Positivity Preservation in the Simulation of Relativistic Laser-Plasma Interaction

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Dipl.-Math. Anke Wortmann

aus
Herford

Tag der mündlichen Prüfung: 17. Juli 2013

Referent: Prof. Dr. Marlis Hochbruck
Korreferent: Prof. Dr. Willy Dörfler

Acknowledgments

I would like to take this opportunity to express my gratitude to all those people who contributed, directly or indirectly, to this work.

I thank my advisor Prof. Dr. Marlis Hochbruck for the opportunity to work on this exciting research project and for taking me with her to her new appointment in Karlsruhe. I would also like to thank Prof. Dr. Willy Dörfler for kindly accepting to be second reviewer of this thesis.

Special thanks also goes to Dr. Götz Lehmann of Heinrich Heine University Düsseldorf for a fruitful collaboration even over the long distance. Also, I would like to thank the Transregio-SFB 18 “Relativistic Laser-Plasma Dynamics”, sub-project B3 “Soliton Formation during Relativistic Laser-Plasma Interaction”, from which the topic of this thesis originated.

Further, my thanks goes to my colleagues who created a pleasant and cheerful atmosphere within our group. Especially I would like to thank PD Dr. Markus Neher for proofreading and valuable hints on my writing.

I thank my family and especially my mother for their boundless support all through my life, without which my pursuit of mathematical studies would not have been possible. I would also like to thank my friends for their company and the diversion after long hours of work.

Finally, I thank Mario Kilies for his love and kind support, which helped me so much through tough times and encouraged me to keep a positive outlook on this thesis and to believe in my scientific abilities.

Contents

1 Introduction

1.1 Motivation

The simulation of relativistic laser-plasma interaction is a very active field of research. The SFB TR18 "Relativistic Laser Plasma Dynamics" investigates the physics of ultra-intense laser interaction with matter. Since experiments are very expensive and time-consuming and the lasers are highly susceptible to minor vibrations from the environment and experiments thus a delicate undertaking, the numerical simulation assists experimental researchers in finding the parameters they are looking for in their experiments. Then the results from simulations and actual experiments can be compared to optimize the model if necessary. That way, many expensive experiments do not have to be carried out because a cheaper computer aided simulation will make a good prediction of what will happen.

While the simulation of electromagnetic phenomena by themselves is rather well understood and also many numerical schemes for nonlinear conservation laws have been developed, the combination of the two for the simulation of laser-plasma interaction is rather fragile. This is the experience that can easily be made in numerical experiments in this field. Hence the appropriate choice as well as a fundamental understanding of the applied numerical methods is extremely important.

The theoretical physics group in Düsseldorf who we are collaborating with, have tried many different approaches to various problems in the simulation of relativistic laser-plasma interaction. The numerical methods have to be chosen according to what effects are to be observed and which model is used. The physicists heavily rely on PIC (particle-in-cell) codes [Puk99] for their simulations because that is an established machinery, which, unfortunately, consume a lot of computing time due to their complexity.

If the density of the plasma is only of minor interest, its numerical solution is often omitted because of the difficulties that arise there. For that case, a promising implementation for high densities using an exponential integrator is presented in [TPLH10]. If only the density modulation, i.e., the difference to some initial background density, is considered, the difficulties we have with the continuity equation do not arise here because negative values are allowed. For the one-dimensional wave-formulation, this has been investigated in [KSH+06].

To gain some insight into the problems involved in the simulation of relativistic laser-plasma interaction, let us shortly go over some attempts of numerically solving such a system from the literature.

In the one-dimensional case, Maxwell's equations can be rewritten in terms of potentials — a vector potential for the magnetic field and a scalar potential for the electric field. With some further model reduction, the equations become much simpler. For this special case, flux-corrected transport (FCT) as introduced in [BB73, Zal79] yields excellent numerical results even when ion movement is considered.[1] Unfortunately, the energy is not conserved very well. The goal of this thesis is thus to enhance energy conservation of the laser-plasma simulations and extend everything to higher dimensions.

After the good results with FCT, we considered relativistic wave breaking. A formulation in Lagrangian coordinates has been described in [LLS07], which leads to a system of ordinary differential equations if the continuity equation is neglected. These can be solved easily, e.g. by the symplectic Euler method. Depending on the parameters

[1] These experiments have been investigated in the Master's thesis [Wor10].

of the equations, there is slow and fast wave-breaking and the Lagrangian system is well suited for these circumstances. When trying to solve the system of partial differential equations — including the continuity equation — directly, we found that FCT does work to the extend that it preserves positivity and remains stable. Unfortunately, the amount of diffusion left in the scheme from the underlying low order method prevents the growth of peaks from the existing waves. These peaks, however, lead to the wave-breaking. So we ended up with a good qualitative recreation of the waves themselves, but not of the height of their peaks. Hence, we cannot find the time when wave-breaking occurs.

One problem that has been of great interest to physicists, is the laser-plasma transition of a laser pulse. The challenge in this case is the time the pulse enters the plasma. The laser excites the plasma and the waves it creates within the transition area, which is usually a rather narrow and steep ascent, can cause tiny wake-breaking. This in turn, leads to severe instabilities in most simulations. The experiences with simulations of relativistic wave-breaking give reason to believe that FCT should yield suitable results here.

1.2 Thesis Outline

The goal of this thesis is a stable, positivity preserving simulation of a hydrodynamic model for relativistic laser-plasma interaction that is competitive with established codes used for more complex models of this problem. The path to describe the theory, problems and solution to this is as follows.

In chapter 2, we shortly review Maxwell's equations and their numerical solution. We are concerned with laser light, which is described by Maxwell's equations. In one space dimension, the wave formulation is a favorable way of stating the equations, while for higher dimensions, the Yee scheme is the most widely used approach for the numerical approximation. For both cases, we examine the extraordinary properties of the so-called *magic time step*.

Chapter 3 is divided into two parts. First we discuss the theory of hyperbolic conservation laws and the many problems that arise. The main problem is the formation of shocks even from smooth initial data. Afterwards we give a brief overview of numerical methods for this class of partial differential equations and their analysis. We review classical stability analysis, illustrate their insufficiency for nonlinear problems and discuss better suitable notions of stability.

Chapter 4 describes models and methods from computational physics: the derivation of the plasma equation from the Vlasov equation, the Boris push for an appropriate treatment of the Lorentz force as well as a discussion of unit systems and relativistic equations. All of these are fundamental for this thesis and are thus collected here for the sake of readability.

In chapter 5, we introduce flux-corrected transport (FCT) algorithms that will help us in the numerical solution of hyperbolic conservation laws. The goal here is to avoid negative densities, which can occur in numerical simulations and lead to instabilities. The idea is to combine monotone diffusive schemes with higher order methods, which are not free of spurious oscillations, into a new method that preserves positivity without the great extent of numerical diffusion. The scheme is easily reformulated for multidimensional problems and enhanced by strong stability preserving Runge-Kutta methods for the time integration.

We then turn to applications in relativistic laser-plasma interaction. We formulate a complete algorithm for the simulation of a vacuum-plasma transition for one- and

two-dimensional examples in chapters 6 and 7, where we also show extensive numerical examples. Comparisons to established codes are carried out to show the accuracy as well as the saving of computational time.

Finally, we give a short conclusion and outlook in chapter 8.

2 Numerical Solution of Maxwell's Equations

We are going to be looking at laser-plasma interaction. Laser is light, which consists of electromagnetic waves. Those are described by *Maxwell's equations*. A detailed description of Maxwell's equations for different kind of media can be found e.g. in [Jac99] or [TH05].

For this thesis, we can restrict ourselves to the case of homogeneous, isotropic media. We assume sufficient smoothness of all functions, so all derivatives are defined. Maxwell's equations in cgs units then read

$$\frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B} - \frac{4\pi}{c}\boldsymbol{j} \tag{2.1a}$$

$$\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E} \tag{2.1b}$$

$$\nabla \cdot \boldsymbol{E} = 4\pi\rho \tag{2.1c}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{2.1d}$$

on some bounded Lipschitz domain $\Omega \subset \mathbb{R}^3$, subject to appropriate initial and boundary conditions. $\boldsymbol{E} : \Omega \times [0,\infty) \to \mathbb{R}^3$ is the electric and $\boldsymbol{B} : \Omega \times [0,\infty) \to \mathbb{R}^3$ the magnetic field intensity, $\boldsymbol{j} : \Omega \times [0,\infty) \to \mathbb{R}^3$ is the electric current density, $\rho : \Omega \times [0,\infty) \to \mathbb{R}$ the electric charge density and c the speed of light.

We denote the components of any vector-valued function by subscripts x, y and z, i.e.,

$$\boldsymbol{F}(\boldsymbol{x},t) = \begin{pmatrix} F_x(\boldsymbol{x},t) \\ F_y(\boldsymbol{x},t) \\ F_z(\boldsymbol{x},t) \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Equation (2.1a) is called *Ampère's law*, (2.1b) *Faraday's law*. They state the interaction between electric and magnetic fields. Equation (2.1c) is *Gauss's law*. It describes the electric flux through a closed surface while (2.1d) states the absence of magnetic monopoles.

Recall that $\nabla = \begin{bmatrix} \partial_x & \partial_y & \partial_z \end{bmatrix}^T$ and $\nabla\cdot$ is the divergence operator, $\nabla\times$ the curl operator.

If we write out the curl operators in (2.1a) and (2.1b), we obtain the following system of coupled scalar equations:

$$\begin{aligned}
\frac{1}{c}\frac{\partial E_x}{\partial t} &= \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} - \frac{4\pi}{c} j_x \\
\frac{1}{c}\frac{\partial E_y}{\partial t} &= \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} - \frac{4\pi}{c} j_y \\
\frac{1}{c}\frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} - \frac{4\pi}{c} j_z \\
\frac{1}{c}\frac{\partial B_x}{\partial t} &= \frac{\partial E_y}{\partial z} - \frac{\partial E_z}{\partial y} \\
\frac{1}{c}\frac{\partial B_y}{\partial t} &= \frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} \\
\frac{1}{c}\frac{\partial B_z}{\partial t} &= \frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x}
\end{aligned}$$

Since we have eight equations for only six unknowns, we have to discuss the existence and uniqueness of solutions for (2.1).

Lemma 2.1. *A smooth solution of* (2.1) *can only exist if the* compatibility condition

$$\partial_t \rho + \nabla \cdot \boldsymbol{j} = 0 \tag{2.2}$$

is satisfied.

Proof.

$$\partial_t \rho = \partial_t \left(\frac{1}{4\pi} \nabla \cdot \boldsymbol{E} \right) = \frac{1}{4\pi} \nabla \cdot (\partial_t \boldsymbol{E}) = \frac{1}{4\pi} \nabla \cdot \left(c \nabla \times \boldsymbol{B} - c \frac{4\pi}{c} \boldsymbol{j} \right) = -\nabla \cdot \boldsymbol{j}$$

because $\nabla \cdot (\nabla \times \boldsymbol{F}) = 0$ for smooth vector fields $\boldsymbol{F}$. □

Lemma 2.2. *Let $\boldsymbol{E}(\boldsymbol{x},t)$ and $\boldsymbol{B}(\boldsymbol{x},t)$ be a smooth solution of* (2.1a)*,* (2.1b) *and* (2.2)*. If the solution satisfies* (2.1c) *and* (2.1d) *at $t = 0$, then this holds for all $t \geq 0$.*

Proof. For Gauss' law, we have

$$\partial_t (\nabla \cdot \boldsymbol{E} - 4\pi\rho) = \nabla \cdot (c \nabla \times \boldsymbol{B} - 4\pi \boldsymbol{j}) - (-4\pi \nabla \cdot \boldsymbol{j}) = c \nabla \cdot (\nabla \times \boldsymbol{B}) = 0,$$

so $\nabla \cdot \boldsymbol{E} - 4\pi\rho$ is constant in time and thus stays zero.
For the magnetic counterpart we have

$$\partial_t (\nabla \cdot \boldsymbol{B}) = -c \nabla \cdot (\nabla \times \boldsymbol{E}) = 0.$$ □

This tells us that we do not have to consider (2.1c) and (2.1d) as long as we ensure they are fulfilled initially.

Finally, we cite a result on the uniqueness of solutions of

$$\frac{1}{c} \frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B} \tag{2.3a}$$

$$\frac{1}{c} \frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E} \tag{2.3b}$$

on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^3$ with initial conditions $\boldsymbol{E}(\boldsymbol{x},0) = \boldsymbol{E}_0(\boldsymbol{x})$ and $B(\boldsymbol{x},0) = B_0(\boldsymbol{x})$ and boundary conditions $\boldsymbol{E}(\boldsymbol{x},t) \times \boldsymbol{\nu}(\boldsymbol{x}) = \boldsymbol{0}$ and $\boldsymbol{B}(\boldsymbol{x},t) \cdot \boldsymbol{\nu}(\boldsymbol{x}) = 0$ on $\Gamma = \partial\Omega$, where $\boldsymbol{\nu}(\boldsymbol{x})$ is the outer normal (cf. [PR00] or [Jah10]):

Theorem 2.3. *If under the above assumptions,*

$$\boldsymbol{B}_0(\boldsymbol{x}) \in H(\mathrm{curl}, \Omega) := \{ \boldsymbol{F} \in L^2(\Omega)^3 : \nabla \times \boldsymbol{F} \in L^2(\Omega)^3 \}$$

and

$$\boldsymbol{E}_0(\boldsymbol{x}) \in H_0(\mathrm{curl}, \Omega) := \{ \boldsymbol{F} \in H(\mathrm{curl}, \Omega) : \boldsymbol{F} \times \boldsymbol{\nu} = \boldsymbol{0} \text{ on } \Gamma \},$$

then (2.3) *has a unique solution $\boldsymbol{E}(\boldsymbol{x},t)$, $\boldsymbol{B}(\boldsymbol{x},t)$.*

2.1 The Wave Equation and the Magic Time Step

In some cases we might not be interested in both fields, but only in the effects they might have. Let us consider free space where $\boldsymbol{j} \equiv \boldsymbol{0}$ and no charge, i.e., $\rho = 0$. If we take the derivative with respect to time t in either of equations (2.1) and then substitute the other, we end up with a wave equation for one of the fields. For example,

$$\frac{1}{c} \frac{\partial}{\partial t} \frac{1}{c} \frac{\partial \boldsymbol{E}}{\partial t} = \frac{1}{c} \frac{\partial}{\partial t} (\nabla \times \boldsymbol{B}) \tag{2.4}$$

$$\iff \quad \frac{1}{c^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = \nabla \times \frac{1}{c} \frac{\partial \boldsymbol{B}}{\partial t} = \nabla \times (-\nabla \times \boldsymbol{E}) \tag{2.5}$$

$$= -\nabla \times (\nabla \times \boldsymbol{E}) = \Delta \boldsymbol{E} - \nabla (\nabla \cdot \boldsymbol{E}) = \Delta \boldsymbol{E}. \tag{2.6}$$

The same can be done for the magnetic field $\boldsymbol{B}$.

In the one-dimensional case where derivatives with respect to y and z are set to zero, this reduces to

$$\frac{\partial^2 \boldsymbol{E}}{\partial t^2} = c^2 \frac{\partial^2 \boldsymbol{E}}{\partial x^2}. \tag{2.7}$$

This way we have reduced the system of two equations to one wave equation.

Numerical Scheme for the Wave Equation

We now discuss the numerical solution of (2.7). Suppose we use central finite differences for the approximation on both sides, i.e.,

$$\frac{\boldsymbol{E}_j^{n+1} - 2\boldsymbol{E}_j^n + \boldsymbol{E}_j^{n-1}}{(\Delta t)^2} + \mathcal{O}\left((\Delta t)^2\right) = c^2 \frac{\boldsymbol{E}_{j+1}^n - 2\boldsymbol{E}_j^n + \boldsymbol{E}_{j-1}^n}{(\Delta x)^2} + \mathcal{O}\left((\Delta x)^2\right), \tag{2.8}$$

where subscripts j denote the spatial grid point $x_j = j\Delta x$ and superscripts n denote the time level $t_n = n\Delta t$. Solving (2.8) for the newest value $\boldsymbol{E}_j^{n+1}$, we obtain

$$\boldsymbol{E}_j^{n+1} = c^2(\Delta t)^2 \frac{\boldsymbol{E}_{j+1}^n - 2\boldsymbol{E}_j^n + \boldsymbol{E}_{j-1}^n}{(\Delta x)^2} + 2\boldsymbol{E}_j^n - \boldsymbol{E}_j^{n-1}, \tag{2.9}$$

a fully explicit second order accurate scheme. It is widely known as the *leap-frog scheme*.

A particularly interesting case is $c\frac{\Delta t}{\Delta x} = 1$, where the scheme reduces to

$$\begin{aligned} \boldsymbol{E}_j^{n+1} &= \boldsymbol{E}_{j+1}^n - 2\boldsymbol{E}_j^n + \boldsymbol{E}_{j-1}^n + 2\boldsymbol{E}_j^n - \boldsymbol{E}_j^{n-1} \\ &= \boldsymbol{E}_{j+1}^n + \boldsymbol{E}_{j-1}^n - \boldsymbol{E}_j^{n-1}. \end{aligned} \tag{2.10}$$

In some non-trivial cases, the numerical scheme reproduces the exact solution (cf. [TH05]):

Lemma 2.4. *Let $\boldsymbol{E}(x,t) = \boldsymbol{F}(x+ct) + \boldsymbol{G}(x-ct)$ be a traveling wave solution of* (2.7) *with $\boldsymbol{E}_0(x) = \boldsymbol{F}(x) + \boldsymbol{G}(x)$ and $\boldsymbol{E}_0'(x) = \boldsymbol{F}'(x) - \boldsymbol{G}'(x)$. Let $\Delta t = \frac{\Delta x}{c}$ and suppose that the starting values of the scheme* (2.10) *are exact, i.e., $\boldsymbol{E}_j^0 = \boldsymbol{E}_0(x_j)$, $\boldsymbol{E}_j^1 = \boldsymbol{E}(x_j, \Delta t) = \boldsymbol{F}(x_j + c\Delta t) + \boldsymbol{G}(x_j - c\Delta t)$. Then the numerical method* (2.10) *yields the exact result, i.e., $\boldsymbol{E}_j^n = \boldsymbol{E}(x_j, t_n)$.*

Remark 2.5. Due to its remarkable properties, the choice $\Delta t = \frac{\Delta x}{c}$ is called the *magic time step*.

Proof. Suppose that $\boldsymbol{E}_j^k = \boldsymbol{F}(x_j + ct_k) + \boldsymbol{G}(x_j - ct_k)$ for all $k = 0, \dots, n$, $j \in \mathbb{Z}$. Then the scheme (2.9) with $\delta := c\Delta t = \Delta x$ reads

$$\begin{aligned} \boldsymbol{E}_j^{n+1} &= \boldsymbol{F}(x_j + \Delta x + cn\Delta t) + \boldsymbol{G}(x_j + \Delta x - cn\Delta t) + \boldsymbol{F}(x_j - \Delta x + cn\Delta t) + \boldsymbol{G}(x_j - \Delta x - cn\Delta t) \\ &\quad - \boldsymbol{F}(x_j + c(n-1)\Delta t) - \boldsymbol{G}(x_j - c(n-1)\Delta t) \\ &= \boldsymbol{F}(x_j + (n+1)\delta) + \boldsymbol{G}(x_j - (n-1)\delta) + \boldsymbol{F}(x_j + (n-1)\delta) + \boldsymbol{G}(x_j - (n+1)\delta) \\ &\quad - \boldsymbol{F}(x_j + (n-1)\delta) - \boldsymbol{G}(x_j - (n-1)\delta) \\ &= \boldsymbol{F}(x_j + (n+1)\delta) + \boldsymbol{G}(x_j - (n+1)\delta) \\ &= \boldsymbol{F}(x_j + t_{n+1}) + \boldsymbol{G}(x_j + t_{n+1}) \\ &= \boldsymbol{E}(x_j, t_{n+1}) \end{aligned}$$

□

Lemma 2.6. *Let $\boldsymbol{E}(x,t) = \boldsymbol{E}_0 e^{i\omega t - ikx}$ with $\frac{\omega}{k} = c$ be a harmonic wave solution of* (2.7)*. The method* (2.9) *yields the exact result if $\boldsymbol{E}_j^0 = \boldsymbol{E}(x_j, t_0)$, $\boldsymbol{E}_j^1 = \boldsymbol{E}(x_j, t_1)$ and $\Delta t = \frac{\Delta x}{c}$ (magic time step).*

Proof. Suppose that $\boldsymbol{E}_j^k = \boldsymbol{E}(x_j, t_k) = \boldsymbol{E}_0 e^{i\omega t_k - i\omega x_j}$ for all $j \in \mathbb{Z}$ and $k = 0, \dots, n$. Then the scheme (2.9) with $\delta := c\Delta t = \Delta x$ reads

$$\begin{aligned}
\boldsymbol{E}_j^{n+1} &= \boldsymbol{E}_{j+1}^n + \boldsymbol{E}_{j-1}^n - \boldsymbol{E}_j^{n-1} \\
&= \boldsymbol{E}_0 e^{i\omega n\Delta t - ik(j+1)\Delta x} + \boldsymbol{E}_0 e^{i\omega n\Delta t - ik(j-1)\Delta x} - \boldsymbol{E}_0 e^{i\omega(n-1)\Delta t - ikj\Delta x} \\
&= \boldsymbol{E}_0 e^{i\omega(n-1)\delta - i\omega j\delta} + \boldsymbol{E}_0 e^{i\omega n\delta - i\omega(j-1)\delta} - \boldsymbol{E}_0 e^{i\omega(n-1)\delta - i\omega j\delta} \\
&= \boldsymbol{E}_0 e^{i\omega n\delta - i\omega(j-1)\delta} \\
&= \boldsymbol{E}\big(x_j, t_{n+1}\big).
\end{aligned}$$

□

Stability

Stability is one of the most important issues in numerical simulations. Very often large time steps lead to numerical instability. We want to understand this phenomenon and find conditions for Δt, for which the numerical solution remains bounded. We will look into this topic in more detail and in a more general setting in section 3.2.2. For now, consider a harmonic wave solution of (2.7), but with a possibly complex $\tilde{\omega}$,

$$E_j^n = E_0 e^{i\tilde{\omega}t_n - i\tilde{k}x_j} = E_0 e^{i(\mathrm{Re}(\tilde{\omega}) + i\,\mathrm{Im}(\tilde{\omega}))n\Delta t - i\tilde{k}j\Delta x} = E_0 e^{-\mathrm{Im}(\tilde{\omega})n\Delta t} e^{i\,\mathrm{Re}(\tilde{\omega})n\Delta t - i\tilde{k}j\Delta x}. \tag{2.11}$$

Written in this form, we can tell that a real-valued $\tilde{\omega}$ will keep the wave amplitude constant with time while the imaginary part changes it. The amplitude will decrease exponentially for $\mathrm{Im}(\tilde{\omega}) > 0$ and increase exponentially for $\mathrm{Im}(\tilde{\omega}) < 0$.

Now substitute $E_j^n \approx E(x_j, t_n) = E_0 e^{i\tilde{\omega}t_n - i\tilde{k}x_j}$ into the numerical scheme (2.9) with $t_n = n\Delta t$ and $x_j = j\Delta x$ and obtain

$$E_0 e^{i\tilde{\omega}t_n - i\tilde{k}x_j} \frac{1}{(\Delta t)^2}\left(e^{i\tilde{\omega}\Delta t} - 2 + e^{-i\tilde{\omega}\Delta t}\right) \overset{!}{=} E_0 e^{i\tilde{\omega}t_n - i\tilde{k}x_j} \frac{c^2}{(\Delta x)^2}\left(e^{i\tilde{k}\Delta x} - 2 + e^{-i\tilde{k}\Delta x}\right).$$

The identity $e^{ia} + e^{-ia} = 2\cos(a)$ yields

$$\frac{1}{(\Delta t)^2}(\cos(\tilde{\omega}\Delta t) - 1) = \frac{c^2}{(\Delta x)^2}(\cos(\tilde{k}\Delta x) - 1).$$

Solving this for $\tilde{\omega}$ yields

$$\tilde{\omega} = \frac{1}{\Delta t}\arccos(\zeta)$$

where

$$\zeta := c^2\frac{\Delta t^2}{\Delta x^2}\big(\cos(\tilde{k}\Delta x) - 1\big) + 1 \;\in\; 1 + c^2\frac{\Delta t^2}{\Delta x^2}[-2, 0].$$

For $\zeta \geq -1$, $\tilde{\omega}$ is real, yielding a constant wave amplitude with time. If $c\frac{\Delta t}{\Delta x} > 1$ then $\zeta < -1$ and $\tilde{\omega}$ will be complex. Using the complex arccosine function

$$\arccos(\zeta) = -i\ln\left(\zeta \pm \sqrt{\zeta^2 - 1}\right),$$

we obtain

$$\tilde{\omega} = -\frac{i}{\Delta t}\ln\left(\zeta \pm \sqrt{\zeta^2 - 1}\right)$$

so that $\mathrm{Re}(\tilde{\omega}) = 0$ and $\mathrm{Im}(\tilde{\omega}) = -\frac{1}{\Delta t}\ln\left(\zeta \pm \sqrt{\zeta^2 - 1}\right)$. Substituting this into (2.11), we obtain

$$\boldsymbol{E}_j^n = \boldsymbol{E}_0 e^{n\ln\left(\zeta + \sqrt{\zeta^2 - 1}\right)} e^{-i\tilde{k}j\Delta x} = \boldsymbol{E}_0\left(\zeta \pm \sqrt{\zeta^2 - 1}\right)^n e^{-i\tilde{k}j\Delta x}.$$

Since $\zeta < -1$, the exponential growth factor $\zeta - \sqrt{\zeta^2 - 1}$ is less than -1 and hence causes numerical instabilities. To avoid them, we have to choose $c\frac{\Delta t}{\Delta x} \leq 1$, that is $\Delta t \leq \frac{\Delta x}{c}$, as stability is guaranteed for time steps smaller than or equal to the magic time step only. So we have stability (in L^∞) if and only if $\left|\zeta \pm \sqrt{\zeta^2 - 1}\right| < 1$.

Dispersion

Now let us consider the case where $\Delta t \leq \frac{\Delta x}{c}$. We want to understand the numerical dispersion (cf. [TH05]).

The harmonic wave $E(x,t) = E_0 e^{i\omega t - ikx}$ solves the wave equation (2.7) if and only if ω and k satisfy the *dispersion relation*

$$\omega = \pm ck. \tag{2.12}$$

ω is called the angular frequency, k the wave number, $\frac{\omega}{k} = \pm c$ the phase velocity and $\frac{d\omega}{dk} = \pm c$ the group velocity.

To understand numerical dispersion, we substitute $E_j^n \approx E(x_j, t_n) = E_0 e^{i\omega t_n - i\tilde{k} x_j}$ with some ω, but $\tilde{k} \neq k$ into the numerical scheme (2.9) with $t_n = n\Delta t$ and $x_j = j\Delta x$, which yields

$$E_0 e^{i\omega t_n - i\tilde{k}x_j} \frac{1}{(\Delta t)^2}\left(e^{i\omega\Delta t} - 2 + e^{-i\omega\Delta t}\right) \overset{!}{=} E_0 e^{i\omega t_n - i\tilde{k}x_j} \frac{c^2}{(\Delta x)^2}\left(e^{i\tilde{k}\Delta x} - 2 + e^{-i\tilde{k}\Delta x}\right).$$

With Euler's formula, we obtain the *numerical dispersion relation*

$$\frac{1}{(\Delta t)^2}(\cos(\omega\Delta t) - 1) = \frac{c^2}{(\Delta x)^2}(\cos(\tilde{k}\Delta x) - 1). \tag{2.13}$$

If $\Delta t = \frac{\Delta x}{c}$ (magic time step), we have

$$\cos(\tilde{k}\Delta x) = \cos(\omega\Delta t)$$

and thus

$$\tilde{k}\Delta x = \pm\omega\Delta t$$

and

$$\tilde{k} = \pm\frac{\omega\Delta t}{\Delta x} = \pm\frac{\omega\Delta t}{c\Delta t} = \pm\frac{\omega}{c} = k.$$

If $\Delta t < \frac{\Delta x}{c}$, the Taylor series expansion of the cosine for small Δt and Δx yields

$$\begin{aligned}
&\frac{1}{(\Delta t)^2}\left(1 - \frac{1}{2}(\omega\Delta t)^2 + \mathcal{O}((\omega\Delta t)^4) - 1\right) \overset{!}{=} \frac{c^2}{(\Delta x)^2}\left(1 - \frac{1}{2}(\tilde{k}\Delta x)^2 + \mathcal{O}((\tilde{k}\Delta x)^4) - 1\right) \\
\iff \quad &\frac{1}{2}\omega^2 + \mathcal{O}(\omega^4\Delta t^2) = \frac{1}{2}c^2\tilde{k}^2 + \mathcal{O}(c^2\tilde{k}^4\Delta x^2) \\
\iff \quad &\tilde{k}^2 = \frac{\omega^2}{c^2} + \mathcal{O}(\Delta t^2 + \Delta x^2).
\end{aligned}$$

This means that for $\Delta t, \Delta x \to 0$, $\tilde{k}$ goes to $\pm\frac{\omega}{c}$, so we have a good approximation for small Δt. For large Δt, the numerical solution still has the correct shape, but it is shifted against the exact solution. This is dispersion. For given time step Δt and mesh size Δx and a given period $\frac{2\pi}{\omega}$ of a propagating wave, (2.13) can be used to compute the numerical wave number $\tilde{k}$, phase velocity $\frac{\omega}{\tilde{k}}$ and group velocity $\frac{d\omega}{d\tilde{k}}$. It can provide some good insight as to why numerical results look as they do.

2.2 The Yee Algorithm

When choosing a grid to represent the domain, usually all quantities are stored at the same grid points. The brilliant idea of Kane S. Yee [Yee66] was to mimic the interleaving of the continuous equations by interleaved grid positions. The arrangement is such that each component is surrounded by four components of the other field — each half a grid step away. These four neighbors are just the ones from the corresponding equation, see figure 2.1. We can imagine this as two interleaved cuboids — one for the electric and one for the magnetic field.

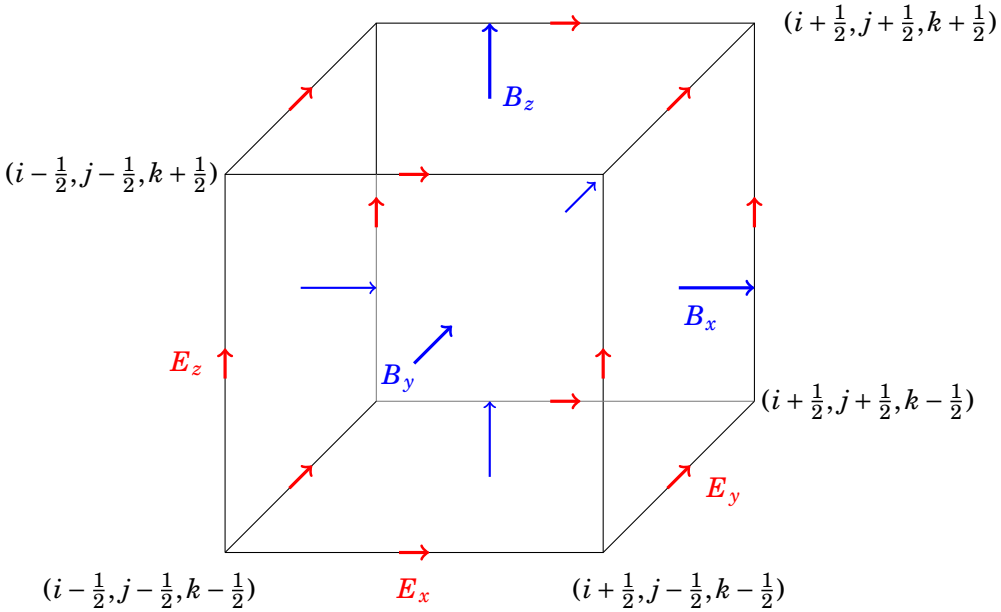

Figure 2.1: *Schematic view of a 3D Yee cell*

Let us take the equation for B_y as an example. It involves the spatial derivatives of E_x and E_z. When we use central finite differences, the four neighbors are just the ones we need for the scheme.

When moving on to the time derivatives, the usual approach is to solve all equations simultaneously to have approximations at some time t_n. The Yee scheme however, also staggers the components in time. That means electric and magnetic components are stored at different times to have a leapfrog-type arrangement. If we consider a one-dimensional example, i.e., we set derivatives with respect to y and z to zero, we have the transverse magnetic (TM) mode

$$\begin{aligned}
\frac{1}{c}\frac{\partial B_x}{\partial t} &= 0 \\
\frac{1}{c}\frac{\partial B_y}{\partial t} &= \frac{\partial E_z}{\partial x} \\
\frac{1}{c}\frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x}
\end{aligned}$$

and the transverse electric mode

$$\begin{aligned}\frac{1}{c}\frac{\partial E_x}{\partial t} &= 0\\ \frac{1}{c}\frac{\partial E_y}{\partial t} &= -\frac{\partial B_z}{\partial x}\\ \frac{1}{c}\frac{\partial B_z}{\partial t} &= -\frac{\partial E_y}{\partial x}.\end{aligned}$$

That means that E_x and B_x are constant in time and we have only four equations for the remaining four unknowns left. They are interleaved such that only B_y and E_z depend on each other as do E_y and B_z. Let us look at the former. Figure 2.2 shows what the staggering in space and time looks like.

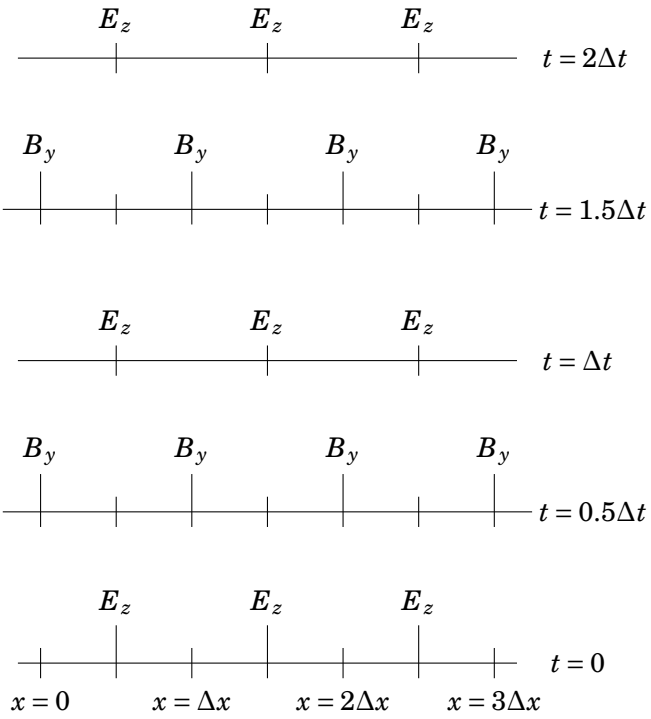

Figure 2.2: *Space-time chart of the Yee algorithm for a one-dimensional example*

With this configuration, two quantities are always Δx apart and we can use central finite differences for the approximation of the spatial derivatives. In the one-dimensional example this leads to

$$B_{y,j}^{n+\frac{1}{2}} = B_{y,j}^{n-\frac{1}{2}} + c\frac{\Delta t}{\Delta x}\left(E_{z,j+\frac{1}{2}}^{n} - E_{z,j-\frac{1}{2}}^{n}\right) \tag{2.14a}$$

$$B_{z,j}^{n+\frac{1}{2}} = B_{z,j}^{n-\frac{1}{2}} - c\frac{\Delta t}{\Delta x}\left(E_{y,j+\frac{1}{2}}^{n} - E_{y,j-\frac{1}{2}}^{n}\right) \tag{2.14b}$$

$$E_{y,j+\frac{1}{2}}^{n+1} = E_{y,j+\frac{1}{2}}^{n} - c\frac{\Delta t}{\Delta x}\left(B_{z,j+1}^{n+\frac{1}{2}} - B_{z,j}^{n+\frac{1}{2}}\right) \tag{2.14c}$$

$$E_{z,j+\frac{1}{2}}^{n+1} = E_{z,j+\frac{1}{2}}^{n} + c\frac{\Delta t}{\Delta x}\left(B_{y,j+1}^{n+\frac{1}{2}} - B_{y,j}^{n+\frac{1}{2}}\right). \tag{2.14d}$$

The same is done for the three-dimensional set of equations. The Yee cell in figure 2.1 is constructed such that each component has four neighbors that are half a mesh width

away. Those are the field components needed for a central finite difference approximation of the spatial derivatives. This is the famous Yee algorithm. It is second order accurate in space and time due to the central finite differences.

Splitting Methods

The Yee scheme is an example of a *splitting method.* Splitting methods have their roots in the field of ordinary differential equations (cf. [HLW06, chapter II.5]), so for a short recollection, let us consider a system

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}).$$

The *exact flow* $\boldsymbol{\varphi}_t$ associates the value $\boldsymbol{y}(t)$ of the solution with initial data $\boldsymbol{y}(0) = \boldsymbol{y}^0$ for any point $\boldsymbol{y}^0$,

$$\boldsymbol{\varphi}_t(\boldsymbol{y}^0) = \boldsymbol{y}(t) \qquad \text{if} \quad \boldsymbol{y}(0) = \boldsymbol{y}^0.$$

When a numerical method is used to approximate the solution $\boldsymbol{y}(t_{n+1})$, this can be represented by another mapping, the *numerical flow*

$$\boldsymbol{\Phi}_{\Delta t} : \boldsymbol{y}_n \mapsto \boldsymbol{y}_{n+1}$$

which maps the already known $\boldsymbol{y}_n$ to the yet unknown $\boldsymbol{y}_{n+1} \approx \boldsymbol{y}(t_{n+1})$, where $\Delta t = t_{n+1} - t_n$ denotes the step size. To any method $\boldsymbol{\Phi}_{\Delta t}$, we can define its *adjoint method* $\boldsymbol{\Phi}^*_{\Delta t} = \boldsymbol{\Phi}^{-1}_{-\Delta t}$ as the inverse map of the original method with reversed time step. The implicit Euler method is the adjoint of the explicit Euler method. A method with $\boldsymbol{\Phi}^*_{\Delta t} = \boldsymbol{\Phi}_{\Delta t}$ is called *symmetric*.

Now we can construct new methods by composition, i.e.,

$$\boldsymbol{\Psi}_{\Delta t} = \boldsymbol{\Phi}_{\alpha_1 \Delta t} \circ \cdots \circ \boldsymbol{\Phi}_{\alpha_s \Delta t}$$

for real numbers $\alpha_1, \ldots, \alpha_s$. A very common technique is to combine a method with its adjoint,

$$\boldsymbol{\Psi}_{\Delta t} = \boldsymbol{\Phi}_{\frac{\Delta t}{2}} \circ \boldsymbol{\Phi}^*_{\frac{\Delta t}{2}}$$

because for any consistent one-step method $\boldsymbol{\Phi}_{\Delta t}$ of order one, $\boldsymbol{\Psi}_{\Delta t}$ is a second order symmetric method (cf. [HLW06, chapter II.5]). For the explicit Euler method, we obtain the implicit midpoint rule by this trick.

These compositions can be applied to "composed" system of equations. Assume that our system takes some split form

$$\boldsymbol{y}' = \boldsymbol{f}^{[1]}(\boldsymbol{y}) + \boldsymbol{f}^{[2]}(\boldsymbol{y}),$$

like a discretization of a two-dimensional partial differential equation or simply two different forces like the electric and magnetic terms in the Lorentz force. Assuming we can explicitly calculate the exact flows $\boldsymbol{\varphi}^{[1]}_t$ and $\boldsymbol{\varphi}^{[2]}_t$ of the systems $\boldsymbol{y}' = \boldsymbol{f}^{[1]}(\boldsymbol{y})$ and $\boldsymbol{y}' = \boldsymbol{f}^{[2]}(\boldsymbol{y})$, we can compute an approximate solution to the whole system in two steps. Starting from $\boldsymbol{y}^0$, we calculate some intermediate value $\boldsymbol{y}^{\frac{1}{2}}$ by solving the first system, which we use as initial data when solving the second one. This yields a numerical method

$$\boldsymbol{\Phi}_{\Delta t} = \boldsymbol{\varphi}^{[2]}_{\Delta t} \circ \boldsymbol{\varphi}^{[1]}_{\Delta t}. \tag{2.15}$$

We obtain the adjoint by reversing the order, in which we solve the systems. These first order methods are called *Lie-Trotter splitting* (cf. [Tro59]).

A symmetric approach is

$$\boldsymbol{\Phi}_{\Delta t}^{[S]} = \boldsymbol{\varphi}_{\Delta t/2}^{[1]} \circ \boldsymbol{\varphi}_{\Delta t}^{[2]} \circ \boldsymbol{\varphi}_{\Delta t/2}^{[1]}, \tag{2.16}$$

the *Strang splitting* (cf. [Str68]), which is the composition of the Lie-Trotter method and its adjoint with step sizes $\frac{\Delta t}{2}$, so we have second order. The Yee scheme is an example of a Strang splitting.

Of course, in practice we do not usually have the means to solve the partial systems exactly, so we proceed by substituting the numerical flow into the splittings.

The Magic Time Step

Now recall the magic time step from the wave formulation of Maxwell's equations. Also with the Yee scheme, the magic time step yields the exact solution to the discrete problem. To show this exemplarily for the one-dimensional TM mode — the coupled B_y and E_z equations—, we consider the solutions to the corresponding wave equations,

$$\begin{aligned} B_y(x,t) &= F_B(x+ct) + G_B(x-ct), \\ E_z(x,t) &= F_E(x+ct) + G_E(x-ct). \end{aligned}$$

Inserting these into the first order equations (2.14a) and (2.14d) for B_y and E_z, we find that $F_B = F_E$ and $G_E = -G_B$, so the solutions are of the form

$$\begin{aligned} B_y(x,t) &= F(x+ct) + G(x-ct), \\ E_z(x,t) &= F(x+ct) - G(x-ct). \end{aligned}$$

We can now use these to check for the magic time step:

Lemma 2.7. *Let*

$$\begin{aligned} B_y(x,t) &= F(x+ct) + G(x-ct) \\ E_z(x,t) &= F(x+ct) - G(x-ct) \end{aligned}$$

be a traveling wave solution of the TM mode and

$$\begin{aligned} B_z(x,t) &= \tilde{F}(x+ct) + \tilde{G}(x-ct) \\ E_y(x,t) &= -\tilde{F}(x+ct) + \tilde{G}(x-ct) \end{aligned}$$

be a traveling wave solution of the TE mode. Then the Yee scheme with $\delta := c\Delta t = \Delta x$ yields the exact results if the initial data are exact, i.e.,

$$\begin{aligned} B_{y,j}^0 &= B_y(x_j, t_0),\ E_{z,j+\frac{1}{2}}^0 = E_z(x_{j+\frac{1}{2}}, t_0), \\ B_{z,j}^0 &= B_z(x_j, t_0),\ E_{y,j+\frac{1}{2}}^0 = E_y(x_{j+\frac{1}{2}}, t_0). \end{aligned}$$

Proof. With $\delta := c\Delta t = \Delta x$, the Yee scheme for the TM mode reads

$$\begin{aligned} B_{y,j}^{n+\frac{1}{2}} &= B_{y,j}^{n-\frac{1}{2}} + E_{z,j+\frac{1}{2}}^n - E_{z,j-\frac{1}{2}}^n \\ E_{z,j+\frac{1}{2}}^{n+1} &= E_{z,j+\frac{1}{2}}^n + B_{y,j+1}^{n+\frac{1}{2}} - B_{y,j}^{n+\frac{1}{2}}. \end{aligned}$$

Inserting the exact initial values yields

$$\begin{aligned}
B_{y,j}^{n+\frac{1}{2}} &= F(x_j + ct_n - \tfrac{\delta}{2}) + G(x_j - (ct_n - \tfrac{\delta}{2})) + F(x_j + \tfrac{\delta}{2} + ct_n) - G(x_j + \tfrac{\delta}{2} - ct_n) \\
&\quad - F(x_j - \tfrac{\delta}{2} + ct_n) + G(x_j - \tfrac{\delta}{2} - ct_n) \\
&= F(x_j + \tfrac{\delta}{2} + ct_n) + G(x_j - \tfrac{\delta}{2} - ct_n) \\
&= F(x_j + ct_{n+\frac{1}{2}}) + G(x_j - ct_{n+\frac{1}{2}}) \\
&= B_y(x_j, t_{n+\frac{1}{2}})
\end{aligned}$$

and

$$\begin{aligned}
E_{z,j+\frac{1}{2}}^{n+1} &= F(x_j + \tfrac{\delta}{2} + ct_n) - G(x_j + \tfrac{\delta}{2} - ct_n) + F(x_j + \delta + ct_n + \tfrac{\delta}{2}) + G(x_j + \delta - (ct_n + \tfrac{\delta}{2})) \\
&\quad - F(x_j + t_n + \tfrac{\delta}{2}) - G(x_j - (t_n + \tfrac{\delta}{2})) \\
&= F(x_j + \tfrac{\delta}{2} + ct_n + \delta) + G(x_j - \tfrac{\delta}{2} + \delta - \delta - ct_n) \\
&= F(x_{j+\frac{1}{2}} + ct_{n+1}) + G(x_{j+\frac{1}{2}} - ct_{n+1}) \\
&= E_z(x_{j+\frac{1}{2}}, t_{n+1}).
\end{aligned}$$

The proof for the TE mode is completely analogous. □

The same property holds again for the harmonic wave solutions:

Lemma 2.8. *Let $B_y(x,t) = e^{i\omega t - ikx}$ and $E_z(x,t) = -e^{i\omega t - ikx}$ be a harmonic wave solution of the TM mode with $\frac{\omega}{k} = c$ and $B_z(x,t) = e^{i\tilde{\omega} t - i\tilde{k}x}$ and $E_y(x,t) = -e^{i\tilde{\omega} t - i\tilde{k}x}$ a harmonic wave solution of the TE mode with $\frac{\omega}{k} = -c$. Then the Yee scheme with $\delta := c\Delta t = \Delta x$ yields the exact results if the initial data are exact, i.e.,*

$$\begin{aligned}
B_{y,j}^0 = B_y(x_j, t_0),\ E_{z,j+\frac{1}{2}}^0 = E_z(x_{j+\frac{1}{2}}, t_0), \\
B_{z,j}^0 = B_z(x_j, t_0),\ E_{y,j+\frac{1}{2}}^0 = E_y(x_{j+\frac{1}{2}}, t_0).
\end{aligned}$$

Proof. With $\delta := c\Delta t = \Delta x$, the Yee scheme for the TM mode reads

$$\begin{aligned}
B_{y,j}^{n+\frac{1}{2}} &= B_{y,j}^{n-\frac{1}{2}} + E_{z,j+\frac{1}{2}}^n - E_{z,j-\frac{1}{2}}^n \\
&= e^{i\omega(t_n - \frac{\delta}{2c}) - i\frac{\omega}{c}x_j} - e^{i\omega t_n - i\frac{\omega}{c}(x_j + \frac{\delta}{2})} + e^{i\omega t_n - i\frac{\omega}{c}(x_j - \frac{\delta}{2})} \\
&= e^{i\omega t_n - i\frac{\omega}{c}x_j - i\omega\frac{\delta}{2c}} - e^{i\omega t_n - i\frac{\omega}{c}x_j - i\omega\frac{\delta}{2c}} + e^{i\omega t_n - i\frac{\omega}{c}x_j + i\omega\frac{\delta}{2c}} \\
&= e^{i\omega(t_n + \frac{\delta}{2c}) - ikx_j} \\
&= B_y(x_j, t_{n+\frac{1}{2}})
\end{aligned}$$

and

$$\begin{aligned}
E_{z,j+\frac{1}{2}}^{n+1} &= E_{z,j+\frac{1}{2}}^n + B_{y,j+1}^{n+\frac{1}{2}} - B_{y,j}^{n+\frac{1}{2}} \\
&= -e^{i\omega t_n - i\frac{\omega}{c}(x_j + \frac{\delta}{2})} + e^{i\omega(t_n + \frac{\delta}{2c}) - i\frac{\omega}{c}(x_j + \delta)} - e^{i\omega(t_n + \frac{\delta}{2c}) - i\frac{\omega}{c}x_j} \\
&= -e^{i\omega t_n - i\frac{\omega}{c}x_j - i\omega\frac{\delta}{2}} + e^{i\omega t_n - i\frac{\omega}{c}x_j - i\omega\frac{\delta}{2c}} - e^{i\omega t_n - i\frac{\omega}{c}x_j + i\omega\frac{\delta}{2c}} \\
&= -e^{i\omega t_n - i\frac{\omega}{c}x_j + i\omega(\frac{\delta}{c} - \frac{\delta}{2c})} \\
&= -e^{i\omega(t_n + \frac{\delta}{c}) - ik(x_j + \frac{\delta}{2})} \\
&= E_z(x_{j+\frac{1}{2}}, t_{n+1}).
\end{aligned}$$

The proof for the TE mode is carried out analogously. □

Dispersion

Now we look again at dispersion. We consider once more the TM mode and a traveling wave solution of the form

$$B_{y,j}^{n+\frac{1}{2}} = B_{y_0} e^{i\omega(n+\frac{1}{2})\Delta t - i\tilde{k}j\Delta x}, \tag{2.17}$$

$$E_{z,j+\frac{1}{2}}^{n} = E_{z_0} e^{i\omega n\Delta t - i\tilde{k}(j+\frac{1}{2})\Delta x}. \tag{2.18}$$

Substituting this into the scheme (2.14), we obtain

$$B_{y_0} = -\frac{\Delta t E_{z_0} \sin(\tilde{k}\frac{\Delta x}{2})}{c\Delta x \sin(\omega\frac{\Delta t}{2})},$$

$$E_{z_0} = -\frac{\Delta t B_{y_0} \sin(\tilde{k}\frac{\Delta x}{2})}{c\Delta x \sin(\omega\frac{\Delta t}{2})}.$$

Substituting B_{y_0} into E_{z_0} then yields

$$\sin^2\left(\omega\frac{\Delta t}{2}\right) = c^2\frac{\Delta t^2}{\Delta x^2}\sin^2\left(\tilde{k}\frac{\Delta x}{2}\right)$$

or

$$\sin\left(\omega\frac{\Delta t}{2}\right) = \pm c\frac{\Delta t}{\Delta x}\sin\left(\tilde{k}\frac{\Delta x}{2}\right) =: \zeta$$

and

$$\omega = \frac{2}{\Delta t}\arcsin(\zeta).$$

For $\Delta x, \Delta t \to 0$, we use the Taylor series expansion of the sine function to see

$$\tilde{k} = \pm\frac{\omega}{c} + \mathcal{O}(\Delta t^2 + \Delta x^2).$$

Thus, as for the wave equation, we have a second order approximation of the true wave number k.

If we consider the full set of three-dimensional Maxwell's equations, we obtain the numerical dispersion relation

$$\sin^2\left(\omega\frac{\Delta t}{2}\right) = c^2\frac{\Delta t^2}{\Delta x^2}\sin^2\left(\tilde{k}_x\frac{\Delta x}{2}\right) + c^2\frac{\Delta t^2}{\Delta y^2}\sin^2\left(\tilde{k}_y\frac{\Delta y}{2}\right) + c^2\frac{\Delta t^2}{\Delta z^2}\sin^2\left(\tilde{k}_z\frac{\Delta z}{2}\right).$$

Stability

We now consider again the possibility of a complex-valued $\tilde{\omega}$ for $|\zeta| > 1$. With the complex arcsine function $\arcsin(\zeta) = -i\ln\left(i\zeta \pm \sqrt{1-\zeta^2}\right)$, we obtain

$$\tilde{\omega} = \frac{\pi}{\Delta t} - \frac{2i}{\Delta t}\ln\left(\zeta \pm \sqrt{\zeta^2 - 1}\right).$$

Substituting this into our trial solution, we obtain

$$V_j^n = V_0\left(\zeta \pm \sqrt{\zeta^2-1}\right)^{2n} e^{i\frac{\pi}{\Delta t}n\Delta t - ikj\Delta x}$$

where V represents either B_y or E_z. We have $\left(\zeta + \sqrt{\zeta^2 - 1}\right)$ greater than one for $\zeta > 1$. Thus we have exponential growth if $c\frac{\Delta t}{\Delta x} > 1$ and our stability bound is again $\Delta t \leq \frac{\Delta x}{c}$.

In the full three-dimensional setting, we have

$$\zeta := \pm c\Delta t \sqrt{\frac{1}{\Delta x^2}\sin^2\left(\tilde{k}_x \frac{\Delta x}{2}\right) + \frac{1}{\Delta y^2}\sin^2\left(\tilde{k}_y \frac{\Delta y}{2}\right) + \frac{1}{\Delta z^2}\sin^2\left(\tilde{k}_z \frac{\Delta z}{2}\right)}.$$

Again, $|\zeta| \leq 1$ is required for stability with the same arguments as in one dimension. The stability bound, however, for $\Delta x = \Delta y = \Delta z =: \Delta$ is now

$$\Delta t \leq \frac{\Delta}{c\sqrt{3}}.$$

In two dimensions, the factor is $\sqrt{2}$, so we have to choose

$$\Delta t \leq \frac{\Delta}{c\sqrt{2}}$$

for stability.

3 Hyperbolic Conservation Laws

The conservation of quantities like mass or energy is of great importance in physics. In this chapter, we first derive the differential equations describing this conservation, and discuss some theory on this kind of equations and the most important classes. One of the major difficulty is that solutions to nonlinear hyperbolic equations can become discontinuous, even if the initial data are smooth. Or we may have infinitely many solutions and then have to find a way to pick the physically relevant one. This will lead us to viscosity solutions and entropy conditions.

Next we turn to the numerical point of view and consider methods for the approximation of hyperbolic conservation laws. This includes a discussion of the numerical analoga of theoretical concepts like entropy. We will review standard (linear) stability analysis and argue why it fails for nonlinear equations. Thus, we will discuss nonlinear stability conditions. To gain even further insight into some standard numerical schemes, we will also consider modified equations, which can help us understand the behavior of the methods.

Extensive sources on the theory and numerics of hyperbolic conservation laws beyond the scope of this thesis are [Kro97] and [LeV11]. Most of the contents of this chapter can be found there. Other important works include [Whi74] or [CF76]. A further reference, especially for convergence results, is [DiP83].

3.1 Theory

To derive a mathematical model for conservation, we consider a quantity $u : \Omega \times \mathbb{R}_+ \to \mathbb{R}$ — let us say some density — on a Lipschitz domain $\Omega \subset \mathbb{R}^d$ and a volume $V \subset \Omega$, also with Lipschitz boundary. The total mass in this volume is then

$$m_V(t) := \int_V u(\boldsymbol{x}, t)\, d\boldsymbol{x}.$$

We now assume this substance to be in motion. It flows in or out through the volume boundary ∂V. This flow is given by a vector field $\boldsymbol{f} : \Omega \times \mathbb{R}_+ \to \mathbb{R}^d$ with $\boldsymbol{f} = \boldsymbol{f}(u(\boldsymbol{x}, t))$. The flow through V is thus

$$F_V(t) := \int_{\partial V} \boldsymbol{f}(u(\boldsymbol{x}, t)) \cdot \boldsymbol{n}_V(\boldsymbol{x})\, d\sigma,$$

where $\boldsymbol{n}_V(\boldsymbol{x})$ is the outer normal of V. The total balance in absence of sources is

$$\frac{d}{dt} m_V(t) = -F_V(t) \qquad \text{for all } V \subset \Omega.$$

The sign is due to the requirement $\frac{d}{dt} m_V \leq 0$ when $\boldsymbol{f}$ points along $\boldsymbol{n}_V$. The equality expresses that no mass is created or disappears, which explains the term *conservation law*. The divergence theorem now yields

$$\int_V \partial_t u(\boldsymbol{x}, t) + \nabla \cdot \boldsymbol{f}(u((\boldsymbol{x}, t))\, d\boldsymbol{x} = 0 \qquad \text{for all } V \subset \Omega \tag{3.1}$$

and since (3.1) must hold for arbitrary V, we obtain the equality

$$\partial_t u(\boldsymbol{x}, t) + \nabla \cdot \boldsymbol{f}(u(\boldsymbol{x}, t)) = 0 \qquad \text{for all } (\boldsymbol{x}, t) \in \Omega \times \mathbb{R}_+. \tag{3.2}$$

Equation (3.2) is called *conservation law in differential form*. Depending on the application, we may also have to provide appropriate initial and/or boundary conditions. In

applications, we often consider several conserved quantities and combine them into one vector $\boldsymbol{u}$.

(3.2) is called *hyperbolic* if the Jacobian $\boldsymbol{f}_u(u(\boldsymbol{x},t))$ has only real eigenvalues and is diagonalizable. It is called *strictly hyperbolic* if the eigenvalues are distinct.

3.1.1 The Advection Equation

The simplest form of conservation law is the *advection equation* where we have $\boldsymbol{f}(u) = u\boldsymbol{v}$ for some velocity $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x},t)$. The general multi-dimensional advection equation thus reads

$$\partial_t u + \nabla \cdot (u\boldsymbol{v}) = 0. \tag{3.3}$$

In the one-dimensional case with $f(u) = au$, for some constant $a \in \mathbb{R}$, the advection equation reads

$$\partial_t u(x,t) + a\partial_x u(x,t) = 0. \tag{3.4}$$

If $u^0(x) = u(x,0)$ is the initial data, then $u(x,t) = u^0(x-at)$ satisfies (3.4) for all $t \geq 0$.[2] So the initial profile is simply advected with velocity a. Based on this simple problem, various schemes can be derived for the numerical solution of conservation laws. The advection equation is a nice test problem for new methods, especially since the exact solution is known and easy to understand and compute.

The advection equation (3.4) is also the foundation for further concepts of stability or the characterization of errors. One important issue is dispersion. Inserting a harmonic wave solution $u(x,t) = e^{i\omega t - ikx}$ into (3.4), we obtain

$$\begin{aligned} i\omega e^{i\omega t - ikx} - aike^{i\omega t - ikx} &= 0 \\ \iff \omega &= ak \end{aligned}$$

This is the exact *dispersion relation*. Our numerical schemes should not only yield a good approximation to the solution, but also to this relation. A poor numerical dispersion can yield unusable results even if the shape of the solution is correct. A simple example is a sine wave that is transported with velocity a. The dispersion error yields a phase error, which can lead to a numerical solution that is just the negative of the exact solution for sufficiently large simulation times. In chapter 2 we saw that both the Yee scheme and the scheme for the wave equation did well here.

For real ω, there is only propagation, while a positive imaginary part $\mathrm{Im}(\omega)$ results in damping or dissipation. The *phase* and *group velocities* are defined by

$$c_p = \frac{\mathrm{Re}(\omega)}{k}, \qquad c_g = \mathrm{Re}\,\frac{d\omega}{dk}.$$

A wave is called *dispersive* if the phase velocity depends on the wave number k. If it is composed of different harmonic waves, it will deform while traveling because faster waves overtake the slower ones. For a non-dispersive wave, phase and group velocity are equal, $c_p = c_g$. For (3.4), we have $c_p = c_g = a$.

[2] The notation u^0 is not to be confused with powers of u. For numerical approximations, the notation is $u(x_j,t_n) \approx u_j^n$, hence we superscribe the zero for the initial condition for consistent notation.

3.1.2 Scalar Conservation Laws

For the description of many phenomena, scalar advection equations are not enough, so we go back to (3.2), but we now consider scalar conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, \quad t > 0,\ x \in \mathbb{R}, \tag{3.5}$$

with a scalar function $u = u(x,t)$ and $f \in C^\infty(\mathbb{R})$ locally bounded. For uniqueness we need an initial condition

$$u(x,0) = u^0(x)$$

with $u^0 \in L^\infty(\mathbb{R})$. Since classical solutions of (3.5) can seldom be obtained, we look for *weak solutions*, which are defined by multiplying (3.5) by a test function $\phi \in C_0^\infty(\mathbb{R} \times [0,T])$ — i.e., $\phi \in C^\infty(\mathbb{R} \times [0,T])$ with compact support — and integration by parts:

Definition 3.1. *(Weak solution)*
Let $u^0 \in L^\infty(\mathbb{R})$. A function $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ is called weak solution *of*

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \qquad \text{for } (x,t) \in \mathbb{R} \times (0,T) \tag{3.6a}$$

$$u(x,0) = u^0(x) \quad \text{for } x \in \mathbb{R} \tag{3.6b}$$

if for all $\phi \in C_0^\infty(\mathbb{R} \times [0,T])$

$$\int_0^T \int_{\mathbb{R}} u \partial_t \phi + f(u) \partial_x \phi \, dx \, dt + \int_{\mathbb{R}} u^0 \phi(x,0) \, dx = 0 \tag{3.7}$$

holds.

It can readily be shown that if a weak solution u is in $C^1(\mathbb{R} \times \mathbb{R}_+)$, then it is in fact a classical solution, i.e., , (3.6) holds pointwise (cf. [Kro97, chapter 2.1]).

If f is differentiable, we can rewrite (3.5) in characteristic form

$$\frac{\partial u}{\partial t} + a(u) \frac{\partial u}{\partial x} = 0, \quad a(u) = \frac{\partial f}{\partial u}. \tag{3.8}$$

Note that

$$\frac{\partial u}{\partial t} + a(u) \frac{\partial u}{\partial x} = \left(a(u), 1\right) \cdot \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial t}\right)^T =: \left(a(u), 1\right) \cdot \nabla u,$$

so we are dealing with a directional derivative in the x-t-plane. Thus (3.8) means there is no change in the solution u in the direction of $(a(u),1)$ in the x-t-plane. Considering a curve $x(t)$ that is everywhere tangent to $(a(u),1)$ and comparing the slopes, we obtain

$$\frac{dx}{dt} = a(u),$$

so

$$u \equiv \text{const.} \quad \text{for} \quad x = a(u)t + \text{const.}$$

These curves are called *characteristics*, u is also called the *characteristic variable* and a the *characteristic speed*. We see that u is constant along the characteristics, which are straight lines.

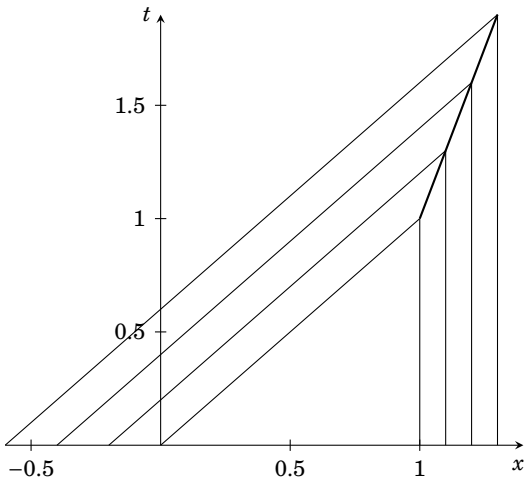

Figure 3.1: *Characteristics forming a shock in the inviscid Burgers equation*

3.1.3 Shocks and Rarefaction Waves

For $a(u) \equiv a$, all the characteristics are parallel and the solution is transported via $u(x,t) = u^0(x-at)$ just as we have seen before. For non-constant $a(u)$, however, the different slopes yield non-parallel characteristics and they may hence intersect — even for smooth initial data. Since u is constant along characteristics, discontinuities arise from the different values along each of them.

To illustrate this behavior, consider the simplest example of a nonlinear scalar conservation law, the inviscid Burgers equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}u^2\right) = 0. \tag{3.9}$$

Now choose smooth, monotone initial conditions with

$$u^0(x) = \begin{cases} 1 & x \le 0 \\ \beta(x) & 0 < x < 1 \\ 0 & x \ge 1 \end{cases}$$

with $\beta(x)$ such that u^0 is monotone and smooth. If we look at the characteristics outside the interval $(0,1)$, we see the following problem: For $x \ge 1$ the characteristics are parallel to the t-axis, while for $x \le 0$ they move to the right so they will eventually intersect. These characteristics are sketched in figure 3.1.

Up to $t = 1$, we can construct a differentiable solution from the characteristics, which then becomes discontinuous, so it is not clear how it can be continued.

The following lemma provides a jump condition for weak solutions that are piecewise smooth (cf. [Kro97, lemma 2.1.4]).

Lemma 3.2 (Rankine-Hugoniot)**.** *Let $\gamma : t \mapsto (x_S(t), t)$ be a smooth curve that separates $\mathbb{R}\times\mathbb{R}_+$ into two parts D_L and D_R. Furthermore, let $u \in L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ such that $u_L := u|_{D_L} \in C^1(\overline{D}_L)$ and $u_R := u|_{D_R} \in C^1(\overline{D}_R)$ and u_L and u_R satisfy* (3.5) *locally in D_L and D_R respectively in the classical sense. Then u is a weak solution of* (3.6) *if and only if*

$$f\big(u_R(x_S(t),t)\big) - f\big(u_L(x_S(t),t)\big) = x_s'(t)\big(u_R(x_S(t),t) - u_L(x_S(t),t)\big) \tag{3.10}$$

for all $t > 0$*. Instead of* (3.10)*, we will often write*

$$f(u_R) - f(u_L) = S(u_R - u_L) \tag{3.11}$$

with $S := x_s'(t)$*.*

This is called the *Rankine-Hugoniot jump condition* and S is the propagation speed of the discontinuity.

Solving the jump condition for the shock speed, we obtain

$$S = \frac{f(u_R) - f(u_L)}{u_R - u_L}$$

which by the mean value theorem implies

$$S = f'(\tilde{u}) = a(\tilde{u})$$

for some $\tilde{u}$ between u_L and u_R. $a(\tilde{u})$ can be regarded as a mean value of $a(u)$ for u between u_L and u_R. In this sense, the shock speed equals an average wave speed.

Characteristics may also behave completely opposite: they can run away from each other causing a *rarefaction wave*. Consider again Burgers equation (3.9), but with initial condition

$$u^0(x) = \begin{cases} -1 & x < 0, \\ 0 & x \geq 0. \end{cases} \tag{3.12}$$

We now see in figure 3.2 that the characteristics leave room for more than one solution. From the Rankine-Hugoniot condition we find

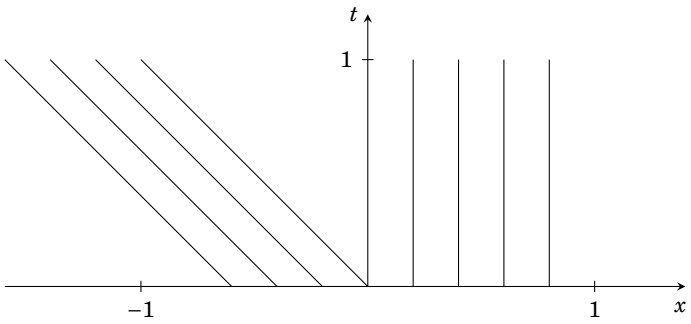

Figure 3.2: *Characteristics forming a rarefaction in the inviscid Burgers equation*

$$u(x,t) = \begin{cases} -1 & x < -\frac{1}{2}t \\ 0 & x \geq -\frac{1}{2}t \end{cases}$$

to be a weak solution.

Note that if u is a weak solution of (3.5), then so is $u_\alpha := u(\alpha x, \alpha t)$ for all $\alpha > 0$. This is called the similarity solution. Hence we try to find a weak solution of the form $u(x,t) = w(x/t)$ and obtain

$$\begin{aligned} &-\frac{x}{t^2} w'\left(\frac{x}{t}\right) + \frac{1}{t} f'\left(w\left(\frac{x}{t}\right)\right) w'\left(\frac{x}{t}\right) = 0 \\ \Leftrightarrow\ &\xi = \frac{x}{t} \quad \text{and} \quad \left(-\xi + f'(w(\xi))\right) w'(\xi) = 0. \end{aligned}$$

Because $f'(u) = u$, this is met by $w(\xi) = \xi$. This yields another solution to the given initial value problem for Burger's equation,

$$u(x,t) = \begin{cases} -1 & x \le -t \\ \frac{x}{t} & -t < x < 0 \\ 0 & x \ge 0. \end{cases}$$

Solutions of this form are called *rarefaction waves*. We can take this even further and have infinitely many solutions by

$$u(x,t) = \begin{cases} -1 & x < s_m t \\ u_m & s_m t \le x \le u_m t \\ \frac{x}{t} & u_m t < x < 0 \\ 0 & x \ge 0 \end{cases}$$

with $s_m = \frac{1}{2}(-1 + u_m)$ for some $u_m \in (-1, 0)$.

An initial value problem with two constant values or *states* as initial condition as in (3.12) is called *Riemann problem*. We have seen two possibilities to connect those states u_L and u_R. We can have a shock with shock speed $S = \frac{f(u_L) - f(u_R)}{u_L - u_R}$ or a rarefaction wave $u(x,t) = w(x/t)$ with $f'(w(\xi)) = \xi$.

3.1.4 Viscosity Solutions

How can we decide, which of those infinitely many solutions is the one we are looking for, the physically correct one? In the derivation of the conservation law, friction has not been considered, but this is a force always present in nature, so $\boldsymbol{f}$ is substituted by

$$\widetilde{\boldsymbol{f}}(u(\boldsymbol{x},t)) \coloneqq \boldsymbol{f}(u(\boldsymbol{x},t)) - \epsilon \nabla u(\boldsymbol{x},t)$$

for some $\epsilon > 0$, since it is well-known that friction is proportional to $-\nabla u$. This is called *Fick's law* (cf. [LeV11, chapter 2]). The differential equation then becomes

$$\partial_t u_\epsilon + \nabla \cdot \boldsymbol{f}(u_\epsilon) = \epsilon \Delta u_\epsilon \qquad \text{on } \mathbb{R}^d \times \mathbb{R}_+.$$

The inner friction of a fluid is called *viscosity*, which is why the above equation is called *viscosity approximation* of the conservation law.

Consider now the scalar Cauchy problem

$$\begin{aligned} \partial_t u_\epsilon + \partial_x f(u_\epsilon) &= \epsilon \partial_x^2 u_\epsilon && \text{on } \mathbb{R} \times \mathbb{R}_+ \\ u_\epsilon(x,0) &= u^0 && \text{on } \mathbb{R} \end{aligned} \tag{3.13}$$

with $u^0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. We want to know if the viscosity solution u_ϵ somehow converges to a weak solution of the conservation law. To state the answer in theorem 3.4, we need some definitions first.

Definition 3.3. *For $v \in L^1(\mathbb{R}^d)$, let*

$$[v]_{BV(\mathbb{R}^d)} \coloneqq \sup \left\{ -\int_{\mathbb{R}^d} v \nabla \cdot \boldsymbol{\Phi} \, d\boldsymbol{x} \ : \ \boldsymbol{\Phi} \in C^\infty(\mathbb{R}^d)^d, \|\boldsymbol{\Phi}\|_{L^\infty(\mathbb{R}^d)^d} = 1 \right\}$$

and

$$BV(\mathbb{R}^d) \coloneqq \left\{ v \in L^1(\mathbb{R}^d) : [v]_{BV(\mathbb{R}^d)} < \infty \right\}.$$

With the norm

$$\|v\|_{BV(\mathbb{R}^d)} := \|v\|_{L^1(\mathbb{R}^d)} + [v]_{BV(\mathbb{R}^d)},$$

$BV(\mathbb{R}^d)$ is a Banach space, the *space of functions of bounded variation on* $\mathbb{R}^d$.

If $v \in C^1(\mathbb{R}^d)$ with $\|\nabla v\|_{L^1(\mathbb{R}^d)} < \infty$, we obtain

$$[v]_{BV(\mathbb{R}^d)} = \sup\left\{ \int_{\mathbb{R}^d} \boldsymbol{\Phi} \cdot \nabla v \, d\boldsymbol{x} \;:\; \boldsymbol{\Phi} \in C^\infty(\mathbb{R}^d)^d, \|\boldsymbol{\Phi}\|_{L^\infty(\mathbb{R}^d)^d} = 1 \right\} = \int_{\mathbb{R}^d} \|\nabla v\| \, d\boldsymbol{x}$$

if we choose a sequence $\{\boldsymbol{\Phi}_k\}$ with $\boldsymbol{\Phi}_k \in C^\infty(\mathbb{R}^d)^d$ and $\boldsymbol{\Phi}_k \to \frac{\nabla v}{\|\nabla v\|}$ in $L^1(\mathbb{R}^d)^d$.

The following theorem is a statement on convergence of the viscosity solution.

Theorem 3.4. *Let $u^0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and $f \in C^2(\mathbb{R})$. Then the sequence of solutions $\{u_{\epsilon_k}\}_k$ of* (3.13) *for any null sequence $\{\epsilon_k\}_k$ converges to a uniquely defined weak solution of the conservation law. This solution is called* viscosity solution of the conservation law *and for all t_1, t_2 with $0 < t_1 < t_2$, the following estimates hold:*

$$\|u\|_{L^\infty(\mathbb{R}\times\mathbb{R}_+)} \le \|u^0\|_{L^\infty(\mathbb{R})},$$

$$\|u(\cdot, t_2)\|_{L^1(\mathbb{R})} \le \|u(\cdot, t_1)\|_{L^1(\mathbb{R})} \le \|u^0\|_{L^1(\mathbb{R})}.$$

In particular, for any two viscosity solutions u, v with initial data u^0, v^0 and $t > 0$, the estimate

$$\|u(\cdot, t) - v(\cdot, t)\|_{L^1(\mathbb{R})} \le \|u^0 - v^0\|_{L^1(\mathbb{R})}$$

holds. This property is called L^1-contractivity. *If in addition, $u^0 \in BV(\mathbb{R})$, then*

$$\|u(\cdot, t_2)\|_{BV(\mathbb{R})} \le \|u(\cdot, t_1)\|_{BV(\mathbb{R})} \le \|u^0\|_{BV(\mathbb{R})}$$

holds.

This theorem and its proof can be found in [Dö12] and in parts in [Lax73], [Kro97, theorem 2.1.7] and [Kru70].

Remark 3.5.

1. With the maximum principle for parabolic equations it actually follows that

   $$u_* := \min_{x\in\mathbb{R}} u^0(x) \le u(x,t) \le \max_{x\in\mathbb{R}} u^0(x) =: u^*$$

   for all $t > 0$ (cf. [Daf72, p. 36]).

2. If f is Lipschitz continuous with Lipschitz constant K on $[u_*, u^*]$, then for $t > 0$ and $x_1 < x_2$

   $$\|u(\cdot, t)\|_{BV([x_1,x_2])} \le \|u^0\|_{BV([x_1 - Kt, x_2 + Kt])}$$

   holds. The same holds for the L^1-norm.

   If we have two sets of initial data u^0, $v^0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$, then we get

   $$\|u(\cdot, t) - v(\cdot, t)\|_{L^1([x_1,x_2])} \le \|u^0 - v^0\|_{L^1([x_1 - Kt, x_2 + Kt])}$$

   for the corresponding solutions u and v (cf. [Daf72, p. 36]).

3. The theorem shows that $u^0 \in BV(\mathbb{R})$ implies $u(\cdot, t) \in BV(\mathbb{R})$ for all $t > 0$. Lax showed that for strictly convex f, even $u^0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ implies $u(\cdot, t) \in BV(\mathbb{R})$. If there exists at least one $z \in [u_*, u^*]$ with $f''(z) = 0$, then this statement is wrong, see [Che83] and the references therein.

3.1.5 Entropy

Unfortunately, it is hardly possible to obtain the correct weak solution by means of theorem 3.4 in practice, so we are looking for a different characterization of our unique solution.

Let $U \in C^2(\mathbb{R})$ be a strictly convex function with $U(0) = 0$. Then define

$$F(z) := \int_0^z U'(\zeta) f'(\zeta)\, d\zeta.$$

(U,F) is called an *entropy pair*. Let u_ϵ be a solution to the viscosity approximation. Now we multiply (3.13) by $U'(u_\epsilon)$ and obtain

$$U'(u_\epsilon)\partial_t u_\epsilon + U'(u_\epsilon) f'(u_\epsilon)\partial_x u_\epsilon = \epsilon U'(u_\epsilon)\partial_x^2 u_\epsilon = \epsilon \partial_x (U'(u_\epsilon)\partial_x u_\epsilon) - \epsilon U''(u_\epsilon)(\partial_x u_\epsilon)^2.$$

The chain rule and the definition of F yield

$$\partial_t U(u_\epsilon) + \partial_x F(u_\epsilon) = \epsilon \partial_x^2 U(u_\epsilon) - \epsilon U''(u_\epsilon)(\partial_x u_\epsilon)^2$$

as $\partial_x F(u_\epsilon) = F'(u_\epsilon)\partial_x u_\epsilon$. The weak formulation of this is

$$\begin{aligned}\int_0^T \int_{\mathbb{R}} U(u_\epsilon)\partial_t \phi + F(u_\epsilon)\partial_x \phi \, dx\, dt + \int_{\mathbb{R}} U(u^0)\phi(0)\, dx \\ = -\epsilon \int_0^T \int_{\mathbb{R}} U(u_\epsilon)\partial_x^2 \phi - U''(u_\epsilon)(\partial_x u_\epsilon)^2 \, dx\, dt.\end{aligned}$$

Now let $\phi \geq 0$ and $\phi(\cdot, 0) = 0$. For $\epsilon \to 0$, we have $u_\epsilon(\cdot,t) \to u(\cdot,t)$ in $L^1(\mathbb{R})$ for fixed $t > 0$. Hence for $\epsilon \to 0$ we obtain

$$\int_0^T \int_{\mathbb{R}} U(u)\partial_t \phi + F(u)\partial_x \phi \, dx\, dt \geq 0$$

for all $\phi \in C_0^\infty(\mathbb{R} \times \mathbb{R}_+)$ with $\phi \geq 0$. This is the weak form of the differential inequality

$$\partial_t U(u) + \partial_x F(u) \leq 0. \tag{3.14}$$

This inequality is called *entropy inequality*. The jump condition for the entropy inequality reads

$$S\big(U(u_L) - U(u_R)\big) \leq F(u_L) - F(u_R).$$

Note that before, we had $U(z) = z$ and $F(z) = f(z)$. For the derivation, we needed $U \in C^2(\mathbb{R})$, but the result can also be shown for the following type of convex functions (cf. [Kru70]). For $a \in \mathbb{R}$ consider

$$U_a(z) := (z-a)_+ = \max\{z-a, 0\},$$

then

$$F_a(z) = (f(z) - f(a))\chi_{\{z \geq a\}}.$$

Assuming $u_R < a < u_L$, it follows that

$$S(u_L - a) \leq f(u_L) - f(a).$$

The Rankine-Hugoniot condition (3.11) now yields

$$\frac{f(u_R)-f(u_L)}{u_R-u_L}(u_L-a) \le f(u_L)-f(a)$$

and

$$\begin{aligned} f(a) &\le f(u_L) - \frac{u_L-a}{u_R-u_L}\big(f(u_R)-f(u_L)\big) \\ &= \left(1+\frac{u_L-a}{u_R-u_L}\right) f(u_L) - \frac{u_L-a}{u_R-u_L} f(u_R) \\ &= \left(1-\frac{u_L-a}{u_L-u_R}\right) f(u_L) + \frac{u_L-a}{u_L-u_R} f(u_R) \end{aligned}$$

for all $a \in [u_R, u_L]$, which is a convex combination. For $u_L < u_R$, we obtain the opposite inequality.

What this means geometrically is that if $u_R < u_L$ ($u_L < u_R$), there exists a shock between u_L and u_R if f is less (greater) than its linear interpolant on $[u_R, u_L]$ ($[u_L, u_R]$). This is called *Oleinik's entropy condition E* (cf. [Ole64]).

Formulated in terms of slopes, if $u_R < u_L$, there exists a shock between u_L and u_R if

$$f'(u_R) \le S \le f'(u_L).$$

For $u_L < u_R$ it reads $f'(u_L) \le S \le f'(u_R)$. This is called the *Lax entropy condition* (cf. [Lax73]). Illustratively, this means there can only be a shock between two states if the characteristics run into it from both sides.

Remark 3.6. If the conservation law is multiplied by $U'(u)$, we obtain

$$\partial_t U(u) + \partial_x F(u) = 0$$

which does not need to hold for non-regular weak solutions. If we multiply Burgers equation by u, we formally obtain

$$\partial_t(u^2) + \partial_x\left(\frac{2}{3}u^3\right) = 0.$$

The jump condition, however, tells us that these two problems are not equivalent because the shock speed is different (cf. [LeV11, chapter 11.12]).

The conservation law and the entropy condition can be combined in a definition of admissible weak solutions. This is due to Hopf (cf. [Hop70]):

Definition 3.7. *$u \in L^\infty(\mathbb{R}\times[0,T])$ is a weak entropy solution if for all non-increasing functions $h:\mathbb{R}\to\mathbb{R}$ and non-negative $\phi \in C_0^\infty(\mathbb{R}\times[0,T))$*

$$\int_0^T \int_{\mathbb{R}} U(u)\partial_t\phi + F(u)\partial_x\phi \, dx\, dt + \int_{\mathbb{R}} U(u^0)\phi(\cdot,0)\, dx \ge 0$$

where

$$U(z) := \int_0^z h(\zeta)\, d\zeta, \qquad F(z) := \int_0^z h(\zeta)\, df(\zeta).$$

For $h \equiv \pm 1$, this yields definition 3.1.

After establishing uniqueness of the viscosity solution, we can cite a result on the uniqueness of the entropy solution (cf. [Eva10, chapter 3, theorem 3]):

Theorem 3.8. *Assume f is convex and smooth. Then there exists — up to a set of measure zero — at most one entropy solution of* (3.5)*.*

For convex f, it is even possible to explicitly state the solution u of the conservation law for $u^0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. This theorem goes back to Peter Lax (cf. [Lax73, theorem 3.2 and §4]).

Theorem 3.9. *Let $f \in C^2(\mathbb{R})$, $f(0) = 0$, be strictly convex and $u^0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. Since f' is monotone, its inverse exists and we define*

$$\begin{aligned} a(z) &:= f'(z) \\ b(z) &:= a^{-1}(z) \\ g(z) &:= zb(z) - f(b(z)) \\ v^0(x) &:= \int_{-\infty}^{x} u^0(y)\,dy \end{aligned}$$

for all $x, z \in \mathbb{R}$. For $(y,x,t) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$, we define

$$\begin{aligned} G(y,x,0) &:= v^0(y) \\ G(y,x,t) &:= v^0(y) + tg\left(\frac{x-y}{t}\right) \end{aligned}$$

for all $t > 0$. Then

$$u(x,t) = b\left(\frac{x - y(x,t)}{t}\right) = u^0(y(x,t)),$$

where $y = y(x,t)$ is given by

$$G(y(x,t);x,t) = \min_{z \in \mathbb{R}} G(z;x,t)$$

for almost all $(x,t) \in \mathbb{R} \times \mathbb{R}_+$, is well-defined and is the uniquely defined viscosity solution of the conservation law (3.5)*. u has at most countably many points of discontinuity at all times.*

This result is known as *Lax representation formula*.

3.1.6 Systems of Conservation Laws

Most of the above results can be carried over to hyperbolic systems. Consider the linear system

$$\partial_t \boldsymbol{u} + \partial_x \boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{0} \tag{3.15}$$

for $\boldsymbol{u} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}^m$ and $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$ with $\boldsymbol{A} \in \mathbb{R}^{m,m}$. By definition of hyperbolicity, there is a matrix $\boldsymbol{R} \in \mathbb{R}^{m,m}$ such that

$$\boldsymbol{A} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1},$$

where $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ and $\lambda_1 \leq \cdots \leq \lambda_m$ are the eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{R}$ contains the eigenvectors $\boldsymbol{r}_l$, $l = 1, \ldots, m$. Now we can conduct a transformation of variables via $\boldsymbol{v} := \boldsymbol{R}^{-1}\boldsymbol{u}$ to obtain

$$\partial_t \boldsymbol{v} + \boldsymbol{D}\partial_x \boldsymbol{v} = \boldsymbol{0},$$

a system of scalar advection equations

$$\partial_t v_l + \lambda_l \partial_x v_l = 0, \qquad l = 1, \dots, m,$$

with solution

$$v_l(x,t) = v_l^0(x - \lambda_l t)$$

where $\boldsymbol{v}^0 = \boldsymbol{R}^{-1}\boldsymbol{u}^0$ and $\boldsymbol{u}^0 = \boldsymbol{u}(x,0)$ is the initial condition. We recover the solution to the original system (3.15) by

$$\boldsymbol{u}(x,t) = \sum_{l=1}^{m} v_l(x,t)\boldsymbol{r}_l = \sum_{l=1}^{m} v_l^0(x - \lambda_l t)\boldsymbol{r}_l.$$

The curve $t \mapsto x_0 + \lambda_l t$ is called the *lth characteristic*. The coefficient v_l is constant along this line.

To understand Riemann problems for linear systems, consider a strictly hyperbolic problem with initial condition

$$\boldsymbol{u}^0(x) = \begin{cases} \boldsymbol{u}_L & x < 0 \\ \boldsymbol{u}_R & x > 0. \end{cases}$$

Now decompose the two states into an eigenbasis

$$\boldsymbol{u}_L = \sum_{l=1}^{m} \alpha_l \boldsymbol{r}_l, \qquad \boldsymbol{u}_R = \sum_{l=1}^{m} \beta_l \boldsymbol{r}_l.$$

In the lth component we have

$$v_l^0(x) = \begin{cases} \alpha_l & x < 0 \\ \beta_l & x > 0, \end{cases}$$

which leads to the solution

$$v_l(x,t) = \begin{cases} \alpha_l & x - \lambda_l t < 0 \\ \beta_l & x - \lambda_l t > 0. \end{cases}$$

This yields

$$\boldsymbol{u}(x,t) = \sum_{x - \lambda_l t > 0} \beta_l \boldsymbol{r}_l + \sum_{x - \lambda_l t < 0} \alpha_l \boldsymbol{r}_l. \tag{3.16}$$

As an illustration, consider an example taken from [LeV06, chapter 6.5] shown in figure 3.3. In this example, $v_1 = \beta_1$ while $v_2 = \alpha_2$ and $v_3 = \alpha_3$. Note that the solution at any point in the wedge between the first and the second characteristic is

$$\boldsymbol{u}(x,t) = \beta_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2 + \alpha_3 \boldsymbol{r}_3.$$

As we cross the lth characteristic, the value of $x - \lambda_l t$ passes through zero and the corresponding v_l jumps from α_l to β_l while the other coefficients remain constant.

The solution is constant in each of the wedges shown in figure 3.4. Denote by

$$[\boldsymbol{u}]_{\boldsymbol{\gamma}_l} = (\beta_l - \alpha_l)\boldsymbol{r}_l$$

the jump across the lth characteristic $\boldsymbol{\gamma}_l$. The jump condition for our system then reads

$$[\boldsymbol{f}(\boldsymbol{u})]_{\boldsymbol{\gamma}_l} = \lambda_l [\boldsymbol{u}]_{\boldsymbol{\gamma}_l}$$

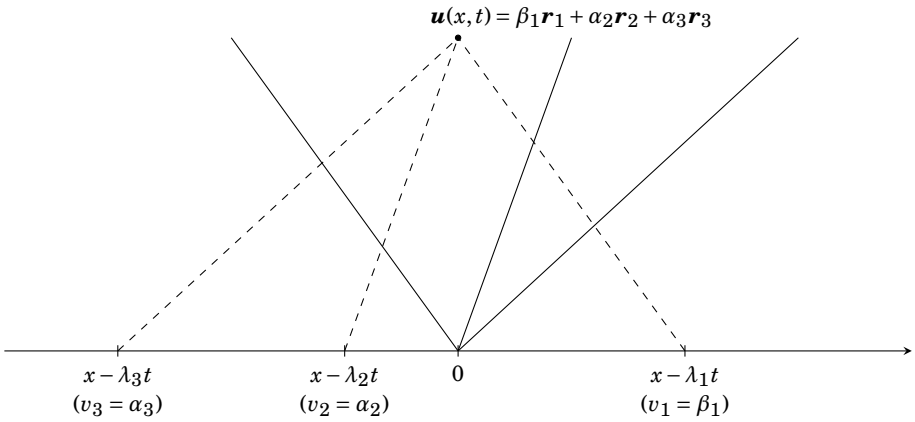

Figure 3.3: *Construction of solution to Riemann problem at* (x,t)

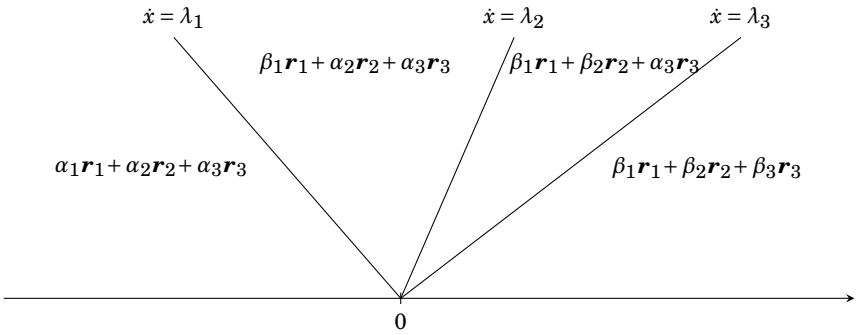

Figure 3.4: *Solution in x-t-plane*

and λ_l is precisely the propagation speed of this jump. We can alternatively write the solution in (3.16) in terms of these jumps as

$$\boldsymbol{u}(x,t) = \boldsymbol{u}_L + \sum_{\lambda_l < \frac{x}{t}} (\beta_l - \alpha_l)\boldsymbol{r}_l = \boldsymbol{u}_R + \sum_{\lambda_l \geq \frac{x}{t}} (\beta_l - \alpha_l)\boldsymbol{r}_l.$$

Unless $\boldsymbol{u}_R - \boldsymbol{u}_L$ is an eigenvector of $\boldsymbol{A}$, the jump cannot propagate as a single discontinuity with any speed without violating the Rankine-Hugoniot condition. An interpretation of the solution of the Riemann problem is finding a decomposition of jumps

$$\boldsymbol{u}_R - \boldsymbol{u}_L = \sum_{l=1}^{m} (\beta_l - \alpha_l)\boldsymbol{r}_l.$$

Each of these can propagate at an appropriate speed λ_l while satisfying the Rankine-Hugoniot condition.

The question arises, which $\boldsymbol{u}_R$ can be reached from a fixed $\boldsymbol{u}_L$ by one shock? We consider a system of two equations for simplicity. A discontinuity with left and right states $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$ can propagate as a single discontinuity only if $\boldsymbol{u}_R - \boldsymbol{u}_L$ is an eigenvector of $\boldsymbol{A}$, so the line segment connecting both states in the phase plane must be parallel to

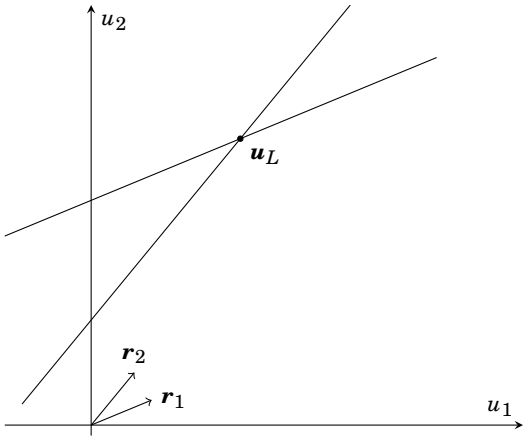

Figure 3.5: *The Hugoniot locus of $\boldsymbol{u}_L$ consists of all states that can be reached from $\boldsymbol{u}_L$ by a scalar multiple of $\boldsymbol{r}_1$ or $\boldsymbol{r}_2$*

$\boldsymbol{r}_1$ or $\boldsymbol{r}_2$. These lines yield the so-called *Hugoniot locus*, the set of all points that can be connected to $\boldsymbol{u}_L$ in the described way.

For a general Riemann problem with arbitrary states $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$, the solution consists of two discontinuities traveling with velocities λ_1 and λ_2. In between, there is a new constant state

$$\boldsymbol{u}_M = \beta_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2$$

so that $\boldsymbol{u}_M - \boldsymbol{u}_L = (\beta_1 - \alpha_1)\boldsymbol{r}_1$ and $\boldsymbol{u}_R - \boldsymbol{u}_M = (\beta_2 - \alpha_2)\boldsymbol{r}_2$.

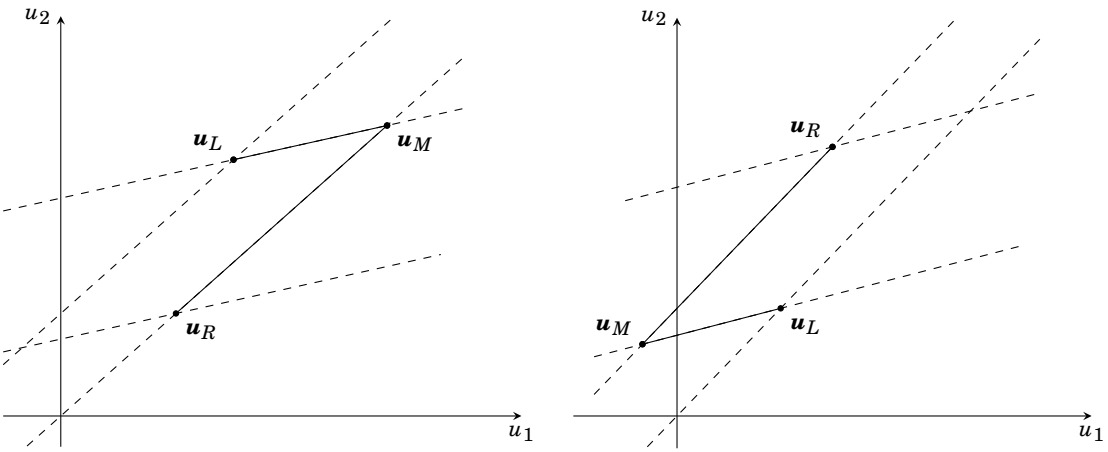

Figure 3.6: *The new state $\boldsymbol{u}_M$ arising in the solution to the Riemann problem for two different choices of $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$*

Nonlinear Systems

Until now, we have considered linear systems of conservation laws. If $\boldsymbol{f}$ in (3.15) in nonlinear, the system can be written in quasilinear form

$$\partial_t \boldsymbol{u} + \boldsymbol{A}(\boldsymbol{u})\partial_x \boldsymbol{u} = \boldsymbol{0}$$

with $\boldsymbol{A}(\boldsymbol{u}) = \boldsymbol{f}'(\boldsymbol{u})$ (cf. [LeV06, chapter 6.4]). The system is hyperbolic if and only if $\boldsymbol{A}(\boldsymbol{u})$ is diagonalizable and has m real eigenvalues $\lambda_1(\boldsymbol{u}) \leq \cdots \leq \lambda_m(\boldsymbol{u})$ for all values of $\boldsymbol{u}$, at

least in some range where the solution is known to lie, and strictly hyperbolic if the eigenvalues are distinct, which we will assume here.

The lth characteristic is defined as the solution to the ordinary differential equation

$$\begin{aligned}\dot{x}(t) &= \lambda_l\big(\boldsymbol{u}(x(t),t)\big)\\ x(0) &= x_0,\end{aligned}$$

where the eigenvalues now depend on the solution $\boldsymbol{u}$, which changes in all but the lth component, so $\dot{x}$ is not constant anymore. Hence it is quite complicated to determine characteristics globally. We can, however, use a local approach, which still yields valuable information about the solution. If we linearize the problem about a constant state $\bar{\boldsymbol{u}}$, we obtain a constant coefficient linear system. Assume an expansion of the solution of the form

$$\boldsymbol{u}^\varepsilon(x,t) = \bar{\boldsymbol{u}} + \varepsilon \boldsymbol{u}^{(1)}(x,t) + \mathcal{O}(\varepsilon^2)$$

for small ε. This yields

$$\partial_t \boldsymbol{u}^\varepsilon + \boldsymbol{A}(\boldsymbol{u}^\varepsilon)\partial_x \boldsymbol{u}^\varepsilon = \Big(\partial_t \boldsymbol{u}^{(1)} + \boldsymbol{A}(\bar{\boldsymbol{u}})\partial_x \boldsymbol{u}^{(1)}\Big)\varepsilon + \mathcal{O}(\varepsilon^2) = 0.$$

For small ε and short times $t \ll \frac{1}{\varepsilon}$ we find that $\boldsymbol{u}^{(1)}$ satisfies

$$\partial_t \boldsymbol{u} + \boldsymbol{A}(\bar{\boldsymbol{u}})\partial_x \boldsymbol{u} = \boldsymbol{0}.$$

Thus, small disturbances propagate approximately along characteristic curves of the form $x_l(t) = x_0 + \lambda_l(\bar{\boldsymbol{u}})t$. Of course, retaining more terms in the expansion yields higher order corrections for nonlinear problems.

As opposed to the linear case where discontinuities propagate only along characteristics, for nonlinear systems we need again the jump condition

$$\boldsymbol{f}(\boldsymbol{u}_L) - \boldsymbol{f}(\boldsymbol{u}_R) = S(\boldsymbol{u}_L - \boldsymbol{u}_R)$$

with the shock velocity S. For weak shocks, however, some linear theory is still applicable. Suppose $\|\boldsymbol{u}_L - \boldsymbol{u}_R\| \le \varepsilon$ then

$$\boldsymbol{f}(\boldsymbol{u}_L) = \boldsymbol{f}(\boldsymbol{u}_R) + \boldsymbol{f}'(\boldsymbol{u}_R)(\boldsymbol{u}_L - \boldsymbol{u}_R) + \mathcal{O}(\varepsilon^2).$$

With the Rankine-Hugoniot condition we obtain

$$\boldsymbol{f}'(\boldsymbol{u}_R)(\boldsymbol{u}_L - \boldsymbol{u}_R) = S(\boldsymbol{u}_L - \boldsymbol{u}_R) + \mathcal{O}(\varepsilon^2).$$

Thus, as $\varepsilon \to 0$, the normalized vector $\frac{1}{\epsilon}(\boldsymbol{u}_L - \boldsymbol{u}_R)$ must approach an eigenvector $\boldsymbol{v}$ of $\boldsymbol{f}'(\boldsymbol{u}_R) = \boldsymbol{A}(\boldsymbol{u}_R)$, say $\boldsymbol{v} = \boldsymbol{r}_l(\boldsymbol{u}_R)$, with S approaching the corresponding eigenvalue $\lambda_l(\boldsymbol{u}_R)$.

But what about the general nonlinear case? For a fixed state $\boldsymbol{u}_L$, we seek all $\boldsymbol{u}$ that can be connected to $\boldsymbol{u}_L$ by a discontinuity. Of course, the Rankine-Hugoniot jump condition

$$\boldsymbol{f}(\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{u}_L) = S(\boldsymbol{u} - \boldsymbol{u}_L)$$

has to be satisfied for some $S \in \mathbb{R}$. That makes $m+1$ unknowns ($\boldsymbol{u}$ and S) in a system of m equations, so we should expect a one parameter family of solutions. If we denote this parameter by ζ and the parametrized solutions by $\boldsymbol{u}(\zeta)$ and $S(\zeta)$ with $\boldsymbol{u}(0) = \boldsymbol{u}_L$, the jump condition reads

$$\boldsymbol{f}(\boldsymbol{u}(\zeta)) - \boldsymbol{f}(\boldsymbol{u}_L) = S(\zeta)(\boldsymbol{u}(\zeta) - \boldsymbol{u}_L).$$

Differentiating with respect to ζ and setting $\zeta = 0$ yields

$$\boldsymbol{f}'(\boldsymbol{u}_L)\boldsymbol{u}'(0) = S(0)\boldsymbol{u}'(0)$$

so $\boldsymbol{u}'(0)$ has to be a scalar multiple of some eigenvector $\boldsymbol{r}_l(\boldsymbol{u}_L)$ of $\boldsymbol{f}'(\boldsymbol{u}_L)$ and $S(0) = \lambda_l(\boldsymbol{u}_L)$. Hence the curve $\boldsymbol{u}(\zeta)$ is tangent to $\boldsymbol{r}_l(\boldsymbol{u}_L)$ in $\boldsymbol{u}_L$. This fits with the linearized approach for weak shocks. By strict hyperbolicity, there are m such linear independent directions. From the implicit function theorem we obtain local existence of these solution curves in a neighborhood of $\boldsymbol{u}_L$. Refer to [Lax73] for more details. These curves are called *Hugoniot curves* and the set of all points on these curves is often called the *Hugoniot locus* for the point $\boldsymbol{u}_L$. Solving a Riemann problem for sufficiently small $\|\boldsymbol{u}_L - \boldsymbol{u}_R\|$ then works the same as in the linear case: we draw the Hugoniot locus for both states and look for the intersections where the intermediate state $\boldsymbol{u}_M$ lies. Recall that we have to follow the path in increasing order of the eigenvalues, just like in the linear case. In general, we cannot hope to get more than this local result.

Recall from the scalar case the concept of entropy conditions and viscosity solutions. We have to think about how we can transfer these concepts to the system case. Until now we have ignored entropy conditions for nonlinear systems. We have only ensured the jump condition, but do not know if the shocks are admissible and could thus exist in a viscosity solution. We also have not discussed rarefaction waves, which were contained in the viscosity solution of scalar equations.

Recall the Lax entropy condition in the scalar case

$$f'(u_L) > S > f'(u_R).$$

This can be generalized to genuinely nonlinear systems. The lth characteristic field is said to be *genuinely nonlinear* if

$$\nabla_{\boldsymbol{u}} \lambda_l(\boldsymbol{u}) \cdot \boldsymbol{r}_l(\boldsymbol{u}) \neq 0 \qquad \text{for all } \boldsymbol{u}$$

where $\nabla_{\boldsymbol{u}}$ is the gradient with respect to $\boldsymbol{u}$. For $m = 1$, this reduces to the requirement $f'' \neq 0$ for all u — a convexity requirement. For a genuinely nonlinear field, Lax's entropy condition says that a direct jump from $\boldsymbol{u}_L$ to $\boldsymbol{u}_R$ is only allowed if

$$\lambda_l(\boldsymbol{u}_L) > S > \lambda_l(\boldsymbol{u}_R).$$

Characteristics of the lth family disappear into the shock just as in the scalar case.

For the discussion of rarefaction waves, recall the special solutions $w(\xi)$ we considered for scalar equations. We differentiate $\boldsymbol{u}(x,t) = \boldsymbol{w}(x/t)$ and obtain

$$\begin{aligned}\partial_t \boldsymbol{w}(\xi) &= -\frac{x}{t^2}\boldsymbol{w}'(\xi) \\ \partial_x \boldsymbol{f}(\boldsymbol{w})) &= \boldsymbol{f}'(\boldsymbol{w})\boldsymbol{w}'(\xi)\frac{1}{t}.\end{aligned}$$

The conservation law then yields

$$0 = \frac{1}{t}\left(-\xi \boldsymbol{w}'(\xi) + \boldsymbol{f}'(\boldsymbol{w}(\xi))\boldsymbol{w}'(\xi)\right)$$

or

$$\boldsymbol{f}'(\boldsymbol{w}(\xi))\boldsymbol{w}'(\xi) = \xi \boldsymbol{w}'(\xi),$$

so $\boldsymbol{w}'(\xi)$ must be some eigenvector $\boldsymbol{r}_l(\boldsymbol{w}(\xi))$ of $\boldsymbol{f}'(\boldsymbol{w}(\xi))$ to the eigenvalue ξ. Hence the values $\boldsymbol{w}(\xi)$ all lie along some integral curve of $\boldsymbol{r}_l$. This is a curve for $\boldsymbol{r}_l(\boldsymbol{u})$ whose tangent at any point $\boldsymbol{u}$ is parallel to $\boldsymbol{r}_l(\boldsymbol{u})$. The states $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$ can be connected by a rarefaction wave if they lie on the same integral curve and if

$$\lambda_l(\boldsymbol{u}_L) < \lambda_l(\boldsymbol{u}_R).$$

Refer to [LeV06] for a more thorough discussion of hyperbolic systems of conservation laws and examples from isothermal gas dynamics.

3.2 Numerical Solution of Conservation Laws

Now let us discuss some techniques that are commonly used for the approximate solution of conservation laws. A thorough discussion of this topic can be found in [Kro97] or [LeV11], for example.

We consider one-dimensional problems

$$\frac{\partial}{\partial t}u + \frac{\partial}{\partial x}f(u) = 0 \qquad \text{on } \Omega \times \mathbb{R}_+ \tag{3.17}$$

for some interval $\Omega \subset \mathbb{R}$ for the course of this section.

From the previous section on the theory of hyperbolic conservation laws we know about the difficulties we have to deal with. Especially the formation of shocks is a big issue. When the solution becomes discontinuous, Taylor series expansions and the order of accuracy are no longer an adequate tool for the analysis of numerical schemes.

The simplest approach for the approximation of partial differential equations is using finite differences. Since we are concerned with conservation laws, a different class of methods is suited much better to accomplish this task.

3.2.1 Finite Volume Methods

For the derivation of finite volume methods, we consider the one-dimensional conservation law in integral form (3.1)

$$\int_I \big(\partial_t u(x,t) + \partial_x f(u)\big)\, dx = 0 \qquad \text{for all } I \subset \Omega. \tag{3.18}$$

If $\Omega = \mathbb{R}$ we can avoid boundary conditions altogether, but have to restrict our computational domain. Hence, we assume $\Omega \subset \mathbb{R}$ to be bounded. For simplicity, we use periodic boundary conditions because we are not interested in any effects at the boundaries. Of course, even if the problem is defined for all $t > 0$, we can only simulate finite time intervals $[0,T] \subset \mathbb{R}_+$.

We divide the domain Ω into so-called cells $D_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ such that $\bigcup_j \overline{D}_j = \overline{\Omega}$ and $D_{j_1} \cap D_{j_2} = \emptyset$ for $j_1 \neq j_2$. The time domain is divided into $0 = t_0 < t_1 < \ldots < t_M = T$ with step size $\Delta t_n = t_n - t_{n-1}$. For ease of notation, we assume $\Delta t_n \equiv \Delta t$, for all $n = 1, \ldots, M$, so $t_n = n\Delta t$.

With $I = D_j$ in (3.18), integrating in time yields

$$\int_{D_j} u(x,t_{n+1})dx - \int_{D_j} u(x,t_n)dx = \int_{t_n}^{t_{n+1}} f\big(u(x_{j-\frac{1}{2}},t)\big)\,dt - \int_{t_n}^{t_{n+1}} f\big(u(x_{j+\frac{1}{2}},t)\big)\,dt,$$

which — assuming constant mesh size $|D_j| = \Delta x_j \equiv \Delta x$ and $x_j = j\Delta x$ for ease of notation — leads to

$$\frac{1}{\Delta x}\int_{D_j} u(x,t_{n+1})\,dx = \frac{1}{\Delta x}\int_{D_j} u(x,t_n)\,dx - \frac{1}{\Delta x}\int_{t_n}^{t_{n+1}} \left[f\big(u(x_{j+\frac{1}{2}},t)\big) - f\big(u(x_{j-\frac{1}{2}},t)\big)\right] dt.$$

The conservation property tells us that

$$\int_\Omega u(x,t)\,dx = \int_\Omega u^0(x)\,dx,$$

so it suggests itself to approximate the cell averages

$$u_j^n \approx \frac{1}{\Delta x}\int_{D_j} u(x,t_n)\,dx, \qquad u_j^0 := \frac{1}{\Delta x}\int_{D_j} u^0(x)\,dx, \qquad x \in D_j$$

in the numerical scheme.

Now we need an approximation for the integrals of f,

$$g(u_j^n, u_{j+1}^n) \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f\big(u(x_{j+\frac{1}{2}},t)\big)\,dt.$$

Definition 3.10. *A function $g : \mathbb{R}^2 \to \mathbb{R}$ is called* numerical flux *to $f : \mathbb{R} \to \mathbb{R}$ if $g(z,z) = f(z)$ for all $z \in \mathbb{R}$ and for some constant K*

$$|g(v,w) - f(z)| = |g(v,w) - g(z,z)| \le K \max\{|v-z|, |w-z|\}.$$

The first property is called consistency.

If a wider stencil is desired, the number of arguments of g can be increased accordingly.

Definition 3.11. *Let g be a numerical flux to f and u_j^0 as defined above. Then we call the scheme*

$$u_j^{n+1} = u_j^n - \lambda\left(F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}\right) \tag{3.19}$$

with $\lambda := \frac{\Delta t}{\Delta x}$ and $F_{j+\frac{1}{2}} := g(u_j^n, u_{j+1}^n)$ to be in conservation form.

The specific method obtained depends on the choice of the numerical fluxes $F_{j+\frac{1}{2}}$. The term *to be in conservation form* is explained by their mimicking the conservation property: Assume we have N cells D_j, $j = j_1, \ldots, j_N$. If we sum up (3.19) over all those cells,

$$\sum_{j=j_1}^{j_N} u_j^{n+1} = \sum_{j=j_1}^{j_N} u_j^n - \lambda \sum_{j=j_1}^{j_N} (F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}) = \sum_{j=j_1}^{j_N} u_j^n - \lambda(F_{j_{N+\frac{1}{2}}} - F_{j_{\frac{1}{2}}})$$

where only the fluxes at the boundaries of Ω remain in the telescopic sum. Since we assume periodic boundary conditions, these just cancel out and we have

$$\sum_{j=j_1}^{j_N} u_j^{n+1} = \sum_{j=j_1}^{j_N} u_j^n,$$

which corresponds to the conservation property

$$\frac{1}{\Delta x}\int_\Omega u(x,t_{n+1})\,dx = \frac{1}{\Delta x}\int_\Omega u(x,t_n)\,dx.$$

The above methods can also be interpreted as finite difference approximations to (3.17). Denote by v_Δ the functions that are piecewise constant on $D_j \times [t_n, t_{n+1})$, $j \in \{j_1,\dots,j_N\}$, $n \geq 0$. Then let $u_\Delta^n(x) = u_\Delta(x,t_n)$ and build a vector $U^n = (u_j^n)_j$ with $u_j^n = u_\Delta^n(x_j) = u_\Delta(x_j,t_{n+1})$. The scheme (3.19) then can be written as $U^{n+1} = H(U^n)$.

By applying a conservative method, we hope to correctly approximate discontinuous solutions of the conservation law. The *Lax-Wendroff theorem* (cf. [Kro97, theorem 2.3.1] or [LeV11, theorem 12.1]) tells us just that:

Theorem 3.12. *Consider sequences Δt, $\Delta x \to 0$, u_Δ a family of approximate solutions discretized in conservation form and*

$$\sup_\Delta \|u_\Delta\|_{L^\infty(\mathbb{R}\times\mathbb{R}_+)} \leq K_1$$
$$\sup_{n,\Delta} \|u_\Delta^n\|_{BV(\mathbb{R})} \leq K_2$$

for positive constants K_1 and K_2 and $u_\Delta \to u$ almost everywhere for some $u : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}$. Then u is a solution of the weak conservation law.

We do not get convergence out of this theorem, but it tells us that if we have convergence, it is towards a weak solution. It is, however, important that the method is in conservation form. Refer to [HL94] for a detailed error analysis of nonconservative schemes.

3.2.2 Stability

We have already seen how difficult the analytical solution of hyperbolic conservation laws is. The numerical solution of partial differential equations is no easy task, either, especially when we have to meet the requirements derived from the theoretical aspects. Before we start thinking about how we can fulfill any entropy conditions, let us start with the classical approach and consider (linear) stability.

Consider a numerical scheme of the form

$$U^{n+1} = H(U^n). \tag{3.20}$$

Now we cite the definitions of consistency and convergence from [Kro97, section 2.4] for all $t \leq T$ for some fixed T:

Definition 3.13. *A scheme* (3.20) *is said to be* consistent of order (q,p) with respect to the norm $\|\cdot\|$ *if the local error satisfies*

$$\|u(\cdot,t_{n+1}) - H(u(\cdot,t_n))\| = \mathcal{O}(\Delta t^{q+1}) + \mathcal{O}(\Delta x^p \Delta t)$$

for any smooth solution u.

Definition 3.14. *Let u be the exact, u_Δ the discrete solution. The scheme* (3.20) converges with respect to the norm $\|\cdot\|$ *if*

$$\|u - u_\Delta\| \to 0 \ \text{ for } \ \Delta t, \Delta x \to 0.$$

It is convergent of order (q,p) *if*

$$\|u(\cdot,t_n)-u_\Delta^n\| = \mathcal{O}(\Delta t^q)+\mathcal{O}(\Delta x^p)$$

uniformly for all $n\in\mathbb{N}$.

For linear problems, i.e., linear f and hence linear H, we call (3.20) *stable* if $\|H^n\|\le C$ for all $n\in\mathbb{N}$ and a constant C that does not depend on Δt or Δx. That means the nth power of the operator H is uniformly bounded. This form of stability is sometimes called *Lax-Richtmyer stability* (cf. [LeV11, section 8.3.2]). If $\|H\|\le 1+\alpha\Delta t$ for some constant α that is independent of Δt as $\Delta t\to 0$, we can take $C=e^{\alpha T}$. Furthermore, we have the famous *Lax equivalence theorem* (cf. [RM67]):

Theorem 3.15. *Given a well-posed initial value problem and a consistent finite difference approximation to it, stability is necessary and sufficient for convergence.*

Remark 3.16 (Order of convergence). More specifically, if $\|u_\Delta^0-u^0\|=\mathcal{O}(\Delta x^p)$, then a stable and order (q,p) consistent scheme is convergent of order (q,p), i.e., $\|u(\cdot,t^n)-u_\Delta^n\| = \mathcal{O}(\Delta t^q)+\mathcal{O}(\Delta x^p)$ uniformly for all $n\le\frac{T}{\Delta t}$.

Von Neumann Analysis

How can we find out if a method is stable? For stability in the 2-norm, a very powerful tool for the analysis of numerical schemes is the Fourier analysis. Refer to [Str04, chapter 2] for a detailed introduction. We consider the linear advection equation (3.4) and a numerical scheme (3.20).

Suppose that u_j^n is bounded so we can apply the inverse Fourier transform

$$u_j^n = \frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{i\xi j\Delta x}\hat{u}^n(\xi)\,d\xi.$$

Substituting this into the linear finite difference method for U^n typically yields an expression of the form

$$u_j^{n+1} = \frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{i\xi j\Delta x} g(\xi,\Delta x,\Delta t)\hat{u}^n(\xi)\,d\xi$$

with some function g. Comparing this to the Fourier inversion formula for u^{n+1},

$$u_j^{n+1} = \frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{\Delta x}} e^{i\xi j\Delta x}\hat{u}^{n+1}(\xi)\,d\xi,$$

and using the uniqueness of the Fourier transform, we obtain

$$\hat{u}^{n+1}(\xi) = g(\xi,\Delta x,\Delta t)\,\hat{u}^n(\xi).$$

This shows that advancing the solution of the scheme by one time step is equivalent to multiplying the Fourier transform of the solution by the *amplification factor* $G(\xi):=g(\xi,\Delta x,\Delta t)$, so we see

$$\hat{u}^n(\xi) = G(\xi)^n\hat{u}^0(\xi)$$

and $|G(\xi)|\le 1$ is a sufficient condition for stability.

The most famous example of an unstable method is the so-called FTCS (forward in time, centered in space) scheme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\Delta x}(f_{j+1} - f_{j-1})$$

i.e., $F_{j+\frac{1}{2}} = \frac{1}{2}(f_{j+1} + f_j)$ with $f_j^n = f(u_j^n)$. For the linear advection equation (3.4), this reads

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}a(u_{j+1}^n - u_{j-1}^n),$$

where

$$\lambda = \frac{\Delta t}{\Delta x},$$

so we have

$$G(\xi) = 1 - \frac{\lambda}{2}ae^{i\xi\Delta x} + \frac{\lambda}{2}ae^{-i\xi\Delta x} = 1 - i\lambda a\sin(\xi\Delta x)$$

by Euler's relation. Then we see

$$|G(\xi)|^2 = |1 - i\lambda a\sin(\xi\Delta x)|^2 = 1 + (\lambda a)^2\sin^2(\xi\Delta x) \geq 1,$$

so the amplification factor is greater than unity and the scheme blows up in finite time, i.e., is unstable. Since we never have $G(\xi) < 1$, FTCS is *unconditionally unstable*.

A first order scheme for the advection equation $u_t + au_x = 0$ is the *upwind method*, which is defined by

$$F_{j+\frac{1}{2}}^{up} = \begin{cases} au_{j+1}^n & a < 0 \\ au_j^n & a \geq 0. \end{cases} \tag{3.21}$$

Applying von Neumann analysis to the upwind method for $a \geq 0$,

$$u_j^{n+1} = u_j^n - \lambda a(u_j^n - u_{j-1}^n)$$

we obtain

$$G(\xi) = 1 - \lambda a + \lambda ae^{i\xi\Delta x} = 1 - \lambda a + \lambda a\big(\cos(\xi\Delta x) + i\sin(\xi\Delta x)\big)$$

and

$$|G(\xi)|^2 = \big(1 - \lambda a + \lambda a\cos(\xi\Delta x)\big)^2 + \sin^2(\xi\Delta x) = 1 - 2\lambda a(1 - \lambda a)\big(1 - \cos(\xi\Delta x)\big)$$

and thus $|G(\xi)| \leq 1$ if and only if

$$2\lambda a(1 - \lambda a) \geq 0$$

or

$$0 \leq \lambda a \leq 1$$

which is the *Courant-Friedrichs-Lewy (CFL) condition*. We also see here why taking the upwind side is important: For $a < 0$, this method is unstable for all Δt, Δx!

Another famous method is the *Lax-Friedrichs scheme*. It uses the flux

$$F_{j+\frac{1}{2}}^{LF} = \frac{1}{2}(f_{j+1}^n + f_j^n) - \frac{1}{2\lambda}(u_{j+1}^n - u_j^n). \tag{3.22}$$

The scheme then reads

$$\begin{aligned} u_j^{n+1} &= \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - \frac{\lambda}{2}(f_{j+1}^n - f_{j-1}^n) \\ &= \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - \frac{\lambda}{2}a(u_{j+1}^n - u_{j-1}^n) \end{aligned}$$

for the linear advection equation, so it can be interpreted as a modified FTCS scheme. For the Lax-Friedrichs scheme, we find the sufficient stability condition

$$|\lambda a| \leq 1.$$

While FTCS is second order accurate in space, we pay the stability of the Lax-Friedrichs method with an order reduction — like the upwind method, it is only first order accurate in space.

The physical interpretation of the stability condition is that the numerical speed of propagation $\frac{\Delta x}{\Delta t}$ is greater than or equal to the propagation speed of the physical problem $|a|$, i.e., the numerical domain of dependence is greater than the physical.

3.2.3 Nonlinear Stability

For the nonlinear effects in hyperbolic conservation laws (3.17), we need a more detailed analysis since order of accuracy is deceiving near steep gradients and shocks. Thus for the nonlinear case, which we are interested in, a different approach is needed. The theoretical concepts of weak solutions, viscosity approximations and entropy conditions we have discussed in section 3.1 shall help us understand the discrete case. [Kro97], [LeV06] and [LeV11] are extensive resources for this topic.

Definition 3.17. *Let (U,F) be an entropy pair as defined in section 3.1.5 and G a numerical flux to F as in definition 3.10. Then*

$$U(u_j^{n+1}) \leq U(u_j^n) - \lambda\left(G^n_{j+\frac{1}{2}} - G^n_{j-\frac{1}{2}}\right) \tag{3.23}$$

is called the discrete entropy condition*.*

With this definition, we obtain a discrete entropy solution:

Theorem 3.18. *If under the assumptions in theorem 3.12, u_Δ satisfies the discrete entropy condition* (3.23)*, then u_Δ converges to a weak solution that fulfills the entropy inequality* (3.14)*.*

The proof can be found in [Kro97, theorem 2.3.4]. Again, we only know that in case of convergence, our scheme converges to the entropy solution, but we do not know if it converges at all. To establish that, we need another property.

Definition 3.19. *The discrete solution is called* BV-stable *or* total variation bounded (TVB) *if there is a constant $R > 0$ such that*

$$\|u_\Delta^n\|_{BV(\mathbb{R})} \leq R$$

for all Δ. A method is called total variation diminishing (TVD) *if*

$$\|u_\Delta(\cdot, t_{n+1})\|_{BV(\mathbb{R})} \leq \|u_\Delta(\cdot, t_n)\|_{BV(\mathbb{R})}$$

for all $n \geq 0$.

The BV-norm measures total variation, so the condition for BV-stability means boundedness of the total variation while for TVD, the total variation of the discrete solution does not increase.

Finally, we can state a result on convergence (cf. [Kro97, theorem 2.3.9]):

Theorem 3.20. *If a method is BV-stable, then there exists a function $u \in L^1_{loc}(\mathbb{R} \times \mathbb{R}_+)$, for which $u_\Delta \to u$ almost everywhere. The requirements of theorem 3.12 imply that u is a weak solution, those of theorem 3.18 imply that u is the unique weak entropy solution.*

Of course, these requirements are hard to ensure in practice, so one might not be able to make use of the theorem and thus other notions of nonlinear stability were explored. Until today, there are many different definitions of stability for nonlinear problems. Most of them have their origin in some kind of behavior that is expected or desired of the solution to the problem at hand. The following definitions all refer to the Cauchy problem to the conservation law (3.17).

When we are dealing with mass and density, the most fundamental physical property is non-negativity. This leads to

Definition 3.21. *A numerical method is called* positivity preserving *if for non-negative initial data $u^0_\Delta \geq 0$, the solution remains non-negative, $u^n_\Delta \geq 0$ for all $n \geq 0$.*

The slight inconsistency in nomenclature results from the need to allow the solutions to be zero in the simulation of laser-plasma interaction like in a vacuum-plasma transition. In the literature, the case $u_\Delta = 0$ is often omitted from the definition because it can be problematic numerically. A value that is greater than zero is less likely to become negative in finite precision arithmetic.

One major problem is the formation of spurious oscillations near discontinuities. We will discuss this in more detail later. An important property that addresses this issue is given by

Definition 3.22. *A method (in conservation form) is called* monotonicity preserving *if the monotonicity (either non-increasing or non-decreasing) of the initial data u^0_Δ implies the same property for u^n_Δ for $n > 0$, i.e., $u^0_j \geq u^0_{j+1}$ implies $u^n_j \geq u^n_{j+1}$ and $u^0_j \leq u^0_{j+1}$ implies $u^n_j \leq u^n_{j+1}$*

Since the Riemann initial data are monotone, oscillations cannot appear near an isolated propagating discontinuity with monotonicity preserving methods. Also, it is easy to see that oscillations increase the total variation and therefore TVD methods are monotonicity preserving (cf. [Kro97, lemma 2.3.13]).

For the advection equation, we can write the method as

$$u^{n+1}_j = \sum_m \gamma_m u^n_{j+m}$$

where the summation is carried out over the finite set of indices $m \in \mathbb{Z}$ that are needed for the stencil of the scheme. Monotonicity preservation is then equivalent to $\gamma_m \geq 0$ for all m and TVD is equivalent to $\gamma_m \geq 0$ for all m and $\sum_m \gamma_m \leq 1$ (cf. [Wes01, section 9.2]). Obviously this is a sufficient condition for positivity preservation.

Another simple way to check if a method is TVD is shown by *Harten's lemma* (cf. [Har83]):

Lemma 3.23. *A method that can be written as*

$$u^{n+1}_i = u^n_i + D^+_{i+\frac{1}{2}}(u^n_{i+1} - u^n_i) - D^-_{i-\frac{1}{2}}(u^n_i - u^n_{i-1})$$

with $D^+_{i+\frac{1}{2}} \geq 0$, $D^-_{i-\frac{1}{2}} \geq 0$ and $D^+_{i+\frac{1}{2}} + D^-_{i+\frac{1}{2}} \leq 1$ is TVD.

This is often used as a definition for TVD even though it is a stronger requirement. But, of course, it is usually easy to check.

Now remember from theorem 3.4 the L^1-*contractivity* property,

$$\|u(\cdot,t_2)-v(\cdot,t_2)\|_{L^1(\mathbb{R})} \le \|u(\cdot,t_1)-v(\cdot,t_1)\|_{L^1(\mathbb{R})}$$

for any two entropy solutions of the same scalar conservation law (possibly with different initial data) and $t_1 < t_2$. The discrete analogon to this is given by

Definition 3.24. *A method is called* L^1-contractive *if for any two grid functions* u_Δ^n *and* v_Δ^n*, for which* $u_\Delta^n - v_\Delta^n$ *has compact support,* $u_\Delta^{n+1} = H(u_\Delta^n)$ *and* $v_\Delta^{n+1} = H(v_\Delta^n)$ *satisfy*

$$\|u_\Delta^{n+1} - v_\Delta^{n+1}\|_{L^1(\mathbb{R})} \le \|u_\Delta^n - v_\Delta^n\|_{L^1(\mathbb{R})}.$$

We know from [LeV06, theorem 15.4] that L^1-contractive methods are TVD. We also find that the Lax-Friedrichs method is L^1-contractive provided that the CFL condition $\left|\lambda f'(u)\right| \le 1$ is satisfied for all $\min_j(u_j^n, v_j^n) \le u \le \max_j(u_j^n, v_j^n)$. For the also L^1-contractive upwind scheme and monotonically increasing f, the CFL condition is $0 \le \lambda f'(u) \le 1$.[3]

Looking again at the weak entropy solution, we find another useful property. For any two sets of initial data u^0 and v^0 with $v^0(x) \ge u^0(x)$ for all x, the respective entropy solutions satisfy $v(x,t) \ge u(x,t)$ for all x, t.

Definition 3.25. *A numerical method is called* monotone *if* $u_\Delta^n \ge v_\Delta^n$ *implies* $u_\Delta^{n+1} \ge v_\Delta^{n+1}$ *for all* $n \ge 0$.

To prove that a method $U^{n+1} = H(U^n)$ is monotone, it suffices to check that

$$\frac{\partial}{\partial U_i^n} H_j(U^n) \ge 0.$$

This is easily done for the Lax-Friedrichs method, provided the CFL condition is satisfied.

Finally, it is known (cf. [LeV06, theorem 15.5]) that monotone methods are L^1-contractive.

In the linear case, monotonicity preserving methods are monotone, so the hierarchy of properties collapses to one single class.

Most importantly, we can now state a strong result on convergence:

Theorem 3.26. *Let* $\{u_\Delta\}_\Delta$ *be a sequence of numerical solutions computed from initial data* $u^0 \in BV(\mathbb{R})$ *with a consistent monotone or* L^1*-contractive method. Assume* $\Delta x,\ \Delta t \to 0$ *with* $\frac{\Delta t}{\Delta x} \le \lambda_0$ *for some* $\lambda_0 > 0$*. Then the sequence of discrete solutions converges in* L^1_{loc} *to the unique entropy solution.*

The proof can be found in [Kro97, theorem 2.3.19] where we also find the following error estimate in theorem 2.3.23:

Theorem 3.27. *Let* $u^0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$ *and* u *be the entropy solution. If* g *is a monotone numerical flux,* $\frac{\Delta t}{\Delta x} \le \lambda_0$ *for some constant* λ_0*, then for any* $\sqrt{\Delta t} \le t \le T$*, the estimate*

$$\|u(\cdot,t) - u_\Delta(\cdot,t)\|_{L^1(\mathbb{R})} \le \|u(\cdot,0) - u_\Delta(\cdot,0)\|_{L^1(\mathbb{R})} + ct\|u_\Delta^0\|_{BV(\mathbb{R})}\sqrt{\Delta t}$$

holds for the numerical solution u_Δ.

[3] Note that the conditions for nonlinear stability are the same as for linear stability. In particular, this means that for nonlinear problems, we cannot avoid the CFL condition by using implicit methods as we would with parabolic or ordinary differential equations.

Unfortunately, there is a great restriction on the class of monotone methods (cf. [LeV06, theorem 15.6]):

Theorem 3.28. *A monotone method is at most first order accurate in space and time.*

Now remember that we do not need a monotone scheme to avoid spurious oscillations. In fact, even TVD might be too restrictive in some applications. But even for monotonicity preserving schemes we are disappointed by *Godunov's order barrier theorem* (cf. [Wes01, theorem 9.2.2]):

Theorem 3.29. *Linear one-step second order accurate numerical schemes for the linear advection equation $u_t + a u_x = 0$ cannot be monotonicity preserving, unless $|a|\,\lambda \in \mathbb{N}$.*

This restriction is quite severe because we usually need $|\lambda a| \leq 1$ for stability. Also, it covers only the linear case. Often, Godunov's theorem is formulated in a slightly different way, namely that linear monotonicity preserving methods are at most first order accurate.

This is actually a simple corollary to theorem 3.28 since we know that monotonicity preserving schemes are monotone in the linear case.

3.2.4 Modified Equations

The proof of the unfruitful result of theorem 3.28 relies on the *modified equation* for the monotone method. By considering the local truncation error and Taylor series expansion, it can be shown that the numerical solution is actually a second order accurate approximation to the solution $v = v(x,t)$ to a *modified equation*

$$\partial_t v + \partial_x f(v) = \Delta x \, \partial_x (b(v) \partial_x v) \tag{3.24}$$

where b is a function of v that depends on the derivatives of the scheme's function H with respect to each argument (cf. [HHLK76]). If H is written in terms of a numerical flux $g(v,w)$, we can express b via

$$b(z) = \frac{1}{2}\big(\partial_1 g(z,z) - \partial_2 g(z,z) - \lambda f'(z)^2\big).$$

(3.24) is an advection-diffusion equation, somewhat similar to the viscosity approximation (3.13). Assuming H describes a monotone method, it can be shown that $b > 0$. The modified equation (3.24) is an $\mathcal{O}(\Delta t)$ perturbation of the original conservation law, so the

With this knowledge, we have a better understanding of the monotone Lax-Friedrichs method, which can be written as

$$\begin{aligned} u_j^{n+1} &= \frac{1}{2}\big(u_{j+1}^n + u_{j-1}^n\big) - \frac{\lambda}{2}\Big(f\big(u_{j+1}^n\big) - f\big(u_{j-1}^n\big)\Big) \\ &= u_j^n - \frac{\lambda}{2}\Big(f\big(u_{j+1}^n\big) - f\big(u_{j-1}^n\big)\Big) + \frac{1}{2}\big(u_{j+1}^n - 2u_j^n + u_{j-1}^n\big). \end{aligned}$$

This last term can be interpreted as artificial viscosity, which establishes the tie to the mentioned advection-diffusion equation. We obtain

$$b(z) = \frac{1}{2\lambda}(1 + \lambda^2 f'(z)^2).$$

For simplicity, let us consider the linear case $f(u) = au$, $a \in \mathbb{R}$. The modified equation for the Lax-Friedrichs method then reads

$$u_t + a u_x = \frac{\Delta x}{2\lambda}\big(1 - (\lambda a)^2\big)\, u_{xx}.$$

For the upwind scheme and $a \geq 0$, the modified equation is

$$u_t + a u_x = \frac{a}{2} \Delta x (1 - \lambda a) u_{xx}.$$

Take, for example, $a = 1$ and $\lambda = \frac{1}{2}$. Then the diffusion coefficient for the Lax-Friedrichs scheme is $\frac{3}{4}\Delta x$, but $\frac{1}{4}\Delta x$ for the upwind scheme, so the numerical diffusion of the Lax-Friedrichs scheme is much stronger than that of the upwind method. This can bee seen in figure 3.8 where the smearing of both the continuous and discontinuous solution is a lot weaker than in figure 3.7. In both cases, the initial data was advected by the length of the domain $\Omega = (0, 1)$ assuming periodic boundary conditions. So the exact solution is just the initial data. As such, we chose a Gaussian and a square wave to compare a smooth and a discontinuous case.

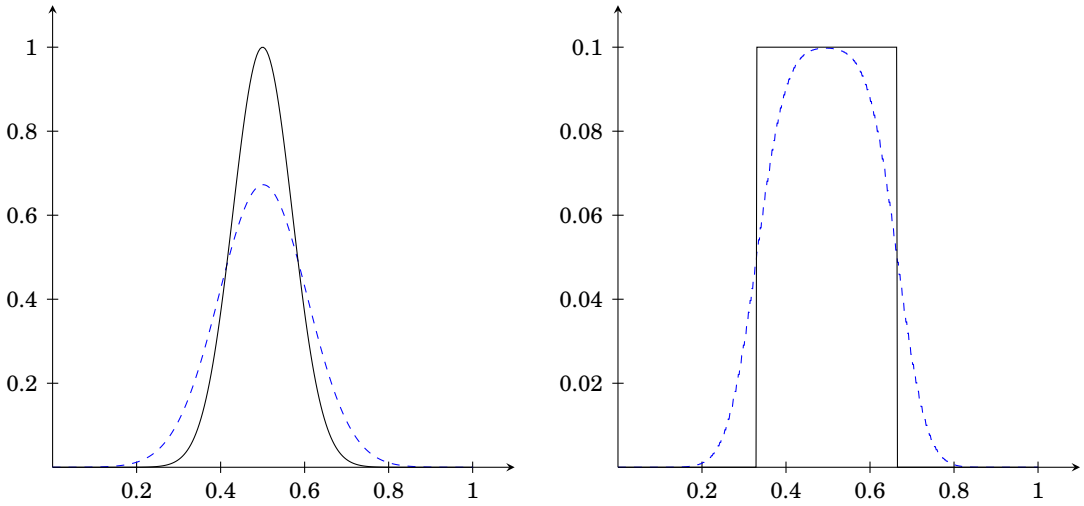

Figure 3.7: *Lax-Friedrichs method (blue) and exact solution (black)*

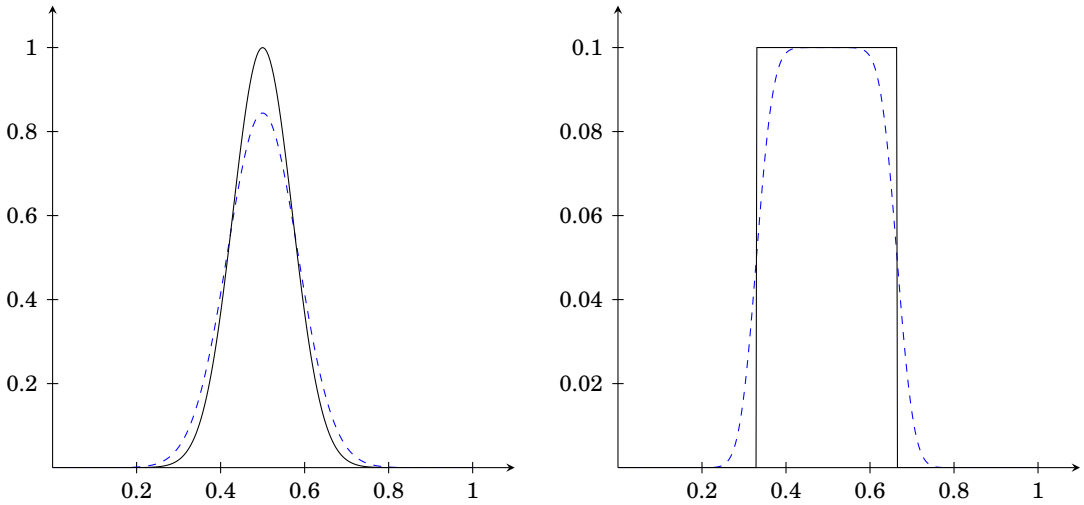

Figure 3.8: *Upwind method (blue) and exact solution (black)*

We also see the relation to stability here: a negative diffusion coefficient yields an unbounded solution. So for stability, the non-negativity condition for the diffusion coef-

ficient yields the stability conditions we saw earlier for these methods. It also confirms the instability of the FTCS method, whose modified equation reads

$$u_t + a u_x = -a^2 \frac{\Delta t}{2} u_{xx}$$

with a negative diffusion coefficient.

Modified equations are not only a tool to evaluate the schemes we already know, but they are also a starting point to create new methods. To get rid of the diffusion that the Lax-Friedrichs and the upwind method produce, the modified equations can be used to find a method that does not have any numerical diffusion. This is how the *Lax-Wendroff method*

$$u_j^{n+1} = u_j^n - \frac{\lambda a}{2}\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{\lambda^2 a^2}{2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right)$$

has been derived for the advection equation. The numerical flux in the general case is

$$F_{j+\frac{1}{2}}^{LW} = \frac{1}{2}\left(f(u_j^n) + f(u_{j+1}^n)\right) - \frac{\lambda}{2}\left|\frac{f(u_{j+1}^n) - f(u_j^n)}{u_{j+1}^n - u_j^n}\right|^2 (u_{j+1}^n - u_j^n).$$

Lax-Wendroff is second order accurate for the advection equation (and thus not monotone) and third order accurate on

$$u_t + a u_x = \frac{a\Delta x^2}{6}\left(\lambda^2 a^2 - 1\right) u_{xxx} =: \mu u_{xxx}, \tag{3.25}$$

which is a dispersive equation (cf. [LeV06, section 11.1]). The meaning of this becomes clearer when we look at the Fourier series solution to this equation,

$$u(x,t) = \int_{-\infty}^{\infty} \hat{u}(k,t) e^{ikx}\, dk.$$

By linearity, it suffices to consider solutions of the form

$$u(x,t) = e^{ikx - i\omega t},$$

i.e., one wave number at a time. Substituting this into (3.25) yields the dispersion relation

$$\omega = ak + \mu k^3.$$

The phase velocity is

$$c_p = \frac{\omega}{k} = a + \mu k^2,$$

which is only close to the original propagation speed a for sufficiently small wave numbers k. The group velocity

$$c_g = \frac{d\omega}{dk} = a + 3\mu k^2$$

looks even worse. What happens is that components with different wave numbers propagate at different speeds, i.e., they disperse. For discontinuous initial data (like a step function), the Fourier coefficients only decay like $\frac{1}{k}$ as $|k| \to 0$ instead of $\frac{1}{k^m}$ in case of C^m functions. Wave number k will be predominantly visible at $x = c_g t$ at time t, so the most oscillatory components are farthest away from the correct location $x = at$. For $|\lambda a| < 1$ and $a > 0$ we have $c_g < a$ for all k, so all components travel too slowly and there are oscillations behind the discontinuity.

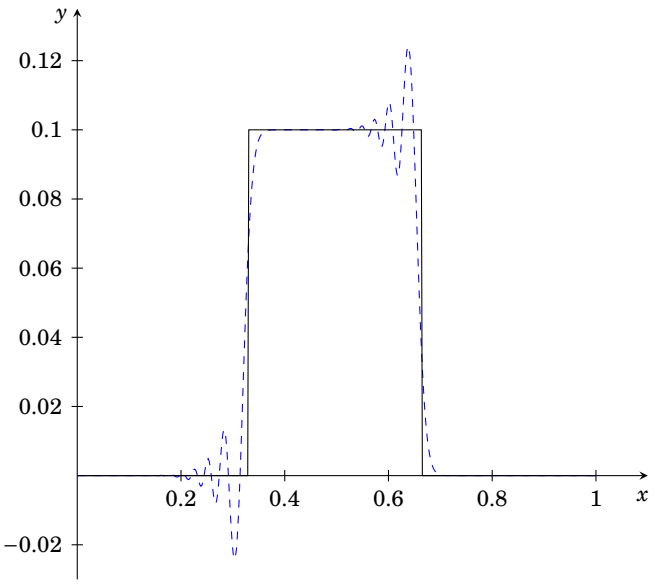

Figure 3.9: *Lax-Wendroff method (blue) and exact solution (black)*

Figure 3.9 shows the solution to the same problem as above, but instead of the diffusion of the upwind and Lax-Friedrichs method, we now observe the predicted oscillations.

This effect is not restricted to the Lax-Wendroff method, but it is the widely known *Gibb's phenomenon* (cf. [Car30]). The peak will rise to a certain level and then it will stay.

Of course, oscillations like these show that the method cannot be monotonicity preserving. Therefore, it also is neither TVD nor BV-stable, but at least for the linear equation, there is a stability result (cf. [Kro97, lemma 2.5.2]:

Theorem 3.30. *In the linear case $f(z) = az$, the Lax-Wendroff method is stable in the sense*

$$\|u_\Delta^n\|_{L^2(\mathbb{R})} \leq C(T) e^{\beta \frac{t_n}{T}} \|u_\Delta^0\|_{L^2(\mathbb{R})}$$

provided that $|a|\lambda < 1$, for some $\beta > 0$ and $C(T) > 0$.

In contrast to the Lax-Wendroff method, oscillations in the FTCS scheme will grow unboundedly. For sufficiently small time steps, however, it produces a numerical solution somewhat similar to the Lax-Wendroff approximation. Better results can be obtained by its fourth order counterpart,

$$F^{H4}_{j+\frac{1}{2}} = \frac{7}{12}\left(f(u^n_{j+1}) + f(u^n_j)\right) - \frac{1}{12}\left(f(u^n_{j+2}) + f(u^n_{j-1})\right),$$

which is used by Zalesak in [Zal79]. Figure 3.10 shows the result to the above test problem with $\lambda = \frac{1}{20}$, so even though these central schemes are easy to implement, they are highly ineffective.

Having understood the behavior of the different methods, the question arises, how we can modify existing methods like Lax-Wendroff to better suit our needs. Godunov's theorem tells us that a linear correction term like the artificial viscosity in Lax-Friedrichs will not work here. That is why we will consider nonlinear corrections in chapter 5.

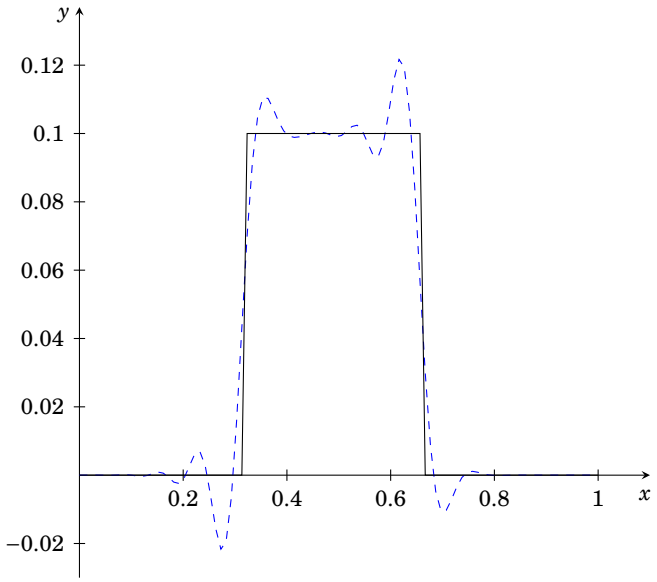

Figure 3.10: *Fourth order method (blue) and exact solution (black)*

4 Physical Models and Methods

In chapter 2, we have already discussed the well-known Maxwell's equations and their numerical solution. Some other techniques in computational physics shall be presented in the following. The standard approach to (laser-)plasma simulations are so-called *particle-in-cell (PIC) codes*. The first such simulations were published in [Bun59] and [Daw62]. A classical monograph on the topic is [BL91] while [Puk99] describes a widely known implementation. Instead of trying to simulate each single particle (like electrons and ions), they are grouped into *macro-particles*. These represent a large number of particles, which is still small, though, compared to the total number of particles. This greatly reduces the computational cost compared to a full kinetic description of a plasma. But even with this trick, those simulations are very expensive in terms of resources and time. Also, a major problem with PIC codes is the amount of noise they produce (cf. [Oku72]).

We are not interested in effects at the boundaries of the computational domain, so we will assume periodic boundary conditions. Mathematically, this means we consider all our equations on some torus $\mathbb{T} \subset \mathbb{R}^{d+1}$ instead of the actual domain $\Omega \subset \mathbb{R}^d$ for $d \in \{1,2,3\}$.

We will follow standard notation in physics: The dyadic product of two vectors $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^n$ is the $m \times n$ matrix

$$\boldsymbol{a}\boldsymbol{b} \equiv \boldsymbol{a} \otimes \boldsymbol{b} \equiv \boldsymbol{a}\boldsymbol{b}^T = \left(a_i b_j\right)_{\substack{i=1,\dots,m \\ j=1,\dots,n}}$$

while $\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^T \boldsymbol{b}$ denotes the scalar product. The divergence of a tensor $\boldsymbol{a}\boldsymbol{b}$ is

$$\nabla \cdot (\boldsymbol{a}\boldsymbol{b}) = (\nabla \cdot \boldsymbol{a})\boldsymbol{b} + (\boldsymbol{a} \cdot \nabla)\boldsymbol{b}.$$

4.1 Plasma Formulation via the Vlasov Equation

Our goal is the simulation of laser-plasma interaction. To understand the governing equations, we have to understand what a plasma is in the first place. Most people are familiar with the three states of aggregation: solid, liquid and gaseous. A very nice every-day example is water: it is liquid at room temperature, but it can freeze to solid ice and boil to steam. This can be done to any material, but the temperatures, at which the transitions occur, are very different, of course. Plasma is sometimes called the fourth state of aggregation. It forms when a gas is heated so much that the atoms ionize. So simply speaking, a plasma is a gas, in which the atoms are split into electrons and ions and thus electric fields induce electric currents within the plasma. The most common examples of plasmas in our daily lives are the sun, fluorescent lamps and plasma televisions.

The derivation of the fluid description of a plasma can be found, e.g. in [Kru03]. Consider a collisionless plasma and the single particle distribution function $f_\alpha(\boldsymbol{x}, \boldsymbol{v}, t)$. This characterizes the location of the particles of species α in phase space as a function of time. We will mainly consider the electrons ($\alpha = e$) and sometimes also the ions ($\alpha = i$) inside a plasma. The *Vlasov equation* for particle species α reads

$$\frac{\partial f_\alpha}{\partial t} + \boldsymbol{v} \cdot \nabla f_\alpha + \frac{q_\alpha}{m_\alpha}\left(\boldsymbol{E} + \frac{\boldsymbol{v}}{c} \times \boldsymbol{B}\right) \cdot \nabla_{\boldsymbol{v}} f_\alpha = 0,$$

where $\nabla_{\boldsymbol{v}} = (\partial_{v_x}, \partial_{v_y}, \partial_{v_z})^T$. This equation tells us that the phase space density is conserved following a dynamical trajectory because $f_\alpha(\boldsymbol{x}(t), \boldsymbol{v}(t), t)$ is constant.

Together with Maxwell's equations, the Vlasov equation is a complete description of a collisionless plasma, the *Vlasov-Maxwell system*.[4] Under certain assumptions, the Cauchy problem to the Vlasov-Maxwell system has a unique C^1-solution for all times $t \geq 0$ (cf. [Gla96]). The numerical simulation, however, turns out to be quite cumbersome (see e.g. [CK76] or [SRBG99]) because of the additional derivatives with respect to $\boldsymbol{v}$. The general single particle phase-space is six-dimensional. If we assume all quantities to vary spatially only along one dimension, the phase-space is reduced to four dimensions. Making use of a symmetry of the 1D Vlasov-Maxwell system (conservation of the canonical momentum) allows a further reduction down to two dimensions. Still, the solution of the Vlasov equation to find $f(x, v_x, t)$ can be far from trivial.

Therefore, a different set of equations, which is easier to deal with, is derived by taking different velocity moments of the Vlasov equation (cf. [Kru03]). Particle density n and mean velocity $\bar{\boldsymbol{v}}$ are determined by averaging the moments of the phase space distribution function over velocities

$$n = \int_{\mathbb{R}^3} f(\boldsymbol{x}, \boldsymbol{v}, t)\, d\boldsymbol{v}$$
$$n\bar{\boldsymbol{v}} = \int_{\mathbb{R}^3} \boldsymbol{v} f(\boldsymbol{x}, \boldsymbol{v}, t)\, d\boldsymbol{v},$$

where we dropped the subscript α for readability. Now we average the Vlasov equation over velocity

$$\int \left(\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla f + \frac{q}{m} \left(\boldsymbol{E} + \frac{\boldsymbol{v}}{c} \times \boldsymbol{B} \right) \cdot \nabla_{\boldsymbol{v}} f \right) d\boldsymbol{v} = 0$$

and after some calculations obtain the continuity equation for the particle density

$$\frac{\partial n}{\partial t} + \nabla \cdot (n\bar{\boldsymbol{v}}) = 0. \tag{4.1}$$

The next moment of the Vlasov equation

$$\int \boldsymbol{v} \left(\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla f + \frac{q}{m} \left(\boldsymbol{E} + \frac{\boldsymbol{v}}{c} \times \boldsymbol{B} \right) \cdot \nabla_{\boldsymbol{v}} f \right) d\boldsymbol{v} = 0$$

yields the equation of motion for the charged fluid

$$\frac{\partial}{\partial t}(n\bar{\boldsymbol{v}}) + \nabla \cdot (n\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}) = \frac{qn}{m} \left(\boldsymbol{E} + \frac{\bar{\boldsymbol{v}}}{c} \times \boldsymbol{B} \right). \tag{4.2}$$

For a detailed derivation, see [Kru03]. Note that (4.2) only holds for so-called cold plasmas, which means that temperature is neglected in the derivation of the model equations. Although this sounds contradictory, the assumption is justified because the effect of pressure is negligible in the cases we are interested in.

So now we have two fluid equations (4.1) and (4.2) for the description of density and momentum of each particle species in the plasma. Combining these with Maxwell's equations

$$\frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B} - \frac{4\pi}{c}\boldsymbol{j}$$
$$\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E}$$
$$\nabla \cdot \boldsymbol{E} = 4\pi\rho$$
$$\nabla \cdot \boldsymbol{B} = 0$$

[4] If collisions between plasma particles shall be modeled, there is an additional term on the right hand side of the Vlasov equation. Comparisons to the influence of the external fields and time scales considered show, however, that neglecting collisions does not have much impact on the simulation (cf. [Taj89]).

where $\rho = \sum_\alpha n_\alpha q_\alpha$ and $\boldsymbol{j} = \sum_\alpha n_\alpha q_\alpha \bar{\boldsymbol{v}}_\alpha$, we have a full description of a collisionless plasma. If we consider a mobile species, the electrons, and an immobile, neutralizing species, the ions, then $\rho = qn$ and $\boldsymbol{j} = \rho\bar{\boldsymbol{v}}$. In our case, this mobile species will be the electrons. Since they are much lighter than ions (by a factor of 1836 in case of hydrogen plasma, which contains the lightest possible ions), which leads to higher inertia and reaction time, ion motion is often neglected to save computational cost.

4.2 Boris Push

We already know how to approximate the solution to Maxwell's equations and to conservation laws without sources. To treat particle motion due to the Lorentz force

$$\boldsymbol{F}_L = q\,(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$$

numerically, we consider the electric and the magnetic term separately. The electric update is straightforward as a simple forward Euler step is used. To treat the vector product as accurately as possible, we analyze it from a geometric point of view. We will see that the product is a rotation,[5] so we can find its angle and use that knowledge to correctly follow the path of this rotation. This method is called *Boris push* after its inventor Jay P. Boris (cf. [Bor70], [BL91]).

To derive the Boris push, consider the particle equations of motion under the Lorentz force,

$$\begin{aligned} m\frac{d\boldsymbol{v}}{dt} &= q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \\ \frac{d\boldsymbol{x}}{dt} &= \boldsymbol{v}. \end{aligned}$$

Centered finite differences lead to

$$\frac{\boldsymbol{v}^{n+\frac{1}{2}} - \boldsymbol{v}^{n-\frac{1}{2}}}{\Delta t} = \frac{q}{m}\left(\boldsymbol{E}^n + \frac{\boldsymbol{v}^{n+\frac{1}{2}} + \boldsymbol{v}^{n-\frac{1}{2}}}{2} \times \boldsymbol{B}^n\right) \tag{4.3a}$$

$$\frac{\boldsymbol{x}^{n+1} - \boldsymbol{x}^n}{\Delta t} = \boldsymbol{v}^{n+\frac{1}{2}}. \tag{4.3b}$$

The averaging of the velocity on the right hand side of (4.3a) is necessary since using the old value $\boldsymbol{v}^{n-\frac{1}{2}}$ results in a wrong particle motion (cf. [BL91]). Instead, we need the velocity at the same time t_n as $\boldsymbol{B}$ and $\boldsymbol{E}$, which is approximated by the mean value of $\boldsymbol{v}^{n-\frac{1}{2}}$ and $\boldsymbol{v}^{n+\frac{1}{2}}$. Solving this implicit set of equations as they are stated above does not respect the rotational movement. A finite difference approach yields a straight instead of a circular path, so a different approach is necessary.

The electric and magnetic forces are to be separated. To achieve this, substitute

$$\boldsymbol{v}^{n-\frac{1}{2}} = \boldsymbol{v}^- - \frac{q\boldsymbol{E}^n}{m}\frac{\Delta t}{2} \tag{4.4}$$

$$\boldsymbol{v}^{n+\frac{1}{2}} = \boldsymbol{v}^+ + \frac{q\boldsymbol{E}^n}{m}\frac{\Delta t}{2} \tag{4.5}$$

into (4.3a), which cancels out the electric field completely and leaves

$$\frac{\boldsymbol{v}^+ - \boldsymbol{v}^-}{\Delta t} = \frac{q}{2m}(\boldsymbol{v}^+ + \boldsymbol{v}^-) \times \boldsymbol{B}^n. \tag{4.6}$$

[5] This is a well-known fact in physics. The rotational nature of the Lorentz force is widely used, for example in mass spectrometry or television tubes.

So now we get from $\boldsymbol{v}^{n-\frac{1}{2}}$ to $\boldsymbol{v}^-$ via (4.4), then obtain $\boldsymbol{v}^+$ from (4.6) and finally add the remaining half of the electric push in (4.5) to reach $\boldsymbol{v}^{n+\frac{1}{2}}$. The electric updates are unproblematic, but to do the magnetic part right, we have to take a look at the problem from a geometric point of view. Equation (4.6) is a pure rotation[6], so this knowledge should be made use of. Figure 4.1 shows that the angle of rotation satisfies

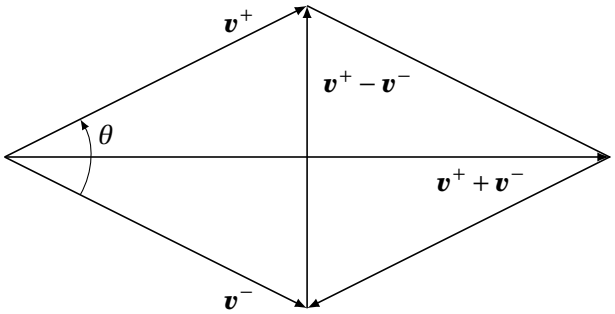

Figure 4.1: *This diagram depicts the plane perpendicular to* $\boldsymbol{B}$ *and shows the rotation in* (4.6)*.*

$$\left|\tan\frac{\theta}{2}\right| = \frac{\|\boldsymbol{v}^+ - \boldsymbol{v}^-\|}{\|\boldsymbol{v}^+ + \boldsymbol{v}^-\|} = \frac{q\|\boldsymbol{B}^n\|}{m}\frac{\Delta t}{2}.$$

To perform the $\boldsymbol{v} \times \boldsymbol{B}$ rotation properly, it is split into two steps. First, define a vector $\boldsymbol{v}'$

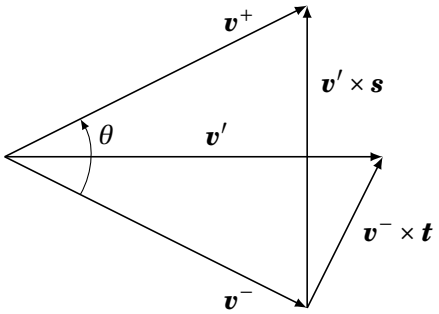

Figure 4.2: *The rotation from* $\boldsymbol{v}^-$ *to* $\boldsymbol{v}^+$*, again in the plane perpendicular to* $\boldsymbol{B}$*.*

that is orthogonal to $\boldsymbol{v}^+ - \boldsymbol{v}^-$ and $\boldsymbol{B}$,

$$\boldsymbol{v}' = \boldsymbol{v}^- + \boldsymbol{v}^- \times \boldsymbol{t}$$

for some $\boldsymbol{t}$ that is parallel to $\boldsymbol{B}$. The angle between $\boldsymbol{v}^-$ and $\boldsymbol{v}'$ is just $\frac{\theta}{2}$, so $\boldsymbol{t} = \frac{q\Delta t}{2m}\boldsymbol{B}$. Finally, $\boldsymbol{v}^+ - \boldsymbol{v}^-$ is parallel to $\boldsymbol{v}' \times \boldsymbol{B}$, so

$$\boldsymbol{v}^+ = \boldsymbol{v}^- + \boldsymbol{v}' \times \boldsymbol{s}$$

with some $\boldsymbol{s}$ that is parallel to $\boldsymbol{B}$. The requirement $\|\boldsymbol{v}^+\| = \|\boldsymbol{v}^-\|$ yields

$$\boldsymbol{s} = \frac{2\boldsymbol{t}}{1 + \|\boldsymbol{t}\|^2}$$

The Boris push is a widely used tool in computational physics as it reproduces the correct rotational particle movement, while finite difference approaches literally throw the particles off their path. The Boris push makes use of the knowledge about rotational movement and therefore keeps the particles on a circular path (cf. [BL91]).

[6] The scalar product with $\boldsymbol{v}^+ + \boldsymbol{v}^-$ yields $\|\boldsymbol{v}^+\| = \|\boldsymbol{v}^-\|$

4.3 Units and Dimensionless Equations

In all previous sections, we have stated the equations in cgs units, which is one of the most widely-used unit systems in plasma physics. In many applications it is common and useful to scale equations to some convenient unit system. This can be another international unit system like mks or SI, but it can also be a custom unit system, which eliminates units altogether by measuring everything in problem specific quantities.

cgs Units

The cgs system is one of the most common unit systems in plasma physics. It is based on centimeters, grams and seconds as base units — hence the name cgs. Another international standard is SI, which is the abbreviation for the French term *système international d'unités*. It is the modern form of the metric system and is based on meters, kilograms and seconds like mks.

Our goal is the simulation of laser-plasma interaction. A laser pulse is an electromagnetic wave. Those waves are described in general by Maxwell's equations. Recall from chapter 2 the formulation for homogeneous media

$$\nabla \cdot \boldsymbol{E} = 4\pi\rho \tag{4.7a}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{4.7b}$$

$$\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E} \tag{4.7c}$$

$$\frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B} - \frac{4\pi}{c}\boldsymbol{j} \tag{4.7d}$$

where c is the speed of light and

$$\boldsymbol{j} = \sum_{\alpha} q_\alpha n_\alpha \boldsymbol{v}_\alpha$$

the current density with charges q_α for particle species α. We consider only electrons ($\alpha = e$) for now, so $q_e = -e$ and $\boldsymbol{j} = -en\boldsymbol{v}$. $\boldsymbol{v}$ is the particle velocity and $\rho = -ne$ is the electric charge density with electric charge e and the particle-number density n.

The plasma description consists of an equation for the density and one for the momentum density as discussed in section 4.1:

$$\frac{\partial n}{\partial t} + \nabla \cdot (n\boldsymbol{v}) = 0 \tag{4.8a}$$

$$\frac{\partial (n\boldsymbol{v})}{\partial t} + \nabla \cdot (n\boldsymbol{v}\boldsymbol{v}) = -\frac{en}{m}\left(\boldsymbol{E} + \frac{\boldsymbol{v}}{c} \times \boldsymbol{B}\right) \tag{4.8b}$$

where m is the mass of the particle.

Depending on the application, it makes sense to scale the above quantities by some physical quantities. We say we *make them dimensionless*.

Laser Units

In the laser (-plasma) physics community, laser-related quantities are often preferred for the conversions, such as the wavelength λ_0 or frequency ω_0 of the laser. They are related via $\lambda_0 = 2\pi\frac{c}{\omega_0}$. The factor 2π can be included into the conversions or not, depending on one's preferences.

The theoretical physicists we are collaborating with choose to measure time in $T_0 = \frac{\omega_0}{2\pi}$ and space in λ_0, so the conversions to be made are

$$\begin{aligned} t_L &= \omega_0 t \\ \boldsymbol{x}_L &= \frac{1}{\lambda_0}\boldsymbol{x} \\ n_L &= \frac{n}{n_0} \\ \boldsymbol{v}_L &= \frac{\boldsymbol{v}}{c} \\ \boldsymbol{j}_L &= \frac{\boldsymbol{j}}{e n_0 c} \\ \boldsymbol{E}_L &= \frac{e}{\omega_{pe} m c}\boldsymbol{E} \\ \boldsymbol{B}_L &= \frac{e}{\omega_{pe} m c}\boldsymbol{B} \end{aligned}$$

That way, we get for (4.7a)

$$\begin{aligned} & \nabla \cdot \boldsymbol{E} = 4\pi\rho \\ \iff & \frac{\omega_0}{2\pi c}\nabla_L \cdot \frac{\omega_0 m c}{e}\boldsymbol{E}_L = -4\pi e n_0 n_L \\ \iff & \nabla_L \cdot \boldsymbol{E}_L = -\frac{2\pi n_L}{\omega_0^2} 4\pi n_0 \frac{e^2}{m} = -2\pi\frac{\omega_{pe}^2}{\omega_0^2} n_L = -2\pi\tilde{n} n_L \end{aligned}$$

with $\omega_{pe} = \sqrt{4\pi n_0 e^2/m}$. Then for (4.7c) we have

$$\begin{aligned} & \frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E} \\ \iff & \frac{1}{c}\frac{\omega_0}{2\pi}\frac{\partial}{\partial t_L}\left(\frac{\omega_0 m c}{e}\boldsymbol{B}_L\right) = -\frac{\omega_0}{2\pi c}\nabla_L \times \left(\frac{\omega_0 m c}{e}\boldsymbol{E}_L\right) \\ \iff & \frac{\partial}{\partial t_L}\boldsymbol{B}_L = -\nabla_L \times \boldsymbol{E}_L \end{aligned}$$

and for (4.7d)

$$\begin{aligned} & \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B} - \frac{4\pi}{c}\boldsymbol{j} \\ \iff & \frac{1}{c}\frac{\omega_0}{2\pi}\frac{\partial}{\partial t_L}\left(\frac{\omega_0 m c}{e}\boldsymbol{E}_L\right) = \frac{\omega_0}{2\pi c}\nabla_L \times \left(\frac{\omega_0 m c}{e}\boldsymbol{B}_L\right) - \frac{4\pi}{c} e n_0 c \boldsymbol{j}_L \\ \iff & \frac{\partial}{\partial t_L}\boldsymbol{E}_L = \nabla_L \times \boldsymbol{B}_L - 4\pi n_0 \frac{e^2}{m}\frac{2\pi}{\omega_0^2}\boldsymbol{j}_L = \nabla_L \times \boldsymbol{B}_L - 2\pi\frac{\omega_{pe}^2}{\omega_0^2}\boldsymbol{j}_L. \end{aligned}$$

The fluid equations (4.8) become

$$\begin{aligned} & \frac{\partial n}{\partial t} + \nabla \cdot (n\boldsymbol{v}) = 0 \\ \iff & \frac{\omega_0}{2\pi}\frac{\partial}{\partial t_L}(n_0 n_L) + \frac{\omega_0}{2\pi c}\frac{\partial}{\partial x_L}(n_0 n_L \boldsymbol{v}_L c) = 0 \\ \iff & \frac{\partial}{\partial t_L} n_L + \frac{\partial}{\partial x_L}(n_L \boldsymbol{v}_L) = 0 \end{aligned}$$

and

$$\frac{\partial(n\boldsymbol{v})}{\partial t} + \nabla\cdot(n\boldsymbol{v}\boldsymbol{v}) = -\frac{en}{m}(\boldsymbol{E} + \frac{\boldsymbol{v}}{c}\times\boldsymbol{B})$$
$$\iff \frac{\omega_0}{2\pi}\frac{\partial}{\partial t_L}(n_0 n_L \boldsymbol{v}_L c) + \frac{\omega_0}{2\pi c}\frac{\partial}{\partial x_L}(n_0 n_L \boldsymbol{v}_L \boldsymbol{v}_L c^2) = -\frac{n_0 n_L e}{m}\frac{\omega_0 mc}{e}(\boldsymbol{E}_L + \boldsymbol{v}_L\times\boldsymbol{B}_L)$$
$$\iff \frac{\partial(n_L\boldsymbol{v}_L)}{\partial t_L} + \nabla_L\cdot(n_L\boldsymbol{v}_L\boldsymbol{v}_L) = -2\pi n_L(\boldsymbol{E}_L + \boldsymbol{v}_L\times\boldsymbol{B}_L)$$

So dropping the subscripts L, we end up with

$$\nabla\cdot\boldsymbol{E} = -2\pi\frac{\omega_{pe}^2}{\omega_0^2}n \tag{4.9a}$$
$$\nabla\cdot\boldsymbol{B} = 0 \tag{4.9b}$$
$$\frac{\partial\boldsymbol{B}}{\partial t} = -\nabla\times\boldsymbol{E} \tag{4.9c}$$
$$\frac{\partial\boldsymbol{E}}{\partial t} = \nabla\times\boldsymbol{B} - 2\pi\frac{\omega_{pe}^2}{\omega_0^2}\boldsymbol{j} \tag{4.9d}$$
$$\frac{\partial n}{\partial t} + \nabla\cdot(n\boldsymbol{v}) = 0 \tag{4.9e}$$
$$\frac{\partial(n\boldsymbol{v})}{\partial t} + \nabla\cdot(n\boldsymbol{v}\boldsymbol{v}) = -2\pi n(\boldsymbol{E} + \boldsymbol{v}\times\boldsymbol{B}). \tag{4.9f}$$

Plasma Units

If our main focus is on the plasma, a possible choice is to measure everything in terms of plasma related quantities. Let ω_{pe} be the plasma frequency. For plasma units, we use $t_P = \omega_{pe}t$ and $\boldsymbol{x}_P = \frac{\omega_{pe}}{c}\boldsymbol{x}$ with $\omega_{pe} = \sqrt{4\pi n_0 e^2/m}$. Using all of the above to transform our set of equations in (4.7) and (4.8), we obtain — dropping again the subscripts P —

$$\nabla\cdot\boldsymbol{E} = -n \tag{4.10a}$$
$$\nabla\cdot\boldsymbol{B} = 0 \tag{4.10b}$$
$$\frac{\partial\boldsymbol{B}}{\partial t} = -\nabla\times\boldsymbol{E} \tag{4.10c}$$
$$\frac{\partial\boldsymbol{E}}{\partial t} = \nabla\times\boldsymbol{B} - \boldsymbol{j} \tag{4.10d}$$
$$\frac{\partial n}{\partial t} + \nabla\cdot(n\boldsymbol{v}) = 0 \tag{4.10e}$$
$$\frac{\partial(n\boldsymbol{v})}{\partial t} + \nabla\cdot(n\boldsymbol{v}\boldsymbol{v}) = -n(\boldsymbol{E} + \boldsymbol{v}\times\boldsymbol{B}). \tag{4.10f}$$

These are particularly nice to read and to work with because there are no factors left in front of any of the terms. That is why we will use plasma units for the description of equations from now on.

When we compare our numerical simulations in chapters 6 and 7 to those of established codes, we have to switch between laser and plasma units. The conversions are quite easy, once there is agreement on what systems to use and what the original equations looked like in cgs units.

4.4 Relativistic Models

Everything discussed so far in this section was non-relativistic. We are going to be considering laser-plasma interaction with large particle velocities. If the velocities of the

plasma particles are small compared to the speed of light, the non-relativistic models work fine, but for greater velocities, this model is no longer adequate. Refer to [Ger10, chapter 13] for a thorough introduction into relativistic physics. We have to substitute relativistic velocities

$$\boldsymbol{v} = \frac{\boldsymbol{p}}{m\gamma}$$

where $\boldsymbol{p}$ is the relativistic momentum and

$$\gamma = \gamma(\boldsymbol{p}) = \frac{1}{\sqrt{1 - \left(\frac{\|\boldsymbol{v}\|}{c}\right)^2}} = \sqrt{1 + \left(\frac{\|\boldsymbol{p}\|}{mc}\right)^2}$$

is the so-called *relativistic γ-factor* or *Lorentz-factor*, m is the *rest mass* of the particle and c the speed of light. Relativistic effects can be neglected for small velocities: If the velocity is small compared to the speed of light, $\|v\| \ll c$, then $\gamma \approx 1$. The speed of light is very high, $c \approx 2.9979 \cdot 10^8 \frac{m}{s}$, so for our daily life, the approximation that γ be one is often justified. We are, however, interested in high velocities and the relativistic effects in the simulation of laser-plasma interaction, so we need to adjust our equations accordingly.

For Maxwell's equations, the current density becomes

$$\boldsymbol{j} = \sum_\alpha n_\alpha q_\alpha \boldsymbol{v}_\alpha = \sum_\alpha n_\alpha q_\alpha \frac{\boldsymbol{p}_\alpha}{m_\alpha \gamma_\alpha}$$

where $\gamma_\alpha = \gamma_\alpha(\boldsymbol{p}_\alpha)$. The propagation speed of the electromagnetic waves is already the speed of light c.

For the derivation of the relativistic continuity and momentum density equation, we have to consider the relativistic Vlasov equation (cf. [BGB+99])

$$\frac{\partial f}{\partial t} + \frac{\boldsymbol{p}}{m\gamma} \cdot \nabla f + \frac{q}{m}\left(\boldsymbol{E} + \frac{1}{c}\frac{\boldsymbol{p}}{m\gamma} \times \boldsymbol{B}\right) \cdot \nabla_{\boldsymbol{p}} f = 0$$

for each particle species to obtain the relativistic fluid equations

$$\frac{\partial n}{\partial t} + \nabla \cdot (n\boldsymbol{v}) = 0$$

and

$$\frac{\partial}{\partial t}(n\boldsymbol{p}) + \nabla \cdot (n\boldsymbol{p}\boldsymbol{v}) = \frac{nq}{m}\left(\boldsymbol{E} + \frac{\boldsymbol{v}}{c} \times \boldsymbol{B}\right).$$

Note that we will usually not use the particle density n in our simulations of a one-species plasma, but the electric charge density ρ. These only differ by the factor of the electric charge, q, which is ± 1 in the dimensionless units we are using. This eliminates a variable from our setting and avoids confusion with the superscript n that we use to indicate the time step in numerical schemes.

5 Flux-Corrected Transport Algorithms

In the numerical simulation of partial differential equations like the continuity equation, the values of the numerical solution can become negative due to the oscillation effect mentioned in section 3.2.4. Density and pressure are examples of quantities that are not allowed to be negative from the physical and/or mathematical point of view. Those negative values can cause severe instabilities, since the system may become non-hyperbolic and therefore ill-posed. Just cutting off the negative part is a bad idea since this changes the overall density and therefore violates the conservation of mass. We have seen that high order methods produce spurious oscillations and monotonicity preservation is limited to first order schemes — as long as we consider linear methods. The idea now is to combine high accuracy with the desirable properties of first order methods. Flux-limiter methods are the most widely-known approach to combine a high and a low order flux in a nonlinear way to elude Godunov's theorem (cf. [LeV11, chapter 6]). In a nutshell, we want to use the high order method where the solution is smooth and switch to a monotone method near sharp gradients.

5.1 Deriving the Algorithm

Probably the first class of methods that combines high accuracy with non-negativity was introduced at a conference in 1971 (cf. [Bor71]) and published in 1973 by Boris and Book in [BB73]. They called these schemes *Flux-Corrected Transport (FCT) algorithms*. The key idea is to correct the flux from one cell to another, such that the density will not become negative and use high order as much as possible. This section will describe the class of flux-corrected transport algorithms in detail.

We consider one-dimensional partial differential equations in conservation form

$$\partial_t u(x,t) + \partial_x f\big(u(x,t)\big) = 0 \tag{5.1}$$

on some bounded domain $\Omega \subset \mathbb{R}$ for $t \geq 0$, subject to appropriate initial and periodic boundary conditions, with functions u and f of independent variables x and t. The one-dimensional continuity equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v)}{\partial x} = 0 \tag{5.2}$$

has this form. All of the following can be generalized for systems — that is, vector valued functions $\boldsymbol{u}$ and $\boldsymbol{f}$ — without major changes. For ease of notation, however, we stick to scalar equations here.

We recall from section 3.2 the approximation in conservation form of equation (5.1),

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}\right) \tag{5.3}$$

where $u_j^n \approx u(x_j, t_n)$ with $t_n = n\Delta t$ and $x_j = j\Delta x$, $j = j_0, \ldots, j_N$, such that $\cup_{l=1}^{N} [x_{j_{l-1}}, x_{j_l}] = \overline{\Omega}$. The choice of the numerical flux $F_{j+\frac{1}{2}}$ defines the integration scheme. The goal is to combine the benefits from both low and high order methods into a new scheme. Note that we only consider uniform meshes for ease of notation. All of the following can easily be done for non-uniform grids, as well.

Boris and Book described their scheme as a predictor-corrector method: First use a low order predictor, like the upwind scheme. They called this the *transport stage*. Since

this introduces a lot of numerical diffusion, the corrector stage is called the *anti-diffusive stage* where as much high order correction as possible is added to the predictor without generating new extrema or introducing negativity. Their approach was revolutionary, yet highly heuristic, especially in the choice of the transport scheme, which they based on the Lax-Wendroff method. In fact, the whole scheme was very rigid as they built it around their particular choice of fluxes. Some years later, Zalesak generalized the concept of flux-corrected transport in [Zal79] for use with arbitrary schemes and also to multidimensional problems. The latter had been an open problem in the tight Boris and Book setting and thus a major drawback of the original approach.

The generalized flux-corrected transport algorithm of Zalesak (cf. [Zal79]) is stated in algorithm 5.1.

Algorithm 5.1 FCT algorithm after Zalesak

for $n = 0, \dots, M-1$ **do**

1. Compute low order (positivity preserving) fluxes $F^L_{j+\frac{1}{2}}$
2. Compute high order fluxes $F^H_{j+\frac{1}{2}}$
3. Define "anti-diffusive fluxes"
$$A_{j+\frac{1}{2}} = F^H_{j+\frac{1}{2}} - F^L_{j+\frac{1}{2}}$$
4. Compute low order update, the "transported and diffused" solution
$$u^{td}_j = u^n_j - \lambda \left(F^L_{j+\frac{1}{2}} - F^L_{j-\frac{1}{2}} \right)$$
5. Limit anti-diffusive fluxes such that no new extrema are created in step 6:
$$A^C_{j+\frac{1}{2}} = C_{j+\frac{1}{2}} A_{j+\frac{1}{2}}, \quad 0 \le C_{j+\frac{1}{2}} \le 1$$
6. Compute new solution with limited anti-diffusive fluxes:
$$u^{n+1}_j = u^{td}_j - \lambda \left(A^C_{j+\frac{1}{2}} - A^C_{j-\frac{1}{2}} \right)$$

end for

So we have again a predictor-corrector scheme, but with great flexibility as to the choice of methods. Inserting u^{td}_j from step three into the update in the last step, we see that the scheme has conservation form (5.3) with numerical flux

$$F^{FCT}_{j+\frac{1}{2}} = C_{j+\frac{1}{2}} F^H_{j+\frac{1}{2}} + (1 - C_{j+\frac{1}{2}}) F^L_{j+\frac{1}{2}},$$

which is a convex combination of the low and high order fluxes. The only unknown step above — and the most critical as well — is the fifth, which computes the limiting factors $C_{j+\frac{1}{2}}$. In the original paper [BB73], Boris and Book used the *minmod limiter*, which chooses the smallest argument in absolute value if they have the same sign and zero if

the signs are different,

$$\text{minmod}(a_1,\dots,a_s) = \begin{cases} \max\limits_j a_j & \text{if } a_j > 0 \text{ for all } j \\ \min\limits_j a_j & \text{if } a_j < 0 \text{ for all } j \\ 0 & \text{otherwise.} \end{cases}$$

For their so-called *SHASTA* scheme, which was designed with fixed high and low order fluxes, the corrective fluxes in flux-limiter notation (cf. [LeV06, chapter 16]) read

$$A^C_{j+\frac{1}{2}} = \frac{1}{8}\phi^n_{j+\frac{1}{2}}\left(u^{td}_{j+1} - u^{td}_j\right)$$

where

$$\phi^n_{j+\frac{1}{2}} = \text{minmod}\left(1, 8r^n_j, \frac{8}{r^n_{j+1}}\right)$$

with

$$r^n_j = \frac{u^{td}_j - u^{td}_{j-1}}{u^{td}_{j+1} - u^{td}_j}.$$

The factor 8 comes from their very own upwind scheme, which yields

$$\lambda A_{j+\frac{1}{2}} = \frac{1}{8}\left(u^{td}_{j+1} - u^{td}_j\right).$$

The minmod limiter is known to be TVD (cf. [LeV11, chapter 6]), so in particular, it is monotonicity and positivity preserving. Moreover, it has been shown in [IN79] that SHASTA is stable in the L^∞-sense and has a convergent subsequence to a weak solution in the L^1_{loc}-sense, while

$$\inf_{x\in\Omega} u^0(x) \le u^n_j \le \sup_{x\in\Omega} u^0(x)$$

holds for any j and $n \ge 0$.

For the generalized Zalesak scheme with arbitrary high and low order fluxes, the anti-diffusive fluxes after Boris and Book can be written as

$$A^C_{j+\frac{1}{2}} = \text{minmod}\left(u^{td}_j - u^{td}_{j-1}, A_{j+\frac{1}{2}}, u^{td}_{j+2} - u^{td}_{j+1}\right).$$

However, as Zalesak pointed out in [Zal79], the Boris and Book scheme is bound to cause a phenomenon he calls “clipping”: peaked profiles will be flattened due to the above limiter.[7] The problem lies in artificial diffusion and the fact that the Boris and Book method does not remember the values from the previous time step and hence cannot correctly reproduce existing extrema. Instead, the limiting process is applied even where it is not necessary. In other words: TVD can be too severe a restriction if positivity preservation is really all we need. Therefore, Zalesak proposes an alternative that also

[7] This holds for many other flux limiters since TVD methods degenerate to first order accuracy at extreme points (cf. [OC84]).

takes into account the previous time step for upper and lower bounds on the new solution u_j^{n+1}. He first defines

$$\begin{aligned} u_j^a &= \max\left(u_j^n, u_j^{td}\right) \\ u_j^{\max} &= \max\left(u_{j-1}^a, u_j^a, u_{j+1}^a\right) \\ u_j^b &= \min\left(u_j^n, u_j^{td}\right) \\ u_j^{\min} &= \max\left(u_{j-1}^b, u_j^b, u_{j+1}^b\right) \end{aligned}$$

to consider extrema of both the predictor approximation and the previous time step, and then the sum of all anti-diffusive fluxes *into* grid point j,

$$P_j^+ = \max\left(0, A_{j-\frac{1}{2}}\right) - \min\left(0, A_{j+\frac{1}{2}}\right)$$

and further

$$\begin{aligned} Q_j^+ &= \left(u_j^{\max} - u_j^{td}\right)\lambda^{-1} \\ R_j^+ &= \begin{cases} \min\left(1, Q_j^+/P_j^+\right) & \text{if } P_j^+ > 0, \\ 0 & \text{if } P_j^+ = 0. \end{cases} \end{aligned}$$

Since $u_j^{\max} \geq u_j^{td}$, none of the above quantities is negative, and R_j^+ is an upper bound for the fraction, by which the anti-diffusive fluxes *into* grid point j are multiplied to guarantee that no new maximum is created. The corresponding quantities regarding minima are defined analogously as

$$\begin{aligned} P_j^- &= \text{sum of all anti-diffusive fluxes } \textit{away from} \text{ grid point } j \\ &= \max\left(0, A_{j+\frac{1}{2}}\right) - \min\left(0, A_{j-\frac{1}{2}}\right) \\ Q_j^- &= \left(u_j^{td} - u_j^{\min}\right)\lambda^{-1} \\ R_j^- &= \begin{cases} \min\left(1, Q_j^-/P_j^-\right) & \text{if } P_j^- > 0 \\ 0 & \text{if } P_j^- = 0. \end{cases} \end{aligned}$$

Now $u_j^{\min} \leq u_j^{td}$, so R_j^- is a lower bound on the fraction, by which the anti-diffusive fluxes *away from* grid point j are multiplied to guarantee that no new minimum is created.

All fluxes are directed from one grid point into a neighboring one. So the limiting of the anti-diffusive fluxes has to prevent undershoot in the source and overshoot in the destination point. To accomplish both, the minimum is taken:

$$C_{j+\frac{1}{2}} = \begin{cases} \min\left(R_{j+1}^+, R_j^-\right) & \text{if } A_{j+\frac{1}{2}} \geq 0 \\ \min\left(R_j^+, R_{j+1}^-\right) & \text{if } A_{j+\frac{1}{2}} < 0. \end{cases} \tag{5.4}$$

It is claimed by Zalesak in [Zal79] that FCT is positivity preserving, but a detailed proof is not to be found anywhere in the literature. Hence we have to check the positivity preservation of the FCT scheme ourselves.

Lemma 5.1. *FCT is positivity preserving.*

Proof. Since the low order method is assumed to preserve positivity, we only have to consider the corrected fluxes $A^C_{j+\frac{1}{2}}$. The new approximation is

$$u_j^{n+1} = u_j^{td} + \lambda \left(C_{j-\frac{1}{2}} A_{j-\frac{1}{2}} - C_{j+\frac{1}{2}} A_{j+\frac{1}{2}} \right).$$

If both fluxes are directed out of cell j, i.e., when $A_{j-\frac{1}{2}} < 0$ and $A_{j+\frac{1}{2}} > 0$, there is the largest possibility to obtain a negative value for u_j^{n+1}. In case both fluxes have the same sign, it is still possible to have negative values if more flows out of the cell than into it. The calculations for those cases can, however, be reduced to the formerly mentioned worst case, so we will omit those considerations here.

It is clear that we cannot exceed $u_j^{\max}$ in this case. To see that we will also not fall below $u_j^{\min}$, we have to go through some lengthy calculations. According to the definition of C in (5.4), we have $C_{j+\frac{1}{2}} = \min(R^+_{j+1}, R^-_j)$ and $C_{j-\frac{1}{2}} = \min(R^+_{j-1}, R^-_j)$. We can ignore the case $C \equiv 1$ because this is just the high order case and only comes into play for quotients of Q and P that are greater than one. That leaves four cases we have to check.

Case I: $C_{j+\frac{1}{2}} = C_{j-\frac{1}{2}} = R^-_j$

$$u_j^{n+1} = u_j^{td} + \frac{(u_j^{td} - u_j^{\min})}{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}} (A_{j-\frac{1}{2}} - A_{j+\frac{1}{2}}) = u_j^{td} - (u_j^{td} - u_j^{\min}) = u_j^{\min}.$$

Case II: $C_{j+\frac{1}{2}} = R^+_{j-1}$, $C_{j-\frac{1}{2}} = R^+_{j+1}$

$$\begin{aligned} u_j^{n+1} &= u_j^{td} + \left(\frac{u_{j-1}^{\max} - u_{j-1}^{td}}{\max(0, A_{j-\frac{3}{2}}) - A_{j-\frac{1}{2}}} A_{j-\frac{1}{2}} - \frac{u_{j+1}^{\max} - u_{j+1}^{td}}{A_{j+\frac{1}{2}} - \min(0, A_{j+\frac{3}{2}})} A_{j+\frac{1}{2}} \right) \\ &\geq u_j^{td} + \left(\frac{u_{j-1}^{\max} - u_{j-1}^{td}}{-A_{j-\frac{1}{2}}} A_{j-\frac{1}{2}} - \frac{u_{j+1}^{\max} - u_{j+1}^{td}}{A_{j+\frac{1}{2}}} A_{j+\frac{1}{2}} \right) \\ &= u_j^{td} - (u_{j-1}^{\max} - u_{j-1}^{td}) - (u_{j+1}^{\max} - u_{j+1}^{td}) \end{aligned}$$

and using again the definition of the C-factors and R^-_j,

$$\geq u_j^{td} - \frac{u_j^{td} - u_j^{\min}}{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}} (A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}) = u_j^{\min}.$$

Case III: $C_{j+\frac{1}{2}} = R^+_{j-1}$, $C_{j-\frac{1}{2}} = R^-_j$

$$\begin{aligned} u_j^{n+1} &= u_j^{td} + \left(\frac{u_{j-1}^{\max} - u_{j-1}^{td}}{\max(0, A_{j-\frac{3}{2}}) - A_{j-\frac{1}{2}}} A_{j-\frac{1}{2}} - \frac{u_j^{td} - u_j^{\min}}{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}} A_{j+\frac{1}{2}} \right) \\ &\geq u_j^{td} + \left(\frac{u_{j-1}^{\max} - u_{j-1}^{td}}{\max(0, A_{j-\frac{3}{2}}) - A_{j-\frac{1}{2}}} A_{j-\frac{1}{2}} - \frac{u_{j+1}^{\max} - u_{j+1}^{td}}{A_{j+\frac{1}{2}} - \min(0, A_{j+\frac{3}{2}})} A_{j+\frac{1}{2}} \right) \\ &\geq u_j^{\min} \quad \text{(see case II).} \end{aligned}$$

Case IV: $C_{j+\frac{1}{2}} = R_j^-$, $C_{j-\frac{1}{2}} = R_{j+1}^+$

$$\begin{aligned} u_j^{n+1} &= u_j^{td} + \frac{(u_j^{td} - u_j^{\min})}{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}} A_{j-\frac{1}{2}} - \frac{u_{j+1}^{\max} - u_{j+1}^{td}}{A_{j+\frac{1}{2}} - \min(0, A_{j+\frac{3}{2}})} A_{j+\frac{1}{2}} \\ &\geq u_j^{td} + \frac{(u_j^{td} - u_j^{\min})}{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}} A_{j-\frac{1}{2}} - \frac{u_j^{td} - u_j^{\min}}{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}} A_{j+\frac{1}{2}} \\ &= u_j^{\min}. \end{aligned}$$

All possibilities are covered now and in all cases we found that $u_j^{n+1} \geq u_j^{\min}$, which in turn has to be greater than zero if the initial data was nonnegative and if the low order scheme is positivity preserving. Then we have shown that Zalesak's FCT scheme is in fact positivity preserving. □

Overall, the weak nonlinear stability condition

$$u_j^{\min} \leq u_j^{n+1} \leq u_j^{\max}$$

holds. The second part of the inequality is shown analogously to the calculations above.

The only thing left now is the definition of the numerical fluxes. In this thesis, for the high order flux the simple second order flux

$$F_{j+\frac{1}{2}} = \frac{1}{2}(f_{j+1} + f_j), \tag{5.5}$$

the fourth order scheme

$$F_{j+\frac{1}{2}} = \frac{7}{12}(f_{j+1} + f_j) - \frac{1}{12}(f_{j+2} + f_{j-1}), \tag{5.6}$$

as in [Zal79] are used as well as the Lax-Wendroff method from chapter 3.2. Surprisingly, numerical experiments show that the fact that the first two schemes are unstable when used on their own is irrelevant in most applications with FCT. The limiting process seems to compensate a lot, so we can employ the computationally easier flux. However, this is just an observation. A detailed analysis of the amplification factor cannot support this for a fact for arbitrary limiting factors $C_{j+\frac{1}{2}}$.

For the low order scheme, we use the Lax-Friedrichs flux

$$F_{j+\frac{1}{2}} = \frac{1}{2}(f_{j+1} + f_j) - \frac{1}{2\lambda}\left(u_{j+1}^n - u_j^n\right) \tag{5.7}$$

or the upwind method.

Dispersion

We have reviewed numerical dispersion of the Yee scheme and of several classical conservative methods. The topic has not been addressed, however, for FCT schemes in the literature. So let us look at it in some detail.

For the dispersion error, we would expect a combination of the dispersion of the two schemes used. And this is exactly what happens. Whenever only the high order scheme is used, we obtain the dispersion relation of the high order scheme. Where only the low order scheme is used, the dispersion relation becomes that of the low order scheme.

Let us look at the combination of the Lax-Friedrichs scheme (5.7) with the fourth order flux (5.6) as an example to gain some insight into numerical dispersion of the FCT scheme.

Lemma 5.2. *For FCT with Lax-Friedrichs and fourth order central fluxes, the dispersion relation yields*

$$\mathrm{Im}(\omega) \approx \big(1-\cos(k\Delta x)\big)\left(\frac{1}{\Delta t} - \frac{C_{j+\frac{1}{2}}}{2\Delta t} - \frac{C_{j-\frac{1}{2}}}{2\Delta t}\right) + \big(1-\cos(2k\Delta x)\big)\left(C_{j+\frac{1}{2}}\frac{a}{12\Delta x} - C_{j-\frac{1}{2}}\frac{a}{12\Delta x}\right).$$

Proof. Inserting $e^{ikx-i\omega t}$ into the FCT scheme for the linear advection equation yields the dispersion relation

$$\begin{aligned}
e^{i\omega\Delta t} &= 1 - \lambda C_{j+\frac{1}{2}}\frac{a}{12}\left(e^{-ik\Delta x} + 1 - e^{-2ik\Delta x} - e^{ik\Delta x}\right) - \frac{1}{2}C_{j+\frac{1}{2}}\left(e^{-ik\Delta x} - 1\right) \\
&\quad + \frac{1}{2}C_{j-\frac{1}{2}}\left(1 - e^{ik\Delta x}\right) + \lambda C_{j-\frac{1}{2}}\frac{a}{12}\left(1 + e^{ik\Delta x} - e^{-ik\Delta x} - e^{2ik\Delta x}\right) \\
&\quad - \frac{\lambda a}{2}\left(e^{-ik\Delta x} - e^{ik\Delta x}\right) + \frac{1}{2}\left(e^{-ik\Delta x} - 2 + e^{ik\Delta x}\right) \\
&= 1 - \lambda C_{j+\frac{1}{2}}\frac{a}{12}\big(1 - \cos(2k\Delta x) + i\sin(2k\Delta x) - 2i\sin(k\Delta x)\big) \\
&\quad - \frac{1}{2}C_{j+\frac{1}{2}}(\cos(k\Delta x) - i\sin(k\Delta x) - 1) + \frac{1}{2}C_{j-\frac{1}{2}}\big(1 - \cos(k\Delta x) - i\sin(k\Delta x)\big) \\
&\quad + \lambda C_{j-\frac{1}{2}}\frac{a}{12}\big(1 - \cos(2k\Delta x) - i\sin(2k\Delta x) + 2i\sin(k\Delta x)\big) \\
&\quad + \lambda a i\sin(k\Delta x) + (\cos(k\Delta x) - 1).
\end{aligned}$$

So by Taylor series expansion of the exponential $e^{i\omega t} = 1 + i\omega t + \mathcal{O}(\Delta t^2)$, we can approximate the imaginary part of ω by

$$\mathrm{Im}(\omega) \approx \big(1-\cos(k\Delta x)\big)\left(\frac{1}{\Delta t} - \frac{C_{j+\frac{1}{2}}}{2\Delta t} - \frac{C_{j-\frac{1}{2}}}{2\Delta t}\right) + \big(1-\cos(2k\Delta x)\big)\left(C_{j+\frac{1}{2}}\frac{a}{12\Delta x} - C_{j-\frac{1}{2}}\frac{a}{12\Delta x}\right).$$

□

$C \equiv 1$ means that only the fourth order scheme is used, and as expected we have $\mathrm{Im}(\omega) = 0$. $C \equiv 0$ means that only the Lax-Friedrichs scheme is used and we obtain

$$\mathrm{Im}(\omega) \approx \frac{1}{\Delta t}\big(1 - \cos(k\Delta x)\big),$$

the dispersion relation of the Lax-Friedrichs scheme. For other values of C, i.e., when we cannot use all of the high order flux, but do not need to drop down to the first order scheme entirely, we have something new: If all our C's take the same value, the new second term vanishes and we have a fraction of the Lax-Friedrichs dispersion left — the smaller C, the higher the dispersion. If the C's are different — which will usually be the case — the last term does not vanish. But if two neighboring C's are close — which is usually the case — it is rather small. In any case it is roughly of the same order as the Lax-Friedrichs term.

A similar analysis of the dispersion relation of FCT can be carried out for any other combination of two schemes. The results are similar to what we have seen here.

To see how FCT compares to the separate schemes, we consider again the classical test of advecting a square wave. In this example, we use Lax-Friedrichs and the fourth order flux. Recall from section 3.2 that the diffusion of the Lax-Friedrichs scheme will smear out the square wave and a high order scheme produces spurious ripples or even instabilities. FCT on the other hand, preserves the square shape nearly perfectly, see figure 5.1.

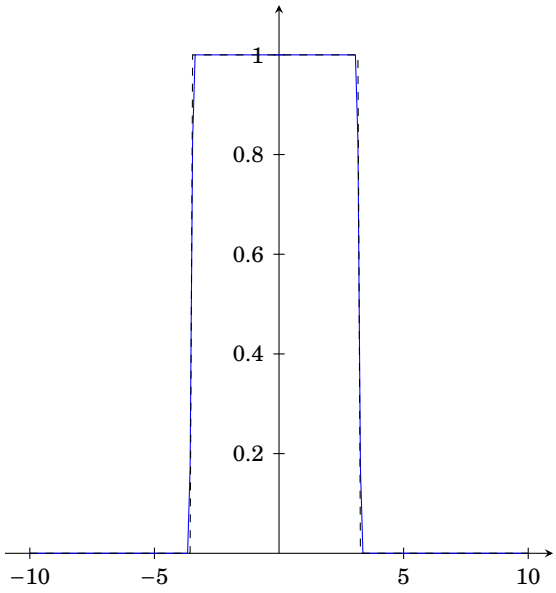

Figure 5.1: *FCT with Lax-Friedrichs and fourth order flux*

5.2 Multidimensional Flux-Corrected Transport

The Zalesak method of flux correction explained above can easily be generalized to higher dimensions. Let us consider a two-dimensional system of conservation laws

$$\partial_t u(x,y,t) + \partial_x f(u,x,y,t) + \partial_y g(u,x,y,t) = 0.$$

If we work on a finite volume coordinate-aligned mesh, we can define our two-dimensional FCT algorithm as

$$u_{ij}^{n+1} = u_{ij}^n - \lambda \left[F_{i+\frac{1}{2},j} - F_{i-\frac{1}{2},j} + G_{i,j+\frac{1}{2}} - G_{i,j-\frac{1}{2}} \right]$$

where in this case, $\lambda = \frac{\Delta t}{\Delta V}$ and the transportive fluxes F and G are basically computed as before. The full procedure is listed in algorithm 5.2.

Again, we need to specify the limiting step 5. This is completely analogous to the one-dimensional case:

$$\begin{aligned}
P_{ij}^+ &= \text{sum of all anti-diffusive fluxes } \textit{into} \text{ grid point } (i,j) \\
&= \max\left(0, A_{i-\frac{1}{2},j}\right) - \min\left(0, A_{i+\frac{1}{2},j}\right) + \max\left(0, A_{i,j-\frac{1}{2}}\right) - \min\left(0, A_{i,j+\frac{1}{2}}\right) \\
Q_{ij}^+ &= \left(u_{ij}^{\max} - u_{ij}^{td}\right) \lambda^{-1} \\
R_{ij}^+ &= \begin{cases} \min\left(1, Q_{i,j}^+ / P_{ij}^+\right) & \text{if } P_{ij}^+ > 0 \\ 0 & \text{if } P_{ij}^+ = 0 \end{cases}
\end{aligned}$$

Algorithm 5.2 Multidimensional FCT

for $n = 0, \ldots, M-1$ **do**

1. Compute the fluxes $F^L_{i+\frac{1}{2},j}$ and $G^L_{i,j+\frac{1}{2}}$ by a low order monotonic scheme
2. Compute the fluxes $F^H_{i+\frac{1}{2},j}$ and $G^H_{i,j+\frac{1}{2}}$ by a high order scheme
3. Define the "anti-diffusive fluxes"

$$A_{i+\frac{1}{2},j} = F^H_{i+\frac{1}{2},j} - F^L_{i+\frac{1}{2},j}$$
$$A_{i,j+\frac{1}{2}} = G^H_{i,j+\frac{1}{2}} - G^L_{i,j+\frac{1}{2}}$$

4. Compute the low order update, the "transported and diffused" solution

$$u^{td}_{ij} = u^n_{ij} - \lambda \left[F^L_{i+\frac{1}{2},j} - F^L_{i-\frac{1}{2},j} + G^L_{i,j+\frac{1}{2}} - G^L_{i,j-\frac{1}{2}} \right]$$

5. Limit the anti-diffusive fluxes

$$A^C_{i+\frac{1}{2},j} = C_{i+\frac{1}{2},j} A_{i+\frac{1}{2},j}, \quad 0 \le C_{i+\frac{1}{2},j} \le 1$$
$$A^C_{i,j+\frac{1}{2}} = C_{i,j+\frac{1}{2}} A_{i,j+\frac{1}{2}}, \quad 0 \le C_{i,j+\frac{1}{2}} \le 1$$

6. Compute the new solution with the limited anti-diffusive fluxes:

$$u^{n+1}_{ij} = u^{td}_{ij} - \lambda \left[A^C_{i+\frac{1}{2},j} - A^C_{i-\frac{1}{2},j} + A^C_{i,j+\frac{1}{2}} - A^C_{i,j-\frac{1}{2}} \right]$$

end for

and

$$\begin{aligned}
P_{ij}^- &= \text{sum of all anti-diffusive fluxes } \textit{away from} \text{ grid point } (i,j) \\
&= \max\left(0, A_{i+\frac{1}{2},j}\right) - \min\left(0, A_{i-\frac{1}{2},j}\right) + \max\left(0, A_{i,j+\frac{1}{2}}\right) - \min\left(0, A_{i,j-\frac{1}{2}}\right) \\
Q_{ij}^- &= \left(u_{ij}^{td} - u_{ij}^{\min}\right)\lambda^{-1} \\
R_{ij}^- &= \begin{cases} \min\left(1, Q_{ij}^-/P_{ij}^-\right) & \text{if } P_{ij}^- > 0 \\ 0 & \text{if } P_{ij}^- = 0 \end{cases}
\end{aligned}$$

and the limiting factors are

$$\begin{aligned}
C_{i+\frac{1}{2},j} &= \begin{cases} \min\left(R_{i+1,j}^+, R_{ij}^-\right) & \text{if } A_{i+\frac{1}{2},j} \geq 0 \\ \min\left(R_{ij}^+, R_{i+1,j}^-\right) & \text{if } A_{i+\frac{1}{2},j} < 0 \end{cases} \\
C_{i,j+\frac{1}{2}} &= \begin{cases} \min\left(R_{i,j+1}^+, R_{ij}^-\right) & \text{if } A_{i,j+\frac{1}{2}} \geq 0 \\ \min\left(R_{ij}^+, R_{i,j+1}^-\right) & \text{if } A_{i,j+\frac{1}{2}} < 0. \end{cases}
\end{aligned}$$

The upper and lower bounds for the computation of $Q_{ij}^{\pm}$ now contain four neighbors,

$$\begin{aligned}
u_{ij}^a &= \max\left(u_{ij}^n, u_{ij}^{td}\right) \\
u_{ij}^{\max} &= \max\left(u_{i-1,j}^a, u_{i+1,j}^a, u_{ij}^a, u_{i,j-1}^a, u_{i,j+1}^a\right) \\
u_{ij}^b &= \min\left(u_{ij}^n, u_{ij}^{td}\right) \\
u_{ij}^{\min} &= \max\left(u_{i-1,j}^b, u_{i+1,j}^b, u_{ij}^b, u_{i,j-1}^b, u_{i,j+1}^b\right).
\end{aligned}$$

This completes the description of multidimensional flux-corrected transport.

Numerical Tests in Two Dimensions

Unfortunately, as DeVore pointed out in [DeV98], the independence of the numerical fluxes into different directions allows the creation of new ripples. Zalesak himself noted this, too, but was not too concerned about the effects as they did not show up in his tests. They can be shown, however, by a simple transport example: Consider linear advection in two dimensions with constant velocities in x and y-direction. Figure 5.2 shows the initial profile that is to be transported to the lower right. Figure 5.3 shows the advected profile with dimensionally applied FCT as described above. The ripples that form perpendicular to the direction of propagation are clearly visible.

The remedy suggested by both authors is to use the original Boris and Book limiter along each coordinate direction before the actual limiting procedure in step 5:

$$\begin{aligned}
A'_{i+\frac{1}{2},j} &= \text{minmod}\left(u_{ij}^{td} - u_{i-1,j}^{td}, A_{j+\frac{1}{2},j}, u_{i+2,j}^{td} - u_{i+1,j}^{td}\right) \\
A'_{i,j+\frac{1}{2}} &= \text{minmod}\left(u_{ij}^{td} - u_{i,j-1}^{td}, A_{i,j+\frac{1}{2}}, u_{i,j+2}^{td} - u_{i,j+1}^{td}\right)
\end{aligned}$$

and continue with A' instead of A.

If we use this pre-limiting step in the above example, we see significant improvement in figure 5.4.

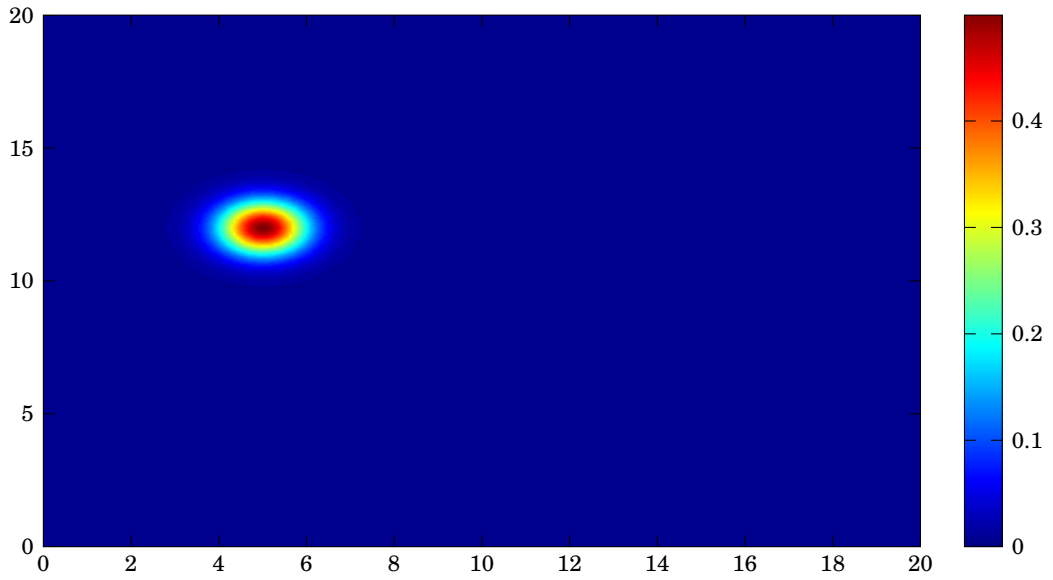

Figure 5.2: *Initial data for multidimensional linear advection*

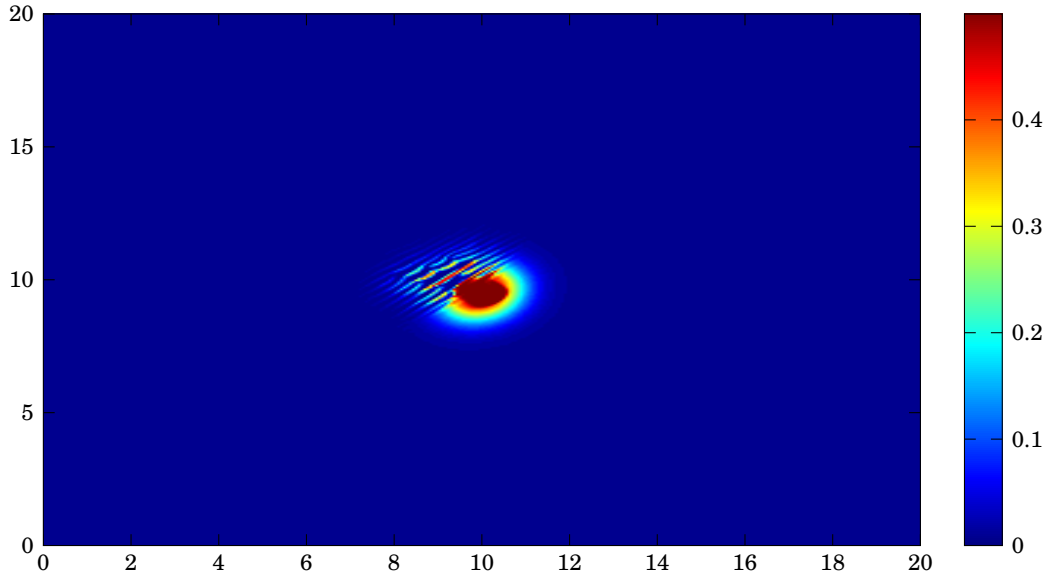

Figure 5.3: *Dimensionally applied FCT for linear advection*

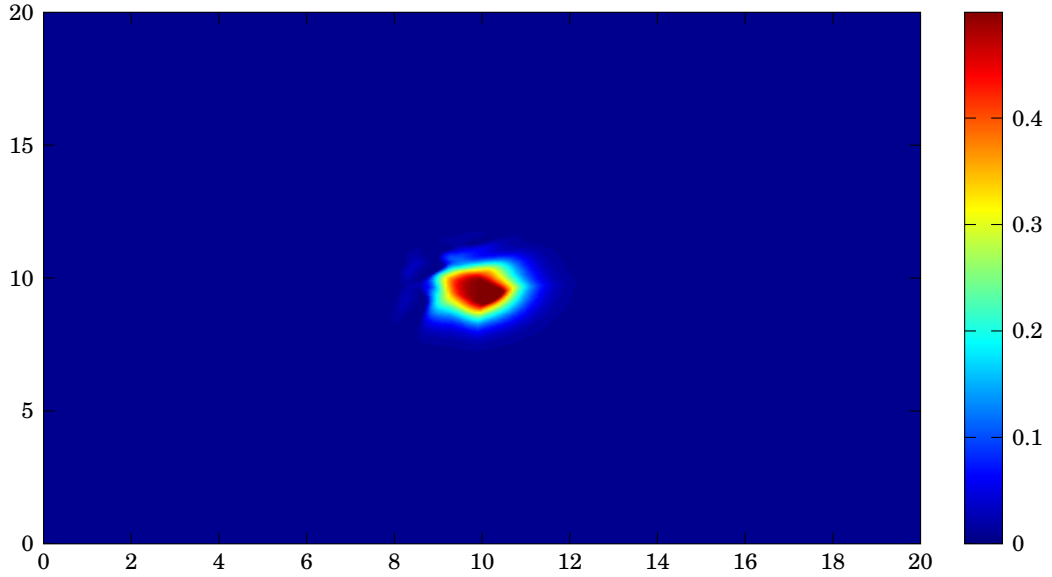

Figure 5.4: *Multidimensional FCT with prelimiting for linear advection*

To properly test FCT in the multidimensional case, we have to consider something more than just constant velocity advection. One of the most widely used examples to test multidimensional algorithms is the rotating cylinder: Consider

$$\partial_t u + \partial_x(uv_x) + \partial_x(uv_y) = 0 \quad \text{on } (0,1)^2$$

for $v_x = -\Omega(y - y_0)$, $v_y = \Omega(x - x_0)$ with some angular frequency Ω and the axis of rotation in $(x_0, y_0) \in (0,1)^2$.

This is somewhat similar to linear advection. The velocity is still constant in time, but not in space anymore. The initial profile is a cylinder whose values range from the bottom plateau at $z = 1$ to the top at $z = 3$. To be able to observe more effects, the cylinder has a slot through the middle (see figure 5.5). This way we can observe not only the advection of the initial data, but also how well its shape is preserved.

With a monotone scheme there is already a lot of diffusion after a quarter revolution in figure 5.6 not only for the cylinder as a whole, but also for the slot, which changed its shape. After a full revolution, the cylinder is unrecognizably diffused (see figure 5.7).

Using the simple second or fourth order scheme is impossible because of their instability. The spurious oscillations are so strong that the cylinder is gone before a quarter revolution. The Lax-Wendroff method, however, works well here and is stable, but as expected, it suffers from oscillations due to Gibb's phenomenon.

With multidimensional FCT, we have a fairly good preservation of the shape in figures 5.10 for the quarter and in figure 5.11 for the full revolution. There is a little diffusion, but there are no spurious oscillations and the maximum value still lies at $z = 3.00$..

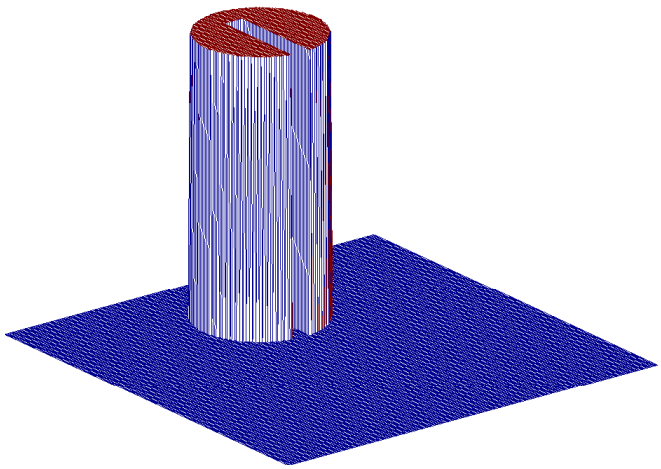

Figure 5.5: *Initial data for the rotating cylinder — values ranging from* $z = 1$ *to* $z = 3$

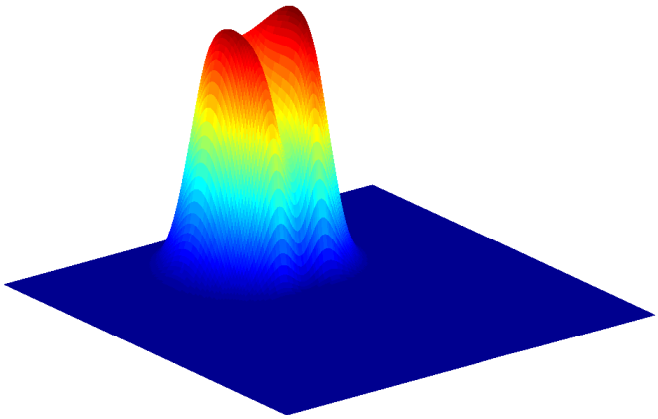

Figure 5.6: *Rotating cylinder after a quarter revolution (rotated view for better comparison) with a low order scheme — the maximum now only lies at* $z = 2.79$

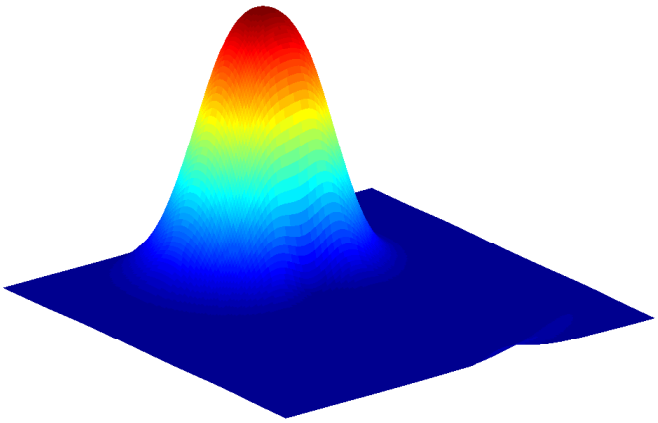

Figure 5.7: *Rotating cylinder after a full revolution with a low order scheme — the maximum now only lies at* $z = 2.38$

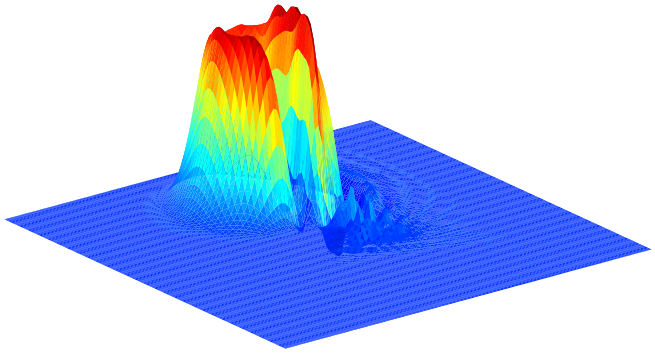

Figure 5.8: *Rotating cylinder after a quarter revolution (rotated view for better comparison) with the Lax-Wendroff scheme*

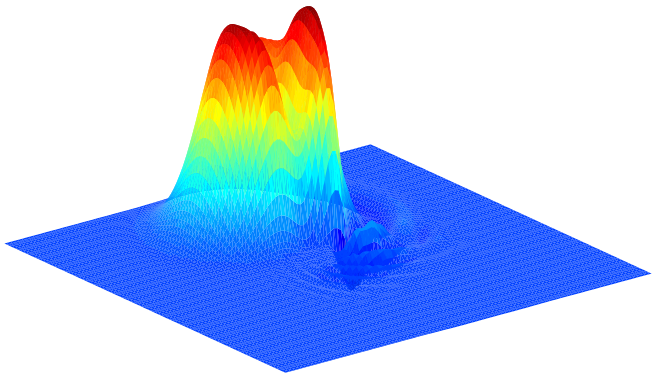

Figure 5.9: *Rotating cylinder after a full revolution with the Lax-Wendroff scheme*

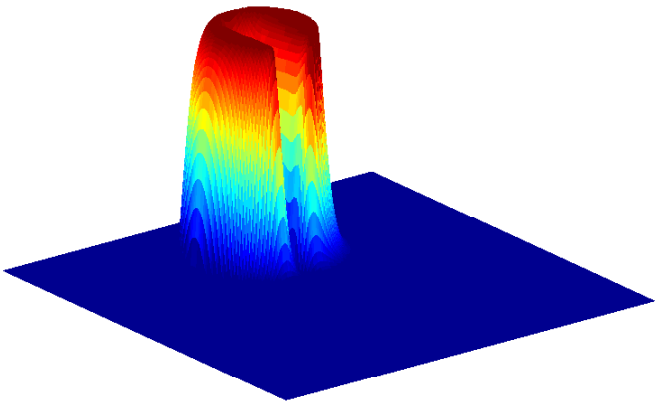

Figure 5.10: *Rotating cylinder after a quarter revolution (rotated view for better comparison) with FCT*

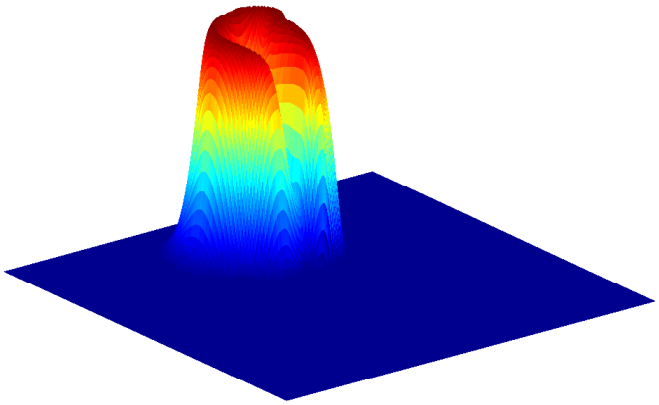

Figure 5.11: *Rotating cylinder after a full revolution with FCT*

5.3 Higher Order Time Discretization — SSP Runge-Kutta Methods

For most time-dependent partial differential equations, a method-of-lines approach is used. That is, a semi-discretization of the spatial derivatives is performed to obtain a system of time-dependent ordinary differential equations. Recall the one-dimensional scalar conservation law

$$\partial_t u + \partial_x f(u) = 0$$

for a space- and time-dependent function $u = u(x,t)$. The method-of-lines approach then yields a system of ordinary differential equations in the time variable t

$$\frac{\partial \boldsymbol{u}_\Delta}{\partial t} = \boldsymbol{L}(\boldsymbol{u}_\Delta) \tag{5.8}$$

where $\boldsymbol{u}_\Delta$ denotes the spatially discretized solution and $\boldsymbol{L}$ denotes the spatial discretization operator. For ease of notation, let us rewrite (5.8) as a scalar equation for u,

$$\frac{\partial u}{\partial t} = L(u). \tag{5.9}$$

We now need a time integrator for (5.9). Until now, we have only considered forward Euler, as it is in any aspect the easiest method for temporal discretization. For general ODEs, we can use, for example, higher order Runge-Kutta methods. However, for nonlinear hyperbolic conservation laws, we need to ensure nonlinear stability. All nonlinearly stable methods we have discussed so far are based on the forward Euler method. It is not clear a priori if and how this stability can be conserved when using a different time integrator. Refer to [GKS11] and the references therein for various examples of higher order Runge-Kutta methods that fail to maintain the nonlinear stability properties introduced by the spatial discretization.

That is why we consider a special class of Runge-Kutta methods first introduced by Shu in [Shu88] and by Shu and Osher in [SO88]. They considered mainly total variation stability (TVD) and therefore called the methods *TVD Runge-Kutta methods*. The more general name now is *strong stability preserving (SSP) Runge-Kutta methods*.

Definition 5.3. *A Runge-Kutta method applied to* (5.8) *is* strong stability preserving (SSP) *with SSP coefficient $\mathscr{C}$ if for time steps $\Delta t \leq \mathscr{C}\Delta t_{FE}$ it holds that*

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

in some (semi-)norm $\|\cdot\|$ *whenever*

$$\|u + \Delta t L(u)\| \le \|u\| \quad \text{for } 0 \le \Delta t \le \Delta t_{FE}, \quad \text{for all } u.$$

SSP Runge-Kutta methods are based on the forward Euler method because of its stability properties. Assuming that the forward Euler time discretization is stable for time steps up to Δt_{FE} under a certain (semi-)norm, $\|u^{n+1}\| \le \|u^n\|$, then SSP methods should maintain this stability under a suitable time step restriction.

Algorithm 5.3 shows an explicit Runge-Kutta method with s stages in so-called *Shu-Osher form* (cf. [GKS11, chapter 2]):

Algorithm 5.3 SSP Runge-Kutta methods

for $n = 0, \dots, M-1$ **do**
  Set $u^{(0)} = u^n$
  for $i = 1, \dots, s$ **do**
    $u^{(i)} = \sum_{j=0}^{i-1} \left(\alpha_{ij} u^{(j)} + \Delta t \beta_{ij} L(u^{(j)})\right)$
  end for
  Set $u^{n+1} = u^{(s)}$
end for

Consistency requires that $\sum_{j=0}^{i-1} \alpha_{ij} = 1$. Of course, SSP Runge-Kutta methods in Shu-Osher form can easily be transformed into the more widely used Butcher array form, but we will shortly see why the Shu-Osher form is more advantageous here.

First note that for non-negative coefficients α_{ij} and β_{ij}, the scheme can be rewritten as convex combinations of forward Euler steps with a modified time step. This motivates

Theorem 5.4. *If the forward Euler method applied to* (5.9) *is strongly stable under the time step restriction* $\Delta t \le \Delta t_{FE}$*, i.e., if*

$$\|u + \Delta t L(u)\| \le \|u\| \ \text{ for } \ 0 \le \Delta t \le \Delta t_{FE} \ \text{ for all } \ u,$$

and if α_{ij}, $\beta_{ij} \ge 0$*, then the solution obtained by an SSP Runge-Kutta method satisfies the strong stability bound*

$$\|u^{n+1}\| \le \|u^n\|, \tag{5.10}$$

under the time step restriction

$$\Delta t \le \mathscr{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \Delta t_{FE}$$

where $\mathscr{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{i,j} \frac{\alpha_{ij}}{\beta_{ij}}$ *and the ratio is understood infinite if* $\beta_{ij} = 0$*.*

Proof. With the non-negativity of the coefficients and the consistency property $\sum_{j=0}^{i-1} \alpha_{ij} = 1$, we see that

$$\begin{aligned}
\|u^{(i)}\| &= \left\| \sum_{j=0}^{i-1} \left(\alpha_{ij} u^{(j)} + \Delta t \beta_{ij} L(u^{(j)})\right) \right\| \\
&= \left\| \sum_{j=0}^{i-1} \alpha_{ij} \left(u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} L(u^{(j)})\right) \right\| \\
&\le \sum_{j=0}^{i-1} \alpha_{ij} \left\| u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} L(u^{(j)}) \right\|
\end{aligned}$$

Since each $\|u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} L(u^{(j)})\| \leq \|u^{(j)}\|$ as long as $\frac{\beta_{ij}}{\alpha_{ij}} \Delta t \leq \Delta t_{FE}$, we can deduce — using again consistency — that $\|u^{(i)}\| \leq \|u^n\|$ for each stage as long as $\frac{\beta_{ij}}{\alpha_{ij}} \Delta t \leq \Delta t_{FE}$ for all i and j, so in particular, $\|u^{n+1}\| \leq \|u^n\|$. □

Remark 5.5. Note that this proof only holds for non-negative coefficients α, β. SSP methods can, however, also be constructed for negative β_{ij} using convex combinations of forward and backward Euler. Due to the somewhat more complex structure, both sets of coefficients are sometimes required to be non-negative in the definition of SSP Runge-Kutta methods (cf. [GKS11]).

Theorem 5.4 provides a sufficient time step restriction for the solution to satisfy the strong stability bound (5.10). We do not know, however, if it is also necessary nor how to find any SSP Runge-Kutta methods or identify the ones with largest possible SSP coefficient $\mathscr{C}$.

To be able to make sensible statements on the SSP coefficient $\mathscr{C}$, let us only consider irreducible methods, i.e., those that cannot be represented by an equivalent method with fewer stages. But even then, the representation of a method is not unique. Consider the the second order Runge-Kutta method, based on the trapezoidal rule,

$$\begin{aligned} u^{(1)} &= u^n + \Delta t L(u^n) \\ u^{n+1} &= u^n + \frac{1}{2}\Delta t L(u^n) + \frac{1}{2}\Delta t L(u^{(1)}). \end{aligned}$$

In this form, we obtain $\mathscr{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$ because $\alpha_{21} = 0$ while $\beta_{21} = \frac{1}{2}$. However, the method can be rewritten as

$$u^{n+1} = \frac{3}{4}u^n + \frac{1}{4}\Delta t L(u^n) + \frac{1}{4}u^{(1)} + \frac{1}{2}\Delta t L(u^{(1)}),$$

which yields $\mathscr{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}$, while

$$u^{n+1} = \frac{1}{2}u^n + \frac{1}{2}u^{(1)} + \frac{1}{2}\Delta t L(u^{(1)})$$

yields $\mathscr{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 1$. All three variants are equivalent representations in Shu-Osher form of the same method. Of course, we are interested in the largest possible SSP coefficient for a given method. To do so, a unique representation of any given method is needed. For irreducible methods, the Butcher form is in fact unique, but does not help us find the SSP coefficient. Therefore, a *canonical Shu-Osher form* has been derived to uniquely represent any method and determine its SSP coefficient. We refer to [GKS11] for the details of this unique representation and how to compute the SSP coefficient and derive methods with optimal SSP coefficients. For the scope of this thesis, we will content ourselves with what we know so far and citing the concerning methods and coefficients.

For the above second order method, $\mathscr{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 1$ is optimal. The most commonly used SSP Runge-Kutta method is the third order three stage method, which is often called *the Shu-Osher method*. It reads

$$\begin{aligned} u^{(1)} &= u^n + \Delta t L(u^n) \\ u^{(2)} &= \frac{3}{4}u^n + \frac{1}{4}u^{(1)} + \frac{1}{4}\Delta t L(u^{(1)}) \\ u^{n+1} &= \frac{1}{3}u^n + \frac{2}{3}u^{(2)} + \frac{2}{3}\Delta t L(u^{(2)}) \end{aligned}$$

and is optimal with $\mathscr{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 1$, so it can be used with the same time step restriction as forward Euler. We found this method to work really well in our numerical experiments, which makes it our method of choice throughout this thesis.

Note that the strong stability property also includes positivity preservation in that a convex combination of positivity preserving Euler steps will still preserve positivity.

Improving Flux-Corrected Transport

Now that we have established some theoretical background on SSP Runge-Kutta methods, we would like to use them in our FCT framework.

From the integral form of our conservation law (3.1) and the derivation of finite volume methods we know that

$$u_j'(t) = -\frac{1}{\Delta x}\left(F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}\right), \quad j = j_1, \ldots, j_N,$$

so we have a formulation we can use to apply a Runge-Kutta method.

So now we want to look at a comparison of classical FCT with forward Euler time discretization and FCT combined with the Shu-Osher method. Since the order of the spatial discretization drops down to first order at steep gradients, we cannot gain much from a very high order time integration. Numerical experiments show, however, that this third order method still yields better results than one with second order. To be able to actually observe order, we cannot take the square wave example here. Instead, we use the Gaussian wave test with periodic boundary conditions as an example. The initial data, a Gaussian, is advected by exactly one interval length. The exact solution is hence equal to the initial data. For this experiment, we chose $\lambda = \frac{\Delta t}{\Delta x} = \frac{1}{2}$.

Figure 5.12 shows the result with classical FCT. There is some diffusion and the numerically computed pulse leans somewhat to the left.

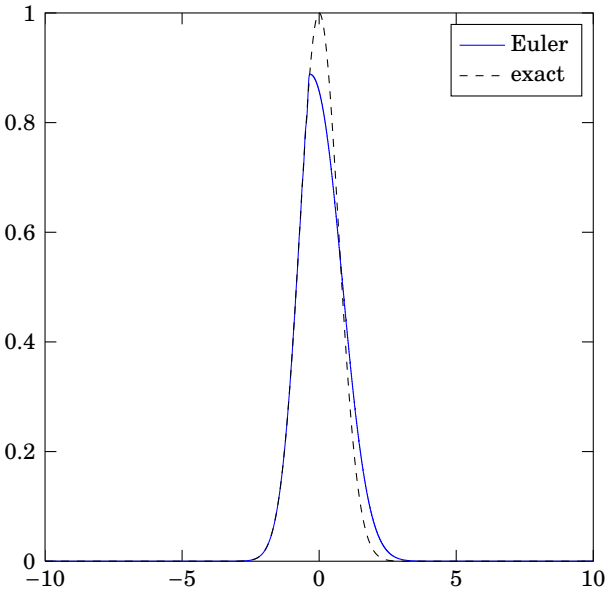

Figure 5.12: *Classical FCT scheme*

Now, we run the same test again, but this time using the third order Shu-Osher method, see figure 5.13. Note how much influence the higher order method has even on

the shape of the Gaussian in figure 5.14. Both do suffer some diffusion, but not nearly as badly as with upwind or Lax-Friedrichs alone.

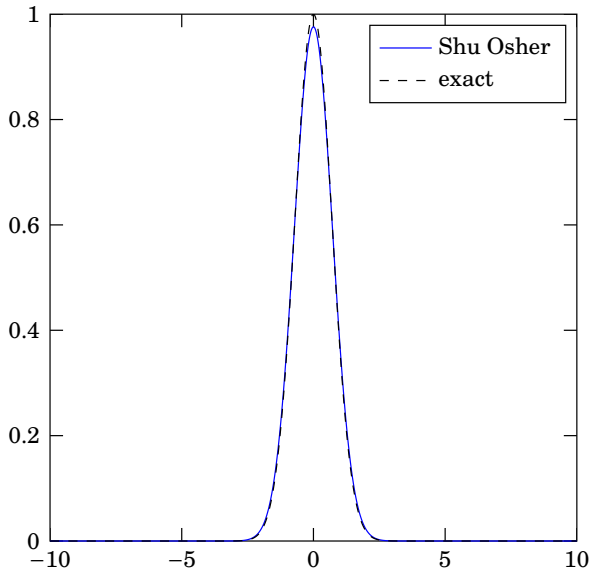

Figure 5.13: *Shu-Osher scheme with FCT flux*

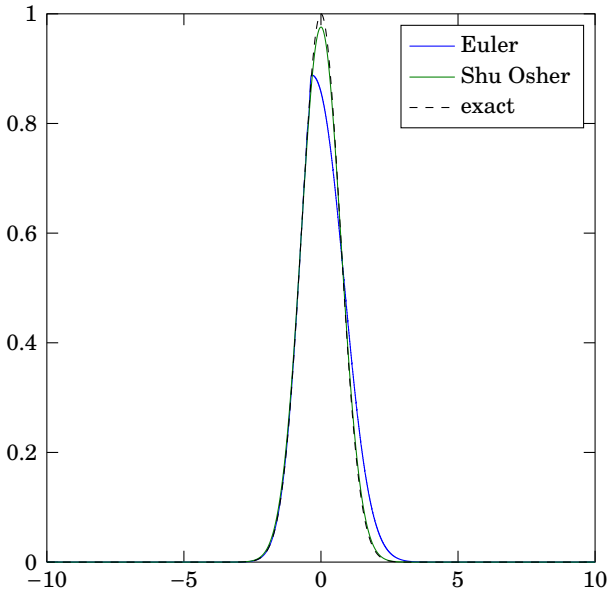

Figure 5.14: *Direct comparison of the Euler and Shu-Osher scheme, both with FCT flux*

6 Relativistic Laser-Plasma Interaction in 1D

In chapter 3, we have discussed how hyperbolic conservation laws may have discontinuous solutions despite smooth initial data. This poses a major problem for the numerical solution. Linear methods of order higher than one produce spurious oscillations. Especially undershoots can be problematic when they lead to negative values of the numerical solution.

In this thesis, we want to simulate a relativistic vacuum-plasma transition of a laser pulse. In the simulation of laser-plasma interaction, we are dealing with plasma density, which cannot be negative physically. Classical schemes, however, cannot guarantee non-negativity, which is why we will be using the FCT scheme from chapter 5.

6.1 Equations

We want to model a laser pulse that starts in a vacuum and enters a plasma. In the one dimensional case, the domain $\Omega \subset \mathbb{R}$ is simply some interval. We are not interested in boundary effects, so we are going to use periodic boundary conditions for an easy implementation.

In most simulations, the equations are scaled by quantities of the problem setting. In this case, measuring everything in terms of the plasma yields the "nicest" equations, as we have seen in chapter 4.

For the laser, we have Maxwell's equations

$$\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B} - \boldsymbol{j} \tag{6.1a}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E}, \tag{6.1b}$$

see chapter 2.

We start by considering the one-dimensional case, i.e., derivatives with respect to y and z are set to zero, which yields $\partial_t B_x = 0$ and hence $B_x \equiv 0$. Thus, (6.1) read

$$\begin{aligned}
\frac{\partial E_x}{\partial t} &= -j_x \\
\frac{\partial E_y}{\partial t} &= -\frac{\partial B_z}{\partial x} - j_y \\
\frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x} - j_z \\
\frac{\partial B_y}{\partial t} &= \frac{\partial E_z}{\partial x} \\
\frac{\partial B_z}{\partial t} &= -\frac{\partial E_y}{\partial x}.
\end{aligned}$$

For the plasma, we use the fluid formulation we have introduced in chapter 4. In the one-dimensional case, they read

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v_x) = 0 \tag{6.2a}$$

$$\frac{\partial (\rho \boldsymbol{p})}{\partial t} + \frac{\partial}{\partial x}(\rho \boldsymbol{p} v_x) = \rho(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \tag{6.2b}$$

where ρ is the density, $\boldsymbol{v}$ the velocity and $\boldsymbol{p}$ the momentum. All quantities except ρ are three-dimensional functions; subscripts indicate the component.

Needless to say, we need appropriate initial conditions to all equations for unique solutions.

6.2 The YeeFCT Algorithm

We have a large amount of equations that need to be solved. We have discussed the numerical solution of Maxwell's equations in chapter 2 and of hyperbolic conservation laws in section 3.2. We even elaborated the Boris push for the correct handling of the rotational Lorentz force. We will now show how to combine them.

To solve (6.1) and (6.2) numerically, we will use a symmetric splitting technique. For our case, we will have three components. First, there are Maxwell's equations without the current $\boldsymbol{j}$,

$$\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B} \tag{6.3a}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E} \tag{6.3b}$$

whose numerical flow we will denote by $\phi^{[1]}$. Its exact definition will be explained later on. For now, we only want to discuss the structure of the splitting, for which we do not yet need to define the flows. $\phi^{[2]}$ describes the numerical flow for the right hand side in (6.2b),

$$\frac{\partial(\rho \boldsymbol{p})}{\partial t} = \rho(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{6.4}$$

Finally, we have the plasma equations (6.2a) and (6.2b) without the Lorentz force on the right hand side plus the Maxwell part with only the current $\boldsymbol{j} = \rho \boldsymbol{v}$,

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v_x) = 0 \tag{6.5a}$$

$$\frac{\partial(\rho \boldsymbol{p})}{\partial t} + \frac{\partial}{\partial x}(\rho \boldsymbol{p} v_x) = \boldsymbol{0} \tag{6.5b}$$

$$\frac{\partial \boldsymbol{E}}{\partial t} = -\boldsymbol{j} \tag{6.5c}$$

for which we use the numerical flow $\phi^{[3]}$.

Of course, we could simply combine into a Lie-Trotter splitting by solving one after the other as in equation (2.15). But as we have seen in section 2.2, a symmetric Strang splitting as in equation (2.16),

$$\phi^{[1]}_{\Delta t/2} \circ \phi^{[2]}_{\Delta t/2} \circ \phi^{[3]}_{\Delta t} \circ \phi^{[2]}_{\Delta t/2} \circ \phi^{[1]}_{\Delta t/2} \tag{6.6}$$

might be better.

Our inner step consists of the plasma equations (6.5). This is the computationally most challenging part.

The first and last step are Maxwell's equations (6.3) without the current $\boldsymbol{j}$, which has already been taken care of in the inner step. In between, we use a Boris push to treat the particle motion due to the Lorentz force in (6.4),

The system in the one-dimensional case now reads

$$\begin{aligned}
\frac{\partial B_y}{\partial t} &= \frac{\partial E_z}{\partial x} \\
\frac{\partial B_z}{\partial t} &= -\frac{\partial E_y}{\partial x} \\
\frac{\partial E_x}{\partial t} &= -j_x \\
\frac{\partial E_y}{\partial t} &= -\frac{\partial B_z}{\partial x} - j_y \\
\frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x} - j_z \\
\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v_x) &= 0 \\
\frac{\partial (\rho \boldsymbol{p})}{\partial t} + \frac{\partial}{\partial x}(\rho \boldsymbol{p} v_x) &= \rho(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})
\end{aligned}$$

where the colors indicate the splitting, which is carried out according to (6.6). Note that both the Yee scheme and the Boris push are splitting schemes themselves, so the algorithm as a whole consists of several sub-algorithms.

To explicitly state our algorithm, we first have to define a grid and the position of all quantities on this grid. To solve Maxwell's equations, we need a staggered grid. Without loss of generality, assume $\Omega = (0, L)$. We divide $\overline{\Omega}$ into subintervals $D_j = [x_j, x_{j+1}]$ where $x_j = j\Delta x$ such that $\overline{\Omega} = \bigcup_{j=1}^{N-1} D_j$. In accordance to the projection of a three-dimensional Yee-cell onto the real line, we approximate the y- and z-components of the magnetic field and the x-component of the electric field on the interfaces, while the y- and z-components of the electric field and the density, velocity and momentum are measured in the middle of each interval, i.e., at $x_{j+\frac{1}{2}}$. See figure 6.1 for a schematic view. Recall that $B_x \equiv 0$ in the one-dimensional case, otherwise it would also be approximated in the cell middle.

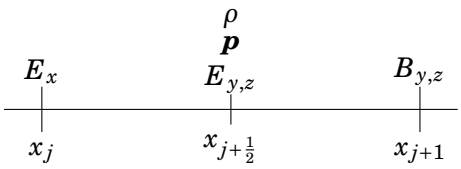

Figure 6.1: *Placement of variables on the one-dimensional grid*

The approach for solving the Maxwell part is the Yee scheme from chapter 2 where half steps for the magnetic fields were combined with whole steps for the electric fields. Since we are doing only half a step of this procedure, the steps reduce to $\frac{\Delta t}{2}$ and $\frac{\Delta t}{4}$. Thus, the first half step with Maxwell's equations looks as follows.

First, a quarter time step with the magnetic fields is performed,

$$\begin{aligned}
B_{y,j}^{n+\frac{1}{4}} &= B_{y,j}^{n} + \frac{\Delta t}{4\Delta x}\left(E_{z,j+\frac{1}{2}}^{n} - E_{z,j-\frac{1}{2}}^{n}\right) \\
B_{z,j}^{n+\frac{1}{4}} &= B_{z,j}^{n} - \frac{\Delta t}{4\Delta x}\left(E_{y,j+\frac{1}{2}}^{n} - E_{y,j-\frac{1}{2}}^{n}\right).
\end{aligned}$$

Next, we use those magnetic updates to compute a half time step for the electric fields,

$$\begin{aligned} E^-_{x,j} &= E^n_{x,j} \\ E^-_{y,j+\frac{1}{2}} &= E^n_{y,j+\frac{1}{2}} - \frac{\Delta t}{2\Delta x}\left(B^{n+\frac{1}{4}}_{z,j+1} - B^{n+\frac{1}{4}}_{z,j}\right) \\ E^-_{z,j+\frac{1}{2}} &= E^n_{z,j+\frac{1}{2}} + \frac{\Delta t}{2\Delta x}\left(B^{n+\frac{1}{4}}_{y,j+1} - B^{n+\frac{1}{4}}_{y,j}\right) \end{aligned}$$

and finally, these are plugged into another quarter step for the magnetic fields,

$$\begin{aligned} B^{n+\frac{1}{2}}_{y,j} &= B^{n+\frac{1}{4}}_{y,j} + \frac{\Delta t}{4\Delta x}\left(E^-_{z,j+\frac{1}{2}} - E^-_{z,j-\frac{1}{2}}\right) \\ B^{n+\frac{1}{2}}_{z,j} &= B^{n+\frac{1}{4}}_{z,j} - \frac{\Delta t}{4\Delta x}\left(E^-_{y,j+\frac{1}{2}} - E^-_{y,j-\frac{1}{2}}\right). \end{aligned}$$

Our next inner step is the Boris push we discussed in section 4.2. This is itself a splitting scheme, which separates the magnetic and electric update in (6.4). We proceed first with an electric update

$$(\rho \boldsymbol{p})^-_{j+\frac{1}{2}} = (\rho \boldsymbol{p})^n_{j+\frac{1}{2}} + \frac{\Delta t}{2}\rho^n_{j+\frac{1}{2}}\boldsymbol{E}^-_{j+\frac{1}{2}}$$

for all three components. Of course, we have to interpolate E_x to the half positions via $E_{x,j+\frac{1}{2}} = (E_{x,j+1} + E_{x,j})/2$. The magnetic part of the push takes two sub-steps, for which we define two quantities

$$\begin{aligned} \boldsymbol{t} &= \frac{\boldsymbol{B}}{\gamma}\frac{\Delta t}{4} \\ \boldsymbol{s} &= \frac{2\boldsymbol{t}}{1 + \|\boldsymbol{t}\|^2}. \end{aligned}$$

The latter are supposed to live on the half positions like γ, so we interpolate again to get our magnetic field onto the correct grid positions via $\boldsymbol{B}_{j+\frac{1}{2}} = (\boldsymbol{B}_{j+1} + \boldsymbol{B}_j)/2$. Then we perform two sub-steps for the cross product $\boldsymbol{v} \times \boldsymbol{B}$, which are

$$\begin{aligned} (\rho \boldsymbol{p})' &= (\rho \boldsymbol{p})^- + (\rho \boldsymbol{p})^- \times \boldsymbol{t} \\ (\rho \boldsymbol{p})^+ &= (\rho \boldsymbol{p})^- + (\rho \boldsymbol{p})' \times \boldsymbol{s}. \end{aligned}$$

Component-wise, these read

$$\begin{aligned} (\rho p_x)'_{j+\frac{1}{2}} &= (\rho p_x)^-_{j+\frac{1}{2}} + (\rho p_y)^-_{j+\frac{1}{2}} t_{z,j+\frac{1}{2}} - (\rho p_z)^-_{j+\frac{1}{2}} t_{y,j+\frac{1}{2}} \\ (\rho p_y)'_{j+\frac{1}{2}} &= (\rho p_y)^-_{j+\frac{1}{2}} - (\rho p_x)^-_{j+\frac{1}{2}} t_{z,j+\frac{1}{2}} \\ (\rho p_z)'_{j+\frac{1}{2}} &= (\rho p_z)^-_{j+\frac{1}{2}} + (\rho p_x)^-_{j+\frac{1}{2}} t_{y,j+\frac{1}{2}} \end{aligned}$$

and

$$\begin{aligned} (\rho p_x)^+_{j+\frac{1}{2}} &= (\rho p_x)^-_{j+\frac{1}{2}} + (\rho p_y)'_{j+\frac{1}{2}} s_{z,j+\frac{1}{2}} - (\rho p_z)'_{j+\frac{1}{2}} s_{y,j+\frac{1}{2}} \\ (\rho p_y)^+_{j+\frac{1}{2}} &= (\rho p_y)^-_{j+\frac{1}{2}} - (\rho p_x)'_{j+\frac{1}{2}} s_{z,j+\frac{1}{2}} \\ (\rho p_z)^+_{j+\frac{1}{2}} &= (\rho p_z)^-_{j+\frac{1}{2}} + (\rho p_x)'_{j+\frac{1}{2}} s_{y,j+\frac{1}{2}} \end{aligned}$$

in the one-dimensional case.

Now we have reached our inner fluid full step. Equations (6.5) read

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v_x) = 0 \tag{6.7a}$$

$$\frac{\partial}{\partial t}(\rho \boldsymbol{p}) + \frac{\partial}{\partial x}(\rho \boldsymbol{p} v_x) = \boldsymbol{0} \tag{6.7b}$$

$$\frac{\partial \boldsymbol{E}}{\partial t} = -\rho \boldsymbol{v} = -\frac{\rho \boldsymbol{p}}{\gamma}. \tag{6.7c}$$

Equations (6.7a) and (6.7b) are approximated by an FCT scheme. So we obtain $\rho^{n+1}_{j+\frac{1}{2}}$ from $\rho^{n}_{j+\frac{1}{2}}$ and $(\rho \boldsymbol{p})^{++}_{j+\frac{1}{2}}$ from $(\rho \boldsymbol{p})^{+}_{j+\frac{1}{2}}$ For the low order flux, we use the Lax-Friedrichs scheme and the central fourth order flux for the high order part. Numerical experiments indicated that this is the best possible combination of all methods discussed here.

We will state the following part of the algorithm with forward Euler for simplicity. In practice, we perform the time integration of the fluid equations by an SSP Runge-Kutta method (cf. section 5.3) to improve accuracy.

For (6.7c), we obtain

$$\boldsymbol{E}^{+}_{j+\frac{1}{2}} = \boldsymbol{E}^{-}_{j+\frac{1}{2}} - \Delta t \frac{(\rho \boldsymbol{p})^{+}_{j+\frac{1}{2}}}{\gamma^{+}_{j+\frac{1}{2}}}$$

for the y- and z-components. For E_x, we interpolate the right hand side to whole positions via

$$\boldsymbol{p}_j = \frac{\boldsymbol{p}_{j+\frac{1}{2}} + \boldsymbol{p}_{j-\frac{1}{2}}}{2}$$

$$\gamma_j = \sqrt{1 + \left\| \boldsymbol{p}_j \right\|_2^2}$$

$$(\rho p)_{x,j} = \frac{(\rho p)_{x,j+\frac{1}{2}} + (\rho p)_{x,j-\frac{1}{2}}}{2}$$

and obtain

$$E^{+}_{x,j} = E^{-}_{x,j} - \Delta t \frac{(\rho p)^{+}_{x,j}}{\gamma^{+}_{j}}.$$

The remaining part is quite obvious, i.e., magnetic push, electric push and Maxwell's equations. Each time we use the latest approximation we have obtained in a previous splitting step as initial value for the next step. So we obtain $(\rho \boldsymbol{p})^{*}_{j+\frac{1}{2}}$ from $(\rho \boldsymbol{p})^{++}_{j+\frac{1}{2}}$ in the magnetic push, $(\rho \boldsymbol{p})^{n+1}_{j+\frac{1}{2}}$ from $(\rho \boldsymbol{p})^{*}_{j+\frac{1}{2}}$ in the electric push, and $\boldsymbol{E}^{n+1}$ and $\boldsymbol{B}^{n+1}$ from $\boldsymbol{E}^{+}$ and $\boldsymbol{B}^{n+\frac{1}{2}}$ in the second half step with the Yee scheme.

Dispersion

We have already looked at dispersion for FCT alone. Now we have a much more complex setting that we wish to analyze. To do so, we make some simplifying assumptions. We only consider the non-relativistic case, i.e., $\gamma \equiv 1$. In this case, the magnetic part of the Lorentz force is negligible because the velocity $\boldsymbol{v}$ is small, $\|\boldsymbol{v}\| \ll c$ and we assume $\boldsymbol{B} \equiv \boldsymbol{0}$. This makes the Yee part of our algorithm superfluous because the derivatives of the magnetic field components that are needed for the electric updates are zero as well.

The algorithm then reduces to half an electric push, fluid step, and another half step of the electric push. With $\gamma \equiv 1$, we have $\boldsymbol{p} \equiv \boldsymbol{v}$. Finally, we assume v_x to be constant[8] and constant density ρ, which eliminates another equation. The field components are approximated on different places on the grid, so we cannot carry out the calculations for all three components at once. Hence, we consider the y- or z-component where we need no interpolation. For ease of notation, we drop the component subscripts for the dispersion considerations.

Lemma 6.1. *Under the above simplifying assumptions, the numerical dispersion relation for the one-dimensional YeeFCT scheme with Lax-Friedrichs and central second order fluxes in the FCT algorithm reads*

$$\begin{aligned} e^{i\omega\Delta t} = \frac{1}{2}\left(C_{j+1} + C_j - \Delta t^2\rho\right) + \frac{\Delta t}{4}\frac{\rho\Delta t}{1-\frac{\Delta t^2}{2}\rho - e^{i\omega\Delta t}}\left(2 + C_{j+1} + C_j - \Delta t^2\rho\right) \\ + \left(\frac{1}{2} + \frac{\Delta t}{4}\frac{\rho\Delta t}{1-\frac{\Delta t^2}{2}\rho - e^{i\omega\Delta t}}\right)\left(e^{-ik\Delta x}(1 - C_{j+1} - \lambda v) + e^{ik\Delta x}(1 - C_j + \lambda v)\right). \end{aligned}$$

Proof. To derive the numerical dispersion relation, we first have to plug everything into the last half electric push, which computes

$$\rho p^{n+1}_{j+\frac{1}{2}} = \rho p^{++}_{j+\frac{1}{2}} + \frac{\Delta t}{2}\rho E^{+}_{j+\frac{1}{2}}$$

where $\rho p^{++}_{j+\frac{1}{2}}$ denotes the fluid step with FCT. Sticking to forward Euler in time and using the calculations from chapter 5 with Lax-Friedrichs and the central second order flux, we obtain

$$F^{FCT}_j = \frac{\rho}{2}\left(v p^{-}_{j+\frac{1}{2}} + v p^{-}_{j-\frac{1}{2}}\right) - (1 - C_j)\frac{\rho}{2\lambda}\left(p^{-}_{j+\frac{1}{2}} - p^{-}_{j-\frac{1}{2}}\right),$$

where v denotes the x-component of the velocity, which we assumed to be constant. Hence we have

$$\begin{aligned} \rho p^{++}_{j+\frac{1}{2}} = \rho p^{-}_{j+\frac{1}{2}} - \rho v\frac{\lambda}{2}\left(p^{-}_{j+\frac{3}{2}} - p^{-}_{j-\frac{1}{2}}\right) + \frac{\rho}{2}\left(p^{-}_{j+\frac{3}{2}} - 2p^{-}_{j+\frac{1}{2}} + p^{-}_{j-\frac{1}{2}}\right) \\ - \frac{\rho C_{j+1}}{2}\left(p^{-}_{j+\frac{3}{2}} - p^{-}_{j+\frac{1}{2}}\right) + \frac{\rho C_j}{2}\left(p^{-}_{j+\frac{1}{2}} - p^{-}_{j-\frac{1}{2}}\right). \end{aligned}$$

To derive the complete formula for ρp^{n+1}, we have to substitute the update for the electric

[8] Note that when considering the x-component of the equation, this is no longer a valid assumption since $v_x = p_x$ in the non-relativistic case, so we have to cope with a quadratic term, which complicates things even further.

field, $E^+ = E^n - \Delta t \rho p^-$, and the electric push $\rho p^- = \rho p^n + \frac{\Delta t}{2}\rho E^n$, which yields

$$\begin{aligned}
\rho p^{n+1}_{j+\frac{1}{2}} &= \rho p^{++}_{j+\frac{1}{2}} + \frac{\Delta t}{2}\rho(E^n_{j+\frac{1}{2}} - \Delta t \rho p^-_{j+\frac{1}{2}}) \\
&= \rho p^-_{j+\frac{1}{2}} - \rho v \frac{\lambda}{2}\left(p^-_{j+\frac{3}{2}} - p^-_{j-\frac{1}{2}}\right) + \frac{\rho}{2}(p^-_{j+\frac{3}{2}} - 2p^-_{j+\frac{1}{2}} + p^-_{j-\frac{1}{2}}) \\
&\quad - \frac{\rho C_{j+1}}{2}(p^-_{j+\frac{3}{2}} - p^-_{j+\frac{1}{2}}) + \frac{\rho C_j}{2}(p^-_{j+\frac{1}{2}} - p^-_{j-\frac{1}{2}}) + \frac{\Delta t}{2}\rho(E^n_{j+\frac{1}{2}} - \Delta t \rho p^-_{j+\frac{1}{2}}) \\
&= \rho p^n_{j+\frac{1}{2}} + \frac{\Delta t}{2}\rho E^n_{j+\frac{1}{2}} - \frac{\lambda}{2}\rho v\left(p^n_{j+\frac{3}{2}} + \frac{\Delta t}{2}E^n_{j+\frac{3}{2}} - p^n_{j-\frac{1}{2}} - \frac{\Delta t}{2}E^n_{j-\frac{1}{2}}\right) \\
&\quad + \frac{\rho}{2}\left(p^n_{j+\frac{3}{2}} + \frac{\Delta t}{2}E^n_{j+\frac{3}{2}} - 2\Big(p^n_{j+\frac{1}{2}} + \frac{\Delta t}{2}E^n_{j+\frac{1}{2}}\Big) + p^n_{j-\frac{1}{2}} + \frac{\Delta t}{2}E^n_{j-\frac{1}{2}}\right) \\
&\quad - \frac{\rho C_{j+1}}{2}\left(p^n_{j+\frac{3}{2}} + \frac{\Delta t}{2}E^n_{j+\frac{3}{2}} - p^n_{j+\frac{1}{2}} - \frac{\Delta t}{2}E^n_{j+\frac{1}{2}}\right) \\
&\quad + \frac{\rho C_j}{2}\left(p^n_{j+\frac{1}{2}} + \frac{\Delta t}{2}E^n_{j+\frac{1}{2}} - p^n_{j-\frac{1}{2}} - \frac{\Delta t}{2}E^n_{j-\frac{1}{2}}\right) + \frac{\Delta t}{2}\rho\left(E^n_{j+\frac{1}{2}} - \Delta t\rho\Big(p^n_{j+\frac{1}{2}} + \frac{\Delta t}{2}E^n_{j+\frac{1}{2}}\Big)\right).
\end{aligned}$$

After some simplifications, we obtain

$$\begin{aligned}
\rho p^{n+1}_{j+\frac{1}{2}} &= \frac{1}{2}\rho p^n_{j+\frac{1}{2}}\left(C_{j+1} + C_j - \Delta t^2\rho\right) + \frac{\Delta t}{4}\rho E^n_{j+\frac{1}{2}}\left(2 + C_{j+1} + C_j - \Delta t^2\rho\right) \\
&\quad + \frac{1}{2}\rho p^n_{j+\frac{3}{2}}(1 - C_{j+1} - \lambda v) + \frac{1}{2}\rho p^n_{j-\frac{1}{2}}(1 - C_j + \lambda v) \\
&\quad + \frac{\Delta t}{4}\rho E^n_{j+\frac{3}{2}}(1 - C_{j+1} - \lambda v) + \frac{\Delta t}{4}\rho E^n_{j-\frac{1}{2}}(1 - C_j + \lambda v).
\end{aligned}$$

Inserting $p = p_0 e^{i\omega t - ikx}$ and $E = E_0 e^{i\omega t - ikx}$ now yields

$$\begin{aligned}
p_0 e^{i\omega\Delta t} &= \frac{1}{2}p_0\left(C_{j+1} + C_j - \Delta t^2\rho\right) + \frac{\Delta t}{4}E_0\left(2 + C_{j+1} + C_j - \Delta t^2\rho\right) \\
&\quad + \left(\frac{1}{2}p_0 + \frac{\Delta t}{4}E_0\right)\left(e^{-ik\Delta x}(1 - C_{j+1} - \lambda v) + e^{ik\Delta x}(1 - C_j + \lambda v)\right).
\end{aligned} \tag{6.8}$$

To obtain the numerical dispersion relation, we also have to consider the electric update

$$E^{n+1}_{j+\frac{1}{2}} = E^+_{j+\frac{1}{2}} = E^n_{j+\frac{1}{2}} - \Delta t\rho p^-_{j+\frac{1}{2}} = \left(1 - \frac{\Delta t^2}{2}\rho\right)E^n_{j+\frac{1}{2}} - \Delta t\rho p^n_{j+\frac{1}{2}}.$$

Inserting the harmonic wave solution yields

$$E_0 e^{i\omega\Delta t} = \left(1 - \frac{\Delta t^2}{2}\rho\right)E_0 - \Delta t\rho p_0,$$

which we solve for E_0. Inserting this into (6.8) yields

$$\begin{aligned}
e^{i\omega\Delta t} &= \frac{1}{2}\left(C_{j+1} + C_j - \Delta t^2\rho\right) + \frac{\Delta t}{4}\frac{\rho\Delta t}{1 - \frac{\Delta t^2}{2}\rho - e^{i\omega\Delta t}}\left(2 + C_{j+1} + C_j - \Delta t^2\rho\right) \\
&\quad + \left(\frac{1}{2} + \frac{\Delta t}{4}\frac{\rho\Delta t}{1 - \frac{\Delta t^2}{2}\rho - e^{i\omega\Delta t}}\right)\left(e^{-ik\Delta x}(1 - C_{j+1} - \lambda v) + e^{ik\Delta x}(1 - C_j + \lambda v)\right).
\end{aligned}$$

□

If we multiply this equation by the denominator $1 - \frac{\Delta t^2}{2}\rho - e^{i\omega\Delta t}$, we obtain an expression in $\eta := e^{i\omega\Delta t}$. For fixed Δt, Δx, ρ, k and C, we can interpret the expression as a problem of finding the roots of a polynomial of degree two in η. For $C \equiv 1$, which corresponds to using only the second order flux in the FCT scheme, we obtain

$$\eta_{1,2} = -\frac{\alpha}{2} \pm \sqrt{\frac{\alpha^2}{4} - 1 - \beta}$$

with

$$\alpha = \Delta t^2 \rho - 2 + \beta, \qquad \beta = i\lambda v \sin(k\Delta x).$$

The high order flux is the simplest case. Also, it covers only one choice of fluxes in the FCT scheme. Therefore and because of all the simplifying assumptions we made in the beginning, we are not in the position to make any general statements about the numerical dispersion of the YeeFCT algorithm. The following numerical examples illustrate the practical performance.

6.3 Numerical Experiments

Having stated the algorithm to our governing equations, we can study some numerical experiments. We want to look at a vacuum-plasma transition, which is a complex example that is hard to handle. In vacuum, the density is zero, which poses the first major problem for most numerical schemes. Somewhere inside this vacuum there will be an area where the plasma is placed and the density is set to one (or some other value). The transition from vacuum to plasma can be modeled linearly or even smoothly by applying a Gaussian at the edges. A jump is also possible numerically, but in reality there will always be some transition area, so we model it with Gaussian edges. All velocities — and therefore also momenta — are set to zero initially.

Now let us go on to the laser. In Maxwell's equations, E_y is interleaved with B_z, while the other couple of equations contains E_z and B_y. E_x is only changed through currents in the one-dimensional case, so we set it to zero initially.

The initial values of the other field components are sine and cosine waves enclosed by a Gaussian, i.e.,

$$\begin{aligned}
E_y^0 &= e^{-\frac{(x-x_m)^2}{\sigma^2}} \cos(k_0 x) \\
B_z^0 &= e^{-\frac{(x-x_m)^2}{\sigma^2}} \cos(k_0 x) \\
E_z^0 &= -e^{-\frac{(x-x_m)^2}{\sigma^2}} \sin(-k_0 x) \\
B_y^0 &= e^{-\frac{(x-x_m)^2}{\sigma^2}} \sin(-k_0 x)
\end{aligned}$$

where $k_0 = \omega_0$ in vacuum.[9] The quotient $\frac{1}{\omega_0^2} = \frac{\rho}{\rho_c} =: \tilde{\rho}$ indicates how close we are to the critical plasma density ρ_c.

If the plasma is over-dense, i.e., $\tilde{\rho} > 1$, the pulse is reflected by the plasma. For low densities, the laser pulse will move through the plasma without exciting it or changing its own shape. We are mainly interested in cases of under-dense plasmas, which do cause visible laser-plasma interaction.

The initial data for $\omega_0 = 3$ ($\tilde{\rho} \approx 0.11$) and Gaussian plasma edges can be seen in figure 6.2.

[9] The relation is $ck_0 = \omega_0$, but in the units used here, $c = 1$.

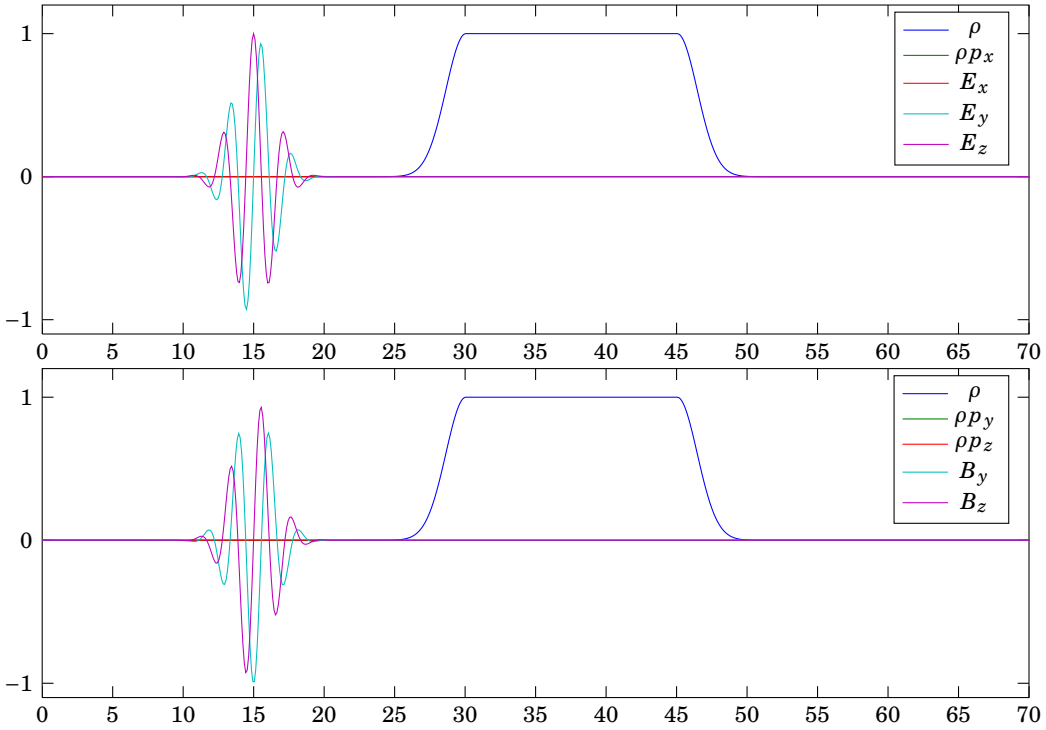

Figure 6.2: *Initial configuration for* $\omega_0 = 3$

To apply the above YeeFCT algorithm, we have to calculate also the relativistic factor

$$\gamma = \sqrt{1 + p_x^2 + p_y^2 + p_z^2}$$

and the velocities $\boldsymbol{v} = \frac{\boldsymbol{p}}{\gamma}$. For the division of $\rho\boldsymbol{p}$ by ρ, we assume the momenta be zero in (or near) vacuum.

The grid has constant step size $\Delta x = 0.1$. After 250 time steps of $\Delta t = \frac{\Delta x}{2}$, the pulse has reached the ramp and we can see the first effects in figure 6.3. The plasma is quite dense, so there is a slight deformation of the pulse and the momenta are starting to grow.

If we look a little further, after 500 time steps, the pulse has fully immerged into the plasma and has excited it. The movement is shown in figure 6.4.

In figure 6.5 we see how the pulse is just leaving the plasma. The momenta in y- and z-direction have gone down again while the plasma is still in motion due to the momentum ρp_x.

As a further comparison, let us look at different values for ω and compare them at time $t = 50T_0$, which corresponds to 500 steps in the above example. In all figures, we scale the density to 1 for easier comparison.

Note that for $\omega = 10$, we only have 1% critical density. The pulse goes through the plasma almost unharmed and the plasma does not move much, either (see figure 6.6). $\omega = 5$ shows a little more interaction (figure 6.7), while for $\omega = 2$ (figure 6.8) or even $\omega = 1.2$ (figure 6.9), we are getting closer to critical density and the pulse deforms completely and is even partially reflected.

The simulations are remarkably good and show the expected behavior for the various parameters. Densities stay non-negative and the whole procedure remains stable during the vacuum-plasma transition, which is usually the point of failure in this kind of simulation — especially for small ω, which corresponds to a high density.

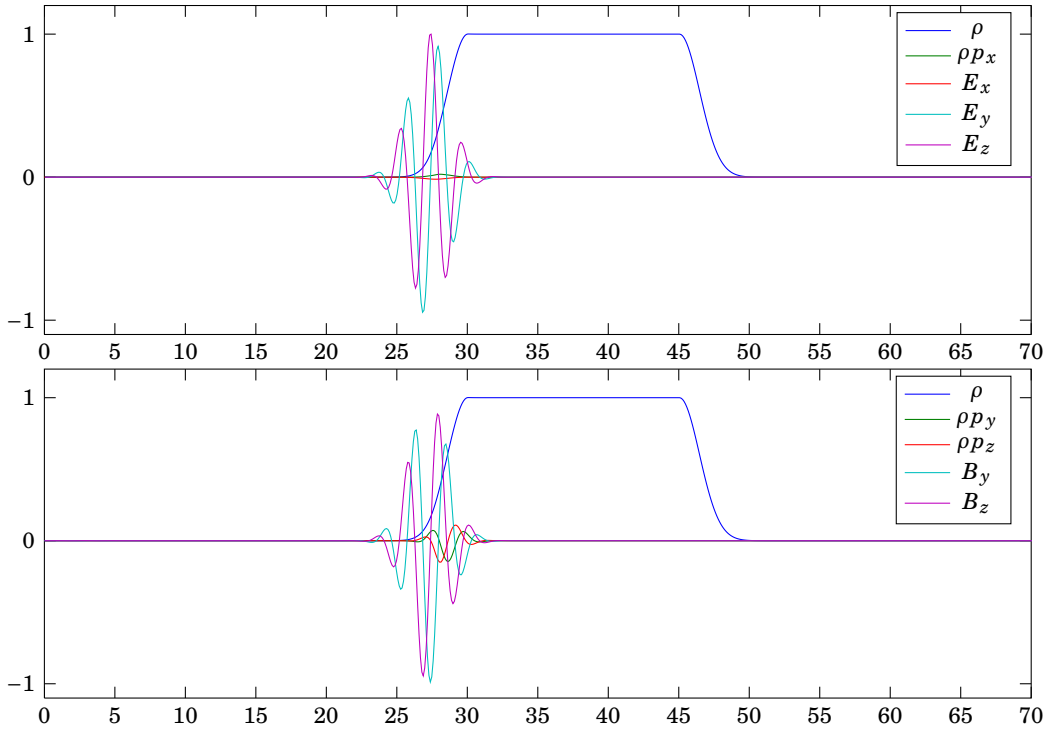

Figure 6.3: *Simulation after 250 steps*

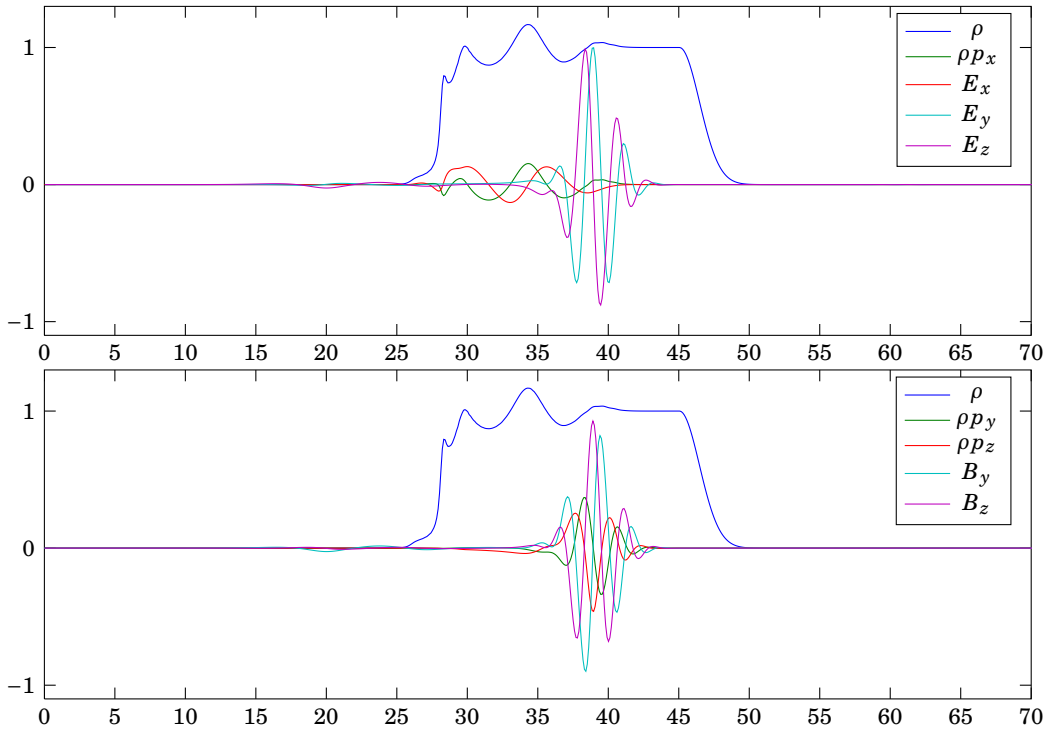

Figure 6.4: *Simulation after 500 steps*

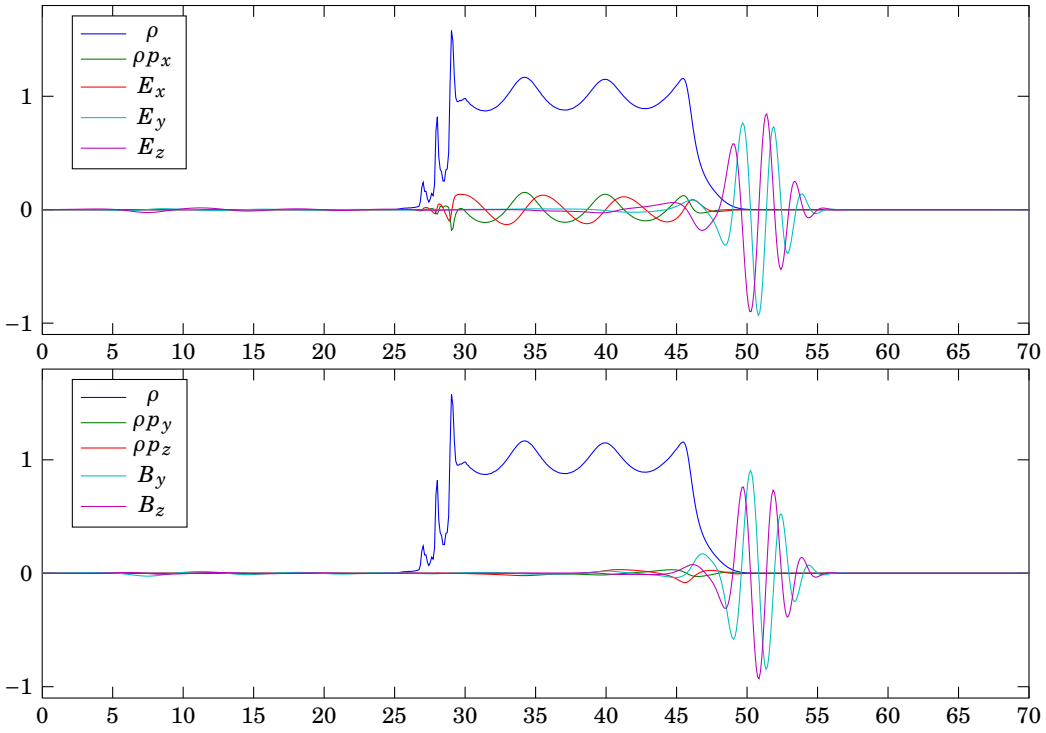

Figure 6.5: *Simulation after 750 steps*

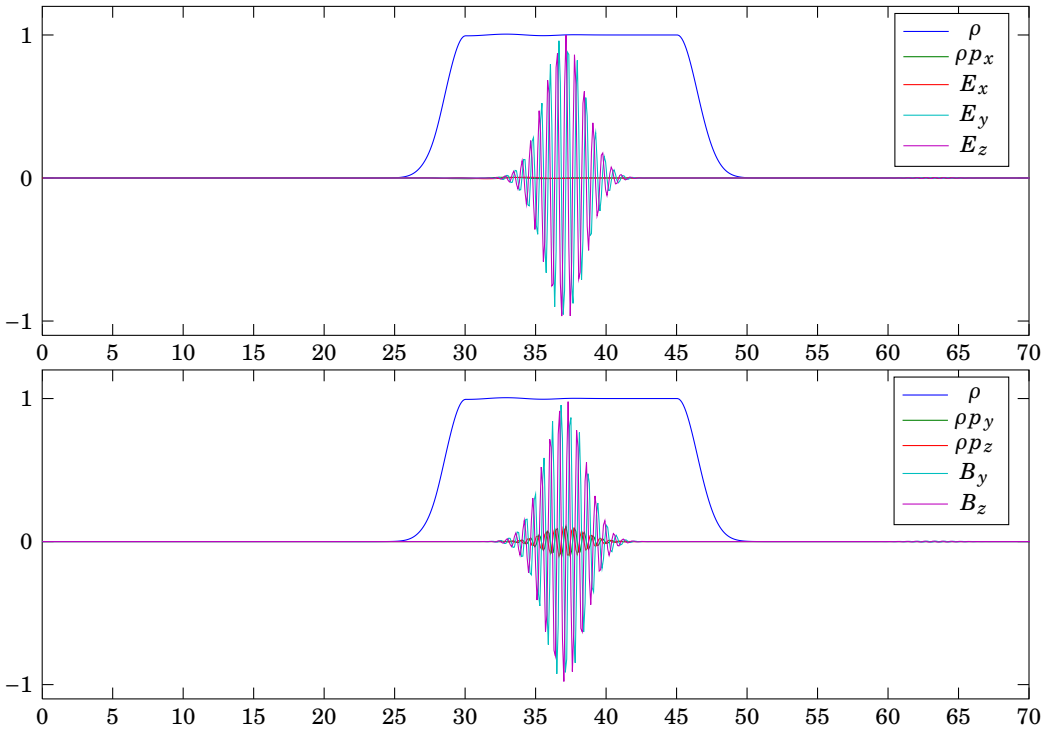

Figure 6.6: *Simulation with* $\omega = 10$ *after 500 steps*

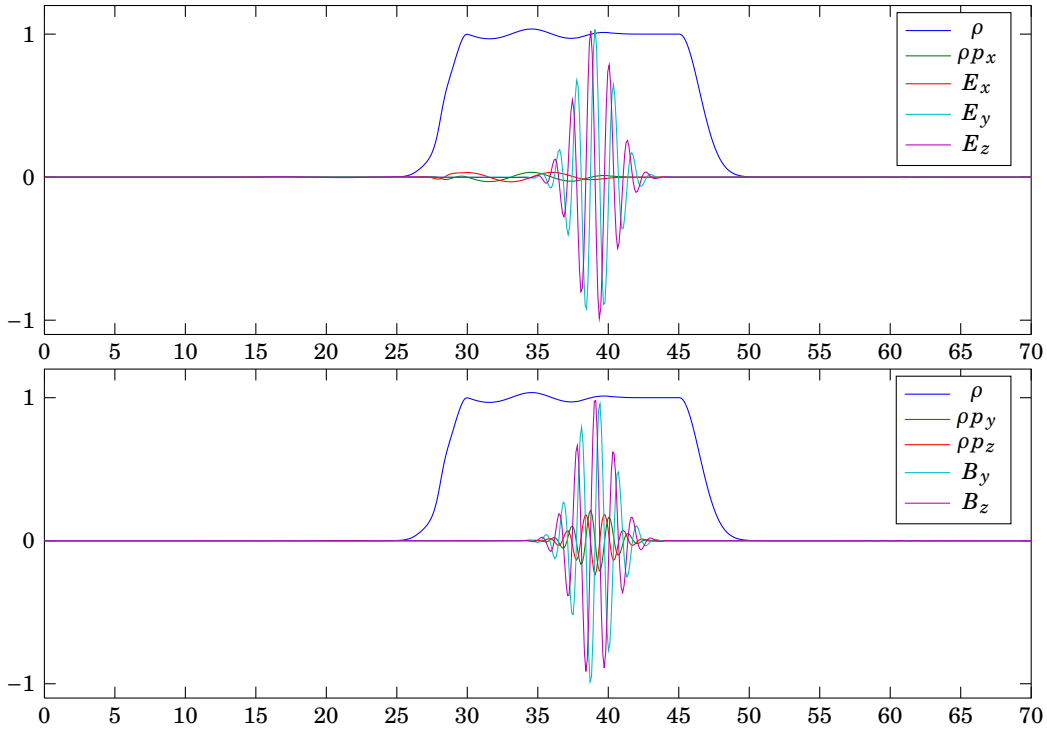

Figure 6.7: *Simulation with* $\omega = 5$ *after 500 steps*

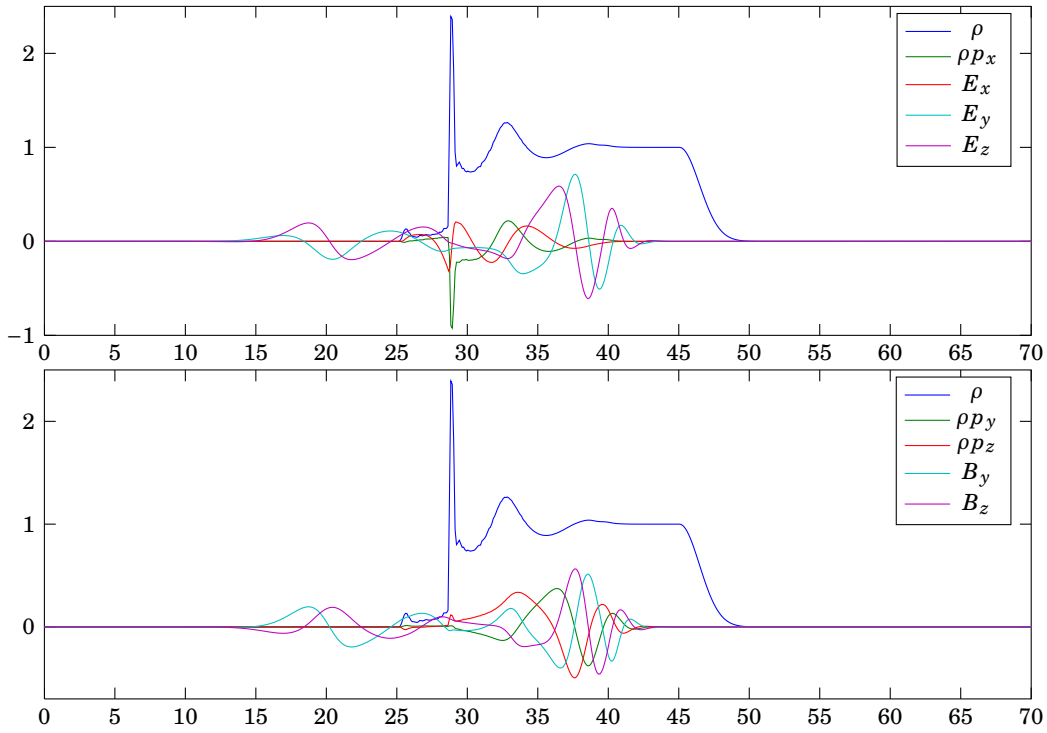

Figure 6.8: *Simulation with* $\omega = 2$ *after 500 steps*

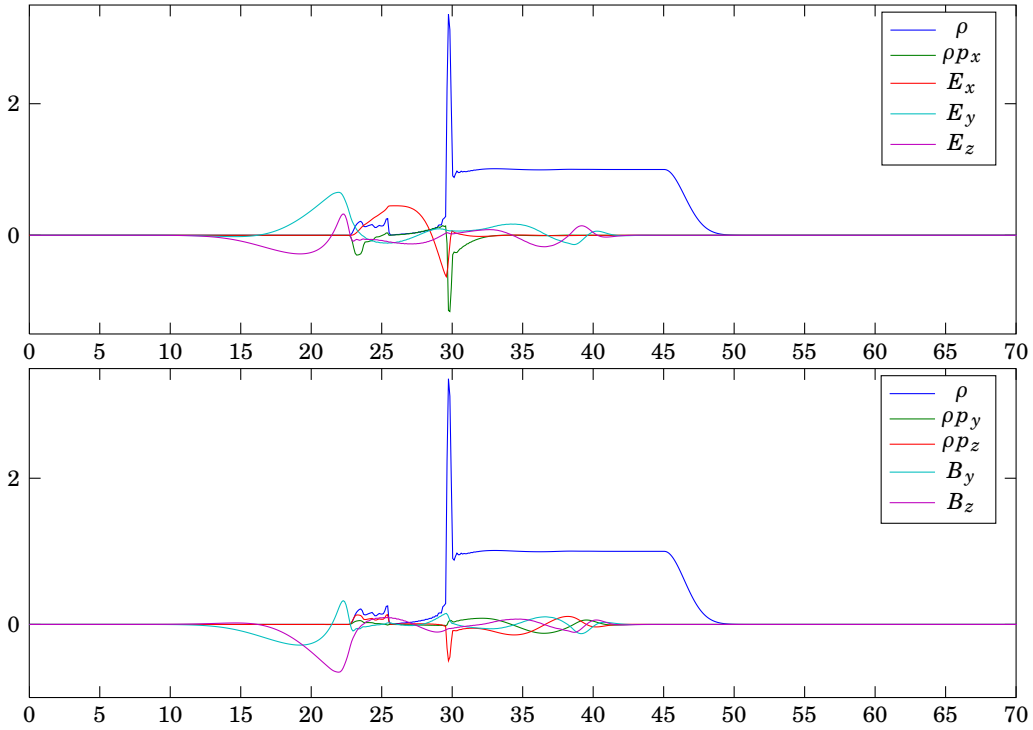

Figure 6.9: *Simulation with $\omega = 1.2$ after 500 steps*

Comparison with Existing Codes

In the following, we compare the results of our code to those of a well-established Vlasov-simulation (see for example [CK76], [SRBG99], [BGB+99] or [GHB03]). Most of these use laser units, i.e., the quantities are measured in laser rather than plasma wavelengths and frequencies. So to be able to compare, we have to rescale the results obtained with the plasma units equations according to section 4.3 by dividing any spatial values by the laser wavelength $\lambda_0 = \frac{2\pi}{k_0}$ and the times by $T_0 = \frac{2\pi}{\omega_0}$. For this test, we use a circularly polarized laser pulse with

$$
\begin{aligned}
B_y^0 &= \frac{k_0}{2\sqrt{2}} e^{-\left(\frac{x-x_m}{\sigma}\right)^2} \sin(-k_0 x) \\
B_z^0 &= \frac{k_0}{2\sqrt{2}} e^{-\left(\frac{x-x_m}{\sigma}\right)^2} \cos(-k_0 x) \\
E_y^0 &= \frac{\omega_0}{2\sqrt{2}} e^{-\left(\frac{x-x_m}{\sigma}\right)^2} \cos(-k_0 x) \\
E_z^0 &= -\frac{\omega_0}{2\sqrt{2}} e^{-\left(\frac{x-x_m}{\sigma}\right)^2} \sin(-k_0 x)
\end{aligned}
$$

so compared to the tests above, the initial values are slightly different — the amplitudes are scaled here. We compare our results to those of a Vlasov simulation carried out by Dr. Götz Lehmann of Heinrich-Heine-Universität Düsseldorf.

Test Case 1: Low Density

We want to look at the case $\tilde{\rho} = 0.04$, which means $\omega_0 = 5$. We choose σ such that the full width at half maximum (FWHM) of the Gaussian is 2π in plasma units. Note that the

Vlasov simulation took a mesh size of $\Delta x = 0.005$, while we used $\Delta x = 0.05$ for YeeFCT — ten times that of the Vlasov code. The simulation worked fine even for $\Delta x = 0.1$, but then the emerging peaks are not resolved well enough. For the time step Δt, we have to make sure it satisfies the CFL condition. Since we normalized our velocities such that the speed of light is one in our units, the absolute value of any velocity is bounded by one and so $\lambda = 1$ satisfies the CFL condition of any method we considered.

Let us take a look at the numerical results at a few points in time. Figures 6.10 through 6.15 show the density and some of the field components from both simulations, which compare really well. The spatial discretization is fine enough to resolve all peaks, but still ten times bigger than what the Vlasov code had to use.

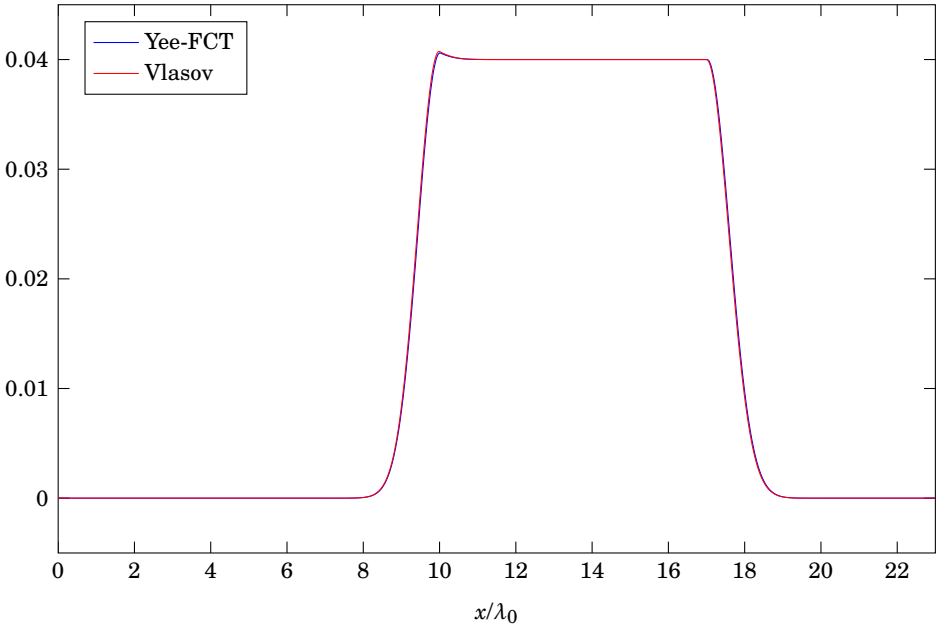

Figure 6.10: *Density with YeeFCT and the reference Vlasov simulation at* $t \approx 5T_0$

Since we are dealing with conservation laws, YeeFCT should conserve the total mass. It turns out that this is the case, as can be seen in figure 6.16. The errors are very close to machine precision, so using a conservative method really paid off.

Another important quantity is energy. The conservation of mass and energy is a fundamental principle in physics. Since YeeFCT does not incorporate any energy conserving techniques — except for the symmetric splitting —, we cannot hope for the conservation of energy to be as neat as the conservation of mass. The total energy W within a given volume V is

$$W = \frac{1}{2}\int_V \boldsymbol{B}^2 + \boldsymbol{E}^2 + 2\rho(\gamma - 1)\,dV.$$

What we observe is that once the pulse enters the plasma, the error in energy increases. But even then, the relative error in this simulation does not grow beyond 10%, see figure 6.17. This is not overwhelming, but better than we have hoped for. Of course, the results are a lot better if we refine our grid. Figure 6.18 shows the relative error in energy at the final time for different Δx. The error decreases with decreasing mesh size roughly like $\mathcal{O}(\Delta x)$.

If we take another moment of the Vlasov equation and do not neglect pressure and temperature like in section 4.1, we end up with an equation for the energy (and a modified

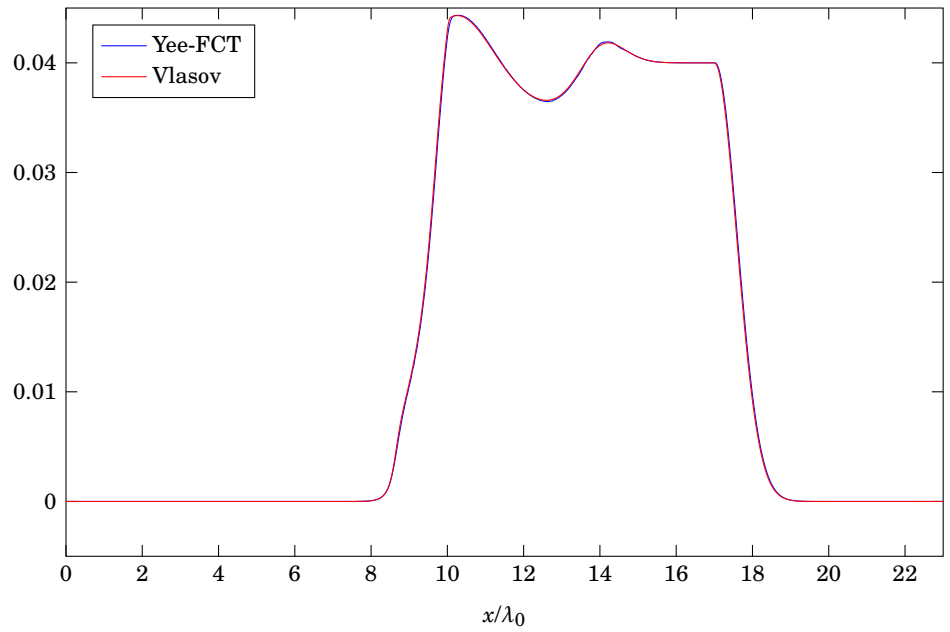

Figure 6.11: *Density with YeeFCT and the reference Vlasov simulation at* $t \approx 10T_0$

equation for the momentum density). That way we could have much better conservation of energy. FCT would be perfectly suited to numerically solve this equation, as well. On the other hand, we would have to deal with yet another equation and more variables. Since energy conservation already works sufficiently well without the additional equation, it might be more sensible to stick to what we have. This is always a decision between accuracy and efficiency.

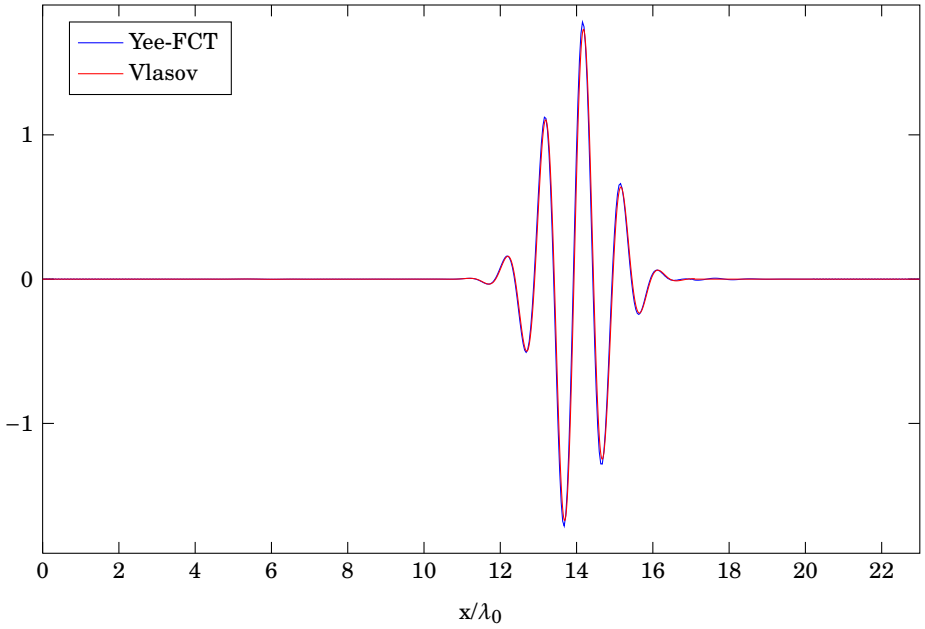

Figure 6.12: *Electric field component E_y with YeeFCT and Vlasov simulation at $t \approx 10T_0$*

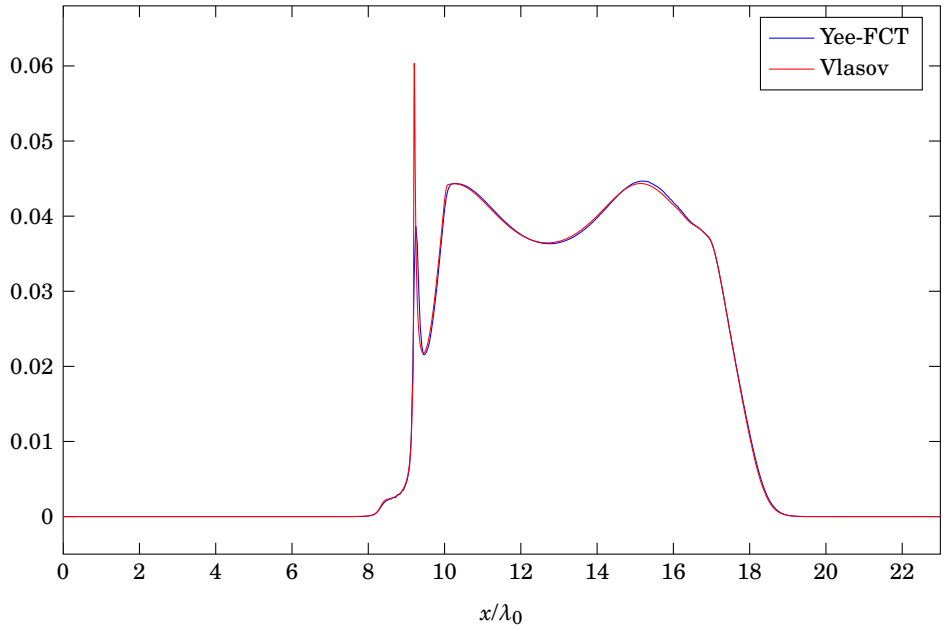

Figure 6.13: *Density with YeeFCT and the reference Vlasov simulation at $t \approx 15T_0$*

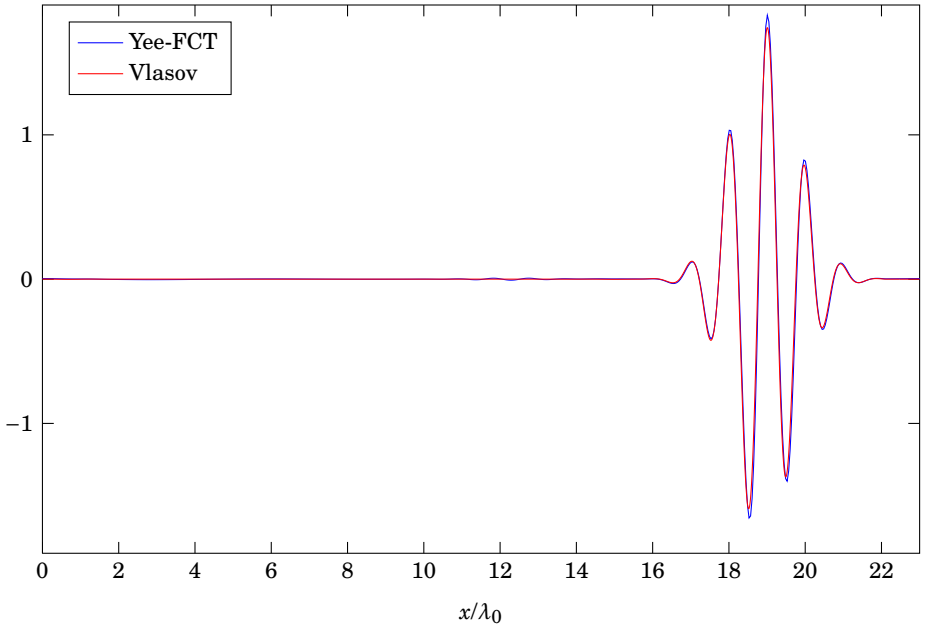

Figure 6.14: *Magnetic field component B_y with YeeFCT and Vlasov simulation at $t \approx 15T_0$*

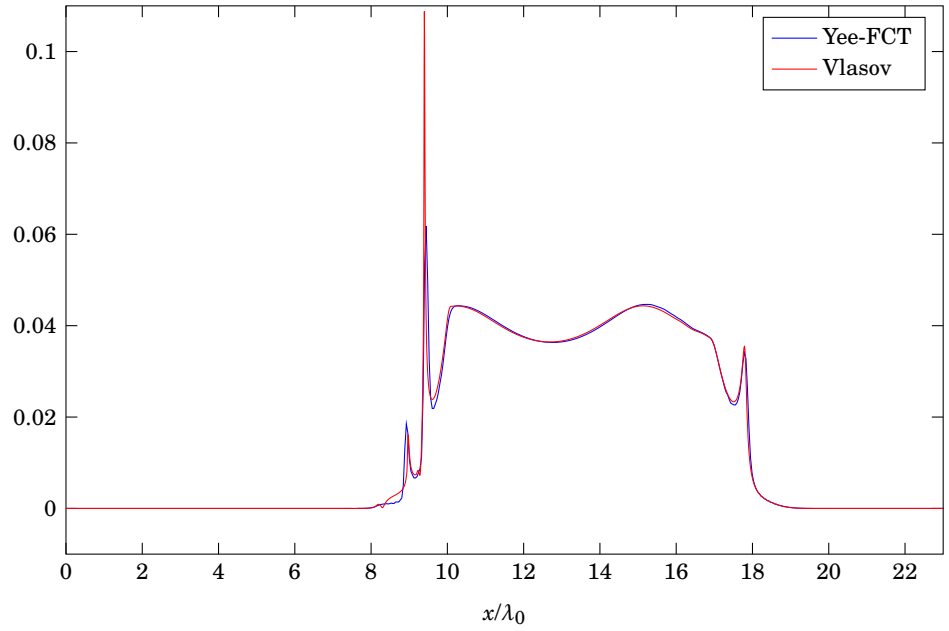

Figure 6.15: *Density with YeeFCT and the reference Vlasov simulation at $t \approx 20T_0$*

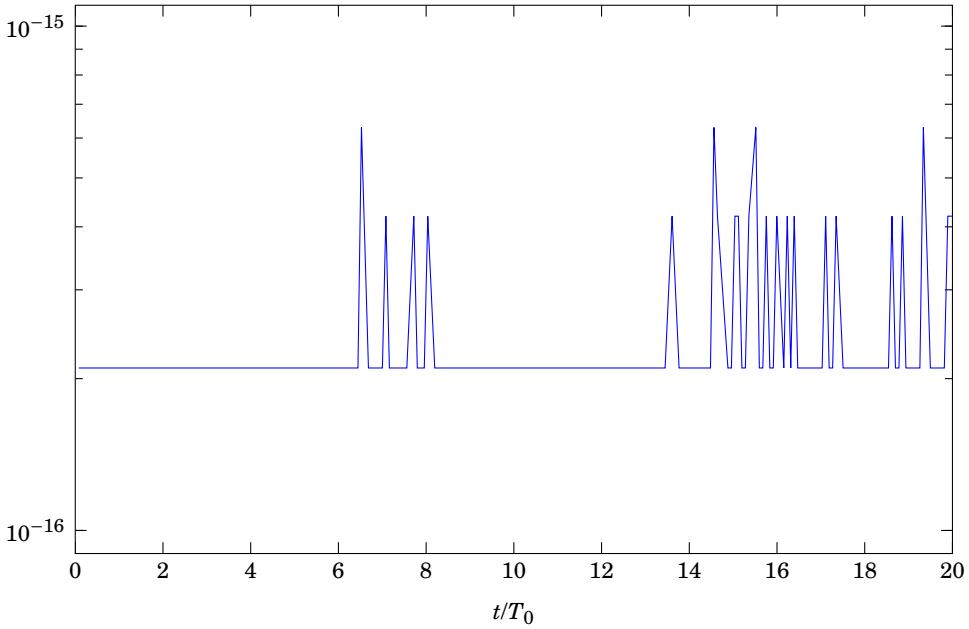

Figure 6.16: *Relative error in total mass*

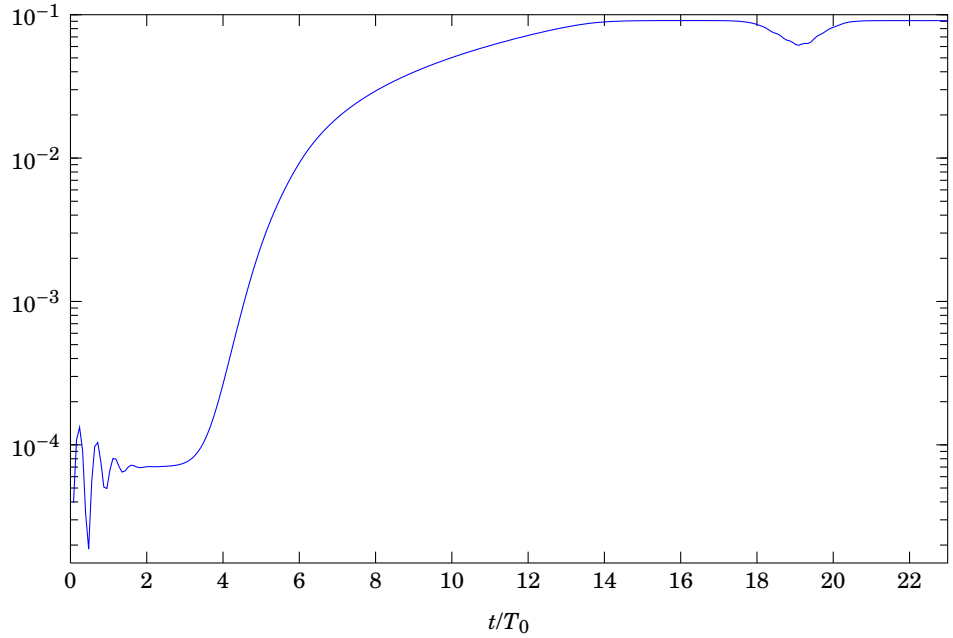

Figure 6.17: *Relative error in energy*

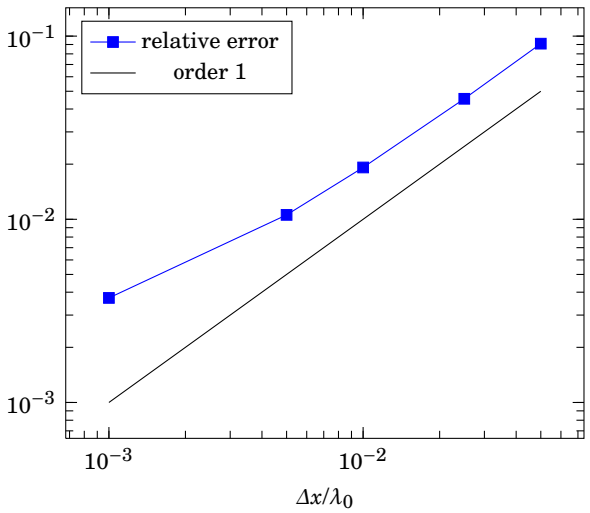

Figure 6.18: *Order plot of the relative error in energy*

Test Case 2: High Density

For the second test, we use the same setting, except that we increase the density to $\tilde{\rho} = 0.6$ — more than half the critical density. The Vlasov code used $\Delta x = 0.0025$, so we choose $\Delta x = 0.025$ — again ten times the Vlasov mesh size.

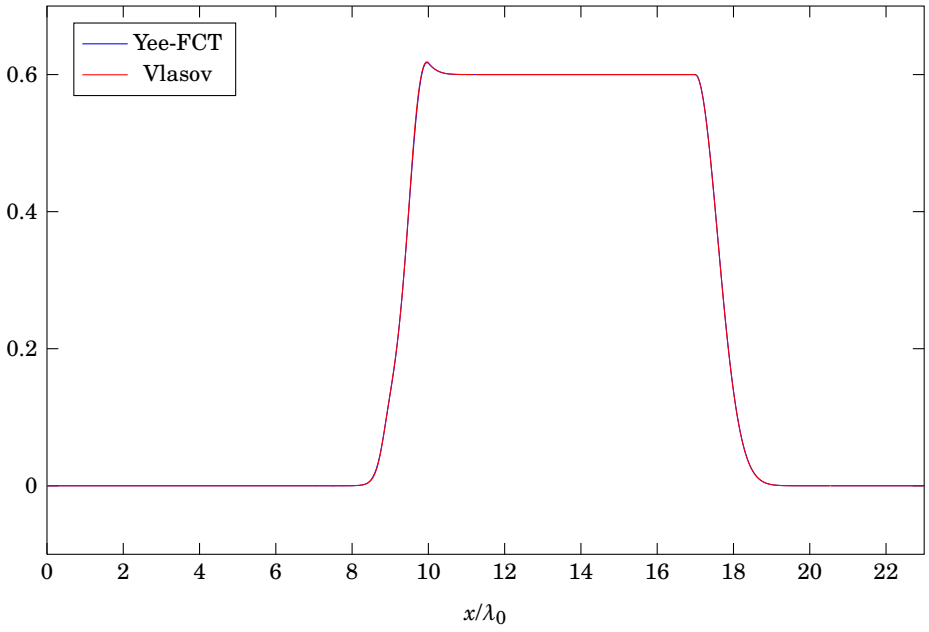

Figure 6.19: *Density with YeeFCT and the reference Vlasov simulation at* $t \approx 5T_0$

Figures 6.19 through 6.24 show the same comparisons as in the first test. The plasma response is much more violent and the simulations do not agree in all details anymore. The qualitative behavior, however, is captured really well again by YeeFCT.

The conservation of mass is also remarkably strong. The relative error of total mass is only about ten times machine precision.

The conservation of energy is similar to the previous example (see figure 6.26), but altogether, the error is bigger now. Figure 6.27 shows an order plot, which again shows order one.

The snapshots of the two simulations show how well the YeeFCT combination works. We can make much larger steps than the Vlasov code and still obtain very accurate results. Only at peaks, the resolution is, of course, not quite as good. Also the comparison of computation time is impressive: Since the Vlasov code is basically two-dimensional, it takes about 80 times longer than YeeFCT!

In summary, we conclude that YeeFCT does have a little weakness in energy conservation, but is more than competitive in all of accuracy, computational cost and time. The qualitative behavior of both the laser and the plasma is reproduced really well — even for high densities.

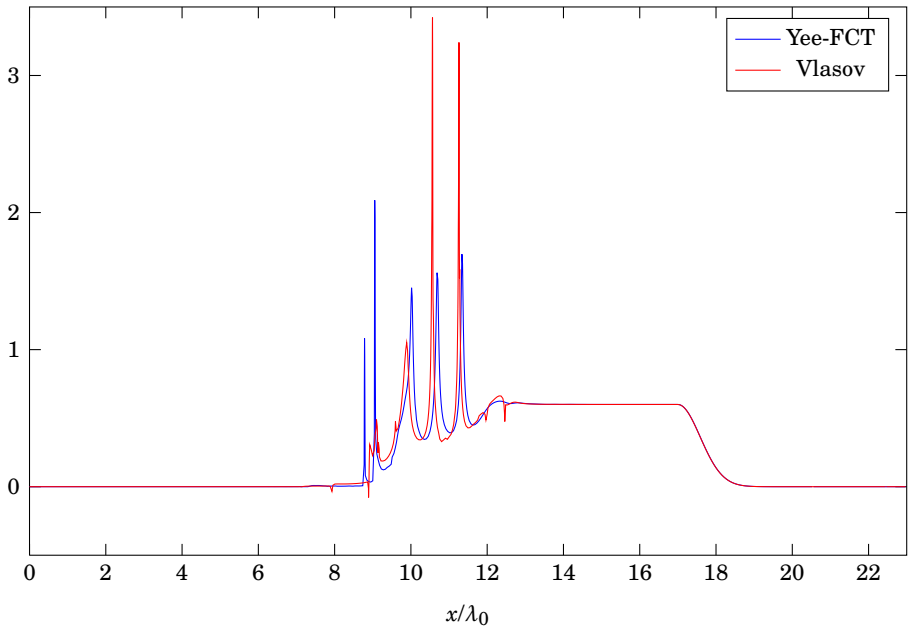

Figure 6.20: *Density with YeeFCT and the reference Vlasov simulation at* $t \approx 10T_0$

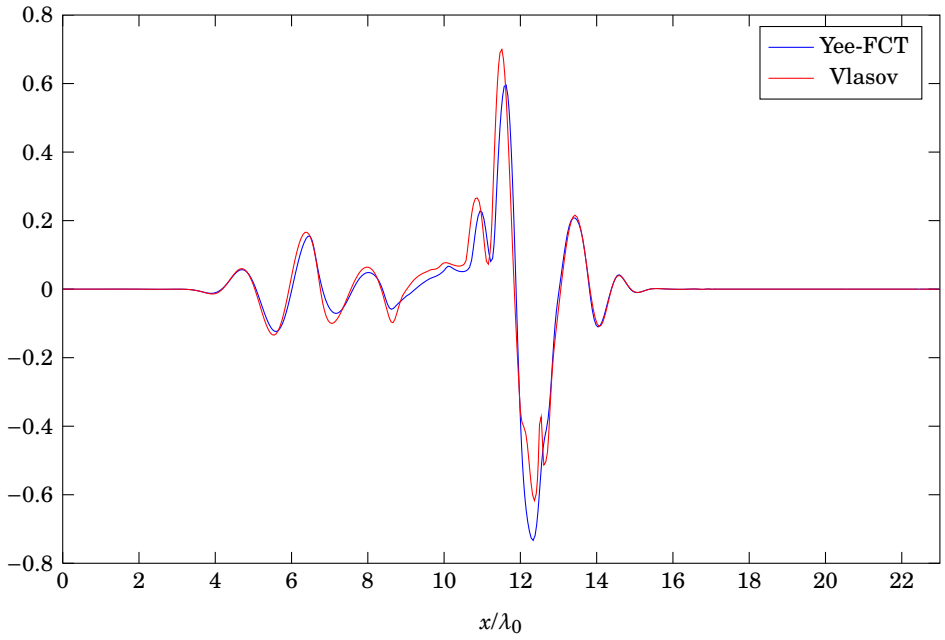

Figure 6.21: *Comparison of the magnetic field component* E_y *with YeeFCT and the reference Vlasov simulation at* $t \approx 10T_0$

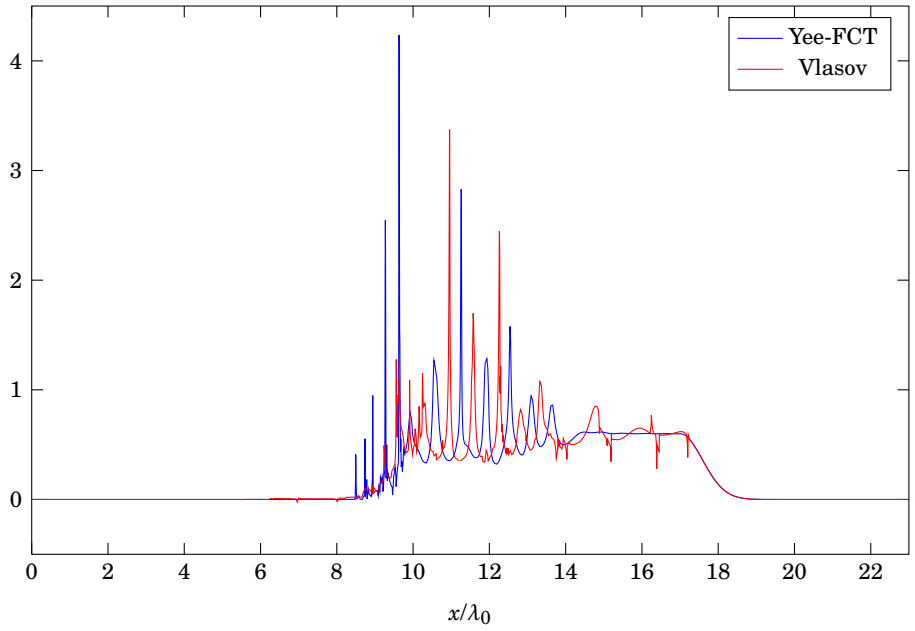

Figure 6.22: *Density with YeeFCT and the reference Vlasov simulation at* $t \approx 15T_0$

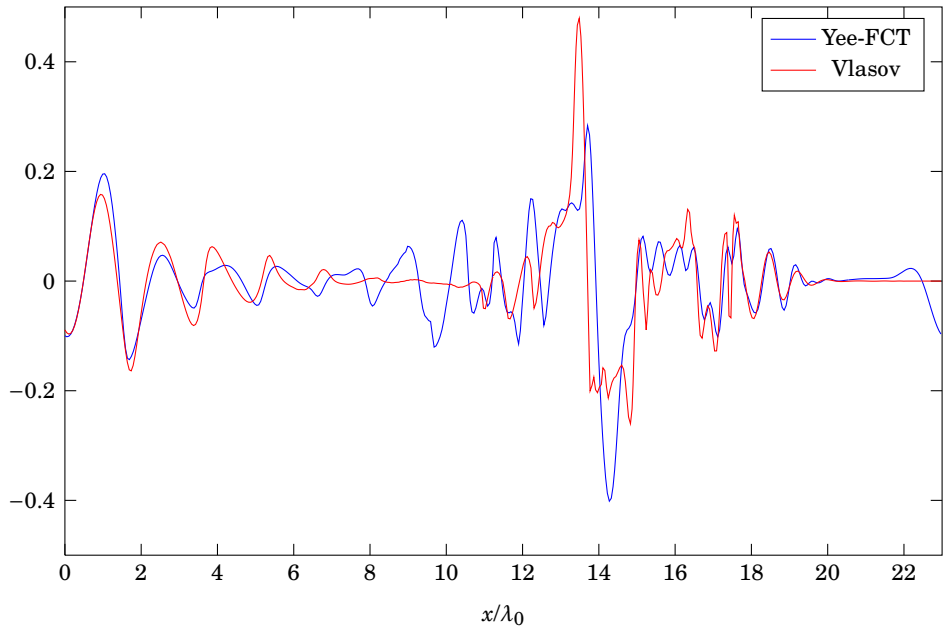

Figure 6.23: *Comparison of the magnetic field component* B_y *with YeeFCT and the reference Vlasov simulation at* $t \approx 15T_0$

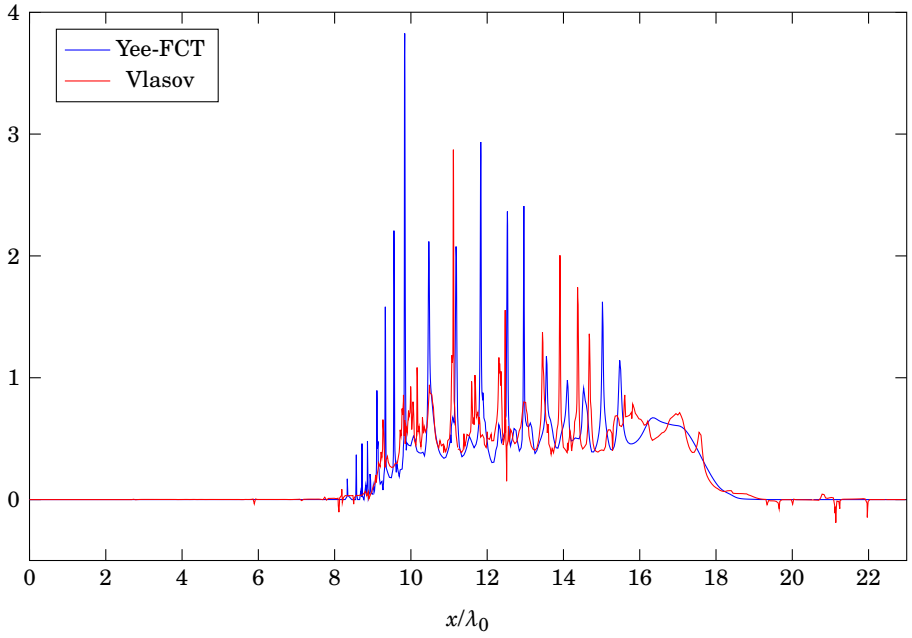

Figure 6.24: *Density with YeeFCT and the reference Vlasov simulation at* $t \approx 20T_0$

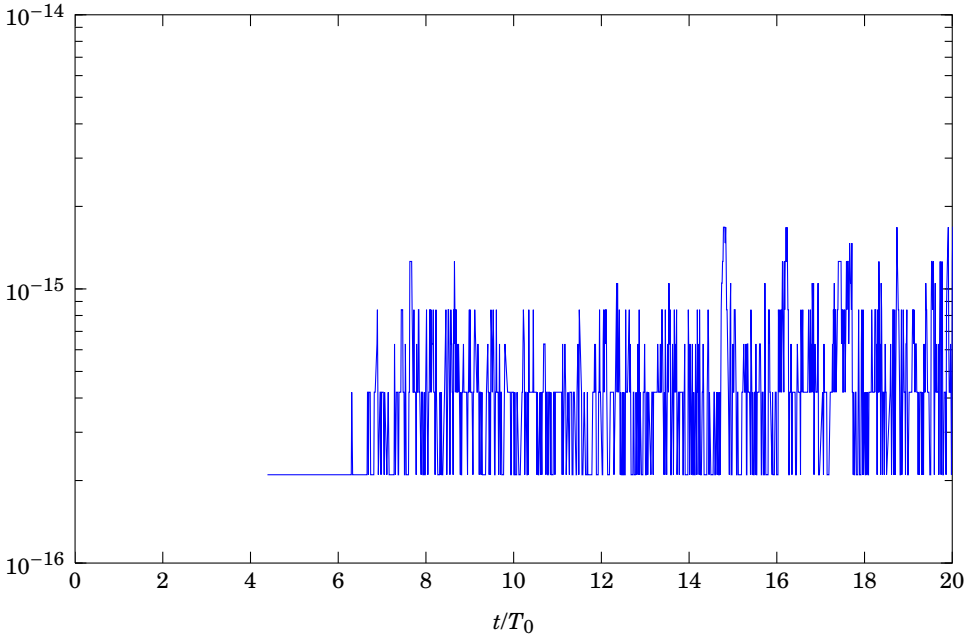

Figure 6.25: *Relative error in total mass*

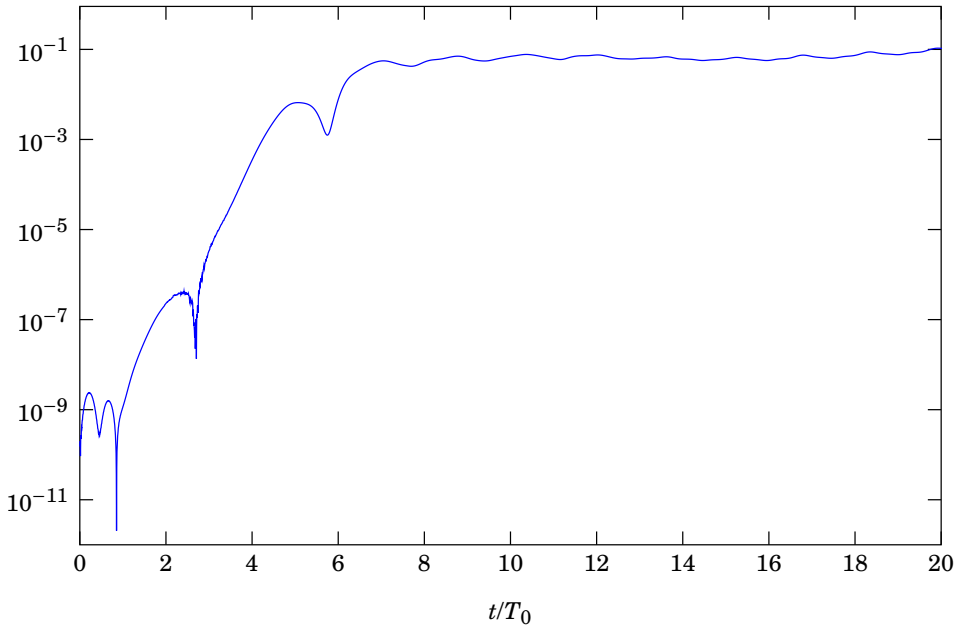

Figure 6.26: *Relative error in energy*

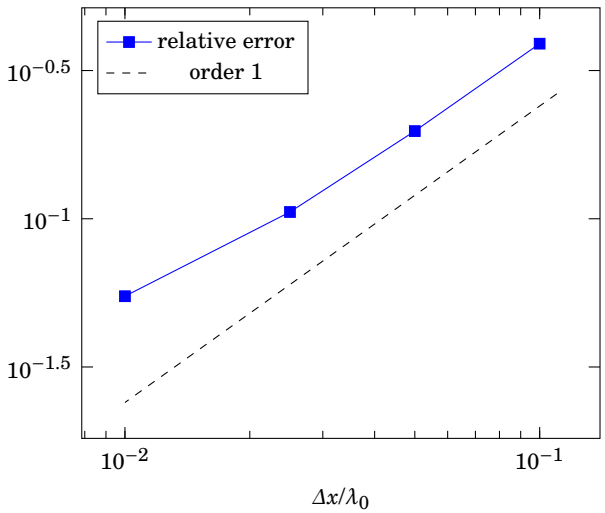

Figure 6.27: *Order plot of the relative error in energy*

7 Relativistic Laser-Plasma Interaction in 2D

After we have taken the first step by successfully applying the presented YeeFCT scheme to the one-dimensional vacuum-plasma transition of a laser pulse, we now want to achieve another important goal — the two-dimensional case. A lot of things that work in 1D are not transferable to the multidimensional case. Recall from chapter 5 that we had to introduce a further limiting step because applying FCT to both coordinate directions simply was not enough. But this problem has been solved successfully. The Yee scheme is well-known to work just fine in the multidimensional case. That leaves the Boris push, but that was constructed for the fully three-dimensional case and adjusted from there to 1D. Now we adjust it to 2D. So we are all set to take the step into the second dimension.

7.1 Equations

The setting is the same as in the previous chapter. We still consider a vacuum-plasma transition of a laser pulse. Think of the domain $\Omega \subset \mathbb{R}^2$ as some rectangle. This is easy to implement together with periodic boundary conditions, so we do not have to deal with other difficulties than the ones already at hand.

In the two-dimensional case, only derivatives with respect to z are set to zero. So now, Maxwell's equations read

$$\begin{aligned}
\frac{\partial E_x}{\partial t} &= \frac{\partial B_z}{\partial y} - j_x \\
\frac{\partial E_y}{\partial t} &= -\frac{\partial B_z}{\partial x} - j_y \\
\frac{\partial E_z}{\partial t} &= \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} - j_z \\
\frac{\partial B_x}{\partial x} &= -\frac{\partial E_z}{\partial y} \\
\frac{\partial B_y}{\partial t} &= \frac{\partial E_z}{\partial x} \\
\frac{\partial B_z}{\partial t} &= \frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x}
\end{aligned}$$

and the two-dimensional plasma equations are

$$\begin{aligned}
\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v_x) + \frac{\partial}{\partial y}(\rho v_y) &= 0 \\
\frac{\partial (\rho \boldsymbol{p})}{\partial t} + \frac{\partial}{\partial x}(\rho \boldsymbol{p} v_x) + \frac{\partial}{\partial y}(\rho \boldsymbol{p} v_y) &= \rho(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}).
\end{aligned}$$

7.2 The YeeFCT Algorithm

There is not much change in the structure of the algorithm compared to the previously discussed one-dimensional case. We take the above equations and split them in the same way as before,

$$\phi^{[1]}_{\Delta t/2} \circ \phi^{[2]}_{\Delta t/2} \circ \phi^{[3]}_{\Delta t} \circ \phi^{[2]}_{\Delta t/2} \circ \phi^{[1]}_{\Delta t/2}$$

where red stands for Maxwell's equations without current $\boldsymbol{j}$,

$$\frac{\partial \boldsymbol{E}}{\partial t} = \nabla \times \boldsymbol{B}$$
$$\frac{\partial \boldsymbol{B}}{\partial t} = -\nabla \times \boldsymbol{E},$$

then green for the Boris push in

$$\frac{\partial(\rho \boldsymbol{p})}{\partial t} = \rho(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}),$$

and finally blue for the fluid equations

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v_x) + \frac{\partial}{\partial y}(\rho v_y) = 0$$
$$\frac{\partial(\rho \boldsymbol{p})}{\partial t} + \frac{\partial}{\partial x}(\rho \boldsymbol{p} v_x) + \frac{\partial}{\partial y}(\rho \boldsymbol{p} v_y) = \boldsymbol{0}$$
$$\frac{\partial \boldsymbol{E}}{\partial t} = -\boldsymbol{j}.$$

We solve these equations on a rectangular finite difference grid. We choose constant mesh sizes Δx and Δy such that $\overline{\Omega} = \bigcup_{i,j} D_{ij}$ where $D_{ij} = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$ and $\boldsymbol{x}_{ij} = (i\Delta x, j\Delta y)^T$.

If we project the three-dimensional Yee-cell onto $\mathbb{R}^2$, we see the placement of the field components in figure 7.1. The density, velocities and momenta are again assumed at the middle of each cell, i.e., at $\boldsymbol{x}_{i+\frac{1}{2},j+\frac{1}{2}}$.

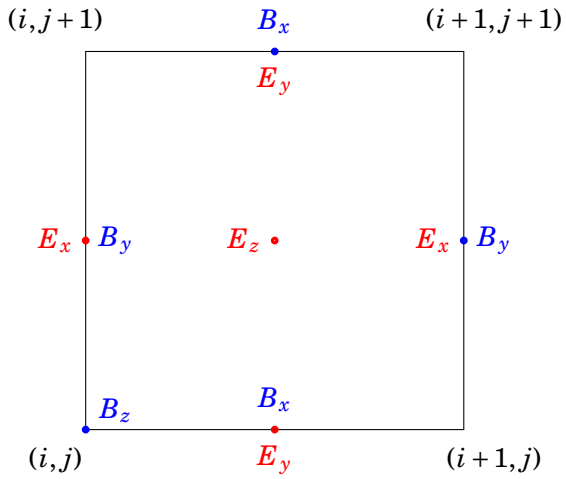

Figure 7.1: *Schematic view of a 2D Yee cell*

The numerical procedure is the same as in the one-dimensional case. We only have to adjust the equations. First, a quarter time step with the magnetic fields is performed,

$$B^{n+\frac{1}{4}}_{x;i+\frac{1}{2},j} = B^{n}_{x;i+\frac{1}{2},j} - \frac{\Delta t}{4\Delta y}\left(E^{n}_{z;i+\frac{1}{2},j+\frac{1}{2}} - E^{n}_{z;i+\frac{1}{2},j-\frac{1}{2}}\right)$$
$$B^{n+\frac{1}{4}}_{y;i,j+\frac{1}{2}} = B^{n}_{y;i,j+\frac{1}{2}} + \frac{\Delta t}{4\Delta x}\left(E^{n}_{z;i+\frac{1}{2},j+\frac{1}{2}} - E^{n}_{z;i-\frac{1}{2},j+\frac{1}{2}}\right)$$
$$B^{n+\frac{1}{4}}_{z;i,j} = B^{n}_{z;i,j} + \frac{\Delta t}{4\Delta y}\left(E^{n}_{x;i,j+\frac{1}{2}} - E^{n}_{x;i,j-\frac{1}{2}}\right) - \frac{\Delta t}{4\Delta x}\left(E^{n}_{y;i+\frac{1}{2},j} - E^{n}_{y;i-\frac{1}{2},j}\right)$$

then we use those magnetic updates to compute a half time step with the electric fields

$$
\begin{aligned}
E^{n+\frac{1}{2}}_{x;i,j+\frac{1}{2}} &= E^{n}_{x;i,j+\frac{1}{2}} + \frac{\Delta t}{2\Delta y}\left(B^{n+\frac{1}{4}}_{z;i,j+1} - B^{n+\frac{1}{4}}_{z;i,j}\right) \\
E^{n+\frac{1}{2}}_{y;i+\frac{1}{2},j} &= E^{n}_{y;i+\frac{1}{2},j} - \frac{\Delta t}{2\Delta x}\left(B^{n+\frac{1}{4}}_{z;i+1,j} - B^{n+\frac{1}{4}}_{z;i,j}\right) \\
E^{n+\frac{1}{2}}_{z;i+\frac{1}{2},j+\frac{1}{2}} &= E^{n}_{z;i+\frac{1}{2},j+\frac{1}{2}} + \frac{\Delta t}{2\Delta x}\left(B^{n+\frac{1}{4}}_{y;i+1,j+\frac{1}{2}} - B^{n+\frac{1}{4}}_{y;i,j+\frac{1}{2}}\right) - \frac{\Delta t}{2\Delta y}\left(B^{n+\frac{1}{4}}_{x;i+\frac{1}{2},j+1} - B^{n+\frac{1}{4}}_{x;i+\frac{1}{2},j}\right)
\end{aligned}
$$

and finally, these are plugged into another quarter step for the magnetic fields

$$
\begin{aligned}
B^{n+\frac{1}{2}}_{x;i+\frac{1}{2},j} &= B^{n+\frac{1}{4}}_{x;i+\frac{1}{2},j} - \frac{\Delta t}{4\Delta y}\left(E^{n+\frac{1}{2}}_{z;i+\frac{1}{2},j+\frac{1}{2}} - E^{n+\frac{1}{2}}_{z;i+\frac{1}{2},j-\frac{1}{2}}\right) \\
B^{n+\frac{1}{2}}_{y;i,j+\frac{1}{2}} &= B^{n+\frac{1}{4}}_{y;i,j+\frac{1}{2}} + \frac{\Delta t}{4\Delta x}\left(E^{n+\frac{1}{2}}_{z;i+\frac{1}{2},j+\frac{1}{2}} - E^{n+\frac{1}{2}}_{z;i-\frac{1}{2},j+\frac{1}{2}}\right) \\
B^{n+\frac{1}{2}}_{z;i,j} &= B^{n+\frac{1}{4}}_{z;i,j} + \frac{\Delta t}{4\Delta y}\left(E^{n+\frac{1}{2}}_{x;i,j+\frac{1}{2}} - E^{n+\frac{1}{2}}_{x;i,j-\frac{1}{2}}\right) - \frac{\Delta t}{4\Delta x}\left(E^{n+\frac{1}{2}}_{y;i+\frac{1}{2},j} - E^{n+\frac{1}{2}}_{y;i-\frac{1}{2},j}\right).
\end{aligned}
$$

Our next step is the Boris push. We proceed first with the electric update

$$
(\rho \boldsymbol{p})^{-}_{i+\frac{1}{2},j+\frac{1}{2}} = (\rho \boldsymbol{p})^{n}_{i+\frac{1}{2},j+\frac{1}{2}} + \frac{\Delta t}{2}\rho^{n}_{i+\frac{1}{2},j+\frac{1}{2}} \boldsymbol{E}^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}}
$$

for all three components. Of course, we have to interpolate E_x and E_y to the half positions via

$$
E_{x;i+\frac{1}{2},j+\frac{1}{2}} = \frac{E_{x;i+1,j+\frac{1}{2}} + E_{x;i,j+\frac{1}{2}}}{2}
$$

and

$$
E_{y;i+\frac{1}{2},j+\frac{1}{2}} = \frac{E_{y;i+\frac{1}{2},j+1} + E_{y;i+\frac{1}{2},j}}{2}.
$$

For the magnetic part of the push we define again the two quantities

$$
\begin{aligned}
\boldsymbol{t} &= \frac{\boldsymbol{B}}{\gamma}\frac{\Delta t}{4} \\
\boldsymbol{s} &= \frac{2\boldsymbol{t}}{1 + \|\boldsymbol{t}\|^2}
\end{aligned}
$$

for which we interpolate the magnetic field onto the grid positions required for the vector products via

$$
\begin{aligned}
B_{x;i+\frac{1}{2},j+\frac{1}{2}} &= \frac{B_{x;i+\frac{1}{2},j+1} + B_{x;i+\frac{1}{2},j}}{2} \\
B_{y;i+\frac{1}{2},j+\frac{1}{2}} &= \frac{B_{y;i+1,j+\frac{1}{2}} + B_{y;i,j+\frac{1}{2}}}{2} \\
B_{z;i+\frac{1}{2},j+\frac{1}{2}} &= \frac{B_{z;i+1,j+1} + B_{z;i+1,j} + B_{z;i,j+1} + B_{z;i,j}}{4}.
\end{aligned}
$$

The two sub-steps for the cross product $\boldsymbol{v} \times \boldsymbol{B}$ are

$$
\begin{aligned}
(\rho \boldsymbol{p})' &= (\rho \boldsymbol{p})^{-} + (\rho \boldsymbol{p})^{-} \times \boldsymbol{t} \\
(\rho \boldsymbol{p})^{+} &= (\rho \boldsymbol{p})^{-} + (\rho \boldsymbol{p})' \times \boldsymbol{s},
\end{aligned}
$$

which — dropping the spatial indices $i+\frac{1}{2}, j+\frac{1}{2}$ for better readability — read componentwise

$$\begin{aligned}(\rho p_x)' &= (\rho p_x)^- + (\rho p_y)^- t_z - (\rho p_z)^- t_y \\ (\rho p_y)' &= (\rho p_y)^- + (\rho p_z)^- t_x - (\rho p_x)^- t_z \\ (\rho p_z)' &= (\rho p_z)^- + (\rho p_x)^- t_y - (\rho p_y)^- t_x\end{aligned}$$

and

$$\begin{aligned}(\rho p_x)^+ &= (\rho p_x)^- + (\rho p_y)' s_z - (\rho p_z)' s_y \\ (\rho p_y)^+ &= (\rho p_y)^- + (\rho p_z)' s_x - (\rho p_x)' s_z \\ (\rho p_z)^+ &= (\rho p_z)^- + (\rho p_x)' s_y - (\rho p_y)' s_x.\end{aligned}$$

The inner step solves the fluid equations

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v_x) + \frac{\partial}{\partial y}(\rho v_y) = 0 \tag{7.2a}$$

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho p_x \\ \rho p_y \\ \rho p_z \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} \rho p_x v_x \\ \rho p_y v_x \\ \rho p_z v_x \end{pmatrix} + \frac{\partial}{\partial y}\begin{pmatrix} \rho p_x v_y \\ \rho p_y v_y \\ \rho p_z v_y \end{pmatrix} = \mathbf{0} \tag{7.2b}$$

$$\frac{\partial \boldsymbol{E}}{\partial t} = -\frac{\rho \boldsymbol{p}}{\gamma}. \tag{7.2c}$$

For (7.2a) and (7.2b), we use FCT, while for (7.2c), we compute

$$\boldsymbol{E}^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}} = \boldsymbol{E}^{n+\frac{1}{2}}_{i+\frac{1}{2}j+\frac{1}{2}} - \Delta t \frac{(\rho \boldsymbol{p})^{n+1}_{i+\frac{1}{2}j+\frac{1}{2}}}{\gamma^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}}}$$

or the appropriate approximation with an SSP Runge-Kutta method. Recall that we have already interpolated all electric fields onto the half-half positions in the electric update of the Boris push, so we can reuse them here as well as for the other half of the electric push before transforming back to the regular positions in the Yee cell for the Maxwell part.

7.3 Numerical Experiments

Now that we have shortly reviewed our equations and discussed the algorithm in 2D, we are ready for some numerical examples. Both the Yee scheme and multidimensional FCT have been tested on their own. We have already seen in chapter 5 that the rotating cylinder was simulated successfully and the performance of the Yee scheme is widely known (cf. [TH05]).

We consider two examples similar to what we have seen in the previous chapter in 1D. The plasma now lies almost like some brick inside our computational domain. The laser pulse starts again in vacuum, traveling towards and into the plasma. All calculations and pictures are in laser units.

We compare our results to those of a PIC code by Götz Lehmann of Heinrich-Heine-Universität Düsseldorf. For this kind of simulation, PIC codes are the standard method of choice.

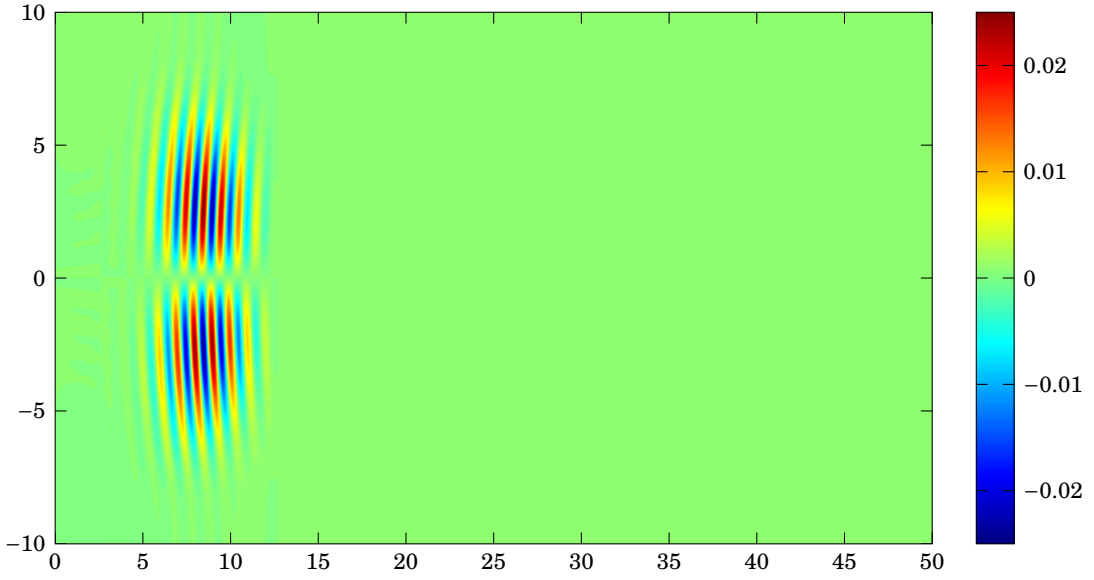

Figure 7.2: *Initial data for E_x*

For better comparison with the PIC code, we extract the initial values of the fields from there. For the PIC code, the pulse is introduced through so-called *emitting boundary conditions*, so our starting point is some time into the PIC simulation — $12.5T_0$ to be precise. The PIC simulation time goes up to $50T_0$ and we compare at PIC times. The PIC data are only available every $2.5T_0$.

Test Case 1: Low Density

We start out with the lower density case where the density ρ maxes at $0.04\rho_c$, that is 4% critical density. Figures 7.2, 7.3 and 7.4 show the initial data for E_x, E_y and B_z. Figure 7.5 shows a slice along the x-axis of all non-zero initial data. Note how this slice compares to the 1D setting from the previous chapter, except that we are now using sharp edges for the density profile instead of smoothing them by a Gaussian.

The PIC results were obtained by mesh sizes $\Delta x = 0.006$ and $\Delta y = 0.039$, while we used $\Delta x = 0.02$ and $\Delta y = 0.03$ for YeeFCT. Remember that we can take time steps as big as the CFL limit $\frac{\Delta x}{\sqrt{2}}$ because our velocities are bounded by one, which is the speed of light in our units.

To compare the results of both methods, we use again a slice along the x-axis and plot the graphs into the same figure.

Figure 7.6 shows the densities computed with both methods at time $t \approx 15T_0$. This is the first comparison after taking the initial data from the PIC simulation. Note how much noise the PIC code has already produced in the density profile. Then in figure 7.7, we see both densities at time $t \approx 20T_0$. Of course, YeeFCT does not reproduce the noise — just as we want. The movement at the left slope where the pulse hits, fits almost perfectly with what PIC did. The results for the electric and magnetic components E_y and B_z at the same time are shown in figures 7.8 and 7.9. Note how well YeeFCT reproduces the shape of the pulses. Only the resolution of peaks is not quite as good due to the bigger mesh size. The differences in peak height are roughly $\mathcal{O}(\Delta x)$.

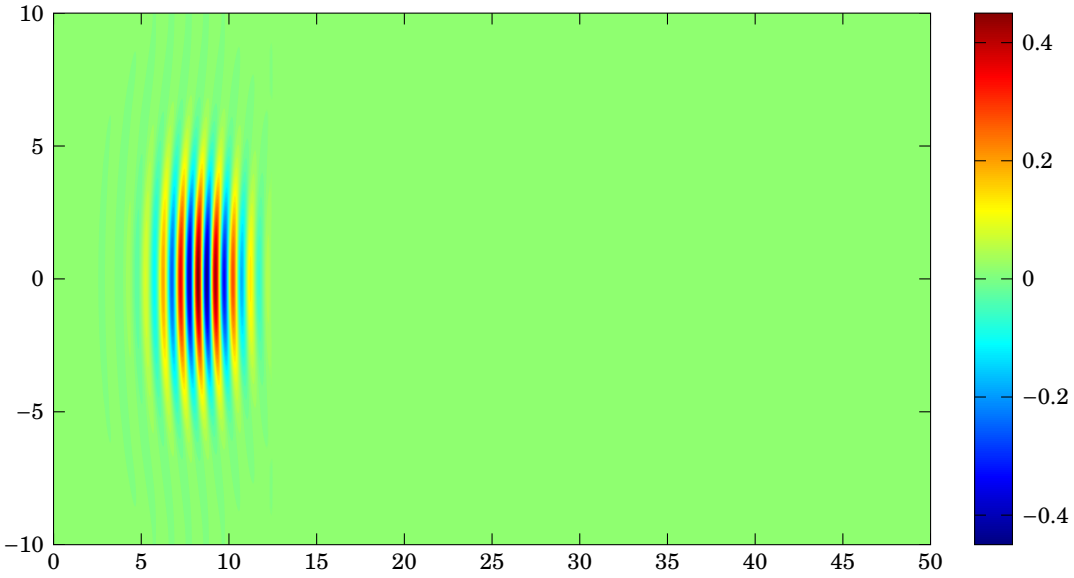

Figure 7.3: *Initial data for E_y*

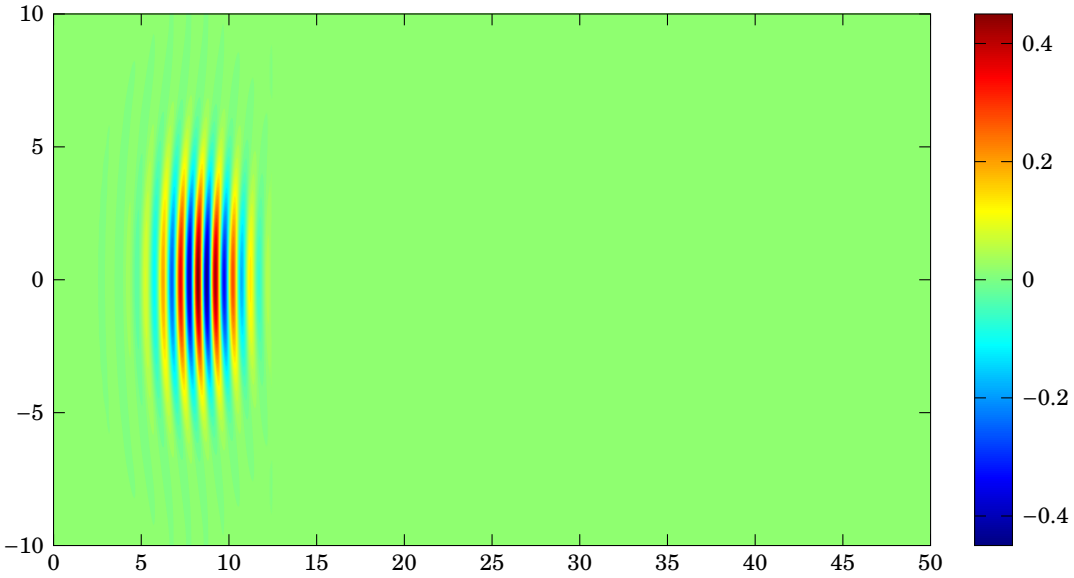

Figure 7.4: *Initial data for B_z*

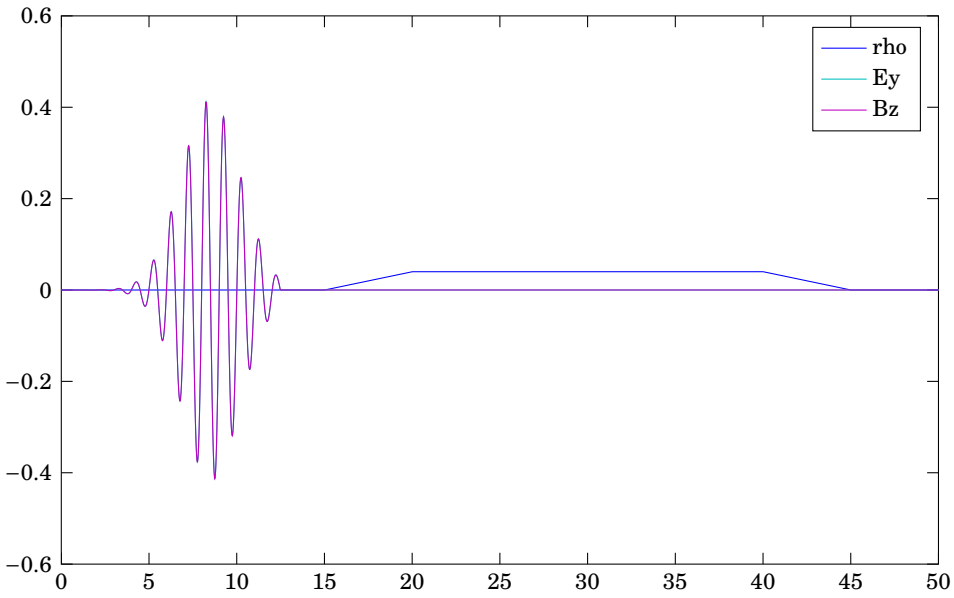

Figure 7.5: *Slice of initial data including plasma ramp*

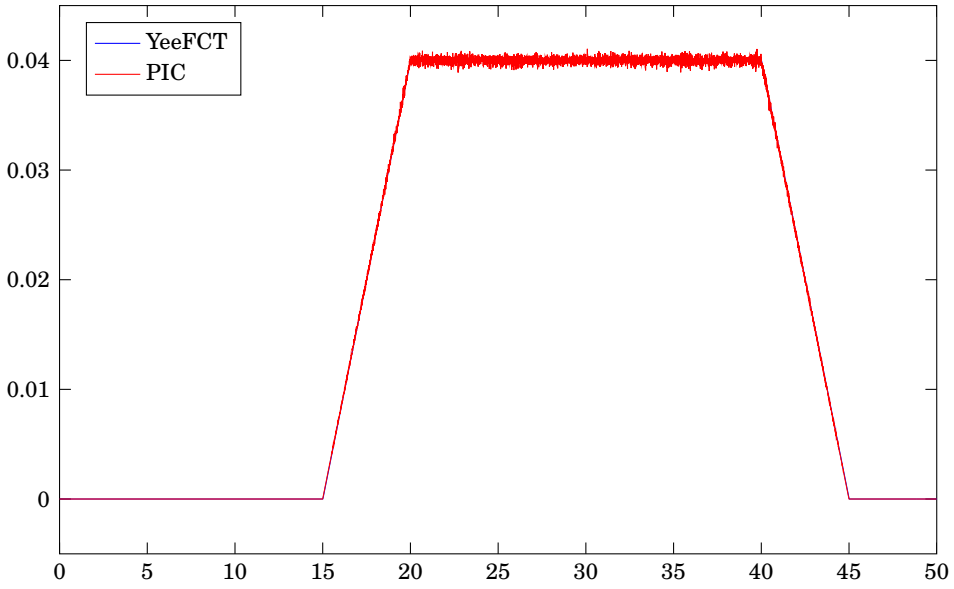

Figure 7.6: *Comparison of YeeFCT and PIC for the density at* $t/T_0 \approx 15$

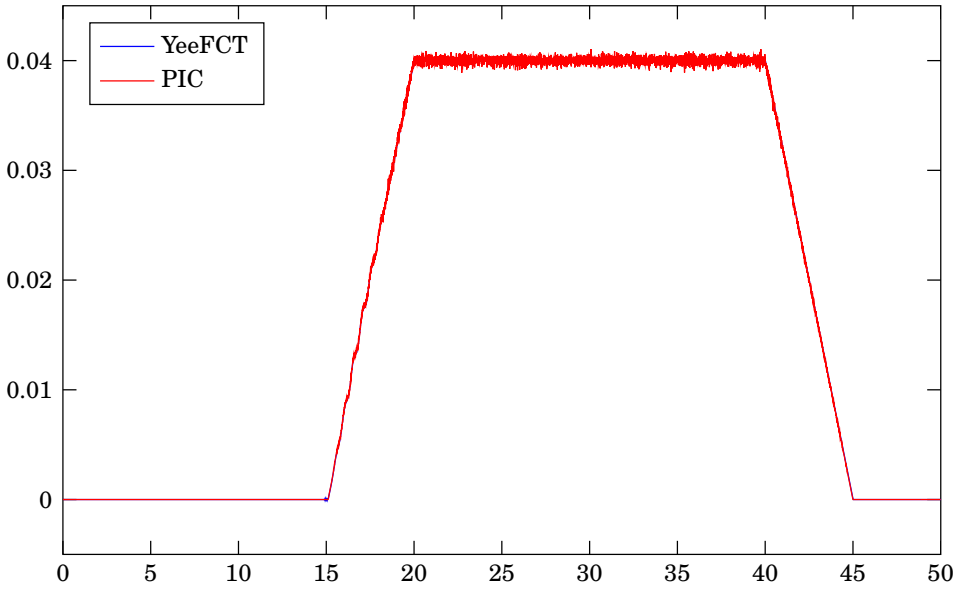

Figure 7.7: *Comparison of YeeFCT and PIC for the density at* $t/T_0 \approx 20$

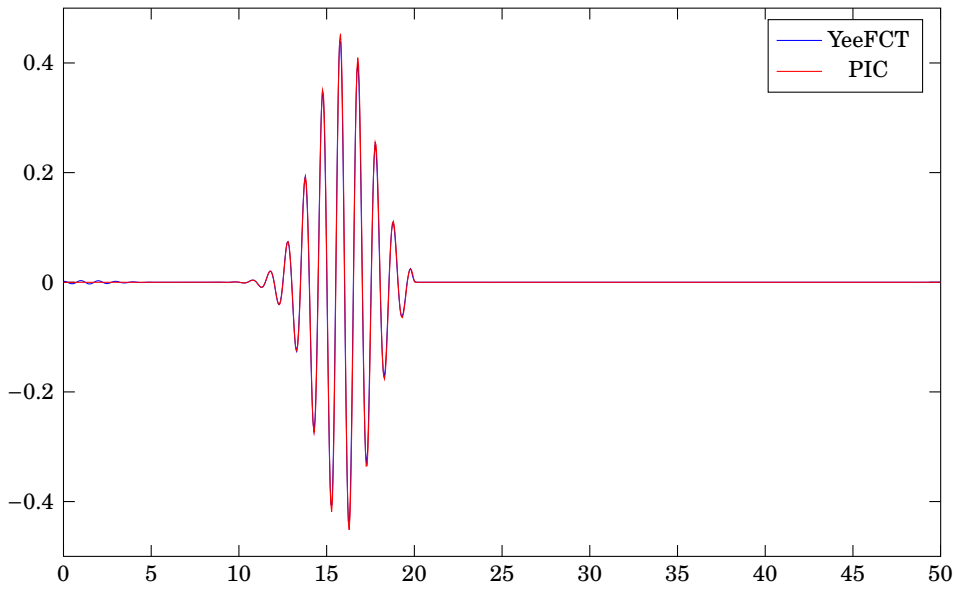

Figure 7.8: *Comparison of YeeFCT and PIC for* E_y *at* $t/T_0 \approx 20$

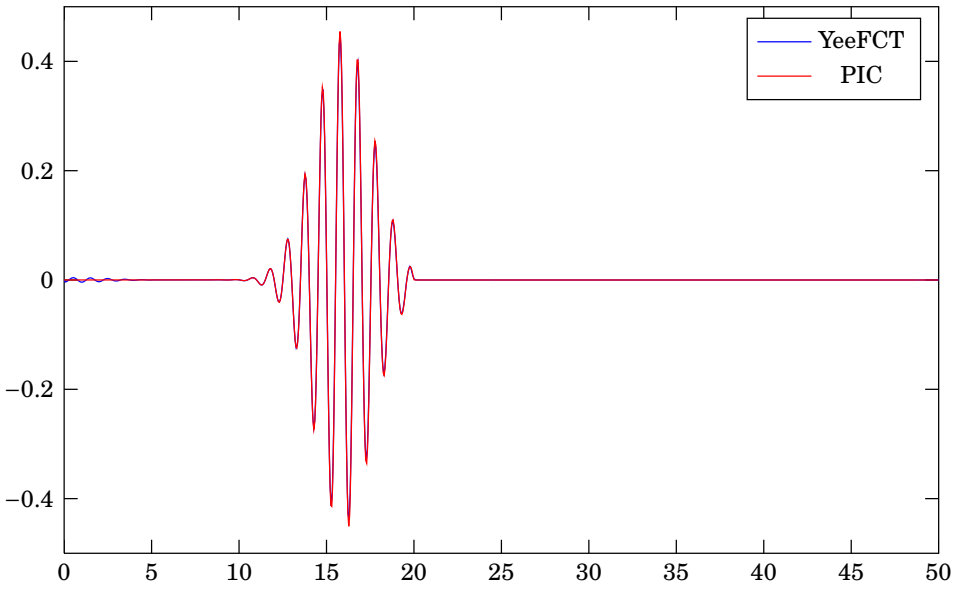

Figure 7.9: *Comparison of YeeFCT and PIC for B_z at $t/T_0 \approx 20$*

Figures 7.10, 7.11 and 7.12 show the same comparison at time $t \approx 30T_0$. The waves that are excited in the plasma are reproduced very accurately by YeeFCT in spite of the noise in the PIC results. Only the narrow peaks are again not resolved perfectly on the coarser mesh.

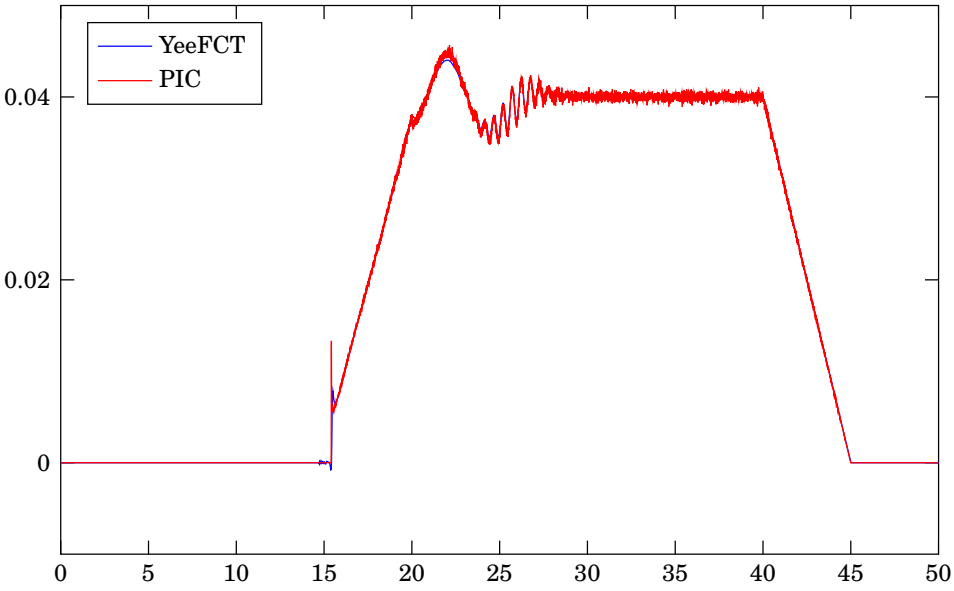

Figure 7.10: *Comparison of YeeFCT and PIC at $t/T_0 \approx 30$*

In figure 7.13 at $t \approx 40T_0$, we see some diffusion effects for the density from the FCT part of our algorithm. The pulse still looks pretty good.

At the final time $t \approx 50T_0$, the pulse in figures 7.17 and 7.18 looks good again. The waves inside the plasma are reproduced quite well by YeeFCT, except for some diffusion. But recall that by means of standard methods, we could not have hoped to achieve any of this. A simple upwind or Lax-Friedrichs scheme would have produced a lot more diffusion

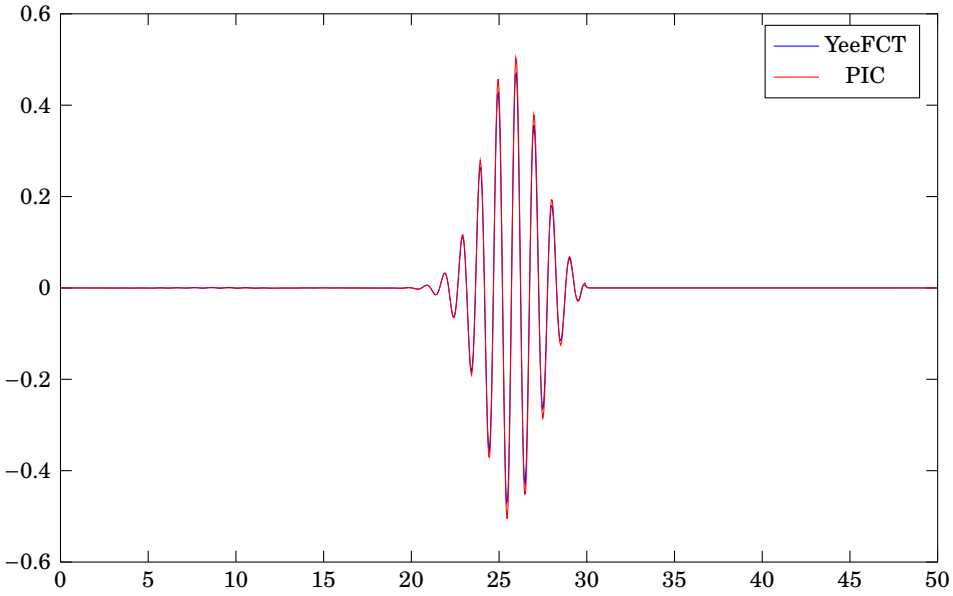

Figure 7.11: *Comparison of YeeFCT and PIC for E_y at $t/T_0 \approx 30$*

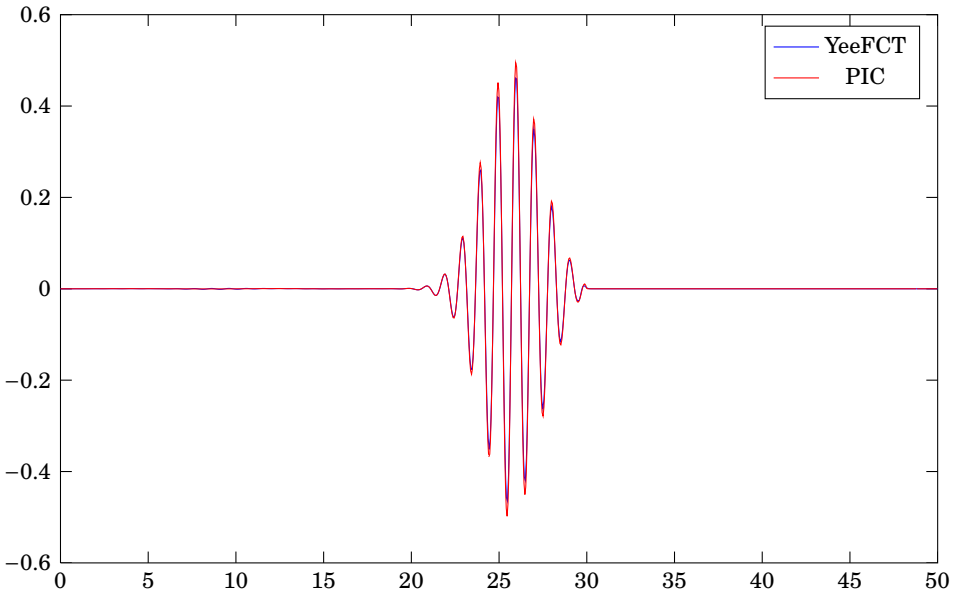

Figure 7.12: *Comparison of YeeFCT and PIC for B_z at $t/T_0 \approx 30$*

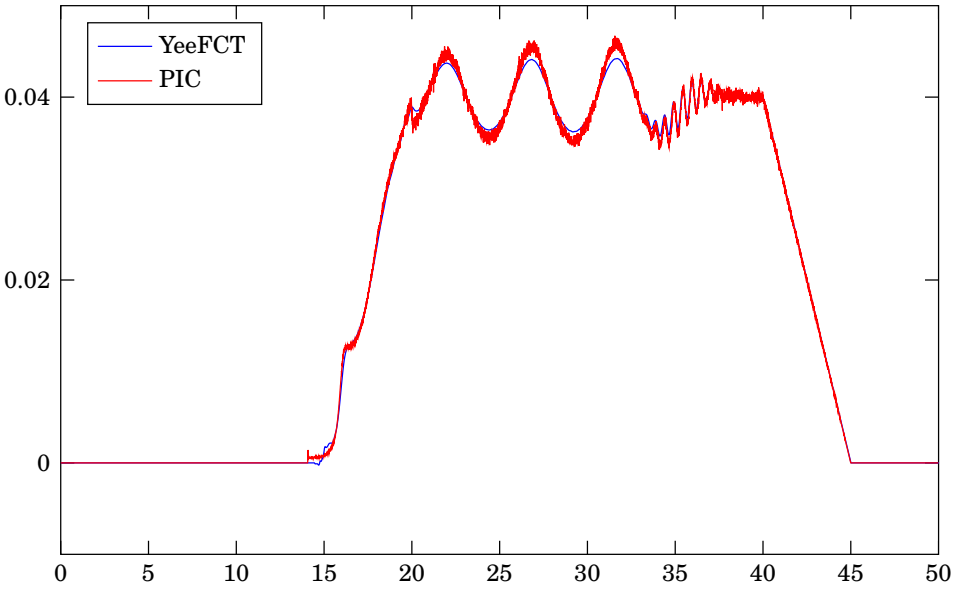

Figure 7.13: *Comparison of YeeFCT and PIC at* $t/T_0 \approx 40$

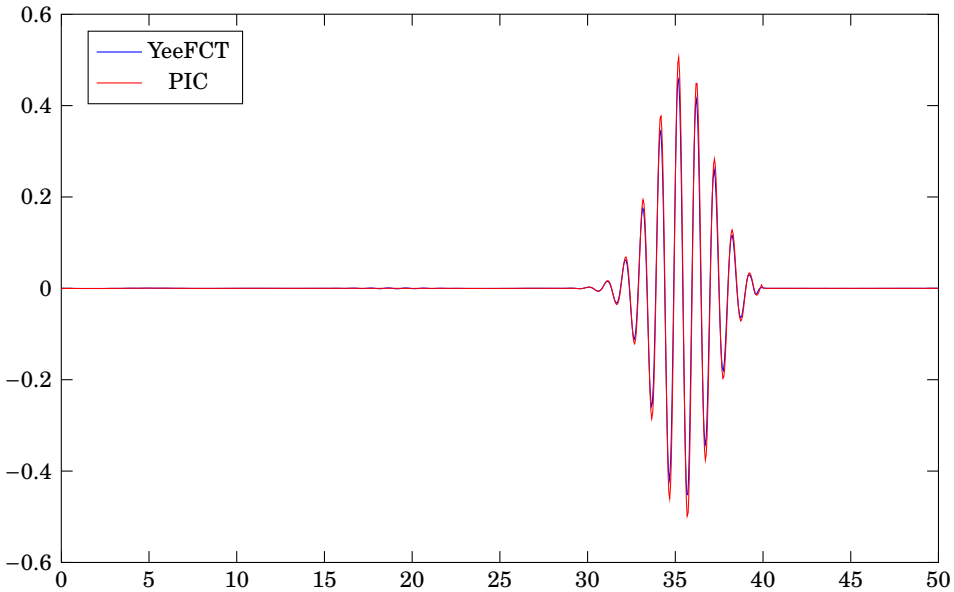

Figure 7.14: *Comparison of YeeFCT and PIC for* E_y *at* $t/T_0 \approx 40$

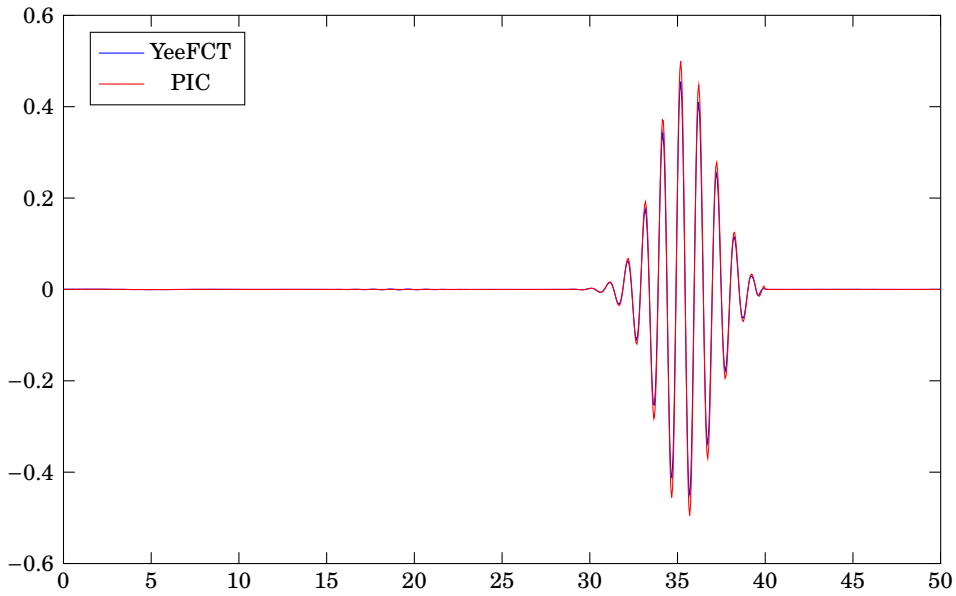

Figure 7.15: *Comparison of YeeFCT and PIC for B_z at $t/T_0 \approx 40$*

while a higher order method like Lax-Wendroff would have introduced spurious ripples even worse than the noise from the PIC code.

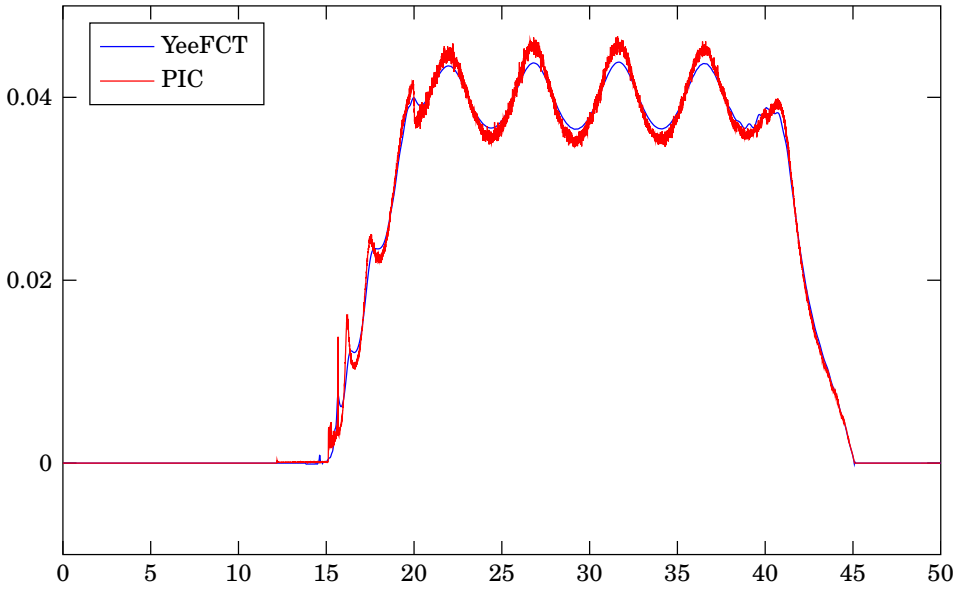

Figure 7.16: *Comparison of YeeFCT and PIC at $t/T_0 \approx 50$*

To compare the fully two-dimensional results, see figures 7.19 and 7.20. The noise makes the PIC image look somewhat smoother than the YeeFCT. But comparing to the slice plot in figure 7.16, we see that YeeFCT actually produced a much smoother density profile. However, we can still conclude that both images of the 2D results are very similar and that YeeFCT worked really well here.

Let us look again at the conservation of mass. In this experiment, the relative error stayed well below 10^{-6}. Figure 7.21 shows a semilogarithmic plot of the relative error in

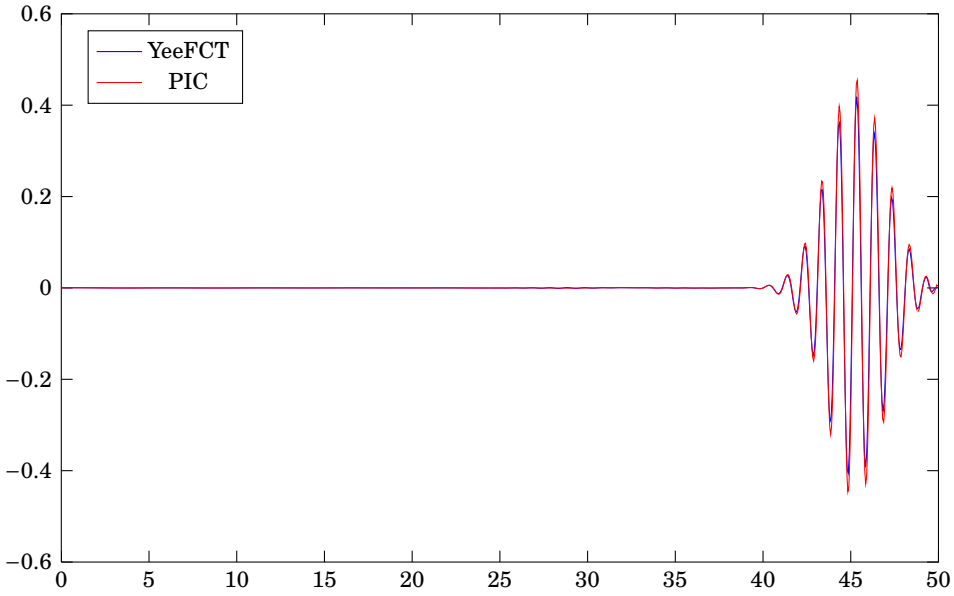

Figure 7.17: *Comparison of YeeFCT and PIC for E_y at $t/T_0 \approx 50$*

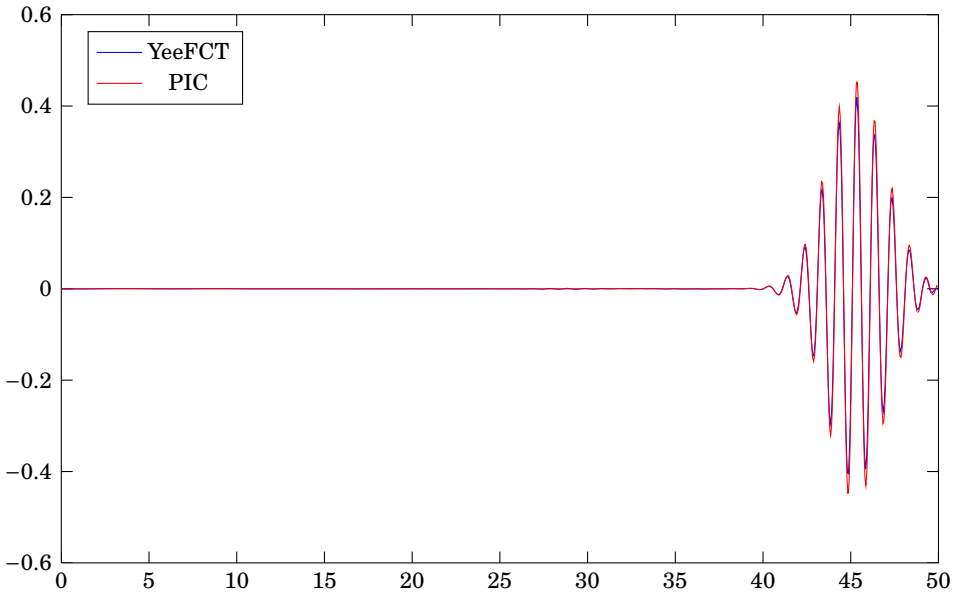

Figure 7.18: *Comparison of YeeFCT and PIC for B_z at $t/T_0 \approx 50$*

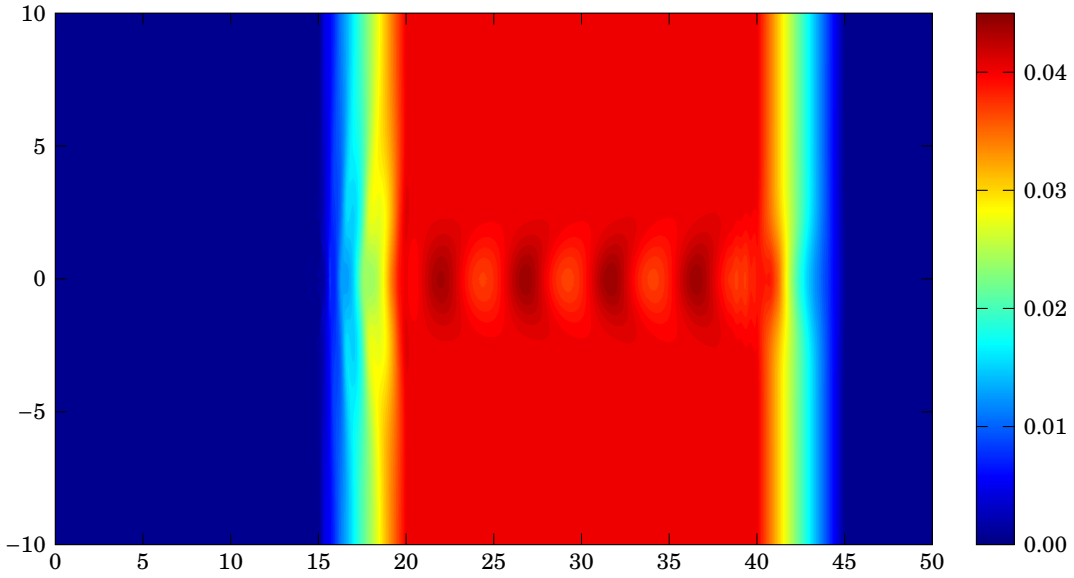

Figure 7.19: *Final density profile with YeeFCT*

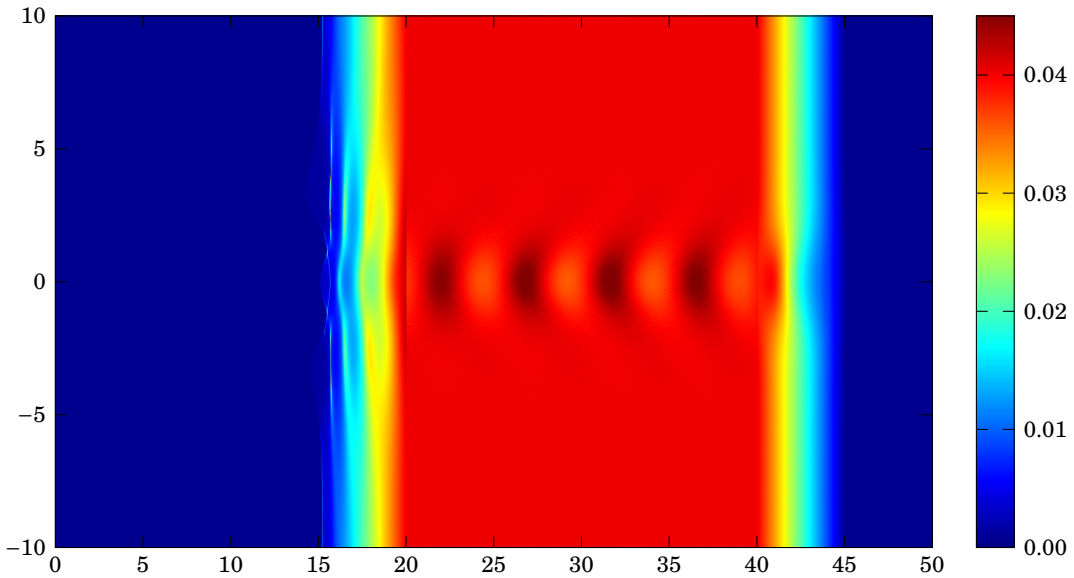

Figure 7.20: *Final density profile with PIC*

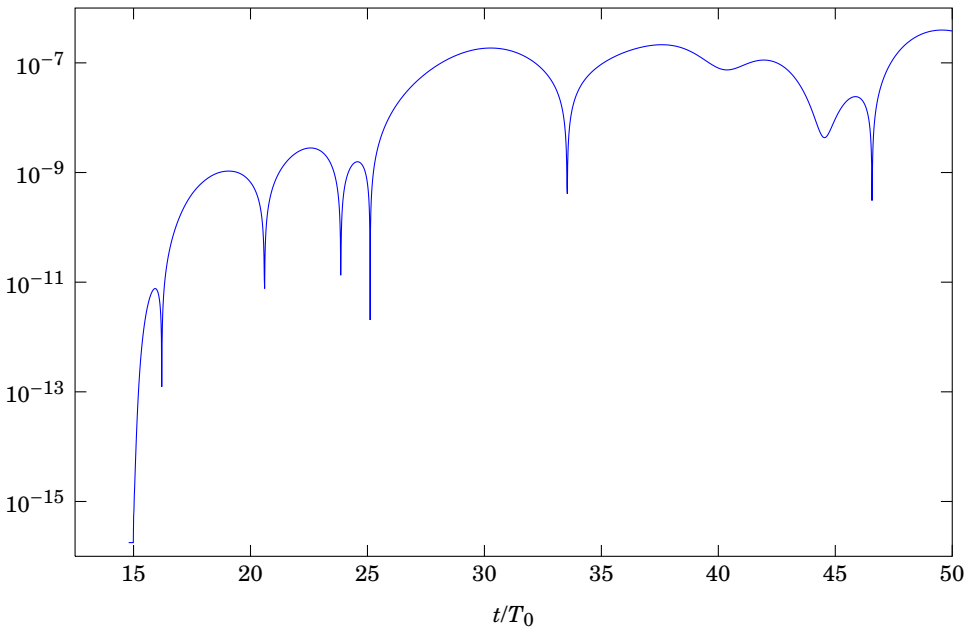

Figure 7.21: *Relative error in total mass with YeeFCT*

total mass for the above simulation.

The other important quantity is energy. We have already found in the one-dimensional case that energy conservation was not optimal. What we observe is that once the pulse enters the plasma, the error in energy increases. But even then, the relative error in this simulation does not exceed 10%, see figure 7.22.

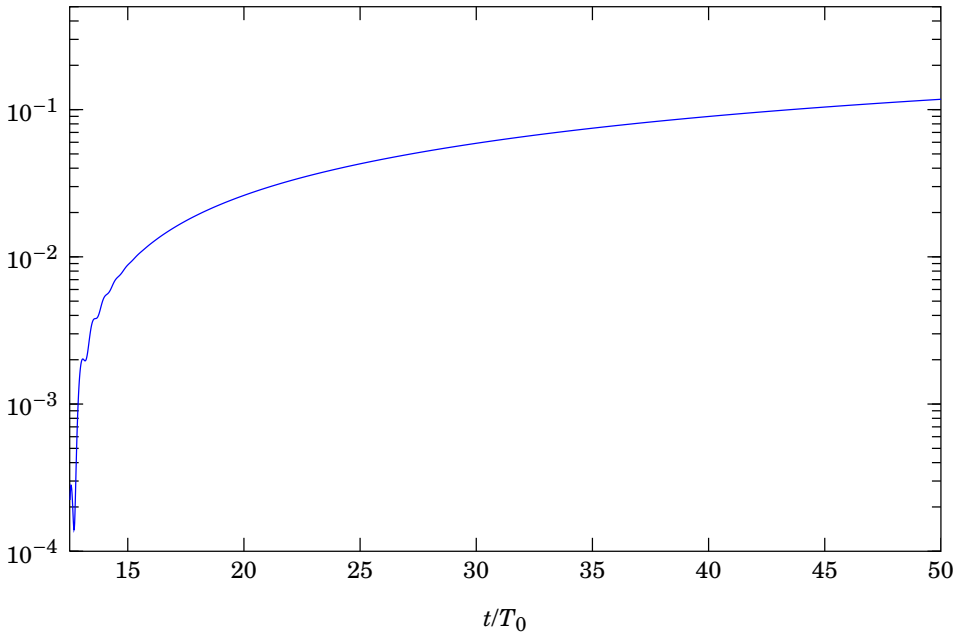

Figure 7.22: *Relative error in energy with YeeFCT*

Test Case 2: High Density

Now, let us carry the high density example with $\tilde{\rho} = 0.6$ from the previous chapter into two dimensions. Like in the one-dimensional case, we reduce the step sizes to $\Delta x = 0.015$ and $\Delta y = 0.02$. Figures 7.23 et seqq. show the same snapshots we had for the comparison in the low density case. Like in 1D, the curves do not fit as well as for low density, but qualitatively, the results are very good again.

When it comes to computational time, YeeFCT, which is programmed in MATLAB, takes less than an hour on a dual core desktop PC to achieve some decent results. The computations shown here only took a few hours. The results from the C-programmed PIC simulation, however, took a few days on a cluster of 190 processors! Of course, the hydrodynamic model has its limits[10], but as long as we are only interested in hydrodynamic effects, we can have even better results in terms of noise and such by using YeeFCT for the simulation. And we should not forget that we are comparing a MATLAB implementation with a highly optimized and parallelized C code. A halfway decent C implementation of YeeFCT should give enough speedup to use a grid that is fine enough to eliminate the remaining diffusion effects.

[10] wave-breaking is an example of what cannot be done using a hydrodynamic model

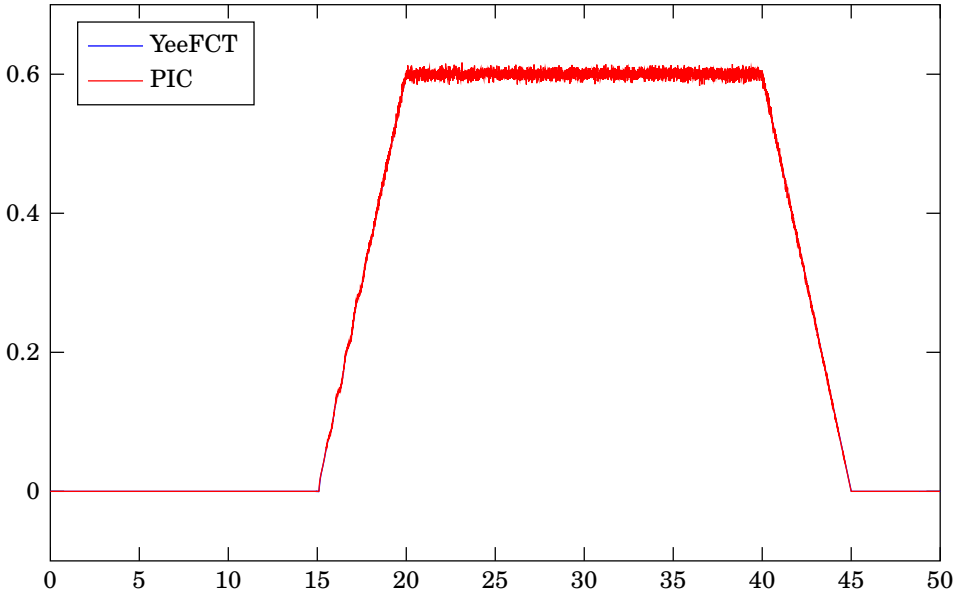

Figure 7.23: *Comparison of YeeFCT and PIC for the density at* $t/T_0 \approx 20$

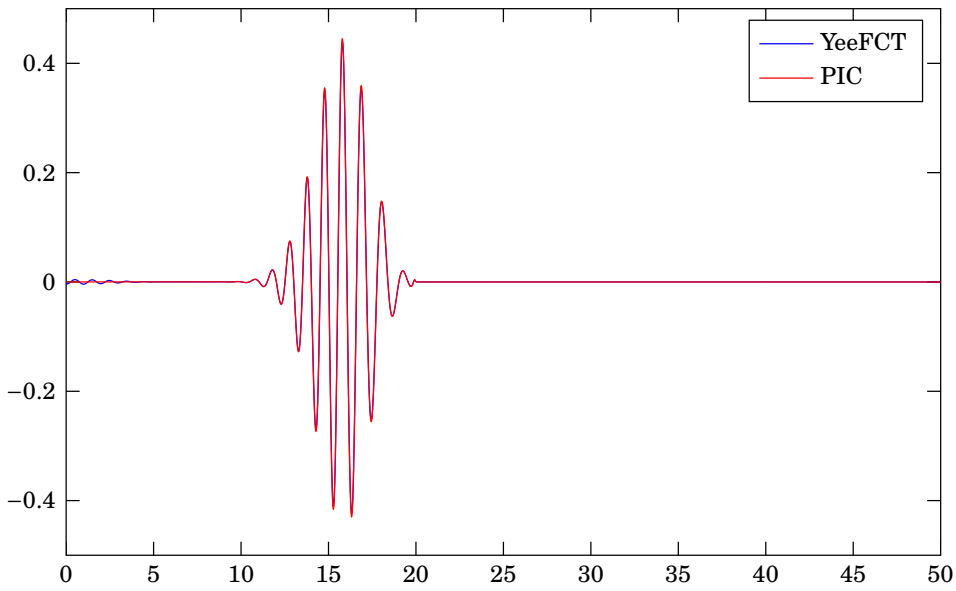

Figure 7.24: *Comparison of YeeFCT and PIC for* E_y *at* $t/T_0 \approx 20$

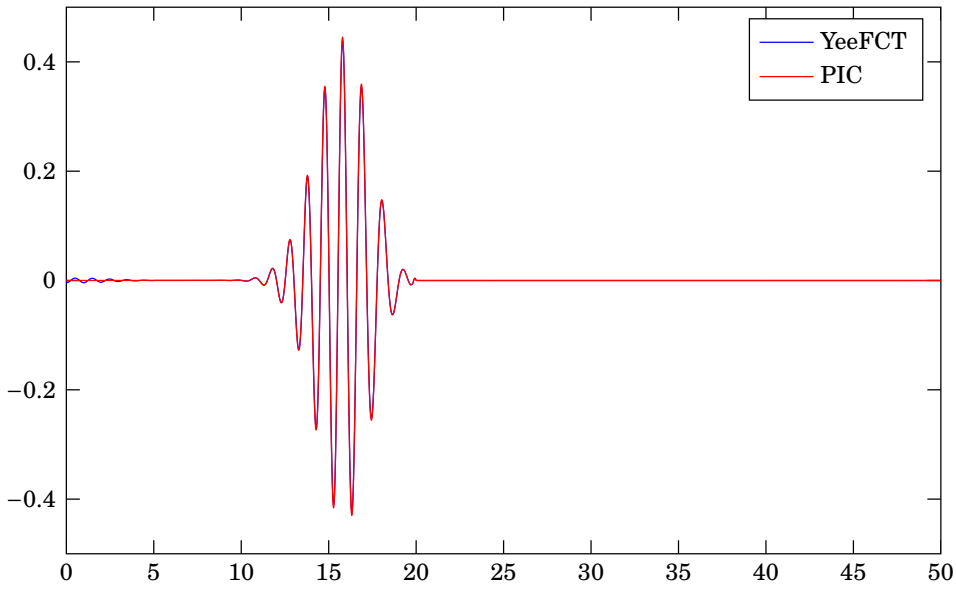

Figure 7.25: *Comparison of YeeFCT and PIC for B_z at $t/T_0 \approx 20$*

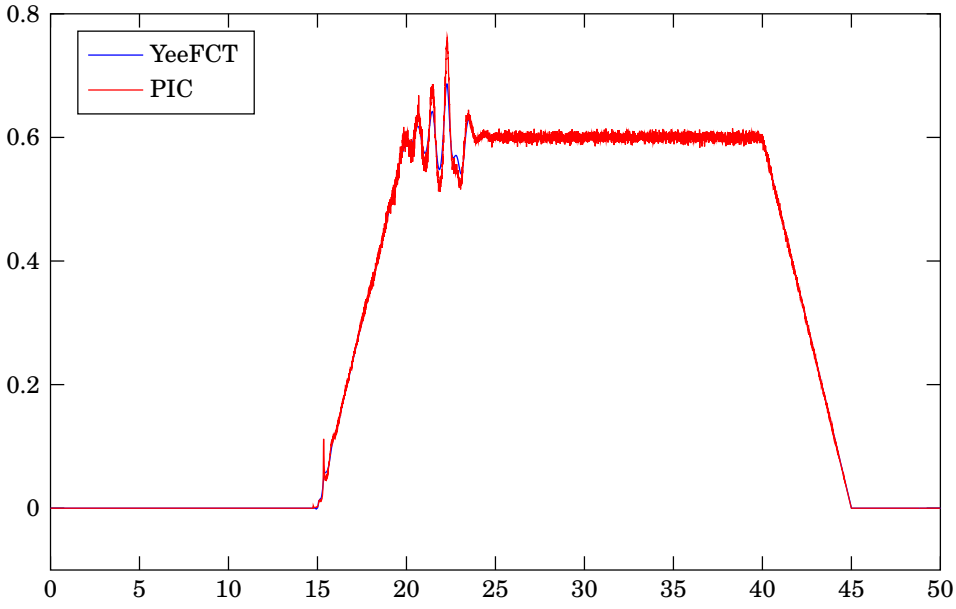

Figure 7.26: *Comparison of YeeFCT and PIC for the density at $t/T_0 \approx 30$*

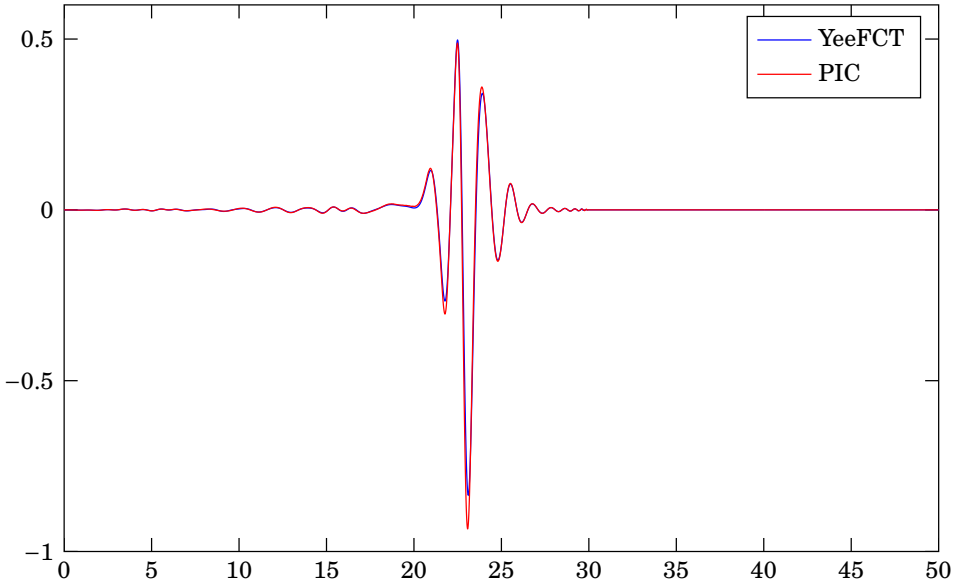

Figure 7.27: *Comparison of YeeFCT and PIC for E_y at $t/T_0 \approx 30$*

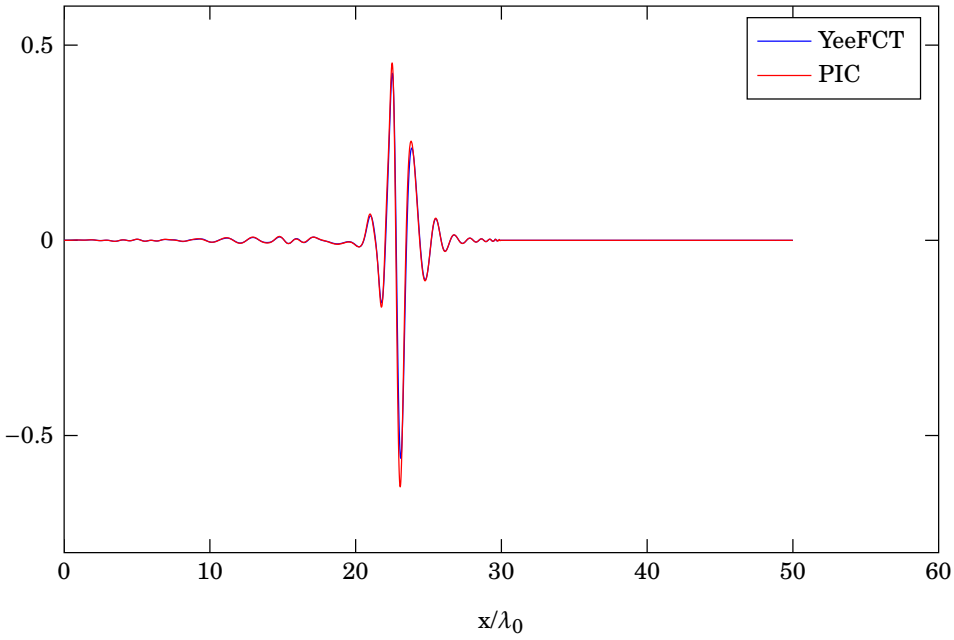

Figure 7.28: *Comparison of YeeFCT and PIC for B_z at $t/T_0 \approx 30$*

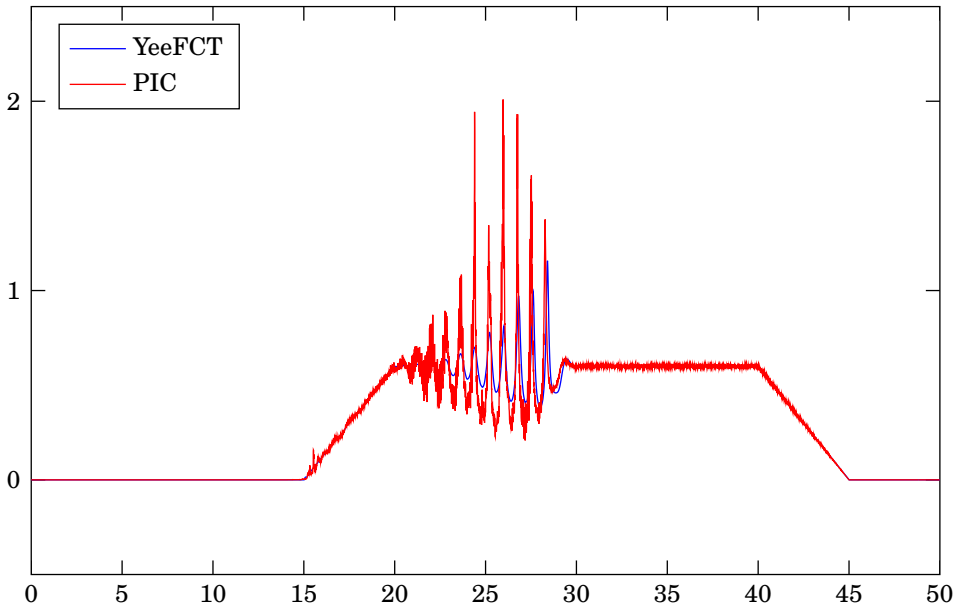

Figure 7.29: *Comparison of YeeFCT and PIC for the density at* $t/T_0 \approx 40$

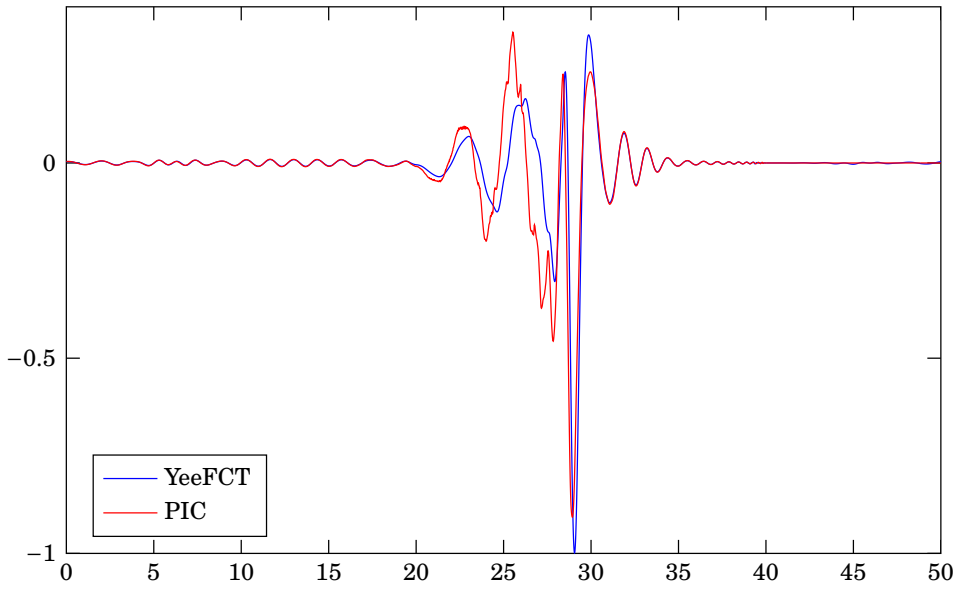

Figure 7.30: *Comparison of YeeFCT and PIC for* E_y *at* $t/T_0 \approx 40$

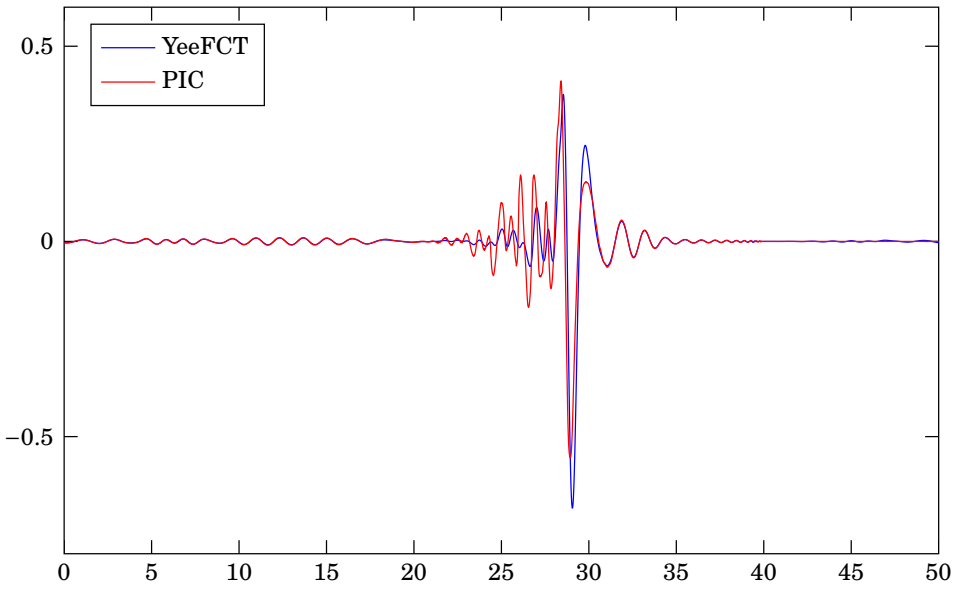

Figure 7.31: *Comparison of YeeFCT and PIC for B_z at $t/T_0 \approx 40$*

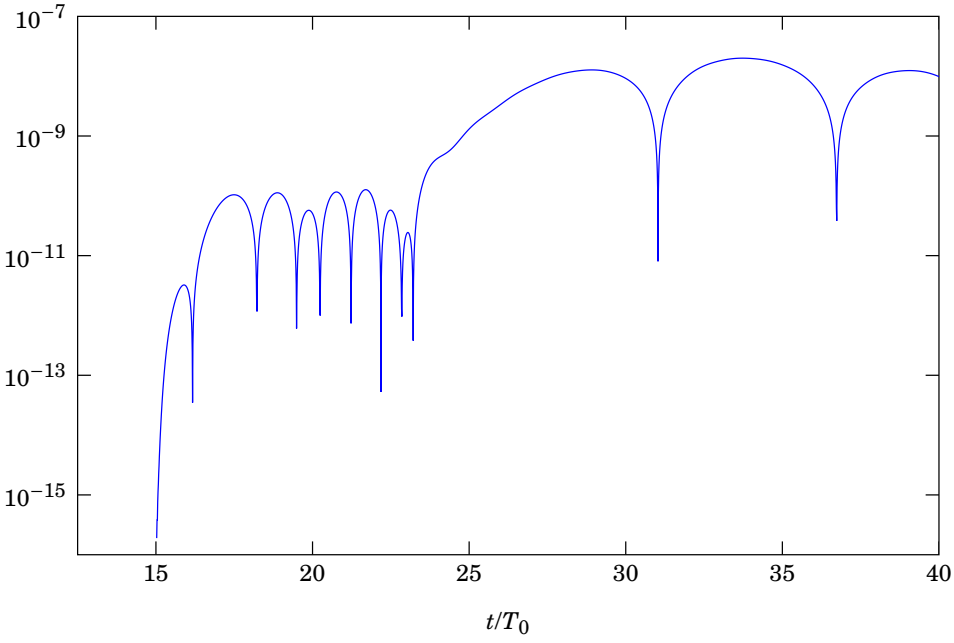

Figure 7.32: *Relative error in total mass with YeeFCT*

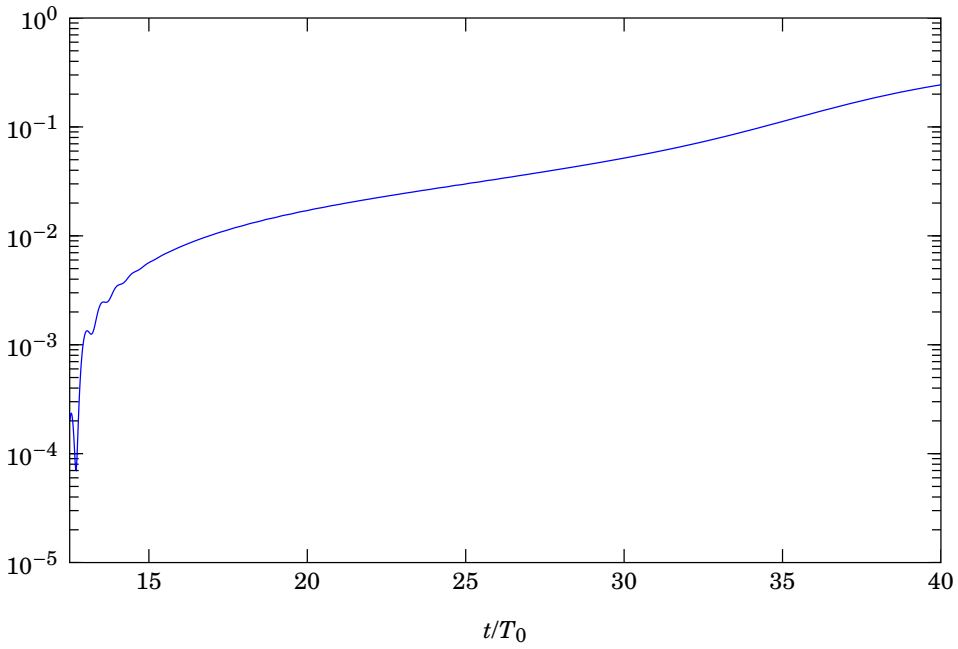

Figure 7.33: *Relative error in energy with YeeFCT*

8 Conclusions and Outlook

Knowing that the solution of hyperbolic conservation laws is quite tricky, the goal to accurately simulate relativistic laser-plasma interaction — especially near vacuum — is ambitious. We have to deal with two conservation laws for the plasma and additionally some equations and terms for the laser.

The electromagnetic laser waves are described by Maxwell's equations, for which the Yee scheme from chapter 2 works really well. In section 3 we learned about the difficulties with nonlinear hyperbolic conservation laws and discussed some classical approaches to their numerical solution and the stability of those methods in section 3.2. The momentum density equation contains a source term — the Lorentz force. In section 4.2, we examined the nature of this force and introduced the Boris push for a correct rotational particle movement. Finally, we presented the flux-corrected transport algorithm in chapter 5, which is constructed to preserve the positivity of the numerical solution of nonlinear conservation laws.

Now we had the ingredients for the numerical solution to every single part of our equations, but the question remained how to combine them to find a numerical solution to the whole set of equations. FCT does not involve source terms and coupling it with the Yee scheme is not straightforward, either. We had to resort to yet another field of numerical analysis and found an appropriate tool in symmetric splitting techniques. That way we were able to combine the famous Yee scheme for Maxwell's equations and the Boris push for the Lorentz force with a positivity preserving FCT scheme into an efficient algorithm for the numerical simulation of relativistic laser-plasma interaction. The numerical results excellently reproduced what established codes like Vlasov or PIC simulations showed, while YeeFCT is much faster — hours on a desktop PC versus days on a cluster with a few hundred cores.

We also addressed the analysis of the scheme, which is somewhat difficult because we do not know the values of the limiting factors as they depend on the solution at a given time step. It might be interesting, though, to further examine the stability and convergence properties. Maybe the knowledge about the properties of the underlying high and low order schemes allows some conclusions about YeeFCT.

One important step towards improving YeeFCT lies in the implementation. While MATLAB offers great possibilities for developing new methods in a comfortable environment, this comes at the price of a huge amount of overhead. A simple C or C++ implementation would eliminate this problem and would lead to an enormous speedup in computation time. This would be a great opportunity especially for fully three-dimensional simulations or in high density simulations where small time steps are required to resolve the formation of large peaks.

The next obvious improvement would be variable step sizes. Adaptive time stepping seems the easy part here since there are very good strategies for Runge-Kutta methods. Adaptive mesh refinement, however, is usually quite complicated, especially in the multidimensional case. As a first step, non-uniform meshes without adaptive refinement should be considered. In the vacuum-plasma transitions discussed in this thesis, it is clear that the plasma ramp is where a finer grid is most necessary while in vacuum, a coarser mesh is sufficient. The theory is easily adapted to non-uniform meshes, so it would be interesting to see the influence on the numerical experiments.

The goal of this thesis was the simulation of a hydrodynamic model for relativistic laser-plasma interaction, which is quite a complex field. The simulation of a kinetic model

with PIC codes works, but takes a lot of time. For the application of a vacuum-plasma transition, the YeeFCT scheme brought a huge speedup with very good results. This advance marks the achievement of this thesis.

References

[BGB+99] Marie L. Bégué, Alain Ghizzo, Pierre Bertrand, Eric Sonnendrücker, and Olivier Couldaud, *Two-dimensional semi-Lagrangian Vlasov simulations of laser-plasma interaction in the relativistic regime*, Journal of Plasma Physics **62** (1999), 367–388.

[BL91] Charles K. Birdsall and A. Bruce Langdon, *Plasma physics via computer simulation*, Hilger, Bristol, 1991.

[Bor70] Jay P. Boris, *Relativistic plasma simulation — optimization of a hybrid code*, Fourth conference on numerical simulation of plasmas (Washington, DC Naval Research Laboratory), 1970.

[Bor71] Jay P. Boris, *A fluid transport algorithm that works*, Proceedings of the Seminar Course on Computing as a Language of Physics (International Centre for Theoretical Physics, Trieste, Italy), 1971.

[BB73] Jay P. Boris and David L. Book, *Flux-Corrected Transport: I. SHASTA, a fluid transport algorithm that works*, Journal of Computational Physics **11** (1973), 38–69.

[Bun59] Oscar Buneman, *Dissipation of currents in ionized media*, Physical Reviews **115** (1959), 503–517.

[Car30] Horatio S. Carslaw, *Introduction to the theory of Fourier's series and integrals*, third rev. and enl. ed., Dover Publ., New York, 1930.

[Che83] Kuo-Shung Cheng, *The space BV is not enough for hyperbolic conservation laws*, J. Math. Anal. Appl. **91** (1983), 559–561.

[CK76] Chio Z. Cheng and Georg Knorr, *The integration of the Vlasov equation in configuration space*, Journal of Computational Physics **22** (1976), 330–351.

[CF76] Richard Courant and Kurt O. Friedrichs, *Supersonic flow and shock waves*, unchanged reprint of the original edition published in 1948 by interscience publishers inc, new york ed., Applied mathematical sciences, vol. 21, Springer, New York, 1976.

[Daf72] Constantine M. Dafermos, *Polygonal approximations of solutions of the initial value problem for a conservation law*, J. Math. Anal. Appl. **38** (1972), 33–41.

[Daw62] John M. Dawson, *One-dimensional plasma model*, Physics of Fluids **5** (1962), 445–459.

[DeV98] C. Richard DeVore, *An improved limiter for multidimensional flux-corrected transport*, 31. December 1998.

[DiP83] Ronald J. DiPerna, *Convergence of approximate solutions to conservation laws*, Archive for Rational Mechanics and Analysis **82** (1983), 27–70.

[Dö12] Willy Dörfler, *Numerical methods for hyperbolic equations*, Lecture Notes, 2012.

[Eva10] Lawrence C. Evans, *Partial differential equations*, Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 2010.

[Ger10] Christian Gerthsen, *Gerthsen physik*, 24th ed., Springer-Lehrbuch, Springer, Berlin, 2010.

[GHB03] Alain Ghizzo, Fabien Huot, and Pierre Bertrand, *A non-periodic 2d semi-lagrangian vlasov code for laser–plasma interaction on parallel computer*, Journal of Computational Physics **186** (2003), 47–69.

[Gla96] Robert T. Glassey, *The Cauchy problem in kinetic theory*, SIAM, Philadelphia, 1996.

[GKS11] Sigal Gottlieb, David Ketcheson, and Chi-Wang Shu, *Strong stability preserving Runge-Kutta and multistep time discretizations*, World Scientific, Singapore, 2011.

[HLW06] Ernst Hairer, Christian Lubich, and Gerhard Wanner, *Geometric numerical integration*, Springer series in computational mathematics, vol. 31, Springer, 2006.

[Har83] Ami Harten, *High resolution schemes for hyperbolic conservation laws*, Journal of Computational Physics **49** (1983), 357–393.

[HHLK76] Ami Harten, James M. Hyman, Peter D. Lax, and Barbara Keyfitz, *On finite-difference approximations and entropy conditions for shocks*, Communications on Pure and Applied Mathematics **29** (1976), 297–322.

[Hop70] Eberhard Hopf, *On the right weak solution of the cauchy problem for a quasi-linear equation of first order*, J. Math. Mech. **19** (1970), 483–487.

[HL94] Thomas Y. Hou and Philippe G. LeFloch, *Why nonconservative schemes converge to wrong solutions: error analysis*, Mathematics of Computation **62** (1994), 497–530.

[IN79] Tsutomu Ikeda and Tomoyasu Nakagawa, *On the shasta fct algorithm for the equation* $\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(v(\rho)\rho) = 0$, Mathematics of Computation **33** (1979), 1157–1169.

[Jac99] John D. Jackson, *Classical electrodynamics*, third ed., Wiley, New York, 1999.

[Jah10] Tobias Jahnke, *Numerical methods for Maxwell's equations*, Lecture Notes, 2010.

[KSH+06] Christoph Karle, Julia Schweitzer, Marlis Hochbruck, Ernst-Wolfgang Laedke, and Karl-Heinz Spatschek, *Numerical solution of nonlinear wave equations in stratified dispersive media*, Journal of Computational Physics **216** (2006), 138–152.

[Kro97] Dietmar Kroener, *Numerical schemes for conservation laws*, Wiley-Teubner series advances in numerical mathematics, Wiley-Teubner, Chichester, 1997.

[Kru03] William L. Kruer, *The physics of laser plasma interactions*, Frontiers in physics, Westview, Boulder, Colo., 2003.

[Kru70] Stanislav N. Kružkov, *First order quasilinear equations in several independent variables*, 217–243.

[Lax73] Peter D. Lax, *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, CBMS-NSF Reg. Conf. Ser. Appl. Math., no. 11, Society for Industrial and Applied Mathematics, 1973.

[LLS07] Götz Lehmann, Ernst W. Laedtke, and Karl-Heinz Spatschek, *Localized wakefield excitation and relativistic wave-breaking*, Physics of plasmas **14** (2007), 1–9.

[LeV06] Randall J. LeVeque, *Numerical methods for conservation laws*, second ed., Lectures in mathematics, Birkhäuser, Basel, 2006, reprint.

[LeV11] Randall J. LeVeque, *Finite volume methods for hyperbolic problems*, Cambridge texts in applied mathematics, Cambridge University Press, 2011.

[Mac11] Andrea Macchi, *An introduction to ultraintense laser-plasma interactions*, Review Article, May 16, 2011, http://www.df.unipi.it/~macchi.

[Oku72] Hideo Okuda, *Nonphysical noises and instabilities in plasma simulation due to a spatial grid*, Journal of Computational Physics **10** (1972), 475–486.

[Ole64] Olga A. Oleinik, *Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation*, AMS Translations **49** (1964).

[OC84] Stanley Osher and Sukumar Chakravarthy, *High resolution schemes and the entropy condition*, SIAM Journal on Numerical Analysis **21** (1984), 955–984.

[PR00] Frédéric Poupaud and Malika Remaki, *Existence et unicité des solutions du système de Maxwell pour des milieux hétérogènes non réguliers*, Comptes Rendus de l'Académie des Sciences - Series I - Mathematics **330** (2000), no. 2, 99–103.

[Puk99] Alexander Pukhov, *Three-dimensional electromagnetic relativistic particle-in-cell code VLPL (Virtual Laser Plasma Lab)*, Journal of Plasma Physics **61** (1999), 425–433.

[RM67] Robert D. Richtmyer and Keith W. Morton, *Difference methods for initial-value problems*, second ed., Interscience tracts in pure and applied mathematics, vol. 4, Interscience Publ., New York, 1967.

[Shu88] Chi-Wang Shu, *Total-variation diminishing time discretizations*, J. Sci. Stat. Comput. **9** (1988), 1073–1084.

[SO88] Chi-Wang Shu and Stanley Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, Journal of Computational Physics **77** (1988), 439–471.

[SRBG99] Eric Sonnendrücker, Jean Roche, Pierre Bertrand, and Alain Ghizzo, *The semi-lagrangian method for the numerical resolution of the Vlasov equation*, Journal of Computational Physics **149** (1999), 201–220.

[Str68] Gilbert Strang, *On the construction and comparison of difference schemes*, SIAM Journal on Numerical Analysis **5** (1968), 506–517.

[Str04] John C. Strikwerda, *Finite difference schemes and partial differential equations*, second ed., SIAM, Philadelphia, Pa., 2004.

[TH05] Allen Taflove and Susan C. Hagness, *Computational electrodynamics: the finite-difference time-domain method*, third ed., Artech House, 2005.

[Taj89] Toshiko Tajima, *Computational plasma physics: With applications to fusion and astrophysics*, Frontiers in physics, vol. 72, Addison-Wesley, Redwood City, 1989.

[Tro59] Hale F. Trotter, *On the product of semi-groups of operators*, Proceedings of the American Mathematical Society **10** (1959), 545–551.

[TPLH10] Tobias Tückmantel, Alexander Pukhov, Jalo Liljo, and Marlis Hochbruck, *Three-dimensional relativistic particle-in-cell hybrid code based on an exponential integrator*, IEEE Transactions on Plasma Science **38** (2010), 2383–2389.

[Wes01] Pieter Wesseling, *Principles of computational fluid dynamics*, Springer Series in Computational Mathematics, vol. 29, Springer-Verlag, Berlin, Heidelberg, 2001.

[Whi74] Gerald B. Whitham, *Linear and nonlinear waves*, Pure and applied mathematics, Wiley-Interscience, New York, 1974.

[Wor10] Anke Wortmann, *Flux-corrected transport algorithms and their application in the simulation of relativistic laser-plasma interaction*, Master's thesis, Heinrich-Heine-Universität Düsseldorf, 2010.

[Yee66] Kane S. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. Antennas and Propagation **14** (1966), 302–307.

[Zal79] Steven T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, Journal of Computational Physics **31** (1979), 335–362.