UCLA

COMPUTATIONAL AND APPLIED MATHEMATICS

Multiphase Computations in Geometrical Optics

Olof Runborg

December 1996

CAM Report 96-52

Department of Mathematics
University of California, Los Angeles
Los Angeles, CA. 90024-1555

Multiphase Computations in Geometrical Optics

Olof Runborg

Royal Institute of Technology
Department of Numerical Analysis and Computing Science
S-100 44 Stockholm, Sweden
Email: olofr@nada.kth.se

Abstract

In this work we propose a new set of partial differential equations (PDEs) which can be seen as a generalization of the classical eikonal and transport equations, to allow for solutions with multiple phases. The traditional geometrical optics pair of equations suffer from the fact that the class of physically relevant solutions is limited. In particular, it does not include solutions with multiple phases, corresponding to crossing waves. Our objective has been to generalize these equations to accommodate solutions containing more than one phase. The new equations are based on the same high frequency approximation of the scalar wave equation as the eikonal and the transport equations. However, they also incorporate a finite superposition principle. The maximum allowed number of intersecting waves in the solution can be chosen arbitrarily, but a higher number means that a larger system of PDEs must be solved. The PDEs form a hyperbolic system of conservation laws with source terms. Although the equations are only weakly hyperbolic, and thus not well-posed in the strong sense, several examples show the viability of solving the equations numerically. The technique we use to capture multi-valued solutions is based on a closure assumption for a system of equations representing the moments.

1 Introduction

In the direct calculation of wave propagation, the computational effort is larger at higher frequencies. With constant accuracy the work grows algebraically with frequency. For sufficiently high frequencies or short wavelengths it is unrealistic to compute the wave field directly. Fortunately, this is often the regime for which high-frequency asymptotic approximations are quite accurate.

Generically, phase and amplitude vary on a much slower scale than the dependent variables in the original wave equations and are thus in principle easier to compute. The geometrical optics type asymptotic expansions are used in many applications, for example in electromagnetic, elastic and acoustic wave propagation.

Traditionally, ray tracing has been the computational method of choice. Recently, however, the geometrical optics approximations are also being solved by partial differential equation (PDE) techniques. This is e.g. done in [5] and within the framework of seismology in [9], [11] and [12]. The PDEs give only one unique phase at each point in space. In this paper we shall derive equations which allow for multiple phases or crossing rays. The equations are based on the closure assumption of a finite number of crossing rays for the kinetic formulation of geometrical optics.

1.1 High-Frequency Asymptotics

When high-frequency waves are treated, the computations can be simplified by considering the asymptotic behavior of the solution as the frequency tends to infinity. There are two strongly related ways to formulate this approximation: the PDEs of geometrical optics and ray tracing. Typical wave phenomena, such as diffraction and interference, are lost in the leading terms of the high-frequency approximation.

Classical geometrical optics is based on the scalar wave equation,

$$u_{tt} + c\nabla^2 u = 0. \tag{1}$$

Here $c = c(\boldsymbol{x})$ is the local wave velocity of the medium. We also define the *index of refraction* as $\eta = c_0/c$ with the reference velocity c_0 (e.g. the speed of light in vacuum). Geometrical optics considers the case when the solution to (1) can be written as a series expansion of the form:

$$u = e^{i\omega\phi(\boldsymbol{x},t)} \sum_{k=0}^{\infty} w_k(\boldsymbol{x},t)(i\omega)^{-k}. \tag{2}$$

Entering this expression into (1) and summing terms of the same order in ω, to zero, we obtain separate equations for the unknown variables in (2). The phase function ϕ will satisfy the *eikonal equation*,

$$\phi_t + c|\nabla\phi| = 0, \tag{3}$$

and the amplitude coefficients w_k solve the *transport equations*,

$$(w_0)_t + c\frac{\nabla\phi \cdot \nabla w_0}{|\nabla\phi|} + \frac{c^2\nabla^2\phi - \phi_{tt}}{2c|\nabla\phi|} w_0 = 0, \tag{4}$$

$$(w_{k+1})_t + c\frac{\nabla\phi \cdot \nabla w_{k+1}}{|\nabla\phi|} + \frac{c^2\nabla^2\phi - \phi_{tt}}{2c|\nabla\phi|} w_{k+1} + \frac{c^2\nabla^2 w_k - (w_k)_{tt}}{2c|\nabla\phi|} = 0. \tag{5}$$

For large ω only the first term in the expansion (2) is significant, and the problem is reduced to computing the phase ϕ and the first amplitude term w_0. Note that once ϕ is known, the transport equations are linear equations with variable coefficients. Solving (3) and (4) can be done by finite difference methods.

The problem with the geometrical optics approach is that the class of solutions which justify an expansion of the type (2), is limited. In particular, it does not include solutions with multiple phases, corresponding to crossing waves. In fact, even in the case of a single phase solution, the series does not necessarily converge, for instance when the geometric boundaries create diffraction effects. We shall concentrate on the multiple phase problem and assume the geometrical optics approximations of (3) and (4).

The eikonal equation is a nonlinear PDE which requires extra conditions to have a unique solution. This solution is known as the *viscosity solution* [3]. Of course, it does not have to agree with the correct physical solution in all cases. At points where the correct solution should have a multi-valued phase, the viscosity solution picks out the phase corresponding to the first arriving wave.

The eikonal equation's inability to capture multi-phase solutions is related to its nonlinear character. In the case of the linear wave equation, that it approximates, a linear combination of solutions is also a solution. For the nonlinear eikonal equation, this *superposition principle* does not hold. An example is shown in Figure 1.

Solving the eikonal equation numerically as a PDE instead of using ray tracing has recently been used in seismology. This technique is demonstrated in [9], [11] and [12]. For these applications it is of direct interest to determine the first arrival.

A second phase, corresponding to crossing rays was calculated in [5] using two separate eikonal equations. Boundary conditions for the second phase was given at the discontinuity of the first phase or at a geometric reflecting boundary. This boundary could be difficult to determine.

Another way to treat high-frequency waves computationally is through ray tracing, which is based on a kinetic formulation. The waves are postulated to be particles (photons) whose trajectories are rays. The *ray vector*, $\boldsymbol{p}$, is defined as the index of refraction multiplied by the unit vector, $\hat{\boldsymbol{s}}$, in the direction of the ray, i.e. $\boldsymbol{p} = \eta\hat{\boldsymbol{s}}$. For simplicity we will henceforth let $c_0 = 1$, so that the velocity vector $\boldsymbol{v} = c\hat{\boldsymbol{s}} = c^2\boldsymbol{p}$. A transport equation for particles in the space $(\boldsymbol{x}, \boldsymbol{p}, t)$ can then be derived. Denoting the density of particles by $f(\boldsymbol{x}, \boldsymbol{p}, t)$ the evolution of f is described by the Vlasov type equation

$$f_t + \boldsymbol{v} \cdot \boldsymbol{\nabla}_x f + c\boldsymbol{\nabla}_x \eta \cdot \boldsymbol{\nabla}_p f = 0. \tag{6}$$

Tracing the particle trajectories of (6) corresponds to ray tracing and also to the method of characteristics for (3) and (6). Since (6) is linear the superposition principle is valid.

Because of the large number of independent variables (six in 3D) it is very hard numerically to solve the full equation (6). If the equation is solved using ray tracing it is difficult to cover the full domain with rays, [11]. There will often be shadow zones where the field cannot be resolved. It is also hard to determine the derivative of ϕ, which is needed when computing the amplitude.

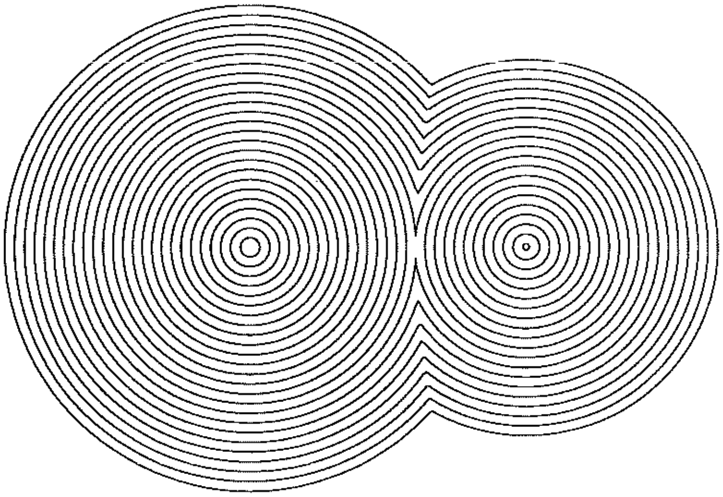


Figure 1: Level curves of ϕ in the solution of the eikonal equation (3) for two interacting waves. Note how the superposition principle does not hold. Instead, the first arriving wave takes precedence over the second at each point.

1.2 Moment Formulation

In this paper we propose a middle way between geometrical optics and the kinetic model. It is a high-frequency approximation through which the whole field can be solved. Moreover, the superposition principle holds up to a point; the maximum allowed number of intersecting waves can be chosen arbitrarily, but a higher number means that a larger system of PDEs must be solved. The technique we use to capture multi-valued solutions is based on a closure assumption for a system of equations representing the moments (see [1]).

The starting point for this approach is the transport equation (6). Instead of solving the full equation in phase space, we observe that when f is of a simple form in $\boldsymbol{p}$, we can transform (6) to a finite system of moment equations in the reduced space $(\boldsymbol{x}, t)$, analogously to the classical approach of the hydrodynamic limit from a kinetic formulation. In particular we are interested in cases where, for given $\boldsymbol{x}$ and t, the density function f is non-zero only for a finite number of $\boldsymbol{p}$. This corresponds to a finite number of rays in different directions at each point.

This paper is organized as follows: In Section 2 the moment equations are derived from the kinetic model for high-frequency waves. They are equivalent to the equations of geometrical optics. We also explore some theoretical issues and find that the resulting hyperbolic equations are not well-posed in the strong sense. Existence of solutions of unbounded variation is indicated. Moreover, we give a proof that the derived system of equations is essentially closed. We also present the solution of the Riemann problem. Next, in Section 3, we describe the numerical approximations we have used to solve these equations. The standard Lax-Friedrich method gives satisfactory results. More elaborate, and less viscous, methods like the Godunov method and the second order TVD Nessyahu-Tadmor scheme, together with dimensional splitting, suffer from problems locally and converge poorly in L_∞, although they are converging well in L_1. A genuinely two-dimensional version of the Nessyahu-Tadmor scheme gives the best result. For the multiple phase system a few additional numerical problems are added, related to the fact that a non-linear system of equations needs to be solved at each point in each time-step. Solving these equations requires some care. When $N = 2$ it can be done analytically. Otherwise, the Lax-Friedrich method works well also in the multiple phase case. We present computational results for homogeneous and inhomogeneous problems in Section 4.

2 Derivation of the Moment Equations

In this section we will derive the system of PDEs that follows from the kinetic model and the assumption that a maximum of N rays pass through any given point in space. The analysis is carried out in two-dimensional space.

The derivation of the moment equations is based on the transport equation (6). This equation comes from the Hamiltonian system

$$\frac{d\boldsymbol{x}}{dt} = \nabla_p H(\boldsymbol{x},\boldsymbol{p}), \qquad \frac{d\boldsymbol{p}}{dt} = -\nabla_x H(\boldsymbol{x},\boldsymbol{p}), \tag{7}$$

where the corresponding density function $f(\boldsymbol{x},\boldsymbol{p},t)$ solves

$$f_t + \nabla_x \cdot (f\nabla_p H) - \nabla_p \cdot (f\nabla_x H) = 0. \tag{8}$$

The equation (6) follows from (8) when

$$H = \frac{1}{2}\left[\frac{|\boldsymbol{p}|^2}{\eta^2} - 1\right]. \tag{9}$$

The form of H, somewhat unusual in the geometrical optics context, was chosen to ensure that the t variable corresponds to actual, unscaled, time.

2.1 The Moment Equations

We start by defining the moments m_{ij}. With $\boldsymbol{p} = (p_1, p_2)$, let

$$m_{ij} = \int_{\mathbb{R}^2} p_1^i p_2^j f d\boldsymbol{p}. \tag{10}$$

Next, we multiply (6) by $p_1^i p_2^j$ and integrate over $\mathbb{R}^2$ with respect to $\boldsymbol{p}$. Using the definition (10) we get the moment equation,

$$(\eta^2 m_{ij})_t + (m_{i+1,j})_x + (m_{i,j+1})_y = i\eta\eta_x m_{i-1,j} + j\eta\eta_y m_{i,j-1}, \qquad (11)$$

where we have used the fact that f has compact support in $\boldsymbol{p}$. Since this equation is valid for all $i, j \geq 0$, we have an infinite system of moment equations. For uniformity in notation we have defined $m_{i,-1} = m_{-1,i} = 0$, $\forall i$.

The system (11) is not closed. If truncated at finite i and j, there are more unknown than equations. To close (11) we use the assumption that for fixed values of $\boldsymbol{x}$ and t, the particle density f is non-zero at a maximum of N points, and only when $|\boldsymbol{p}| = \eta(\boldsymbol{x})$. Thus f can be written

$$f(\boldsymbol{x}, \boldsymbol{p}, t) = \sum_{k=1}^{N} g_k \cdot \delta(|\boldsymbol{p}| - \ \eta, \arg \boldsymbol{p} - \theta_k). \qquad (12)$$

The new variables that we have introduced here are $g_k = g_k(\boldsymbol{x}, t)$, which corresponds to the strength (particle density) of ray k, and $\theta_k = \theta_k(\boldsymbol{x}, t)$ which is the direction of the same ray. Inserting (12) into (10) yields

$$m_{ij} = \sum_{k=1}^{N} \eta^{i+j} g_k \cos^i \theta_k \sin^j \theta_k, \qquad (13)$$

which is the expression for the moments that we will use.

A system describing N phases, needs $2N$ equations, corresponding to the N ray strengths g_k and their directions θ_k. It is not immediately clear which equations to select among the candidates in (11). Given the equations for a set of $2N$ moments, it must be possible to write the remaining moments of these equations in terms of the leading ones. This is not always true. For instance, with the choice of m_{20} and m_{02}, for $N = 1$, the quadrant of the angle θ cannot be recovered, and therefore in general not the sign of the moments.

We choose here the equations for the moments $m_{2k-1,0}$ and $m_{0,2k-1}$ with $k = 1, \ldots, N$. This system can be essentially closed for all N (see Section 2.3). After scaling the moments, $\eta^{i+j}\tilde{m}_{ij} = m_{ij}$, those equations take the following form (η is assumed to be smooth):

$$\begin{aligned}
&(\eta^2 \tilde{m}_{2k-1,0})_t + (\eta \tilde{m}_{2k,0})_x + (\eta \tilde{m}_{2k-1,1})_y \\
&\qquad\qquad = (2k-1)(\eta_x \tilde{m}_{2k-2,0} - \eta_x \tilde{m}_{2k,0} - \eta_y \tilde{m}_{2k-1,1}), \qquad (14) \\
&(\eta^2 \tilde{m}_{0,2k-1})_t + (\eta \tilde{m}_{1,2k-1})_x + (\eta \tilde{m}_{0,2k})_y \\
&\qquad\qquad = (2k-1)(\eta_y \tilde{m}_{0,2k-2} - \eta_x \tilde{m}_{1,2k-1} - \eta_y \tilde{m}_{0,2k}). \qquad (15)
\end{aligned}$$

To simplify notation, we will henceforth write m_{ij} for $\tilde{m}_{ij}$. We introduce new variables,

$$\begin{aligned}
\boldsymbol{u} &= (u_1, u_2, u_3, u_4 \ldots, u_{2N-1}, u_{2N})^T \\
&:= (g_1 \cos\theta_1, g_1 \sin\theta_1, g_2 \cos\theta_2, g_2 \sin\theta_2, \ldots, g_N \cos\theta_N, g_N \sin\theta_N)^T. \qquad (16)
\end{aligned}$$

The $\boldsymbol{u}$ variables have a physical interpretation; the vector (u_{2k-1}, u_{2k}) shows the direction and strength of ray k. These variables permit us to write the equations as a system of nonlinear conservation laws with source terms,

$$\boldsymbol{F}_0(\eta^2\boldsymbol{u})_t + \boldsymbol{F}_1(\eta\boldsymbol{u})_x + \boldsymbol{F}_2(\eta\boldsymbol{u})_y = \boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y), \tag{17}$$

where the functions $\boldsymbol{F}_k$ and $\boldsymbol{K}$ are rather complicated nonlinear functions, which depend on the particular choice of moments above. For other choices they would be different. Equivalently, with

$$\hat{\boldsymbol{m}} = (m_{10}, m_{01}, m_{30}, m_{03}, \dots, m_{2N-1,0}, m_{0,2N-1})^T, \tag{18}$$

we could write

$$(\eta^2\hat{\boldsymbol{m}})_t + \boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}(\eta\hat{\boldsymbol{m}})_x + \boldsymbol{F}_2 \circ \boldsymbol{F}_0^{-1}(\eta\hat{\boldsymbol{m}})_y = \boldsymbol{K}(\boldsymbol{F}_0^{-1}(\hat{\boldsymbol{m}}), \eta_x, \eta_y). \tag{19}$$

Since the angles θ_k remain unaffected when $\boldsymbol{u}$ is scaled by a constant, all $\boldsymbol{F}_k$ and $\boldsymbol{K}$ are homogeneous, $\boldsymbol{F}_k(\alpha\boldsymbol{u}) = \alpha\boldsymbol{F}_k(\boldsymbol{u})$, $\boldsymbol{K}(\alpha\boldsymbol{u}, \eta_x, \eta_y) = \alpha\boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y)$ for all $\alpha \in \mathbb{R}$.

To find a concise expression for the functions $\boldsymbol{F}_k$ and $\boldsymbol{K}$, we need a few definitions. Let I be the 2×2 identity matrix, and

$$D_k = \begin{pmatrix} \cos\theta_k & 0 \\ 0 & \sin\theta_k \end{pmatrix}, \qquad A = \begin{pmatrix} I & I & \cdots & I \\ D_1^2 & D_2^2 & \cdots & D_N^2 \\ D_1^4 & D_2^4 & \cdots & D_N^4 \\ \vdots & \vdots & \ddots & \vdots \\ D_1^{2N-2} & D_2^{2N-2} & \cdots & D_N^{2N-2} \end{pmatrix}. \tag{20}$$

Moreover,

$$\begin{aligned} D &= \operatorname{diag}(D_1, D_2, \dots, D_N), & (21)\\ C &= \operatorname{diag}(\cos\theta_1 I, \cos\theta_2 I, \dots, \cos\theta_N I), & (22)\\ S &= \operatorname{diag}(\sin\theta_1 I, \sin\theta_2 I, \dots, \sin\theta_N I), & (23)\\ T &= \operatorname{diag}(I, 3I, \dots, (2N-1)I), & (24)\\ R &= \operatorname{diag}(\underbrace{\eta_x, \eta_y, \dots, \eta_x, \eta_y}_{2N \text{ elements}}). & (25) \end{aligned}$$

Noting that R and A commute, $RA = AR$, we get

$$\boldsymbol{F}_0(\boldsymbol{u}) = A\boldsymbol{u}, \qquad \boldsymbol{F}_1(\boldsymbol{u}) = AC\boldsymbol{u}, \qquad \boldsymbol{F}_2(\boldsymbol{u}) = AS\boldsymbol{u}, \tag{26}$$

$$\boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y) = TA(RD^{-1} - \eta_x C - \eta_y S)\boldsymbol{u}. \tag{27}$$

It should be observed that the source term, $\boldsymbol{K}$, always vanishes for constant η. In the most simple case, $N = 1$, the function $\boldsymbol{F}_0$ is the identity and

$$\boldsymbol{F}_1 = \begin{pmatrix} \frac{u_1^2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}, \quad \boldsymbol{F}_2 = \begin{pmatrix} \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_2^2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}, \quad \boldsymbol{K} = \begin{pmatrix} \frac{\eta_x u_2^2 - \eta_y u_1 u_2}{\sqrt{u_1^2+u_2^2}} \\ \frac{\eta_y u_1^2 - \eta_x u_1 u_2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}. \tag{28}$$

In Section 3.2 the functions are given explicitly also for the case $N = 2$.

2.2 A Comparison with Geometrical Optics

To see how the moment equations (17) are equivalent formulations of the equations of geometrical optics, we present the following derivation for smooth solutions.

The additional definitions

$$
\begin{aligned}
\Theta &= \mathrm{diag}(\theta_1 I, \theta_2 I, \dots, \theta_N I), && (29)\\
\hat{\boldsymbol{g}} &= \underbrace{(g_1, g_1, g_2, g_2, \dots, g_N, g_N)}_{2N \text{ elements}}{}^T, && (30)
\end{aligned}
$$

will help us write a separated form of (17). We insert (26, 27) in (17) and note that $\boldsymbol{u} = D\hat{\boldsymbol{g}}$. Since each element of the matrices only depend on one variable (one of the θ_ks), we can let the prime sign denote elementwise differentiation of a matrix. Using the identities $(AD)' = TAD'$ and $R = (\eta_x D + \eta_y D')C + (\eta_y D - \eta_x D')S$, the equations (17) can be written as

$$
\begin{aligned}
AD(\eta^2 \hat{g}_t + (\eta C\hat{g})_x + (\eta S\hat{g})_y)&\\
+ (AD)'(\eta^2\Theta_t + (\eta S)_x - (\eta C)_y)\hat{g} = 0.& \qquad (31)
\end{aligned}
$$

Noting that C, S and Θ are diagonal and that they all, together with $\hat{g}$, have their elements ordered pairwise, a solution is given by solving the N separated systems

$$
\left.
\begin{aligned}
\eta^2(\theta_k)_t + (\eta \sin\theta_k)_x - (\eta\cos\theta_k)_y &= 0,\\
\eta^2(g_k)_t + (\eta g_k \cos\theta_k)_x + (\eta g_k \sin\theta_k)_y &= 0.
\end{aligned}
\right.
\qquad k = 1, \dots, N. \qquad (32)
$$

On the other hand, after some algebraic manipulations of (3) and (4) we get

$$
\begin{aligned}
\eta^2 \frac{\partial}{\partial t} \arctan\left(\frac{\phi_y}{\phi_x}\right) + \frac{\partial}{\partial x}\left(\eta \frac{\phi_y}{|\boldsymbol{\nabla}\phi|}\right) - \frac{\partial}{\partial y}\left(\eta \frac{\phi_x}{|\boldsymbol{\nabla}\phi|}\right) &= 0,\\
\eta^2 \frac{\partial}{\partial t} w_0^2 + \frac{\partial}{\partial x}\left(\eta w_0^2 \frac{\phi_x}{|\boldsymbol{\nabla}\phi|}\right) + \frac{\partial}{\partial y}\left(\eta w_0^2 \frac{\phi_y}{|\boldsymbol{\nabla}\phi|}\right) &= 0.
\end{aligned}
\qquad (33)
$$

If we identify the variables of (32) as

$$
g_k = w_{0,k}^2, \qquad \cos\theta_k = \frac{(\phi_k)_x}{|\boldsymbol{\nabla}\phi_k|}, \qquad \sin\theta_k = \frac{(\phi_k)_y}{|\boldsymbol{\nabla}\phi_k|}, \qquad k = 1, \dots, N, \quad (34)
$$

they will solve (32) and hence (17). As expected the vector $(u_{2k-1}, u_{2k})^T$ points in the direction of the gradient of ϕ_k. The length of the vector corresponds to the first amplitude coefficient squared.

2.3 Closure of the Moment Equations

In this section we will show some conditions under which the system (17) is closed. The choice of moment equations is a deciding factor, and the result of the analysis serves to motivate our particular choice.

We start by observing that the system is closed if and only if the flux functions

$$\boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}(\hat{\boldsymbol{m}}) \qquad \text{and} \qquad \boldsymbol{F}_2 \circ \boldsymbol{F}_0^{-1}(\hat{\boldsymbol{m}}) \tag{35}$$

are well defined for all solutions, $\hat{\boldsymbol{m}}$, to (19) at all times. Here, $\hat{\boldsymbol{m}}$ is the vector of moments whose corresponding equations we have chosen, not necessarily the same as in (18). It is not clear that these rather strict requirements can be satisfied for any choice of moment equations (which determines the functions $\boldsymbol{F}_k$). For the choice we made in Section 2.1, that is the moments in (18), we can, however, prove that the system is closed as long as no two rays meet head-on, or more precisely, it is closed for all t such that

$$\theta_k(\boldsymbol{x}, t) \neq \theta_\ell(\boldsymbol{x}, t) + \pi, \qquad \forall \boldsymbol{x}, k, \ell. \tag{36}$$

To show this, we will first prove a lemma. In what follows, we will need the definition of the function $\boldsymbol{\sigma}_n$:

$$\boldsymbol{\sigma}_n : \mathbb{C} \to \mathbb{C}^n, \qquad \boldsymbol{\sigma}_n(z) = \begin{pmatrix} z \\ z^{-3} \\ \vdots \\ z^{(2n-1)(-1)^{n+1}} \end{pmatrix}, \qquad \text{for } z \in \mathbb{C}, \tag{37}$$

Lemma 1. *Let z_k, $k = 1, \ldots, M$ be complex numbers such that $|z_k| = 1$ Define the corresponding vectors $\boldsymbol{z}_k = \boldsymbol{\sigma}_N(z_k)$ with $M \leq 2N$ and $\boldsymbol{\sigma}_N$ as in (37). Then $\boldsymbol{z}_k \in \mathbb{C}^N$ are linearly independent over $\mathbb{R}$ if and only if*

$$z_k^2 \neq z_\ell^2, \qquad k \neq \ell. \tag{38}$$

Proof. The necessity is obvious, since if $z_k^2 = z_\ell^2$ for some k, ℓ, then $\boldsymbol{z}_k = \pm \boldsymbol{z}_\ell$. To show that (38) is a sufficient condition, let (38) hold and take $M = 2N$. Suppose that $\boldsymbol{z}_k$ are linearly dependent. Then we can find α_k such that

$$\sum_{k=1}^{2N} \alpha_k \boldsymbol{z}_k = 0, \qquad \alpha_k \in \mathbb{R}. \tag{39}$$

where not all α_k are zero. By forming the real matrix

$$A = \begin{pmatrix} | & | & & | \\ \Re(\boldsymbol{z}_1) & \Re(\boldsymbol{z}_2) & \cdots & \Re(\boldsymbol{z}_{2N}) \\ | & | & & | \\ | & | & & | \\ \Im(\boldsymbol{z}_1) & \Im(\boldsymbol{z}_2) & \cdots & \Im(\boldsymbol{z}_{2N}) \\ | & | & & | \end{pmatrix}, \qquad A \in \mathbb{R}^{2N \times 2N} \tag{40}$$

we have the equivalent formulation

$$A\boldsymbol{\alpha} = 0, \qquad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{2N})^T, \tag{41}$$

and we see that since $\boldsymbol{\alpha} \neq 0$, the matrix A must be singular. Then so is A^T and we can find a vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{2N})^T \neq 0$ such that $A^T\boldsymbol{\beta} = 0$. Now, since $|z_k| = 1$, we have that $\bar{z}_k = 1/z_k$ and

$$\Re(z_k) = \frac{1}{2}(\boldsymbol{\sigma}_N(z_k) + \boldsymbol{\sigma}_N(\frac{1}{z_k})), \qquad \Im(z_k) = \frac{1}{2i}(\boldsymbol{\sigma}_N(z_k) - \boldsymbol{\sigma}_N(\frac{1}{z_k})). \tag{42}$$

Hence, $A^T\boldsymbol{\beta} = 0$ can be written

$$\frac{1}{2}\sum_{\ell=1}^{N} \beta_\ell (z_k^{2\ell-1} + \frac{1}{z_k^{2\ell-1}}) + \frac{1}{2i}\sum_{\ell=1}^{N} (-1)^{\ell+1}\beta_{\ell+N}(z_k^{2\ell-1} - \frac{1}{z_k^{2\ell-1}}) = 0, \tag{43}$$

for $k = 1, \dots, 2N$. Since $z_k \neq 0$ we can multiply (43) by z_k^{2N-1} and by defining the polynomial

$$P_{\boldsymbol{\beta}}(z) = \frac{1}{2}\sum_{\ell=1}^{N} \beta_\ell (z^{\ell+N-1} + z^{N-\ell}) + \frac{1}{2i}\sum_{\ell=1}^{N} (-1)^{\ell+1}\beta_{\ell+N}(z^{\ell+N-1} - z^{N-\ell}), \tag{44}$$

we have that (41) implies

$$P_{\boldsymbol{\beta}}(z_k^2) = 0, \qquad k = 1, \dots, 2N \tag{45}$$

for some $\boldsymbol{\beta}$. But since the degree of $P_{\boldsymbol{\beta}}$ is at most $2N - 1$ it cannot have $2N$ distinct zeros, regardless of the choice of $\boldsymbol{\beta}$. Therefore, there must exist k, ℓ such that $z_k^2 = z_\ell^2$, a contradiction. Hence, $\boldsymbol{z}_k$ are linearly independent over $\mathbb{R}$ when $M = 2N$ and (38) holds.

Finally, if $M < 2N$ and (38) holds, we can always find $2N - M$ additional z_k such that (38) still holds. A subset of a set of linearly independent vectors are also linearly independent, from which the lemma follows. □

We now introduce complex versions of our variables

$$z_k = \cos\theta_k + i\sin\theta_k, \qquad \boldsymbol{m} = \begin{pmatrix} m_{10} + im_{01} \\ m_{30} + im_{03} \\ \vdots \\ m_{2N-1,0} + im_{0,2N-1} \end{pmatrix}, \tag{46}$$

so that

$$g_k z_k = u_{2k-1} + iu_{2k}. \tag{47}$$

Furthermore, let

$$Z = \begin{pmatrix} | & | & & | \\ \boldsymbol{\sigma}_N(z_1) & \boldsymbol{\sigma}_N(z_2) & \dots & \boldsymbol{\sigma}_N(z_N) \\ | & | & & | \end{pmatrix}, \qquad \boldsymbol{g} = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_N \end{pmatrix}. \tag{48}$$

In general we have

$$(\cos\theta + (-1)^{k+1} i \sin\theta)^{2k-1} = \sum_{\ell=0}^{k-1} w_{k,k-\ell}(\cos^{2(k-\ell)-1}\theta + i\sin^{2(k-\ell)-1}\theta), \quad (49)$$

where $w_{k\ell}$ is the $(2\ell-1)^{\text{th}}$ coefficient of the $(2k-1)^{\text{th}}$ degree Chebyshev polynomial. Setting $w_{k\ell} = 0$ for $\ell > k$ we can form a lower triangular matrix $W = \{w_{k\ell}\}$. This matrix is non-singular since $w_{kk} = 4^{k-1} > 0$. Thereby, the two equations

$$\boldsymbol{F}_0(\boldsymbol{u}) = \hat{\boldsymbol{m}} \qquad \Leftrightarrow \qquad Z\boldsymbol{g} = W\boldsymbol{m}, \qquad (50)$$

are equivalent if the variables are identified as in (47). We are now ready to prove the following theorem.

Theorem 1. *Let* $\boldsymbol{F}_k$ *be the functions in (17), corresponding to the moment vector* $\hat{\boldsymbol{m}}$ *defined in (18). Let* $\boldsymbol{F}_0|D_f$ *be the restriction of* $\boldsymbol{F}_0$ *to the domain*

$$D_f = \{\boldsymbol{u} \in \mathbb{R}^{2N} \mid z_k + z_\ell \neq 0,\ \forall k, \ell\}, \qquad (51)$$

with z_k *defined as in (46). For* $\hat{\boldsymbol{m}} \in \boldsymbol{F}_0(D_f)$ *the flux functions*

$$\boldsymbol{F}_1 \circ (\boldsymbol{F}_0|D_f)^{-1}(\hat{\boldsymbol{m}}), \qquad \boldsymbol{F}_2 \circ (\boldsymbol{F}_0|D_f)^{-1}(\hat{\boldsymbol{m}}) \qquad (52)$$

are well defined.

Proof. Let $\boldsymbol{u}, \tilde{\boldsymbol{u}} \in D_f$ be two solutions to (50), so that

$$\boldsymbol{F}_0(\boldsymbol{u}) = \hat{\boldsymbol{m}} = \boldsymbol{F}_0(\tilde{\boldsymbol{u}}), \qquad \hat{\boldsymbol{m}} \in \boldsymbol{F}_0(D_f). \qquad (53)$$

We need to show that $\boldsymbol{F}_1(\boldsymbol{u}) = \boldsymbol{F}_1(\tilde{\boldsymbol{u}})$ and $\boldsymbol{F}_2(\boldsymbol{u}) = \boldsymbol{F}_2(\tilde{\boldsymbol{u}})$. Using the corresponding z, Z and g-variables in (46) and (48), (53) is equivalent to

$$Z\boldsymbol{g} = \tilde{Z}\tilde{\boldsymbol{g}}, \qquad (54)$$

by (50). With $\boldsymbol{z}_k = \boldsymbol{\sigma}_N(z_k)$ and $\tilde{\boldsymbol{z}}_k = \boldsymbol{\sigma}_N(\tilde{z}_k)$, (54) implies that

$$\sum_{k=1}^{M} g_k \boldsymbol{z}_k = \sum_{k=1}^{\tilde{M}} \tilde{g}_k \tilde{\boldsymbol{z}}_k. \qquad (55)$$

Here M and $\tilde{M}$ denote the number of nonzero g_k and $\tilde{g}_k$ in the respective solutions. Now, let J and $\tilde{J}$ be the number of distinct z_k and $\tilde{z}_k$ in the solutions. Without loss of generality we order the variables such that $z_{\ell_j} = \ldots = z_{\ell_{j+1}-1}$, with $1 = \ell_1 < \ldots < \ell_{J+1} = M+1$, and similar for the second solution. With this notation we get

$$\sum_{j=1}^{J} \left(\sum_{k=\ell_j}^{\ell_{j+1}-1} g_k \right) \boldsymbol{z}_{\ell_j} = \sum_{j=1}^{\tilde{J}} \left(\sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} \tilde{g}_k \right) \tilde{\boldsymbol{z}}_{\tilde{\ell}_j}. \qquad (56)$$

The sets of numbers $\{z_{\ell_j}\}_{j=1}^{J}$ and $\{\tilde{z}_{\tilde{\ell}_j}\}_{j=1}^{\tilde{J}}$ both satisfy (38) and the corresponding vectors are thus linearly independent by Lemma 1. But since $J + \tilde{J} \leq 2N$, again by Lemma 1, there must exist j and k such that $z_{\ell_j}^2 = \tilde{z}_{\tilde{\ell}_k}^2$. Since $g_k, \tilde{g}_k > 0$, in fact we must have that $z_{\ell_j} = \tilde{z}_{\tilde{\ell}_k}$. By induction it follows that $J = \tilde{J}$ and, possibly after some reordering,

$$\ell_j = \tilde{\ell}_j, \qquad \boldsymbol{z}_{\ell_j} = \tilde{\boldsymbol{z}}_{\tilde{\ell}_j}, \qquad \sum_{k=\ell_j}^{\ell_{j+1}-1} g_k = \sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} \tilde{g}_k, \qquad \forall j. \tag{57}$$

Thus, the solutions are identical up to permutations and to the individual g_k values. However, this ambiguity is resolved when we apply $\boldsymbol{F}_1$ or $\boldsymbol{F}_2$ to the solution. For $\boldsymbol{F}_1$ we have from (26) that $\boldsymbol{F}_1(\boldsymbol{u}) = \boldsymbol{F}_0(C\boldsymbol{u})$. Using (50), substituting g_k for $g_k \cos\theta_k$, we get via (54), (56) and (57) that $\boldsymbol{F}_1(\boldsymbol{u}) = \boldsymbol{F}_1(\tilde{\boldsymbol{u}})$ if

$$\sum_{j=1}^{J} \left(\sum_{k=\ell_j}^{\ell_{j+1}-1} g_k \cos\theta_k \right) \boldsymbol{z}_{\ell_j} = \sum_{j=1}^{J} \left(\sum_{k=\ell_j}^{\ell_{j+1}-1} \tilde{g}_k \cos\tilde{\theta}_k \right) \boldsymbol{z}_{\ell_j}. \tag{58}$$

This holds indeed, since

$$\cos\tilde{\theta}_k = \cos\tilde{\theta}_{\ell_j} = \cos\theta_{\ell_j} = \cos\theta_k, \qquad \text{for } \ell_j \leq k < \ell_{j+1}. \tag{59}$$

That $\boldsymbol{F}_2(\boldsymbol{u}) = \boldsymbol{F}_2(\tilde{\boldsymbol{u}})$ follows in the same way. □

Remark. With a different choice of $\hat{\boldsymbol{m}}$, Theorem 1 does not necessarily hold. For instance, with $N = 2$ and

$$\hat{\boldsymbol{m}} = (m_{10}, m_{01}, m_{20}, m_{02})^T, \tag{60}$$

there are in general two unrelated solutions to (50), which $\boldsymbol{F}_1$ does not map to the same point. This is exemplified by

$$\boldsymbol{u} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ -1 \end{pmatrix}, \ \tilde{\boldsymbol{u}} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix} \qquad \Rightarrow \qquad \boldsymbol{F}_0(\boldsymbol{u}) = \boldsymbol{F}_0(\tilde{\boldsymbol{u}}) = \begin{pmatrix} 1 \\ 0 \\ \frac{1}{\sqrt{2}} \\ 1 + \frac{1}{\sqrt{2}} \end{pmatrix} \tag{61}$$

but

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = \boldsymbol{F}_1(\boldsymbol{u}) \neq \boldsymbol{F}_1(\tilde{\boldsymbol{u}}) = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}. \tag{62}$$

The function $\boldsymbol{F}_2 \circ \boldsymbol{F}_0^{-1}$ is ill defined in the same way.

2.4 Analysis of the Conservation Laws

For simplicity we will mainly deal with the single phase, $N = 1$, one-dimensional case where the medium is vacuum, $\eta = \text{const} = 1$. With the $\boldsymbol{u}$ variables defined in (16) as conserved quantities, the system can be written on the standard form of a conservation law,

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \qquad \boldsymbol{f}(\boldsymbol{u}) = \begin{pmatrix} \frac{u_1^2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}. \tag{63}$$

The Jacobian of $\boldsymbol{f}$ with respect to $\boldsymbol{u}$ has the following form:

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} = G \begin{pmatrix} \cos\theta & -\sin\theta \\ 0 & \cos\theta \end{pmatrix} G^{-1}, \qquad G = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{64}$$

Thus, the linearized problem has a double real eigenvalue, $\cos\theta$, and an incomplete set of eigenvectors; the system (63) is only *weakly hyperbolic*. In general this means that (63) is not well-posed in the strongly hyperbolic sense. The system is likely to be much more sensitive than regular hyperbolic systems. The solution of the linearized problem with frozen coefficients loses one derivative. The L_2 norm of the solution at time $t > t_0$ can be estimated in terms of the H_1 norm of the initial data at time $t = t_0$. The sensitivity of the equations is reflected in difficulties in finding stable numerical methods to solve them (see Section 3).

The existence of solutions to (63) is also an open question. It appears that solutions cannot be expected to be of bounded variation. In fact, analytic and numerical evidence suggest that (63) can have measure solutions, i.e. of delta function type (cf. Figure 7). An extended solution concept is needed to accommodate measure solutions. This problem was addressed in [2] and [4], where also existence of such solutions for certain conservation laws was proved. Entropy conditions and uniqueness of solutions to (63) are even more uncertain.

The appearance of a delta function is closely linked to when the physically correct solution passes outside the class of solutions that the system (17) describe. If initial data dictates a physical solution with N phases for $t > T$, the system (17) with $M < N$ phases will have a measure solution for $t > T$. In the case of (63), a delta function will appear in the solution when multiple phases are present.

The statements above are supported by our numerical simulations. We will consider a one-dimensional example. In vacuum, $\eta = 1$, the separated system (32) can be rewritten as:

$$\begin{aligned} (\theta_k)_t + (\sin\theta_k)_x &= 0, \\ (g_k)_t + \cos\theta_k (g_k)_x &= -g_k(\cos\theta_k)_x, \end{aligned} \qquad k = 1, \dots, N. \tag{65}$$

The equation for θ_k is known to develop shocks in finite time. The angle θ_k will be constant along characteristics, which are straight lines corresponding to rays. The shock develops where characteristics cross, i.e. where two wave fields meet. The equation for g_k is an ordinary transport equation with a source term involving the derivative of $\cos\theta_k$. Along characteristics, which are the same for both equations, the source term is zero, except at a shock where it becomes a delta function. The resulting solution for g_k is a delta function where the phase "should" have split into two new phases.

It is interesting again to compare the moment equations with the eikonal and transport equations, (3) and (4). The latter also form a weakly hyperbolic system with the same eigenvalue as (63). As was mentioned before, the viscosity solution picks out the phase corresponding to first arrival where the physically correct phase is multi-valued. When wave fields meet, there will therefore in general be a jump in ϕ. Because of the term $\nabla^2\phi$ in the source term of (4), the first amplitude coefficient w_0 has a measure at these points. Hence, the two different formulations are similar also in this respect.

For the two-dimensional case, another function, $\boldsymbol{g}$, is added to (63),

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x + \boldsymbol{g}(\boldsymbol{u})_y = 0, \qquad \boldsymbol{g}(\boldsymbol{u}) = \begin{pmatrix} \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_2^2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}. \tag{66}$$

Taking a linear combination of the Jacobians for $\boldsymbol{f}$ and $\boldsymbol{g}$ we get

$$\begin{aligned} J(\theta,\alpha_1,\alpha_2) &:= \alpha_1 \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} + \alpha_2 \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}} \\ &= G\left[\alpha_1 \begin{pmatrix} \cos\theta & -\sin\theta \\ 0 & \cos\theta \end{pmatrix} + \alpha_2 \begin{pmatrix} \sin\theta & \cos\theta \\ 0 & \sin\theta \end{pmatrix}\right] G^{-1}, \end{aligned} \tag{67}$$

with the same rotation G as in (64). Regardless of the choice of (α_1, α_2), we still only have one eigenvalue and an incomplete set of eigenvectors.

In the general case with N phases, homogeneous medium, the governing equations (17) reads

$$\boldsymbol{F}_0(\boldsymbol{u})_t + \boldsymbol{F}_1(\boldsymbol{u})_x + \boldsymbol{F}_2(\boldsymbol{u})_y = 0. \tag{68}$$

Denoting the Jacobians of $\boldsymbol{F}_k$ with J_k, the following relationship can be derived:

$$\alpha_1 J_1 + \alpha_2 J_2 = J_0 \cdot \operatorname{diag}(J(\theta_1,\alpha_1,\alpha_2),\ldots,J(\theta_N,\alpha_1,\alpha_2)). \tag{69}$$

This shows that the eigenvalues of the general system are simply the union of the eigenvalues of N systems of the type (66). It also shows that there will not be more than N eigenvectors, for the $2N \times 2N$ system. Hence, we have shown that the general system (17) is weakly hyperbolic.

2.5 The Riemann Problem

We study the one-dimensional Riemann problem for (63),

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \qquad \boldsymbol{u}(x,0) = \begin{cases} \boldsymbol{u}_l & x < 0, \\ \boldsymbol{u}_r & x > 0. \end{cases} \tag{70}$$

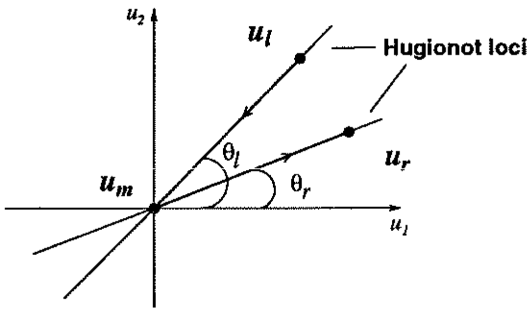


Figure 2: Hugionot loci for left and right state, and solution to the Riemann problem for the first type of discontinuity, plotted in phase space.

At a discontinuity the conservation form gives the Rankine-Hugionot jump condition,

$$\boldsymbol{f}(\boldsymbol{u}_l) - \boldsymbol{f}(\boldsymbol{u}_r) = s(\boldsymbol{u}_l - \boldsymbol{u}_r), \tag{71}$$

where s represents the propagation speed of the discontinuity. Since $\boldsymbol{f}(\boldsymbol{u}) = \cos\theta\boldsymbol{u}$, this simplifies to

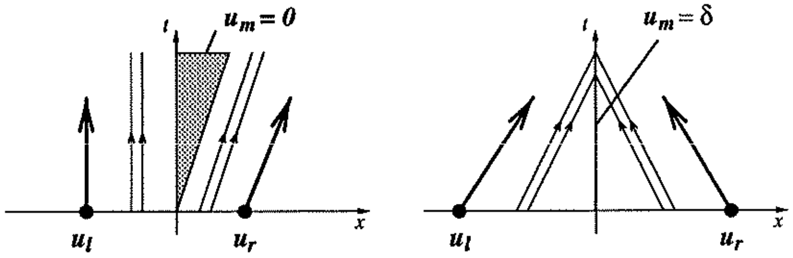
$$\cos\theta_l\boldsymbol{u}_l - \cos\theta_r\boldsymbol{u}_r = s(\boldsymbol{u}_l - \boldsymbol{u}_r). \tag{72}$$

For a given state, $\boldsymbol{u}_l$, we can construct the Hugionot locus consisting of all states to which we can connect with a jump. In our case the Hugionot locus of state $\boldsymbol{u}_l$ is simply $\alpha\boldsymbol{u}_l$, $\alpha \in \mathbb{R}$, hence, in phase space, a straight line shooting out from the origin at angle θ_l. The speed of propagation is $s = \cos\theta_l$. It follows that two non-zero states $\boldsymbol{u}_l$ and $\boldsymbol{u}_r$ can only be connected with a discontinuity, should they be parallel, $\boldsymbol{u}_l \parallel \boldsymbol{u}_r$. Otherwise there must be an intermediate state, $\boldsymbol{u}_m$, in between. What this state will be is dictated by the Lax entropy condition, saying that the left discontinuity must move slower than the right one. There will be two types of discontinuities. If $\cos\theta_l < \cos\theta_r$, we can use the origin of the phase space as the intermediate state, hence $\boldsymbol{u}_m = 0$. The Hugionot loci and the solution for this type of discontinuity is illustrated in Figure 2. If $\cos\theta_l > \cos\theta_r$, on the other hand, we do not have a solution in a classical sense. Formally, however, a weak solution to the conservation law with this initial data is given by setting $\boldsymbol{u}_m = t\tilde{\boldsymbol{u}}_m\delta(x - st)$. The conservation form gives a slightly modified jump condition for this case,

$$\cos\theta_l\boldsymbol{u}_l - \cos\theta_r\boldsymbol{u}_r = \cos\tilde{\theta}_m(\boldsymbol{u}_l - \boldsymbol{u}_r) + \tilde{\boldsymbol{u}}_m, \tag{73}$$

with the propagation speed $s = \cos\tilde{\theta}_m$. This situation corresponds to two meeting wave fields.

It is easily verified from (64). that $\boldsymbol{u}$ itself is an eigenvector of the Jacobian of $\boldsymbol{f}$. Therefore, the Hugionot locus will coincide with the integral curves of the system's characteristic fields. Since the (double) eigenvalue of (64), $\cos\theta$, is constant along the curves, the fields are linearly degenerate. From this we



(a) Contact discontinuity

(b) Overcompressive shock

Figure 3: The two different types of discontinuities.

conclude that the first type of discontinuity is a linear, contact discontinuity; characteristics run parallel to the discontinuity. The linear degeneracy also excludes the possibility of rarefaction wave solutions.

The second type of discontinuity will always have two characteristics incident to the discontinuity at each side, because of the double eigenvalue. These discontinuities are thus of overcompressive shock type. The two different discontinuities, plotted in (x, t)-space, are shown in Figure 3.

3 Numerical Approximations

This section includes some results on the numerical treatment of (17). As was discussed in the previous section, the system (17) is very sensitive, and this creates problems for the numerical methods. The sensitivity derives from the fact that the system is only weakly hyperbolic.

For the numerical methods we will use the following notation. Space and time is discretized uniformly with step sizes Δx, Δy and Δt. The grid function $\boldsymbol{U}_{ij}^n$ approximates the analytic solution,

$$\boldsymbol{U}_{ij}^n \approx \boldsymbol{u}(i\Delta x, j\Delta y, n\Delta t), \tag{74}$$

where $\boldsymbol{u}$ are the variables introduced in (16). Similarly, for the index of refraction,

$$\eta_{ij} = \eta(i\Delta x, j\Delta y), \qquad \partial_x \eta_{ij} = \eta_x(i\Delta x, j\Delta y), \qquad \partial_y \eta_{ij} = \eta_y(i\Delta x, j\Delta y). \tag{75}$$

3.1 Single Phase

The point of departure for our numerical approximations is the basic first order accurate *Lax-Friedrichs* finite difference method. For the one phase system in an inhomogeneous medium, it takes the form

$$\begin{aligned}\eta_{ij}^2 \boldsymbol{U}_{ij}^{n+1} &= \frac{\eta_{ij}^2}{4}\left(\boldsymbol{U}_{i-1,j}^n + \boldsymbol{U}_{i+1,j}^n + \boldsymbol{U}_{i,j-1}^n + \boldsymbol{U}_{i,j+1}^n\right) \\ &\quad -\frac{\eta_{ij}\Delta t}{2\Delta x}\left(\boldsymbol{F}_1(\boldsymbol{U}_{i+1,j}^n) - \boldsymbol{F}_1(\boldsymbol{U}_{i-1,j}^n)\right) \\ &\quad -\frac{\eta_{ij}\Delta t}{2\Delta y}\left(\boldsymbol{F}_2(\boldsymbol{U}_{i,j+1}^n) - \boldsymbol{F}_2(\boldsymbol{U}_{i,j-1}^n)\right) + \Delta t \boldsymbol{K}(\boldsymbol{U}_{ij}^n, \partial_x \eta_{ij}, \partial_y \eta_{ij}),\end{aligned} \tag{76}$$

with $\boldsymbol{F}_1$, $\boldsymbol{F}_2$ and $\boldsymbol{K}$ as defined in (28). Even if the Lax-Friedrichs method is only of first order, it works quite well and remains stable despite the sensitivity of the equations. Most of our results are produced using this method. The purpose of the numerical experiments is just to show the feasibility of the moment closure technique and for this purpose a first order method is sufficient. The reason for the Lax-Friedrich scheme's stability is that it introduces a substantial amount of viscosity, which implies that discontinuities in the solution are smeared out.

Less smearing of shocks is obtained with the *Godunov* method (see e.g. [7]), another first order method which adds a smaller measure of viscosity than Lax-Friedrichs. It is based on the exact solution of local Riemann problems. The two-dimensional Godunov method is constructed by applying an ordinary splitting approach.

Even though the Godunov method applied to the single phase system converges in L_1 (see Table 1) there are large L_∞ errors also for smooth problems (see Table 2 and Figure 5).

The reason for the method's behavior in Figure 5 can be found in the analysis of the Riemann problem in Section 2.5. Along the line $y = 1$ the Riemann problem in the y-direction corresponds to the situation in Figure 3a. Since there is no rarefaction wave solution, there will be no flux in the y direction, and the method reduces to the one-dimensional Godunov method in the x-direction along this line. Hence, along $y = 1$ there will be pure transport, and no damping. A similar distortion occurs when the source is not located exactly in the middle of a grid cell.

A second order accurate scheme introduces less artificial viscosity and can therefore be expected to be more sensitive. To avoid oscillations at discontinuities, so called *TVD* methods are desirable [7]. These nonlinear methods use limiters to ensure that new artificial extrema are not introduced in the solution. At an extrema, TVD methods are at most first order accurate.

We have implemented the *Nessyahu-Tadmor* method, [8]. It is a second order TVD method based on the Lax-Friedrichs structure. No Riemann problem needs to be solved, and the solution never has to be split up into characteristic fields, since the limiter is applied componentwise for systems. To compute the numerical derivatives in the scheme we use a variant of the *minmod* limiter,

$$u_j' = \mathrm{MM}\Big(\theta(u_j - u_{j-1}),\ \frac{1}{2}(u_{j+1} - u_{j-1}),\ \theta\Delta(u_{j+1} - u_j)\Big), \tag{77}$$

L_1	Lax-Friedrichs				Godunov		Nessyahu-Tadmor			
	unsplitted		splitted		splitted		unsplitted		splitted	
N	error	order	error	order	error	order	error	order	error	order
10	7.78e-3		3.60e-2		1.13e-2		7.98e-3		1.02e-2	
		0.85		0.56		0.80		1.47		1.23
20	4.33e-3		2.44e-2		6.50e-3		2.89e-3		4.35e-3	
		0.92		0.72		0.69		1.74		1.20
40	2.29e-3		1.48e-2		4.04e-3		8.66e-4		1.89e-3	
		0.96		0.82		0.78		1.88		1.03
80	1.18e-3		8.39e-3		2.35e-3		2.35e-4		9.24e-4	
		0.98		0.89		0.85		1.89		0.76
160	5.99e-4		4.53e-3		1.30e-3		6.32e-5		5.45e-4	

Table 1: L_1 norm of the errors at time $t = 0.85$ for test case A (see Section 4), using the single phase equations. Here $\Delta x = \Delta y = 1/N$ and CFL=0.65.

L_∞	Lax-Friedrichs				Godunov		Nessyahu-Tadmor			
	unsplitted		splitted		splitted		unsplitted		splitted	
N	error	order	error	order	error	order	error	order	error	order
10	9.49e-2		1.78e-1		3.04e-1		6.64e-2		8.99e-2	
		1.26		0.85		0.06		1.26		0.91
20	3.97e-2		9.87e-2		2.91e-1		2.76e-2		4.78e-2	
		1.21		0.55		0.02		1.71		1.07
40	1.71e-2		6.73e-2		2.87e-1		8.46e-3		2.28e-2	
		1.15		0.73		0.02		1.70		0.80
80	7.71e-3		4.06e-2		2.83e-1		2.61e-3		1.31e-2	
		1.09		0.85		0.01		1.57		0.91
160	3.63e-3		2.26e-2		2.82e-1		8.75e-4		6.98e-3	

Table 2: L_∞ norm of the errors at time $t = 0.85$ for test case A (see Section 4), using the single phase equations. Here $\Delta x = \Delta y = 1/N$ and CFL=0.65.

where

$$\mathrm{MM}(u_1, u_2, \ldots) = \begin{cases} \min_k u_k & \text{if } u_k > 0, \forall k, \\ \max_k u_k & \text{if } u_k < 0, \forall k, \\ 0 & \text{otherwise.} \end{cases} \tag{78}$$

We let $\theta = 2$ in our test cases. This gives better results than the classical choice $\theta = 1$. Derivatives are computed componentwise, without using the exact Jacobians.

A two-dimensional Nessyahu-Tadmor scheme was obtained using the one-dimensional method together with dimensional splitting. *Strang splitting*, [10], was used to preserve second order accuracy. The result is however not perfectly satisfactory. The convergence rate for the splitted Nessyahu-Tadmor scheme turns out to be somewhat slower than Lax-Friedrichs in L_∞ and only marginally higher in L_1 (see Table 1 and Table 2). A likely cause of this is the dimensional splitting. As a comparison, we also include results for a splitted version of the Lax-Friedrichs scheme, which are clearly inferior to the results for the unsplitted scheme. The failure of the Godunov scheme could also be attributed to the dimensional splitting. See [6] for a further discussion.

In contrast, when we use a genuinely two-dimensional version of the Nessyahu-Tadmor scheme, [6], we get a considerable improvement of the results. Almost full second order accuracy is achieved (again, see Table 1 and Table 2).

3.2 Multiple Phases

In the multiple phase case, we have mainly considered the case of two phases, $N = 2$. It is more difficult to get reliable calculations when solving (17) with multiple phases, than in the case of a single phase. A few new problems are added to the numerical methods. In each time step a nonlinear system of equations must be solved. The Jacobian of this system can be singular. It may even happen that it does not have a solution. Being careful when solving the system, however, it seems possible to compute solutions for most configurations. Where not otherwise stated, we assume below that the choice of moments are those in (18).

We have only used the Lax-Friedrichs method to solve the the multiple phase systems. With an inhomogeneous media, it can be written as

$$\begin{aligned} \eta_{ij}^2 \boldsymbol{F}_0(\boldsymbol{U}_{ij}^{n+1}) &= \frac{\eta_{ij}^2}{4}\left(\boldsymbol{F}_0(\boldsymbol{U}_{i-1,j}^n) + \boldsymbol{F}_0(\boldsymbol{U}_{i+1,j}^n) + \boldsymbol{F}_0(\boldsymbol{U}_{i,j-1}^n) + \boldsymbol{F}_0(\boldsymbol{U}_{i,j+1}^n)\right) \\ &\quad -\frac{\eta_{ij}\Delta t}{2\Delta x}\left(\boldsymbol{F}_1(\boldsymbol{U}_{i+1,j}^n) - \boldsymbol{F}_1(\boldsymbol{U}_{i-1,j}^n)\right) \\ &\quad -\frac{\eta_{ij}\Delta t}{2\Delta y}\left(\boldsymbol{F}_2(\boldsymbol{U}_{i,j+1}^n) - \boldsymbol{F}_2(\boldsymbol{U}_{i,j-1}^n)\right) \\ &\quad +\Delta t \boldsymbol{K}(\boldsymbol{U}_{ij}^n, \partial_x \eta_{ij}, \partial_y \eta_{ij}). \end{aligned} \tag{79}$$

For $N = 2$ the functions are

$$\boldsymbol{F}_0 = \begin{pmatrix} u_1 + u_3 \\ u_2 + u_4 \\ \frac{u_1^3}{u_1^2+u_2^2} + \frac{u_3^3}{u_3^2+u_4^2} \\ \frac{u_2^3}{u_1^2+u_2^2} + \frac{u_4^3}{u_3^2+u_4^2} \end{pmatrix}, \qquad \boldsymbol{F}_1 = \begin{pmatrix} \frac{u_1^2}{\sqrt{u_1^2+u_2^2}} + \frac{u_3^2}{\sqrt{u_3^2+u_4^2}} \\ \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} + \frac{u_3 u_4}{\sqrt{u_3^2+u_4^2}} \\ \frac{u_1^4}{(u_1^2+u_2^2)^{3/2}} + \frac{u_3^4}{(u_3^2+u_4^2)^{3/2}} \\ \frac{u_1 u_2^3}{(u_1^2+u_2^2)^{3/2}} + \frac{u_3 u_4^3}{(u_3^2+u_4^2)^{3/2}} \end{pmatrix},$$

$$\boldsymbol{F}_2 = \begin{pmatrix} \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} + \frac{u_3 u_4}{\sqrt{u_3^2+u_4^2}} \\ \frac{u_2^2}{\sqrt{u_1^2+u_2^2}} + \frac{u_4^2}{\sqrt{u_3^2+u_4^2}} \\ \frac{u_1^2 u_2}{(u_1^2+u_2^2)^{3/2}} + \frac{u_3^2 u_4}{(u_3^2+u_4^2)^{3/2}} \\ \frac{u_2^3}{(u_1^2+u_2^2)^{3/2}} + \frac{u_4^3}{(u_3^2+u_4^2)^{3/2}} \end{pmatrix}, \quad \boldsymbol{K} = \begin{pmatrix} \frac{\eta_x u_2^2 - \eta_y u_1 u_2}{\sqrt{u_1^2+u_2^2}} + \frac{\eta_x u_4^2 - \eta_y u_3 u_4}{\sqrt{u_3^2+u_4^2}} \\ \frac{\eta_y u_1^2 - \eta_x u_1 u_2}{\sqrt{u_1^2+u_2^2}} + \frac{\eta_y u_3^2 - \eta_x u_3 u_4}{\sqrt{u_3^2+u_4^2}} \\ \frac{\eta_x u_1^2 u_2^2 - \eta_y u_1^3 u_2}{(u_1^2+u_2^2)^{3/2}} + \frac{\eta_x u_3^2 u_4^2 - \eta_y u_3^3 u_4}{(u_3^2+u_4^2)^{3/2}} \\ \frac{\eta_y u_1^2 u_2^2 - \eta_x u_1 u_2^3}{(u_1^2+u_2^2)^{3/2}} + \frac{\eta_y u_3^2 u_4^2 - \eta_x u_3 u_4^3}{(u_3^2+u_4^2)^{3/2}} \end{pmatrix}. \tag{80}$$

We see from (79) that for each iteration, at each point, it is necessary to solve a nonlinear system of equations of the type

$$\boldsymbol{F}_0(\boldsymbol{U}_{ij}^{n+1}) = \hat{\boldsymbol{m}}_{ij}^n. \tag{81}$$

Actually, solving (81) is the first part of the evaluation of the flux functions $\boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}$ and $\boldsymbol{F}_2 \circ \boldsymbol{F}_0^{-1}$. The second part of the composition is computed in the subsequent time step.

There are a number of problems associated with this evaluation. First, solving (81) can be difficult. In the general case an iterative solver must be used, which is expensive and requires good initial values. For the $N = 2$ case it can, however, be done analytically, see Appendix A.1.

Second, (81) may not have a solution. Although, for the exact solution of the PDE (81) should always be satisfied, truncation errors in the numerical scheme may have perturbed the solution so that $\hat{\boldsymbol{m}}_{ij}^n \notin \boldsymbol{F}_0(D_f)$. Thus, the conditions in Theorem 1 are not satisfied. In this case we use the least squares solution of (81).

Third, in general, the Jacobian of $\boldsymbol{F}_0$ is singular at some points in the computational domain. For $N = 2$, the Jacobian

$$J_0 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 3\cos^2\theta_1 - 2\cos^4\theta_1 & -2\sin\theta_1\cos^3\theta_1 & 3\cos^2\theta_2 - 2\cos^4\theta_2 & -2\sin\theta_2\cos^3\theta_2 \\ -2\cos\theta_1\sin^3\theta_1 & 3\sin^2\theta_1 - 2\sin^4\theta_1 & -2\cos\theta_2\sin^3\theta_2 & 3\sin^2\theta_2 - 2\sin^4\theta_2 \end{pmatrix}, \tag{82}$$

is singular when

$$\cos\theta_1 = \pm\cos\theta_2 \qquad \text{and} \qquad \sin\theta_1 = \pm\sin\theta_2, \tag{83}$$

or, equivalently, $z_1^2 = z_2^2$, using the variables defined in (46). (Note that J_0 only depends on the angles θ_k, not on g_k.) For iterative methods that use the Jacobian, this is a problem. Furthermore, to ensure the Lipschitz continuity of the numerical flux function $\boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}(\hat{\boldsymbol{m}}_{ij}^n)$, there should be a constant M such that

$$||J_1 J_0^{-1}|| = ||J_0 \Lambda_1 J_0^{-1}|| \leq M \text{ a.e.}, \qquad \Lambda_1 = \text{diag}(J(\theta_1, 1, 0), \ldots, J(\theta_N, 1, 0)), \tag{84}$$

where we also used (69). A similar inequality should hold for $\boldsymbol{F}_2 \circ \boldsymbol{F}_0^{-1}$. Since J_0 can be singular it is not clear that the condition in (84) holds. For $N = 2$ (84) holds in any closed subset of D_f, that is essentially whenever $z_1 + z_2 \neq 0$, see Theorem 2 in Appendix B.

For the $N = 2$ case we have also used an alternative choice of moments, namely those in (60). As was seen in the remark after Theorem 1, the flux functions are in general two-valued for this choice. For simple problems this ambiguity can be resolved by for instance choosing the value that maximizes the smoothness of the solution in some sense. Also for this choice (81) can be solved analytically, see Appendix A.2. The expression in (84) is unbounded when either

$$\cos\theta_1 = \cos\theta_2 \qquad \text{or} \qquad \sin\theta_1 = \sin\theta_2, \tag{85}$$

but not both. This permits us to solve problems where $z_1^2 = z_2^2$, including the case $z_1 + z_2 = 0$ (which cannot be solved with the choice used above). On the other hand, two new singular branches are introduced, which depend on the choice of coordinate system. Solutions cannot be computed on these branches.

Another feature of the system (81) should also be noted. It has always at least $N!$ solutions, since $\boldsymbol{F}_k$ is invariant under the action of S_N, the group of permutations of N elements, on the vectors $(u_{2k-1}, u_{2k})^T$. For instance, when $N = 2$,

$$\boldsymbol{F}_k(u_1, u_2, u_3, u_4) = \boldsymbol{F}_k(u_3, u_4, u_1, u_2), \qquad k = 0, 1, 2. \tag{86}$$

The phases are thus interchangeable, which from a physical standpoint is quite natural. Numerically, it has the effect that we cannot be certain which of the $N!$ roots our method finds. Therefore, the numerically calculated variables $\boldsymbol{u}$ can be very discontinuous over the domain, even though the moments, which we get by applying the $\boldsymbol{F}_k$ to the variables, are smooth.

4 Numerical Results

In this section we show results from eight different test cases. We have considered both homogeneous and inhomogeneous media. Sources are located outside the computational domain. The test cases for homogeneous problems are

A) the rectangle $0 \leq x \leq 1$ and $0 \leq y \leq 2$; one source located at coordinates (−0.2, 1); smooth point source with exact solution $g = \max(0, t - r)^3/r$,

B) the rectangle $0 \leq x \leq 1$ and $0 \leq y \leq 1$; two sources located at coordinates (−0.2, −0.2) and (1.2, 1.2); smooth point sources with exact solution $g_k = \max(0, t - r_k)^2/r_k$, $k = 1, 2$,

C) the same rectangle as in B but with two sources located at coordinates $(-0.3,\ 0.65)$ and $(1.3,\ 0.35)$; discontinuous point sources with exact solution $g_k = H(t - r_k)/r_k$, $k = 1, 2$.

D) the same rectangle as in B but with three sources located at coordinates $(-0.4,\ 0.8)$, $(0.2,\ 1.4)$ and $(-0.15,\ 1.15)$; discontinuous point sources with exact solution $g_k = H(t - r_k)/r_k$, $k = 1, 2, 3$.

The variable $r_k = r_k(x, y)$ is the distance to source k. In the homogeneous case we use the value of the exact solution as a Dirichlet boundary condition on all boundaries. A CFL number of 0.65 was used for all computations.

General results for test case A is shown in Figure 4, where the Lax-Friedrich method was used to solve the $N = 1$ system (76). The difficulties with using the Godunov and the splitted Nessyahu-Tadmor methods for the same problem are highlighted in Figure 5 and Figure 6 respectively. Convergence for the different methods are summarized in Table 1 and Table 2.

For test case B we only used the Lax-Friedrich method. In Figure 7 the single phase system was solved, even though the physically correct solution contains two phases. A measure solution is suggested. In Figure 8 we used the $N = 2$ system (79, 80) for the same problem and it captures both phases.

Also for test case C and D all solutions were computed using the Lax-Friedrich method. We present the results for the $N = 2$ system on case C in Figure 9 and the results for the $N = 3$ system on case D in Figure 10.

Test cases B and C both contain waves meeting head-on. For the $N = 2$ system the choice of moments in (60) was therefore used. In all other cases, including those below, the moments in (18) were used.

The inhomogeneous test cases are

E) the same rectangle as in B; a plane wave entering obliquely from the left, at an angle of 45 degrees to the line $x = 0$, index of refraction modeling a smooth interface,

$$\eta(x, y) = 1.5 + \frac{1}{\pi}\arctan(40(x - \frac{1}{2})). \tag{87}$$

F) the rectangle $0 \le x \le 2$ and $0 \le y \le 2$; a plane wave entering from the left, orthogonal to the line $x = 0$, index of refraction modeling a smooth concave lens,

$$\eta(x, y) = \begin{cases} 1 & D > 1, \\ \frac{4}{3 - \cos\frac{\pi y}{s}} & D \le 1, \end{cases} \tag{88}$$

with

$$D = -\left(\frac{x-1}{0.8}\right)^2 + \left(\frac{y}{0.3}\right)^2, \qquad s = 0.3\sqrt{1 + \left(\frac{x-1}{0.8}\right)^2}. \tag{89}$$

G) the rectangle $0 \leq x \leq 1.5$ and $0 \leq y \leq 2$; a plane wave entering from the left, orthogonal to the line $x = 0$, index of refraction modeling a smooth wedge,

$$\eta(x,y) = 1.5 + \frac{1}{\pi}\arctan(\frac{20}{\sqrt{2}}(x - 0.2 - |y - 1|)). \tag{90}$$

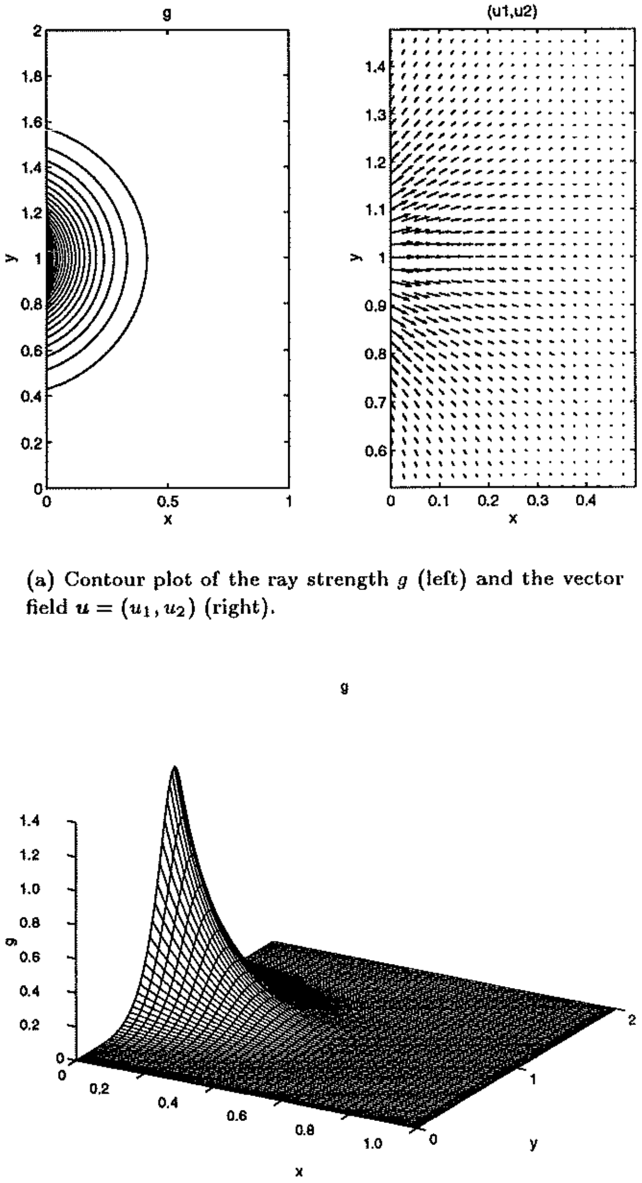
H) the same rectangle as in F; a plane wave entering from the left, orthogonal to the line $x = 0$, index of refraction modeling a smooth convex lens,

$$\eta(x,y) = \begin{cases} 1 & d > 1, \\ \frac{4}{3-\cos(\pi d)} & d \leq 1, \end{cases} \qquad d = \left(\frac{x - 0.3}{0.8}\right)^2 + \left(\frac{y}{0.3}\right)^2. \tag{91}$$

Test cases F and H were taken from [5]. In the inhomogeneous case we use a Dirichlet boundary condition on the left boundary, $x = 0$. On the remaining boundaries we use simple extrapolation, $\boldsymbol{U}_{n+1,j} = \boldsymbol{U}_{nj}$, etc. All computations use the Lax-Friedrich method with CFL=0.65.

For test cases E and F we have only solved the $N = 1$ system, which gives the correct physical solution. Results are in Figure 11 and Figure 12.

Test cases G and H were solved with both the one and the two phase system. Figure 13 and Figure 14 show the results for these cases. One phase is not sufficient to describe the physically correct solution, so a measure solution is indicated in Figure 13a and Figure 14a.



(a) Contour plot of the ray strength g (left) and the vector field $\boldsymbol{u} = (u_1, u_2)$ (right).

(b) Ray strength g.

Figure 4: Solution at time $t = 0.85$ of the single phase system for test case A, using Lax-Friedrich with 40 x 80 points.

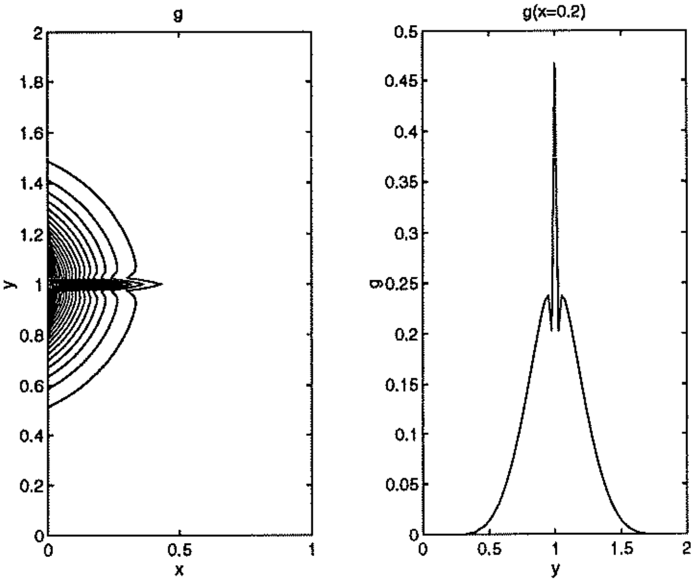


Figure 5: Solution at time $t = 0.85$ of the single phase system for test case A, using the Godunov method with 40×80 points. Left figure is a contour plot of the ray strength g. Right figure shows g in a vertical cut at $x = 0.2$.

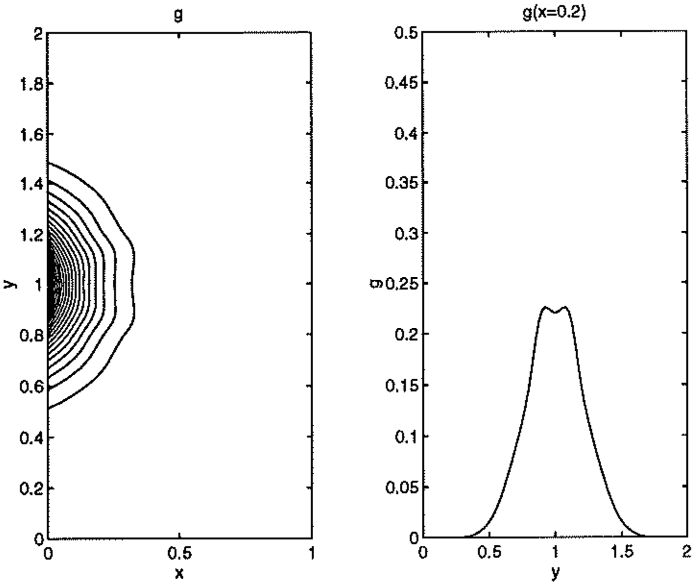


Figure 6: Solution at time $t = 0.85$ of the single phase system for test case A, using the splitted Nessyahu-Tadmor method with 40×80 points. Left figure is a contour plot of the ray strength g. Right figure shows g in a vertical cut at $x = 0.2$.

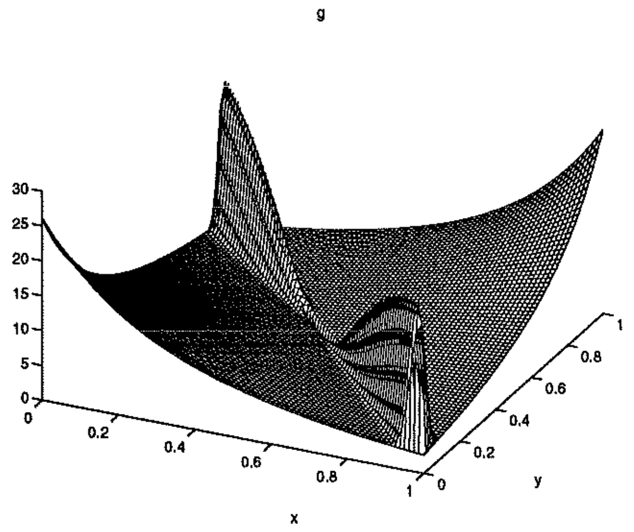


Figure 7: Solution at time $t = 3.0$ of the single phase system for test case B, using the Lax-Friedrich method with 80×80 points. Figure shows ray strength g.

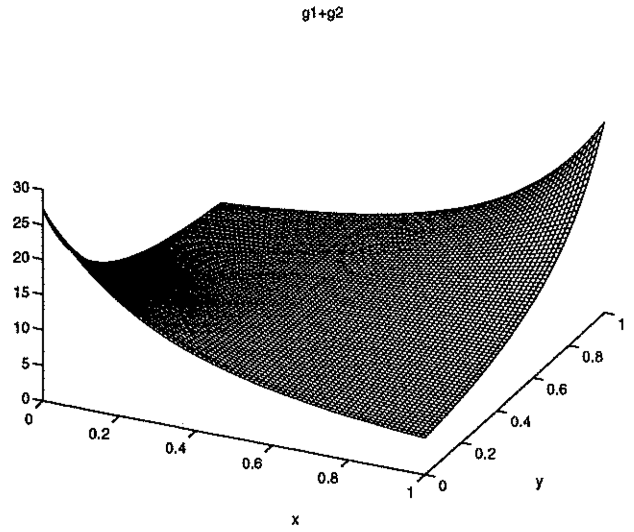
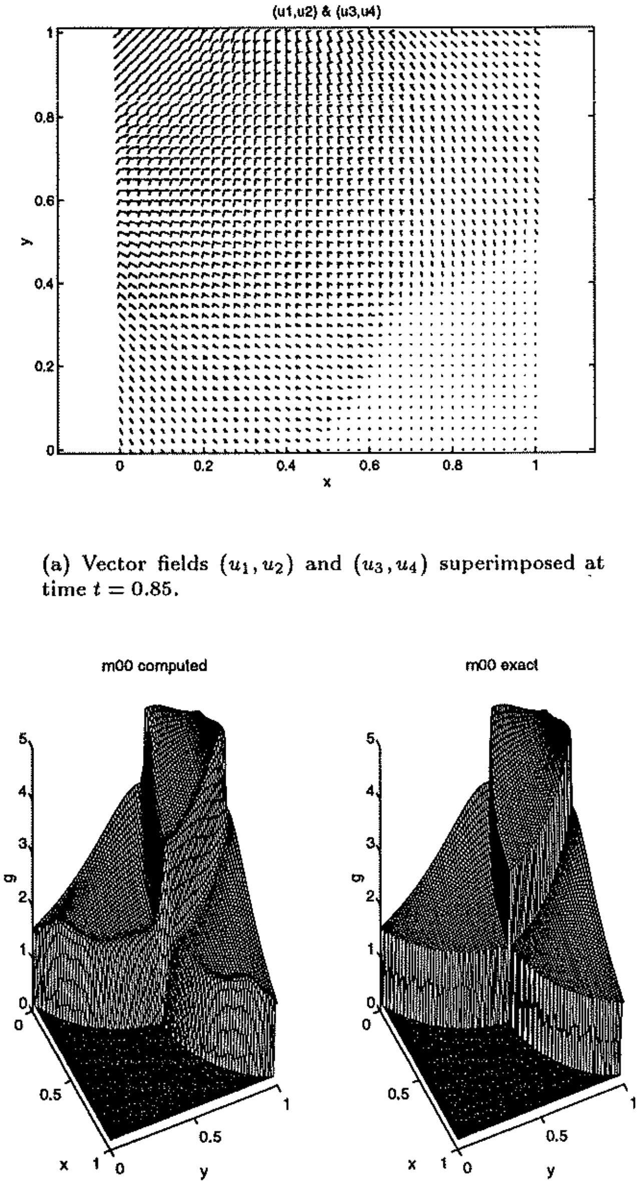


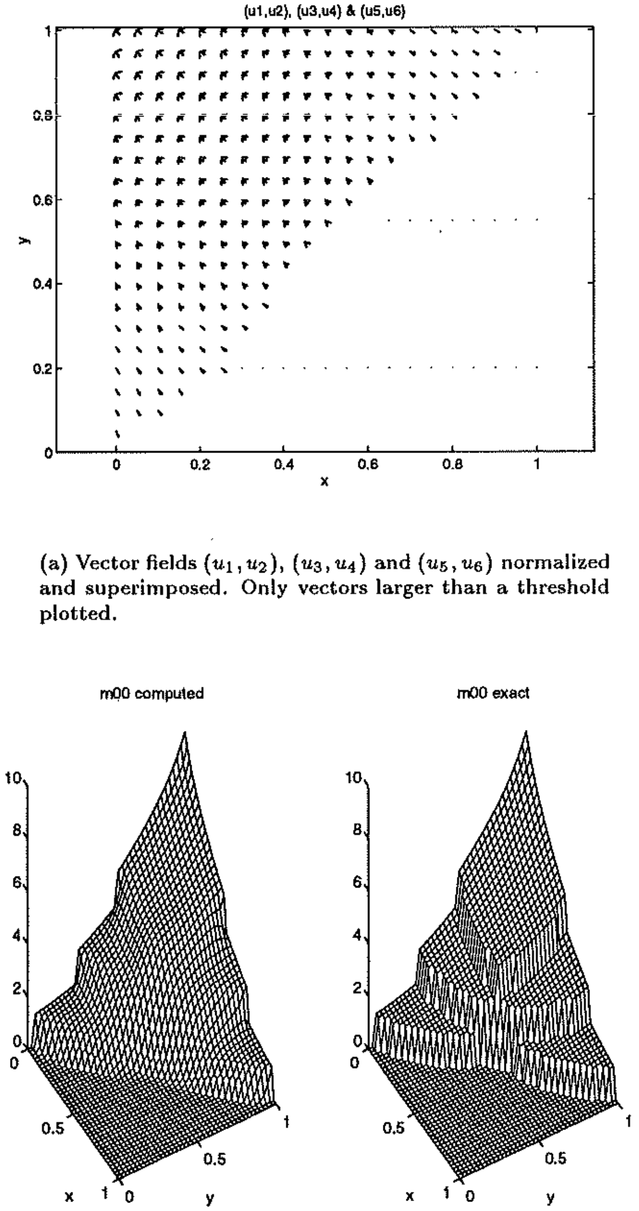
Figure 8: Solution at time $t = 3.0$ of the two phase system for test case B, using the Lax-Friedrich method with 80×80 points. Figure shows the combined ray strengths $g_1 + g_2 = m_{00}$.



(a) Vector fields (u_1, u_2) and (u_3, u_4) superimposed at time $t = 0.85$.

(b) Combined strengths of the two waves, $g_1 + g_2 = m_{00}$, at time $t = 0.7$, computed (left) and exact (right).

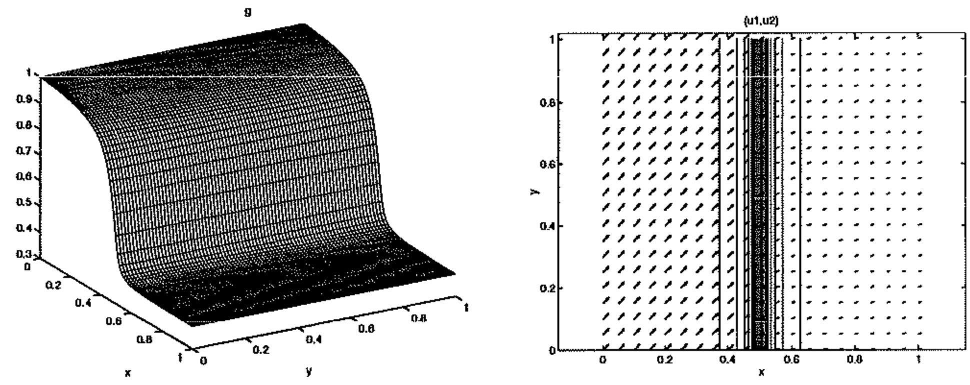
Figure 9: Solution of the two phase system for test case C, using the Lax-Friedrich method with 80 × 80 points.



(a) Vector fields (u_1, u_2), (u_3, u_4) and (u_5, u_6) normalized and superimposed. Only vectors larger than a threshold plotted.

(b) Combined strengths of the three waves, $g_1 + g_2 + g_3 = m_{00}$, computed (left) and exact (right).

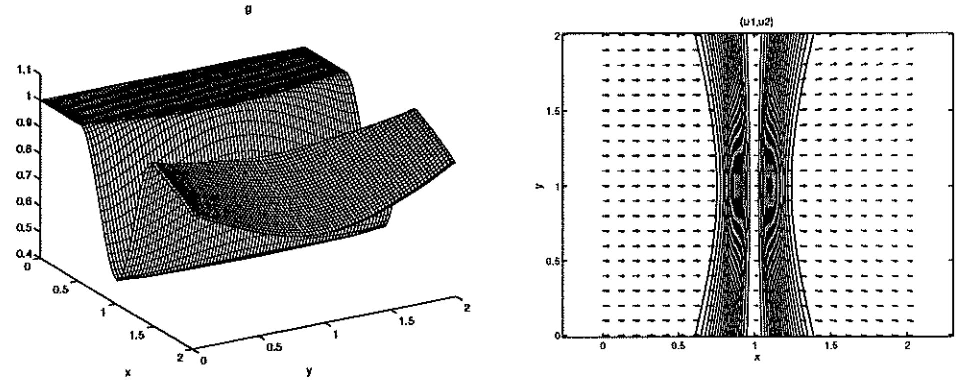
Figure 10: Solution of the three phase system for test case D, using the Lax-Friedrich method with 40 × 40 points.



(a) Ray strength g.

(b) Vector field $\boldsymbol{u} = (u_1, u_2)$ with contour lines of the index of refraction η overlayed.

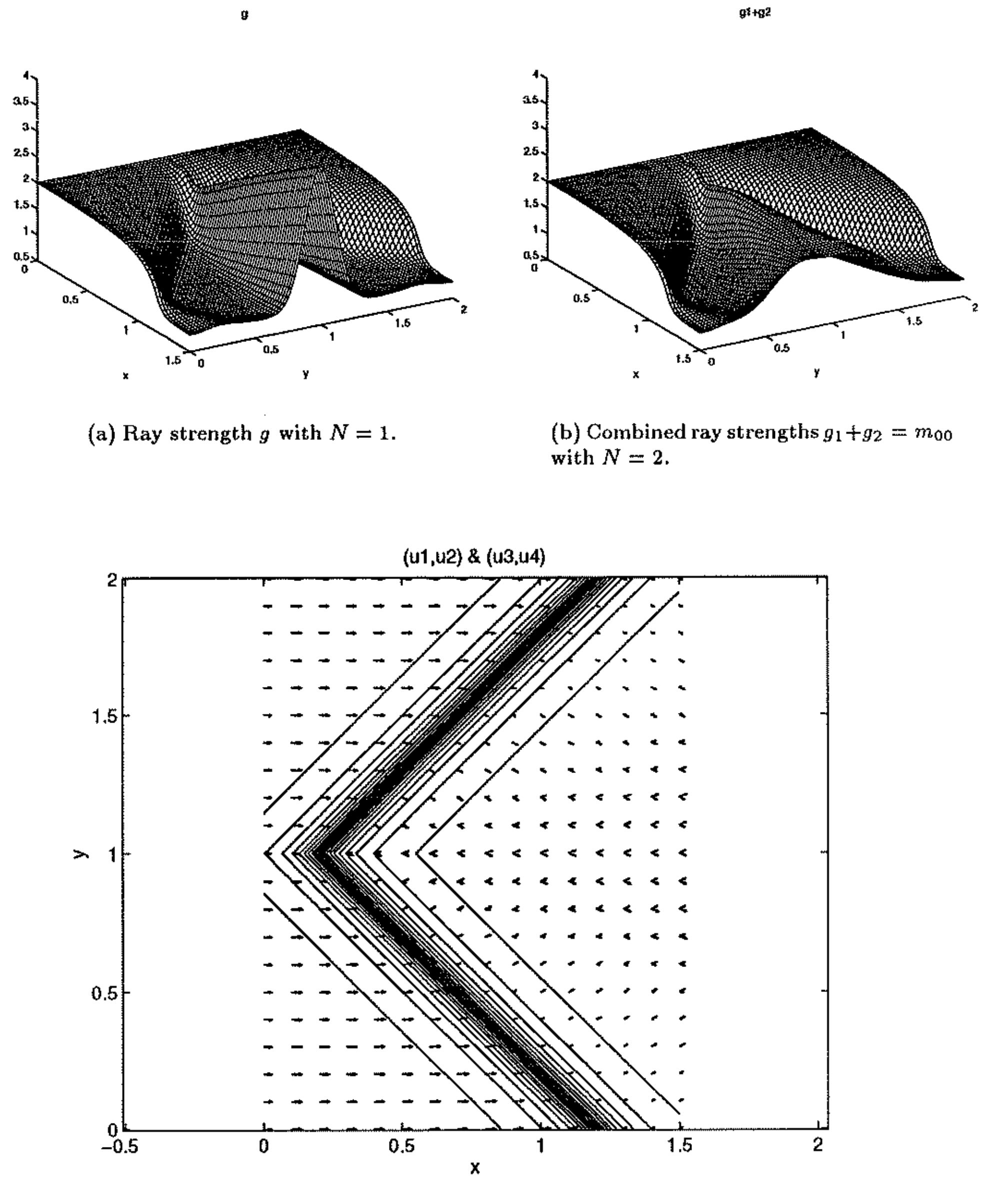
Figure 11: Solution at time $t = 3.0$ of the single phase system for test case E, using Lax-Friedrich with 80 × 80 points.



(a) Ray strength g.

(b) Vector field $\boldsymbol{u} = (u_1, u_2)$ with contour lines of the index of refraction η overlayed.

Figure 12: Solution at time $t = 3.0$ of the single phase system for test case F, using Lax-Friedrich with 80 × 80 points.

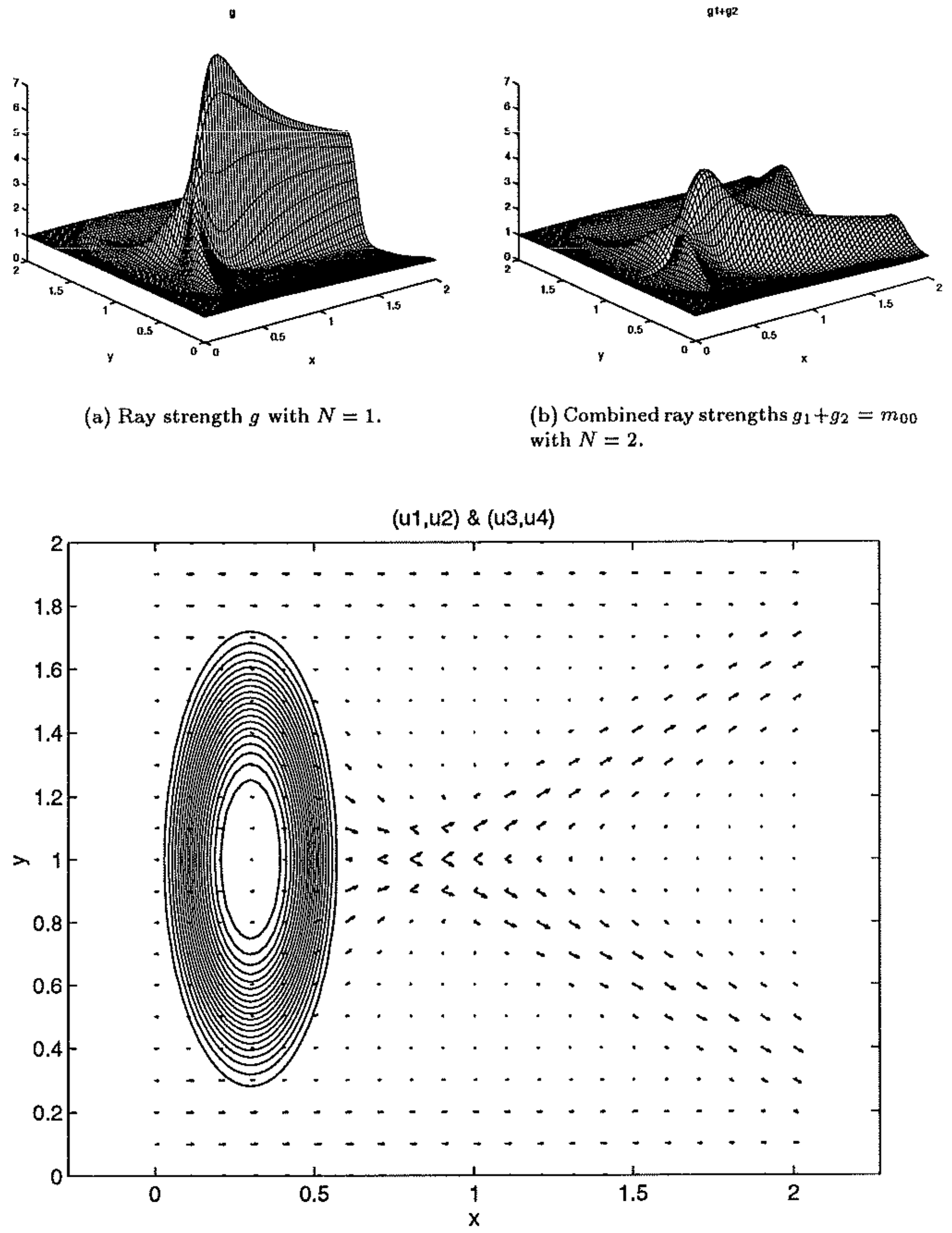


(a) Ray strength g with $N = 1$.

(b) Combined ray strengths $g_1 + g_2 = m_{00}$ with $N = 2$.

(c) Vector fields (u_1, u_2), (u_3, u_4) and contour lines of the index of refraction η superimposed, with $N = 2$.

Figure 13: Solution at time $t = 5.0$ of the one and two phase systems for test case G, using Lax-Friedrich with 60×80 points.



(a) Ray strength g with $N = 1$.

(b) Combined ray strengths $g_1+g_2 = m_{00}$ with $N = 2$.

(c) Vector fields (u_1, u_2), (u_3, u_4) and contour lines of the index of refraction η superimposed, with $N = 2$.

Figure 14: Solution at time $t = 3.5$ of the one and two phase systems for test case H, using Lax-Friedrich with 80 × 80 points.

References

[1] Y. Brenier and L. Corrias. Capturing multivalued solutions. CAM Report 94-46, Department of Mathematics, UCLA, 1994.

[2] Y. Brenier and E. Grenier. On the model of pressureless gases with sticky particles. CAM Report 94-45, Department of Mathematics, UCLA, 1994.

[3] M. Crandall and P. Lions. Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 277:1–42, 1983.

[4] W. E, Yu. G. Rykov, and Ya. G. Sinai. Generalized variational principles, global weak solutions and behavior with random initial data for systems of conservation laws arising in adhesion particle dynamics, to appear.

[5] B. Engquist, E. Fatemi, and S. Osher. Numerical solution of the high frequency asymptotic expansion for the scalar wave equation. CAM Report 93-05, Department of Mathematics, UCLA, 1993.

[6] G.-S. Jiang and E. Tadmor. Non-oscillatory central schemes for multidimensional hyperbolic conservation laws. CAM Report 96-36, Department of Mathematics, UCLA, 1996.

[7] R. J. LeVeque. *Numerical Methods for Conservation Laws.* Birkhäuser, 1992.

[8] H. Nessyahu and E. Tadmor. Non-oscillatory central differencing for hyperbolic conservation laws. *Journal of Computational Physics*, 87(2):408–463, 1990.

[9] F. Qin et al. Finite-difference solution of the eikonal equation along expanding wavefronts. *Geophysics*, 57(3):478–487, March 1992.

[10] G. Strang. On the construction and comparison of difference schemes. *SIAM Journal of Numerical Analysis*, 5:506–517, 1968.

[11] J. van Trier and W. W. Symes. Upwind finite-difference calculation of traveltimes. *Geophysics*, 56(6):812–821, June 1991.

[12] J. Vidale. Finite-difference calculation of traveltimes. *Bulletin of the Seismological Society of America*, 78(6):2062–2076, December 1988.

A Inverting F_0

We will here show how to solve the system of equations (81) analytically, when $N = 2$, for two different choices of moments. We use the ansatz

$$u_1 = \frac{1}{2}m_{10} + \frac{1}{2b}(am_{10}\cos\phi - cm_{01}\sin\phi), \qquad u_3 = m_{10} - u_1, \tag{92}$$
$$u_2 = \frac{1}{2}m_{01} + \frac{1}{2b}(am_{01}\cos\phi + cm_{10}\sin\phi), \qquad u_4 = m_{01} - u_2,$$

with the variables a, b and c defined as

$$a = g_1 + g_2, \qquad b = \sqrt{m_{10}^2 + m_{01}^2}, \qquad a^2 = b^2 + c^2, \tag{93}$$

so that

$$g_1 = \frac{1}{2}(a + b\cos\phi), \qquad g_2 = \frac{1}{2}(a - b\cos\phi). \tag{94}$$

A.1 The m_{30} case

For the choice of moments in (18), the nonlinear system of equations (81) reads

$$\begin{array}{rcl} u_1 + u_3 &=& m_{10}, \\ u_2 + u_4 &=& m_{01}, \\ \frac{u_1^3}{u_1^2+u_2^2} + \frac{u_3^3}{u_3^2+u_4^2} &=& m_{30}, \\ \frac{u_2^3}{u_1^2+u_2^2} + \frac{u_4^3}{u_3^2+u_4^2} &=& m_{03}. \end{array} \tag{95}$$

We observe that

$$\begin{array}{rcl} m_{30} &=& m_{10}\cos^2\alpha - R\cos(\beta + 3\alpha)(1 + \cos\beta), \\ m_{03} &=& m_{01}\sin^2\alpha + R\sin(\beta + 3\alpha)(1 + \cos\beta). \end{array} \tag{96}$$

Here,

$$\tan\alpha = \frac{m_{01}}{m_{10}}, \qquad \beta = \theta_1 + \theta_2 - 2\alpha, \qquad R = \frac{b}{2}(1 - \frac{b^2}{a^2}), \tag{97}$$

with a to be determined. The values of β and R follows from (96),

$$\tan(\beta + 3\alpha) = -\frac{m_{03} - m_{01}\sin^2\alpha}{m_{30} - m_{10}\cos^2\alpha}, \tag{98}$$
$$R^2(1 + \cos\beta)^2 = (m_{30} - m_{10}\cos^2\alpha)^2 + (m_{03} - m_{01}\sin^2\alpha)^2. \tag{99}$$

From R we can compute a. With β known we now use the relationship

$$\cos(\beta + 2\alpha) = \cos(\theta_1 + \theta_2) = \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 = \frac{u_1u_3}{g_1g_2} - \frac{u_2u_4}{g_1g_2}. \tag{100}$$

Multiplying (100) by g_1g_2 and inserting (92) and (94) yields an equation of the form

$$q_1\cos(2\phi) + q_2\sin(2\phi) = q_3, \tag{101}$$

which can be solved exactly. The coefficients are

$$\begin{aligned} q_1 &= \tfrac{1}{b^2}(a^2+c^2)(m_{10}^2-m_{01}^2)-b^2\cos(\beta+2\alpha), \\ q_2 &= -\tfrac{4ac}{b^2}m_{10}m_{01}, \\ q_3 &= (b^2-2a^2)\cos(\beta+2\alpha)+m_{10}^2-m_{01}^2. \end{aligned} \tag{102}$$

A.2 The m_{20} case

If we choose the moments in (60) the equations (81) will be

$$\begin{aligned} u_1+u_3 &= m_{10}, \\ u_2+u_4 &= m_{01}, \\ \tfrac{u_1^2}{\sqrt{u_1^2+u_2^2}}+\tfrac{u_3^2}{\sqrt{u_3^2+u_4^2}} &= m_{20}, \\ \tfrac{u_2^2}{\sqrt{u_1^2+u_2^2}}+\tfrac{u_4^2}{\sqrt{u_3^2+u_4^2}} &= m_{02}. \end{aligned} \tag{103}$$

We get immediately that

$$a = m_{20}+m_{02}. \tag{104}$$

We have left to solve the angle ϕ. To do this, we use equation three of (103),

$$\frac{u_1^2}{g_1}+\frac{u_3^2}{g_2} = m_{20}. \tag{105}$$

Like in Appendix A.1 we multiply (105) by g_1g_2 and insert (92) and (94) to get an equation of the form

$$r_1\cos(2\phi)+r_2\sin(2\phi) = r_3. \tag{106}$$

This equations can be solved exactly. Here, the coefficients are

$$\begin{aligned} r_1 &= \tfrac{a}{2b^2}\left((a^2-2b^2)m_{10}^2-c^2m_{01}^2\right)+\tfrac{b^2}{2}m_{20}, \\ r_2 &= -\tfrac{c^3}{b^2}m_{10}m_{01}, \\ r_3 &= \tfrac{1}{2}(2a^2-b^2)m_{20}-\tfrac{a}{2b^2}(a^2m_{10}^2+c^2m_{01}^2). \end{aligned} \tag{107}$$

B Stability of $\boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}$

We here prove a theorem showing the stability of the numerical flux function $\boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}$, in the sense that its Jacobian is bounded.

Theorem 2. *Let K be any closed subset of D_f, defined in (51). For $N=2$ there exists a constant M, only depending on K, such that the matrix functions $J_0(\boldsymbol{u})$ and $\Lambda_1(\boldsymbol{u})$, from (82) and (84) respectively, satisfy*

$$||J_0\Lambda_1J_0^{-1}|| \leq M, \qquad \forall \boldsymbol{u}\in K. \tag{108}$$

Proof. We first recall that both J_0 and Λ are actually functions of only θ_1 and θ_2. Interchangeably with θ_k we will use the complex z_k. The relation between these variables and $\boldsymbol{u}$ are defined in (46) and (47).

We now start by showing that J_0 is singular if and only if condition (83) holds. That it is sufficient condition is obvious, since it makes column one (two) and three (four) of J_0 equal. To see that it is also necessary, suppose there are z_1 and z_2 such that (83) does not hold, that is $z_1^2 \neq z_2^2$, and that $J_0(z_1, z_2)$ is singular. Then there is a vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_4)^T \neq 0$ satisfying $J_0^T \boldsymbol{\beta} = 0$, or equivalently

$$\begin{aligned} 4\beta_1 + \beta_3(3 + 2\cos 2\theta_k - \cos 4\theta_k) + \beta_4(-2\sin 2\theta_k + \sin 4\theta_k) &= 0, \\ 4\beta_2 + \beta_3(-2\sin 2\theta_k - \sin 4\theta_k) + \beta_4(3 - 2\cos 2\theta_k - \cos 4\theta_k) &= 0, \end{aligned} \qquad (109)$$

for $k = 1, 2$. We multiply the second of these equations by the imaginary unit and add them to get

$$4\alpha_1 + 3\alpha_2 + 2\bar{\alpha}_2 \frac{1}{z_k^2} - \alpha_2 z_k^4 = 0, \qquad k = 1, 2, \qquad (110)$$

where we also introduced $\alpha_1 = \beta_1 + i\beta_2$ and $\alpha_2 = \beta_3 + i\beta_4$. Since $|z_k| = 1$ we can multiply (110) by z_k^2. Thus, we have that z_1^2 and z_2^2 are two roots of the polynomial

$$P_\alpha(z) = 2\bar{\alpha}_2 + (4\alpha_1 + 3\alpha_2)z - \alpha_2 z^3. \qquad (111)$$

We note that this implies that $\alpha_2 \neq 0$. Denoting the third root of P_α by γ we have the identity

$$2\bar{\alpha}_2 + (4\alpha_1 + 3\alpha_2)z - \alpha_2 z^3 = -\alpha_2(z - z_1^2)(z - z_2^2)(z - \gamma). \qquad (112)$$

Identification gives for the constant term

$$2\bar{\alpha}_2 = \alpha_2 z_1^2 z_2^2 \gamma \qquad \Rightarrow \qquad 2|\bar{\alpha}_2| = |\alpha_2||\gamma| \qquad \Rightarrow \qquad |\gamma| = 2, \qquad (113)$$

and the coefficient in front of the z^2 term, ($|z_k^2| = 1$)

$$\alpha_2(z_1^2 + z_2^2 + \gamma) = 0 \qquad \Rightarrow \qquad |z_1^2 + z_2^2| = 2 \qquad \Rightarrow \qquad z_1^2 = z_2^2, \qquad (114)$$

a contradiction.

Now, we define the set $K_\epsilon \subset K$ as

$$K_\epsilon = \{\boldsymbol{u} \in K \mid |z_1 - z_2| < \epsilon\} \qquad (115)$$

and note that J_0 is nonsingular in $K \backslash K_\epsilon$ by the result above, and the definition of D_f. Hence, J_0, J_0^{-1} and Λ_1 are continuous matrix functions of $\boldsymbol{u}$ on the compact set $K \setminus K_\epsilon$ and there is thus a $\boldsymbol{u}^*$ such that

$$\sup_{\boldsymbol{u} \in K \backslash K_\epsilon} ||J_0 \Lambda_1 J_0^{-1}|| = ||J_0 \Lambda_1 J_0^{-1}|| \Big|_{\boldsymbol{u} = \boldsymbol{u}^*} \equiv M_1 < \infty. \qquad (116)$$

If $K_\epsilon = \emptyset$ we are done. Otherwise, take $\boldsymbol{u} \in K_\epsilon$. Then the difference between the corresponding angles can be estimated, $|\theta_2 - \theta_1| \equiv |\Delta\theta| \leq C_0\epsilon$, for some constant C_0. Let $J_0(\theta_1, \theta_1 + \Delta\theta)\boldsymbol{y} = \boldsymbol{x}$ and define

$$J_0 = \begin{pmatrix} I & I \\ J(0) & J(\Delta\theta) \end{pmatrix}, \ \Lambda_1 = \begin{pmatrix} \Lambda(0) & 0 \\ 0 & \Lambda(\Delta\theta) \end{pmatrix}, \ \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}, \ \boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}. \tag{117}$$

Substituting $\boldsymbol{y}_2 = \boldsymbol{x}_1 - \boldsymbol{y}_1$ we get

$$J(0)\boldsymbol{y}_1 + J(\Delta\theta)\boldsymbol{y}_2 = \boldsymbol{x}_2, \Rightarrow -\delta J \equiv -(J(\Delta\theta) - J(0))\boldsymbol{y}_1 = \boldsymbol{x}_2 - J(\Delta\theta)\boldsymbol{x}_1. \tag{118}$$

Moreover,

$$\begin{aligned} J_0\Lambda_1\boldsymbol{y} &= \begin{pmatrix} \Lambda(0) & \Lambda(\Delta\theta) \\ J(0)\Lambda(0) & J(\Delta\theta)\Lambda(\Delta\theta) \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 \\ -\boldsymbol{y}_1 + \boldsymbol{x}_1 \end{pmatrix} \\ &= -\begin{pmatrix} [\Lambda(\Delta\theta) - \Lambda(0)]\boldsymbol{y}_1 \\ [J(\Delta\theta)\Lambda(\Delta\theta) - J(0)\Lambda(0)]\boldsymbol{y}_1 \end{pmatrix} + \begin{pmatrix} \Lambda(\Delta\theta)\boldsymbol{x}_1 \\ J(\Delta\theta)\Lambda(\Delta\theta)\boldsymbol{x}_1 \end{pmatrix} \\ &\equiv \begin{pmatrix} \delta\Lambda(\delta J)^{-1}(\boldsymbol{x}_2 - J(\Delta\theta)\boldsymbol{x}_1) \\ \delta(J\Lambda)(\delta J)^{-1}(\boldsymbol{x}_2 - J(\Delta\theta)\boldsymbol{x}_1) \end{pmatrix} + \begin{pmatrix} \Lambda(\Delta\theta)\boldsymbol{x}_1 \\ J(\Delta\theta)\Lambda(\Delta\theta)\boldsymbol{x}_1 \end{pmatrix}. \end{aligned} \tag{119}$$

Thus,

$$\begin{aligned} ||J_0\Lambda_1 J_0^{-1}|| &= \sup_{\boldsymbol{x}\in\mathbb{R}^4} \frac{||J_0\Lambda_1 J_0^{-1}\boldsymbol{x}||}{||\boldsymbol{x}||} = \sup_{\boldsymbol{x}\in\mathbb{R}^4} \frac{||J_0\Lambda_1\boldsymbol{y}||}{||\boldsymbol{x}||} \\ &\leq \sup_{\boldsymbol{x}\in\mathbb{R}^4} \frac{(||\delta\Lambda(\delta J)^{-1}|| + ||\delta(J\Lambda)(\delta J)^{-1}||)||\boldsymbol{x}_2 - J(\Delta\theta)\boldsymbol{x}_1||}{||\boldsymbol{x}||} \\ &\quad + \frac{(||\Lambda(\Delta\theta)|| + ||J(\Delta\theta)\Lambda(\Delta\theta)||)||\boldsymbol{x}_1||}{||\boldsymbol{x}||}. \end{aligned} \tag{120}$$

It is clear that $||\Lambda(\Delta\theta)||$ and $||J(\Delta\theta)||$ can both be bounded independently of θ_1 and $\Delta\theta$. Also, $||\boldsymbol{x}_k|| \leq ||\boldsymbol{x}||$ for $k = 1, 2$. Therefore, there are constants C_1, C_2 such that

$$||J_0\Lambda_1 J_0^{-1}|| \leq C_1(||\delta\Lambda(\delta J)^{-1}|| + ||\delta(J\Lambda)(\delta J)^{-1}||) + C_2 \tag{121}$$

Define $\delta J'$ by

$$\delta J' = \begin{pmatrix} j_{22} & -j_{12} \\ -j_{21} & j_{11} \end{pmatrix}, \qquad \text{where} \qquad \delta J = \begin{pmatrix} j_{11} & j_{12} \\ j_{21} & j_{22} \end{pmatrix} \tag{122}$$

so that $(\delta J)^{-1} = \delta J'/\det\delta J$. Since $\delta\Lambda\delta J'$ is an analytic function of $\Delta\theta$ we can write it as a Taylor series round $\Delta\theta = 0$. We have

$$\delta\Lambda\delta J' = \sum_{k=4}^{\infty} A_k(\theta_1)(\Delta\theta)^k. \tag{123}$$

This shows that also $\delta\Lambda\delta J'/(\Delta\theta)^4$ is an analytic function of $\Delta\theta$. For $\Delta\theta \leq C_0\epsilon$ we can now get the estimate $||\delta\Lambda\delta J'|| < C_3(\Delta\theta)^4$, with C_3 a constant independent of θ_1. The uniformity in θ_1 follows from the fact that $\delta\Lambda\delta J'$ is also analytic as a function of θ_1 for all $\Delta\theta$. Similarly,

$$\delta(J\Lambda)\delta J' = \sum_{k=4}^{\infty} B_k(\theta_1)(\Delta\theta)^k, \tag{124}$$

gives $||\delta(J\Lambda)\delta J'|| < C_4(\Delta\theta)^4$. Finally,

$$\frac{|\Delta\theta|^4}{|\det \delta J|} = \frac{|\Delta\theta|^4}{|-1+2\cos^2\Delta\theta - \cos^4\Delta\theta|} = \frac{|\Delta\theta|^4}{|-(\Delta\theta)^4 + \mathcal{O}((\Delta\theta)^6)|} \leq C_5. \tag{125}$$

We get

$$\sup_{u\in K_\epsilon} ||J_0\Lambda_1 J_0^{-1}|| \leq C_1C_5(C_3+C_4) + C_2 \equiv M_2. \tag{126}$$

With $M = \max(M_1, M_2)$ the result in the theorem follows. □

Remark. The first terms in the Taylor series (123) and (124) are

$$A_4 = \begin{pmatrix} \cos\theta(2\cos^4\theta - \frac{7}{2}\cos^2\theta + 2) & \sin\theta(-2\cos^4\theta + \frac{5}{2}\cos^2\theta - 1) \\ \sin\theta(2\cos^4\theta - \frac{3}{2}\cos^2\theta - \frac{1}{2}) & \cos\theta(2\cos^4\theta - \frac{5}{2}\cos^2\theta) \end{pmatrix} \tag{127}$$

and

$$B_4 = \begin{pmatrix} -10\cos^7\theta + \frac{39}{2}\cos^5\theta - 12\cos^3\theta + 4\cos\theta) & \sin\theta(10\cos^6\theta - \frac{29}{2}\cos^4\theta + 6\cos^2\theta) \\ \sin\theta(10\cos^6\theta - \frac{23}{2}\cos^4\theta + 3\cos^2\theta - \frac{3}{2}) & 10\cos^7\theta - \frac{33}{2}\cos^5\theta + \frac{15}{2}\cos^3\theta \end{pmatrix} \tag{128}$$

respectively.