (\nu_y, g) = (\delta_{u(y)}, g) \) for all \( g \in C_0(\mathbb{R}) \). Hence, \( \nu_y = \delta_{u(y)} \). \( \square \)

Hence, to show that the convergence \( u_k \to u \) is strong, it suffices to show that \( \nu_y \) is a Dirac measure for a.e. \( y \). The following auxiliary lemma will prove convenient in showing this.

**Lemma A.3** (Vecchi [MNRR96]). Let \( \Omega \subset \mathbb{R}^m \) be open and let \( \nu : \Omega \to \text{Prob}(\mathbb{R}^n) \) be a Young measure. Then the following are equivalent:

1. There is a \( u \in L^\infty(\Omega, \mathbb{R}^n) \) such that \( \nu_y = \delta_{u(y)} \) for a.e. \( y \in \Omega \).
2. For every \( y \in \Omega \) there is an \( H \in L^1(\mathbb{R}^n \times \mathbb{R}^n; \nu_y \times \nu_y) \) such that

\[
H(a, b) \begin{cases} > 0 & \text{if } a \neq b \\ = 0 & \text{if } a = b \end{cases} \quad \text{and} \quad \int_{\mathbb{R}^n \times \mathbb{R}^n} H \, d(\nu_y \times \nu_y) = 0 \quad \forall \ y \in \Omega.
\]

**Proof.** If \( \nu = \delta_a \) then \( H(a, b) = |a - b| \) satisfies (A.1).

Let \( H \) satisfy (A.1), and assume conversely that there exist distinct points \( a, b \in \text{supp}(\nu_y) \) for a \( y \in \Omega \). Let \( U \) and \( V \) be disjoint neighbourhoods of \( a \) and \( b \), respectively. Then \( H(y_1, y_2) > 0 \) for all \( (y_1, y_2) \in U \times V \). Hence,

\[
0 = \int_{\mathbb{R}^n \times \mathbb{R}^n} H \, d(\nu_y \times \nu_y) \geq \int_{U \times V} H \, d(\nu_y \times \nu_y) > 0,
\]

a contradiction. Hence, \( \text{supp}(\nu_y) \) can contain at most one point, so either \( \nu_y = 0 \) or \( \nu_y = \delta_{u(y)} \) for some \( u(y) \in \mathbb{R}^n \). But since \( \nu_y(\mathbb{R}^n) = 1 \), it cannot vanish everywhere, so we must have \( \nu_y = \delta_{u(y)} \). The function \( u \) is measurable by the weak-* measurability of \( \nu \), and is in \( L^\infty(\Omega, \mathbb{R}^n) \) since \( |u(y)| \leq \sup_z \sup_{w \in \text{supp}(\nu_y)} |w| < \infty \). \( \square \)
Bibliography


