

OLOF RUNBORG

**Some new results in multiphase geometrical optics**

*ESAIM: Modélisation mathématique et analyse numérique*, tome 34, n° 6 (2000),  
p. 1203-1231

[http://www.numdam.org/item?id=M2AN\\_2000\\_\\_34\\_6\\_1203\\_0](http://www.numdam.org/item?id=M2AN_2000__34_6_1203_0)

© SMAI, EDP Sciences, 2000, tous droits réservés.

L'accès aux archives de la revue « ESAIM: Modélisation mathématique et analyse numérique » (<http://www.esaim-m2an.org/>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques  
<http://www.numdam.org/>

## SOME NEW RESULTS IN MULTIPHASE GEOMETRICAL OPTICS\*

OLOF RUNBORG<sup>1</sup>

**Abstract.** In order to accommodate solutions with multiple phases, corresponding to crossing rays, we formulate geometrical optics for the scalar wave equation as a kinetic transport equation set in phase space. If the maximum number of phases is finite and known *a priori* we can recover the exact multiphase solution from an associated system of moment equations, closed by an assumption on the form of the density function in the kinetic equation. We consider two different closure assumptions based on delta and Heaviside functions and analyze the resulting equations. They form systems of nonlinear conservation laws with source terms. In contrast to the classical eikonal equation, these equations will incorporate a “finite” superposition principle in the sense that while the maximum number of phases is not exceeded a sum of solutions is also a solution. We present numerical results for a variety of homogeneous and inhomogeneous problems.

**Résumé.** Afin d'exhiber des solutions possédant des phases multiples, et dans l'objectif de traiter le cas de rayons qui se croisent, nous formulons l'optique géométrique pour l'équation d'ondes scalaire comme une équation cinétique de transport posée dans l'espace des phases. Si le nombre maximum de phases est fini et connu *a priori*, nous reconstruisons la solution multivaluée exacte en résolvant un système associé d'équations de moments. Nous fermons ce système en faisant deux hypothèses différentes sur la forme particulière de la fonction densité dans l'équation cinétique, basée sur des fonctions de Dirac et de Heaviside. Nous analysons les équations résultantes. Elles forment des systèmes de lois de conservation non linéaires avec termes source. Contrairement à l'équation eikonale classique, ces équations permettent d'obtenir un principe de superposition “fini”, dans le sens suivant : tant que le nombre maximum de phases n'est pas excédé, une somme de solutions du système obtenu demeure une solution. Nous présentons des résultats numériques pour un certain nombre de problèmes homogènes et non homogènes.

**Mathematics Subject Classification.** 35L65, 65M06, 78A05.

Received: March 3, 2000. Revised: August 14, 2000.

---

*Keywords and phrases.* Geometrical optics, multivalued traveltimes, eikonal equation, kinetic equations, conservation laws, moment equations, finite difference methods, nonstrictly hyperbolic system.

\* The author was partially supported by the NSF KDI grant DMS-9872890. Part of this work was carried out at the Department of Numerical Analysis and Computing Science, KTH, Stockholm, Sweden and at Laboratoire d'Analyse Numérique, Université Paris VI, Paris, France, within the EU TMR network on Hyperbolic Systems of Conservation Laws, H.C.L. ERBFM-RXCT960033.

<sup>1</sup> Program in Applied and Computational Mathematics, Fine Hall, Princeton University, Princeton, NJ 08544, USA.  
e-mail: orunborg@math.princeton.edu

## INTRODUCTION

We consider the linear scalar wave equation,

$$u_{tt} - c(\mathbf{x})^2 \Delta u = 0, \quad (t, \mathbf{x}) \in \mathbb{R}^+ \times \mathbb{R}^d \quad (1)$$

together with initial and/or boundary data that generate high-frequency solutions. The exact form of the data will not be important, but a typical example would be  $u(t, x_1, 0) = A(t, x_1)e^{i\omega(t+x_1)}$  with the frequency  $\omega \gg 1$ . In the direct numerical simulation of (1) the accuracy of the solution is determined by the number of grid points used per wavelength and dimension. The computational cost to maintain constant accuracy grows algebraically with the frequency and for sufficiently high frequencies a direct approach is no longer feasible. Approximate methods are needed.

In this paper we consider geometrical optics, which is the asymptotic approximation obtained when the frequency tends to infinity. This approximation is widely used in applications such as computational electromagnetics, acoustics, optics and geophysics. Instead of the oscillating wave field  $u$ , the unknowns in the geometrical optics equations are the phase  $\phi$  and the amplitude  $A$ , which both vary on a much slower scale than  $u$ . They should hence in principle be easier to compute numerically.

The derivation of the geometrical optics equations in the linear case is classical. See for instance the book by Whitham [35]. Formally, they follow if we write  $u$  as a series expansion of the form

$$u(t, \mathbf{x}) = e^{i\omega\phi(t, \mathbf{x})} \sum_{k=0}^{\infty} A_k(t, \mathbf{x})(i\omega)^{-k}. \quad (2)$$

Entering this expression into (1) and summing terms of the same order in  $\omega$  to zero, we obtain separate equations for the unknown variables in (2). The phase function  $\phi$  will satisfy the Hamilton-Jacobi type *eikonal equation*,

$$\phi_t + c(\mathbf{x}) |\nabla\phi| = 0, \quad (3)$$

and  $A_0$  solves the *transport equation*,

$$(A_0)_t + c \frac{\nabla\phi \cdot \nabla A_0}{|\nabla\phi|} + \frac{c^2 \Delta\phi - \phi_{tt}}{2c |\nabla\phi|} A_0 = 0. \quad (4)$$

For large  $\omega$  we can discard the remaining terms in (2).

Some typical wave phenomena, such as diffraction and interference, are lost in the infinite frequency approximation. Moreover, the approximation breaks down at caustics, where the amplitude,  $A_0$ , blows up. For some situations correction terms can be derived, such as those given by Keller [21], in the 1960s, with his pioneering geometrical theory of diffraction (GTD), further developed by for instance Kouyoumjian and Pathak [22]. A closer study of the solution's asymptotic behavior close to caustics was done by Ludwig [27], and Kravtsov [23], among others. We will however not treat these refined theories in this paper.

The traditional way to compute traveltimes of high-frequency waves is through *ray tracing*. The traveltime of a wave is given directly by the phase function  $\phi$ , and ray tracing corresponds to solving the eikonal equation (3) through the method of characteristics, *i.e.* solving the system of ordinary differential equations (ODE),

$$\frac{d\mathbf{x}}{dt} = \nabla_p H(\mathbf{x}, \mathbf{p}), \quad \frac{d\mathbf{p}}{dt} = -\nabla_x H(\mathbf{x}, \mathbf{p}), \quad H(\mathbf{x}, \mathbf{p}) = c(\mathbf{x})|\mathbf{p}|. \quad (5)$$

There are also ODEs for the amplitude. Very many rays are often required to cover the full computational domain, which can make ray tracing a rather inefficient method [33, 34]. Also for smoothly varying  $c(\mathbf{x})$  there may be shadow zones where the field is hard to resolve. In addition, it is difficult to compute the amplitude and to find the minimum traveltime in regions where rays cross.

Recently several new methods based on partial differential equations (PDE) formulations have been proposed to avoid some of the drawbacks of ray tracing. The interest has focused on solving the eikonal equation (3). One problem with (3) is that it cannot produce solutions with multiple phases, corresponding to crossing rays. There is no superposition principle. With extra conditions given in [11] there is, however, always a unique, generalized, solution to (3), known as the *viscosity solution*. At points where the correct physical solution should have a multivalued phase, the viscosity solution picks out the phase corresponding to the first arriving wave. Hence, the eikonal equation gives the first arrival traveltimes.

Upwind finite difference methods were used in [33, 34] to compute the viscosity solution of (3). High resolution methods of ENO type for Hamilton-Jacobi equations were introduced in [28] and used for the eikonal equation in [13]. In the latter a first attempt was also made to compute multivalued traveltimes. A second phase, corresponding to the second arrival time was calculated 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. Other interesting approaches to finding multivalued traveltimes, also based on the eikonal equation, are the big ray tracing method of Benamou [1, 2], which uses unstructured grids, and the slowness matching algorithm of Symes [31]. See also [3], where the multivalued solutions are computed directly using an accompanying equation that detects where the the phase field becomes multivalued.

Brenier and Corrias [8], proposed a different way of finding the multivalued solutions in the one-dimensional homogeneous case, subsequently adapted for two-dimensional inhomogeneous problems by Engquist and Runborg [14, 15]. Geometrical optics can be reformulated as a “kinetic” Liouville equation in phase space and the method is based on a closure assumption for the equations representing the moments of this Liouville equation. This is a classical way to approximate kinetic transport equations, the standard example being the compressible Euler approximation of the Boltzmann equation. See for instance [18] and more recently [25]. In the geometrical optics case, however, the closure assumption and the resulting equations are typically *exact* for a large class of problems.

This paper is a continuation of the investigations in [14, 15]. We give a theoretical motivation for the choice of moments and some additional analysis of the resulting equations, both for the closure assumption based on delta functions and for a new closure assumption based on Heaviside functions. We also include a numerical study of some homogeneous and inhomogeneous problems. We restrict ourselves to the two-dimensional case throughout the paper and for much of the analysis this restriction is crucial. It should be noted, however, that the delta function closure has a direct three-dimensional equivalent. Our approach would have to be modified to analyze its properties, though. The paper is organized as follows. We first derive the general moment equations in Section 1. Next, in Sections 2 and 3 we discuss the two different closure assumptions and the equations they result in. Finally, in Section 4 we show numerical results.

## 1. MOMENT EQUATIONS FOR GEOMETRICAL OPTICS

The kinetic formulation of geometrical optics is based on the interpretation that rays are trajectories of particles (photons) following the Hamiltonian dynamics in (5). We introduce  $\eta = 1/c$ , the *index of refraction* and  $f(t, \mathbf{x}, \mathbf{p}) \geq 0$ , the density of particles in phase space. We note that if we let  $f$  initially only be supported where  $|\mathbf{p}| = \eta$ , it will remain so for all times  $t > 0$ , since  $dH/dt = 0$  in (5). Using the Liouville theorem and the constraint  $|\mathbf{p}| = \eta$ , we get a Liouville equation describing the evolution of  $f$ ,

$$f_t + \frac{1}{\eta^2} \mathbf{p} \cdot \nabla_x f + \frac{1}{\eta} \nabla_x \eta \cdot \nabla_p f = 0. \quad (6)$$

This equation could also be derived directly from (1) using H-measures or Wigner measures, see [16, 26, 32]. The precise relationship between the stationary version of (6) and the high frequency limit of the Helmholtz equation with a source term was recently studied in [4, 10].

In order to solve the full equation (6) by direct numerical methods all independent variables (six in 3D) need to be discretized. This would require unrealistic computational time. Instead, it was observed in [8], when  $f$  is of a simple form in  $\mathbf{p}$ , the transport equation (6) can be transformed into a finite system of moment equations in the reduced space  $(t, \mathbf{x})$ . In this paper we consider the two-dimensional case. Let us define the moments  $m_{ij}$ , with  $\mathbf{p} = (p_1, p_2)^T$ , as

$$m_{ij}(t, \mathbf{x}) = \frac{1}{\eta(\mathbf{x})^{i+j}} \int_{\mathbb{R}^2} p_1^i p_2^j f(t, \mathbf{x}, \mathbf{p}) d\mathbf{p}. \quad (7)$$

Next, multiply (6) by  $\eta^{2-i-j} p_1^i p_2^j$  and integrate over  $\mathbb{R}^2$  with respect to  $\mathbf{p}$ . Assuming that  $\eta(x)$  is smooth, it follows that the moments  $m_{ij}$  formally satisfy the infinite system of moment equations,

$$(\eta^2 m_{ij})_t + (\eta m_{i+1,j})_x + (\eta m_{i,j+1})_y = i\eta_x m_{i-1,j} + j\eta_y m_{i,j-1} - (i+j)(\eta_x m_{i+1,j} + \eta_y m_{i,j+1}), \quad (8)$$

valid for all  $i, j \geq 0$ . For uniformity in notation we have defined  $m_{i,-1} = m_{-1,i} = 0$ ,  $\forall i$ .

The system (8) is not closed. If truncated at finite  $i$  and  $j$ , there are more unknowns than equations. To close the system we will make specific assumptions on the form of the density function  $f$ . First we will consider the case when  $f$  is a weighted sum of delta functions in  $\mathbf{p}$  and second, the case when  $f$  is a sum of Heaviside functions. Both cases correspond to the assumption of a finite number of rays at each point in time and space.

## 2. CLOSURE WITH DELTA FUNCTIONS

To close (8) we assume in this section that  $f$  can be written

$$f(t, \mathbf{x}, \mathbf{p}) = \sum_{k=1}^N g_k \cdot \delta(\mathbf{p} - \mathbf{p}_k), \quad \mathbf{p}_k = \eta \begin{pmatrix} \cos \theta_k \\ \sin \theta_k \end{pmatrix}. \quad (9)$$

Hence, for fixed values of  $\mathbf{x}$  and  $t$ , the particle density  $f$  is non-zero at a maximum of  $N$  points, and only when  $|\mathbf{p}| = \eta(\mathbf{x})$ . The new variables that we have introduced here are  $g_k = g_k(t, \mathbf{x})$ , which corresponds to the strength (particle density) of ray  $k$ , and  $\theta_k = \theta_k(t, \mathbf{x})$  which is the direction of the same ray. Inserting (9) into (7) yields

$$m_{ij} = \sum_{k=1}^N g_k \cos^i \theta_k \sin^j \theta_k. \quad (10)$$

A system describing  $N$  phases, needs  $2N$  equations, corresponding to the  $N$  ray strengths  $g_k$  and their directions  $\theta_k$ . It is not immediately clear which equations to select among the candidates in (8). Given the equations for a set of  $2N$  moments one should be able to write the remaining moments in these equations in terms of the leading ones. This is not always possible. For instance, with the choice of  $m_{20}$  and  $m_{02}$ , for  $N = 1$ , the quadrant of the angle  $\theta$  cannot be recovered, and therefore in general not the sign of the moments. We choose here the equations for the moments  $m_{2\ell-1,0}$  and  $m_{0,2\ell-1}$  with  $\ell = 1, \dots, N$ ,

$$\begin{aligned} (\eta^2 m_{2\ell-1,0})_t + (\eta m_{2\ell,0})_x + (\eta m_{2\ell-1,1})_y &= (2\ell-1)(\eta_x m_{2\ell-2,0} - \eta_x m_{2\ell,0} - \eta_y m_{2\ell-1,1}), \\ (\eta^2 m_{0,2\ell-1})_t + (\eta m_{1,2\ell-1})_x + (\eta m_{0,2\ell})_y &= (2\ell-1)(\eta_y m_{0,2\ell-2} - \eta_x m_{1,2\ell-1} - \eta_y m_{0,2\ell}), \end{aligned}$$

and collect those moments in a vector,

$$\mathbf{m} = (m_{10}, m_{01}, m_{30}, m_{03}, \dots, m_{2N-1,0}, m_{0,2N-1})^T. \quad (11)$$

As we will show below in Section 2.1, this system of equations for  $\mathbf{m}$  can be essentially closed, for all  $N$ . We introduce new variables,

$$\mathbf{u} = (u_1, u_2, \dots, u_{2N-1}, u_{2N})^T := (g_1 \cos \theta_1, g_1 \sin \theta_1, \dots, g_N \cos \theta_N, g_N \sin \theta_N)^T, \quad (12)$$

which have a physical interpretation; the vector  $(u_{2k-1}, u_{2k})$  shows the direction and strength of ray  $k$ . The new variables together with (10) define a function  $\mathbf{F}_0$  through the equation

$$\mathbf{F}_0(\mathbf{u}) = \mathbf{m}. \quad (13)$$

Similarly, they define the functions

$$\mathbf{F}_1(\mathbf{u}) = (m_{20}, m_{11}, \dots, m_{2N,0}, m_{1,2N-1})^T, \quad \mathbf{F}_2(\mathbf{u}) = (m_{11}, m_{02}, \dots, m_{2N-1,1}, m_{0,2N})^T, \quad (14)$$

$$\mathbf{K}(\mathbf{u}, \eta_x, \eta_y) = \begin{pmatrix} \eta_x m_{00} - \eta_x m_{2,0} - \eta_y m_{1,1} \\ \eta_y m_{00} - \eta_x m_{1,1} - \eta_y m_{0,2} \\ \vdots \\ (2N-1)(\eta_x m_{2N-2,0} - \eta_x m_{2N,0} - \eta_y m_{2N-1,1}) \\ (2N-1)(\eta_y m_{0,2N-2} - \eta_x m_{1,2N-1} - \eta_y m_{0,2N}) \end{pmatrix}.$$

These functions permit us to write the equations as a system of nonlinear conservation laws with source terms,

$$\mathbf{F}_0(\eta^2 \mathbf{u})_t + \mathbf{F}_1(\eta \mathbf{u})_x + \mathbf{F}_2(\eta \mathbf{u})_y = \mathbf{K}(\mathbf{u}, \eta_x, \eta_y). \quad (15)$$

Equivalently, we can write (15) as

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

The functions  $\mathbf{F}_j$  and  $\mathbf{K}$  are rather complicated nonlinear functions. In Appendix A.1 they are given for the cases  $N = 1, 2$ . Since the angles  $\theta_k$  remain unaffected when  $\mathbf{u}$  is scaled by a constant, all  $\mathbf{F}_j$  and  $\mathbf{K}$  are homogeneous of degree one,  $\mathbf{F}_j(\alpha \mathbf{u}) = \alpha \mathbf{F}_j(\mathbf{u})$ ,  $\mathbf{K}(\alpha \mathbf{u}, \eta_x, \eta_y) = \alpha \mathbf{K}(\mathbf{u}, \eta_x, \eta_y)$  for all  $\alpha \in \mathbb{R}$ . The source term  $\mathbf{K}$  always vanishes for constant  $\eta$ .

## 2.1. Properties of the flux functions

In this section we analyze the flux functions and source

$$\mathbf{F}_1 \circ \mathbf{F}_0^{-1}(\mathbf{m}), \quad \mathbf{F}_2 \circ \mathbf{F}_0^{-1}(\mathbf{m}), \quad \mathbf{K}(\mathbf{F}_0^{-1}(\mathbf{m}), \eta_x, \eta_y).$$

In order for them to be well defined we must restrict their domain to the case when there are no rays meeting head-on. With this restriction they are also continuous. We have

**Theorem 2.1.** *Let  $\mathbf{F}_0$  be the function in (13) and let  $\mathbf{F}_0|_{U_N}$  be its restriction to the domain*

$$U_N = \{\mathbf{u} \in \mathbb{R}^{2N} \mid 1 + \cos(\theta_k - \theta_\ell) \neq 0, \text{ whenever } g_k g_\ell > 0, \forall k, \ell\},$$

and  $M_N = \mathbf{F}_0(U_N)$ . The composition  $m \circ (\mathbf{F}_0|_{U_N})^{-1} : M_N \mapsto \mathbb{R}$  is well-defined and continuous for all maps of the form

$$m : U_N \mapsto \mathbb{R}, \quad m(\mathbf{u}) = \sum_{k=1}^N g_k h(\theta_k), \quad (16)$$

where  $h : \mathbb{S} \mapsto \mathbb{R}$  is continuous.

Since  $\mathbf{F}_1, \mathbf{F}_2$  and  $\mathbf{K}$  are all of the form (16) we have:

**Corollary 2.2.** *Let  $\mathbf{F}_j$  and  $\mathbf{K}$  be the functions in (13, 14) and let  $\mathbf{F}_0|U_N$  and  $M_N$  be as in Theorem 2.1. Then the functions*

$$\mathbf{F}_1 \circ (\mathbf{F}_0|U_N)^{-1}(\mathbf{m}), \quad \mathbf{F}_2 \circ (\mathbf{F}_0|U_N)^{-1}(\mathbf{m}), \quad \mathbf{K}((\mathbf{F}_0|U_N)^{-1}(\mathbf{m}), \eta_x, \eta_y)$$

are well defined and depend continuously on  $\mathbf{m} \in M_N$ .

**Remark 2.3.** If we do not restrict  $\mathbf{F}_0$  to  $U_N$  the result is false. Take for instance  $\mathbf{u} = (-1 \ 0 \ 1 \ 0)^T$  and  $\tilde{\mathbf{u}} = 2\mathbf{u}$  for  $N = 2$  so that  $\mathbf{F}_0(\mathbf{u}) = \mathbf{F}_0(\tilde{\mathbf{u}}) = 0$ , but  $\mathbf{F}_1(\tilde{\mathbf{u}}) = 2\mathbf{F}_1(\mathbf{u}) \neq 0$ . Furthermore, with a different choice of moment equations the result does not necessarily hold either. For instance, if instead of (11) we use the equations for

$$\mathbf{m} = (m_{10}, m_{01}, m_{20}, m_{02})^T,$$

when  $N = 2$ , the functions  $\mathbf{F}_j$  change and in general there are two unrelated solutions to  $\mathbf{F}_0(\mathbf{u}) = \mathbf{m}$  which  $\mathbf{F}_1$  does not map to the same point. For example, if  $\mathbf{u} = (1 \ 1 \ 0 \ -1)^T$  and  $\tilde{\mathbf{u}} = (1 \ -1 \ 0 \ 1)^T$  then  $\mathbf{F}_0(\mathbf{u}) = \mathbf{F}_0(\tilde{\mathbf{u}})$ , but  $\mathbf{F}_1(\mathbf{u}) = \mathbf{F}_1(\tilde{\mathbf{u}}) + (0 \ \sqrt{2} \ 0 \ 0)^T$ . The function  $\mathbf{F}_2 \circ \mathbf{F}_0^{-1}$  is ill defined in the same way.

*Proof of Theorem 2.1*

It will be convenient to work with complex versions of our variables and we start by introducing the isometry  $\mathcal{A} : \mathbb{R}^{2N} \mapsto \mathbb{C}^N$ ,

$$\mathcal{A}(x_1, \dots, x_{2N})^T = (x_1 + ix_2, \dots, x_{2N-1} + ix_{2N})^T.$$

Set  $\mathbf{w} = (w_1, \dots, w_N)^T := \mathcal{A}\mathbf{u}$  and

$$z_k := \cos \theta_k + i \sin \theta_k, \quad \mathbf{z}_k := \left( z_k, z_k^{-3}, \dots, z_k^{(2N-1)(-1)^{N+1}} \right)^T, \tag{17}$$

so that  $w_k = g_k z_k$ . Furthermore, define the continuous mapping  $\mathbf{Q} : \mathbb{C}^N \mapsto \mathbb{C}^N$ ,

$$\mathbf{Q}(\mathbf{w}) = \begin{pmatrix} | & | & & | \\ \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_N \\ | & | & & | \end{pmatrix} \begin{pmatrix} g_1 \\ \vdots \\ g_N \end{pmatrix}.$$

To relate  $\mathbf{w}$  to  $\mathbf{m}$  via this function, we use the trigonometric identity

$$\mathbf{z}_k = B \begin{pmatrix} \cos \theta_k + i \sin \theta_k \\ \vdots \\ \cos^{2N-1} \theta_k + i \sin^{2N-1} \theta_k \end{pmatrix}, \tag{18}$$

where  $B = \{b_{k\ell}\} \in \mathbb{R}^{N \times N}$  is a lower triangular matrix with  $b_{k\ell}$  equal to the  $(2\ell - 1)$ th coefficient of the  $(2k - 1)$ th degree Chebyshev polynomial, for  $k \leq \ell$ . The matrix is non-singular since  $b_{kk} = 4^{k-1} > 0$ . From the definition of  $\mathbf{Q}$  and the identity (18) it then follows that

$$\mathbf{Q}(\mathbf{w}) = \mathbf{Q}(\mathcal{A}\mathbf{u}) = B\mathbf{A}\mathbf{m}, \tag{19}$$

where we also recall that  $\mathbf{F}_0(\mathbf{u}) = \mathbf{m}$ . Before continuing we show the following lemma.

**Lemma 2.4.** *Let  $\{z_k\}$  be  $N'$  complex numbers such that  $|z_k| = 1$  and let  $\{\mathbf{z}_k\}$  be the corresponding vectors as defined in (17). If  $N' \leq 2N$  then  $\mathbf{z}_k \in \mathbb{C}^N$  are linearly independent over  $\mathbb{R}$  if and only if*

$$z_k^2 \neq z_\ell^2, \quad k \neq \ell. \tag{20}$$

*Proof.* The necessity is obvious. To show that (20) is a sufficient condition, we only need to consider the case  $N' = 2N$ , since we can always find  $2N - N'$  additional  $z_k$  such that (20) still holds if  $N' < 2N$ . Suppose therefore that  $\{\mathbf{z}_k\}_{k=1}^{2N}$  are linearly dependent over  $\mathbb{R}$ , and that (20) is true. Then the real matrix

$$A = \begin{pmatrix} \Re(\mathbf{z}_1) & \Re(\mathbf{z}_2) & \cdots & \Re(\mathbf{z}_{2N}) \\ \Im(\mathbf{z}_1) & \Im(\mathbf{z}_2) & \cdots & \Im(\mathbf{z}_{2N}) \end{pmatrix}, \quad A \in \mathbb{R}^{2N \times 2N},$$

is singular and we can find a vector  $\beta = (\beta_1, \dots, \beta_{2N})^T \neq 0$  such that  $A^T \beta = 0$ . Using the fact that  $|z_k| = 1$  and  $\bar{z}_k = 1/z_k$ , this implies

$$P_\beta(z_k^2) = 0, \quad k = 1, \dots, 2N,$$

where

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

But since the degree of  $P_\beta$  is at most  $2N - 1$ , regardless of  $\beta$ , it cannot have  $2N$  distinct zeros if  $\beta \neq 0$ . Therefore, there must exist  $k, \ell$  such that  $z_k^2 = z_\ell^2$ , a contradiction.  $\square$

Let  $\bar{m}(\mathbf{w}) := m(\mathcal{A}^{-1}\mathbf{w})$  and let  $\bar{\mathbf{Q}}$  be the restriction of  $\mathbf{Q}$  to  $\mathcal{AU}_N$ . We now want to prove that  $\bar{m} \circ \bar{\mathbf{Q}}^{-1}$  is well defined on  $\bar{\mathbf{Q}}(\mathcal{AU}_N)$  and we do this by showing that  $\bar{\mathbf{Q}} \circ \bar{m}^{-1}$  is injective on  $\bar{m}(\mathcal{AU}_N)$ . Let  $\mathbf{w}, \tilde{\mathbf{w}} \in \mathcal{AU}_N$  be such that  $\bar{\mathbf{Q}}(\mathbf{w}) = \bar{\mathbf{Q}}(\tilde{\mathbf{w}})$ . We need to show that  $\bar{m}(\mathbf{w}) = \bar{m}(\tilde{\mathbf{w}})$  and we use the variables introduced in (17). A tilde indicates that a variable relates to  $\tilde{\mathbf{w}}$ . Let  $N'$  and  $\tilde{N}'$  respectively be the number of distinct  $z_k$  and  $\tilde{z}_k$  with  $g_k, \tilde{g}_k > 0$ . Without loss of generality we order the variables such that  $z_{\ell_j} = \dots = z_{\ell_{j+1}-1}$ , with  $1 = \ell_1 < \dots < \ell_{N'+1} = N + 1$ , and similar for  $\{\tilde{z}_k\}$ . With this notation we get

$$\bar{\mathbf{Q}}(\mathbf{w}) = \sum_{j=1}^{N'} \left( \sum_{k=\ell_j}^{\ell_{j+1}-1} g_k \right) \mathbf{z}_{\ell_j} = \sum_{j=1}^{\tilde{N}'} \left( \sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} \tilde{g}_k \right) \tilde{\mathbf{z}}_{\tilde{\ell}_j} = \bar{\mathbf{Q}}(\tilde{\mathbf{w}}).$$

The sets of numbers  $\{z_{\ell_j}\}_{j=1}^{N'}$  and  $\{\tilde{z}_{\tilde{\ell}_j}\}_{j=1}^{\tilde{N}'}$  both satisfy (20), because  $\mathbf{w}, \tilde{\mathbf{w}} \in \mathcal{AU}_N$ . Therefore, since  $N' + \tilde{N}' \leq 2N$ , there must exist  $j$  and  $k$  such that  $z_{\ell_j}^2 = \tilde{z}_{\tilde{\ell}_k}^2$  by Lemma 2.4. By induction it follows that  $N' = \tilde{N}'$  and, possibly after some reordering,

$$\ell_j = \tilde{\ell}_j, \quad z_{\ell_j} = s_j \tilde{z}_{\tilde{\ell}_j}, \quad \sum_{k=\ell_j}^{\ell_{j+1}-1} g_k = s_j \sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} \tilde{g}_k, \quad s_j = \pm 1, \quad \forall j.$$

But  $g_k, \tilde{g}_k$  are positive, and we can conclude that  $s_j = 1$  for all  $j$ . Thus,  $\mathbf{w}$  and  $\tilde{\mathbf{w}}$  are identical up to permutations and to the individual  $g_k$  values. We now apply  $\bar{m}$  to them.

$$\bar{m}(\mathbf{w}) = \sum_{j=1}^{N'} \sum_{k=\ell_j}^{\ell_{j+1}-1} g_k h(z_k) = \sum_{j=1}^{N'} h(z_{\ell_j}) \sum_{k=\ell_j}^{\ell_{j+1}-1} g_k = \sum_{j=1}^{N'} h(\tilde{z}_{\tilde{\ell}_j}) \sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} \tilde{g}_k = \sum_{j=1}^{N'} \sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} \tilde{g}_k h(\tilde{z}_k) = \bar{m}(\tilde{\mathbf{w}}).$$

Hence,  $\bar{m} \circ \bar{\mathbf{Q}}^{-1}$  is well-defined on its domain of definition. Now, (19) and the fact that  $\mathbf{F}_0(\mathbf{u}) = \mathbf{m}$  show that  $m \circ (\mathbf{F}_0|_{U_N})^{-1}(\mathbf{m}) = \bar{m} \circ \bar{\mathbf{Q}}^{-1}(B\mathcal{A}\mathbf{m})$ , which implies that  $m \circ (\mathbf{F}_0|_{U_N})^{-1}$  is well defined on  $M_N$ . The continuity follows by approximating  $U_N$  by compact sets, and using the following lemma from elementary analysis, which we do not prove.

**Lemma 2.5.** *Let  $U$  be a compact metric space and  $f, g$  two continuous functions on  $U$ . Suppose  $f : U \mapsto X$  and  $g : U \mapsto Y$  where  $X, Y$  are metric spaces and  $X = f(U)$ . If the composition  $f \circ g^{-1} : g(U) \mapsto X$  is injective then  $g \circ f^{-1} : X \mapsto Y$  is continuous.*

### 2.2. Analysis of the conservation laws

In [14, 15] it was shown that the general system (15) is nonstrictly hyperbolic for all states  $\mathbf{u}$  and  $N$ . The systems are thus not well-posed in the strong sense, and they are more sensitive to perturbations than strictly hyperbolic systems. The Jacobian has a Jordan type degeneracy and there will never be more than  $N$  linearly independent eigenvectors for the  $2N \times 2N$  system. For a general study of this type of degenerate systems of conservation laws, see [36].

A distinguishing feature of the system (15) is that it typically has measure solutions of delta function type, even for smooth and compactly supported initial data. These appear when the physically correct solution passes outside the class of solutions that the system (15) describes. If initial data dictates a physical solution with  $M$  phases for  $t > T$ , the system (15) with  $N < M$  phases will have a measure solution for  $t > T$ .

For smooth solutions, (15) with  $N$  phases is equivalent to  $N$  pairs of eikonal and transport equations (3, 4) if the variables are identified as

$$g_k = A_{0,k}^2, \quad (\cos \theta_k, \sin \theta_k)^T = \frac{\nabla \phi_k}{|\nabla \phi_k|}, \quad k = 1, \dots, N,$$

see [14]. The pair (3, 4) form a nonstrictly hyperbolic system, just like (15), with the same eigenvalue. Where wave fields meet, the viscosity solution of (3) is in general discontinuous. Because of the term  $\Delta \phi$  in the source term of (4), the first amplitude coefficient  $A_0$  has a concentration of mass at these points. Hence, the two different formulations are similar also in this respect.

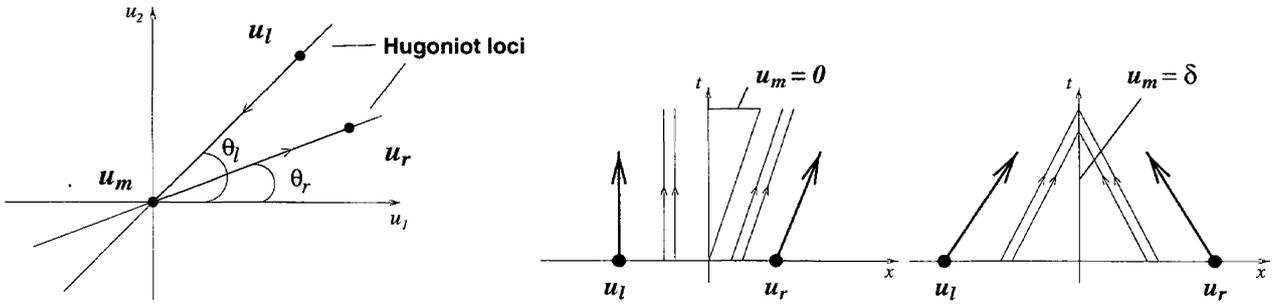
There is a close relationship between (15) with  $N = 1$  and  $\eta \equiv 1$ ,

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x + \mathbf{g}(\mathbf{u})_y = 0, \quad \mathbf{f}(\mathbf{u}) = u_1 \frac{\mathbf{u}}{|\mathbf{u}|}, \quad \mathbf{g}(\mathbf{u}) = u_2 \frac{\mathbf{u}}{|\mathbf{u}|}, \tag{21}$$

and the equations of pressureless gases,

$$\begin{aligned} \rho_t + (\rho u)_x &= 0, \\ (\rho u)_t + (\rho u^2)_x &= 0. \end{aligned} \tag{22}$$

In fact, the steady state version of (21) is precisely (22) if we identify  $\rho = g \cos^2 \theta$  and  $u = \tan \theta$ . Moreover, the one-dimensional version of (21) corresponds to (22) with relativistic effects added if we identify  $\rho = g \sin \theta$  and  $u = \cos \theta$ . In the context of non-relativistic pressureless gases this problem was addressed by Bouchut [5] and later Brenier and Grenier [9, 19], and E *et al.* [12], who independently proved global existence of measure solutions to (22). The uniqueness question was settled in [7]. For linear transport equations related results have been obtained by Bouchut and James [6] and Poupaud and Rascle [29]. The questions of existence and uniqueness for (21) and its one-dimensional version are still open.



(a) Hugoniot loci of states and solution for contact discontinuity.

(b) Contact discontinuity.

(c) Overcompressive shock.

FIGURE 1. The Riemann problem, with Hugoniot loci for the left and right states in phase space and the two different types of discontinuities in  $(t, \mathbf{x})$  space.

2.2.1. The Riemann problem

Since standard numerical schemes are based on solving one-dimensional Riemann problems [24], we consider this problem for (21),

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \quad \mathbf{f}(\mathbf{u}) = u_1 \frac{\mathbf{u}}{|\mathbf{u}|}, \quad \mathbf{u}(0, x) = \begin{cases} \mathbf{u}_\ell & x < 0, \\ \mathbf{u}_r & x > 0. \end{cases} \quad (23)$$

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

$$\mathbf{f}(\mathbf{u}_\ell) - \mathbf{f}(\mathbf{u}_r) = s(\mathbf{u}_\ell - \mathbf{u}_r), \quad (24)$$

where  $s$  represents the propagation speed of the discontinuity. Since  $\mathbf{f}(\mathbf{u}) = \cos \theta \mathbf{u}$ , the jump condition (24) simplifies to

$$\cos \theta_\ell \mathbf{u}_\ell - \cos \theta_r \mathbf{u}_r = s(\mathbf{u}_\ell - \mathbf{u}_r).$$

The states to which a given non-zero state  $\mathbf{u}_\ell$  can connect with a discontinuity, *i.e.* its Hugoniot locus, is simply  $\alpha \mathbf{u}_\ell$  for  $\alpha \in \mathbb{R}$ , with speed of propagation  $s = \cos \theta_\ell$  when  $\alpha \geq 0$  and  $s = \cos \theta_\ell(1 + \alpha)/(1 - \alpha)$  for  $\alpha < 0$ . It follows that, unless they are parallel, two non-zero states  $\mathbf{u}_\ell$  and  $\mathbf{u}_r$  can only be connected *via* the intermediate state  $\mathbf{u}_m = 0$ . There will be two types of discontinuities. If  $\cos \theta_\ell < \cos \theta_r$ , the solution with  $\mathbf{u}_m = 0$ , satisfies the Lax entropy condition (the left discontinuity moves slower than the right one). The states' Hugoniot loci and the solution for this type of discontinuity is illustrated in Figure 1a. If  $\cos \theta_\ell > \cos \theta_r$ , on the other hand, we do not have a solution in the usual weak sense. This situation corresponds to two meeting wave fields. Formally, however,  $\mathbf{u}_m = t \tilde{\mathbf{u}}_m \delta(x - st)$  is a weak solution to the conservation law with this initial data. The conservation form gives a slightly modified jump condition,

$$\cos \theta_\ell \mathbf{u}_\ell - \cos \theta_r \mathbf{u}_r = \cos \tilde{\theta}_m (\mathbf{u}_\ell - \mathbf{u}_r) + \tilde{\mathbf{u}}_m,$$

with the propagation speed  $s = \cos \tilde{\theta}_m$ . This construction, a delta function solution to the Riemann problem leading to a modified Rankine-Hugoniot condition, is found also in [36] for more general equations.

It is easily verified that  $\mathbf{u}$  itself is an eigenvector of the Jacobian of  $\mathbf{f}$  and that the Jacobian has a double eigenvalue equaling  $\cos \theta$ . Therefore, the Hugoniot locus will coincide with the integral curves of the system's characteristic fields and, since  $\cos \theta$  remains constant along the curves, the fields are linearly degenerate. From

this we 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  $(t, \mathbf{x})$ -space, are shown in Figures 1b and 1c.

2.2.2. Entropy

For the analysis of (21) it would be useful to find a strictly convex entropy pair for the one-dimensional system. This is, however, not possible since the system is nonstrictly hyperbolic. There do however exist nonstrictly convex entropy pairs, which can be characterized as follows.

**Theorem 2.6.** *Let  $U \in C^2$  be convex. There exists a function  $F \in C^2$  such that  $U(\mathbf{u})_t + F(\mathbf{u})_x = 0$  for all smooth solutions  $\mathbf{u} = g(\cos \theta, \sin \theta)$  to*

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \quad \mathbf{f}(\mathbf{u}) = u_1 \frac{\mathbf{u}}{|\mathbf{u}|}, \tag{25}$$

if and only if  $U$  is of the form

$$U = gh(\theta) + \text{const}, \quad h \in C^2(\mathbb{S}), \quad h + h'' \geq 0.$$

*Proof.* We must show that there exists  $F \in C^2$  such that that  $\nabla U \frac{d\mathbf{f}}{d\mathbf{u}} = \nabla F$ , which is equivalent to the condition

$$\nabla \times \left( \nabla U \frac{d\mathbf{f}}{d\mathbf{u}} \right) = \cos \theta \frac{\partial^2 U}{\partial g^2} = 0.$$

Hence,  $F$  exists if and only if  $U = gh_1(\theta) + h_2(\theta)$  for some  $h_1, h_2 \in C^2$ . The determinant of the Hessian matrix of  $U$  then equals  $-h_2'(\theta)^2/g^4$ . Since  $U$  is convex, this must be nonnegative, showing that  $h_2 = \text{const}$ . With this restriction, we finally compute the eigenvalues of the Hessian, which turn out to be  $\lambda_1 = 0$  and

$$\lambda_2 = \frac{h_1(\theta) + h_1''(\theta)}{g}.$$

This proves the theorem. □

2.2.3. Superposition

The multiple phase systems possess a finite superposition principle in the sense that a sum of  $N$  solutions to the single phase system, is a solution to the  $N$ -phase system. This follows from a trivial computation if the solutions are smooth. Physical solutions can, however, well have discontinuities in  $g$ . On the other hand, a discontinuous  $\theta$  would typically not be physical, generating a delta shock type solution, as seen in Section 2.2.1. In fact, if we introduce weak solutions we can for instance show that a sufficient condition for the superposition principle to hold is just that  $g$  is bounded and that  $\theta$  is continuous and has locally bounded variation.

**Theorem 2.7.** *Suppose  $\{\mathbf{u}_k\}_{k=1}^N$  are  $N$  weak solutions to the homogeneous single phase system (21) in the sense that  $\mathbf{u}_k \in L^\infty((0, \infty) \times \mathbb{R}^2)$  and*

$$\iint_{t \geq 0} \mathbf{u}_k \phi_t + \mathbf{f}(\mathbf{u}_k) \phi_x + \mathbf{g}(\mathbf{u}_k) \phi_y \, dt \, d\mathbf{x} = 0, \quad \forall \phi \in C_c^1((0, \infty) \times \mathbb{R}^2). \tag{26}$$

Moreover, suppose that for each  $k$  and each point in  $(0, \infty) \times \mathbb{R}^2$ , there is an open neighborhood on which we can define a continuous function  $\theta_k(t, \mathbf{x})$  with locally bounded variation such that  $\mathbf{u}_k = |\mathbf{u}_k|(\cos \theta_k, \sin \theta_k)^T$  on that neighborhood. Then  $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_N)^T$  is a weak solution to the homogeneous  $N$ -phase system (15) in the same sense as (26).

*Proof.* We start by showing that if  $\mathbf{v} = (v_1, v_2)^T$  is a weak solution to (21) in the sense of Theorem 2.7, then  $m_{i,0}$  and  $m_{0,i}$ , with  $i > 1$ , are weak solutions in the same sense to the corresponding moment equations, under the given hypotheses. Take  $\phi \in C_c^1((0, \infty) \times \mathbb{R}^2)$  and assume without loss of generality that  $\theta$  is continuous and that  $\mathbf{v} = g(\cos \theta, \sin \theta)^T$  on  $\text{supp } \phi$ . (We can always obtain such a  $\theta$  after a partition of unity.) Let  $M \in C_c^\infty(\mathbb{R}^3)$  be a mollifier with  $\int M \, dt \, d\mathbf{x} = 1$  and set  $\theta_\epsilon = \theta \star M_\epsilon$ , where  $M_\epsilon = M(t/\epsilon, \mathbf{x}/\epsilon)/\epsilon^3$ . Furthermore, set

$$\psi_s^\epsilon = \phi \left( \cos^{i-1} \theta_\epsilon - \frac{d \cos^{i-1} \theta_\epsilon}{d\theta_\epsilon} \sin \theta_\epsilon \cos \theta_\epsilon \right), \quad \psi_c^\epsilon = \phi \frac{d \cos^{i-1} \theta_\epsilon}{d\theta_\epsilon} \cos^2 \theta_\epsilon. \tag{27}$$

We observe that  $\phi_t \cos^i \theta_\epsilon = (\psi_s^\epsilon)_t \cos \theta_\epsilon + (\psi_c^\epsilon)_t \sin \theta_\epsilon$ , and similar for the partial derivatives with respect to  $x$  and  $y$ . Also,  $m_{i,0} = g \cos^i \theta$  on the support of  $\phi$ . This shows that, for all  $\epsilon$ ,

$$\iint_{t \geq 0} m_{i,0} \phi_t \, dt \, d\mathbf{x} = \iint_{t \geq 0} (m_{i,0} - m_{i,0}^\epsilon) \phi_t + (v_1^\epsilon - v_1) (\psi_s^\epsilon)_t + (v_2^\epsilon - v_2) (\psi_c^\epsilon)_t + v_1 (\psi_s^\epsilon)_t + v_2 (\psi_c^\epsilon)_t \, dt \, d\mathbf{x},$$

where a superscripted  $\epsilon$  denotes that a function depends on  $\theta_\epsilon$  instead of  $\theta$ . The first term of the right-hand side tends to zero by the dominated convergence theorem. Since  $\theta \in \text{BV}_{\text{loc}}$  the expression  $\|\phi \partial_t \theta_\epsilon\|_{L^1}$  is bounded independently of  $\epsilon$ , and therefore

$$\iint_{t \geq 0} |v_1^\epsilon - v_1| |(\psi_s^\epsilon)_t| \, dt \, d\mathbf{x} \leq C \sup_{(t, \mathbf{x}) \in \text{supp } \phi} |v_1^\epsilon - v_1| \leq C \|\mathbf{v}\|_{L^\infty} \sup_{(t, \mathbf{x}) \in \text{supp } \phi} |\cos \theta_\epsilon - \cos \theta| \rightarrow 0,$$

by the continuity of  $\theta$ . Using the same argument for the remaining terms, we arrive at

$$\begin{aligned} \iint_{t \geq 0} m_{i,0} \phi_t + m_{i+1,0} \phi_x + m_{i,1} \phi_y \, d\mathbf{x} \, dt &= \iint_{t \geq 0} v_1 (\psi_s^\epsilon)_t + \frac{v_1^2}{|\mathbf{v}|} (\psi_s^\epsilon)_x + \frac{v_1 v_2}{|\mathbf{v}|} (\psi_s^\epsilon)_y \, d\mathbf{x} \, dt \\ &+ \iint_{t \geq 0} v_2 (\psi_c^\epsilon)_t + \frac{v_2 v_1}{|\mathbf{v}|} (\psi_c^\epsilon)_x + \frac{v_2^2}{|\mathbf{v}|} (\psi_c^\epsilon)_y \, d\mathbf{x} \, dt + R^\epsilon, \end{aligned} \tag{28}$$

where  $R^\epsilon \rightarrow 0$ . But  $\psi_c^\epsilon, \psi_s^\epsilon \in C_c^1((0, \infty) \times \mathbb{R}^2)$  and  $\mathbf{v}$  is a weak solution, so by letting  $\epsilon \rightarrow 0$  we see that (28) in fact equals zero. After replacing  $\cos^{i-1} \theta$  with  $\sin^{i-1} \theta$  in (27) we get the same result for  $m_{0,i}$ . We can now conclude that with  $m_{i,j}^k = g_k \cos^i \theta_k \sin^j \theta_k$ ,

$$\sum_{k=1}^N \iint_{t \geq 0} m_{2\ell-1,0}^k \phi_t + m_{2\ell,0}^k \phi_x + m_{2\ell-1,1}^k \phi_y \, d\mathbf{x} \, dt = 0, \quad \ell = 1, \dots, N.$$

The same is true for  $m_{0,2\ell-1}^k$ . But these are just the componentwise statements of

$$\iint_{t \geq 0} \mathbf{F}_0(\mathbf{u}) \phi_t + \mathbf{F}_1(\mathbf{u}) \phi_x + \mathbf{F}_2(\mathbf{u}) \phi_y = 0,$$

and  $\|\mathbf{u}\|_{L^\infty}$  is bounded by  $\sum_{k=1}^N \|\mathbf{u}_k\|_{L^\infty}$ . □

Of course some of the  $\mathbf{u}_k$  solutions in Theorem 2.7 can be identically zero, so that in particular a weak solution of the single phase system is also a solution of the  $N$ -phase system, under the above assumptions.

### 3. CLOSURE WITH HEAVISIDE FUNCTIONS

We will now consider a different way to close (8). We discard the amplitude information carried by  $g_k$  used in Section 2 and only solve it for the  $\theta_k$ . In this way we get fewer and less singular equations. The ‘‘correct’’

values of the unknowns  $\theta_k$  are, however, not well defined when the physically motivated amplitude is zero. In particular, this is the case at time  $t = 0$  for the typical initial value problem with sources given through boundary values (like the problems in Section 4). In order to reduce the initialization problem we will therefore only consider steady state solutions to (8) where no rays go in the negative  $x$ -direction. The equations can then be “time-stepped” in the positive  $x$ -direction and data only need to be given on the line  $x = 0$ . In other words, we put an additional restriction on  $f$  that  $f(t, \mathbf{x}, \mathbf{p}) = 0$  when  $\mathbf{p} \cdot \mathbf{e}_x \leq 0$  where  $\mathbf{e}_x$  is the unit vector in the  $x$ -direction. We then consider density functions of the form

$$f(t, \mathbf{x}, \mathbf{p}) = \begin{cases} \frac{1}{\eta} \delta(|\mathbf{p}| - \eta) \sum_{k=1}^N (-1)^{k+1} H(\theta - \theta_k(t, \mathbf{x})), & \mathbf{p} \cdot \mathbf{e}_x > 0, \\ 0, & \text{otherwise,} \end{cases} \quad \mathbf{p} = |\mathbf{p}| \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}, \quad (29)$$

with the convention that  $-\pi/2 < \theta_1 \leq \dots \leq \theta_N < \pi/2$ . For the time being we will assume that  $N$  is even. The general formula for the moments follows directly from (29) together with (7)

$$m_{ij}(t, \mathbf{x}) = \sum_{k=1}^N (-1)^{k+1} \int_{\theta_k(t, \mathbf{x})}^{\pi/2} \cos^i \theta \sin^j \theta \, d\theta. \quad (30)$$

For fixed  $(t, \mathbf{x})$ , the density function  $f$  is supported by a set of intervals on the sphere  $\{|\mathbf{p}| = \eta\}$ . The intervals correspond to fans of rays whose edges are given by the unknown angles  $\theta_k$ . The transport equation (6) governs the propagation of all these rays, and in particular the rays at the edges, which will propagate just like ordinary rays as long as  $f$  stays of the form (29). The values of the  $N$  angles  $\theta_k$  will then coincide with those of a problem with  $N$  rays crossing at each point, as long as the assumption (29) holds.

Among the equations in (8) we choose the ones for the moments  $\{m_{0,\ell}\}$  with  $\ell = 0, \dots, N - 1$ . This leads to the steady state equations

$$(\eta m_{1,\ell})_x + (\eta m_{0,\ell+1})_y = \ell(\eta_y m_{0,\ell-1} - \eta_x m_{1,\ell} - \eta_y m_{0,\ell+1}), \quad \ell = 0, \dots, N - 1. \quad (31)$$

Next, we introduce the new variables,

$$\mathbf{u} = (u_1, \dots, u_N)^T, \quad u_k = \sin \theta_k. \quad (32)$$

By evaluating the integrals in (30) we then get

$$m_{1,\ell} = \sum_{k=1}^N \frac{(-1)^k u_k^{\ell+1}}{\ell + 1}, \quad m_{0,\ell} = \sum_{k=1}^N (-1)^k R_\ell(u_k), \quad R_\ell = \begin{cases} \arcsin(u), & \ell = 0, \\ -\sqrt{1 - u^2}, & \ell = 1, \\ \frac{\ell-1}{\ell} R_{\ell-2} - \frac{1}{\ell} u^{\ell-1} \sqrt{1 - u^2}, & \ell \geq 2. \end{cases} \quad (33)$$

Although these expressions were derived for even  $N$ , we will in fact use them to define the moments also for odd  $N$ . See below for a motivation of this. In the subsequent analysis it will be convenient to use an alternative, equivalent, expression for the moments. We let

$$\tilde{f} = \sum_{k=1}^N (-1)^{k+1} H(u - u_k).$$

Then, if  $|u_k| \leq L < 1$ , we can express the moments as

$$m_{1,\ell} = \begin{cases} \langle \tilde{f}, u^\ell \rangle, & N \text{ even,} \\ \langle \tilde{f} - H(u), u^\ell \rangle, & N \text{ odd,} \end{cases} \quad m_{0,\ell} = \begin{cases} \left\langle \tilde{f}, \frac{u^\ell}{\sqrt{1-u^2}} \right\rangle, & N \text{ even,} \\ \left\langle \tilde{f}, \frac{u^\ell}{\sqrt{1-u^2}} \right\rangle - R_\ell(L), & N \text{ odd,} \end{cases} \quad (34)$$

where  $\langle \cdot, \cdot \rangle$  denotes the standard inner product on  $L^2[-L, L]$ .

Like in Section 2 we collect the moments in a vector,  $\mathbf{m} = (m_{10}, \dots, m_{1,N-1})^T$ . We define the function  $\mathbf{F}_1$  by  $\mathbf{F}_1(\mathbf{u}) = \mathbf{m}$  together with (33) or (34), and similarly for  $\mathbf{F}_2$  and  $\mathbf{K}$ . We can then finally write (31) as

$$(\eta \mathbf{F}_1(\mathbf{u}))_x + (\eta \mathbf{F}_2(\mathbf{u}))_y = \mathbf{K}(\mathbf{u}, \eta_x, \eta_y), \tag{35}$$

or, in terms of  $\mathbf{m}$ ,

$$(\eta \mathbf{m})_x + (\eta \mathbf{F}_2 \circ \mathbf{F}_1^{-1}(\mathbf{m}))_y = \mathbf{K}(\mathbf{F}_1^{-1}(\mathbf{m}), \eta_x, \eta_y).$$

The functions  $\mathbf{F}_j$  and  $\mathbf{K}$  are again rather complicated nonlinear functions. See Appendix A.2 for their precise form when  $N = 1, 2, 3$ .

If  $u_k < u_{k+1}$  for all  $k$ , we can compute the gradient of  $m_{0,\ell}(\mathbf{m})$  explicitly,

$$\nabla_{\mathbf{m}} m_{0,\ell} = V^{-1} \Theta_{\ell}. \tag{36}$$

Here  $V = \{v_{k,\ell}\} \in \mathbb{R}^{N \times N}$  is the Vandermonde matrix associated to the points  $\mathbf{u}$ , i.e.  $v_{k,\ell} = u_k^{\ell-1}$  (nonsingular by the assumption on  $\mathbf{u}$ ), and

$$\Theta_{\ell} = \left\{ \frac{u_k^{\ell}}{\sqrt{1-u_k^2}} \right\}_{k=1}^N = \{u_k^{\ell-1} \tan \theta_k\}_{k=1}^N \in \mathbb{R}^N.$$

By using (36) we get an expression for the Jacobian of  $\mathbf{F}_2 \circ \mathbf{F}_1^{-1}$ ,

$$\frac{d}{d\mathbf{m}} \mathbf{F}_2 \circ \mathbf{F}_1^{-1} = V^T \text{diag}(\{\tan \theta_k\}) V^{-T}.$$

We see that this system is strictly hyperbolic as long as  $\theta_k \neq \theta_{\ell}$  for all  $k, \ell$ . The theory for the system is standard and we will not further discuss its structure. We just note that since  $\tan \theta \rightarrow \infty$  when  $|\theta| \rightarrow \pi/2$  or  $|u| \rightarrow 1$  the Jacobian will blow up at these points. This is expected by the assumption that the equation can be time-stepped in the  $x$ -direction.

It remains to motivate the moment equations when  $N$  is odd. For this case, let  $(u_1, \dots, u_{N+1})^T$  be the solution to the, well defined, equations with  $N + 1$  phases. By making the change of variables  $u = \frac{\phi_y}{|\nabla \phi|}$  in (3) when  $\phi$  is smooth, we get the scalar conservation law

$$(\eta u)_x - \left( \eta \sqrt{1-u^2} \right)_y = 0, \tag{37}$$

for the steady state solution. Assume now that there is a solution  $u^*$  to (37), such that  $u^*(x, y) > u_N(x, y)$  for all  $(x, y)$  and all solutions  $u_N$  that we are interested in. For sufficiently smooth solutions (see Th. 3.1) the superposition principle is valid and we can substitute the fixed  $u^*$  for  $u_{N+1}$  and the  $N + 1$  equations are still satisfied. We can therefore reduce the  $N + 1$  equations to only  $N$  equations with  $N$  unknowns. For example, in the homogeneous case when  $\eta = \text{const.}$ , we can take  $u^* \equiv 1$ . In practice the reduction amounts to simply doing what we mentioned above: use the expression of the moments in (33) also when  $N$  is odd.

We close this section by establishing the same superposition principle as for the delta equations in Section 2.

**Theorem 3.1.** *Suppose  $\{u_k\}_{k=1}^M$  are  $M$  weak solutions to (35) with  $N = 1$  in the sense of Theorem 2.7, and  $\eta \in C^1$ . If  $u_k$  are continuous functions with locally bounded variation, then  $\mathbf{u} = (u_1, \dots, u_M)^T$  is a weak solution to (35) with  $N = M$  in the same sense.*

The proof is essentially the same as for Theorem 2.7 and we leave it out.

**Remark 3.2.** The kinetic version of (37) is

$$f_x + a_v f_y - a_y f_v = 0, \quad f = f(x, y, v), \quad a = a(x, y, v) = -\sqrt{\eta^2 - v^2}.$$

If we dictate that  $f = 0$  whenever  $|v| > \eta(\mathbf{x})$  and treat this equation according to the method in [8], we arrive at exactly the same moment equations as those in (31).

### 3.1. Properties of the flux function

We will here show that the functions

$$\mathbf{F}_2 \circ \mathbf{F}_1^{-1}(\mathbf{m}) \quad \text{and} \quad \mathbf{K}(\mathbf{F}_1^{-1}(\mathbf{m}), \eta_x, \eta_y) \tag{38}$$

are well defined and regular on their domains of definition. The results will be given for a slightly more general class of functions than those in (38). We start by introducing some notation. For a closed interval  $I \subset \mathbb{R}$ , let  $F_N(I) \subset L^1(I)$  be defined by

$$F_N(I) = \left\{ f \in L^1(I) \mid f(u) = \sum_{k=1}^N (-1)^{k+1} H(u - u_k), \quad u_1 \leq \dots \leq u_N, \quad u_k \in I \right\}.$$

Note that these functions correspond to the function  $\tilde{f}$  in Section 3. For simplicity we drop the tilde in this section. Also introduce the (compact) set of attainable moments,  $M_N \subset \mathbb{R}^N$ ,

$$M_N(I) = \{m(f) \mid f \in F_N(I)\}, \quad m(f) = \begin{cases} \{\langle f, u^{k-1} \rangle\}_{k=1}^N, & N \text{ even,} \\ \{\langle f - H(u), u^{k-1} \rangle\}_{k=1}^N, & N \text{ odd,} \end{cases}$$

where  $\langle \cdot, \cdot \rangle$  is the  $L^2$  inner product on  $I$ . In [8] it was shown that if  $I = [0, \infty)$ , the mapping  $m(f)$  is one-to-one and functions of the type  $\langle m^{-1}(\mathbf{m}), \psi \rangle$  are continuous on  $M_N[0, L] \subset M_N(I)$  for each  $0 < L < \infty$ , when  $\psi$  has a strictly positive and bounded  $N$ th distributional derivative. This type of functions were identified as entropies for the moment system in [8]. We consider the compact interval  $I = [-L, L]$ , with also negative velocities, and show a slightly more complete regularity result for a larger class of  $\psi$ , including the regularity of (38).

Let  $P^n$  be the space of polynomials of degree at most  $n$  and let  $C^{0,\alpha}$  be the set of Hölder continuous functions of exponent  $\alpha$  with  $0 < \alpha \leq 1$ . In general  $J_\psi$  is Hölder continuous, but not continuously differentiable, as seen in:

**Theorem 3.3.** *The mapping  $J_\psi : M_N(I) \mapsto \mathbb{R}$  given by*

$$J_\psi(\mathbf{m}) = \langle \psi, f \rangle, \quad \mathbf{m} = m(f), \tag{39}$$

*is well defined. If  $\psi \in L^p(I)$ ,  $1 \leq p \leq \infty$  then  $J_\psi \in C^{0, \frac{p-1}{pN}}(M_N(I))$  (and  $C^0$  for  $p = 1$  and  $C^{0, 1/N}$  for  $p = \infty$ ). If  $N = 1$  and  $\psi \in C^M(I)$  then  $J_\psi \in C^{M+1}(M_N(I))$ . If  $N > 1$  and  $\psi \in C^M(I)$  with  $M < N$ , then  $J_\psi \in C^{0, \frac{1}{N-M}}(M_N(I))$ . If  $N > 1$  and  $\psi \in C^0(I)$ , then  $\nabla J_\psi$  is continuous almost everywhere, but it is discontinuous at  $\mathbf{m} = 0$ , unless  $\psi \in P^{N-1}(I)$ . If  $\psi \in P^{N-1}(I)$  then  $J_\psi \in C^\infty(M_N(I))$ .*

When  $|u_k| \leq L < 1$  it follows from (34) that, up to a constant, each element of the flux function is of the form (39) with  $\psi = u^\ell / \sqrt{1 - u^2}$ . The source function  $\mathbf{K}$  is of a similar form. Hence, we have:

**Corollary 3.4.** *The flux and source functions (38) are well defined and depend Lipschitz continuously on  $\mathbf{m} \in M_N[-L, L]$  when  $0 < L < 1$ . They are not continuously differentiable.*

*Proof of Theorem 3.3*

Without loss of generality we prove the theorem for  $L = 1$ . In the proof we will use the coefficient mapping  $\mathcal{C} : \mathbb{P}^{n-1} \mapsto \mathbb{C}^n$ ,

$$\mathcal{C}(c_0 + c_1x + \dots + c_{n-1}x^{n-1}) = (c_0, \dots, c_{n-1})^T.$$

The cases  $N = 1$  and  $\psi \in \mathbb{P}^{N-1}$  are trivial. For the remaining cases we need the following two lemmas.

**Lemma 3.5.** *Let  $h \geq 0$  belong to  $L^\infty[0, b]$ . For  $a \in [0, b]$  and integers  $m, n \geq 0$ , we have the sharp estimate*

$$\left( \int_0^a h(u) du \right)^{m+n+1} \leq \frac{(m+n+1)!}{m!n!} \|h\|_{L^\infty}^{m+n} \int_0^a h(u) u^m (a-u)^n du. \quad (40)$$

*Proof.* Denote the left and right-hand sides by  $f_L(m, n, a)$  and  $f_R(m, n, a)$  respectively. Assume  $h$  is continuous. Then  $\partial_a f_L$  and  $\partial_a f_R$  exist everywhere in  $[0, b]$  and since  $f_L(m, n, 0) = f_R(m, n, 0)$  we can prove (40) by showing that  $\partial_a f_L \leq \partial_a f_R$  in  $[0, b]$ . Since  $h \geq 0$ ,

$$\frac{\partial f_L(m, 0, a)}{\partial a} = (m+1)h(a) \left( \int_0^a h(u) du \right)^m \leq (m+1)h(a)a^m \|h\|_{L^\infty}^m = \frac{\partial f_R(m, 0, a)}{\partial a},$$

proving (40) when  $n = 0$ . Assume now that (40) holds for  $0 \leq n < p$ . Then it holds also for  $n = p$ , because

$$\frac{\partial f_L(m, p, a)}{\partial a} = (m+p+1)h(a)f_L(m, p-1, a) \leq \frac{(m+p+1)!}{m!p!} \|h\|_{L^\infty}^{m+p} \int_0^a h(u) u^m (a-u)^{p-1} du = \frac{\partial f_R(m, p, a)}{\partial a}.$$

The general case (40) follows by induction, and by approximating  $h$  with continuous functions. The sharpness of (40) is shown by taking  $h$  constant.  $\square$

**Lemma 3.6.** *For  $1 \leq n \leq N$  let*

$$P_n(u) = \prod_{k=2}^n (u_k - u) \in \mathbb{P}^{n-1}[-1, 1], \quad P_1 = 1. \quad (41)$$

*There is a constant  $C$  only depending on  $n$  and  $N$  such that*

$$\|(f_1 - f_2)P_n\|_{L^1} \leq C|m(f_1) - m(f_2)|^{1/(N-n+1)}, \quad \forall f_1, f_2 \in F_N[-1, 1].$$

*Proof.* Let  $f_1$  be defined by the points  $\{u_k\}$  and assume without loss of generality that  $f_2(u) = 0$  for  $u \leq u_1$ , so that  $(f_1 - f_2)P_n \geq 0$ . For some increasing sequence of points  $\{a_j\}$  and signs  $s_j \in \{-1, 1\}$ , we can write

$$f_1 - f_2 = \sum_{j=1}^{2N} s_j H(u - a_j), \quad \{u_k\}_{k=1}^N \subset \{a_j\}_{j=1}^{2N}.$$

Therefore,

$$\|(f_1 - f_2)P_n\|_{L^1}^{N-n+1} = \left( \sum_{j=1}^N \int_{a_{2j-1}}^{a_{2j}} |P_n| du \right)^{N-n+1} \leq C \sum_{j=1}^N \left( \int_{a_{2j-1}}^{a_{2j}} \tilde{s}_j P_n du \right)^{N-n+1}, \quad (42)$$

with  $\tilde{s}_j \in \{-1, 1\}$  and  $\tilde{s}_j P_n \geq 0$  in the interval. Let  $n_j$  be the largest non-negative number such that  $u_{n_j+n} \leq a_{2j-1}$ , or 0 if no such number exist. Since  $P_n$  is bounded in  $[-1, 1]$  independently of  $\{u_k\}$  we can use Lemma 3.5

and show that (42) is bounded by

$$\begin{aligned} &\leq C \sum_{j=1}^N \int_{a_{2j-1}}^{a_{2j}} \tilde{s}_j P_n(u) (u - a_{2j-1})^{n_j} (a_{2j} - u)^{N-n-n_j} du \leq C \sum_{j=1}^N \int_{a_{2j-1}}^{a_{2j}} |P_N(u)| du \\ &= C \langle f_1 - f_2, P_N \rangle = C(m(f_1) - m(f_2))^T \mathcal{C}P_N \leq C|m(f_1) - m(f_2)|, \end{aligned}$$

which proves the lemma. □

Lemma 3.6 assures that  $m$  is one-to-one and, consequently, that  $J_\psi$  is well defined on  $M_N(I)$ . When  $\psi \in L^1$ , let  $\mathbf{m}^n \rightarrow \mathbf{m} \in M_N$ . By Lemma 3.6 the corresponding  $f^n \rightarrow f$  in  $L^1$  and hence a subsequence  $f^{n_j} \rightarrow f$  a.e. By the dominated convergence theorem  $J_\psi(\mathbf{m}^{n_j}) \rightarrow J_\psi(\mathbf{m})$ , which remains true for  $\mathbf{m}^n$  since  $|f^n| \leq 1$ . When  $\psi \in L^p(I)$  and  $1 < p \leq \infty$ , let  $q = p/(p - 1)$  for  $p < \infty$  and  $q = 1$  for  $p = \infty$ . Then

$$|J_\psi(\mathbf{m}_1) - J_\psi(\mathbf{m}_2)| \leq \|\psi\|_p \|f_1 - f_2\|_q = \|\psi\|_p \|f_1 - f_2\|_{L^1}^{1/q} \leq C|\mathbf{m}_1 - \mathbf{m}_2|^{1/qN}, \quad \forall \mathbf{m}_1, \mathbf{m}_2 \in M_N,$$

by Hölder’s inequality and taking  $n = 1$  in Lemma 3.6.

Next, assume that  $\psi \in C^M(I)$  and  $1 \leq M < N$ . Take  $f_1, f_2 \in F_N(I)$ . Let  $f_1$  be defined by the points  $\{u_k\}$  and assume as above without loss of generality that  $f_2(u) = 0$  for  $u \leq u_1$ . Furthermore, let  $Q$  be the unique polynomial in  $P^{M-1}(I)$  which interpolates  $\psi$  at the points  $\{u_k\}_{k=2}^{M+1}$ . (If  $n$  points coincide, the first  $n - 1$  derivatives of  $Q$  and  $\psi$  should also agree at this point.) Then from standard results in approximation theory the error term in the interpolation has the form

$$\psi(u) - Q(u) = (-1)^M \frac{\psi^M(\xi)}{M!} P_{M+1}(u), \quad \xi \in I,$$

with  $P_{M+1}$  as in (41). Moreover,  $|\mathcal{C}Q|$  is bounded by the max norm of  $\psi$  and its first  $M - 1$  derivatives. Using also Lemma 3.6 and the boundedness of  $\mathbf{m}$ , we get for  $\mathbf{m}_1 = m(f_1)$  and  $\mathbf{m}_2 = m(f_2)$ ,

$$\begin{aligned} |J_\psi(\mathbf{m}_1) - J_\psi(\mathbf{m}_2)| &\leq C \sup_{\xi \in I} |\psi^M(\xi)| \cdot \|(f_1 - f_2)P_{M+1}\|_{L^1} + |\langle f_1 - f_2, Q \rangle| \\ &\leq C|\mathbf{m}_1 - \mathbf{m}_2|^{1/(N-M)} + |(\mathbf{m}_1 - \mathbf{m}_2)^T \mathcal{C}Q| \leq C|\mathbf{m}_1 - \mathbf{m}_2|^{1/(N-M)}. \end{aligned}$$

Finally, consider the compact set  $U^\epsilon = \{\mathbf{u} \in \mathbb{R}^N \mid u_k \leq u_{k+1} - \epsilon, u_k \in I\}$  for  $\epsilon > 0$ . Let  $g$  be the natural map from  $U^\epsilon$  to  $F_N$ , so that  $F_N = g(U^0)$ , and set  $M_N^\epsilon = m(g(U^\epsilon))$ . Since  $g$  is continuous and injective on  $g(U^\epsilon)$  when  $\epsilon > 0$ , the inverse map  $g^{-1} \circ m^{-1} : M_N^\epsilon \mapsto U^\epsilon$  is well defined and continuous by Lemma 3.6. As in (36) the gradient of  $J_\psi$  on  $M_N^\epsilon$  is explicitly given by  $\nabla_m J_\psi(\mathbf{m}) = V^{-1}(\mathbf{u})\Theta(\mathbf{u})$  for  $\mathbf{m} = m(g(\mathbf{u}))$ , with the same  $V$  as in (36) and  $\Theta = \{\psi(u_k)\} \in \mathbb{R}^N$ . Both  $V^{-1}$  and  $\Theta$  are continuous on  $U^\epsilon$ , when  $\psi$  is continuous. Since we can take arbitrarily small  $\epsilon$ , this shows the continuity of  $\nabla J_\psi(\mathbf{m})$  almost everywhere. It also implies  $\nabla J_\psi(\mathbf{m}) = \mathcal{C}P^*$ , where  $P^*$  is the unique polynomial in  $P^{N-1}(I)$  that interpolates  $\psi$  at all the points  $\mathbf{u} = \{u_k\}_{k=1}^N$ , given by

$$P^*(u) = \psi(u_1) + \sum_{\ell=2}^N \psi[u_1, \dots, u_\ell](u - u_1) \cdots (u - u_{\ell-1}). \tag{43}$$

Here  $[\cdot]$  denotes divided differences. To show that  $\nabla J_\psi(\mathbf{m})$  is not continuous at  $\mathbf{m} = 0$  when  $N$  is even, let  $\mathbf{u}^n = \{u_k^n\}$  be a sequence such that  $u_k^n \rightarrow a \in I$  for all  $k$ . Then  $f^n := g(\mathbf{u}^n) \rightarrow 0$  a.e. and consequently  $\mathbf{m}^n := m(f^n) \rightarrow 0$ . From (43) we deduce that  $(\nabla_m J_\psi(\mathbf{m}_n))_N = \psi[u_1^n, \dots, u_N^n]$ . The limit of this as  $n \rightarrow \infty$  either does not exist, depends on the specific sequence or tends to  $\psi^{(N-1)}(a)/(N-1)!$  by results in approximation theory. If  $\psi \notin P^{N-1}$  its  $(N - 1)$ th derivative is not constant, and since  $a$  was arbitrary,  $\nabla J_\psi$  is discontinuous at zero in all cases. When  $N$  is odd we use the same sequence  $\{u_k^n\}$  except that we let  $u_N^n \rightarrow 0$  so that  $\mathbf{m}_n \rightarrow 0$

as before. One can then show that  $0 = \partial_a \psi(0) = \partial_a (\lim_{n \rightarrow \infty} P^*(u_N^n))$  implies that  $\lim_{n \rightarrow \infty} (\nabla_m J_\psi(\mathbf{m}_n))_N$  is again independent of  $a$  only if  $\psi \in P^{N-1}$ .

#### 4. NUMERICAL RESULTS

In this section we show results of applying the equations derived in Sections 2 and 3 to a number of different test problems. We consider both homogeneous ( $\eta \equiv 1$ ) and inhomogeneous ( $\eta = \eta(\mathbf{x})$ ) media and use closures corresponding to  $N = 1, 2, 3$  crossing rays at each point. The equations closed with delta functions, (15), are set in two-dimensional space while the Heaviside equations, (35), are reduced to a one-dimensional space by treating  $x$  as a time-like variable. As a shorthand we will refer to the equations as the  $\delta$ - and the  $H$ -equations.

As we remarked in Section 2.2 the  $\delta$ -equations (15) are nonstrictly hyperbolic with linearly degenerate fields. This is reflected in their sensitivity to numerical treatment. It was shown in [14, 15] that even for smooth problems some standard numerical methods such as the Godunov scheme and the Nessyahu-Tadmor scheme with dimensional splitting converge poorly in  $L^1$  and may fail to converge in  $L^\infty$ . In the Godunov case this could be explained by the solution of the Riemann problem (23). In [20] it was shown that with a genuinely two-dimensional version of Nessyahu-Tadmor the expected second order convergence rate is obtained for smooth problems. This was confirmed in [30], where it was also demonstrated that a splitted version of Lax-Friedrichs had bad convergence properties. It appears that the dimensional splitting aggravates the numerical errors, although for the Godunov scheme, it was observed in [17], the same type of failure to converge in  $L^\infty$  can also occur in the much simpler case of a linear one-dimensional equation with variable coefficients.

Another difficulty for the  $\delta$ -equations is to evaluate the flux functions  $\mathbf{F}_1 \circ \mathbf{F}_0^{-1}$  and  $\mathbf{F}_2 \circ \mathbf{F}_0^{-1}$ . In both cases it is necessary to solve a nonlinear system of equations

$$\mathbf{F}_0(\mathbf{u}) = \mathbf{m}, \quad (44)$$

for each time step, at each grid point. Solving (44) can be difficult. An iterative solver must be used when  $N > 2$ , which is expensive and requires good initial values. In general, the Jacobian of  $\mathbf{F}_0$  is singular at some points in the computational domain. This happens when two rays are parallel. For iterative methods that use the Jacobian, this is a problem. When  $N = 1, 2$  we have found an analytical way to invert  $\mathbf{F}_0$ , see Appendix B.1. Furthermore, (44) may not have a solution. Although, for the exact solution of the PDE, (44) should always be satisfied, truncation errors in the numerical scheme may have perturbed the solution so that  $\mathbf{m}$  is not in  $M_N$ , the range of  $\mathbf{F}_0$ . We use the least squares solution of (44) when  $\mathbf{m} \notin M_N$ .

The  $H$ -equations (35) are strictly hyperbolic and numerical schemes are not as sensitive as for the  $\delta$ -equations. The evaluation of the flux functions is also easier, since it can be reduced to solving polynomial equations of low degree, see Appendix B.2. By accepting also complex roots of those polynomial equations,  $M_N$  the domain of definition of the flux function can be continuously extended, avoiding most problems when (44) does not have a solution.

The difficulties with the  $H$ -equations are of a different type. When the number of physically relevant phases is less than the number of phases supported by the system we must still give initial data for the nonexistent phases. In the delta case a near zero value can be given. (It is practical though not to use exactly zero since the flux functions have a weak singularity at zero.) Alternatively, the phase can be initialized to the same as another, physically relevant, phase. In the Heaviside case this is not as easy. The fictitious phases can obviously not be set to zero. Moreover, they cannot be set to the same as another phase. That would eliminate them from the equations. For instance, suppose there is only one relevant phase and a system with an even  $N > 1$  is solved. If all fictitious phases were initialized at  $x = 0$  to the same as the relevant phase, we would get  $m_{ij}(0, y) = 0$  for all  $i, j$  and the system would hence only give the trivial solution. If  $N$  had been odd the system would have been reduced to the  $N = 1$  case. Instead, the non-physical phases must be initialized to some other values. The solution can be sensitive to the right choice.

The main purpose of this section is to study the behavior of the exact solution to the PDEs, rather than their numerical properties. Therefore, in order to avoid problems with numerical artifacts we have used the

TABLE 1. Specifications of parameters used for the test problems. An asterisk indicates that a convergence study was also made for this problem with different grid sizes.

Problem	$\delta$ -equations				$H$ -equations		
	$N$	Grid Size	CFL	Sample Time	$N$	Grid Size	CFL
<i>Three point sources</i>	1	$80 \times 160$	0.65	1.0			
	2	$160 \times 320$	0.65	1.0			
	3	$40 \times 80$	0.65	1.0			
<i>Simple Caustic</i>	1	$256 \times 256$	0.65	3.0	2	1024*	0.4
	2	$512 \times 512$	0.65	3.0			
<i>Focus</i>	1	$80 \times 80$	0.65	4.0			
<i>Interface</i>	2	$80 \times 80^*$	0.65	5.0	3	256*	1.0
<i>Wedge</i>	1	$512 \times 512$	0.65	4.0	2	1024	1.0
	2	$512 \times 512$	0.65	5.0	3	2048	0.9
<i>Convex Lens</i>	1	$512 \times 512$	0.65	5.0	2	1024	1.0
	2	$512 \times 512$	0.65	5.0	3	2048	0.6

Lax-Friedrichs method. Although only first order accurate, it has proved to be the most robust and predictable method for these equations. For the numerical methods we will use the following notation. Space and time is discretized uniformly with step sizes  $\Delta x$ ,  $\Delta y$  and  $\Delta t$ . The grid functions  $\mathbf{U}_j^n$  (1D) and  $\mathbf{U}_{ij}^n$  (2D) approximates the exact solutions,

$$\mathbf{U}_{ij}^n \approx \mathbf{u}(n\Delta t, i\Delta x, j\Delta y), \quad \mathbf{U}_j^n \approx \mathbf{u}(n\Delta x, j\Delta y),$$

where  $\mathbf{u}$  are the variables introduced in (12) and (32) respectively. To measure the error of the computations we use the discrete  $L^1$ - and  $L^\infty$ -norms, defined as

$$\begin{aligned} \|\mathbf{U}\|_1 &= \Delta y \sum_i |\mathbf{U}_i|, & \|\mathbf{U}\|_\infty &= \max_i |\mathbf{U}_i|, & (1D), \\ \|\mathbf{U}\|_1 &= \Delta x \Delta y \sum_{i,j} |\mathbf{U}_{ij}|, & \|\mathbf{U}\|_\infty &= \max_{i,j} |\mathbf{U}_{ij}|, & (2D). \end{aligned}$$

In the computations we typically apply the exact solution as a Dirichlet boundary condition on the line  $x = 0$ . Where nothing else is said we use simple extrapolation,  $\mathbf{U}_{n+1,j} = \mathbf{U}_{nj}$ , etc. on the remaining boundaries. More detailed specifications of parameters used is listed in Table 1. For a more complete numerical study, see [30].

## 4.1. Homogeneous problems

### 4.1.1. Three point sources

We begin with a problem with three point sources located at coordinates  $\mathbf{s}_1 = (-0.5, 0.5)$ ,  $\mathbf{s}_2 = (-0.5, 1.0)$  and  $\mathbf{s}_3 = (-0.5, 1.5)$ . The exact solutions are  $\mathbf{w}_k(t, \mathbf{x}) = A_k(\mathbf{x} - \mathbf{s}_k)H(t - r_k)/r_k^2$ ,  $r_k = \|\mathbf{x} - \mathbf{s}_k\|$ ,  $k = 1, 2, 3$  where  $A_1 = 1.25$ ,  $A_2 = 0.75$  and  $A_3 = 1.0$ . The solution is computed in the rectangle  $[0, 1] \times [0, 2]$ . Figure 2 shows the solution at  $t = 1.0$  of the  $\delta$ -equations with  $N = 1, 2, 3$ . For the  $N = 3$  system, the exact solution was given at  $x = 0$ . For the  $N = 2$  system, the first two arriving waves were given at  $x = 0$ , *i.e.*

$$\mathbf{u}_1 = \mathbf{w}_2, \quad \mathbf{u}_2 = \begin{cases} \mathbf{w}_1 & r_1 < r_3, \\ \mathbf{w}_3 & r_1 \geq r_3. \end{cases}$$

TABLE 2. *Simple caustic*.  $L^1$ -norm of the errors in the sense of (45) at  $x = 0.2$ ,  $x = 0.55$  and  $x = 1.0$  for the  $H$ -system with  $N = 2$ . Here  $\Delta y = 2/n$ .

$L^1$	$x = 0.2$		$x = 0.55$		$x = 1.0$	
$n$	error	order	error	order	error	order
128	$1.61 \times 10^{-2}$		$3.12 \times 10^{-2}$		$8.08 \times 10^{-2}$	
		0.87		0.68		0.41
256	$8.79 \times 10^{-3}$		$1.95 \times 10^{-2}$		$6.10 \times 10^{-2}$	
		0.89		0.75		0.45
512	$4.75 \times 10^{-3}$		$1.16 \times 10^{-2}$		$4.45 \times 10^{-2}$	
		0.92		0.77		0.48
1024	$2.52 \times 10^{-3}$		$6.81 \times 10^{-3}$		$3.18 \times 10^{-2}$	
		0.94		0.91		0.44
2048	$1.31 \times 10^{-3}$		$3.62 \times 10^{-3}$		$2.34 \times 10^{-2}$	

Finally, for the  $N = 1$  system, the first arriving wave was given at  $x = 0$ ,

$$\mathbf{u} = \begin{cases} \mathbf{w}_1 & r_1 < r_2, r_1 < r_3, \\ \mathbf{w}_2 & r_2 \leq r_1, r_2 < r_3, \\ \mathbf{w}_3 & r_3 \leq r_1, r_3 \leq r_2. \end{cases}$$

As expected, the  $N = 3$  system is the only one solving this problem correctly. Delta functions appear in the solutions of the  $N = 1, 2$  systems. Note that the  $N = 2$  system, somewhat surprisingly, gives the waves of the outer two sources ( $k = 1, 3$ ) in the area along the line  $y = 1$ , cf. Figure 2e.

4.1.2. *Simple caustic*

In this problem we give data at  $x = 0$  such that a caustic forms,

$$g(0, y) = 1, \quad \sin \theta(0, y) = \begin{cases} -\sin 2\tilde{\theta}(y), & y > 1, \\ 0, & y \leq 1, \end{cases} \quad \sin \tilde{\theta}(y) = \frac{y - 1}{\sqrt{(y - 1)^2 + 4 \cos^2 \frac{\pi}{5}}}.$$

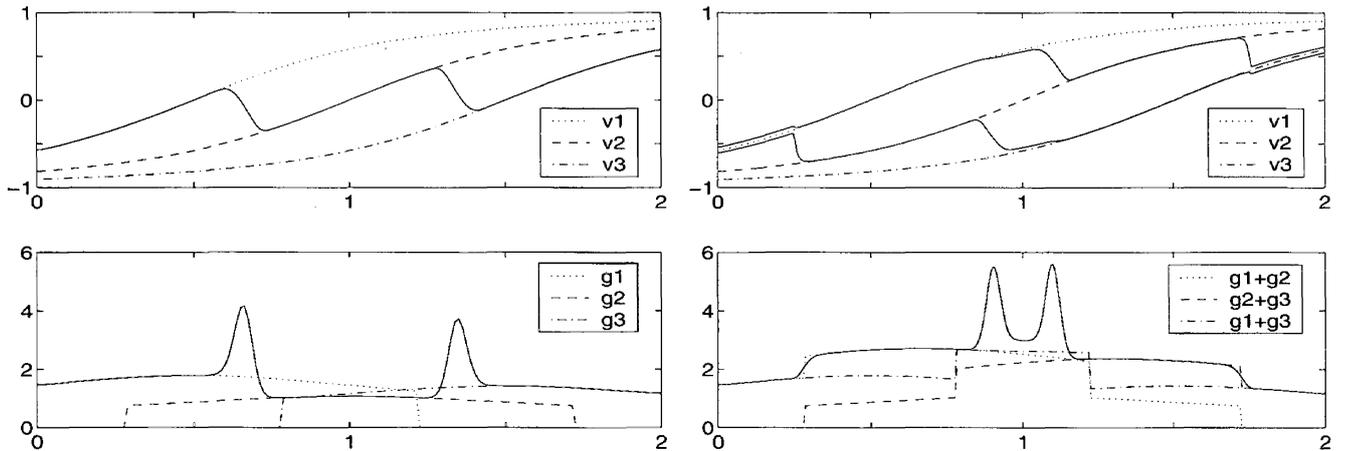
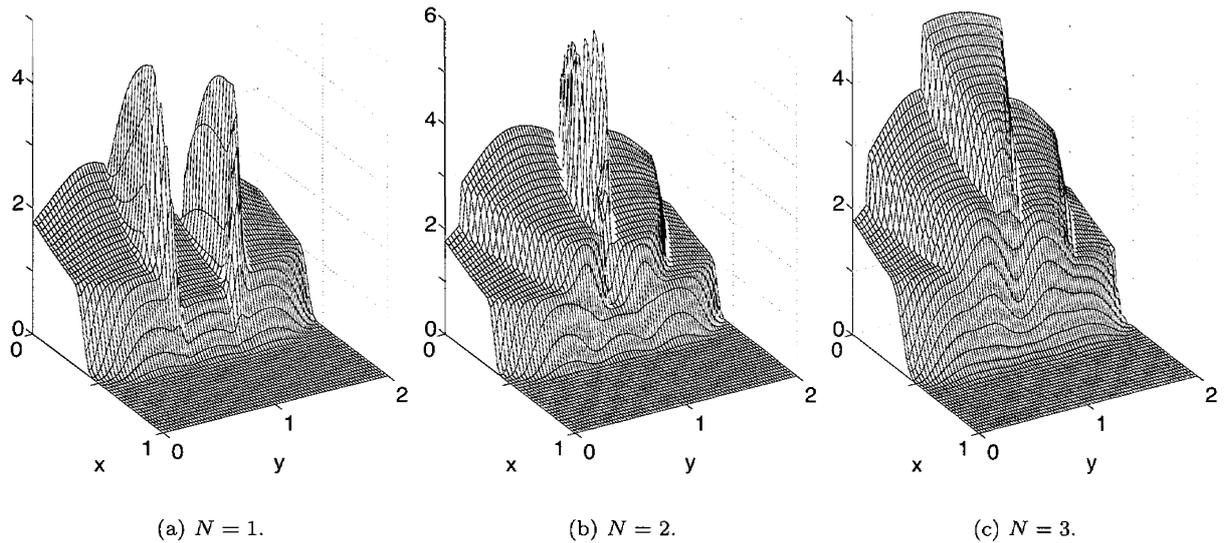
The exact ray traced solution is shown in Figure 3c. Although the rays concentrate along the caustic, there are at most two crossing rays at each spatial point.

The solution was computed in the square  $[0, 2] \times [0, 2]$  for the  $\delta$ -equations with  $N = 1, 2$  and the  $H$ -equations with  $N = 2$ . For the  $H$ -equations, initial data for  $\theta_1$  was as above, and  $\theta_2(0, y) = \theta_1(0, 2)$ . In Figure 3 the result is shown. At this level of resolution the  $\delta$ -equations with  $N = 2$  does not manage to capture the second phase just after the caustic. Away from the caustic, at  $x > 1.8$  it gives a better solution, though. The  $H$ -equations with  $N = 2$ , on the other hand, computes both phases accurately also close to the caustic.

For the  $H$ -equations we have made a convergence study. The solution is sampled before ( $x = 0.2$ ), inside ( $x = 0.55$ ) and just after ( $x = 1.0$ ) the caustic for different grid sizes. To measure the error against the exact solution, which can be multivalued, we let  $(u_{\text{ex}}(x, s), y(x, s))$  with  $s \in [0, 1]$  be the parameterized exact solution at  $x = n\Delta x$ . The discrete  $L^1$ -norm of the error, in the sense

$$\|\mathbf{U}\|_1 = \Delta y \sum_i \sum_{y(x,s)=i\Delta y} \min(|u_{\text{ex}}(x, s) - (\mathbf{U}_i^n)_1|, |u_{\text{ex}}(x, s) - (\mathbf{U}_i^n)_2|), \tag{45}$$

is tabulated in Table 2. We obtain full first order convergence rate before and inside the caustic. After the caustic the solution of the conservation law has a discontinuity at  $y = 1$ , cf. Figure 3f, and the convergence rate drops to one half, which is the expected rate for discontinuous solutions.



(d)  $N = 1$ , sine of ray angle,  $v_k = \sin \theta_k$ , (above), ray strength  $g$  (below).

(e)  $N = 2$ , sine of ray angles,  $v_k = \sin \theta_k$ , (above), total ray strength  $g_1 + g_2$  (below).

FIGURE 2. *Three point sources*. Solution of the  $\delta$ -equations with  $N = 1, 2, 3$ . Top pictures show total ray strength, *i.e.*  $g$ ,  $g_1 + g_2$  and  $g_1 + g_2 + g_3$  respectively. Bottom pictures show solution in a cut at  $x = 0.2$ , computed (solid) and exact (dotted, dashed, dashdotted).

4.1.3. *Focus*

We consider data at  $x = 0$  that generates a clean focus at coordinates  $(1, 1)$ ,

$$\tan \theta(0, y) = 1 - y, \quad g(0, y) = \frac{1}{1 + 4y^2}.$$

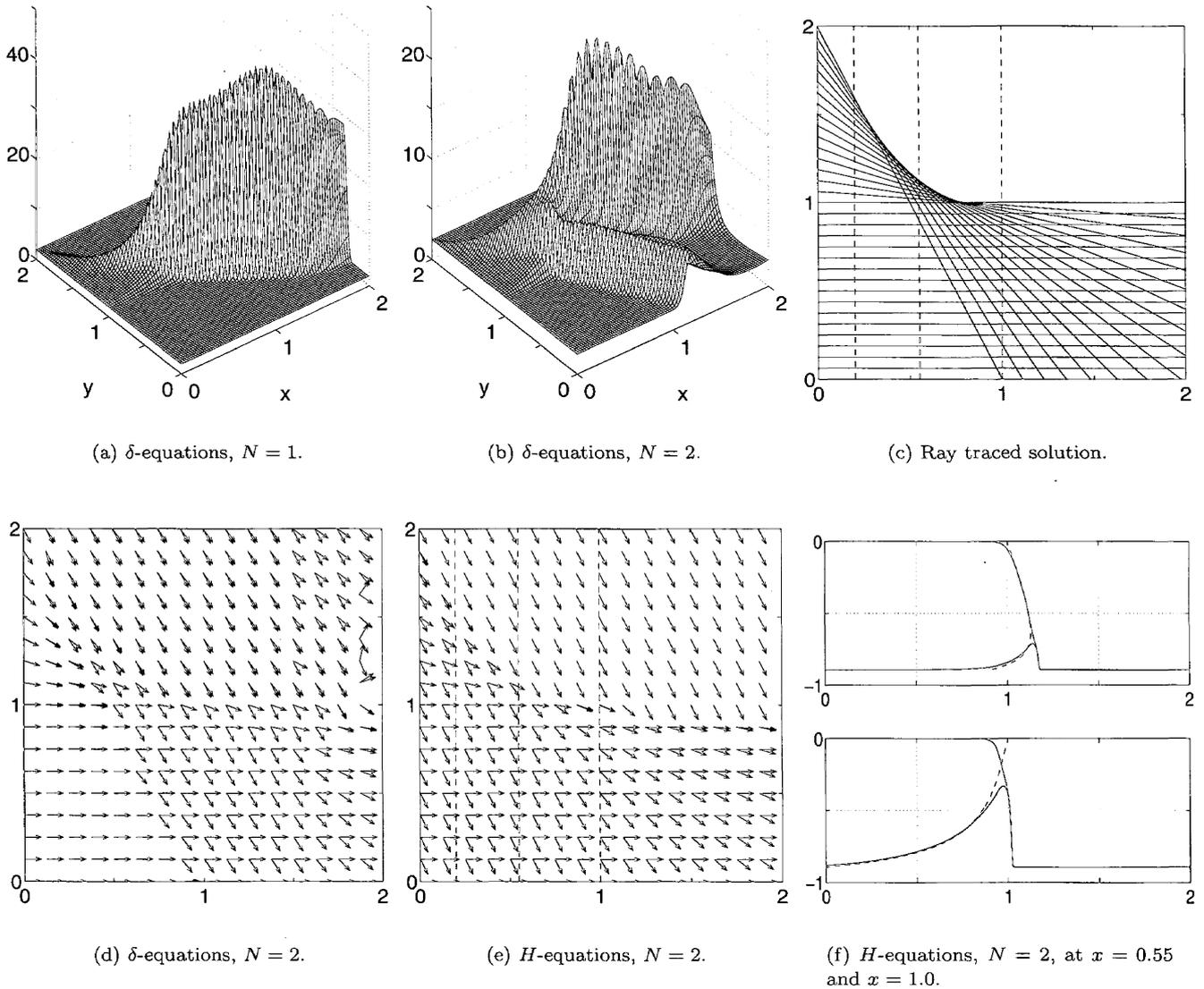


FIGURE 3. *Simple caustic*. Top left and middle figures show total ray strength, *i.e.*  $g$  and  $g_1 + g_2$  respectively, given by the  $\delta$ -equations with  $N = 1, 2$ . Top right figure shows a ray traced solution. Bottom left and middle figures contain quiver plots of ray angles for solution to  $\delta$ - and  $H$ -equations with  $N = 2$ . Bottom right figure shows sine of ray angles (solid) in cuts at  $x = 0.55$  (above) and  $x = 1.0$  (below) together with the corresponding values for a ray traced solution (dashed).

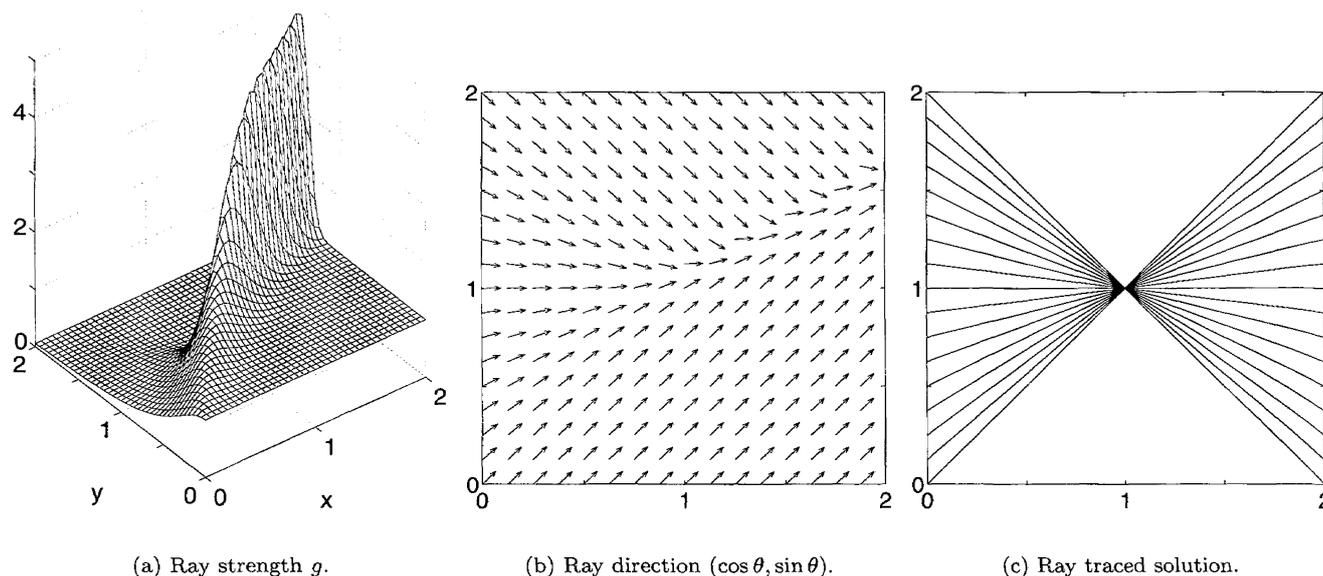


FIGURE 4. *Focus*. Solution of the  $\delta$ -equations with  $N = 1$ , together with a ray traced solution.

This is a difficult problem and none of the equations manages to solve it correctly. Although, almost everywhere, there is only one phase in the exact solution, the focus point formally contains an infinite number of phases.

We show the solution in the square  $[0, 2] \times [0, 2]$  for the  $\delta$ -equations with  $N = 1$  in Figure 4. The behavior of these equations can be explained by their relationship to the pressureless gas equations (22) mentioned in Section 2.2. On a discrete level (22) describes the so-called sticky particle dynamics. This signifies particles that move with constant speed until they collide. Colliding particles stick together and form a new particle with mass and velocity given by the conservation of mass and momentum. The focus point can be seen as a point where an infinite number of particles collide. The result is that all particles stick together, yielding the mass concentration, and move from the collision point along a line determined by their total momentum in the  $y$ -direction. The total momentum is given by the amplitude distribution of the injected wave, at  $x = 0$ . Incidentally, this shows that the conservation of mass and momentum produces a solution that in general does not agree with the viscosity solution for the corresponding eikonal equation, in which the line of discontinuity in the phase would have been  $y = 1$  regardless of the amplitude.

## 4.2. Inhomogeneous problems

### 4.2.1. Interface

This test problem uses an index of refraction that models a slightly tilted interface,

$$\eta(x, y) = \begin{cases} 1 & t \leq 0, \\ 1 + 4t^2(3 - 4t) & 0 < t \leq \frac{1}{2}, \\ 2 & \frac{1}{2} < t, \end{cases} \quad t = (x - 0.8) \cos \frac{\pi}{16} - y \sin \frac{\pi}{16}.$$

A number of plane waves with different angles are injected at  $x = 0$  by giving the exact solution on this boundary. The solution is computed in the square  $[0, 2] \times [0, 2]$ .

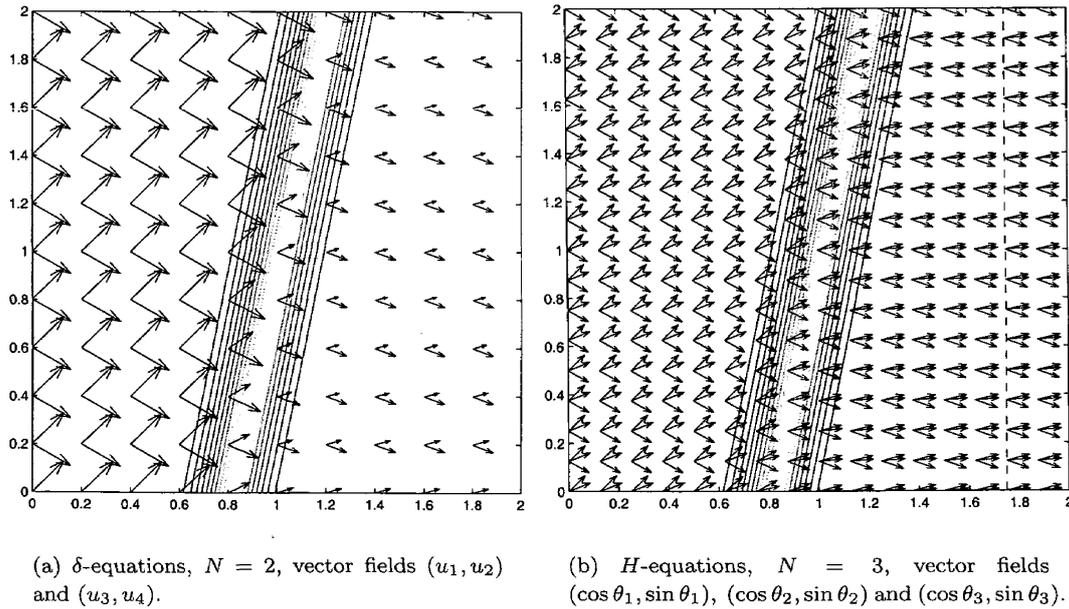


FIGURE 5. *Interface*. Result with two plane waves for the  $\delta$ -equations with  $N = 2$  and three plane waves for the  $H$ -system with  $N = 3$ . Contour lines of the index of refraction  $\eta$  is superimposed.

In the first test we used two waves with angles  $\theta_1 = \pi/4$ ,  $\theta_2 = -\pi/6$  and  $g_1(0, y) = g_2(0, y) = 1$ . The solution of the  $\delta$ -system with  $N = 2$  is shown in Figure 5a. Next, we add a third wave with angle  $\theta_3 = \pi/8$  and solve the  $H$ -system with  $N = 3$ , see Figure 5b.

The exact solutions of these problems are given by Snell's law,

$$\eta_\ell \sin \theta_\ell = \eta_r \sin \theta_r, \quad \eta_\ell g_\ell \cos \theta_\ell = \eta_r g_r \cos \theta_r.$$

The indices indicates the value to the left ( $\ell$ ) and right ( $r$ ) of the interface. We study how the computed solutions converge to this solution when we refine the mesh. In the delta case we sample the solution at time  $t = 5.0$  and  $x = 2$ . To avoid boundary effects we only look at the interval  $1.5 \leq y \leq 2$  for the first wave and  $0 \leq y \leq 0.5$  for the second. The errors in these intervals are given in Table 3. The tabulated errors are the sum of the discrete  $L^1$ -norms of the errors in the components  $u_k$ , and the total discrete  $L^\infty$ -norm of the component errors. The Heaviside case was treated the same way, only that the solution was sampled at  $x = 1.75$  and that we took the solution in the interval  $0.5 \leq y \leq 1.5$  for all three waves. The errors, computed as in the delta case, are also shown in Table 3. In both cases we obtain the expected first order convergence rate.

#### 4.2.2. Wedge

In this test problem a plane wave, injected at  $x = 0$  with  $\theta(0, y) = 0$  and  $g(0, y) = 2$ , is refracted by a smooth wedge, modeled by the index of refraction

$$\eta(x, y) = 1.5 - \frac{1}{\pi} \arctan(20((y - 1)^2 - 0.3(x - 0.5))).$$

When it is refracted in the interface a second and third phase appear. A caustic develops around the point  $(1, 1)$ , fanning out to the right, see Figure 6c. We have computed the solution in the square  $[0, 2] \times [0, 2]$

TABLE 3. *Interface*. The  $L^1$ - and  $L^\infty$ -norm of the errors for the  $\delta$ -system with  $N = 2$  and the  $H$ -system with  $N = 3$ . Error measured against the values predicted by Snell's law. Here  $\Delta y = 2/n$  and  $\Delta x = \Delta y$  in the delta case.

$n$	$\delta$ -equations, $N = 2$				$n$	$H$ -equations, $N = 3$			
	$L^1$		$L^\infty$			$L^1$		$L^\infty$	
	error	order	error	order		error	order	error	order
20	$5.28 \times 10^{-2}$		$5.07 \times 10^{-2}$		128	$1.95 \times 10^{-2}$		$9.66 \times 10^{-3}$	
		0.65		0.73			1.03		1.03
40	$3.37 \times 10^{-2}$		$3.05 \times 10^{-2}$		256	$9.52 \times 10^{-3}$		$4.72 \times 10^{-3}$	
		0.80		0.81			1.02		1.02
80	$1.94 \times 10^{-2}$		$1.73 \times 10^{-2}$		512	$4.71 \times 10^{-3}$		$2.33 \times 10^{-3}$	
		0.90		0.89			1.01		1.01
160	$1.04 \times 10^{-2}$		$9.38 \times 10^{-3}$		1024	$2.34 \times 10^{-3}$		$1.16 \times 10^{-3}$	
		0.94		0.94			1.00		1.00
320	$5.40 \times 10^{-3}$		$4.89 \times 10^{-3}$		2048	$1.17 \times 10^{-3}$		$5.78 \times 10^{-4}$	

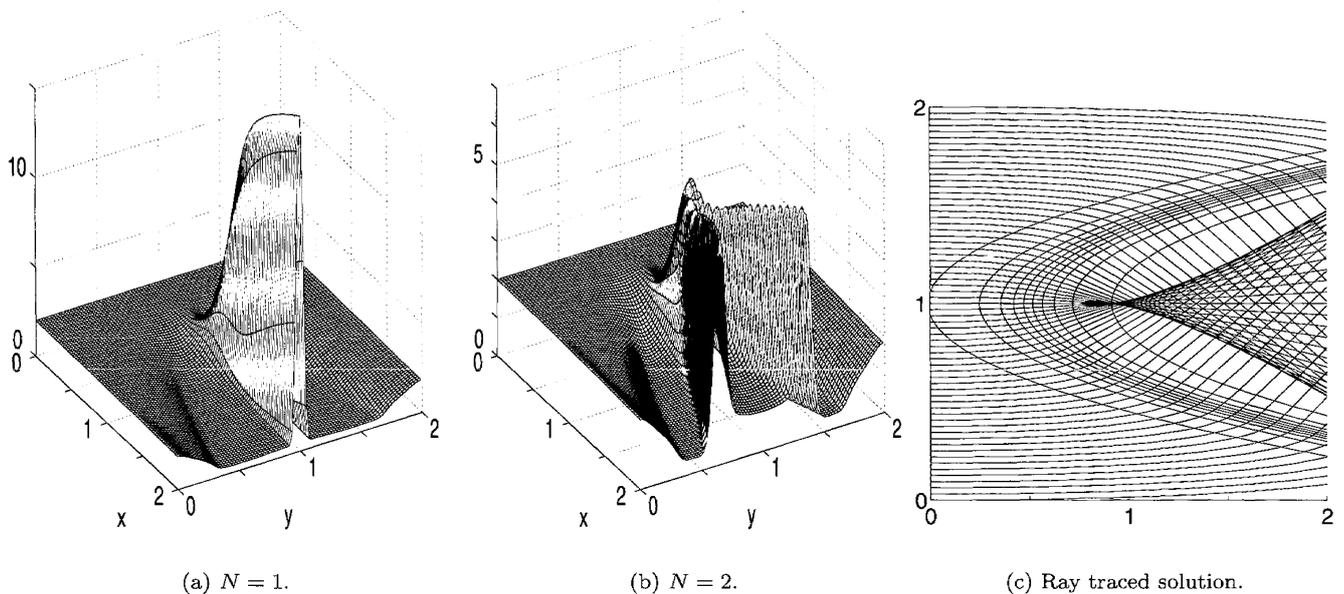


FIGURE 6. *Wedge*. Amplitude results for  $\delta$ -equations with  $N = 1, 2$ . Left and middle pictures show total ray strength, *i.e.*  $g$  and  $g_1 + g_2$  respectively. Right figure shows a ray traced solution with contour lines of the index of refraction superimposed.

for the  $\delta$ -equations with  $N = 1, 2$  and the  $H$ -equations with  $N = 2, 3$ . In the Heaviside case, initial data was  $\theta_3(0, y) = \frac{\pi}{4}H(y)$ ,  $\theta_1 = \theta_2 - \pi/4$  and  $\theta_2 = \frac{3}{4}(\theta_1 + \theta_2)$ . Different aspects of the solutions are shown in Figure 6 and Figure 7. The  $\delta$ -equations with  $N = 1$  only captures one of the phases as expected. The  $N = 2$  system captures quite well both the second phase and the caustic. The existence of a third phase does not have a markedly adverse effect on the solution, presumably because it carries little energy. In contrast, the  $H$ -equations, which do not include the amplitude, cannot correctly capture the second phase when  $N = 2$ . However, when  $N = 3$  all three phases are captured.

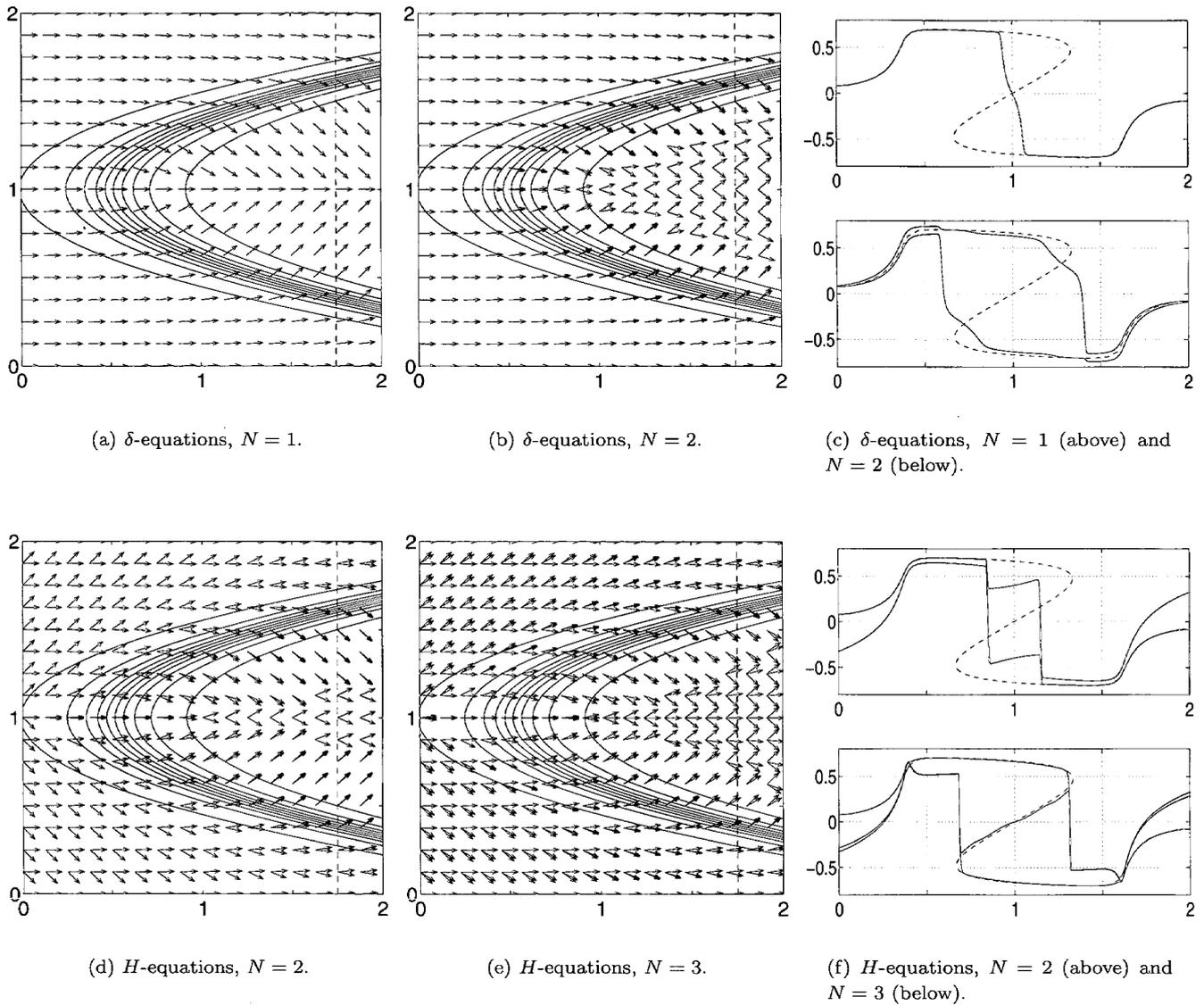


FIGURE 7. *Wedge*. Left and middle figures show quiver plots of ray angles for  $\delta$ - and  $H$ -equations with different  $N$ . A contour plot of the index of refraction is overlaid on the solution. Right figures show sine of ray angles (solid) in a cut at  $x = 1.75$  together with the corresponding values for a ray traced solution (dashed).

4.2.3. *Convex lens*

In this last test problem a plane wave is sent through a smooth convex lens, given by the index of refraction

$$\eta(x, y) = \begin{cases} 1 & d > 1, \\ \left(\frac{4}{3 - \cos(\pi d)}\right)^2 & d \leq 1, \end{cases} \quad d = \left(\frac{x - 0.5}{0.2}\right)^2 + \left(\frac{y - 1}{0.8}\right)^2.$$

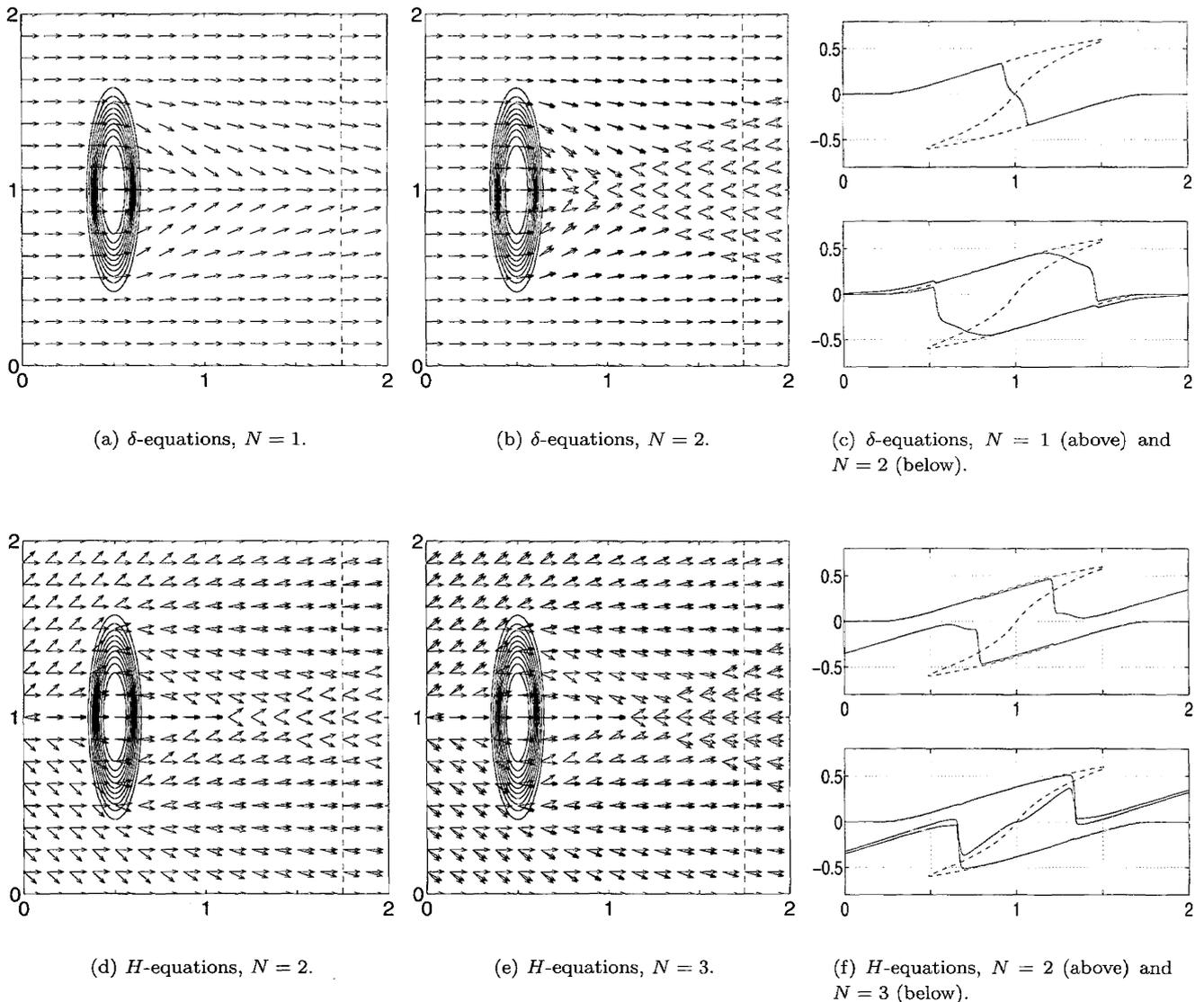


FIGURE 8. *Convex lens*. Left and middle figures show quiver plots of ray angles for  $\delta$ - and  $H$ -equations with different  $N$ . A contour plot of the index of refraction is overlaid on the solution. Right figures show sine of ray angles (solid) in a cut at  $x = 1.75$  together with the corresponding values for a ray traced solution (dashed).

This problem was taken from [13]. As in the previous problem the  $\delta$ - and  $H$ -equations were solved in the square  $[0, 2] \times [0, 2]$ . The same initial data as in the Wedge problem above was used. Figure 8 shows various features of the solutions. In this problem the exact solution develops up to five phases in the focus area around  $(1, 1)$  and settles with three phases behind this point. None of the equations manages to capture fully all three phases, although in general the higher  $N$  the better they perform.

APPENDIX A. FORM OF THE FLUX FUNCTIONS

A.1. Delta function case

In the most simple case,  $N = 1$ , the function  $\mathbf{F}_0$  is the identity and

$$\mathbf{F}_1 = \frac{u_1}{|\mathbf{u}|} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \mathbf{F}_2 = \frac{u_2}{|\mathbf{u}|} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \mathbf{K} = \frac{\eta_x u_2 - \eta_y u_1}{|\mathbf{u}|} \begin{pmatrix} u_2 \\ -u_1 \end{pmatrix}.$$

For  $N = 2$ , let  $\mathbf{w} = (w_1, w_2)^T$  and

$$\mathbf{f}_0 = \begin{pmatrix} w_1 \\ w_2 \\ w_1^3/|\mathbf{w}|^2 \\ w_2^3/|\mathbf{w}|^2 \end{pmatrix}, \quad \mathbf{f}_1 = \frac{w_1}{|\mathbf{w}|} \mathbf{f}_0, \quad \mathbf{f}_2 = \frac{w_2}{|\mathbf{w}|} \mathbf{f}_0, \quad \mathbf{k} = \frac{\eta_x w_2 - \eta_y w_1}{|\mathbf{w}|} \begin{pmatrix} w_2 \\ -w_1 \\ w_1^2 w_2 / |\mathbf{w}|^2 \\ -w_1 w_2^2 / |\mathbf{w}|^2 \end{pmatrix}.$$

Then  $\mathbf{F}_j = \mathbf{f}_j(u_1, u_2) + \mathbf{f}_j(u_3, u_4)$  for  $j = 0, 1, 2$  and  $\mathbf{K} = \mathbf{k}(u_1, u_2) + \mathbf{k}(u_3, u_4)$ .

A.2. Heaviside function case

For  $N = 1$ , the functions are simple,

$$\mathbf{F}_1(u_1) = -u_1, \quad \mathbf{F}_2(u_1) = \sqrt{1 - u_1^2}, \quad \mathbf{K} = 0.$$

For  $N = 2$ , let

$$\mathbf{f}_1 = \begin{pmatrix} w \\ \frac{1}{2}w^2 \end{pmatrix}, \quad \mathbf{f}_2 = \begin{pmatrix} -\sqrt{1 - w^2} \\ \frac{1}{2}(\arcsin(w) - w\sqrt{1 - w^2}) \end{pmatrix}, \quad \mathbf{k} = \begin{pmatrix} 0 \\ \frac{\eta_y}{2}(\arcsin(w) + w\sqrt{1 - w^2}) - \frac{1}{2}\eta_x w^2 \end{pmatrix}.$$

Then  $\mathbf{F}_j = -\mathbf{f}_j(u_1) + \mathbf{f}_j(u_2)$  for  $j = 1, 2$  and  $\mathbf{K} = -\mathbf{k}(u_1) + \mathbf{k}(u_2)$ . Finally, for  $N = 3$ , let

$$\mathbf{f}_1 = \begin{pmatrix} w \\ \frac{1}{2}w^2 \\ \frac{1}{3}w^3 \end{pmatrix}, \quad \mathbf{f}_2 = \begin{pmatrix} -\sqrt{1 - w^2} \\ \frac{1}{2}(\arcsin(w) - w\sqrt{1 - w^2}) \\ -\frac{1}{3}(2 + w^2)\sqrt{1 - w^2} \end{pmatrix}, \quad \mathbf{k} = \begin{pmatrix} 0 \\ \eta_y(\arcsin(w) + w\sqrt{1 - w^2}) - \frac{1}{2}\eta_x w^2 \\ -\frac{2}{3}\eta_y(1 - w^2)\sqrt{1 - w^2} - \frac{2}{3}\eta_x w^3 \end{pmatrix}.$$

Then  $\mathbf{F}_j = -\mathbf{f}_j(u_1) + \mathbf{f}_j(u_2) - \mathbf{f}_j(u_3)$  for  $j = 1, 2$  and  $\mathbf{K} = -\mathbf{k}(u_1) + \mathbf{k}(u_2) - \mathbf{k}(u_3)$ .

APPENDIX B. EVALUATING THE FLUX FUNCTIONS

B.1. Delta function case

We will here show how to solve the system of equations (44) analytically, when  $N = 2$ , for the choice of moments in (11). The nonlinear system reads

$$\begin{aligned} u_1 + u_3 &= m_{10}, & \frac{u_1^3}{u_1^2 + u_2^2} + \frac{u_3^3}{u_3^2 + u_4^2} &= m_{30}, \\ u_2 + u_4 &= m_{01}, & \frac{u_2^3}{u_1^2 + u_2^2} + \frac{u_4^3}{u_3^2 + u_4^2} &= m_{03}. \end{aligned} \tag{46}$$

We introduce the two new unknowns  $a = g_1 + g_2$  and  $\beta = \theta_1 + \theta_2$  and observe that

$$\begin{aligned} m_{30} &= m_{10} \cos^2 \alpha - R(a) \cos(\beta + \alpha)(1 + \cos(\beta - 2\alpha)), \\ m_{03} &= m_{01} \sin^2 \alpha + R(a) \sin(\beta + \alpha)(1 + \cos(\beta - 2\alpha)), \end{aligned} \tag{47}$$

where

$$R(a) = \frac{b}{2} \left( 1 - \frac{b^2}{a^2} \right), \quad b = \sqrt{m_{10}^2 + m_{01}^2}, \quad \tan \alpha = \frac{m_{01}}{m_{10}}.$$

From (47) we obtain  $\beta$  and  $a$ ,

$$\begin{aligned} \tan(\beta + \alpha) &= -\frac{m_{03} - m_{01} \sin^2 \alpha}{m_{30} - m_{10} \cos^2 \alpha}, \\ R(a)^2 (1 + \cos(\beta - 2\alpha))^2 &= (m_{30} - m_{10} \cos^2 \alpha)^2 + (m_{03} - m_{01} \sin^2 \alpha)^2. \end{aligned}$$

Next, we define the new unknown  $\phi$  by  $g_1 = (a + b \cos \phi)/2$  and use the relationship  $b^2 = g_1^2 + g_2^2 + 2g_1g_2 \cos \beta$ , obtained by squaring and summing the two leftmost equations in (46). After setting  $a^2 = b^2 + c^2$ , we arrive at an equation of the form

$$\left[ \frac{(a^2 + c^2)(m_{10}^2 - m_{01}^2)}{b^2} - b^2 \cos \beta \right] \cos(2\phi) - \frac{4acm_{10}m_{01}}{b^2} \sin(2\phi) = m_{10}^2 - m_{01}^2 - (a^2 + c^2) \cos \beta,$$

which can be solved exactly to get  $\phi$ . The solution is finally given by

$$\begin{aligned} u_1 &= \frac{1}{2}m_{10} + \frac{1}{2b}(am_{10} \cos \phi - cm_{01} \sin \phi), & u_3 &= m_{10} - u_1, \\ u_2 &= \frac{1}{2}m_{01} + \frac{1}{2b}(am_{01} \cos \phi + cm_{10} \sin \phi), & u_4 &= m_{01} - u_2. \end{aligned}$$

## B.2. Heaviside function case

We show here how to solve the equation  $\mathbf{F}_1(\mathbf{u}) = (m_{10}, \dots, m_{1,N-1})^T$  analytically when  $N = 1, 2, 3$ . For  $N = 1$  it is trivial,  $u_1 = -m_{10}$ . For  $N = 2$  the solution is

$$u_1 = \frac{m_{11}}{m_{10}} - \frac{m_{10}}{2}, \quad u_2 = \frac{m_{11}}{m_{10}} + \frac{m_{10}}{2}.$$

For  $N = 3$ , the solution is given as the roots of the polynomials

$$\begin{aligned} (6m_{10}^2 + 12m_{11})u^2 + (4m_{10}^3 - 12m_{12})u + m_{10}^4 + 12m_{11}^2 - 12m_{12}m_{10} &= 0, & (u_1, u_3) \\ 3(m_{10}^2 + 2m_{11})u - m_{10}^3 - 6m_{12} - 6m_{10}m_{11} &= 0. & (u_2) \end{aligned}$$

For  $N = 3, \dots, 6$  the solutions are given by solving two polynomial equations of degree  $N/2$ , when  $N$  is even, and of degree  $(N+1)/2$  and  $(N-1)/2$  when  $N$  is odd. The coefficients of the polynomials are rational functions of the moments  $(m_{10}, \dots, m_{1,N-1})$ . We conjecture this holds for all  $N$ .

*Acknowledgements.* The author would like to thank Professor Björn Engquist for many valuable discussions.

## REFERENCES

- [1] R. Abgrall and J.-D. Benamou, Big ray tracing and eikonal solver on unstructured grids: Application to the computation of a multivalued traveltime field in the Marmousi model. *Geophysics* **64** (1999) 230–239.
- [2] J.-D. Benamou, Big ray tracing: Multivalued travel time field computation using viscosity solutions of the eikonal equation. *J. Comput. Phys.* **128** (1996) 463–474.
- [3] J.-D. Benamou, Direct solution of multivalued phase space solutions for Hamilton-Jacobi equations. *Comm. Pure Appl. Math.* **52** (1999) 1443–1475.
- [4] J.-D. Benamou, F. Castella, T. Katsaounis and B. Perthame, High frequency limit of the Helmholtz equation. Research report DMA-99-25, Département de Mathématiques et Applications, École Normale Supérieure, Paris (1999).

- [5] F. Bouchut, On zero pressure gas dynamics, in *Advances in kinetic theory and computing, Ser. Adv. Math. Appl. Sci.* **22**, World Sci. Publishing, River Edge, NJ (1994) 171–190.
- [6] F. Bouchut and F. James, Équations de transport unidimensionnelles à coefficients discontinus. *C. R. Acad. Sci. Paris Sér. I Math.* **320** (1995) 1097–1102.
- [7] F. Bouchut and F. James, Duality solutions for pressureless gases, monotone scalar conservation laws and uniqueness. *Comm. Partial Differential Equations* **24** (1999) 2173–2189.
- [8] Y. Brenier and L. Corrias, A kinetic formulation for multibranch entropy solutions of scalar conservation laws. *Ann. Inst. H. Poincaré* **15** (1998) 169–190.
- [9] Y. Brenier and E. Grenier, Sticky particles and scalar conservation laws. *SIAM J. Numer. Anal.* **35** (1998) 2317–2328.
- [10] F. Castella, O. Runborg and B. Perthame, High frequency limit of the Helmholtz equation II: Source on a general smooth manifold. Research report, Département de Mathématiques et Applications, École Normale Supérieure, Paris (2000).
- [11] M. Crandall and P. Lions, Viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.* **277** (1983) 1–42.
- [12] 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. *Comm. Math. Phys.* **177** (1996) 349–380.
- [13] B. Engquist, E. Fatemi and S. Osher, Numerical solution of the high frequency asymptotic expansion for the scalar wave equation. *J. Comput. Phys.* **120** (1995) 145–155.
- [14] B. Engquist and O. Runborg, Multiphase computations in geometrical optics. *J. Comput. Appl. Math.* **74** (1996) 175–192.
- [15] B. Engquist and O. Runborg, Multiphase computations in geometrical optics, in *Hyperbolic Problems: Theory, Numerics, Applications*, M. Fey and R. Jeltsch Eds., *Internat. Ser. Numer. Math.* **129**, ETH Zentrum, Zürich, Switzerland (1998).
- [16] P. Gérard, P.A. Markowich, N.J. Mauser and F. Poupaud, Homogenization limits and Wigner transforms. *Comm. Pure Appl. Math.* **50** (1997) 323–379.
- [17] L. Gosse and F. James, Numerical approximations of one-dimensional linear conservation equations with discontinuous coefficients. *Math. Comp.* **69** (2000) 987–1015.
- [18] H. Grad, On the kinetic theory of rarefied gases. *Comm. Pure Appl. Math.* **2** (1949) 331–407.
- [19] E. Grenier, Existence globale pour le système des gaz sans pression. *C. R. Acad. Sci. Paris Sér. I Math.* **321** (1995) 171–174.
- [20] G.-S. Jiang and E. Tadmor, Nonoscillatory central schemes for multidimensional hyperbolic conservation laws. *SIAM J. Sci. Comput.* **19** (1998) 1892–1917.
- [21] J. Keller, Geometrical theory of diffraction. *J. Opt. Soc. Amer.* **52** (1962) 116–130.
- [22] R.G. Kouyoumjian and P.H. Pathak, A uniform theory of diffraction for an edge in a perfectly conducting surface. *Proc. IEEE* **62** (1974) 1448–1461.
- [23] Yu.A. Kravtsov, On a modification of the geometrical optics method. *Izv. Vyssh. Uchebn. Zaved. Radiofiz.* **7** (1964) 664–673.
- [24] R.J. LeVeque, *Numerical Methods for Conservation Laws*. Birkhäuser (1992).
- [25] C.D. Levermore, Moment closure hierarchies for kinetic theories. *J. Stat. Phys.* **83** (1996) 1021–1065.
- [26] P.-L. Lions and T. Paul, Sur les mesures de Wigner. *Rev. Mat. Iberoamericana* **9** (1993) 553–618.
- [27] D. Ludwig, Uniform asymptotic expansions at a caustic. *Comm. Pure Appl. Math.* **19** (1966) 215–250.
- [28] S. Osher and C.-W. Shu, High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.* **28** (1991) 907–922.
- [29] F. Poupaud and M. Rascle, Measure solutions to the linear multi-dimensional transport equation with non-smooth coefficients. *Comm. Partial Differential Equations* **22** (1997) 337–358.
- [30] O. Runborg, *Multiscale and Multiphase Methods for Wave Propagation*. Ph.D. thesis, Department of Numerical Analysis and Computing Science, KTH, Stockholm (1998).
- [31] W.W. Symes, A slowness matching finite difference method for traveltimes beyond transmission caustics. Preprint, Dept. of Computational and Applied Mathematics, Rice University (1996).
- [32] L. Tartar,  $H$ -measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations. *Proc. Roy. Soc. Edinburgh Sect. A* **115** (1990) 193–230.
- [33] J. van Trier and W.W. Symes, Upwind finite-difference calculation of traveltimes. *Geophysics* **56** (1991) 812–821.
- [34] J. Vidale, Finite-difference calculation of traveltimes. *Bull. Seismol. Soc. Amer.* **78** (1988) 2062–2076.
- [35] G.B. Whitham, *Linear and Nonlinear Waves*. John Wiley & Sons (1974).
- [36] Y. Zheng, Systems of conservation laws with incomplete sets of eigenvectors everywhere, in *Advances in Nonlinear Partial Differential Equations and Related Areas*, World Sci. Publishing, River Edge, NJ (1998) 399–426.