KUNGL TEKNISKA HÖGSKOLAN

Multiscale and Multiphase Methods for Wave Propagation

Olof Runborg

Stockholm 1998

Doctoral Dissertation
Royal Institute of Technology
Department of Numerical Analysis and Computing Science

Akademisk avhandling som med tillstånd av Kungl Tekniska Högskolan framläggges till offentlig granskning för avläggande av teknisk doktorsexamen fredagen den 16 oktober 1998 kl 14.00 i Kollegiesalen, Administrationsbyggnaden, Kungl Tekniska Högskolan, Valhallavägen 79, Stockholm.

ISBN 91-7170-318-7

TRITA-NA-9818
ISSN 0348-2952
ISRN KTH/NA/R--98/18--SE

Abstract

In this thesis we consider multiple scale problems in numerical computations of propagating waves.

We first study the case of high frequency waves, where fine scales are present in the initial or the boundary data. The traditional geometrical optics approximation cannot handle solutions with multiple phases, corresponding to crossing rays. We derive a new family of equations that accommodate solutions containing more than one phase. These equations are based on the same high frequency approximation of the scalar wave equation as the classical eikonal and transport equations. However, they also incorporate a finite superposition principle. The maximum allowed number of intersecting waves in the solution can be chosen arbitrarily, but a higher number means that a larger system of PDEs must be solved. The PDEs form a hyperbolic system of conservation laws with source terms. A numerical study shows that the equations perform well on a variety of homogeneous and inhomogeneous problems. The technique we use to capture multi-valued solutions is based on two different closure assumptions for a system of equations representing the moments.

Next, we analyze the case when there are fine scales in the differential operator, typically a rapidly varying material coefficient. We describe a systematic technique for the derivation of subgrid scale models in the numerical solution of partial differential equations, in particular for wave propagation problems modeled by the Helmholtz equation. The technique is based on sparse approximation of the numerically homogenized operator, obtained through wavelet projections of the discrete operator. As numerical testing suggests, the resulting numerical method will accurately represent subgrid scale phenomena on a coarse grid. Applications to numerical homogenization and wave propagation in materials with subgrid inhomogeneities are presented.

Finally, we apply the numerical homogenization techniques to a problem in fiber optics, where high frequency waves interact with a small scale structure of gratings in a waveguide. Fine scales are thus present both in the operator and in the boundary data. We show that the homogenized operator can be analyzed and identified with a one-dimensional model of the waveguide. We also demonstrate the use of the homogenized operator as a subgrid model.

ISBN 91-7170-318-7 • TRITA-NA-9818 • ISSN 0348-2952 • ISRN KTH/NA/R--98/18--SE

Contents

Acknowledgment

The first one to be thanked is my dynamic advisor Prof. Björn Engquist for his stimulating scientific guidance and support throughout my PhD. His wide knowledge and sharp intellect have always impressed me and been a great source of inspiration. It has been a privilege to draw on his intuition and insights in applied mathematics. He has also arranged several valuable visits for me at the University of California, Los Angeles (UCLA).

I would also like to thank Prof. Jesper Oppelstrup. During all my time at NADA he has given me a huge amount of encouraging help and showed a never failing enthusiasm for my and everybody's work.

I am also grateful to Dr. Anders Szepessy for always being helpful and interested in my mathematical questions.

I am indebted to my co-authors Ulf Andersson, Gunnar Ledfelt and Per-Olof Persson.

A big thanks also to my nice colleagues and friends at NADA with whom I have had the pleasure to work, and to all people at the Department of Mathematics, UCLA, for receiving me so many times.

Finally, I want to extend a warm thanks to my parents, Claes and Ingrid, for all their support and encouragement over the years.

Financial support has been provided by TFR (C^2M^2), FRN (PDC), NUTEK (PSCI), EIT (doktorandtjänst, resestipendium), KTH (Wallenbergstiftelsen, internationaliseringsmedel, Björns, Danielsons, Lindstrands, Styffes och Nyströmers stipendiefonder), EU TMR network "Hyperbolic Systems of Conservation Laws", NSF grant DMS 9706827 and AFOSR STTR F49620-97-0070. All financial supporters are gratefully acknowledged.

Chapter 1

Introduction

1.1 Motivation

A general problem in the numerical analysis of differential equations is the existence of multiple scales in their solutions. By "scale" we loosely mean the size of the interval, temporal or spatial, over which a typical change in a dependent variable occurs. The scale is closely related to the local frequency contents of the variable such that high frequencies correspond to small scales. From a numerical point of view it is the ratio between the smallest and the largest scales of the solution that is interesting. In general, the larger the ratio, the harder it is to solve the problem directly by numerical methods. The "large" scale in these cases is usually the size of the computational domain.

Numerical approximations of the exact solution must necessarily be sought in a space of finite dimensionality. This approximation space must be rich enough to contain functions of close resemblance to the exact solution. Without a priori knowledge of the solution, fine details, i.e. small scales, require a richer space, of higher dimensionality. The most common numerical methods discretize the physical domain by means of a grid and use local basis functions or pointwise values at the grid points to generate the approximation space. In this setting, the number of grid points needed to resolve the solution increases with the ratio between large and small scales. The sampling theorem gives a precise answer to the question of how many grid points are necessary to reconstruct exactly a function with a certain frequency contents on a uniform mesh. Similarly, one can estimate how many grid points are needed to achieve a given accuracy on a more general grid. Generally, at least

$$\#\text{ gridpoints} \sim \left(\frac{\text{large scale}}{\text{small scale}}\right)^d \tag{1.1}$$

where d is the dimension. The computational cost grows linearly or faster with the number of grid points. See [30] for estimates of the smallest scale and resolution requirement in the classic case of the Navier-Stokes equations.

A good resolution of the fine scales is not only important for the immediate approximation of the solution and the differential operator. Even if the small scales in the solution are also small in absolute size, such that discarding them would still yield a good approximation, they are in general important because they can influence the coarse scales via the differential operator. Fine scales in the solution may for example interact or resonate with fine scales in the material properties and produce coarse scale contributions to the solution.

There are several reasons for the presence of fine scales in the solution. Small scales in the initial and boundary *data* often persist in the solution as well. For instance, fine scale initial data that generate high-frequency waves is the generic behaviour for most linear hyperbolic equations, such as the scalar wave equations, Maxwell's equation and the equations of elastic waves.

Small scales can also be independently generated by *nonlinearities* in the equations. Turbulence, boundary layers and shock waves are for instance typical for the solutions of the ubiquitous Navier-Stokes equations. In fact, shock wave solutions are typical for all nonlinear hyperbolic conservation laws.

There can also be small scales in the surrounding physical media, i.e. in the *differential operator*. A rapidly varying material induces rapid variation in the solution. One example is flow in porous media where a fine grained distribution of domains with large and small permeability gives rise to large fluctuations in the flow velocity. Also an electromagnetic wave propagating in a composite material will have a very complicated small scale structure. Closely related are small scales generated by small scales in the geometry of a problem. Although the geometry can often be seen as a discontinuous material coefficient, computationally the discontinuity is usually handled via boundary and interface conditions. Small details in the geometry are characteristic for many electromagnetic problems, where e.g. thin wires, small holes and slits significantly influence the solution.

In this thesis our main interest is the multiple scale problem in the case of wave propagation. We study problems with small scales in the data, in the operator and in both.

1.2 Background

The problem of multiple scales has been approached from different angles. Analytically, asymptotic techniques have been used to study the limiting case when the scales become infinitely small and to derive differential equations which describe the limiting solution. For small scales in initial data, *geometrical optics* is one such approach. Under certain conditions, it gives the solution of the wave equation when the frequency tends to infinity, i.e. when the wavelength goes to zero. Geometrical optics can be seen as a special case of *homogenization* which is a general method used to derive equations for the limiting solution, mainly in the case of fine scales in the differential operator, but also in the initial data.

The limiting solution will in many cases be a good approximation of the solution when its true scales are very small. Numerically, the scale problem will typically be reduced. In some cases, for instance elliptic problems with periodic coefficients, the small scales simply does not exist in the limiting solution. In other cases, such as geometrical optics, the scales of the limiting solution have a simple structure, making it possible to choose an approximation space of low dimensionality.

1.2.1 Geometrical Optics

The derivation of the geometrical optics equations in the linear case is classical. See for instance the book by Whitham, [60]. The most simple form is based on the scalar wave equation,

$$u_{tt} - c^2 \Delta u = 0. \tag{1.2}$$

Generically, phase and amplitude vary on a much slower scale than the dependent variable in (1.2) and they are thus in principle easier to compute. It is therefore natural to consider solutions that can be written as a series expansion of the form

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

Entering this expression into (1.2) and summing terms of the same order in ω, to zero, we obtain separate equations for the unknown variables in (1.3). The phase function ϕ will satisfy the *eikonal equation*,

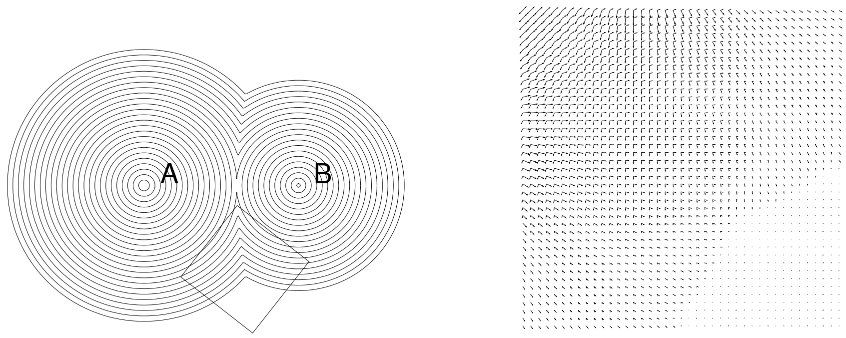
$$\phi_t + c|\nabla\phi| = 0, \tag{1.4}$$

and A_0 solves the *transport equation*,

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

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

The weakness of the infinite frequency approximation is the loss of some typical wave phenomena in the solution, such as diffraction and interference. In addition, the approximation breaks down at caustics, where the amplitude, A_0, blows up. Accepting these shortcomings, there still remain problems with the particular ansatz (1.3). It does not include solutions with multiple phases, corresponding to crossing waves. Even in the case of a single phase solution, the series does not necessarily converge, for instance when the geometric boundaries create diffraction effects. However, the eikonal equation (1.4) has a solution regardless of these problems. With some extra conditions, introduced by Crandall and Lions, [17], for general Hamilton-Jacobi type equations, the solution is also unique. Of course, this *viscosity solution* does not have to agree with the correct physical solution in all cases. At points where the correct solution should have a multi-valued phase, the viscosity solution picks out the phase corresponding to the first arriving wave, see Figure 1.1.



(a) Eikonal equation

(b) Moment equations with δ-closure for $N = 2$

Figure 1.1. 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. The left figure shows level curves of ϕ in the solution of the eikonal equation (1.4) for the two interacting waves. Note how the superposition principle does not hold. Instead, the first arriving wave takes precedence over the second at each point. The right figure shows the solution of the same problem in the square region indicated to the left. This time the moment equations with delta function closure for two phases were used and the correct solution with two overlapping wave fields is obtained.

Another way to formulate the geometrical optics approximation is by adopting a kinetic viewpoint. The waves are postulated to be particles (photons) whose trajectories are rays. Let $\eta = 1/c$ and define the the *ray vector* $\boldsymbol{p} = \eta\hat{\boldsymbol{s}}$, where $\hat{\boldsymbol{s}}$ is the unit vector in the direction of the ray. A transport equation for particles in the space $(t, \boldsymbol{x}, \boldsymbol{p})$ can then be derived. Denoting the density of particles by $f(t, \boldsymbol{x}, \boldsymbol{p})$ the evolution of f is described by the Vlasov type equation

$$f_t + \frac{1}{\eta^2}\boldsymbol{p} \cdot \nabla_x f + \frac{1}{\eta}\nabla_x \eta \cdot \nabla_p f = 0. \tag{1.6}$$

Tracing the particle trajectories of (1.6) corresponds to ray tracing and also to the method of characteristics for (1.4) and (1.6).

The geometrical optics type approximation is used in many applications, for example electromagnetic, elastic and acoustic high frequency wave propagation. Traditionally, ray tracing has been the computational method of choice. Because of the large number of independent variables (six in 3D) it is very hard numerically to solve the full equation (1.6). If the equation is solved using ray tracing it is difficult to cover the full domain with rays, [58]. There will often be shadow zones where the field cannot be resolved. It is also hard to determine the derivative of ϕ, which is needed when computing the amplitude.

Solving the eikonal equation (1.4) numerically as a partial differential equation (PDE) instead of using ray tracing has recently been used in seismology. This technique is demonstrated in [52], [58] and [59]. For these applications it is of direct interest to determine the first arrival. A second phase, corresponding to crossing rays was calculated in [22] using two separate eikonal equations. Boundary conditions for the second phase was given at the discontinuity of the first phase or at a geometric reflecting boundary. This boundary could be difficult to determine. Other interesting approaches to the multiple phase problem, also based on the eikonal equation, is the big ray tracing method of Benamou, [4, 5], and the slowness matching algorithm of Symes, [54].

In many applications the infinite frequency approximation is too crude, and correction terms are needed. Such correction terms were given by Keller in the 1960s, [34], with his pioneering geometrical theory of diffraction (GTD), further developed by e.g. Kouyoumjian and Pathak, [36]. In parallel, work on the solution's behaviour close to caustics, where the geometrical optics approximation fails, was carried out by among others Ludwig, [45], and Kravtsov [37].

Geometrical optics has also been developed in the case of nonlinear hyperbolic conservation laws. For small perturbations this is close to the linear case, but new phenomena in the form of resonance is added. A review of this field can be found in [46].

1.2.2 Homogenization

Homogenization was first developed as an analytical tool to derive effective material coefficients for equations with rapidly varying materials. Almost all results concern variations of either periodic or special stochastic nature. A systematic treatment of the general linear case was given by a number of authors in the 1970s, all summed up in the seminal book by Bensoussan, J. L. Lions and Papanicolaou, [6].

The classical homogenization problem can be stated as follows. Consider a family of differential operators L_ϵ, indexed by the small parameter ϵ, and study

$$L_\epsilon u_\epsilon = f_\epsilon, \tag{1.7}$$

where f_ϵ is given. The operator L_ϵ has coefficients which are periodic or stochastic on the spatial ϵ-scale. Show convergence, $u_\epsilon \to \bar{u}$ as $\epsilon \to 0$, and find the *homogenized operator* $\bar{L}$ and right hand side $\bar{f}$ such that

$$\bar{L}\bar{u} = \bar{f}. \tag{1.8}$$

In practice one often wants to solve (1.7) for a small but fixed ϵ. Since u_ϵ is close to $\bar{u}$ we may take $\bar{u}$ as an approximation. The homogenized equation does not contain the ϵ-scale and it can be solved on a coarse grid. In some cases $\bar{L}$ has a closed form analytic expression, [6]. If that is not the case, but the general structure of $\bar{L}$ is known, a numerical approximation is possible, [20].

The method of compensated compactness, a general mathematical tool also applicable for homogenization in the nonlinear case, was developed by Tartar and Murat in the end of the 1970s, [56], [47] and [48]. More recently, interest for the limit of more general quantities, usually related to the energy, has led to the development of new mathematical objects such as defect measures, [24], H-measures, [57] and a renewed interest in the classical Wigner transform, [42], [25].

1.2.3 Numerical Techniques

From the numerical side, the most direct way to mitigate the scale problem is to use *local mesh refinement* where the solution varies a lot. Hence, by drawing on some a priori information about the solution a more optimal grid can be obtained. Similarly, adaptive methods use a posteriori information to optimize the grid. See [7] and [3] for the finite difference and finite element setting, respectively. A drawback with these methods is that they might lead to reflections at mesh size discontinuities, increased complexity of the algorithm and reduced time steps for explicit methods due to stability constraints.

Another numerical approach is to make a local model of the effect of the scales that the grid cannot resolve. These *subgrid models* can be based on physical considerations for a specific application such as turbulence models in computational fluid dynamics, [61], and analytically derived local subcell models in computational electromagnetics, [55]. Another example is the immersed interface method, [41] used on Cartesian grids. In this method the subgrid variations of the shape of a material interface is taken into consideration by changing the updating stencils near the interfaces. Shock tracking or shock fitting can also be seen as subgrid models, [1].

Ideas for a more general approach to the construction of subgrid models were introduced by Brewster and Beylkin in [15] and further developed by Dorobantu and Engquist, [19]. Their method is based on multiresolution analysis with wavelet projections and approximation of the discrete operator. For simple problems the method can be seen as a discrete analogue of analytical homogenization.

The problem formulation begins with a finite-dimensional approximation of a differential equation,

$$Lu = f, \tag{1.9}$$

where L approximates the operator and u the solution. The operator L can be written on the form

$$L = P(\Delta, A, h), \tag{1.10}$$

where Δ is a collection of difference operators, A are discretized variable coefficients, typically diagonal matrices, and h represents the grid size, which is assumed to be sufficiently small to resolve all scales. We seek an operator $\bar{L}$ and a right hand side $\bar{f}$ with the following properties. First, $\bar{L}\bar{u} = \bar{f}$ and $\bar{u}$ is a projection of u onto a lower dimensional subspace. Second, $\bar{L}$ can be written on the same form as L,

$$\bar{L} = P(\Delta, H, \bar{h}), \tag{1.11}$$

but with $\bar{h} \gg h$ and the structure of H close to the structure of A, typically diagonal dominant. The matrix H is interpreted as the subgrid model of A. If A corresponds to a material coefficient, H can be seen as the effective material coefficient. The procedure outlined above resembles homogenization in the continuous case. In view of this, $\bar{L}$ is called the homogenized operator.

The suggestion in [15] was to use wavelet spaces for the lower dimensional subspace. Let V_j and W_j refer to scaling and wavelet spaces of coarseness level j for some wavelet system. Suppose u, f in (1.9) belongs to V_{j+1}. By definition $V_{j+1} = V_j \oplus W_j$, and the coarse scale of u, f are their projections onto V_j. We introduce the orthogonal transformation

$$\mathcal{W}_j : V_{j+1} \to W_j \times V_j, \quad \mathcal{W}_j u \equiv \begin{pmatrix} u_h \\ u_\ell \end{pmatrix} \qquad u_h \in W_j, \qquad u_\ell \in V_j, \tag{1.12}$$

and note that the linear operator $\mathcal{W}_j L \mathcal{W}_j^\top$ can be decomposed into four operators $A_j + B_j + C_j + L_j$, acting between the subspaces V_j and W_j, such that (1.9) becomes

$$\begin{pmatrix} A_j & B_j \\ C_j & D_j \end{pmatrix} \begin{pmatrix} u_h \\ u_\ell \end{pmatrix} = \begin{pmatrix} f_h \\ f_\ell \end{pmatrix}, \qquad u_h, f_h \in W_j, \qquad u_\ell, f_\ell \in V_j. \tag{1.13}$$

when we apply $\mathcal{W}_j$ from the left. Block Gaussian elimination now gives an equation for u_ℓ, the coarse part of the solution,

$$\bar{L}_j u_\ell = \bar{f}_j, \quad \bar{L}_j = D_j - C_j A_j^{-1} B_j, \quad \bar{f}_j = f_\ell - C_j A_j^{-1} f_h. \tag{1.14}$$

Hence, our homogenized operator $\bar{L}_j$ is the Schur complement of $\mathcal{W}_j L \mathcal{W}_j^\top$. We also get the homogenized right hand side, $\bar{f}_j$. The homogenization procedure can be applied recursively on $\bar{L}_j$ to get $\bar{L}_{j-1}$ and so on. It should be noted that in general $\bar{L}_j$ will not be sparse even if L is. For the method to be efficient we must be able to approximate $\bar{L}_j$ with a sparse matrix. Theoretical and numerical evidence that this can indeed be done is given in [15] and [19].

In some important cases the homogenized operator has been proven to remain in the same form as the original operator. In [19] it was shown that with standard discretization the one-dimensional elliptic operator $-(au_x)_x$ will preserve its discrete divergence form during homogenization. More precisely,

$$L = \frac{1}{h^2} \Delta_+ \mathrm{diag}(a) \Delta_- \qquad \Rightarrow \qquad \bar{L}_j = \frac{1}{\bar{h}^2} \Delta_+ H_j \Delta_-, \tag{1.15}$$

where the subgrid model H_j is a strongly diagonal dominant matrix. Analogously, for the first order operator au_x and the two dimensional elliptic operator $-\nabla \cdot (a \nabla u)$, the discretized form is preserved during homogenization, although in the latter case cross-derivatives must also be included.

The great advantage of this procedure is its generality. It can be used on any system of differential equations and does not require separation into the distinct $\mathcal{O}(\epsilon)$ and $\mathcal{O}(1)$ scales or periodic coefficients. It can also be used to test if it is physically reasonable to represent fine scales effects on a coarse scale grid with a local operator.

See [26] for a comparison with analytical homogenization and analysis in the one-dimensional case. There are also similarities with numerical homogenization based on techniques from algebraic multigrid, [35, 50] and from the use of special purpose basis functions in finite element approximations, [31].

The homogenization procedure outlined above is still in an early stage of development. We envisage, however, that it will be most useful for computations on Cartesian grids. A subgrid model would mitigate the main drawback of Cartesian grids, their poor ability to describe geometry accurately. It will still allow a solver to capitalize on the Cartesian grids' advantages: computational speed and small errors due to artificial dispersion and reflections. Typical applications would be oil reservoir simulations and computational electromagnetics.

1.3 Results

This thesis makes contributions to the multiple scale problem in the case of wave propagation. The first section treats an aspect of the problem where small scales in the initial or boundary data generate high frequency waves. This is related to the PDE formulation of geometrical optics. The second section concerns small scales in the differential operator, and the construction of subgrid models for such problems. In the last section the subgrid technique is applied to a problem in fiber optics, where small scales are present both in the data and in the operator.

1.3.1 Multiphase Geometrical Optics

As seen in Section 1.2.1 a drawback with the traditional PDE formulation of geometrical optics is that it only describes a limited class of physically relevant solutions. In particular, this class does not include solutions with multiple phases. Also, solving the full transport equation (1.6) is too expensive. We propose a middle way between the PDE formulation and the kinetic model. A new set of equations is derived. They can be seen as a generalization of the classical eikonal and transport equations, to allow for solutions with multiple phases. The new equations are based on the same high frequency approximation of the scalar wave equation as the eikonal and the transport equations. However, they also incorporate a finite superposition principle. The maximum allowed number of intersecting waves in the solution can be chosen arbitrarily, but a higher number means that a larger system of PDEs must be solved. The technique we use to capture multi-valued solutions is based on a closure assumption for a system of equations representing the moments, in the spirit of Brenier and Corrias, [12].

The starting point for this approach is the transport equation (1.6). We define the moments m_{ij}, with $\boldsymbol{p} = (p_1, p_2)^T$, as

$$m_{ij} = \int_{\mathbb{R}^2} p_1^i p_2^j f \, d\boldsymbol{p}. \tag{1.16}$$

Next, we multiply (1.6) by $p_1^i p_2^j$ and integrate over $\mathbb{R}^2$ with respect to $\boldsymbol{p}$. Using the definition (1.16) we see that formally m_{ij} will satisfy the infinite system of moment equations

$$(\eta^2 m_{ij})_t + (m_{i+1,j})_x + (m_{i,j+1})_y = i\eta\eta_x m_{i-1,j} + j\eta\eta_y m_{i,j-1}, \tag{1.17}$$

valid for all $i, j \geq 0$, if f has compact support in $\boldsymbol{p}$.

The system (1.17) is not closed. We observe, however, that when f is of a simple form in $\boldsymbol{p}$, we can close (1.17), analogously to the classical approach of the hydrodynamic limit from a kinetic formulation. In particular we are interested in forms of f which correspond to a finite number of rays in different directions at each point. This requires at least that the support of $f(t, \boldsymbol{x}, \cdot)$ is restricted to the circle $|\boldsymbol{p}| = \eta(\boldsymbol{x})$.

We consider first the case when f is a weighted sum of delta functions in $\boldsymbol{p}$,

$$f(t, \boldsymbol{x}, \boldsymbol{p}) = \sum_{k=1}^{N} g_k \cdot \delta(\boldsymbol{p} - \boldsymbol{p}_k), \qquad \boldsymbol{p}_k \equiv \eta \begin{pmatrix} \cos\theta_k \\ \sin\theta_k \end{pmatrix}. \tag{1.18}$$

For a system describing N phases we get $2N$ equations corresponding to the N ray strengths, $g_k = g_k(t, \boldsymbol{x})$, and directions $\theta_k = \theta_k(t, \boldsymbol{x})$. The resulting PDEs will be a system of conservation laws with source terms. For $N = 1$, $\eta \equiv 1$, the equations read

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_t + \begin{pmatrix} \frac{u_1^2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}_x + \begin{pmatrix} \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_2^2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}_y = 0, \qquad \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = g \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}. \tag{1.19}$$

In Figure 1.1 the corresponding equations with $N = 2$ were used. The equations are only weakly hyperbolic, and thus not well-posed in the strong sense. Existence of solutions of unbounded variation are also indicated when the allowed number of phases is exceeded. These facts imply that stable numerical methods can be hard to find. However, the standard first order Lax Friedrichs scheme gives satisfactory results. More elaborate, and less viscous, methods like the Godunov method and the second order TVD Nessyahu-Tadmor scheme, together with dimensional splitting, suffer from problems locally and converge poorly in L_∞, although they are converging well in L_1. A genuinely two-dimensional version of the Nessyahu-Tadmor scheme gives the best result. For the multiple phase system a few additional numerical problems are added, related to the fact that a non-linear system of equations needs to be solved at each point in each time-step. When $N = 2$ it can be done analytically.

Second, we consider the case when f is a sum of Heaviside functions. Essentially the ansatz is

$$f(t,\boldsymbol{x},\boldsymbol{p}) = \delta(|\boldsymbol{p}|-\eta)\sum_{k=1}^{N}(-1)^{k+1}H(\theta-\theta_k), \qquad \boldsymbol{p} \equiv |\boldsymbol{p}|\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}. \tag{1.20}$$

Here we can eliminate the ray strengths from the equations and only get N equations for an N phase system. We also confine ourselves to the steady state case. The PDEs will in general be strictly hyperbolic. We thus get fewer and less singular equations, but instead there are problems related to how a zero phase should be initialized and how to remain within the assumed class of solutions. For $N = 2$ and $\eta \equiv 1$ the equations are

$$\begin{aligned} \Big(\sin\theta_2 - \sin\theta_1\Big)_x + \Big(\cos\theta_1 - \cos\theta_2\Big)_y &= 0, \\ \Big(\sin^2\theta_2 - \sin^2\theta_1\Big)_x + \Big(\theta_2 - \theta_1 - \tfrac{1}{2}(\sin 2\theta_2 - \sin 2\theta_1)\Big)_y &= 0. \end{aligned} \tag{1.21}$$

We present a numerical study of how the equations obtained by the different closure assumptions perform on a variety of problems, including both homogeneous and inhomogeneous examples.

1.3.2 Wavelet Based Homogenization

In the second part of this thesis we continue the work initiated in [19]. The aim is to give a proof of concept that the subgrid technique of [19] produces useful subgrid models, H, also for wave-propagation problems. Although the homogenization procedure in fact amounts to a direct block Gaussian elimination, H can be computed efficiently when $\bar{L}_j$ is well approximated by sparse matrices in each homogenization step. This is true for many important cases, [15, 19].

Maximum efficiency has however not been our main concern. In our setting, H is supposed to be reused, either in multiple computations with different initial data, or in multiple places within the computational domain. Moreover, H itself can be of interest in the analysis of a physical problem. The homogenized operator could also be used as a multigrid coarse grid operator, [23].

We start by defining the homogenization algorithm in the context of Hilbert spaces and prove some general facts about it. The special case of wavelet spaces is then introduced and used as in [19]. In [19] only elliptic problems with stochastic and periodic variations of the material coefficients were studied. We extend this to wave propagation problems. First, we apply the method to the Helmholtz equation in one and two dimensions. This is still an elliptic equation, but a model for the propagation of time-harmonic waves. Second, we turn to the time-dependent hyperbolic model equation $u_t + a_\epsilon u_x = 0$ in one dimension where a_ϵ is periodic with period ϵ. For analytical homogenization of this equation, the influence of the time-derivative on the homogenization vanishes as $\epsilon \to 0$. For the

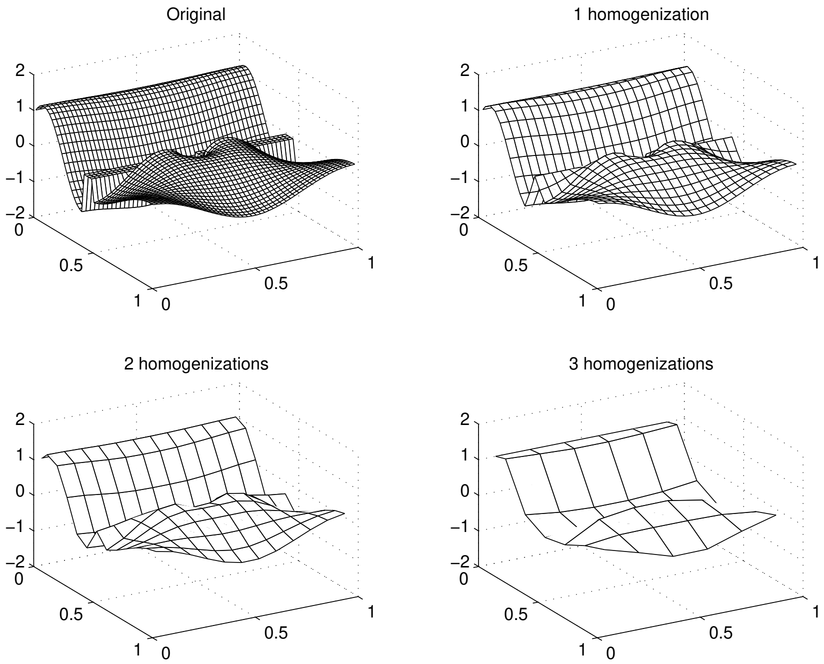


Figure 1.2. Simulation of a plane wave hitting a wall with a small slit where a part of the wave pass through. Solution computed using the 2D Helmholtz equation. The wave amplitude shown for $0,\ldots,3$ homogenization steps. (See also Section 3.2.3.)

numerical homogenization, however, we show that a correction for the finiteness of ϵ is needed to get good accuracy. Hence, for the semi-discrete approximation,

$$\frac{\partial u}{\partial t} + Lu = f, \tag{1.22}$$

we get a homogenized equation of the form

$$\frac{\partial u_\ell}{\partial t} + \tilde{L}_j u_\ell = \tilde{f}_j. \tag{1.23}$$

The operator $\tilde{L}_j$ and right hand side $\tilde{f}_j$ are corrected versions of the $\bar{L}_j$ and $\bar{f}_j$ given in (1.14),

$$\tilde{L}_j = K_j^{-1}\bar{L}_j, \qquad \tilde{f}_j = K_j^{-1}\bar{f}_j + K_j^{-1}C_jA_j^{-2}\frac{\partial f_h}{\partial t}. \tag{1.24}$$

with f_h as in (1.13) and the correction matrix

$$K_j = I + C_jA_j^{-2}B_j. \tag{1.25}$$

For the wave propagation problems, and some elliptic problems, we also make numerical experiments with *local* inhomogeneities. The results suggest that the method also works for this case. For instance, in Figure 1.2 we show that the effect of a small geometric detail can be captured on a coarse grid.

As was remarked in Section 1.2.3, it is important to be able to approximate the homogenized operator $\bar{L}_j$ with a sparse matrix. We introduce a new truncation algorithm where $\bar{L}_j$ is projected onto banded form in such a way that the projected matrix and the original matrix are equivalent when restricted to a given subspace. Hence, if band$(\cdot, \nu)$ represents this *band projection* to a bandwidth of ν,

$$M\boldsymbol{x} = \text{band}(M, \nu)\boldsymbol{x}, \qquad \forall \boldsymbol{x} \in \text{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_\nu\}. \tag{1.26}$$

where we choose the $\boldsymbol{v}_j$ vectors to be the first ν polynomials,

$$\boldsymbol{v}_j = \{1^{j-1}, 2^{j-1}, \ldots, N^{j-1}\}^T. \tag{1.27}$$

A smooth solution $\bar{u}$ to the homogenized problem should be well approximated by these vectors and hence $\bar{L}_j\bar{u}$ by band$(\bar{L}_j, \nu)\bar{u}$. The smoothness assumption is natural in connection to elliptic equations. This technique is similar to probing techniques suggested by Chan et al., [16], and Axelsson, Pohlman and Wittum, [2]. The method can also be used for compression of the wavelet based representation of Calderon-Zygmund operators in [15]. In numerical experiments we show that the band projection is superior to the crude truncation used in [19], in particular when applied to the subgrid model H_j instead of $\bar{L}_j$. We also extend the truncation algorithm to two dimensional problems and show numerical results.

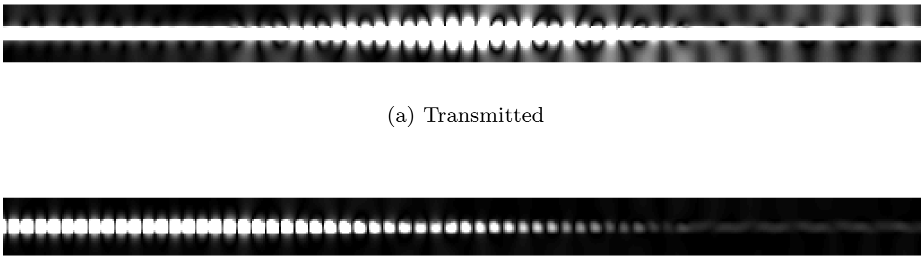
1.3.3 Simulation of Optical Filter

In the final part of the thesis we apply the subgrid technique to a problem in fiber optics. We study propagation of high frequency electromagnetic waves in a waveguide with a small scale structure of gratings. In this case the problems of small scales stem from both data and the geometry.

The waveguide structure is a narrow-band filter used as a component in a variety of communication applications, to selectively transmit, reflect, or detect signals. See e.g. [40, 29].

We model the problem with the Helmholtz equation in one and two dimensions and show that our numerical solution behaves physically correctly. See Figure 1.3 for two typical solutions. We then use the homogenization technique presented in Section 1.3.2 and show that the homogenized operator can be truncated while maintaining good accuracy. Moreover, we show that in order to compute a good approximation of the filter's power transmission spectrum, it suffices to know the homogenized solutions.

Next, we exemplify how the homogenized operator can be analyzed and how it can be reused in multiple places in the computational domain.



(a) Transmitted

(b) Reflected

Figure 1.3. Optical filter with 32 grating slits. Signal is injected from the left and propagates to the right in the middle part of the waveguide. Gray scales indicate the amplitude, where dark stands for small amplitude and light means high amplitude. Solution shown at resonance when signal is transmitted (above), and in the stop band, when the filter reflects the wave (below).

We analyse the structure of the two-dimensional homogenized operator when we project the solution to just a few points in the y-direction. The operator can be identified with a discretization of a system of differential equations in one dimension. We derive two different models through this interpretation, compare them to the simple one-dimensional model used originally and conclude that they both give better result.

Finally, we show how to use the homogenized operator to derive subgrid models for the gratings and other parts of the waveguide. This is done for a small problem, with only a few gratings. Subsequently, a much larger problem is solved on a coarse mesh by using the subgrid models.

1.4 Contributions

The thesis is partially based on the following papers.

(i) B. Engquist and O. Runborg. Multi-phase computations in geometrical optics. *Journal of Computational and Applied Mathematics*, 74:175–192, 1996.

(ii) O. Runborg. Multi-phase computations in geometrical optics. *Licentiate's Thesis*. TRITA-NA-9609. NADA, KTH, 1996.

(iii) U. Andersson, B. Engquist, G. Ledfelt and O. Runborg. A Contribution to Wavelet-Based Subgrid Modeling. 1996. To appear in *Applied and Computational Harmonic Analysis*.

(iv) B. Engquist, P.-O. Persson and O. Runborg. Simulation of Optical Resonance Filter Using Wavelet-Based Homogenization. To appear.

Chapter 2

Multiphase Methods in Geometrical Optics

2.1 Geometrical Optics

In this chapter we study the Cauchy problem for the scalar wave equation

$$u_{tt} - c^2\Delta u \quad = \quad 0, \qquad (t, \boldsymbol{x}) \in \mathbb{R}^+ \times \mathbb{R}^2 \tag{2.1}$$

together with initial data. Here $c = c(\boldsymbol{x})$ is the local wave velocity of the medium. We also define the *index of refraction* as $\eta = c_0/c$ with the reference velocity c_0 (e.g. the speed of light in vacuum). When c is constant (2.1) admits the simple plane wave solution

$$u(t, \boldsymbol{x}) = Ae^{i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)}, \qquad |\boldsymbol{k}| = \frac{\omega}{c} \tag{2.2}$$

where $\boldsymbol{k}$ is the wave vector giving the direction of propagation and A is a constant representing the amplitude (both determined by initial data). For more complicated waves and when c is not constant, we need to replace $\boldsymbol{k} \cdot \boldsymbol{x}/\omega - t$ by a general *phase function* ϕ and also permit the amplitude to depend on time and space. Hence, (2.1) has solutions of the type

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

The level curves of ϕ corresponds to what we can interpret as the wavefront of a propagating wave, cf. Figure 2.1. Since (2.1) is linear the superposition principle is valid and a sum of solutions is itself a solution.

We are interested in the case when the solution to (2.1) is a finite sum of terms like (2.3), and when the frequency ω is very large. These solutions contain length and timescales that become very small as the frequency increases. In the direct numerical solution of (2.1) a fixed number of grid points per wavelength and dimension is needed to maintain constant accuracy. The work hence 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 we accept the loss of typical wave phenomena in the solution, such as diffraction and interference, 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. They correspond to a PDE, an ODE and a kinetic viewpoint. We will subsequently use the kinetic formulation to treat the multiple phase problem.

2.1.1 PDE Model

With this approach we derive PDEs for the phase and the amplitude functions 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 in principle be easier to compute. In the homogeneous case (2.2), for instance, $\phi = \boldsymbol{k} \cdot \boldsymbol{x}/\omega - t$ stays nonoscillating and bounded independently of ω, since $|\boldsymbol{k}| = \omega/c$.

To begin with we assume that a so called WKB-expansion can be used to describe the solution to (2.1),

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

This ansatz is a slight generalization of the form (2.3) that also includes a series expansion in powers of $1/\omega$ of the amplitude. We now enter the expression (2.4) into (2.1) and sum terms of the same order in 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 consider the one with a plus sign in what follows. 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 for the remaining amplitude terms,

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

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 scalar wave equation (2.1) we could use the Helmholtz equation

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

and enter the ansatz (2.4) into (2.8). In this way we would get an alternative version of the pair (2.5, 2.6),

$$|\nabla\phi| = \pm 1/c, \qquad 2\nabla\phi\cdot\nabla A_0 + \Delta\phi A_0 = 0. \tag{2.9}$$

We will, however, mostly use the form (2.5, 2.6).

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. Similar to the case of hyperbolic conservation laws, extra conditions are needed for this type of equations to have a unique solution. These were given in [17] 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 Hamilton-Jacobi equations can develop discontinuities in the gradient of the solution, just like shocks appear in the solution to conservation laws. In the case of the eikonal equation the points of discontinuity corresponds to points where the physically correct phase solution should become multivalued, cf. Figure 2.1. We notice that since the transport equation (2.6) has a lower order term involving $\Delta\phi$, we can expect blowup of A_0 at these points.

2.1.2 ODE Model

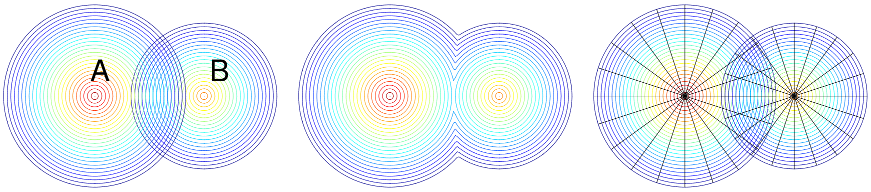
Another formulation of geometrical optics is *ray tracing* which gives the solution via ODEs. This is closely related to the method of characteristics for (2.5). Let $\boldsymbol{x}(t)$ be a characteristic of (2.5) defined by

$$0 = \frac{d}{dt}\phi(t,\boldsymbol{x}(t)) = \phi_t + \nabla\phi\cdot\frac{d\boldsymbol{x}}{dt}. \tag{2.10}$$

Let us introduce the *ray vector* $\boldsymbol{p} \equiv \nabla\phi$. From (2.10) and by differentiating (2.5) with respect to x and y we get

$$\frac{d\boldsymbol{x}}{dt} = \frac{c}{|\boldsymbol{p}|}\boldsymbol{p}, \tag{2.11}$$

$$\frac{d\boldsymbol{p}}{dt} = -|\boldsymbol{p}|\nabla c. \tag{2.12}$$



(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. The left figure shows the physically correct solution with two superimposed waves. Level curves of their phase functions are plotted. The middle figure shows level curves of ϕ in the viscosity solution of the eikonal equation (1.4). Note how the superposition principle does not hold. Instead, the first arriving wave takes precedence over the second at each point. The right figure shows a ray traced solution.

From this, and some tedious algebra, follows that $d(\boldsymbol{p}\cdot d\boldsymbol{x}/dt)/dt = 0$ and hence $\boldsymbol{p}\cdot d\boldsymbol{x}/dt = \text{const} = c_0$, which is the same constant as in the definition of the index of refraction. Multiplication of (2.11) by $\boldsymbol{p}$ shows that $|\boldsymbol{p}| = c_0/c = \eta$. For simplicity we will henceforth let $c_0 = 1$. By scaling time $t = \eta(\boldsymbol{x}(t))^2\tau$ we get a simple ODE for $\boldsymbol{x}$,

$$\frac{d^2\boldsymbol{x}}{d\tau^2} = \frac{1}{2}\nabla\eta^2. \tag{2.13}$$

As was discussed in Section 2.1.1 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 characteristics, however, does not have this problem and we can extend their validity to all t, see Figure 2.1. The characteristics are usually called *rays*. Note that since $d\boldsymbol{x}/dt \parallel \boldsymbol{p} = \nabla\phi$ the rays are always orthogonal to the level curves of ϕ. If $\eta = \text{const}$ the rays are just straight lines by (2.13).

There is another interpretation of (2.13). It is equivalent to

$$\frac{d\boldsymbol{x}}{d\tau} = \nabla_p H(\boldsymbol{x},\boldsymbol{p}), \qquad \frac{d\boldsymbol{p}}{d\tau} = -\nabla_x H(\boldsymbol{x},\boldsymbol{p}), \qquad H = \frac{|\boldsymbol{p}|^2 - \eta^2}{2}. \tag{2.14}$$

This is a Hamiltonian system and we can hence identify the rays with the trajectories of particles following the dynamics given by H in (2.14). Since $d\boldsymbol{x}/dt \parallel \boldsymbol{p}$ and $|\boldsymbol{p}| = 1/|d\boldsymbol{x}/dt|$ this viewpoint explains why $\boldsymbol{p}$ is also often called the *slowness* vector.

Suppose now that η is only a function of x and that $\boldsymbol{p} = \eta(\cos\theta, \sin\theta)$. Then, dividing (2.12) by dx/dt gives $d(\eta^2\cos^2\theta)/dx = d(\eta^2)/dx$, or $|\eta\sin\theta| = \text{const}$

(as long as $dx/dt \neq 0$). This is Snell's law of refraction, usually written in the form

$$\eta_1 \sin\theta_1 = \eta_2 \sin\theta_2 \tag{2.15}$$

and it is valid also in the limit of a discontinuous η. By entering (2.10) into (2.6), one get the corresponding law for the amplitude,

$$\eta_1 A_0^2 \cos\theta_1 = \eta_2 A_0^2 \cos\theta_2. \tag{2.16}$$

Solving (2.10) or (2.13) is called ray tracing. Rays can cross without problems but the solution is only obtained along the rays. It can be difficult to cover the full domain with rays, [58]. There will often be shadow zones where the field cannot be resolved. It can also be hard to determine the amplitude accurately.

2.1.3 Kinetic Model

Finally, we can adopt a purely kinetic viewpoint which will prove to be useful in the next section when we consider the multiple phase problem. This is based on the interpretation that rays are trajectories of particles following an Hamiltonian dynamics. We introduce the phase space $(t, \boldsymbol{x}, \boldsymbol{p})$ where $\boldsymbol{p}$ is the slowness vector defined above in Section 2.1.2. The evolution of a particle in this space is governed by

$$\frac{d\boldsymbol{x}}{dt} = \nabla_p H(\boldsymbol{x}, \boldsymbol{p}), \qquad \frac{d\boldsymbol{p}}{dt} = -\nabla_x H(\boldsymbol{x}, \boldsymbol{p}), \qquad H = \frac{1}{2}\Big(\frac{|\boldsymbol{p}|^2}{\eta^2} - 1\Big) \tag{2.17}$$

which is the unscaled version of (2.14). Now, let $f(t, \boldsymbol{x}, \boldsymbol{p})$ be a particle density function. By the Liouville theorem f will satisfy the Liouville equation

$$f_t + \nabla_p H \cdot \nabla_x f - \nabla_x H \cdot \nabla_p f = 0. \tag{2.18}$$

Using (2.17), we get

$$f_t + \frac{1}{\eta^2}\boldsymbol{p} \cdot \nabla_x f + \frac{|\boldsymbol{p}|^2}{\eta^3}\nabla_x \eta \cdot \nabla_p f = 0 \tag{2.19}$$

and since $|\boldsymbol{p}| = \eta$ this reduces to the Vlasov type equation

$$f_t + \frac{1}{\eta^2}\boldsymbol{p} \cdot \nabla_x f + \frac{1}{\eta}\nabla_x \eta \cdot \nabla_p f = 0. \tag{2.20}$$

Note, that with these definition $f(t, \boldsymbol{x}, \boldsymbol{p})$ must be zero whenever $|\boldsymbol{p}| \neq \eta$. Also, if $\eta = \text{const} = 1$, the equation (2.20) 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).

The transport equation (2.20) can also be derived directly from the scalar wave equation (2.1) using H-measures, see Section 3.3 of [57].

2.2 Derivation of the Moment Equations

In this section we will derive the system of PDEs that follows from the kinetic model and the assumption that a maximum of N rays pass through any given point in space. In order to solve the full equation (2.20) by direct numerical methods all independent variables (six in 3D) need to be discretized. This would require unrealistic computational time. Instead, we observe that when f is of a simple form in $\boldsymbol{p}$, (2.20) can be transformed into a finite system of moment equations in the reduced space $(t, \boldsymbol{x})$.

2.2.1 Background

The technique to approximate a kinetic transport equation by a finite system of moment equations is classical. See for instance [27] and more recently [39]. 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.

The opposite approach was taken by Lions, Perthame and Tadmor [43, 44] who demonstrated that a wide class of scalar multidimensional conservation laws, and a smaller class of 2×2 systems, can be seen as the first (and second in the system case) moment equation of a certain kinetic equation, with some constraint on the form of the dependent variable. They also showed the equivalence between the solution of the kinetic formulation and the entropy solution of the conservation laws.

Brenier and Corrias, [12], made the connection with the concept of multivalued solutions of a conservation law. They note that the multivalued solutions of the inviscid Burgers equation,

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \qquad u(0,x) = u_0(x), \tag{2.21}$$

is given by the solution to the free transport equation

$$f_t + v f_x = 0, \qquad f = f(t,x,v) \tag{2.22}$$

with initial data $f(0,x,v) = H(u_0(x) - v)H(v)$. The solution f will be on the form

$$f(t,x,v) = \sum_{j=1}^{K(t)} (-1)^{j-1} H(u_j(t,x) - v) \tag{2.23}$$

where $K(t)$ is an integer representing the number of branches of the solution at time t and $u_j(t,x)$ are the branches of the multivalued solution. Their key observation is that if the maximum number of branches for a problem is known a priori, say $K(t) \le N$, then the exact solution of (2.22) can be recovered from the solution of the first N moment equations. The moments are defined as

$$m_k(t,x) = \int_{\mathbb{R}_v} v^k f(t,x,v)\,dv, \qquad k \in \mathbb{Z}^+. \tag{2.24}$$

Multiplying (2.22) by v^k and integrating over all v we get the moment equations

$$(m_k)_t + (m_{k+1})_x = 0, \tag{2.25}$$

which is valid for all $k \in \mathbb{Z}^+$. By the special form of f in (2.23) it is possible to write m_N as a function of $m_0, \ldots, m_{N-1}$ and thus close the system (2.25). This closure can also be seen as an entropy maximization over all functions f of a particular class with prescribed moments $m_0, \ldots, m_{N-1}$.

Brenier and Corrias go on to generalize the results in [43, 44] by defining the *entropy K-multivalued* solutions as the solution to the system of the first K moment equations, closed as above, together with a set of entropy inequalities. Then the solution is equivalent to the solution of the transport equation

$$f_t + v f_x = (-1)^{K-1} \frac{\partial^K \mu}{\partial v^K}, \tag{2.26}$$

with the constraint that f stays on the form given in (2.23) with $K(t) = K$. Here $\mu(t, x, v)$ is some nonnegative Radon measure. When $K = 1, 2$ this is the result in [43, 44].

The entropy K-multivalued solution is the same as the classical multivalued solution as long as the number of branches is less than or equal to K. For $K = 1$ this is when shocks appear. Also the measure μ is zero for classical solutions.

2.2.2 Moment Equations for Geometrical Optics

We will now treat the kinetic transport equation (2.20) in a way similar to the treatment of (2.22) in [12]. There are two major differences. First, $\boldsymbol{p}$, which corresponds to the v-variable in [12], is two-dimensional. Second, the natural closure assumption, motivated by the physics of the problem, is to use delta functions in (2.23) instead of Heaviside functions. However, first, since $f{=}0$ when $|\boldsymbol{p}| \neq \eta$, we can think of $\boldsymbol{p}$ as an element of S^1, thus one-dimensional, and second, it is also possible to use Heaviside functions as we do in Section 2.4.

Let us define the moments m_{ij}, with $\boldsymbol{p} = (p_1, p_2)^T$, as

$$m_{ij} = \int_{\mathbb{R}^2_p} p_1^i p_2^j f \, d\boldsymbol{p}. \tag{2.27}$$

Next, multiply (2.20) by $\eta^2 p_1^i p_2^j$ and integrate over $\mathbb{R}^2$ with respect to $\boldsymbol{p}$, so that

$$\begin{aligned} \eta^2 \partial_t \int p_1^i p_2^j f \, d\boldsymbol{p} + \partial_x \int p_1^{i+1} p_2^j f \, d\boldsymbol{p} + \partial_y \int p_1^i p_2^{j+1} f \, d\boldsymbol{p} & \\ - \eta \eta_x \int i p_1^{i-1} p_2^j f \, d\boldsymbol{p} - \eta \eta_y \int j p_1^i p_2^{j-1} f \, d\boldsymbol{p} &= 0, \end{aligned} \tag{2.28}$$

if f has compact support in $\boldsymbol{p}$. Using the definition (2.27) we see that formally m_{ij} will satisfy the infinite system of moment equations

$$(\eta^2 m_{ij})_t + (m_{i+1,j})_x + (m_{i,j+1})_y = i \eta \eta_x m_{i-1,j} + j \eta \eta_y m_{i,j-1}, \tag{2.29}$$

valid for all $i, j \geq 0$. For uniformity in notation we have defined $m_{i,-1} = m_{-1,i} = 0$, $\forall i$. To be a little bit more precise, let $f \in C(\mathbb{R}^+; \mathcal{M}(\mathbb{R}_x^2 \times \mathbb{R}_p^2))$ satisfy (2.20) in $\mathcal{D}'(\mathbb{R}^+ \times \mathbb{R}_x^2 \times \mathbb{R}_p^2)$ and assume that there is a compact set $K \subset \mathbb{R}^2$ such that

$$\mathrm{supp}_p f(t, \boldsymbol{x}, \cdot) \subset K, \qquad \forall t, \boldsymbol{x} \tag{2.30}$$

and that $\eta(\boldsymbol{x}) \in C^\infty(\mathbb{R}_x^2)$. Now, take $\phi \in C_c^1(\mathbb{R}^+ \times \mathbb{R}_x^2)$, and a sequence of approximating functions $\phi^n \in C_c^\infty(\mathbb{R}^+ \times \mathbb{R}_x^2)$ such that $\phi^n \to \phi$ uniformly. By Urysohn's Lemma, there is a $\psi \in C_c^\infty(\mathbb{R}_p^2)$ with $\psi = 1$ on K. Then for $i, j \in \mathbb{Z}^+$

$$\phi^n p_1^i p_2^j \psi(\boldsymbol{v}) \in C_c^\infty(\mathbb{R}^+ \times \mathbb{R}_x^2 \times \mathbb{R}_p^2), \tag{2.31}$$

and by assumption

$$\langle f, \eta^2 \phi_t^n p_1^i p_2^j \psi \rangle + \langle f, \boldsymbol{p} \cdot \nabla_x \phi^n p_1^i p_2^j \psi \rangle + \langle f, \phi^n \eta \nabla_x \eta \cdot \nabla_p (p_1^i p_2^j \psi) \rangle = 0. \tag{2.32}$$

Since $\langle f, \cdot \rangle$ is continuous on C_0 we can pass to the limit $\phi^n \to \phi$ in (2.32). Hence, after dropping ψ,

$$\begin{aligned} 0 \quad &= \quad \int \eta^2 \phi_t p_1^i p_2^j f + \phi_x p_1^{i+1} v_2^j f + \phi_y p_1^i p_2^{j+1} f \\ &\qquad + \phi i \eta_x p_1^{i-1} p_2^j f + \phi j \eta_y p_1^i p_2^{j-1} f d\boldsymbol{p} d\boldsymbol{x} dt \\ &= \quad \int \phi_t \eta^2 m_{ij} + \phi_x m_{i+1,j} + \phi_y m_{i,j+1} \\ &\qquad + \phi i \eta_x m_{i-1,j} + \phi j m_{i,j-1} d\boldsymbol{p} d\boldsymbol{x} dt. \end{aligned} \tag{2.33}$$

As ϕ was arbitrary, $m_{ij} \in C(\mathbb{R}^+; \mathcal{M}(\mathbb{R}_x^2))$ is a weak solution (in the sense of measures) to (2.29).

The system (2.29) is not closed. If truncated at finite i and j, there are more unknowns than equations. To close the system we will make specific assumptions on the form of the density function f. First we will consider the case when f is a weighted sum of delta functions in $\boldsymbol{p}$ and second, the case when f is a sum of Heaviside functions. Both cases correspond to the assumption of a finite number of rays at each point in time and space.

In what follows we will assume that η is a smooth function. Since its derivative enters in the right hand side of all the equations in (2.29) (except the one when $i = j = 0$), this is necessary to enable us to treat the equations numerically.

2.3 Closure with Delta Functions

To close (2.29) we assume that for fixed values of $\boldsymbol{x}$ and t, the particle density f is non-zero at a maximum of N points, and only when $|\boldsymbol{p}| = \eta(\boldsymbol{x})$. Thus f can be written

$$f(t, \boldsymbol{x}, \boldsymbol{p}) = \sum_{k=1}^N g_k \cdot \delta(\boldsymbol{p} - \boldsymbol{p}_k), \qquad \boldsymbol{p}_k \equiv \eta \begin{pmatrix} \cos\theta_k \\ \sin\theta_k \end{pmatrix}. \tag{2.34}$$

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.

2.3.1 Equations

Inserting (2.34) into (2.27) yields

$$m_{ij} = \sum_{k=1}^{N} \eta^{i+j} g_k \cos^i \theta_k \sin^j \theta_k. \tag{2.35}$$

Since η was assumed to be smooth, we can also introduce the scaling $\eta^{i+j}\tilde{m}_{ij} = m_{ij}$,

$$\tilde{m}_{ij} = \sum_{k=1}^{N} g_k \cos^i \theta_k \sin^j \theta_k \tag{2.36}$$

which transforms (2.29) to

$$\begin{aligned}(\eta^2 \tilde{m}_{ij})_t &+ (\eta \tilde{m}_{i+1,j})_x + (\eta \tilde{m}_{i,j+1})_y \\ &= i\eta_x \tilde{m}_{i-1,j} + j\eta_y \tilde{m}_{i,j-1} - (i+j)(\eta_x \tilde{m}_{i+1,j} + \eta_y \tilde{m}_{i,j+1}).\end{aligned} \tag{2.37}$$

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

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 (2.37). Given the equations for a set of $2N$ moments one should be able to write the remaining moments of 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 not the sign of the moments.

We choose here the equations for the moments $m_{2k-1,0}$ and $m_{0,2k-1}$ with $k = 1, \ldots, N$. This system can be essentially closed for all N (see Section 2.3.2). Those equations take the following form

$$\begin{aligned}(\eta^2 m_{2k-1,0})_t &+ (\eta m_{2k,0})_x + (\eta m_{2k-1,1})_y \\ &= (2k-1)(\eta_x m_{2k-2,0} - \eta_x m_{2k,0} - \eta_y m_{2k-1,1}),\end{aligned} \tag{2.38}$$

$$\begin{aligned}(\eta^2 m_{0,2k-1})_t &+ (\eta m_{1,2k-1})_x + (\eta m_{0,2k})_y \\ &= (2k-1)(\eta_y m_{0,2k-2} - \eta_x m_{1,2k-1} - \eta_y m_{0,2k}).\end{aligned} \tag{2.39}$$

We introduce new variables,

$$\begin{aligned}\boldsymbol{u} &= (u_1, u_2, u_3, u_4 \ldots, u_{2N-1}, u_{2N})^T \\ &\equiv (g_1 \cos\theta_1, g_1 \sin\theta_1, g_2 \cos\theta_2, g_2 \sin\theta_2, \ldots, g_N \cos\theta_N, g_N \sin\theta_N)^T.\end{aligned} \tag{2.40}$$

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

$$\boldsymbol{F}_0(\eta^2 \boldsymbol{u})_t + \boldsymbol{F}_1(\eta \boldsymbol{u})_x + \boldsymbol{F}_2(\eta \boldsymbol{u})_y = \boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y), \tag{2.41}$$

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

$$\hat{\boldsymbol{m}} = (m_{10}, m_{01}, m_{30}, m_{03}, \dots, m_{2N-1,0}, m_{0,2N-1})^T, \tag{2.42}$$

we could write (2.41) as

$$(\eta^2\hat{\boldsymbol{m}})_t + \boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}(\eta\hat{\boldsymbol{m}})_x + \boldsymbol{F}_2 \circ \boldsymbol{F}_0^{-1}(\eta\hat{\boldsymbol{m}})_y = \boldsymbol{K}(\boldsymbol{F}_0^{-1}(\hat{\boldsymbol{m}}), \eta_x, \eta_y). \tag{2.43}$$

Since the angles θ_k remain unaffected when $\boldsymbol{u}$ is scaled by a constant, all $\boldsymbol{F}_k$ and $\boldsymbol{K}$ are homogeneous, $\boldsymbol{F}_k(\alpha\boldsymbol{u}) = \alpha\boldsymbol{F}_k(\boldsymbol{u})$, $\boldsymbol{K}(\alpha\boldsymbol{u}, \eta_x, \eta_y) = \alpha\boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y)$ for all $\alpha \in \mathbb{R}$. The source term, $\boldsymbol{K}$, always vanishes for constant η. Also note that if we let I be the 2×2 identity matrix, and

$$\begin{aligned} C &= \operatorname{diag}(\cos\theta_1 I, \cos\theta_2 I, \dots, \cos\theta_N I), && (2.44)\\ S &= \operatorname{diag}(\sin\theta_1 I, \sin\theta_2 I, \dots, \sin\theta_N I), && (2.45)\\ & && (2.46)\end{aligned}$$

we get

$$\boldsymbol{F}_1(\boldsymbol{u}) = \boldsymbol{F}_0(C\boldsymbol{u}), \qquad \boldsymbol{F}_2(\boldsymbol{u}) = \boldsymbol{F}_0(S\boldsymbol{u}). \tag{2.47}$$

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

$$\boldsymbol{F}_1 = \begin{pmatrix} \frac{u_1^2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}, \quad \boldsymbol{F}_2 = \begin{pmatrix} \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_2^2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}, \quad \boldsymbol{K} = \begin{pmatrix} \frac{\eta_x u_2^2 - \eta_y u_1 u_2}{\sqrt{u_1^2+u_2^2}} \\ \frac{\eta_y u_1^2 - \eta_x u_1 u_2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}. \tag{2.48}$$

For $N = 2$ the functions are

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

$$\boldsymbol{F}_2 = \begin{pmatrix} \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} + \frac{u_3 u_4}{\sqrt{u_3^2+u_4^2}} \\ \frac{u_2^2}{\sqrt{u_1^2+u_2^2}} + \frac{u_4^2}{\sqrt{u_3^2+u_4^2}} \\ \frac{u_1^2 u_2}{(u_1^2+u_2^2)^{3/2}} + \frac{u_3^2 u_4}{(u_3^2+u_4^2)^{3/2}} \\ \frac{u_2^3}{(u_1^2+u_2^2)^{3/2}} + \frac{u_4^3}{(u_3^2+u_4^2)^{3/2}} \end{pmatrix}, \quad \boldsymbol{K} = \begin{pmatrix} \frac{\eta_x u_2^2 - \eta_y u_1 u_2}{\sqrt{u_1^2+u_2^2}} + \frac{\eta_x u_4^2 - \eta_y u_3 u_4}{\sqrt{u_3^2+u_4^2}} \\ \frac{\eta_y u_1^2 - \eta_x u_1 u_2}{\sqrt{u_1^2+u_2^2}} + \frac{\eta_y u_3^2 - \eta_x u_3 u_4}{\sqrt{u_3^2+u_4^2}} \\ \frac{\eta_x u_1^2 u_2^2 - \eta_y u_1^3 u_2}{(u_1^2+u_2^2)^{3/2}} + \frac{\eta_x u_3^2 u_4^2 - \eta_y u_3^3 u_4}{(u_3^2+u_4^2)^{3/2}} \\ \frac{\eta_y u_1^2 u_2^2 - \eta_x u_1 u_2^3}{(u_1^2+u_2^2)^{3/2}} + \frac{\eta_y u_3^2 u_4^2 - \eta_x u_3 u_4^3}{(u_3^2+u_4^2)^{3/2}} \end{pmatrix}. \tag{2.49}$$

Remark 2.1. 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 (2.20), use the Boltzmann equation

$$f_t + \boldsymbol{p} \cdot \nabla_x f = Q(f, f), \qquad (\boldsymbol{x}, \boldsymbol{p}) \in \mathbb{R}^2 \times \mathbb{R}^2, \tag{2.50}$$

where $Q(f, f)$ is the collision operator. Moreover, instead of the closure assumption (2.34), assume f is a *Maxwellian*,

$$f(t, \boldsymbol{x}, \boldsymbol{p}) = \frac{\rho(t, \boldsymbol{x})}{2\pi T(t, \boldsymbol{x})} e^{-\frac{|\boldsymbol{p} - \boldsymbol{u}(t,\boldsymbol{x})|^2}{2T(t,\boldsymbol{x})}}, \qquad \boldsymbol{u} = \begin{pmatrix} u \\ v \end{pmatrix}. \tag{2.51}$$

The first moment equations of (2.50) are the same as (2.20) with $\eta \equiv 1$ by the special form of Q. If we now pick the following equations from (2.37)

$$\begin{pmatrix} m_{00} \\ m_{10} \\ m_{01} \\ m_{20} + m_{02} \end{pmatrix}_t + \begin{pmatrix} m_{10} \\ m_{20} \\ m_{11} \\ m_{30} + m_{12} \end{pmatrix}_x + \begin{pmatrix} m_{01} \\ m_{11} \\ m_{02} \\ m_{21} + m_{03} \end{pmatrix}_y = 0 \tag{2.52}$$

and write them in terms of the unknowns $\rho, \boldsymbol{u}, T$ in (2.51) and $E \equiv \rho(\frac{1}{2}|\boldsymbol{u}|^2 + \frac{3}{2}T)$, we get

$$\begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix}_t + \begin{pmatrix} \rho u \\ \rho(u^2 + T) \\ \rho uv \\ u(E + \rho T) \end{pmatrix}_x + \begin{pmatrix} \rho v \\ \rho uv \\ \rho(v^2 + T) \\ v(E + \rho T) \end{pmatrix}_y = 0, \tag{2.53}$$

the compressible Euler equations for a perfect monoatomic gas.

Also note that if we formally let $T \to 0$ in (2.51) we recover the assumption (2.34) with $N = 1$ and without the restriction on $|\boldsymbol{p}_1|$. The same, formal, limit of (2.53) gives the equations of pressureless gases,

$$\begin{pmatrix} \rho \\ \rho u \\ \rho v \end{pmatrix}_t + \begin{pmatrix} \rho u \\ \rho u^2 \\ \rho uv \end{pmatrix}_x + \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 \end{pmatrix}_y = 0. \tag{2.54}$$

In Section 2.5.3 we will further discuss the relationship between (2.41) with $N = 1$ and (2.54).

2.3.2 Properties of the Flux Function

In this section we analyze the flux functions and source

$$\boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}(\hat{\boldsymbol{m}}), \qquad \boldsymbol{F}_2 \circ \boldsymbol{F}_0^{-1}(\hat{\boldsymbol{m}}), \qquad \boldsymbol{K}(\boldsymbol{F}_0^{-1}(\hat{\boldsymbol{m}}), \eta_x, \eta_y). \tag{2.55}$$

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 we also show that they are continuous. It will be convenient to work with complex versions of our variables and we start by introducing the isomorphism $\mathcal{A} : \mathbb{R}^{2N} \mapsto \mathbb{C}^N$ as

$$\mathcal{A}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{2N-1} \\ x_{2N} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_3 \\ \vdots \\ x_{2N-1} \end{pmatrix} + i\begin{pmatrix} x_2 \\ x_4 \\ \vdots \\ x_{2N} \end{pmatrix}. \tag{2.56}$$

We also define two useful continuous functions, namely $\boldsymbol{\sigma} : \mathbb{C} \to \mathbb{C}^N$, given by

$$\boldsymbol{\sigma}(z) = \begin{pmatrix} z \\ z^{-3} \\ \vdots \\ z^{(2N-1)(-1)^{N+1}} \end{pmatrix}, \qquad \text{for } z \in \mathbb{C}, \tag{2.57}$$

and $Q_0 : \mathbb{C}^N \mapsto \mathbb{C}^N$, with $\boldsymbol{w} = (w_1, \dots, w_N)^T$,

$$Q_0(\boldsymbol{w}) = \begin{pmatrix} | & | & & | \\ \boldsymbol{\sigma}\left(\frac{w_1}{|w_1|}\right) & \boldsymbol{\sigma}\left(\frac{w_2}{|w_2|}\right) & \dots & \boldsymbol{\sigma}\left(\frac{w_N}{|w_N|}\right) \\ | & | & & | \end{pmatrix} \begin{pmatrix} |w_1| \\ \vdots \\ |w_N| \end{pmatrix} \tag{2.58}$$

With this notation we can rewrite the equation $\boldsymbol{F}_0(\boldsymbol{u}) = \hat{\boldsymbol{m}}$ as follows. In general we have

$$(\cos\theta + (-1)^{k+1} i \sin\theta)^{2k-1} = \sum_{\ell=0}^{k-1} w_{k,k-\ell}(\cos^{2(k-\ell)-1}\theta + i\sin^{2(k-\ell)-1}\theta), \tag{2.59}$$

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

$$Q_0(\mathcal{A}\boldsymbol{u}) = W\mathcal{A}\hat{\boldsymbol{m}} = W\mathcal{A}\boldsymbol{F}_0(\boldsymbol{u}). \tag{2.60}$$

There is hence an equivalence between for instance $\boldsymbol{F}_0(\boldsymbol{u}) = \hat{\boldsymbol{m}}$ and $Q_0(\boldsymbol{w}) = \boldsymbol{m}$, where $\boldsymbol{w} = \mathcal{A}\boldsymbol{u}$ and $\boldsymbol{m} = W\mathcal{A}\hat{\boldsymbol{m}}$. In what follows will we use the complex equations. Before we go on we need two lemmas.

Lemma 2.2. *Let z_k, $k = 1, \dots, N'$ be complex numbers such that $|z_k| = 1$ Define the corresponding vectors $\boldsymbol{z}_k = \boldsymbol{\sigma}(z_k)$ with $N' \leq 2N$ and $\boldsymbol{\sigma}$ as in (2.57). Then $\boldsymbol{z}_k \in \mathbb{C}^N$ are linearly independent over $\mathbb{R}$ if and only if*

$$z_k^2 \neq z_\ell^2, \qquad k \neq \ell. \tag{2.61}$$

Proof. The necessity is obvious, since if $z_k^2 = z_\ell^2$ for some k, ℓ, then $\boldsymbol{z}_k = \pm \boldsymbol{z}_\ell$. To show that (2.61) is a sufficient condition, let (2.61) hold and take $N' = 2N$. Suppose that $\boldsymbol{z}_k$ are linearly dependent. Then we can find α_k such that

$$\sum_{k=1}^{2N} \alpha_k \boldsymbol{z}_k = 0, \qquad \alpha_k \in \mathbb{R}. \tag{2.62}$$

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

$$A = \begin{pmatrix} | & | & & | \\ \Re(\boldsymbol{z}_1) & \Re(\boldsymbol{z}_2) & \cdots & \Re(\boldsymbol{z}_{2N}) \\ | & | & & | \\ | & | & & | \\ \Im(\boldsymbol{z}_1) & \Im(\boldsymbol{z}_2) & \cdots & \Im(\boldsymbol{z}_{2N}) \\ | & | & & | \end{pmatrix}, \qquad A \in \mathbb{R}^{2N \times 2N} \tag{2.63}$$

we have the equivalent formulation

$$A\boldsymbol{\alpha} = 0, \qquad \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{2N})^T, \tag{2.64}$$

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

$$\Re(\boldsymbol{z}_k) = \frac{1}{2}(\boldsymbol{\sigma}(z_k) + \boldsymbol{\sigma}({z_k}^{-1})), \qquad \Im(\boldsymbol{z}_k) = \frac{1}{2i}(\boldsymbol{\sigma}(z_k) - \boldsymbol{\sigma}({z_k}^{-1})). \tag{2.65}$$

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

$$\frac{1}{2}\sum_{\ell=1}^{N} \beta_\ell (z_k^{2\ell-1} + \frac{1}{z_k^{2\ell-1}}) + \frac{1}{2i}\sum_{\ell=1}^{N} (-1)^{\ell+1} \beta_{\ell+N} (z_k^{2\ell-1} - \frac{1}{z_k^{2\ell-1}}) = 0, \tag{2.66}$$

for $k = 1, \dots, 2N$. Since $z_k \neq 0$ we can multiply (2.66) by z_k^{2N-1} and by defining the polynomial

$$P_{\boldsymbol{\beta}}(z) = \frac{1}{2}\sum_{\ell=1}^{N} \beta_\ell (z^{\ell+N-1} + z^{N-\ell}) + \frac{1}{2i}\sum_{\ell=1}^{N} (-1)^{\ell+1} \beta_{\ell+N} (z^{\ell+N-1} - z^{N-\ell}), \tag{2.67}$$

we have that (2.64) implies

$$P_{\boldsymbol{\beta}}(z_k^2) = 0, \qquad k = 1, \dots, 2N \tag{2.68}$$

for some $\boldsymbol{\beta} \neq 0$. But since the degree of $P_{\boldsymbol{\beta}}$ is at most $2N - 1$ it cannot have $2N$ distinct zeros if $\boldsymbol{\beta} \neq 0$. Therefore, there must exist k, ℓ such that $z_k^2 = z_\ell^2$, a contradiction. Hence, $\boldsymbol{z}_k$ are linearly independent over $\mathbb{R}$ when $N' = 2N$ and (2.61) holds.

Finally, if $N' < 2N$ and (2.61) holds, we can always find $2N - N'$ additional z_k such that (2.61) still holds. A subset of a set of linearly independent vectors are also linearly independent, from which the lemma follows. □

Moreover, we have this lemma from elementary analysis.

Lemma 2.3. *Let U be a compact metric space and f, g two continuous functions on U. Suppose $f : U \mapsto X$ and $g : U \mapsto Y$ where X, Y are metric spaces and $X = f(U)$. If the composition $f \circ g^{-1} : g(U) \mapsto X$ is injective then $g \circ f^{-1} : X \mapsto Y$ is continuous.*

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

$$f \circ g^{-1}(A) \cup f \circ g^{-1}(A^c) = f(g^{-1}(A)) \cup f(g^{-1}(A)^c) = f(U) = X. \tag{2.69}$$

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

We can now state the following theorem, showing that all moments are well defined by and depend continuously on the finite set of moments in (2.42), provided the moments do not correspond to a situation where rays meet head-on.

Theorem 2.4. *Let*

$$U_N = \{\boldsymbol{w} \in \mathbb{C}^N \bigm| \boldsymbol{w} = \{w_k\}_{k=1}^N,\ \frac{w_k}{|w_k|} + \frac{w_\ell}{|w_\ell|} \neq 0, \forall k, \ell \text{ where } w_k w_\ell \neq 0\} \tag{2.70}$$

and $M_N = Q_0(U_N)$, where Q_0 is the restriction to U_N of the Q_0 defined in (2.58). Moreover, set $m : U_N \mapsto \mathbb{C}$

$$m(\boldsymbol{w}) = \sum_{k=1}^{N} |w_k| h\left(\frac{w_k}{|w_k|}\right) \tag{2.71}$$

where $h : \mathbb{C} \mapsto \mathbb{C}$ is continuous. Then the composition $m \circ Q_0^{-1} : M_N \mapsto \mathbb{C}$ is well-defined and continuous.

Proof. We start by showing that $Q_0 \circ m^{-1}$ is injective. Let $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in U_N$ be such that $Q_0(\boldsymbol{w}) = Q_0(\tilde{\boldsymbol{w}})$. We need to show that $m(\boldsymbol{w}) = m(\tilde{\boldsymbol{w}})$. For $w_k, \tilde{w}_k \neq 0$ set $z_k = w_k/|w_k|$, $\boldsymbol{z}_k = \boldsymbol{\sigma}(z_k)$ and $\tilde{z}_k = \tilde{w}_k/|\tilde{w}_k|$, $\tilde{\boldsymbol{z}}_k = \boldsymbol{\sigma}(\tilde{z}_k)$. Then

$$\sum_{|w_k|>0} |w_k| \boldsymbol{z}_k = \sum_{|\tilde{w}_k|>0} |\tilde{w}_k| \tilde{\boldsymbol{z}}_k. \tag{2.72}$$

Now, let N' and $\tilde{N}'$ respectively be the number of distinct z_k and $\tilde{z}_k$ with $|w_k|, |\tilde{w}_k| > 0$. Without loss of generality we order the variables such that $z_{\ell_j} = \ldots = z_{\ell_{j+1}-1}$, with $1 = \ell_1 < \ldots < \ell_{N'+1}$, and similar for $\{\tilde{z}_k\}$. With this notation we get

$$\sum_{j=1}^{N'} \left(\sum_{k=\ell_j}^{\ell_{j+1}-1} |w_k| \right) \boldsymbol{z}_{\ell_j} = \sum_{j=1}^{\tilde{N}'} \left(\sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} |\tilde{w}_k| \right) \tilde{\boldsymbol{z}}_{\tilde{\ell}_j}. \tag{2.73}$$

The sets of numbers $\{z_{\ell_j}\}_{j=1}^{N'}$ and $\{\tilde{z}_{\tilde{\ell}_j}\}_{j=1}^{\tilde{N}'}$ now both satisfy (2.61) and the corresponding vectors are thus linearly independent over $\mathbb{R}$ by Lemma 2.2. But since $N' + \tilde{N}' \le 2N$, again by Lemma 2.2, there must exist j and k such that $z_{\ell_j}^2 = \tilde{z}_{\tilde{\ell}_k}^2$. Since $|w_k|, |\tilde{w}_k| > 0$, in fact we must have that $z_{\ell_j} = \tilde{z}_{\tilde{\ell}_k}$. By induction it follows that $N = \tilde{N}'$ and, possibly after some reordering,

$$\ell_j = \tilde{\ell}_j, \qquad z_{\ell_j} = \tilde{z}_{\tilde{\ell}_j}, \qquad \sum_{k=\ell_j}^{\ell_{j+1}-1} |w_k| = \sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} |\tilde{w}_k|, \qquad \forall j. \tag{2.74}$$

Thus, $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ are identical up to permutations and to the individual $|w_k|$ values. We now apply m to them.

$$\begin{aligned} m(\boldsymbol{w}) &= \sum_{j=1}^{N'} \sum_{k=\ell_j}^{\ell_{j+1}-1} |w_k| h(z_k) = \sum_{j=1}^{N'} h(z_{\ell_j}) \sum_{k=\ell_j}^{\ell_{j+1}-1} |w_k| = \sum_{j=1}^{\tilde{N}'} h(\tilde{z}_{\tilde{\ell}_j}) \sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} |\tilde{w}_k| \\ &= \sum_{j=1}^{\tilde{N}'} \sum_{k=\tilde{\ell}_j}^{\tilde{\ell}_{j+1}-1} |\tilde{w}_k| h(\tilde{z}_k) = m(\tilde{\boldsymbol{w}}). \end{aligned} \tag{2.75}$$

Hence, $Q_0 \circ m^{-1} : m(U_N) \mapsto M_N$ is injective and therefore $m \circ Q_0^{-1}$ is well-defined. Now, let $U_N^n \subset U_N$ be given by

$$\begin{aligned} U_N^n &= \{\boldsymbol{w} \in \mathbb{C}^N \mid \boldsymbol{w} = \{w_k\}_{k=1}^N,\ |w_k| \le n, \\ &\qquad \left| \frac{w_k}{|w_k|} + \frac{w_\ell}{|w_\ell|} \right| \ge 1/n, \forall k, \ell \text{ where } w_k w_\ell \ne 0\} \end{aligned} \tag{2.76}$$

and $M_N^n = Q_0(U_N^n)$. Clearly, $Q_0 \circ m^{-1}$ is injective also on this smaller space. Then, since U_N^n is compact, Q_0 and m are continuous, it follows from Lemma 2.3 that $m \circ Q_0^{-1}$ is a continuous function from M_N^n to $\mathbb{C}$. But this holds for all n and $M_N = \cup_{n=1}^{\infty} M_N^n$, giving the stated result. $\square$

Corollary 2.5. *Let $\boldsymbol{F}_k$ and $\boldsymbol{K}$ be the functions in (2.41), corresponding to the moment vector $\hat{\boldsymbol{m}}$ defined in (2.42). Let $\boldsymbol{F}_0|D_N$ be the restriction of $\boldsymbol{F}_0$ to the domain*

$$\begin{aligned} D_N = \mathcal{A}^{-1} U_N &= \{\boldsymbol{u} \in \mathbb{R}^{2N} \mid \boldsymbol{w} = \mathcal{A}\boldsymbol{u} = \{w_k\}_{k=1}^N, \\ &\qquad \frac{w_k}{|w_k|} + \frac{w_\ell}{|w_\ell|} \ne 0, \forall k, \ell \text{ where } w_k w_\ell \ne 0\} \end{aligned} \tag{2.77}$$

where $\mathcal{A}$ is given by (2.56). For $\hat{\boldsymbol{m}} \in \boldsymbol{F}_0(D_N)$ the flux functions

$$\boldsymbol{F}_1 \circ (\boldsymbol{F}_0|D_N)^{-1}(\hat{\boldsymbol{m}}), \qquad \boldsymbol{F}_2 \circ (\boldsymbol{F}_0|D_N)^{-1}(\hat{\boldsymbol{m}}) \tag{2.78}$$

are well defined and continuous. The same is true for $\boldsymbol{K}((\boldsymbol{F}_0|D_N)^{-1}(\hat{\boldsymbol{m}}), \eta_x, \eta_y)$ if $\eta \in C^1(\mathbb{R}^2)$.

Proof. Consider the general moment m_{ij} as a function of $\boldsymbol{u}$. Let Q_0 be the restriction to U_N in (2.70) of the Q_0 defined in (2.58) and let $W, \mathcal{A}$ be as in (2.60). Then, by (2.60),

$$m_{ij}((\boldsymbol{F}_0|D_N)^{-1}(\hat{\boldsymbol{m}})) = m_{ij}(\mathcal{A}^{-1}Q_0^{-1}(W\mathcal{A}\hat{\boldsymbol{m}})) \equiv m'_{ij} \circ Q_0^{-1}(\boldsymbol{m}) \tag{2.79}$$

where $\boldsymbol{m}(\hat{\boldsymbol{m}}) = W\mathcal{A}\hat{\boldsymbol{m}}$ and by (2.36)

$$m'_{ij}(\boldsymbol{w}) = \sum_{k=1}^{N} |w_k| \Re\left(\frac{w_k}{|w_k|}\right)^i \Im\left(\frac{w_k}{|w_k|}\right)^j \equiv \sum_{k=1}^{N} |w_k| h\left(\frac{w_k}{|w_k|}\right). \tag{2.80}$$

We can now use Theorem 2.4 since h defined above is continuous and conclude that $m'_{ij} \circ Q_0^{-1}(\boldsymbol{m})$ is well-defined and continuous on $M_N = Q_0(U_N)$. But (2.60) gives that $\boldsymbol{m}(\hat{\boldsymbol{m}})$ is a continuous map between $\boldsymbol{F}_0(D_N)$ and $Q_0(\mathcal{A}D_N) = M_N$. Hence, $m_{ij}((\boldsymbol{F}_0|D_N)^{-1}(\hat{\boldsymbol{m}}))$ is well defined and continuous on $\boldsymbol{F}_0(D_N)$. Since the elements of $\boldsymbol{F}_k$ are all equal to some m_{ij} the same is true for the compositions involving them. Finally, the same thing holds for $\boldsymbol{K}$ if the gradient of η is continuous, since the elements of $\boldsymbol{K}$ is a linear combination of m_{ij} involving η_x and η_y (see (2.38)). □

Remark 2.6. If we do not restrict $\boldsymbol{F}_0$ to D_N in Corollary 2.5 the result is false. Take for instance $\boldsymbol{u} = (-1\ 0\ 1\ 0)$ and $\tilde{\boldsymbol{u}} = (-2\ 0\ 2\ 0)$ for $N = 2$ so that $\boldsymbol{F}_0(\boldsymbol{u}) = \boldsymbol{F}_0(\tilde{\boldsymbol{u}}) = 0$. Then

$$\begin{pmatrix} 2 \\ 0 \\ 2 \\ 0 \end{pmatrix} = \boldsymbol{F}_1(\boldsymbol{u}) \neq \boldsymbol{F}_1(\tilde{\boldsymbol{u}}) = \begin{pmatrix} 4 \\ 0 \\ 4 \\ 0 \end{pmatrix}. \tag{2.81}$$

Furthermore, with a different choice of moments Corollary 2.5 does not necessarily hold either. For instance, with $N = 2$ and

$$\hat{\boldsymbol{m}} = (m_{10}, m_{01}, m_{20}, m_{02})^T, \tag{2.82}$$

there are in general two unrelated solutions to $\boldsymbol{F}_0(\boldsymbol{u}) = \hat{\boldsymbol{m}}$ which $\boldsymbol{F}_1$ does not map to the same point. This is exemplified by

$$\boldsymbol{u} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ -1 \end{pmatrix}, \ \tilde{\boldsymbol{u}} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix} \qquad \Rightarrow \qquad \boldsymbol{F}_0(\boldsymbol{u}) = \boldsymbol{F}_0(\tilde{\boldsymbol{u}}) = \begin{pmatrix} 1 \\ 0 \\ \frac{1}{\sqrt{2}} \\ 1 + \frac{1}{\sqrt{2}} \end{pmatrix} \tag{2.83}$$

but

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = \boldsymbol{F}_1(\boldsymbol{u}) \neq \boldsymbol{F}_1(\tilde{\boldsymbol{u}}) = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}. \tag{2.84}$$

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

Remark 2.7. In (ii) (see Section 1.4), it was proved by direct calculation that in the case $N = 2$ the Jacobian of $\boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}$, given below in (2.92), is uniformly bounded in any closed subset of (2.77). Together with Corollary 2.5 and the mean value theorem this implies that $\boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1}$ is Lipschitz continuous in any closed and convex subset of (2.77) when $N = 2$.

2.3.3 Structure of the Conservation Laws

For simplicity we will mainly deal with the single phase, $N = 1$, one-dimensional case where the medium is vacuum, $\eta = \text{const} = 1$. With the $\boldsymbol{u}$ variables defined in (2.40) as conserved quantities, the system can be written on the standard form of a conservation law,

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \qquad \boldsymbol{f}(\boldsymbol{u}) = \begin{pmatrix} \frac{u_1^2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}. \tag{2.85}$$

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

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} = G \begin{pmatrix} \cos\theta & -\sin\theta \\ 0 & \cos\theta \end{pmatrix} G^{-1}, \qquad G = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{2.86}$$

Thus, the linearized problem has a double real eigenvalue, $\cos\theta$, and an incomplete set of eigenvectors for all $\boldsymbol{u}$; the system (2.85) is only *weakly hyperbolic*. In general this means that (2.85) is not well-posed in the strongly hyperbolic sense. The system is likely to be much more sensitive than regular hyperbolic systems. The solution of the linearized problem with frozen coefficients loses one derivative. The L_2 norm of the solution at time $t > t_0$ can be estimated in terms of the H_1 norm of the initial data at time $t = t_0$. The sensitivity of the equations is reflected in difficulties in finding stable numerical methods to solve them (see Section 2.6). For a general study of this type of degenerate systems of conservation laws, see [62].

The existence of solutions to (2.85) is an open question. It clearly appears that solutions cannot be expected to be of bounded variation. In fact, numerical results in Section 2.7, for instance Figure 2.8, suggest that (2.85) can have measure solutions, i.e. of delta function type. Also, as shown in Section 2.5.3, there is a close relationship between (2.85) and the equations of pressureless gases, which are known to have measure solutions. The same indication is given by the solution to the Riemann problem below, Section 2.3.3. An extended solution concept is needed to accommodate such solutions. In the context of pressureless gases this problem was addressed by Bouchut [9] and later Brenier and Grenier, [13, 28, 14] and E, Rykov and Sinai, [21] who independently proved global existence of measure solutions. For linear transport equations related results have been obtained by Bouchut and James [10] and Poupaud and Rascle [51]. Entropy conditions and uniqueness of solutions to (2.85) are even more uncertain. For the pressureless gas equations, see [11].

The appearance of a delta function is closely linked to when the physically correct solution passes outside the class of solutions that the system (2.41) describes. If initial data dictates a physical solution with N phases for $t > T$, the system (2.41) with $M < N$ phases will have a measure solution for $t > T$. In the case of (2.85), a delta function will appear in the solution when multiple phases are present.

The statements above are supported by our numerical simulations. We will consider a one-dimensional example. In vacuum, $\eta = 1$, the system can be written in separated form when solutions are smooth,

$$\begin{aligned}(\theta_k)_t + (\sin\theta_k)_x &= 0, \\ (g_k)_t + \cos\theta_k (g_k)_x &= -g_k(\cos\theta_k)_x,\end{aligned} \qquad k = 1, \dots, N. \tag{2.87}$$

(See (2.154) in Section 2.5.1 below.) The equation for θ_k is known to develop shocks in finite time. The angle θ_k will be constant along characteristics, which are straight lines corresponding to rays. The shock develops where characteristics cross, i.e. where two wave fields meet. The equation for g_k is an ordinary transport equation with a source term involving the derivative of $\cos\theta_k$. Along characteristics, which are the same for both equations, the source term is zero, except at a shock where it becomes a delta function. The resulting solution for g_k is a delta function where the phase "should" have split into two new phases.

It is interesting again to compare the moment equations with the eikonal and transport equations, (2.5) and (2.6). The latter also form a weakly hyperbolic system with the same eigenvalue as (2.85). As was mentioned before, the viscosity solution picks out the phase corresponding to first arrival where the physically correct phase is multi-valued. When wave fields meet, there will therefore in general be a jump in ϕ. Because of the term $\Delta\phi$ 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 similar also in this respect.

For the two-dimensional case, another function, $\boldsymbol{g}$, is added to (2.85),

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x + \boldsymbol{g}(\boldsymbol{u})_y = 0, \qquad \boldsymbol{g}(\boldsymbol{u}) = \begin{pmatrix} \frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}} \\ \frac{u_2^2}{\sqrt{u_1^2+u_2^2}} \end{pmatrix}. \tag{2.88}$$

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

$$\begin{aligned}\alpha_1 \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} + \alpha_2 \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}} &= G\left[\alpha_1 \begin{pmatrix} \cos\theta & -\sin\theta \\ 0 & \cos\theta \end{pmatrix} + \alpha_2 \begin{pmatrix} \sin\theta & \cos\theta \\ 0 & \sin\theta \end{pmatrix}\right] G^{-1}, \\ &\equiv \alpha_1 \tilde{\Lambda}_1(\theta) + \alpha_2 \tilde{\Lambda}_2(\theta)\end{aligned} \tag{2.89}$$

with the same rotation G as in (2.86). Regardless of the choice of (α_1, α_2), we still only have one eigenvalue and an incomplete set of eigenvectors. In the general case with N phases, homogeneous medium, the governing equations (2.41) reads

$$\boldsymbol{F}_0(\boldsymbol{u})_t + \boldsymbol{F}_1(\boldsymbol{u})_x + \boldsymbol{F}_2(\boldsymbol{u})_y = 0. \tag{2.90}$$

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

$$\begin{aligned} \alpha_1 J_1 + \alpha_2 J_2 &= \alpha_1 J_0 \Lambda_1 + \alpha_2 J_0 \Lambda_2, \\ \Lambda_j &= \operatorname{diag}(\tilde{\Lambda}_j(\theta_1), \dots, \tilde{\Lambda}_j(\theta_N)), \qquad j = 1, 2, \end{aligned} \tag{2.91}$$

with $\tilde{\Lambda}_j$ as in (2.89). Thus the Jacobians of the flux functions (2.55) are given by

$$\frac{\partial}{\partial \hat{\boldsymbol{m}}} \boldsymbol{F}_1 \circ \boldsymbol{F}_0^{-1} = J_0 \Lambda_1 J_0^{-1}, \qquad \frac{\partial}{\partial \hat{\boldsymbol{m}}} \boldsymbol{F}_2 \circ \boldsymbol{F}_0^{-1} = J_0 \Lambda_2 J_0^{-1}. \tag{2.92}$$

This shows that the eigenvalues of the general system are simply the union of the eigenvalues of N systems of the type (2.88). It also shows that there will not be more than N eigenvectors, for the $2N \times 2N$ system. Hence, we have shown that the general system (2.41) is weakly hyperbolic.

The Riemann Problem

We study the one-dimensional Riemann problem for (2.85),

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \qquad \boldsymbol{u}(x, 0) = \begin{cases} \boldsymbol{u}_l & x < 0, \\ \boldsymbol{u}_r & x > 0. \end{cases} \tag{2.93}$$

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

$$\boldsymbol{f}(\boldsymbol{u}_l) - \boldsymbol{f}(\boldsymbol{u}_r) = s(\boldsymbol{u}_l - \boldsymbol{u}_r), \tag{2.94}$$

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

$$\cos\theta_l \boldsymbol{u}_l - \cos\theta_r \boldsymbol{u}_r = s(\boldsymbol{u}_l - \boldsymbol{u}_r). \tag{2.95}$$

For a given state, $\boldsymbol{u}_l$, we can construct the Hugoniot locus consisting of all states to which we can connect with a jump. In our case the Hugoniot locus of state $\boldsymbol{u}_l$ is simply $\alpha \boldsymbol{u}_l$, $\alpha \in \mathbb{R}$, hence, in phase space, a straight line shooting out from the origin at angle θ_l. The speed of propagation is $s = \cos\theta_l$. It follows that two non-zero states $\boldsymbol{u}_l$ and $\boldsymbol{u}_r$ can only be connected with a discontinuity, should they be parallel, $\boldsymbol{u}_l \parallel \boldsymbol{u}_r$. Otherwise there must be an intermediate state, $\boldsymbol{u}_m$, in between. What this state will be is dictated by the Lax entropy condition, saying that the left discontinuity must move slower than the right one. There will be two types of discontinuities. If $\cos\theta_l < \cos\theta_r$, we can use the origin of the phase space as the intermediate state, hence $\boldsymbol{u}_m = 0$. The Hugoniot loci and the solution for this type of discontinuity is illustrated in Figure 2.2. If $\cos\theta_l > \cos\theta_r$, on the other hand, we do not have a solution in a classical sense. This situation corresponds to two meeting wave fields. Formally, however, a weak solution to the conservation law with this initial data is given by setting $\boldsymbol{u}_m = t\tilde{\boldsymbol{u}}_m \delta(x - st)$. The conservation form gives a slightly modified jump condition for this case,

$$\cos\theta_l \boldsymbol{u}_l - \cos\theta_r \boldsymbol{u}_r = \cos\tilde{\theta}_m (\boldsymbol{u}_l - \boldsymbol{u}_r) + \tilde{\boldsymbol{u}}_m, \tag{2.96}$$

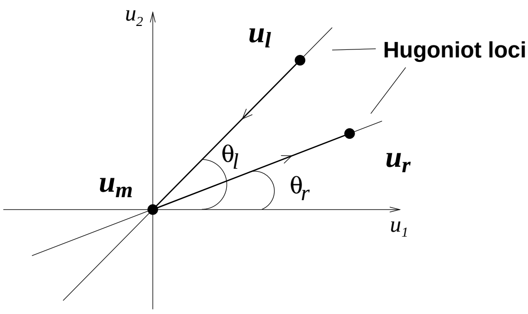
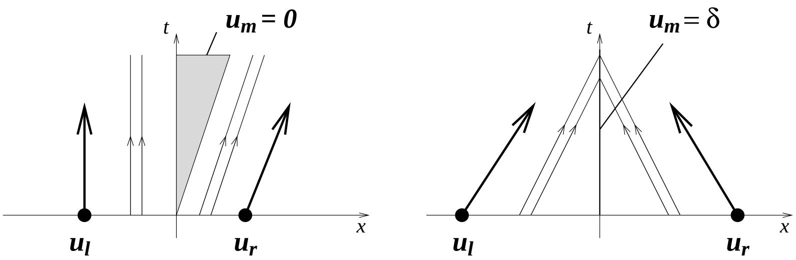


Figure 2.2. Hugoniot loci for left and right state, and solution to the Riemann problem for the first type of discontinuity, plotted in phase space.



(a) Contact discontinuity

(b) Overcompressive shock

Figure 2.3. The two different types of discontinuities.

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 [62] for more general equations.

It is easily verified from (2.86). that $\boldsymbol{u}$ itself is an eigenvector of the Jacobian of $\boldsymbol{f}$. Therefore, the Hugoniot locus will coincide with the integral curves of the system's characteristic fields. Since the (double) eigenvalue of (2.86), $\cos\theta$, is 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 (x, t)-space, are shown in Figure 2.3.

Entropy

For the analysis of (2.85) it would be useful to find a convex entropy pair corresponding to the system. This is, however, not possible.

Lemma 2.8. *Let $\boldsymbol{u}$ be a smooth solution to (2.85). Then there exist functions $U(\boldsymbol{u})$ and $F(\boldsymbol{u})$, both in C^2, such that*

$$U_t + F_x = 0 \tag{2.97}$$

if and only if U is the sum of eigenfunctions corresponding to eigenvalues 0 and 1 of the operator $\mathcal{L} \equiv \boldsymbol{u} \cdot \nabla_u$.

Proof. Let $U \in C^2$. Multiplication of (2.85) by ∇U gives

$$U_t + \nabla U \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} \boldsymbol{u}_x = 0. \tag{2.98}$$

We must show that there exists $F \in C^2$ such that $\nabla U \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} = \nabla F$. This is equivalent to the condition $\nabla \times (\nabla U \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}) = 0$, which in turn implies

$$\nabla \times (\nabla U \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}) = \frac{u_1}{|\boldsymbol{u}|^3} \mathcal{L}(\mathcal{L}U - U) = 0. \tag{2.99}$$

Obviously, this is satisfied for the stated kind of U. Conversely, if $u_1 \neq 0$ and (2.99) holds, let $U = U_0 + U_1$ where $U_0 = U - \mathcal{L}U$. Then U_0 is in the nullspace of $\mathcal{L}$ and $U_1 = \mathcal{L}U$ satisfies $\mathcal{L}U_1 = U_1$. By the continuity of $U_0, U_1, \mathcal{L}U_0, \mathcal{L}U_1$, which holds since $U \in C^2$, this is true also for $u_1 = 0$. Hence, U_0, U_1 are eigenfunctions of $\mathcal{L}$ corresponding to eigenvalues $0, 1$ respectively. □

Remark 2.9. If $U = U_0 + U_1$ satisfy the conditions in the theorem above, the flux F is given by

$$F(\boldsymbol{u}) = U_0 + \frac{u_1}{|\boldsymbol{u}|} U_1 + u_1 \int_0^{u_2} \frac{\frac{\partial U_0(u_1, u_2')}{\partial u_2}}{\sqrt{u_1^2 + (u_2')^2}} du_2'. \tag{2.100}$$

Example of U_0 is u_2/u_1 and of U_1 for instance $U_1 = u_1 + u_2$ or $U_1 = \sqrt{u_1 u_2}$. In fact for U_0 we can take any function of the form $\tilde{U_0}(\lambda)$, and for U_1 any function $|\boldsymbol{u}|\tilde{U_1}(\lambda)$, where $\lambda = \lambda(\boldsymbol{u})$ is the sole eigenvalue of the system and $\tilde{U_j} \in C^2$. This is clear since the operator $\mathcal{L}$ can be transformed to $\mathcal{L} = |\boldsymbol{u}|\partial_{|\boldsymbol{u}|}$. It was shown in (2.86) that $\lambda(\boldsymbol{u})$ is independent of $|\boldsymbol{u}|$ so $\mathcal{L}\tilde{U_0} = 0$ and $\mathcal{L}|\boldsymbol{u}|\tilde{U_1} = \tilde{U_1}\mathcal{L}|\boldsymbol{u}| = |\boldsymbol{u}|\tilde{U_1}$. Another way of seeing that $\tilde{U_0}(\lambda)$ can be used as U_0 is by noting that $\mathcal{L}\tilde{U_0}(\lambda) = \tilde{U}_0' \boldsymbol{u} \cdot \nabla_u \lambda$. Since $\boldsymbol{u}$ is an eigenvector of the system corresponding to λ and since the system is linearly degenerate, this expression must be zero.

Theorem 2.10. *There exists no twice continuously differentiable convex entropy pair for the system (2.85).*

Proof. Let $U, F \in C^2$ satisfy (2.97). For U to be convex, the determinant of its Hessian matrix should be positive. We have

$$\det \begin{pmatrix} U_{u_1,u_1} & U_{u_1,u_2} \\ U_{u_1,u_2} & U_{u_2,u_2} \end{pmatrix} = U_{u_1,u_1} U_{u_2,u_2} - U_{u_1,u_2}^2. \tag{2.101}$$

By Lemma 2.8,

$$\begin{aligned} 0 &= \mathcal{L}^2 U - \mathcal{L} U = u_1^2 U_{u_1,u_1} + 2u_1u_2 U_{u_1,u_2} + u_2^2 U_{u_2,u_2} \qquad \Rightarrow \\ U_{u_1,u_2} &= -\frac{1}{2}\left(\frac{u_1}{u_2} U_{u_1,u_1} + \frac{u_2}{u_1} U_{u_2 u_2}\right). \end{aligned} \tag{2.102}$$

Therefore,

$$U_{u_1,u_1} U_{u_2,u_2} - U_{u_1,u_2}^2 = -\frac{1}{4}\left(\frac{u_1}{u_2} U_{u_1,u_1} - \frac{u_2}{u_1} U_{u_2,u_2}\right)^2 \leq 0. \tag{2.103}$$

□

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. A sufficient condition is that $\cos\theta \in C^1$ for the single phase solutions, which we show here, beginning with a lemma.

Lemma 2.11. *Suppose $\boldsymbol{u}(t, \boldsymbol{x})$ is a weak solution of (2.88) such that $\cos\theta(t, \boldsymbol{x}) \in C^1$, where $\boldsymbol{u} = g(\cos\theta, \sin\theta)^T$. Then $m_{ij} = g\cos^i\theta\sin^j\theta$ is a weak solution to*

$$(m_{ij})_t + (m_{i+1,j})_x + (m_{i,j+1})_y = 0, \qquad \forall i, j \geq 0. \tag{2.104}$$

Proof. First note that since the equation for m_{00} is obtained by adding the equations for m_{20} and m_{02}, we only have to prove the statement for $i+j > 1$. Let $\phi \in C_c^1(\mathbb{R}^+ \times \mathbb{R}_x^2)$. Suppose $i > 0$ and set $T = \cos^{i-1}\theta\sin^j\theta$. So $m_{ij} = gT\cos\theta$ is bounded, $|m_{ij}| \leq |T||\boldsymbol{u}|$, and

$$\begin{aligned} \phi T_t \cos\theta &= \phi\frac{dT}{d\theta}(\cos\theta(\sin\theta)_t - \sin\theta(\cos\theta)_t)\cos\theta \\ &= \cos\theta\left(\phi\frac{dT}{d\theta}\sin\theta\cos\theta\right)_t - \sin\theta\left(\phi\frac{dT}{d\theta}\cos^2\theta\right)_t \end{aligned} \tag{2.105}$$

and similar for the partial derivative with respect to x and y. Therefore, setting $\psi_s = \phi \frac{dT}{d\theta} \sin\theta\cos\theta$ and $\psi_c = \phi \frac{dT}{d\theta} \cos^2\theta$,

$$\begin{aligned}
&\iint_{t\geq 0} m_{ij}\phi_t + m_{i+1,j}\phi_x + m_{i,j+1}\phi_y d\boldsymbol{x}dt \qquad (2.106)\\
&= \iint_{t\geq 0} g\cos\theta T\phi_t + g\cos^2\theta T\phi_x + g\sin\theta\cos\theta T\phi_y d\boldsymbol{x}dt\\
&= \iint_{t\geq 0} g\cos\theta (T\phi)_t + g\cos^2\theta (T\phi)_x + g\sin\theta\cos\theta (T\phi)_y d\boldsymbol{x}dt\\
&\quad - \iint_{t\geq 0} g\cos\theta (\psi_s)_t + g\cos^2\theta (\psi_s)_x + g\sin\theta\cos\theta (\psi_s)_y d\boldsymbol{x}dt\\
&\quad + \iint_{t\geq 0} g\sin\theta (\psi_c)_t + g\sin\theta\cos\theta (\psi_c)_x + g\sin^2\theta (\psi_c)_y d\boldsymbol{x}dt.
\end{aligned}$$

By assumption $\psi_c, \psi_s, T\phi \in C_c^1(\mathbb{R}^+ \times \mathbb{R}_x^2)$ and $\boldsymbol{u}$ is a weak solution to (2.88). Consequently,

$$\begin{aligned}
&\iint_{t\geq 0} m_{ij}\phi_t + m_{i+1,j}\phi_x + m_{i,j+1}\phi_y d\boldsymbol{x}dt\\
&= -\int_{t=0} g(0,\boldsymbol{x})\cos\theta(0,\boldsymbol{x})(T(0,\boldsymbol{x})\phi(0,\boldsymbol{x}) - \psi_s(0,\boldsymbol{x}))\\
&\quad + g(0,\boldsymbol{x})\sin\theta(0,\boldsymbol{x})\psi_c(0,\boldsymbol{x})dt\\
&= -\int_{t=0} m_{ij}(0,\boldsymbol{x})\phi(0,\boldsymbol{x})dt. \qquad (2.107)
\end{aligned}$$

So m_{ij} is indeed a weak solution to (2.104) if $i > 0$. If $i = 0$, use instead $T = \cos^i\theta\sin^{j-1}\theta$. The proof is identical. □

Using this lemma we can show the following theorem.

Theorem 2.12. *Suppose $\{\boldsymbol{u}_k\}_k^N$ are N weak solutions to the homogeneous single phase system (2.88) such that $\cos\theta_k \in C^1$ where $\boldsymbol{u}_k = g_k(\cos\theta_k, \sin\theta_k)^T$. Then $\boldsymbol{u} = \sum_k^N \boldsymbol{u}_k$ is a weak solution to the homogeneous N-phase system (2.90).*

Proof. Let $\phi \in C_c^1(\mathbb{R}^+ \times \mathbb{R}_x^2)$. We have by Lemma 2.11

$$\begin{aligned}
&\textstyle\sum_{k=1}^N \iint_{t\geq 0} g_k\cos^{2\ell-1}\theta_k\phi_t + g_k\cos^{2\ell}\theta_k\phi_x + g_k\cos^{2\ell-1}\theta_k\sin\theta_k\phi_y d\boldsymbol{x}dt\\
&\qquad = -\textstyle\sum_{k=1}^N \int_{t=0} g_k(0,\boldsymbol{x})\cos^{2\ell-1}\theta_k(0,\boldsymbol{x})\phi(0,\boldsymbol{x}) \qquad (2.108)
\end{aligned}$$

for all ℓ. Similarly, for all ℓ,

$$\begin{aligned}
&\textstyle\sum_{k=1}^N \iint_{t\geq 0} g_k\sin^{2\ell-1}\theta_k\phi_t + g_k\sin^{2\ell-1}\theta_k\cos\theta_k\phi_x + g_k\sin^{2\ell}\theta_k\phi_y d\boldsymbol{x}dt\\
&\qquad = -\textstyle\sum_{k=1}^N \int_{t=0} g_k(0,\boldsymbol{x})\sin^{2\ell-1}\theta_k(0,\boldsymbol{x})\phi(0,\boldsymbol{x}). \qquad (2.109)
\end{aligned}$$

But (2.108) and (2.109) are just the componentwise statements of

$$\iint_{t\geq 0} \boldsymbol{F}_0(\boldsymbol{u})\phi_t + \boldsymbol{F}_1(\boldsymbol{u})\phi_x + \boldsymbol{F}_2(\boldsymbol{u})\phi_y = -\int_{t=0} \boldsymbol{F}_0(\boldsymbol{u}(0,\boldsymbol{x}))\phi(0,\boldsymbol{x}) \tag{2.110}$$

and $|\boldsymbol{u}|$ is bounded by $\sum_k^N |\boldsymbol{u}_k|$. □

Of course some of the $\boldsymbol{u}_k$ solutions in Theorem 2.12 can be identically zero.

2.4 Closure with Heaviside Functions

We will now consider a different way to close (2.29). In this case we discard the information carried by g_k used in Section 2.3, that is the amplitude, and only solve for the θ_k. In this way we get fewer and less singular equations. Since we do not obtain any information about the amplitude in this way, the "correct" values of the unknowns are 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 2.7). In order to reduce the initialization problem we will therefore only consider steady state solutions to (2.29), such that can be "timestepped" in the x-direction. In other words, we put an additional restriction on f, namely that $f(t,\boldsymbol{x},\boldsymbol{p}) = 0$ when $\boldsymbol{p}\cdot\boldsymbol{e}_x \leq 0$ where $\boldsymbol{e}_x$ is the unit vector in the x-direction.

We consider the case when the density function f can be written on the form

$$f(t,\boldsymbol{x},\boldsymbol{p}) = \begin{cases} \delta(|\boldsymbol{p}|-\eta)\sum_{k=1}^N (-1)^{k+1} H(\theta-\theta_k), & \boldsymbol{p}\cdot\boldsymbol{e}_x > 0, \\ 0, & \text{otherwise,} \end{cases} \qquad \boldsymbol{p} \equiv |\boldsymbol{p}| \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{2.111}$$

with the convention that $-\pi/2 \leq \theta_1 \leq \ldots \leq \theta_N \leq \pi/2$. Hence, for fixed $t, \boldsymbol{x}$ the density function f is supported by a set of intervals on the circle $|\boldsymbol{p}| = \eta$. The intervals correspond to fans of rays, whose edges are given by the unknown angles θ_k. The edges will propagate just like ordinary rays, Figure 2.4, as long as f stays on the form (2.111). When N is odd the edge of the last interval is fixed to $\pi/2$. For inhomogeneous problems this will not be a solution to the transport equation (2.20) and the correct f does not stay on the form in (2.111). For the time being we will therefore assume N is even.

2.4.1 Equations

After the same rescaling as in Section 2.3, we get the following general expression for the moments

$$\begin{aligned} m_{ij} &= \frac{1}{\eta^{i+j}} \int_{\mathbb{R}^2} p_1^i p_2^j f dp_1 dp_2 = \sum_{k=1}^N \int_{-\pi/2}^{\pi/2} (-1)^{k+1} H(\theta-\theta_k) \cos^i \theta_k \sin^j \theta_k d\theta_k \\ &= \sum_{k=1}^N \int_{\theta_k}^{\pi/2} (-1)^{k+1} \cos^i\theta \sin^j \theta d\theta. \end{aligned} \tag{2.112}$$

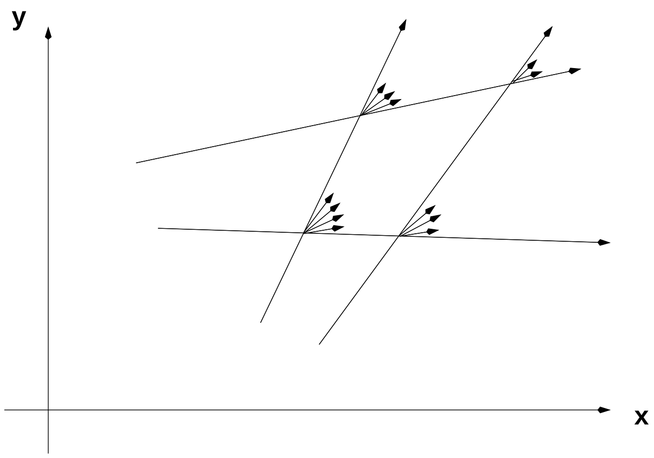


Figure 2.4. Ray fans, when $N = 2$. The edges of the fans propagate as rays.

We choose the equations for the moments $\{m_{0,k}\}$ with $k = 0, \dots, N-1$ which leads to the steady state equations

$$(\eta m_{1,k})_x + (\eta m_{0,k+1})_y = k(\eta_y m_{0,k-1} - \eta_x m_{1,k} - \eta_y m_{0,k+1}). \tag{2.113}$$

Next, we introduce the new variables

$$u_k = \sin\theta_k, \qquad \boldsymbol{u} = (u_1, \dots, u_N) \tag{2.114}$$

and the function

$$\tilde{f} = \sum_{k=1}^{N} (-1)^{k+1} H(u - u_k), \qquad -1 \le u_1 \le \dots \le u_N \le 1. \tag{2.115}$$

From (2.112) it then follows that

$$m_{1,k} = \sum_{j=1}^{N} (-1)^{j+1} \int_{u_j}^{1} u^k dx = \langle \tilde{f}, u^k \rangle = \frac{1}{k+1} \sum_{j=1}^{N} (-1)^j u_j^{k+1}, \tag{2.116}$$

$\langle \cdot, \cdot \rangle$ denoting the standard inner product on $L_2[-1,1]$. Similarly,

$$m_{0,k+1} = \sum_{j=1}^{N} (-1)^{j+1} \int_{u_j}^{1} \frac{u^{k+1}}{\sqrt{1-u^2}} dx = \left\langle \tilde{f}, \frac{u^{k+1}}{\sqrt{1-u^2}} \right\rangle = \sum_{j=1}^{N} (-1)^j R_k(u_j) \tag{2.117}$$

where R_k is the primitive function for $u^{k+1}/\sqrt{1-u^2}$. With $\boldsymbol{u}$ in (2.114) we can, like in Section 2.3, write (2.113) on the form

$$(\eta \boldsymbol{F}_1(\boldsymbol{u}))_x + (\eta \boldsymbol{F}_2(\boldsymbol{u}))_y = \boldsymbol{K}(\boldsymbol{u}, \eta_x, \eta_y), \tag{2.118}$$

where the functions $\boldsymbol{F}_k$ and $\boldsymbol{K}$ again are rather complicated nonlinear functions, given by (2.116) and (2.117). With

$$\hat{\boldsymbol{m}} = (m_{10}, \dots, m_{1,N-1})^T \tag{2.119}$$

we could write (2.118) as

$$(\eta\hat{\boldsymbol{m}})_x + (\eta\boldsymbol{F}_2 \circ \boldsymbol{F}_1^{-1}(\hat{\boldsymbol{m}}))_y = \boldsymbol{K}(\boldsymbol{F}_1^{-1}(\hat{\boldsymbol{m}}), \eta_x, \eta_y). \tag{2.120}$$

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

$$\nabla_{\hat{\boldsymbol{m}}} m_{0,k+1}^T = V^{-1}\Theta. \tag{2.121}$$

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

$$\Theta = \{u_k^{j+1}/\sqrt{1-u_k^2}\}_{j=1}^N = \{\tan\theta_j\}_{j=1}^N \in \mathbb{R}^N. \tag{2.122}$$

By using (2.121) we get an expression for the Jacobian of $\boldsymbol{F}_2 \circ \boldsymbol{F}_1^{-1}$,

$$\begin{aligned}
J^T &= V^{-1}\begin{pmatrix} \frac{u_1}{\sqrt{1-u_1^2}} & \cdots & \frac{u_1^N}{\sqrt{1-u_1^2}} \\ \vdots & \ddots & \vdots \\ \frac{u_N}{\sqrt{1-u_N^2}} & \cdots & \frac{u_N^N}{\sqrt{1-u_N^2}} \end{pmatrix} = V^{-1}\begin{pmatrix} \frac{u_1}{\sqrt{1-u_1^2}} & & \\ & \ddots & \\ & & \frac{u_N}{\sqrt{1-u_N^2}} \end{pmatrix} V \\
&= V^{-1}\begin{pmatrix} \tan\theta_1 & & \\ & \ddots & \\ & & \tan\theta_N \end{pmatrix} V.
\end{aligned} \tag{2.123}$$

We see that this system is strictly hyperbolic as long as $\theta_i \neq \theta_j$ for all i, j. The theory for the system is standard and we will not further discuss its structure. We just 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 by the assumption that the equation can be time-stepped in the x-direction, cf. Remark 2.17.

Up to now we have only considered N even. Under some assumptions the derivations above hold also for N odd. To show this we start with the equations for $N+1$ phases. We assume that there is a smooth solution $\boldsymbol{u}^*$ to the $N = 2$ system such that $u_2^*(x) > u_N(x)$ for all x and all solutions u_N that we are interested in. If the superposition principle is valid, we can substitute the fixed u_2^* for u_{N+1} and the $N+1$ equations are still satisfied. We can therefore reduce the $N+1$ equations to only N equations with N unknowns. For example, in the homogeneous case we can take $u_2^* \equiv \sin\pi/2 = 1$. In practice the reduction amounts to simply using the expression of the moments (2.116, 2.117) also when N is odd.

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

$$\boldsymbol{F}_1(u_1) = -u_1, \qquad \boldsymbol{F}_2(u_1) = \sqrt{1-u_1^2}, \qquad \boldsymbol{K} = 0. \tag{2.124}$$

For $N = 2$,

$$
\begin{aligned}
\boldsymbol{F}_1 &= \begin{pmatrix} u_2 - u_1 \\ \frac{1}{2}(u_2^2 - u_1^2) \end{pmatrix}, \\
\boldsymbol{F}_2 &= \begin{pmatrix} \sqrt{1-u_1^2} - \sqrt{1-u_2^2} \\ \frac{1}{2}\big(\arcsin(u_2) - \arcsin(u_1) + u_1\sqrt{1-u_1^2} - u_2\sqrt{1-u_2^2}\big) \end{pmatrix}, \\
\boldsymbol{K} &= \begin{pmatrix} 0 \\ \frac{1}{2}\big(\eta_y(\arcsin(u_2) - \arcsin(u_1) - u_1\sqrt{1-u_1^2} \\ \quad +u_2\sqrt{1-u_2^2}\big) - \eta_x(u_2^2 - u_1^2) \end{pmatrix},
\end{aligned}
\tag{2.125}
$$

and for $N = 3$,

$$
\begin{aligned}
\boldsymbol{F}_1 &= \begin{pmatrix} -u_1 + u_2 - u_3 \\ \frac{1}{2}(-u_1^2 + u_2^2 - u_3^2) \\ \frac{1}{3}(-u_1^3 + u_2^3 - u_3^3) \end{pmatrix}, \\
\boldsymbol{F}_2 &= \begin{pmatrix} \sqrt{1-u_1^2} - \sqrt{1-u_2^2} + \sqrt{1-u_3^2} \\ \frac{1}{2}(\arcsin(u_2) - \arcsin(u_1) - \arcsin(u_3)) \\ \quad +\frac{1}{2}\big(u_1\sqrt{1-u_1^2} - u_2\sqrt{1-u_2^2} + u_3\sqrt{1-u_3^2}\big) \\ \frac{1}{3}((2+u_1^2)\sqrt{1-u_1^2} - (2+u_2^2)\sqrt{1-u_2^2} + (2+u_3^2)\sqrt{1-u_3^2}) \end{pmatrix}, \\
\boldsymbol{K} &= \begin{pmatrix} 0 \\ \eta_y\big(\arcsin(u_2) - \arcsin(u_1) - \arcsin(u_3) \\ \quad -u_1\sqrt{1-u_1^2} + u_2\sqrt{1-u_2^2} - u_3\sqrt{1-u_3^2}\big) \\ \quad -\frac{1}{2}\eta_x(u_2^2 - u_1^2 - u_3^2) \\ \frac{1}{3}\eta_y\big((1-u_1^2)\sqrt{1-u_1^2} - (1-u_2^2)\sqrt{1-u_2^2} \\ \quad +(1-u_3^2)\sqrt{1-u_3^2}\big) - \frac{2}{3}\eta_x(-u_1^3 + u_2^3 - u_3^3) \end{pmatrix}.
\end{aligned}
\tag{2.126}
$$

2.4.2 Properties of the Flux Function

We will here show that the functions

$$\boldsymbol{F}_2 \circ \boldsymbol{F}_1^{-1}(\boldsymbol{m}) \qquad \text{and} \qquad \boldsymbol{K}(\boldsymbol{F}_1^{-1}(\boldsymbol{m}), \eta_x, \eta_y) \tag{2.127}$$

are well defined and regular on their domains of definition. The results will be given for a slightly more general class of functions than those in (2.127). We start by introducing some notation. For an interval $I \subset \mathbb{R}$, let $U_N(I) \subset \mathbb{R}^N$ be defined by

$$U_N(I) = \big\{u_k \in I \;\big|\; u_1 \le \ldots \le u_N\big\} \tag{2.128}$$

and $F_N(I) \subset L_1(I)$ by

$$F_N(I) = \Big\{f \in L_1(I) \;\Big|\; f(u) = \sum_{k=1}^{N} (-1)^{k+1} H(u - u_k),\ \{u_k\} \in U_N(I)\Big\}. \tag{2.129}$$

Also introduce the set of attainable moments, $M_N \subset \mathbb{R}^N$, with $\langle \cdot\,,\cdot \rangle$ being the L_2 inner product on I,

$$M_N(I) = \big\{m(f) \;\big|\; f \in F_N(I)\big\}, \qquad m(f) = \{\langle f, u^k \rangle\}_{k=1}^N. \tag{2.130}$$

In [12] it was shown that if $I = [0, \infty)$, the mapping $m(f)$ is one-to-one and functions of the type $\langle m^{-1}(\boldsymbol{m}), \psi\rangle$, where $\psi \in L_1(I)$ (e.g. $\boldsymbol{F}_2 \circ \boldsymbol{F}_1^{-1}$ above) are continuous on $M_N[0, L] \subset M_N(I)$ for each $0 < L < \infty$. We will consider any fixed bounded interval I and show a slightly more precise continuity result. In what follows, let L be the length of I and $0 < L < \infty$. For $f, f' \in F_N(I)$, defined by $\{u_k\}, \{u'_k\} \in U_N$,

$$||f-f'||_1 = ||\sum_{k=1}^{N}(-1)^{k+1}(H(u-u_k)-H(u-u'_k))||_1 \leq \sum_{k=1}^{N}|u_k-u'_k| = |\{u_k\}-\{u'_k\}|_1 \tag{2.131}$$

where $||\cdot||_1$ and $|\cdot|_1$ are the L_1 and ℓ_1 norms respectively. The natural map from U_N to F_N is thus continuous. Also

$$|\langle v^k, f - f'\rangle| \leq ||v^k||_1 \cdot ||f - f'||_1 \leq \text{const}||f - f'||_1, \tag{2.132}$$

so $|m(f) - m(f')|_1 \leq \text{const}||f - f'||_1$, showing that $m : F_N \mapsto M_N$ is continuous. We note that since the mappings $U_N \mapsto F_N \mapsto M_N$ are continuous, a closed (and hence compact) interval I implies that U_N, F_N and M_N are all compact. Before proceeding we show a simple lemma.

Lemma 2.13. *Let $f \in F_N(I)$ be defined by the points $\{u_k\}$. Then the polynomial*

$$P(u) = \begin{cases} \prod_{k=2}^{N}(u_k - u), & N > 1 \\ 1, & N = 1, \end{cases} \tag{2.133}$$

satisfies $P(f - f') \geq 0$ for all $u \in I$ and all $f' \in F_N$ such that

$$f'(u) = 0 \qquad \textit{for} \qquad u \leq u_1. \tag{2.134}$$

Proof. First, $f(u) - f'(u) \geq 0$ and $P(u) \geq 0$ for all $u \in \text{supp}(f)$ or $u \leq u_1$. Conversely, $f(u) - f'(u) \leq 0$ and $P(u) < 0$ for all $u \in \text{supp}(f)^c$ and $u \geq u_1$. $\square$

Before the next lemma we introduce some more notation. Let $\mathcal{P}_n(I)$ be the space of polynomials of degree at most n and let $\mathcal{C} : \mathcal{P}_{n-1} \mapsto \mathbb{C}^n$ be the coefficient mapping,

$$\mathcal{C}(c_0 + c_1 x + \ldots + c_{n-1}x^{n-1}) = \begin{pmatrix} c_0 \\ \vdots \\ c_{n-1} \end{pmatrix}. \tag{2.135}$$

For the inverse of the mapping m we now have

Lemma 2.14. *For each N and I, there is a constant $C_{N,I}$ such that*

$$||f - f'||_1 \leq C_{N,I}|m(f) - m(f')|_1^{\alpha}, \qquad \forall f, f' \in F_N(I), \tag{2.136}$$

if and only if $0 \leq \alpha \leq 1/N$, i.e. $m^{-1} : M_N \mapsto F_N$ is Hölder continuous with critical exponent $1/N$.

Proof. As in Lemma 2.13, let $f \in F_N(I)$ be defined by the points $\{u_k\}$ and P be the polynomial in (2.133). Take $f' \in F_N(I)$. Without loss of generality we can assume (2.134). (Otherwise, exchange the prime signs.) We can write

$$f - f' = \sum_{k=1}^{2N} s_k H(u - a_k), \qquad \{u_k\} \subset \{a_k\} \in U_{2N}(I), \tag{2.137}$$

$s_k \in \{-1, 1\}$. Then, with $u_k \le a_j \le a_{j+1} \le u_{k+1}$,

$$\begin{aligned}
\int_{a_j}^{a_{j+1}} (f - f')P du &= \int_{a_j}^{a_{j+1}} \prod_{\ell=2}^{N} |u - u_\ell| du \\
&\ge \int_{a_j}^{a_{j+1}} |u - a_j|^k |u - a_{j+1}|^{N-1-k} du \\
&\ge \int_{a_j}^{\frac{a_j + a_{j+1}}{2}} (a_{j+1} - u)^{N-1} du + \int_{\frac{a_j + a_{j+1}}{2}}^{a_{j+1}} (u - a_k)^{N-1} du \\
&= 2\frac{(a_{j+1} - a_j)^N}{N}(1 - 2^{-N}) \\
&\ge \frac{\alpha}{L^N}(a_{j+1} - a_j)^{1/\alpha},
\end{aligned} \tag{2.138}$$

for $\alpha \le 1/N$. Therefore,

$$\begin{aligned}
||f - f'||_1^{1/\alpha} &= \Big(\sum_{j=1}^{N} \int_{a_{2j-1}}^{a_{2j}} |f - f'| du\Big)^{1/\alpha} = \Big(\sum_{j=1}^{N} |a_{2j-1} - a_{2j}|\Big)^{1/\alpha} \\
&\le \sum_{j=1}^{N} (a_{2j} - a_{2j-1})^{1/\alpha} \le \frac{L^N}{\alpha} \sum_{j=1}^{N} \int_{a_{2j-1}}^{a_{2j}} (f - f')P du \\
&= \frac{L^N}{\alpha} \langle f - f', P \rangle = \frac{L^N}{\alpha} (\mathcal{C}P)^T (m(f) - m(f')) \\
&\le \frac{L^N}{\alpha} |m(f) - m(f')|_1.
\end{aligned} \tag{2.139}$$

We get the stated inequality (2.136) with $C_{N,I} = L^{\alpha N} \alpha^{-\alpha}$. Suppose now that $\alpha > 1/N$ and that a neighborhood of zero is contained in I. Let f_ϵ be defined by the Chebyshev abscissae, $u_j = -2\epsilon \cos((j-1)\pi/N)$ and f'_ϵ by $u'_j = u_{j+1}$, where we set $u_{N+1} = 2\epsilon$, Then by construction $||f_\epsilon - f'_\epsilon||_1 = 4\epsilon$ and

$$\begin{aligned}
\langle u^{k-1}, f_\epsilon - f'_\epsilon \rangle &= \sum_{j=1}^{N} \frac{(-1)^{j-1}(u_{j+1}^k - u_j^k)}{k} \\
&= \frac{(-2\epsilon)^k}{k} \sum_{j=1}^{N} (-1)^{j-1} \Big(\cos^k \frac{j\pi}{N} - \cos^k \frac{(j-1)\pi}{N}\Big)
\end{aligned} \tag{2.140}$$

The sum can be explicitly computed to be zero for $k < N$ and $-N/2^{N-2}$ for $k = N$. Hence, $|m(f_\epsilon) - m(f'_\epsilon)|_1 = 4\epsilon^N$, and

$$\frac{||f_\epsilon - f'_\epsilon||_1}{|m(f_\epsilon) - m(f'_\epsilon)|_1^\alpha} = \frac{4\epsilon}{(4\epsilon^N)^\alpha} = 4^{1-\alpha}\frac{1}{\epsilon^{1-\alpha N}} \to \infty \tag{2.141}$$

when $\epsilon \to 0$. Now, if instead I contains a neighborhood of v, then we can consider $f_\epsilon(u - v) - f'_\epsilon(u - v)$ and get

$$\int_I u^{k-1}(f_\epsilon(u-v) - f'_\epsilon(u-v))du = \int_{-2\epsilon}^{2\epsilon}(u+v)^{k-1}(f_\epsilon(u) - f'_\epsilon(u))du. \tag{2.142}$$

Since the moments are zero for $k < N$ we get the same expression as in (2.140) and again obtain (2.141). □

We can now show the following regularity result.

Theorem 2.15. *Let the mapping* $J_\psi : M_N(I) \mapsto \mathbb{R}$ *be defined by*

$$J_\psi(\boldsymbol{m}) = \langle \psi, f\rangle, \qquad \boldsymbol{m} = m(f). \tag{2.143}$$

If $\psi \in L_1(I)$ *then* $J_\psi \in C^{0,1/N}(M_N(I))$. *If* $\psi \in C^{N-1}(I)$ *then* $J_\psi \in C^{0,1}(M_N(I))$. *If* $\psi \in C^{M-1}(I)$ *and* $N = 1$ *then* $J_\psi \in C^{M,1}(M_N(I))$. *If* $\psi \in \mathcal{P}_{N-1}(I)$ *then* $J_\psi \in C^\infty(M_N(I))$. *If* $\psi \notin \mathcal{P}_{N-1}(I)$, $N > 1$, N *even, then* ∇J_ψ *is discontinuous at* $\boldsymbol{m} = 0$.

Proof. The case when $\psi \in \mathcal{P}_{N-1}$ is trivial, since the function $J_\psi(\boldsymbol{m}) = (\mathcal{C}\psi)^T\boldsymbol{m}$, is just a linear combination of the elements of $\boldsymbol{m}$. Also if $N = 1$, we have $J_\psi(m) = \int_{m-1}^1 \psi du$, showing this case. For $\psi \in L_1(I)$,

$$\begin{aligned} |J_\psi(\boldsymbol{m}) - J_\psi(\boldsymbol{m}')| &= |\langle \psi, m^{-1}(\boldsymbol{m}) - m^{-1}(\boldsymbol{m}')\rangle| \\ &\le ||\psi||_1 \cdot ||m^{-1}(\boldsymbol{m}) - m^{-1}(\boldsymbol{m}')||_1 \\ &\le \text{const}||\boldsymbol{m} - \boldsymbol{m}'||_1^{1/N}, \qquad \forall \boldsymbol{m}, \boldsymbol{m}' \in M_N \end{aligned} \tag{2.144}$$

by Lemma 2.14. Now, assume that $\psi \in C^{N-1}(I)$ and $N > 1$. Let $f \in F_N(I)$ be defined by the points $\{u_k\} \in U_N$ and P be the polynomial in (2.133). Furthermore, let H be the unique polynomial which interpolates ψ at the points $\{u_k\}_{k=2}^N$. Then from standard results in approximation theory H can be written on the form

$$H(u) = \psi(\xi_1) + \sum_{j=2}^{N-1}\frac{\psi^{(j-1)}(\xi_j)}{(j-1)!}\prod_{k=2}^{j}(u - u_k), \qquad \xi_j \in [u_2, u_{j+1}]. \tag{2.145}$$

We note that $|\mathcal{C}P|_1 \le$ const and since $\psi \in C^{N-1}$ we also have $|\mathcal{C}H|_1 \le$ const. The constants depend on N, I and on the max norm of ψ and its derivatives. Moreover, the error term in the interpolation has the form

$$\psi(u) - H(u) = \frac{\psi^{(N-1)}(\xi_N)}{(N-1)!}\prod_{k=2}^{N}(u - u_k) = (-1)^{N-1}\frac{\psi^{(N-1)}(\xi_N)}{(N-1)!}P(u) \tag{2.146}$$

for some $\xi_N \in [u_2, u_N]$. Hence, if $\boldsymbol{m} = m(f)$ and $\boldsymbol{m}' = m(f')$,

$$\begin{aligned}
|J_\psi(\boldsymbol{m}) - J_\psi(\boldsymbol{m}')| &= |\langle \psi, f - f' \rangle| \leq |\langle \psi - H, f - f' \rangle| + |\langle H, f - f' \rangle| \\
&= \left| \left\langle \frac{\psi^{(N-1)}(\xi_N)}{(N-1)!} P, f - f' \right\rangle \right| + |(\mathcal{C}H)^T(\boldsymbol{m} - \boldsymbol{m}')| \\
&\leq \frac{\sup_{\xi \in I} |\psi^{(N-1)}(\xi)|}{(N-1)!} \langle |P|, |f - f'| \rangle + \text{const}|\boldsymbol{m} - \boldsymbol{m}'| \\
&\leq \text{const}\Big(\langle P, f - f' \rangle + |\boldsymbol{m} - \boldsymbol{m}'| \Big) \\
&= \text{const}\Big((\mathcal{C}P)^T(\boldsymbol{m} - \boldsymbol{m}') + |\boldsymbol{m} - \boldsymbol{m}'| \Big) \\
&\leq \text{const}|\boldsymbol{m} - \boldsymbol{m}'|.
\end{aligned} \tag{2.147}$$

(Again, without loss of generality we assumed (2.134).) Finally, as in (2.121) if $u_k < u_{k+1}$ for all k, the gradient of J_ψ is explicitly given by

$$\nabla_{\boldsymbol{m}} J_\psi^T = V^{-1}\Theta. \tag{2.148}$$

with the same V as in (2.121) and $\Theta = \{\psi(u_k)\} \in \mathbb{R}^N$. We see that the elements of the gradient are the coefficients of the unique polynomial interpolating ψ at all the points $\{u_k\}$. This polynomial can be written on the form

$$\bar{H}(u) = \psi(\xi_1) + \sum_{j=1}^{N-1} \frac{\psi^{(j)}(\xi_{j+1})}{j!} \prod_{k=1}^{j} (u - u_k), \qquad \xi_j \in [u_1, u_j]. \tag{2.149}$$

Let $\{f_\epsilon\} \subset F_N$ be a sequence defined by $\{u_k^\epsilon\}$ such that $u_k^\epsilon \to a \in I$ for all k. Then if N is even, by (2.132), $m(f_\epsilon) \to 0$ and the $(N-1)^{\text{th}}$ coefficient of $\bar{H}_\epsilon$, tends to $\psi^{(N-1)}(a)/(N-1)!$ by (2.149). But a was arbitrary, and if $\psi \notin \mathcal{P}_{N-1}$ its $(N-1)^{\text{th}}$ derivative is not constant. □

Corollary 2.16. *The flux and source functions (2.127) are well defined and belong to* $C^{0,1}(M_N(-1,1))$.

Proof. The estimate (2.136) of course implies that for $\boldsymbol{m} \in M_N$ there is only one function $f \in F_N$ such that $m(f) = \boldsymbol{m}$, so the functions are well defined. As seen in (2.117) the functions are given by using $\psi(u) = \frac{u^{k+1}}{\sqrt{1-u^2}} \in C^\infty(-1,1)$. □

Remark 2.17. We cannot take the interval $[-1,1]$ in Corollary 2.16 since the ψ functions goes to infinity at $-1, 1$. In fact also ∇J_ψ is unbounded there. These points correspond to rays going parallel to the y-axis, and we could not expect a good behaviour for this situation as pointed out in Section 2.4.1

2.5 Relationship to Other Equations

2.5.1 Eikonal and Transport Equation

To see how the moment equations with delta function closure (2.41) are equivalent formulations of the traditional equations of geometrical optics (2.5) and (2.6), we present the following derivation for smooth solutions.

We use the complex versions of the variables introduced in Section 2.3.2. Let $\boldsymbol{w} = \mathcal{A}\boldsymbol{u}$ and define z_k, g_k by $g_k = |w_k|$ and $w_k = g_k z_k = g_k \exp(i\theta_k)$. This is consistent with our usual notation from (2.40). Moreover, denote the matrix in the definition of $Q_0(\boldsymbol{w})$ in (2.58) by Z and the vector $\boldsymbol{g} = \{g_k\} \in \mathbb{R}^N$ so that $Q_0(\boldsymbol{w}) = Z\boldsymbol{g}$. Finally we need the diagonal matrices $C = \text{diag}(\{\cos\theta_k\})$, $S = \text{diag}(\{\sin\theta_k\})$ and $\Theta = \text{diag}(\{\theta_k\})$.

Applying $W\mathcal{A}$ to (2.41) from the left, also using (2.60) and (2.47), gives

$$(\eta^2 Z\boldsymbol{g})_t + (\eta ZC\boldsymbol{g})_x + (\eta ZS\boldsymbol{g})_y = W\mathcal{A}\boldsymbol{K}. \tag{2.150}$$

For the source $\boldsymbol{K}$ we use the relationship $m_{k+2,\ell} + m_{k,\ell+2} = m_{k\ell}$ (which easily follows from (2.36)) and get for component k of $\mathcal{A}\boldsymbol{K}$,

$$\begin{aligned}(\mathcal{A}\boldsymbol{K})_k &= (2k-1)(\eta_x(m_{2k-2,2} - im_{1,2k-1}) - \eta_y(m_{2k-1,1} - im_{2,2k-2})) \\ &= -\sum_{j=1}^{N} g_j(\eta_x \sin\theta_j - \eta_y\cos\theta_j)\frac{d}{d\theta}(\cos^{2k-1}\theta + i\sin^{2k-1}\theta)\Big|_{\theta=\theta_j}\end{aligned} \tag{2.151}$$

Since each element of the Z matrix only depend on one variable (one of the θ_ks), we can let the prime sign denote elementwise differentiation with respect to the θ_k. We get

$$\mathcal{A}\boldsymbol{K} = -(W^{-1}Z)'(\eta_x S - \eta_y C)\boldsymbol{g}. \tag{2.152}$$

The equations (2.150) can then be written as

$$\begin{aligned}&Z(\eta^2\boldsymbol{g}_t + (\eta C\boldsymbol{g})_x + (\eta S\boldsymbol{g})_y) \\ &\qquad + Z'(\eta^2\Theta_t + (\eta S)_x - (\eta C)_y)\boldsymbol{g} = 0.\end{aligned} \tag{2.153}$$

A solution to (2.41) is hence given by solving the N separated systems

$$\begin{aligned}\eta^2(\theta_k)_t + (\eta\sin\theta_k)_x - (\eta\cos\theta_k)_y &= 0, \\ \eta^2(g_k)_t + (\eta g_k\cos\theta_k)_x + (\eta g_k\sin\theta_k)_y &= 0.\end{aligned} \qquad k = 1,\dots,N. \tag{2.154}$$

On the other hand, after some algebraic manipulations of (2.5) and (2.6) we get

$$\eta^2\frac{\partial}{\partial t}\arctan\left(\frac{\phi_y}{\phi_x}\right) + \frac{\partial}{\partial x}\left(\eta\frac{\phi_y}{|\nabla\phi|}\right) - \frac{\partial}{\partial y}\left(\eta\frac{\phi_x}{|\nabla\phi|}\right) = 0, \tag{2.155}$$

$$\eta^2\frac{\partial}{\partial t}A_0^2 + \frac{\partial}{\partial x}\left(\eta A_0^2\frac{\phi_x}{|\nabla\phi|}\right) + \frac{\partial}{\partial y}\left(\eta A_0^2\frac{\phi_y}{|\nabla\phi|}\right) = 0 \tag{2.156}$$

for $k = 1, \dots, N$. If we identify the variables of (2.154) as

$$g_k = A_{0,k}^2, \qquad \cos\theta_k = \frac{(\phi_k)_x}{|\nabla\phi_k|}, \qquad \sin\theta_k = \frac{(\phi_k)_y}{|\nabla\phi_k|}, \qquad k = 1, \dots, N, \tag{2.157}$$

they will solve (2.154) and hence (2.41). As expected the vector $(u_{2k-1}, u_{2k})^T$ points in the direction of the gradient of ϕ_k. The length of the vector corresponds to the first amplitude coefficient squared.

For the $N = 1$ Heaviside equation (2.124) we note that it is exactly the the steady state version of (2.155). Also when the index of refraction only depends on x and it has a discontinuity the Rankine-Hugoniot condition gives

$$\eta_l \sin\theta_l - \eta_r \sin\theta_r = s(\eta_r \cos\theta_r - \eta_l \cos\theta_l). \tag{2.158}$$

For a shock location orthogonal to the x-axis, $s = 0$ and we recover Snell's law, (2.15).

2.5.2 Entropy K-multivalued Solutions

By making the change of variables,

$$u = \eta \frac{\phi_y}{|\nabla\phi|} \tag{2.159}$$

in (2.155) we get the scalar conservation law

$$u_x - \left(\sqrt{\eta^2 - u^2}\right)_y = 0. \tag{2.160}$$

for the steady state solution. Note that by definition $|u| \le \eta$. We want to compare the entropy K-multivalued solutions (see Section 2.2.1) for this equation with the solutions of (2.118). For this purpose we consider the kinetic version of (2.160),

$$f_x + a_v f_y - a_y f_v = 0, \qquad f = f(x, y, v), \qquad a = a(x, y, v) = -\sqrt{\eta^2 - v^2}. \tag{2.161}$$

With the definition (2.24) we get the moment equations,

$$(m_k)_x + \partial_y \int_{-\infty}^{\infty} \frac{v^{k+1}}{\sqrt{\eta^2 - v^2}} f dv = k\eta\eta_y \int_{-\infty}^{\infty} \frac{v^{k-1}}{\sqrt{\eta^2 - v^2}} f dv. \tag{2.162}$$

For the moment equations in Section 2.2.2, we assume that the density function $\bar{f} = \bar{f}(\boldsymbol{x}, \boldsymbol{p})$ is on the form $\bar{f} = \delta(|\boldsymbol{p}| - \eta)\tilde{f}(\boldsymbol{x}, \theta)$ with θ defined as in (2.111). Also assume that $\tilde{f} = 0$ when $|\theta| > \pi/2$. This includes (2.111) as a special case. We have the following expression for the unscaled moments

$$m_{ij} = \int_{\mathbb{R}^2} p_1^i p_2^j \bar{f} d\boldsymbol{p} = \int_{-\pi/2}^{\pi/2} \eta^{i+j} \cos^i\theta \sin^j\theta \tilde{f} d\theta. \tag{2.163}$$

Now, we go back to (2.162) and assume that $f = 0$ whenever $|v| > \eta(\boldsymbol{x})$. After making the change of variables $v = \eta \sin\theta$ in (2.162) we obtain

$$\begin{aligned}\partial_x \int_{-\pi/2}^{\pi/2} \eta^{k+1} \sin^k \theta \cos\theta f d\theta + \partial_y \int_{-\pi/2}^{\pi/2} \eta^{k+1} \sin^{k+1} \theta f d\theta \\ = k\eta\eta_y \int_{-\pi/2}^{\pi/2} \eta^{k-1} \sin^{k-1} \theta f d\theta. \end{aligned} \tag{2.164}$$

Identifying $\tilde{f}(\boldsymbol{x}, \theta) = f(\boldsymbol{x}, \eta \sin\theta)$ and using (2.163), we get

$$(m_{1,k})_x + (m_{0,k+1})_y = k\eta\eta_y m_{0,k-1}. \tag{2.165}$$

After scaling the moments, this is nothing but (2.113). Hence, the two formulations give the same result.

2.5.3 Pressureless Gas

The model of pressureless gases with sticky particles in one dimension can be written as

$$\begin{aligned}\rho_t + (\rho u)_x &= 0, \\ (\rho u)_t + (\rho u^2)_x &= 0\end{aligned} \tag{2.166}$$

where ρ is the density and u the velocity. This system, representing conservation of mass and momentum, has many similarities to the single phase delta function system (2.85) in a homogeneous medium. It is weakly hyperbolic with one real eigenvalue and an incomplete set of eigenvectors $\forall \rho, u$. The system has been studied in [9], [13], [28], [14] and [21]. The two-dimensional form of (2.85) with $\eta \equiv 1$ has the form

$$\begin{aligned}(u_1)_t + \left(\frac{u_1^2}{\sqrt{u_1^2+u_2^2}}\right)_x + \left(\frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}}\right)_y &= 0, \\ (u_2)_t + \left(\frac{u_1 u_2}{\sqrt{u_1^2+u_2^2}}\right)_x + \left(\frac{u_2^2}{\sqrt{u_1^2+u_2^2}}\right)_y &= 0.\end{aligned} \tag{2.167}$$

For a stationary solution of this equation, $(u_i)_t = 0$, we can identify

$$\rho = \frac{u_1^2}{\sqrt{u_1^2+u_2^2}} = g\cos^2\theta, \qquad u = \frac{u_2}{u_1} = \tan\theta. \tag{2.168}$$

Hence, (2.166) is equivalent to the stationary solution of the $N = 1$ delta equations (2.41) with the restriction $|u| \leq 1$.

The two dimensional version of (2.166), given in (2.54) can be obtained directly from (2.167) supplemented with the moment equation for m_{00},

$$\left(\sqrt{u_1^2+u_2^2}\right)_t + (u_1)_x + (u_2)_y = 0. \tag{2.169}$$

If we identify

$$\rho = \sqrt{u_1^2 + u_2^2} = g, \qquad u = u_1/g = \cos\theta, \qquad v = u_2/g \sin\theta \tag{2.170}$$

(2.167) and (2.169) are the same as (2.54). Note that in our identification there is an implied relationship between u and v such that $u^2 + v^2 = 1$.

On a discrete level (2.166) describes the so called sticky particle dynamics. This signifies particles that move with constant speed until they collide and at a collision they stick together and form a new particle with mass and velocity given by the conservation of mass and momentum. Mathematically a particle with mass m and velocity v is represented by $\rho(t,x) = m\delta(x - vt)$ and $u(t, vt) = v$ which is a solution to (2.166) in the sense of measures. For a collision between two particles with masses m_1, m_2 and velocities v_1, v_2, the corresponding measure solution is (see [28])

$$\rho = m_1\delta(x - v_1 t) + m_2\delta(x - v_2 t), \qquad u(t, v_1 t) = v_1, \qquad u(t, v_2 t) = v_2 \tag{2.171}$$

for $t \le 0$ (before the collision) and

$$\rho = (m_1 + m_2)\delta(x - v't), \qquad u(t, v't) = v' = \frac{m_1 v_1 + m_2 v_2}{m_1 + m_2} \tag{2.172}$$

for $t > 0$ (after the collision). Solutions of the continuous equations (2.166) are given as the limit when the number of particles goes to infinity with total mass bounded. For the geometrical optics interpretation the particle trajectories are light rays. The density ρ represents the strength of the ray, $\sim g$, and the velocity u represents the direction of the ray, $\tan\theta$. In this simple case we hence have a description of what happens when two rays meet and the system does not support enough phases to accommodate a crossing ray solution. In particular (2.171, 2.172) show that the $N = 1$ delta function system in general does *not give the first arriving wave* as the stationary solution away from discontinuities. Suppose we have two meeting wave fields of strengths g_1, g_2, such that at $x = 0$ and $y > 0$ the rays are going in direction $-\pi/4$ and for $x = 0$, $y < 0$ they have direction $\pi/4$. Whenever two rays meet they will continue as one, according to (2.171, 2.172) with $m_1 = g_1/2$, $m_2 = g_2/2$, $v_1 = -1$ and $v_2 = 1$. If $g_1 = g_2$ we get $v' = 0$ and all rays collide at and continue along the line $y = 0$. Away from $y = 0$ the solution gives the first arriving wave. However, if $g_1 \ne g_2$ rays will collide along a different line, determined by v', and the solution will not give the first arriving wave at all points away from this line. Hence, the conservation of mass and momentum means in this case that we do not get the viscosity solution for the corresponding eikonal equation.

The same conclusion can be drawn from the formal solution to the Riemann problem (2.93) with data generating the overcompressive shock solution. If $\cos\theta_l = -\cos\theta_r$ in (2.96) the shock speed $s = \cos\tilde{\theta}_m$ is not necessarily zero, which would have corresponded to a solution giving the first arriving wave.

2.6 Numerical Approximations

This section includes some results on the numerical treatment of (2.41) and (2.118). The equations closed with delta functions, (2.41), are set in two-dimensional space while the Heaviside equations (2.118) are reduced to a one-dimensional space by treating x as a time-like variable. For the numerical methods we will use the following notation. Space and time is discretized uniformly with step sizes Δx, Δy and Δt. The grid functions U_j^n (1D) and U_{ij}^n (2D) approximates the exact solutions,

$$U_{ij}^n \approx \boldsymbol{u}(i\Delta x, j\Delta y, n\Delta t), \qquad U_j^n \approx \boldsymbol{u}(n\Delta x, j\Delta y), \tag{2.173}$$

where $\boldsymbol{u}$ are the variables introduced in (2.40) and (2.114) respectively. Similarly, for the index of refraction,

$$\eta_{ij} = \eta(i\Delta x, j\Delta y), \qquad \partial_x \eta_{ij} = \eta_x(i\Delta x, j\Delta y), \qquad \partial_y \eta_{ij} = \eta_y(i\Delta x, j\Delta y). \tag{2.174}$$

2.6.1 Single Phase

For the case of a single phase we only consider the delta equations (2.41). As was discussed in Section 2.3.3, this system is sensitive, which creates problems for the numerical methods. The sensitivity derives from the fact that the system is only weakly hyperbolic.

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

$$\begin{aligned} \eta_{ij}^2 U_{ij}^{n+1} \quad = \quad & \frac{1}{4}\left(\eta_{i-1,j}^2 U_{i-1,j}^n + \eta_{i+1,j}^2 U_{i+1,j}^n + \eta_{i,j-1}^2 U_{i,j-1}^n + \eta_{i,j+1}^2 U_{i,j+1}^n\right) \\ & -\frac{\Delta t}{2\Delta x}\left(\boldsymbol{F}_1(\eta_{i+1,j}U_{i+1,j}^n) - \boldsymbol{F}_1(\eta_{i-1,j}U_{i-1,j}^n)\right) \\ & -\frac{\Delta t}{2\Delta y}\left(\boldsymbol{F}_2(\eta_{i,j+1}U_{i,j+1}^n) - \boldsymbol{F}_2(\eta_{i,j-1}U_{i,j-1}^n)\right) \\ & +\Delta t \boldsymbol{K}(U_{ij}^n, \partial_x \eta_{ij}, \partial_y \eta_{ij}), \end{aligned} \tag{2.175}$$

with $\boldsymbol{F}_1$, $\boldsymbol{F}_2$ and $\boldsymbol{K}$ as defined in (2.48). Even if the Lax-Friedrichs method is only of first order, it works quite well and remains stable despite the sensitivity of the equations. Most of our results are produced using this method. The reason for the Lax-Friedrichs scheme's stability is that it introduces a substantial amount of viscosity, which implies that discontinuities in the solution are smeared out.

Less smearing of shocks is obtained with the *Godunov* method, another first order method which adds a smaller measure of viscosity than Lax-Friedrichs. It is based on the exact solution of local Riemann problems. The two-dimensional

Godunov method is constructed by applying an ordinary splitting approach. For the homogeneous case it has the form

$$\begin{aligned} U_{ij}^{n+1/2} &= U_{ij}^n - \frac{\Delta t}{\Delta x}\left(\boldsymbol{f}(U_{ij}^n, U_{i+1,j}^n) - \boldsymbol{f}(U_{i-1,j}^n, U_{ij}^n)\right) \\ U_{ij}^{n+1} &= U_{ij}^{n+1/2} - \frac{\Delta t}{\Delta y}\left(\boldsymbol{g}(U_{ij}^{n+1/2}, U_{i,j+1}^{n+1/2}) - \boldsymbol{g}(U_{i,j-1}^{n+1/2}, U_{ij}^{n+1/2})\right). \end{aligned} \tag{2.176}$$

The flux functions $\boldsymbol{f}, \boldsymbol{g}$ are given as follows. Let θ be the angle corresponding to $\boldsymbol{u}$ as in Section 2.3.3 and c, s be short for cos, sin so that e.g. $c_r = \cos\theta_r$. Then

$$\boldsymbol{f}(\boldsymbol{u}_l, \boldsymbol{u}_r) = \begin{cases} \boldsymbol{F}_1(\boldsymbol{u}_r) & c_l \leq c_r < 0, \\ 0 & c_l \leq 0 \leq c_r, \\ \boldsymbol{F}_1(\boldsymbol{u}_l) & 0 < c_l \leq c_r, \\ \text{undefined} & c_l > c_r, \end{cases} \qquad \boldsymbol{g}(\boldsymbol{u}_l, \boldsymbol{u}_r) = \begin{cases} \boldsymbol{F}_2(\boldsymbol{u}_r) & s_l \leq s_r < 0, \\ 0 & s_l \leq 0 \leq s_r, \\ \boldsymbol{F}_2(\boldsymbol{u}_l) & 0 < s_l \leq s_r, \\ \text{undefined} & s_l > s_r. \end{cases} \tag{2.177}$$

The undefined cases correspond to where the Riemann problem has a delta function solution, that is where two wave fields meet. In the computations we have not used the Godunov method for such situations.

Even though the Godunov method applied to the single phase system converges in L_1 (see Table 2.1) there are large L_∞ errors also for smooth problems (see Table 2.2 and Figure 2.6). The reason for the method's behavior in Figure 2.6 can be found in the analysis of the Riemann problem in Section 2.3.3. Along the line $y = 1$ the Riemann problem in the y-direction corresponds to the situation in Figure 2.3a. Since there is no rarefaction wave solution, there will be no flux in the y-direction, and the method reduces to the one-dimensional Godunov method in the x-direction along this line. Hence, along $y = 1$ there will be pure transport, and no damping. A similar distortion occurs when the source is not located exactly in the middle of a grid cell. In [32] it was observed that the Godunov method can generate this type of error also in the much simpler case of a linear one-dimensional equation with variable coefficients.

A second order accurate scheme introduces less artificial viscosity and can therefore be expected to be more sensitive. To avoid oscillations at discontinuities, so called *TVD* methods are desirable [38]. These nonlinear methods use limiters to ensure that new artificial extrema are not introduced in the solution. At an extrema, TVD methods are at most first order accurate. We have implemented the *Nessyahu-Tadmor* method, [49]. It is a second order TVD method based on the Lax-Friedrichs structure. No Riemann problem needs to be solved, and the solution never has to be split up into characteristic fields, since the limiter is applied componentwise for systems. For a one-dimensional grid function V_j and flux function $\boldsymbol{f}$, it can be written

$$\begin{aligned} V_j^{n+1/2} &= U_j^n - \frac{\Delta t}{2\Delta x}(\boldsymbol{f}_j^n)' \\ V_j^{n+1} &= \frac{1}{2}(V_{j-1}^n + V_{j+1}^n) + \frac{1}{4}((V_{j+1}^n)' - (V_{j-1}^n)') \\ &\quad - \frac{\Delta t}{2\Delta x}\Big(\boldsymbol{f}(V_{j+1}^{n+1/2}) - \boldsymbol{f}(V_{j-1}^{n+1/2})\Big) \end{aligned} \tag{2.178}$$

The numerical derivatives $(V_j^n)'$ and $(\boldsymbol{f}_j^n)'$ are computed componentwise using a variant of the *minmod* limiter,

$$(V_j^n)' = \mathrm{MM}\Big(\beta(V_j^n - V_{j-1}^n),\ \frac{1}{2}(V_{j+1}^n - V_{j-1}^n),\ \beta(V_{j+1}^n - V_j^n)\Big), \tag{2.179}$$

and

$$\begin{aligned} (\boldsymbol{f}_j^n)' &= \mathrm{MM}\Big(\beta(\boldsymbol{f}(V_j^n) - \boldsymbol{f}(V_{j-1}^n)),\ \frac{1}{2}(\boldsymbol{f}(V_{j+1}^n) - \boldsymbol{f}(V_{j-1}^n)), \\ &\qquad \beta(\boldsymbol{f}(V_{j+1)}^n - \boldsymbol{f}(V_j^n))\Big), \end{aligned} \tag{2.180}$$

where

$$\mathrm{MM}(u_1, u_2, \dots) = \begin{cases} \min_k u_k & \text{if } u_k > 0, \forall k, \\ \max_k u_k & \text{if } u_k < 0, \forall k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.181}$$

We let $\beta = 2$ in our test cases. This gives better results than the classical choice $\beta = 1$.

A two-dimensional Nessyahu-Tadmor scheme for the homogeneous single phase system was obtained using the one-dimensional method together with dimensional splitting. *Strang splitting*, [53], was used to preserve second order accuracy. This means the that the scheme outlined above in each time step is first applied to $V_i = U_{ij}, \boldsymbol{f} = \boldsymbol{F}_1$ with $\Delta t/2$, then to $V_j = U_{ij}, \boldsymbol{f} = \boldsymbol{F}_2$ with Δt, and finally to $V_i = U_{ij}, \boldsymbol{f} = \boldsymbol{F}_1$ with $\Delta t/2$ again.

The result is not perfectly satisfactory. The convergence rate for the splitted Nessyahu-Tadmor scheme turns out to be somewhat slower than Lax-Friedrichs in L_∞ and only marginally higher in L_1 (see Table 2.1 and Table 2.2). A likely cause of this is the dimensional splitting. As a comparison, we also include results for a splitted version of the Lax-Friedrichs scheme, which are clearly inferior to the results for the unsplitted scheme. Also for the Godunov scheme the dimensional splitting aggravates the error, although, as seen in [32], the problem exists already in one dimension. See [33] for a further discussion.

In contrast, when we use a genuinely two-dimensional version of the Nessyahu-Tadmor scheme, [33], we get a considerable improvement of the results. Almost full second order accuracy is achieved (again, see Table 2.1 and Table 2.2). The two-dimensional scheme for the homogeneous case is written as

$$
\begin{aligned}
U_{ij}^{n+1/2} &= U_{ij}^{n} - \frac{\Delta t}{2\Delta x}(\boldsymbol{f}_{ij}^{n})_x - \frac{\Delta t}{2\Delta y}(\boldsymbol{g}_{ij}^{n})_y \qquad (2.182)\\
U_{ij}^{n+1} &= \frac{1}{4}(U_{i,j+1}^{n} + U_{i-1,j}^{n}U_{i+1,j}^{n} + U_{i,j-1}^{n})\\
&+ \frac{1}{8}\big((U_{i-1,j-1}^{n})_x - (U_{i+1,j-1}^{n})_x + (U_{i-1,j+1}^{n})_x - (U_{i+1,j+1}^{n})_x\big)\\
&- \frac{\Delta t}{4\Delta x}\Big(\boldsymbol{f}(U_{i+1,j-1}^{n+1/2}) - \boldsymbol{f}(U_{i-1,j-1}^{n+1/2}) + \boldsymbol{f}(U_{i+1,j+1}^{n+1/2}) - \boldsymbol{f}(U_{i-1,j+1}^{n+1/2})\Big)\\
&+ \frac{1}{8}\big((U_{i-1,j-1}^{n})_y - (U_{i-1,j+1}^{n})_y + (U_{i+1,j-1}^{n})_y - (U_{i+1,j+1}^{n})_y\big)\\
&- \frac{\Delta t}{4\Delta x}\Big(\boldsymbol{g}(U_{i-1,j+1}^{n+1/2}) - \boldsymbol{g}(U_{i-1,j-1}^{n+1/2}) + \boldsymbol{g}(U_{i+1,j+1}^{n+1/2}) - \boldsymbol{g}(U_{i+1,j-1}^{n+1/2})\Big)
\end{aligned}
$$

with $\boldsymbol{f} = \boldsymbol{F}_1$ and $\boldsymbol{g} = \boldsymbol{F}_2$. The derivative terms are computed dimensionwise in the same way as in the one-dimensional case, e.g. $(U_{ij})_x$ is given by using $V_i = U_{ij}$ in (2.179) and $(\boldsymbol{g}_{ij})_y$ by using $V_j = U_{ij}$ and $\boldsymbol{f} = \boldsymbol{F}_2$ in (2.180).

2.6.2 Multiple Phases

In the multiple phase case, we have mainly considered $N = 2$ for the delta equations and $N = 2, 3$ for the Heaviside equations. A few new problems are added to the numerical methods. In each time step a nonlinear system of equations must be solved. The Jacobian of this system can be singular. It may even happen that it does not have a solution.

We have only used the Lax-Friedrichs method to solve the the multiple phase systems. For the two-dimensional delta equations with inhomogeneous medium, (2.41), it can be written as

$$
\begin{aligned}
\eta_{ij}^2\boldsymbol{F}_0(U_{ij}^{n+1}) &= \frac{1}{4}\big(\eta_{i-1,j}^2\boldsymbol{F}_0(U_{i-1,j}^{n}) + \eta_{i+1,j}^2\boldsymbol{F}_0(U_{i+1,j}^{n})\\
&\quad +\eta_{i,j-1}^2\boldsymbol{F}_0(U_{i,j-1}^{n}) + \eta_{i,j+1}^2\boldsymbol{F}_0(U_{i,j+1}^{n})\big)\\
&\quad -\frac{\Delta t}{2\Delta x}\big(\boldsymbol{F}_1(\eta_{i+1,j}U_{i+1,j}^{n}) - \boldsymbol{F}_1(\eta_{i-1,j}U_{i-1,j}^{n})\big)\\
&\quad -\frac{\Delta t}{2\Delta y}\big(\boldsymbol{F}_2(\eta_{i,j+1}U_{i,j+1}^{n}) - \boldsymbol{F}_2(\eta_{i,j-1}U_{i,j-1}^{n})\big)\\
&\quad +\Delta t\boldsymbol{K}(U_{ij}^{n}, \partial_x\eta_{ij}, \partial_y\eta_{ij}). \qquad (2.183)
\end{aligned}
$$

The functions $\boldsymbol{F}_k$ and $\boldsymbol{K}$ are those in (2.49). The corresponding one-dimensional method for the Heaviside equations, (2.118), is

$$\begin{aligned}\eta_{nj}\boldsymbol{F}_1(U_j^{n+1}) \quad = \quad & \frac{1}{2}\left(\eta_{n,j-1}\boldsymbol{F}_1(U_{j-1}^n)+\eta_{n,j+1}\boldsymbol{F}_1(U_{j+1}^n)\right)\\ & -\frac{\Delta x}{2\Delta y}\left(\eta_{n,j+1}\boldsymbol{F}_2(U_{j+1}^n)-\eta_{n,j-1}\boldsymbol{F}_2(U_{j-1}^n)\right)\\ & +\Delta x\boldsymbol{K}(U_j^n,\partial_x\eta_{nj},\partial_y\eta_{nj}) \end{aligned} \tag{2.184}$$

where $\boldsymbol{F}_1$, $\boldsymbol{F}_2$ and $\boldsymbol{K}$ are given by (2.125) and (2.126). We see that in both cases it is necessary to solve a nonlinear system of equations of the type

$$\boldsymbol{F}(U^{n+1}) = \hat{\boldsymbol{m}}^n. \tag{2.185}$$

for each iteration, at each grid point. Actually, solving (2.185) is the first part of the evaluation of the flux functions, $\boldsymbol{F}_1\circ\boldsymbol{F}_0^{-1}$ and $\boldsymbol{F}_2\circ\boldsymbol{F}_0^{-1}$ in the delta case and $\boldsymbol{F}_2\circ\boldsymbol{F}_1^{-1}$ in the Heaviside case. The second part of these compositions is computed in the subsequent time step.

There are a number of problems associated with this evaluation. First, solving (2.185) can be difficult. For the delta case, an iterative solver must be used when $N > 2$. This is expensive and requires good initial values. When $N = 1, 2$ it can, however, be done analytically, see Appendix A.1.1. In the Heaviside case solving (2.185) can be reduced to solving polynomial equations of degree $N/2$ when $N = 1, \dots, 4$, see Section A.2. For the delta function case we should remark that if we can invert the function $\boldsymbol{F}_0$ we can also invert $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$, by drawing on the relationship between the functions in (2.47). This would be useful for instance when solving the steady state version of the delta equations.

Second, (2.185) may not have a solution. Although, for the exact solution of the PDE (2.185) should always be satisfied, truncation errors in the numerical scheme may have perturbed the solution so that $\hat{\boldsymbol{m}}^n$ is not in the range of $\boldsymbol{F}$. In this case the flux function must be continued outside its domain of definition. For the Heaviside equations this can be done by extending the domain of $\boldsymbol{F}_1$ and allow $\boldsymbol{u}$ to have an imaginary part. For the delta equations we use the least squares solution of (2.185) when $\hat{\boldsymbol{m}}^n \notin \boldsymbol{F}_0(D_f)$.

Third, in general, the Jacobian of $\boldsymbol{F}$ is singular at some points in the computational domain. This happens when two rays are parallel. For iterative methods that use the Jacobian, this is a problem.

When solving the delta equations we primarily use the choice of moments in (2.42). For $N = 2$ we have, however, also used an alternative choice, namely the moments in (2.82). As was seen in Remark 2.6, the flux functions are in general two-valued for this choice. For simple problems this ambiguity can be resolved by for instance choosing the value that maximizes the smoothness of the solution in some sense. Also for this choice (2.185) can be solved analytically, see Appendix A.1.2. The advantage with this choice is that it permits us to solve problems where waves meet head-on, i.e. where $w_1/|w_1| + w_2/|w_2| = 0$ with the notation of Corollary 2.5. On the other hand, two new singular branches are introduced, which depend on the choice of coordinate system. The Jacobians of the flux functions, (2.92), are unbounded when either

$$\cos\theta_1 = \cos\theta_2 \qquad \text{or} \qquad \sin\theta_1 = \sin\theta_2, \tag{2.186}$$

but not both. Solutions cannot be computed on these branches.

When the number 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.

The Heaviside function case is more difficult. The fictitious phases can obviously not be set to zero. Moreover, they cannot be set to the same as another phase. That would eliminate them from the equations. For instance, suppose there is only one relevant phase and a system with an even $N > 1$ is solved. If all fictitious phases were initialized at $x = 0$ to the same as the relevant phase, we would get $m_{ij}(0, y) = 0$ for all i, j and the system would hence only give the trivial solution. If N had been odd the system would have been reduced to the $N = 1$ case. Instead, the non-physical phases must be initialized to some other values. The solution can be sensitive to the right choice.

2.7 Numerical Results

In this section we show results of applying the equations derived in Section 2.3 and Section 2.4 to a number of different test cases. We consider both homogeneous and inhomogeneous media and use closures corresponding to $N = 1, 2, 3$ crossing rays at each point.

In the numerical computations the approximations described in Section 2.6 were used. The main purpose is to study the behavior of the exact solution to the PDEs, rather than their numerical properties. Therefore, in order to avoid problems with numerical artifacts we have primarily used the Lax-Friedrichs method. Although only first order accurate, it is the most robust and predictable method. Whenever another method was used this will be explicitly pointed out in the description of the cases below.

To measure the error of the computations we use the discrete L_1 and L_∞ norms, defined as

$$||U||_{1,h} = \Delta y \sum_i |U_i|, \qquad ||U||_{\infty,h} = \max_i |U_i|. \qquad (1D), \tag{2.187}$$

$$||U||_{1,h} = \Delta x \Delta y \sum_{i,j} |U_{ij}|, \quad ||U||_{\infty,h} = \max_{i,j} |U_{ij}|. \quad (2D). \tag{2.188}$$

In the computations we typically apply the exact solution as a Dirichlet boundary condition on the line $x = 0$. Where nothing else is said we use simple extrapolation, $U_{n+1,j} = U_{nj}$, etc. on the remaining boundaries.

2.7.1 Homogeneous 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 2.5, where the Lax-Friedrichs method was used to solve the $N = 1$ delta system with 40×80 grid points. The difficulties with using the Godunov and the splitted Nessyahu-Tadmor methods for the same problem are highlighted in Figure 2.6.

Convergence for the different methods are summarized in Table 2.1 and Table 2.2. The error in $u_1 = m_{10}$ is shown measured against the exact solution.

Finally, we considered the case when the source is inside the computational domain. We let the source be located at $\boldsymbol{s} = (0.5, 0.5)$ and solve the $N = 1$ δ-equations in the unit square. For the outer boundaries we use simple extrapolation as boundary condition. The (exact) source data is applied to all grid points satisfying $|R - ||(i\Delta x, j\Delta y) - \boldsymbol{s}||| < \Delta x/2$, i.e. on a circle of radius R around the source point. The radius R is reduced as the grid is refined, according to the relation $R = R_0(\Delta x/R_0)^p$ for some R_0 and $0 \leq p \leq 1$. Convergence for these two approaches are summarized in Table 2.3 where the errors at time $t = 1.0$ in the rectangle $[0, 0.4] \times [0, 1]$ is tabulated. Shown is the discrete L_1 and L_∞ norm of the components u_1, u_2 (summed for the L_1 norm). We use $R = 0.1$ and $p = 0, 0.5, 1$. For $p = 1$ the solution does not converge. Data is only given in the eight grid points surrounding the source and the solution becomes highly influenced by the grid orientation, cf. Figure 2.7 where 80×80 points were used. For $p = 0.5$ the method converges in L_1, with a rate of approximately 0.5. With $p = 0$, corresponding to a fixed radius $R = R_0$, we get first order convergence in both L_1 and L_∞.

For all computations a CFL number of 0.65 was used and the solution is shown/measured at time $t = 0.85$.

Two Point Sources

For this case we use two point sources. In the first computation the sources are located at coordinates $\boldsymbol{s}_1 = (-0.2,\, -0.2)$ and $\boldsymbol{s}_2 = (1.2,\, 1.2)$. They are smooth with exact solutions $\boldsymbol{w}_k = (\boldsymbol{x} - \boldsymbol{s}_k) \max(0, t - r_k)^2/r_k^2$, $r_k = ||\boldsymbol{x} - \boldsymbol{s}_k||$, $k = 1, 2$. The solution was computed in the unit square with 80×80 points and CFL=0.65. The exact solution was applied as a Dirichlet condition on all boundaries. In Figure 2.8 the solution at time $t = 3.0$ is shown for the δ-equations with $N = 1, 2$. The physically correct solution contains two phases and a measure solution is therefore suggested for the $N = 1$ system. The $N = 2$ system captures both phases. In this setup the solution contains waves meeting head-on. For the $N = 2$ system the choice of moments in (2.82) was therefore used.

In the second test the sources are located at coordinates $\boldsymbol{s}_1 = (-0.5,\, 0.5)$ and $\boldsymbol{s}_2 = (-0.5,\, 1.5)$ and they are discontinuous, with exact solutions $\boldsymbol{w}_k(t, \boldsymbol{x}) =$

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

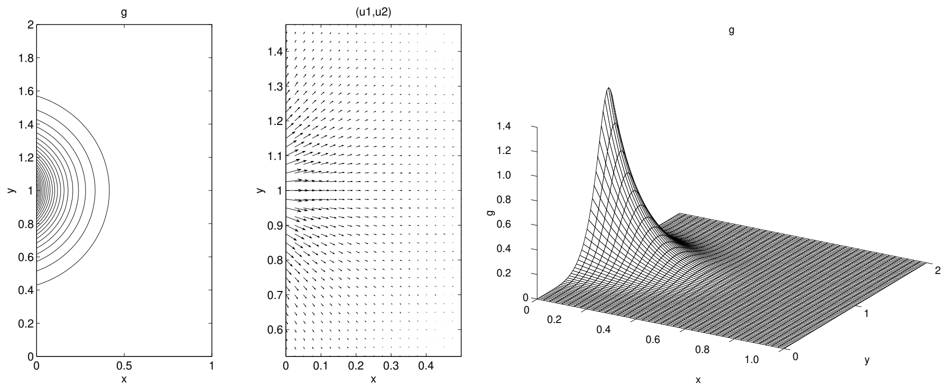
Table 2.1. *One point source.* L_1 norm of the errors for the δ-equations with $N = 1$. Here $\Delta x = \Delta y = 1/n$.

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

Table 2.2. *One point source.* L_∞ norm of the errors for the δ-equations with $N = 1$. Here $\Delta x = \Delta y = 1/n$.

	$p=0$				$p=0.5$				$p=1$	
	L_1		L_∞		L_1		L_∞		L_1	
n	error	order	error	order	error	order	error	order	error	order
20	2.99e-2		3.21e-1		6.83e-2		1.33		6.83e-2	
		1.17		1.59		1.63		1.69		0.40
40	1.33e-2		1.07e-1		2.21e-2		4.13e-1		5.19e-2	
		1.01		1.51		0.66		1.43		0.27
80	6.60e-3		3.76e-2		1.39e-2		1.53e-1		4.30e-2	
		1.08		1.24		0.53		0.01		0.03
160	3.13e-3		1.59e-2		9.67e-3		1.52e-1		4.21e-2	
		1.02		1.04		0.47		0.05		-0.07
320	1.54e-3		7.73e-3		7.00e-3		1.47e-1		4.41e-2	

Table 2.3. *One point source.* L_1 and L_∞ norm of the errors for the δ-equations with $N=1$ when source data is applied on a circle of radius $R = 0.1(\Delta x/0.1)^p$. Here $\Delta x = \Delta y = 1/n$.



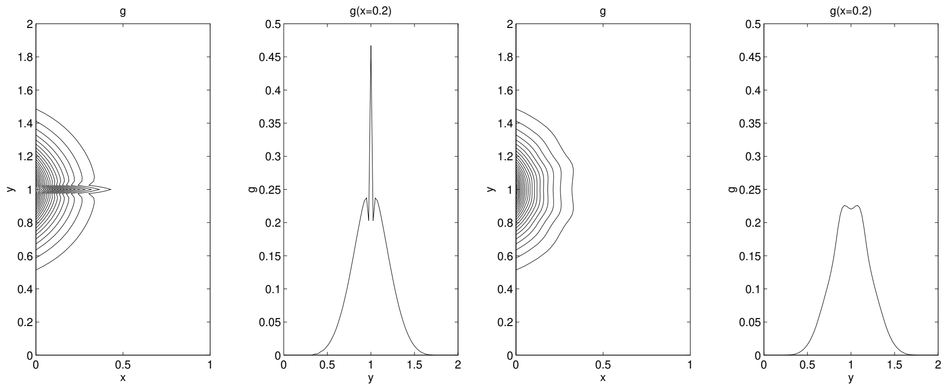
(a) Contour plot of the ray strength g (left) and the vector field $\boldsymbol{u} = (u_1, u_2)$ (right).

(b) Ray strength g.

Figure 2.5. *One point source.*

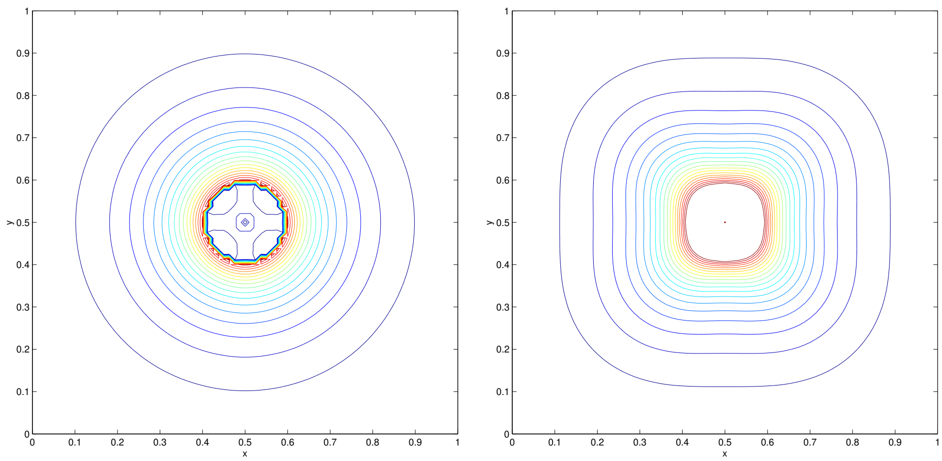
$A_k(\boldsymbol{x} - \boldsymbol{s}_k)H(t - r_k)/r_k^2$, $r_k = ||\boldsymbol{x} - \boldsymbol{s}_k||$, $k = 1, 2$. $A_1 = 0.75$, $A_2 = 1.25$. The solution is computed in the rectangle $[0,1] \times [0,2]$. In Figure 2.9 the solution at $t = 1.0$ of the δ-equations with $N = 1, 2$ is presented. This computation used CFL=0.65 and 80×160 grid points. For the $N = 2$ system, the exact solution was given at $x = 0$. For the $N = 1$ system, the first arriving wave was given at $x = 0$, i.e.

$$\boldsymbol{u} = \begin{cases} \boldsymbol{w}_1 & r_1 < r_2, \\ \boldsymbol{w}_2 & r_1 \geq r_2. \end{cases} \tag{2.189}$$



(a) Godunov (b) Nessyahu-Tadmor

Figure 2.6. *One point source.* Solution of the δ-equations with $N = 1$ using the Godunov and the Nessyahu-Tadmor schemes. Ray strength g shown as a contour plot and in a vertical cut at $x = 0.2$.



(a) $p = 0$ (b) $p = 1$

Figure 2.7. *One point source.* Contour plots of solution to the $N = 1$ delta equations with the source inside the computational domain. Data applied on a circle of radius $R = 0.1(\Delta x/0.1)^p$, with $\Delta x = 1/80$.

The $N = 1$ system cannot capture the two overlapping wave fields and compensates for this with a delta function. The $N = 2$ equations solves the problem

	$N=1$				$N=2$			
	L_1		L_∞		L_1		L_∞	
n	error	order	error	order	error	order	error	order
10	1.00e-1		2.35e-1		1.42e-1		2.36e-1	
		1.42		0.55		1.35		0.98
20	3.76e-2		1.61e-1		5.60e-2		1.20e-1	
		1.37		1.37		1.11		0.94
40	1.45e-2		6.23e-2		2.59e-2		6.25e-2	
		1.22		0.97		1.01		0.97
80	6.25e-3		3.18e-2		1.28e-2		3.19e-2	
		1.10		0.98		0.95		0.98
160	2.92e-3		1.61e-2		6.65e-3		1.62e-2	

Table 2.4. *Two point sources.* L_1 and L_∞ norm of the errors for the δ-equations with $N = 1, 2$. Here $\Delta x = \Delta y = 1/n$.

correctly, see Figure 2.9. Note, since $A_1 \neq A_2$, the delta function in the $N = 1$ case does not remain at $y = 1$. Hence, the solution does not give the first arriving wave, cf. Section 2.5.3.

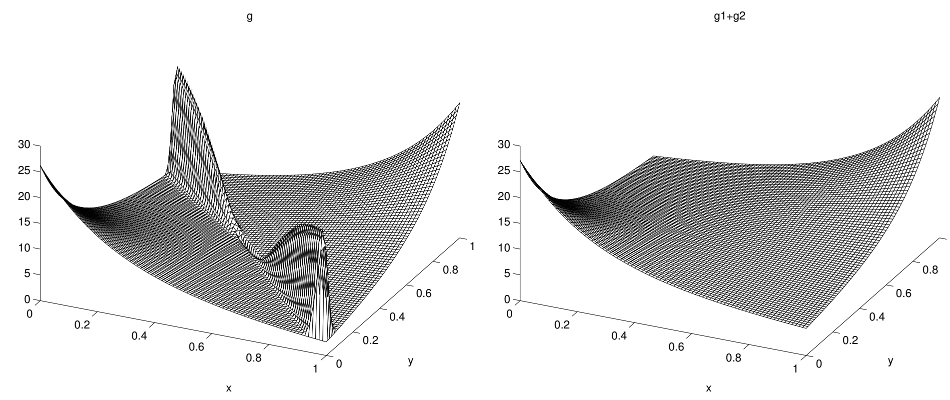
Convergence for the last problem is shown in Table 2.4. In these tests the sample time is $t = 1.5$ and the error is the discrete L_1 norms of the solution's components, u_k, summed, and the discrete L_∞ norm over all components. Also, $A_1 = A_2 = 1$, so that in the $N = 1$ case the delta function stays at $y = 1$. The errors are then measured against the exact solution of the closest source, and only in the area $[0, 2] \times [0, 2] \setminus [0, 2] \times [0.9, 1.1]$.

Finally, the Heaviside system with $N = 2$ was solved using 256 points, CFL=0.3 and the same source location as in the previous test. Solution is shown in Figure 2.10.

Three Point Sources

We now treat the case of 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 $\boldsymbol{w}_k(t, \boldsymbol{x}) = A_k(\boldsymbol{x} - \boldsymbol{s}_k)H(t - r_k)/r_k^2$, $r_k = ||\boldsymbol{x} - \boldsymbol{s}_k||$, $k = 1, 2, 3$. $A_1 = 1.25$, $A_2 = 0.75$ and $A_3 = 1.0$. The solution is computed in the rectangle $[0, 1] \times [0, 2]$. Figure 2.11 shows the solution at $t = 1.0$ of the δ-equations with $N = 1, 2, 3$. For these computations CFL=0.65 and the grid sizes were 80×160 ($N = 1$), 160×320 ($N = 2$) and 40×80 ($N = 3$). For the $N = 3$ system, the exact solution was given at $x = 0$. For the $N = 2$ system, the first two arriving waves were given at $x = 0$, i.e.

$$\boldsymbol{u}_1 = \boldsymbol{w}_2, \qquad \boldsymbol{u}_2 = \begin{cases} \boldsymbol{w}_1 & r_1 < r_3, \\ \boldsymbol{w}_3 & r_1 \geq r_3. \end{cases} \tag{2.190}$$



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

Figure 2.8. *Two point sources.* Solution of the δ-equations with $N = 1, 2$. Pictures show total ray strength, i.e. g and $g_1 + g_2$ respectively.

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

$$\boldsymbol{u} = \begin{cases} \boldsymbol{w}_1 & r_1 < r_2,\ r_1 < r_3, \\ \boldsymbol{w}_2 & r_2 \leq r_1,\ r_2 < r_3, \\ \boldsymbol{w}_3 & r_3 \leq r_1,\ r_3 \leq r_2. \end{cases} \tag{2.191}$$

As expected, the $N = 3$ system is the only one solving this problem correctly. Delta functions appear in the solutions of the $N = 1, 2$ systems. Note that the $N = 2$ system, somewhat surprisingly, gives the waves of the outer two sources ($k = 1, 3$) in the area along the line $y = 1$, cf. Figure 2.11e.

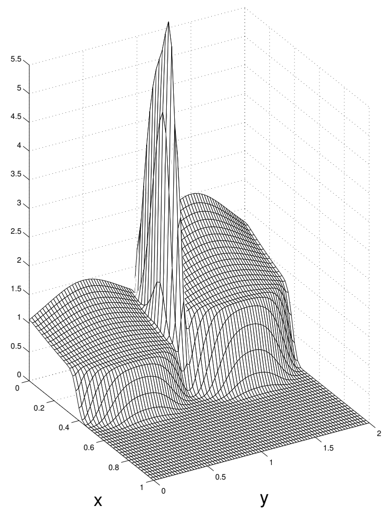
We also solved this problem with the $N = 3$ Heaviside system using 256 points and CFL=0.3, see Figure 2.12. For this case we also made a convergence study. The errors, measured at $x = 0.5$ in the discrete L_1 (sum of the errors in all three components) and L_∞ norm, are tabulated in Table 2.5.

Simple Caustic

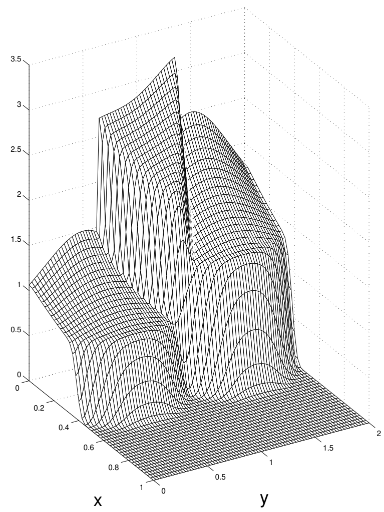
In this case we give data at $x = 0$ such that a caustic forms,

$$\sin\theta(0, y) = \begin{cases} \frac{4\cos\frac{\pi}{5}(y-1)}{(y-1)^2 + 4\cos^2\frac{\pi}{5}}, & y > 1, \\ 0, & y \leq 1, \end{cases} \tag{2.192}$$

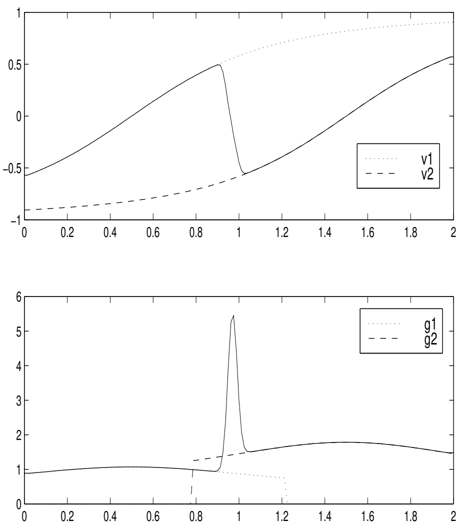
cf. the exact ray traced solution in Figure 2.13a. The solution was computed in the square $[0, 2] \times [0, 2]$ for the δ-equations with $N = 1, 2$ and the Heaviside equations with $N = 2$. For the δ-equations, the solution was sampled at $t = 3.0$,



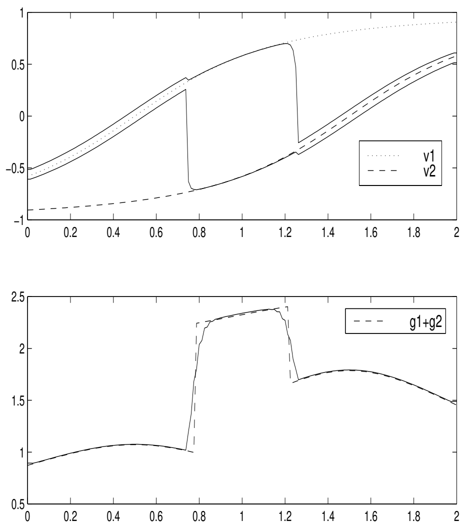
(a) $N = 1$



(b) $N = 2$

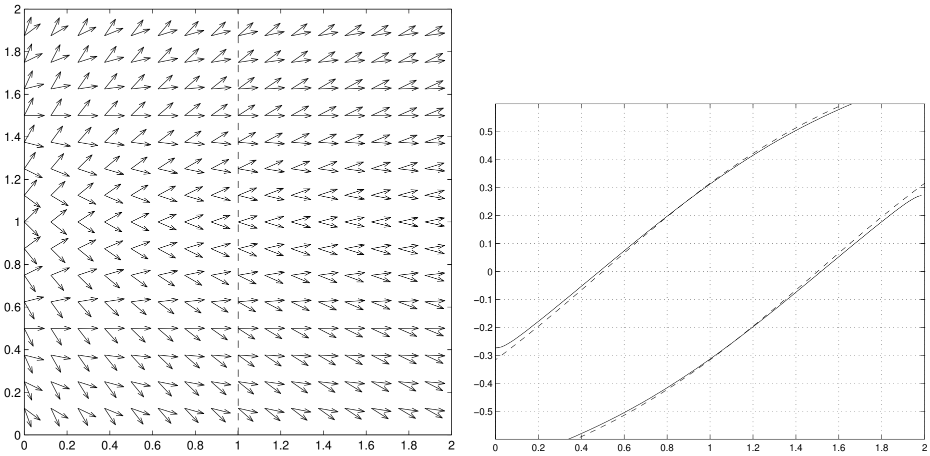


(c) $N = 1$, sine of ray angle (above), ray strength g (below)



(d) $N = 2$, sine of ray angles (above), total ray strength $g_1 + g_2$ (below)

Figure 2.9. *Two point sources.* Solution of the δ-equations with $N = 1, 2$. Top pictures show total ray strength, i.e. g and $g_1 + g_2$ respectively. Bottom pictures show solution in a cut at $x = 0.2$, computed (solid) and exact (dotted, dashed).



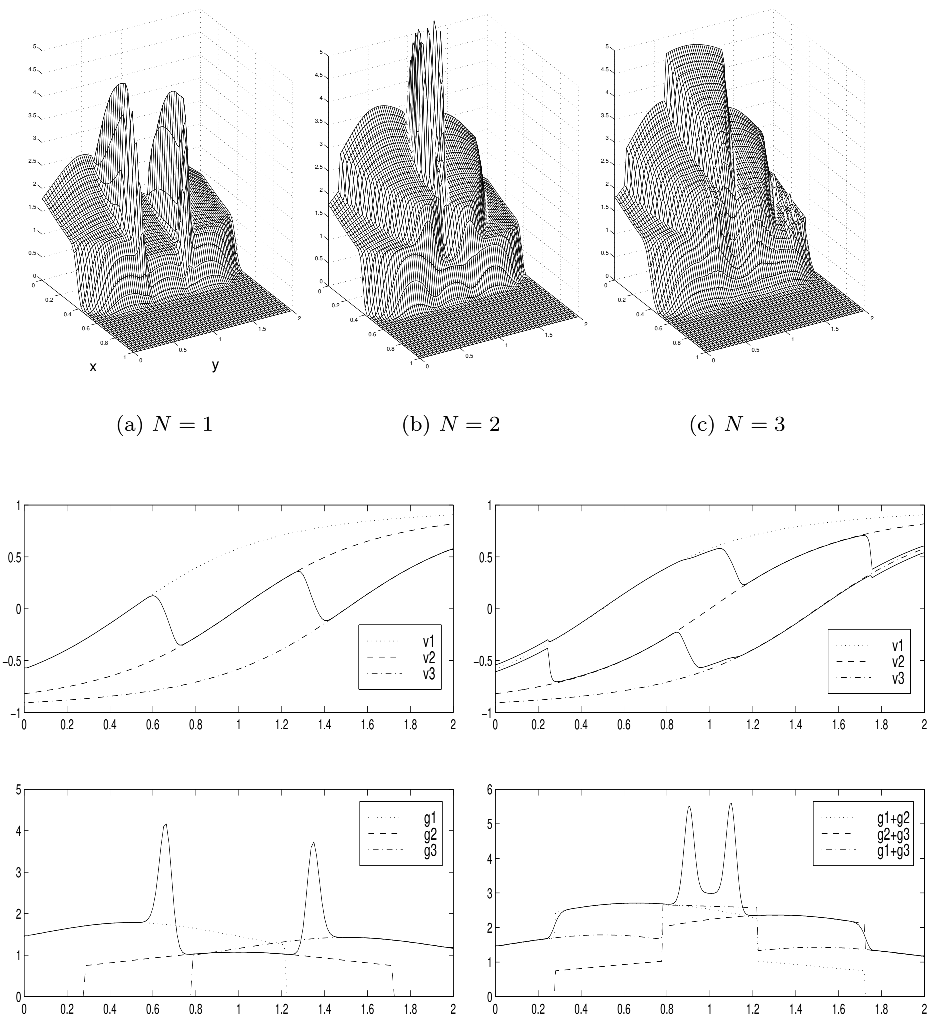
(a) Vector fields $(\cos\theta_1, \sin\theta_1)$ and $(\cos\theta_2, \sin\theta_2)$.

(b) Sine of ray angles in a cut at $x = 1.0$. Dashed line shows ray traced solution.

Figure 2.10. *Two point sources.* Solution to $N = 2$ Heaviside system.

	L_1		L_∞	
n	error	order	error	order
128	3.05e-1		1.26e-1	
		0.92		0.89
256	1.61e-1		6.80e-2	
		0.96		0.94
512	8.29e-2		3.55e-2	
		0.98		0.97
1024	4.21e-2		1.81e-2	
		0.99		0.98
2048	2.12e-2		9.16e-3	

Table 2.5. *Three point sources.* L_1 and L_∞ norm of the errors for the $N = 3$ Heaviside system. Here $\Delta y = 2/n$.

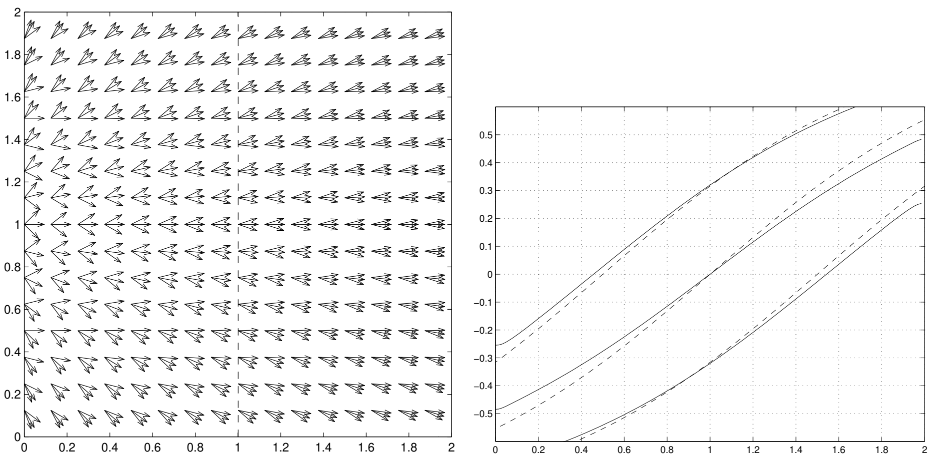


(a) $N = 1$ (b) $N = 2$ (c) $N = 3$

(d) $N = 1$, sine of ray angle (above), ray strength g (below)

(e) $N = 2$, sine of ray angles (above), total ray strength $g_1 + g_2$ (below)

Figure 2.11. *Three point sources.* Solution of the δ-equations with $N = 1, 2, 3$. Top pictures show total ray strength, i.e. g, $g_1 + g_2$ and $g_1 + g_2 + g_3$ respectively. Bottom pictures show solution in a cut at $x = 0.2$, computed (solid) and exact (dotted, dashed, dashdotted).



(a) Vector fields $(\cos\theta_1, \sin\theta_1)$, $(\cos\theta_2, \sin\theta_2)$ and $(\cos\theta_3, \sin\theta_3)$.

(b) Sine of ray angles in a cut at $x = 1.0$. Dashed line shows ray traced solution.

Figure 2.12. *Three point sources.* Solution to $N = 3$ Heaviside system

CFL=0.65, and the grid size was 256×256 ($N = 1$) and 512×512 ($N = 2$). The Heaviside equations used 1024 points and CFL=0.4. Initial data was $\theta_1(0, y)$ as above and $\theta_2(0, y) = \theta_1(0, 2)$. In Figure 2.13 and Figure 2.14 the result is shown. Although the rays concentrate along the caustic, at each spatial point there are at most two crossing rays. The δ-equations perform relatively poorly on this case, and the $N = 2$ does not manage to capture the second phase at this level of resolution (away from the caustic, at $x > 1.8$ it gives a better solution, though). The Heaviside equations with $N = 2$, on the other hand, computes the second phase accurately also close to the caustic.

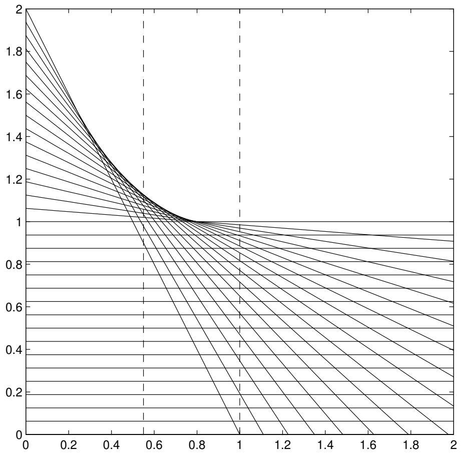
For the Heaviside equations we have made a convergence study. The solution is sampled before ($x = 0.2$), inside ($x = 0.55$) and just after ($x = 1.0$) the caustic for different grid sizes. To measure the error against the exact solution, which can be multivalued, let $(u_{ex}(x, t), y(x, t))$ with $t \in [0, 1]$ be the parameterized exact solution at x. We look at the residual,

$$\mathrm{res}(x, t) = \min(|u_{ex}(x, t) - u_1(x, y(t))|, |u_{ex}(x, t) - u_2(x, y(t))|). \tag{2.193}$$

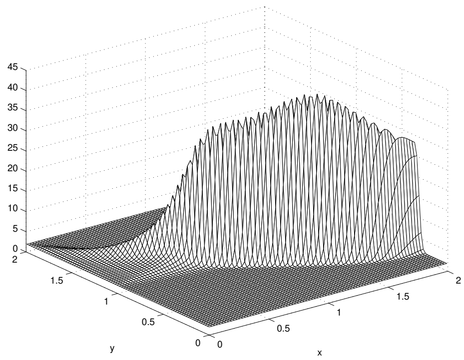
where $u_k = \sin\theta_k$ is the computed solution. The discrete L_1 norm of $\mathrm{res}(x, t)$, in the sense

$$||\cdot||_{1,h} = \Delta y \sum_j \sum_{y(t)=j\Delta y} \mathrm{res}(x, t), \tag{2.194}$$

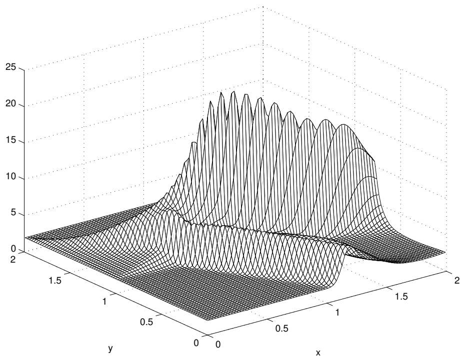
is tabulated in Table 2.6. Only after the caustic did we obtain less than full first order convergence rate.



(a) Ray traced solution.

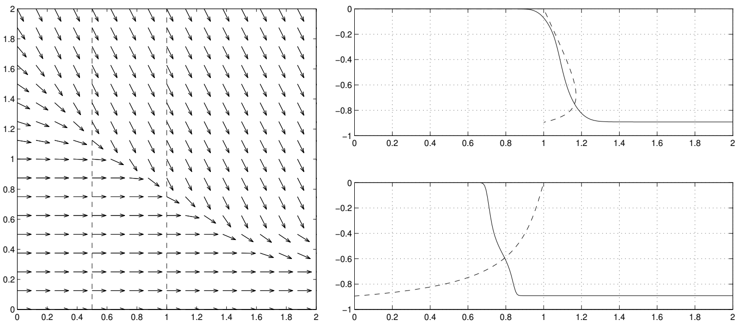


(b) $N = 1$

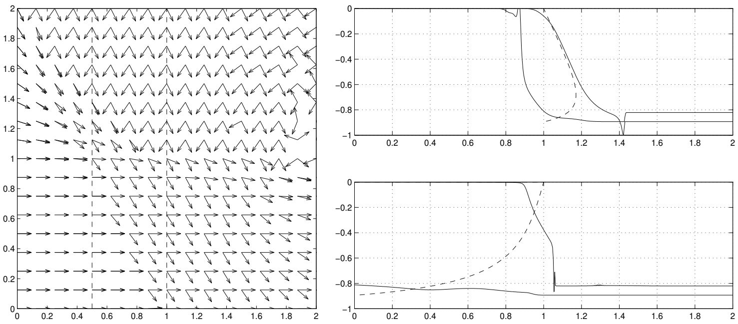


(c) $N = 2$

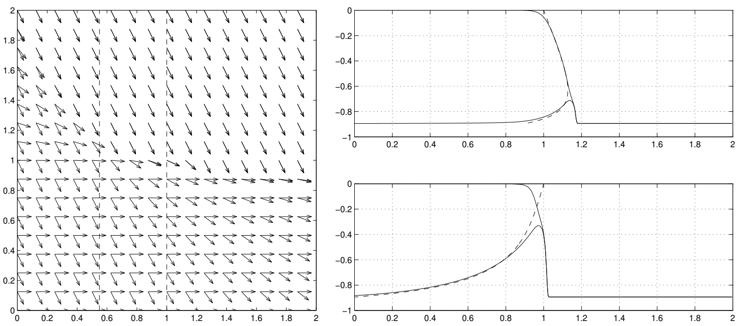
Figure 2.13. *Simple caustic.* Solution of the δ-equations with $N = 1, 2$. Bottom pictures show total ray strength, i.e. g and $g_1 + g_2$ respectively.



(a) δ-equations, $N = 1$



(b) δ-equations, $N = 2$



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

Figure 2.14. *Simple caustic.* Left figures contain quiver plots of ray angles for solution to delta and Heaviside equations. Right figures show sine of ray angles (solid) in cuts at $x = 0.55$ (above) and $x = 1.0$ (below) together with the corresponding values for a ray traced solution (dashed).

L_1	$x = 0.2$		$x = 0.55$		$x = 1.0$	
n	error	order	error	order	error	order
128	1.61e-2		3.12e-2		8.08e-2	
		0.87		0.68		0.41
256	8.79e-3		1.95e-2		6.10e-2	
		0.89		0.75		0.45
512	4.75e-3		1.16e-2		4.45e-2	
		0.92		0.77		0.48
1024	2.52e-3		6.81e-3		3.18e-2	
		0.94		0.91		0.44
2048	1.31e-3		3.62e-3		2.34e-2	

Table 2.6. *Simple caustic.* L_1 norm of the errors of the multivalued solution at $x = 0.2$, $x = 0.55$ and $x = 1.0$ for $N = 2$ Heaviside system. Here $\Delta y = 2/n$.

Focus

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

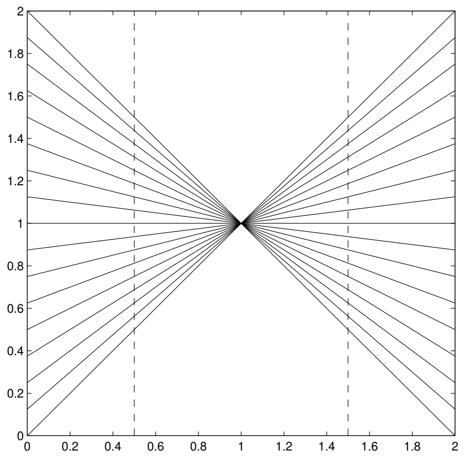
$$\tan\theta(0, y) = 1 - y. \tag{2.195}$$

Just like in the caustic case above the δ-equations with $N = 1, 2$ and the Heaviside equations with $N = 2$ were solved in the square $[0, 2] \times [0, 2]$. For the δ-equations, the solution was sampled at $t = 4.0$, CFL=0.65, and the grid size was 80×80. The Heaviside equations used 512 points and CFL=1.0. Initial data was

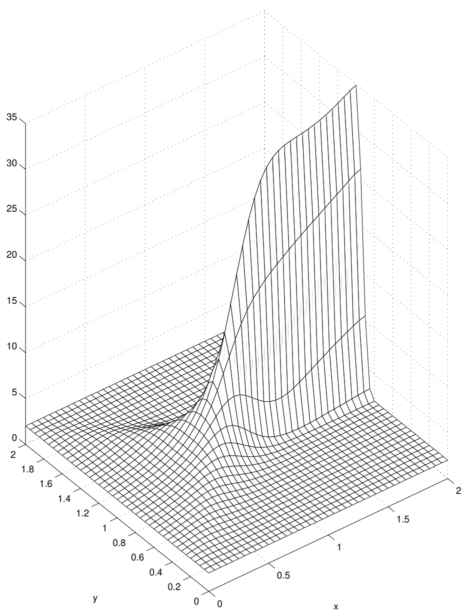
$$\begin{aligned} \theta_1(0, y) &= \min(\theta_f(y), \theta_b(y)), \\ \theta_2(0, y) &= \max(\theta_f(y), \theta_b(y)), \end{aligned} \tag{2.196}$$

with θ_f given by (2.195) and $\theta_b(y) = \mathrm{sgn}(y - 1)\pi/4$. Results are in Figure 2.15 and Figure 2.16. The focus point in the exact solution should formally contain an infinite number of phases. All equations give the right solution up to the focus at $x = 1$, but none manages to give the correct solution beyond the focus point.

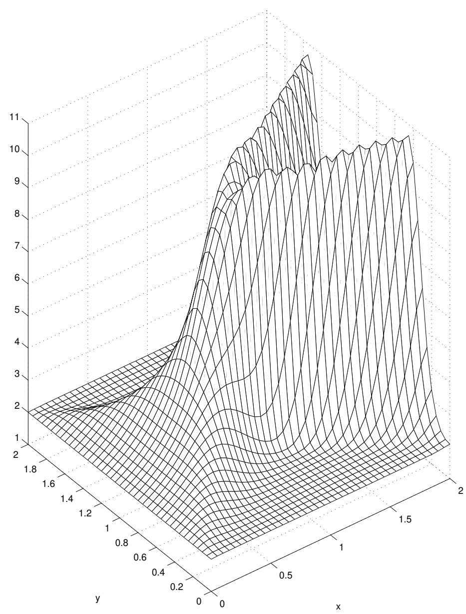
The behavior of the $N = 1$ delta equations can be explained by the discussion in Section 2.5.3. The focus point can be seen as a point where an infinite number of particles collide. Their total momentum in the y-direction is zero so the result is that all particles stick together, yielding the mass concentration, and move from the collision point along the line $y = 1$.



(a) Ray traced solution.

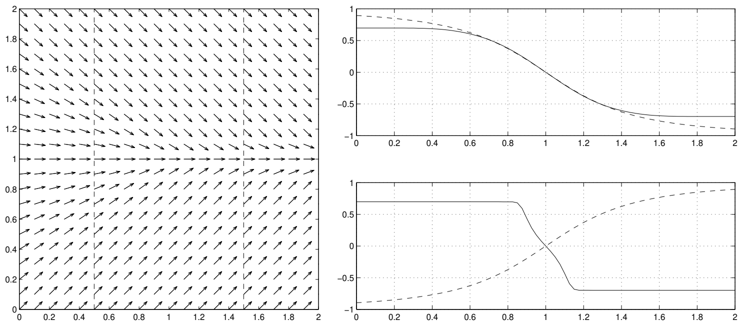


(b) $N = 1$

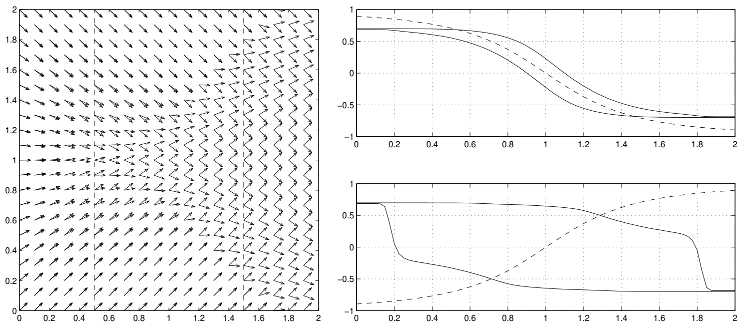


(c) $N = 2$

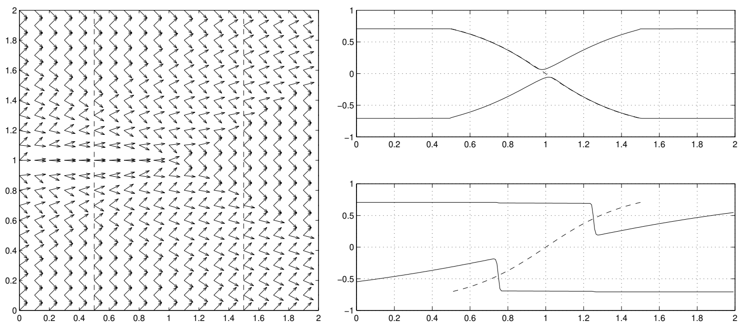
Figure 2.15. *Focus.* Solution of the δ-equations with $N = 1, 2$. Pictures show total ray strength, i.e. g and $g_1 + g_2$ respectively.



(a) δ-equations, $N = 1$



(b) δ-equations, $N = 2$



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

Figure 2.16. *Focus.* Left figures contain quiver plots of ray angles for delta and Heaviside equations. Right figures show sine of ray angles (solid) in cuts at $x = 0.5$ (above) and $x = 1.5$ (below) together with the corresponding values for a ray traced solution (dashed).

2.7.2 Inhomogeneous Problems

Tilted Interface 1

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

$$\eta(x,y) = \begin{cases} 1 & t \leq 0, \\ 1 + 4t^2(3-4t) & 0 < t \leq \frac{1}{2}, \\ 2 & \frac{1}{2} < t, \end{cases} \qquad t = (x-0.8)\cos\frac{\pi}{16} - y\sin\frac{\pi}{16}. \tag{2.197}$$

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

In the first test we used two waves with angles $\theta_1 = \pi/4$, $\theta_2 = -\pi/6$. The delta system with $N = 2$ was solved using 80×80 grid points and CFL=0.65. The solution at time $t = 5.0$ is shown in Figure 2.17.

Next, we add a third wave with angle $\theta_3 = \pi/8$ and solve the Heaviside system with $N = 3$. This time we use 256 points and CFL=1.0. Again, solution is shown in Figure 2.17.

The exact solutions of these problems are given by Snell's law, (2.15, 2.16). We study how the computed solutions converge to this solution when we refine the mesh. In the delta case we sample the solution at time $t = 5.0$ and $x = 2$. To avoid boundary effects we only look at the interval $1.5 \leq y \leq 2$ for the first wave and $0 \leq y \leq 0.5$ for the second. In Table 2.7 the errors in these intervals are given. The tabulated errors are the the sum of the discrete L_1 norms of the errors in the components u_k, and the total discrete L_∞ norm of the component errors. The Heaviside case was treated the same way, only that the solution was sampled at $x = 1.75$ and that we took the solution in the interval $0.5 \leq y \leq 1.5$ for all three waves. The errors, computed as in the delta case, are shown in Table 2.7. In both cases we obtain the expected first order convergence rate.

Concave Lens

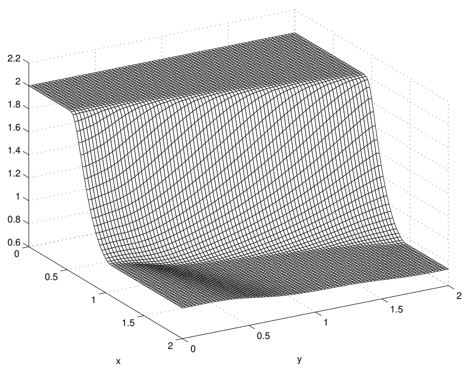
In this test case we have an index of refraction modeling a smooth concave lens,

$$\eta(x,y) = \begin{cases} 1 & D > 1, \\ \frac{4}{3-\cos\frac{\pi y}{s}} & D \leq 1, \end{cases} \tag{2.198}$$

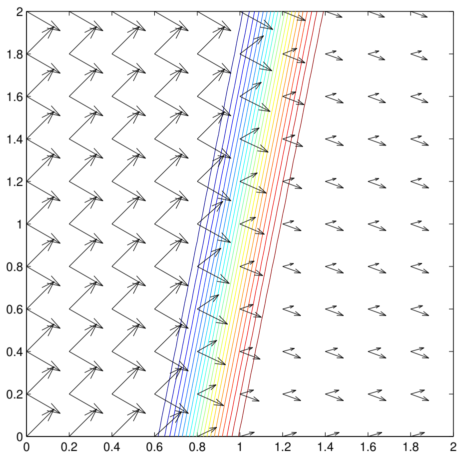
with

$$D = -\left(\frac{x-1}{0.3}\right)^2 + \left(\frac{y}{0.8}\right)^2, \qquad s = 0.3\sqrt{1 + \left(\frac{x-1}{0.3}\right)^2}. \tag{2.199}$$

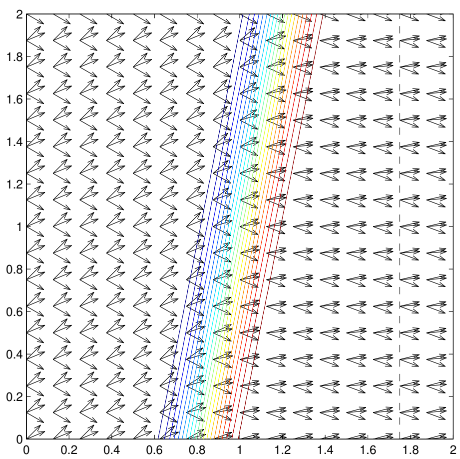
A plane wave, orthogonal to the line $x = 0$, is used as source. The δ-equations with $N = 1$ are solved in the square $[0,2] \times [0,2]$ with 80×80 points and CFL=0.65. Solution at time $t = 3.0$ is shown in Figure 2.18.



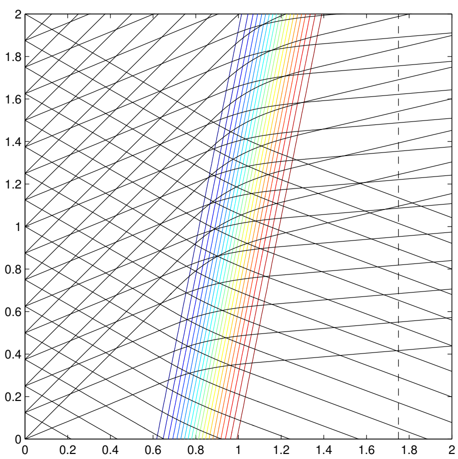
(a) Total ray strength $g_1 + g_2$.



(b) Vector fields (u_1, u_2) and (u_3, u_4) with contour lines of the index of refraction η overlayed.

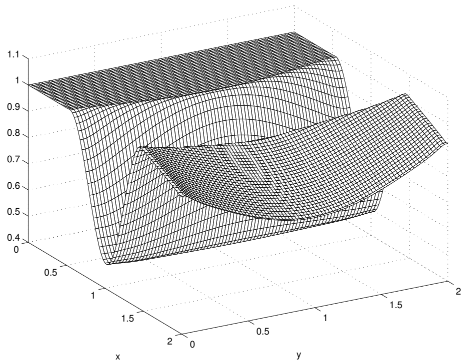


(c) Vector fields $(\cos\theta_1, \sin\theta_1)$, $(\cos\theta_2, \sin\theta_2)$ and $(\cos\theta_3, \sin\theta_3)$ with contour lines of index of refraction η superimposed.

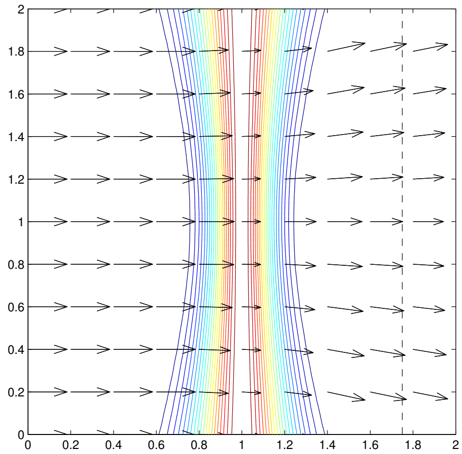


(d) Ray traced solution with contour lines of index of refraction η superimposed.

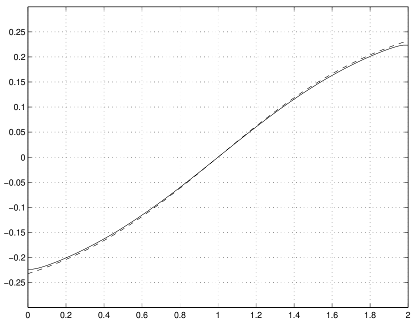
Figure 2.17. *Tilted interface 1.* Result with two plane waves for the δ-equations with $N = 2$ above. Below, result with three plane waves for the Heaviside system with $N = 3$, together with a raytraced solution.



(a) Ray strength g.



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



(c) Sine of ray angle in a cut at $x = 1.75$. Dashed line shows ray traced solution.

Figure 2.18. *Concave lens.* Solution of δ-equations with $N = 1$.

	δ-equations, $N=2$				H-equations, $N=3$			
	L_1		L_∞			L_1		L_∞	
n	error	order	error	order	n	error	order	error	order
20	5.28e-2		5.07e-2		128	1.95e-2		9.66e-3	
		0.65		0.73			1.03		1.03
40	3.37e-2		3.05e-2		256	9.52e-3		4.72e-3	
		0.80		0.81			1.02		1.02
80	1.94e-2		1.73e-2		512	4.71e-3		2.33e-3	
		0.90		0.89			1.01		1.01
160	1.04e-2		9.38e-3		1024	2.34e-3		1.16e-3	
		0.94		0.94			1.00		1.00
320	5.40e-3		4.89e-3		2048	1.17e-3		5.78e-4	

Table 2.7. *Tilted interface 1.* The L_1 and L_∞ norm of the errors for the $N=2$ delta system and the $N=3$ Heaviside system. Error measured against the values predicted by Snell's law. Here $\Delta y = 2/n$ and $\Delta x = \Delta y$ in the delta case.

Tilted Interface 2

We here use a smoothed interface similar to the one above. This time it is sharper and more tilted. It is also translated so that it crosses the line $x=0$. The index of refraction is given by

$$\eta(x,y) = 1.5 + \frac{1}{\pi}\arctan(40t), \qquad t = (x+1)\cos\frac{\pi}{4} - y\sin\frac{\pi}{4}. \tag{2.200}$$

A plane wave with $\theta(0,y)=0$ enters at $x=0$. When it is refracted in the interface a second and third phase appears. We have computed the solution in the square $[0,2]\times[0,2]$ for the δ-equations with $N=1,2$ and the Heaviside equations with $N=2,3$. Different aspects of the solutions are shown in Figure 2.19, Figure 2.20 and Figure 2.21. The $N=2$ systems manage to capture the second phase quite well, and the $N=3$ also captures the weak third phase.

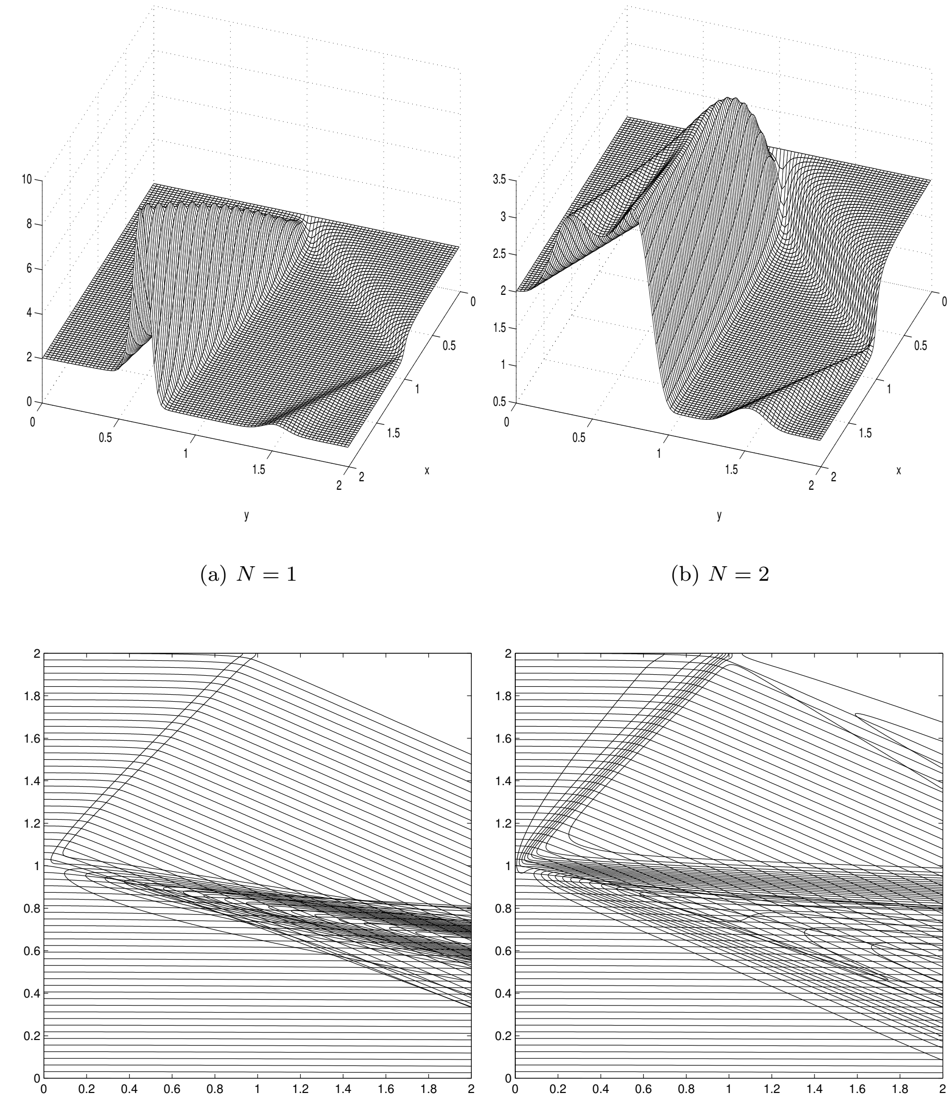
The δ-equations were computed on a 512×512 grid with CFL=0.65. The solution is shown at $t=5.0$. For the Heaviside equations 1024 points were used for $N=2$ and 2048 points for $N=3$. In both cases CFL=1.0. The initial data was

$$\begin{aligned} \theta_1(0,y) &= \min(0,\theta_b(y)), \\ \theta_2(0,y) &= \max(0,\theta_b(y)), \\ \theta_3(0,y) &= \frac{3}{4}(\theta_1(0,y)+\theta_2(0,y)). \end{aligned} \tag{2.201}$$

where $\theta_b(y) = \mathrm{sgn}(y-1)\pi/4$.

Wedge

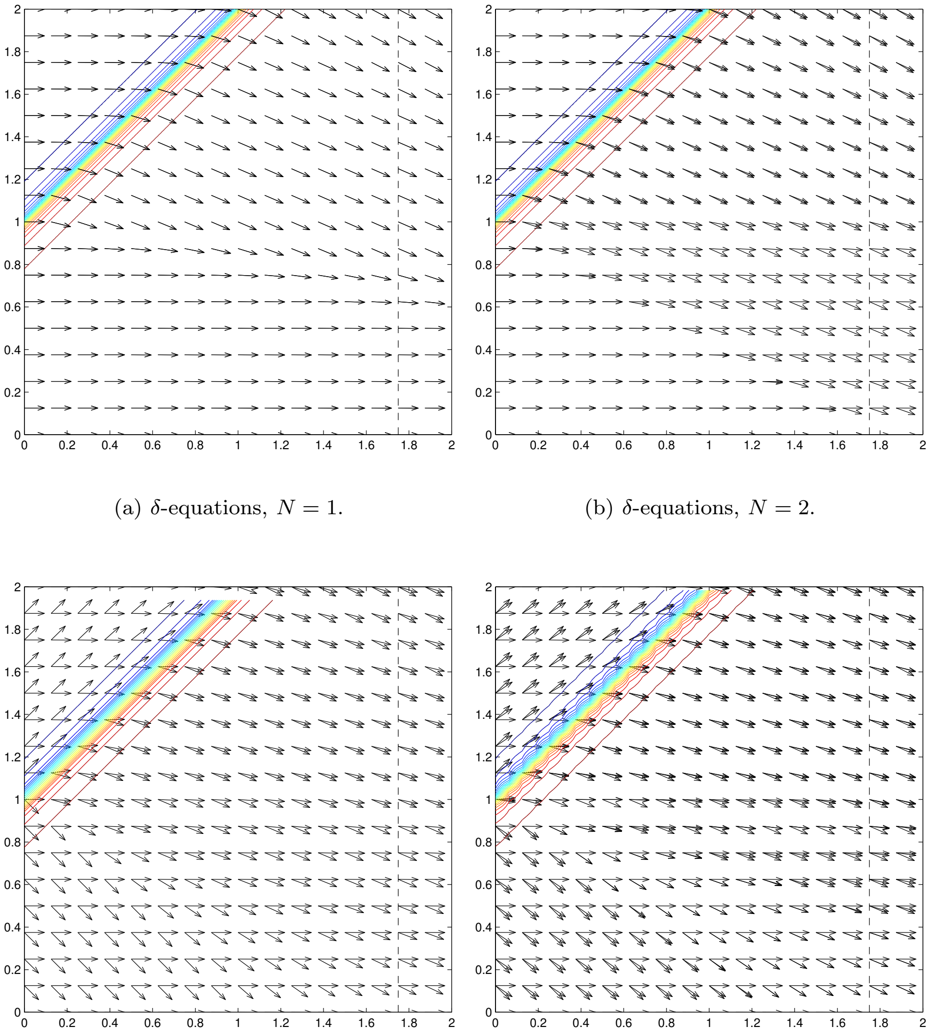
This test case is similar to the previous one. The same experiments are carried out, with a plane wave entering at $x=0$ and the solution of the δ- and the



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

(c) $N = 1$ (d) $N = 2$

Figure 2.19. *Tilted interface 2.* Amplitude results for δ-equations with $N = 1, 2$. Top pictures show total ray strength, i.e. g and $g_1 + g_2$ respectively. Bottom two pictures show level curves of the same quantities, with a ray traced solution superimposed.



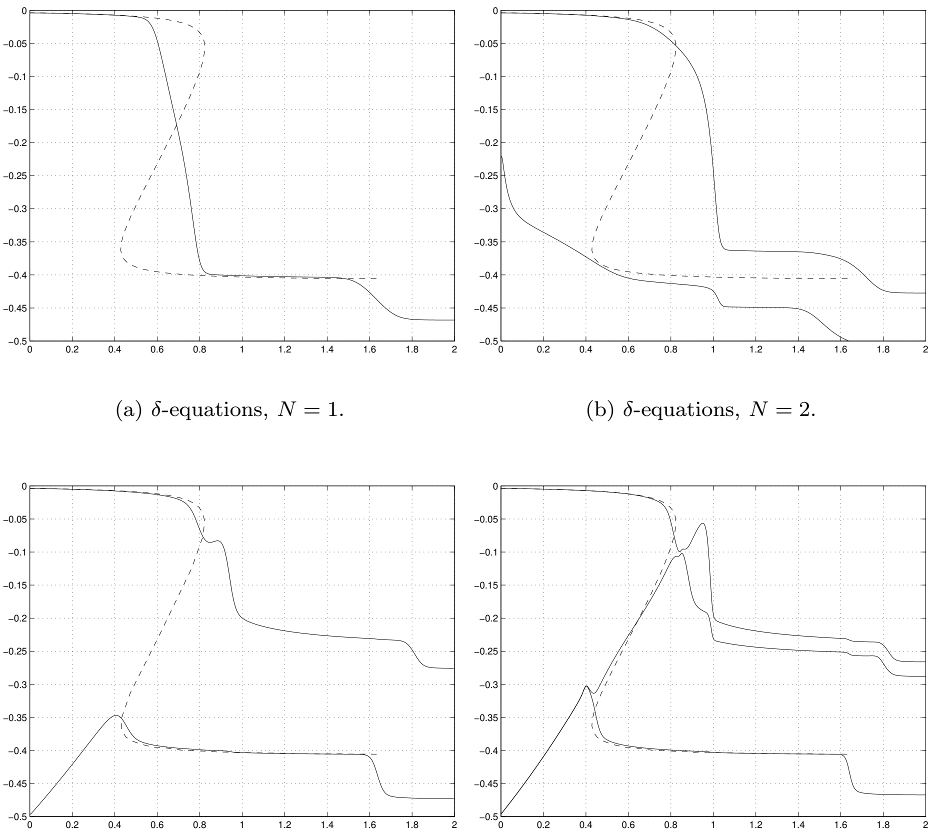
(a) δ-equations, $N = 1$.

(b) δ-equations, $N = 2$.

(c) H-equations, $N = 2$.

(d) H-equations, $N = 3$.

Figure 2.20. *Tilted interface 2.* Quiver plots of ray angles for delta and Heaviside equations with different N. A contour plot of the index of refraction is overlayed on the solution.



(a) δ-equations, $N = 1$.

(b) δ-equations, $N = 2$.

(c) H-equations, $N = 2$.

(d) H-equations, $N = 3$.

Figure 2.21. *Tilted interface 2.* Solution in a cut at $x = 1.75$ for delta and Heaviside equations with different N. Shown is sine of the ray angles (solid) and the corresponding values for a ray traced solution (dashed).

Heaviside equations is computed in the square $[0,2] \times [0,2]$. However, another index of refraction is used,

$$\eta(x, y) = 1.5 - \frac{1}{\pi} \arctan(k((y-1)^2 - 0.3(x-0.5))), \tag{2.202}$$

a model of a smooth wedge. Results are in Figure 2.22, Figure 2.23 and Figure 2.24. Again, two new phases appear when the plane wave is focused by the wedge and they are captured by the $N = 2, 3$ equations. Here the two caustics that appear are stronger than in the tilted interface case, and the solution to the δ-equations with $N = 2$ appears to include them correctly.

For the δ-equations with $N = 1$ solution is sampled at $t = 4.0$ For the $N = 3$ Heaviside system, CFL=0.9. Otherwise, the same grids, initial data, CFL numbers and sample times as in the *Tilted interface 2* case above were used.

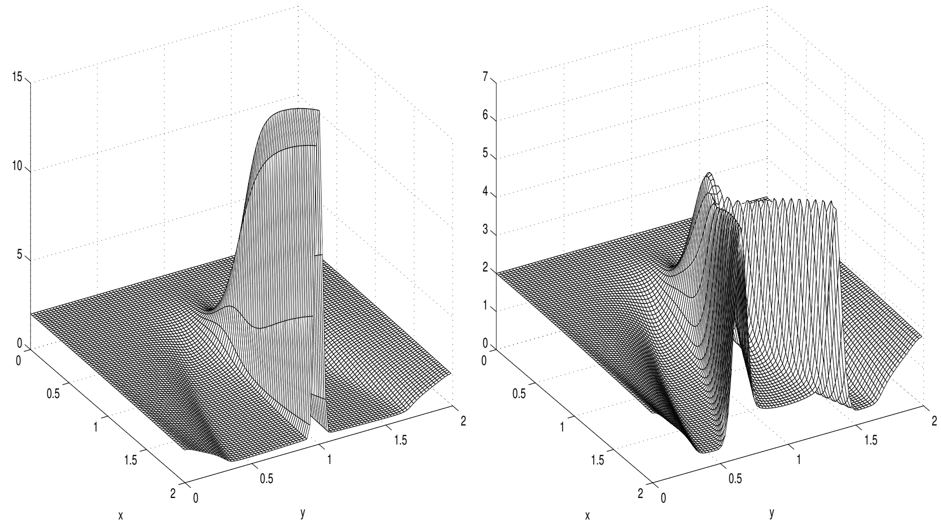
Convex Lens

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

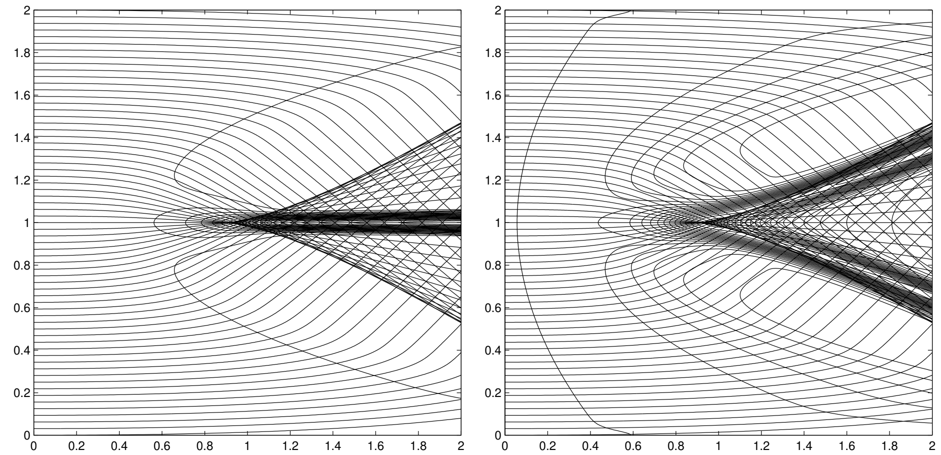
$$\eta(x, y) = \begin{cases} 1 & d > 1, \\ \left(\frac{4}{3-\cos(\pi d)}\right)^2 & d \leq 1, \end{cases} \qquad d = \left(\frac{x-0.5}{0.2}\right)^2 + \left(\frac{y-1}{0.8}\right)^2. \tag{2.203}$$

As in the previous two cases the δ- and Heaviside equations were solved in the square $[0,2] \times [0,2]$. Figure 2.25, Figure 2.26 and Figure 2.27 show various features of the solutions. In this case the exact solution develops up to five phases in the focus area around $(1,1)$ and settles with three phases behind this point. None of the equations manages to capture fully all three phases, although in general the higher N the better they perform.

For the $N = 3$ Heaviside system, CFL=0.6. Otherwise, the same grids, initial data, CFL numbers and sample times as in the tilted interface 2 case above were used.

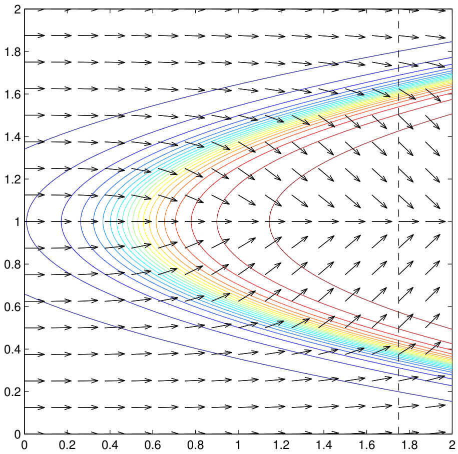


(a) $N = 1$ (b) $N = 2$

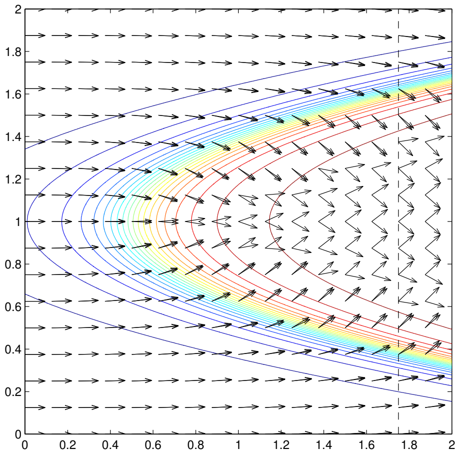


(c) $N = 1$ (d) $N = 2$

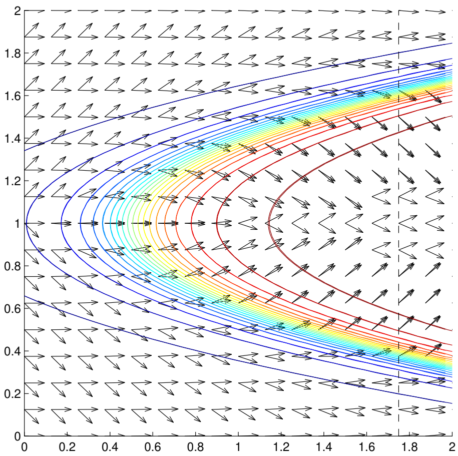
Figure 2.22. *Wedge.* Amplitude results for δ-equations with $N = 1, 2$. Top pictures show total ray strength, i.e. g and $g_1 + g_2$ respectively. Bottom two pictures show level curves of the same quantities, with a ray traced solution superimposed.



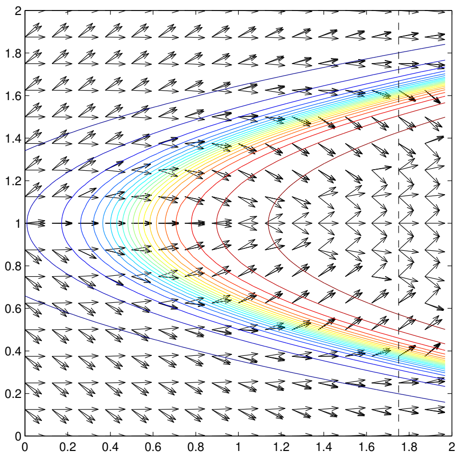
(a) δ-equations, $N = 1$.



(b) δ-equations, $N = 2$.

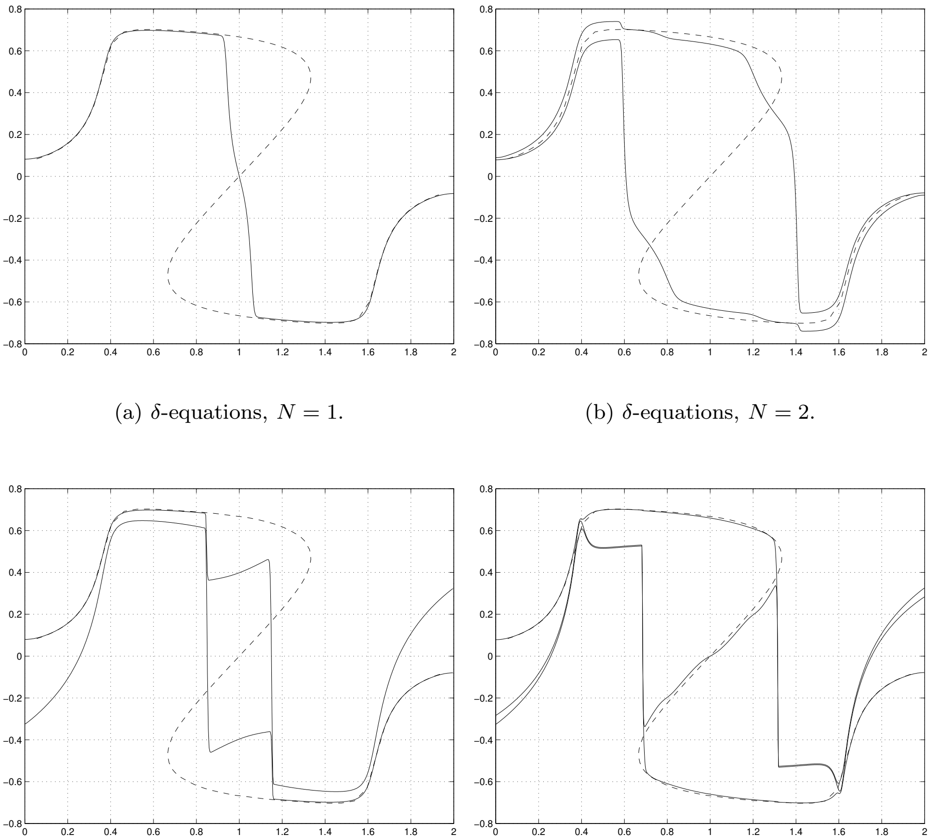


(c) H-equations, $N = 2$.



(d) H-equations, $N = 3$.

Figure 2.23. *Wedge.* Quiver plots of ray angles for delta and Heaviside equations with different N. A contour plot of the index of refraction is overlayed on the solution.



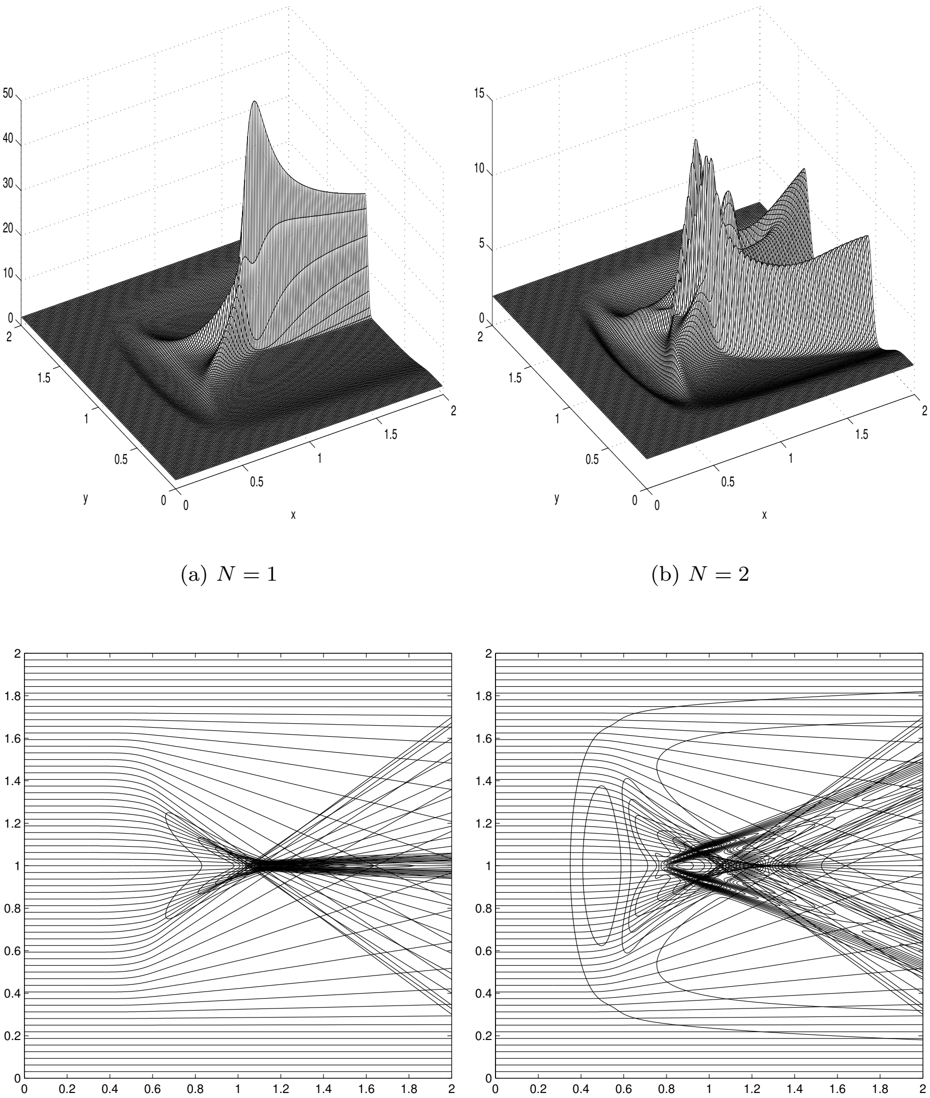
(a) δ-equations, $N = 1$.

(b) δ-equations, $N = 2$.

(c) H-equations, $N = 2$.

(d) H-equations, $N = 3$.

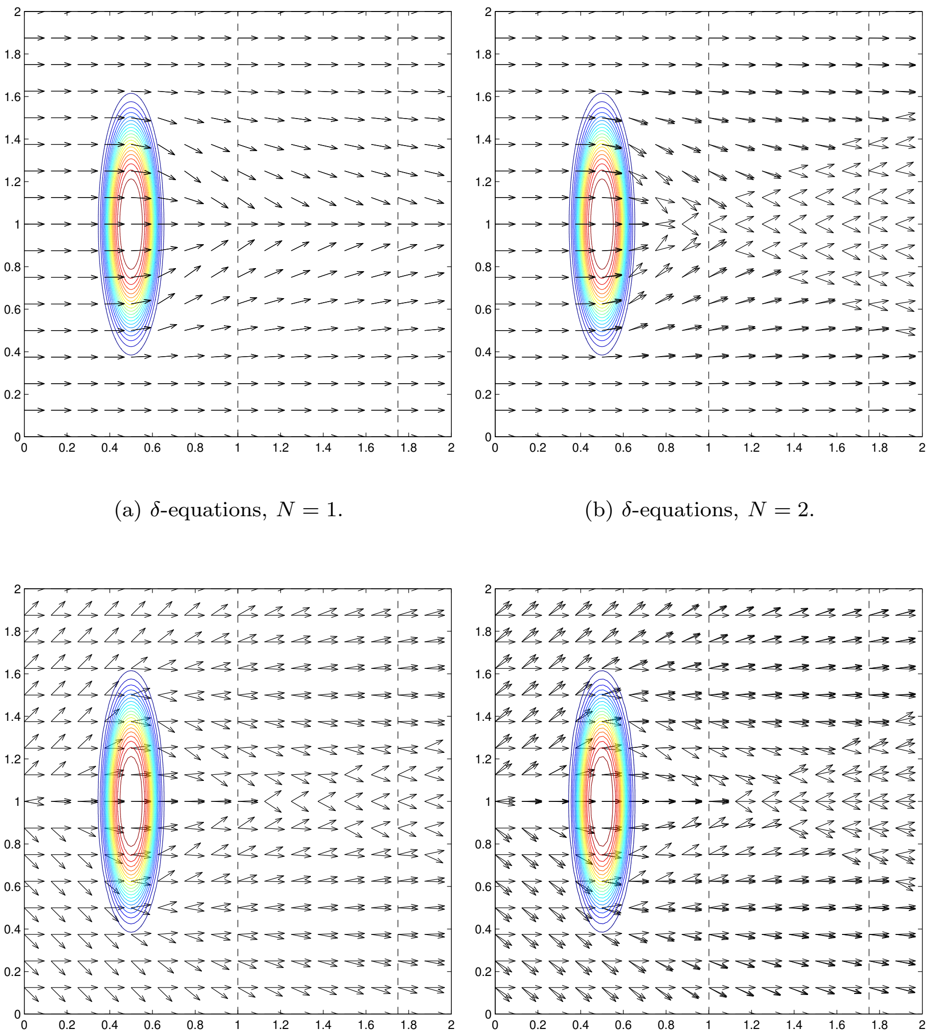
Figure 2.24. *Wedge.* Solution in a cut at $x = 1.75$ for delta and Heaviside equations with different N. Shown is sine of the ray angles (solid) and the corresponding values for a ray traced solution (dashed).



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

(c) $N = 1$ (d) $N = 2$

Figure 2.25. *Convex lens.* Amplitude results for δ-equations with $N = 1, 2$. Top pictures show total ray strength, i.e. g and $g_1 + g_2$ respectively. Bottom two pictures show level curves of the same quantities, with a ray traced solution superimposed.



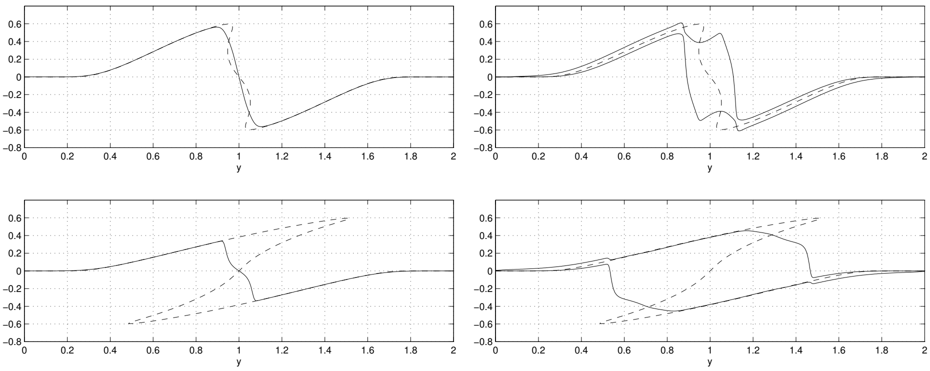
(a) δ-equations, $N = 1$.

(b) δ-equations, $N = 2$.

(c) H-equations, $N = 2$.

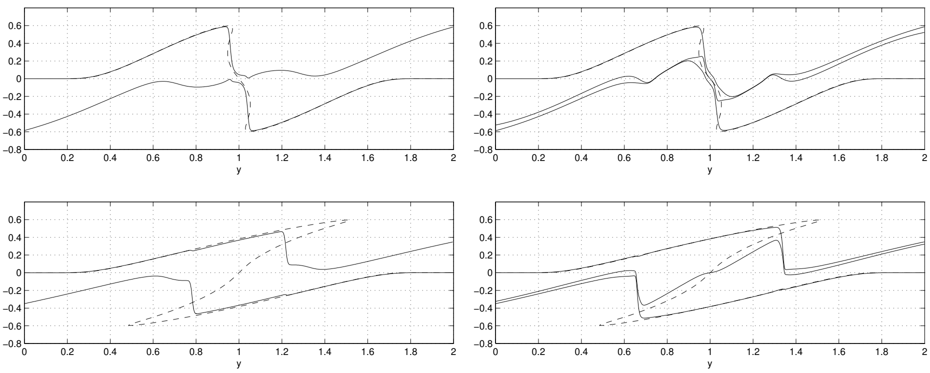
(d) H-equations, $N = 3$.

Figure 2.26. *Convex lens.* Quiver plots of ray angles for delta and Heaviside equations with different N. A contour plot of the index of refraction is overlayed on the solution.



(a) δ-equations, $N = 1$.

(b) δ-equations, $N = 2$.



(c) H-equations, $N = 2$.

(d) H-equations, $N = 3$.

Figure 2.27. *Convex lens.* Solution in a cut at $x = 1$ (above) and $x = 1.75$ (below) for delta and Heaviside equations with different N. Shown is sine of the ray angles (solid) and the corresponding values for a ray traced solution (dashed).

Chapter 3

Wavelet Based Numerical Homogenization

This chapter will describe the following approach to homogenization. Given an equation $Lu = f$ where L is a linear operator, f a right hand side and u a solution that contains fine scales, find the *homogenized* operator $\bar{L}$ and right hand side $\bar{f}$ such that $\bar{L}Pu = \bar{f}$ where P is a projection operator onto a subspace where the fine scales in the original solution does not exist. We will begin this section by defining homogenization in the setting of Hilbert spaces and from this deduce some general properties. We go on to review the theory of multiresolution analysis and then start considering the special case of homogenization in wavelet spaces. Eventually, we let $Lu = f$ be a finite-dimensional approximation of a partial differential equation and P a projection onto a coarse scale wavelet space. The resulting $\bar{L}$ can be approximated by a sparse matrix and used for computing an approximate solution to the differential equation on a coarse grid. We show examples of this for elliptic and hyperbolic problems.

3.1 Theory

3.1.1 Homogenization in Hilbert Spaces

Let V be a Hilbert space and $V_0 \subset V$ a closed subspace. We denote by P the projection in V onto V_0 and let $Q = I - P$, where I is the identity operator in V. We also introduce the unitary operator $\mathcal{W}$ on V defined by

$$\mathcal{W} : V \mapsto V_0 \times V_0^{\perp}, \qquad \mathcal{W}u = \begin{pmatrix} Qu \\ Pu \end{pmatrix}. \tag{3.1}$$

Let $\mathcal{L}(X,Y)$ denote the set of bounded linear maps from X to Y. For an arbitrary operator $L \in \mathcal{L}(V,V)$, we have

$$\begin{aligned} \mathcal{W}L\mathcal{W}^* \begin{pmatrix} u \\ v \end{pmatrix} &= \mathcal{W}L(P+Q)(u+v) = \begin{pmatrix} QL(P+Q)(u+v) \\ PL(P+Q)(u+v) \end{pmatrix} \\ &= \begin{pmatrix} QL(Pv+Qu) \\ PL(Pv+Qu) \end{pmatrix} \equiv \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \end{aligned} \tag{3.2}$$

where

$$\begin{array}{ll} A = QLQ \in \mathcal{L}(V_0^\perp, V_0^\perp), & B = QLP \in \mathcal{L}(V_0, V_0^\perp), \\ C = PLQ \in \mathcal{L}(V_0^\perp, V_0), & D = PLP \in \mathcal{L}(V_0, V_0). \end{array} \tag{3.3}$$

With this decomposition we can now state

Definition 3.1. The *homogenizing* operator, $\mathcal{H}$, with respect to V and V_0 maps an operator in $\mathcal{L}(V,V)$ to its Schur complement with respect to the decomposition in (3.2). Hence,

$$\mathcal{H} : \mathcal{L}(V,V) \mapsto \mathcal{L}(V_0,V_0), \qquad \mathcal{H}L = D - CA^{-1}B. \tag{3.4}$$

To an operator $L \in \mathcal{L}(V,V)$ is associated the linear *right hand side homogenizing* operator, $\mathcal{R}$, with respect to V_0, defined as

$$\mathcal{R} : V \mapsto V_0, \qquad \mathcal{R}u = Pu - CA^{-1}Qu. \tag{3.5}$$

We will write $\mathcal{H}_{V,V_0}$ and $\mathcal{R}^L_{V,V_0}$ when we need to show explicitly between which spaces the homogenizing operators act and to which L that $\mathcal{R}$ is associated. We also note that $\mathcal{H}$ and $\mathcal{R}$ are not defined for all elements of $\mathcal{L}(V,V)$. We will require that A is invertible to consider $\mathcal{H}$ and $\mathcal{R}$ well defined. From Definition 3.1 we immediately have

Proposition 3.2. *Suppose $Lu = f$, where $L \in \mathcal{L}(V,V)$ and $u, f \in V$. If $\mathcal{H}L$ is well defined*

$$(\mathcal{H}L)Pu = \mathcal{R}f. \tag{3.6}$$

Proof. Since $Lu = f$ we get

$$\mathcal{W}L\mathcal{W}^*\mathcal{W}u = \mathcal{W}f \qquad \Rightarrow \qquad \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} Qu \\ Pu \end{pmatrix} = \begin{pmatrix} Qf \\ Pf \end{pmatrix}. \tag{3.7}$$

Moreover, the fact that $\mathcal{H}L$ is well defined implies that A^{-1} exists and we can reduce this system with Gaussian elimination. It yields (3.6). ☐

Some properties that will be kept by the homogenized operator is self-adjointness and positivity, as seen by this proposition.

Proposition 3.3. *Let V be a Hilbert space and V_0 a closed subspace. Furthermore, let $L \in \mathcal{L}(V,V)$ be such that $\mathcal{H}L$ is well defined. If L is self-adjoint $\mathcal{H}L$ is self-adjoint. If also*

$$c_1 ||u||_V^2 \le \langle Lu, u \rangle \le c_2 ||u||_V^2, \qquad \forall u \in V. \tag{3.8}$$

then for the same constants c_1, c_2,

$$c_1||v||_{V_0}^2 \leq \langle \mathcal{H}Lv, v\rangle \leq c_2||v||_{V_0}^2, \qquad \forall v \in v \in V_0 \tag{3.9}$$

if $||v||_{V_0} = ||v||_V$ for $v \in V_0$. If the dimension of the nullspace of B in (3.2) is zero, and if $v \neq 0$ the right inequality in (3.9) is strict.

Proof. Let L be self-adjoint. Then $(\mathcal{H}L)^* = D^* - B^*(A^{-1})^*C^*$. Since L is self-adjoint, $\mathcal{W}L\mathcal{W}^*$ is self-adjoint and therefore $A = A^*$, $D = D^*$ and $C = B^*$. Hence, $(\mathcal{H}L)^* = D - B^*(A^{-1})B = \mathcal{H}L$.

Next, take $v \in V_0$ and set $u = -A^{-1}Bv \in V_0^{\perp}$. We get, since $\mathcal{W}$ is unitary,

$$\begin{aligned} c_1||v||_{V_0} &\leq c_1||v+u||_V \leq \langle L(u+v), (u+v)\rangle = \left\langle \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle \\ &= \left\langle \begin{pmatrix} 0 \\ (-CA^{-1}B + D)v \end{pmatrix}, \begin{pmatrix} -A^{-1}Bv \\ v \end{pmatrix} \right\rangle = \langle \mathcal{H}Lv, v\rangle. \end{aligned} \tag{3.10}$$

This proves the left inequality in (3.9). Since $C = B^*$ and A^{-1} is positive by the positivity of L, we have for the right inequality

$$\begin{aligned} \langle \mathcal{H}Lv, v\rangle &= \langle (D - CA^{-1}B)v, v\rangle = \langle Lv, v\rangle - \langle A^{-1}Bv, Bv\rangle \\ &\leq c_2||v||_V^2 = c_2||v||_{V_0}^2. \end{aligned} \tag{3.11}$$

If the dimension of B's nullspace is zero and $v \neq 0$, the inequality becomes strict. □

We continue by giving two lemmas that will be helpful for showing some further properties of the homogenizing operator.

Lemma 3.4. *Let V, V' be two Hilbert spaces with bases of the same cardinality, and with the closed subspaces V_0, V_0' respectively. Moreover, denote by P, P' the projection operators in V, V' onto the respective subspaces. Let $M : V \mapsto V'$ be an isomorphism and M_0 its restriction to V_0. Then*

$$P'M = MP. \tag{3.12}$$

if and only if M_0 is an isomorphism between V_0 and V_0'.

Proof. First, assume (3.12) holds. Take $v \in V_0$. Then by (3.12), $M_0v = MPv = P'Mv \in V_0'$. On the other hand, for any $v' \in V_0'$, there is a unique $v \in V$ such that $v' = Mv$ and $v = M^{-1}v' = M^{-1}P'MM^{-1}v' = PM^{-1}v' \in V_0$. Hence, we have in fact that $M_0 : V_0 \mapsto V_0'$ is one-to-one and onto. On the other hand, suppose $M_0 : V_0 \mapsto V_0'$ is an isomorphism. Take $u \in V$ and decompose it as $u = u_0 + u_1$ where $u_0 \in V_0$ and $u_1 \in V_0^{\perp}$. Then if $Mu_1 \in V_0'$ we must have $u_1 = M^{-1}Mu_1 = M_0^{-1}Mu_1 \in V_0$, a contradiction, so $Mu_1 \in V_0'^{\perp}$. Then $P'Mu = P'M_0u_0 + P'Mu_1 = M_0u_0 = MPu$. □

Lemma 3.5. *With the same spaces and operators as in Lemma 3.4, let (3.12) be true and suppose that*

$$\bar{L} \in \mathcal{L}(V_0, V_0),\ \bar{L}' \in \mathcal{L}(V_0', V_0') \text{ are invertible,} \tag{3.13}$$
$$R \in \mathcal{L}(V, V_0) \text{ satisfies } Ru = Pu,\ \forall u \in V_0, \tag{3.14}$$
$$R' \in \mathcal{L}(V', V_0') \text{ satisfies } R'v = P'v,\ \forall v \in V_0', \tag{3.15}$$
$$\forall f \in V\ \exists u \text{ such that } \bar{L}u = Rf \text{ and } \bar{L}'M_0 u = R'Mf. \tag{3.16}$$

Then

$$\bar{L} = M_0^{-1}\bar{L}'M_0, \qquad R = M_0^{-1}R'M. \tag{3.17}$$

Proof. First, note that Lemma 3.4 ensures that (3.16) makes sense. Take $f \in V_0$ and u according to (3.16). Then, by Lemma 3.4, $M_0 f \in V_0'$ and in view of (3.14) and (3.15)

$$M_0\bar{L}u = M_0 Rf = M_0 f = R'M_0 f = R'Mf = \bar{L}'M_0 u. \tag{3.18}$$

Therefore, since $\bar{L}, \bar{L}'$ and M_0 are invertible by (3.13) and Lemma 3.4,

$$(M_0\bar{L})^{-1}M_0 f = (\bar{L}'M_0)^{-1}M_0 f, \qquad \forall f \in V_0, \tag{3.19}$$

or

$$(\bar{L}^{-1} - M_0^{-1}(\bar{L}')^{-1}M_0)f = 0, \qquad \forall f \in V_0, \tag{3.20}$$

from which we get $\bar{L} = M_0^{-1}\bar{L}'M_0$. Using this we finally have,

$$M_0 Rf = M_0\bar{L}u = \bar{L}'M_0 u = R'Mf, \qquad \forall f \in V, \tag{3.21}$$

or $R = M_0^{-1}R'M$. □

We can now show that homogenization is "independent of the path".

Proposition 3.6. *Let V be a Hilbert space and V_0, V_1 two closed subspaces nested as $V_0 \subset V_1 \subset V$. Take a surjective operator $L \in \mathcal{L}(V, V)$ such that that $\mathcal{H}_{V_1,V_0}\mathcal{H}_{V,V_1}L$ and $\mathcal{H}_{V,V_0}L$ are both well defined and invertible. Then,*

$$\mathcal{H}_{V_1,V_0}\mathcal{H}_{V,V_1}L = \mathcal{H}_{V,V_0}L, \qquad \mathcal{R}^{\mathcal{H}L}_{V_1,V_0}\mathcal{R}^{L}_{V,V_1} = \mathcal{R}^{L}_{V,V_0}. \tag{3.22}$$

Proof. Let P_j be the projection onto V_j and let L be as in the proposition text. Take $f \in V$. Since L is surjective there is a $u \in V$ such that $Lu = f$. Then, by Proposition 3.2,

$$(\mathcal{H}_{V,V_0}L)P_0 u = \mathcal{R}^{L}_{V,V_0} f \tag{3.23}$$

and,

$$(\mathcal{H}_{V,V_1}L)P_1 u = \mathcal{R}^{L}_{V,V_1} f \qquad \Rightarrow \qquad (\mathcal{H}_{V_1,V_0}\mathcal{H}_{V,V_1}L)P_0 u = \mathcal{R}^{\mathcal{H}L}_{V_1,V_0}\mathcal{R}^{L}_{V,V_1} f \tag{3.24}$$

where we used that $P_0P_1 = P_0$ since $V_0 \subset V_1$. Moreover, for $u \in V_0$, we have

$$\mathcal{R}^{L}_{V,V_0} u = P_0 u - CA^{-1}(I - P_0)u = P_0 u \tag{3.25}$$

and, similarly,

$$\mathcal{R}^{\mathcal{H}L}_{V_1,V_0}\mathcal{R}^{L}_{V,V_1}u = \mathcal{R}^{\mathcal{H}L}_{V_1,V_0}(P_1 u) = P_0 u - CA^{-1}(I-P_0)P_1 u = P_0 u. \tag{3.26}$$

The result now follows from Lemma 3.5 with $V = V'$, $M = I$, $\bar{L} = \mathcal{H}_{V,V_0}L$, $\bar{L}' = \mathcal{H}_{V_1,V_0}\mathcal{H}_{V,V_1}L$, $R = \mathcal{R}^{L}_{V,V_0}$ and $R' = \mathcal{R}^{\mathcal{H}L}_{V_1,V_0}\mathcal{R}^{L}_{V,V_1}$. □

Finally, we show that an operator that can be split "dimensionwise" is also homogenized "dimensionwise".

Proposition 3.7. *Suppose $\boldsymbol{V} = V \otimes U$. Let $L = L_v \otimes L_u$ be an invertible operator on $\mathcal{L}(\boldsymbol{V},\boldsymbol{V})$ where $L_v \in \mathcal{L}(V,V)$ and $L_u \in \mathcal{L}(U,U)$. Let V_0, U_0 be closed subspaces of U, V respectively and $\boldsymbol{V}_0 = V_0 \otimes U_0$. If $\mathcal{H}_{\boldsymbol{V},\boldsymbol{V}_0}(L_v \otimes L_u)$, $\mathcal{H}_{V,V_0}L_v$ and $\mathcal{H}_{U,U_0}L_u$ are well defined and invertible,*

$$\mathcal{H}_{\boldsymbol{V},\boldsymbol{V}_0}(L_v \otimes L_u) = \Big(\mathcal{H}_{V,V_0}L_v\Big)\otimes\Big(\mathcal{H}_{U,U_0}L_u\Big). \tag{3.27}$$

$$\mathcal{R}^{L_v\otimes L_u}_{\boldsymbol{V},\boldsymbol{V}_0} = \mathcal{R}^{L_v}_{V,V_0}\otimes\mathcal{R}^{L_u}_{U,U_0}. \tag{3.28}$$

Proof. We subindex the operators introduced in (3.1) and (3.3) by v and u according to which one of the spaces V and U they are associated. All three of $V_0 \otimes U$, $V \otimes U_0$ and $V_0 \otimes U_0$ are closed subspaces of $\boldsymbol{V}$ by the definition of a tensor product space. We note that $\mathcal{W}_v \otimes I_u : \boldsymbol{V} \mapsto V_0 \otimes U \times (V_0 \otimes U)^{\perp}$ and

$$(\mathcal{W}_v \otimes I_u)(L_v \otimes L_u)(\mathcal{W}_v^T \otimes I_u^T) = \begin{pmatrix} A_v \otimes L_u & B_v \otimes L_u \\ C_v \otimes L_u & D_v \otimes L_u \end{pmatrix}. \tag{3.29}$$

Consequently,

$$\begin{aligned}
\mathcal{H}_{\boldsymbol{V},V_0\otimes U}(L_v\otimes L_u) &= D_v\otimes L_u - (C_v\otimes L_u)(A_v\otimes L_u)^{-1}(B_v\otimes L_u)\\
&= D_v\otimes L_u - (C_v\otimes L_u)(A_v^{-1}\otimes L_u^{-1})(B_v\otimes L_u)\\
&= (D_v - C_vA_v^{-1}B_v)\otimes L_u\\
&= \Big(\mathcal{H}_{V,V_0}L_v\Big)\otimes L_u. \qquad (3.30)
\end{aligned}$$

Similarly, if $\tilde{L}_v$ is an operator in $\mathcal{L}(V_0, V_0)$

$$\mathcal{H}_{V_0\otimes U,\boldsymbol{V}_0}(\tilde{L}_v\otimes L_u) = \tilde{L}_v\otimes(\mathcal{H}_{U,U_0}L_u). \tag{3.31}$$

From (3.30) and (3.31) it then follows that

$$\begin{aligned}
\mathcal{H}_{V_0\otimes U,\boldsymbol{V}_0}\mathcal{H}_{\boldsymbol{V},V_0\otimes U}(L_v\otimes L_u) &= \mathcal{H}_{V_0\otimes U,\boldsymbol{V}_0}((\mathcal{H}_{V,V_0}L_v)\otimes L_u)\\
&= (\mathcal{H}_{V,V_0}L_v)\otimes(\mathcal{H}_{U,U_0}L_u). \qquad (3.32)
\end{aligned}$$

For the right hand side homogenization operator, take $f_v \in V$ and $f_u \in U$. Then

$$\begin{aligned}
\mathcal{R}^{L_v\otimes L_u}_{\boldsymbol{V},V_0\otimes U}(f_v\otimes f_u) &= (P_{V_0\otimes U} - (C_v\otimes L_u)(A_v\otimes L_u)^{-1}P_{V_0^{\perp}\otimes U})(f_v\otimes f_u)\\
&= \Big((P_{V_0} - C_vA_v^{-1}P_{V_0^{\perp}})\otimes I_u\Big)(f_v\otimes f_u)\\
&= (\mathcal{R}^{L_v}_{V,V_0}f_v)\otimes f_u. \qquad (3.33)
\end{aligned}$$

And, with $\tilde{f}_v \in V$,

$$\mathcal{R}^{\tilde{L}_v \otimes L_u}_{V_0 \otimes U, \boldsymbol{V}_0}(\tilde{f}_v \otimes f_u) = \tilde{f}_v \otimes (\mathcal{R}^{L_u}_{U,U_0} f_u). \tag{3.34}$$

Using (3.30) together with (3.33) and (3.34) gives

$$\begin{aligned} \mathcal{R}^{\mathcal{H}L_v \otimes L_u}_{V_0 \otimes U, \boldsymbol{V}_0} \mathcal{R}^{L}_{\boldsymbol{V}, V_0 \otimes U}(f_v \otimes f_u) &= (\mathcal{R}^{L_v}_{V,V_0} f_v) \otimes (\mathcal{R}^{L_u}_{U,U_0} f_u) \\ &= (\mathcal{R}^{L_v}_{V,V_0} \otimes \mathcal{R}^{L_u}_{U,U_0})(f_v \otimes f_u). \end{aligned} \tag{3.35}$$

This is true for all product vectors in $\boldsymbol{V}$. Since linear combinations of product vectors are dense in $\boldsymbol{V}$ and $\mathcal{R}$ is a bounded linear operator, (3.35) extends to all of $\boldsymbol{V}$.

We can now use Proposition 3.6 and (3.32, 3.35) to deduce the result. We just have to notice that by (3.32) $\mathcal{H}_{V_0 \otimes U, \boldsymbol{V}_0} \mathcal{H}_{\boldsymbol{V}, V_0 \otimes U}(L_v \otimes L_u)$ is well defined and invertible if $\mathcal{H}_{V,V_0} L_v$ and $\mathcal{H}_{U,U_0} L_u$ are. $\square$

3.1.2 Multiresolution Analysis

In this section we will give a brief outline of the homogenization spaces that we will use when we develop subgrid scale models. In general we want to homogenize to a subspace that represents the coarse scales of a function space. Therefore, it is close at hand to consider *multiresolution analysis* (MRA) and its related wavelet spaces. A wavelet representation lends itself naturally to analyzing the fine and coarse scales as well as the localization properties of a function. For a detailed description of MRA we refer the reader e.g. to the book by Daubechies, [18].

An MRA is made up of a ladder of closed function spaces,

$$\ldots \subset V_j \subset V_{j+1} \subset \ldots, \tag{3.36}$$

satisfying

$$\overline{\bigcup_{j \in \mathbb{Z}} V_j} = L_2(\mathbb{R}), \qquad \bigcap_{j \in \mathbb{Z}} V_j = \{0\}. \tag{3.37}$$

Moreover, the spaces V_j represent different scales in the sense that

$$f(x) \in V_j \Leftrightarrow f(2^{-j} x) \in V_0. \tag{3.38}$$

The zero space should also be invariant under integer translations,

$$f(x) \in V_0 \Rightarrow f(x-k) \in V_0, \qquad \forall k \in \mathbb{Z}, \tag{3.39}$$

and there should exist a *shape function*, $\varphi \in V_0$, such that

$$\{\varphi(x-k) : \ k \in \mathbb{Z}\}, \tag{3.40}$$

is an L_2-orthonormal basis for V_0.

From the definition above follows a number of important properties. Dilates and translates of the shape function span the V_j spaces,

$$V_j = \operatorname{span}\{\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)\}, \qquad j, k \in \mathbb{Z}. \tag{3.41}$$

Denote the orthogonal complement of V_j in V_{j+1} by W_j,

$$V_{j+1} = V_j \oplus W_j, \qquad V_j \perp W_j. \tag{3.42}$$

Then, $L_2(\mathbb{R}) = \overline{\bigoplus_{j\in\mathbb{Z}} W_j}$ and there exists a *mother wavelet*, $\psi \in W_0$, with properties corresponding to that of φ,

$$W_j = \operatorname{span}\{\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)\}, \qquad j, k \in \mathbb{Z}. \tag{3.43}$$

This mother wavelet can be constructed explicitly from a given MRA. The functions $\varphi_{j,k}$ form an L_2-orthonormal basis in V_j and so does the functions $\psi_{j,k}$ in W_j.

The simplest MRA is the one built on the Haar basis, where the shape function and the mother wavelet are given by

$$\varphi(x) = \begin{cases} 1, & \text{if } 0 \le x \le 1, \\ 0, & \text{otherwise,} \end{cases} \qquad \psi(x) = \begin{cases} 1, & \text{if } 0 \le x \le 1/2, \\ -1, & \text{if } 1/2 < x \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.44}$$

This is the basis we will mainly use in the thesis.

In more than one dimension the MRA is extended by considering tensor product spaces. Each dimension can be treated separately, using the one-dimensional analysis. The resulting multi-dimensional MRA is a tensor product of one-dimensional MRAs. In two dimensions we define the tensor product spaces

$$\boldsymbol{V}_j = V_j \otimes V_j, \qquad \ldots \subset \boldsymbol{V}_j \subset \boldsymbol{V}_{j+1} \subset \ldots, \qquad j \in \mathbb{Z}, \tag{3.45}$$

with V_j being the one-dimensional spaces introduced above. The spaces $\boldsymbol{V}_j$ will satisfy the same conditions, (3.36)-(3.40), as the V_j spaces, with $\mathbb{R}, \mathbb{Z}$ replaced by $\mathbb{R}^2, \mathbb{Z}^2$. The shape function of this two-dimensional MRA can be written in terms of the one-dimensional shape function,

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \varphi(x_1) \otimes \varphi(x_2) = \varphi(x_1)\varphi(x_2). \tag{3.46}$$

It follows, like in the one-dimensional case, that

$$\{\boldsymbol{\varphi}_{j,k,l}(\boldsymbol{x}) = \varphi_{j,k}(x_1) \otimes \varphi_{j,l}(x_2)\}, \qquad k, l \in \mathbb{Z}, \tag{3.47}$$

is an orthonormal basis for $\boldsymbol{V}_j$. Furthermore, denoting the orthogonal complement of $\boldsymbol{V}_j$ in $\boldsymbol{V}_{j+1}$ by $\boldsymbol{W}_j$, we have that

$$\boldsymbol{V}_{j+1} = \boldsymbol{V}_j \oplus \boldsymbol{W}_j, \qquad \boldsymbol{V}_j \perp \boldsymbol{W}_j, \qquad L_2(\mathbb{R}^2) = \overline{\bigoplus_{j\in\mathbb{Z}} \boldsymbol{W}_j}. \tag{3.48}$$

The wavelet space $\boldsymbol{W}_j$ is composed of three parts,

$$\boldsymbol{W}_j = (W_j \otimes W_j) \oplus (V_j \otimes W_j) \oplus (W_j \otimes V_j). \tag{3.49}$$

Here, the W_j spaces are those of the one-dimensional case, given by (3.42). Similar tensor product extensions can be made also for higher dimensions.

3.1.3 Wavelet Projection of Difference Operators

We will now discuss homogenization in a more concrete setting. The spaces we use are finite dimensional wavelet spaces introduced in Section 3.1.2 and the typical operators that we homogenize are discrete approximations of differential operators. Given such an operator, defined on a fine grid, the underlying idea of the homogenization approach is to find an operator defined on a lower dimensional space that extracts only the coarse scales of the solution. For a function in the wavelet space V_{j+1}, the coarse scale is represented by V_j, and we are thus in the first step interested in the homogenized operator given by $\mathcal{H}_{V_{j+1},V_j}$. We will mostly use the standard Haar basis. In the examples of this chapter, numerical tests have not shown any substantial advantage with using more sophisticated wavelets.

We should also remark that of course other than wavelet spaces could be used. For instance Knapek, [35], uses grid function spaces defined on a nested sequence of grids in a multigrid fashion. In this setting the coarse subspace is simply the grid function space given by removing every other grid point. There are however drawbacks with this choice, see Remark 3.8.

Let us now consider the equation

$$L_{j+1}U = F, \qquad U, F \in V_{j+1}, \qquad L_{j+1} \in \mathcal{L}(V_{j+1}, V_{j+1}). \tag{3.50}$$

This equation may originate from a finite difference, finite element or finite volume discretization of a given differential equation. We identify U as a piecewise constant approximation of $u(x)$, the solution to the continuous problem. Let $\mathcal{W}_j : V_{j+1} \mapsto V_j \times W_j$ be the unitary operator corresponding to (3.1). (By (3.42) $V_j^{\perp} = W_j$.) As in the infinite dimensional case (3.7), we apply the transformation $\mathcal{W}_j$ on L_{j+1} from the left and get

$$\mathcal{W}_j L_{j+1} \mathcal{W}_j^{\top} (\mathcal{W}_j U) = \mathcal{W}_j F \tag{3.51}$$

or if we decompose $\mathcal{W}_j L_{j+1} \mathcal{W}_j^{\top}$ according to (3.2),

$$\begin{pmatrix} A_j & B_j \\ C_j & D_j \end{pmatrix} \begin{pmatrix} U_h \\ U_l \end{pmatrix} = \begin{pmatrix} F_h \\ F_l \end{pmatrix}, \qquad U_h, F_h \in W_j, \qquad U_l, F_l \in V_j. \tag{3.52}$$

Here subindex h means projection onto the fine scale subspace, W_j in this case, and subindex l stands for projection onto the coarse scale subspace, V_j here. This is a convention we will use frequently.

The homogenized "coarse grid operator" is now given by block Gaussian elimination of (3.52),

$$(D_j - C_j A_j^{-1} B_j) U_l = F_l - C_j A_j^{-1} F_h. \tag{3.53}$$

For simplicity we will use the notation $\mathcal{H}_{V_{j+1},V_j} L_{j+1} \equiv \bar{L}_j$ and $\mathcal{R}_{V_{j+1},V_j} F \equiv \bar{F}_j$. Hence,

$$\bar{L}_j = D_j - C_j A_j^{-1} B_j, \qquad \bar{F}_j = F_l - C_j A_j^{-1} F_h. \tag{3.54}$$

We can interpret $C_j A_j^{-1} B_j$ as a correction term, which includes subgrid phenomena in $\bar{L}_j$. Had it been zero, $\bar{L}_j$ would have been just a projection of L_{j+1} onto $\mathcal{L}(V_j, V_j)$.

In the finite dimensional case we will represent the operators using their matrix representations with respect to the bases given by (3.41) and (3.43). For instance in the Haar basis we get the following matrix representation of $\mathcal{W}_j$,

$$\mathcal{W}_j = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 & \cdots & & \\ 0 & 0 & 1 & -1 & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \\ 0 & 0 & \cdots & 0 & 1 & -1 \\ 1 & 1 & 0 & \cdots & & \\ 0 & 0 & 1 & 1 & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \\ 0 & 0 & \cdots & 0 & 1 & 1 \end{pmatrix}. \tag{3.55}$$

The matrix representation of (3.52) is a block matrix decomposition of the same form as the operator decomposition.

Before we go on we should also note that in general $\bar{L}_j$ will not be represented by a sparse matrix even if L_{j+1} is. For the homogenization method to be efficient we must be able to approximate $\bar{L}_j$ with a sparse matrix. This sparse matrix can be seen as a discretization of a local differential operator acting on the coarse space. In Section 3.1.5 we will get back to this issue and in subsequent sections we will show that this can indeed be done for many applications.

The procedure to obtain $\bar{L}_j$ described above can be applied recursively on $\bar{L}_j$ itself to get $\bar{L}_{j-1}$ and so on. This can easily be verified when L_{j+1} is symmetric positive definite. Then A_j is always invertible and Proposition 3.3 shows that also $\bar{L}_j$ is symmetric positive definite. It can thus be homogenized. Moreover, for the first step an improvement in the condition number can often be estimated from (cf. (3.11))

$$\langle \bar{L}_j v, v\rangle = \left\langle (D_j - C_j A_j^{-1} B_j) v, v \right\rangle \leq \langle D_j v, v\rangle \tag{3.56}$$

and Proposition 3.3. Typically for standard discretizations $\max_{||v||=1}\langle D_j v, v\rangle < \max_{||u||=1}\langle L_{j+1} u, u\rangle$. Also note that Proposition 3.6 gives $\bar{L}_k = \mathcal{H}_{V_{j+1},V_k} L_{j+1}$ and $\bar{F}_k = \mathcal{R}_{V_{j+1},V_k} F$ for $k \leq j$.

The two-dimensional wavelet transform

$$\boldsymbol{\mathcal{W}}_j : \boldsymbol{V}_{j+1} \mapsto \boldsymbol{W}_j \times \boldsymbol{V}_j \tag{3.57}$$

can be written as a tensor product of one-dimensional transforms,

$$\boldsymbol{\mathcal{W}}_j = \mathcal{W}_j \otimes \mathcal{W}_j. \tag{3.58}$$

A linear operator L_{j+1} that acts on the space $\boldsymbol{V}_{j+1}$ can be decomposed in a way similar to the one-dimensional case. The equation

$$L_{j+1}U = F, \qquad U, F \in \boldsymbol{V}_{j+1}, \qquad L_{j+1} \in \mathcal{L}(\boldsymbol{V}_{j+1}, \boldsymbol{V}_{j+1}) \tag{3.59}$$

can then be transformed to

$$\begin{pmatrix} A_j & B_j \\ C_j & D_j \end{pmatrix}\begin{pmatrix} U_h \\ U_l \end{pmatrix} = \begin{pmatrix} F_h \\ F_l \end{pmatrix}, \qquad U_h, F_h \in \boldsymbol{W}_j, \qquad U_l, F_l \in \boldsymbol{V}_j, \tag{3.60}$$

and the coarse grid operator is again the Schur complement,

$$\mathcal{H}_{\boldsymbol{V}_{j+1},\boldsymbol{V}_j} L \equiv \bar{L}_j = D_j - C_j A_j^{-1} B_j. \tag{3.61}$$

To get the matrix representation of (3.60), we take the one-dimensional matrix representations of $\mathcal{W}_j$ and, following (3.58), compute its Kronecker tensor product with itself. We must thereafter also apply a suitable permutation matrix to the product in order to get the same block structure of the matrix as in (3.60).

Note that the coarse part of U can be decomposed as

$$U_h = \begin{pmatrix} U_{hh} \\ U_{lh} \\ U_{hl} \end{pmatrix}, \qquad U_{hh} \in W_j \otimes W_j, \qquad U_{lh} \in V_j \otimes W_j, \qquad U_{hl} \in W_j \otimes V_j. \tag{3.62}$$

In some important cases the homogenized operator has proven to remain in the same form as the original operator. Let the forward and backward undivided differences be defined as

$$\Delta_+ u_i = u_{i+1} - u_i, \qquad \Delta_- u_i = u_i - u_{i-1}. \tag{3.63}$$

In [19] it was shown that the one-dimensional elliptic model equation $-(au')' = f$ discretized with

$$-\Delta_+ \mathrm{diag}(a) \Delta_- U = h^2 F \tag{3.64}$$

will preserve its divergence form during homogenization. That is, we will get

$$\bar{L}_j = \frac{1}{h^2} \Delta_+ H_j \Delta_-, \tag{3.65}$$

where the subgrid model H_j is a strongly diagonal dominant matrix. Analogously, for the first order differential operator $a(x)\frac{\partial}{\partial x}$ the discretized form $\mathrm{diag}(a)\Delta_-/h$ is preserved during homogenization,

$$\bar{L}_j = \frac{1}{h} H_j \Delta_-. \tag{3.66}$$

In two dimensions, the elliptic model equation $-\nabla(a(x,y)\nabla u) = f$ can be discretized as

$$L_{j+1} = -\frac{1}{h^2}\Delta_+^x A \Delta_-^x - \frac{1}{h^2}\Delta_+^y A \Delta_-^y, \qquad L_{j+1}U = F. \tag{3.67}$$

Then we can no longer find a subgrid model H of exactly the same form. The cross-derivatives must also be included. We get

$$\bar{L}_j = -\frac{1}{h^2}\Delta_+^x H^{xx}\Delta_-^x - \frac{1}{h^2}\Delta_+^y H^{yx}\Delta_-^x - \frac{1}{h^2}\Delta_+^x H^{xy}\Delta_-^y - \frac{1}{h^2}\Delta_+^y H^{yy}\Delta_-^y. \tag{3.68}$$

3.1.4 Comparison with Homogenization

Homogenization is a classical analytical way to approximate the effect of some classes of periodic or stochastic subcell scales. The problem is often formulated as follows. Consider a set of operators L_ϵ, indexed by the small parameter ϵ, and a right hand side f. Find the *homogenized operator* $\bar{L}$ defined by

$$L_\epsilon u_\epsilon = f, \qquad \lim_{\epsilon\to 0} u_\epsilon = \bar{u}, \qquad \bar{L}\bar{u} = f. \tag{3.69}$$

In certain cases the convergence above and existence of the homogenized operator can be proved, [6].

For simple model problems, with coefficients periodic on the fine scale, exact closed form solutions can be obtained. A one-dimensional case is

$$L_\epsilon = -\frac{\partial}{\partial x}\left(a(x/\epsilon)\frac{\partial}{\partial x}\right), \tag{3.70}$$

where the variable coefficient $a(y)$ is positive, one-periodic and bounded away from zero. In this case the homogenized operator is

$$\bar{L} = -\bar{a}\frac{\partial^2}{\partial x^2}, \qquad \bar{a} = \left(\int_0^1 \frac{dy}{a(y)}\right)^{-1}. \tag{3.71}$$

We see that the variable coefficient operator is replaced by a constant coefficient operator, the constant being the harmonic average of $a(x)$.

Similarly, in the case of the time-dependent one-dimensional hyperbolic model problem,

$$L_\epsilon = \frac{\partial}{\partial t} + a(x/\epsilon)\frac{\partial}{\partial x}, \tag{3.72}$$

the solution is

$$\bar{L} = \frac{\partial}{\partial t} + \bar{a}\frac{\partial}{\partial x}. \tag{3.73}$$

Hence, in the limit only the spatial part of the original operator is affected by the homogenization, and that part is homogenized in exactly the same way as in the elliptic case, (3.70). The model examples (3.70, 3.71) and (3.72, 3.73) are used in Sections 3.2 and 3.3 respectively.

In higher dimensions, the solution to (3.69) is a little more complicated, although for model problems the structure of the homogenized operator can still be written down, as long as the coefficients are periodic. In the d-dimensional elliptic case, let $A(y) \in \mathbb{R}^{d\times d}$ be one-periodic in each of its arguments and let I_d denote the unit square. It can then be shown, [6], that for

$$L_\epsilon = -\nabla \cdot (A\left(\frac{x}{\epsilon}\right)\nabla) \tag{3.74}$$

the homogenized operator is given by

$$\bar{L} = -\nabla \cdot (H\nabla), \qquad H = \int_{I_d} A(y) - A(y)\frac{\partial \chi(y)}{\partial y}dy, \tag{3.75}$$

where $\partial\chi/\partial y$ is the jacobian of the vector valued function $\chi(y) \in \mathbb{R}^d$, whose components χ_k are given by solving the so called cell problems

$$\nabla \cdot (A(y)\nabla\chi_k) = \sum_{i=1}^{d} \frac{\partial}{\partial y_i} a_{ik}(y), \qquad k = 1, \dots , d, \tag{3.76}$$

with periodic boundary conditions for χ_k. Note that H is a constant matrix. See [20] for a practical application of this.

There is a striking relationship between the analytically homogenized operator in (3.75) and the the Schur complement $\bar{L}_j$ in (3.54), repeated here for convenience,

$$\bar{L}_j = D_j - C_j A_j^{-1} B_j. \tag{3.77}$$

The first terms in (3.75) and (3.77) both represent averaged operators, L_j in a discrete sense and

$$-\nabla \cdot \left(\int_{I_d} A(y)dy\right)\nabla, \tag{3.78}$$

in an integral sense. In both formulations a correction term is subtracted from the average. Furthermore, in the correction term, χ is the solution of an elliptic equation and A_j^{-1} is a discrete positive definite solution operator.

Remark 3.8. In the classical homogenization setting (3.74) homogenization gives the asymptotic expansion

$$u_\epsilon(x) = \bar{u}(x) + \epsilon u_1(x, x/\epsilon) + \mathcal{O}(\epsilon^2) \tag{3.79}$$

for the solution, see [6]. The techniques described in [35], [50] and [31] give homogenized operators with coarse grid solutions which directly samples (3.79). The oscillatory term $\epsilon u_1(x, x/\epsilon)$ is contained in the solution. In the wavelet homogenization the solution is a projection of u_ϵ onto a coarse scale V_j space. The influence of the u_1 term, which oscillates with mean zero [6], is then significantly reduced.

3.1.5 Compact Representation of Projected Operators

When the operator L_{j+1} is derived from a finite difference, finite element or finite volume discretization, it is sparse and of a certain structure. In one dimension it might for instance be tridiagonal. However, as remarked in Section 3.1.3, the matrix $\bar{L}_j$ is not sparse since A_j^{-1} is usually dense. Computing all components of $\bar{L}_j$ would be inefficient. Fortunately, $\bar{L}_j$ will be diagonal dominant in many important cases. We should then be able to find a sparse matrix that is a close approximation of $\bar{L}_j$.

One simple way to approximate $\bar{L}_j$ is to set all components outside a prescribed bandwidth ν equal to zero. Let us define

$$\operatorname{trunc}(M,\nu)_{ij} = \begin{cases} M_{ij}, & \text{if } 2\,|i-j| \le \nu - 1 \\ 0, & \text{otherwise.} \end{cases} \tag{3.80}$$

For $\nu = 1$ we will get a diagonal matrix, for $\nu = 3$ we will get a tridiagonal matrix and so on. This procedure was introduced in [8] and used in [19]. We will use it in this paper in order to control the structure of $\bar{L}_j$ and we refer to it as *truncation*. A theoretical motivation for this form of approximation is given in [19]. For a class of elliptic problems it is proved that the matrix elements of $\bar{L}_j$ decay exponentially away from the diagonal. For lower triangular matrices we define truncation to lower triangular form as

$$\operatorname{trunc}_{\mathrm{L}}(M,\nu)_{ij} = \begin{cases} M_{ij}, & \text{if } 0 \le i-j \le \nu - 1 \\ 0, & \text{otherwise.} \end{cases} \tag{3.81}$$

Another possible strategy is the basic truncation method used in [8],

$$\operatorname{trunc}_{\mathrm{BCR}}(M,\epsilon)_{ij} = \begin{cases} M_{ij}, & \text{if } |M_{ij}| > \epsilon, \\ 0, & \text{otherwise} \end{cases} \tag{3.82}$$

but it is not practical here since the location of the non-zero elements cannot be controlled. We propose that $\bar{L}_j$ be projected onto banded form in a more effective manner. The aim is that the projected matrix should give the same result as the original matrix on a given subspace, e.g. when applied to vectors representing smooth functions. Let $\{\boldsymbol{v}_j\}_{j=1}^{\nu}$ be a set of linearly independent vectors in $\mathbb{R}^N$. Denote the subspace of $\mathbb{R}^{N\times N}$ with matrices essentially[1] of bandwidth ν by $\mathcal{T}_\nu$. Moreover, let

$$\mathcal{L}_\nu = \{M \in \mathbb{R}^{N\times N} :\ \operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_\nu\} \subset \mathsf{N}(M)\}, \tag{3.83}$$

where $\mathsf{N}(M)$ represents the null space of M. Then

$$\mathbb{R}^{N\times N} = \mathcal{T}_\nu \oplus \mathcal{L}_\nu \tag{3.84}$$

[1] We must require that each row of the matrices in $\mathcal{T}_\nu$ has the same number of elements. Therefore, the first and last $\nu - 1$ rows will have additional elements located immediately to the right and left of the band, respectively.

and we define the *band projection* of a matrix $M \in \mathbb{R}^{N\times N}$ as the projection of M onto $\mathcal{T}_\nu$ along $\mathcal{L}_\nu$, with the notation

$$\mathrm{band}(M,\nu) = \mathrm{Proj}_{\mathcal{T}_\nu} M. \tag{3.85}$$

As a consequence,

$$M\boldsymbol{x} = \mathrm{band}(M,\nu)\boldsymbol{x}, \qquad \forall \boldsymbol{x} \in \mathrm{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_\nu\}. \tag{3.86}$$

In our setting M will usually operate on vectors representing smooth functions, for instance solutions to elliptic equations, and a natural choice for $\boldsymbol{v}_j$ vectors are thus the first ν polynomials,

$$\boldsymbol{v}_j = \{1^{j-1}, 2^{j-1}, \dots, N^{j-1}\}^T, \qquad j = 1, \dots, \nu. \tag{3.87}$$

Smooth solutions to the homogenized problem should be well approximated by these vectors. For the case $\nu = 1$ we should remark that we get the standard "masslumping" of a matrix, often used in the context of finite element methods.

This technique is similar to the probing technique used by Chan et al., [16]. In that case the vectors $\boldsymbol{v}_j$ are sums of unit vectors. Other probing techniques have been suggested by Axelsson, Pohlman and Wittum, see Chapter 8 in [2]. The choice of $\boldsymbol{v}_j$ vectors could be optimized if there is some a priori knowledge of the homogenized solution. In some cases the band projection technique only gives improvements for small values of ν, see Figure 3.1. The probing technique of [16] could be used for larger ν in our examples. Numerical evidence indicate that for small values of ν, the band projection technique is more efficient.

The focus in this paper is on the principle and the truncation $\mathrm{trunc}(\bar{L}_j, \nu)$ is done after computing the full inverse of A_j. In large scale applications, to be reported elsewhere, incomplete or thresholded LU-decomposition will be used. The computation of $\mathrm{band}(\bar{L}_j, \nu)$ can be based on $\mathrm{trunc}(\bar{L}_j, \mu)$, $\mu > \nu$. The additional computational cost is proportional to $(\nu^3 + \mu\nu)N$. The $\nu^3 N$ term corresponds to solving N $\nu \times \nu$ systems and $\mu\nu N$ to computing the right hand sides, see also [2].

The truncation methods described above are even more efficient when applied to the subgrid model H_j instead of directly to the homogenized operator. The following proposition shows that when the solution to the homogenized problem belongs to the Sobolev space H_1, the accuracy is one order higher.

Proposition 3.9. *Suppose $L = \Delta_+ H \Delta_-$ and $LU = h^2 f$. Let $\kappa = ||L^{-1}|| \cdot ||H||$ and*

$$r = \frac{||\delta L||}{||L||} = \frac{||\delta H||}{||H||} < \frac{1}{4\kappa}, \tag{3.88}$$

where $||\cdot||$ denotes the l_2-norm. Furthermore, let u be any H_1-function such that $U_j = u(jh)$. Then, if

$$(L + \delta L)(U + \delta U_L) = h^2 f, \tag{3.89}$$
$$\Delta_+(H + \delta H)\Delta_-(U + \delta U_H) = h^2 f \tag{3.90}$$

the following bounds on the perturbed solutions hold $\forall f$

$$||\delta U_L|| \leq \frac{4r\kappa}{1-4r\kappa}||U||, \tag{3.91}$$

$$||\delta U_H|| \leq \frac{1}{2}\cdot\frac{4r\kappa}{1-4r\kappa}||\Delta_- U|| \leq \frac{h}{2}\cdot\frac{4r\kappa}{1-4r\kappa}||u||_{H_1}. \tag{3.92}$$

The bound (3.91) is sharp.

Proof. Using $||\Delta_+|| = ||\Delta_-|| = 2$ and (3.88) we have

$$||L^{-1}\delta L|| \leq \frac{||L^{-1}||\cdot||\delta H||\cdot||\Delta_+ H\Delta_-||}{||H||} \leq 4r||L^{-1}||\cdot||H|| = 4r\kappa < 1. \tag{3.93}$$

Therefore,

$$||\delta U_L|| = ||(I + L^{-1}\delta L)^{-1}L^{-1}\delta LU|| \leq \frac{||L^{-1}\delta L||}{1-||L^{-1}\delta L||}||U|| \leq \frac{4r\kappa}{1-4r\kappa}||U||, \tag{3.94}$$

proving the first bound. In the second case,

$$||L^{-1}\Delta_+\delta H\Delta_-|| \leq 4||L^{-1}||\cdot||\delta H|| \leq 4r\kappa < 1, \tag{3.95}$$

giving

$$\begin{aligned} ||\delta u_H|| &= ||(I + L^{-1}\Delta_+\delta H\Delta_-)^{-1}L^{-1}\Delta_+\delta H\Delta_- U|| \\ &\leq \frac{||L^{-1}\Delta_+\delta H||}{1-||L^{-1}\Delta_+\delta H\Delta_-||}||\Delta_- U|| \leq \frac{2r\kappa}{1-4r\kappa}||\Delta_- U|| \end{aligned} \tag{3.96}$$

which proves the second bound. The choice $L = -\Delta_+\Delta_-$, $\delta L = -rI||L||$ and f the eigenvector of L corresponding to its smallest eigenvalue shows the sharpness of (3.91). The final step, $||\Delta_- U|| \leq h||u||_{H_1}$, follows from Cauchy-Schwarz' inequality. Note that (3.91) cannot be improved to $\mathcal{O}(h)$ be replacing $||U||$ with $||u||_{H_1}$. □

In two dimensions truncation to simple banded form is in general not adequate, since the exact operator will typically be block banded. However, both the crude truncation and the band projection generalizes easily to treat block banded form instead of just banded. Let M be the tensor product of two $N \times N$ matrices. Then we define truncation as

$$\text{trunc}_2(M,\nu)_{ij} = \begin{cases} M_{ij}, & \text{if } 2\,|i-j-rN| \leq \nu - 1 - |2r|, \\ 0, & \text{otherwise}, \end{cases} \qquad |2r| + 1 \leq \nu. \tag{3.97}$$

The motivation for this is Proposition 3.7. If $M = M_v \otimes M_u$ then $\text{trunc}_2(M) = \text{trunc}(M_v) \otimes \text{trunc}(M_u)$. See Section 4.3.2 for an example where block banded truncation is compared to simple banded truncation. For the band projection, the space $\mathcal{T}_\nu$ of banded matrices in the one-dimensional definition, is simply replaced by the space of matrices with the block banded sparsity pattern defined for two-dimensional truncation in (3.97).

In two dimensions, untangling the various H components of (3.68) from $\bar{L}$ is more complicated than finding the H in (3.65) and (3.66) for one-dimensional problems. There is however no unsurmountable difficulty, and truncating H instead of $\bar{L}$ can be done also for two-dimensional problems.

3.2 Elliptic Problems

In this section we present numerical results for the algorithms described above. We consider first the one-dimensional elliptic model equation,

$$-\frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right) = f(x), \qquad u(0) = u'(1) = 0, \tag{3.98}$$

with $a(x) > 0$ and then the Helmholtz equation,

$$-\sum_{i=1}^{d} \frac{\partial}{\partial x_i}(a(x)\frac{\partial u}{\partial x_i}) - \omega^2 u(x) = 0, \tag{3.99}$$

in one and two dimensions ($d = 1, 2$). The Helmholtz equation can be derived from the wave equation $v_{tt} = \nabla \cdot a(x)\nabla v$ under the assumption that $v(x,t) = u(x)e^{i\omega t}$. Hence $u(x)$ represents the amplitude and ω is the frequency. See also the derivation of the scalar Helmholtz equation from the Maxwell equations in Section 4.1.

We will use a uniform finite volume discretization in the numerical experiments. The computational domain will be the unit interval (square in 2D). The grid size will be denoted n, and the cell size $h = 1/n$. For a given function, $u(x)$, the corresponding grid function will be written u_i (or u_{ij}). The grid function approximates cell averages,

$$u_i \sim \frac{1}{h}\int_{(i-1)h}^{ih} u(x)dx, \qquad i = 1, \ldots, n, \tag{3.100}$$

in one dimension and

$$u_{ij} \sim \frac{1}{h^2}\int_{(i-1)h}^{ih}\int_{(j-1)h}^{jh} u(x_1, x_2)dx_1 dx_2, \qquad i, j = 1, \ldots, n, \tag{3.101}$$

in two dimensions. For the whole vector $(u_1, \ldots, u_n)^T$ we will write U. In two dimensions, $U = (u_{11}, \ldots, u_{1n}, u_{21}, \ldots, u_{2n}, \ldots, u_{nn})^T$.

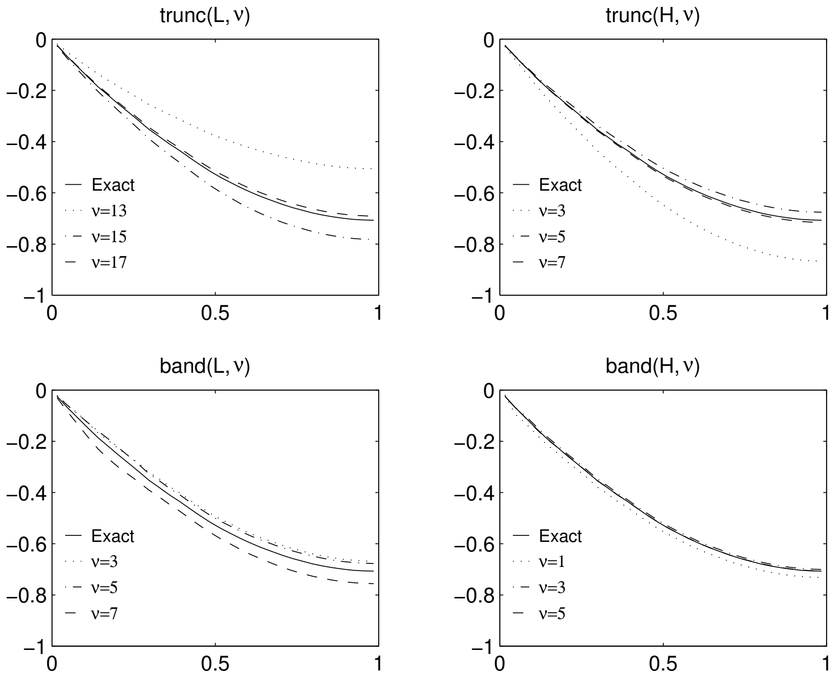


Figure 3.1. Result for the elliptic model problem, $a(x)$ random, when the homogenized operator is approximated in different ways. The "exact" solution refers to the solution with the full 32×32 homogenized operator.

3.2.1 The 1D Elliptic Model Equation

We approximate (3.98) with the discretization

$$-\frac{1}{h^2}\Delta_+ a_i \Delta_- u_i = f_i, \qquad i = 1, \dots, n, \qquad u_0 = -u_1, \qquad u_{n+1} = u_n. \quad (3.102)$$

First we set $f = -1$ and let the coefficients of $a(x)$ have a uniform random distribution in the interval $[0.5, 1]$. We take $n = 256$ and make three homogenization steps, i.e. the coarsest level contains 32 grid points.

In Figure 3.1 we compare different truncation strategies. In the top two subplots we use crude truncation, (3.80). In the bottom two subplots we use the band projection described in Section 3.1.5. The approximation is performed on H, see (3.65), and on $\bar{L}$ after all three homogenizations. We see that band projection gives a better approximation. We also see that it is more efficient to truncate H than to truncate $\bar{L}$. The exact reference solution here is the numerically homogenized solution at the coarsest level without any truncation. This is equivalent to the projection onto the coarse scale of the solution on the finest scale.

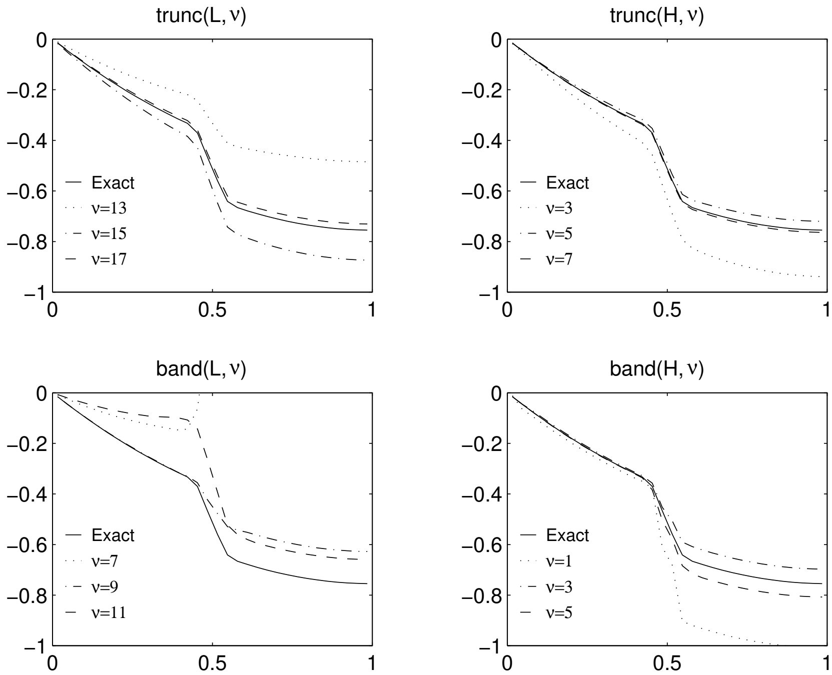


Figure 3.2. Result for the elliptic model problem, $a(x)$ a slit, when the homogenized operator is approximated in different ways. The “exact” solution refers to the solution with the full 32×32 homogenized operator.

Next we change the coefficient in the differential equation to be

$$a(x) = \begin{cases} 1/6 & 0.45 < x < 0.55 \\ 1 & \text{otherwise.} \end{cases} \tag{3.103}$$

We keep all other characteristics. The result is given in Figure 3.2 and it shows that the relative merit of the different methods are more or less the same. The structure of the untruncated $\bar{L}$ and H matrices is shown in Figure 3.3. It should be noted that the local inhomogeneity of the full operator has spread out over a larger area, but it is still essentially local.

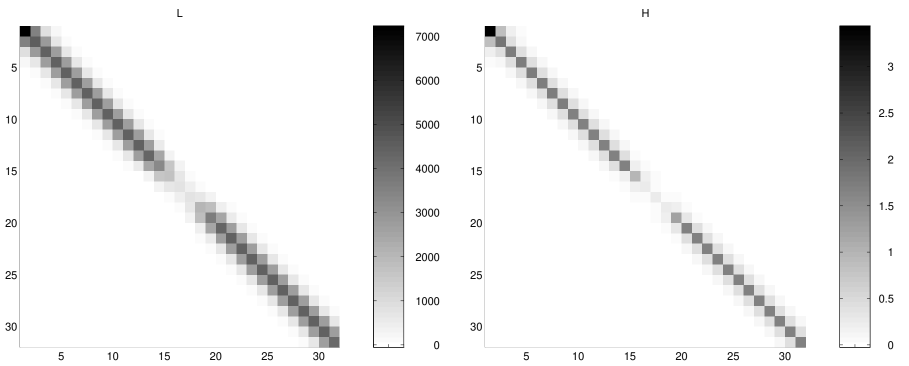


Figure 3.3. Structure of the untruncated homogenized operators $\bar{L}$ (left) and H (right) for the elliptic model problem, $a(x)$ a slit.

3.2.2 The 1D Helmholtz Equation

We discretize (3.99) in one dimension, with the boundary conditions $u(0) = 1$ and $u'(1) = 0$,

$$\begin{aligned} -\frac{1}{h^2}\Delta_+ a_i \Delta_- u_i - \omega^2 u_i = 0, \quad i = 1, \dots, n, \\ u_0 = 2u(0) - u_1, \\ u_{n+1} = u_n. \end{aligned} \tag{3.104}$$

We use $\omega = 2\pi$ and the same $a(x)$ as in (3.103) and again we take $n = 256$ and use three homogenizations. We get

$$\bar{L}_j u = (\tilde{L}_j - \omega^2 I)u = 0. \tag{3.105}$$

Truncation is performed on $\tilde{L}_j$ (or $\tilde{H}_j$) and not on $\bar{L}_j$. The result is in Figure Figure 3.4. We see that Helmholtz equation gives results similar to those of the model equation. Again band projection is more efficient than truncation and approximating H_j is more efficient than approximating $\bar{L}_j$.

3.2.3 The 2D Helmholtz Equation

We consider the two-dimensional version of (3.99), with periodic boundary conditions in the y-direction, and $u(0, y) = 1$, $u_x(1, y) = 0$ at the left and right boundaries respectively. This is a simple model of a plane time-harmonic wave of amplitude one entering the computational domain from the left and flowing out at the right. The discretization that we use is

$$\begin{aligned} -\frac{1}{h^2}\Delta_+^x a_{i\ell} \Delta_-^x u_{i\ell} - \frac{1}{h^2}\Delta_+^y a_{i\ell} \Delta_-^y u_{i\ell} - \omega^2 u_{i\ell} = 0, \qquad i, \ell = 1, \dots, n, \\ u_{i,0} = u_{i,n}, \quad u_{i,n+1} = u_{i,1}, \quad u_{n+1,\ell} = u_{n,\ell} \quad u_{0,\ell} = 2 - u_{1,\ell}. \end{aligned} \tag{3.106}$$

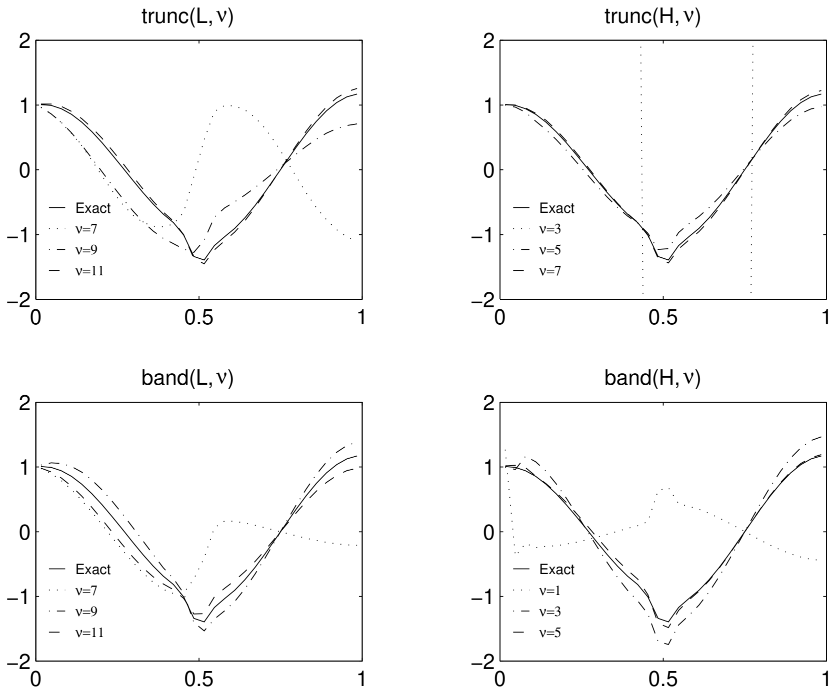
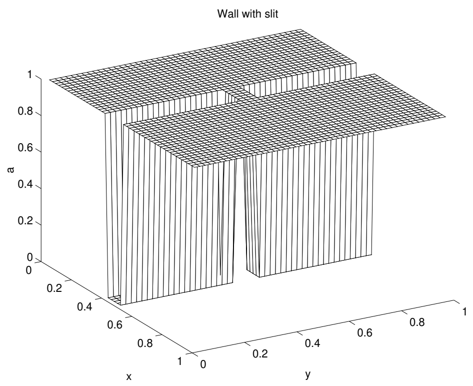


Figure 3.4. Result for the Helmholtz equation, $a(x)$ a slit, when the homogenized operator is approximated in different ways. The "exact" solution refers to the solution with the full 32×32 homogenized operator.



$$a(x,y) = \begin{cases} 10^{-4}, & \text{if } 0.4 < x < 0.5 \text{ and} \\ & |y - 0.5| > 0.05, \\ 1, & \text{otherwise.} \end{cases}$$

Figure 3.5. The variable coefficient $a(x,y)$ used in the 2D Helmholtz example.

This leads to the matrix equation

$$L_{j+1}U = F, \qquad U, F \in \boldsymbol{V}_{j+1}, \qquad n = m2^{j+1}, \tag{3.107}$$

where m is a positive integer. L_{j+1} is homogenized following the theory for two-dimensional problems in Section 3.1.3.

As an example we choose the $a(x,y)$ shown in Figure 3.5, which models a wall with a small slit where the incoming wave can pass through. With $\omega = 3\pi$ and $n = 48$, we obtained the results presented in Figure 3.6.

The operator after one homogenization step is truncated according to (3.97). We show the results of truncation in Figure 3.7, for various values of ν. The case $\nu = 9$ corresponds to a compression to approximately 7% of the original size. The structure of this operator is shown in Figure 3.8.

3.3 Hyperbolic Problems

In this section we consider the time-dependent hyperbolic model equation in one space dimension,

$$\frac{\partial u}{\partial t} + a_\epsilon(x)\frac{\partial u}{\partial x} = f(x,t), \qquad u(0,t) = 0, \quad u(x,0) = g(x). \tag{3.108}$$

The variable coefficient $a_\epsilon(x)$ is positive and bounded.

We will use three different methods for the numerical homogenization of (3.108). First, we consider a semi-discrete form of the problem, and only homogenize the spatial part of the operator. Second, we make a correction for the extra errors introduced when neglecting the time-derivative. Third, we homogenize the full two-dimensional operator.

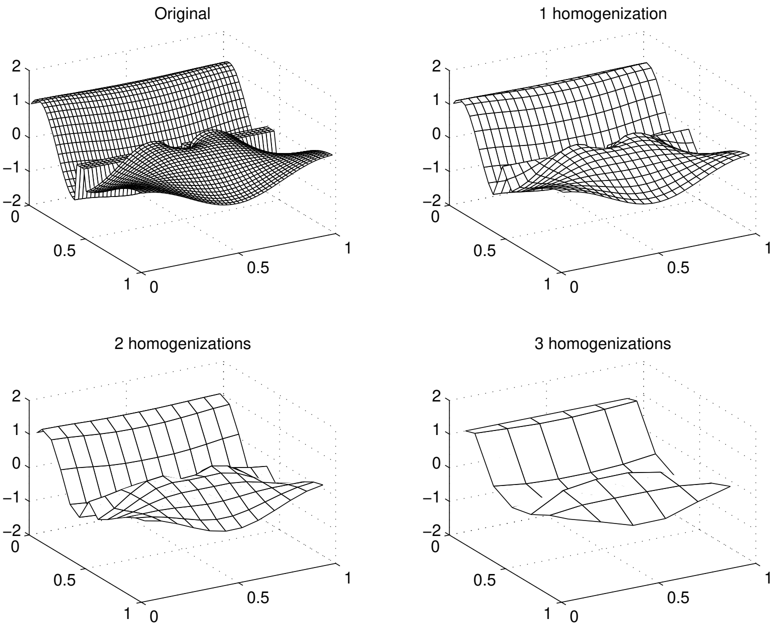


Figure 3.6. Result for the 2D Helmholtz example. Solution shown for $0, \ldots, 3$ homogenization steps.

The most straightforward way to homogenize (3.108) is by a semi-discrete approximation,

$$\frac{\partial u_i}{\partial t} + \frac{1}{h} a_i \Delta_- u_i = f_i(t), \qquad i = 1, \ldots, n, \qquad u_0 = -u_1, \quad u_i(0) = g_i, \quad (3.109)$$

with $a_i = a_\epsilon((i - 1/2)h)$. In matrix form we will have

$$U_t + L_{j+1} U = F, \qquad 2^{j+1} = n, \qquad (3.110)$$

which can be expressed in a wavelet basis as

$$\frac{\partial}{\partial t}\begin{pmatrix} U_h \\ U_l \end{pmatrix} + \begin{pmatrix} A_j & B_j \\ C_j & D_j \end{pmatrix}\begin{pmatrix} U_h \\ U_l \end{pmatrix} = \begin{pmatrix} F_h \\ F_l \end{pmatrix}, \qquad U_h, F_h \in W_j, \qquad U_l, F_l \in V_j. \qquad (3.111)$$

We now claim that we can homogenize L_{j+1} in the same way as in the elliptic case, using the Schur complement. The motivation for this is (3.73), which shows that when the scales of $a_\epsilon(x)$ become very small, $\epsilon \to 0$, the effect of the time-derivative vanishes, $\frac{\partial}{\partial t} U_h \to 0$. As a limiting process in classical homogenization, this is fully justified, compare [6]. This argument implies that in the first set of equations of (3.111),

$$\frac{\partial U_h}{\partial t} + A_j U_h + B_j U_l = F_h \quad \Rightarrow \quad U_h = A_j^{-1}\left(-B_j U_l + F_h - \frac{\partial U_h}{\partial t}\right), \qquad (3.112)$$

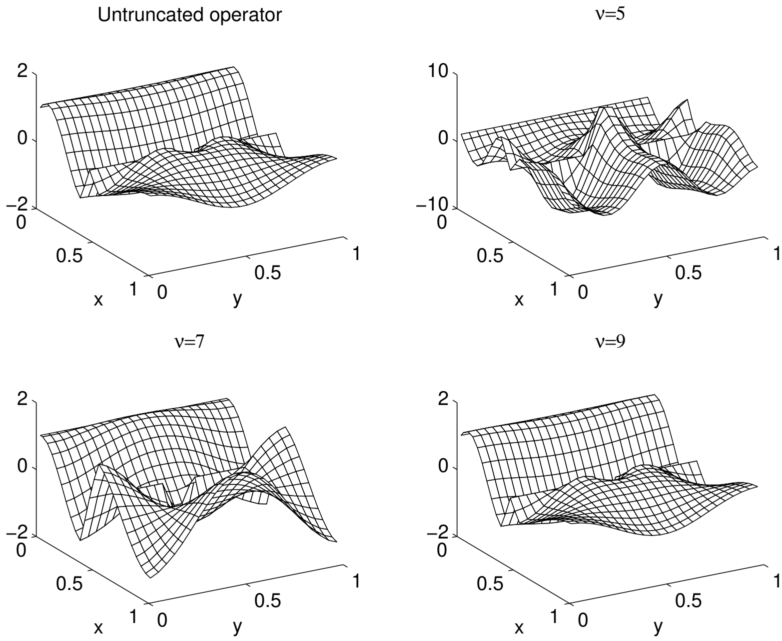


Figure 3.7. Results for the 2D Helmholtz example, using the one step homogenized operator, truncated with different ν.

the last term, $-A_j^{-1}\frac{\partial}{\partial t}U_h$ is eliminated. Substitution into the second set of equations of (3.111) yields

$$\frac{\partial U_l}{\partial t} + \bar{L}_j U_l = \bar{F}_j, \tag{3.113}$$

with $\bar{L}_j$ and $\bar{F}_j$ defined by (3.54).

We found experimentally that using an approximation of the term $-A_j^{-1}\frac{\partial}{\partial t}U_h$ improved the homogenized operator. The correction is derived from (3.112) after differentiation,

$$\frac{\partial U_h}{\partial t} = A_j^{-1}(-B_j\frac{\partial U_l}{\partial t} + \frac{\partial F_h}{\partial t} - \frac{\partial^2 U_h}{\partial t^2}). \tag{3.114}$$

After a transient mode we have for the model problem (3.108) with $a_\epsilon = a(x/\epsilon)$ and $a(y)$ 1-periodic,

$$u_\epsilon(x,t) = \bar{u}(x,t) + \epsilon u_1(x, x/\epsilon, t) + \mathcal{O}(\epsilon^2), \tag{3.115}$$

compare [6]. Hence, $\frac{\partial^2}{\partial t^2}U_h$ is of the same order as $\frac{\partial}{\partial t}U_h$ since the oscillations on the ϵ-scale are not functions of time and we here choose to neglect $-A_j^{-1}\frac{\partial^2}{\partial t^2}U_h$ in (3.114) instead of $-A_j^{-1}\frac{\partial}{\partial t}U_h$ in (3.112). Substituting (3.114) after elimination, into (3.111) gives

$$K_j\frac{\partial U_l}{\partial t} + (D_j - C_jA_j^{-1}B_j)U_l = F_l - C_jA_j^{-1}F_h + C_jA_j^{-2}\frac{\partial F_h}{\partial t}, \tag{3.116}$$

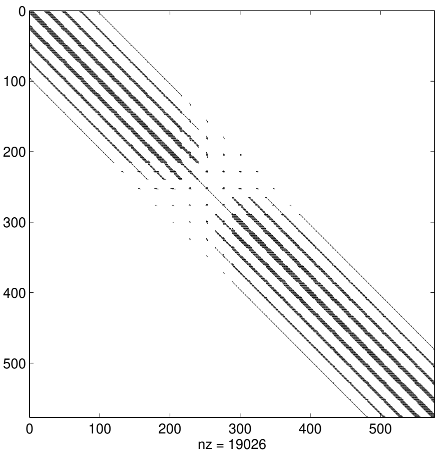


Figure 3.8. Structure of the homogenized operator $\bar{L}$, after one homogenization step, for the 2D Helmholtz example. Elements larger than 0.1% of max value shown.

with the correction matrix

$$K_j = I + C_j A_j^{-2} B_j. \tag{3.117}$$

Hence, the corrected homogenized operator $\tilde{L}_j$ and right hand side $\tilde{F}_j$ are

$$\tilde{L}_j = K_j^{-1} \bar{L}_j, \qquad \tilde{F}_j = K_j^{-1} \bar{F}_j + K_j^{-1} C_j A_j^{-2} \frac{\partial F_h}{\partial t} \tag{3.118}$$

and we will solve (3.113), with $\tilde{L}_j$, $\tilde{F}_j$ substituted for $\bar{L}_j$, $\bar{F}_j$ respectively.

In principle, better approximations could be obtained by reiterating the steps above. After repeated differentiations of (3.114) we could choose to eliminate terms involving successively higher order time derivatives of U_h. However, the condition numbers of the correction matrices produced in this manner rapidly deteriorate.

Let $n = 128$, $f \equiv 0$ and

$$g(x) = \begin{cases} \sin^2(4\pi x), & 0 \le x \le 0.25 \\ 0, & 0.25 < x \le 1. \end{cases} \tag{3.119}$$

The result at time $t = 1$, is given in Figure 3.9 for the case of $a_\epsilon(x)$ uniformly random distributed in $[0.1, 2]$ and $a_\epsilon(x)$ as in (3.103). The effect of the first order correction is shown to the right. In these calculations, the time-integration was replaced by a fourth-order Runge-Kutta method.

Because of causality the full operator must be lower triangular in the hyperbolic case. The homogenized operator will keep this property in the Haar case. After two homogenization steps the operator, with $a_\epsilon(x)$ a slit, has the structure

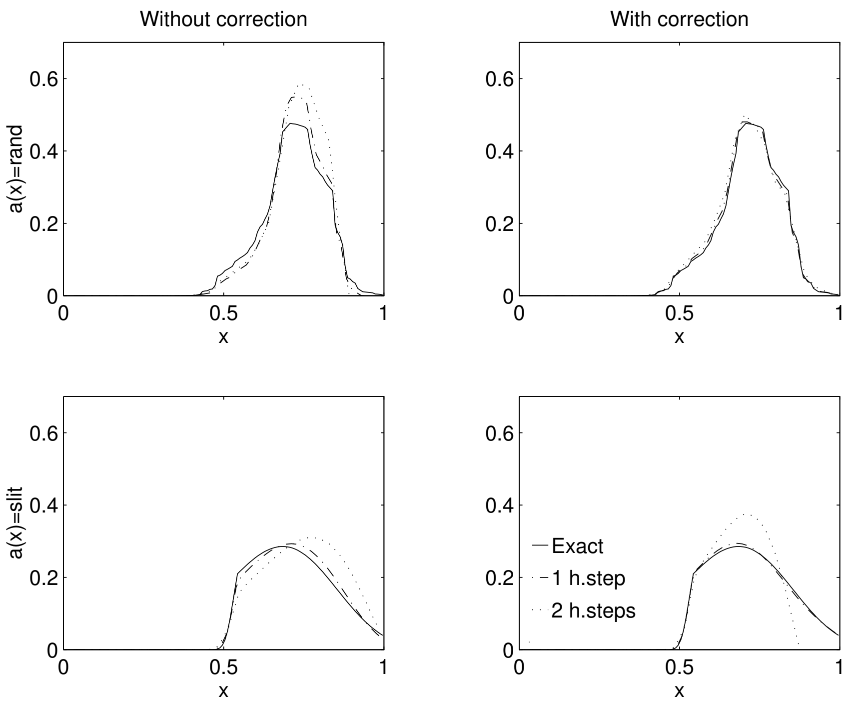


Figure 3.9. The solution $u(x)$ at $t = 1$ for the hyperbolic case, using different $a_\epsilon(x)$, with and without first order correction. Solutions shown are the exact solution, computed with the full operator, and solutions computed with the operator homogenized one and two steps.

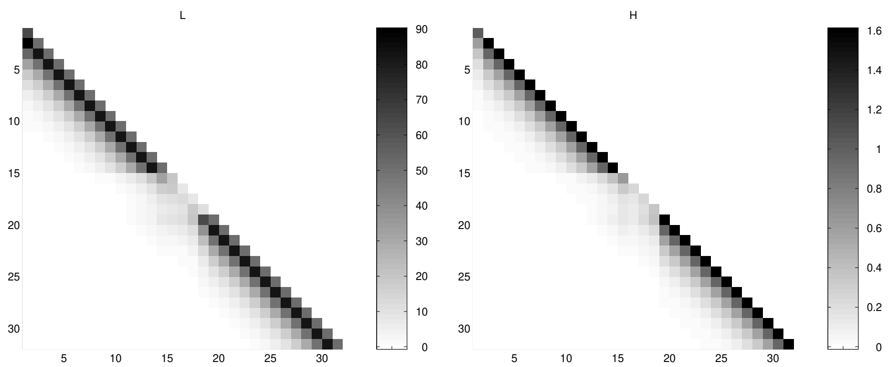


Figure 3.10. Structure of the untruncated homogenized operators $\bar{L}$ (left) and H (right) for the hyperbolic problem, $a_\epsilon(x)$ a slit.

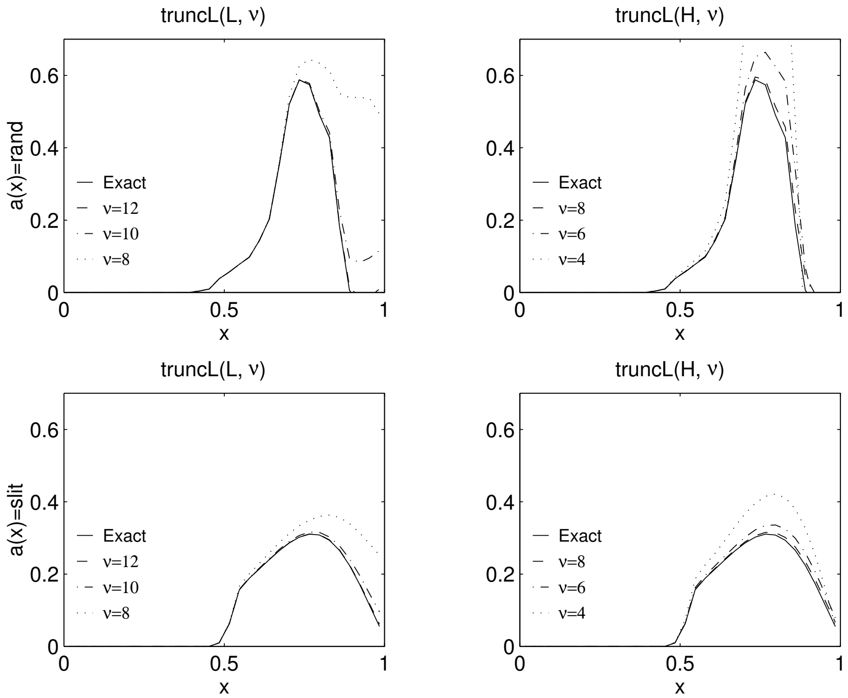


Figure 3.11. Solution $u(x)$ for the hyperbolic case, using different $a_\epsilon(x)$ and different approximations of the homogenized operator. The "exact" solution refers to the solution with the full 32×32 homogenized operator.

shown in Figure 3.10. In view of this we approximate $\bar{L}$ and H, see (3.66), by truncation to lower triangular form. In Figure 3.11 results using this truncation is displayed. The same two types of a_ϵ as in Figure 3.9 were used. Like in the elliptic case, truncating the H matrix is more efficient than truncating $\bar{L}$. When the correction method is used it is harder to approximate the resulting homogenized operator, $\tilde{L}$, with a sparse matrix. It will not be as diagonal dominant as in the non-corrected case.

In our third approach we leave the semi-discrete form and instead make a full two-dimensional discretization of (3.108) in (x,t) space, using the implicit Euler approximation in time,

$$\begin{aligned}\frac{1}{h}\Delta_-^t u_{i\ell} + \frac{1}{h} a_i \Delta_-^x u_{i\ell} &= f_i, \qquad i,\ell = 1,\dots,n,\\ u_{0,\ell} &= -u_{1,\ell},\\ u_{i,0} &= 2g_i - u_{i,1}.\end{aligned} \tag{3.120}$$

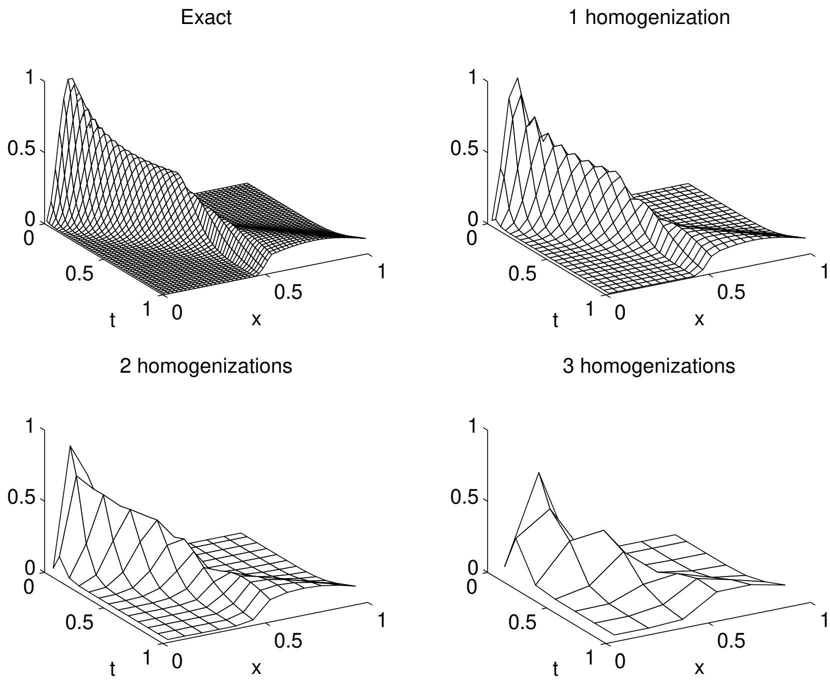


Figure 3.12. Result for the 2D hyperbolic case, with $a_\epsilon(x)$ a slit.

Like in the two-dimensional Helmholtz case, Section 3.2.3, we arrive at a linear system of equations,

$$L_{j+1}U = F, \qquad U, F \in \boldsymbol{V}_{j+1}, \qquad n = 2^{j+1}, \tag{3.121}$$

which is homogenized in the standard way, with the Schur complement.

Using the same data as in the semi-discrete case, but with $n = 48$, we get for the case of a_ϵ as in (3.103) the results in Figure 3.12. We used the untruncated homogenized operator, but its structure indicates that, like in the 2D Helmholtz case, good sparse approximations can be found.

Chapter 4

Simulation of Optical Filter

In this chapter we will study a specific problem to which we apply some of the homogenization techniques described in the previous chapter. The goal is to simulate an optical filter, where small scales in the geometry interact with small scales in the boundary data. In connection to this we will give some examples of how the homogenized operator can be used.

The filter is composed of a straight waveguide with small gratings engraved in one of its sides at regular intervals. At one point the distance between two grating teeth is increased by 50%, see Figure 4.1. This *quarter-wave step* gives the filter effect and causes waves of one resonant frequency to pass through, while reflecting adjacent frequencies. Its good ability to separate frequencies makes the filter a useful component in many communication applications. Such narrow-band filters allow an optical communication link to be partitioned into many channels, where signals can be selectively transmitted and detected using the filters. For some applications, see for instance [40] and [29].

4.1 Mathematical Model

The Maxwell equations govern the behavior of electromagnetic waves. Let us assume that the waves are time-harmonic, propagating in a linear, isotropic, homogeneous and non-conducting medium. Then the Maxwell equations can be reduced to a system of Helmholtz equations for the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{H}$,

$$\nabla^2 \boldsymbol{E} + \mu\epsilon\omega^2 \boldsymbol{E} = 0, \qquad \nabla^2 \boldsymbol{H} + \mu\epsilon\omega^2 \boldsymbol{H} = 0, \tag{4.1}$$

where μ and ϵ represent the medium's permeability and permittivity respectively, and ω is the (fixed) frequency of the wave. The equations in (4.1) are coupled by boundary conditions. In the rest of the chapter we will only be interested in the z-component of the $\boldsymbol{H}$-field and we will denote it with u. It is clear from (4.1) that u satisfies the scalar Helmholtz equation.

In our simulation of optical filters we will consider a waveguide oriented along the xy-plane, infinite in the z direction. Since the fields will be independent of the z-coordinate, we arrive at a two-dimensional problem, see Figure 4.1. The filter is piecewise homogeneous and we need to know the boundary conditions at the interface between the media. Suppose there is a curve in the xy-plane along which the material coefficients ϵ, μ has a discontinuity. Let $\partial/\partial n$, $\hat{n}$, $\hat{t}$ be the normal derivative, the normal and the tangent corresponding to this curve. Then, using the relationship $\nabla \times \boldsymbol{H} = i\omega\epsilon\boldsymbol{E}$ dictated by the Maxwell equations,

$$\frac{\partial u}{\partial n} = \hat{n}\cdot\nabla u = \hat{n}\cdot\begin{pmatrix} \frac{\partial H_z}{\partial x} \\ \frac{\partial H_z}{\partial y} \end{pmatrix} = \hat{n}\cdot\begin{pmatrix} -\frac{\partial H_x}{\partial z} + \frac{\partial H_z}{\partial x} \\ \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} \end{pmatrix} = i\omega\epsilon\hat{n}\cdot\begin{pmatrix} -E_y \\ E_x \end{pmatrix} = i\omega\epsilon\boldsymbol{E}\cdot\hat{t}. \quad (4.2)$$

It is well-known for Maxwell equations that at the interface between two lossless media the tangential components of both the $\boldsymbol{E}$ and the $\boldsymbol{H}$ fields are continuous. Moreover, for the non-magnetic materials that we consider here, μ vary very little and we will assume it to be constant. After introducing the speed of light $c = 1/\sqrt{\mu\epsilon}$, we must therefore have that both u and $c^2\partial u/\partial n$ are continuous across the interface. In conclusion we seek a weak solution $u \in C(\mathbb{R}^2)$ of

$$\nabla \cdot (c^2\nabla u) + \omega^2 u = 0. \quad (4.3)$$

The internal boundary condition will automatically be satisfied with this formulation.

In Figure 4.1 a more detailed description of the actual problem is given. The figure shows the variables that define the dimensions of the waveguide. We will usually fix the length $w = 1$ and use $a, d, , \Lambda, \delta$ and τ as design parameters. Further parameters are S, the number of gratings, and η_0, η_1, the indices of refraction outside and inside the waveguide respectively. As in Chapter 2 the index of refraction is defined as $\eta(x, y) = 1/c(x, y)$. Throughout this chapter we assume the material of the waveguide is GaAs, with $\eta = 3.3$ and that it is surrounded by vacuum, $\eta = 1.0$.

4.1.1 One-Dimensional Model

The two-dimensional model above can be further simplified if we assume that $c(x, y)$ can be replaced by an effective speed of propagation c_{eff} which does not depend on y. Then we get

$$\partial_x(c_{\text{eff}}(x)^2\partial_x u) + \omega^2 u = 0 \quad (4.4)$$

where u should be a weak solution in $C([0, w])$.

The problem here is how to obtain $c_{\text{eff}}(x)$ from $c(x, y)$. A simple approach is to relate $c_{\text{eff}}(x)$ to the local width of the waveguide, provisionally denoted $d(x)$. For a fixed x and $d(x)$, we will let $c_{\text{eff}}(x)$ be the speed of the lowest propagating mode in an infinite straight waveguide of (constant) width $d(x)$ $(w = \infty, S = 0)$

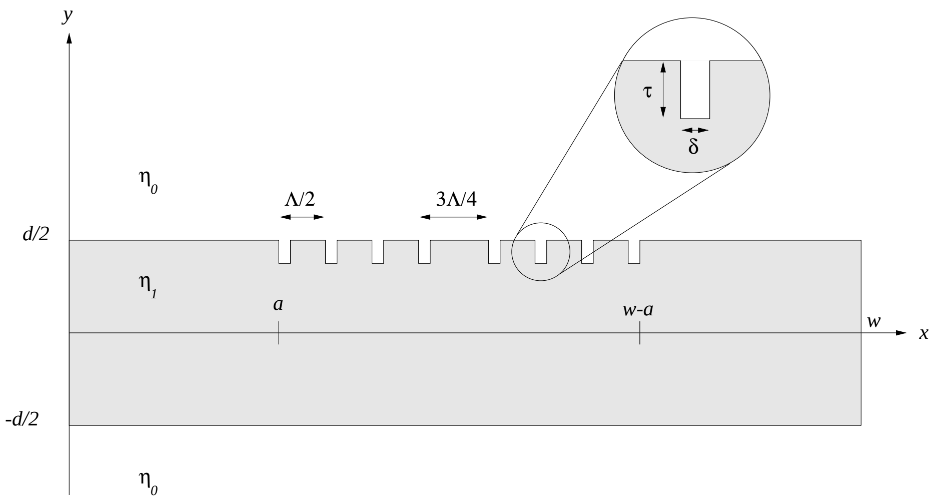


Figure 4.1. Specifications of the waveguide under study.

at frequency ω. The infinite waveguide will have the same indices of refraction as the original problem.

Let d be fixed and consider an infinite straight waveguide. The solution of (4.3) must be of the form $u(x,y) = f(y)\exp(i\lambda x)$ with $\lambda \in \mathbb{R}$ in order to be a propagating mode. Moreover,

$$c(x,y) = c(y) = \begin{cases} c_0 & |y| > d/2, \\ c_1 & |y| \le d/2. \end{cases} \tag{4.5}$$

This gives us a one-dimensional eigenvalue problem,

$$\frac{1}{c^2}\partial_y(c^2\partial_y f) + \frac{\omega^2}{c^2} f = \lambda^2 f, \qquad f \in C(\mathbb{R}). \tag{4.6}$$

The problem (4.6) has a discrete spectrum, where the eigenmodes are

$$f(y) = \begin{cases} \cos\gamma y & |y| \le d/2, \\ \cos\frac{\gamma d}{2} e^{-\alpha(|y|-d/2)} & |y| > d/2, \end{cases} \qquad \lambda^2 = \eta_1^2\omega^2 - \gamma^2 = \eta_0^2\omega^2 + \alpha^2. \tag{4.7}$$

The unknowns α, γ are given by the nonlinear equations

$$\alpha^2 + \gamma^2 = (\eta_1^2 - \eta_0^2)\omega^2, \qquad \alpha = \gamma\frac{\eta_0^2}{\eta_1^2}\tan\frac{\gamma d}{2}. \tag{4.8}$$

In general there are only a finite number of real eigenvalues. We now let λ_0 denote the largest real eigenvalue. For the fixed width d we set $c_{\text{eff}} = \omega/\lambda_0$. Since d actually depends on x, we get $c_{\text{eff}} = c_{\text{eff}}(x)$ for fixed ω, η_0 and η_1.

4.2 Numerical Approximation

4.2.1 Discretization of the Helmholtz Equation

For the one-dimensional model (4.4) we introduce a grid of $N+2$ equally spaced points $\{x_k\}_{k=0}^{N+1}$, such that

$$x_k = (k - 1/2)h, \qquad h = w/N. \tag{4.9}$$

We also define grid functions corresponding to $u(x)$ and $c_{\text{eff}}(x)$,

$$\begin{aligned} u_k &\approx u(x_k), & k &= 0, \dots, N+1, \\ c_{k+1/2} &= c_{\text{eff}}(x_k + h/2), & k &= 0 \dots, N. \end{aligned} \tag{4.10}$$

Moreover, let

$$\Delta_+ u_k = u_{k+1} - u_k, \qquad \Delta_- u_k = u_k - u_{k-1}, \tag{4.11}$$

be the forward and backward undivided differences. Using these to approximate the derivatives in (4.4), we get the following second order scheme,

$$\frac{1}{h^2}\Delta_+ c_{k-1/2}^2 \Delta_- u_k + \omega^2 u_k = 0, \qquad k = 1, \dots, N. \tag{4.12}$$

We hence use a discretization that corresponds to the same divergence form as the continuous equation. In the two-dimensional case, we start by truncating the physical domain to the square $[0, w] \times [-b, b]$ where $b > d/2$ is an additional parameter. The grid will have $M + 1$ grid points in the y-direction, denoted (x_k, y_ℓ), where $y_\ell = -b + (\ell - 1/2)h$ for $\ell = 0, \dots, M+1$. Since we have square cells, the variables will satisfy the relationship by $M = 2Nb$. We will only use b and d values such that M is an integer and $y_\ell = d/2$ for some ℓ. The grid functions are

$$u_{k\ell} \approx u(x_k, y_\ell), \qquad k = 0, \dots, N+1, \qquad \ell = 0, \dots, M+1 \tag{4.13}$$

and letting x^+ (x^-) denote the limiting values from above (below),

$$\begin{aligned} c_{k+1/2,\ell} &= c(x_k^- + h/2, y_\ell^*), & k &= 0, \dots, N, & \ell &= 0, \dots, M+1, \\ c_{k,\ell+1/2} &= c(x_k^-, y_\ell^* + h/2), & k &= 0, \dots, N+1, & \ell &= 0, \dots, M \end{aligned} \tag{4.14}$$

where $y_\ell^* = y_\ell^+$ for positive y_ℓ and $y_\ell^* = y_\ell^-$ for negative y_ℓ. In the two dimensional case, the difference operators are defined,

$$\begin{aligned} \Delta_+^x u_{k\ell} &= u_{k+1,\ell} - u_{k\ell}, & \Delta_-^x u_{k\ell} &= u_{k\ell} - u_{k-1,\ell}, \\ \Delta_+^y u_{k\ell} &= u_{k,\ell+1} - u_{k\ell}, & \Delta_-^y u_{k\ell} &= u_{k\ell} - u_{k,\ell-1}. \end{aligned} \tag{4.15}$$

A second order scheme approximating (4.3) can then be written

$$\frac{1}{h^2}\Delta_+^x c_{k-1/2,\ell}^2 \Delta_-^x u_{k\ell} + \frac{1}{h^2}\Delta_+^y c_{k,\ell-1/2}^2 \Delta_-^y u_{k\ell} + \omega^2 u_{k\ell} = 0 \tag{4.16}$$

for $k = 1, \dots, N$ and $\ell = 1, \dots, M$. Again we let the discretization be of the same divergence form as the continuous equation.

4.2.2 Boundary Conditions

To complete the discretization, extra conditions must be given at the boundary of the computational domain. Since the wave solution that we are looking for is propagating along the x-axis, we will use a type of non-reflecting boundary conditions at $x = 0$ and $x = w$, in both the one and two-dimensional cases. In the two-dimensional case we also need a boundary condition at $y = \pm b$.

The non-reflecting boundary condition is based on the assumption that, in a neighborhood of the boundary, the outgoing part of the continuous solution is approximately a plane wave. Close to $x = w$ we thus make the ansatz

$$u(x, y) = f(y)e^{-i\beta x}, \tag{4.17}$$

which gives

$$\frac{\partial u}{\partial x} + i\beta u = 0, \qquad x = w. \tag{4.18}$$

At $x = 0$ we approximate the solution by a superposition of two plane waves, an unknown outgoing wave and an incoming wave with known amplitude,

$$u(x, y) = f(y)e^{i\beta x} + u_{\text{in}}(y)e^{-i\beta x}. \tag{4.19}$$

This gives

$$\frac{\partial u}{\partial x} - i\beta u = -2i\beta u_{\text{in}}(y)e^{-i\beta x}, \qquad x = 0. \tag{4.20}$$

At $y = \pm b$ we use a Neumann boundary condition,

$$\frac{\partial u}{\partial y} = 0, \qquad y = \pm b. \tag{4.21}$$

This is motivated by the fact that the solution should be almost constant zero far away from the waveguide.

Before we can discretize (4.18) and (4.20) we need to determine β. A natural choice would be $\beta = \omega/c_{\text{eff}}$, the wave number of an infinite straight waveguide. However, that discretization causes numerical dispersion which results in large phase errors if β is not compensated for this. We will use a modified β_h which also depends on the grid size h. It is computed in much the same way as c_{eff} in Section 4.1. Suppose the waveguide is of width d and make the ansatz $u_{k\ell} = f_\ell \exp(i\beta_h x_k)$ where $f_\ell \in \ell^2(\mathbb{R})$. Since $c(x, y)$ in this case is independent of x, we let

$$c_\ell = c(y_\ell^*), \qquad c_{\ell+1/2} = c(y_\ell^* + h/2), \qquad \ell \in \mathbb{Z} \tag{4.22}$$

where y_ℓ^* is defined as after equation (4.14) and $\{y_\ell\}$ is extended to an infinite grid in the obvious way. This leads to the discrete eigenvalue problem

$$\frac{1}{c_\ell^2 h^2}\Delta_+ c_{\ell-1/2}^2 \Delta_- f_\ell + \frac{\omega^2}{c_\ell^2} f_\ell = \lambda^2 f_\ell, \qquad \ell \in \mathbb{Z}. \tag{4.23}$$

The eigenmodes are

$$f_\ell = \begin{cases} \cos\gamma y_\ell & |y_\ell| \le d/2, \\ \cos\frac{\gamma d}{2} e^{-\alpha(|y_\ell|-d/2)} & |y_\ell| > d/2, \end{cases} \qquad \begin{aligned} \lambda^2 &= \eta_1^2\omega^2 - \tfrac{4}{h^2}\sin^2\tfrac{\gamma h}{2}, \\ &= \eta_0^2\omega^2 + \tfrac{4}{h^2}\sinh^2\tfrac{\alpha h}{2} \end{aligned} \tag{4.24}$$

where α, γ are given by

$$\frac{4}{h^2}\Big(\sin^2\frac{\gamma h}{2} + \sinh^2\frac{\alpha h}{2}\Big) = (\eta_1^2 - \eta_0^2)\omega^2 \tag{4.25}$$

and

$$e^{\alpha h} - 1 = \frac{\eta_0^2}{\eta_1^2}\left(\sin(\gamma h)\tan\left(\frac{\gamma d}{2}\right) - 2\sin^2\frac{\gamma h}{2}\right). \tag{4.26}$$

Like in the continuous case, we let λ_0 be the largest real eigenvalue. Its corresponding eigenvector will be denoted $\{f_\ell^0\}$ The wave number β_h is related to λ_0 as

$$\lambda_0^2 = \frac{4}{h^2}\sin^2\frac{\beta_h h}{2}. \tag{4.27}$$

In our experiments below we are only interested in the behavior of the lowest mode of propagation. For the two-dimensional case, we will therefore always use $u_{\text{in}}(y_\ell) = f_\ell^0$ in order to inject this mode at $x = 0$.

It should be noted that equations (4.24)–(4.27) all transform into their continuous equivalents when $h \to 0$, as expected. In particular $\lim_{h\to 0}\beta_h = \omega/c_{\text{eff}}$.

We continue with the discretization of the boundary conditions (4.18) and (4.20). A simple second order discretization would be

$$\frac{1}{h}\Delta_+^x u_{0,\ell} - i\beta_h\frac{u_{0,\ell}+u_{1,\ell}}{2} = -2i\beta_h u_{\text{in}}(y_\ell), \tag{4.28}$$

$$\frac{1}{h}\Delta_+^x u_{N,\ell} - i\beta_h\frac{u_{N,\ell}+u_{N+1,\ell}}{2} = 0. \tag{4.29}$$

for $\ell = 1, \dots, M$. Again, this discretization introduces phase errors. We will use an alternative second order discretization which is exact for (4.17) and (4.19) when $\beta = \beta_h$ and $f(y_\ell) = u_{\text{in}}(y_\ell) = f_\ell^0$. Since

$$\Delta_+ e^{i\beta_h x_k} = (e^{i\beta_h h} - 1)e^{i\beta_h x_k} \tag{4.30}$$

we get

$$\begin{aligned} \frac{1}{h}\Delta_+ u_{0,\ell} + \frac{1-e^{i\beta_h h}}{h}u_{0,\ell} &= -2i\frac{\sin(\beta_h h)}{h}e^{i\frac{\beta_h h}{2}}u_{\text{in}}(y_\ell), \\ \frac{1}{h}\Delta_+ u_{N,\ell} + \frac{1-e^{-i\beta_h h}}{h}u_{N,\ell} &= 0, \end{aligned} \tag{4.31}$$

for $\ell = 1, \dots, M$. The boundary condition at $y = \pm b$ is not as sensitive to phase error as the ones at $x = 0$ and $x = w$. We can use simple central differencing to get the second order discretization

$$\frac{1}{h}\Delta_+^y u_{k,0} = 0, \qquad \frac{1}{h}\Delta_+^y u_{k,M} = 0, \qquad k = 1, \dots, N. \tag{4.32}$$

The one-dimensional model is treated in the same way as the two-dimensional model at the left and right boundaries. The equations corresponding to (4.31) read

$$\begin{aligned} \frac{1}{h}\Delta_+ u_0 + \frac{1-e^{i\beta_h h}}{h}u_0 &= -2i\frac{\sin(\beta_h h)}{h}e^{i\frac{\beta_h h}{2}}u_{\text{in}}, \\ \frac{1}{h}\Delta_+ u_N + \frac{1-e^{-i\beta_h h}}{h}u_N &= 0. \end{aligned} \tag{4.33}$$

4.3 Results

4.3.1 Simulation

Figure 4.2 shows examples of how the solution looks for the one- and the two-dimensional models, both in the case of total reflection and total transmission of the injected wave. For the one-dimensional problem (4.12) and (4.33) were used as the numerical approximation. For the two-dimensional problem (4.16) and (4.31) were used.

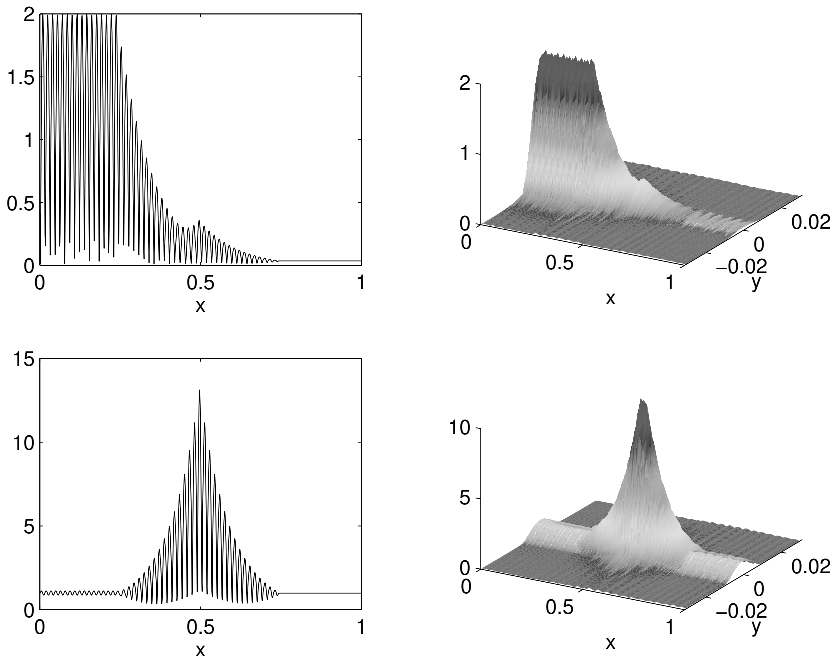


Figure 4.2. Plots of $|u|$ for frequencies in the stopband (top) and at resonance (bottom) for the one-dimensional (left) and two-dimensional (right) model. In this example $w = 1$, $a = 1/4$, $S = 32$, $d = 17/1024$, $b = 32/1024$, $\Lambda = 32/1024$, $\tau = 4/1024$, $\delta = 4/1024$ and $N = 1024$, $M = 64$.

The power transmission of the filter is shown in Figure 4.3 for a range of frequencies. This coefficient is defined and computed as

$$T_{\text{pow}} \equiv \frac{\int_{-\infty}^{\infty} |u(L,y)|^2 dy}{\int_{-\infty}^{\infty} |u_{\text{in}}(y)|^2 dy} \approx \frac{\sum_{\ell=1}^{M} |u_{N,\ell}|^2}{\sum_{\ell=1}^{M} |f_\ell^0|^2} = \frac{||u_{N,\cdot}||_{2,h}^2}{||f^0||_{2,h}^2} \tag{4.34}$$

where we introduced the discrete L_2 norm, defined as

$$||u||_{2,h} = \sqrt{h \sum_i |u_i|^2}, \qquad (1D) \tag{4.35}$$

$$||u||_{2,h} = h\sqrt{\sum_{i,j} |u_{ij}|^2}, \qquad (2D) \tag{4.36}$$

for a grid function u in one and two dimensions, as given by (4.10) and (4.13).

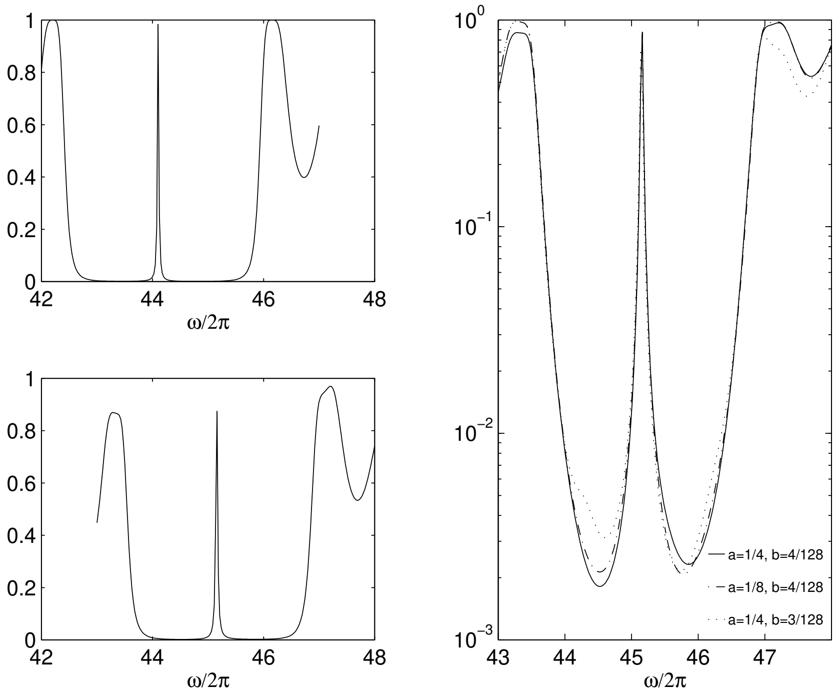


Figure 4.3. Plots of T_{pow} as a function of ω for the one-dimensional (top left) and two-dimensional (bottom left) models. Note the difference in resonance frequency. The same parameters as in Figure 4.2 were used to define the waveguide. In the right picture, results for the two-dimensional model with perturbed a and b are shown.

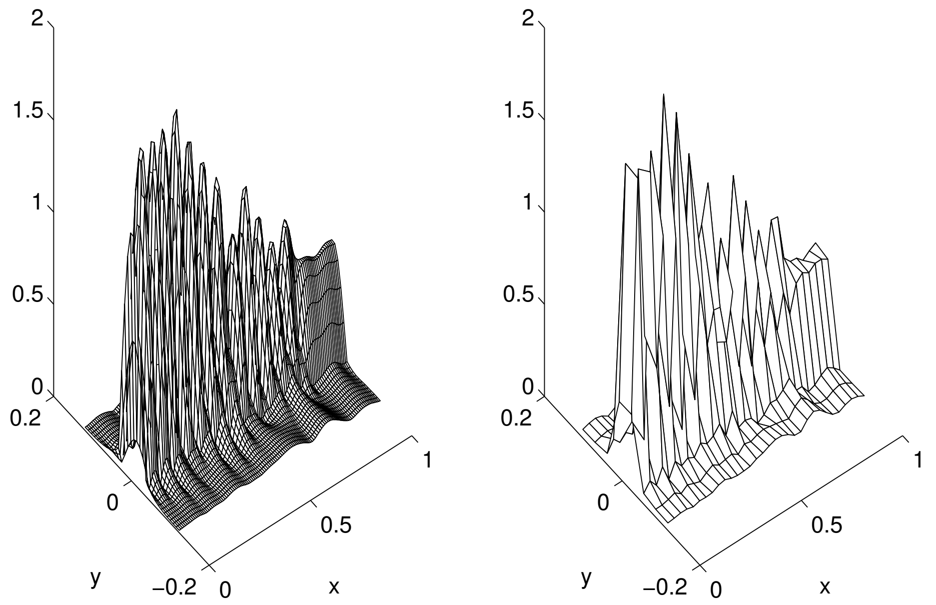


Figure 4.4. Solution of test problem (4.38). Full solution (left) and two steps homogenized solution in both x- and y-directions (right).

4.3.2 Homogenized Solutions

In both one and two dimensions the discretizations in Section 4.2 will lead to a linear system of equations,

$$Lu = f, \tag{4.37}$$

where $u = (u_1, \dots, u_N)^T$ and f is a vector determined by the boundary conditions. We apply the technique described in Section 3.1.3 to these systems and study the result after two recursive homogenization steps.

We will consider the test problem defined by the constants

$$\begin{array}{lllll} w = 1, & a = 3/16, & S = 10, & d = 9/32, & b = 3/16, \\ \Lambda = 1/8, & \tau = 3/128, & \delta = 3/128, & M = 32, & N = 128. \end{array} \tag{4.38}$$

In Figure 4.4 the solution of this problem at frequency $\omega/2\pi = 11$ is shown, together with the solution of the homogenized problem. Homogenization was done two steps in both the x- and y-directions. Note, the homogenized operator was not approximated, so the homogenized solution is just the projection of the full solution onto the coarse subspace. The effect of approximating the operator using the truncation is shown in Figure 4.5. We do both crude truncation according to (3.80) and block banded truncation according to (3.97). In the figure we plot the error between the solution obtained using an operator truncated to a varying number of diagonals and the solution obtained using the full homogenized operator. (To better be able to compare the results we show the number of diagonals, not the ν in (3.80) and (3.97).) We give relative L_2 error and relative error in T_{pow} for both truncation strategies, and conclude that the the block truncation method is better as expected. We also show the strongly diagonal dominant structure of the homogenized operator in Figure 4.6.

Next, we examine how the power transmission spectrum, (4.34), is affected by homogenization. When homogenizing in the y-direction, f_ℓ in (4.34) is replaced by the projection of f_ℓ onto the coarser subspace. The spectrum changes very little when homogenizing in the y-direction. After homogenization in the x-direction, on the other hand, the spectrum is shifted downwards. This can be explained by noting that close to $x = w$, the solution should be approximately a right going wave $f(y)e^{-i\beta_h x}$. What is more,

$$|f(y_\ell)| \approx \sqrt{T_{\text{pow}}}|f_\ell^0|, \qquad \forall \ell \tag{4.39}$$

since the shape of the outgoing wave should be the same as the injected wave. Let $v = (u_{N,1}, \ldots, u_{N,N})^T$. In the case of n homogenization steps in the y-direction,

$$\begin{aligned} T_{\text{pow}}^y &= \frac{||\mathcal{W}_{j-n+1}\cdots\mathcal{W}_j v||_{2,h}^2}{||\mathcal{W}_{j-n+1}\cdots\mathcal{W}_j f^0||_{2,h}^2} \\ &\approx \frac{||\sqrt{T_{\text{pow}}}e^{-i\beta_h w}\mathcal{W}_{j-n+1}\cdots\mathcal{W}_j f^0||_{2,h}^2}{||\mathcal{W}_{j-n+1}\cdots\mathcal{W}_j f^0||_{2,h}^2} = T_{\text{pow}}. \end{aligned} \tag{4.40}$$

Hence, T_{pow} changes little. In the Haar basis, n homogenization steps in the x-direction give

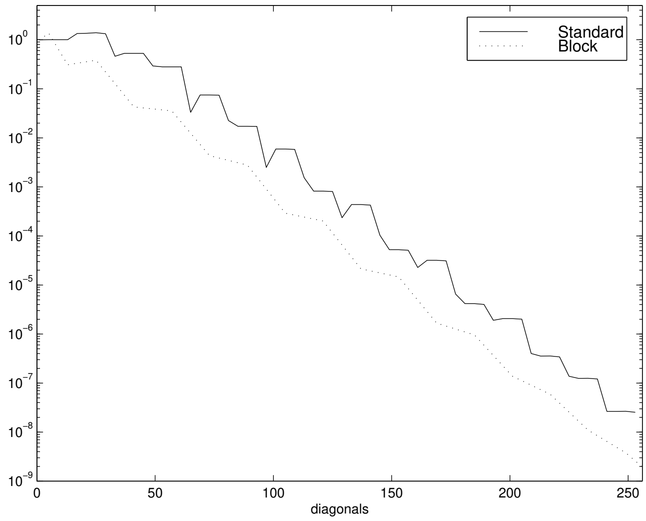
$$\begin{aligned} T_{\text{pow}}^x &= \frac{\sum_{\ell=1}^M |\frac{1}{2^n}\sum_{k=1}^{2^n} u_{N-k+1,\ell}|^2}{||f^0||_{2,h}^2} \approx \frac{\sum_{\ell=1}^M |\frac{f_\ell^0}{2^n}\sum_{k=1}^{2^n} e^{-i\beta_h(w-(k-1)h)}|^2}{||f^0||_{2,h}^2} \\ &= \frac{1}{2^{2n}}\Big|\sum_{k=0}^{2^n-1} e^{i\beta_h kh}\Big|^2 T_{\text{pow}} \equiv \alpha_n(\omega)T_{\text{pow}}. \end{aligned} \tag{4.41}$$

Hence, T_{pow} is scaled by the frequency dependent parameter $\alpha_n(\omega) < 1$. For $\beta_h h \ll 1$ the coefficient $\alpha_n \approx 1 - (2^{2n} - 1)\beta_h^2 h^2/12$.

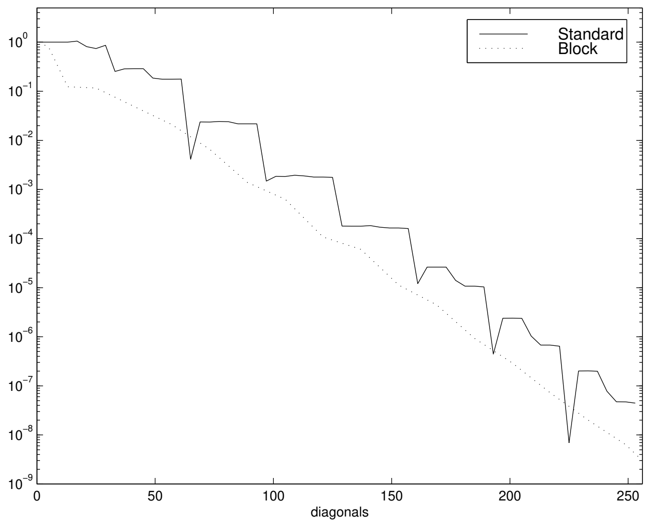
The spectrum is plotted in Figure 4.7. This figure also shows the relative error in T_{pow} incurred by homogenizing two steps in the x-direction, y-direction and both. Here, α_2 was used to correct the computed T_{pow} after homogenization in the x-direction. The error is a few percent, increasing with frequency.

4.3.3 One-Dimensional Models Through Homogenization

The homogenization technique described above can be used to derive one-dimensional models of the waveguide from the two-dimensional model. For this



(a) Relative L_2 error



(b) Relative T_{pow} error

Figure 4.5. Relative error of homogenized solution as a function of numbers of diagonals kept, for the two different truncation strategies.

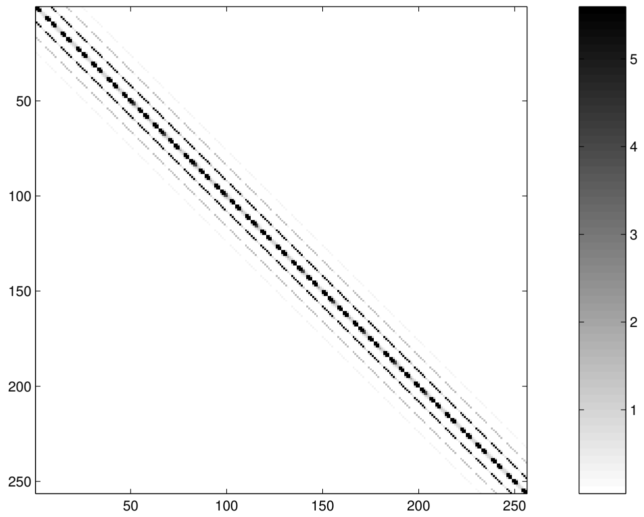


Figure 4.6. Structure of the homogenized operator. Absolute values of elements shown.

purpose the full two-dimensional operator is homogenized multiple steps only in the y-direction. The resulting operator gives the full solution's projection onto a subspace spanned by only a few grid points in the y-direction. Taken to its extreme only one grid point is retained, in which case the homogenized solution is the projection onto the x-axis of the two-dimensional solution.

The operator obtained in this way is truncated and identified with a discretization of a one-dimensional differential operator. This operator will represent a system of equations with a size equaling the number of grid points left in the y-direction. Hence, the extreme case would be a scalar operator. Truncation to a larger bandwidth will correspond to a wider stencil of the discretization and a higher order differential operator.

These one-dimensional models yield an approximation of the solution homogenized to only a few points in the y-direction. In view of the results in Section 4.3.2 this would give a quantitative picture of the full solution and in particular a possibility to compute a good approximation of the power transmission T_{pow}. As was seen in Figure 4.3 the simple one-dimensional model of Section 4.1.1 does not give us the right T_{pow}.

Scalar Model

A scalar model derived through homogenization corresponds to the extreme case of keeping only one grid point in the y-direction. There are problems associated with this case. Since by definition this homogenized operator gives the two-dimensional solution's projection onto the x-axis for all possible initial data, it must take all possible modes of the waveguide into consideration. This includes

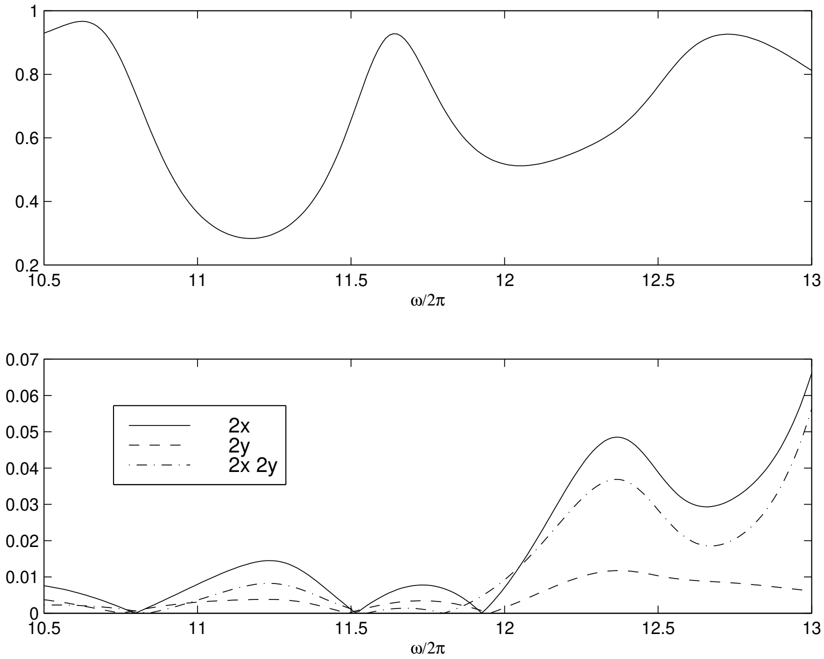


Figure 4.7. Above, spectrum of test problem (4.38). Below, relative error of power transmission spectrum of the test problem in (4.38). Result shown for homogenization two steps in the x-direction, y-direction and both in relation to the result without homogenization.

both propagating modes and damped modes. A one-dimensional model of the type (4.4) can never mimic this behavior, since it cannot have multiple modes. The practical consequence is that the homogenized operator cannot be truncated without a substantial loss of accuracy. It is thus not a local operator in this case.

A natural idea would be to consider an L_2 projection of the two-dimensional Helmholtz operator onto the x-axis. If the projection is onto an eigenfunction of the one-dimensional Helmholtz operator the resulting operator would only contain one mode, and thus be well defined. One could therefore in principle build a wavelet system with a scaling function that approximates such an eigenfunction and expect a better behaved homogenized operator. However, for a large class of wavelet bases, the one point homogenized operator is independent of the choice of basis when we work with finite length functions extended by periodicity. We have the following proposition.

Proposition 4.1. *Let* $\{V_j^{\text{per}}\}$ *and* $\{V_j'^{\text{per}}\}$ *be the scaling spaces corresponding to two different multiresolution analyses on* $L^2(S^1)$*. Suppose their corresponding unperiodized shape functions* φ, φ' *are continuous and decay as* $|\varphi(x)|, |\varphi'(x)| \leq C(1+|x|)^{-2-\epsilon}$ *for some* $\epsilon > 0$*. Then for a surjective* $L \in \mathcal{L}(V_j^{\text{per}}, V_j^{\text{per}})$ *there* *exists an isomorphism* $M : V_j^{\mathrm{per}} \mapsto V_j'^{\mathrm{per}}$ *such that its restriction to* V_0 *is the identity and*

$$\mathcal{H}_{V_j^{\mathrm{per}},V_0^{\mathrm{per}}}L = \mathcal{H}_{V_j'^{\mathrm{per}},V_0'^{\mathrm{per}}}(MLM^{-1}), \qquad \mathcal{R}_{V_j^{\mathrm{per}},V_0^{\mathrm{per}}} = \mathcal{R}_{V_j'^{\mathrm{per}},V_0'^{\mathrm{per}}}M \quad (4.42)$$

whenever the expressions are well defined and the homogenized operators are invertible. Moreover, let $\boldsymbol{V}_j^{\mathrm{per}} = V_j{}^{\mathrm{per}} \otimes V_j{}^{\mathrm{per}}$, $\boldsymbol{V}_0^{\mathrm{per}} = V_0{}^{\mathrm{per}} \otimes V_j{}^{\mathrm{per}}$, $\boldsymbol{V}_j'{}^{\mathrm{per}} = V_j'^{\mathrm{per}} \otimes V_j'^{\mathrm{per}}$, *and* $\boldsymbol{V}_0'{}^{\mathrm{per}} = V_0'^{\mathrm{per}} \otimes V_j'^{\mathrm{per}}$. *Define the isomorphism* $\boldsymbol{M} : \boldsymbol{V}_j^{\mathrm{per}} \mapsto \boldsymbol{V}_j'{}^{\mathrm{per}}$ *by* $\boldsymbol{M} = M \otimes M$ *and denote its restriction to* $\boldsymbol{V}_0^{\mathrm{per}}$ *by* $\boldsymbol{M}_0$. *Then* $\boldsymbol{M}_0 = I \otimes M$ *and if* $L \in \mathcal{L}(\boldsymbol{V}_j^{\mathrm{per}}, \boldsymbol{V}_j^{\mathrm{per}})$ *is surjective,*

$$\begin{aligned} \mathcal{H}_{\boldsymbol{V}_j^{\mathrm{per}},\boldsymbol{V}_0^{\mathrm{per}}}L &= \boldsymbol{M}_0^{-1}\mathcal{H}_{\boldsymbol{V}_j'{}^{\mathrm{per}},\boldsymbol{V}_0'{}^{\mathrm{per}}}(\boldsymbol{M}L\boldsymbol{M}^{-1})\boldsymbol{M}_0, \\ \mathcal{R}_{\boldsymbol{V}_j^{\mathrm{per}},\boldsymbol{V}_0^{\mathrm{per}}} &= \boldsymbol{M}_0^{-1}\mathcal{R}_{\boldsymbol{V}_j'{}^{\mathrm{per}},\boldsymbol{V}_0'{}^{\mathrm{per}}}\boldsymbol{M} \end{aligned} \quad (4.43)$$

whenever the expressions are well defined and the homogenized operators are invertible.

Proof. The periodized shape functions are defined by

$$\varphi_{j,k}^{\mathrm{per}}(x) = \sum_{\ell \in \mathbb{Z}} \varphi_{j,k}(x+\ell) \quad (4.44)$$

where $\varphi_{j,k} = 2^{j/2}\varphi(2^j x - k)$. Under the assumptions on φ in the proposition $\varphi_{0,k}^{\mathrm{per}}(t) = 1$, constant, independent of the basis. (For definitions of periodized multiresolution analysis and proof of this statement, see e.g. Section 9.3 in [18].) It follows that $V_0^{\mathrm{per}} = V_0'^{\mathrm{per}}$. Since V_j^{per} and $V_j'^{\mathrm{per}}$ have the same dimensions there is an isomorphism between them, $M : V_j^{\mathrm{per}} \mapsto V_j'^{\mathrm{per}}$, whose restriction to V_0^{per} is the identity. Let P and P' be the projections onto V_0^{per} from the two spaces. By Lemma 3.4 $P'M = MP$. Now take $f \in V_j^{\mathrm{per}}$. Since L is surjective there is a u such that $Lu = f$. For simplicity, set $\bar{L} = \mathcal{H}_{V_j^{\mathrm{per}},V_0^{\mathrm{per}}}L$, $\bar{L}' = \mathcal{H}_{V_j'^{\mathrm{per}},V_0^{\mathrm{per}}}(MLM^{-1})$, $R = \mathcal{R}_{V_j{}^{\mathrm{per}},V_0^{\mathrm{per}}}$ and $R' = \mathcal{R}_{V_j'^{\mathrm{per}},V_0^{\mathrm{per}}}$. Then by Proposition 3.2

$$\bar{L}Pu = Rf, \qquad \bar{L}'P'Mu = \bar{L}'Pu = R'Mf, \qquad \forall f \in V_j^{\mathrm{per}}. \quad (4.45)$$

It also follows from the definition of $\mathcal{R}$ that if $f \in V_0^{\mathrm{per}}$ we have $Rf = R'f = f = Pf$. Therefore, Lemma 3.5 gives the result (4.42).

For the two-dimensional result take a surjective $L \in \mathcal{L}(\boldsymbol{V}_j, \boldsymbol{V}_j)$ and let $\boldsymbol{P} = P \otimes I$, $\boldsymbol{P}' = P' \otimes I$ and $\boldsymbol{M} = M \otimes M$. Then

$$\boldsymbol{P}'\boldsymbol{M} = (P'M) \otimes M = (MP) \otimes M = \boldsymbol{M}\boldsymbol{P}. \quad (4.46)$$

Here, let $\bar{L} = \mathcal{H}_{\boldsymbol{V}_j^{\mathrm{per}},\boldsymbol{V}_0^{\mathrm{per}}}L$, $\bar{L}' = \mathcal{H}_{\boldsymbol{V}_j'{}^{\mathrm{per}},\boldsymbol{V}_0^{\mathrm{per}}}(\boldsymbol{M}L\boldsymbol{M}^{-1})L$, $R = \mathcal{R}_{\boldsymbol{V}_j{}^{\mathrm{per}},\boldsymbol{V}_0^{\mathrm{per}}}$ and $R' = \mathcal{R}_{\boldsymbol{V}_j'{}^{\mathrm{per}},\boldsymbol{V}_0^{\mathrm{per}}}$. As above assuming $f \in \boldsymbol{V}_j$ there is a u satisfying $Lu = f$. Proposition 3.2 then gives

$$\bar{L}\boldsymbol{P}u = Rf, \qquad \bar{L}'\boldsymbol{P}'\boldsymbol{M}u = \bar{L}'\boldsymbol{M}_0\boldsymbol{P}u = R'\boldsymbol{M}f \quad (4.47)$$

with $\boldsymbol{M}_0 = I \otimes M$. Also, by definition, $Rf = \boldsymbol{P}f$ and $R'f = \boldsymbol{P}'f'$ for $f \in \boldsymbol{V}_0^{\mathrm{per}}$ and $f' \in \boldsymbol{V}_0'{}^{\mathrm{per}}$. Expression (4.43) now follows from Lemma 3.5. $\square$

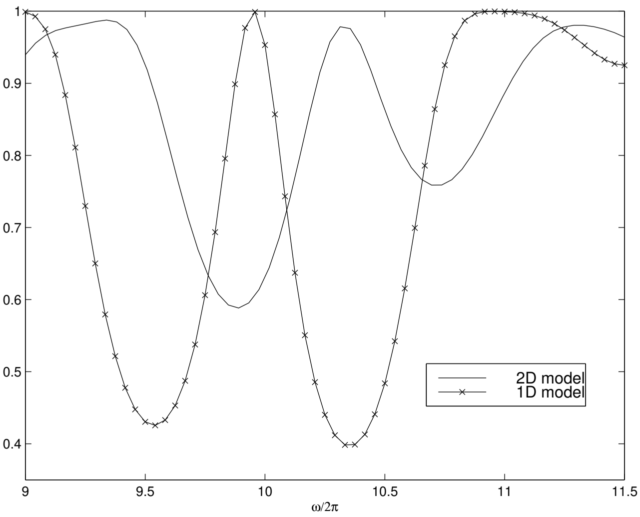


Figure 4.8. Power transmission spectrum of test problem in (4.48) computed with the full two-dimensional model and the one-dimensional model of Section 4.1.1.

System Model

Instead of homogenizing to only one point in the y-direction we could keep a few more points. This gives us a system of one-dimensional equations where the number of unknowns is given by the number of points retained. For this case the result in Proposition 4.1 does not hold and the choice of wavelet basis is important.

We will derive two one-dimensional models for the test problem given by

$$\begin{array}{lllll} w = 1, & a = 3/16, & S = 10, & d = 5/16, & b = 3/16, \\ \Lambda = 1/8, & \tau = 1/64, & \delta = 1/16, & M = 16, & N = 64. \end{array} \tag{4.48}$$

In Figure 4.8 the spectrum for this problem is compared with the one given by the simple one-dimensional model of Section 4.1.1. The simple model does not give the right shape of the spectrum. In particular the resonance peak is at the wrong frequency. We shall see that multipoint models derived through homogenization yield better results. In general there is a trade-off between the number of grid points kept, that is the size of the system, and the number of diagonals needed in the truncated operator to maintain accuracy, which corresponds to the order of the differential operator in the model. The first model keeps four points in the y-direction and hence gives a one-dimensional system of four unknowns. The homogenized operator is identified with a discretization of a second order differential operator. The second model only keeps two unknowns but uses a fourth order differential operator.

The continuous 4×4 system model with the unknowns $u = (u_1, u_2, u_3, u_4) \in \mathbb{C}^4$ is

$$\partial_x(B(x)\partial_x u) + C(x)u + \omega^2 u = 0, \tag{4.49}$$

where $\omega \in \mathbb{R}$ is the same as in the original two-dimensional model and $B(x), C(x) \in \mathbb{R}^{4\times 4}$ are to be determined. The discrete two-dimensional problem is homogenized using Daubechies wavelets with four vanishing moments. The equations and unknowns in the discrete homogenized equation $\bar{L}\bar{u} = \bar{f}$ can be reordered to the blocked form

$$\begin{pmatrix} \bar{L}_{11} & \bar{L}_{12} & \bar{L}_{13} & \bar{L}_{14} \\ \bar{L}_{21} & \bar{L}_{22} & \bar{L}_{23} & \bar{L}_{24} \\ \bar{L}_{31} & \bar{L}_{32} & \bar{L}_{33} & \bar{L}_{34} \\ \bar{L}_{41} & \bar{L}_{42} & \bar{L}_{43} & \bar{L}_{44} \end{pmatrix} \begin{pmatrix} \bar{u}_1 \\ \bar{u}_2 \\ \bar{u}_3 \\ \bar{u}_4 \end{pmatrix} = \begin{pmatrix} \bar{f}_1 \\ \bar{f}_2 \\ \bar{f}_3 \\ \bar{f}_4 \end{pmatrix}. \tag{4.50}$$

Each of the suboperators $\bar{L}_{ij}$ is truncated to tridiagonal form, which essentially corresponds to truncating the full operator $\bar{L}$ to 15-diagonal form. We now identify the unknowns $\bar{u}_j$ and $\bar{f}_j$ with one-dimensional grid functions, $U_i^j \approx u_j(x_i)$ and $F_i^j \approx f_j(x_i)$ for some f_j. Furthermore, the truncated operators $\bar{L}_{ij}$ can be identified with the discretization

$$\frac{1}{h^2}\bar{L}_{ij} = \frac{1}{h^2}\Delta_+ b_{ij}(x_{k-1/2})\Delta_- + \tilde{c}_{ij}(x_k) + \frac{1}{2}\gamma_{ij}(x_k)(E + E^{-1}) + \omega^2 \tag{4.51}$$

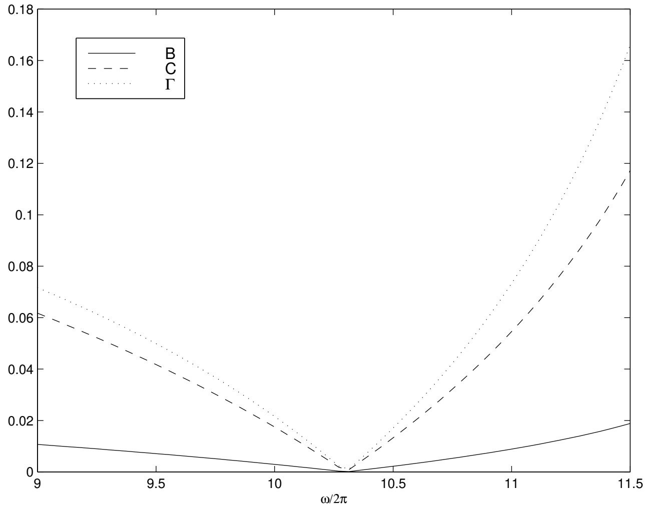
where E is the displacement operator defined by $EU_k = U_{k+1}$. The γ_{ij} terms are needed since $\bar{L}_{ij}$ is in general not symmetric. However, since $\bar{L}_{ii}$ is symmetric (which follows from the fact that L and, by Proposition 3.3, $\bar{L}$ is symmetric) $\gamma_{ii} = 0$. Each $\bar{L}_{ij}$ can be seen as a discretization of a second order differential operator acting on the grid function U^j. Assembling the suboperators by setting $B = \{b_{ij}\}$, $\tilde{C} = \{\tilde{c}_{ij}\}$ and $\Gamma = \{\gamma_{ij}\}$ we get

$$\frac{1}{h^2}\Delta_+ B(x_{k-1/2})\Delta_- U_k + \tilde{C}(x_k)U_k + \frac{1}{2}\Gamma(x_k)(U_{k-1} + U_{k+1}) + \omega^2 U_k = F_k. \tag{4.52}$$

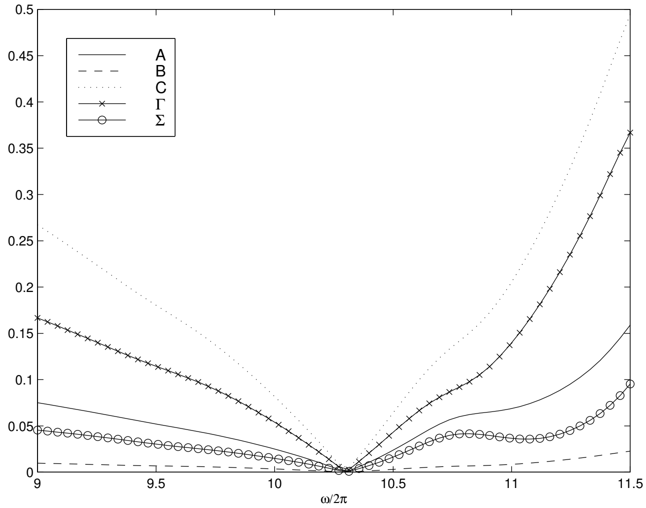
where we also assembled the grid functions $U = \{U^j\}$, $F = \{F^j\}$, If $C(x_k) = \tilde{C}(x_k) + \Gamma(x_k)$ and $B(x_k)$ were real and independent of ω and if $F = 0$ this could be identified with a second order discretization of (4.49). The computed coefficients are in fact satisfying these requirements to a fairly good level of approximation. Let $\omega_0/2\pi = 10.3$. In Figure 4.9a the frequency dependency of the coefficients is illustrated. The coefficients are computed for a range of ω and the relative difference, measured in the discrete L_1-norm, compared to the coefficients computed at $\omega = \omega_0$, is plotted as a function of ω. The $B(x)$ coefficients vary on the order of only 1% in the interval.

In Figure 4.10 we show the spectrum computed in four different ways, and compare it with the spectrum for the two-dimensional model. The four methods are:

A) Recompute coefficients for each new frequency,

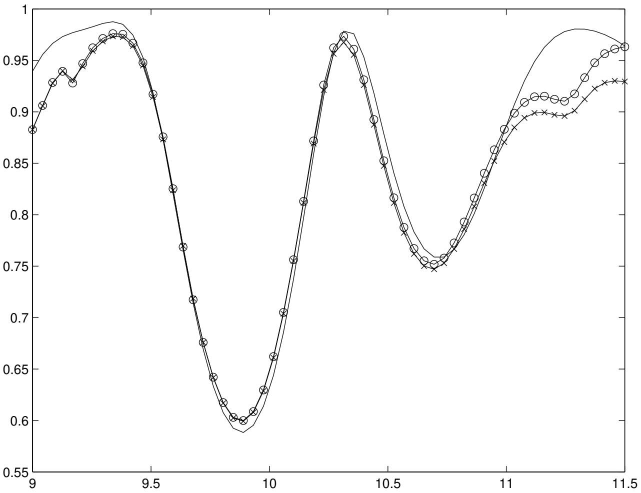


(a) 4×4 model

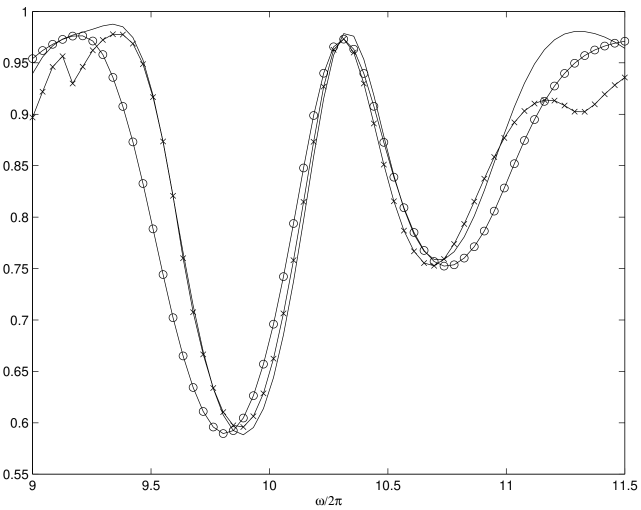


(b) 2×2 model

Figure 4.9. Frequency dependency of coefficients. Relative L_1 difference between coefficients computed at ω and at $\omega_0/2\pi = 10.3$.



(a) Method A (crosses) and B (circles).



(b) Method C (crosses) and D (circles).

Figure 4.10. Power transmission spectrum of test problem (4.48) computed using the 4×4 model and the four different ways (A-D) given in the text. Solid line without marker shows spectrum computed with the full two-dimensional model.

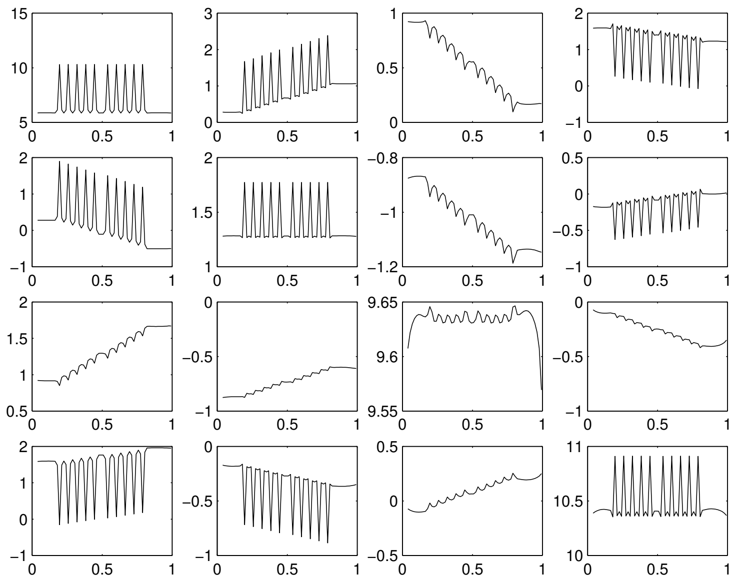


Figure 4.11. The components of the coefficient matrix $B(x)$ (real part) at $\omega_0/2\pi = 10.3$.

B) As in A), but discard imaginary part of all coefficients and set right hand side to zero,

C) As in B), except $B(x)$ is computed at $\omega = \omega_0$ and kept constant for all ω,

D) All coefficients are computed at $\omega = \omega_0$ and kept constant for all ω.

The first and last rows of each $\bar{L}_{ij}$ and $\bar{f}_j$ corresponds to the boundary conditions of the operator. To get the right boundary condition we have simply left the first and last two rows of each (truncated) $\bar{L}_{ij}$ and f_j intact in all methods A-D.

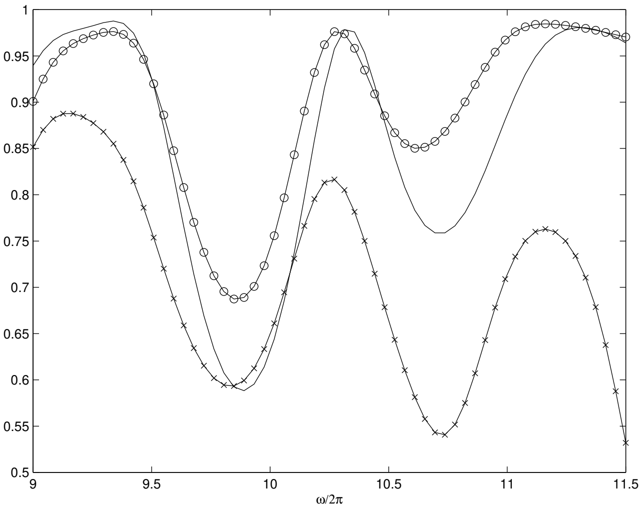
As seen in Figure 4.10 the model captures the true spectrum to a fairly high degree, even when all coefficients are kept constant in method D. The real part of the coefficient matrix $B(x)$ computed at $\omega = \omega_0$ is plotted in Figure 4.11. Note how most components of the matrix follow the geometric variation of the gratings, but that some, surprisingly, are almost constant.

The second model is a 2×2 system with the unknowns $u = (u_1, u_2) \in \mathbb{C}^2$ and

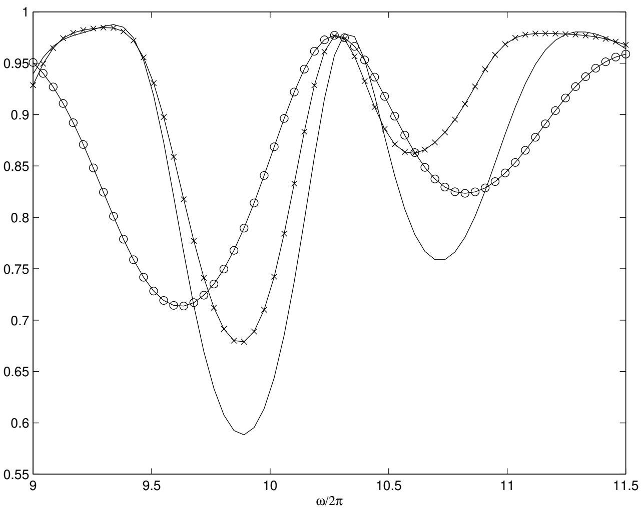
$$\partial_{xx}(A(x)\partial_{xx}u) + \partial_x(B(x)\partial_x u) + C(x)u + \omega^2 u = 0. \tag{4.53}$$

Again, ω is the same as in the original equation and $A(x), B(x), C(x) \in \mathbb{R}^{2\times 2}$. This time we use Daubechies wavelets with three vanishing moments. The homogenized equation is reordered and decomposed into the blocks

$$\begin{pmatrix} \bar{L}_{11} & \bar{L}_{12} \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix} \begin{pmatrix} \bar{u}_1 \\ \bar{u}_2 \end{pmatrix} = \begin{pmatrix} \bar{f}_1 \\ \bar{f}_2 \end{pmatrix}. \tag{4.54}$$



(a) Method A (crosses) and B (circles).



(b) Method C (crosses) and D (circles).

Figure 4.12. Power transmission spectrum of test problem (4.48) computed using the 2×2 model and the four ways (A-D) given in the text. Solid line without marker shows spectrum computed with the full two-dimensional model.

The suboperators $\bar{L}_{ij}$ are truncated to five diagonals, corresponding to an 11-diagonal $\bar{L}$. Like in the 4×4 model we identify the unknowns and right hand side with grid functions U^j, F^j. The suboperators are this time identified with the discretization,

$$\begin{aligned}\frac{1}{h^2}\bar{L}_{ij} &= \frac{1}{h^4}\Delta_+\Delta_- a_{ij}(x_k)\Delta_+\Delta_- + \frac{1}{h^2}\Delta_+ b_{ij}(x_{k-1/2})\Delta_- + \tilde{c}_{ij}(x_k) \\ &\quad + \frac{1}{2}\sigma_{ij}(x_k)(E^2 + E^{-2}) + \frac{1}{2}\gamma_{ij}(x_k)(E + E^{-1}) + \omega^2. \qquad (4.55)\end{aligned}$$

Again γ_{ij}, σ_{ij} is needed since $\bar{L}_{ij}$ is not always symmetric, but $\gamma_{ii} = \sigma_{ii} = 0$. Setting $A = \{a_{ij}\}$, $B = \{b_{ij}\}$, $\tilde{C} = \{\tilde{c}_{ij}\}$, $\Gamma = \{\gamma_{ij}\}$ and $\Sigma = \{\sigma_{ij}\}$ gives

$$\begin{aligned}&\tfrac{1}{h^4}\Delta_+\Delta_- A(x_k)\Delta_+\Delta_- U_k + \tfrac{1}{h^2}\Delta_+ B(x_{k-1/2})\Delta_- U_k + \tilde{C}(x_k)U_k \\ &+ \tfrac{1}{2}\Sigma(x_k)(U_{k-2} + U_{k+2}) + \tfrac{1}{2}\Gamma(x_k)(U_{k-1} + U_{k+1}) + \omega^2 U_k = F_k. \qquad (4.56)\end{aligned}$$

The frequency dependency of the coefficients of this model is higher than for the 4×4 model as can be seen in Figure 4.9b. The coefficients $B(x)$ and $A(x)$ however still vary very little. The same four methods for computing the spectrum as above are used also for this model, except that for method C both coefficients $A(x)$ and $B(x)$ are held constant. For the boundary conditions we kept the three first and last rows of each (truncated) $\bar{L}_{ij}$ and f_j intact. The result is in Figure 4.12. In this case discarding the imaginary part of the coefficients proves to be important. Because of the rather large relative frequency dependency of $C(x)$ in this case, the method D gives a dilated spectrum.

4.3.4 The Homogenized Operator as a Subgrid Model

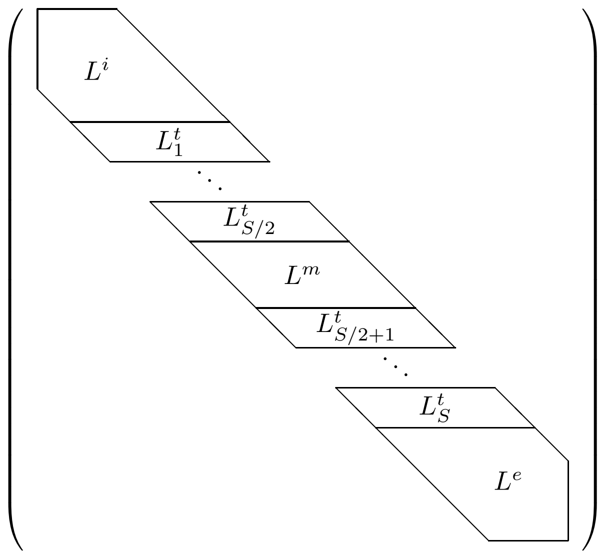
In this section we will demonstrate how the homogenization technique can be used to compute approximate subgrid models for small geometrical details in a problem. The aim is then to use the subgrid model to solve a large problem on a coarse mesh that does not resolve the small scales of the geometry. This can be seen as an analogy to solving the cell problem in classical homogenization and using the computed effective material coefficient in the homogenized equation.

Our target problem is the same as in (4.38), only that we extend the waveguide by increasing the length w and the number of gratings S. We use the same frequency as before, $\omega/2\pi = 11$. In order to solve this problem on a coarse grid, we would need the homogenized operator and right hand side. The key observation here is that a good approximation of these can be obtained *directly from a much smaller problem*. We will now show how to do that.

In the waveguide example we identify four distinct parts of the problem: the initial part, a grating tooth, the quarter-wave step in the middle and the end part. Let $\bar{u}^i$, $\bar{u}^m$ and $\bar{u}^e$ denote the solution at grid values in the initial segment, middle segment and end segment of the waveguide. Moreover, let $\bar{u}_k^t$ be the solution at tooth k. Then we can write $\bar{L}\bar{u} = \bar{F}$ as

$$(\tilde{L} + \delta L) \begin{pmatrix} \bar{u}^i \\ \bar{u}_1^t \\ \vdots \\ \bar{u}_{S/2}^t \\ \bar{u}^m \\ \bar{u}_{S/2+1}^t \\ \vdots \\ \bar{u}_S^t \\ \bar{u}^e \end{pmatrix} = \begin{pmatrix} \bar{F}^i \\ \bar{F}_1^t \\ \vdots \\ \bar{F}_{S/2}^t \\ \bar{F}^m \\ \bar{F}_{S/2+1}^t \\ \vdots \\ \bar{F}_S^t \\ \bar{F}^e \end{pmatrix} \tag{4.57}$$

where $\bar{F}$ follows the same partioning as $\bar{u}$ and

$\tilde{L} =$  (4.58)

with elements being zero outside the delineated areas. The matrix δL is the residual between $\bar{L}$ and $\tilde{L}$, with all elements zero within in the delineated areas.

With this notation we now observe that since $\bar{L}$ is essentially a local operator, δL will be small. What is more,

$$L_k^t \approx L_\ell^t, \qquad \bar{F}_k^t \approx \bar{F}_\ell^t, \qquad \forall k, \ell. \tag{4.59}$$

We will interpret L^i, L^t, L^m and L^e together with $\bar{F}^i$, $\bar{F}^t$, $\bar{F}^m$ and $\bar{F}^e$ as the subgrid models for the corresponding parts of the geometry. (Since L_k^t and $\bar{F}_k^t$ are approximately the same for all k we will henceforth drop the subindices.) This is a slightly simplified interpretation compared to the one in Chapter 3, since we do not untangle the equivalent of the H matrix, which we do not need here.

The strategy for constructing an approximation of the homogenized operator of the full problem is then clear. First, we compute an approximation of L^i, L^t, L^m and L^e by homogenizing a much smaller problem. At the same time we also obtain approximations of $\bar{F}^i$, $\bar{F}^t$, $\bar{F}^m$ and $\bar{F}^e$. Second, we assemble the subgrid models according to the block structures of (4.57) and (4.58) to obtain an approximation of $\tilde{L}$ and $\bar{F}$. Finally, we neglect δL.

In our example we use (4.38) with $w = 1$ and $S = 10$ as our small problem. This is a well resolved problem with eight gridpoints between the gratings and approximately 15 gridpoints per effective wavelength in the x-direction ($\beta_h/2\pi \approx 8.75$). We homogenize the operator three steps in the y-direction and one step in the x-direction, corresponding to a grid size of $M = 4$ and $N = 64$. This gives a very coarse resolution in the y-direction, in particular no resolution of the gratings, and a reasonable resolution in the x-direction. This grid size should be enough to represent the solution in a quantitatively correct way.

For the large problems we denote the corresponding parameters with a prime sign. We will have $w' > 1$ and $S' > 10$ and we set $L^{i'} = L_i$, $L^{m'} = L^m$, $L^{e'} = L^e$ and

$$L_k^{t\,\prime} = \begin{cases} L_k^t & 1 \le k \le 2, \\ L_3^t & 3 \le k \le S'/2 - 2, \\ L_{k+(S-S')/2}^t & S'/2 - 1 \le k \le S'/2 + 2, \\ L_{S/2+3}^t & S'/2 + 3 \le k \le S' - 2, \\ L_{k+S-S'}^t & S' - 1 \le k \le S'. \end{cases} \tag{4.60}$$

So far we have deliberately been vague about the exact size of the areas in (4.58), and hence the size of the subgrid models. We now introduce a parameter ν, the band width of $\tilde{L}$, with which we can define all sizes at the same time. The sizes, i.e the value of ν, will determine the accuracy of the solution. The results of this test is shown in Figure 4.13, where the relative L_2-error between the solution computed using the subgrid technique and the exact solution, projected onto the coarse subspace, is plotted as a function of ν, for five different problem sizes. Our "small" problem, $w = 1, S = 10$ is also included, for comparison, in which case ν just indicates the bandwidth of the truncated operator. The error stays small even for a problem 16 times as large as the small problem.

There are two different error sources for this technique. Let $\delta\tilde{L}$ be the difference between the exact $\tilde{L}$ and the one constructed using subgrid models. Then the difference between the exact homogenized operator $\bar{L}'$ and the approximation, $\bar{L}'_{app}$, is

$$\bar{L}' - \bar{L}'_{app} = \tilde{L} + \delta L - (\tilde{L} - \delta\tilde{L}) = \delta L + \delta\tilde{L}. \tag{4.61}$$

The error attributed to δL will follow the same pattern as in Section 4.3.2 for truncated operators, while the error ascribed to $\delta\tilde{L}$ is independent of ν. In Figure 4.13 we see that all error curves, except the one for the small problem where $\delta\tilde{L}$ is zero, level out after a certain ν value. This point, ν^*, is where the effect of $\delta\tilde{L}$ starts to dominate. From an accuracy point of view it is therefore meaningless to use $\nu > \nu^*$. In this case $\nu^* \approx 45$ independently of the problem

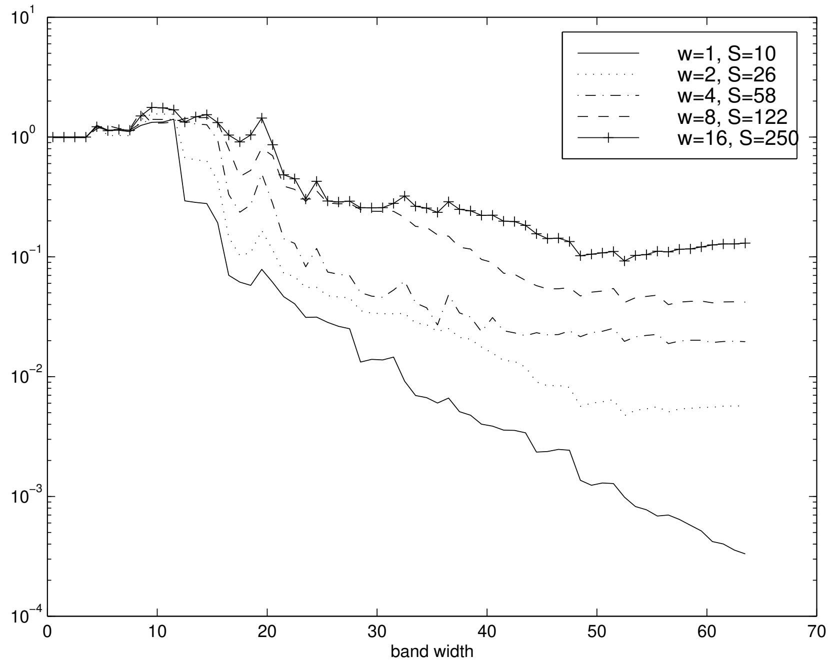


Figure 4.13. Result of the subgrid technique. Relative L_2-error of solution as a function of the bandwidth.

size. Also note that already with $\nu = 22$ a good approximation of the solution is obtained even for the largest problem, see Figure 4.14

There can be a big computational gain using the subgrid technique. We compare it to solving the large problem on a fine grid. Let N' and M' denote the grid size of the large problem. Then, since the band width of the exact operator is M', we get an approximate cost of $cN'M'^3$ flops, using a direct solver. Suppose our coarse grid is of the size $N'/2^n$ and $M'/2^m$, corresponding to homogenization n times in the x-direction and m times in the y-direction. Let $r = 2^{-m-n}$. The direct solution using the subgrid model then costs $crN'M'\nu^2$ giving a ratio between the costs

$$r\left(\frac{\nu}{M'}\right)^2, \tag{4.62}$$

In our largest case $M' = 32$, $N' = 2048$, $m = 3$, $n = 1$ and with $\nu = \nu^* = 45$, we get an approximate ratio of $1/8$.

We also need to compute the subgrid model, although this only needs to be done once. Suppose M,N is the size of the "small" problem. A direct computa-

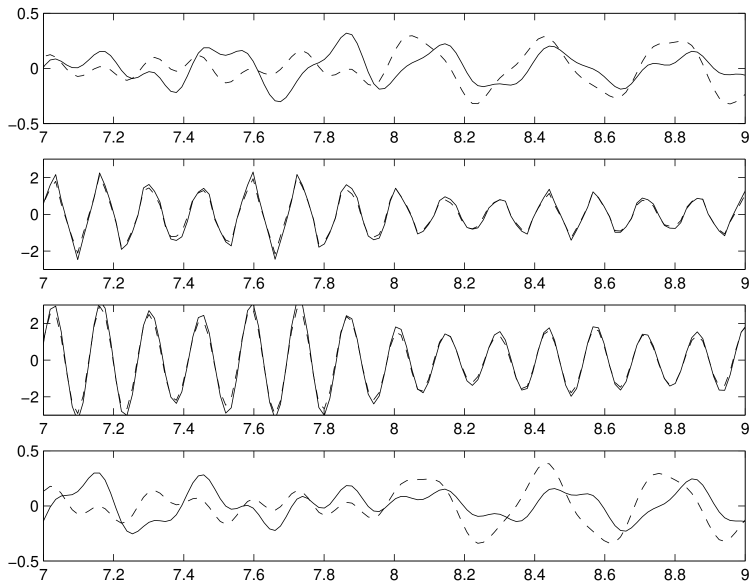


Figure 4.14. Result of the large $w = 16$ problem approximated with $\nu = 22$. Plot of exact solution (solid) and approximate solution (dashed) as a function of x for four y-values corresponding to the four grid points in the y-direction of the coarse mesh. Real part of u shown in the x-interval $[7, 9]$.

tion of the exact homogenized operator would then start by evaluating $CA^{-1}B$ where

$$A \in \mathbb{R}^{(1-r)MN\times(1-r)MN}, \qquad B \in \mathbb{R}^{(1-r)MN\times rMN}, \\ C \in \mathbb{R}^{rMN\times(1-r)MN}. \tag{4.63}$$

The bandwidth of A can be shown to be $2^n(1-r)M$. Hence, it can be LU-decomposed to a cost of $c2^{2n}(1-r)^3NM^3$. The matrices L and U will have the same bandwidth as A. Computing $L^{-1}B$ will then cost

$$c(1-r)NM2^n(1-r)MrMN/2 = c2^{2n-1}r(1-r)^2N^2M^3 \tag{4.64}$$

flops. The factor $1/2$ stems from the fact that also B is sparse. The computation of $U^{-1}L^{-1}B$ has the same cost. In all the subgrid model costs

$$c2^n(rN + 2^n(1-r))(1-r)^2NM^3. \tag{4.65}$$

The ratio to the cost of solving the large problem is

$$2^n(rN + 2^n(1-r))(1-r)^2\frac{N}{N'}\left(\frac{M}{M'}\right)^3 \approx (\frac{N}{2^m} + 2^{2n})\frac{N}{N'}\left(\frac{M}{M'}\right)^3. \tag{4.66}$$

This ratio may not always be small. However, in cases where the subgrid model is reused in multiple computations (as in our case where we used it in all five computations) the cost of computing the subgrid model matters less. More importantly, in earlier work, [8] and [19], it was shown that for many practical cases the matrix A and the resulting subgrid model will be strongly diagonal dominant, with an exponential decay of the elements off the diagonal. A less naive approach than the direct computation given above, involving for instance incomplete LU-decomposition, would allow us to bypass the exact subgrid model and to obtain the small bandwidth approximation directly at a much lower cost.

Appendix A

Evaluating the Flux Functions

A.1 Delta Function Case

We will here show how to solve the system of equations (2.185) analytically, when $N = 2$, for two different choices of moments. We use the ansatz

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

with the variables a, b and c defined as

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

so that

$$g_1 = \frac{1}{2}(a + b\cos\phi), \qquad g_2 = \frac{1}{2}(a - b\cos\phi). \tag{A.3}$$

A.1.1 The m_{30}, m_{03} Case

For the choice of moments in (2.42), the nonlinear system of equations (2.185) reads

$$\begin{aligned} u_1 + u_3 &= m_{10}, \\ u_2 + u_4 &= m_{01}, \\ \tfrac{u_1^3}{u_1^2+u_2^2} + \tfrac{u_3^3}{u_3^2+u_4^2} &= m_{30}, \\ \tfrac{u_2^3}{u_1^2+u_2^2} + \tfrac{u_4^3}{u_3^2+u_4^2} &= m_{03}. \end{aligned} \tag{A.4}$$

We observe that

$$\begin{array}{rcl} m_{30} & = & m_{10}\cos^2\alpha - R\cos(\beta+3\alpha)(1+\cos\beta), \\ m_{03} & = & m_{01}\sin^2\alpha + R\sin(\beta+3\alpha)(1+\cos\beta). \end{array} \tag{A.5}$$

Here,

$$\tan\alpha = \frac{m_{01}}{m_{10}}, \qquad \beta = \theta_1+\theta_2-2\alpha, \qquad R = \frac{b}{2}(1-\frac{b^2}{a^2}), \tag{A.6}$$

with a to be determined. The values of β and R follow from (A.5),

$$\tan(\beta+3\alpha) = -\frac{m_{03}-m_{01}\sin^2\alpha}{m_{30}-m_{10}\cos^2\alpha}, \tag{A.7}$$

$$R^2(1+\cos\beta)^2 = (m_{30}-m_{10}\cos^2\alpha)^2 + (m_{03}-m_{01}\sin^2\alpha)^2. \tag{A.8}$$

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

$$\cos(\beta+2\alpha) = \cos(\theta_1+\theta_2) = \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 = \frac{u_1u_3}{g_1g_2} - \frac{u_2u_4}{g_1g_2}. \tag{A.9}$$

Multiplying (A.9) by g_1g_2 and inserting (A.1) and (A.3) yields an equation of the form

$$q_1\cos(2\phi) + q_2\sin(2\phi) = q_3, \tag{A.10}$$

which can be solved exactly. The coefficients are

$$\begin{array}{rcl} q_1 & = & \frac{1}{b^2}(a^2+c^2)(m_{10}^2-m_{01}^2) - b^2\cos(\beta+2\alpha), \\ q_2 & = & -\frac{4ac}{b^2}m_{10}m_{01}, \\ q_3 & = & (b^2-2a^2)\cos(\beta+2\alpha) + m_{10}^2 - m_{01}^2. \end{array} \tag{A.11}$$

A.1.2 The m_{20}, m_{02} Case

If we choose the moments in (2.82) the equations (2.185) will be

$$\begin{array}{rcl} u_1+u_3 & = & m_{10}, \\ u_2+u_4 & = & m_{01}, \\ \frac{u_1^2}{\sqrt{u_1^2+u_2^2}} + \frac{u_3^2}{\sqrt{u_3^2+u_4^2}} & = & m_{20}, \\ \frac{u_2^2}{\sqrt{u_1^2+u_2^2}} + \frac{u_4^2}{\sqrt{u_3^2+u_4^2}} & = & m_{02}. \end{array} \tag{A.12}$$

We get immediately that

$$a = m_{20} + m_{02}. \tag{A.13}$$

We have left to solve the angle ϕ. To do this, we use equation three of (A.12),

$$\frac{u_1^2}{g_1} + \frac{u_3^2}{g_2} = m_{20}. \tag{A.14}$$

Like in Appendix A.1.1 we multiply (A.14) by $g_1 g_2$ and insert (A.1) and (A.3) to get an equation of the form

$$r_1 \cos(2\phi) + r_2 \sin(2\phi) = r_3. \tag{A.15}$$

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

$$\begin{array}{rcl} r_1 & = & \frac{a}{2b^2}\left((a^2-2b^2)m_{10}^2 - c^2 m_{01}^2\right) + \frac{b^2}{2} m_{20}, \\ r_2 & = & -\frac{c^3}{b^2} m_{10} m_{01}, \\ r_3 & = & \frac{1}{2}(2a^2-b^2)m_{20} - \frac{a}{2b^2}(a^2 m_{10}^2 + c^2 m_{01}^2). \end{array} \tag{A.16}$$

A.2 Heaviside Function Case

We show here how to solve the equation $\boldsymbol{F}_1(\boldsymbol{u}) = \boldsymbol{d}$ analytically when $N = 1, \ldots, 4$. The solution is given by solving two polynomial equations of degree $N/2$.

A.2.1 $N = 1$

$$u_1 = m_{10} \tag{A.17}$$

A.2.2 $N = 2$

The system is

$$-u_1 + u_2 = m_{10}, \tag{A.18}$$

$$\frac{1}{2}(-u_1^2 + u_2^2) = m_{11}, \tag{A.19}$$

and the solution is

$$u_1 = \frac{m_{11}}{m_{10}} - \frac{m_{10}}{2}, \tag{A.20}$$

$$u_2 = \frac{m_{11}}{m_{10}} + \frac{m_{10}}{2}. \tag{A.21}$$

A.2.3 $N = 3$

The system is

$$-u_1 + u_2 - u_3 = m_{10}, \tag{A.22}$$

$$\frac{1}{2}(-u_1^2 + u_2^2 - u_3^2) = m_{11}, \tag{A.23}$$

$$\frac{1}{3}(-u_1^3 + u_2^3 - u_3^3) = m_{12}, \tag{A.24}$$

and the solution is given as roots to the polynomials

$$\begin{aligned} (6{m_{10}}^2+12m_{11})u^2+(4{m_{10}}^3-12m_{12})u & & \\ +{m_{10}}^4+12{m_{11}}^2-12m_{12}m_{10} &= 0, \quad (u_1,u_3) \\ 3({m_{10}}^2+2m_{11})u-{m_{10}}^3-6m_{12}-6m_{10}m_{11} &= 0, \quad (u_2) \end{aligned} \tag{A.25}$$

A.2.4 $N=4$

The system is

$$-u_1+u_2-u_3+u_4 = m_{10}, \tag{A.26}$$

$$\frac{1}{2}(-u_1^2+u_2^2-u_3^2+u_4^2) = m_{11}, \tag{A.27}$$

$$\frac{1}{3}(-u_1^3+u_2^3-u_3^3+u_4^3) = m_{12}, \tag{A.28}$$

$$\frac{1}{4}(-u_1^4+u_2^4-u_3^4+u_4^4) = m_{13}. \tag{A.29}$$

The solution of u_1 and u_3 is given as the roots of the following polynomial

$$\begin{aligned} &(144\,m_{11}^2+12\,m_{10}^4-144\,m_{12}\,m_{10})u^2+ \\ &(-24\,m_{11}\,m_{10}^3+72\,m_{11}^2\,m_{10} \\ &-144\,m_{11}\,m_{12}+6\,m_{10}^5-72\,m_{12}\,m_{10}^2+144\,m_{13}\,m_{10})u \\ &+72\,m_{11}^3+36\,m_{11}^2\,m_{10}^2+72\,m_{13}\,m_{10}^2-24\,m_{12}\,m_{10}^3+m_{10}^6 \\ &-144\,m_{11}\,m_{12}\,m_{10}-144\,m_{11}\,m_{13}-6\,m_{11}\,m_{10}^4+144\,m_{12}^2 \quad = \quad 0. \end{aligned} \tag{A.30}$$

Similarly, u_2 and u_4 are given by

$$\begin{aligned} &(144\,m_{11}^2+12\,m_{10}^4-144\,m_{12}\,m_{10})u^2+ \\ &(-24\,m_{11}\,m_{10}^3-72\,m_{11}^2\,m_{10} \\ &-144\,m_{11}\,m_{12}-6\,m_{10}^5+72\,m_{12}\,m_{10}^2+144\,m_{13}\,m_{10})u \\ &-72\,m_{11}^3+36\,m_{11}^2\,m_{10}^2-72\,m_{13}\,m_{10}^2-24\,m_{12}\,m_{10}^3+m_{10}^6 \\ &+144\,m_{11}\,m_{12}\,m_{10}-144\,m_{11}\,m_{13}+6\,m_{11}\,m_{10}^4+144\,m_{12}^2 \quad = \quad 0. \end{aligned} \tag{A.31}$$

Bibliography

[1] J. Anderson. *Computational Fluid Dynamics, The Basics with Applications.* McGraw-Hill, 1995.

[2] O. Axelsson. *Iterative Solution Methods.* Cambridge University Press, 1994.

[3] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978.

[4] J.-D. Benamou. Big ray tracing: Multivalued travel time field computation using viscosity solutions of the eikonal equation. *J. Comp. Physics*, 128(4):463–474, 1996.

[5] J.-D. Benamou. Multivalued solution and viscosity solutions of the eikonal equation. Research report 3281, Projet Ondes, INRIA Rocquencourt, 1997.

[6] A. Bensoussan, J.-L. Lions, and G. Papanicolau. *Asymptotic Analysis for Periodic Structures.* North-Holland Publ. Comp., The Netherlands, 1978.

[7] M. Berger and R. LeVeque. Cartesian meshes and adaptive refinement for hyperbolic partial differential equations. In *Third International Conference on Hyperbolic Problems, Vol. I, II (Uppsala, 1990)*, pages 67–73. Studentlitteratur, Lund, 1991.

[8] G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Comm. Pure Appl. Math.*, 44:141–183, 1991.

[9] F. Bouchut. On zero pressure gas dynamics. In *Advances in kinetic theory and computing*, volume 22 of *Ser. Adv. Math. Appl. Sci.*, pages 171–190. World Sci. Publishing, River Edge, NJ, 1994.

[10] F. Bouchut and F. James. Équations de transport unidimensionnelles à coefficients discontinus. *C. R. Acad. Sci. Paris Sér. I Math.*, 320(9):1097–1102, 1995.

[11] F. Bouchut and F. James. Duality solutions for pressureless gases, monotone scalar conservation laws and uniqueness. *Comm. Partial Differential Equations*, 1998. To appear.

[12] Y. Brenier and L. Corrias. A kinetic formulation for multibranch entropy solutions of scalar conservation laws. *Ann. Inst. Henri Poincaré*, 15(2):169–190, 1998.

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

[14] Y. Brenier and E. Grenier. Sticky particles and scalar conservation laws. *SIAM J. Num. Analysis*, 1998. To appear.

[15] M. Brewster and G. Beylkin. A multiresolution strategy for numerical homogenization. *Applied and Computational Harmonic Analysis*, 2:327–349, 1995.

[16] T. Chan and T. Mathew. The interface probing technique in domain decomposition. *SIAM J. Matrix Anal. Appl.*, 13(1):212–238, January 1992.

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

[18] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, 1991.

[19] M. Dorobantu and B. Engquist. Wavelet-based numerical homogenization. *SIAM J. on Num. Anal.*, 35(2):540–559, April 1998.

[20] L. Durlofsky. Numerical calculation of equivalent grid block permeability tensors for heterogeneous porous media. *Water Resour. Res.*, 27:699–708, 1991.

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

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

[23] B. Engquist and E. Luo. New coarse grid operators for highly oscillatory coefficient elliptic problems. *J. Comput. Phys.*, 129(2):296–306, 1996.

[24] P. Gérard. Microlocal defect measures. *Comm. Partial Differential Equations*, 16(11):1761–1794, 1991.

[25] P. Gérard, P. A. Markowich, N. J. Mauser, and F. Poupaud. Homogenization limits and Wigner transforms. *Comm. Pure Appl. Math.*, 50(4):323–379, 1997.

[26] A. C. Gilbert. A comparison of multiresolution and classical one-dimensional homogenization schemes. *Appl. Comput. Harmon. Anal.*, 5(1):1–35, 1998.

[27] H. Grad. On the kinetic theory of rarefied gases. *Comm. Pure Appl. Math.*, 2:331–407, 1949.

[28] E. Grenier. Existence globale pour le système des gaz sans pression. *C. R. Acad. Sci. Paris Sér. I Math.*, 321(2):171–174, 1995.

[29] H. A. Haus and Y. Lai. Narrow-band optical channel-dropping filter. *J. Lightwave Technol.*, 10(1):57–61, 1992.

[30] W. D. Henshaw, H. Kreiss, and L. G. Reyna. On the smallest scale for the incompressible Navier-Stokes equations. *Theoretical and Computational Fluid Dynamics*, 1:65–95, 1989.

[31] T. Y. Hou and X. H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.

[32] F. James and L. Gosse. Numerical approximations of one-dimensional linear conservation equations with discontinuous coefficients. Preprint 98.6, FORTH/IACM, Heraklion, Crete, July 1998.

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

[34] J. Keller. Geometrical theory of diffraction. *J. Optical Society of America*, 52, 1962.

[35] S. Knapek. Matrix-dependent multigrid-homogenization for diffusion problems. *SIAM J. Sci. Stat. Comp.*, 1998. To appear.

[36] R. G. Kouyoumjian and P. H. Pathak. A uniform theory of diffraction for an edge in a perfectly conducting surface. *Proc. IEEE*, 62(11):1448–1461, 1974.

[37] Yu. A. Kravtsov. On a modification of the geometrical optics method. *Izv. VUZ Radiofiz.*, 7(4):664–673, 1964.

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

[39] C. D. Levermore. Moment closure hierarchies for kinetic theories. *J. Statist. Phys.*, 83(5-6):1021–1065, 1996.

[40] M. Levy, L. Eldata, R. Scarmozzino, R. M. Osgood Jr, P. S. D. Lin, and F. Tong. Fabrication of narrow-band channel-dropping filters. *IEEE Photonics Technology Letters*, 4(12):1378–1381, 1992.

[41] Z. Li. *The Immersed Interface Method—A Numerical Approach for Partial Differential Equations with Interfaces.* PhD thesis, Department of Applied Mathematics, Univ. of Washington, 1994.

[42] P.-L. Lions and T. Paul. Sur les mesures de Wigner. *Rev. Mat. Iberoamericana*, 9(3):553–618, 1993.

[43] P.-L. Lions, B. Perthame, and E. Tadmor. A kinetic formulation of multidimensional scalar conservation laws and related equations. *J. Amer. Math. Soc.*, 7(1):169–191, 1994.

[44] P.-L. Lions, B. Perthame, and E. Tadmor. Kinetic formulation of the isentropic gas dynamics and p-systems. *Comm. Math. Phys.*, 163(2):415–431, 1994.

[45] D. Ludwig. Uniform asymptotic expansions at a caustic. *Comm. Pure Appl. Math.*, 19:215–250, 1966.

[46] A. Majda. Nonlinear geometric optics for hyperbolic systems of conservation laws. In *Oscillation Theory, Computation, and Methods of Compensated Compactness*, IMA Vol. Math. Appl., 2, pages 115–165. Springer, New York, 1986.

[47] F. Murat. Compacité par compensation. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 5(3):489–507, 1978.

[48] F. Murat and L. Tartar. Calcul des variations et homogenisation. In *Homogenization Methods: Theory and Applications in Physics*, Collect. Dir. Etudes Rech. Elec. France, 57, pages 316–369. Eyrolles, Paris, 1985.

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

[50] N. Neuss. *Homogenisierung und Mehrgitter.* PhD thesis, Fakultät Mathematik Universität Heidelberg, 1995.

[51] F. Poupaud and M. Rascle. Measure solutions to the linear multidimensional transport equation with non-smooth coefficients. *Comm. Partial Differential Equations*, 22(1-2):337–358, 1997.

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

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

[54] W. Symes. A slowness matching finite difference method for traveltimes beyond transmission caustics. Preprint, Dept. of Computational and Applied Mathematics, Rice University, 1996.

[55] A. Taflove. *Computational Electromagnetics, The Finite-Difference Time-Domain Method*, chapter 10. Artech House, 1995.

[56] L. Tartar. Compensated compactness and applications to partial differential equations. In R. J. Knops, editor, *Nonlinear Analysis and Mechanics*, Heriot-Watt Symposium IV, Research Notes in Mathematics, pages 136–212. Pitman, London, 1979.

[57] L. Tartar. H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations. *Proc. Roy. Soc. Edinburgh Sect. A*, 115(3-4):193–230, 1990.

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

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

[60] G. B. Whitham. *Linear and Nonlinear Waves.* John Wiley & Sons, 1974.

[61] D. Wilcox. *Turbulence Modeling for CFD.* DCW Industries, Inc., 1993.

[62] Y. Zheng. Systems of conservation laws with incomplete sets of eigenvectors everywhere. Preprint 9612, Institute for Scientific Computing and Applied Mathematics, Indiana University, 1996.