NTNU – Trondheim
Norwegian University of
Science and Technology

High-Resolution Large Time-Step Schemes for Hyperbolic Conservation Laws

Sigbjørn Løland Bore

Master of Science in Physics and Mathematics
Submission date: June 2015
Supervisor: Jon Andreas Støvneng, IFY

Norwegian University of Science and Technology
Department of Physics

Summary

This thesis is concerned with numerical methods for solving hyperbolic conservation laws. A generalization of large time-step schemes (LTSS) to high resolution is presented. The generalization is based on the previous work in [Lindquist, 2014; Harten, 1986]. Starting from a general LTSS, a set of sufficient conditions for conservative, consistent, and total variation diminishing (TVD) LTSS are derived. Second-order accuracy away from discontinuities is achieved by a modified flux approach. Such an approach is shown to be TVD whenever a supplementary condition is satisfied. The full set of criteria constitutes a new framework of sufficient conditions for high-resolution LTSS. By application of this framework on the large time-step Roe scheme (LTS-Roe1), a new second-order version (LTS-Roe2) is proposed. Further, to overcome the problems of LTS-Roe1 and LTS-Roe2 with transonic rarefaction, a hybrid scheme of LTS-Roe and Lax-Friedrichs is proposed (Hybrid). The methods are investigated and compared against the second-order LTS-Harten. This is done by numerical studies on Burgers' equation and on the Euler equations. Numerical tests for continuous initial conditions show second-order convergence for all methods. For discontinuous initial conditions LTS-Roe2 has better accuracy than LTS-Roe1- however, this difference becomes small for high CFL-numbers. LTS-Roe2 is shown to have a very good resolution of discontinuities, but for high CFL-numbers it produces spurious oscillations for the Euler equations. Hybrid is more diffusive, but has no problems with transonic rarefaction. Tests show that LTS-Harten consistently gives good results with less oscillations than LTS-Roe2, but it has, however, a tendency to smear out discontinuities when the CFL-number is increased.

Sammendrag

Denne avhandlingen omhandler numeriske metoder for å løse hyperbolske konserveringslover. En generalisering av skjema med store tidssteg (LTSS) til høy oppløsning blir presentert. Generaliseringen er basert på tidligere arbeid i [Lindqvist, 2014; Harten, 1986]. Tilstrekkelige betingelser for et konservativt, konsistent, og totalvariasjonsforminskende LTSS utledes. Andreordens nøyaktighet borte fra diskontinuiteter oppnås med en modifisert-fluks-tilnærming. Denne tilnærmingen blir vist å være totalvariasjonsforminskende når en ekstra betingelse er tilfredsstilt. Det hele settet med kriterier gir et nytt rammeverk av tilstrekkelige betingelser for høyoppløsnings-LTSS. Ved annvendelse av dette rammeverket på Roe med store tidssteg (LTS-Roe1), blir en ny andreordensversjon foreslått (LTS-Roe2). Videre, for å overvinne problemene til LTS-Roe2 med transsonisk fortynning, blir et nytt hybridskjema bestående av LTS-Roe2 og Lax-Friedrichs foreslått (Hybrid). Metodene blir undersøkt og sammenlignet med LTS-Harten [Harten, 1986]. Dette gjøres ved numeriske eksperiment på Burgers ligning og Euler-likningene. Numeriske tester for kontinuerlige initialbetingelser viser andreordenskonvergens for alle andreordensmetodene. For diskontinuerlige initialbetingelser har LTS-Roe2 bedre nøyaktighet enn LTS-Roe1; imidlertid er forskjellen liten for høye CFL-tall. LTS-Roe2 har god oppløsning av diskontinuiteter, men sliter for høye CFL-tall med falske oscillasjoner på Eulerligningene. Testene viser at LTS-Harten gir gjennomgående bedre resultater, men har en tendens til smøre ut løsninger ettersom CFL-tallet økes.

Table of Contents

Chapter 1

Introduction

In this master thesis we consider numerical methods for solving hyperbolic conservation laws

$$\mathbf{q}_t + \mathbf{f}_x(\mathbf{q}) = 0, \quad \mathbf{q}(x,0) = \boldsymbol{\phi}(x), \quad -\infty < x < \infty. \tag{1.1}$$

Here $\mathbf{q} = \mathbf{q}(x,t)$ is the conserved quantity of the system, $\mathbf{f}(\mathbf{q})$ is the flux of $\mathbf{q}$ through the walls of a control volume and $\boldsymbol{\phi}(x)$ is the initial configuration. Subscripts indicate derivatives with respect to x and t. Hyperbolic conservation laws are useful in describing systems where conserved quantities are transported. Phenomena ranging from gas flow, liquid flow, to even traffic flow [1], can all be described by hyperbolic conservation laws.

An important concept in hyperbolic conservation laws is that information or solutions travel at finite speeds. If the law is nonlinear, then the speed will be spatially dependent and discontinuous solutions called *shocks* can form. *Shock formation* makes it hard to analyze such equations analytically. For complex problems, numerical solution is the only tractable way of treating such systems. We will in this thesis focus on the explicit *finite volume method* (FVM). In this method we take advantage of the fact that information travels at a finite speed. This is done by using local differences to approximate the derivative of the flux and integrate explicitly in such a way as for the method to be conservative. A fundamental limit to the explicit FVM is the *Courant-Friedrichs-Lewy condition* (CFL-condition). It states that the time step must be small enough so that information can not travel further than the stencil used (the set of points used). For instance, basic FVMs like Lax–Friedrichs, Upwind and Godunov use a 3-point stencil to approximate the flux between cells. That is, they use information from the cell and its two neighbours to compute the state at the next time level. For such methods, if the fastest information travels at speed a, then the time step is limited by

$$\Delta t \leq \frac{\Delta x}{a}. \tag{1.2}$$

If the time step is too large, then it will be impossible for the numerical method to capture all the dynamics. This is often reflected in an unstable method. This condition can also be

expressed as

$$CFL \leq 1, \tag{1.3}$$

where we have defined the CFL-number as

$$CFL = \frac{a\Delta t}{\Delta x}. \tag{1.4}$$

The CFL-number can be interpreted as the number of cells the fastest moving information can travel during a single time step Δt. Methods that use a $(2N+1)$-point stencil are in principle limited by

$$CFL \leq N. \tag{1.5}$$

However in practice, high resolution methods using limiters [5] and ENO- and WENO-methods [8], all use a wider stencil, but are limited by (1.2). Methods that are stable for any time step are called *large time-step schemes* (LTSS) and are the topic of this thesis.

There are many motivations for developing LTSS. Most important is computational speed. Not having to adhere to (1.2) means fewer time steps. If the stencil is widened from $N = 1$ to $N = 10$, the simulation could run 10 times faster. However in practice, widening the stencil means more work for each iteration. This cost may or may not be compensated for by fewer iterations and is problem dependent. In many cases, most of the time is not used on the iteration itself, but by the Riemann solver at each cell wall. In such cases the added computational cost of the wider stencil is likely to be well compensated for. Another type of problem where LTSS are likely to be useful, is for hyperbolic relaxation systems. Typically such systems are solved by a fractional step approach, dividing the partial differential equation into two separate equations and numerically integrating them separately. Often the relaxation step is computationally very costly (for instance a very complex equation of state). In such cases, LTSS can significantly reduce the number of relaxation steps needed and increase computational speed.

1.1 Previous work on large time-step schemes

In the early 80s LeVeque introduced the first LTSS [11–13]. His approach was based on the wave interpretation of the Godunov method. He generalized the existing Godunov method by letting waves travel beyond a single cell without any interaction. Projecting the waves, he obtained the method referred to as *LTS-Godunov*. Initial results on scalar equations were promising. However there were problems with oscillations for systems of conservation laws. Inspired by LeVeque, Harten proposed in[6] a new method based on dividing the time step into N equal steps. Doing so he was able to derive a method which was unconditionally *total variation diminishing* (TVD) with an entropy fix (an addition to the method to ensure that the numerical solution converges to the correct solution). Further he generalized this method to second-order accuracy away from discontinuities. Numerical tests tended to show smearing of the solution as the time step is increased. This was contrary to LeVeque who observed oscillations. Some follow-up work has been done on both of these methods. In [16] the work by Leveque is expanded by taking interactions

between waves into account. They show that the new method is less prone to entropy mistakes. In [17] Harten's LTSS is further studied. Here the authors show that Harten's scheme can be made more stable by a very small modification. Further, other linearizations than Roe are used and simulations are done for two dimensions with dimensional splitting. More fundamental work was done by Sofia Lindqvist during two summer internships at SINTEF [14]. The approach differed from that of Harten and LeVeque. Instead of using a wave interpretation or existing schemes as a building block, a general framework was made for developing LTSS. In this framework conditions are derived for which a method is consistent and TVD. The modified equation was also derived for a general LTSS. Analyzing the modified equation, LTS-Roe (LTS-Roe1) was derived. This scheme corresponds to the LTS-Godunov done with Roe-linearization [19]. LTS-Roe1 was found to be the least diffusive scheme possible. Further, the most diffusive scheme, LTS-Lax-Friedrichs was found. Another important result was the discovery that by using random CFL-numbers one could remove entropy mistakes (the numerical solution converging to the wrong solution) produced by LTS-Roe1 and increase accuracy.

1.2 Organization of thesis

The thesis is organized into five parts. In the first two parts we give a short review of hyperbolic conservation laws and the explicit FVM. Here we introduce important concepts like characteristics, the Riemann problem, and shock formation. Further we show how these are related to the FVM. We also introduce the Godunov method and show how a system of conservation laws can be treated with the FVM. The fourth part is dedicated to the theory of high-resolution LTSS. Here we propose and prove fundamental sufficient conditions for a conservative, consistent, and total variation diminishing LTSS. We prove that a general LTSS can be made second order by a modified flux approach and find a supplementary sufficient condition on the coefficients of LTSS such that it is also TVD. The fourth part of this thesis contains numerical investigations of LTSS. Here we present and evaluate results from the application of LTSS on the Burgers' equation and the Euler equations. Finally in the fifth part we summarize the most important conclusions of this thesis and present future prospects.

Chapter 2

Hyperbolic conservation laws

Conservation laws are derived by considering conserved quantities of a system. These quantities can be anything from the total mass, momentum, energy, to the number of cars on a highway. The change of a conserved quantity $\mathbf{q}$ inside the control volume Ω with surface $\partial\Omega$ is given by

$$\frac{\partial}{\partial t}\int_{\Omega}\mathbf{q}\mathrm{d}V + \int_{\partial\Omega}\mathbf{f}(\mathbf{q})\cdot\hat{\mathbf{n}}\mathrm{d}A = 0. \tag{2.1}$$

Here $\mathbf{f}(\mathbf{q})$ is the flux of $\mathbf{q}$, $\mathrm{d}V$ is a differential volume, $\mathrm{d}A$ is a differential area and $\hat{\mathbf{n}}$ its the normal vector. Equation (2.1) can be interpreted as the change of $\mathbf{q}$ inside the control volume being equal to the flux of $\mathbf{q}$ through its boundary. This form is the most general, but impractical for solving problems. A partial differential equation formulation is obtained by using the Gauss theorem

$$\frac{\partial}{\partial t}\int_{\Omega}\mathbf{q}\mathrm{d}V + \int_{\Omega}\boldsymbol{\nabla}\cdot\mathbf{f}(\mathbf{q})\mathrm{d}V = 0. \tag{2.2}$$

Using the fundamental theorem of calculus (2.2) can be written as

$$\frac{\partial\mathbf{q}}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{f}(\mathbf{q}) = 0, \tag{2.3}$$

a partial differential equation.

Remark It is important to note that in obtaining this equation we have assumed that $\mathbf{q}$ is a continuous function. However, this is not always the case. For nonlinear conservation equations, discontinuous solutions can emerge from continuous initial conditions. In such situations we are forced to return to the original integral formulation.

2.1 The concept of characteristics

Limiting to one dimension, (2.3) can be written in quasilinear form as

$$\frac{\partial \mathbf{q}}{\partial t} + \mathbf{A}(\mathbf{q})\mathbf{q}_x = 0, \tag{2.4}$$

where

$$\mathbf{A}(\mathbf{q}) \equiv \frac{\partial \mathbf{f}\,(\mathbf{q})}{\partial \mathbf{q}}, \tag{2.5}$$

a matrix referred to as the *Jacobian* of the system. A m-order partial differential system is termed *hyperbolic* when $\mathbf{A}(\mathbf{q})$ can be diagonalized and all its eigenvalues $a^p(\mathbf{q})$ are real. This implies that $\mathbf{A}(\mathbf{q})$ can be written as

$$\mathbf{A}(\mathbf{q}) = \mathbf{T}^{-1}\mathbf{D}\mathbf{T}. \tag{2.6}$$

Here $\mathbf{D} = \text{diag}\left(a(\mathbf{q})^1, \ldots, a^m(\mathbf{q})\right)$, and $\mathbf{T}$ and $\mathbf{T}^{-1}$ are a similarity matrix and its inverse. Multiplying (2.4) by $\mathbf{T}$ and using (2.6), we find

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{D}\frac{\partial \mathbf{v}}{\partial x} = 0, \tag{2.7}$$

where

$$\partial \mathbf{v} = \mathbf{T}\partial \mathbf{q}, \tag{2.8}$$

the *characteristic* variables of the system. As $\mathbf{D}$ is diagonal we can rewrite (2.7) as

$$\frac{\partial v^p}{\partial t} + a^p(\mathbf{q})\frac{\partial v^p}{\partial x} = 0. \tag{2.9}$$

Thus the evolution of each of the characteristic variables can be described in terms of wave equations with the eigenvalue as the speed of propagation. The full solution can thus be thought of in terms of a combination of characteristic variables travelling at corresponding eigenspeeds. This is best illustrated by considering some examples of different conservation laws.

2.1.1 Advection equation

The simplest conservation law is the advection equation

$$q_t + (aq)_x = 0, \tag{2.10}$$

where a is the constant advection speed. The solution of this equation is

$$q = q_0(x - at), \tag{2.11}$$

thus the initial configuration is constant on $x = x_0 + at$. Figure 2.1 shows how the initial configuration is advected by traveling on characteristic lines in the time-position space.

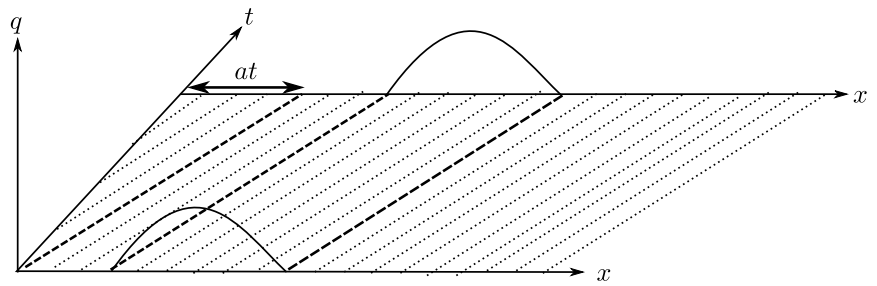


Figure 2.1: Illustration of how an initial solution travels on characteristic lines for the advection equation.

2.1.2 Scalar conservation law

For a scalar conservation law we have that

$$q_t + a(q)q_x = 0. \tag{2.12}$$

We can interpret this equation as q being unchanged on lines $x = x_0 + a(q)t$. Therefore the solution (before the onset of shocks) is given similarly to (2.11)by

$$q(x,t) = q_0(x - a(q)t). \tag{2.13}$$

2.1.3 Linear system

For a linear system, (2.9) reduces to m independent advection equations, the solutions of which are given by

$$v^p(x,t) = v^p(x - a^p t, 0). \tag{2.14}$$

The solution in terms of the conserved variable is obtained by

$$\mathbf{q} = T^{-1}\mathbf{v}. \tag{2.15}$$

2.1.4 The Euler equations

The Euler equations is a set of conservation equations that describes inviscid fluid flow. In conservation form they are given by

$$\begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix}_t + \begin{pmatrix} \rho u \\ \rho u^2 + P \\ u(E+P) \end{pmatrix}_x = 0. \tag{2.16}$$

Here ρ is the density of the fluid, ρu is the momentum density, E is density of energy, and P is the pressure. For an ideal gas

$$P = (\gamma - 1)\left(E - \frac{1}{2}\rho u^2\right), \tag{2.17}$$

where $\gamma = c_\mathrm{p}/c_\mathrm{v}$, the ratio of specific heats (for air $\gamma = 1.4$). The eigenvalues of the Jacobi matrix $\mathbf{A}(\mathbf{q}) = \partial \mathbf{f}(\mathbf{q})/\partial \mathbf{q}$ are given by

$$a^1(\mathbf{q}) = u - c, \tag{2.18a}$$

$$a^2(\mathbf{q}) = u, \tag{2.18b}$$

$$a^3(\mathbf{q}) = u + c, \tag{2.18c}$$

where $c = (\gamma P/\rho)^{1/2}$ is the speed of sound. The corresponding characteristic variables are given by

$$\partial v^1 = \partial P - \rho c \partial u, \tag{2.19a}$$

$$\partial v^2 = \partial P - c^2 \partial \rho, \tag{2.19b}$$

$$\partial v^3 = \partial P + \rho c \partial u. \tag{2.19c}$$

We can interpret the equations above as ∂v^p being zero on lines

$$\frac{\mathrm{d}x_p}{\mathrm{d}t} = a^p(q). \tag{2.20}$$

Figure 2.2 shows characteristic lines on which the characteristic variables are constant. Note that unlike for a linear system or a scalar nonlinear conservation law, the lines no longer have a constant slopes.

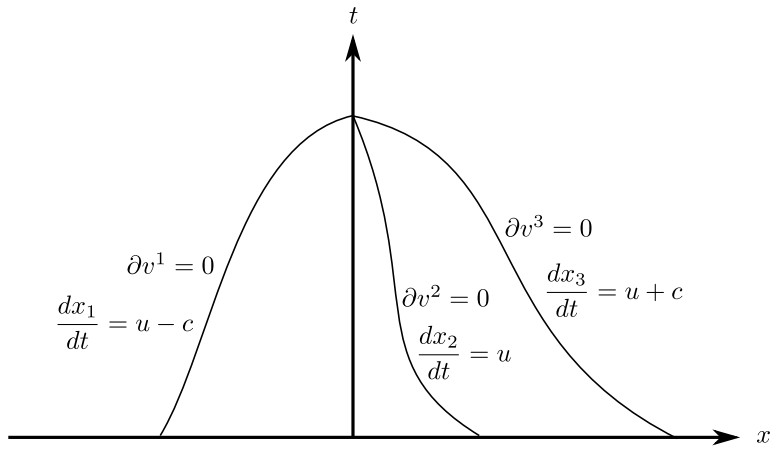


Figure 2.2: Characteristic lines for the characteristic variables of the Euler equations.

2.2 Nonlinear conservation laws and shock formation

For nonlinear conservation laws, the eigenvalues will be dependent on $\mathbf{q}$. This means that how fast the characteristic variables travel, depends on the local configuration. The

simplest example of what can occur is with the *Burgers' equation*

$$q_t + (\frac{1}{2}q^2)_x = 0,$$

and in quasilinear form,

$$q_t + qq_x = 0. \tag{2.21}$$

The solution is given implicitly by

$$q = q(x - qt, 0), \tag{2.22}$$

where the eigenvalue is q. This means that the solution is constant on lines of $x = x_0 + qt$. Thus starting from initial values $q_0(x)$, solutions travel on lines $x = x_0 + q_0 t$. The best way to visualize this is by drawing lines in the (x, t)–plane as in Figure 2.3. First, note that

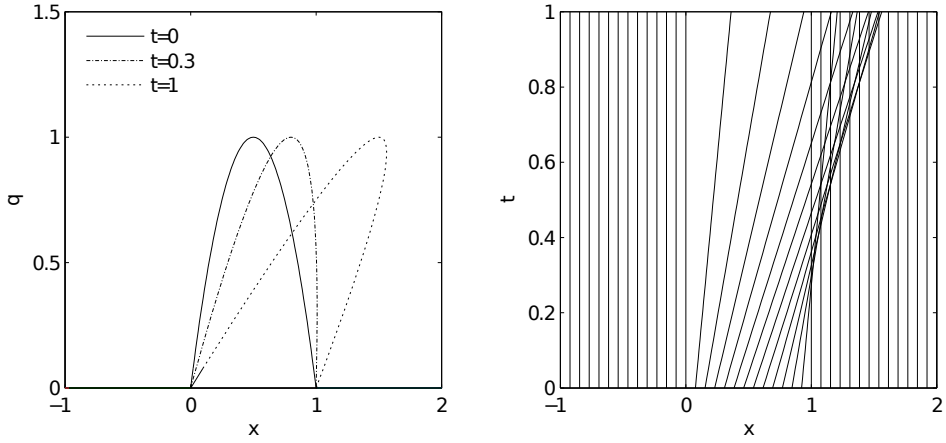


Figure 2.3: (Left) Initial condition and solution of Burgers' equation at $t = 0$, $t = 0.3$, and $t = 1$. (Right) Characteristic lines on which the solution is constant.

as for the advection equation, the lines are straight. However, due to the nonlinearity the slope differ dependent on the initial velocity. We see in the right graph that solutions with high initial velocity will travel faster to the right than solutions with low initial velocity. This results in the creation of a *shock* at $t = 0.3$, shown in Figure 2.3 – a discontinuity in the solution. The time at which the sharp discontinuity appears coincides with the time the characteristic lines start intersecting. At time $t = 1$ we see that the solution is no longer unique (at some positions in space there are two values of q). This can also be seen in the right graph where characteristic lines cross. For a physical problem there can only be one solution, but which one? The problem appears when the lines starts to intersect. At this time the derivative is ill–defined and hence (2.21) is not valid. Keep in mind that Burgers' equation we wrote is a special case of the more general integral version and for the integral version the equations are defined for discontinuous solutions. The problem of finding the correct behaviour at these discontinuities is strongly related to the *Riemann problem*, and solving it is the key to good numerical methods for hyperbolic systems.

2.3 The Riemann problem

We saw in the previous section that the method of characteristics works well until the lines start intersecting. At this point a discontinuity forms and the method of characteristics is no longer valid. Our problem is thus only in dealing with the discontinuity. A model problem with the same challenge is the Riemann Problem (for an illustration see Figure 2.4)

$$\begin{cases} \mathbf{q}_t + (\mathbf{f}(\mathbf{q}))_x = 0, \\ \mathbf{q}(x,0) = \begin{cases} \mathbf{q}_L & x < 0 \\ \mathbf{q}_R & 0 \leq x \end{cases} \end{cases} . \tag{2.23}$$

A possible solution of this problem is for the discontinuity to move without change in

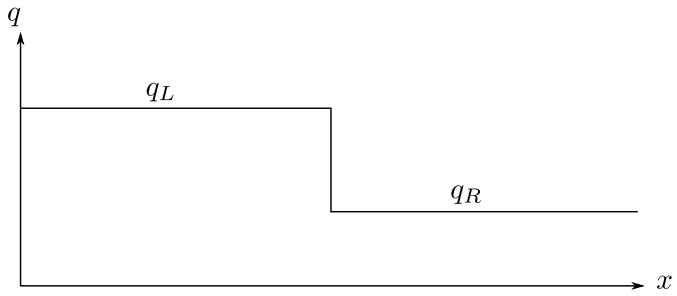


Figure 2.4: Illustration of the Riemann problem.

form, this referred to as a shock. Using the integral formulation, we can derive the speed s at which this discontinuity should travel (also called the Rankine-Hugoniot conditions). The speed and conditions are derived by integrating (2.1) over $[x_-, x_+]$ $(x_- < 0 < x_+)$

$$\frac{\partial}{\partial t} \int_{x_-}^{x_+} \mathbf{q} \mathrm{d}x + \mathbf{f}(\mathbf{q}_R) - \mathbf{f}(\mathbf{q}_L) = 0,$$

$$\frac{\partial}{\partial t} \left(\mathbf{q}_L \cdot (x_{\text{shock}} - x_-) + \mathbf{q}_R \cdot (x_+ - x_{\text{shock}}) \right) = - \left(\mathbf{f}(\mathbf{q}_R) - \mathbf{f}(\mathbf{q}_L) \right),$$

giving the Rankine Hugoniot conditions

$$(\mathbf{q}_R - \mathbf{q}_L)s = (\mathbf{f}(\mathbf{q}_R) - \mathbf{f}(\mathbf{q}_L)) \tag{2.24}$$

where x_{shock} is the position of the shock discontinuity and $s \equiv \partial x_{\text{shock}}/\partial t$ its speed. Note that s has to satisfy this equation for all components. We now examine some specific examples of this.

Scalar In the scalar case we are left with the shock satisfying the following condition

$$s = \frac{f(q_R) - f(q_L)}{q_R - q_L}. \tag{2.25}$$

For the advection equation ($f(q) = au$) we have

$$s = \frac{aq_R - aq_L}{q_R - q_L} = a. \tag{2.26}$$

Not surprisingly the shock moves at the advection speed a. For the Burgers' equation (2.21)

$$s = \frac{\frac{1}{2}q_R^2 - \frac{1}{2}q_L^2}{q_R - q_L} = \frac{1}{2}\left(q_R + q_L\right), \tag{2.27}$$

i.e. the arithmetic average of the two velocities.

Linear system For a linear system we have $\mathbf{f}(\mathbf{q}) = \mathbf{A}\mathbf{q}$ and

$$\mathbf{A}\left(\mathbf{q}_R - \mathbf{q}_L\right) = s\left(\mathbf{q}_R - \mathbf{q}_L\right). \tag{2.28}$$

Thus for a linear system one may interpret s and

$$\Delta\mathbf{q} = \mathbf{q}_R - \mathbf{q}_L \tag{2.29}$$

as being an eigenvalue and eigenvector of $\mathbf{A}$. Thus the discontinuity moves at the same speed as one of the eigenvalues. This is exactly what we see for the scalar advection equation where the shock speed is equal to a. If m eigenvalues of $\mathbf{A}$ are unique, then the initial discontinuity will be decomposed into $m + 1$ constant states separated by m discontinuities.

So far we have stressed that from the integral formulation of conservation laws, one may obtain solutions that allow for discontinuities and we have obtained the speed at which these discontinuities travel. However, in using the integral formulation great care must be taken. Unlike for the partial differential formulation, this formulation does not have a unique solution, but many generalized solutions (solutions of the integral equations). Thus the challenge of the Riemann problem reduces to finding the solution that corresponds to the physical solution (also referred to as the strong solution). A guiding principle for finding the correct solution is to add a small amount of viscosity. For Burgers' equation this is done as follows

$$q_t + (\frac{1}{2}q^2)_x = \nu q_{xx}. \tag{2.30}$$

This equation can no longer have discontinuities. Taking the limit at which $\nu \to 0$ we obtain the true solution[1]. In practice, adding viscosity in such a way for systems is a cumbersome way to find the correct solution. The addition of viscosity makes an already complex problem even more complicated. However, we can show that this is equivalent to the method of choice, which is to check if the solution obeys the *Entropy conditions*. This method originates from the Euler equations, where the correct solution of the Riemann problem was determined by finding which solution increases the entropy of the system.

[1] Numerical techniques for solving such problems often introduce such a viscosity; this helps to explain why methods with large viscosity (like Lax–Friedrichs) converge to the correct solution.

For a scalar convex ($f'(q) > 0$) conservation law we have the Lax entropy conditions: A shock is the correct solution given that

$$f'(q_L) > s > f'(q_R). \tag{2.31}$$

That is if the left side of the discontinuity has a higher characteristic speed than the shock speed and the shock speed has a higher speed than the right side, then the solution will be a shock. For Burgers' equation, we have that $s = 0.5(q_R + q_L)$. Thus we have for the Riemann problem the following cases

$$q_L < s < q_R, \tag{2.32a}$$

$$q_L > s > q_R. \tag{2.32b}$$

$q_R < q_L$ satisfies the entropy condition and thus for this case we get a shock. As $q_L < q_R$ does not satisfy (2.31), this means that a shock is not a physical solution. In this case the physical solution is the rarefaction fan solution shown in Figure (2.5). Why this is the correct solution can be seen by considering vanishing viscosity. If we have viscosity we expect the discontinuity to be of finite width. Drawing the characteristics we see that there are none intersecting lines. Taking the limit of vanishing viscosity we get the rarefaction fan shown in Figure (2.5).

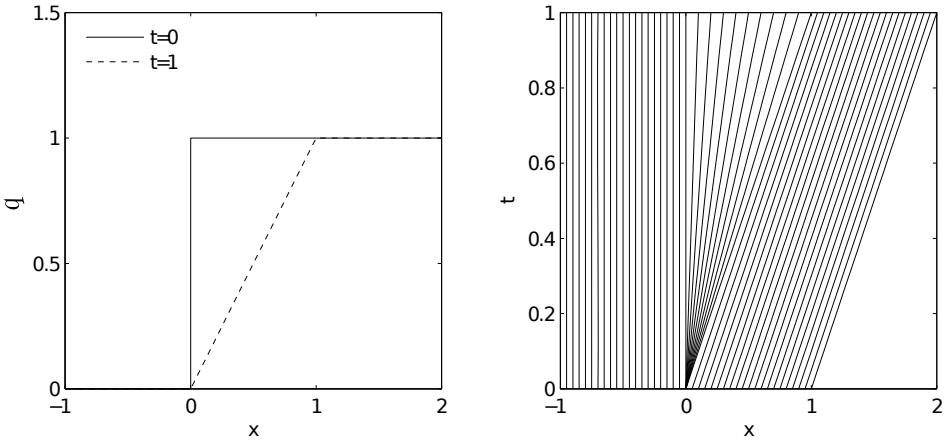


Figure 2.5: Illustration of rarefaction.

The behavior of a system of nonlinear conservation laws is complex. For the special case of a Riemann problem, some properties of the solution structure can be inferred theoretically. In particular, for a system of m nonlinear conservation laws, the Riemann problem will result in (at most) $m + 1$ constant states separated by shocks, rarefaction waves and discontinuities. For a thorough analysis we refer to [10] and [21].

Chapter 3

The finite volume method

Beyond linear and scalar conservation laws, analytic treatment is difficult. To treat nonlinear systems further numerical methods are needed, and in this report we will focus on the *finite volume method* (FVM). In the 1D–FVM we consider a conservation law on a line. We divide this line into $\mathcal{N}$ cells of width Δx, with center position x_i (see Figure 3.1) and cell walls $x_{i\pm1/2}$. We are interested in finding the time evolution of the average amount of $\mathbf{q}$ in each cell during a small time step Δt. This is obtained by integrating (2.1) over time

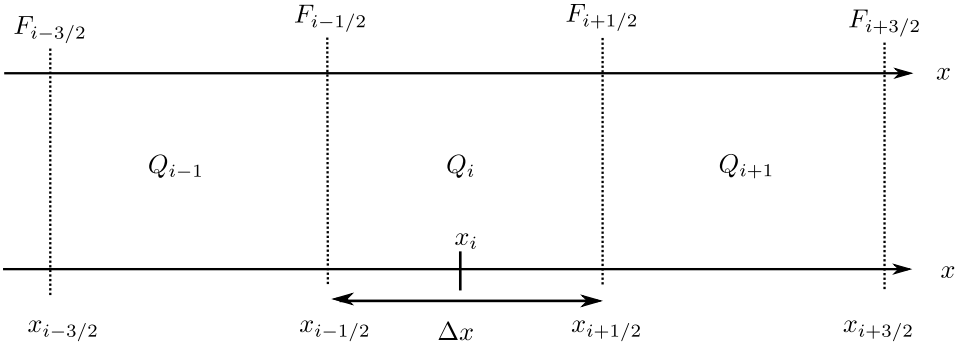


Figure 3.1: Finite volume computational grid.

from t_n to $t_{n+1} = t_n + \Delta t$ as follows

$$\int_{t_n}^{t_n+\Delta t} \mathrm{d}t \left(\int_{x_{i-1/2}}^{x_{i+1/2}} \frac{\partial \mathbf{q}}{\partial t} \mathrm{d}x + \mathbf{f}(\mathbf{q})_{i+1/2} - \mathbf{f}(\mathbf{q})_{i-1/2} \right) = 0.$$

Integrating over time and space we find that

$$\mathbf{Q}_i^{n+1} = \mathbf{Q}_i^n - \lambda \left(\hat{\mathbf{F}}_{i+1/2}^n - \hat{\mathbf{F}}_{i-1/2}^n \right), \tag{3.1}$$

where we have used the following definitions

$$\mathbf{Q}_i^n \equiv \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{q}(x_i, t_n) \mathrm{d}x, \tag{3.2a}$$

$$\hat{\mathbf{F}}_{i\pm1/2}^n \equiv \frac{1}{\Delta t} \int_{t_n}^{t_n+\Delta t} \mathbf{F}(\mathbf{q}(x_{i\pm1/2}, t)) \mathrm{d}t, \tag{3.2b}$$

and $\lambda = \frac{\Delta t}{\Delta x}$. Equation (3.1) tells us that the change in average conserved quantity $\mathbf{Q}_i$ is equal to the average flux through the boundary of the cell from time t_n to $t_n + \Delta t$. The goal of the FVM is to find good approximations to $\hat{\mathbf{F}}_{i\pm1/2}$. There are two challenges to this. Firstly, the fluxes are computed on the edges of the cells and not inside the cells. Secondly, these values change during the time integration.

Before going into specific methods we give some general remarks on the FVM.

Remark 1: Conservative method As we derived the method from the conservation equation the method is conservative. This can be shown by summation over the whole computational domain ($\mathcal{N}$ is the number of cells) as follows

$$\Delta \mathbf{Q}_{\text{tot}} = \sum_{i=1}^{\mathcal{N}} (\mathbf{Q}_i^{n+1} - \mathbf{Q}_i^n) = -\lambda \sum_{i=1}^{\mathcal{N}} (\hat{\mathbf{F}}_{i+1/2}^n - \hat{\mathbf{F}}_{i-1/2}^n) = -\lambda \left(\hat{\mathbf{F}}_{\mathcal{N}+1/2}^n - \hat{\mathbf{F}}_{1/2}^n \right).$$

The change $\Delta \mathbf{Q}_{\text{tot}}$ is equal to the total amount of flux going out at the right boundary minus the flux coming in at the left boundary. Methods where the flux is written in a consistent manner (the flux of left cell for the right cell border equals the flux for the right cell at left cell border) are always conservative.

Remark 2: Discontinuous solutions We could have started from the differential equation formulation and easily derived finite difference methods. However, we saw that a partial differential formulation had problems treating shocks. Only by using the integral formulation, we were able to treat these shocks. In the same way a great advantage of the FVM is that since it is derived from the integral formulation, we are able to describe these discontinuous solutions. Note that how well these are approximated depend on how $\hat{\mathbf{F}}_{i\pm1/2}^n$ is approximated.

Remark 3: Close relationship to the Riemann problem By doing cell averages, we go from a smooth configuration Figure 3.2 (Left) into discrete values shown in Figure 3.2 (Right). Zooming in on two consecutive cells we see that the problem we are dealing with is the Riemann problem. Solving it yields the value of $\mathbf{Q}$ during the time step at the edge and hence the average flux at the cell edge. How we solve or approximate this problem is what distinguishes between the different FVMs.

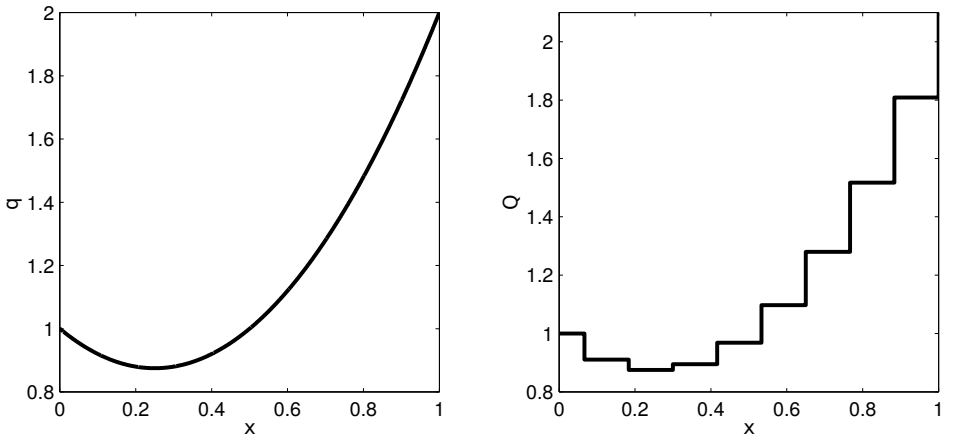


Figure 3.2: (Left) Smooth configuration. (Right) Cell average configuration.

Remark 4: Domain of dependence and domain of influence For hyperbolic conservation laws, a solution or information travels at finite speeds given by the eigenvalues. This means that if points are far enough apart, then they can not interact. For instance for the linear conservation law and the point (x, t) the domain of dependence is given by

$$D = \{x - a^p t, \quad p = 1, \ldots, m \tag{3.3}$$

only these points will determine the solution at (x, t). Similarly, a point at (x, t) has a domain of influence

$$I = \{x + a^p t, \quad p = 1, \ldots, m\}. \tag{3.4}$$

For a nonlinear conservation law, eigenvalues will depend on $\mathbf{Q}$ and the domain of dependence, and influence will be a bounded area rather than a set of points. Most FVMs only consider two cells per edge. For this to be correct one needs to make sure that information does not travel beyond one cell. The simplest example is shown in Figure 3.3 for the advection equation. Here the time step is chosen small enough so that cells only can influence their neighbours. This limitation on the time step is called the CFL-condition. For a system of equations this can be generalized to

$$\frac{\max_p (a^p)\, \Delta t}{\Delta x} < 1, \tag{3.5}$$

which means that the time step needs to be limited so that the distance the fastest information travels is less than the cell width. Note that the CFL–condition is a necessary condition for stability, however it is not sufficient. Individual methods and problems can require a smaller time step than that of the CFL–condition for the FVM to be stable.

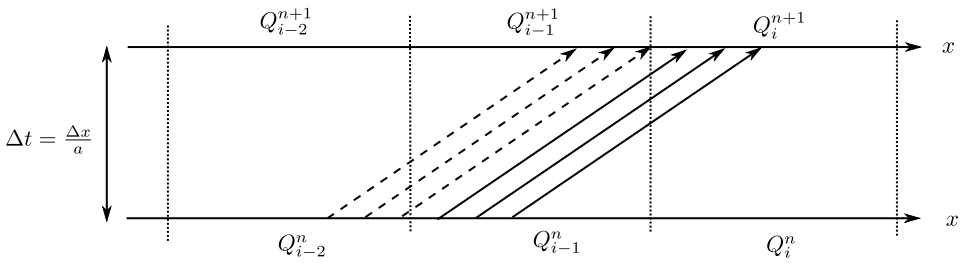


Figure 3.3: Illustration of the CFL–condition.

3.1 The Godunov Method

In the Godunov method [4] we solve the Riemann problem exactly at each edge and use the solution to compute $\hat{\mathbf{F}}(\mathbf{Q}^n_{i\pm 1/2})$. In the Godunov method this value is given by

$$\hat{\mathbf{F}}(\mathbf{Q}) = \mathbf{F}(\mathbf{Q}_\downarrow), \tag{3.6}$$

where $\mathbf{Q}_\downarrow$ is the value of $\mathbf{Q}$ at the position of the left/right edge during the time step. This process is visualized in Figure 3.4. Here we have two Riemann problems between two cells each. In the left figure the solution is a shock while in the right figure the solution is a rarefaction fan. Despite this difference, in both situations during the whole time step, we have that the solution at $x = 0$ is given by $Q_\downarrow = Q_L$. Using this value in the flux we obtain the Godunov method in both cases. For a nonlinear system the solution is obtained in the

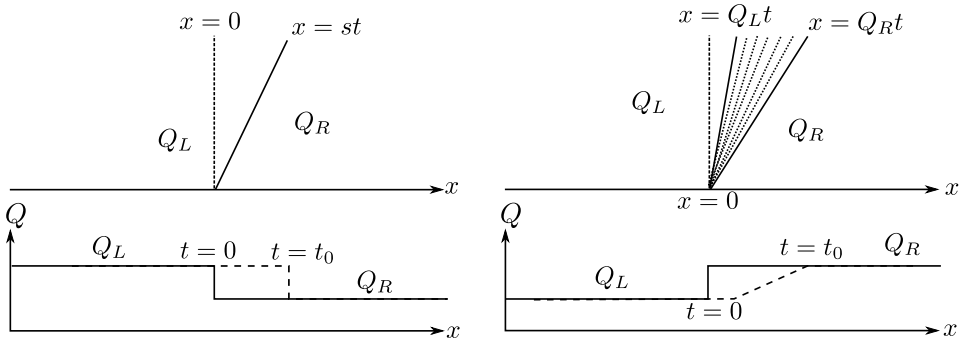


Figure 3.4: Illustration of how FVM works for the Godunov–method for two Riemann problems. (Left) Shock. (Right) Rarefaction.

same way by solving the Riemann problem exactly. If this can not be done analytically, then it is done numerically, for instance by Newton's method. This can be very costly. Another approach is to solve the Riemann problem approximately. Many techniques are based on linearization of the system. That is to assume

$$\mathbf{F}(\mathbf{Q}) \simeq \tilde{\mathbf{A}}(\mathbf{Q})\mathbf{Q}, \tag{3.7}$$

where $\tilde{\mathbf{A}}(\mathbf{Q})$ is a local linearization and is dependent on which approximate Riemann solver is used. The problem is then solved using the Godunov method as if the problem was linear. Motivated by this we seek a method for linear systems, i.e. systems written as

$$\mathbf{F}(\mathbf{Q}) = \mathbf{A}\mathbf{Q}.$$

Since the system is linear and hyperbolic we can decompose

$$\mathbf{Q}_i^n = \sum_{p=1}^{m} \beta_i^p \mathbf{r}^p, \tag{3.8}$$

where $\mathbf{r}^p$ are the eigenvectors of $\mathbf{A}$ and β_i^p their strength. We can write the change between two neighbouring cells

$$\mathbf{Q}_i^n - \mathbf{Q}_{i-1}^n = \sum_{p=1}^{m} \beta_i^p \mathbf{r}^p - \sum_{p=1}^{m} \beta_{i-1}^p \mathbf{r}^p = \sum_{p=1}^{m} \alpha_{i-1/2}^p \mathbf{r}^p = \sum_{p=1}^{m} \mathbf{W}_{i-1/2}^p, \tag{3.9}$$

where

$$\alpha_{i-1/2}^p \equiv \beta_i^p - \beta_{i-1}^p \tag{3.10}$$

and

$$\mathbf{W}_{i-1/2}^p \equiv \alpha_{i-1/2}^p \mathbf{r}^p. \tag{3.11}$$

This means that each discontinuity can be decomposed into waves $\mathbf{W}_{i-1/2}^p$. To each wave there corresponds an a^p – the speed at which the wave travels. How each wave $\mathbf{W}_{i-1/2}^p$ affects neighbouring cells can be understood by interpreting the Godunov method in terms of projections. Suppose we instead had a scalar equation (see Figure 3.5) and that the eigenvalue is given by $a > 0$. The discontinuity wave will move into the right cell and

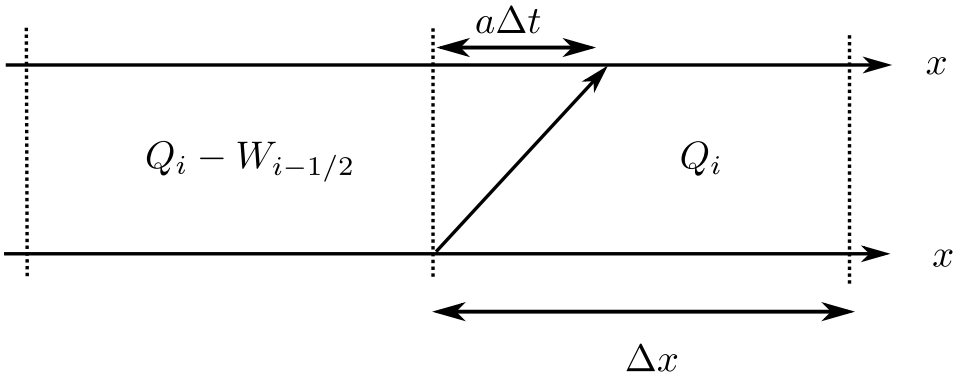


Figure 3.5: Illustration of projection Godunov interpretation.

cover a portion $a\Delta t$ thereby changing it from $Q_i \to Q_i - W_{i-1/2}$. The remaining part of

the cell is unaffected. Averaging or projecting the solution we obtain the following

$$Q_i^{n+1} = Q_i^n \frac{\Delta x - a\Delta t}{\Delta x} + (Q_i^n - W_{i-1/2}) \frac{a\Delta t}{\Delta x},$$

or

$$Q_i^{n+1} = Q_i^n - \frac{a\Delta t}{\Delta x} W_{i-1/2} = 0. \tag{3.12}$$

For $a < 0$, cell i will only be affected by cell $i+1$ and we will have

$$Q_i^{n+1} = Q_i^n - \frac{a\Delta t}{\Delta x} W_{i+1/2} = 0. \tag{3.13}$$

This is easily generalized to any a as follows

$$Q_i^{n+1} - Q_i^n + \frac{\Delta t}{\Delta x}(a^+ W_{i-1/2} + a^- W_{i+1/2}) = 0, \tag{3.14}$$

where $a^+ = \max(a, 0)$ and $a^- = \min(a, 0)$. The same projection-interpretation works for a system of linear conservation laws, however instead of one discontinuity wave there are multiple waves, each with its own advection speed. The generalization is straightforward,

$$\mathbf{Q}_i^{n+1} = \mathbf{Q}_i^n - \lambda \sum_{p=1}^{m} ((a^p)^+ \mathbf{W}_{i-1/2}^p + (a^p)^- \mathbf{W}_{i+1/2}^p) = 0. \tag{3.15}$$

Remark When we say that we solve the Riemann problem exactly using the Godunov method this does not mean that the numerical method is exact. No matter how well the Godunov method solves the Riemann problem, it considers only two cells and can just get first order accuracy. To obtain higher accuracy, extrapolation from next neighbouring cells needs to be used. However, this must be done with care using *limiters*, as straightforward extrapolations will lead to large oscillations near the discontinuities.

3.2 Approximate Riemann solvers

To apply the Godunov method to a system of equations we only need to determine $\mathbf{Q}_\downarrow$ for each of the Riemann problems. In doing so we also compute the whole structure of the Riemann problem, whether or not we have shocks or rarefaction waves. However, in most cases $\mathbf{Q}_\downarrow$ lies on the intermediate states in between the shocks and rarefaction waves. A good example of this is shown in Figure 3.4. Here the value of $\mathbf{Q}_\downarrow$ is independent of whether we have rarefaction or not. Only for transonic rarefaction (a rarefaction fan spanning from negative to positive values) do we need to know the details. Thus in most cases a lot of detailed information is obtained without being used. Approximate Riemann solvers can reduce the number of costly computations and in many cases obtain exactly the same result as Godunov. We will in the following recall some of the most important approximate Riemann solvers.

3.2.1 The Roe method

For a linear system we could easily solve the Riemann problem. In this spirit Roe [18] proposed, rather than approximating the flux, to approximate $\mathbf{A}(\mathbf{q})$ by a constant matrix $\tilde{\mathbf{A}}$, a constant matrix locally in time for each cell edge, and then solve

$$\begin{cases} \mathbf{q}_t + \tilde{\mathbf{A}}\mathbf{q}_x = 0 \\ \mathbf{q}(x,0) = \begin{cases} \mathbf{q}_L & \text{if} \quad x \leq 0 \\ \mathbf{q}_R & \text{if} \quad x \geq 0 \end{cases} \end{cases} \tag{3.16}$$

exactly for each edge. Two possibilities could be $\tilde{\mathbf{A}} = \mathbf{A}(\mathbf{q}_L)$ and $\tilde{\mathbf{A}} = \mathbf{A}(\mathbf{q}_R)$. Roe proposed to use an average or a combination of the two. Further the method requires this matrix to satisfy the following properties:

Property A: Hyperbolicity of the problem. This is equivalent to requiring that $\tilde{\mathbf{A}}$ only has real eigenvalues and that all eigenvectors are linearly independent.

Property B: Consistency with the exact Jacobian

$$\tilde{\mathbf{A}}(\mathbf{q},\mathbf{q}) = \mathbf{A}(\mathbf{q}). \tag{3.17}$$

Property C: Conservation across discontinuities

$$\mathbf{F}(\mathbf{q}_R) - \mathbf{F}(\mathbf{q}_L) = \tilde{\mathbf{A}}(\mathbf{q}_R - \mathbf{q}_L). \tag{3.18}$$

Property (A) ensures that the approximate Riemann problem has the same mathematical character as the original problem. Secondly it also guarantees that we can solve the problem using the wave structure. Property (B) ensures that (3.16) is consistent with the original problem. Property (C) ensures that the method is conservative.

The Roe–matrix is dependent on the specific system of conservation laws and is in general cumbersome to derive. Assuming we have the Roe–matrix, we automatically obtain the eigenstructure. Projecting the discontinuity onto waves $\mathbf{W}^p_{i-1/2}$ going at speed $a^p_{i-1/2}$ we solve it by

$$\mathbf{Q}_i^{n+1} - \mathbf{Q}_i^n + \frac{\Delta t}{\Delta x} \sum_{p=1}^{m} \left(\left(a^p_{i-1/2} \right)^+ \mathbf{W}^p_{i-1/2} + \left(a^p_{i+1/2} \right)^- \mathbf{W}^p_{i+1/2} \right) = 0. \tag{3.19}$$

Note that we now use lower indices on eigenvalues, because unlike for a linear system, the Roe–matrix is different at each edge. For some problems, algebraic relations can be found for the eigenvalues, strengths and eigenvectors. This saves computing time that would otherwise have gone into finding the Roe-matrix.

Chapter 4

Theory of high resolution large time-step schemes

This chapter can roughly be divided into four parts. In the first part we consider first order large time-step schemes (LTSS) for scalar conservation laws. Here we present sufficient conditions for conservative, consistent and total variation diminishing LTSS. Further, we present some LTSS that satisfy these conditions. In the second part we show how and when a first-order LTSS can be generalized to second order away from discontinuities. The third part is devoted to the generalization of the methods of the previous sections to systems of conservation laws. Finally we end this chapter by a discussion on the numerical diffusion introduced by the different LTSS.

4.1 First order LTSS for scalar conservation laws

To make methods that are not limited by the CFL-condition we have to utilize information beyond the nearest cell. We start by considering scalar conservation laws. In the following we will consider explicit $(2N+1)$-point schemes (referred to as wave schemes) of the following form[1]:

$$Q_j^{n+1} = Q_j^n - \sum_{i=0}^{N-1} \left(C^{+i}_{j-1/2-i} \Delta_{j-1/2-i} + C^{-i}_{j+1/2+i} \Delta_{j+1/2+i} \right). \tag{4.1}$$

Here

$$\Delta_{j+1/2} = Q^n_{j+1} - Q^n_j, \tag{4.2}$$

is the change in Q from cell j to cell $j+1$. $C^{\pm i}_{j\mp 1/2\mp i}$ corresponds to the contribution from the Riemann problem at face $j \mp 1/2 \mp i$ to cell j. We will assume that this is only a

[1] The form we use here is different from the one used in [14]. Here the author starts from conservative form with a $2N$-point stencil for the flux, giving an overall method of $(2N+1)$-point stencil.

function of the local CFL-number D at the cell face and is given by

$$C^{\pm i}_{j\mp 1/2\mp i} = C^{\pm i}(D_{j\mp 1/2\mp i}), \tag{4.3a}$$

$$D_{j+1/2} = \lambda \cdot \frac{F^n_{j+1} - F^n_j}{Q^n_{j+1} - Q^n_j}, \tag{4.3b}$$

where $\lambda = \Delta t/\Delta x$.

4.1.1 Conservative wave schemes

A general wave scheme is not necessarily a conservative scheme. We could establish that a specific wave scheme is conservative by writing it in conservative form

$$Q^{n+1}_j = Q^n_j - \lambda \left(\hat{F}^n_{j+1/2} - \hat{F}^n_{j-1/2} \right), \tag{4.4}$$

like is done in [6, 14]. However for the schemes we will consider, the following condition is sufficient.

Proposition 1. *A sufficient condition for a wave scheme to be conservative is*

$$\sum_{i=0}^{N-1} \left(C^{+i}(D) + C^{-i}(D) \right) = D. \tag{4.5}$$

Proof. We prove this by computing the change in the conserved quantity $\mathbf{Q}$ during a single time step

$$\begin{aligned}
\sum_{j=1}^{\mathcal{N}} (Q^{n+1}_j - Q^n_j) &= -\sum_{j=1}^{\mathcal{N}} \sum_{i=0}^{N-1} \left(C^{+i}_{j-1/2-i} \Delta_{j-1/2-i} + C^{-i}_{j+1/2+i} \Delta_{j+1/2+i} \right) \\
&= -\sum_{j=1}^{\mathcal{N}} \left(\sum_{i=0}^{N-1} \left(C^{+i}_{j+1/2} + C^{-i}_{j+1/2} \right) \right) \Delta_{j+1/2} \\
&\overset{(4.5)}{=} -\sum_{j=1}^{\mathcal{N}} D_{j+1/2} \Delta_{j+1/2} = -\lambda \sum_{j=1}^{\mathcal{N}} (F^n_{j+1} - F^n_j) \\
&= \lambda (F^n_1 - F^n_{\mathcal{N}+1}).
\end{aligned}$$

Thus the change in the total amount of Q is given by the amount going in and out at the border and hence the scheme is conservative. ∎

4.1.2 Generalization of Harten's theorem

Proving stability of numerical methods is hard for nonlinear conservation laws. Von Neumann stability analysis [2] can give some information as to when we can expect a method to be stable (often in the form of a CFL-like condition). However this is not sufficient to

guarantee stability. A very strong condition which guarantees stable numerical techniques (see chapter 12.12 in [10]) is to have schemes that are *total variation diminishing* (TVD)

$$\mathrm{TV}\left(\left\{Q^{n+1}\right\}\right) \leq \mathrm{TV}\left(\{Q^n\}\right) \tag{4.6}$$

where

$$\mathrm{TV}\left(\{Q^n\}\right) = \sum_{j=0}^{\mathcal{N}} \left|Q_{j+1}^n - Q_j^n\right|. \tag{4.7}$$

The TVD condition is not only sufficient to guarantee stability, it is stronger as it also guarantees that no spurious oscillations occur (as observed in the Lax-Wendroff scheme). Central in development of TVD-schemes was a theorem proposed in [5] (referred to as Harten's theorem). We present here a generalization of this theorem for wave schemes[2]. The generalization we present to $(2N+1)$-point schemes has with different notation also been given in [6, 9, 14].

Proposition 2. *A wave scheme is TVD whenever the following inequalities are satisfied*

$$C_{j+1/2}^{+(N-1)} \geq 0, \tag{4.8a}$$

$$C_{j+1/2}^{+i} \geq C_{j+1/2}^{+(i+1)}, \tag{4.8b}$$

$$1 - C_{j+1/2}^{+0} + C_{j+1/2}^{-0} \geq 0, \tag{4.8c}$$

$$C_{j+1/2}^{-(i+1)} \geq C_{j+1/2}^{-i}, \tag{4.8d}$$

$$0 \geq C_{j+1/2}^{-(N-1)}. \tag{4.8e}$$

Proof. Using (4.1) we find that

$$\Delta_{j+1/2}^{n+1} = \Delta_{j+1/2} - \sum_{i=0}^{N-1} \Big(C_{j+1/2-i}^{+i} \Delta_{j+1/2-i} + C_{j+3/2+i}^{-i} \Delta_{j+3/2+i} - C_{j-1/2-i}^{+i} \Delta_{j-1/2-i} - C_{j+1/2+i}^{-i} \Delta_{j+1/2+i} \Big). \tag{4.9}$$

Writing out the sum gives the following expression

$$\begin{aligned} \Delta_{j+1/2}^{n+1} = {} & C_{j-N+1/2}^{+(N-1)} \Delta_{j-N+1/2} + \cdots + \left(C_{j-i+1/2}^{+(i-1)} - C_{j-i+1/2}^{+i} \right) \Delta_{j-i+1/2} \\ & + \cdots + \left(1 - C_{j+1/2}^{+0} + C_{j+1/2}^{-0} \right) \Delta_{j+1/2} + \cdots + \\ & \left(C_{j+i+3/2}^{-(i+1)} - C_{j+i+3/2}^{-i} \right) \Delta_{j+i+3/2} + \cdots - C_{j+N+1/2}^{-(N-1)} \Delta_{j+N+1/2}. \end{aligned} \tag{4.10}$$

Next we use (4.10) to compute the total variation

$$\mathrm{TV}\left(\left\{Q^{n+1}\right\}\right) = \sum_{j=1}^{\mathcal{N}} \left| C_{j-N+1/2}^{+(N-1)} \Delta_{j-N+1/2} + \cdots + \right.$$

[2] Harten's original theorem is formulated for methods in the form of a wave scheme with $N = 1$.

$$
\begin{aligned}
&\left(C^{+(i-1)}_{j-i+1/2} - C^{+i}_{j-i+1/2}\right)\Delta_{j-i+1/2} + \\
&\cdots + \left(1 - C^{+0}_{j+1/2} + C^{-0}_{j+1/2}\right)\Delta_{j+1/2} + \cdots + \\
&\left(C^{-(i+1)}_{j+i+3/2} - C^{-i}_{j+i+3/2}\right)\Delta_{j+i+3/2} + \cdots - C^{-(N-1)}_{j+N+1/2}\Delta_{j+N+1/2}\Big| \\
\leq &\sum_{j=1}^{\mathcal{N}} \Bigg(\left|C^{+(N-1)}_{j-N+1/2}\Delta_{j-N+1/2}\right| + \cdots + \\
&\left|\left(C^{+(i-1)}_{j-i+1/2} - C^{+i}_{j-i+1/2}\right)\Delta_{j-i+1/2}\right| + \\
&\cdots + \left|\left(1 - C^{+0}_{j+1/2} + C^{-0}_{j+1/2}\right)\Delta_{j+1/2}\right| + \cdots + \\
&\left|\left(C^{-(i+1)}_{j+i+3/2} - C^{-i}_{j+i+3/2}\right)\Delta_{j+i+3/2}\right| + \cdots + \\
&\left|-C^{-(N-1)}_{j+N+1/2}\Delta_{j+N+1/2}\right| \Bigg) \\
= &\sum_{j=1}^{\mathcal{N}} \left|C^{+(N-1)}_{j+1/2}\Delta_{j+1/2}\right| + \cdots + \left|\left(C^{+(i-1)}_{j+1/2} - C^{+i}_{j+1/2}\right)\Delta_{j+1/2}\right| + \\
&\cdots + \left|\left(1 - C^{+0}_{j+1/2} + C^{-0}_{j+1/2}\right)\Delta_{j+1/2}\right| + \cdots + \\
&\left|\left(C^{-(i+1)}_{j+1/2} - C^{-i}_{j+1/2}\right)\Delta_{j+1/2}\right| + \cdots + \left|-C^{-(N-1)}_{j+1/2}\Delta_{j+1/2}\right|\Big) \\
= &\sum_{j=1}^{\mathcal{N}} \Big(\left|C^{+(N-1)}_{j+1/2}\right| + \cdots + \left|C^{+(i-1)}_{j+1/2} - C^{+i}_{j+1/2}\right| + \cdots + \\
&\left|1 - C^{+0}_{j+1/2} + C^{-0}_{j+1/2}\right| + \cdots + \left|C^{-(i+1)}_{j+1/2} - C^{-i}_{j+1/2}\right| + \cdots \\
&+ \left|-C^{-(N-1)}_{j+1/2}\right|\Big)\left|\Delta_{j+1/2}\right|.
\end{aligned}
$$

When all expressions inside the absolute signs are positive they add up to 1 and we have (4.6). This is true when (4.8) is satisfied. ■

4.1.3 Modified equation

When using FVM, errors are introduced. Typically these errors will lead to some diffusion not present in the original partial differential equation that will smear out the solution. In the modified equation approach we ask: is there another partial differential equation that the numerical method approximates better than the original? In this section we find the modified equation for a general LTSS.

Proposition 3. *A wave scheme is consistent when*[3]

$$
\sum_{i=0}^{N-1} \left(C^{+i}(D) + C^{-i}(D)\right) = D, \tag{4.11}
$$

[3] A modified equation for a general large time-step scheme in conservative form was derived in [14].

and the modified equation for a wave scheme is given by

$$Q_t + (f(Q) - g(Q))_x = \mathcal{O}\left(\Delta x^2\right), \tag{4.12}$$

where

$$g(Q) = \frac{\Delta x}{\lambda}\sigma(D)Q_x. \tag{4.13}$$

with diffusion coefficient

$$\sigma(D) = \frac{1}{2}\sum_{i=0}^{N-1}(2i+1)\left(C^{+i} - C^{-i}\right) - \frac{1}{2}D^2. \tag{4.14}$$

Proof. To find the modified equation we start by considering a general wave scheme and Taylor expand (4.1) around $[x = x_j, t = t_n]$ (when the index is omitted it is implied that $x = x_j$ and $t = t_n$). To do so we will use the following:

$$Q_j^{n+1} = Q(x, t+\Delta t) = Q + \Delta t Q_t + \frac{\Delta t^2}{2}Q_{tt} + \mathcal{O}\left(\Delta t^3\right), \tag{4.15a}$$

$$\Delta_{j-1/2-i} = \Delta x Q_x - \frac{\Delta x^2}{2}\left(2i+1\right)Q_{xx} + \mathcal{O}\left(\Delta x^3\right), \tag{4.15b}$$

$$\Delta_{j+1/2+i} = \Delta x Q_x + \frac{\Delta x^2}{2}(2i+1)Q_{xx} + \mathcal{O}\left(\Delta x^3\right), \tag{4.15c}$$

$$C^{+i}\left(D_{j-i-1/2}\right) = C^{+i} - C_D^{+i}\Delta x\left(i+\frac{1}{2}\right)D_x + \mathcal{O}(\Delta x^2), \tag{4.15d}$$

$$C^{-i}\left(D_{j+i+1/2}\right) = C^{-i} + C_D^{-i}\Delta x\left(i+\frac{1}{2}\right)D_x + \mathcal{O}(\Delta x^2). \tag{4.15e}$$

Combining relations (4.15) we find that

$$C^{+i}_{j-1/2-i}\Delta_{j-1/2-i} = \Delta x C^{+i}Q_x - \frac{\Delta x^2}{2}(2i+1)\partial_x\left(C^{+i}Q_x\right) + \mathcal{O}\left(\Delta x^3\right), \tag{4.16a}$$

$$C^{-i}_{j+1/2+i}\Delta_{j+1/2+i} = \Delta x C^{-i}Q_x + \frac{\Delta x^2}{2}(2i+1)\partial_x\left(C^{-i}Q_x\right) + \mathcal{O}\left(\Delta x^3\right). \tag{4.16b}$$

Insertion of relations (4.15a) and (4.16) into (4.1) results in, to second order,

$$Q_t + \frac{1}{\lambda}\sum_{i=0}^{N-1}\left(C^{+i} + C^{-i}\right)Q_x = \frac{\Delta x}{2\lambda}\sum_{i=0}^{N-1}(2i+1)\partial_x\left(\left(C^{+i} - C^{-i}\right)Q_x\right) - \frac{\Delta t}{2}Q_{tt}. \tag{4.17}$$

Next we replace Q_{tt} by derivatives[4] with respect to x. We do this by taking the t–derivative of (4.17) to first order as follows

$$Q_{tt} = -\frac{1}{\lambda}\sum_{i=0}^{N-1}\left(\left(C^{+i} + C^{-i}\right)Q_{xt} + \left(C_D^{+i} + C_D^{-i}\right)D_t Q_x\right). \tag{4.18}$$

[4] An alternative way of doing this is by using the original partial differential equation. In this case we would get the same result. However, if we were to compute higher order corrections this is not necessarily true.

Next we note that

$$D_t = -\frac{1}{\lambda}\sum_{i=0}^{N-1}\left(C^{+i}+C^{-i}\right)\partial_x\left(D\right)+\mathcal{O}(\Delta x) \tag{4.19a}$$

$$Q_{xt} = -\frac{1}{\lambda}\sum_{i=0}^{N-1}\partial_x\left(\left(C^{+i}+C^{-i}\right)Q_x\right) \tag{4.19b}$$

Inserting (4.19) into (4.18) we find that

$$Q_{tt} = \frac{1}{\lambda^2}\sum_{i=0}^{N-1}\partial_x\left(\left(C^{+i}+C^{-i}\right)\left(\sum_{k=0}^{N-1}\left(C^{+k}+C^{-k}\right)Q_x\right)\right). \tag{4.20}$$

Replacing Q_{tt} by (4.20) in (4.17) we obtain the modified equation:

$$\begin{aligned} Q_t + \frac{1}{\lambda}\sum_{i=0}^{N-1}\left(C^{+i}+C^{-i}\right)Q_x = \frac{\Delta x}{2\lambda}\Bigg(&\sum_{i=0}^{N-1}(2i+1)\partial_x\left(\left(C^{+i}+C^{-i}\right)Q_x\right)- \\ &\partial_x\sum_{i=0}^{N-1}\left(\left(C^{+i}+C^{-i}\right)\left(\sum_{k=0}^{N-1}\left(C^{+k}+C^{-k}\right)Q_x\right)\right)\Bigg). \end{aligned} \tag{4.21}$$

To obtain a consistent method we require (4.11). Using this requirement we can write the modified equation as

$$Q_t + a(Q)Q_x = \frac{\Delta x}{2\lambda}\partial_x\left(\left(\sum_{i=0}^{N-1}(2i+1)\left(C^{+i}-C^{-i}\right)-D^2\right)Q_x\right), \tag{4.22}$$

or (4.12) with definition (4.13) and (4.14). ■

4.1.4 First order large-time step schemes

In the previous sections we derived conditions on the coefficients for a conservative, TVD and consistent LTSS. Further, we derived the diffusion coefficient $\sigma(D)$ of a general wave scheme. This coefficient tells us how much smearing one might expect from a LTSS. Thus we have a recipe for making LTSS. In this section we will consider some LTSS.

Least diffusive TVD scheme: LTS-Roe

The least diffusive scheme is obtained by having $\pm C^{\pm i}$ as big as possible for the lower indices and as low as possible for the higher indices. In fact the least diffusive scheme (that is still TVD) has all $C^{+i}=1$ and $C^{-i}=0$ or $C^{+i}=0$, and $C^{-i}=-1$, except for $C^{\pm(N-1)}$ which contains the remaining part needed for consistency. A compact way of writing this is

$$C^{+i}(D) = \max(0,\min(D-i,1)) \tag{4.23a}$$

$$C^{-i}(D) = \min(0,\max(D+i,-1)) \tag{4.23b}$$

Next we compute $\sigma^{\text{LTS-Roe}}$. Assuming $N-1 < D < N$, we find that

$$\begin{aligned}\sum_{i=0}^{N-1} i\left(C^{+i} - C^{-i}\right) &= \sum_{i=0}^{N-1} iC^{+i} = \sum_{i=0}^{N-1} i\max\left(0, \min\left(D-i, 1\right)\right) \\ &= \sum_{i=0}^{N-2} i + (N-1)\left(D-(N-1)\right) \\ &= \frac{(N-2)(N-1)}{2} + (N-1)\left(D-(N-1)\right) \\ &= -\frac{1}{2}N^2 + \frac{1}{2}N + (N-1)D\end{aligned}$$

and for $N-1 < -D < N$

$$\sum_{i=0}^{N-1} i\left(C^{+i} - C^{-i}\right) = -\frac{1}{2}N^2 + \frac{1}{2}N - (N-1)D,$$

or simply

$$\sum_{i=0}^{N-1} i\left(C^{+i} - C^{-i}\right) = -\frac{1}{2}N^2 + \frac{1}{2}N - (N-1)\left|D\right|. \tag{4.24}$$

Inserting into (4.14) we find the diffusion coefficient

$$\sigma^{\text{LTS-Roe}}(D) = \frac{1}{2}(|D| - (N-1))(N - |D|). \tag{4.25}$$

LTS-Lax-Friedrichs TVD scheme

The most diffusive scheme is obtained by having all $\pm C^{\pm i} = \pm C^{\pm(i+1)}$ and as large as possible (still satisfying Harten's generalized theorem). The following scheme is the most diffusive

$$C^{\pm i}(D) = \frac{1}{2N}(D \pm N). \tag{4.26}$$

Here we need to be careful in what we mean by N. If $N = \text{ceil}(\max_j(D))$ we have LTS-LF. However if $N = \text{ceil}(D)$, we have a local large time-step Lax-Friedrichs (Local-LTS-LF). Both methods are perfectly fine, however Local-LTS-LF has less diffusion. To find the modified equation of LTS-LF we compute

$$\begin{aligned}\sum_{i=0}^{N-1}(2i+1)\left(C^{+i} - C^{-i}\right) &= \sum_{i=0}^{N-1}(2i+1) \\ &= (N-1)N + N = N^2.\end{aligned}$$

Thus the modified flux is

$$\sigma^{\text{LTS-LF}}(D) = \frac{1}{2}(N^2 - D^2)Q_x. \tag{4.27}$$

Hybrid

In some cases a wrong solution is obtained by LTS-Roe1 because it looks upon everything as shocks and ignores rarefaction waves. Lax-Friedrichs on the other hand introduces viscosity and the correct solution is obtained. A special case where entropy mistakes are made is for $D \simeq 0$. All our conditions are such that if two schemes are TVD and consistent, then this is also true for a combination of the two. We can thus make the following hybrid scheme (Hybrid)

$$C^{\pm i}(D) = \begin{cases} \alpha\left(\frac{D}{\epsilon}\right) C^{\pm i}_{\text{LTS-Roe}}(D) + \left(1 - \alpha\left(\frac{D}{\epsilon}\right)\right) C^{\pm i}_{\text{LTS-LF}}(D), & \text{if } |D| \leq \epsilon, \\ C^{\pm i}_{\text{LTS-Roe}}, & \text{if } \quad \epsilon < |D|, \end{cases} \tag{4.28}$$

where $0 \leq \alpha(\frac{D}{\epsilon}) \leq 1$. There is no a priori choice for $\alpha(x)$, however

$$\alpha(x) = \sin^2\left(\frac{\pi}{2}x\right) \tag{4.29}$$

has some nice features. When $D \simeq 0$, only Local-LF is used and when $D \simeq \epsilon$ only LTS-Roe is used. Further, it has continuous derivatives of $C^{\pm i}$ at $D \simeq \epsilon$. The diffusion coefficient is simply given by the sum

$$\sigma(D) = \begin{cases} \alpha\left(\frac{D}{\epsilon}\right) \sigma^{\text{LTS-Roe}}(D) + \left(1 - \alpha\left(\frac{D}{\epsilon}\right)\right) \sigma^{\text{LTS-LF}}(D), & \text{if } |D| \leq \epsilon, \\ \sigma^{\text{LTS-Roe}}(D), & \text{if } \quad \epsilon < |D|. \end{cases} \tag{4.30}$$

4.2 High resolution large time-step schemes

In the modified equation

$$Q_t + (f(Q) - g(Q))_x = \mathcal{O}\left(\Delta x^2\right) \tag{4.31}$$

the extra $g(q)$-term makes it so that the numerical method only approximates the conservation law to first order. In order to increase the accuracy to second order, this term needs to be removed. To see how this can be done, we consider instead the following equation

$$q_t + (f(q) + g(q))_x = \mathcal{O}(\Delta x^2). \tag{4.32}$$

If a numerical method approximates the $f(q)$ in (4.32) as before and $g(q)$ to second order accuracy, the second term will cancel out $g(Q)$ in (4.31), giving a second order accurate method. There are infinitely many ways of doing this. A clever way was suggested by Harten in [5] and later in [6]. Instead of approximating the two terms separately, we combine the two terms into one term

$$f^M(q) = f(q) + g(q), \tag{4.33}$$

and apply the old method on $f^M(q)$. Doing so results in the following modified equation

$$q_t + f(q)_x = \mathcal{O}(\Delta x^2), \tag{4.34}$$

thus achieving second order.

We will only focus on wave schemes. To use such schemes we need to compute the local CFL-numbers D at the cell faces. Because we have a modified flux, these values will be altered as follows

$$D^M_{j+1/2} = \lambda \cdot \frac{F^M_{j+1} - F^M_j}{Q_{j+1} - Q_j} = \lambda \cdot \frac{F_{j+1} - F_j}{Q_{j+1} - Q_j} + \lambda \cdot \frac{G_{j+1} - G_j}{Q_{j+1} - Q_j} = D_{j+1/2} + \gamma_{j+1/2}, \tag{4.35}$$

where G_j is a discretization of $g(Q(x_j))$ and

$$\gamma_{j+1/2} \equiv \lambda \cdot \frac{G_{j+1} - G_j}{Q_{j+1} - Q_j}. \tag{4.36}$$

Thus we have two contributions to the local wave speed. From this a problem arises. How should we discretize the G_j when they depend on derivatives of Q and how do we keep them from diverging at discontinuities?

4.2.1 Smoothing of modified flux

We use the same smoothing method as in [5, 7]

$$G_j = s_{1+1/2} \cdot \max\left[0, \min\left(\left|\tilde{G}_{j+1/2}\right|, s_{1+1/2} \cdot \tilde{G}_{j-1/2}\right)\right], \tag{4.37}$$

where

$$\tilde{G}_{j+1/2} = \frac{1}{\lambda}\sigma(D_{j+1/2})\Delta_{j+1/2} \tag{4.38}$$

and $s_{1+1/2} \equiv \operatorname{sgn}\left(G_{j+1/2}\right)$. How the smoothing works is best understood from

$$G_j = \begin{cases} s_{j+1/2} \min\left(\left|\tilde{G}_{j+1/2}\right|, \left|\tilde{G}_{j-1/2}\right|\right), & \text{if } \tilde{G}_{j+1/2}\tilde{G}_{j-1/2} \geq 0 \\ 0 & \text{if } \tilde{G}_{j+1/2}\tilde{G}_{j-1/2} < 0 \end{cases}. \tag{4.39}$$

When the two are of equal sign, the smallest correction is chosen. When they are of opposite sign, neither is chosen. This way, corrections are limited and kept from being used near extrema.

Before using this smoothing process we need to show that it is sufficiently accurate to obtain second order. In particular, to prove second order accuracy we will use the following relations.

Proposition 4. *When the smoothing procedure in* (4.37) *is used, the following relations hold true:*

$$\frac{G_{j+1} + G_j}{2} = G(Q(x_{j+1/2})) + \mathcal{O}\left(\Delta x^2\right) = \tilde{G}_{j+1/2} + \mathcal{O}\left(\Delta x^2\right), \tag{4.40a}$$

$$\gamma_{j+1/2}\Delta_{j+1/2} = G_{j+1} - G_j = \mathcal{O}\left(\Delta x^2\right). \tag{4.40b}$$

Proof. To show that (4.40a) holds true we assume $g(q)$ to be a continuous function. First we examine the case of $\tilde{G}_{j+1/2}\tilde{G}_{j-1/2} > 0$, thus we can write

$$G_j = \frac{1}{2}\left[\tilde{G}_{j-1/2} + \tilde{G}_{j+1/2} - s_{j+1/2}\left|\tilde{G}_{j+1/2} - \tilde{G}_{j+1/2}\right|\right] \tag{4.41}$$

$$= \tilde{G}_{j\pm1/2} + \frac{1}{2}\left[\mp(\tilde{G}_{j+1/2} - \tilde{G}_{j-1/2}) - s_{j+1/2}\left|\tilde{G}_{j+1/2} - \tilde{G}_{j-1/2}\right|\right]. \tag{4.42}$$

We note that $\tilde{G}_{j+1/2}$ is a continuous function and that $\tilde{G}_{j+1/2} = \mathcal{O}(\Delta x)$, and it is a continuous function. Thus we have that

$$\tilde{G}_{j+1/2} - \tilde{G}_{j-1/2} = \mathcal{O}\left(\Delta x^2\right) \tag{4.43}$$

which implies that

$$G_j = \tilde{G}_{j\pm1/2} + \mathcal{O}\left(\Delta x^2\right). \tag{4.44}$$

If $\tilde{G}_{j+1/2}\tilde{G}_{j-1/2} > 0$ we have a change of sign in $\Delta_{j+1/2}$ and hence we are either at a maximum or minimum. Assuming continuous derivatives we have $\Delta_{j\pm1/2} = \mathcal{O}\left(\Delta x^2\right)$ and therefore $G_{j\pm1/2} = \mathcal{O}\left(\Delta x^2\right)$. Hence in both cases we have

$$G_i = \tilde{G}_{j\pm1/2} + \mathcal{O}\left(\Delta x^2\right). \tag{4.45}$$

From this relation, relations (4.40) follow immediately. ■

4.2.2 Proof of second order accuracy

We now turn to proving that Harten's prescription gives a second order accurate method[5]. We will use the following relations:

$$\frac{G_{j+1} + G_j}{2} = \tilde{G}_{j+1/2} + \mathcal{O}\left(\Delta x^2\right), \tag{4.46a}$$

$$\Delta x C^{\pm i}(a_{j+1/2} + \gamma_{j+1/2})\Delta_{j+1/2} = \Delta x C^{\pm i}(a_{j+1/2})\Delta_{j+1/2} + \mathcal{O}(\Delta x^3), \tag{4.46b}$$

$$\Delta x C^{\pm i}(D_{j+1/2})\Delta_{j+1/2} = \frac{\Delta x}{2}\big(C^{\pm i}(D_{j-1/2})\Delta_{j-1/2} + C^{\pm i}(D_{j+1/2})\Delta_{j+1/2}\big) + \mathcal{O}\left(\Delta x^3\right). \tag{4.46c}$$

The first relation we have already proven. The second relation is a consequence of $\gamma_{j+1/2}$ being of $\mathcal{O}(\Delta x)$. The third is a consequence of $C^{\pm i}(D)$ being a continuous function of D.

Proposition 5. *The modified flux approach with the smoothing procedure in* (4.37) *gives a second order accurate solution away from discontinuities.*

[5] Our proof is similar to the one presented in [6]. However, Harten considered a specific scheme, while we consider all possible wave schemes.

Proof.

$$
\begin{aligned}
Q_j^{n+1} &= Q_j^n - \sum_{i=0}^{N-1} \left(\left[C^{+i}(D^M)\Delta\right]_{j-1/2-i} + \left[C^{-i}(D^M)\Delta\right]_{j+1/2+i} \right) \\
&= Q_j^n - \frac{1}{2}\sum_{i=0}^{N-1} \left[C^{+i}(D^M)\Delta - i\Delta x\partial_x \left(C^{+i}(D^M)\Delta\right) + \right. \\
&\quad \left. C^{-i}(D^M)\Delta + (i+1)\Delta x\partial_x \left(C^{-i}(D^M)\Delta\right)\right]_{j-1/2} - \\
&\quad \frac{1}{2}\sum_{i=0}^{N-1} \left[C^{+i}(D^M)\Delta - (i+1)\Delta x\partial_x \left(C^{+i}(D^M)\Delta\right) + \right. \\
&\quad \left. C^{-i}(D^M)\Delta + i\Delta x\partial_x C^{-i}(D^M)\Delta\right]_{j+1/2} + \mathcal{O}\left(\Delta x^3\right) \\
&\overset{(4.46c)}{=} Q_j^n - \frac{1}{2}\sum_{i=0}^{N-1} \left(\left(C^{+i}(D^M_{j+1/2}) + C^{-i}(D^M_{j+1/2})\right)\Delta_{j+1/2} + \right. \\
&\quad \left. \left(C^{+i}(D^M_{j-1/2}) + C^{-i}(D^M_{j-1/2})\right)\Delta_{j-1/2}\right) - \\
&\quad \frac{\Delta x}{2}\partial_x \sum_{i=0}^{N-1}(2i+1)\frac{1}{2}\left(\left(C^{-i}(D^M_{j+1/2}) - C^{+i}(D^M_{j+1/2})\right)\Delta_{j+1/2} + \right. \\
&\quad \left. \left(C^{+i}(D^M_{j-1/2}) - C^{+i}(D^M_{j-1/2})\right)\Delta_{j-1/2}\right) \\
&\overset{(4.46b)}{=} Q_j^n - \frac{\Delta t}{2\Delta x}\left(a_{j-1/2}\Delta_{j-1/2} + a_{j+1/2}\Delta_{j+1/2}\right) - \\
&\quad \frac{\Delta t}{2\Delta x}\left(\gamma_{j-1/2}\Delta_{j-1/2} + \gamma_{j+1/2}\Delta_{j+1/2}\right) - \\
&\quad \frac{\Delta x}{2}\partial_x \sum_{i=0}^{N-1}(2i+1)\frac{1}{2}\left(\left(C^{-i}(a_{j+1/2}) - C^{+i}(a_{j+1/2})\right)\Delta_{j+1/2} + \right. \\
&\quad \left. \left(C^{-i}(a_{j-1/2}) - C^{+i}(a_{j-1/2})\right)\Delta_{j-1/2}\right) \\
&= Q_j^n - \Delta x\left[a(Q)Q_x\right]_{x=x_j} - \frac{\Delta t}{2}\left(G_{j+1} - G_j + G_j - G_{j-1}\right) \\
&\quad - \frac{\Delta x^2}{2}\partial_x \sum_{i=0}^{N-1}(2i+1)\left[\left(C^{-i}(a(Q)) - C^{+i}(a(Q))\right)Q_x\right]_{x=x_j} \\
&\overset{(4.46a)}{=} Q_j^n - \Delta t\left[a(Q)Q_x\right]_{x=x_j} + \frac{\Delta t^2}{2}\partial_x\left[a(Q)^2 Q_x\right]_{x=x_j} + \\
&\quad \frac{\Delta x^2}{2}\partial_x\left[\sum_{i=0}^{N}(2i+1)\left(C^{+i}(a(Q)) - C^{-i}(a(Q))\right)Q_x\right]_{x=x_j} - \\
&\quad \frac{\Delta x^2}{2}\partial_x \sum_{i=0}^{N-1}(2i+1)\left[\left(C^{-i}(a(Q)) - C^{+i}(a(Q))\right)Q_x\right]_{x=x_j}
\end{aligned}
$$

$$= Q(x_j, t) + \Delta t Q_t(x_j, t) + \frac{\Delta t^2}{2} Q_{tt}(x_j, t) + \mathcal{O}\left(\Delta x^3\right)$$
$$= Q(x_j, t + \Delta t) + \mathcal{O}\left(\Delta x^3\right)$$

■

4.2.3 Supplementary condition for TVD

In the previous sections we showed that a first order method can be made second order by using a modified flux. By design the method is TVD whenever

$$\left|D^M\right| \leq N, \tag{4.47}$$

or

$$|D + \gamma| \leq N. \tag{4.48}$$

γ will in general be dependent on N, and thus we cannot choose N to be sufficiently big. Instead we find a sufficient condition for the method to be TVD.

Proposition 6. *When smoothing in* (4.37) *is used, the following equation holds true*

$$\left|\gamma_{j+1/2}\right| = \frac{|G_{j+1} - G_j|}{\left|\Delta_{j+1/2}\right|} \leq \sigma(D_{j+1/2}). \tag{4.49}$$

Proof. We show this as follows (G_j and G_{j+1} must be of same sign)

$$|G_{j+1} - G_j| \leq \max\left(|G_j|, |G_{j+1}|\right) \tag{4.50}$$
$$\leq \max\left[\min\left(\left|\tilde{G}_{j-1/2}\right|, \left|\tilde{G}_{j+1/2}\right|\right), \min\left(\left|\tilde{G}_{j+1/2}\right|, \left|\tilde{G}_{j+3/2}\right|\right)\right] \tag{4.51}$$
$$\leq \left|\tilde{G}_{j+1/2}\right| \tag{4.52}$$

which directly implies (4.49). ■

Proposition 7. *A sufficient condition for the method to be TVD is*

$$|D| + \sigma(D) \leq N. \tag{4.53}$$

Proof. We have that

$$\left|D_{j+1/2} + \gamma_{j+1/2}\right| \leq \left|D_{j+1/2}\right| + \left|\gamma_{j+1/2}\right| \leq N.$$

Next, using (4.49) we find that

$$\left|D_{j+1/2}\right| + \sigma(D_{j+1/2}) \leq N. \tag{4.54}$$

This has to hold for all $D_{j+1/2}$ and thus (4.53). ■

General wave scheme

For a general wave scheme we require that

$$|D| + \frac{1}{2}\sum_{i=0}^{N-1}(2i+1)(C^{+i} - C^{-i}) - \frac{1}{2}D^2 \leq N, \tag{4.55}$$

or

$$\sum_{i=0}^{N-1}(2i+1)(C^{+i} - C^{-i}) \leq 2(N - |D|) + D^2. \tag{4.56}$$

Thus there is a limit on the coefficients to how big they can be and still yield a method that is TVD, independent of Harten's generalized theorem.

High resolution LTS-Roe

For LTS-Roe, $N = \text{ceil}(|D|)$, thus we can write $D = N - \alpha$, where $0 \leq \alpha \leq 1$. Using this notation we find that (assuming $D > 0$)

$$D + \sigma^{\text{Roe}}(D) = N - \alpha + \frac{1}{2}\alpha(1-\alpha) = N - \frac{\alpha(1+\alpha)}{2} \leq N,$$

and thus the method is TVD whenever $D \leq N$.

High resolution LTS-LF

Again we assume $N = \text{ceil}(|D|)$ and $0 < D$. Then we can write

$$\begin{aligned} D + \sigma^{\text{LF}}(D) &= N - \alpha + \frac{1}{2}\left(N^2 - (N-\alpha)^2\right) \\ &= N - \alpha + \frac{1}{2}\left(N^2 - N^2 + 2N\alpha - \alpha^2\right) \\ &= N - \alpha + N\alpha - \frac{1}{2}\alpha^2, \end{aligned}$$

which is only less than N for $N = 1$. Thus a high resolution scheme cannot be made for LTS-LF or local-LTS-LF. However, it can be used in Hybrid as long as $\epsilon < 1$.

4.2.4 Procedure for second order LTSS

In summary, if a wave scheme satisfies conditions for consistency, conservation, Harten's generalized theorem and (4.55), then a high-resolution version is given by

$$Q_j^{n+1} = Q_j^n - \sum_{i=0}^{N-1}\Big(C^{+i}_{j-i-1/2}(D+\gamma)\Delta_{j-i-1/2} + C^{-i}_{j+i+1/2}(D+\gamma)\Delta_{j+i+1/2}\Big). \tag{4.57}$$

Here $D_{j+1/2}$ is as before, and

$$\gamma_{j+1/2} \equiv \lambda \cdot \frac{G_{j+1} - G_j}{Q_{j+1} - Q_j}. \tag{4.58}$$

where

$$G_j = s_{1+1/2} \cdot \max\left[0, \min\left(\left|\tilde{G}_{j+1/2}\right|, s_{1+1/2} \cdot \tilde{G}_{j-1/2}\right)\right], \tag{4.59}$$

$$\tilde{G}_{j+1/2} = \frac{1}{\lambda}\sigma(D_{j+1/2})\Delta_{j+1/2}, \tag{4.60}$$

and $s_{1+1/2} \equiv \text{sgn}\left(G_{j+1/2}\right)$.

4.3 Generalization to systems

In this section we describe how the scalar LTSS can be generalized to a system of conservation laws. Our extension is based on the Roe-linearization in [19] and the extension done by Harten in [6].

The idea of the generalization is to linearize the conservation equation and then decompose the problem onto eigenvectors $\mathbf{r}^p_{j+1/2}$ with corresponding strength $\alpha^p_{j+1/2}$ and component dependent local CFL-number $D^p_{j+1/2}$. Next we apply the LTSS to each component. As we are using the Roe-linearization, special care is needed to obtain the correct scheme. We start from the same second order conservative form as in [6]

$$\begin{aligned}\lambda\hat{\mathbf{F}}_{j+1/2} = \frac{\lambda}{2}(\mathbf{F}_{j+1} + \mathbf{F}_j) + \sum_{p=1}^{m}\Bigg[\frac{\lambda}{2}\mathbf{r}^p_{j+1/2}\left(G^p_{j+1} + G^p_j\right) - \\ \sum_{i=-N+1}^{N-1}\left[K_l(D+\gamma)\alpha\mathbf{r}\right]^p_{j+i+1/2}\Bigg].\end{aligned} \tag{4.61}$$

Here

$$G^p_j = s^p_{j+1/2}\max\left[0, \min\left(\left|\tilde{G}^p_{j+1/2}\right|, s_{j+1/2}\tilde{G}^p_{j-1/2}\right)\right] \tag{4.62}$$

with

$$\tilde{G}^p_{j+1/2} = \frac{1}{\lambda}\sigma(D^p_{j+1/2})\alpha^p_{j+1/2}, \tag{4.63}$$

$$\gamma^p_{j+1/2} = \lambda\frac{G^p_{j+1} - G^p_j}{\alpha^p_{j+1/2}} \tag{4.64}$$

and $s^p_{j+1/2} = \text{sgn}\left(\alpha^p_{j+1/2}\right)$. The bracket notation implies that all elements inside brackets has the superscript p and subscript $j+1/2$. Equation (4.61) gives the following scheme

$$\begin{aligned}\mathbf{Q}_j^{n+1} = \mathbf{Q}_j^n - \frac{\lambda}{2}\left(\mathbf{F}_{j+1} - \mathbf{F}_j + \mathbf{F}_j - \mathbf{F}_{j-1}\right) - \\ \sum_{p=1}^{m}\left(\frac{\lambda}{2}\left((G^p_{j+1} + G^p_j)\mathbf{r}^p_{j+1/2} - (G^p_j + G^p_{j-1})\mathbf{r}^p_{j-1/2}\right) - \right. \\ \left. \sum_{i=-N+1}^{N-1}\left([K_l(D+\gamma)\alpha\mathbf{r}]^p_{j+i+1/2} - [K_l(D+\gamma)\alpha\mathbf{r}]^p_{j+i-1/2}\right)\right).\end{aligned} \tag{4.65}$$

In order to find how (4.65) is related to the wave schemes presented earlier, we need to separate the part that is described by local information at each cell face and that which can not. With the Roe linearization we have that

$$\lambda\left(\mathbf{F}_{j+1} - \mathbf{F}_j + \mathbf{F}_j - \mathbf{F}_{j-1}\right) = \sum_{p=1}^{m}\left([D\alpha\mathbf{r}]^p_{j+1/2} + [D\alpha\mathbf{r}]^p_{j-1/2}\right), \tag{4.66a}$$

$$\lambda\left(G^p_{j+1} + G^p_j\right)\mathbf{r}^p_{j+1/2} = [\gamma\alpha\mathbf{r}]^p_{j+1/2} + 2\lambda G^p_j\mathbf{r}^p_{j+1/2}, \tag{4.66b}$$

$$\lambda\left(G^p_j + G^p_{j-1}\right)\mathbf{r}^p_{j-1/2} = [\gamma\alpha\mathbf{r}]^p_{j-1/2} + 2\lambda G^p_j\mathbf{r}^p_{j-1/2}. \tag{4.66c}$$

Inserting (4.66) into (4.65) we find

$$\begin{aligned}\mathbf{Q}_j^{n+1} = \mathbf{Q}_j^n - \sum_{p=1}^{m}\left(\frac{1}{2}\left([(D+\gamma)\,\alpha\mathbf{r}]^p_{j+1/2} + [(D+\gamma)\,\alpha\mathbf{r}]^p_{j-1/2}\right) - \right. \\ \sum_{i=-N+1}^{N-1}\left([K_l(D+\gamma)\alpha\mathbf{r}]^p_{j+i+1/2} - [K_l(D+\gamma)\alpha\mathbf{r}]^p_{j+i-1/2}\right) + \\ \left.\lambda G^p_j\left(\mathbf{r}^p_{j+1/2} - \mathbf{r}^p_{j-1/2}\right)\right).\end{aligned} \tag{4.67}$$

We identify two contributions. The first contribution is of the form consistent with the wave schemes we presented earlier, while the second is not. Thus we cannot apply wave schemes on each component. We have to add a correction

$$\begin{aligned}\mathbf{Q}_j^{n+1} = \mathbf{Q}_J^n - \sum_{p=1}^{m}\left(\sum_{i=0}^{N-1}\left(\left[C^{+i}(D+\gamma)\alpha\mathbf{r}\right]^p_{j+1/2-i} + \right.\right. \\ \left.\left.\left[C^{-i}\,(D+\gamma)\,\alpha\mathbf{r}\right]^p_{j+1/2+i}\right) + \lambda G^p_j\left(\mathbf{r}^p_{j+1/2} - \mathbf{r}^p_{j-1/2}\right)\right)\end{aligned} \tag{4.68}$$

Remark 1: Roe linearization The correction is a consequence of using Roe-linearization. The Roe linearization we use is designed so that the flux $\mathbf{f}(\mathbf{q})$ is conserved, not the modified flux: $\mathbf{f}^M(\mathbf{q})$. Thus the consistency condition we derived for scalar conservation laws

becomes

$$\sum_{i=0}^{N-1}\left(\left[C^{+i}(D+\gamma)\alpha\mathbf{r}\right]^p_{j+1/2}+\left[C^{-i}(D+\gamma)\alpha\mathbf{r}\right]^p_{j+1/2}\right)=\left[(D+\gamma)\alpha\mathbf{r}\right]^p_{j+1/2} \tag{4.69}$$

and

$$\sum_{p=1}^{m}\left[(D+\gamma)\alpha\mathbf{r}\right]^p_{j+1/2}=\lambda\left(\mathbf{F}_{j+1}-\mathbf{F}_{j+1}\right)+\lambda\sum_{p=1}^{m}(G^p_{j+1}-G^p_j)\mathbf{r}^p_{j+1/2}, \tag{4.70}$$

which is dependent on the cell wall value and thus will not give a conservative method.

Remark 2: System of conservation laws and TVD For scalar hyperbolic conservation laws it is very powerful to have schemes that are TVD as it guarantees stability and no oscillations. However, a system of nonlinear hyperbolic conservation laws can have solutions where the total variation increases. Thus the TVD concept is not generalizable to systems. Nonetheless, as is shown in [6], the particular generalization we use here retains some of the TVD aspect of scalar wave schemes by being TVD in each of characteristic components for linear systems.

4.3.1 Harten's LTS High Resolution scheme

A high resolution LTSS has already been suggested by Harten (LTS-Harten). LTS-Harten is based on dividing the time step into N equal time steps and applying Upwind N times in order to get a first order accurate scheme that is TVD for $CFL < N$. Second order is then obtained by using the modified flux approach we previously followed. The method with a small modification[6] is

$$\begin{aligned}\lambda\hat{\mathbf{f}}_{j+1/2}=&\frac{\lambda}{2}(\mathbf{f}_{j+1}+\mathbf{f}_j)+\\&\sum_{p=1}^{m}\left(\frac{1}{2}\left(G^p_{j+1}+G^p_j\right)\mathbf{r}^p_{j+1/2}-\sum_{i=-N+1}^{N-1}\left[K_l(D+\gamma)\alpha\mathbf{r}\right]^p_{j+i+1/2}\right)\end{aligned} \tag{4.71}$$

where G^p_j is as before with

$$\sigma^{\text{LTS-Harten}}(D)=\frac{N}{2}\left\{\mathcal{Q}\left(\frac{D}{N}\right)\left[1+\frac{N-1}{2}\mathcal{Q}\left(\frac{D}{N}\right)\right]-\frac{N+1}{2}\left(\frac{D}{N}\right)^2\right\}, \tag{4.72}$$

where

$$\mathcal{Q}(D)=\begin{cases}\dfrac{1}{2}\left(\dfrac{D^2}{\epsilon}+\epsilon\right) & \text{for } |D|<\epsilon,\\[2ex] |D| & \text{for } |D|\geq\epsilon,\end{cases} \tag{4.73}$$

[6] In [6], Harten uses $\mathbf{r}_{j+1/2}$ instead of $\mathbf{r}_{j+i+1/2}$. This might be a conscious choice by Harten or simply a typo. As in [17], we experienced greater stability with the natural modification without loss of accuracy, thus we will only use the latter.

is an entropy fix where ϵ should be sufficiently small (see next paragraph). Further, the coefficients are given by

$$K_{\pm i}(D) = \begin{cases} c_i\left(\frac{1}{2}\left[\mathcal{Q}\left(\frac{D}{N}\right) \mp \frac{D}{N}\right]\right) & \text{for } 1 \leq i \leq N, \\ \frac{N}{2}\mathcal{Q}\left(\frac{D}{N}\right) & \text{for } i = 0, \end{cases} , \tag{4.74}$$

where

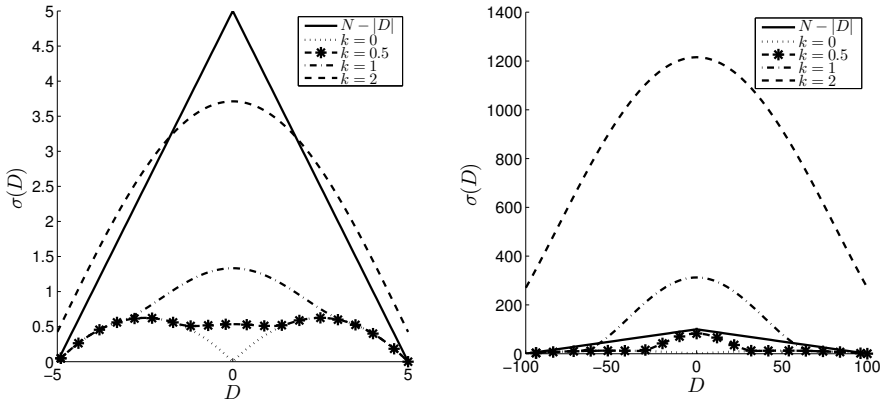
$$c_i(x)(D) = \begin{cases} -\binom{N}{i} x^i \sum_{k=1}^{N-i} \binom{N-i}{k} \frac{ki}{(k+i-1)(k+i)} (-x)^k & \text{for } i \geq 1, \\ \frac{N}{2}x & \text{for } i = 0. \end{cases} \tag{4.75}$$

Warning Harten proved that this scheme was TVD for a scalar conservation equation for any CFL-number. However great care needs to be taken in the actual implementation of this scheme for high CFL-number to avoid number overflow. At high CFL binomials become huge and the powers of x have to be very accurately computed for the scheme to yield correct $c_i(x)$. Above $CFL > 30$, we recommend using multiprecision variables, for instance the multiprecision implementation in the boost library.

Inaccuracy in [6] It is claimed that the entropy fix $\mathcal{Q}(D)$ gives a scheme that is TVD whenever $\epsilon \leq 2N(1 - 2^{-1/N})$. It is not mentioned that this limit is not sufficient for the second order version. We see this by considering if $\sigma^{\text{Harten}}(D) \leq N - |D|$. In Figure 4.1 we show the diffusion coefficient with the maximum allowed diffusion for $N = 5$, and $N = 100$ for different ϵ. In Figure 4.1a all the ϵ are within maximum allowed diffusion except for $\epsilon \leq 2N(1 - 2^{-1/N})$. Further, in Figure 4.1b, even the more conservative $\epsilon \leq N(1 - 2^{-1/N})$ does not give a scheme that is TVD. From these figures we conclude that one should use a more conservative value of ϵ. If the simulations are run with $CFL \leq 100$, then it is safe to use $\epsilon \leq \frac{1}{2}N(1 - 2^{-1/N})$.

4.4 Analysis of diffusion coefficients and dispersion CFL-numbers

We end this chapter by comparing the diffusion coefficient of the different methods. Figure 4.2a shows the diffusion coefficients for various methods at $N = 1$. First we note that all methods are within the maximum allowed diffusion. When $N = 1$, LTS-Harten and LTS-Roe should reduce to normal Roe when the entropy fix is not used. We see that this is indeed the case for sufficiently high $|D|$ ($|(|D) > \epsilon$) by the diffusion coefficients completely overlapping. Next we examine the level of diffusion in the middle for the different methods. In order to avoid entropy mistakes the diffusion should be nonzero.



(a) $N = 5$ **(b)** $N = 100$

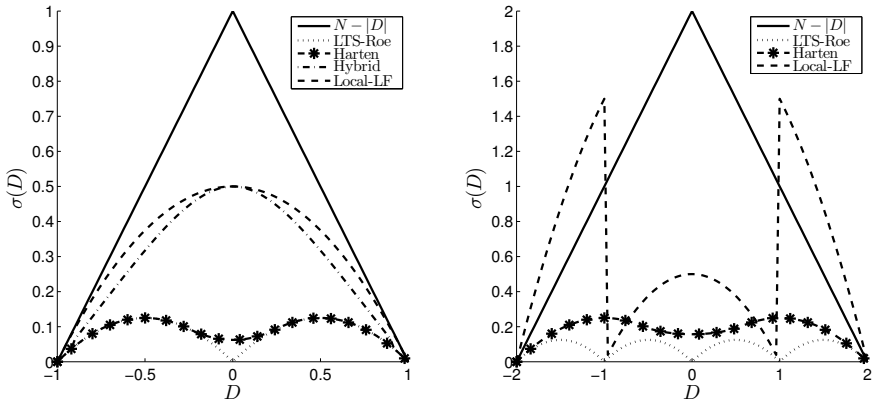
Figure 4.1: Diffusion coefficient $\sigma^{\text{Harten}}(D)$ against maximum allowed diffusion for $\epsilon = kN(1 - 2^{-1/N})$.

We see that all methods except LTS-Roe have nonzero diffusion coefficients. Of the three LTS-Harten has the lowest diffusion. In light of this figure, Hybrid might introduce an unnecessary large amount of diffusion, and it might be better to use a higher weight on LTS-Roe. Next we investigate the diffusion coefficients for $N = 2$ in Figure 4.2b. First we note that Local-LF is not within the maximum allowed diffusion (see Proposition 7), thus it cannot be made second order. Next we see that LTS-Roe and LTS-Harten have completely different form. LTS-Roe repeats the pattern of $N = 1$ periodically, while LTS-Harten has more or less the same pattern as before, but wider. When the CFL-number is increased in the LTS-Roe the level of diffusion is not increased, while LTS-Harten becomes more diffusive. This indicates that LTS-Harten will tend to smear solutions as the CFL-number is increased, while LTS-Roe will not.

LTS-Roe and LTS-LF have many points where there is zero diffusion, while LTS-Harten has always a nonzero diffusion. These points can cause entropy mistakes. This was also observed in [14] for LTS-Roe and Local-LTS-LF. Local-LTS-LF has zero diffusion when $D = N = \text{ceil}(D)$. Lets consider the Burgers equation. In this equation we have $D = \frac{\Delta t}{\Delta x} q$. We expect entropy mistakes wherever $q_{\text{mistake}} = N\frac{\Delta x}{\Delta t}$. We have $CFL = q_{\max}\frac{\Delta t}{\Delta x}$. Thus

$$q_{\text{mistake}} = \frac{N}{CFL} q_{\max}. \tag{4.76}$$

At such points diffusion is zero. When diffusion is zero, steps will not diffuse. It is similar to entropy mistakes for upwind when $D = 0$. It was suggested in [14] that entropy mistakes can be significantly reduced by using random time steps. From our simple analysis we can get an intuitive understanding of why this works. When using random time steps, the CFL-numbers become dispersed while $q_{\max}$ is unchanged. This means that



(a) $N = 1$ **(b)** $N = 2$

Figure 4.2: Plot of diffusion coefficient for different schemes and CFL-numbers.

q_{mistakes} also becomes dispersed (except for $N = 0$). Thus instead of having entropy mistakes propagate between time steps, they get dispersed. Another interesting observation in [14] was that most LTSS had a tendency to produce step-like solutions. These step-like solutions were not entropy mistakes since the solutions converged with grid refinement. Nevertheless, this phenomena impaired the accuracy of the LTSS studied (LTS-Harten was not studied). It was shown that errors caused by this step phenomena could be significantly reduced by random time steps, thus increasing the accuracy. A full understanding of the step phenomena has yet to be developed.

Chapter 5

Numerical investigations

To assess and compare the methods we have developed, we will examine how well they approximate the exact solution. Our tests will focus on Burgers' equation and the Euler equations.

5.1 Error estimate and order of accuracy

When measuring the accuracy of the methods, we need to specify which norm we use to measure the error. If Q denotes the numerical solution and q the exact, then a measure of the error is

$$L_b(\{Q\},\{q\}) = \left(\sum_{j=1}^{\mathcal{N}} \Delta x \left|Q_j - q(x_j)\right|^b\right)^{\frac{1}{b}}. \tag{5.1}$$

L_b is referred to as the b–norm of the error. In general, the higher the b, the higher the contribution from points with low accuracy. We will use the standard

$$\epsilon(\Delta x) \equiv L_1(\{Q\},\{q\}) = \sum_{j=1}^{\mathcal{N}} \Delta x \left|Q_j - q(x_j)\right|. \tag{5.2}$$

As the grid size Δx is refined we have that

$$\epsilon(\Delta x) = L_1(\{Q\},\{q\}) \simeq C\Delta x^p, \tag{5.3}$$

where p is the rate of convergence (also called the order of the method). We find the local order of convergence by

$$p = \frac{\log\left(\epsilon(\Delta x_1)\right)/\log\left(\epsilon(\Delta x_1)\right)}{\log(\Delta x_1)/\log(\Delta x_2)}. \tag{5.4}$$

5.2 Burgers' equation

A C++ program with implementation of LTS-Roe, Hybrid and LTS-Harten for Burgers' equation (2.21) was made. All simulations were done with ghost cells (an extension of numerical domain by the inclusion of additional cells at either side of the numerical domain) as boundary conditions using zeroth-order extrapolation.

5.2.1 Gauss pulse

The first problem we consider is a Gauss pulse that propagates for a short enough time that no shocks are created. Figure 5.1 shows the initial configuration with the exact solution of Burgers' equation at $t = 0.1$ (the exact solution is obtained by the method of characteristics).

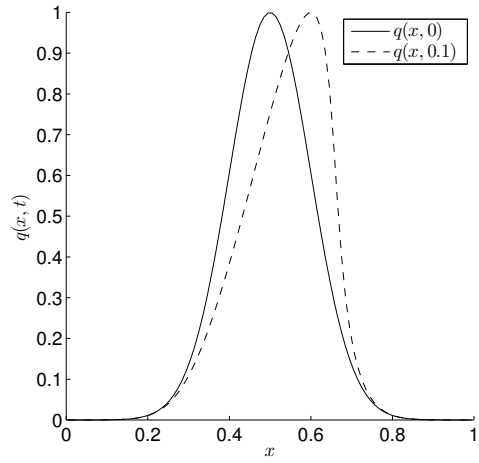
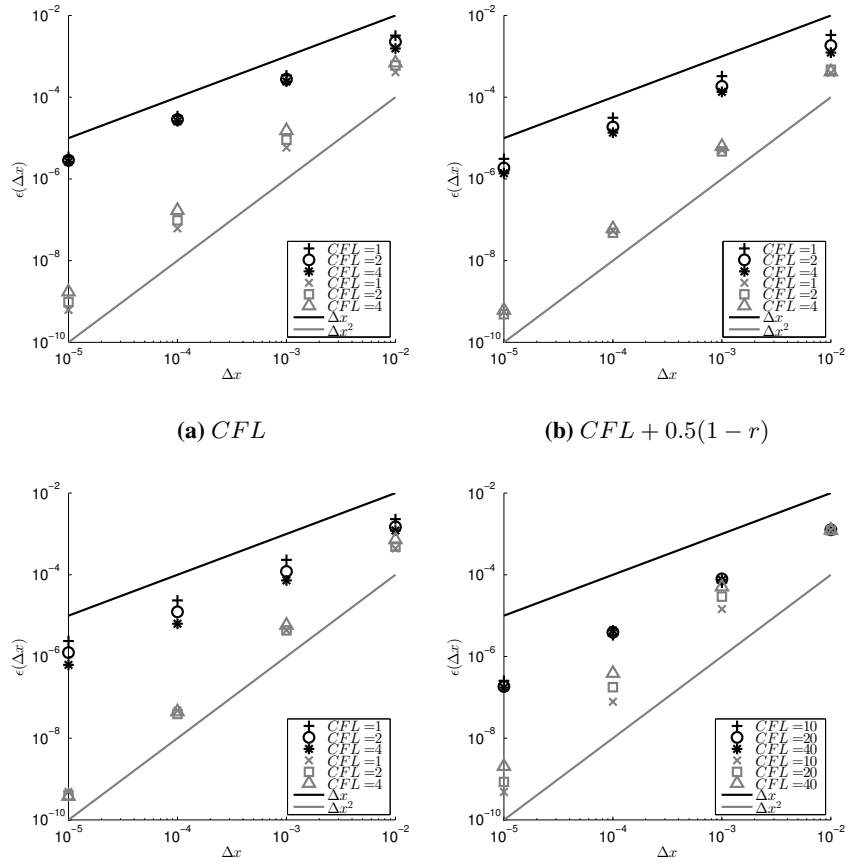


Figure 5.1: Initial conditions and corresponding exact solution for Burgers' equation at $t = 0.1$.

Figure 5.2a shows $\epsilon(\Delta x)$ for the first order LTS-Roe (LTS-Roe1) and the second-order LTS-Roe (LTS-Roe2) at various CFL-numbers and expected slope for first- and second-order methods Δx and Δx^2. We start by analyzing LTS-Roe1. $\epsilon(\Delta x)$ for all CFL-numbers has a slope close to that which is expected for first order convergence (confirmed in Table 5.1a for constant CFL). Interestingly, the error is reduced as the CFL-number is increased. This is understandable as larger time steps means fewer projections and thus a less diffusive method. Next we examine the LTS-Roe2. For this scheme we have close to second-order convergence for all CFL-numbers (see table 5.1b). As the cell size is decreased, LTS-Roe2 is more accurate than LTS-Roe1 by several magnitudes. Contrary to LTS-Roe1, accuracy is decreased as the CFL-number is increased.

In Section 4.4 we discussed how using random CFL-numbers could reduce entropy mistakes and increase accuracy by reducing a step-like solution phenomena. Taking $CFL_{\text{rand}} = CFL + (0.5 - r)$ where r is a random number between 0 and 1 gives on average CFL. Figure 5.2b shows $\epsilon(\Delta x)$ using random CFL-numbers. For both methods, the accuracy



(a) CFL **(b)** $CFL + 0.5(1 - r)$

(c) $[0, 2CFL]$ **(d)** $[0, 2CFL]$

Figure 5.2: Log-Log scatter plot of $\epsilon(\Delta x)$ for LTS-Roe1 (black) and LTS-Roe2 (gray) at different CFL-numbers with lines showing the expected convergence of first- and second-order methods.

is better. This is also reflected in Table 5.1a and 5.1b by higher rate of convergence p for coarser grid. Another way of taking a random CFL-number is to take a random $CFL_{\text{rand}} = [0, 2 \cdot CFL]$. Doing so covers a wider range of CFL-numbers. One might expect that very large steps will reduce the accuracy, however as we see in Figure 5.2c; accuracy is increased significantly. All methods converge faster and especially simulations done with the largest time steps start performing for LTS-Roe1 better and for LTS-Roe2 almost as accurate as simulations done with smaller steps. Finally we examine Figure 5.2d. Here simulations are done with very large average time steps. Despite this, as the grid is refined, the LTSS perform well. For LTS-Roe1 we get better accuracy than we did in Figure 5.2c. However at high enough CFL-number, the largest step simulation has lower

Table 5.1: Order of convergence p for different CFL-numbers with various step types.

(a) LTS-Roe1

$\Delta x_1/\Delta x_2$	Constant CFL			$CFL-(0.5-r)$			[0,2 CFL]		
	1	2	4	1	2	4	1	2	4
1.00e-02/1.00e-03	0.98	0.92	0.81	1.01	1.00	0.96	1.00	1.09	1.22
1.00e-03/1.00e-04	1.00	0.99	0.97	1.02	1.00	1.00	0.99	0.99	1.06
1.00e-04/1.00e-05	1.00	1.00	1.00	1.01	1.00	0.99	0.99	1.00	1.01

(b) LTS-Roe2

$\Delta x_1/\Delta x_2$	Constant CFL			$CFL-(0.5-r)$			[0,2 CFL]		
	1	2	4	1	2	4	1	2	4
1.00e-02/1.00e-03	1.84	1.82	1.66	1.88	2.00	1.83	2.00	2.05	2.09
1.00e-03/1.00e-04	1.98	1.98	1.96	1.98	1.99	2.01	1.97	2.05	2.12
1.00e-04/1.00e-05	2.00	2.00	1.99	1.98	1.99	2.00	1.99	1.99	2.07

(c) LTS-Harten

$\Delta x_1/\Delta x_2$	Constant CFL			$CFL-(0.5-r)$			[0,2 CFL]		
	1	2	4	1	2	4	1	2	4
1.00e-02/1.00e-03	1.84	1.91	1.88	1.84	1.89	1.93	1.89	1.93	1.94
1.00e-03/1.00e-04	1.98	1.98	1.99	1.97	1.96	1.98	1.96	2.01	1.97
1.00e-04/1.00e-05	2.00	2.00	2.00	1.98	1.98	1.99	1.99	1.99	2.01

accuracy than simulations done with smaller steps. For LTS-Roe2 we see a deterioration of accuracy as the CFL-numbers are increased. increased[1].

Finally, we examine LTS-Harten and compare its performance with LTS-Roe2. Figure 5.3 shows $\epsilon(\Delta x)$ for various CFL-numbers for both schemes. First we note that LTS-Harten has a second-order convergence. This is also reflected in the Table 5.1c with order of convergence p approaching 2 as the grid is refined. Note that unlike for LTS-Roe2, there is no significant advantage in using random CFL-number. This is understandable as our analysis of the diffusion coefficient showed it to be nonzero for all CFL-numbers. Although both schemes are second order, LTS-Roe2 has a significant better accuracy than LTS-Harten.

5.2.2 Square pulse

We now turn to the problem of a square pulse as initial configuration shown in Figure 5.4 with corresponding exact solution at $t = 0.2$. This problem is more challenging as it involves a discontinuous solution consisting of a rarefaction fan and a shock. This makes a square pulse a good case for not only testing the accuracy, but also whether the method is TVD and respects the entropy condition.

[1] In the Figure 5.2d all points for $N = 100$ collapse onto a single point. This is an artifact of the time step being larger than simulation time and thus limited to the simulation time.

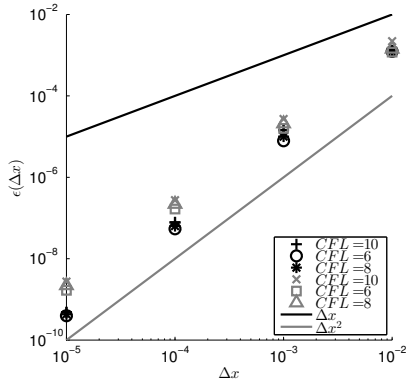


Figure 5.3: Error of numerical solution for LTS-Roe2 (black) and LTS-Harten at various CFL-numbers.

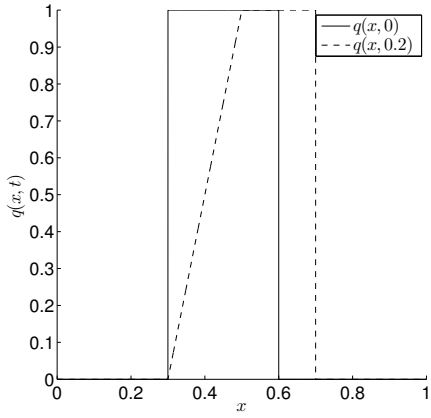
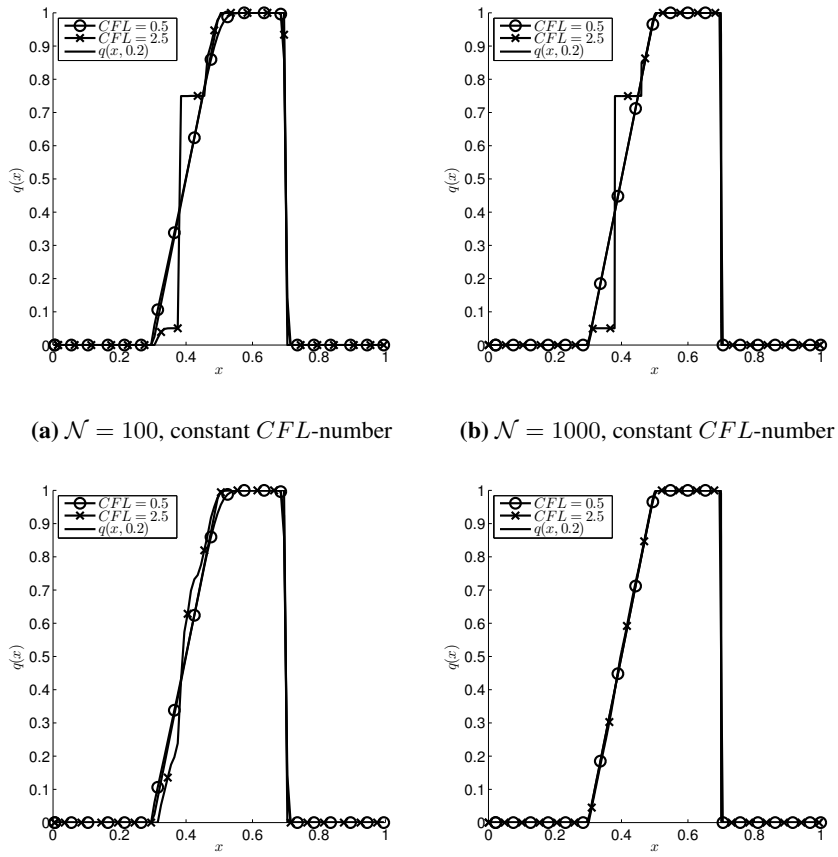


Figure 5.4: Initial condition and exact solution of square pulse at $t = 0.2$.

Before doing any convergence analysis, we examine the numerical solutions obtained with constant and random CFL-numbers. Figures 5.5a and 5.5b show numerical simulations done at constant CFL-number for $\mathcal{N} = 100$ and 1000. For $CFL = 0.5$ the numerical solution converges nicely, whereas for $CFL = 2.5$ the numerical solution does not converge to the correct solution. Figure 5.5c and 5.5d shows the same simulations, but with random CFL-numbers. With random $CFL = 2.5$, the solution for $\mathcal{N} = 100$ is not perfect. However as the grid is refined, the numerical solution converges towards the correct solution. We will thus use random time step in the following simulations[2]

[2]It will become clear in later simulations that high CFL-numbers gives less accurate solutions and for sys-



(a) $\mathcal{N} = 100$, constant CFL-number

(b) $\mathcal{N} = 1000$, constant CFL-number

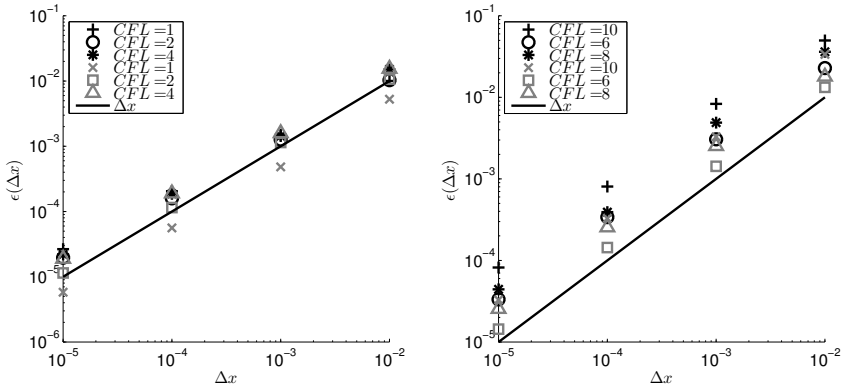
(c) $\mathcal{N} = 100$, random CFL-number

(d) $\mathcal{N} = 1000$, random CFL-number

Figure 5.5: Simulations of square pulse with LTS-Roe.

Next we examine the convergence of LTS-Roe1 and LTS-Roe2. Figure 5.6 shows $\epsilon(\Delta x)$ for simulations done at various random CFL-numbers. Unlike for the Gauss-pulse, we no longer have second-order convergence for LTS-Roe2. The first order convergence is to be expected as the methods we use are only first order near discontinuities. Consequently, this is by no means an artifact of us using LTSS. For $CFL = 1$, there is an significant improvement in accuracy by using LTS-Roe2 over LTS-Roe1. As the CFL-number is raised, LTS-Roe2 still gives more accurate results, but only by a small margin. Next we compare with LTS-Harten. The general trend is the same for both methods - the higher the CFL-number, the higher error. At all CFL-numbers LTS-Harten outperforms the LTS-Roe2 scheme.

tems oscillations. Thus we will use the more conservative choice of $CFL + (0.5 - r)$.



(a) LTS-Roe1 (black) and LTS-Roe2 (gray) **(b)** LTS-Roe2 (black) and LTS-Harten (gray)

Figure 5.6: Error of numerical solution obtained with LTSS.

5.2.3 Transonic rarefaction

It is well known that the Roe-scheme has problems with transonic rarefaction. Instead of producing a rarefaction fan like in figure 5.7, Roe produces a stationary shock (see figure 5.8a). This problem is caused by the Roe method having zero viscosity when the wave speed is zero and is also a problem for LTS-Roe. With Hybrid, we add diffusion by using Lax-Friedrichs when $D \simeq 0$. Figure 5.8b shows that Hybrid converges towards the correct solution. Thus in the following we will only focus on Hybrid.

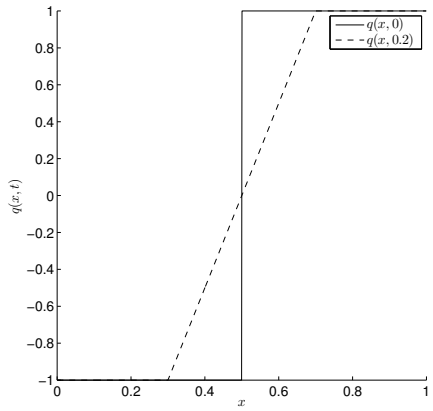
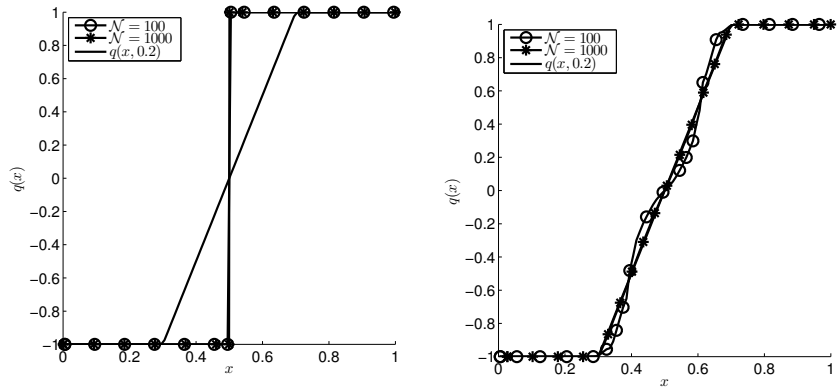


Figure 5.7: Initial condition and exact solution of transonic rarefaction at $t = 0.2$.

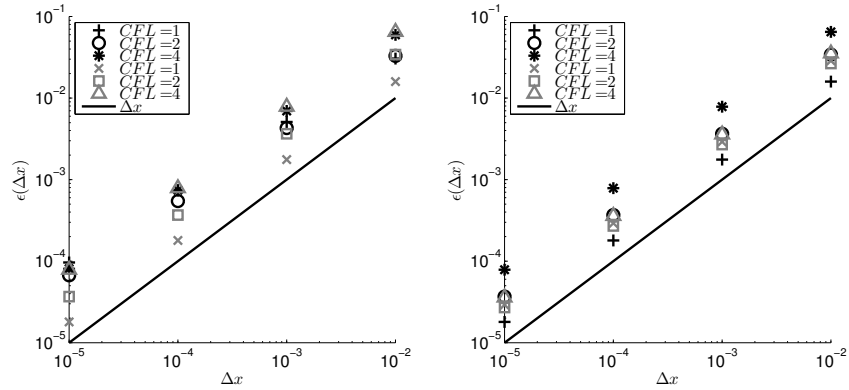
Next we examine the convergence of Hybrid. Figure 5.9 shows that we get first or-



(a) LTS-Roe2 **(b)** Hybrid

Figure 5.8: Simulations with the LTS-Roe2 and Hybrid using random time step.

der convergence for first (Hybrid1), second-order Hybrid (Hybrid2) and LTS-Harten. This is analogous to the previous case where we started with a discontinuous solution. For $CFL = 1$ and 2 Hybrid2 has the best performance, while for $CFL = 4$ there is little difference between Hybrid1 and Hybrid2. Finally compared to LTS-Harten, Hybrid2 only has the highest accuracy for $CFL = 1$. For higher CFL, LTS-Harten has better performance.



(a) Hybrid1 (black) and Hybrid2 (gray) **(b)** Hybrid2 (black) and LTS-Harten (gray)

Figure 5.9: Comparison of $\epsilon(\Delta x)$ for Hybrid2 and LTS-Harten.

5.2.4 Analysis of accuracy and random CFL-numbers

We have previously shown that LTS-Roe has the lowest possible diffusion of all LTSS. Therefore it was not unexpected that LTS-Roe2 obtained the highest accuracy of all the schemes for the continuous initial conditions. However, this accuracy was only achieved by using random CFL-numbers. LTS-Roe1 and LTS-Roe2 converged with constant CFL-numbers, thus there were no entropy problems. The increased accuracy observed is thus likely related to reduced error associated with the step phenomena.

For discontinuous initial conditions LTS-Harten had the best accuracy. LTS-Roe1 and LTS-Roe2 only converged for random CFL-numbers. Thus the entropy correction aspect of random CFL-numbers played an important role. This might help to explain why LTS-Roe2 did not achieve better accuracy than LTS-Harten. Random CFL-numbers helps to disperse entropy mistakes and thus they don't get reinforced from time-step to time-step. However mistakes do occur, and thus impair the accuracy at each time step. A natural conclusion is that random CFL-numbers are sufficient to ensure that LTS-Roe1 and LTS-Roe2 converge to the correct solution, but the accuracy is still impaired by entropy mistakes.

5.3 The Euler equations

To test how well LTSS performs for a system of conservation equations we apply the methods on the Euler equations (2.16). All of the presented results are obtained using the Roe-linearization. For a simple implementation of the Roe-linearization for the Euler equations, we recommend [7], and for a more in depth analysis we recommend [21]. The exact solutions for the Riemann problems are computed using a `FORTRAN` program written by Toro and provided in [21]. If not specified, simulations are done with ghost cells using zeroth-order extrapolation as boundary conditions.

5.3.1 Continuous initial conditions

First we examine if the LTSS are second order for a system of nonlinear conservation laws. We check this by using continuous initial conditions

$$\mathbf{q}_0 = \begin{pmatrix} 1 + 0.1\sin(2x\pi) \\ 1 + 0.1\sin((2x + 0.5)\pi) \\ 1 + 0.1\sin((2x + 0.25)\pi) \end{pmatrix}, \tag{5.5}$$

with periodic boundary conditions for short enough time, $t = 0.5$ that no shocks are created. The "exact solution" is estimated by a LTS-Roe2 simulation at $CFL = 1$ using $\mathcal{N} = 24300$ cells[3].

Tables 5.2a, 5.2b and 5.2c show order of convergence p for density ρ, velocity u and pressure P for LTS-Roe2, LTS-Harten and Hybrid. All methods have a convergence rate p close to 2 for all components. This confirms that all the second-order methods have second-order convergence for continuous initial conditions.

[3]This is a good choice, as simulations $\mathcal{N} = 100 \cdot 3^n$ for $0 \leq n \leq 5$ will have cell centers that coincides exactly with the "exact solution".

Table 5.2: The order convergence p for second-order methods for continuous solution at various CFL-numbers.

(a) LTS-Roe2

$\Delta x_1/\Delta x_2$	$CFL=1$			$CFL=2$			$CFL=4$			$CFL=8$		
	ρ	u	P	ρ	u	P	ρ	u	P	ρ	u	P
1.00e-02/3.33e-03	1.82	1.88	1.85	1.81	1.88	1.88	1.90	2.00	1.96	1.94	1.99	1.98
3.33e-03/1.11e-03	1.89	1.94	1.92	1.85	1.91	1.91	1.88	1.94	1.92	1.94	1.93	1.98
1.11e-03/1.23e-04	1.81	1.82	1.84	1.84	1.86	1.88	1.90	1.91	1.92	1.94	1.90	1.95

(b) LTS-Harten

$\Delta x_1/\Delta x_2$	$CFL=1$			$CFL=2$			$CFL=4$			$CFL=8$		
	ρ	u	P	ρ	u	P	ρ	u	P	ρ	u	P
1.00e-02/3.33e-03	1.82	1.88	1.87	1.80	1.88	1.87	1.80	1.89	1.87	1.83	1.91	1.88
3.33e-03/1.11e-03	1.88	1.92	1.91	1.87	1.92	1.91	1.86	1.93	1.93	1.85	1.95	1.94
1.11e-03/1.23e-04	1.79	1.81	1.82	1.79	1.82	1.83	1.80	1.83	1.84	1.83	1.87	1.87

(c) Hybrid

$\Delta x_1/\Delta x_2$	$CFL=1$			$CFL=2$			$CFL=4$			$CFL=8$		
	ρ	u	P	ρ	u	P	ρ	u	P	ρ	u	P
1.00e-02/3.33e-03	1.66	1.68	1.65	1.92	1.92	1.95	1.93	2.03	2.01	1.93	2.05	1.98
3.33e-03/1.11e-03	1.92	1.91	1.91	1.90	1.94	1.95	1.88	1.94	1.93	1.94	1.93	1.98
1.11e-03/1.23e-04	1.92	1.96	1.95	1.89	1.93	1.94	1.90	1.91	1.92	1.94	1.90	1.95

5.3.2 The shocktube problem

Next we consider the shocktube problem explored in [5–7]. Here, two initial states

$$\begin{aligned} \mathbf{q}_{\mathrm{L}} &= (0.455, 0.311, 8.928)^{\mathrm{T}}, \\ \mathbf{q}_{\mathrm{R}} &= (0.5, 0, 1.4275)^{\mathrm{T}}, \end{aligned} \tag{5.6}$$

inside a one dimensional tube (of length 1) are separated by a diaphragm at $x_0 = 0.5$. The diaphragm is then removed and the final state is observed at $t = 0.1$. This is a good test as the exact solution involves rarefaction fans and shocks and it allows for direct comparison of our results to those obtained by LTS-Harten in [6].

We start by comparing the results obtained by LTS-Roe1 and LTS-Roe2. Figure 5.10 shows numerical results for $CFL = 0.9$. As expected and well documented in [5], significant improvement in the resolution of shocks and rarefaction waves is obtained by LTS-Roe2 over LTS-Roe2. Next, the same simulation for $CFL = 1.8$ is shown in figure 5.11. LTS-Roe1 is sharper than for $CFL = 0.9$. LTS-Roe2 is a bit less sharp than for $CFL = 0.9$, but still sharper than the first order method. Further, we see the onset of small oscillations for both methods. As the CFL-number is increased to $CFL = 3.6$ (figure 5.12), the accuracy of the two is very similar, with both starting to develop more oscillations. Interestingly, the accuracy of LTS-Roe1 significantly improves and show high-resolution behavior. For $CFL = 5.4$ (Figure 5.13), oscillations become even more severe and deteriorate the solution significantly near the rarefaction fan. Thus for this problem the LTS-Roe1 improves in accuracy as CFL-number is increased until the onset of oscillations at $CFL = 3.6$. The accuracy of LTS-Roe2 however deteriorates as the CFL-number is increased and the difference in accuracy between LTS-Roe1 and LTS-Roe2 becomes very small.

Further, we examine LTS-Harten. Figures 5.14 and 5.15 show results obtained with

LTS-Harten for $CFL = 0.9, 1.8, 3.6$ and 5.4. The results replicate perfectly the ones presented in [6]. The trend for LTS-Harten is for the solution to diffuse more and more as CFL-number is increased. As with LTS-Roe2, when the CFL-number is high, oscillations appear near the rarefaction fan. This oscillation-like pattern is of much lower frequency than that of LTS-Roe. To explore further the nature of these oscillations we do simulations with $\mathcal{N} = 1800$. Figure 5.16 shows that for LTS-Roe2 the oscillations are still present and with a higher frequency than for $\mathcal{N} = 180$. Due to these oscillations the solution does not properly converge. The oscillations observed in the LTS-Harten scheme on the other hand are nearly non-existent. Thus oscillations can be removed by grid refinement. The oscillations observed for the Roe-scheme are the same observed in [13].

Next we compare the accuracy of LTS-Roe2 and LTS-Harten. Figures 5.17 and 5.18 show the error of the numerical solutions for LTS-Roe2 and LTS-Harten for various CFL-numbers. The corresponding local rate of convergence p is given in Table 5.3a and 5.3b. The convergence of $\epsilon_\rho(\Delta x)$ for both schemes is very similar. This is also reflected in the local rate of convergence, which for both is $p \simeq 0.6$, that is a sub first order convergence. However for $\epsilon_u(\Delta x)$ and $\epsilon_P(\Delta x)$, LTS-Harten approaches first order convergence while for LTS-Roe, the convergence rate is closer to $p \simeq 0.5$. Thus LTS-Harten has a better convergence than LTS-Roe2 for this numerical test.

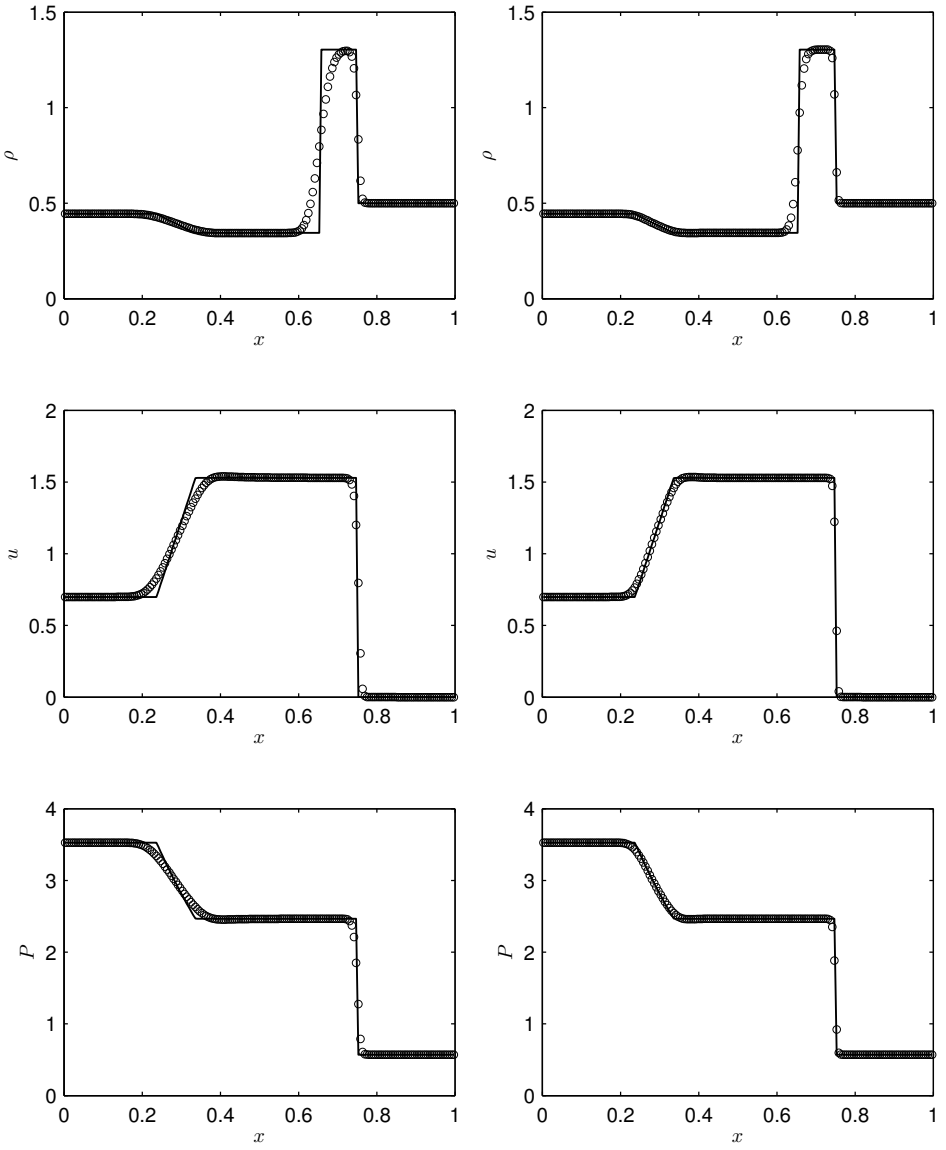
Table 5.3: Convergence order p for second-order methods at various CFL-numbers.

(a) LTS-Roe2

$\Delta x_1/\Delta x_2$	$CFL=1$			$CFL=2$			$CFL=4$			$CFL=8$		
	ρ	u	P	ρ	u	P	ρ	u	P	ρ	u	P
1.00e-02/5.00e-03	0.75	1.04	0.93	0.60	0.69	0.55	0.76	1.43	1.10	1.25	1.69	1.63
5.00e-03/2.50e-03	0.68	0.64	0.85	0.68	0.35	0.74	0.74	0.22	0.64	0.90	0.86	0.95
2.50e-03/1.25e-03	0.68	0.81	0.82	0.66	0.63	0.62	0.64	0.48	0.47	0.93	0.99	0.97
1.25e-03/6.25e-04	0.79	1.37	1.18	0.70	1.11	0.83	0.86	1.32	1.02	0.86	1.35	1.18
6.25e-04/3.13e-04	0.59	0.32	0.43	0.59	0.29	0.42	0.52	-0.09	0.07	0.76	0.50	0.54
3.13e-04/1.56e-04	0.75	1.05	0.96	0.76	0.97	0.87	0.72	0.69	0.62	0.81	0.85	0.83
1.56e-04/7.80e-05	0.65	0.76	0.66	0.62	0.59	0.54	0.57	0.45	0.39	0.65	0.52	0.48
7.80e-05/3.90e-05	0.65	0.66	0.62	0.61	0.53	0.48	0.60	0.53	0.46	0.64	0.40	0.39

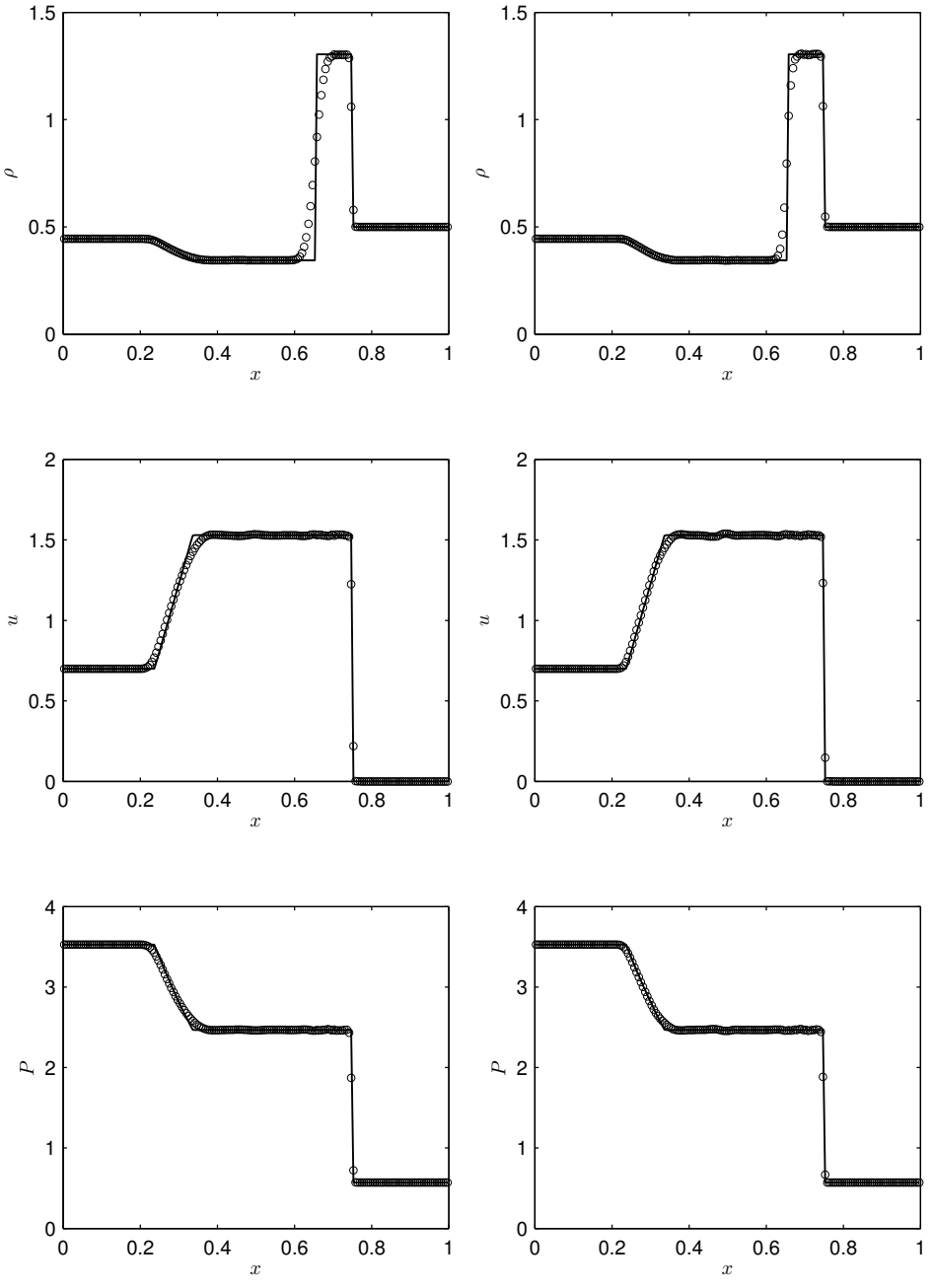
(b) Hartens LTSS

$\Delta x_1/\Delta x_2$	$CFL=1$			$CFL=2$			$CFL=4$			$CFL=8$		
	ρ	u	P	ρ	u	P	ρ	u	P	ρ	u	P
1.00e-02/5.00e-03	0.77	1.10	0.99	0.78	1.00	0.94	0.87	1.15	1.07	1.02	1.30	1.28
5.00e-03/2.50e-03	0.66	0.64	0.87	0.71	0.73	0.88	0.75	0.67	0.82	0.87	0.96	1.03
2.50e-03/1.25e-03	0.68	0.86	0.88	0.69	0.85	0.88	0.75	0.89	0.92	0.85	1.12	1.10
1.25e-03/6.25e-04	0.79	1.43	1.27	0.79	1.24	1.10	0.80	1.21	1.10	0.86	1.24	1.16
6.25e-04/3.13e-04	0.63	0.53	0.70	0.65	0.62	0.74	0.69	0.74	0.85	0.74	0.75	0.84
3.13e-04/1.56e-04	0.75	1.40	1.30	0.75	1.27	1.17	0.76	1.24	1.16	0.79	1.17	1.12
1.56e-04/7.80e-05	0.69	1.08	1.04	0.70	1.09	1.05	0.71	1.05	1.03	0.74	1.01	1.00
7.80e-05/3.90e-05	0.68	1.10	1.02	0.69	1.14	1.07	0.70	1.05	1.01	0.72	1.06	1.04



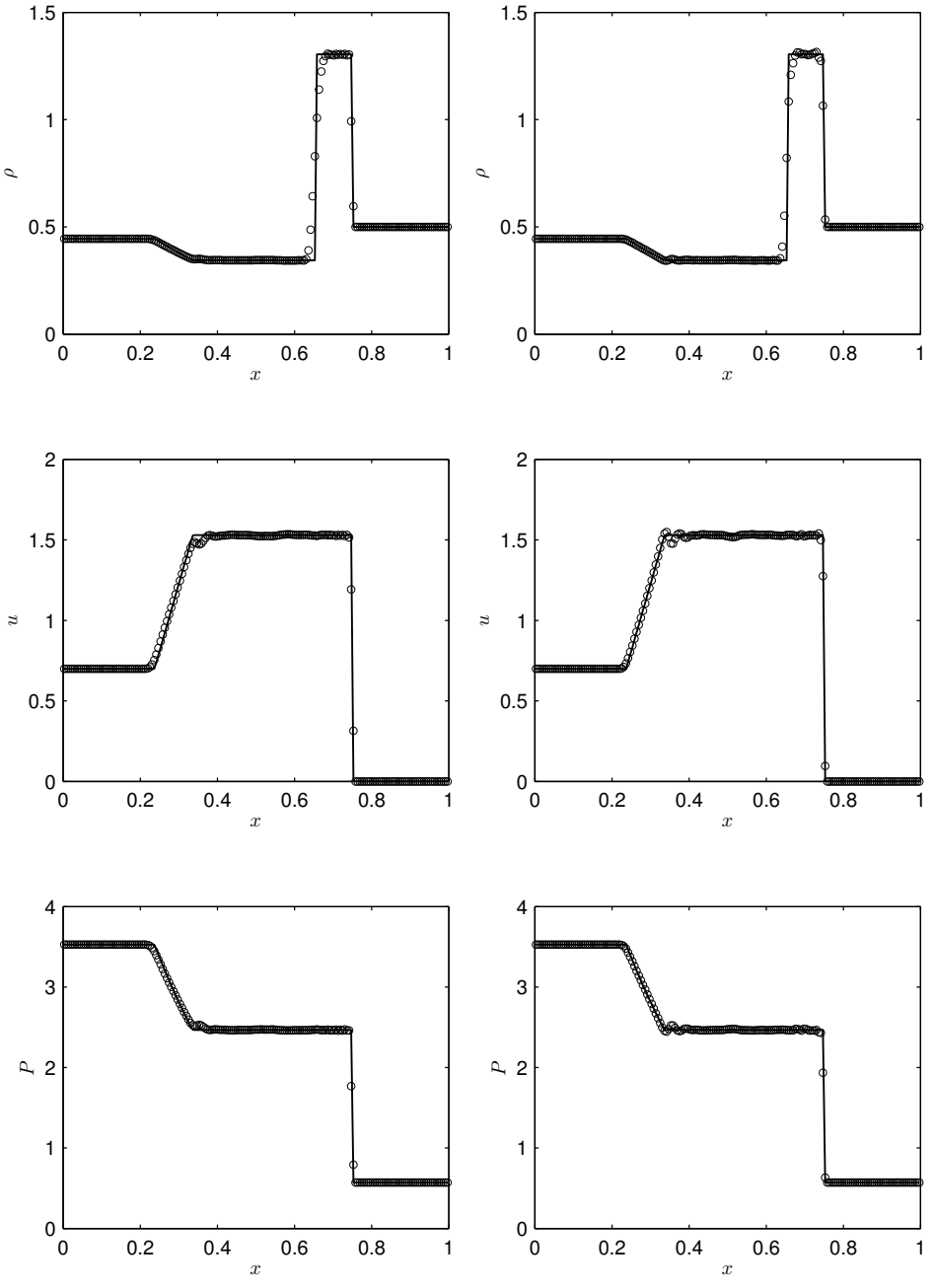
(a) LTS-Roe1 **(b)** LTS-Roe2

Figure 5.10: Numerical results obtained with random $CFL = 0.9$ and $\mathcal{N} = 180$. Exact solution given by black line.



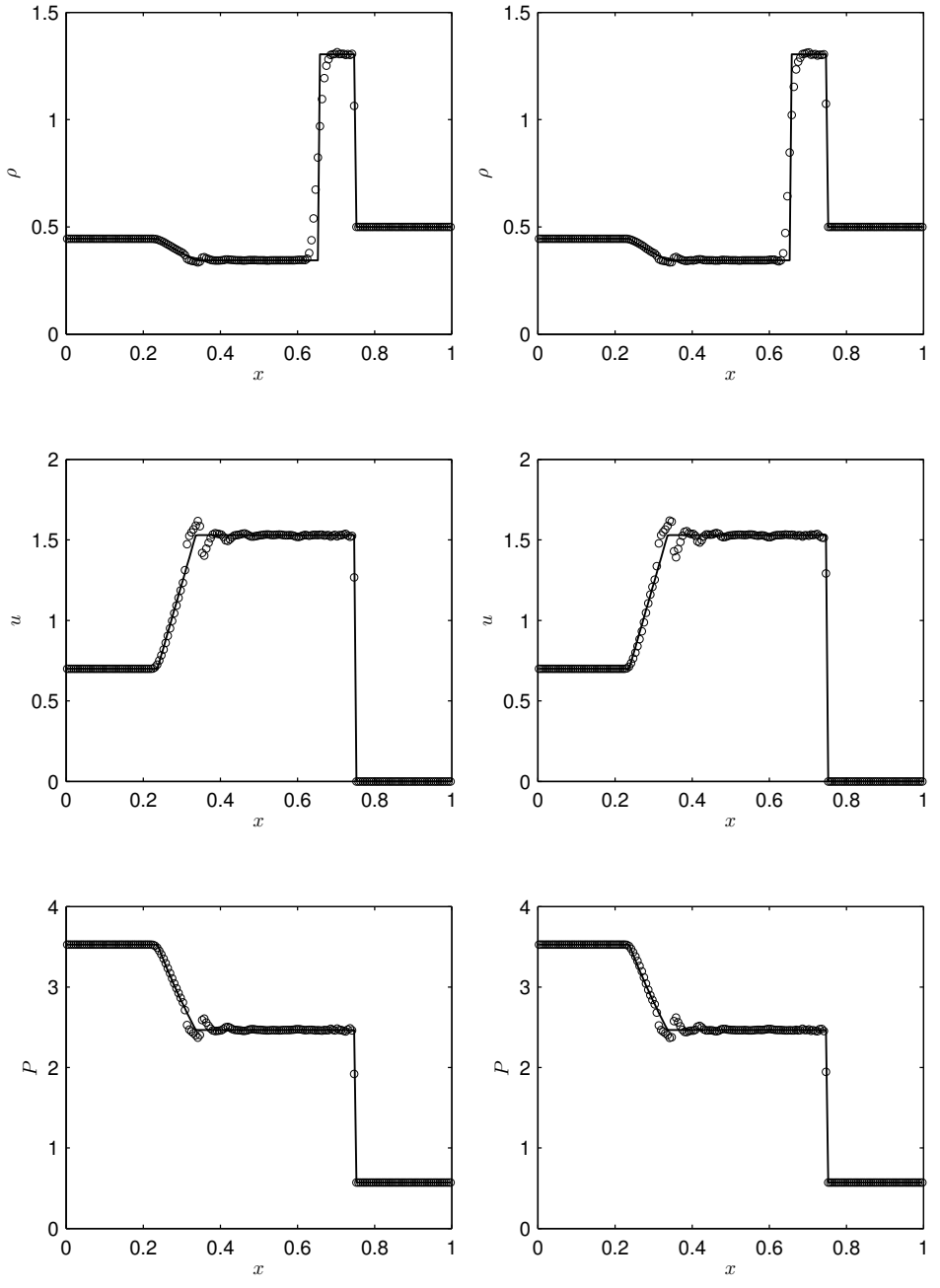
(a) LTS-Roe1 **(b)** LTS-Roe2

Figure 5.11: Numerical results obtained with random $CFL = 1.8$ and $\mathcal{N} = 180$. Exact solution given by black line.



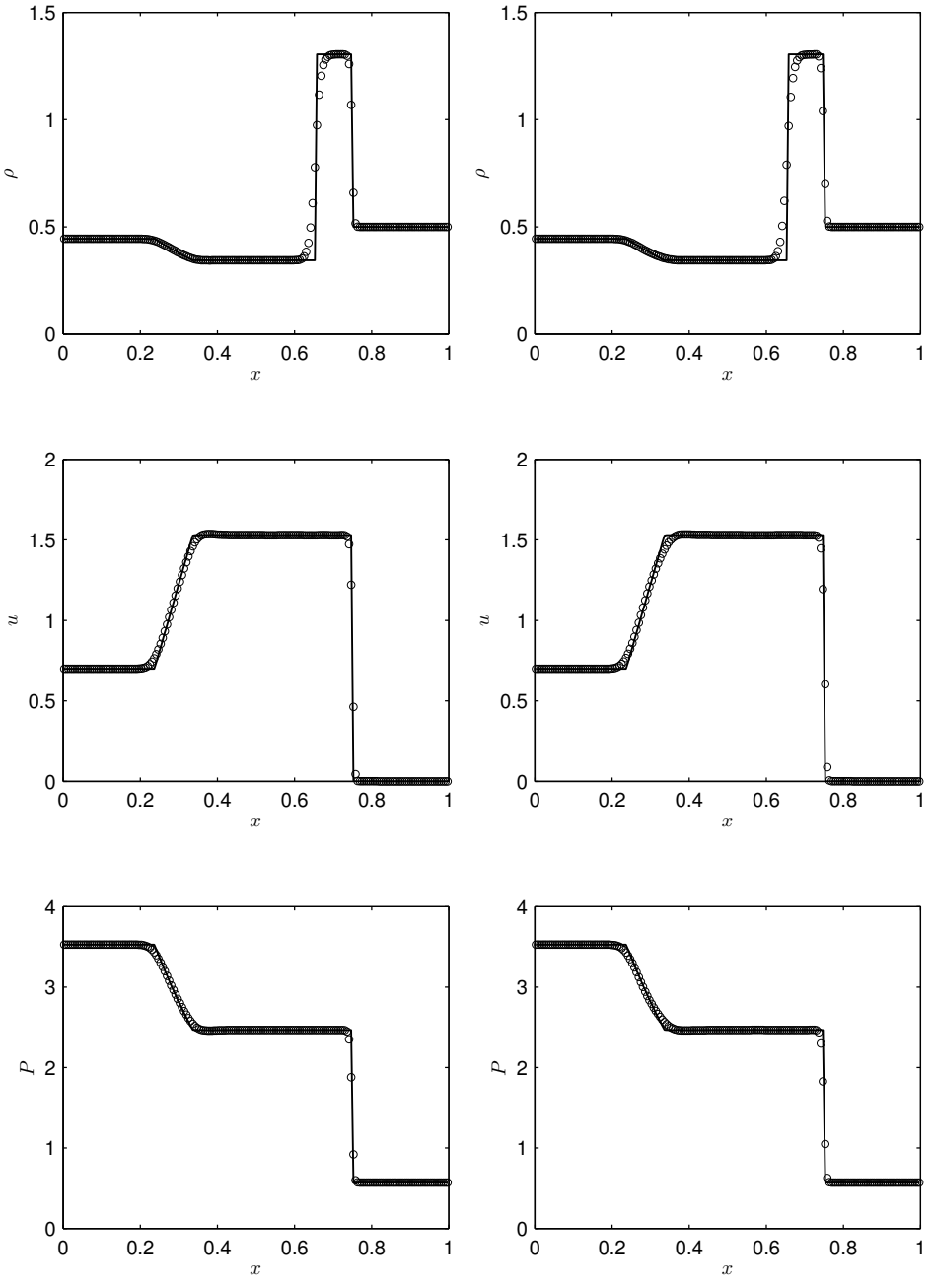
(a) LTS-Roe1 **(b)** LTS-Roe2

Figure 5.12: Numerical results obtained with random $CFL = 3.6$ and $\mathcal{N} = 180$. Exact solution given by black line.



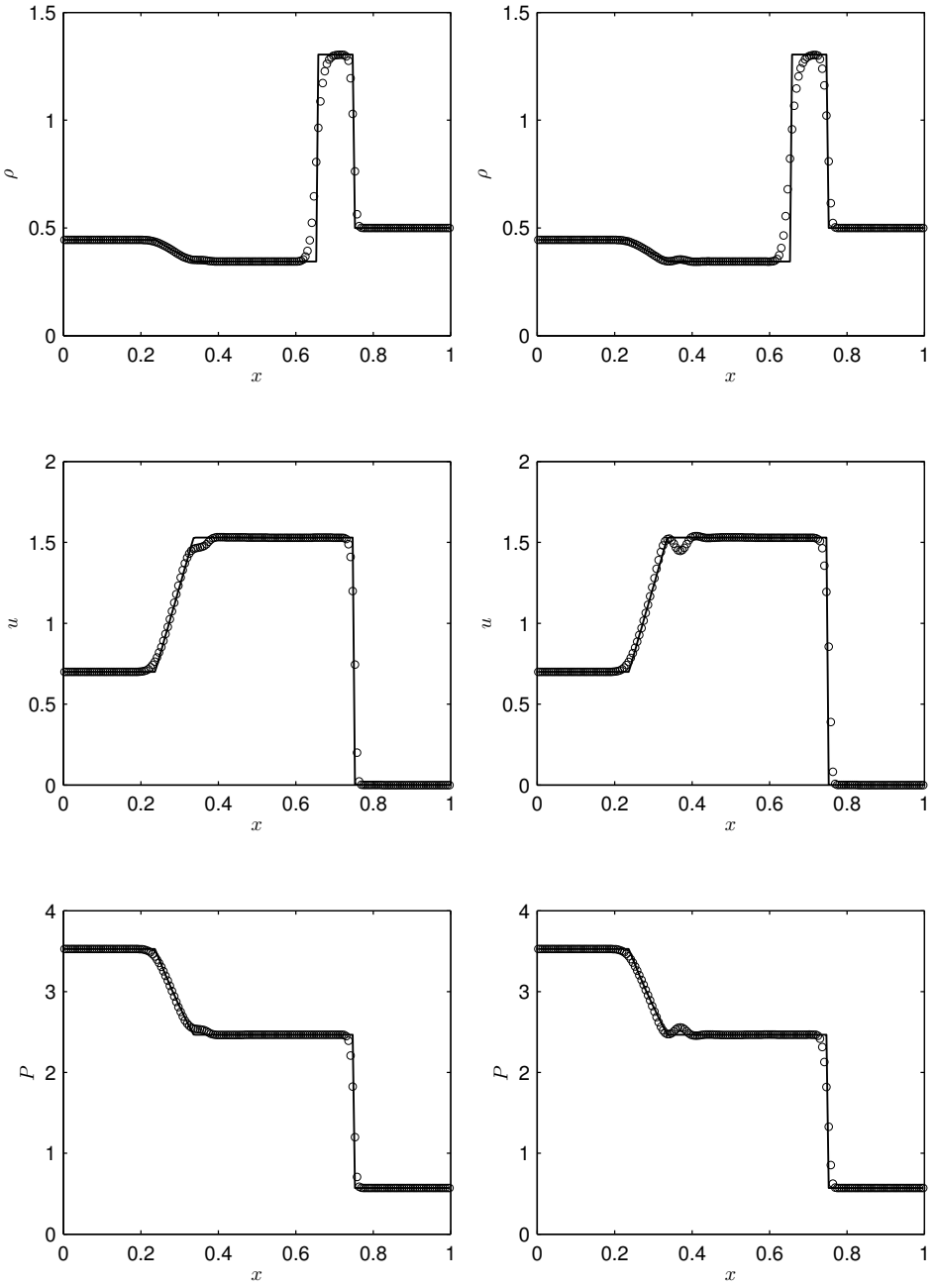
(a) LTS-Roe1 **(b)** LTS-Roe2

Figure 5.13: Numerical results obtained with random $CFL = 5.4$ and $\mathcal{N} = 180$. Exact solution given by black line.



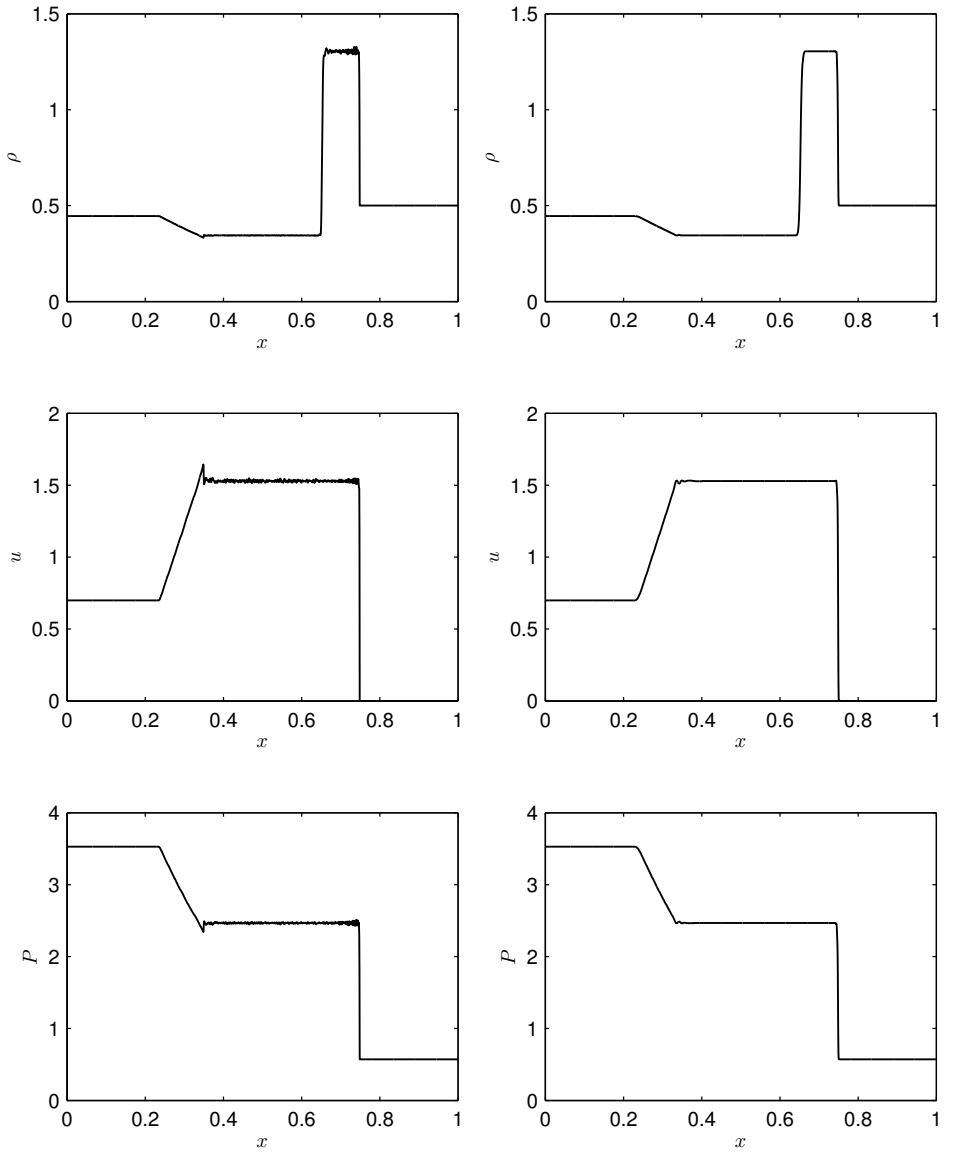
(a) $CFL = 0.9$ **(b)** $CFL = 1.8$

Figure 5.14: Numerical results obtained using LTS-Harten with constant CFL-number and $\mathcal{N} = 180$. Exact solution given by black line.



(a) $CFL = 3.6$ **(b)** $CFL = 5.4$

Figure 5.15: Numerical results obtained using LTS-Harten with constant CFL-number and $\mathcal{N} = 180$. Exact solution given by black line.



(a) LTS-Roe **(b)** Hartens LTSS

Figure 5.16: Numerical results obtained with $\mathcal{N} = 1800$ for LTS-Roe2 and LTS-Harten at $CFL = 5.4$.

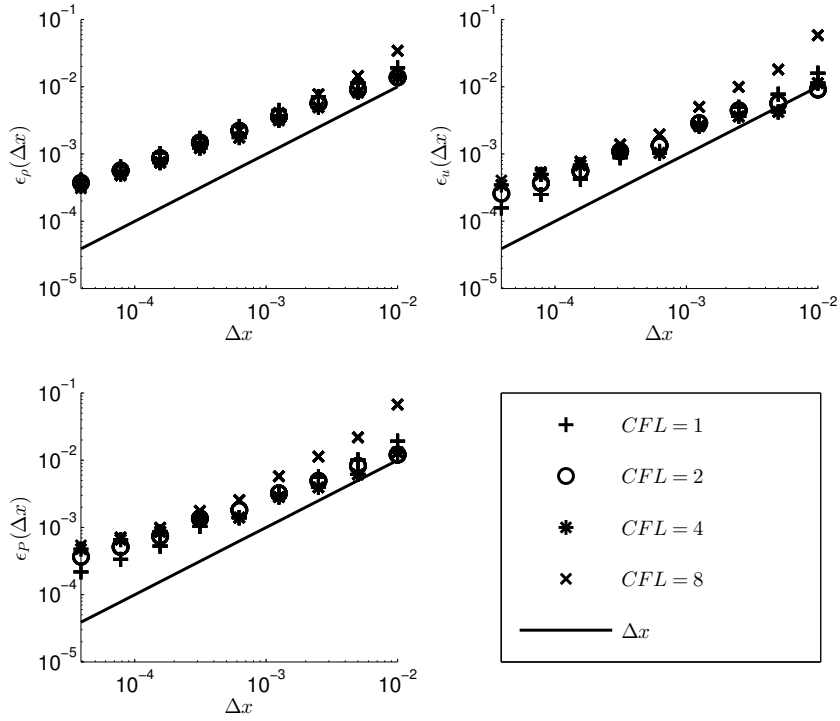


Figure 5.17: $\epsilon(\Delta x)$ for LTS-Roe2 at various CFL-numbers

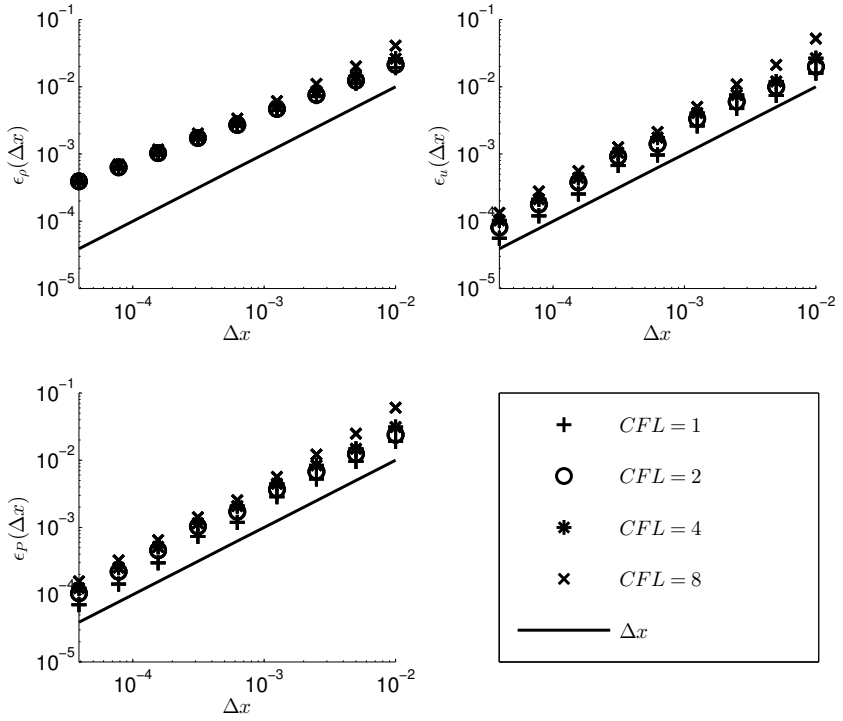


Figure 5.18: $\epsilon(\Delta x)$ for LTS-Harten at various CFL-numbers

5.3.3 Toro's five test problems

To further test the LTS-Roe2 and LTS-Harten, we focus on a set of problems proposed by Toro in [21] and discussed in [15]. These problems are designed to give a thorough test of the accuracy, robustness and entropy satisfaction. The input parameters of the tests are given in Table 5.4.

Test	ρ_L	u_L	P_L	ρ_R	u_R	P_R	x_0	t
#2	1.0	0.75	1	0.125	0.0	0.1	0.3	0.2
#2	1.0	-2.0	0.4	1.0	2.0	0.4	0.5	0.15
#3	1.0	0.0	1000.0	1.0	0.0	0.01	0.5	0.012
#4	5.99924	19.5975	460.894	5.99924	-6.19633	46.095	0.5	0.035
#5	1.0	-19.5975	1000.0	1.0	-19.59745	0.01	0.8	0.012

Table 5.4: Toro's five test cases.

Test 1

This test is a variation of Sod's shocktube problem [20]. It differs in that there is a sonic point in the rarefaction. The solution is composed of a right moving shock wave, a right travelling contact wave and a left sonic rarefaction wave. This test is particularly good for testing whether the method has good entropy satisfaction.

Figures 5.19 and 5.20 show numerical results for this test with $CFL = 1, 2, 4$ and 8 for LTS-Roe2. For $CFL = 1$ there is a entropy violation on the rarefaction wave. This violation becomes weaker however as the CFL-number is increased. Especially impressive is the solution for $CFL = 4$. Here the gap for $CFL = 1$ is closed and shocks are even sharper, without much oscillations. For $CFL = 8$, the numerical solution is severely deteriorated by oscillations. It is likely that the entropy mistake is caused by zero diffusion at the sonic point, thus it is interesting to test Hybrid on this problem. In Figures 5.21 and 5.22 we show the same experiment using Hybrid. As in the case of transonic rarefaction of Burgers' equation, Hybrid obtains the correct solution. However, for $CFL = 1$ the rarefaction fan is not very well resolved. Increasing the CFL-number gives better resolution. However, as for LTS-Roe2 at $CFL = 8$, oscillations impair accuracy.

Figures 5.23 and 5.24 show the same experiment for LTS-Harten. For $CFL = 1, 2$ and 4, LTS-Harten performs well. However unlike LTS-Roe2, LTS-Harten tends to become more diffusive and start having bad resolution of shocks and rarefaction waves as the CFL-number is increased. For this specific problem, the accuracy of LTS-Roe2 is comparable to LTS-Harten.

Test 2

In this test a near vacuum is created in the middle of the tube. This problem is well suited for testing entropy violations due to transonic rarefaction. This problem is very challenging to the stability. Normal time step methods need to be run at low CFL-numbers ($CFL < 1$). The same goes for our LTSS and thus they fail to be stable for this problem

[3]. In [16], the authors also incorporate interactions between waves and are able to get good results up to $CFL = 4$.

Test 3

This test is similar to the first shocktube test we did, but with stronger shocks. Figures 5.25, 5.26, 5.27 and 5.28 show numerical results for $CFL = 1, 2, 4,$ and 8 for LTS-Roe2 and LTS-Harten. LTS-Roe2 has very sharp resolution of the shocks. Especially impressive is its ability to resolve the peak in ρ for $CFL = 8$. LTS-Harten gets more diffusive as CFL is raised and by $CFL = 8$, the peak is severely reduced in height. However, more severely than the other tests, major oscillations occur for LTS-Roe2, accompanied by a big overshoot of the rarefaction fan. A similar overshoot occur for LTS-Harten, but with less oscillations.

Test 4

In this test two strong shock waves travelling in opposite directions interact resulting in three discontinuities. This test is particularly good for evaluating the method's ability to resolve shocks.

Figures 5.29, 5.30, 5.31 and 5.32 show numerical results for $CFL = 1, 2, 4$ and 8 for LTS-Roe2 and LTS-Harten. As CFL is raised, LTS-Roe2 is able to give sharper and sharper resolution of the discontinuities. However for high CFL-number, there is much oscillation. LTS-Harten has poorer resolution of discontinuities, but no oscillations.

Test 5

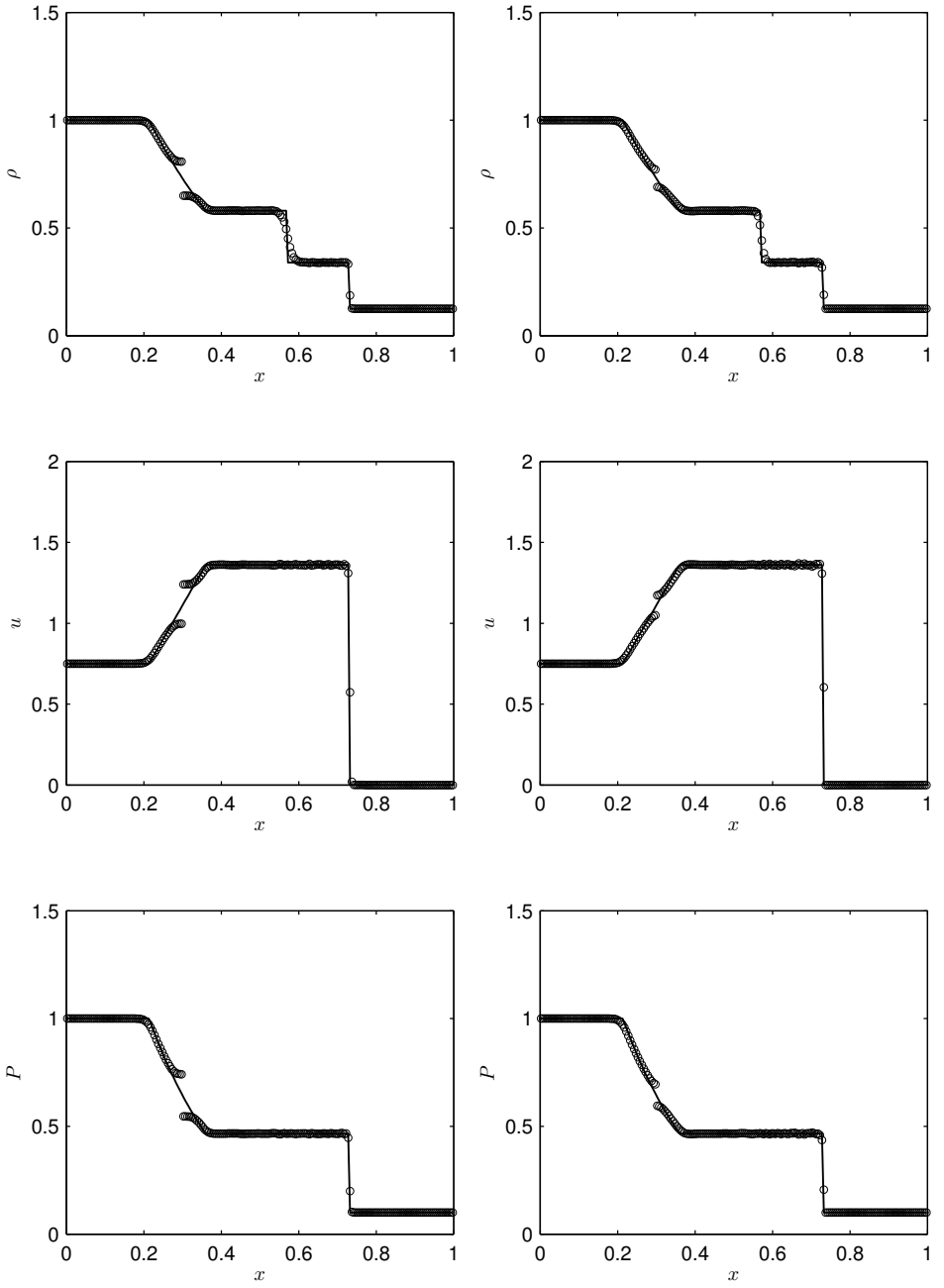
Test 5 is similar to test 3, only for this test, there is a uniform, negative background speed giving a virtually stationary contact discontinuity. This test is designed to test the numerical method's ability to resolve slow moving shocks or stationary discontinuities in addition to overall robustness.

Figure 5.33 and 5.34 show numerical simulations for LTS-Roe2 at $CFL = 1, 2, 4$ and 8. For $CFL = 1$ and 2, the solution is very accurate, with little diffusion. At $CFL = 4$ we start having an overshoot of the rarefaction wave, and oscillations, which become even bigger for $CFL = 8$. Despite this, LTS-Roe2 resolves excellently the peak in ρ. LTS-Harten was only stable for $CFL = 1$, and unstable for higher values.

5.3.4 Thoughts on the oscillations

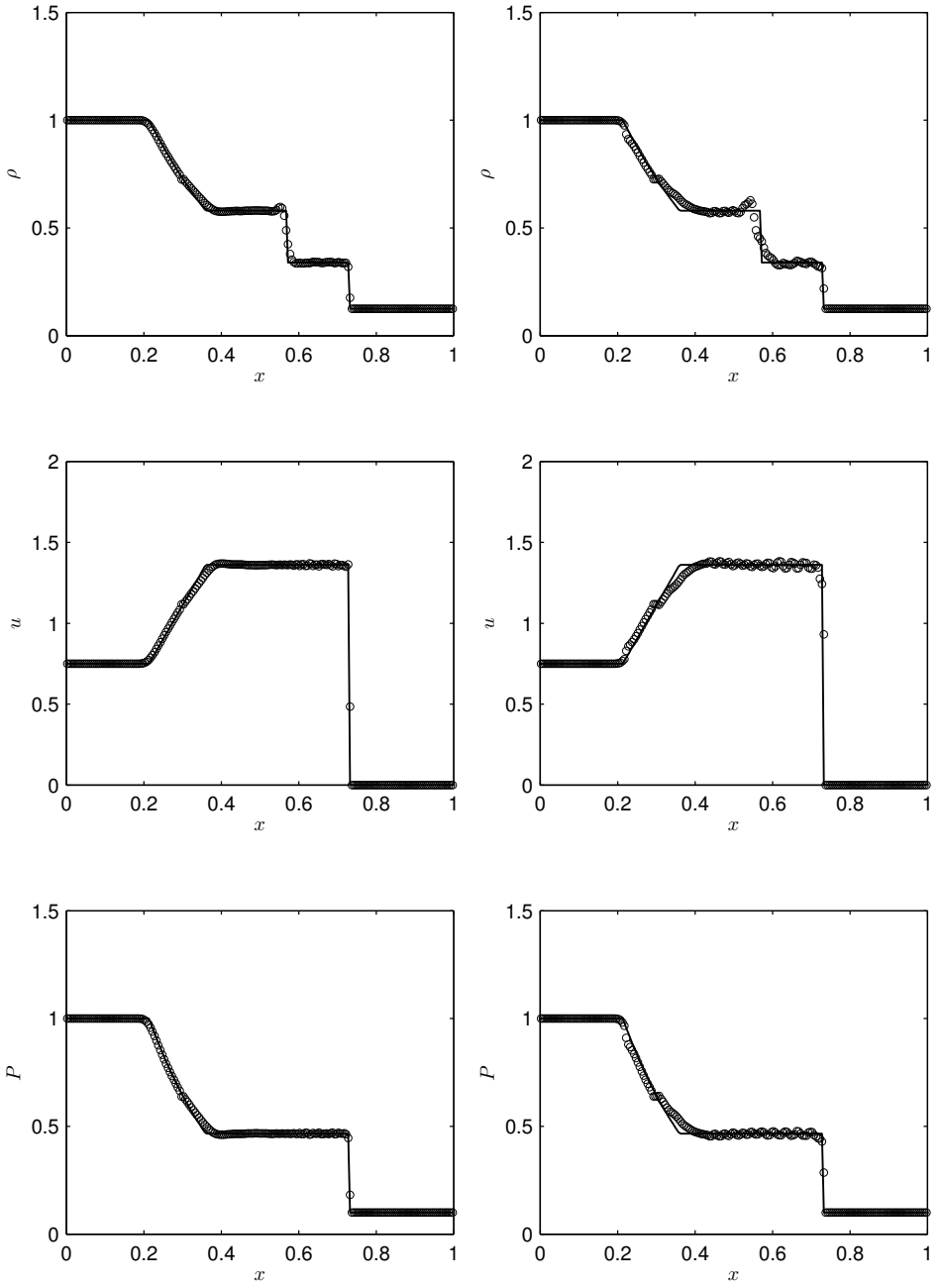
A recurring theme throughout the numerical experiments were the oscillations for high CFL-numbers. These were most severe for LTS-Roe1 and LTS-Roe2. As is noted in [13], the oscillations are likely a consequence of inadequate treatment of interaction between waves. Some attempts at accounting for this have been suggested and done in [12, 13, 16]. A problem with such approaches is that they will invariably make the methods more complex. Another approach to reducing the oscillations, is to forget their origin. Even though LTS-Roe2 and LTS-Harten both linearize interactions between waves, LTS-Harten has much less oscillations. Further, as the grid is refined the oscillations of

LTS-Harten disappear altogether. The key difference between LTS-Harten and LTS-Roe2 lies in the amount of numerical diffusion. Even though interactions are not accounted for, LTS-Harten introduces enough diffusion to damp the oscillations. Thus a possible way of remedying the oscillations of LTS-Roe2, is to introduce numerical diffusion, for instance in a similar way as was done with Hybrid.



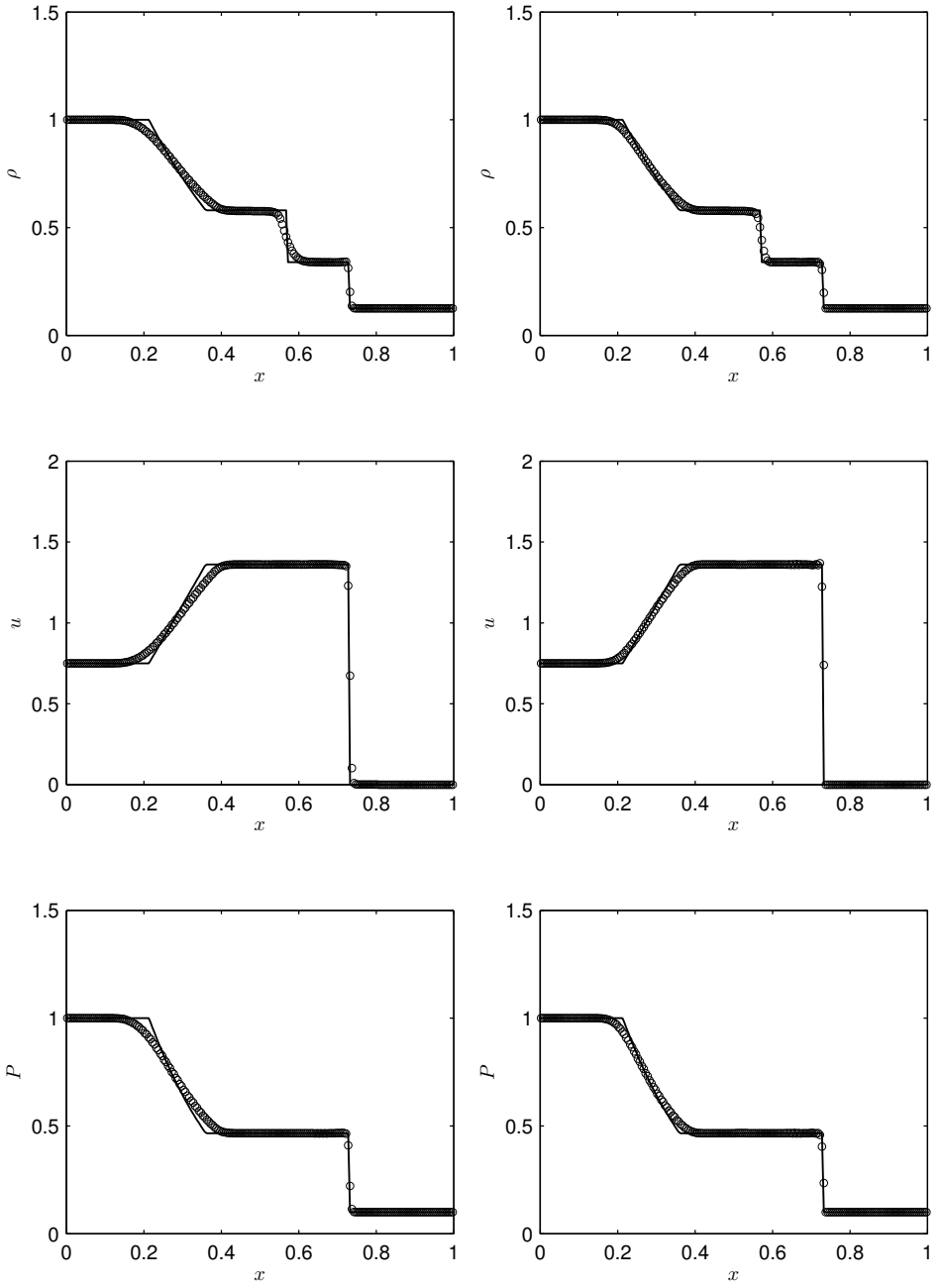
(a) $CFL = 1$ **(b)** $CFL = 2$

Figure 5.19: Numerical solution of Toro's test 1 for $CFL = 1$ and 2 using LTS-Roe2 with $\mathcal{N} = 200$.



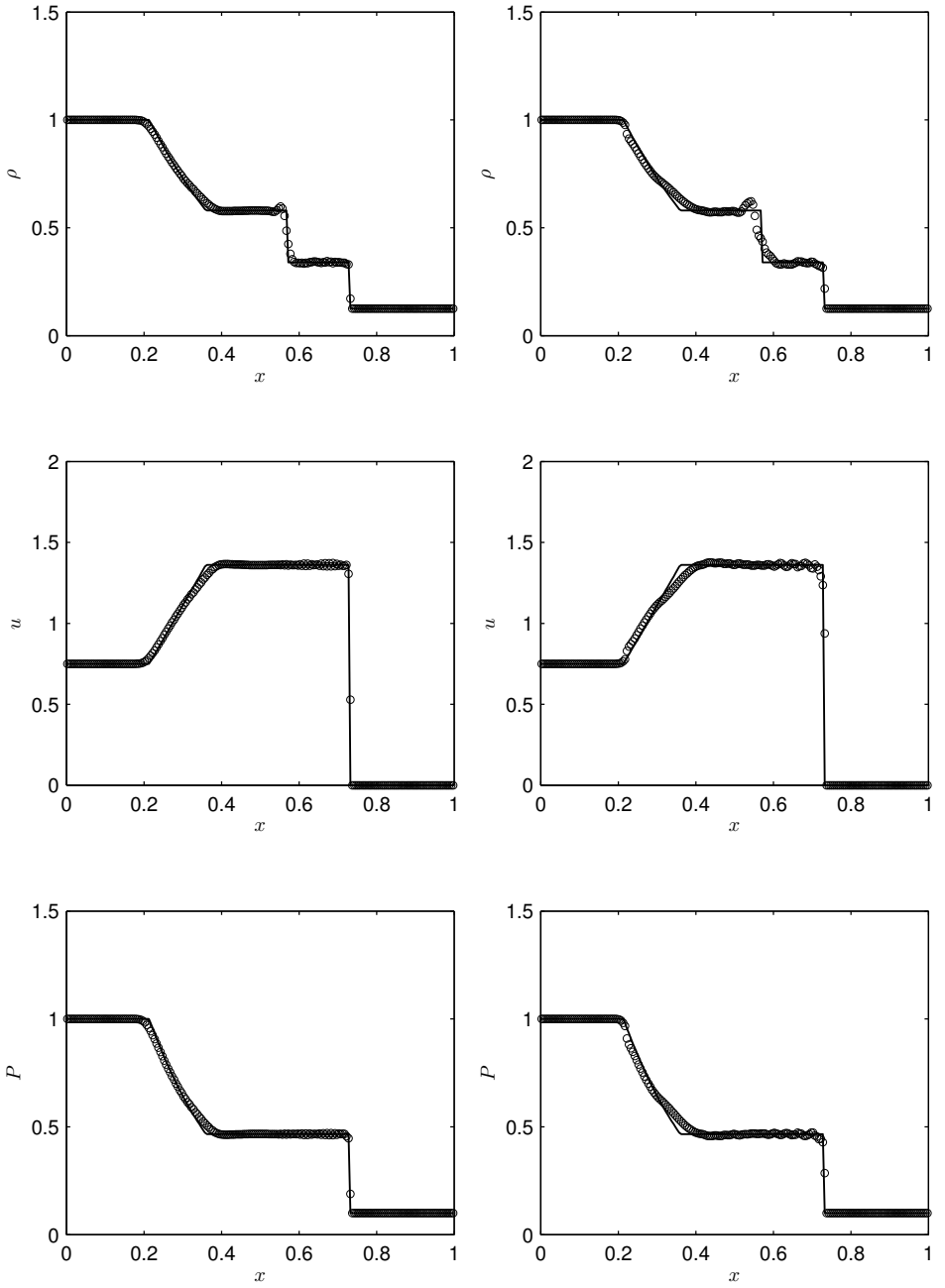
(a) $CFL = 4$ **(b)** $CFL = 8$

Figure 5.20: Numerical solution of Toro's test 1 for $CFL = 4$ and 8 using LTS-Roe2 with $\mathcal{N} = 200$.



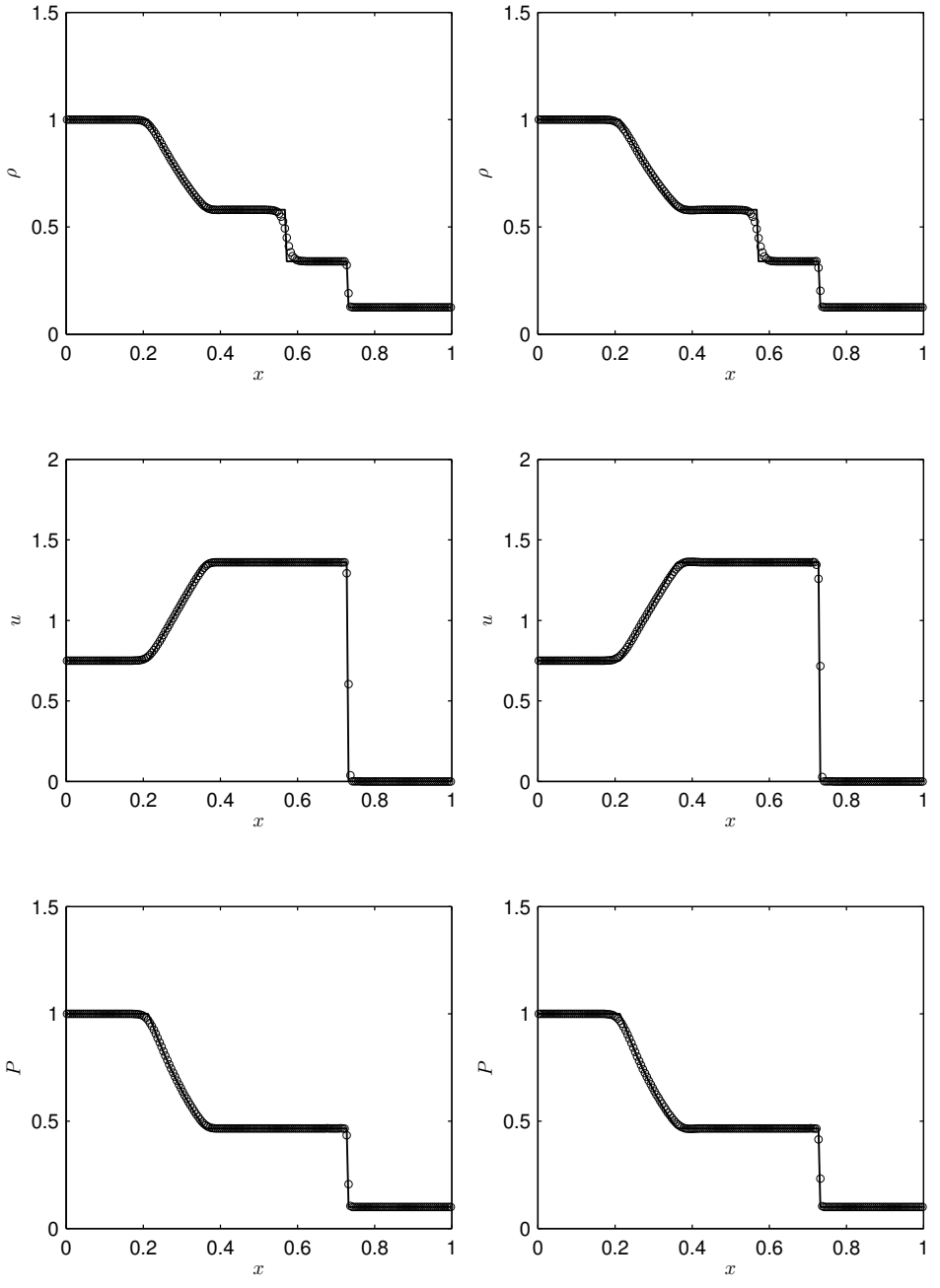
(a) $CFL = 1$ **(b)** $CFL = 2$

Figure 5.21: Numerical solution of Toro's test 1 for $CFL = 1$ and 2 using Hybrid with $\mathcal{N} = 200$.



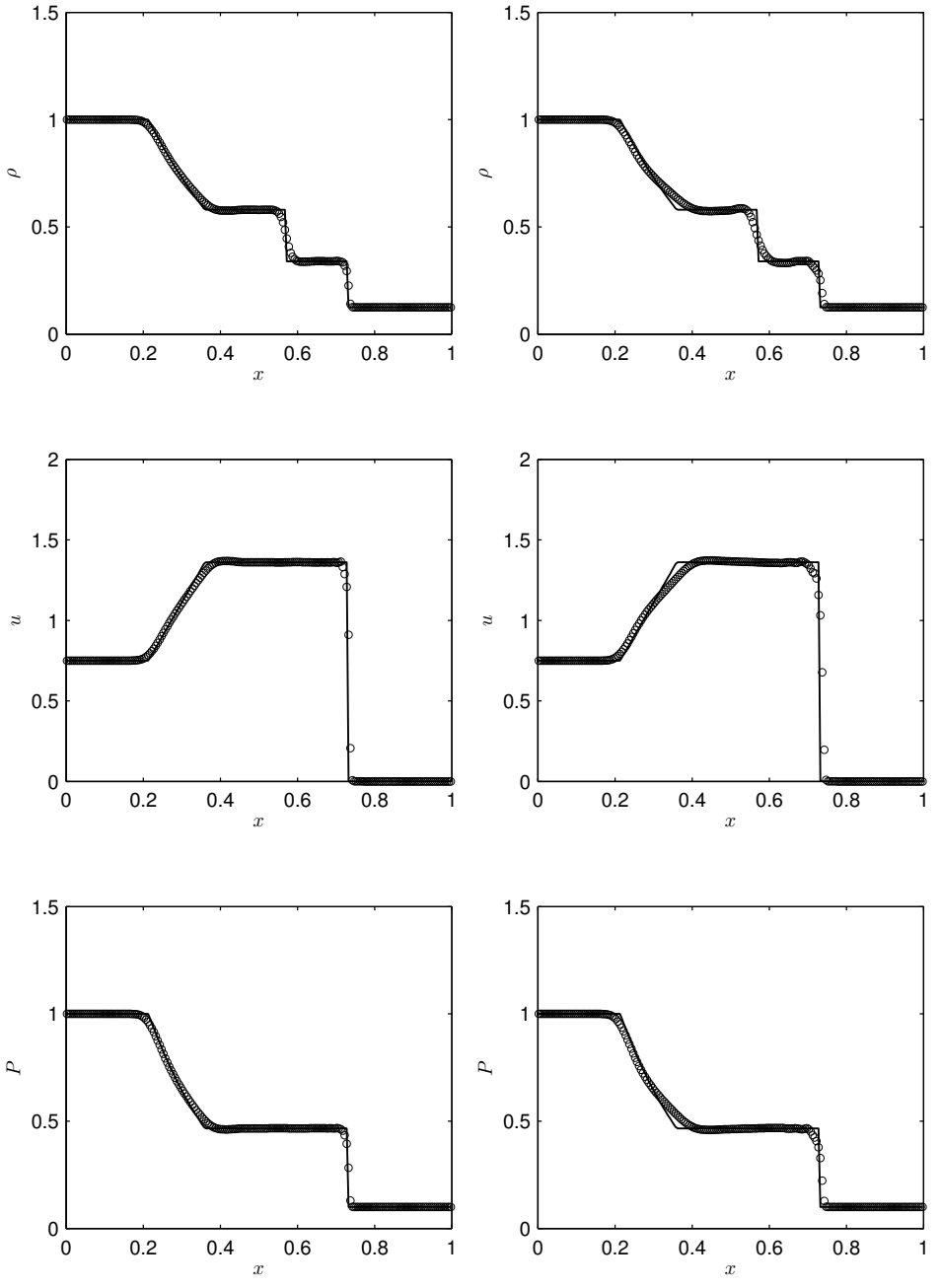
(a) $CFL = 4$ **(b)** $CFL = 8$

Figure 5.22: Numerical solution of Toro's test 1 for $CFL = 4$ and 8 using Hybrid with $\mathcal{N} = 200$.



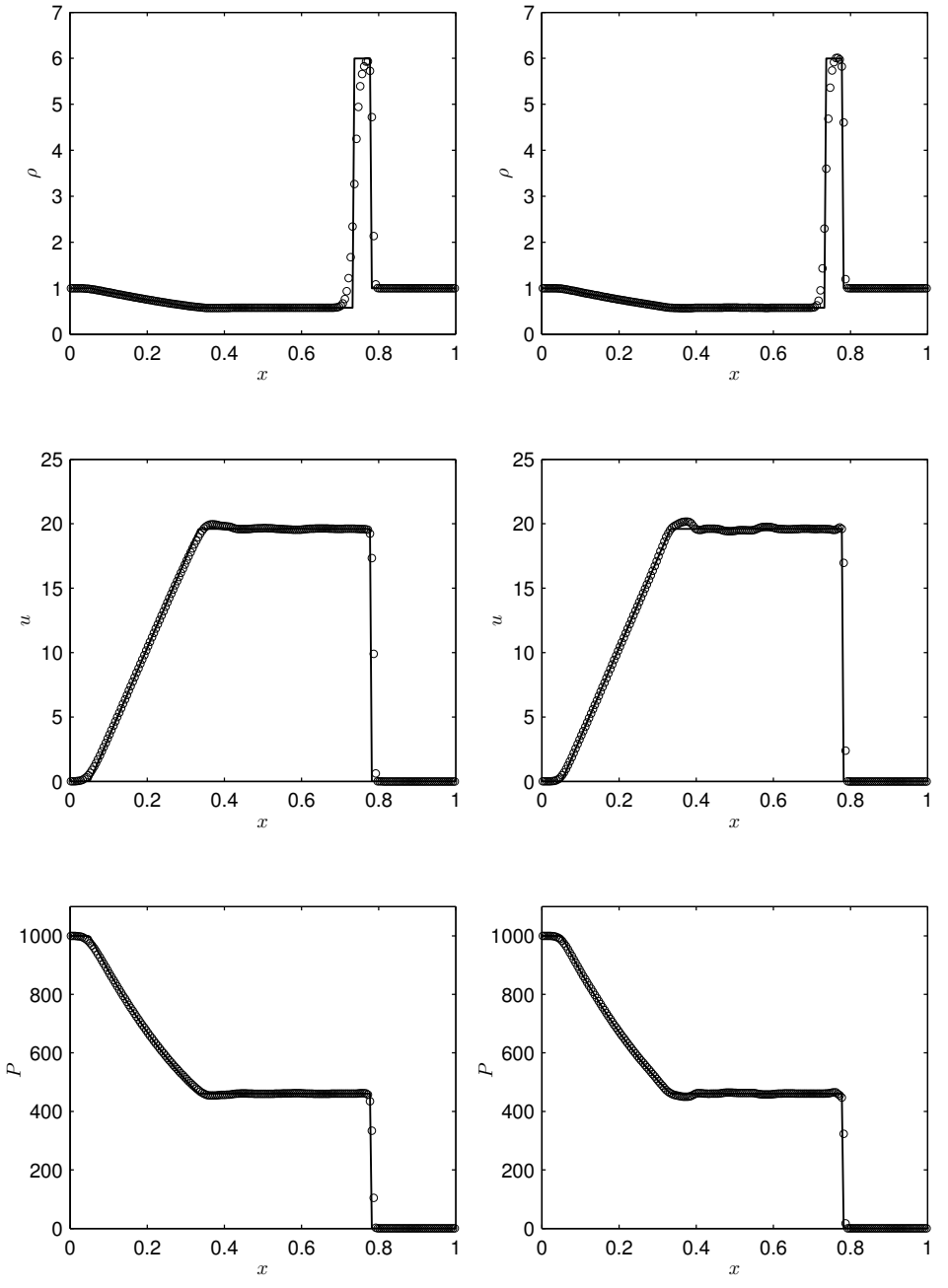
(a) $CFL = 1$ **(b)** $CFL = 2$

Figure 5.23: Numerical solution of Toro's test 1 for $CFL = 1$ and 2 using LTS-Harten with $\mathcal{N} = 200$.



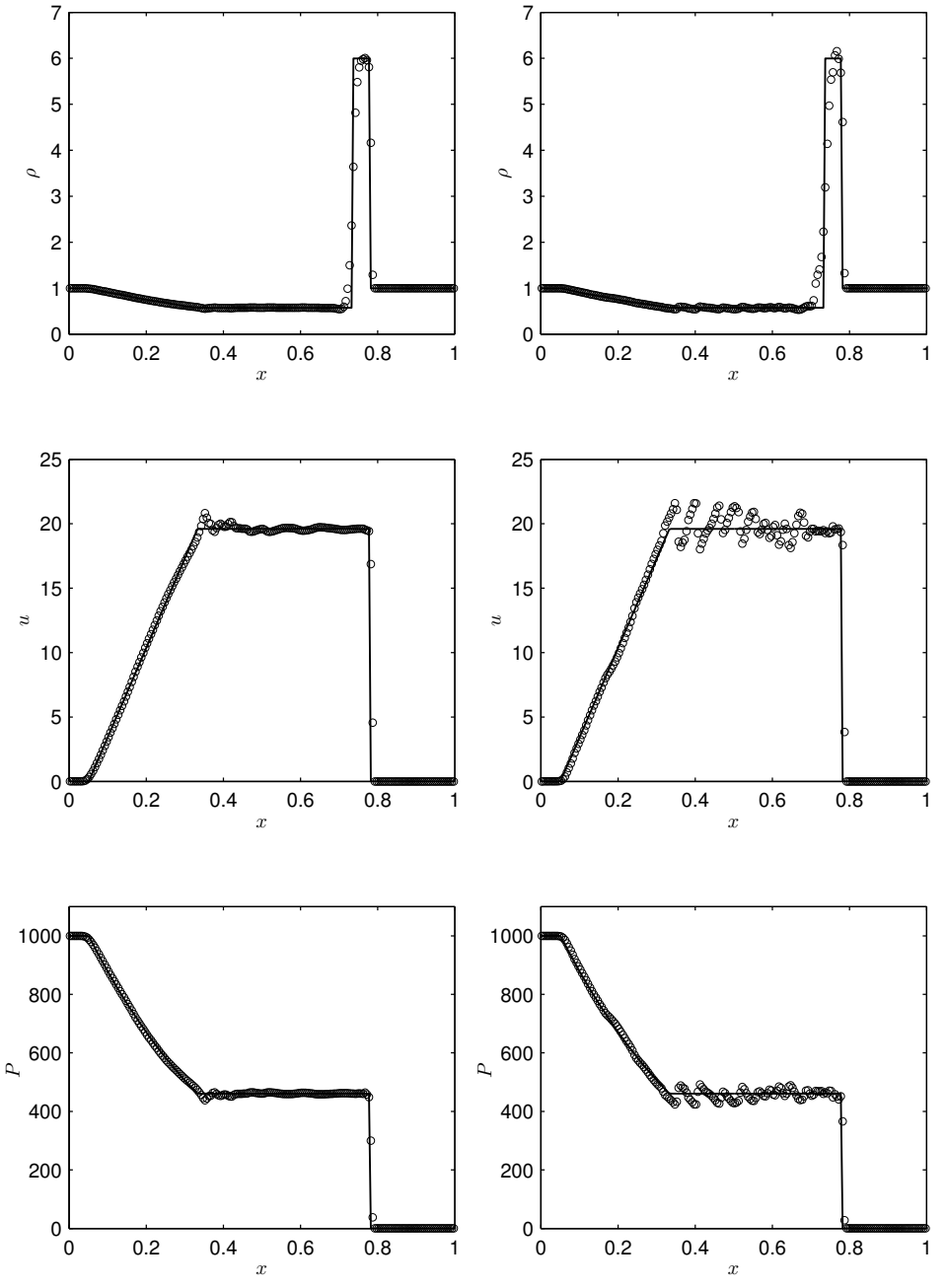
(a) $CFL = 4$ **(b)** $CFL = 8$

Figure 5.24: Numerical solution of Toro's test 1 for $CFL = 4$ and 8 using LTS-Harten with $\mathcal{N} = 200$.



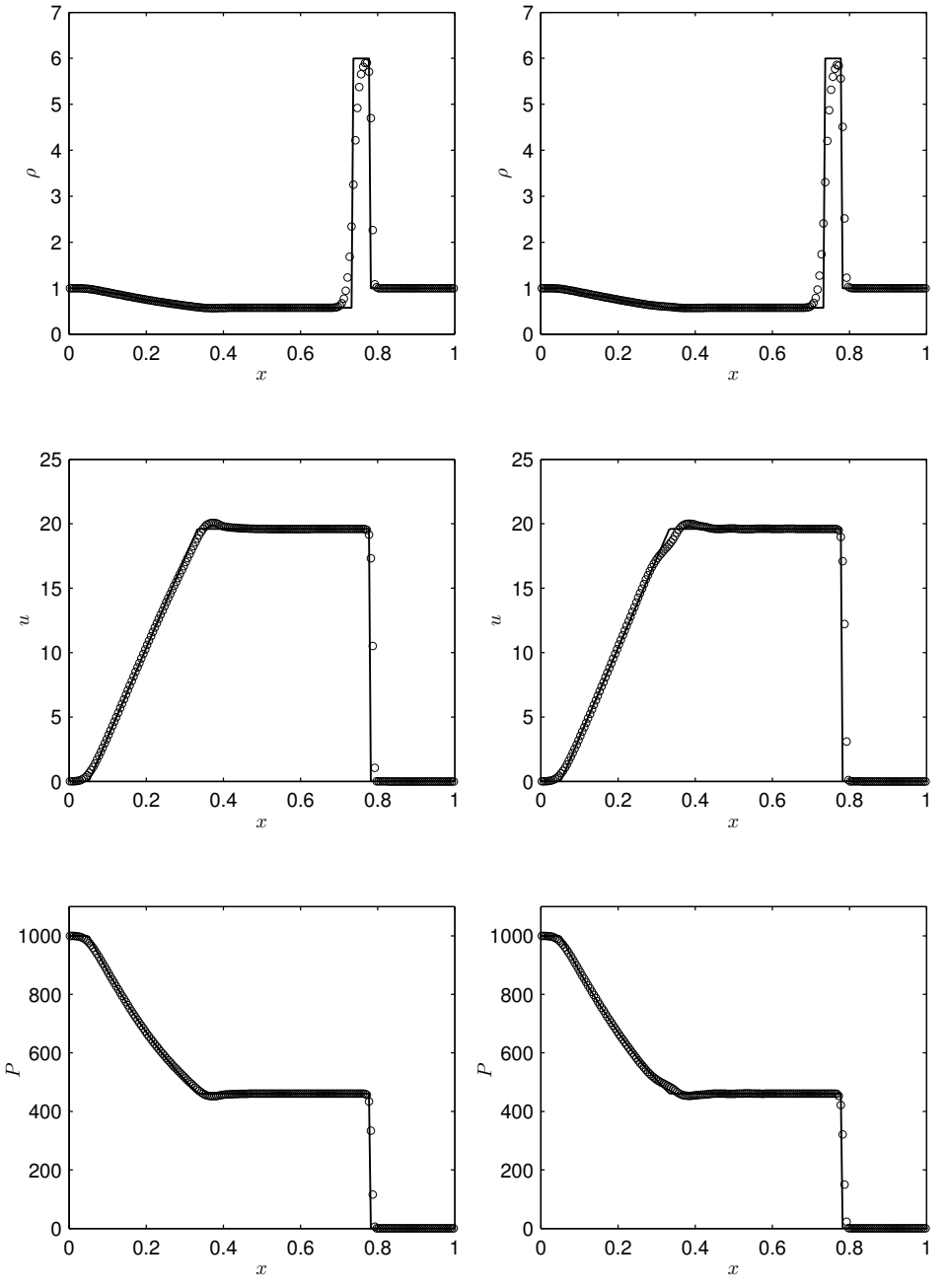
(a) $CFL = 1$ **(b)** $CFL = 2$

Figure 5.25: Numerical solution of Toro's test 3 for $CFL = 1$ and 2 using the LTS-Roe2 scheme with $\mathcal{N} = 200$.



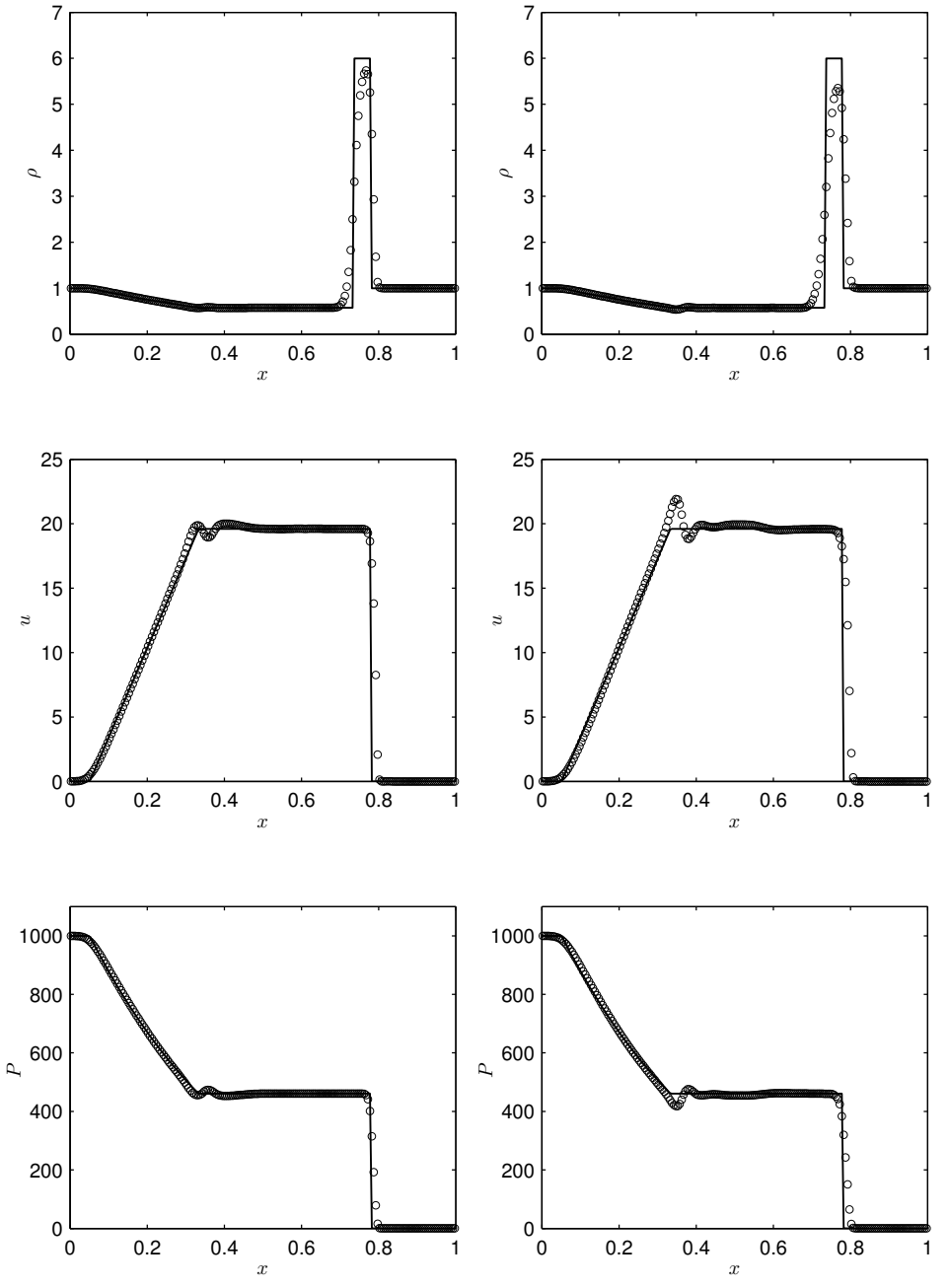
(a) $CFL = 4$ **(b)** $CFL = 8$

Figure 5.26: Numerical solution of Toro's test 3 for $CFL = 4$ and 8 using LTS-Roe2 with $\mathcal{N} = 200$.



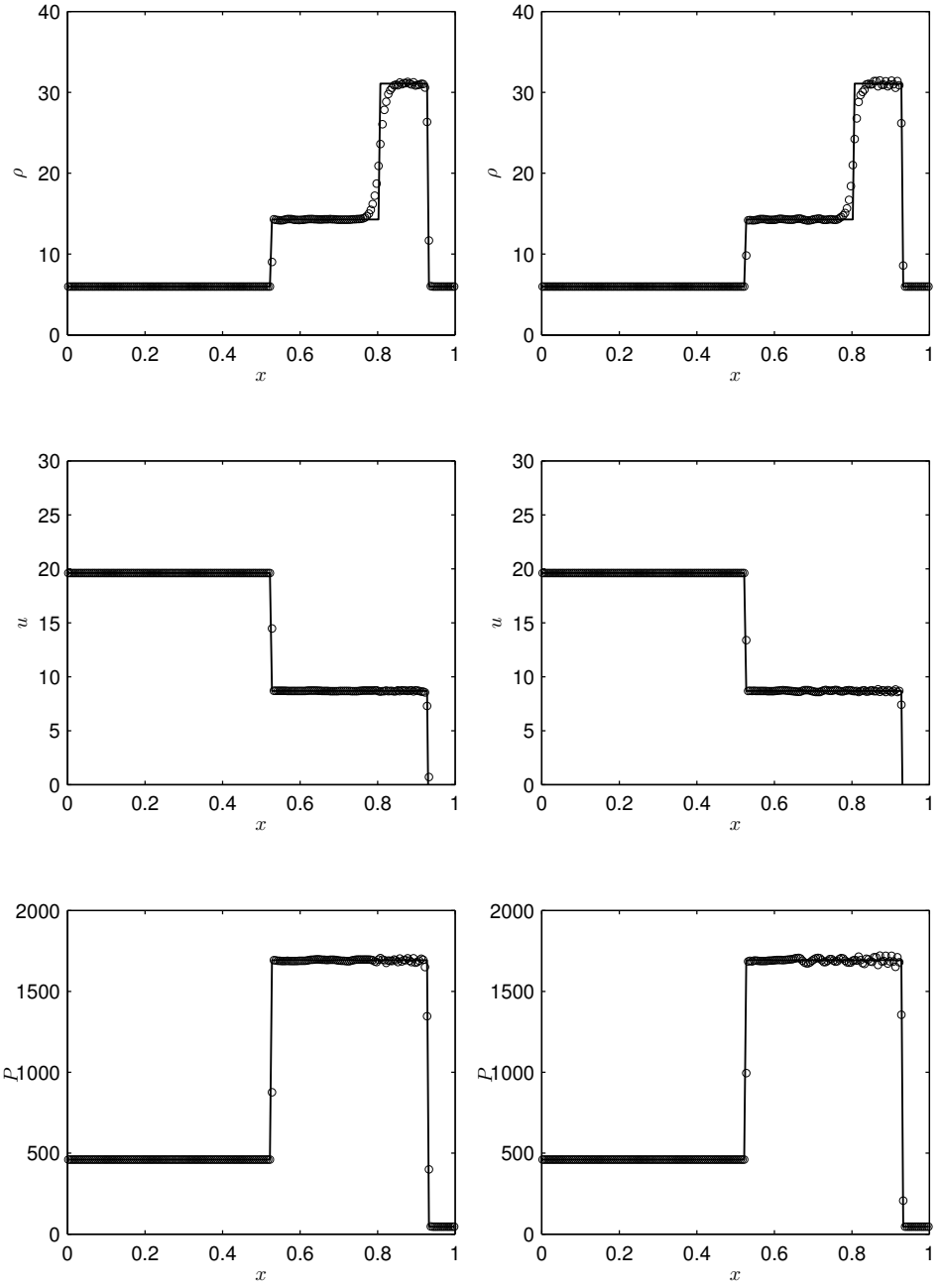
(a) $CFL = 1$ **(b)** $CFL = 2$

Figure 5.27: Numerical solution of Toro's test 3 for $CFL = 1$ and 2 using the LTS-Harten scheme with $\mathcal{N} = 200$.



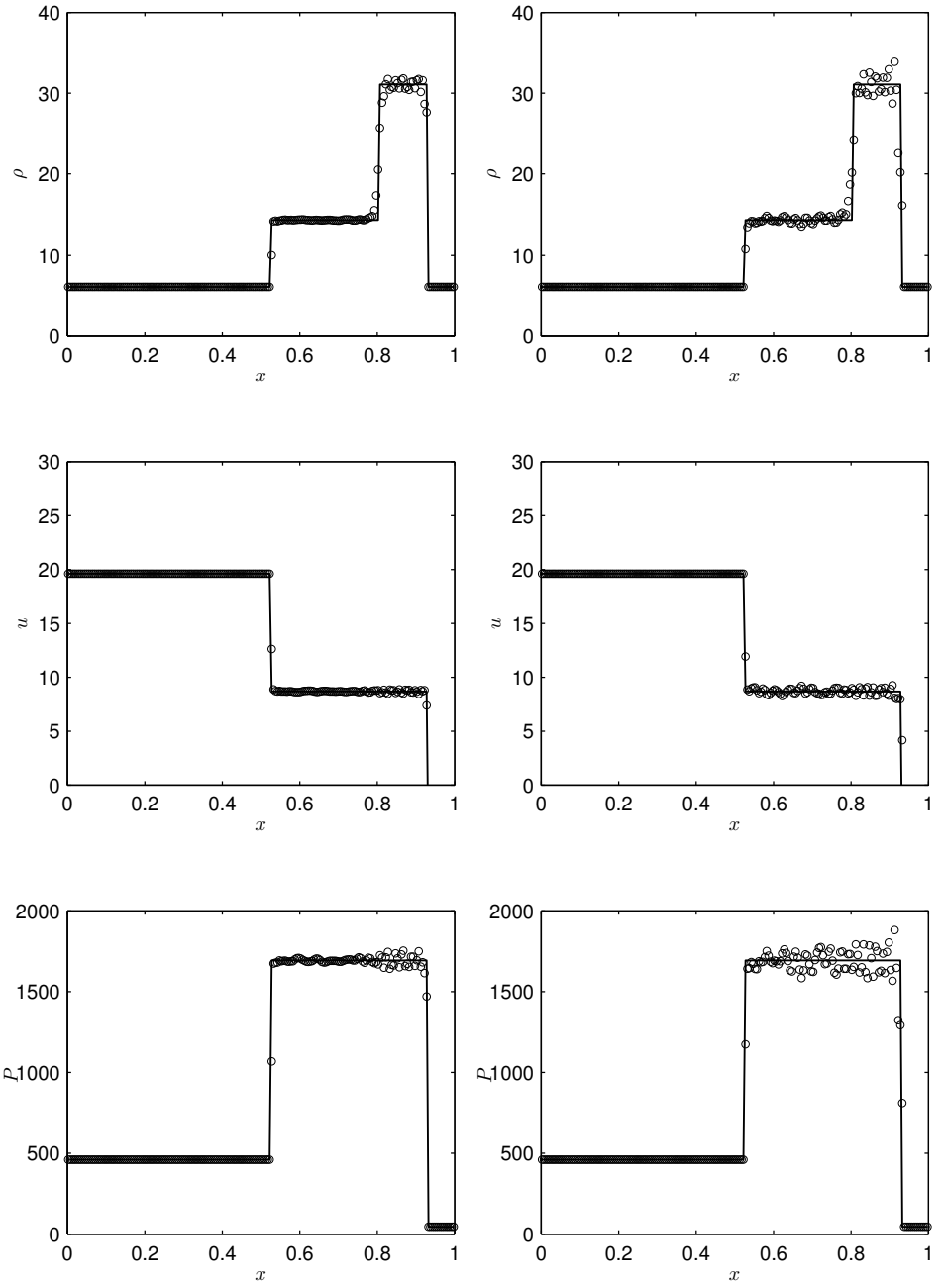
(a) $CFL = 4$ **(b)** $CFL = 8$

Figure 5.28: Numerical solution of Toro's test 3 for $CFL = 4$ and 8 using LTS-Harten with $\mathcal{N} = 200$.



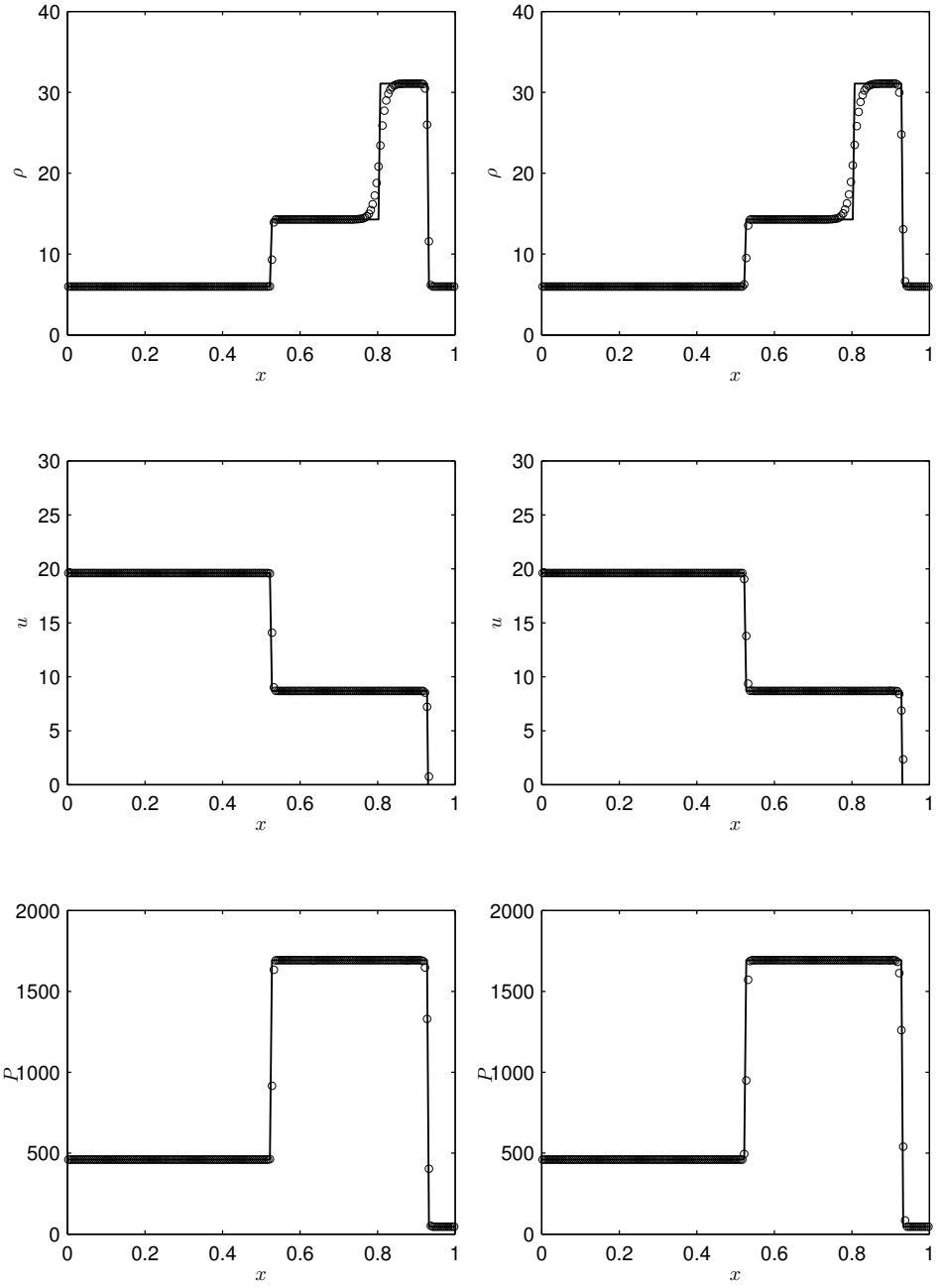
(a) $CFL = 1$ **(b)** $CFL = 2$

Figure 5.29: Numerical solution of Toro's test 4 for $CFL = 1$ and 2 using the LTS-Roe2 scheme with $\mathcal{N} = 200$.



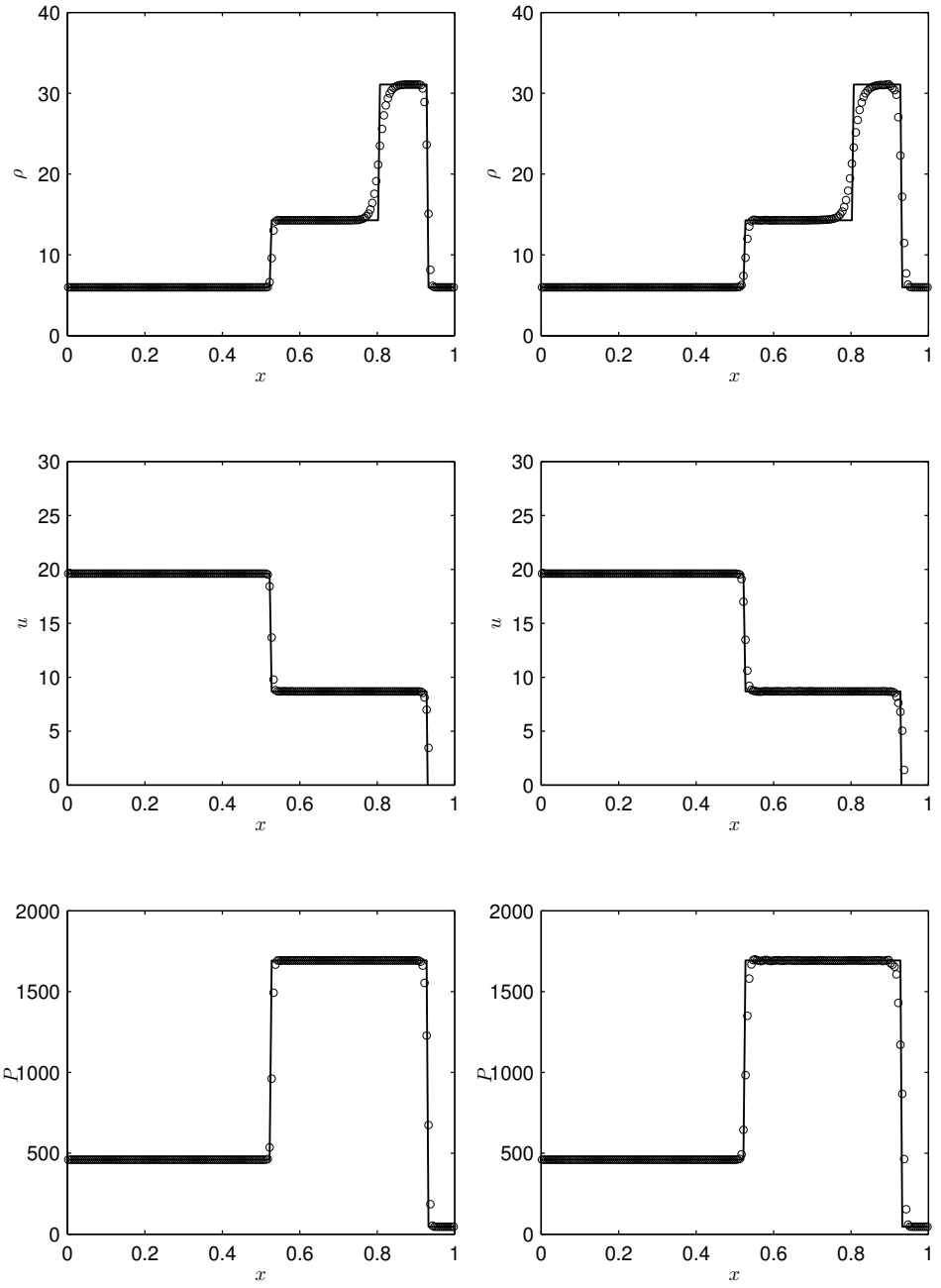
(a) $CFL = 4$ **(b)** $CFL = 8$

Figure 5.30: Numerical solution of Toro's test 4 for $CFL = 4$ and 8 using LTS-Roe2 with $\mathcal{N} = 200$.



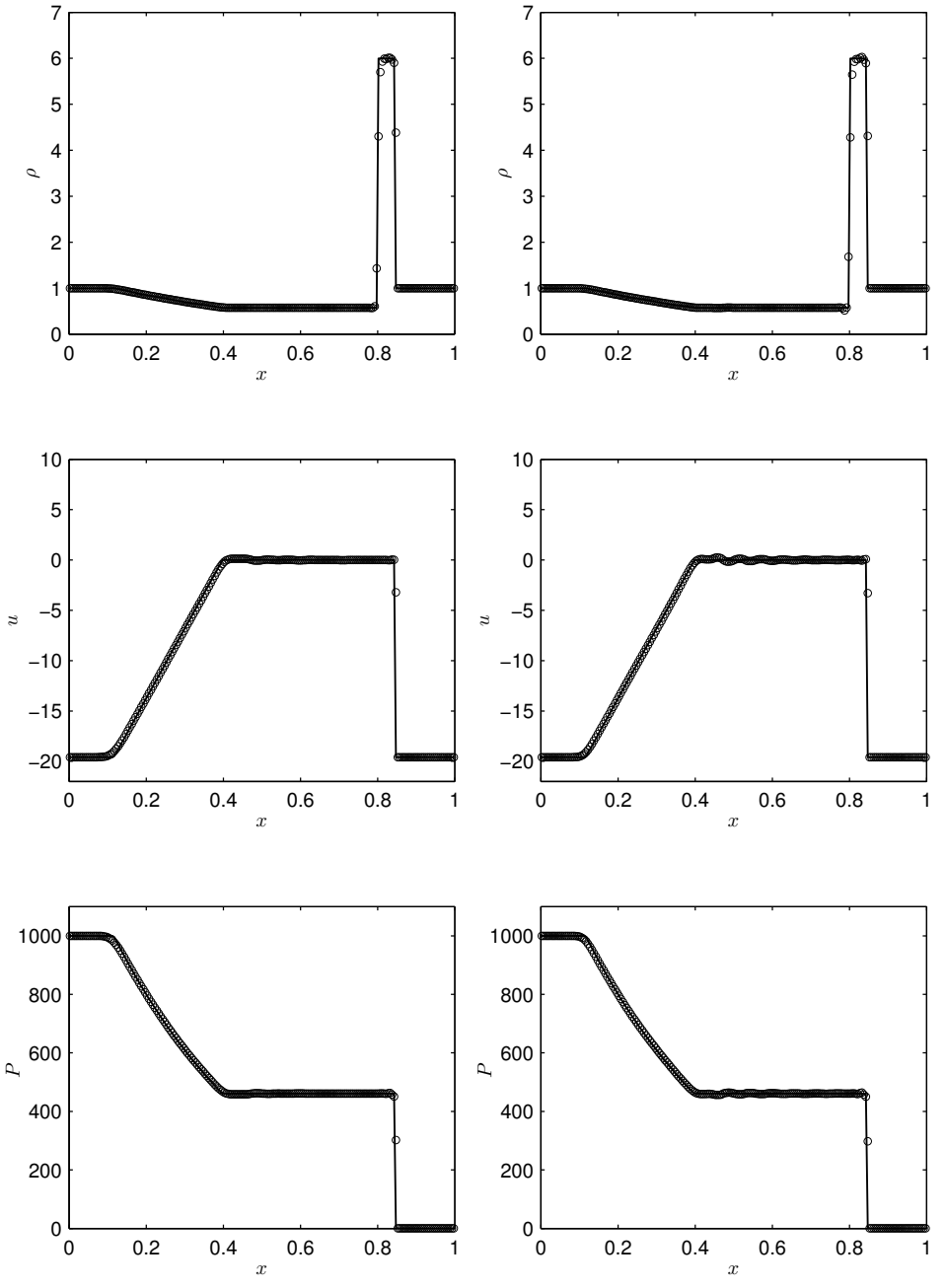
(a) $CFL = 1$ **(b)** $CFL = 2$

Figure 5.31: Numerical solution of Toro's test 4 for $CFL = 1$ and 2 using LTS-Harten with $\mathcal{N} = 200$.



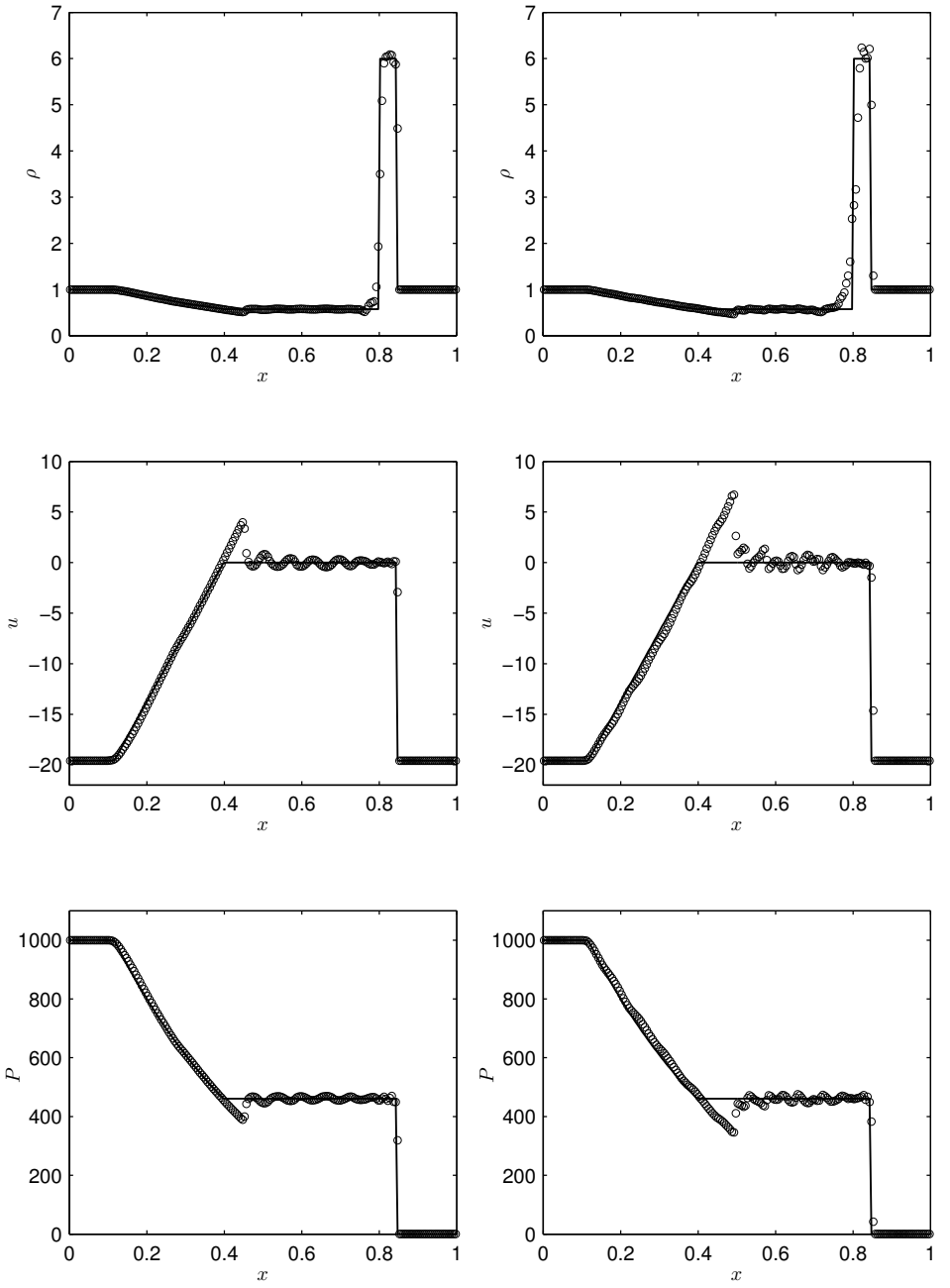
(a) $CFL = 4$ **(b)** $CFL = 8$

Figure 5.32: Numerical solution of Toro's test 4 for $CFL = 4$ and 8 using LTS-Harten with $\mathcal{N} = 200$.



(a) $CFL = 1$ **(b)** $CFL = 2$

Figure 5.33: Numerical solution of Toro's test 5 for $CFL = 1$ and 2 using the LTS-Roe2 scheme with $\mathcal{N} = 200$.



(a) $CFL = 4$ **(b)** $CFL = 8$

Figure 5.34: Numerical solution of Toro's test 5 for $CFL = 4$ and 8 using LTS-Roe2 with $\mathcal{N} = 200$.

Chapter 6

Conclusions and future prospects

6.1 Conclusions

This thesis has treated high-resolution large time-step schemes for hyperbolic conservation laws. The main conclusions of this work are given below.

6.1.1 Framework for high-resolution large time-step schemes

Conditions for a conservative, total variation diminishing, and consistent large time-step scheme were derived, conditions based on previous literature [6, 9, 14]. Further it was proved that all consistent and conservative LTSS can be generalized to second-order accuracy away from discontinuities by a modified flux approach. Such an approach was shown to be TVD under a supplementary condition on the coefficients of the LTSS. The full set of criteria constitutes a new framework of sufficient conditions for general high-resolution large time-step schemes.

6.1.2 Second order large-time step schemes

By application of the framework of conditions, two new second-order large time-step schemes were successfully developed: LTS-Roe2 and Hybrid. Of all TVD LTSS, LTS-Roe2 has the lowest possible numerical diffusion. Moreover, it has zero diffusion for integer local CFL-numbers, making it vulnerable to entropy mistakes. Motivations were given for how random CFL-numbers can increase the accuracy and correct for entropy mistakes, except for transonic rarefaction. Hybrid, being a combination Lax-Friedrichs and LTS-Roe, has nonzero diffusion CFL-numbers close to 0, and it is thus well suited to resolve transonic rarefaction. LTS-Harten, proposed in [6], was examined. Analysis shows that this scheme has nonzero diffusion for all local CFL-numbers and an inaccuracy in the entropy fix suggested in [6] was found.

6.1.3 Assessment of numerical results

Simulations of Burgers' equation with continuous initial conditions showed second-order convergence for all second-order methods. Further, it was shown that using random CFL-numbers increases the accuracy of LTS-Roe1 and LTS-Roe2. Of all methods, LTS-Roe2 has the best accuracy. For discontinuous solutions, LTS-Roe1 and LTS-Roe2 only converge for random CFL-numbers. Generally, LTS-Roe2 has better accuracy than LTS-Roe1, however this difference in accuracy becomes small for large CFL-numbers. As was expected, Hybrid proved able to resolve transonic rarefaction. For the tests with discontinuous initial conditions at high CFL-numbers, LTS-Harten consistently outperforms LTS-Roe2 and Hybrid. The low accuracy of LTS-Roe2 and Hybrid is likely related to the random CFL-numbers not sufficiently correcting for entropy mistakes made at each time-step.

Simulations of the Euler equations show improvement of accuracy with LTS-Roe2 over LTS-Roe1. However, as with the Burgers' equation, when the CFL-number is increased, this difference becomes small. LTS-Roe2 performs well, but at sufficiently high CFL-number, oscillations deteriorate significantly the solution. LTS-Harten has also some oscillations, but of lower frequency. Refinement of the grid shows that these oscillations for LTS-Harten become smaller, while for LTS-Roe they get higher frequency. The general trend is for LTS-Roe to have excellent resolution of sharp peaks and shocks. Our test with LTS-Harten shows that it is very robust. However as was observed in [6], it has a tendency to smear solutions as the CFL-number is increased.

Overall, of all the LTSS, LTS-Harten was found to produce the most accurate results. However, it has a tendency to smear solutions as the CFL-number increased. Thus to achieve the same accuracy of a normal method ($CFL < 1$), more cells are needed. This is probably the reason why normal time-step schemes ($CFL < 1$) are generally preferred and LTS-Harten is rarely used in the literature.

6.2 Future prospects

A large portion of this thesis was dedicated the development of a new framework for high-resolution LTSS. The natural extension of this thesis lies in the application of the framework and the experimentation in development of new LTSS. A concrete research project is to develop a systematic way of adding diffusion to LTS-Roe2 (similar to what was done with Hybrid). Special focus should be put into adding diffusion for local CFL-numbers close to integer values. Doing so will make LTS-Roe far more robust, less dependent on random CFL-numbers, and most likely, more accurate. A big question is how much more diffusion is needed to reduce the spurious oscillations observed for LTS-Roe2. If very little diffusion is needed, then a modified version is likely to be a very efficient and useful scheme.

The thesis has not focused on the efficiency of LTSS. To do so in any meaningful way requires detailed analysis. This should be done by assessing the computational time needed to achieve a certain accuracy at a certain CFL-number using the necessary amount of cells (needed to achieve the accuracy).

Finally, there were stability problems with Toro's second test due to the Roe-linearization.

Thus an avenue for further research is to use or develop other linearizations that are more robust. For instance, it would be interesting to apply the HLLC-method [22] which is known for it robustness.

Bibliography

[1] Aw, A. and Rascle, M. (2000). Resurrection of "second order" models of traffic flow. *SIAM Journal on Applied Mathematics*, 60(3):916–938.

[2] Charney, J. G., Fjörtoft, R., and Von Neumann, J. (1950). Numerical integration of the barotropic vorticity equation. *Tellus A*, 2(4).

[3] Einfeldt, B., Munz, C.-D., Roe, P. L., and Sjögreen, B. (1991). On godunov-type methods near low densities. *Journal of Computational Physics*, 92(2):273–295.

[4] Godunov, S. K. (1959). Difference methods for the numerical calculations of discontinuous solutions of the equations of fluid dynamics. *Matematicheskii Sbornik*, 47:271–306. In Russian, translation in: *US Joint Publ. Res. Service, JPRS*, 7226 (1969).

[5] Harten, A. (1983). High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357 – 393.

[6] Harten, A. (1986). On a large time-step high resolution scheme. *Mathematics of computation*, 46(174):379–399.

[7] Harten, A. (1997). High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 135(2):260–278.

[8] Harten, A., Engquist, B., Osher, S., and Chakravarthy, S. R. (1987). Uniformly high order accurate essentially non-oscillatory schemes, iii. *Journal of Computational Physics*, 71(2):231–303.

[9] Jameson, A. and Lax, P. D. (1986). Conditions for the construction of multi-point total variation diminishing difference schemes. *Applied Numerical Mathematics*, 2(35):335 – 345. Special Issue in Honor of Milt Rose's Sixtieth Birthday.

[10] LeVeque, R. (2002). *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press.

[11] LeVeque, R. J. (1982). Large time step shock-capturing techniques for scalar conservation laws. *SIAM Journal on Numerical Analysis*, 19(6):1091–1109.

[12] LeVeque, R. J. (1984). Convergence of a large time step generalization of godunov's method for conservation laws. *Communications on pure and applied mathematics*, 37(4):463–477.

[13] LeVeque, R. J. (1985). A large time step generalization of godunovs method for systems of conservation laws. *SIAM Journal on Numerical Analysis*, 22(6):1051–1073.

[14] Lindqvist, S. (2014). Large time step schemes, summer report. *Internal SINTEF memo*.

[15] Liska, R. and Wendroff, B. (2003). Comparison of several difference schemes on 1d and 2d test problems for the euler equations. *SIAM Journal on Scientific Computing*, 25(3):995–1017.

[16] Qian, Z. and Lee, C.-H. (2011). A class of large time step godunov schemes for hyperbolic conservation laws and applications. *Journal of Computational Physics*, 230(19):7418–7440.

[17] Qian, Z. and Lee, C.-H. (2012). On large time step tvd scheme for hyperbolic conservation laws and its efficiency evaluation. *Journal of Computational Physics*, 231(21):7415–7430.

[18] Roe, P. L. (1981a). Approximate riemann solvers, parameter vectors and difference schemes. *Journal of Computational Physics*, 43:357–372.

[19] Roe, P. L. (1981b). Approximate riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43(2):357–372.

[20] Sod, G. A. (1978). A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *Journal of Computational Physics*, 27(1):1–31.

[21] Toro, E. F. (2009). *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*. Springer Science & Business Media.

[22] Toro, E. F., Spruce, M., and Speares, W. (1994). Restoration of the contact surface in the hll-riemann solver. *Shock waves*, 4(1):25–34.