Acta Numerica (2003), pp. 181–266
DOI: 10.1017/S0962492902000119

Computational high frequency wave propagation

Björn Engquist
PACM, Department of Mathematics,
Princeton University,
Princeton, NJ 08544, USA
E-mail: engquist@math.princeton.edu

Olof Runborg
Department of Numerical Analysis
and Computer Science,
Royal Institute of Technology (KTH),
10044 Stockholm, Sweden
E-mail: olofr@nada.kth.se

Numerical simulation of high frequency acoustic, elastic or electro-magnetic wave propagation is important in many applications. Recently the traditional techniques of ray tracing based on geometrical optics have been augmented by numerical procedures based on partial differential equations. Direct simulations of solutions to the eikonal equation have been used in seismology, and lately approximations of the Liouville or Vlasov equation formulations of geometrical optics have generated impressive results. There are basically two techniques that follow from this latter approach: one is wave front methods and the other moment methods. We shall develop these methods in some detail after a brief review of more traditional algorithms for simulating high frequency wave propagation.

CONTENTS

1. Introduction

The numerical approximation of high frequency wave propagation is important in many applications. Examples are the simulation of seismic, acoustic and optical waves, and microwaves. When the *essential frequencies* in the wave field are relatively high, and thus the wavelengths are short compared to the overall size of the computational domain, direct simulation using the standard wave equations will be very costly, and approximate models for wave propagation must be used. Fortunately, there exist good approximations of many wave equations precisely for very high frequency solutions. Even for linear wave equations these approximations are often nonlinear. It is the goal of this paper to discuss numerical simulations based on such high frequency approximations.

We consider the linear scalar wave equation

$$u_{tt} - c(\boldsymbol{x})^2 \Delta u = 0, \qquad (t, \boldsymbol{x}) \in \mathbb{R}^+ \times \Omega, \qquad \Omega \subset \mathbb{R}^d, \tag{1.1}$$

where $c(\boldsymbol{x})$ is the local speed of wave propagation of the medium. We complement (1.1) with initial or boundary data that generate high frequency solutions. The exact form of the data will not be important here, but a typical example would be $u(t, \boldsymbol{x}) = A(t, \boldsymbol{x}) \exp(\mathrm{i}\omega(c(\boldsymbol{x})t - \boldsymbol{k} \cdot \boldsymbol{x}))$ at $t = 0$ and with $|\boldsymbol{k}|^2 = \sum_{j=1}^d k_j^2 = 1$ and the frequency $\omega \gg 1$. With $u(t, \boldsymbol{x}) = \exp(\mathrm{i}\omega t) v(\boldsymbol{x})$, the solution in frequency domain is given by the Helmholtz equation

$$\Delta v + \frac{\omega^2}{c(\boldsymbol{x})^2} v = 0, \qquad \boldsymbol{x} \in \Omega. \tag{1.2}$$

We shall continue with the time domain formulation in the Introduction and later come back to approximations in frequency domain.

In the direct numerical simulation of (1.1) the accuracy of the solution is determined by the number of grid points or elements per wavelength. The computational cost of maintaining constant accuracy grows algebraically with the frequency, and for sufficiently high frequencies a direct numerical simulation of (1.1) is no longer feasible. Numerical methods based on approximations of (1.1) are needed.

In this paper we consider variants of geometrical optics, which are asymptotic approximations obtained when the frequency tends to infinity. These approximations are widely used in applications such as computational electromagnetics, acoustics, optics and geophysics. Instead of the oscillating wave field u, the unknowns in standard geometrical optics equations are the phase ϕ and the amplitude A, neither of which depends on the parameter ω, and typically vary on a much coarser scale than u. Hence they should 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 (1974). Formally, the

equations follow if we assume a series expansion of the form

$$u(t,\boldsymbol{x}) = \mathrm{e}^{\mathrm{i}\omega\phi(t,\boldsymbol{x})} \sum_{k=0}^{\infty} A_k(t,\boldsymbol{x})(\mathrm{i}\omega)^{-k}. \tag{1.3}$$

Entering this expression into (1.1) and summing terms of the same order in ω to zero, we obtain separate equations for the unknown dependent variables in (1.3). The $\mathcal{O}(\omega^2)$ terms give the equation for the phase function ϕ. It satisfies the Hamilton–Jacobi-type *eikonal equation*

$$\phi_t + c(\boldsymbol{x})\,|\nabla\phi| = 0, \tag{1.4}$$

where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^d$, $|\boldsymbol{x}| = \left(\sum_{j=1}^d x_j^2\right)^{1/2}$, for $\boldsymbol{x} = (x_1,\ldots,x_d)^T \in \mathbb{R}^d$. For the $\mathcal{O}(\omega)$ terms we get the *transport equation* for A_0:

$$(A_0)_t + c(\boldsymbol{x})\frac{\nabla\phi\cdot\nabla A_0}{|\nabla\phi|} + \frac{c(\boldsymbol{x})^2\Delta\phi - \phi_{tt}}{2c(\boldsymbol{x})\,|\nabla\phi|}A_0 = 0. \tag{1.5}$$

For large ω we can discard the remaining terms in (1.3).

Some typical wave phenomena, such as diffraction, are lost in the infinite frequency approximation. Moreover, the approximation breaks down at caustics, where the amplitude A_0 is unbounded. For these situations, correction terms can be derived, such as those given by Keller (1962) in his *geometrical theory of diffraction* (GTD), further developed by Kouyoumjian and Pathak (1974), for instance. The geometry of Ω and boundary conditions are accounted for in GTD. A closer study of the solution's asymptotic behaviour close to caustics was made by Ludwig (1966) and Kravtsov (1964), among others. Generalizations of the series expansion (1.3), also valid at caustics and when the solution contains several crossing waves with different phase functions, were studied by Maslov (1965) and Duistermaat (1974), for example. For a rigorous treatment of propagation of singularities in linear partial differential equations, see Hörmander (1983–1985). We will briefly comment on some of these techniques in Section 2.5.

The traditional way to compute travel times of high frequency waves is through *ray tracing*. See Section 2 for a derivation of the ray equations (1.6) and the equations (1.7), (1.8) and (1.9). The travel time of a wave is given directly by the phase function ϕ, and ray tracing corresponds to solving the eikonal equation (1.4) through the method of characteristics, *i.e.*, solving the system of ordinary differential equations (ODEs)

$$\begin{gathered}\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \nabla_p H(\boldsymbol{x},\boldsymbol{p}), \qquad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla_x H(\boldsymbol{x},\boldsymbol{p}), \\ H(\boldsymbol{x},\boldsymbol{p}) = c(\boldsymbol{x})|\boldsymbol{p}|, \qquad \boldsymbol{x},\boldsymbol{p}\in\mathbb{R}^d,\end{gathered} \tag{1.6}$$

where the momentum variable $\boldsymbol{p}$ is usually called the 'slowness' vector, and

∇_p and ∇_x are the gradients taken with respect to $\boldsymbol{p}$ and $\boldsymbol{x}$, respectively, that is,

$$\nabla_p = \left(\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_d}\right)^T, \qquad \nabla_x = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d}\right)^T.$$

There are also ODEs for the amplitude. Suppose the source is a curve $\boldsymbol{x}_0(r)$ in $\mathbb{R}^2$ with $\phi(\boldsymbol{x}_0(r)) \equiv 0$. Let $(\boldsymbol{x}(t,r), \boldsymbol{p}(t,r))$ be the solution of (1.6) with $\boldsymbol{x}(0,r) = \boldsymbol{x}_0(r)$, and $\boldsymbol{p}(0,r) = \nabla\phi(\boldsymbol{x}_0(r))$. Then

$$A_0(\boldsymbol{x}(t,r)) = A_0(\boldsymbol{x}_0(r))\sqrt{\frac{|\boldsymbol{x}_{0r}(r)|c(\boldsymbol{x}(t,r))}{|\boldsymbol{x}_r(t,r)|c(\boldsymbol{x}_0(r))}}. \tag{1.7}$$

The vector $\boldsymbol{x}_r$ is obtained by solving the auxiliary ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\boldsymbol{x}_r\\ \boldsymbol{p}_r\end{pmatrix} = \begin{pmatrix} D^2_{px}H & D^2_{pp}H \\ -D^2_{xx}H & -D^2_{px}H\end{pmatrix}\begin{pmatrix}\boldsymbol{x}_r\\ \boldsymbol{p}_r\end{pmatrix}, \quad \begin{pmatrix}\boldsymbol{x}_r(0,r)\\ \boldsymbol{p}_r(0,r)\end{pmatrix} = \begin{pmatrix}\boldsymbol{x}_{0r}(r)\\ \partial_r\nabla\phi(\boldsymbol{x}_0(r))\end{pmatrix}. \tag{1.8}$$

The initial data in (1.8) represent the local shape of the ray's source, which is an additional piece of information needed to compute the amplitude along rays.

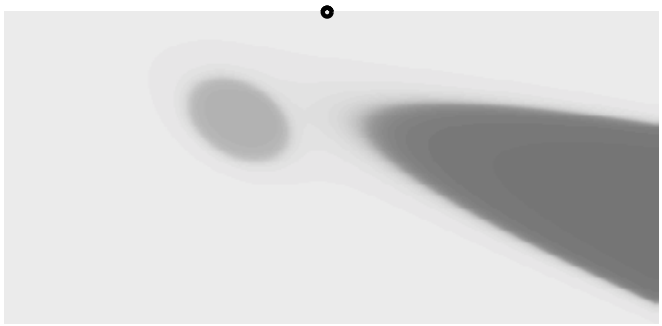
Finally, we can adopt a purely kinetic viewpoint, which will prove to be useful as a basis for some new numerical techniques. The kinetic model is based on the interpretation that rays are trajectories of particles following Hamiltonian dynamics. We introduce the phase space $(t, \boldsymbol{x}, \boldsymbol{p})$, where $\boldsymbol{p}$ is the slowness vector defined above. The evolution of a particle in this space is governed by (1.6). Letting $f(t, \boldsymbol{x}, \boldsymbol{p})$ be a particle density function, it will satisfy the Liouville equation

$$f_t + \nabla_p H \cdot \nabla_x f - \nabla_x H \cdot \nabla_p f = 0. \tag{1.9}$$

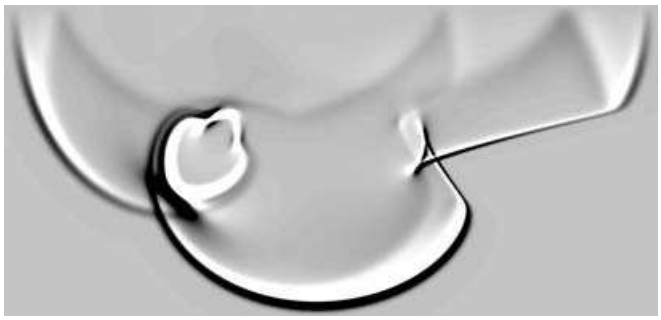
In Figure 1.1(b) we see a snapshot of a wave front propagating in a heterogeneous medium. The front can be accurately followed by using equation (1.9), Figure 1.1(c). The faint fronts in the upper part of these figures represent reflections and are not captured by (1.9) but they vanish in the limit as $\omega \to \infty$. Figure 1.2(c) shows that ray tracing may produce diverging rays that fail to cover the domain. With ray tracing it is also difficult to compute the amplitude and to find the minimum travel time in regions where rays cross.

Recently, new computational methods based on partial differential equations (PDEs) have been proposed to avoid some of the drawbacks of ray tracing. Interest was initially focused on solving the eikonal equation (1.4) numerically, and different types of upwind finite difference methods have been used to compute the viscosity solution of (1.4).

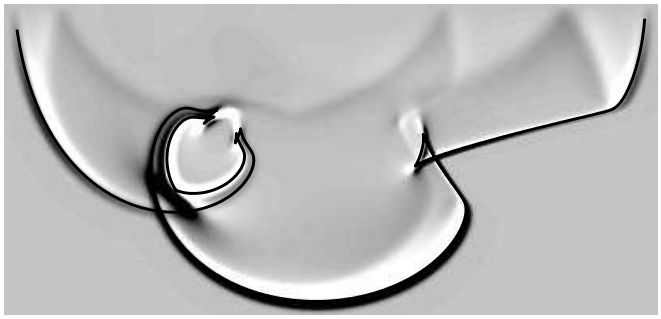
One problem with (1.4) is that it cannot produce solutions with multiple phases, corresponding to crossing rays. There is no super-position principle.



(a) Index of refraction and source point, marked by circle

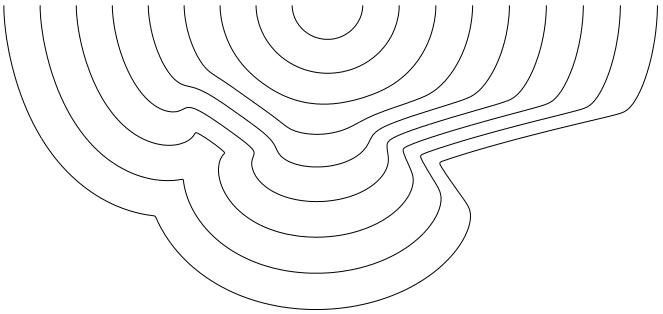


(b) Wave equation solution

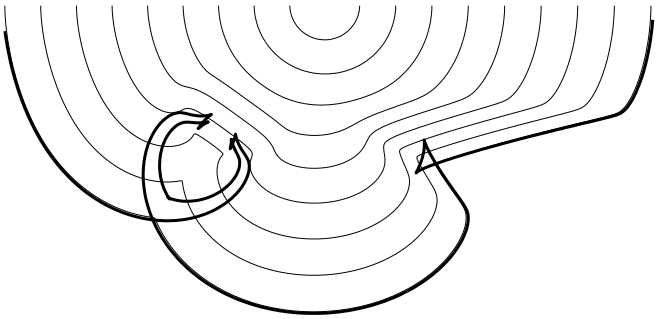


(c) Wave equation solution and wave front

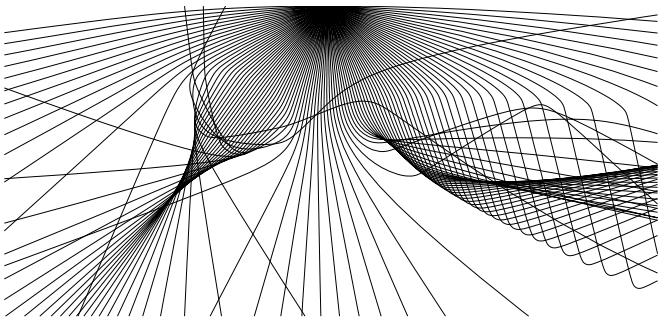
Figure 1.1. Comparison between different techniques for the same problem. A wave propagates from a point source through a hetereogeneous medium. (a) Source and index of refraction of the medium. Dark and light areas represent high and low index of refraction, respectively. (b) Snapshot of a resolved numerical solution of the wave equation, where the solution is represented by grey scale levels. (c) The same solution with a wave front construction solution overlaid.



(a) Eikonal equation solution



(b) Eikonal equation solution and wave front



(c) Ray-traced solution

Figure 1.2. Comparison between different techniques for the same problem. (a) Iso curves of a solution to the eikonal equation. (b) The same solution with a wave front construction solution overlaid. (c) A ray-traced 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 (Crandall and Lions 1983); see Figure 1.2(a). Hence, the eikonal equation only gives the first arrival travel time. A multivalued solution, however, can be constructed by patching together the solutions of several eikonal equations: see Section 3.2. See also Benamou (2003) for another survey of Eulerian methods for geometrical optics.

In this paper we shall mainly focus on numerical techniques based on the Liouville equation (1.9). It has the advantage of the linear superposition property of the ray equations and, like the eikonal equation, the solution is defined by a PDE and can easily be computed on a uniform Eulerian grid.

There is, however, a serious drawback with direct numerical approximation of the Liouville equation. It has the phase included in the set of independent variables, and straightforward simulation would be computationally very costly. We will discuss two ways to remedy this problem and focus the presentation on techniques that have not been surveyed before. In one, special wave front solutions are computed. The other is based on reducing the number of independent variables by introducing equations for moments.

Wave front methods are related to ray tracing. The evolution of a wave front is tracked in the physical or the phase space. We shall mainly follow the presentation in Engquist, Runborg and Tornberg (2002), in which the front evolution is defined directly by the Liouville equation and the tracking is done by the segment projection method. Level set and fast marching methods will also be discussed.

The moment method relies on the closure assumption that only a finite number of rays cross at each point in time and space. Then f in the Liouville equation (1.9) is of a special form, and it can be transformed into a finite system of equations representing the moments of f, set in the reduced space $(t, \boldsymbol{x})$.

Finally, let us mention that simulation of high frequency wave propagation is practically important in many applications. Classical optics will clearly require formulations other than the wave equation as the basis for computations. Visible light has wavelengths of the order of hundreds of nanometers. Thus, in numerical simulation, the number of unknowns would be prohibitively large for simulation over dimensions of metres.

A modern application of geometrical optics is computer visualization. The rendering of images is based on geometrical optics together with different *radiosity* boundary conditions. Usually the wave velocity is constant, resulting in straight rays, and the computational problem is mainly geometric.

For acoustic problems the computational domain is often smaller compared to the wavelength; the wave equation can be directly approached by a numerical method, and no geometrical optics is needed. High frequency

techniques become interesting, however, for very large distances, which, for instance, may occur in underwater acoustics.

High frequency approximation techniques also apply to other wave equations, and we shall give two examples of practical importance. The first is Maxwell's equations for electromagnetic waves in a lossless medium, that is,

$$\begin{aligned} \varepsilon(\boldsymbol{x})E_t &= \nabla \times H - J_e(\boldsymbol{x}), \\ \mu(\boldsymbol{x})H_t &= -\nabla \times E \\ \nabla \cdot (\varepsilon(\boldsymbol{x})E) &= \rho(\boldsymbol{x}), \\ \nabla \cdot (\mu(\boldsymbol{x})H) &= 0, \end{aligned} \tag{1.10}$$

where $E(t, \boldsymbol{x})$ and $H(t, \boldsymbol{x})$ are the electric and magnetic fields, respectively, ε and μ are the electric permittivity and magnetic permeability, respectively, J_e is the electric current density, and ρ is the electric charge density. Direct simulation based on (1.10) is common when the wavelength is not too short relative to the size of the computational domain. The geometrical theory of diffraction is the method of choice when the relative wavelength is very short. The latter is, for example, the case in the study of locations of base stations for cell phones in a city.

The other example is elastic wave propagation, given, for example, by

$$\rho(\boldsymbol{x})\boldsymbol{u}_{tt} = \nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}, \nabla \boldsymbol{u}), \tag{1.11}$$

where $\boldsymbol{u}(t, \boldsymbol{x})$ is the displacement vector, ρ is the density, and $\boldsymbol{\sigma}$ is the stress tensor. Seismic wave propagation is a challenging problem of this type. Both the forward and the inverse problems are of great interest, and may require geometrical optics-type approximations when the relative wavelength is short.

2. Mathematical background

In this chapter we derive the equations that are used in geometrical optics. We thus study the Cauchy problem for the scalar wave equation (1.1):

$$\begin{aligned} u_{tt}(\boldsymbol{x}, t) - c(\boldsymbol{x})^2 \Delta u(\boldsymbol{x}, t) &= 0, && \boldsymbol{x} \in \mathbb{R}^d,\ t > 0, \\ u(\boldsymbol{x}, 0) &= u_0(\boldsymbol{x}), && \boldsymbol{x} \in \mathbb{R}^d, \\ u_t(\boldsymbol{x}, 0) &= u_1(\boldsymbol{x}), && \boldsymbol{x} \in \mathbb{R}^d. \end{aligned} \tag{2.1}$$

Here $c(\boldsymbol{x})$ is the local wave velocity of the medium. We also define the *index of refraction* as $\eta(\boldsymbol{x}) = c_0/c(\boldsymbol{x})$ with the reference velocity c_0 (*e.g.*, the speed of light in a vacuum). For simplicity we will henceforth let $c_0 = 1$. When c is constant, equation (1.1) admits the simple plane wave solution

$$u(t, \boldsymbol{x}) = A\mathrm{e}^{\mathrm{i}\omega(ct - \boldsymbol{k}\cdot\boldsymbol{x})}, \qquad |\boldsymbol{k}| = 1, \tag{2.2}$$

where $\boldsymbol{k}$ is the wave vector giving the direction of propagation and A is a constant representing the amplitude. Both $\boldsymbol{k}$ and A are determined by appropriate initial data. For more complicated waves and when c is not constant, we need to replace $ct - \boldsymbol{k} \cdot \boldsymbol{x}$ by a general *phase function* ϕ, and also permit the amplitude to depend on time and space. Hence, (1.1) has solutions of the type

$$u(t, \boldsymbol{x}) = A(t, \boldsymbol{x}) \mathrm{e}^{\mathrm{i}\omega\phi(t,\boldsymbol{x})}. \tag{2.3}$$

The level curves of ϕ correspond to the wave fronts of a propagating wave: *cf.* Figure 2.1.

Since (1.1) is linear, the superposition principle is valid and a sum of solutions is itself a solution. The generic solution to (1.1) is, at least locally, described by a finite sum of terms like (2.3), with the amplitudes and phases being smooth functions that depend only mildly on the frequency ω. Typically this setting only breaks down at a small set of points, namely focus points, caustica and discontinuities in $c(\boldsymbol{x})$.

The solutions contain length and time scales that become very small as the frequency increases. In the direct numerical solution of (1.1) a substantial number of grid points per wavelength and dimension is needed to maintain constant accuracy. The work therefore grows algebraically with frequency. For sufficiently high frequencies or short wavelengths, it is unrealistic to compute the wave field directly. Fortunately, this is often the regime for which high frequency asymptotic approximations are quite accurate.

We will assume the geometrical optics approximation that $\omega \to \infty$. This means that, for the moment, we accept the loss of diffraction phenomena in the solution, and that the approximation of the wave amplitude breaks down at caustics. There are three strongly related formulations of geometrical optics, which we will review here. In Section 2.5 we consider some other approximations besides the pure geometrical optics.

2.1. Eikonal equations

Let us now derive Eulerian PDEs for the phase and the amplitude functions that are formally valid in the limit when $\omega \to \infty$. This is motivated by the observation that the phase and amplitudes of (2.3) generically vary on a much larger scale than the solution u itself, and should therefore be easier to compute. In the homogeneous case (2.2), for instance, $\phi = ct - \boldsymbol{k} \cdot \boldsymbol{x}$ stays nonoscillating and bounded independently of ω.

To begin with, we assume that the solution to (1.1) can be described by the asymptotic WKB expansion (Hörmander 1983–1985),

$$u = \mathrm{e}^{\mathrm{i}\omega\phi(t,\boldsymbol{x})} \sum_{k=0}^{\infty} A_k(t, \boldsymbol{x}) (\mathrm{i}\omega)^{-k}. \tag{2.4}$$

This form is a slight generalization of (2.3) that also includes a series expansion in powers of $1/\omega$ of the amplitude. We now substitute the expression (2.4) into (1.1) and, following the procedure outlined in the Introduction, equate coefficients of powers of ω to zero. For ω^2, this gives the *eikonal equation*,

$$\phi_t \pm c\,|\nabla\phi| = 0. \tag{2.5}$$

In fact, because of the sign ambiguity, we get two eikonal equations. Without loss of generality we will henceforth consider the one with a plus sign. For ω^1, we get the *transport equation* for the first amplitude term,

$$(A_0)_t + c\frac{\nabla\phi\cdot\nabla A_0}{|\nabla\phi|} + \frac{c^2\Delta\phi - \phi_{tt}}{2c\,|\nabla\phi|}A_0 = 0. \tag{2.6}$$

For higher-order terms of $1/\omega$, we get additional transport equations

$$(A_{k+1})_t + c\frac{\nabla\phi\cdot\nabla A_{k+1}}{|\nabla\phi|} + \frac{c^2\Delta\phi - \phi_{tt}}{2c\,|\nabla\phi|}A_{k+1} + \frac{c^2\Delta A_k - (A_k)_{tt}}{2c\,|\nabla\phi|} = 0 \tag{2.7}$$

for the remaining amplitude terms. When ω is large, only the first term in the expansion (2.4) is significant, and the problem is reduced to computing the phase ϕ and the first amplitude term A_0. Note that, once ϕ is known, the transport equations are linear equations with variable coefficients.

Instead of the time-dependent wave equation (1.1) we can consider the frequency domain problem. Setting $u(t,\boldsymbol{x}) = v(\boldsymbol{x})\exp(\mathrm{i}\omega t)$, with ω fixed, v satisfies the Helmholtz equation

$$c^2\Delta v + \omega^2 v = 0. \tag{2.8}$$

Substituting the series

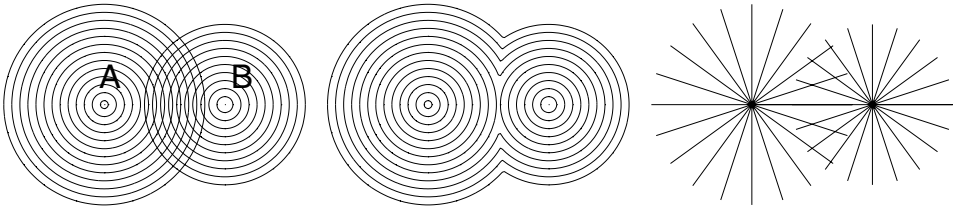
$$v = \mathrm{e}^{\mathrm{i}\omega\tilde{\phi}(\boldsymbol{x})}\sum_{k=0}^{\infty}\tilde{A}_k(\boldsymbol{x})(\mathrm{i}\omega)^{-k} \tag{2.9}$$

into (2.8), we get an alternative, frequency domain, version of the pair (2.5) and (2.6),

$$|\nabla\tilde{\phi}| = 1/c = \eta, \qquad 2\nabla\tilde{\phi}\cdot\nabla\tilde{A}_0 + \Delta\tilde{\phi}\tilde{A}_0 = 0. \tag{2.10}$$

With consistent initial and boundary data, $\phi(t,\boldsymbol{x}) = \tilde{\phi}(\boldsymbol{x}) - t$. We note that, since the family of curves $\{\boldsymbol{x} \mid \phi(t,\boldsymbol{x}) = \tilde{\phi}(\boldsymbol{x}) - t = 0\}$, parametrized by $t \geq 0$, describes a propagating wave front in (2.9), we often directly interpret the frequency domain phase $\tilde{\phi}(\boldsymbol{x})$ as the *travel time* of a wave; the difference in phase between two points on the same characteristic signifies the time it takes for a wave to travel between them.

We will drop the zero index in what follows and simply denote A_0 by A. We also drop the tilde for the frequency domain quantities.



(a) Correct solution (b) Eikonal equation (c) Ray tracing

Figure 2.1. Solution after some time to a homogeneous problem with two point sources, A and B, where the A source began transmitting slightly before the B source. Figure (a) shows the physically correct solution with two superimposed wave fields. Level curves of their phase functions are plotted. Figure (b) shows level curves of ϕ in the viscosity solution of the eikonal equation (2.5). Note that the superposition principle does not hold. Instead, the first arriving wave takes precedence over the second at each point. Figure (c) shows a ray-traced solution.

One problem with the eikonal and transport equations is that they do not accept solutions with multiple phases. There is no superposition principle for the nonlinear eikonal equation: *cf.* Figure 2.1. A finite sum of solutions of the form (2.3), with slowly varying A and ϕ, can in general not be well approximated by the first term in the ansatz (2.4) at high frequencies.

The eikonal equation is a nonlinear Hamilton–Jacobi-type equation with Hamiltonian $H(\boldsymbol{x}, \boldsymbol{p}) = c(\boldsymbol{x})|\boldsymbol{p}|$. As in the case of hyperbolic conservation laws, extra conditions are needed for this type of equation to have a unique solution. These were given in Crandall and Lions (1983) and the solution is known as the *viscosity solution*, which is the analogue of the entropy solution for conservation laws. As can be deduced from the previous paragraph, the viscosity solution does not have to agree with the correct physical solution in all cases. At points where the correct solution should have a multivalued phase, the viscosity solution picks out the phase corresponding to the first arriving wave.

It is well known that solutions of Hamilton–Jacobi equations can develop kinks, *i.e.*, discontinuities in the gradient, just as shocks appear in the solutions of conservation laws. In the case of the eikonal equation, the kinks are located where the physically correct phase solution should become multivalued: *cf.* Figure 2.1. We notice that the transport equation (2.6) has a factor involving $\Delta\phi$, which is not bounded at kinks, and therefore we can expect blow-up of A_0 at these points.

2.2. Ray equations

Another formulation of geometrical optics is *ray tracing*, which gives the solution via ODEs. This Lagrangian formulation is closely related to the method of characteristics for (2.5). Let $(\boldsymbol{x}(t), \boldsymbol{p}(t))$ be a bicharacteristic pair related to the Hamiltonian $H(\boldsymbol{x}, \boldsymbol{p}) = c(\boldsymbol{x})|\boldsymbol{p}|$, hence

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \nabla_p H(\boldsymbol{x}, \boldsymbol{p}) = c(\boldsymbol{x})\frac{\boldsymbol{p}}{|\boldsymbol{p}|}, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{2.11}$$

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla_x H(\boldsymbol{x}, \boldsymbol{p}) = -|\boldsymbol{p}|\nabla c(\boldsymbol{x}), \qquad \boldsymbol{p}(0) = \boldsymbol{p}_0. \tag{2.12}$$

In d dimensions the bicharacteristics are curves in $2d$-dimensional *phase space* $(\boldsymbol{x}, \boldsymbol{p}) \in \mathbb{R}^{d\times d}$. It follows immediately that H is constant along them, $H(\boldsymbol{x}(t), \boldsymbol{p}(t)) = H(\boldsymbol{x}_0, \boldsymbol{p}_0)$. We are interested in solutions for which $H \equiv 1$. In this case the projections on physical space, $\boldsymbol{x}(t)$, are usually called *rays*, and we can reduce (2.11) and (2.12) to

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \frac{1}{\eta^2}\boldsymbol{p}, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{2.13}$$

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \frac{\nabla\eta}{\eta}, \qquad \boldsymbol{p}(0) = \boldsymbol{p}_0, \quad |\boldsymbol{p}_0| = \eta(\boldsymbol{x}_0). \tag{2.14}$$

Solving (2.13) and (2.14) is called *ray tracing*. It should be noted here that if $\eta = \text{const}$ the rays are just straight lines.

We use the frequency domain version of the eikonal equation, (2.10), to explain the significance of the bicharacteristics when the solution ϕ is smooth. It can be written as

$$H(\boldsymbol{x}, \nabla\phi(\boldsymbol{x})) = 1, \tag{2.15}$$

with H as above. By differentiating (2.15) with respect to $\boldsymbol{x}$, we get

$$\nabla_x H(\boldsymbol{x}, \nabla\phi(\boldsymbol{x})) + D^2\phi(\boldsymbol{x})\nabla_p H(\boldsymbol{x}, \nabla\phi(\boldsymbol{x})) = 0.$$

Here D^2 represents the Hessian. Then for any curve $\boldsymbol{y}(t)$ we have the identity

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(\boldsymbol{y}(t)) &= D^2\phi(\boldsymbol{y}(t))\frac{\mathrm{d}\boldsymbol{y}(t)}{\mathrm{d}t}\\ &= D^2\phi(\boldsymbol{y}(t))\left[\frac{\mathrm{d}\boldsymbol{y}(t)}{\mathrm{d}t} - \nabla_p H(\boldsymbol{y}(t), \nabla\phi(\boldsymbol{y}(t)))\right]\\ &\quad - \nabla_x H(\boldsymbol{y}, \nabla\phi(\boldsymbol{y}(t))).\end{aligned}$$

Taking $\boldsymbol{x}(t)$ to be the curve for which the expression in brackets vanishes, we see that $(\boldsymbol{x}(t), \nabla\phi(\boldsymbol{x}(t)))$ is a bicharacteristic. By the uniqueness of solutions to (2.11), (2.12), we therefore have that $\boldsymbol{p}(t) \equiv \nabla\phi(\boldsymbol{x}(t))$ if we take $\boldsymbol{p}_0 = \nabla\phi(\boldsymbol{x}_0)$. Hence, with this initialization, the rays are therefore

always orthogonal to the level curves of ϕ, since $\mathrm{d}\boldsymbol{x}/\mathrm{d}t$ is parallel to $\boldsymbol{p} = \nabla\phi$ by (2.13). Moreover, for our particular H,

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(\boldsymbol{x}(t)) = \nabla\phi(\boldsymbol{x}(t)) \cdot \frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{p}(t) \cdot \nabla_p H(\boldsymbol{x}(t), \boldsymbol{p}(t)) = H(\boldsymbol{x}(t), \boldsymbol{p}(t)) = 1. \tag{2.16}$$

Thus, as long as ϕ is smooth, the solution to (2.15) along the ray is given by the simple expression

$$\phi(\boldsymbol{x}(t)) = \phi(\boldsymbol{x}_0) + t. \tag{2.17}$$

Since ϕ corresponds to travel time, this also shows that the parametrization t in (2.11) and (2.12) actually corresponds to unscaled time; the ray $\boldsymbol{x}(t)$ traces one point on a propagating wave front at time t. The absolute value of its time derivative $|\,\mathrm{d}\boldsymbol{x}/\mathrm{d}t|$ is precisely the local speed of propagation $c(\boldsymbol{x})$ by (2.13), and since $\boldsymbol{p}$ is parallel to $\mathrm{d}\boldsymbol{x}/\mathrm{d}t$, while $|\boldsymbol{p}| = H(\boldsymbol{x}, \boldsymbol{p})c(\boldsymbol{x})^{-1} = c(\boldsymbol{x})^{-1}$ by (2.15), the vector $\boldsymbol{p}$ is often called the *slowness* vector.

As was discussed in Section 2.1, the solution of the eikonal equation (2.5) is valid up to the point where discontinuities appear in the gradient of ϕ. This is where the phase should become multivalued but, by the construction, cannot. The bicharacteristics, however, do not have this problem, and we can extend their validity to all t: see Figure 2.1.

The ODEs for the bicharacteristics are sometimes solved using another parametrization than time. Setting $\mathrm{d}t = \eta(\boldsymbol{x}(t))^2\,\mathrm{d}\tau$, we get a simple ODE for $\boldsymbol{x}$,

$$\frac{\mathrm{d}^2\boldsymbol{x}}{\mathrm{d}\tau^2} = \frac{1}{2}\nabla\eta(\boldsymbol{x})^2. \tag{2.18}$$

This can still be interpreted as a Hamiltonian system, with a different H, but in this case an accompanying ODE must be solved to obtain the solution ϕ, *i.e.*, the travel time, along the ray,

$$H = \frac{|\boldsymbol{p}|^2 - \eta^2}{2}, \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}\phi(\boldsymbol{x}(\tau)) = \eta(\boldsymbol{x}(\tau))^2. \tag{2.19}$$

The rays can also be derived from the calculus of variations, using Fermat's principle. By analogy with the least action principle in classical mechanics, it says that the rays between two points are stationary curves of the functional

$$I[\gamma] = \int_\gamma \eta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

taken over all curves γ starting and ending at the points in question. The Euler–Lagrange equations for this optimization problem give the same bicharacteristics as (2.11) and (2.12), but the formulation is also well defined for non-differentiable η. The integral represents the length of γ under the measure $\eta\,\mathrm{d}s$ and therefore we often describe the rays as the *shortest optical path* between two points.

In order to compute the amplitude along a ray we also need information about the local shape of the ray's source. Let $(\boldsymbol{x}(t,\boldsymbol{x}_0),\boldsymbol{p}(t,\boldsymbol{x}_0))$ denote the bicharacteristic originating in $\boldsymbol{x}_0$ with $\boldsymbol{p}(0,\boldsymbol{x}_0)=\nabla\phi(\boldsymbol{x}_0)$, hence $\boldsymbol{x}(0,\boldsymbol{x}_0)=\boldsymbol{x}_0$. Let $J(t,\boldsymbol{x}_0)$ be the Jacobian of $\boldsymbol{x}$ with respect to initial data, $J=D_{x_0}\boldsymbol{x}(t,\boldsymbol{x}_0)$. By differentiating (2.13) we get

$$\begin{aligned}\frac{\partial J}{\partial t}=D_{x_0}\frac{\partial \boldsymbol{x}(t,\boldsymbol{x}_0)}{\partial t}&=D_{x_0}c(\boldsymbol{x}(t,\boldsymbol{x}_0))^2\boldsymbol{p}(t,\boldsymbol{x}_0)\\&=D_{x_0}c^2(\boldsymbol{x}(t,\boldsymbol{x}_0))\nabla\phi(\boldsymbol{x}(t,\boldsymbol{x}_0))\\&=\big(D_xc^2\nabla\phi\big)\,J.\end{aligned}$$

Assume that J is nonsingular and let $J=S\Lambda S^{-1}$ be a Jordan decomposition, so that the diagonal entries of Λ are the eigenvalues $\{\lambda_j\}$ of J. Setting $q=\det J\prod_j\lambda_j$, and using the fact that $\operatorname{tr}(T^{-1}AT)=\operatorname{tr}A$, we have

$$\begin{aligned}\frac{\partial q}{\partial t}&=q\,\operatorname{tr}\big(\Lambda^{-1}\Lambda_t\big)\\&=q\,\operatorname{tr}\big(S\Lambda^{-1}S^{-1}S\Lambda_tS^{-1}+(S\Lambda^{-1})S^{-1}S_t(S\Lambda^{-1})^{-1}+(S^{-1})_tS\big)\\&=q\,\operatorname{tr}\big(J^{-1}J_t\big)\\&=q\,\operatorname{tr}\big(J^{-1}\left(Dc^2\nabla\phi\right)J\big)\\&=q\,\operatorname{tr}\big(Dc^2\nabla\phi\big)=q\nabla\cdot c^2\nabla\phi.\end{aligned}$$

Therefore differentiation along the ray gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big[A^2(\boldsymbol{x}(t,\boldsymbol{x}_0))\eta(\boldsymbol{x}(t,\boldsymbol{x}_0))^2q(t,\boldsymbol{x}_0)\big]&=q(\nabla A^2\eta^2)\cdot\frac{\partial\boldsymbol{x}}{\partial t}+qA^2\eta^2\nabla\cdot c^2\nabla\phi\\&=q\nabla\cdot(A^2\nabla\phi)\\&=qA\big[2\nabla A\cdot\nabla\phi+\Delta\phi A\big]=0,\end{aligned}$$

using (2.10) in the last step. It follows that the amplitude is given by the expression

$$A(\boldsymbol{x}(t,\boldsymbol{x}_0))=A(\boldsymbol{x}_0)\frac{\eta(\boldsymbol{x}_0)}{\eta(\boldsymbol{x}(t,\boldsymbol{x}_0))}\sqrt{\left|\frac{q(0,\boldsymbol{x}_0)}{q(t,\boldsymbol{x}_0)}\right|}.\tag{2.20}$$

For example, an outgoing spherical wave in homogeneous medium with $\eta\equiv1$ is given by $\boldsymbol{x}(t)=\boldsymbol{x}_0+t\boldsymbol{x}_0/|\boldsymbol{x}_0|$. Then $J=I+t(I/|\boldsymbol{x}_0|-\boldsymbol{x}_0\boldsymbol{x}_0^T/|\boldsymbol{x}_0|^3)$ and $q=\det J=(1+t/|\boldsymbol{x}_0|)^{d-1}=(|\boldsymbol{x}|/|\boldsymbol{x}_0|)^{d-1}$ in d dimensions. Consequently, by (2.20), we get the well-known amplitude decay of such waves, $A\sim|\boldsymbol{x}|^{-(d-1)/2}$.

The determinant q is often called the *geometrical spreading*, since it measures the amplification of an infinitesimal area transported by the rays. It vanishes at caustics, and we see clearly from this expression that the amplitude is unbounded close to these points. (Strictly speaking we have only shown (2.20) as long as J is nonsingular, and then, by continuity, $q(0)$ and

$q(t)$ have the same sign. The expression is, however, also valid after caustic points, with the absolute values placed under the root sign.)

In order to compute A we thus need q, the determinant of $D_{x_0}\boldsymbol{x}$. The elements of this matrix are given by another ODE system. After differentiating (2.11) and (2.12) with respect to $\boldsymbol{x}_0$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} D_{x_0}\boldsymbol{x} \\ D_{x_0}\boldsymbol{p} \end{pmatrix} = \begin{pmatrix} D^2_{px}H & D^2_{pp}H \\ -D^2_{xx}H & -D^2_{px}H \end{pmatrix}\begin{pmatrix} D_{x_0}\boldsymbol{x} \\ D_{x_0}\boldsymbol{p} \end{pmatrix}, \tag{2.21}$$

with initial data

$$D_{x_0}\boldsymbol{x}(0,\boldsymbol{x}_0) = I, \qquad D_{x_0}\boldsymbol{p}(0,\boldsymbol{x}_0) = D^2\phi(\boldsymbol{x}_0).$$

We note that the system matrix here only depends on $\boldsymbol{x}$ and $\boldsymbol{p}$.

Since we have the constraint $H(\boldsymbol{x},\boldsymbol{p}) = 1$, or $|\boldsymbol{p}| = \eta(\boldsymbol{x})$, the dimension of the phase space $(\boldsymbol{x},\boldsymbol{p})$ can actually be reduced by one. We have not done this reduction in the equations above, and (2.13), (2.14), (2.20) and (2.21) are in this sense all overdetermined. We will here show the reduced equations in two dimensions.

Setting $\boldsymbol{p} = \eta(\cos\theta, \sin\theta)$, we can use θ as a dependent variable in (2.13) and (2.14) instead of $\boldsymbol{p}$. We then get, with $\boldsymbol{x} = (x, y)$,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = c(x,y)\cos\theta, \tag{2.22}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = c(x,y)\sin\theta, \tag{2.23}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\partial c}{\partial x}\sin\theta - \frac{\partial c}{\partial y}\cos\theta. \tag{2.24}$$

Suppose the source is a curve $\boldsymbol{x}_0(r)$ in $\mathbb{R}^2$ parametrized by r, and $\phi(\boldsymbol{x}_0(r)) \equiv 0$. Set $\tilde{\boldsymbol{x}}(t,r) := \boldsymbol{x}(t,\boldsymbol{x}_0(r))$ and $\tilde{\boldsymbol{p}}(t,r) := \boldsymbol{p}(r,\boldsymbol{x}_0(r))$. Then $\phi(\tilde{\boldsymbol{x}}(t,r)) = t$ by (2.17) and $\tilde{\boldsymbol{x}}_t \perp \tilde{\boldsymbol{x}}_r$ for all time, since

$$0 = \frac{\partial}{\partial r}\phi(\tilde{\boldsymbol{x}}(t,r)) = \nabla\phi(\tilde{\boldsymbol{x}})\cdot\tilde{\boldsymbol{x}}_r = \boldsymbol{p}\cdot\tilde{\boldsymbol{x}}_r = \eta^2\boldsymbol{x}_t\cdot\tilde{\boldsymbol{x}}_r.$$

We can then introduce the orthogonal matrix $R := [\tilde{\boldsymbol{x}}_r\, \tilde{\boldsymbol{x}}_t]$, with determinant $|\det R| = |\tilde{\boldsymbol{x}}_r||\tilde{\boldsymbol{x}}_t| = |\tilde{\boldsymbol{x}}_r|/\eta(\tilde{\boldsymbol{x}})$. By definition, for $0 \le s \le t$, we have $\boldsymbol{x}(t,\boldsymbol{x}_0) = \boldsymbol{x}(s,\boldsymbol{x}(t-s,\boldsymbol{x}_0))$, and, by differentiating both sides,

$$\boldsymbol{x}_t(t,\boldsymbol{x}_0) = D_{x_0}\boldsymbol{x}(s,\boldsymbol{x}(t-s,\boldsymbol{x}_0))\boldsymbol{x}_t(t-s,\boldsymbol{x}_0).$$

Evaluating at $s = t$ gives

$$\boldsymbol{x}_t(t,\boldsymbol{x}_0) = D_{x_0}\boldsymbol{x}(t,\boldsymbol{x}_0)\boldsymbol{x}_t(0,\boldsymbol{x}_0).$$

Therefore $D_{\boldsymbol{x}_0}\boldsymbol{x}(t,\boldsymbol{x}_0(r))R(0,r) = R(t,r)$ and

$$|q(t,\boldsymbol{x}_0(r))| = |\det D_{x_0}\boldsymbol{x}(t,\boldsymbol{x}_0(r))| = \frac{|\det R(t,r)|}{|\det R(0,r)|} = \frac{|\tilde{\boldsymbol{x}}_r(t,r)|\eta(\boldsymbol{x}_0(r))}{|\partial_r\boldsymbol{x}_0(r)|\eta(\tilde{\boldsymbol{x}}(t,r))},$$

so that

$$A(\boldsymbol{x}(t,r)) = A(\boldsymbol{x}_0(r))\sqrt{\frac{|\partial_r \boldsymbol{x}_0(r)|\eta(\boldsymbol{x}_0(r))}{|\tilde{\boldsymbol{x}}_r(t,r)|\eta(\tilde{\boldsymbol{x}}(t,r))}}. \tag{2.25}$$

We then only need to compute $\tilde{\boldsymbol{x}}_r$ to get the amplitude, which reduces (2.21) to

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\tilde{\boldsymbol{x}}_r\\ \tilde{\boldsymbol{p}}_r\end{pmatrix} = \begin{pmatrix} D^2_{px}H & D^2_{pp}H \\ -D^2_{xx}H & -D^2_{px}H\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{x}}_r\\ \tilde{\boldsymbol{p}}_r\end{pmatrix}. \tag{2.26}$$

2.3. Kinetic equations

Finally, we can adopt a purely kinetic viewpoint. This is based on the interpretation that rays are trajectories of particles following the Hamiltonian dynamics of (2.11) and (2.12). We introduce the phase space $(t, \boldsymbol{x}, \boldsymbol{p})$, where $\boldsymbol{p}$ is the slowness vector defined above in Section 2.2, and we let $f(t, \boldsymbol{x}, \boldsymbol{p})$ be a particle ('photon') density function. It will satisfy the Liouville equation,

$$f_t + \nabla_p H \cdot \nabla_x f - \nabla_x H \cdot \nabla_p f = 0, \tag{2.27}$$

or, with $H(\boldsymbol{x}, \boldsymbol{p}) = c(\boldsymbol{x})|\boldsymbol{p}|$,

$$f_t + \frac{c(\boldsymbol{x})}{|\boldsymbol{p}|}\boldsymbol{p}\cdot\nabla_x f + \frac{|\boldsymbol{p}|}{\eta^2}\nabla_x\eta\cdot\nabla_p f = 0. \tag{2.28}$$

We are only interested in solutions to (2.11) and (2.12) for which $H \equiv 1$, meaning that f only has support on the sphere $|\boldsymbol{p}| = \eta(\boldsymbol{x})$ in phase space. Because of this we can simplify (2.28) to the Vlasov-type equation

$$f_t + \frac{1}{\eta^2}\boldsymbol{p}\cdot\nabla_x f + \frac{1}{\eta}\nabla_x\eta\cdot\nabla_p f = 0, \tag{2.29}$$

with initial data $f_0(\boldsymbol{x}, \boldsymbol{p})$ vanishing whenever $|\boldsymbol{p}| \neq \eta$. We note that, if $\eta \equiv 1$, the equation (2.29) is just a free transport equation with solution $f(t, \boldsymbol{x}, \boldsymbol{p}) = f_0(\boldsymbol{x} - t\boldsymbol{p}, \boldsymbol{p})$ which corresponds to straight line ray solutions of (2.13) and (2.14).

The Wigner transform provides a direct link between the density function f in (2.29) and the solution to the scalar wave equation (1.1) and Helmholtz equation (1.2). It is an important tool in the study of high frequency, homogenization and random medium limits of these and many other equations, such as the Schrödinger equation (Lions and Paul 1993, Gérard, Markowich, Mauser and Poupaud 1997, Ryzhik, Papanicolaou and Keller 1996, Benamou, Castella, Katsaounis and Perthame 2002). The Wigner transform of $u^\varepsilon(\boldsymbol{x})$ is defined by

$$f^\varepsilon(t, \boldsymbol{x}, \boldsymbol{p}) := \int_{\mathbb{R}^d} \exp\bigl(-\mathrm{i}\boldsymbol{y}\cdot\boldsymbol{p}\bigr) u^\varepsilon(t, \boldsymbol{x} + \varepsilon\boldsymbol{y}/2)\overline{u^\varepsilon(t, \boldsymbol{x} - \varepsilon\boldsymbol{y}/2)}\,\mathrm{d}y.$$

If $\{u^\varepsilon\}$ is bounded in $L^2(\mathbb{R}^d)$ (for instance), then a subsequence of $\{f^\varepsilon\}$

converges weakly in $\mathcal{S}'(\mathbb{R}^d)$, the space of tempered distributions (Lions and Paul 1993). The limit is a locally bounded nonnegative measure, called the Wigner measure or semiclassical measure, which in our case agrees with the density function f in (2.29) above. An important property of the Wigner transform is that, when u^ε is a simple wave,

$$u^\varepsilon(t,\boldsymbol{x}) = A(t,\boldsymbol{x})\mathrm{e}^{\mathrm{i}\phi(t,\boldsymbol{x})/\varepsilon}, \tag{2.30}$$

then $f^\varepsilon \to f$ weakly in $\mathcal{S}'$, and the Wigner measure f represents a 'particle' in phase space of the form

$$f(t,\boldsymbol{x},\boldsymbol{p}) = A^2(t,\boldsymbol{x})\delta(\boldsymbol{p} - \nabla\phi(t,\boldsymbol{x})). \tag{2.31}$$

Even though f^ε is not linear in u^ε, a sum of simple wave solutions to (1.1) of the type (2.30) converges to a sum of 'particle' solutions to (2.29) of the type (2.31): see, *e.g.*, Jin and Li (200x) and Sparber, Mauser and Markowich (2003). Some other references dealing with the rigorous study of the convergence $f^\varepsilon \to f$, and proving that the limiting Wigner measure f satisfies a transport equation such as (2.29), are Castella, Perthame and Runborg (2002), Miller (2000) and Bal, Papanicolaou and Ryzhik (2002). We can also derive (2.29) directly from the wave equation (1.1) using so-called H-measures (Tartar 1990) or microlocal defect measures (Gérard 1991).

From (2.31) it follows that the amplitude at a point $\boldsymbol{x}$ is given as the integral of f over the phase variable,

$$A^2(t,\boldsymbol{x}) = \int_{\mathbb{R}^d} f(t,\boldsymbol{x},\boldsymbol{p})\,\mathrm{d}p.$$

Equation (2.29) can in fact be further reduced by drawing on the constraint $|\boldsymbol{p}| = \eta(\boldsymbol{x})$. Let us use polar coordinates for $\boldsymbol{p}$ in two dimensions, setting $\boldsymbol{p} = r(\cos\theta, \sin\theta)$. We then make the substitution

$$f(t,\boldsymbol{x},r,\theta) = \frac{1}{\eta(\boldsymbol{x})}\delta(r - \eta(\boldsymbol{x}))\tilde{f}(t,\boldsymbol{x},\theta),$$

and integrate (2.28) over all positive r. This gives a similar transport equation for $\tilde{f}$,

$$\tilde{f}_t + \frac{1}{\eta}\cos\theta\tilde{f}_x + \frac{1}{\eta}\sin\theta\tilde{f}_y + \frac{1}{\eta^2}(\eta_y\cos\theta - \eta_x\sin\theta)\tilde{f}_\theta = 0, \tag{2.32}$$

with $\boldsymbol{x} = (x,y)$. Also, with this scaling, that the integral over all phases gives the amplitude

$$\int_0^{2\pi}\tilde{f}(t,\boldsymbol{x},\theta)\,\mathrm{d}\theta = \int_0^{2\pi}\int_0^\infty f(t,\boldsymbol{x},r,\theta)r\,\mathrm{d}r\,\mathrm{d}\theta = \int_{\mathbb{R}^2} f(t,\boldsymbol{x},\boldsymbol{p})\,\mathrm{d}p = A^2(t,\boldsymbol{x}).$$

2.4. Boundary conditions

When a wave hits a sharp interface between two materials there will in general be one reflected and one transmitted wave. The interface is modelled by a rapid variation in the index of refraction η. For simplicity we assume that $\eta(\boldsymbol{x})$, with $\boldsymbol{x} = (x, y)$, only depends on x in two dimensions, and that $\eta = \eta_L$ to the left and $\eta = \eta_R$ to the right of the interface: *cf.* Figure 2.2. In the high frequency limit the solution depends on the limiting ratio between the width δ of the interface and the wavelength $\lambda = 2\pi c/\omega$ of the incident wave.

If $\lambda \ll \delta$ as $\lambda, \delta \to 0$, the geometrical optics equations are also valid at the interface, and if $\boldsymbol{p} = (p_x, p_y) = \eta(\cos\theta, \sin\theta)$ then

$$\frac{\mathrm{d}p_y}{\mathrm{d}t} = 0,$$

by (2.14), so that $\eta \sin\theta$ is constant along the ray. In the limit of a sharp interface this is Snell's law of refraction, usually written in the form

$$\eta_L \sin\theta_{\mathrm{inc}} = \eta_R \sin\theta_{\mathrm{tr}}. \tag{2.33}$$

Similarly, for a plane wave ($A_y = \phi_{yy} = 0$) hitting the interface, the transport equation (2.10) gives

$$(A^2\phi_x)_x = 0,$$

and since $\phi_x = \eta\cos\theta$, we get the corresponding law for the amplitudes of the incident and transmitted waves

$$\eta_L A_{\mathrm{inc}}^2 \cos\theta_{\mathrm{inc}} = \eta_R A_{\mathrm{tr}}^2 \cos\theta_{\mathrm{tr}}. \tag{2.34}$$

In this scaling limit there is no reflected wave.

The most common situation, however, is when $\lambda \gg \delta$ as $\lambda, \delta \to 0$. In this case boundary conditions must be derived directly from the wave equation before passing to the high frequency limit. They follow from assuming that the incident, reflected and transmitted waves have smooth phase functions.

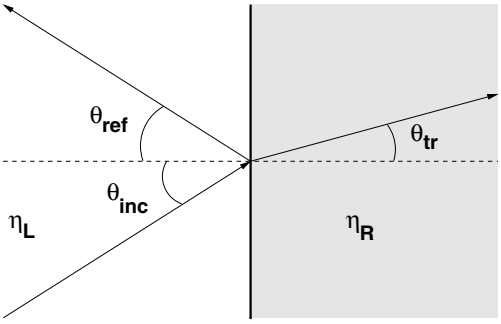


Figure 2.2. Reflection and transmission of a ray at a sharp interface when $\eta_L < \eta_R$.

Continuity of the solution across the interface gives Snell's law (2.33) and the reflection law

$$\theta_{\text{ref}} = \theta_{\text{inc}}.$$

There are also other interface conditions for the solution that depend on the type of wave equation, physics for the problem in question, and the local shape of the interface. Those give expressions for A_{ref} and A_{tr} in terms of A_{inc} as well as the corresponding quantities in (2.21) related to the geometrical spreading, often represented by the wave front's principal radii of curvature. In the case of the scalar wave equation (1.1), with a plane incident wave and a planar interface, continuity of the solution's normal derivative at the interface implies

$$A_{\text{ref}} = \frac{\eta_L \cos\theta_{\text{inc}} - \eta_R \cos\theta_{\text{tr}}}{\eta_L \cos\theta_{\text{inc}} + \eta_R \cos\theta_{\text{tr}}} A_{\text{inc}}, \qquad A_{\text{tr}} = \frac{2\eta_L \cos\theta_{\text{inc}}}{\eta_L \cos\theta_{\text{inc}} + \eta_R \cos\theta_{\text{tr}}} A_{\text{inc}}.$$

For the systems of wave equations (1.10) and (1.11), the interface and boundary conditions typically couple the amplitudes of different components, and one incident wave may generate several transmitted and reflected waves.

2.5. Other models

In this section we shall comment on a few different techniques that are related to our main focus on geometrical optics. These techniques handle high frequencies more efficiently than direct numerical approximation of the wave equation (1.1), and they include some phenomena that are not described by geometrical optics.

Paraxial approximations

Paraxial wave equation approximations are used to study waves propagating in a preferred direction. These approximations allow for numerical approximations of moderately higher frequencies than for regular wave equation methods even if the wave field is approximated directly without introducing phase or amplitude functions. A simple derivation follows from introducing a moving coordinate frame. Assume c to be constant and the waves propagating mainly in the positive x-direction,

$$\tilde{x} = x - ct, \qquad \frac{\partial}{\partial x} = \frac{\partial}{\partial \tilde{x}}, \tag{2.35}$$

$$\tilde{y} = y, \qquad \frac{\partial}{\partial y} = \frac{\partial}{\partial \tilde{y}}, \tag{2.36}$$

$$\tilde{t} = t, \qquad \frac{\partial}{\partial t} = \frac{\partial}{\partial \tilde{t}} - c\frac{\partial}{\partial \tilde{x}}. \tag{2.37}$$

This implies

$$u_{\tilde{t}\tilde{t}} - 2cu_{\tilde{x}\tilde{t}} = c^2 u_{\tilde{y}\tilde{y}},$$

and for waves moving essentially in the positive x-direction, $u_{\tilde{t}\tilde{t}}$ is small and set to zero. Dropping the tilde, the paraxial equation takes the form

$$u_{xt} = -\frac{c}{2} u_{yy}. \tag{2.38}$$

Equation (2.38) is well posed as an evolution equation both in the x- and the t-direction. The slow variation of u with respect to t reduces the computational complexity in numerical approximations.

Higher-order paraxial approximations can be derived from the dispersion relation of the wave equation or from the calculus of pseudo-differential operators. Paraxial approximations are, for example, used in underwater acoustics, in the inverse migration technique in seismology (Claerbout 1976) and as absorbing boundary conditions (Engquist and Majda 1977).

In geometrical optics a paraxial approximation often signifies another simplification, which can be made when there is one preferred coordinate direction. This means that all rays propagate in one direction and do not turn back. The slowness vector component in this direction is always positive. In those cases, time can be replaced by the preferred coordinate direction in the evolution equations, which reduces the dimension of the problem by one in the time-dependent case. Note that the expression is a misnomer in this case, since there is no approximation involved if the assumptions hold.

In two dimensions, $\boldsymbol{x} = (x, y)$, suppose that the x-axis can be used as evolution direction. Time is thus not explicitly needed in the calculation and θ, y and ϕ can be computed as a function of x directly. The phase ϕ (which is also the travel time) must be computed by a separate ODE. Dividing (2.23), (2.24) and (2.16) by (2.22), we get

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{pmatrix} y \\ \theta \end{pmatrix} = \begin{pmatrix} \tan\theta \\ \eta^{-1}(\eta_y - \eta_x \tan\theta) \end{pmatrix} =: \boldsymbol{u}(y, \theta), \tag{2.39}$$

$$\frac{\mathrm{d}\phi}{\mathrm{d}x} = \frac{\eta}{\cos\theta}. \tag{2.40}$$

These equations are valid as long as there are no turning rays, by which we mean that there is a constant C such that $|\theta| \le C < \pi/2$.

Let $(y(x, r), \theta(x, r))$ be the ray originating at $x = x_0$, $y = y_0(r)$ and $\theta = \theta_0(r)$, where r is some parametrization of the initial data. Then the amplitude is reduced to

$$A(x, y(x, r)) = A(x_0, y_0(r)) \sqrt{\frac{\eta(x_0, y_0(r)) |\partial_r y_0(r)|}{\eta(x, y(x, r)) |y_r(x, r)|}}. \tag{2.41}$$

Here y_r can be computed through the ODEs

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{pmatrix} y_r \\ \theta_r \end{pmatrix} = \begin{pmatrix} u_y & u_\theta \\ v_y & v_\theta \end{pmatrix}\begin{pmatrix} y_r \\ \theta_r \end{pmatrix}, \qquad \begin{pmatrix} y_r(0) \\ \theta_r(0) \end{pmatrix} = \frac{\mathrm{d}}{\mathrm{d}r}\begin{pmatrix} y_0(r) \\ \theta_0(r) \end{pmatrix}, \tag{2.42}$$

where $\boldsymbol{u} = (u, v)$ was defined above in (2.39) and $y_r(0, r) = \partial_r y_0(r)$, $\theta_r(0, r) = \partial_r \theta_0(r)$.

The eikonal equation (2.10) can be similarly reduced. In two dimensions we can evolve the equation in the x-direction, giving

$$\phi_x - \sqrt{\eta^2 - \phi_y^2} = 0,$$

which is now a one-dimensional evolution equation, valid as long as $|\phi_y| < \eta$. By the simple modification

$$\phi_x - \sqrt{\max(\eta^2 - \phi_y^2,\ \eta^2 \cos^2 \theta^*)} = 0, \tag{2.43}$$

with $\theta^* < \pi/2$, the equation is also well defined for problems with turning rays. This *paraxial eikonal equation* ignores rays with a propagation angle larger than θ^*, and its solution represents the first arrival time among the remaining rays (Gray and May 1994). See also Symes and Qian (2003) for a rigorous statement and proof of this.

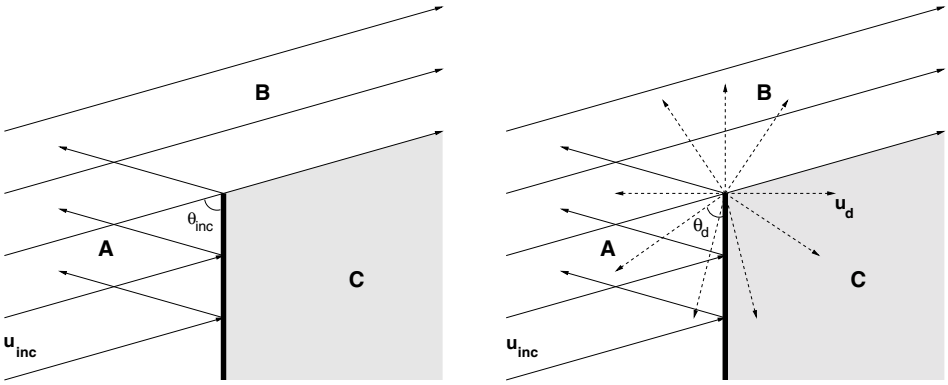
Geometrical theory of diffraction

The geometrical theory of diffraction (GTD) can be seen as a generalization of geometrical optics. It was pioneered by J. Keller in the 1960s (Keller 1962), and provides a systematic technique for adding diffraction effects to the geometrical optics approximation.

Standard geometrical optics excludes diffraction phenomena, which may be too crude an approximation for a scattering problem at moderate frequencies. The derivation of (2.5) and (2.6) in Section 2.1 does not take into account the effects of geometry and boundary conditions, which often gives rise to geometrical optics solution that are discontinuous: see Figure 2.3. In this case the series expansion (2.4) is not adequate. Extra terms must be added to the expansion to match the solution to the boundary conditions. One typical such expansion is

$$u = \mathrm{e}^{\mathrm{i}\omega\phi} \sum_{k=0}^{\infty} A_k (\mathrm{i}\omega)^{-k} + \mathrm{e}^{\mathrm{i}\omega\phi_\mathrm{d}} \sum_{k=0}^{\infty} B_k (\mathrm{i}\omega)^{-k-1/2}, \tag{2.44}$$

which is similar to the standard geometrical optics ansatz (2.4), only that a new diffracted wave scaled by $\sqrt{\omega}$ has been added (index d). For high frequencies, the term B_0 is also retained, together with A_0. The *local* geometry of the boundary determines the first B_k coefficients. More elaborate expansions must sometimes be used, such as those given by the *uniform theory of diffraction* (UTD) (Kouyoumjian and Pathak 1974).



(a) GO (b) GO + GTD

Figure 2.3. A typical geometrical optics solution in two dimensions and in a constant medium ($c \equiv 1$) around a perfectly reflecting halfplane (a), and the same problem augmented with diffracted waves given by GTD (b). In the geometrical optics case, region A contains two phases (incident and reflected), region B one phase (incident), and region C is in shadow, with no phases and hence a zero solution. On the boundaries between the regions the solution is discontinuous.

In general, diffracted rays are induced by rays that form discontinuities in the standard geometrical optics solution. Those rays produce an infinite set of diffracted rays that obey the usual geometrical optics equations. The main computational task, even for GTD, is thus based on the standard GO approximation, which is the central topic of this article. In Figure 2.3 the incident ray hitting the tip of the halfplane splits into a reflected ray that divides regions A and B, and another one that continues past the tip, dividing regions B and C. This ray gives rise to infinitely many diffracted rays shooting out in all directions from the tip of the wedge, which thus acts as an (anisotropic) point source.

The amplitude of each diffracted ray is proportional to the amplitude of the inducing ray and a diffraction coefficient D ($\sim B_0$). The coefficient D depends on the directions of the inducing and diffracted rays, on the frequency and on the local boundary geometry and index of refraction. In a two-dimensional homogeneous medium, the diffraction coefficient D for a halfplane is

$$D(\theta_{\rm d}, \theta_{\rm inc}, \omega) = \frac{{\rm e}^{{\rm i}\pi/4}}{2\sqrt{2\pi\omega}} \left(\frac{1}{\cos\frac{\theta_{\rm d}-\theta_{\rm inc}}{2}} \pm \frac{1}{\cos\frac{\theta_{\rm d}+\theta_{\rm inc}}{2}} \right), \tag{2.45}$$

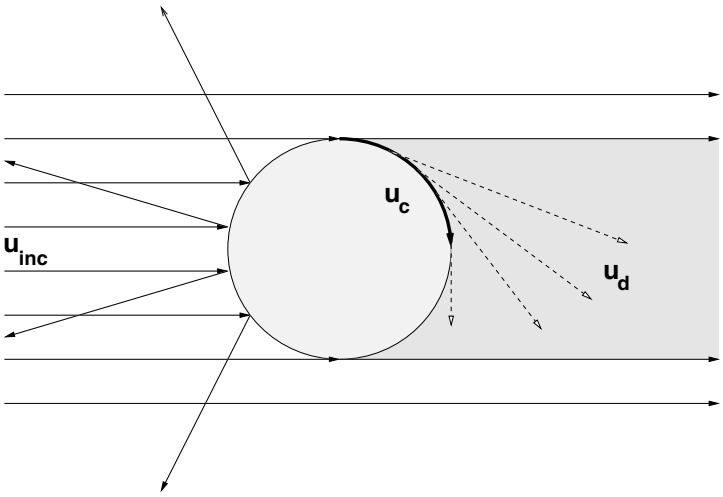


Figure 2.4. Diffraction by a smooth cylinder. The incident field u_{inc} induces a creeping ray u_{c} at the north (and south) pole of the cylinder. As the creeping ray propagates along the surface, it continuously emits surface-diffracted rays u_{d} with exponentially decreasing initial amplitude.

with the definition of the angles as in Figure 2.3(a) and Figure 2.3(b). The expression for the diffracted wave is then

$$u_{\mathrm{d}} = \frac{u_{\mathrm{inc}}}{\sqrt{r}} D(\theta_{\mathrm{d}}, \theta_{\mathrm{inc}}, \omega) \mathrm{e}^{-\mathrm{i}\omega r}, \tag{2.46}$$

where r is the distance to the tip of the halfplane.

It is important to note that diffraction coefficients only depend on the local geometry of the boundary. Relatively few types of coefficients are therefore sufficient for a systematic use of GTD. Diffraction coefficients have been computed for many different canonical geometries, such as wedges, slits and apertures, different wave equations, in particular Maxwell equations, and different materials and boundary conditions.

Another type of diffraction is generated even from smooth scatterers. When an incident field hits a smooth body such that some rays are tangent to the body surface, there will be a shadow zone behind it. The geometrical optics solution will again be discontinuous, and the curve (point in 2D) dividing the shadow part and the illuminated part of the body, will act as a source for surface rays, or *creeping rays*, that propagate along geodesics on the scatterer surface, if the surrounding medium is homogeneous, $\eta \equiv 1$. The creeping ray carries an amplitude proportional to the amplitude of the inducing ray. The amplitude decays exponentially along the creeping ray's trajectory. In three dimensions, the amplitude also changes through geometrical spreading on the surface. At each point on a convex surface, the

creeping ray emits surface-diffracted rays in the tangential direction, with its current amplitude. Those rays then follow the usual geometrical optics laws. See Figure 2.4 for an example. Other well-known surface waves are the Rayleigh waves in the elastic wave equation (1.11).

Physical optics

The physical optics (PO) method, also known as Kirchhoff's approximation, combines the geometrical optics (GO) solution with a boundary integral formulation of the solution to the Helmholtz equation. It is often used for scattering problems in, *e.g.*, computational electromagnetics. Let Ω be a perfectly reflecting scatterer in $\mathbb{R}^3$ and divide the solution into an incident and scattered part, $u = u_{\text{inc}} + u_{\text{s}}$. Then, in a homogeneous medium with $c \equiv 1$,

$$\Delta u_{\text{s}} + \omega^2 u_{\text{s}} = 0, \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \overline{\Omega}, \tag{2.47}$$

$$u_{\text{s}} = -u_{\text{inc}}, \qquad \boldsymbol{x} \in \partial\Omega, \tag{2.48}$$

together with an outgoing radiation condition. The solution outside Ω is given by the integral

$$u_{\text{s}}(\boldsymbol{x}) = -\oint_{\partial\Omega} u_{\text{inc}}(\boldsymbol{x}') \frac{\partial G(\boldsymbol{x}, \boldsymbol{x}')}{\partial n} + G(\boldsymbol{x}, \boldsymbol{x}') \frac{\partial u_{\text{s}}(\boldsymbol{x}')}{\partial n} \, \mathrm{d}x', \tag{2.49}$$

where G is the free space Green's function in three dimensions:

$$G(\boldsymbol{x}, \boldsymbol{x}') = \frac{\mathrm{e}^{\mathrm{i}\omega|\boldsymbol{x}-\boldsymbol{x}'|}}{4\pi|\boldsymbol{x} - \boldsymbol{x}'|}. \tag{2.50}$$

The unknown in this, exact, expression for the solution is $\partial u_{\text{s}}/\partial n$ on the boundary of Ω. In physical optics, this unknown is simply replaced by the geometrical optics solution. For example, if the incident field is a plane wave $u_{\text{inc}} = \exp(-\mathrm{i}\omega \boldsymbol{k} \cdot \boldsymbol{x})$, with $|\boldsymbol{k}| = 1$, then we would use $\partial u_{\text{s}}/\partial n = -\mathrm{i}\omega \boldsymbol{k} \cdot \hat{\boldsymbol{n}} u_{\text{inc}}$ for $\boldsymbol{x} \in \partial\Omega$, where $\hat{\boldsymbol{n}}$ is the normal of $\partial\Omega$ at $\boldsymbol{x}$. In the so-called *physical theory of diffraction* (PTD), the GTD extension of the geometrical optics solution is used.

Expression (2.49) gives a rigorous solution to Helmholtz in free space. The PO approximation is made at the boundary. PO can be regarded as a high frequency approximation in the sense that the accuracy increases with frequency. The computational cost is lower than the direct solution of the boundary integral formulation of the wave equation. Unlike GO, however, the cost is typically not frequency-independent, but grows drastically with frequency. Note also that PO is not self-consistent for finite frequencies. The resulting $\partial u_{\text{s}}(\boldsymbol{x} \in \partial\Omega)/\partial n$ is not equivalent to the applied GO solution. Iterative schemes to obtain this consistency can be used.

3. Overview of numerical methods

High frequency wave propagation is well approximated by asymptotic formulations like geometrical optics and the geometrical theory of diffraction. These formulations can be the basis of computations or they can be used analytically for the understanding of high frequency phenomena. In this section we shall describe different classes of computational techniques, based on the three different mathematical models for geometrical optics discussed in Section 2 above: see Figure 3.1.

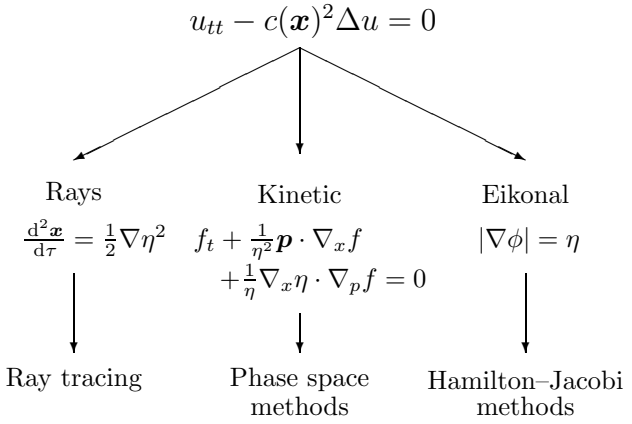


Figure 3.1. Mathematical models and numerical methods.

3.1. Ray tracing

The ray equations derived in Section 2.2 are the basis of ray tracing. The ray $\boldsymbol{x}(t)$ and slowness vector $\boldsymbol{p}(t) = \nabla\phi(\boldsymbol{x}(t))$ are governed by the ODE system (2.13) and (2.14). This system can be augmented by another ODE system for the amplitude, (2.21). Solving those ODEs is called ray tracing and it can be regarded as the method of characteristics applied to the eikonal equation. Some general references on ray tracing are Červený, Molotkov and Psencik (1977), Julian and Gubbins (1977), Langan, Lerche and Cutler (1985) and Thurber and Ellsworth (1980).

Ray tracing is typically not used to solve the complete Cauchy problem, with arbitrary initial and boundary data. Rather, the interest is to find the travel time of a wave from one source point to all points in a domain, or to a limited set of receiver points, together with the corresponding amplitudes in those points. The initial data are thus a single point source. In applications the same information is often needed for many source points, such as all points on a curve. The procedure is then repeated for each source point.

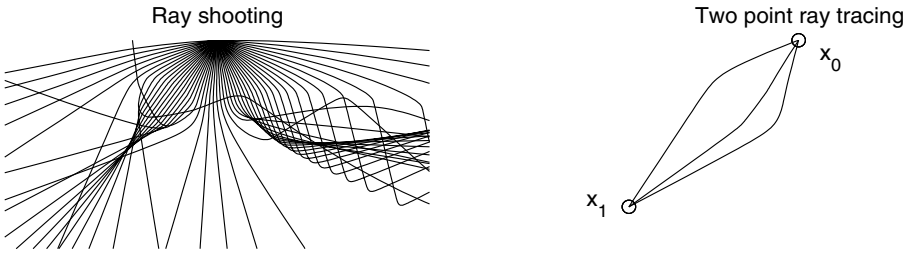


Figure 3.2. Ray shooting and two-point ray tracing. In this case there are three solutions to the two-point ray tracing problem.

Ray tracing gives the phase and the amplitude along the rays, and there is no *a priori* control of which points the ray passes through. One way of obtaining the solution at the particular points of interest is to use *ray shooting*: see Figure 3.2, left. A great many rays are shot from the source point in different directions. The result at the desired receiver points is interpolated from the solutions along the rays. This method is preferred when the travel time is sought for many receiver points, such as the grid points of a discretized domain. The ODEs are solved with standard numerical methods, for instance second- or fourth-order Runge–Kutta methods. The index of refraction is often only given on a grid, and it must be interpolated for the method to work. The interpolation can be smooth, such that the gradient of η in (2.14) exists everywhere, but simple piecewise constant or linear interpolation is also used. The rays are then straight lines or circular arcs within the grid cells, and they can be propagated exactly without an ODE solver. Snell's law of refraction is used at cell boundaries. Interpolating the ray solutions to a uniform grid from a large number of rays is difficult, in particular in shadow zones where few rays penetrate, and in regions where many families of rays cross: *cf.* Figure 3.2.

Another strategy to obtain the solution at a particular point is *two-point ray tracing*, also known as *ray bending*: see Figure 3.2, right. It is often used when there is only a limited number of receiver points. In this setting the ODEs are regarded as a nonlinear elliptic boundary value problem. From (2.18) we get

$$\begin{aligned} \frac{\mathrm{d}^2\boldsymbol{x}}{\mathrm{d}\tau^2} &= \frac{1}{2}\nabla\eta(\boldsymbol{x}(\tau))^2, \qquad (3.1)\\ \boldsymbol{x}(0) &= \boldsymbol{x}_0,\\ \boldsymbol{x}(\tau^*) &= \boldsymbol{x}_1, \end{aligned}$$

where $\boldsymbol{x}_0$ is the source point and $\boldsymbol{x}_1$ is the point of interest: see, *e.g.*, Pereyra, Lee and Keller (1980). Note that τ^*, the parameter value at the end point $\boldsymbol{x}_1$, is an additional unknown that must be determined together with the

solution. The equation (3.1) can be solved by a standard shooting method. It can also be discretized and turned into a nonlinear system of equations that can be solved with, for instance, variants of Newton's method. Initial data for the iterative solver can be difficult to find, in particular if there are multiple solutions (arrivals). Also, for two-point ray tracing the index of refraction must be interpolated.

In most problems in computational electromagnetics (CEM) the medium is piecewise homogeneous. This simplifies the calculations, since the solution of (2.13) and (2.14) is trivial given the solution at the boundaries and on the interfaces between media. Rays are straight lines satisfying the reflection law and Snell's law at interfaces. Ray tracing then reduces to the geometrical problem of finding points where rays are reflected and refracted. In the electromagnetic community, ray shooting is often referred to as *shooting and bouncing rays* (SBR), and two-point ray tracing as *ray tracing*: see, *e.g.*, (Ling, Chou and Lee 1989).

Note that the source and receiver points may be at infinity, corresponding to incident and scattered plane waves. For instance, a common problem in CEM is to compute the radar cross section (RCS) of an object. In this case both the source and receiver points are typically at infinity.

3.2. Hamilton–Jacobi methods

To avoid the problem of diverging rays, several PDE-based methods have been proposed for the eikonal and transport equations (2.5), (2.6) and (2.10). When the solution is sought in a domain, this is also computationally a more efficient and robust approach. The equations are solved directly, using numerical methods for PDEs, on a uniform Eulerian grid to control the resolution.

Viscosity solutions

The eikonal equation is a Hamilton–Jacobi-type equation and it has a unique viscosity solution which represents the first arrival travel time (Crandall and Lions 1983). This is also the solution to which monotone numerical finite difference schemes converge, and computing it was the starting point for a number of PDE-based methods. Vidale (1988) and van Trier and Symes (1991) used upwind methods to compute the viscosity solution of the frequency domain eikonal equation

$$|\nabla\phi| = \eta. \tag{3.2}$$

Upwind methods are stable, monotone methods that give good resolution of the kinks usually appearing in a viscosity solution. Importantly, the methods of Vidale (1988) and van Trier and Symes (1991) are explicit: computing the solution at a new grid point only involves previously computed

solutions at adjacent grid points. The methods make one sweep over the computational domain, finding the solution at one grid point after another, following an imagined expanding 'grid wave front', propagating out from the source: see Figure 3.3. To ensure causality and to obtain the correct viscosity solution from an explicit scheme, the grid points must be updated in a certain order. Those early methods used a grid wave front with fixed shape (rectangular in Vidale (1988) and circular in van Trier and Symes (1991)) and fail in this respect when there are rays in the exact solution that run parallel to the grid wave front, much in the same way as a paraxial approximation fails when there are turning rays. To avoid failure, the grid wave front could systematically be advanced from the grid point that has the smallest current solution value (minimum travel time). This ensures causality and guarantees a correct result, which was recognized by Qin, Luo, Olsen, Cai and Schuster (1992). The method presented by Qin *et al.* (1992) included simple sorting of the points on the grid wave front according to solution value. The sorting was improved in Cao and Greenhalgh (1994), where an efficient heap sort algorithm was proposed to maintain the right ordering of the points on the grid wave front, as it is advanced. The method of Cao and Greenhalgh (1994) bears a close resemblance to the *fast marching method* (Tsitsiklis 1995, Sethian 1996, Sethian 1999). This is an upwind-based method for efficient evaluation of distances or generalized distance functions such as the phase ϕ in (3.2). It also uses a heap sort algorithm allowing for computationally efficient choices of marching directions. Those methods can be seen as versions of Dijkstra's algorithm

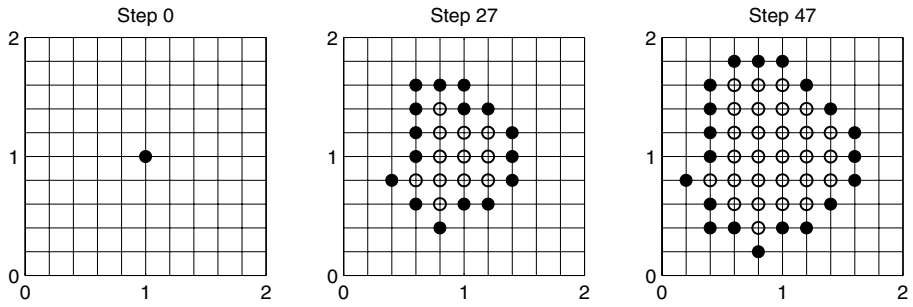


Figure 3.3. An explicit solver for the frequency domain eikonal equation (3.2). Starting from one source point (left), the grid points are updated one at a time in a certain order (middle, right). The outermost points (filled circles) constitute the grid wave front, which propagates outwards from the source, leaving behind it points where the solution is already established (circles). Note that the grid wave front is not necessarily close to an actual wave front.

for finding the shortest path in a network, adapted to a grid-based setting. The overall computational complexity for solving a problem with N grid points is $\mathcal{O}(N \log N)$. Another recent fast method is the *group marching method* (Kim 2000), whose complexity is merely $\mathcal{O}(N)$.

In parallel, high-resolution methods of ENO and WENO type, which had for some time been used in the numerical analysis of nonlinear conservation laws, were adapted to Hamilton–Jacobi equations (Osher and Shu 1991). Those methods were used for the time-dependent eikonal equation (2.5) in Fatemi, Engquist and Osher (1995). Constructing higher-order schemes for methods that use an expanding grid wave front is difficult if the shape of the front changes, as in the fast marching method. For paraxial approximations and methods with fixed-shape grid fronts, the high-resolution methods can be applied directly to obtain higher-order schemes. *Post sweeping* is a technique for avoiding the failures that are associated with turning rays in these methods. The problem at hand is solved in several 'sweeps', using different preferred directions. For each sweep, at each grid point, the smallest of the new and the previously computed solution value is selected (Schneider, Ranzinger, Balch and Kruse 1992, Kim and Cook 1999, Tsai, Cheng, Osher and Zhao 2003).

Multivalued solutions

The eikonal and transport equations only describe one unique wave (phase) at a time. There is no superposition principle in the nonlinear eikonal equation. At points where the correct solution should have a multivalued phase, the viscosity solution picks out the phase corresponding to the first arriving wave. When later arriving waves are also of interest, the viscosity solution is not enough. In inverse seismic problems, for instance, it is recognized that first arrival travel times are often not sufficient to give a good migration image (Geoltrain and Brac 1993). This is a particular problem in complicated inhomogeneous media, where caustics that generate new phases appear in the interior of the computational domain for any type of source. The problem is related to the fact that the first arrival wave is not always the most energetic one (*cf.* the example in Figures 1.1 and 1.2).

One way to obtain more than the first arrival solution is to geometrically decompose the computational domain, and solve the the eikonal solution, with appropriate boundary conditions, in each of the subdomains. The viscosity solutions thus obtained can be pieced together to reconstruct a larger part of the full multibranch solution.

A simple decomposition strategy can be based on detecting kinks in the viscosity solution. The kinks appear where two different branches of the full solution meet: *cf.* Section 2.1. Fatemi *et al.* (1995) made an attempt to compute multivalued travel times with this approach. A second phase, corresponding to the second arrival time, was calculated using two separate

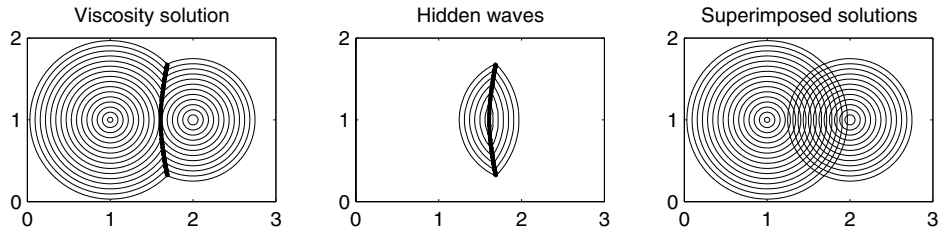


Figure 3.4. Geometrical decomposition by detecting kinks. Bold lines indicate location of the (first) viscosity solution kink. The middle figure shows second viscosity solution where the first solution (left) was applied as boundary condition at the kink.

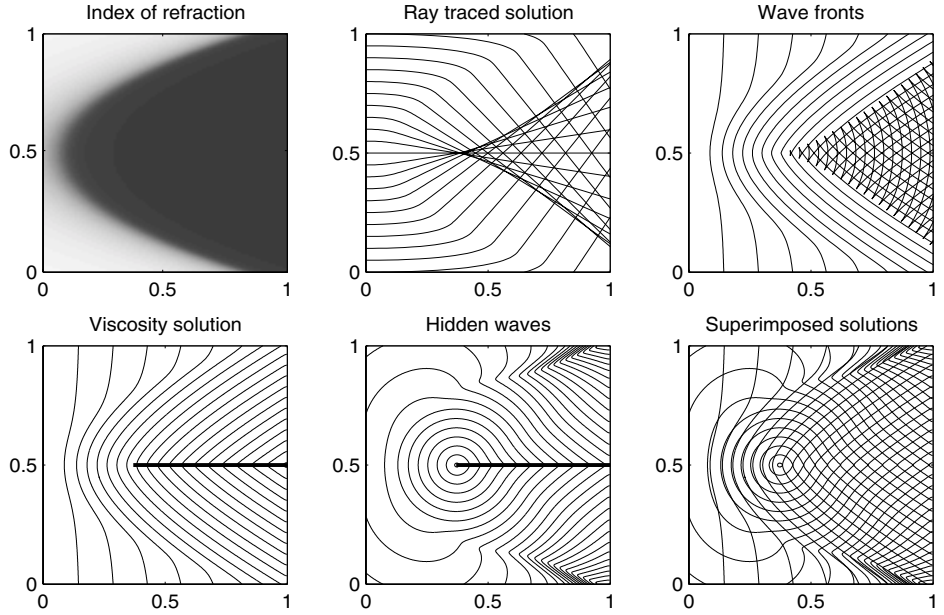


Figure 3.5. Geometrical decomposition by detecting kinks in a problem with a caustic. The top row shows the index of refraction and exact solution. The bottom row shows computed solutions. Bold lines indicate location of the (first) viscosity solution kink.

viscosity solutions of the eikonal equation, with boundary conditions for the second phase given at the location of the kink that had appeared in the first viscosity solution: see Figure 3.4. The same technique was also used at geometric reflecting boundaries. In principle, the same procedure could be repeated, using kinks in the second solution as boundary data for a third phase, and so on.

It is difficult, however, to find a robust way of detecting a kink and to distinguish it from rapid, but smooth, gradient shifts at strong refractions. For more complicated problems, such as the one shown in Figure 3.5, there are difficulties even if the kink could be detected perfectly. In this example, there is no obvious way to find boundary data for the third phase using the singularities in the second viscosity solution (bottom row, middle figure). Moreover, only the part of the second solution lying to the right of the caustic curve that develops (see the ray-traced solution), corresponds to a physical wave. The rest of the solution should be disregarded, including the kinks near the top and bottom right corners.

Another, more *ad hoc*, way of dividing the domain is used in the *big ray tracing* method: see Figure 3.6. It was introduced by Benamou (1996), and extended for use with unstructured grids by Abgrall and Benamou (1999). A limited number of rays are shot from the source point in different directions. The domains bounded by two successive rays are the 'big rays'. In each

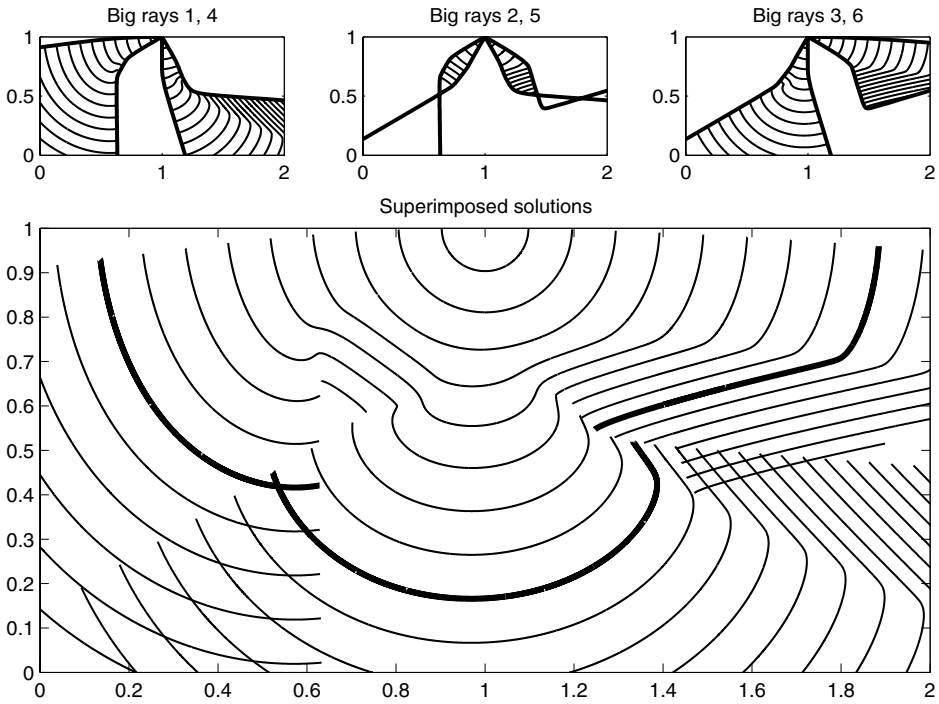


Figure 3.6. Big ray tracing using six big rays for the problem in Figure 1.1. The top row shows viscosity solutions in each ray. The bottom figure shows solutions superimposed; the bold line corresponds to points on the same wave front that was indicated in Figure 1.2(b).

big ray, the viscosity solution is computed. Since the big rays may overlap (*e.g.*, big rays 1 and 3 in Figure 3.6), multivalued solutions can be obtained, although in general the method will not capture all phases. In the presence of caustics the basic method is not so reliable, and it needs to be modified. Then there is, for instance, no guarantee that it includes the viscosity solution among its branches: *cf.* the example in Figure 3.6.

Benamou (1999) introduced a more natural decomposition of the computational domain, which ensures that all phases in the multibranch solution are captured. In his method, the domain is cut along caustic curves. The caustics are detected by solving an accompanying PDE that enables a continuous monitoring of the geometrical spreading. The geometrical spreading

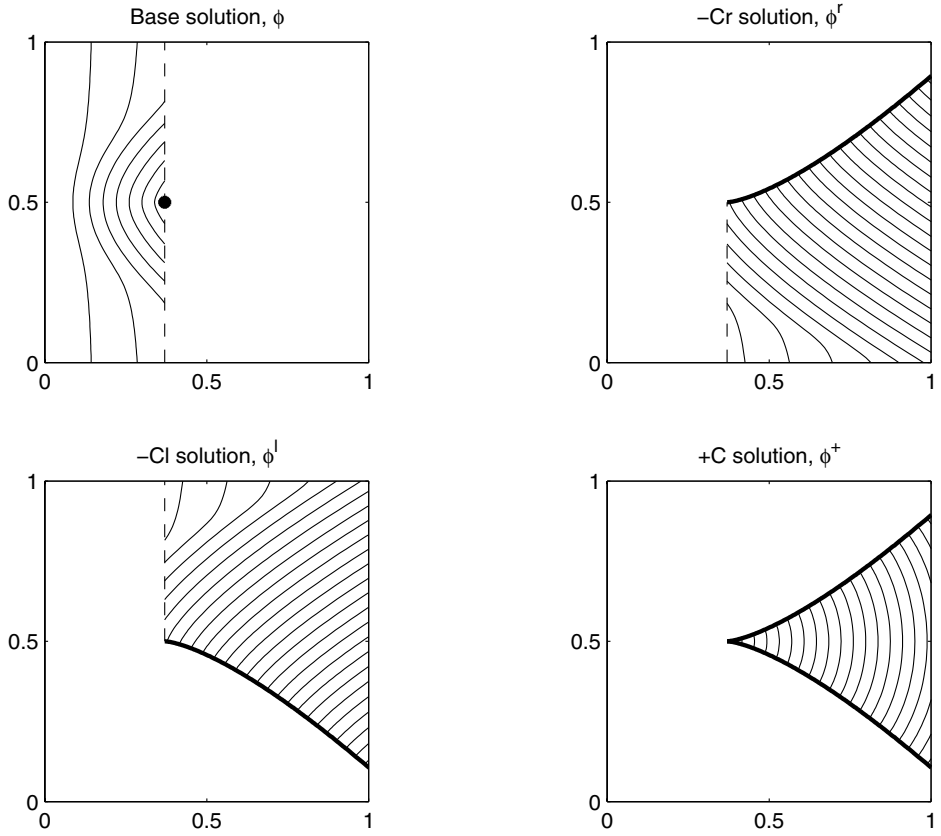


Figure 3.7. Direct computation of multivalued solutions for the problem in Figure 3.5. The computational domain is cut along caustic curves. The filled circle in the top left figure indicates point (x^*, y^*), where the caustic is first detected. Bold lines indicate the caustic curve.

vanishes at caustics, which can therefore be found numerically by checking sign changes in the computed geometrical spreading.

Figure 3.7 exemplifies the method for the problem described in Figure 3.5. In this case the paraxial approximation

$$\begin{aligned} \phi_x - \sqrt{\eta^2 - {\phi_y}^2} &= 0, \\ \phi(0,y) &= \phi_0(y), \end{aligned}$$

is used. The accompanying PDE is the Eulerian version of (2.42). Let the initial data satisfy $\partial_y \phi_0(y) = \eta(0,y)\cos\theta_0(y)$, and, in the notation of Section 2.5 on paraxial approximations (see page 200), set $\delta(x, y(x,r)) = (y_r(x,r), \theta_r(x,r))^T$, with $y_0(r) = r$ and $\theta_0(r)$ as above. Then, by (2.39) and the chain rule, $\delta(x,y)$ satisfies

$$\delta_x + (\tan\theta)\delta_y = \begin{pmatrix} u_y & u_\theta \\ v_y & v_\theta \end{pmatrix}\delta, \qquad \delta(0,y) = \begin{pmatrix} 1 \\ \frac{\mathrm{d}\theta_0(y)}{\mathrm{d}y} \end{pmatrix},$$

where

$$u = \tan\theta, \qquad v = \eta(x,y)^{-1}(\eta_y(x,y) - \eta_x(x,y)\tan\theta), \qquad \tan\theta = \frac{\phi_y}{\phi_x}.$$

The geometrical spreading is the first component of $\delta = (\delta_1, \delta_2)$, and a sign change in $\delta_1 \sim y_r$ indicates a caustic point. When such a point is discovered (call it (x^*, y^*)), the solution is split into three separate branches: $-Cr$, $-C\ell$ and $+C$. The first two are the outer branches, and the last one is the middle branch: see Figure 3.7. Let variables related to the $-Cr$, $-C\ell$ and $+C$ branches be superscripted by r, ℓ and $+$, respectively. For $-Cr$, define the domain

$$\Omega^r(\delta^r) = \big\{(x,y) \in (x^*, \infty) \times \mathbb{R} \mid \delta_1^r(x,y) > 0\big\}.$$

This represents the domain below the top caustic curve. The viscosity solution in the $-Cr$ branch is given by the free boundary problem

$$\begin{aligned} \phi_x^r - \sqrt{\eta^2 - {\phi_y^r}^2} &= 0, && (x,y) \in \Omega^r(\delta^r), \\ \delta_x^r + (\tan\theta^r)\delta_y^r &= \begin{pmatrix} u_y & u_\theta \\ v_y & v_\theta \end{pmatrix}\delta^r, && (x,y) \in \Omega^r(\delta^r), \\ \phi^r(x^*, y) &= \phi(x^*, y), && y \le y^*, \\ \delta^r(x^*, y) &= \delta(x^*, y), && y \le y^*, \end{aligned}$$

where we note that the domain Ω^r depends on the solution. The $-C\ell$ branch is treated similarly. For the middle branch, a Dirichlet problem, coupled by

boundary conditions to the other two systems, is solved:

$$\begin{aligned}
\phi_x^+ - \sqrt{\eta^2 - \phi_y^{+2}} = 0, \qquad & (x,y) \in \Omega^\ell(\delta^\ell) \cap \Omega^r(\delta^r), \\
\phi^+(x,y) = \phi^\ell(x,y), \qquad & (x,y) \in \partial\Omega^\ell(\delta^\ell), \\
\phi^+(x,y) = \phi^r(x,y), \qquad & (x,y) \in \partial\Omega^r(\delta^r).
\end{aligned}$$

The three solutions ϕ^r, ϕ^ℓ and ϕ^+ together make up the full multibranch solution. The strategy can be used recursively, detecting new caustics in the three solutions, and decomposing them into new branches if they appear. For fold caustics, the caustic can be traced more accurately by solving an ODE coupled to the eikonal equations (Benamou, Lafitte, Sentis and Solliec 2003, Solliec 2003). Let the top caustic curve be given by $y^{c\ell}(x) = y(x, s(x))$, where $s(x)$ is an unknown function and $y(x,r)$ is as on page 200. Then, since the geometrical spreading $y_r \equiv 0$ at the caustic,

$$\frac{\mathrm{d}y^{c\ell}}{\mathrm{d}x} = y_x + s_x y_r = y_x = \tan\theta = \frac{\phi_y^r(x, y^{c\ell})}{\phi_x^r(x, y^{c\ell})}.$$

See also Benamou and Solliec (2000).

The *slowness matching method* of Symes (Symes 1996, Symes and Qian 2003), is another method for finding multivalued solutions to the eikonal equation. It is based on the travel time map $\tau(\boldsymbol{x}_{\mathrm{src}}, \boldsymbol{x}_{\mathrm{rcv}})$, which gives the travel time of a wave from a source point $\boldsymbol{x}_{\mathrm{src}}$ to a receiver point $\boldsymbol{x}_{\mathrm{rcv}}$. Hence, if $(\boldsymbol{x}(t), \boldsymbol{p}(t))$ is a bicharacteristic going from $\boldsymbol{x}(0) = \boldsymbol{x}_1$ to $\boldsymbol{x}(T) = \boldsymbol{x}_2$, then $\tau(\boldsymbol{x}_1, \boldsymbol{x}_2) = T$. This function may of course be multivalued if there is more than one such bicharacteristic, and we distinguish between values by their associated arrival slowness, $\boldsymbol{p}(T)$. There is, however, always a neighbourhood of $\boldsymbol{x}_{\mathrm{src}}$ for which $\tau(\boldsymbol{x}_{\mathrm{src}}, \cdot\,)$ is smooth and single-valued. We denote this neighbourhood by $\mathcal{N}(\boldsymbol{x}_{\mathrm{src}})$. For fixed $\boldsymbol{x}_{\mathrm{src}}$, the map satisfies the eikonal equation with respect to $\boldsymbol{x}_{\mathrm{rcv}}$ in $\mathcal{N}(\boldsymbol{x}_{\mathrm{src}})$,

$$|\nabla_{x_{\mathrm{rcv}}} \tau(\boldsymbol{x}_{\mathrm{src}}, \boldsymbol{x}_{\mathrm{rcv}})| = \eta(\boldsymbol{x}_{\mathrm{rcv}}), \quad \boldsymbol{x}_{\mathrm{rcv}} \in \mathcal{N}(\boldsymbol{x}_{\mathrm{src}}), \quad \tau(\boldsymbol{x}_{\mathrm{src}}, \boldsymbol{x}_{\mathrm{src}}) = 0. \quad (3.3)$$

The slowness matching method draws on the following observation. Suppose there is a travel time $\tau(\boldsymbol{x}_1, \boldsymbol{x}_2)$ between the points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ with arrival slowness $\boldsymbol{p}$. If there is a third point $\boldsymbol{x}_3$ for which $\boldsymbol{x}_2 \in \mathcal{N}(\boldsymbol{x}_3)$ and the *slowness matching condition* holds at $\boldsymbol{x}_2$,

$$\boldsymbol{p} + \nabla_{x_{\mathrm{rcv}}} \tau(\boldsymbol{x}_3, \boldsymbol{x}_2) = 0, \quad (3.4)$$

then the travel times are additive, that is,

$$\tau(\boldsymbol{x}_1, \boldsymbol{x}_3) = \tau(\boldsymbol{x}_1, \boldsymbol{x}_2) + \tau(\boldsymbol{x}_2, \boldsymbol{x}_3) \quad (3.5)$$

is a travel time between $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$ with arrival slowness $\nabla_{x_r}\tau(\boldsymbol{x}_2, \boldsymbol{x}_3)$. Conversely, if a ray from $\boldsymbol{x}_1$ to $\boldsymbol{x}_3$ passes through a point $\boldsymbol{x}_2 \in \mathcal{N}(\boldsymbol{x}_3)$, then (3.4) and (3.5) hold. This follows from the uniqueness of solutions to the ray equations (2.13) and (2.14), and the fact that $\boldsymbol{p}(t) = \nabla\phi(\boldsymbol{x}(t))$ when ϕ is a smooth solution to the eikonal equation and $(\boldsymbol{x}(t), \boldsymbol{p}(t))$ is a bicharacteristic.

The method divides the domain into M layers of width Δx, with layer boundaries at constant x-coordinates, $x_n = n\Delta x$, $n = 0, \ldots, M$. At each point on layer boundary n, data are stored about the travel times $\tau_n(y)$ and arrival slownesses $\boldsymbol{p}_n(y)$ of rays crossing the boundary: see Figure 3.8, left. These functions will typically be multivalued. In order to compute the corresponding data for layer boundary $n+1$, the travel time map is used to piece together the multivalued solution via the slowness matching principle (3.4). For each point (x_{n+1}, y) on boundary $n+1$, find all points $(x_n, \tilde{y})$ on boundary n with an arrival slowness $\boldsymbol{p}_n(\tilde{y})$ such that

$$\boldsymbol{p}_n(\tilde{y}) + \nabla_{x_r}\tau(x_{n+1}, y;\, x_n, \tilde{y}) = 0. \tag{3.6}$$

Then, for each $\tilde{y}$ satisfying this condition, $\tau_{n+1}(y) = \tau_n(\tilde{y}) + \tau(x_n, \tilde{y};\, x_{n+1}, y)$ is a travel time at (x_{n+1}, y) with arrival slowness

$$\boldsymbol{p}_{n+1}(y) = \nabla_{x_r}\tau(x_n, \tilde{y};\, x_{n+1}, y).$$

See the example in Figure 3.8, right.

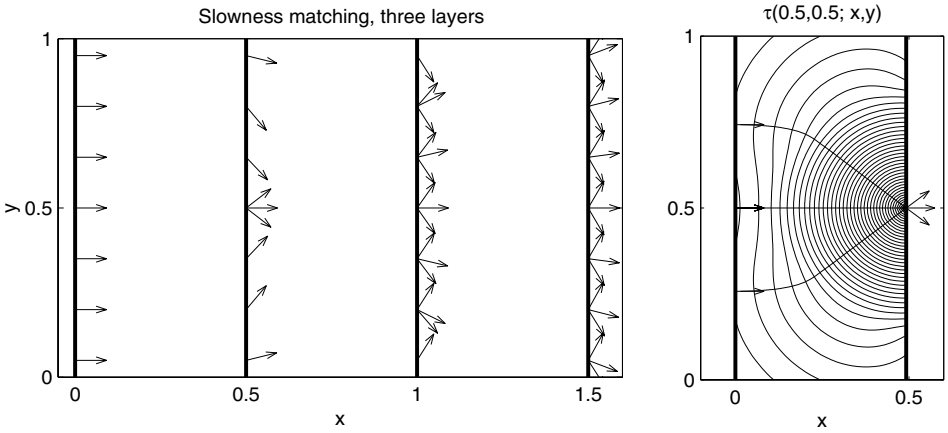


Figure 3.8. Slowness matching method. The left frame shows results for problem in Figure 3.5 with $\Delta x = 0.5$. Arrows indicate the arrival slownesses $\boldsymbol{p}_n$, bold lines indicate layer boundaries. The right frame shows the slowness matching condition and its three solutions at the point $(0.5, 0.5)$ on the second layer boundary. Iso curves of travel time map $\tau(0.5, 0.5;\, x, y)$, and rays associated to the three solutions are shown.

Numerically, the y-coordinate is discretized with a uniform grid, $\{y_j\}$. For each grid point (x_n, y_j), a list is maintained that contains one or more associated travel times and arrival slownesses. The travel time map τ is computed by solving the paraxial eikonal equation (2.43) with $\boldsymbol{x}_{\rm src} = (x_{n+1}, y_j)$ for all j. The slowness matching involves interpolating the travel time map on a regular grid and root finding.

Under the paraxial approximation the solution $\tilde{y}$ to (3.6) satisfies $|y - \tilde{y}| \leq C\,\Delta x$. Therefore, when Δx is small enough, $(x_n, \tilde{y}) \in \mathcal{N}(x_{n+1}, y)$ and the travel time map $\tau(x_{n+1}, y; \cdot)$ is smooth, single-valued around $(x_n, \tilde{y})$. Multivalued solutions can still be obtained, however, since there may be multiple solutions to the slowness matching condition (3.6): *cf.* Figure 3.8, right frame. In fact, all branches of the complete multivalued solution will be found in this way, if Δx is small enough and the paraxial approximation holds.

The cost of the slowness matching method is quite high when it is used as described above. Let N_x and N_y be the number of grid points in the x- and y-coordinate directions, respectively. Computing the travel time map τ involves $\mathcal{O}(N_x^2 N_y/M)$ operations, and it is the dominating cost when $M \ll N_x/\sqrt{N_y}$. It should be compared with the $\mathcal{O}(N_x N_y)$ cost of computing the viscosity solution with the paraxial approximation. The travel time map τ can, however, be re-used when the solution is sought for multiple sources, a case for which the slowness matching method is competitive. In this respect it falls in the same category as the fast phase space method of Fomel and Sethian (2002), described below in Section 4.5.

3.3. Phase space methods

This class of methods is based on the kinetic formulation in Section 2.3, and it can be seen as a compromise between ray tracing and Hamilton–Jacobi-based methods. The techniques try to keep the linear superposition principle of ray tracing and the regular representation of the solution over the computational domain that can be achieved by the approximation of a PDE.

We mentioned that the computational drawback of the Liouville equation was the large number of independent variables. To overcome this difficulty with computational complexity, we can either consider special solutions or modify the equations. The first approach leads to wave front methods and the second to moment-based methods.

In wave front methods, an interface representing a wave front is evolved following the kinetic formulation. There are different ways of representing interfaces, leading to different techniques. Lagrangian front tracking has been used, and it is closest to traditional ray tracing. Eulerian methods are based on the segment projection method, the level set method or the fast marching method for interface evolution.

For the other approach, with moment-based methods, new equations with fewer unknowns are derived from the kinetic formulation. A finite number of nonlinear partial differential equations for the moments of the kinetic density function f in (2.29) is obtained using a closure assumption that allows for a limited superposition principle.

We will discuss these methods in more detail below in Section 4 and Section 5, respectively.

3.4. Dynamic surface extension

The method of *dynamic surface extension* was introduced by Steinhoff and collaborators in Steinhoff, Wenren, Underhill and Puskas (1995) and Steinhoff, Fan and Wang (2000), and further refined by Ruuth, Merriman and Osher (2000). There are a few variants of the method, but the dependent variables in this technique are essentially the coordinates of the closest point on a wave front from a given $\boldsymbol{x}$-coordinate. This clever choice of representation and a time-stepping scheme following the rules of geometrical optics allow for linear superposition in an Eulerian representation. The present forms of the method have higher complexity than ray tracing, and certain cases will not be correctly described. One such example is a wave front given by a collapsing circle with two parallel tangent lines. At the time when the circle and the tangent lines have been reduced to one line, the information of the circle is lost and cannot be recovered. The method is quite straightforward and is being further developed.

We may illustrate the capability of the dynamic surface extension method to handle crossing wave fronts by the following simple one-dimensional algorithm. Let $X(x_j, t_0)$ be the location of the front which is closest to the grid point $x_j = j\Delta x$, at the initial time t_0, and let $|c(X)|$ be the velocity at X, with $c > 0$ if the front propagates in the positive x-direction and otherwise negative. The algorithm consists of two steps. In the first step, the location of the fronts are updated,

$$\tilde{X}(x_j, t_{n+1}) = X(x_j, t_n) + \Delta t c(X(x_j, t_n)), \qquad t_n = t_0 + n\Delta t,$$

and in the second step, the fronts are assigned to the appropriate grid points,

$$X(x_j, t_{n+1}) = \tilde{X}(x_{j+\ell}, t_{n+1}),$$

where $\ell \in \{-1, 0, 1\}$ is chosen such that

$$|\tilde{X}(x_{j+\ell}, t_{n+1}) - x_j|$$

is minimal.

It is easy to see how fronts are allowed to cross each other. For example, let $|c| = 1$, $\Delta x = 1$, and let a front at $X = 1/4$ be moving in the positive x-direction and one at $X = 3/4$ be moving in the negative x-direction.

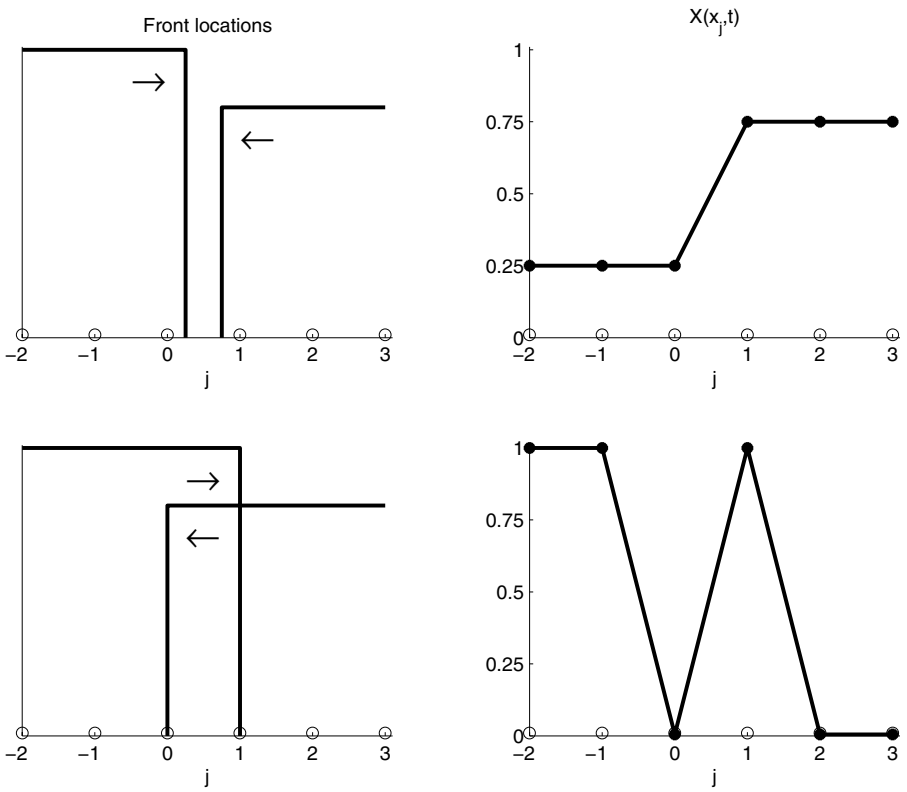


Figure 3.9. Dynamic surface extension. The top row shows the initial front locations and values of $X(x_j, t_0)$. The bottom row shows the state after one step of the algorithm with $\Delta t = 0.75$.

This example is illustrated in Figure 3.9. The algorithm will then give, for $1/2 < \Delta t < 1$,

$$X(x_j, t_0) = \begin{cases} 1/4, & j \le 0, \\ 3/4, & j > 0, \end{cases} \qquad \tilde{X}(x_j, t_1) = \begin{cases} 1/4 + \Delta t, & j \le 0, \\ 3/4 - \Delta t, & j > 0, \end{cases}$$

$$X(x_j, t_1) = \begin{cases} 1/4 + \Delta t, & j \le -1, \\ 3/4 - \Delta t, & j = 0, \\ 1/4 + \Delta t, & j = 1, \\ 3/4 - \Delta t, & j > 1. \end{cases}$$

The grid points at $j = 0, 1$ have registered the new front location, and this information will spread to the other j-values at later times. The extension to multi-dimensional problems also requires an interpolation step, but unfortunately not all cases of front propagation are well represented.

4. Wave front methods

Wave front methods are related to standard ray tracing, but instead of computing a sequence of individual rays a wave front is evolved in physical or phase space. This can be based on the ODE formulation (2.11) and (2.12) or the PDE Liouville equation (2.29).

The propagation of a wave front in the xy-plane is given by the velocity $c(\boldsymbol{x})$ in its normal direction $\hat{\boldsymbol{n}}$. The velocity $\boldsymbol{u} = (u, v)$ of the wave front in the xy-plane is thus

$$(u, v) = c(\boldsymbol{x})\hat{\boldsymbol{n}} = c(\boldsymbol{x})\,(\cos\theta, \sin\theta), \tag{4.1}$$

where θ is the angle between the normal vector and the x-axis. At caustic and focus points, the normal direction is not defined, and front tracking methods based on (4.1) break down.

The tracing of the wave fronts in phase space facilitates problems including the formation of caustics, as is seen in the following simple example. In Figure 4.1, left frame, an initial circular wave front is given in the xy-plane. This frame also displays the phase plane curve γ in $\mathbb{R}^3$ together with its $x\theta$- and $y\theta$-projections. Let the circular wave front contract with time in

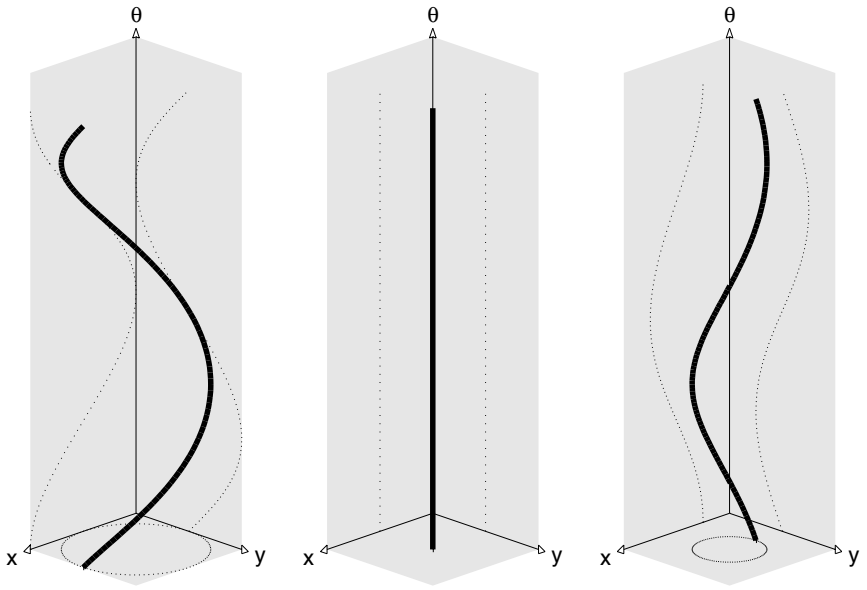


Figure 4.1. Phase plane curve γ: thick line, with projections onto xy-, $x\theta$- and $y\theta$-planes: dotted lines. The left frame shows the initial circular wave front at $t = 0$. The middle frame shows the focus at $t = 1$. The right frame shows the wave front after the focus at $t = 1.5$.

a constant medium, $c(\boldsymbol{x}) \equiv 1$, and be focused to a point $(x, y) = (1,\ 1)$ at time $t = 1$. Although degenerate in the xy-plane, the representations of γ at $t = 1$ in the $x\theta$- and $y\theta$-planes are smooth and the evolution, as well as the computation of amplitudes, can easily be continued to $t > 1$.

For the PDE-based wave front methods in phase space, the evolution of the front is given by (2.29) and the front is represented by some interface propagation technique. We shall here discuss the application of the segment projection method (Engquist *et al.* 2002, Tornberg and Engquist 2003), and also briefly outline level set techniques (Osher and Sethian 1988) and methods based on fast marching (Fomel and Sethian 2002). The segment projection method uses an explicit representation of the wave front, while the level set and fast marching methods use implicit representations: see below. These classes of techniques are based on Eulerian grids and thus there is no need for redistribution of marker points.

4.1. Wave front construction

Wave front construction is a front tracking method in which Lagrangian markers on the phase space wave front are propagated according to the ray equations (2.13) and (2.14). To maintain an accurate description of the front, new markers are adaptively inserted by interpolation when the resolution of the front deteriorates, *e.g.*, in shadow zones. The method was introduced by Vinje, Iversen and Gjøystdal (1992, 1993).

Let us consider the two-dimensional case. As in Section 2.2, we assume that the wave front in phase space is described by $(\boldsymbol{x}(t, r), \boldsymbol{p}(t, r))$ at time t, where r is the parametrization induced by the parametrization of the source. The markers $(\boldsymbol{x}_j^n, \boldsymbol{p}_j^n)$ are initialized uniformly in r at $t = 0$, $(\boldsymbol{x}_j^0, \boldsymbol{p}_j^0) = (\boldsymbol{x}(0, j\Delta r), \boldsymbol{p}(0, j\Delta r))$. Each marker is updated by a standard ODE-solver, such as a fourth order Runge–Kutta method, applied to the ray equations (2.13) and (2.14). Thus the markers approximately trace rays, and

$$\boldsymbol{x}_j^n \approx \boldsymbol{x}(n\Delta t, j\Delta r), \qquad \boldsymbol{p}_j^n \approx \boldsymbol{p}(n\Delta t, j\Delta r), \qquad \forall n > 0, j.$$

See Figure 4.2, left.

When the resolution of the wave front worsens, new markers must be inserted. The location in phase space of the new points is found via interpolation from the old points. A new marker $(\boldsymbol{x}_{j+1/2}^n,\ \boldsymbol{p}_{j+1/2}^n)$ between markers j and $j + 1$ would satisfy

$$\boldsymbol{x}_{j+1/2}^n \approx \boldsymbol{x}(n\Delta t, j\Delta r + \Delta r/2), \qquad \boldsymbol{p}_{j+1/2}^n \approx \boldsymbol{p}(n\Delta t, j\Delta r + \Delta r/2).$$

See Figure 4.2, middle. When deciding on whether to add new markers, it is not sufficient only to look at the distance in physical space between the old markers, because it degenerates at caustics and focus points. The distance in the phase variable should also be taken into account (Sun 1992). A useful

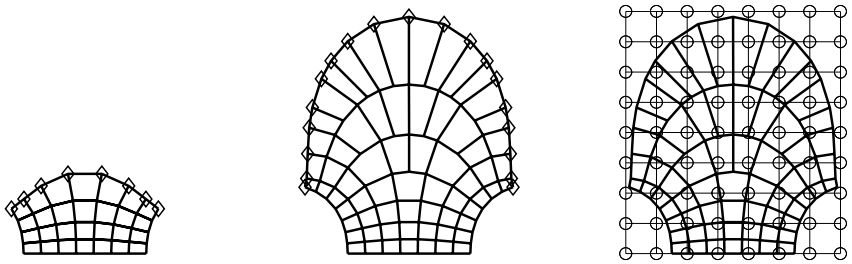


Figure 4.2. Wave front construction. Markers ($\diamond$) on the wave front are propagated as ordinary rays (left). The grid approximates the wave front in physical space, $\boldsymbol{x}(t,r)$ at constant t- and r-values. When the markers move too wide apart to accurately describe the front, new markers are inserted via interpolation (middle). The travel times and possibly amplitudes on the wave front are interpolated onto a regular grid as the front propagates (right).

criterion is to add a new marker between markers j and $j+1$ if

$$|\boldsymbol{x}_{j+1}^n - \boldsymbol{x}_j^n| \geq \text{TOL} \qquad \text{or} \qquad |\boldsymbol{p}_{j+1}^n - \boldsymbol{p}_j^n| \geq \text{TOL}.$$

for some tolerance TOL. This criterion ensures that the phase wave front remains fairly uniformly sampled. Lambaré, Lucio and Hanyga (1996) introduced another criterion, where more points are added when the curvature of the phase space wave front is large. For each marker, they compute the additional quantities

$$\boldsymbol{X}_j^n \approx \boldsymbol{x}_r(n\Delta t, j\Delta r), \qquad \boldsymbol{P}_j^n \approx \boldsymbol{p}_r(n\Delta t, j\Delta r),$$

via the ODE system (2.26). Based on the fact that

$$\begin{aligned} |\boldsymbol{x}(t,r+\Delta r) - \boldsymbol{x}(t,r) - \Delta r \boldsymbol{x}_r(t,r)| &\approx \frac{1}{2}(\Delta r)^2 |\boldsymbol{x}_{rr}| \geq \frac{1}{2}(\Delta r |\boldsymbol{x}_r|)^2 \kappa(r) \\ &\approx \frac{1}{2}|\boldsymbol{x}(t,r+\Delta r) - \boldsymbol{x}(t,r)|^2 \kappa(r), \end{aligned}$$

where $\kappa(r)$ is the curvature, the criterion for adding a new marker is taken as

$$|\boldsymbol{x}_{j+1}^n - \boldsymbol{x}_j^n - \Delta r \boldsymbol{X}_j^n| \geq \text{TOL} \qquad \text{or} \qquad |\boldsymbol{p}_{j+1}^n - \boldsymbol{p}_j^n - \Delta r \boldsymbol{P}_j^n| \geq \text{TOL}.$$

The computed variables $\boldsymbol{X}_j^n$, which is the geometrical spreading, and $\boldsymbol{P}_j^n$ are also used for computing the amplitude and to simplify high-order interpolation when inserting new markers, and in the grid interpolation below.

Finally, the interesting quantities carried by the markers on the wave front, such as travel time and amplitude, are interpolated down on a regular Cartesian grid: see Figure 4.2, right. The wave front construction covers the physical space by quadrilateral 'ray cells'. The interpolation step involves mapping the grid points to the right ray cells, in order to find the markers

and marker positions from which to interpolate. This can be complicated. See, *e.g.*, Bulant and Klimeš (1999).

In three dimensions the wave front is a two-dimensional surface. The method generalizes by using a triangulated wave front, and performing the same steps as above. Interpolation can be done in essentially the same way as in two dimensions, but the ray cells are now triangular prism-like 'ray tubes'. The topology of the triangulation may change with time, and there is no simple parametrization for general surfaces.

4.2. Segment projection method

Let us first consider the segment projection method for general interfaces and then apply the technique to geometrical optics. In order to make the presentation more clear, we shall discuss the two-dimensional case, and this is also the case for which the most general software has been developed (Tornberg and Engquist 2003).

The segment projection method is a computational method for tracking the dynamic evolution of interfaces (Tornberg 2000, Tornberg and Engquist 2000). The basic idea is to represent a curve or surface as a union of segments. Each segment is chosen such that it can be given as a function of the independent variables. The representation is thus analogous to a manifold being defined by an atlas of charts. The motions of the individual segments are given by partial differential equations based on the physics describing the evolution of the interfaces.

The segments representing a curve γ in $\mathbb{R}^2$ are here given by functions $Y_j(x)$ and $X_k(y)$. The domains of the independent variables of these functions are projections of the segments onto the coordinate axis. The coordinates of the points on γ are given by $(x, y) = (x, Y_j(x))$ or $(x, y) = (X_k(y), y)$. For each point on a curve γ, there is at least one segment defining the curve. To make the description complete, information about the connectivity of segments must also be provided. For each segment in one variable there is information regarding which part of the curve has overlap with segments in the other variable, as well as pointers to these segments.

The number of segments needed to describe a curve depends on the shape of the curve. An extremum of a function $Y_j(x)$ defines a separation point for the y-segments, as no segment given as a function of y can continue past this point. Similarly, an extremum of a function $X_k(y)$ defines a separation point for the x-segments. A sketch of a distribution of segments is shown in Figure 4.3. For moving interfaces, $Y_j = Y_j(x, t)$ and $X_k = X_k(y, t)$ are also functions of time.

The segments are moved by equations of motion, and after each numerical advection step, the segment representation is re-initialized. Dynamic creation and elimination of segments are employed to follow the evolution

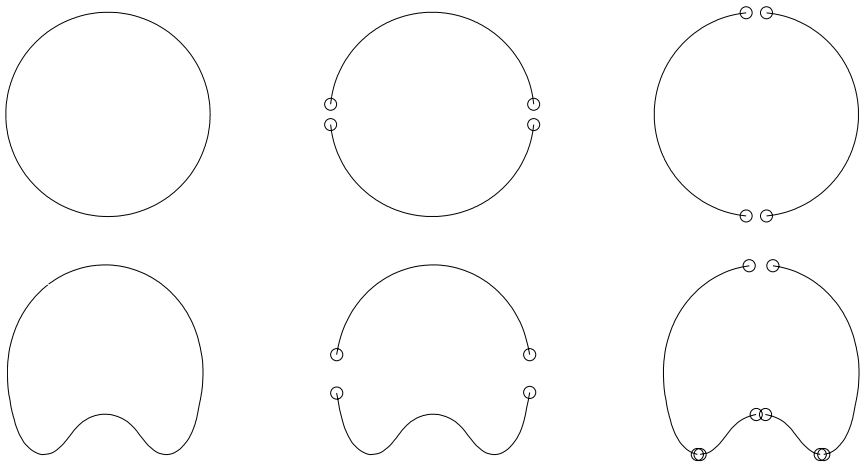


Figure 4.3. Segment structure for circle and deformed circle: curve γ (left), x-segments (middle), y-segments (right).

of the curves. New segments are created if necessary, and segments are removed when they are no longer needed. The connectivity of segments must be kept updated in such a way that the pointers relating segments represent the current configuration. If we assume that the lower deformed circle in Figure 4.3 has evolved from the circle above, a new maximum and two new minima have appeared in the lower x-segment. The number of y-segments should then be increased, as is seen in the figure.

For each segment, the domain of the independent variable must be defined. These segments are numerically given by arrays for Y_j and X_k. The domains of the independent variables, the arrays and information about connectivity between the segments define the structure that represents the curve.

From the definition of an x-segment, an ordered set of numbers is created that contains the start and end points of the segment, together with the extremum points of the segment. The intervals between these points correspond to different segments of the other variable. It is necessary to keep track of the connections between these segments.

Let a velocity field $\boldsymbol{u} = (u, v)^T$ be given, by which the curve should move. The segments $y = Y(x, t)$ and $x = X(y, t)$ are updated according to the partial differential equations

$$\frac{\partial Y}{\partial t} + u\frac{\partial Y}{\partial x} = v, \tag{4.2}$$

$$\frac{\partial X}{\partial t} + v\frac{\partial X}{\partial y} = u. \tag{4.3}$$

Note that there is only one spatial variable present in each of these equations. Quantities that are transported by the velocity field can also be defined as functions on the segments. Let $\mathcal{F}^t(x,y)$ be the flow generated by $\boldsymbol{u}$. If $r(x,y,t)$ evolves according to the ODE

$$\frac{\mathrm{d}r(\mathcal{F}^t,t)}{\mathrm{d}t} = h(\mathcal{F}^t, r(\mathcal{F}^t,t),t),$$

for fixed (x,y), then

$$\frac{\partial R^x}{\partial t} + u\frac{\partial R^x}{\partial x} = h(x,Y,R^x,t), \tag{4.4}$$

$$\frac{\partial R^y}{\partial t} + v\frac{\partial R^y}{\partial y} = h(X,y,R^y,t), \tag{4.5}$$

where $R^x(x,t) = r(x,Y(x,t),t)$ and $R^y(y,t) = r(Y(y,t),y,t)$.

Boundary conditions must be defined for the segments. They are either given in the original problem formulation or interpolated from an overlapping segment in the other coordinate direction. Note that this interpolation is well defined as an interpolation on an irregular mesh from the discrete segment in the other coordinate direction.

After the numerical advection step based on (4.2) and (4.3), we need to review the segment structure. If no new extrema have appeared and no old ones have disappeared, no change needs to be made in the structure of the segments.

Any moving curve γ is represented by overlapping segments. These segments evolve individually and may separate slightly in the overlapping regions due to numerical errors. A re-initialization is applied in every time step to realign the segments. This is done by a weighted interpolation. A segment will typically yield the most accurate description of the curve if its slope is small.

When different parts of γ cross each other, geometric rules for the segment interaction must be given. Examples are the merging of two bubbles in multiphase flow and the reflection of a wave front γ_1 meeting a curve γ_2, representing a perfect reflector.

The advection and re-initialization process for a structure of segments, representing a curve γ, can be summarized as follows.

(1) Advect all x- and y-segments from a velocity field $\boldsymbol{u} = (u,v)$ and evolve associated quantities defined on the segments by numerical approximations of (4.2)–(4.5).
(2) Update the segment structure.
(3) For each segment whose domain of definition has increased, new values need to be defined. These are interpolated from the corresponding segment in the other coordinate direction.

(4) Interpolate the segment between overlapping parts of the x- and y-segments. The new values are assigned using a weight function based on the slopes of the segments.

(5) Rearrange the segment structure from the rules of segment interactions.

These steps are generic and essentially the same for different applications. Common software will thus apply to different problems with only minor modifications, for example, in the advection and the interaction algorithms.

Curves of co-dimension two in $\mathbb{R}^3$ are approximated by their projections onto the two-dimensional coordinate planes. If a curve

$$\gamma(t) : \{X_1(s,t), X_2(s,t), X_3(s,t)\},$$

is parametrized by s and evolves by the velocity field

$$\boldsymbol{u}(\boldsymbol{x},t) = (u_1(\boldsymbol{x},t), u_2(\boldsymbol{x},t), u_3(\boldsymbol{x},t)),$$

we have

$$\frac{\mathrm{d}X_j}{\mathrm{d}t} = u_j(\boldsymbol{x},t), \quad j = 1,2,3.$$

Let the projection of γ onto the x_jx_k-plane be represented by a set of segment functions of the type $x_j = X^{jk}(x_k,t)$ and $x_k = X^{kj}(x_j,t)$. The evolution of the segment functions in all three projection planes is then given by the equations

$$\frac{\partial X^{jk}}{\partial t} + u_k \frac{\partial X^{jk}}{\partial x_k} = u_j, \quad j = 1,2,3;\ k = 1,2,3;\ j \neq k. \tag{4.6}$$

The projections in the x_jx_k-planes are updated in time following the steps (1)–(5) above. A sixth step is then added with interpolation between the representations in the three coordinate planes. This step is similar to step (4) and is in general needed in order to define u_k and u_j in (4.6). The simulations presented in this paper do not require such interpolations, however.

There is as yet no general software for two-dimensional surfaces in $\mathbb{R}^3$. The principle is analogous to the lower-dimensional case. The surface Σ is represented by functions defined on the three coordinate planes. The functions $x_\ell = X_\ell^{jk}(x_j,x_k,t)$ define the segments as in $\mathbb{R}^2$ and the union of segments defines Σ. Given the velocity field $\boldsymbol{u}(\boldsymbol{x},t)$, the motion of the segments is given by

$$\frac{\partial X_\ell^{jk}}{\partial t} + u_j \frac{\partial X_\ell^{jk}}{\partial x_j} + u_k \frac{\partial X_\ell^{jk}}{\partial x_k} = u_\ell,$$
$$j = 1,2,3, \quad k = 1,2,3, \quad \ell = 1,2,3, \quad j \neq k,\ j \neq \ell,\ k \neq \ell.$$

4.3. Segment projection method for geometrical optics

The segment projection method will be applied to track the evolution in phase space of fronts that are given by geometrical optics. For two space dimensions a curve γ in $\mathbb{R}^3$ is tracked, and for three space dimensions a surface Σ in $\mathbb{R}^5$ is evolved. We shall mainly discuss the two-dimensional case and only give one three-dimensional example on page 230. In the presentation of the method, we may assume that γ or Σ and their projections are functions. The segment projection technique will reduce the general case to a set of segments that are functions of some of the independent variables.

Let the independent variables be x, y and θ. The orthogonal projections of γ onto the xy-, $x\theta$- and $y\theta$-planes are denoted by γ_{xy}, $\gamma_{x\theta}$ and $\gamma_{y\theta}$, respectively. The evolution of $\gamma = \gamma(t)$ will be determined by the two-dimensional segment projection method, as presented in Section 4.2.

From the general equations (4.2) and (4.3) and the velocity field (4.1), we get the Eulerian form of the evolution equations for the x- and y-segments in the xy-plane, respectively:

$$\begin{cases} \dfrac{\partial Y^x}{\partial t} + c(x, Y^x(x,t))\,\cos\theta\,\dfrac{\partial Y^x}{\partial x} = c(x, Y^x(x,t))\sin\theta, \\ \dfrac{\partial X^y}{\partial t} + c(X^y(y,t), y)\,\sin\theta\,\dfrac{\partial X^y}{\partial y} = c(X^y(y,t), y)\cos\theta, \end{cases} \tag{4.7}$$

The approach of tracking the front only in the xy-plane, computing θ from the segments, breaks down at caustics. Therefore, the front should be tracked in phase space, and the other two projections are needed. From (2.22), (2.23) and (2.24) we get the velocity field needed to apply (4.2) and (4.3) to the segment equations in the $x\theta$- and $y\theta$-planes. Let the x- and θ-segments in the $x\theta$-plane be denoted by Θ^x and X^θ, and let the y- and θ-segments in the $y\theta$-plane be Θ^y and Y^θ, respectively. The segment equations are

$$\begin{cases} \dfrac{\partial \Theta^x}{\partial t} + c\cos\theta\,\dfrac{\partial \Theta^x}{\partial x} = \alpha, \\ \dfrac{\partial X^\theta}{\partial t} + \alpha\,\dfrac{\partial X^\theta}{\partial \theta} = c\cos\theta, \end{cases} \qquad \begin{cases} \dfrac{\partial \Theta^y}{\partial t} + c\sin\theta\,\dfrac{\partial \Theta^y}{\partial y} = \alpha, \\ \dfrac{\partial Y^\theta}{\partial t} + \alpha\,\dfrac{\partial Y^\theta}{\partial \theta} = c\sin\theta, \end{cases} \tag{4.8}$$

$$\alpha = \frac{\partial c(\boldsymbol{x})}{\partial x}\,\sin\theta - \frac{\partial c(\boldsymbol{x})}{\partial y}\,\cos\theta. \tag{4.9}$$

The one-dimensional hyperbolic equations above are easily solved by standard numerical methods. Note that the representation of the phase plane curve γ may be degenerate for the projection onto one of the coordinate planes, but there will always be two projections for which γ is well represented.

When η is constant the amplitude on the curve can easily be calculated by post-processing of the results from (4.8). (Below we will compute the amplitude for problems with variable η.) Consider, for instance, an initial curve $(x_0(r), y_0(r))$ with amplitude $A_0(r)$ moving in the normal direction $(\cos\theta_0(r), \sin\theta_0(r))^T$. We let r be the parametrization defined such that $\theta_0(r) = r$. Then by (2.24), since $\alpha \equiv 0$, we will have $\theta(t, r) = r$ also for $t > 0$ and we see that r and θ are therefore the same parametrization for all times. By (2.25), the amplitude at time t is given by

$$A^2(t,\theta) = \frac{A_0^2(\theta) q(\theta, 0)}{q(t,\theta)}, \qquad q(t,\theta) = \left((x_\theta(t,\theta))^2 + (y_\theta(t,\theta))^2\right)^{1/2}. \quad (4.10)$$

We note finally that q can be computed from $X^\theta(t,\theta)$ and $Y^\theta(t,\theta)$ in (4.8).

We will also make use of the paraxial approximation, discussed in Section 2.5 (see page 200) in order to reduce two-dimensional problems to one dimension. Time is thus not explicitly needed in the calculation and θ can be computed as a function of x and y, and y as a function of x and θ. From (2.39) we get the velocity field

$$\boldsymbol{u} = (u,\, v)^T = \left(\tan\theta,\ \frac{1}{\eta}(\eta_y - \eta_x \tan\theta)\right)^T$$

for this setting, and the partial differential equations for the segments $\theta = \Theta(x, y)$ and $y = Y(x, \theta)$ are

$$\frac{\partial \Theta}{\partial x} + u\frac{\partial \Theta}{\partial y} = v,$$

$$\frac{\partial Y}{\partial x} + v\frac{\partial Y}{\partial \theta} = u.$$

The travel time T is now a quantity defined on the phase plane curve, and can be computed according to (4.4, 4.5). Let $\mathcal{F}^x(y,\theta) = (\mathcal{F}^x_y, \mathcal{F}^x_\theta)^T$ be the flow generated by $\boldsymbol{u}$. Keeping in mind that T is given by the phase ϕ, we see from (2.40) that $\mathrm{d}T(x, \mathcal{F}^x)/\mathrm{d}x = \eta/\cos\theta$. This yields the following differential equations, defined for the y-segments and θ-segments respectively:

$$\frac{\partial T^y}{\partial x} + u\frac{\partial T^y}{\partial y} = \frac{\eta}{\cos\theta}, \quad (4.12)$$

$$\frac{\partial T^\theta}{\partial x} + v\frac{\partial T^\theta}{\partial \theta} = \frac{\eta}{\cos\theta}.$$

To compute the amplitude we use equations (2.41) and (2.42). Suppose the initial data are given as $y(0, r) = y_0(r)$ and $\theta(0, r) = \theta_0(r)$ with amplitude $A_0(y_0(r), \theta_0(r))$. When Θ is well defined, the amplitude on the segment

is given by (2.41):

$$A(x,y) = A_0(\mathcal{F}^{-x}(y,\Theta(x,y)))\sqrt{\frac{\eta(0,\mathcal{F}_y^{-x}(y,\Theta(x,y)))|y_r(0,\mathcal{F}^{-x}(y,\Theta(x,y)))|}{\eta(x,y)|y_r(x,y,\Theta(x,y))|}}.$$

When Y is well defined, the same equality holds after replacing $(y,\Theta(x,y))$ by $(Y(x,\theta),\theta)$. In order to compute $A(x,y)$ we must hence also evolve $\mathcal{F}^{-x}$ and y_r as quantities on the curve. Let J be the Jacobian of $\boldsymbol{u}(x,y,\theta)$ with respect to (y,θ), and set $\boldsymbol{z} = (y_r,\theta_r)^T$. Then, by the definition of $\mathcal{F}^x$ and (2.42),

$$\frac{\mathrm{d}\mathcal{F}^{-x}(x,\mathcal{F}^x)}{\mathrm{d}x} = 0, \qquad \frac{\mathrm{d}\boldsymbol{z}(x,\mathcal{F}^x)}{\mathrm{d}x} = J(x,\mathcal{F}^x)\boldsymbol{z}(x,\mathcal{F}^x), \tag{4.13}$$

for fixed (y,θ). Note that both $\mathcal{F}^{-x}$ and $\boldsymbol{z}$ remain bounded and smooth also at caustics, where the amplitude A becomes infinite. The quantities are then given by the PDEs

$$\frac{\partial F^y}{\partial x} + u\frac{\partial F^y}{\partial y} = 0, \qquad \frac{\partial Z^y}{\partial x} + u\frac{\partial Z^y}{\partial y} = JZ^y,$$

$$\frac{\partial F^\theta}{\partial x} + v\frac{\partial F^\theta}{\partial \theta} = 0, \qquad \frac{\partial Z^\theta}{\partial x} + v\frac{\partial Z^\theta}{\partial \theta} = JZ^\theta.$$

Here, F^y, F^θ give $\mathcal{F}^{-x}$ and Z^y, Z^θ give $\boldsymbol{z}$ on the segments $\Theta(x,y)$ and $Y(x,\theta)$ respectively. Initial data for those equations are $F^y(0,y) = (y,\Theta(0,y))^T$, $F^\theta(0,\theta) = (Y(0,\theta),\theta)^T$ and $Z^y(0,y) = (\partial_r y_0(r)\ ,\partial_r\theta_0(r))^T$ for r given by $y_0(r) = y$ and $Z^\theta(0,\theta) = (\partial_r y_0(r)\ ,\partial_r\theta_0(r))^T$, with $\theta_0(r) = \theta$.

We shall present three computational examples in order to describe different aspects of the method. All of the examples involve caustics and superposition.

Contracting elliptical and ellipsoidal wave front

The initial values in the first example correspond to a one-dimensional elliptical wave front in $\mathbb{R}^2$: see Figure 4.4. The initial motion is contraction and the index of refraction is constant. The projections of the front in phase space onto the $x\theta$- and $y\theta$-planes are smooth even through caustics: see Figure 4.5. The function α defined in (4.9) vanishes, which simplifies the calculations. No differential equation needs to be solved in the xy-plane. The equations (4.8) are sufficient and the xy-location of the wave front is given by (X^θ, Y^θ). With constant index of refraction in these examples, there is no need to perform the interpolation discussed in Section 4.2 as a sixth step in the segment projection process. There is no problem in

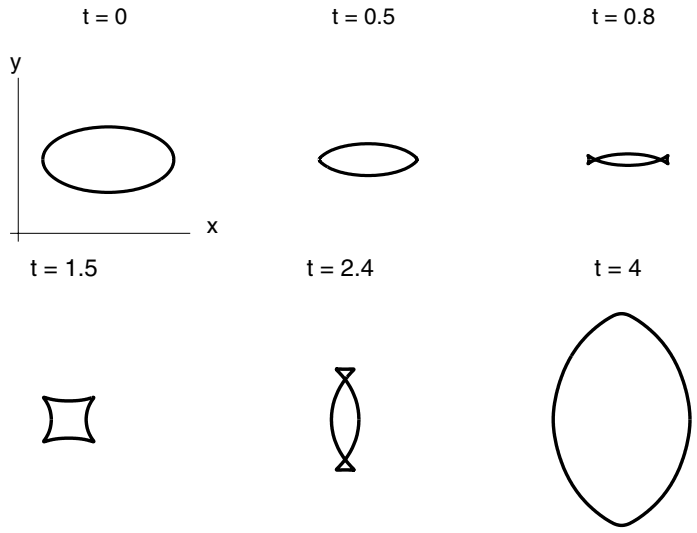


Figure 4.4. Evolution of an initially elliptical wave front in the xy-plane.

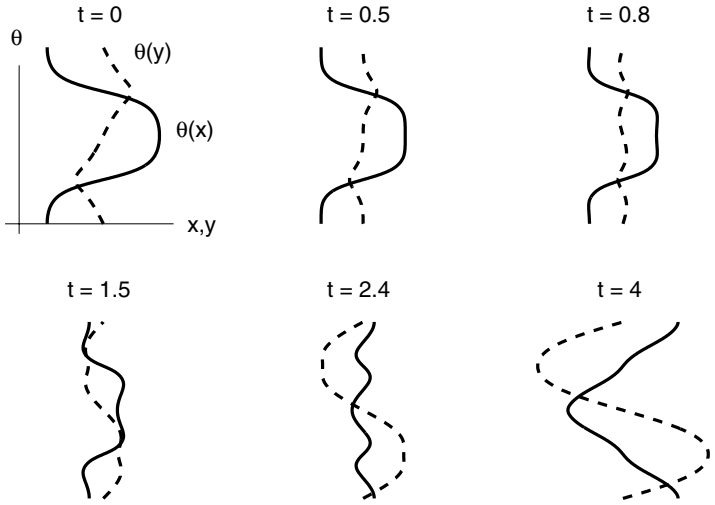


Figure 4.5. Projection of phase plane curves γ: solid lines show $x\theta$-plane, dashed lines show $y\theta$-plane.

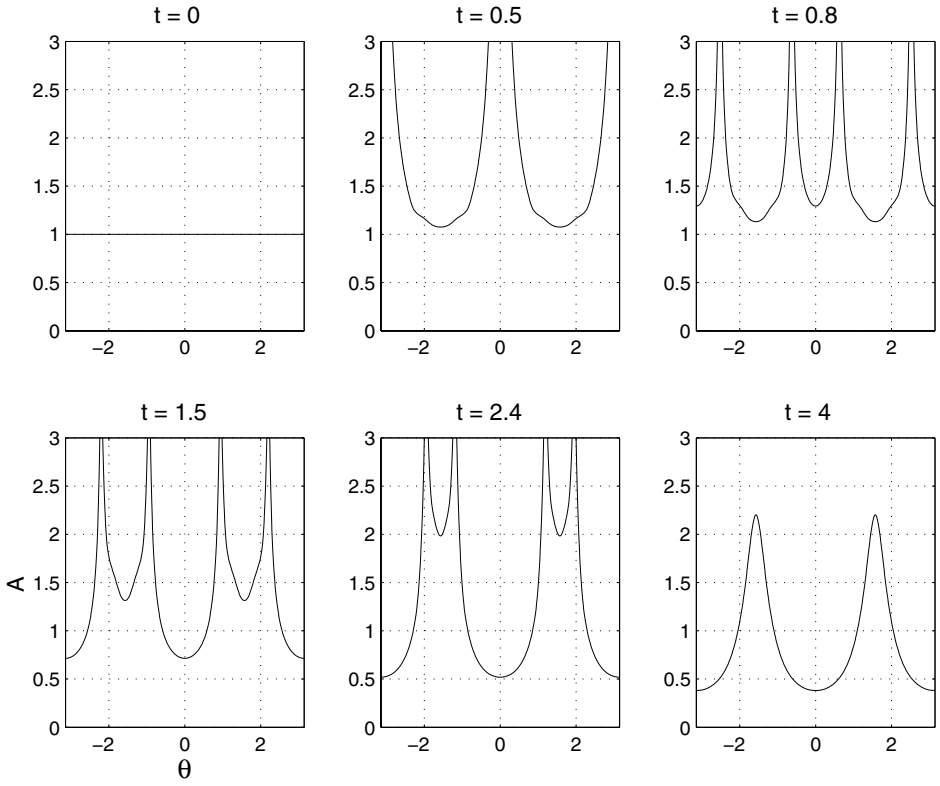


Figure 4.6. Evolution of amplitude as function of θ.

calculating the amplitude through the formation of caustics using (4.10) (see Figure 4.6), even on a coarse grid; the quantities that we solve for are smooth. However, in order to resolve the spikes in the post-processed amplitude for the presentation, we used a fairly dense grid, $\Delta\theta = 2\pi/512 \approx 0.01$.

The previous example can easily be extended to a surface in $\mathbb{R}^3$. Even though general software for the three-dimensional segment projection method has not yet been developed, this simulation can be done. The reason is that the index of refraction is constant, and then only one segment is needed in each of the coordinate planes given in Figure 4.7, which displays the projections of the initial surface in phase space onto the $x\theta_1\theta_2$-, $y\theta_1\theta_2$- and $z\theta_1\theta_2$-spaces. In Figure 4.8 we see the evolution of the wave front in xyz-space at different times. For the general case of variable index of refraction, a larger number of segments could be required. See Tornberg and Engquist (2003) for a simple three-dimensional example with several overlapping segments. The grid resolution used in the computations was $\Delta\theta_1 = \Delta\theta_2 = 2\pi/60 \approx 0.1$.

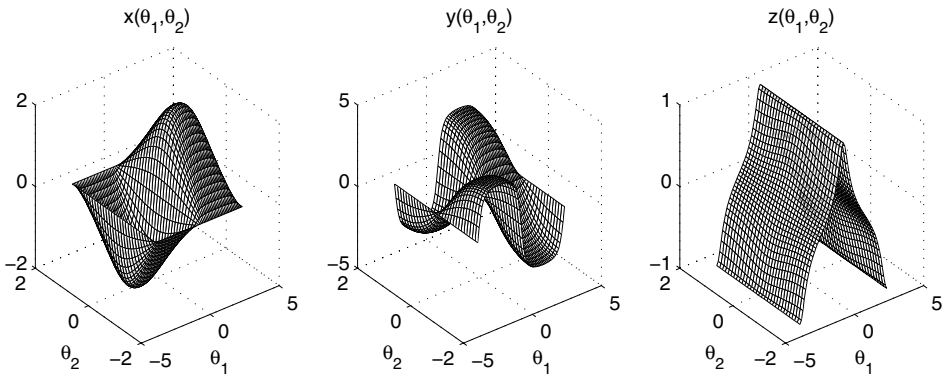


Figure 4.7. Projections of phase space surface Σ onto different coordinate planes at $t = 0$.

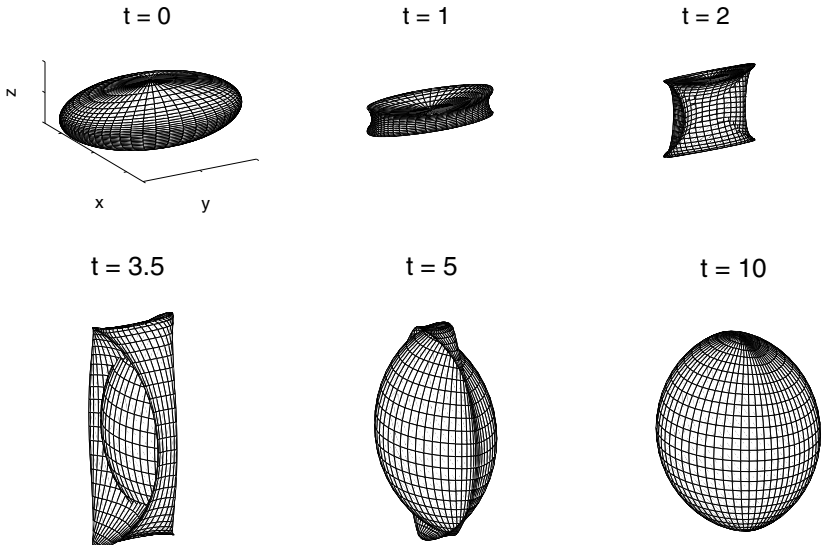


Figure 4.8. Evolution of wave front in xyz-space.

Wave guide

In this simulation an incoming plane wave at $x = 0$, with constant amplitude $A = 1$, enters a wave guide. The variable index of refraction in the wave guide, $\eta(x, y) = 1+\exp(-y^2)$, causes the rays to bend. A ray-traced solution is shown in Figure 4.9, together with amplitude and wave fronts as computed by the segment projection method.

Since all rays go in the positive x-direction in this simulation, we can use the paraxial approximation discussed in Section 4.3. The travel time T is a well-defined quantity on the phase plane curve. It is computed via (4.12), and the y-segment functions $T^y(x, y)$ are used to plot the wave fronts in Figure 4.9. Note that, since we use a constant Δx and a uniform y grid, T is in fact obtained on a uniform $x \times y$ grid. At each point the number of y-segments corresponds to the number of crossing wave fronts.

The phase space curve in the $y\theta$-plane becomes complicated at larger x-values but it is still handled well by the segment projection method: see Figure 4.10. The grid resolution was high ($\Delta y = 7/4096 \approx 0.002$) to get an accurate rendering of the amplitude at the far end of the wave guide.

4.4. Level set methods for geometrical optics

The level set method was introduced by Osher and Sethian (1988) as a general technique for the simulation of moving interfaces. Level set methods for special applications had been introduced earlier. The method uses an implicit representation of an interface in $\mathbb{R}^d$ as the zero level set of a function $\phi(t, \boldsymbol{x})$. The motion of the interface following a velocity field $\boldsymbol{u}(t, \boldsymbol{x})$ is given by a PDE for the level set function ϕ,

$$\phi_t + \boldsymbol{u} \cdot \nabla\phi = 0. \tag{4.14}$$

This technique has been successfully applied to many different types of problems. Examples are multiphase flow, etching, epitaxial growth, image processing and visualization, described in the two books by Osher and Fedkiw (2002) and Sethian (1999). An attractive property is that equation (4.14) can be applied without modifications even if the topology of the interface changes as, for example, when merging occurs in multiphase flow.

For the location of the interface to be well defined, the gradient of ϕ in the direction normal to the interface should be bounded away from zero. In practice, the level set function ϕ is re-initialized at regular time intervals such that it is approximately a signed distance function to the interface.

If $\phi(t, \boldsymbol{x}) = 0$ represents an evolving wave front given by geometrical optics, the velocity is $\boldsymbol{u}(t, \boldsymbol{x}) = c(\boldsymbol{x})\hat{\boldsymbol{n}}(\boldsymbol{x})$, where $\hat{\boldsymbol{n}}$ is the normal vector at the interface, that is,

$$\hat{\boldsymbol{n}}(t, \boldsymbol{x}) = \frac{\nabla\phi}{|\nabla\phi|}.$$

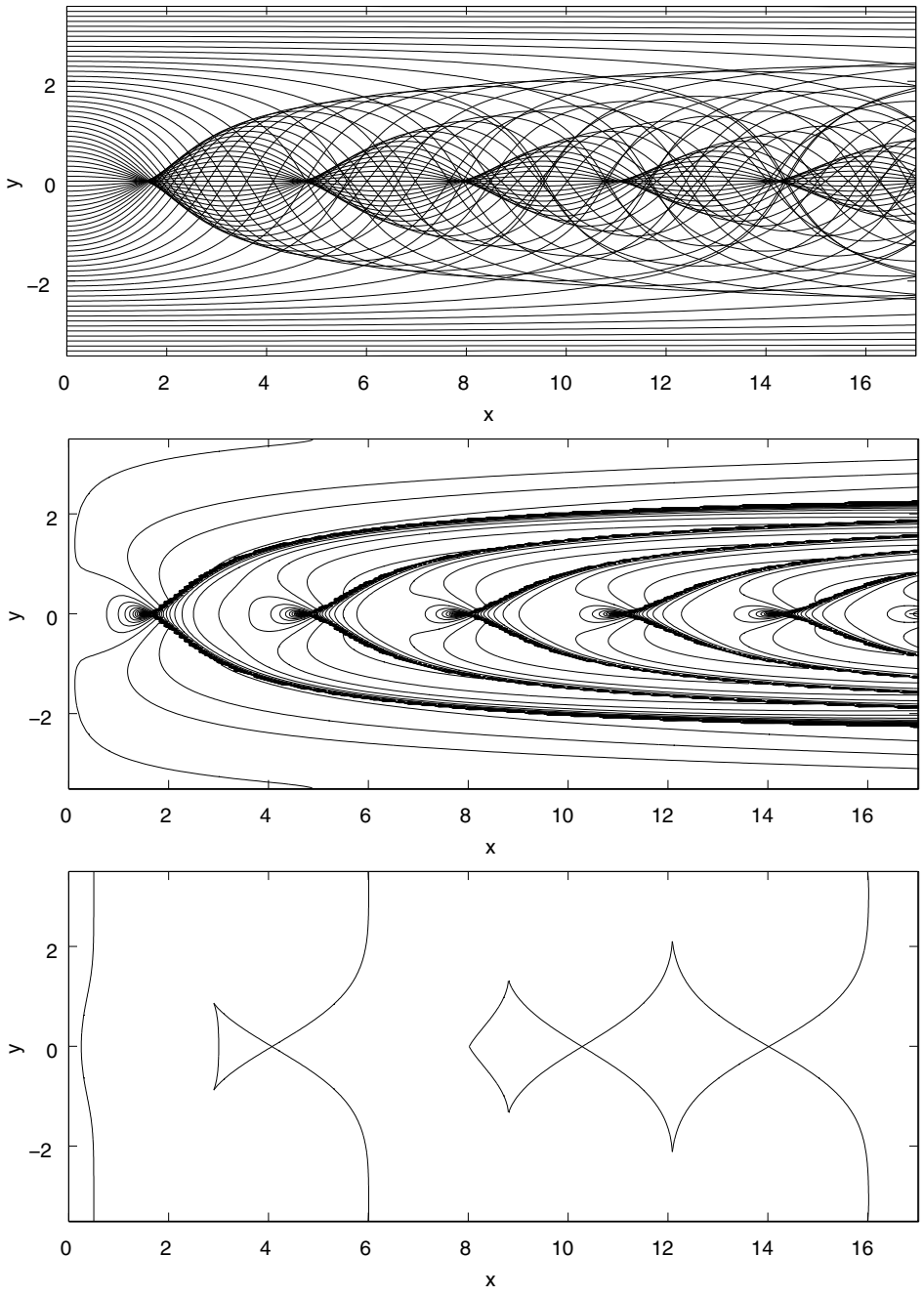


Figure 4.9. Results for the wave guide simulation. The top frame shows rays from initial plane wave. The middle frame shows amplitude, contour lines of $\min(A, 4)$. The bottom frame shows wave fronts in the xy-plane at $T = 0.5$, 6, 16.

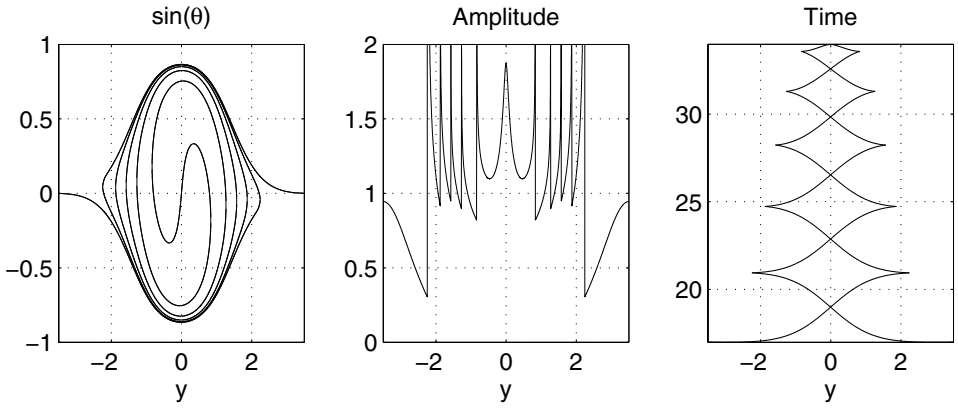


Figure 4.10. Results for wave guide at $x = 17$ as a function of y. At this point there are 11 y-segments and 13 θ-segments. The left frame shows $\sin(\theta)$. The middle frame shows amplitude A. The right frame shows time T.

This results in the eikonal equation,

$$\phi_t + c(\boldsymbol{x})\hat{\boldsymbol{n}} \cdot \nabla\phi = \phi_t + c(\boldsymbol{x})|\nabla\phi| = 0.$$

A direct application will thus clearly not satisfy the linear superposition principle. The method can, however, still be used if we approximate the wave front in phase space and evolve the front using the Liouville equation (2.29), as was done in the segment projection method.

The wave front in the kinetic formulation (2.29) is of higher codimension, and such geometrical objects can be represented by the intersection of interfaces that are given by different level set functions (Osher, Cheng, Kang, Shim and Tsai 2002, Cheng, Osher and Qian 2002). The helix in Figure 4.1 can be defined by the intersection of two regular surfaces: see Figure 4.11. The evolutions of both level set functions are defined by the same velocity vector given by (2.22), (2.23) and (2.24). The advantage of the kinetic formulation is that the superposition principle is valid, and this is also true for the corresponding level set formulation.

A practical problem with this approach is that the evolution of the one-dimensional object representing the wave front requires approximation of the evolution of two level set functions in three dimensions. For wave fronts in $\mathbb{R}^3$, three level set functions in five independent variables and time are required. The computational burden can be reduced by restricting the computation to a small neighbourhood of the wave front. In order to have a well-functioning algorithm, a number of special techniques are useful. Reinitialization of the level set functions ϕ_j, $j = 1, \ldots, d$ in d dimensions should

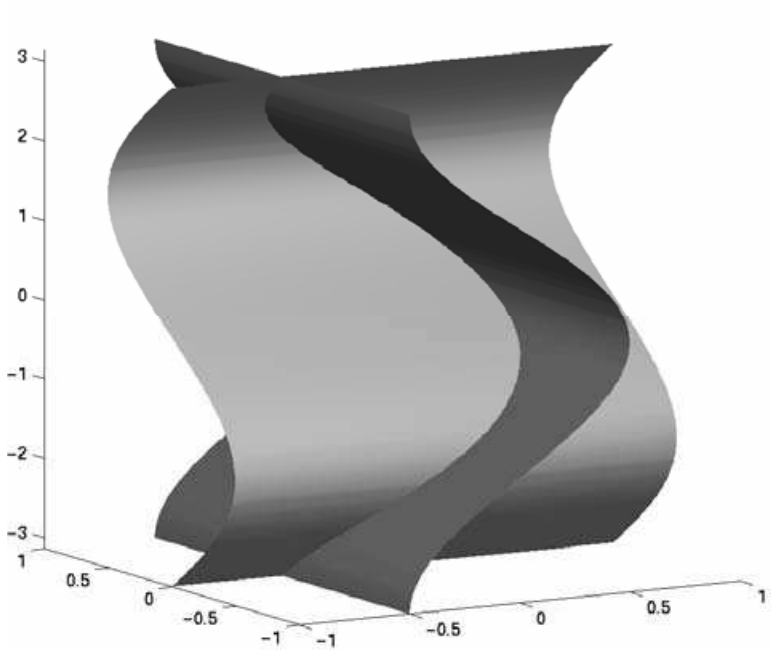


Figure 4.11. Using two level set functions to describe the phase space front. From Osher *et al.* (2002), reproduced with permission.

be performed at regular time intervals, such that

$$|\nabla\phi_j| \approx 1, \qquad \nabla\phi_j \cdot \nabla\phi_k \approx 0, \quad j \neq k,$$

at the interface. In Figure 4.12 we see the result of a level set simulation of a propagating wave front in an inhomogeneous medium simulating high-frequency seismic waves.

4.5. Fast marching in phase space

In Section 3.2 on viscosity solutions (see page 207) the fast marching method was briefly mentioned in the context of viscosity solutions for the frequency domain eikonal equation (2.15). The method can also be applied to a transport equation in phase space, enabling it to capture multivalued solutions. This idea was put forth by Fomel and Sethian (2002).

The transport equation in question (4.15) is an 'escape' equation set in a subdomain Ω of phase space. The unknown, $\hat{\boldsymbol{y}}(\boldsymbol{x}, \boldsymbol{p})$, represents the point on the boundary $\partial\Omega$ (in phase space) where a bicharacteristic originating in $(\boldsymbol{x}, \boldsymbol{p}) \in \Omega$ crosses the boundary. We note that $\hat{\boldsymbol{y}}(\boldsymbol{x}(t), \boldsymbol{p}(t))$ is constant along a bicharacteristic $(\boldsymbol{x}(t), \boldsymbol{p}(t))$. Therefore, after differentiation with respect to t and multiplication by η^2, the chain rule together with the ray

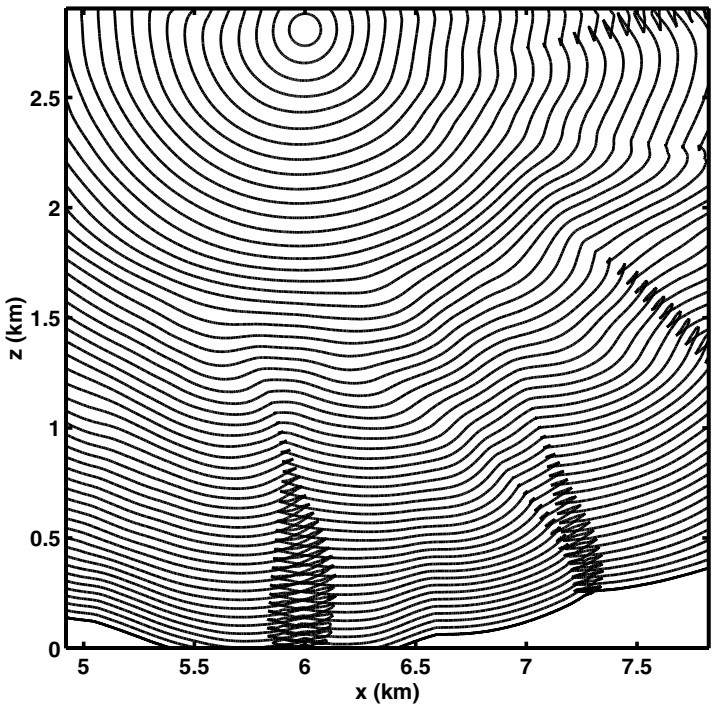


Figure 4.12. Numerical result for Marmousi test problem. From Cheng *et al.* (2002), reproduced with permission.

equations (2.13) and (2.14) give the transport equation

$$\begin{aligned} D_x\hat{\boldsymbol{y}}\ \boldsymbol{p} + \eta D_p\hat{\boldsymbol{y}}\ \nabla\eta = 0, \qquad & (\boldsymbol{x},\boldsymbol{p}) \in \Omega, \\ \hat{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{p}) = (\boldsymbol{x},\boldsymbol{p}), \qquad & (\boldsymbol{x},\boldsymbol{p}) \in \partial\Omega, \end{aligned} \tag{4.15}$$

where $D_x\hat{\boldsymbol{y}}$ and $D_p\hat{\boldsymbol{y}}$ are the Jacobians of $\hat{\boldsymbol{y}}$ with respect to $\boldsymbol{x}$ and $\boldsymbol{p}$, respectively. Note that this is the stationary version of (2.29) with the scalar density function f replaced by the vector $\hat{\boldsymbol{y}}$. There is also an accompanying transport equation for the travel time $T(\boldsymbol{x},\boldsymbol{p})$ from the point $(\boldsymbol{x},\boldsymbol{p})$ to the first boundary crossing. Similarly, since $\partial_t T(\boldsymbol{x}(t),\boldsymbol{p}(t)) = 1$, we get

$$\begin{aligned} \boldsymbol{p}\cdot\nabla_x T + \eta\nabla\eta\cdot\nabla_p T = \eta^2, \qquad & (\boldsymbol{x},\boldsymbol{p}) \in \Omega, \\ T(\boldsymbol{x},\boldsymbol{p}) = 0, \qquad & (\boldsymbol{x},\boldsymbol{p}) \in \partial\Omega. \end{aligned} \tag{4.16}$$

Once the solutions to (4.15) and (4.16) have been found, travel times between any two points $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ can be computed. First, solve

$$\hat{\boldsymbol{y}}(\boldsymbol{x}_0,\boldsymbol{p}_0) = \hat{\boldsymbol{y}}(\boldsymbol{x}_1,\boldsymbol{p}_1) \tag{4.17}$$

for $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$. Then the travel time is $|T(\boldsymbol{x}_0,\boldsymbol{p}_0) - T(\boldsymbol{x}_0,\boldsymbol{p}_1)|$. There may be multiple solutions to (4.17), giving multiple travel times. If $\boldsymbol{x}_1 \in \partial\Omega$, the

expression simplifies. Setting $\hat{\boldsymbol{y}} = (\hat{\boldsymbol{x}}, \hat{\boldsymbol{p}})$, we can solve

$$\hat{\boldsymbol{x}}(\boldsymbol{x}_0, \boldsymbol{p}) = \boldsymbol{x}_1$$

for $\boldsymbol{p}$, to get the travel time $T(\boldsymbol{x}_0, \boldsymbol{p})$. To find the travel time at $\boldsymbol{x}_0$ of a wave front that starts at the boundary of Ω in physical space, we instead need to find $\boldsymbol{p}$ such that

$$\hat{\boldsymbol{p}}(\boldsymbol{x}_0, \boldsymbol{p}) = \eta \hat{\boldsymbol{n}}(\hat{\boldsymbol{x}}(\boldsymbol{x}_0, \boldsymbol{p})),$$

where $\hat{\boldsymbol{n}}(\hat{\boldsymbol{x}})$ is the normal of the boundary at $\hat{\boldsymbol{x}}$. Again, the travel time is $T(\boldsymbol{x}_0, \boldsymbol{p})$.

The amplitude can also be obtained directly through post-processing of the solution. Let us consider a point source at $\boldsymbol{x}_0$ in two dimensions. In the notation of Section 2.2, we have

$$A^{-2}(\tilde{\boldsymbol{x}}(t, r)) \sim |\tilde{\boldsymbol{x}}_r(t, r)| \eta(\tilde{\boldsymbol{x}}(t, r)),$$

after assuming that $\boldsymbol{x}_0(r) = \boldsymbol{x}_0 + \varepsilon(\cos r, \sin r)$ and $\varepsilon \to 0$. Set $\boldsymbol{p}_0(r) = \eta(\boldsymbol{x}_0)(\cos r, \sin r)^T$. Then there is a function $t(r)$ such that $\hat{\boldsymbol{x}}(\boldsymbol{x}_0, \boldsymbol{p}_0(r)) = \tilde{\boldsymbol{x}}(t(r), r)$, and, after differentiation with respect to r,

$$D_p \hat{\boldsymbol{x}}\ \boldsymbol{p}_0^{\perp} = \tilde{\boldsymbol{x}}_t t'(r) + \tilde{\boldsymbol{x}}_r.$$

But $\hat{\boldsymbol{p}}(\boldsymbol{x}_0, \boldsymbol{p}_0) \parallel \tilde{\boldsymbol{x}}_t \perp \tilde{\boldsymbol{x}}_r$, and since $|\hat{\boldsymbol{p}}| = \eta(\hat{\boldsymbol{x}})$,

$$A^{-2}(\hat{\boldsymbol{x}}(\boldsymbol{x}_0, \boldsymbol{p}_0)) \sim \hat{\boldsymbol{p}}^{\perp} D_p \hat{\boldsymbol{x}}\ \boldsymbol{p}_0^{\perp}.$$

The method computes the travel time to the boundary and escape location for rays with all possible starting points and starting directions in the domain Ω, at the expense of solving the transport equations (4.15) and (4.16) set in the full phase space. It does so in an Eulerian framework, on a fixed grid. In two dimensions, the phase space is three-dimensional, and the cost for fast marching is $\mathcal{O}(N^3 \log N)$ when each dimension is discretized with N grid points. The corresponding cost for three-dimensional problems is $\mathcal{O}(N^5 \log N)$. This is expensive if only one set of initial data is of interest. In many applications, however, we are interested in solving a problem with the same index of refraction $\eta(\boldsymbol{x})$ for many different initial data (sources). Examples include the inverse problem in geophysics and the computation of bistatic radar cross sections. Then the cost is competitive: *cf.* the slowness matching method in Section 3.2 on multivalued solutions (see page 216).

5. Moment-based methods

In the kinetic formulation of geometrical optics presented in Section 2.3, we interpret rays as particle trajectories governed by the Hamiltonian system (1.6). We let $f(t, \boldsymbol{x}, \boldsymbol{p}) \geq 0$ be the density of particles in phase space.

It satisfies the Liouville equation

$$f_t + \frac{1}{\eta^2}\boldsymbol{p}\cdot\nabla_x f + \frac{1}{\eta}\nabla_x\eta\cdot\nabla_p f = 0. \tag{5.1}$$

Like kinetic equations in general, solving the full equation (5.1) by direct numerical methods would be very expensive, because of the large number of independent variables (six in 3D). Instead we use the classic technique of approximating a kinetic transport equation set in high-dimensional phase space $(t, \boldsymbol{x}, \boldsymbol{p})$, by a finite system of moment equations in the reduced space $(t, \boldsymbol{x})$. See, for instance, Grad (1949) and more recently Levermore (1996). In general the moment equations form a system of conservation laws that gives an approximation of the true solution. The classical example is the compressible Euler approximation of the Boltzmann equation (see remark below). In our setting, the moment system is, however, typically *exact* under the closure assumption that at most N rays cross at any given point in time and space. In fact, this moment system solution is equivalent to N disjoint pairs of eikonal and transport equations (2.5) and (2.6) when the solution is smooth.

Brenier and Corrias (1998) originally proposed this approach for finding multivalued solutions to geometrical optics problems in the one-dimensional homogeneous case. It was subsequently adapted for two-dimensional inhomogeneous problems by Engquist and Runborg (Engquist and Runborg 1996, 1998, Runborg 2000). See also Gosse (2002). More recently, the same technique has been applied to the Schrödinger equation by Jin and Li (200x), Gosse, Jin and Li (200x), and Sparber *et al.* (2003).

In this section we derive and analyse the system of PDEs that follows from the kinetic model in two dimensions, together with the closure assumption that a maximum of N rays passes through any given point in space and time. We consider two different versions of the closure assumption and present some numerical examples.

5.1. Moment equations

We start by defining the moments m_{ij}, with $\boldsymbol{p} = (p_1, p_2)^T$, as

$$m_{ij}(t,\boldsymbol{x}) = \frac{1}{\eta(\boldsymbol{x})^{i+j}}\int_{\mathbb{R}^2} p_1^i p_2^j f(t,\boldsymbol{x},\boldsymbol{p})\,\mathrm{d}\boldsymbol{p}. \tag{5.2}$$

Next, multiply (5.1) by $\eta^{2-i-j}p_1^i p_2^j$ and integrate over $\mathbb{R}^2$ with respect to $\boldsymbol{p}$, so that

$$\begin{aligned}\frac{\eta^2}{\eta^{i+j}}\partial_t \int p_1^i p_2^j f\,\mathrm{d}\boldsymbol{p}&\\ + \partial_x \frac{1}{\eta^{i+j}}\int p_1^{i+1}p_2^j f\,\mathrm{d}\boldsymbol{p} &+ \frac{(i+j)\eta_x}{\eta^{i+j+1}}\int p_1^{i+1}p_2^j f\,\mathrm{d}\boldsymbol{p}\end{aligned} \tag{5.3}$$

$$+\partial_y \frac{1}{\eta^{i+j}} \int p_1^i p_2^{j+1} f \,\mathrm{d}\boldsymbol{p} + \frac{(i+j)\eta_y}{\eta^{i+j+1}} \int p_1^i p_2^{j+1} f \,\mathrm{d}\boldsymbol{p}$$
$$- \frac{\eta_x}{\eta^{i+j-1}} \int i p_1^{i-1} p_2^j f \,\mathrm{d}\boldsymbol{p} - \frac{\eta_y}{\eta^{i+j-1}} \int j p_1^i p_2^{j-1} f \,\mathrm{d}\boldsymbol{p} = 0,$$

after using integration by parts and the fact that f has compact support in $\boldsymbol{p}$ for the last term. From definition (5.2) we see that formally m_{ij} will satisfy the infinite system of moment equations

$$(\eta^2 m_{ij})_t + (\eta m_{i+1,j})_x + (\eta m_{i,j+1})_y = \\ i\eta_x m_{i-1,j} + j\eta_y m_{i,j-1} - (i+j)(\eta_x m_{i+1,j} + \eta_y m_{i,j+1}), \tag{5.4}$$

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 (5.4) 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, in Section 5.2 we consider the case when f is a weighted sum of delta functions in $\boldsymbol{p}$,

$$f(t, \boldsymbol{x}, \boldsymbol{p}) = \sum_{k=1}^{N} g_k \cdot \delta(\boldsymbol{p} - \boldsymbol{p}_k), \qquad \boldsymbol{p}_k = \eta \begin{pmatrix} \cos\theta_k \\ \sin\theta_k \end{pmatrix}, \tag{5.5}$$

and second, in Section 5.3, the case when f is a sum of Heaviside functions on the sphere in phase space,

$$f(t, \boldsymbol{x}, \boldsymbol{p}) = \frac{1}{\eta}\delta(|\boldsymbol{p}| - \eta) \sum_{k=1}^{N} (-1)^{k+1} H(\theta - \theta_k), \qquad \boldsymbol{p} = |\boldsymbol{p}| \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \tag{5.6}$$

where H is the Heaviside function. Both cases correspond to the assumption of a finite number of rays at each point in time and space. Since f here only depends on a finite number of unknowns ($2N$ for (5.5) and N for (5.6)) the infinite system (5.4) can be reduced to a finite system.

Remark. The derivation above of the multiphase equations for geometrical optics is completely analogous to the derivation of the hydrodynamical limit from a kinetic formulation of gas dynamics. Instead of (5.1), we use the Boltzmann equation

$$f_t + \boldsymbol{p} \cdot \nabla_x f = Q(f, f), \qquad (\boldsymbol{x}, \boldsymbol{p}) \in \mathbb{R}^2 \times \mathbb{R}^2, \tag{5.7}$$

where $Q(f, f)$ is the collision operator. Moreover, instead of the closure assumptions (5.5) and (5.6), we assume f is a *Maxwellian*, that is,

$$f(t, \boldsymbol{x}, \boldsymbol{p}) = \frac{\rho(t, \boldsymbol{x})}{2\pi T(t, \boldsymbol{x})} \exp\left(-\frac{|\boldsymbol{p} - \boldsymbol{u}(t, \boldsymbol{x})|^2}{2T(t, \boldsymbol{x})} \right), \qquad \boldsymbol{u} = \begin{pmatrix} u \\ v \end{pmatrix}. \tag{5.8}$$

The lowest moments of Q are zero by its special form, and therefore the first moment equations of (5.7) are the same as those of (5.1) with $\eta \equiv 1$. If we pick the following equations from (5.4),

$$\begin{pmatrix} m_{00} \\ m_{10} \\ m_{01} \\ m_{20}+m_{02} \end{pmatrix}_t + \begin{pmatrix} m_{10} \\ m_{20} \\ m_{11} \\ m_{30}+m_{12} \end{pmatrix}_x + \begin{pmatrix} m_{01} \\ m_{11} \\ m_{02} \\ m_{21}+m_{03} \end{pmatrix}_y = 0, \tag{5.9}$$

and write them in terms of the unknowns $\rho, \boldsymbol{u}, T$ in (5.8) and

$$E \equiv \rho\left(\frac{1}{2}|\boldsymbol{u}|^2 + \frac{3}{2}T\right),$$

we get

$$\begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix}_t + \begin{pmatrix} \rho u \\ \rho(u^2+T) \\ \rho uv \\ u(E+\rho T) \end{pmatrix}_x + \begin{pmatrix} \rho v \\ \rho uv \\ \rho(v^2+T) \\ v(E+\rho T) \end{pmatrix}_y = 0, \tag{5.10}$$

the compressible Euler equations for a perfect monoatomic gas.

5.2. Closure with delta functions

To close (5.4) we assume in this section that f can be written as

$$f(t,\boldsymbol{x},\boldsymbol{p}) = \sum_{k=1}^{N} g_k \cdot \delta(\boldsymbol{p}-\boldsymbol{p}_k), \qquad \boldsymbol{p}_k = \eta\begin{pmatrix}\cos\theta_k \\ \sin\theta_k\end{pmatrix}. \tag{5.11}$$

Hence, for fixed values of $\boldsymbol{x}$ and t, the particle density f is nonzero at a maximum of N points, and only when $|\boldsymbol{p}| = \eta(\boldsymbol{x})$. The new variables that we have introduced here are $g_k = g_k(t,\boldsymbol{x})$, which corresponds to the strength (particle density) of ray k, and $\theta_k = \theta_k(t,\boldsymbol{x})$, which is the direction of the same ray. Inserting (5.11) into the definition of the moments (5.2) yields

$$m_{ij} = \sum_{k=1}^{N} g_k \cos^i\theta_k \sin^j\theta_k. \tag{5.12}$$

A system describing N phases needs $2N$ equations, corresponding to the N ray strengths g_k and their directions θ_k. It is not immediately clear which equations to select among the candidates in (5.4). Given the equations for a set of $2N$ moments, we 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 θ cannot be recovered, and therefore, in general, the sign of the moments cannot be determined. Here we choose the equations for the

moments $m_{2\ell-1,0}$ and $m_{0,2\ell-1}$

$$\begin{aligned}(\eta^2 m_{2\ell-1,0})_t + (\eta m_{2\ell,0})_x + (\eta m_{2\ell-1,1})_y = \\ (2\ell-1)(\eta_x m_{2\ell-2,0} - \eta_x m_{2\ell,0} - \eta_y m_{2\ell-1,1}),\\ (\eta^2 m_{0,2\ell-1})_t + (\eta m_{1,2\ell-1})_x + (\eta m_{0,2\ell})_y = \\ (2\ell-1)(\eta_y m_{0,2\ell-2} - \eta_x m_{1,2\ell-1} - \eta_y m_{0,2\ell}),\end{aligned} \tag{5.13}$$

for $\ell = 1, \ldots, N$, and we collect these moments in a vector,

$$\boldsymbol{m} = (m_{10}, m_{01}, m_{30}, m_{03}, \ldots, m_{2N-1,0}, m_{0,2N-1})^T. \tag{5.14}$$

As we will show below, this system of equations for $\boldsymbol{m}$ can be essentially closed, for all N, meaning that, for almost all $\boldsymbol{m}$, we can uniquely determine the remaining moments in (5.13). We introduce new variables,

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{2N-1} \\ u_{2N} \end{pmatrix} := \begin{pmatrix} g_1 \cos\theta_1 \\ g_1 \sin\theta_1 \\ \vdots \\ g_N \cos\theta_N \\ g_N \sin\theta_N \end{pmatrix}, \tag{5.15}$$

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 (5.12) define a function $\boldsymbol{F}_0$ through the equation

$$\boldsymbol{F}_0(\boldsymbol{u}) = \boldsymbol{m}. \tag{5.16}$$

Similarly, they define the functions

$$\begin{aligned}\boldsymbol{F}_1(\boldsymbol{u}) &= (m_{20}, m_{11}, \ldots, m_{2N,0}, m_{1,2N-1})^T,\\ \boldsymbol{F}_2(\boldsymbol{u}) &= (m_{11}, m_{02}, \ldots, m_{2N-1,1}, m_{0,2N})^T,\end{aligned} \tag{5.17}$$

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

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

$$\boldsymbol{F}_0(\eta^2 \boldsymbol{u})_t + \boldsymbol{F}_1(\eta \boldsymbol{u})_x + \boldsymbol{F}_2(\eta \boldsymbol{u})_y = \boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y). \tag{5.18}$$

Equivalently, we can write (5.18) as

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

The functions $\boldsymbol{F}_j$ and $\boldsymbol{K}$ are rather complicated nonlinear functions. In the most simple case, $N = 1$, the function $\boldsymbol{F}_0$ is the identity, and

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

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

$$\boldsymbol{f}_0 = \begin{pmatrix} w_1 \\ w_2 \\ w_1^3/|\boldsymbol{w}|^2 \\ w_2^3/|\boldsymbol{w}|^2 \end{pmatrix}, \qquad \boldsymbol{f}_1 = \frac{w_1}{|\boldsymbol{w}|}\boldsymbol{f}_0, \qquad \boldsymbol{f}_2 = \frac{w_2}{|\boldsymbol{w}|}\boldsymbol{f}_0,$$

$$\boldsymbol{k} = \frac{\eta_x w_2 - \eta_y w_1}{|\boldsymbol{w}|}\begin{pmatrix} w_2 \\ -w_1 \\ w_1^2 w_2/|\boldsymbol{w}|^2 \\ -w_1 w_2^2/|\boldsymbol{w}|^2 \end{pmatrix}.$$

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

Since the angles θ_k remain unaffected when $\boldsymbol{u}$ is scaled by a constant for all N, the $\boldsymbol{F}_j$ and $\boldsymbol{K}$ are always homogeneous of degree one, $\boldsymbol{F}_j(\alpha\boldsymbol{u}) = \alpha\boldsymbol{F}_j(\boldsymbol{u})$, $\boldsymbol{K}(\alpha\boldsymbol{u}, \eta_x, \eta_y) = \alpha\boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y)$ for all $\alpha \in \mathbb{R}$. Moreover, the source term $\boldsymbol{K}$ always vanishes for constant η.

Properties of the flux functions

In this section we analyse the flux functions and source

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

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 the following result.

Theorem 5.1. Let $\boldsymbol{F}_0$ be the function in (5.16) and let $\boldsymbol{F}_0|U_N$ be its restriction to the domain

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

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

$$m : U_N \to \mathbb{R}, \qquad m(\boldsymbol{u}) = \sum_{k=1}^{N} g_k h(\theta_k), \tag{5.19}$$

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

Since $\boldsymbol{F}_1$, $\boldsymbol{F}_2$ and $\boldsymbol{K}$ are all of the form (5.19) we have the following result.

Corollary 5.2. Let $\boldsymbol{F}_j$ and $\boldsymbol{K}$ be the functions in (5.16) and (5.17) and let $\boldsymbol{F}_0|U_N$ and M_N be as in Theorem 5.1. Then the functions

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

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

Remark. If we do not restrict $\boldsymbol{F}_0$ to U_N the result is false. Take, for instance, $\boldsymbol{u} = (-1\ 0\ 1\ 0)^T$ and $\tilde{\boldsymbol{u}} = 2\boldsymbol{u}$ for $N = 2$ so that $\boldsymbol{F}_0(\boldsymbol{u}) = \boldsymbol{F}_0(\tilde{\boldsymbol{u}}) = 0$, but $\boldsymbol{F}_1(\tilde{\boldsymbol{u}}) = 2\boldsymbol{F}_1(\boldsymbol{u}) \neq 0$. Furthermore, with a different choice of moment equations the result does not necessarily hold either. For instance, if instead of (5.14) we use the equations for

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

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

Proof of Theorem 5.1. It will be convenient to work with complex versions of our variables, and we start by introducing the isometry $\mathcal{A} : \mathbb{R}^{2N} \to \mathbb{C}^N$,

$$\mathcal{A}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{2N} \end{pmatrix} = \begin{pmatrix} x_1 + \mathrm{i}x_2 \\ \vdots \\ x_{2N-1} + \mathrm{i}x_{2N} \end{pmatrix},$$

identifying $\mathbb{R}^{2N}$ with $\mathbb{C}^N$. We set $\boldsymbol{w} = (w_1, \ldots, w_N)^T := \mathcal{A}\boldsymbol{u}$ and

$$z_k := \cos\theta_k + \mathrm{i}\sin\theta_k, \qquad \boldsymbol{z}_k := \left(z_k, z_k^{-3}, \ldots, z_k^{(2N-1)(-1)^{N+1}}\right)^T, \tag{5.20}$$

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

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

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

$$\boldsymbol{z}_k = B \begin{pmatrix} \cos\theta_k + \mathrm{i}\sin\theta_k \\ \vdots \\ \cos^{2N-1}\theta_k + \mathrm{i}\sin^{2N-1}\theta_k \end{pmatrix}, \tag{5.21}$$

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 $k \le \ell$. The matrix is nonsingular since $b_{kk} = 4^{k-1} > 0$. From the definition of $\boldsymbol{Q}$ and the identity (5.21) it then follows that

$$\boldsymbol{Q}(\boldsymbol{w}) = \boldsymbol{Q}(\mathcal{A}\boldsymbol{u}) = B\mathcal{A}\boldsymbol{m}, \tag{5.22}$$

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

Lemma 5.3. Let $\{z_k\}$ be N' complex numbers such that $|z_k| = 1$, and let $\{\boldsymbol{z}_k\}$ be the corresponding vectors as defined in (5.20). If $N' \le 2N$ then $\boldsymbol{z}_k \in C^N$ are linearly independent over $\mathbb{R}$ if and only if

$$z_k^2 \ne z_\ell^2, \qquad k \ne \ell. \tag{5.23}$$

Proof. The necessity is obvious. To show that (5.23) 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 (5.23) still holds if $N' < 2N$. Suppose therefore that $\{\boldsymbol{z}_k\}_{k=1}^{2N}$ are linearly dependent over $\mathbb{R}$, and that (5.23) is true. Then the real matrix

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

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

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

where

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

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

Let $\bar{m}(\boldsymbol{w}) := m(\mathcal{A}^{-1}\boldsymbol{w})$ and let $\bar{\boldsymbol{Q}}$ be the restriction of $\boldsymbol{Q}$ to $\mathcal{A}U_N$. We now want to prove that $\bar{m} \circ \bar{\boldsymbol{Q}}^{-1}$ is well defined on $\bar{\boldsymbol{Q}}(\mathcal{A}U_N)$, and we do this by showing that $\bar{\boldsymbol{Q}} \circ \bar{m}^{-1}$ is injective on $\bar{m}(\mathcal{A}U_N)$. Let $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \mathcal{A}U_N$ be such that $\bar{\boldsymbol{Q}}(\boldsymbol{w}) = \bar{\boldsymbol{Q}}(\tilde{\boldsymbol{w}})$. We need to show that $\bar{m}(\boldsymbol{w}) = \bar{m}(\tilde{\boldsymbol{w}})$ and we use the variables introduced in (5.20). A tilde indicates that a variable relates to $\tilde{\boldsymbol{w}}$. 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 $z_{\ell_j} = \cdots = z_{\ell_{j+1}-1}$, with $1 = \ell_1 < \cdots < \ell_{N'+1} = N+1$, and similarly

for $\{\tilde{z}_k\}$. With this notation we get

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

The sets of numbers $\{z_{\ell_j}\}_{j=1}^{N'}$ and $\{\tilde{z}_{\tilde{\ell}_j}\}_{j=1}^{\tilde{N}'}$ both satisfy (5.23), because $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \mathcal{A}U_N$. 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 5.3. By induction it follows that $N' = \tilde{N}'$ and, possibly after some reordering,

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

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

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

Hence, $\bar{m} \circ \bar{\boldsymbol{Q}}^{-1}$ is well defined on its domain of definition. Now, (5.22) and the fact that $\boldsymbol{F}_0(\boldsymbol{u}) = \boldsymbol{m}$ show that $m \circ (\boldsymbol{F}_0|U_N)^{-1}(\boldsymbol{m}) = \bar{m} \circ \bar{\boldsymbol{Q}}^{-1}(B\mathcal{A}\boldsymbol{m})$, which implies that $m \circ (\boldsymbol{F}_0|U_N)^{-1}$ is well defined on M_N. The continuity follows by approximating U_N by compact sets, and using the following lemma from elementary analysis.

Lemma 5.4. Let K be a compact metric space and suppose $f : K \to X$ and $g : K \to Y$ are continuous functions, and $X = f(K), Y$ are metric spaces. If the composition $f \circ g^{-1} : g(U) \to X$ is injective, then $g \circ f^{-1} : X \to Y$ is continuous. (The function inverses should be interpreted as set functions here.)

Proof. We need to show that, for any open set $U \subset Y$, the set $f \circ g^{-1}(U)$ is open. Since g is continuous and K compact, $g^{-1}(U)^c$ is compact, and consequently $f(g^{-1}(U)^c)$ is also compact by the continuity of f. But $g^{-1}(U^c) = g^{-1}(U)^c$, and hence $f \circ g^{-1}(U^c)$ is compact. Moreover,

$$f \circ g^{-1}(U) \cup f \circ g^{-1}(U^c) = f(g^{-1}(U)) \cup f(g^{-1}(U)^c) = f(K) = X.$$

Since $f \circ g^{-1}$ is injective, $f \circ g^{-1}(U) \cap f \circ g^{-1}(U^c) = \emptyset$, and therefore $f \circ g^{-1}(U) = (f \circ g^{-1}(U^c))^c$, which is open. □

Analysis of the conservation laws

Engquist and Runborg (1996, 1998) showed that the general system (5.18) is nonstrictly hyperbolic for all states $\boldsymbol{u}$ and N. Jin and Li (200x) showed the same for the Schrödinger equation case. 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 Zheng (1998).

A distinguishing feature of the system (5.18) 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 (5.18) describes. If initial data dictate a physical solution with M phases for $t > T$, the system (5.18) with $N < M$ phases will have a measure solution for $t > T$; *cf.* Figure 5.3(a, b) and Figure 5.5(a).

For smooth solutions, (5.18) with N phases is equivalent to N pairs of eikonal and transport equations (2.5) and (2.6) if the variables are identified as

$$g_k = A_{0,k}^2, \qquad \begin{pmatrix} \cos\theta_k \\ \sin\theta_k \end{pmatrix} = \frac{\nabla\phi_k}{|\nabla\phi_k|}, \qquad k = 1, \ldots, N,$$

(Engquist and Runborg 1996). Note that this is expected from the relationship between equations (2.30) and (2.31) and the remark thereafter. The pair (2.5) and (2.6) form a nonstrictly hyperbolic system, just like (5.18), with the same eigenvalue. Where wave fields meet, the viscosity solution of (2.5) is in general discontinuous. Because of the term $\Delta\phi$ in the source term of (2.6), the first amplitude coefficient A_0 has a concentration of mass at these points. Hence, the two different formulations are also similar in this respect.

There is a close relationship between (5.18) with $N = 1$ and $\eta \equiv 1$,

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x + \boldsymbol{g}(\boldsymbol{u})_y = 0, \qquad \boldsymbol{f}(\boldsymbol{u}) = u_1 \frac{\boldsymbol{u}}{|\boldsymbol{u}|}, \qquad \boldsymbol{g}(\boldsymbol{u}) = u_2 \frac{\boldsymbol{u}}{|\boldsymbol{u}|}, \tag{5.24}$$

and the equations of pressureless gases:

$$\rho_t + (\rho u)_x = 0, \qquad (\rho u)_t + (\rho u^2)_x = 0. \tag{5.25}$$

Indeed, the steady state version of (5.24) is precisely (5.25) if we identify $\rho = g\cos^2\theta$ and $u = \tan\theta$. Moreover, the one-dimensional version of (5.24) corresponds to (5.25), with relativistic effects added if we identify $\rho = g\sin\theta$ and $u = \cos\theta$. We also note that, if we formally let $T \to 0$ in (5.8),

we recover (5.11) with $N = 1$ and without the restriction on $|\boldsymbol{p}_1|$. The same formal limit of (5.10) gives the two-dimensional pressureless gas equations.

In the context of non-relativistic pressureless gases, this problem was addressed by Bouchut (1994) and later Brenier and Grenier (Grenier 1995, Brenier and Grenier 1998), and E, Rykov and Sinai (1996), who independently proved global existence of measure solutions to (5.25). The uniqueness question was settled in Bouchut and James (1999). For linear transport equations, related results have been obtained by Bouchut and James (1995) and Poupaud and Rascle (1997). The questions of existence and uniqueness for (5.24) and its one-dimensional version are still open.

The Riemann problem. Since standard numerical schemes are based on solving one-dimensional Riemann problems (LeVeque 1992), we consider this problem for (5.24):

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \qquad \boldsymbol{f}(\boldsymbol{u}) = u_1 \frac{\boldsymbol{u}}{|\boldsymbol{u}|}, \qquad \boldsymbol{u}(0,x) = \begin{cases} \boldsymbol{u}_\ell & x < 0, \\ \boldsymbol{u}_r & x > 0. \end{cases} \tag{5.26}$$

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

$$\boldsymbol{f}(\boldsymbol{u}_\ell) - \boldsymbol{f}(\boldsymbol{u}_r) = s(\boldsymbol{u}_\ell - \boldsymbol{u}_r), \tag{5.27}$$

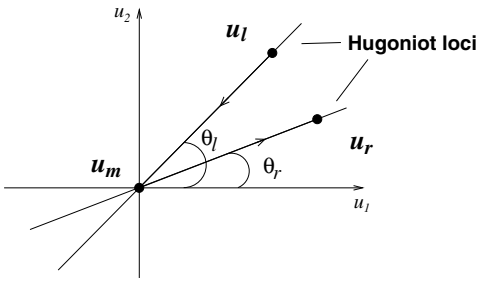
where s represents the propagation speed of the discontinuity. Since $\boldsymbol{f}(\boldsymbol{u}) = \cos\theta \boldsymbol{u}$, the jump condition (5.27) simplifies to

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

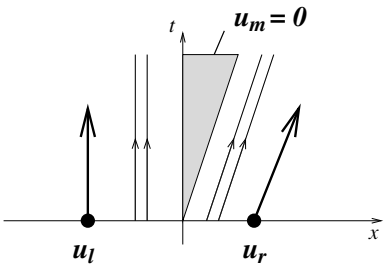
The states to which a given nonzero state $\boldsymbol{u}_\ell$ can connect with a discontinuity, *i.e.*, its Hugoniot locus, is simply $\alpha \boldsymbol{u}_\ell$ for $\alpha \in \mathbb{R}$, with speed of propagation $s = \cos\theta_\ell$ when $\alpha \geq 0$ and $s = \cos\theta_\ell (1+\alpha)/(1-\alpha)$ for $\alpha < 0$. It follows that, unless they are parallel, two nonzero states $\boldsymbol{u}_\ell$ and $\boldsymbol{u}_r$ can only be connected via the intermediate state $\boldsymbol{u}_m = 0$. There will be two types of discontinuity. If $\cos\theta_\ell < \cos\theta_r$, the solution with $\boldsymbol{u}_m = 0$, satisfies the Lax entropy condition (the left discontinuity moves more slowly than the right one). The states' Hugoniot loci and the solution for this type of discontinuity is illustrated in Figure 5.1(a). 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, $\boldsymbol{u}_m = t\tilde{\boldsymbol{u}}_m \delta(x - st)$ is a weak solution to the conservation law with these initial data. The conservation form gives a slightly modified jump condition,

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

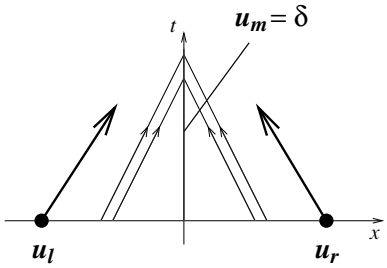
with the propagation speed $s = \cos\tilde{\theta}_m$. This construction, a delta function solution to the Riemann problem leading to a modified Rankine–Hugoniot condition, is found also in Zheng (1998) for more general equations.



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



(b) Contact discontinuity



(c) Overcompressive shock

Figure 5.1. The Riemann problem, with Hugoniot loci for the left and right states in phase space and the two different types of discontinuity in $(t, \boldsymbol{x})$ space.

It is easily verified that $\boldsymbol{u}$ itself is an eigenvector of the Jacobian of $\boldsymbol{f}$ and that the Jacobian has a double eigenvalue equalling $\cos\theta$. 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 $(t, \boldsymbol{x})$-space, are shown in Figure 5.1(b)–(c).

Entropy. For the analysis of (5.24) 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. However, there do

exist nonstrictly convex entropy pairs, which can be characterized as follows (Runborg 2000).

Theorem 5.5. Let $U \in C^2$ be convex. There exists a function $F \in C^2$ such that $U(\boldsymbol{u})_t + F(\boldsymbol{u})_x = 0$ for all smooth solutions $\boldsymbol{u} = g(\cos\theta, \sin\theta)$ to

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \qquad \boldsymbol{f}(\boldsymbol{u}) = u_1 \frac{\boldsymbol{u}}{|\boldsymbol{u}|}, \tag{5.28}$$

if and only if U is of the form

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

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, have discontinuities in g. If we introduce weak solutions, we can show that a sufficient condition for the superposition principle to hold is just that g is bounded and that θ is continuous and has locally bounded variation. A discontinuous θ would typically not be physical, generating a delta shock-type solution, as seen above. We have the following result.

Theorem 5.6. Suppose $\{\boldsymbol{u}_k\}_{k=1}^N$ are N weak solutions to the homogeneous single phase system (5.24) in the sense that $\boldsymbol{u}_k \in L^\infty((0,\infty)\times\mathbb{R}^2)$ and

$$\iint_{t\geq 0} \boldsymbol{u}_k\phi_t + \boldsymbol{f}(\boldsymbol{u}_k)\phi_x + \boldsymbol{g}(\boldsymbol{u}_k)\phi_y \,\mathrm{d}t\,\mathrm{d}\boldsymbol{x} = 0, \qquad \forall \phi \in C_c^1((0,\infty)\times\mathbb{R}^2). \tag{5.29}$$

Moreover, suppose that, for each k and each point in $(0,\infty)\times\mathbb{R}^2$, there is an open neighbourhood on which we can define a continuous function $\theta_k(t,\boldsymbol{x})$ with locally bounded variation such that $\boldsymbol{u}_k = |\boldsymbol{u}_k|(\cos\theta_k, \sin\theta_k)^T$ on that neighbourhood. Then $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)^T$ is a weak solution to the homogeneous N-phase system (5.18) in the same sense as (5.29).

Proof. We start by showing that, if $\boldsymbol{v} = (v_1, v_2)^T$ is a weak solution to (5.24) in the sense of Theorem 5.6, 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 θ is continuous and that $\boldsymbol{v} = g(\cos\theta, \sin\theta)^T$ on supp ϕ. (We can always obtain such a θ after a partition of unity.) Let $M \in C_c^\infty(\mathbb{R}^3)$ be a mollifier with $\int M \,\mathrm{d}t\,\mathrm{d}\boldsymbol{x} = 1$ and set $\theta_\epsilon = \theta \star M_\epsilon$, where $M_\epsilon = M(t/\epsilon, \boldsymbol{x}/\epsilon)/\epsilon^3$. Furthermore, set

$$\psi_s^\epsilon = \phi\left(\cos^{i-1}\theta_\epsilon - \frac{\mathrm{d}\cos^{i-1}\theta_\epsilon}{\mathrm{d}\theta_\epsilon}\sin\theta_\epsilon\cos\theta_\epsilon\right), \qquad \psi_c^\epsilon = \phi\frac{\mathrm{d}\cos^{i-1}\theta_\epsilon}{\mathrm{d}\theta_\epsilon}\cos^2\theta_\epsilon. \tag{5.30}$$

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

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

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

$$\begin{aligned}\iint_{t\geq 0} |v_1^\epsilon - v_1|\,|(\psi_s^\epsilon)_t| \,\mathrm{d}t\,\mathrm{d}\boldsymbol{x} &\leq C \sup_{(t,\boldsymbol{x})\in \mathrm{supp}\ \phi} |v_1^\epsilon - v_1| \\ &\leq C\|\boldsymbol{v}\|_{L^\infty} \sup_{(t,\boldsymbol{x})\in\mathrm{supp}\ \phi} |\cos\theta_\epsilon - \cos\theta| \to 0,\end{aligned}$$

by the continuity of θ. Using the same argument for the remaining terms, we arrive at

$$\begin{aligned}&\iint_{t\geq 0} m_{i,0}\phi_t + m_{i+1,0}\phi_x + m_{i,1}\phi_y \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \qquad (5.31)\\ &\quad= \iint_{t\geq 0} v_1(\psi_s^\epsilon)_t + \frac{v_1^2}{|\boldsymbol{v}|}(\psi_s^\epsilon)_x + \frac{v_1v_2}{|\boldsymbol{v}|}(\psi_s^\epsilon)_y \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\ &\qquad+ \iint_{t\geq 0} v_2(\psi_c^\epsilon)_t + \frac{v_2v_1}{|\boldsymbol{v}|}(\psi_c^\epsilon)_x + \frac{v_2^2}{|\boldsymbol{v}|}(\psi_c^\epsilon)_y \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + R^\epsilon,\end{aligned}$$

where $R^\epsilon \to 0$. But $\psi_c^\epsilon, \psi_s^\epsilon \in C_c^1((0,\infty)\times\mathbb{R}^2)$ and $\boldsymbol{v}$ is a weak solution, so by letting $\epsilon\to 0$ we see that (5.31) in fact equals zero. After replacing $\cos^{i-1}\theta$ with $\sin^{i-1}\theta$ in (5.30) we get the same result for $m_{0,i}$. We can now conclude that, with $m_{i,j}^k = g_k\cos^i\theta_k\sin^j\theta_k$,

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

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

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

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

Of course, some of the $\boldsymbol{u}_k$ solutions in Theorem 5.6 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.

5.3. Closure with Heaviside functions

We will now consider a different way to close (5.4). We discard the amplitude information carried by g_k used in Section 5.2, and only solve for the angles θ_k. In this way we get fewer and less singular equations. The 'correct' values of the unknowns θ_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 5.4). In order to reduce the initialization problem we make the paraxial approximation discussed in Section 2.5, assuming that no rays go in the negative x-direction. This means that data only need to be given on the line $x = 0$. We then consider density functions of the form

$$f(t,\boldsymbol{x},\boldsymbol{p}) = \frac{1}{\eta(\boldsymbol{x})}\delta(|\boldsymbol{p}|-\eta(\boldsymbol{x}))\sum_{k=1}^{N}(-1)^{k+1}H(\theta-\theta_k(t,\boldsymbol{x})), \quad \boldsymbol{p} = |\boldsymbol{p}|\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}. \tag{5.32}$$

For fixed $(t,\boldsymbol{x})$, the density function f is supported by a set of intervals on the sphere $\{|\boldsymbol{p}| = \eta\}$. The intervals correspond to fans of rays whose edges are given by the unknown angles θ_k. The transport equation (5.1) 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 (5.32). The values of the N angles θ_k will then coincide with those of a problem with N rays crossing at each point, as long as the assumption (5.32) holds.

The paraxial approximation amounts to the additional assumption that $f(t,\boldsymbol{x},\boldsymbol{p}) = 0$ when $\boldsymbol{p}\cdot\boldsymbol{e}_x \le 0$, where $\boldsymbol{e}_x$ is the unit vector in the x-direction, and that the boundary data at $x = 0$ is time-independent. We also adopt the convention that $-\pi/2 < \theta_1 \le \cdots \le \theta_N < \pi/2$. The general formula for the moments then follows from (5.32) together with (5.2), namely

$$m_{ij}(t,\boldsymbol{x}) = \sum_{k=1}^{N}(-1)^{k+1}\int_{\theta_k(t,\boldsymbol{x})}^{\pi/2}\cos^i\theta\sin^j\theta\,\mathrm{d}\theta. \tag{5.33}$$

Among the equations in (5.4) we choose the ones for the moments $\{m_{0,\ell}\}$ with $\ell = 0,\dots,N-1$. By the paraxial approximation, this leads to the steady state equations

$$(\eta m_{1,\ell})_x + (\eta m_{0,\ell+1})_y = \ell(\eta_y m_{0,\ell-1} - \eta_x m_{1,\ell} - \eta_y m_{0,\ell+1}), \qquad \ell = 0,\dots,N-1. \tag{5.34}$$

Next, we introduce the new variables:

$$\boldsymbol{u} = (u_1,\dots,u_N)^T, \qquad u_k = \sin\theta_k. \tag{5.35}$$

By evaluating the integrals in (5.33), we then get, for N even,

$$m_{1,\ell} = \sum_{k=1}^{N} \frac{(-1)^k u_k^{\ell+1}}{\ell+1}, \qquad m_{0,\ell} = \sum_{k=1}^{N} (-1)^k R_\ell(u_k),$$

$$R_\ell = \begin{cases} \arcsin(u), & \ell = 0, \\ -\sqrt{1-u^2}, & \ell = 1, \\ \frac{\ell-1}{\ell} R_{\ell-2} - \frac{1}{\ell} u^{\ell-1}\sqrt{1-u^2}, & \ell \geq 2. \end{cases} \tag{5.36}$$

These expressions can in fact also be used to define the moments for odd N (Runborg 2000).

As in Section 5.2 we let $\boldsymbol{m} = (m_{10}, \ldots, m_{1,N-1})^T$. We define the function $\boldsymbol{F}_1$ by $\boldsymbol{F}_1(\boldsymbol{u}) = \boldsymbol{m}$ together with (5.3), and similarly for $\boldsymbol{F}_2$ and $\boldsymbol{K}$. We can then finally write (5.34) as

$$(\eta \boldsymbol{F}_1(\boldsymbol{u}))_x + (\eta \boldsymbol{F}_2(\boldsymbol{u}))_y = \boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y), \tag{5.37}$$

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

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

The functions $\boldsymbol{F}_j$ and $\boldsymbol{K}$ are again rather complicated nonlinear functions. For $N = 1$, the functions are simple:

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

For $N = 2$, let

$$\boldsymbol{f}_1 = \begin{pmatrix} w \\ \frac{1}{2}w^2 \end{pmatrix}, \qquad \boldsymbol{f}_2 = \begin{pmatrix} -\sqrt{1-w^2} \\ \frac{1}{2}\big(\arcsin(w) - w\sqrt{1-w^2}\big) \end{pmatrix},$$

$$\boldsymbol{k} = \begin{pmatrix} 0 \\ \frac{\eta_y}{2}\big(\arcsin(w) + w\sqrt{1-w^2}\big) - \frac{1}{2}\eta_x w^2 \end{pmatrix}.$$

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

$$\boldsymbol{f}_1 = \begin{pmatrix} w \\ \frac{1}{2}w^2 \\ \frac{1}{3}w^3 \end{pmatrix}, \qquad \boldsymbol{f}_2 = \begin{pmatrix} -\sqrt{1-w^2} \\ \frac{1}{2}\big(\arcsin(w) - w\sqrt{1-w^2}\big) \\ -\frac{1}{3}(2+w^2)\sqrt{1-w^2} \end{pmatrix},$$

$$\boldsymbol{k} = \begin{pmatrix} 0 \\ \eta_y\big(\arcsin(w) + w\sqrt{1-w^2}\big) - \frac{1}{2}\eta_x w^2 \\ -\frac{2}{3}\eta_y(1-w^2)\sqrt{1-w^2} - \frac{2}{3}\eta_x w^3 \end{pmatrix}.$$

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

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

$$\nabla_m m_{0,\ell} = V^{-1}\Theta_\ell. \tag{5.38}$$

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

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

By using (5.38) we get an expression for the Jacobian of $\boldsymbol{F}_2 \circ \boldsymbol{F}_1^{-1}$,

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

We see that this system is strictly hyperbolic as long as $\theta_k \neq \theta_\ell$ for all k, ℓ. See Gosse (2002) for further discussion of the theory for this system and how to couple it with equations for the amplitudes. Here, we simply note that, since $\tan\theta \to \infty$ when $|\theta| \to \pi/2$ or $|u| \to 1$, the Jacobian will blow up at these points. This is expected under the paraxial assumption.

We close this section by establishing the same superposition principle as for the delta equations in Section 5.2.

Theorem 5.7. Suppose $\{u_k\}_{k=1}^M$ are M weak solutions to (5.37) with $N = 1$ in the sense of Theorem 5.6, and $\eta \in C^1$. If u_k are continuous functions with locally bounded variation, then $\boldsymbol{u} = (u_1, \ldots, u_M)^T$ is a weak solution to (5.37) with $N = M$ in the same sense.

Properties of the flux functions
Also in this case, the functions

$$\boldsymbol{F}_2 \circ \boldsymbol{F}_1^{-1}(\boldsymbol{m}) \qquad \text{and} \qquad \boldsymbol{K}(\boldsymbol{F}_1^{-1}(\boldsymbol{m}), \eta_x, \eta_y) \tag{5.39}$$

are well defined and regular on their domains of definition. We consider a slightly more general class of functions than those in (5.39). For a closed interval $I \subset \mathbb{R}$, define the (compact) set of attainable moments, $M_N \subset \mathbb{R}^N$,

$$M_N(I) = \left\{\boldsymbol{m} \in \mathbb{R}^N \;\middle|\; \boldsymbol{m} = \boldsymbol{F}_1(\boldsymbol{u}),\ u_1 \leq \cdots \leq u_N,\ u_k \in I\right\},$$

and introduce the class of mappings from $M_N(I)$ to $\mathbb{R}$ given by

$$J_\psi(\boldsymbol{m}) = \int_I \psi(t) f_m(t)\,\mathrm{d}t, \tag{5.40}$$

where $f_m(t)$ and $\boldsymbol{m}$ are related by

$$\boldsymbol{m} = \boldsymbol{F}_1(\boldsymbol{u}), \qquad f_m(t) = \sum_{k=1}^N (-1)^{k+1} H(t-u_k), \qquad u_1 \leq \cdots \leq u_N, \quad u_k \in I. \tag{5.41}$$

Brenier and Corrias (1998) showed that, if $I = [0, L]$, the mappings given by (5.40) and (5.41) are well defined and continuous on $M_N(I)$ for each $0 < L < \infty$, when ψ has a strictly positive and bounded Nth distributional derivative. These functions were identified as entropies for the moment system in Brenier and Corrias (1998). In general J_ψ is Hölder-continuous, but not continuously differentiable, as seen in the following result, from Runborg (2000).

Theorem 5.8. Let $I = [-L, L]$ for some positive $L < \infty$. The mapping $J_\psi : M_N(I) \to \mathbb{R}$ is well defined by (5.40) and (5.41). If $\psi \in L^p(I)$, with $1 \le p \le \infty$, then

$$J_\psi \in \begin{cases} C^0, & p = 1, \\ C^{0,\frac{p-1}{pN}}, & 0 < p < \infty, \\ C^{0,1/N}, & p = \infty, \end{cases}$$

where $C^{0,\alpha}$ is the set of Hölder-continuous functions with exponent α. If $\psi \in C^M(I)$, then

$$J_\psi \in \begin{cases} C^{M+1}, & N = 1, \\ C^{0,1/\max(N-M,1)}, & N > 1. \end{cases}$$

If $N > 1$ and $\psi \in C^0(I)$, then ∇J_ψ is continuous almost everywhere. It is discontinuous at $\boldsymbol{m} = 0$, unless ψ is a polynomial of degree at most $N - 1$, in which case $J_\psi \in C^\infty$.

When $|u_k| \le L < 1$ then, up to a constant, each element of the flux function $\boldsymbol{F}_2 \circ \boldsymbol{F}_1^{-1}$ is of the form (5.40) and (5.41) with $\psi = u^\ell/\sqrt{1-u^2}$, $\ell = 1, \ldots, N$. The source function $\boldsymbol{K}$ is of a similar form. Hence we have the following corollary.

Corollary 5.9. The flux and source functions (5.39) are well defined and depend Lipschitz-continuously on $\boldsymbol{m} \in M_N[-L, L]$ when $0 < L < 1$. They are not continuously differentiable.

We refer to Runborg (2000, Section 3.1) for the proofs.

5.4. Numerical results

In this section we show results of applying the equations derived in Sections 5.2 and 5.3 to a few different test problems. We consider both homogeneous ($\eta \equiv 1$) and inhomogeneous ($\eta = \eta(\boldsymbol{x})$) media and use closures corresponding to $N = 1, 2, 3$ crossing rays at each point. The equations closed with delta functions, (5.18), are set in two-dimensional space, while the Heaviside equations, (5.37), are reduced to a one-dimensional space by the paraxial approximation. As a shorthand we will refer to the equations

as the δ- and the H-equations. For a more complete numerical study, see Runborg (1998, 2000) and Gosse (2002).

As we remarked on page 246, the δ-equations (5.18) are nonstrictly hyperbolic with linearly degenerate fields. This is reflected in their sensitivity to numerical treatment. Even for smooth problems, some standard numerical schemes, such as Godunov, Lax–Friedrichs and Nessyahu–Tadmor with dimensional splitting, converge poorly in L^1 and may fail to converge in L^∞ (Engquist and Runborg 1996, 1998). The standard unsplit Lax–Friedrichs scheme converges well, and Jiang and Tadmor (1998) showed that, with an unsplit, genuinely two-dimensional version of Nessyahu–Tadmor, the expected second-order convergence rate is obtained for smooth problems. This is illustrated by Figure 5.2, Table 5.1 and Table 5.2. It appears that the dimensional splitting aggravates the numerical errors, although for the Godunov scheme James and Gosse (2000) observed that the same type of failure to converge in L^∞ can also occur in the much simpler case of a linear one-dimensional equation with variable coefficients. Kinetic schemes have been recognized to handle nonstrictly hyperbolic problems better. They were used with success for the δ-equations in Jin and Li (200x) and Gosse *et al.* (200x). Also note the importance of treating the source term correctly in heterogeneous media, where it may be very stiff because of large gradients in the index of refraction, for example by using so-called well-balanced schemes (Gosse 2002, Gosse *et al.* 200x).

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

$$\boldsymbol{F}_0(\boldsymbol{u}) = \boldsymbol{m}, \tag{5.42}$$

for each time step, at each grid point. Solving (5.42) 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 $\boldsymbol{F}_0$ is singular when two rays are parallel. For iterative methods that use the Jacobian, this is a problem. When $N = 1, 2$ there is an analytical way to invert $\boldsymbol{F}_0$: see Runborg (2000). Furthermore, (5.42) may not have a solution. Although, for the exact solution of the PDE, (5.42) should always be satisfied, truncation errors in the numerical scheme may have perturbed the solution so that $\boldsymbol{m}$ is not in M_N, the range of $\boldsymbol{F}_0$.

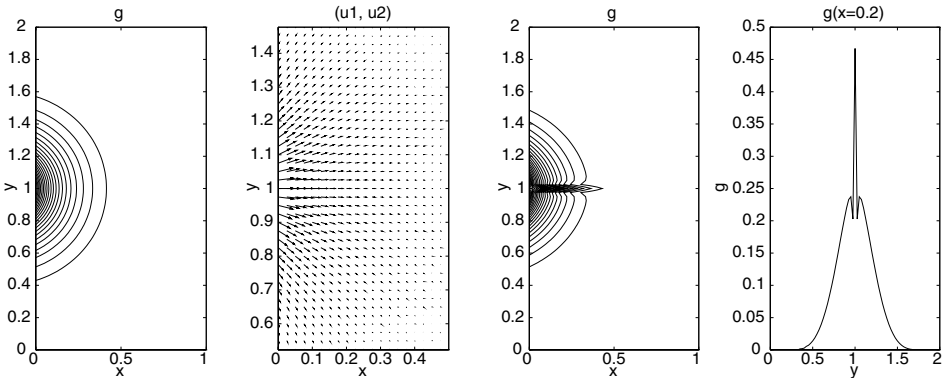
The H-equations (5.37) are strictly hyperbolic and numerical schemes are not as sensitive as for the δ-equations. The evaluation of the flux functions is also easier, since it can be reduced to solving polynomial equations of low degree (Runborg 2000). By also accepting complex roots of those polynomial equations, M_N, the domain of definition of the flux function can be continuously extended, avoiding the problem of (5.42) not having a solution.

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 the fictitious phases can obviously not be set to zero. Moreover, setting them to the same as another phase would eliminate them from the equations. For the H-equations with $N = 2$, for instance, $u_1 \equiv u_2$ is a trivial solution. However, Gosse (2002) pointed out that simply initializing $u_{k+1} = u_k + \epsilon$, with a small ϵ, often works well when u_k and u_{k+1} are physically relevant and nonrelevant phases, respectively.

Test problems

One point source. We consider one point source located at $\boldsymbol{s} = (-0.2, 1)$ and compute the solution in the rectangle $[0, 1] \times [0, 2]$. The source is smooth with exact solution $\boldsymbol{u}(t, \boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{s}) \max(0, t - r)^3 / r^2$, $r = \|\boldsymbol{x} - \boldsymbol{s}\|$, which we apply as boundary value at $x = 0$. General results are shown in Figure 5.2(a), where the Lax–Friedrichs method was used to solve the δ-system with $N = 1$ and 40×80 grid points. The difficulties with using the Godunov method for the same problem are highlighted in Figure 5.2(b).

Convergence for the different methods are summarized in Tables 5.1 and 5.2. The numerical error in $u_1 = m_{10}$ is shown measured in the L^1- and L^∞-norms.



(a) Lax–Friedrichs scheme: contour plot of g (left) and quiver plot of vector field $\boldsymbol{u} = (u_1, u_2)$ (right)

(b) Godunov scheme: contour plot of g (left), and g in a vertical cut at $x = 0.2$ (right)

Figure 5.2. *One point source.* Snapshot of solution of δ-equations with $N = 1$ at time $t = 0.85$ using Lax–Friedrichs and Godunov schemes.

Table 5.1. *One point source.* L^1-norm of the numerical errors for the single-phase δ-equations with different methods. Here n is the number of grid points in the x-direction.

L^1	Lax–Friedrichs unsplit		Lax–Friedrichs split		Godunov split		Nessyahu–Tadmor unsplit		Nessyahu–Tadmor split	
n	error	order	error	order	error	order	error	order	error	order
10	7.78e-3		3.60e-2		1.13e-2		7.98e-3		1.02e-2	
		0.85		0.56		0.80		1.47		1.23
20	4.33e-3		2.44e-2		6.50e-3		2.89e-3		4.35e-3	
		0.92		0.72		0.69		1.74		1.20
40	2.29e-3		1.48e-2		4.04e-3		8.66e-4		1.89e-3	
		0.96		0.82		0.78		1.88		1.03
80	1.18e-3		8.39e-3		2.35e-3		2.35e-4		9.24e-4	
		0.98		0.89		0.85		1.89		0.76
160	5.99e-4		4.53e-3		1.30e-3		6.32e-5		5.45e-4	

Table 5.2. *One point source.* L^∞-norm of the numerical errors for the single-phase δ-equations with different methods. Here n is the number of grid points in the x-direction.

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

Three point sources. We now consider a problem with three point sources located at coordinates $\boldsymbol{s}_1 = (-0.5, 0.5)$, $\boldsymbol{s}_2 = (-0.5, 1.0)$ and $\boldsymbol{s}_3 = (-0.5, 1.5)$. The exact solutions are

$$\boldsymbol{w}_k(t, \boldsymbol{x}) = A_k(\boldsymbol{x} - \boldsymbol{s}_k)H(t - r_k)/r_k^2,$$
$$r_k = \|\boldsymbol{x} - \boldsymbol{s}_k\|, \quad k = 1, 2, 3,$$

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 5.3 shows the solution at $t = 1.0$ of the δ-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 $x = 0$, that is,

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

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

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

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, where rays should cross, but cannot, since the systems describe too few phases; *cf.* the analysis of the conservation law in Section 5.2.

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

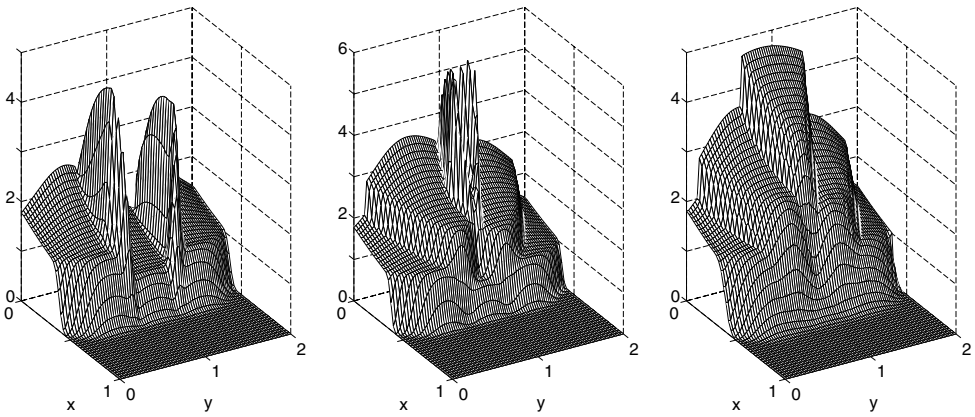
$$\eta(x, y) = \begin{cases} 1 & d^2 > 1, \\ \left(\frac{4}{3 - \cos(\pi d^2)}\right)^2 & d^2 \leq 1, \end{cases} \qquad d^2 = \left(\frac{x - 0.5}{0.2}\right)^2 + \left(\frac{y - 1}{0.8}\right)^2.$$

We have computed the solution in the square $[0, 2] \times [0, 2]$ for the δ-equations with $N = 1, 2$ and the H-equations with $N = 2, 3$. Figure 5.4 shows the ray angles of the solutions. Here initial data only contain one phase, but the focusing of the lens creates additional phases, which are captured automatically by the multiphase systems.

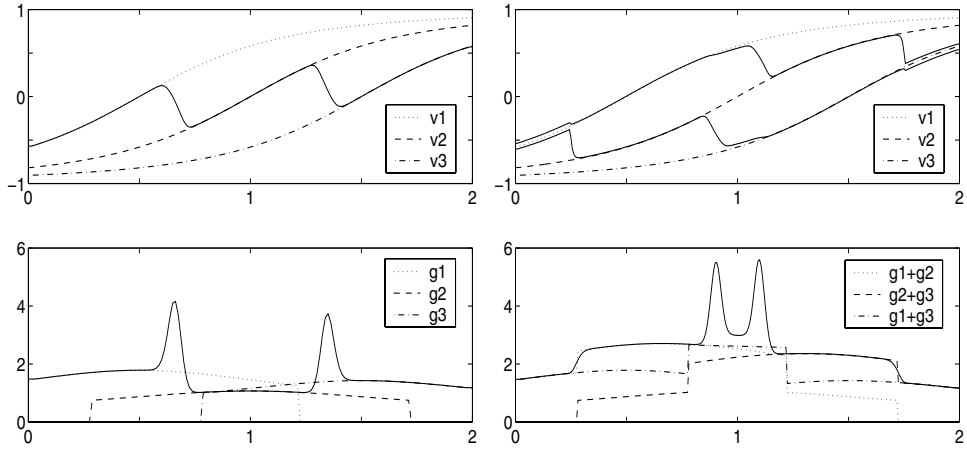
Wedge. In this test problem a plane wave, injected at $x = 0$ with $\theta(0, y) = 0$ and $g(0, y) = 2$, is refracted by a smooth wedge, modelled by the index of refraction

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

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 5.5(c).



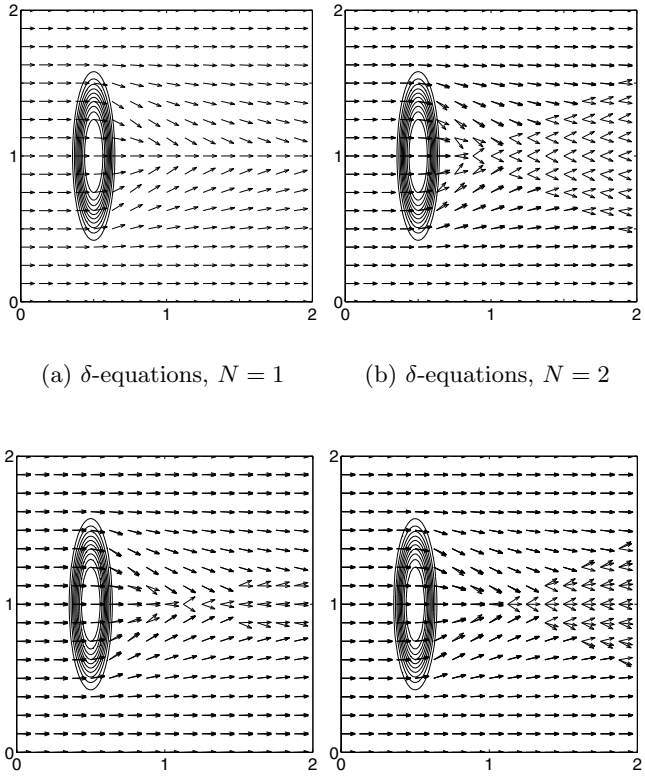
(a) $N = 1$ (b) $N = 2$ (c) $N = 3$



(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 5.3. *Three point sources.* Solution of the δ-equations with $N = 1, 2, 3$. Figures (a), (b), (c) show total ray strength, that is, g, $g_1 + g_2$ and $g_1 + g_2 + g_3$, respectively. Figures (d), (e) show solution in a cut at $x = 0.2$, computed (solid) and exact (dotted, dashed, dash-dotted).

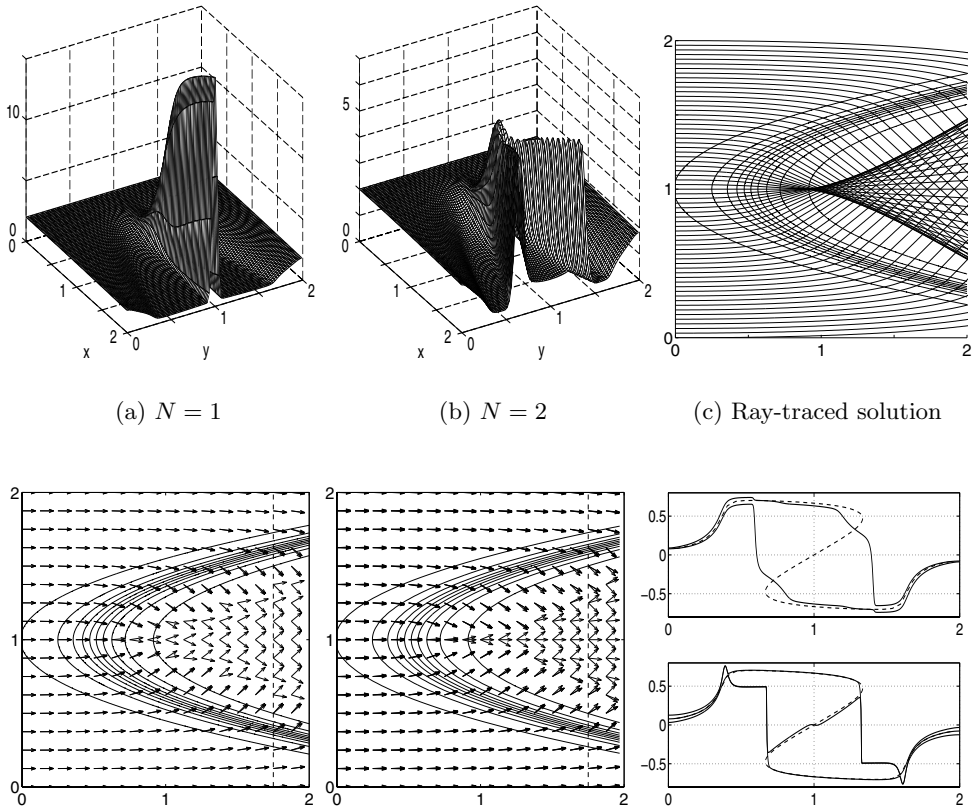


(a) δ-equations, $N = 1$ (b) δ-equations, $N = 2$

(c) H-equations, $N = 2$ (d) H-equations, $N = 3$

Figure 5.4. *Convex lens.* Ray angles for δ- and H-equations with different N. A contour plot of the index of refraction is overlaid on the solution.

As in the previous problem the δ- and H-equations were solved in the square $[0, 2]\times[0, 2]$. Different aspects of the solutions are shown in Figure 5.5. The δ-equations with $N = 1$ only capture one of the phases, as expected. A delta function appears where rays try to cross. The $N = 2$ system captures both the second phase and the caustic quite well. The H-equations cannot correctly capture the second phase when $N = 2$. However, when $N = 3$ all three phases are captured.



(a) $N = 1$ (b) $N = 2$ (c) Ray-traced solution

(d) δ-equations, $N = 2$ (e) H-equations, $N = 3$ (f) δ-equations, $N = 2$ (above), H-equations, $N = 3$ (below)

Figure 5.5. *Wedge.* Amplitude results, (a), (b), (c), for δ-equations with $N = 1, 2$. Figures (a), (b) show total ray strength, that is, g and $g_1 + g_2$ respectively. Figure (c) shows a ray-traced solution with contour lines of the index of refraction superimposed. Figures (d), (e) show quiver plots of ray angles for δ- and H-equations with $N = 2, 3$. A contour plot of the index of refraction is overlaid on the solution. Figure (f) shows sine of ray angles (solid) in a cut at $x = 1.75$ together with the corresponding values for a ray-traced solution (dashed).

REFERENCES

R. Abgrall and J.-D. Benamou (1999), 'Big ray tracing and eikonal solver on unstructured grids: Application to the computation of a multivalued traveltime field in the Marmousi model', *Geophysics* **64**, 230–239.

G. Bal, G. Papanicolaou and L. Ryzhik (2002), 'Radiative transport limit for the random Schrödinger equation', *Nonlinearity* **15**, 513–529.

J.-D. Benamou (1996), 'Big ray tracing: Multivalued travel time field computation using viscosity solutions of the eikonal equation', *J. Comput. Phys.* **128**, 463–474.

J.-D. Benamou (1999), 'Direct computation of multivalued phase space solutions for Hamilton–Jacobi equations', *Comm. Pure Appl. Math.* **52**, 1443–1475.

J.-D. Benamou (2003), 'An introduction to Eulerian geometrical optics (1992–2002)', *SIAM J. Sci. Comp.* To appear.

J.-D. Benamou and I. Solliec (2000), 'An Eulerian method for capturing caustics', *J. Comput. Phys.* **162**, 132–163.

J.-D. Benamou, F. Castella, T. Katsaounis and B. Perthame (2002), 'High frequency limit of the Helmholtz equations', *Rev. Mat. Iberoamericana* **18**, 187–209.

J.-D. Benamou, O. Lafitte, R. Sentis and I. Solliec (2003), 'A geometric optics based numerical method for high frequency electromagnetic fields computations near fold caustics, Part I', *J. Comput. Appl. Math.* To appear.

F. Bouchut (1994), On zero pressure gas dynamics, in *Advances in Kinetic Theory and Computing*, Vol. 22 of *Ser. Adv. Math. Appl. Sci.*, World Scientific Publishing, River Edge, NJ, pp. 171–190.

F. Bouchut and F. James (1995), 'Equations de transport unidimensionnelles à coefficients discontinus', *C. R. Acad. Sci. Paris Sér. I Math.* **320**, 1097–1102.

F. Bouchut and F. James (1999), 'Duality solutions for pressureless gases, monotone scalar conservation laws, and uniqueness', *Comm. Part. Diff. Eqns* **24**, 2173–2189.

Y. Brenier and L. Corrias (1998), 'A kinetic formulation for multibranch entropy solutions of scalar conservation laws', *Ann. Inst. Henri Poincaré* **15**, 169–190.

Y. Brenier and E. Grenier (1998), 'Sticky particles and scalar conservation laws', *SIAM J. Numer. Anal.* **35**, 2317–2328 (electronic).

P. Bulant and L. Klimeš (1999), 'Interpolation of ray theory traveltimes within ray cells', *Geophys. J. Int.* **139**, 273–282.

S. Cao and S. Greenhalgh (1994), 'Finite-difference solution of the eikonal equation using an efficient, first-arrival, wavefront tracking scheme', *Geophysics* **59**, 632–643.

F. Castella, B. Perthame and O. Runborg (2002), 'High frequency limit of the Helmholtz equation II: Source on a general smooth manifold', *Comm. Part. Diff. Eqns* **27**, 607–651.

V. Červený, I. A. Molotkov and I. Psencik (1977), *Ray Methods in Seismology*, University of Karlova Press.

L.-T. Cheng, S. J. Osher and J. Qian (2002), 'Level set based Eulerian methods for multivalued traveltimes in both isotropic and anisotropic media'. Preprint.

J. Claerbout (1976), *Fundamentals of Geophysical Data Processing*, McGraw-Hill.

M. G. Crandall and P.-L. Lions (1983), 'Viscosity solutions of Hamilton–Jacobi equations', *Trans. Amer. Math. Soc.* **277**, 1–42.

J. J. Duistermaat (1974), 'Oscillatory integrals, Lagrange immersions and unfolding of singularities', *Comm. Pure Appl. Math.* **27**, 207–281.

W. E, Y. G. Rykov and Y. G. Sinai (1996), '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**, 349–380.

B. Engquist and A. Majda (1977), 'Absorbing boundary conditions for the numerical simulation of waves', *Math. Comp.* **31**, 629–651.

B. Engquist and O. Runborg (1996), 'Multiphase computations in geometrical optics', *J. Comput. Appl. Math.* **74**, 175–192.

B. Engquist and O. Runborg (1998), Multiphase computations in geometrical optics, in *Hyperbolic Problems: Theory, Numerics, Applications* (M. Fey and R. Jeltsch, eds), Vol. 129 of *Internat. Ser. Numer. Math.*, ETH Zentrum, Zürich, Switzerland.

B. Engquist, O. Runborg and A.-K. Tornberg (2002), 'High frequency wave propagation by the segment projection method', *J. Comput. Phys.* **178**, 373–390.

E. Fatemi, B. Engquist and S. J. Osher (1995), 'Numerical solution of the high frequency asymptotic expansion for the scalar wave equation', *J. Comput. Phys.* **120**, 145–155.

S. Fomel and J. A. Sethian (2002), 'Fast-phase space computation of multiple arrivals', *Proc. Natl. Acad. Sci. USA* **99**, 7329–7334 (electronic).

S. Geoltrain and J. Brac (1993), 'Can we image complex structures with first-arrival traveltime?', *Geophysics* **58**, 564–575.

P. Gérard (1991), 'Microlocal defect measures', *Comm. Part. Diff. Eqns* **16**, 1761–1794.

P. Gérard, P. A. Markowich, N. J. Mauser and F. Poupaud (1997), 'Homogenization limits and Wigner transforms', *Comm. Pure Appl. Math.* **50**, 323–379.

L. Gosse (2002), 'Using K-branch entropy solutions for multivalued geometric optics computations', *J. Comput. Phys.* **180**, 155–182.

L. Gosse, S. Jin and X. Li (200x), 'On two moment systems for computing multiphase semiclassical limits of the Schrödinger equation'. To appear.

H. Grad (1949), 'On the kinetic theory of rarefied gases', *Comm. Pure Appl. Math.* **2**, 331–407.

S. Gray and W. May (1994), 'Kirchhoff migration using eikonal equation traveltimes', *Geophysics* **59**, 810–817.

E. Grenier (1995), 'Existence globale pour le système des gaz sans pression', *CR Acad. Sci. Paris Sér. I Math.* **321**, 171–174.

L. Hörmander (1983–1985), *The Analysis of Linear Partial Differential Operators, I–IV*, Springer, Berlin.

F. James and L. Gosse (2000), 'Numerical approximations of one-dimensional linear conservation equations with discontinuous coefficients', *Math. Comp.* **69**, 987–1015.

G.-S. Jiang and E. Tadmor (1998), 'Nonoscillatory central schemes for multidimensional hyperbolic conservation laws', *SIAM J. Sci. Comput.* **19**, 1892–1917.

S. Jin and X. Li (200x), 'Multi-phase computations of the semiclassical limit of the Schrödinger equation and related problems: Whitham vs. Wigner', *Physica D*. To appear.

B. R. Julian and D. Gubbins (1977), 'Three-dimensional seismic ray tracing', *J. Geophys. Res.* **43**, 95–114.

J. Keller (1962), 'Geometrical theory of diffraction', *J. Opt. Soc. Amer.*

S. Kim (2000), 'An $\mathcal{O}(N)$ level set method for eikonal equations', *SIAM J. Sci. Comput.* **22**, 2178–2193.

S. Kim and R. Cook (1999), '3-D traveltime computation using second-order ENO scheme', *Geophysics* **64**, 1867–1876.

R. G. Kouyoumjian and P. H. Pathak (1974), 'A uniform theory of diffraction for an edge in a perfectly conducting surface', *Proc. IEEE* **62**, 1448–1461.

Y. A. Kravtsov (1964), 'On a modification of the geometrical optics method', *Izv. VUZ Radiofiz.* **7**, 664–673.

G. Lambaré, P. S. Lucio and A. Hanyga (1996), 'Two-dimensional multivalued traveltime and amplitude maps by uniform sampling of ray field', *Geophys. J. Int.* **125**, 584–598.

R. T. Langan, I. Lerche and R. T. Cutler (1985), 'Tracing of rays through heterogeneous media: An accurate and efficient procedure', *Geophysics* **50**, 1456–1465.

R. J. LeVeque (1992), *Numerical Methods for Conservation Laws*, Birkhäuser.

C. D. Levermore (1996), 'Moment closure hierarchies for kinetic theories', *J. Statist. Phys.* **83**, 1021–1065.

H. Ling, R. Chou and S. W. Lee (1989), 'Shooting and bouncing rays: Calculating the RCS of an arbitrarily shaped cavity', *IEEE T. Antenn. Propag.* **37**, 194–205.

P.-L. Lions and T. Paul (1993), 'Sur les mesures de Wigner', *Rev. Mat. Iberoamericana* **9**, 553–618.

D. Ludwig (1966), 'Uniform asymptotic expansions at a caustic', *Comm. Pure Appl. Math.* **19**, 215–250.

V. P. Maslov (1965), *Theory of Perturbations and Asymptotic Methods*.

L. Miller (2000), 'Refraction of high-frequency waves density by sharp interfaces and semiclassical measures at the boundary', *J. Math. Pures Appl. IX* **79**, 227–269.

S. J. Osher and R. P. Fedkiw (2002), *Level Set Methods and Dynamic Implicit Surfaces*, Springer.

S. J. Osher and J. A. Sethian (1988), 'Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations', *J. Comput. Phys.* **79**, 12–49.

S. J. Osher and C.-W. Shu (1991), 'High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations', *SIAM J. Numer. Anal.* **28**, 907–922.

S. J. Osher, L.-T. Cheng, M. Kang, H. Shim and Y.-H. Tsai (2002), 'Geometric optics in a phase-space-based level set and Eulerian framework', *J. Comput. Phys.* **179**, 622–648.

V. Pereyra, W. H. K. Lee and H. B. Keller (1980), 'Solving two-point seismic ray-tracing problems in a heterogeneous medium', *Bull. Seism. Soc. Amer.* **70**, 79–99.

F. Poupaud and M. Rascle (1997), 'Measure solutions to the linear multi-dimensional transport equation with non-smooth coefficients', *Comm. Part. Diff. Eqns* **22**, 337–358.

F. Qin, Y. Luo, K. B. Olsen, W. Cai and G. T. Schuster (1992), 'Finite-difference solution of the eikonal equation along expanding wavefronts', *Geophysics* **57**, 478–487.

O. Runborg (1998), Multiscale and multiphase methods for wave propagation, PhD thesis, NADA, KTH, Stockholm.

O. Runborg (2000), 'Some new results in multiphase geometrical optics', *M2AN Math. Model. Numer. Anal.* **34**, 1203–1231.

S. J. Ruuth, B. Merriman and S. J. Osher (2000), 'A fixed grid method for capturing the motion of self-intersecting wavefronts and related PDEs', *J. Comput. Phys.* **163**, 1–21.

L. Ryzhik, G. Papanicolaou and J. B. Keller (1996), 'Transport equations for elastic and other waves in random media', *Wave Motion* **24**, 327–370.

W. A. Schneider, K. A. Ranzinger, A. H. Balch and C. Kruse (1992), 'A dynamic programming approach to first arrival traveltime computation in media with arbitrary distributed velocities', *Geophysics* **57**, 39–50.

J. A. Sethian (1996), 'A fast marching level set method for monotonically advancing fronts', *Proc. Nat. Acad. Sci. USA* **93**, 1591–1595.

J. A. Sethian (1999), *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, 2nd edn, Cambridge University Press, Cambridge.

I. Solliec (2003), Calcul optique multivalué et calcul d'energie électromagnétique en présence d'une caustique de type pli, PhD thesis, Université Pierre et Marie Curie, Paris 6, France.

C. Sparber, N. Mauser and P. A. Markowich (2003), 'Wigner functions vs. WKB-techniques in multivalued geometrical optics', *J. Asympt. Anal.* **33**, 153–187.

J. Steinhoff, M. Fan and L. Wang (2000), 'A new Eulerian method for the computation of propagating short acoustic and electromagnetic pulses', *J. Comput. Phys.* **157**, 683–706.

J. Steinhoff, Y. Wenren, D. Underhill and E. Puskas (1995), Computation of short acoustic pulses, in *Proceedings, 6th International Symposium on CFD, Lake Tahoe NV.*

Y. Sun (1992), Computation of 2D multiple arrival traveltime fields by an interpolative shooting method, in *Soc. Expl. Geophys.*, pp. 1320–1323.

W. W. Symes (1996), A slowness matching finite difference method for traveltimes beyond transmission caustics, Preprint, Department of Computational and Applied Mathematics, Rice University.

W. W. Symes and J. Qian (2003), 'A slowness matching Eulerian method for multivalued solutions of eikonal equations', *SIAM J. Sci. Comp.* To appear.

L. Tartar (1990), 'H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations', *Proc. Roy. Soc. Edinburgh Sect. A* **115**, 193–230.

C. H. Thurber and W. L. Ellsworth (1980), 'Rapid solution of ray tracing problems in hetereogeneous media', *Bull. Seism. Soc. Amer.* **70**, 1137–1148.

A.-K. Tornberg (2000), Interface tracking methods with applications to multiphase flows, PhD thesis, NADA, KTH, Stockholm.

A.-K. Tornberg and B. Engquist (2000), Interface tracking in multiphase flows, in *Multifield Problems*, Springer, Berlin, pp. 58–65.

A.-K. Tornberg and B. Engquist (2003), 'The segment projection method for interface tracking', *Comm. Pure Appl. Math.* **56**, 47–79.

Y. R. Tsai, L. T. Cheng, S. Osher and H. K. Zhao (2003), 'Fast sweeping algorithms for a class of Hamilton–Jacobi equations', *SIAM J. Numer. Anal.* To appear.

J. N. Tsitsiklis (1995), 'Efficient algorithms for globally optimal trajectories', *IEEE Trans. Automat. Control* **40**, 1528–1538.

J. van Trier and W. W. Symes (1991), 'Upwind finite-difference calculation of traveltimes', *Geophysics* **56**, 812–821.

J. Vidale (1988), 'Finite-difference calculation of traveltimes', *Bull. Seism. Soc. Amer.* **78**, 2062–2076.

V. Vinje, E. Iversen and H. Gjøystdal (1992), Traveltime and amplitude estimation using wavefront construction, in *Eur. Assoc. Expl. Geophys.*, pp. 504–505.

V. Vinje, E. Iversen and H. Gjøystdal (1993), 'Traveltime and amplitude estimation using wavefront construction', *Geophysics* **58**, 1157–1166.

G. B. Whitham (1974), *Linear and Nonlinear Waves*, Wiley.

Y. Zheng (1998), Systems of conservation laws with incomplete sets of eigenvectors everywhere, in *Advances in Nonlinear Partial Differential Equations and Related Areas*, World Scientific Publishing, River Edge, NJ, pp. 399–426.