KTH Computer Science
and Communication

Topics in Analysis and Computation of Linear Wave Propagation

MOHAMMAD MOTAMED

Doctoral Thesis in Numerical Analysis
Stockholm, Sweden 2008

TRITA-CSC-A 2008:07
ISSN-1653-5723 KTH School of Computer Science and Communication
ISRN-KTH/CSC/A–08/07–SE SE-100 44 Stockholm
ISBN 978-91-7178-961-7 SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framläggs till offentlig granskning för avläggande av teknologie doktorsexamen tisdagen den 20 maj 2008 klockan 10.15 i D2, Huvudbyggnaden, Kungl Tekniska högskolan, Lindstedsvägen 3, Stockholm.

Tryck: Universitetsservice US AB

To my sister, POUNEH,
for bringing me back enthusiasm for life.

Abstract

This thesis concerns the analysis and numerical simulation of wave propagation problems described by systems of linear hyperbolic partial differential equations.

A major challenge in wave propagation problems is numerical simulation of high frequency waves. When the wavelength is very small compared to the overall size of the computational domain, we encounter a multiscale problem. Examples include the forward and the inverse seismic wave propagation, radiation and scattering problems in computational electromagnetics and underwater acoustics. In direct numerical simulations, the accuracy of the approximate solution is determined by the number of grid points or elements per wavelength. The computational cost to maintain constant accuracy grows algebraically with the frequency, and for sufficiently high frequency, direct numerical simulations are no longer feasible. Other numerical methods are therefore needed. Asymptotic methods, for instance, are good approximations for very high frequency waves. They are based on constructing asymptotic expansions of the solution. The accuracy increases with increasing frequency for a fixed computational cost. Most asymptotic techniques rely on geometrical optics equations with frequency independent unknowns. There are however two deficiencies in the geometrical optics solution. First, it does not include diffraction effects. Secondly, it breaks down at caustics. Geometrical theory of diffraction provides a technique for adding diffraction effects to the geometrical optics approximation by introducing diffracted rays. In papers 1 and 2 we present a numerical algorithm for computing an important type of diffracted rays known as creeping rays. Another asymptotic model which is valid also at caustics is based on Gaussian beams. In papers 3 and 4, we present an error analysis of Gaussian beams approximation and develop a new numerical algorithm for computing Gaussian beams, respectively.

Another challenge in computation of wave propagation problems arises when the system of equations consists of second order hyperbolic equations involving mixed space-time derivatives. Examples include the harmonic formulation of Einstein's equations and wave equations governing elasticity and acoustics. The classic computational treatment of such second order hyperbolic systems has been based on reducing the systems to first order differential forms. This treatment has however the disadvantage of introducing auxiliary variables with their associated constraints and boundary conditions. In paper 5, we treat the problem in the second order differential form, which has advantages for both computational efficiency and accuracy over the first order formulation.

Finally, paper 6 concerns the concept of well-posedness for a class of linear hyperbolic initial boundary value problems which are not boundary stable. The well-posedness is well established for boundary stable hyperbolic systems for which we can obtain sharp estimates of the solution including estimates at boundaries. There are, however, problems which are not boundary stable but are well-posed in a weaker sense, i.e., the problems for which an energy estimate can be obtained in the interior of the domain but not on the boundaries. We analyze a model problem of this type. Possible applications arise in elastic wave equations and Maxwell's equations describing glancing and surface waves.

Preface

This thesis consists of six papers and an introduction.

Paper I: M. Motamed and O. Runborg, *"A Fast Phase Space Method for Computing Creeping Rays"*, Journal of Computational Physics, vol. 219, issue 1, pp. 276–295, 2006.
The author of this thesis contributed to the ideas and developing the numerical algorithm, performed the numerical computations, and wrote parts of the manuscript.
This paper is also part of the licentiate thesis [44].

Paper II: M. Motamed and O. Runborg, *"A Multiple-patch Phase Space Method for Computing Trajectories on Manifolds with Applications to Wave Propagation Problems"*, Communications in Mathematical Sciences, vol. 5, no. 3, pp. 617–648, 2007.
The author of this thesis contributed to the ideas and developing the numerical algorithm, performed the numerical computations, and wrote parts of the manuscript.
This paper is also part of the licentiate thesis [44].

Paper III: M. Motamed and O. Runborg, *"Taylor Expansion Errors in Gaussian Beam Summation"*, Preprint, 2008.
The author of this thesis contributed to the ideas and formulation and proof of theorems and lemmas, performed the numerical computations, and wrote parts of the manuscript.

Paper IV: M. Motamed and O. Runborg, *"A Wave Front-based Gaussian Beam Method for Computing High Frequency Waves"*, Preprint, 2008.
The author of this thesis contributed to the ideas and developing the numerical algorithm, performed the numerical computations, and wrote the manuscript.

Paper V: M. Motamed, M. Babiuc, B. Szilagyi, H-O. Kreiss and J. Winicour, *"Finite Difference Schemes for Second Order Systems Describing Black Holes"*, Journal of Physical Review D, vol. 73, issue 12, 2006.
The author of this thesis contributed to the ideas, developing the numerical algorithms and formulation and proof of theorems, performed the numerical computations in Section 5, and wrote Sections 3 and 5 of the manuscript.

Paper VI: M. Motamed and H-O. Kreiss, *"Hyperbolic Initial Boundary Value Problems which are not Boundary Stable"*, Preprint, 2008.
The author of this thesis contributed to the ideas and formulation and proof of theorems and lemmas, and wrote the manuscript.

Acknowledgments

Let me do not follow the standard routine and start by thanking *Lennart Edsberg* and *Gunilla Kreiss* who introduced me to the beautiful world of numerical differential equations. This is however not the only reason I am indebted to them, I am also grateful for their kind support when I needed most.

Next, I would like to thank my principal supervisor, *Olof Runborg*, for his continuous encouragement and support. His deep understanding of mathematics and ability to simplify difficult problems have been a great help for me. Olof, you have taught me many things necessary to be a good researcher, and I am very happy and proud for having you as adviser and will always be thankful to you.

I wish to thank my co-adviser, *Heinz-Otto Kreiss*, for his strong enthusiasm for mathematics which has been an invaluable source of inspiration for me. Heinz, I have learned so much from you that goes far beyond science. I really enjoyed our discussions on both mathematics and life. Also thanks to your wife, Barbro, for having me in several occasions at your home and your Träskö-Storö Institute of Mathematics.

The fifth paper of this thesis was done in a collaboration with *Jeffrey Winicour* at Albert Einstein Institute and University of Pittsburgh. I am thankful to him and his wife, Susane, for the pleasant time spending and working together.

There are still many who have been a source of encouragement. I wish to thank *Björn Engquist*, *Jesper Oppelstrup* and *Axel Ruhe*, just to name a few.

I would like to thank my former and present colleagues and friends here at KTH. I have had a pleasant time working and being with you.

Finally, thank to my family and especially my mother, for all love and support they have given me over the years.

Financial support from the Swedish Research Council (VR), the Swedish Foundation for Strategic Research (SSF), IPAM (for a three-month visit at UCLA) and the Lars Hierta Memorial Foundation (for a two-month visit at Newton Institute) are gratefully acknowledged.

Contents

Chapter 1

Introduction

Many physical problems are formulated as systems of partial differential equations (PDEs). Accurate treatment of such problems requires a careful combination of analysis and computation. The existence of solution to PDEs is investigated by theoretical studies. For most PDEs it is however not possible to derive explicit formulas for solutions. Numerical studies are therefore needed to compute approximate solutions.

In the theoretical study of PDEs, a fundamental concept is *well-posedness*. A given problem for a PDE is said to be well-posed if it has a solution, the solution is unique and the solution depends continuously on the data given in the problem. Well-posedness is a desirable requirement for physical problems. The first two conditions are minimal requirements for a reasonable problem, and the last condition ensures that small perturbations, such as small errors in measurements or interpolation of data, do not change the solution unduly. A second important concept is *robustness*. A PDE is said to be robust if the qualitative behavior of the solution is unaffected by the addition of lower-order terms in the equation or by small changes in the coefficients. The robustness property is important because almost all PDEs modeling physical processes are derived based on some simplifying assumptions and ignoring certain effects. We want this simplification to not affect the conclusions of the analysis.

Numerical studies of PDEs concern the construction and implementation of accurate and efficient numerical algorithms for computing approximate solutions. A PDE is usually solved by first discretizing the equation on a grid or mesh, bringing it into a finite dimensional subspace, and then solving the resulting system of equations in this finite dimensional space. The first stage is usually done by the finite difference method, the finite element method or the finite volume method. The most basic property that a numerical algorithm must have is that its solutions approximate the solution of the corresponding PDE and that the approximation improves as the grid

spacings or the size of elements tend to zero. Such an algorithm is called *convergent*. Usually, it is not easy to verify convergence for a given algorithm. However, there are two related concepts that are easier to investigate, *consistency* and *stability*. Consistency implies that the solution of the PDE, if it is smooth, is an approximate solution of the numerical scheme. Stability, on the other hand, implies that the numerical solution is bounded in some sense. A fundamental theorem of numerical analysis, known as the Lax-Richtmyer equivalence theorem, states that a consistent approximation to a well-posed linear problem is convergent if and only if it is stable. In solving PDEs, the primary challenge is therefore to construct algorithms which are numerically stable.

This thesis concerns mainly some challenging problems in the analysis and computation of wave propagations described by systems of linear hyperbolic equations. A major challenge is numerical simulation of high frequency waves. When the wavelength is very small compared to the overall size of the computational domain, we encounter a multiscale problem. Examples include the forward and the inverse seismic wave propagation, radiation and scattering problems in computational electromagnetics and underwater acoustics. In direct numerical simulations, the accuracy of the approximate solution is determined by the number of grid points or elements per wavelength. The computational cost to maintain constant accuracy grows algebraically with the frequency, and for sufficiently high frequency, direct numerical simulations are no longer feasible. Other numerical methods are therefore needed. Papers 1-4 concern geometrical optics and Gaussian beams which are computationally much less costly models based on asymptotic approximations of the equations.

Another challenge in computation of wave propagation problems arises when the system of equations consists of second order hyperbolic PDEs involving mixed space-time derivatives. Examples include the harmonic formulation of Einstein's equations and wave equations governing elasticity and acoustics. The classic computational treatment of such second order hyperbolic systems has been based upon reducing the systems to first order differential forms. This treatment has however the disadvantage of introducing auxiliary variables with their associated constraints and boundary conditions. Paper 5 treats the problem in the second order differential form, which has advantages for both computational efficiency and accuracy over the first order formulation.

Finally, Paper 6 concerns the concept of well-posedness for a class of linear hyperbolic initial boundary value problems which are not boundary stable. The well-posedness is well established for two classes of problems: symmetric systems with maximally dissipative boundary conditions and boundary stable hyperbolic systems. For the first class of problems, an energy estimate can be derived using integration by parts. For the second class, the mode analysis and symmetrizer technique are used to

obtain sharp estimates of the solution including estimates at boundaries. Existence of such estimates imply that the problem is boundary stable. There are, however, problems which are not boundary stable but are well-posed in a weaker sense, i.e., the problems for which an energy estimate can be obtained in the interior of the domain but not on the boundaries. We analyze a model problem of this type. Possible applications arise in elastic wave equations and Maxwell's equations describing glancing and surface waves.

Chapter 2

Linear Hyperbolic Equations

Hyperbolic partial differential equations are in general interpreted as equations supporting "wave-like" solutions. In this chapter we briefly review the definition and properties of linear hyperbolic equations and address some challenging problems in theory and numerics of such equations which are topics of the papers in this thesis.

2.1 Initial Value Problems

The simplest hyperbolic equation is the one-way wave equation

$$u_t + a\, u_x = 0, \qquad (x,t) \in \mathbb{R} \times \mathbb{R}_+, \tag{2.1}$$

where $a \in \mathbb{R}$ is a constant, $t > 0$ denotes time, $x \in \mathbb{R}$ represents the spatial variable, and $u : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ is the unknown, $u = u(x,t)$. We specify the initial condition

$$u(x,0) = f(x), \tag{2.2}$$

where the function $f : \mathbb{R} \to \mathbb{R}$ is given. Equation (2.1) together with (2.2) is called an *initial value problem* or *Cauchy problem*. It is easy to show that the solution to this initial value problem is

$$u(x,t) = f(x - at),$$

and can be regarded as a wave that propagates with speed a without any change of shape. The solution at (x,t) depends only on the value of $\xi = x - at$. The lines in the (x,t) plane for which $x - at$ is constant are called *characteristics*. The solution is constant along characteristics. In general, when a is not constant, characteristics are curves and give important information about the solution of hyperbolic equations.

Another hyperbolic equation involving second-order derivatives is the scalar wave equation

$$u_{tt} - \Delta u = 0, \qquad (\boldsymbol{x}, t) \in \mathbb{R}^n \times \mathbb{R}_+, \tag{2.3}$$

where the Laplacian Δ is taken with respect to the spatial variables $\boldsymbol{x} = (x_1, \ldots, x_n)$. For $n = 1, 2$, the wave equation (2.3) is a simplified model for a vibrating string and a membrane, respectively. In these physical interpretations, $u(\boldsymbol{x}, t)$ represents the displacement of the point $\boldsymbol{x}$ at time t. We augment (2.3) with the initial data

$$u(\boldsymbol{x}, 0) = f(\boldsymbol{x}), \qquad u_t(\boldsymbol{x}, 0) = g(\boldsymbol{x}), \tag{2.4}$$

where the functions f and g are given. For $n = 1$, the wave equation can be rewritten in the form

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x_1}\right)\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x_1}\right) u = 0,$$

or as a system of two equations

$$\begin{aligned} v_t + v_{x_1} &= 0, \\ u_t - u_{x_1} &= v, \end{aligned}$$

which are one-way wave equations. The solution to this initial value problem consists of two waves that propagate with finite speeds and is given by d'Alembert's formula, [14],

$$u(x_1, t) = \frac{1}{2}\left[f(x_1 + t) + f(x_1 - t)\right] + \frac{1}{2}\int_{x_1 - t}^{x_1 + t} g(y)dy.$$

Many practical problems in science and engineering are described by systems of differential equations, not only by a single equation. One important class of such systems consists of linear first-order equations which is a natural generalization of the one-way wave equation (2.1),

$$u_t + \sum_{j=1}^{n} A_j(\boldsymbol{x}, t)\, u_{x_j} = f(\boldsymbol{x}, t), \qquad (\boldsymbol{x}, t) \in \mathbb{R}^n \times \mathbb{R}_+, \tag{2.5}$$

subject to the initial condition

$$u(\boldsymbol{x}, 0) = g(\boldsymbol{x}). \tag{2.6}$$

The unknown is $u : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}^m$, $u = (u_1, \ldots, u_m)^\top$, and the functions $A_j : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}^{m \times m}$, $f : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}^m$, and $g : \mathbb{R}^n \to \mathbb{R}^m$ are given.

Definition 1. *The system of PDEs (2.5) is called hyperbolic if the symbol*

$$P(\boldsymbol{x}, t; \omega) := \sum_{j=1}^{n} A_j(\boldsymbol{x}, t)\, \omega_j,$$

has real eigenvalues and is uniformly diagonalizable for each $\boldsymbol{x}$ and $\omega = (\omega_1, \ldots, \omega_n)$ in $\mathbb{R}^n$ and $t \geq 0$. The system is called strictly hyperbolic if the eigenvalues of the symbol are real and distinct.

Another important class is the system of second-order hyperbolic equations,

$$u_{tt} - \sum_{j,k=1}^{n} A_{jk}(\boldsymbol{x}, t)\, u_{x_j x_k} + \sum_{j=1}^{n} B_j(\boldsymbol{x}, t)\, u_{x_j} + C(\boldsymbol{x}, t)u = f(\boldsymbol{x}, t), \quad (\boldsymbol{x}, t) \in \mathbb{R}^n \times \mathbb{R}_+, \tag{2.7}$$

with the initial condition

$$u(\boldsymbol{x}, 0) = g(\boldsymbol{x}), \qquad u_t(\boldsymbol{x}, 0) = h(\boldsymbol{x}). \tag{2.8}$$

Such systems can be seen as natural generalization of the wave equation (2.3).

Definition 2. *The system of PDEs (2.7) is called hyperbolic if there is a positive constant δ such that*

$$\sum_{j,k=1}^{n} A_{jk}(\boldsymbol{x}, t)\, \omega_j\, \omega_k \geq \delta |\omega|^2\, I,$$

for all $\boldsymbol{x}$ and $\omega = (\omega_1, \ldots, \omega_n)$ in $\mathbb{R}^n$ and $t \geq 0$. Here, I is the $m \times m$ identity matrix.

We now quantify the concept of well-posedness introduced in the introduction. For two C^∞-functions $u, v : \mathbb{R}^n \to \mathbb{C}^m$ which are 1-periodic in $\boldsymbol{x}$, we define the L_2-inner product and norm by

$$(u, v) = \int_0^1 \ldots \int_0^1 < u(\boldsymbol{x}), v(\boldsymbol{x}) > \, dx_1 \ldots dx_n, \qquad ||u|| = (u, u)^{1/2},$$

where

$$< u, v > = \sum_{j=1}^{m} \bar{u}_j\, v_j.$$

Consider the initial boundary value problem (2.5), (2.6) with C^∞-coefficients and data which are 1-periodic in every spatial dimension.

Definition 3. *The initial boundary value problem (2.5), (2.6) is well-posed if:*

(1) for every 1-periodic and C^∞ data f, g, there exists a unique solution $u(\boldsymbol{x}, t) \in C^\infty(\boldsymbol{x}, t)$, which is 1-periodic in every spatial dimension;

(2) for each $T > 0$, there is a constant $K(T)$, independent of f and g, such that

$$||u(.,t)||^2 \leq K(T) \left(||g(.)||^2 + \int_0^t ||f(.,\tau)||^2 d\tau \right). \tag{2.9}$$

In order to motivate the above definition, we change the initial data (2.6) to

$$\tilde{u}(\boldsymbol{x}, 0) = g(\boldsymbol{x}) + \delta\, h(\boldsymbol{x}), \qquad 0 < \delta \ll 1, \qquad ||h(.)|| = 1.$$

The difference $w(\boldsymbol{x}, t) = \tilde{u}(\boldsymbol{x}, t) - u(\boldsymbol{x}, t)$ is then a solution of (2.5) with $f \equiv 0$ and initial data

$$w(\boldsymbol{x}, 0) = \delta\, h(\boldsymbol{x}).$$

If the estimate (2.9) holds, then

$$||\tilde{u}(.,t) - u(.,t)||^2 \leq K(T)\, \delta^2\, ||h(.)||^2 = K(T)\, \delta^2.$$

Therefore, (2.9) guarantees that for any finite time interval $0 \leq t \leq T$, small perturbations of the initial data results in small changes in the solution, i.e, the solution depends continuously in the initial data.

By reducing second-order systems to first-order systems, similar to reducing the wave equation to two one-way wave equations, we can also define well-posedness for second order systems. From the theory of linear hyperbolic systems, [14, 37], it is well known that the initial value problems for first and second order hyperbolic systems are well-posed. Moreover, the wave solutions have finite propagation speed in the sense that the solution at a given point $(\boldsymbol{x}_0, t_0)$ depends only on the data in $(\boldsymbol{x}, t) \in \Omega$, where Ω is a finite region of space and time. In other words, we can change the data outside the region Ω without affecting the solution at $(\boldsymbol{x}_0, t_0)$.

2.2 Boundary Conditions

Most physical applications of partial differential equations involve domains with boundaries, and interesting phenomena frequently occur near these boundaries. The formulation of boundary conditions therefore play an important role. The problem of determining a solution to a partial differential equation when both initial data and boundary data are present is called an *initial boundary value problem*. One fundamental question is then how to impose proper boundary conditions such that the problem becomes well-posed.

In the case of hyperbolic equations, the characteristics play an important role in determining correctly posed boundary conditions. To illustrate this, we consider the one-way wave equation

$$u_t + a\, u_x = 0,$$

in the strip $0 \le x \le 1$, $t \ge 0$. If $a > 0$, the characteristics in this region propagate from the left to the right. The solution must therefore be specified on the boundary at $x = 0$, in addition to the initial data, in order to be defined for all time. Moreover, no data need to be supplied at the other boundary at $x = 1$, since otherwise the solution will be overdetermined. In general, values for the ingoing characteristic variables must be provided at the boundaries.

Consider the initial boundary value problems for (2.5) with initial conditions (2.6) in the half-space

$$R_0 = \{\boldsymbol{x}\,|x_1 \ge 0,\ -\infty < x_j < \infty,\ j = 2, \ldots, n\}, \tag{2.10}$$

and boundary conditions, at $x_1 = 0$,

$$S\, u(0, \boldsymbol{x}_-, t) = h(\boldsymbol{x}_-, t), \qquad \boldsymbol{x}_- = (x_2, \ldots, x_n), \tag{2.11}$$

where $S \in \mathbb{R}^{r\times m}$ is a rectangular matrix, with r being the number of ingoing characteristic variables.

Definition 4. *Let $f \equiv g \equiv 0$. We call the half-space problem (2.5), (2.6), (2.11) boundary stable if for all smooth boundary data h, there is a unique solution u, and in each time interval $0 \le t \le T$ there is constant K_T independent of the data such that*

$$\int_0^t ||u(0, \boldsymbol{x}_-, \tau)||^2_{R_-}\, d\tau \le K_T \int_0^t ||h(0, \boldsymbol{x}_-, \tau)||^2_{R_-}\, d\tau.$$

Here, $||.||_{R_-}$ denote the L_2-norm over the space

$$R_- = \{\boldsymbol{x}_- \,| -\infty < x_j < \infty, j = 2, \ldots, n\}.$$

The theory of linear hyperbolic initial boundary value problems is well developed in the case when the problem is *boundary stable*, i.e., when there exist proper estimates of the solution based on the data at the boundaries. In the general theory, the following concept of well-posedness is introduced.

Definition 5. *Let $g \equiv 0$. The half-space problem (2.5), (2.6), (2.11) is called strongly well-posed in the generalized sense if for all smooth compatible data, f and h, there is a unique solution u, and in each time interval $0 \le t \le T$, there is a constant K_T independent of the data such that*

$$\int_0^t ||u(\boldsymbol{x}, \tau)||^2_{R_0} + ||u(0, \boldsymbol{x}_-, \tau)||^2_{R_-}\, d\tau \le K_T \int_0^t ||F(\boldsymbol{x}, \tau)||^2_{R_0} + ||g(0, \boldsymbol{x}_-, \tau)||^2_{R_-}\, d\tau.$$

We then have

Theorem 1. *Assume that the half-space problem is boundary stable. Then it is strongly well-posed in the generalized sense.*

There are, however, problems which are not boundary stable but are well-posed in a weaker sense. Examples include surface waves and glancing waves in electromagnetic and elastic wave propagation problems described by Maxwell's equations and elastic wave equations with certain types of boundary conditions. It is therefore necessary to develop a theory for such types of problems. In paper 6, [46], we consider a model problem which may not be boundary stable and extend the theory of boundary stable problems to this case. We consider (2.5) with $n = m = 2$ and

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \tag{2.12}$$

in the half-space (2.10). We augment this system with the initial condition (2.6) and the boundary condition at $x_1 = 0$,

$$u_1(0, x_2, t) = \alpha\, u_2(0, x_2, t) + h(x_2, t), \qquad \alpha \in \mathbb{C}. \tag{2.13}$$

Here $u(x_1, x_2, t) = (u_1, u_2)^\top$ is a vector-valued function, and the data f, g, h are assumed to be compatible smooth functions with compact support. Different values of α in the boundary condition result in different behavior of the problem. Here, we summarize the result, [46]:

1) if $|\alpha| < 1$, then the problem is boundary stable and therefore strongly well-posed in the generalized sense.

2) if $|\alpha| = 1$, then the problem is not boundary stable but is well-posed in the sense that there are proper energy estimates inside the domain, but not at the boundary.

3) if $|\alpha| > 1, \alpha \in \mathbb{R}$, then the problem is ill-posed in the sense that the solution "looses" one derivative at each reflection from the boundary.

4) if $|\alpha| > 1, \alpha \notin \mathbb{R}$, then the problem is ill-posed in the sense that there are solutions which grow exponentially, arbitrarily fast.

We formulate a theorem for this model problem and conjecture that the theorem holds also for the more general initial boundary value problem (2.5), (2.6), (2.11), which is the topic of future work.

2.3 Numerical Methods

For most hyperbolic initial boundary value problems, it is not possible to derive explicit formulas for solutions. Numerical methods are employed to compute approximate solutions. There is a wide range of different methods for solving linear hyperbolic problems. The most commonly used numerical methods include the finite difference method, the finite volume method, the finite element method, spectral methods and the boundary element method.

The finite difference method, [19, 61] is one of the oldest numerical methods. In this method, the PDE is discretized on a grid by approximating the derivatives of the solution in terms of the values of the solution on a set of discrete grid points. This gives a large algebraic system of equations which needs to be solved. For systems of first-order hyperbolic equations, upwind-type methods based on the direction of characteristics are frequently used. For second-order hyperbolic systems, leapfrog schemes on staggered grids are more attractive. For example, a widely used class of this type is the finite difference time-domain method (Yee scheme) for solving Maxwell's equations of electromagnetics, [74].

In the finite volume method, [39], instead of calculating the solution at discrete grid points, the total integral of the solution is approximated over grid cells which are small volumes surrounding each grid point. It is based on the integral form of the PDE. Finite volume methods are particularly useful for solving nonlinear hyperbolic problems. One advantage of these methods is that they are easily formulated to allow for unstructured grids. We should emphasize that the computational treatment of nonlinear PDEs is more difficult than that of linear equations, due to possible discontinuity and non-uniqueness of the solutions. For such problems essentially non oscillatory (ENO) and weighted essentially non oscillatory (WENO) schemes are employed, [60].

The finite element method, [13], is based on discretizing the weak form of the boundary value problem in a finite dimensional space. This method is particularly useful for solving PDEs over complex domains. The domain is decomposed into small elements, which may be simply triangles or more complicated curvilinear polygons. The solution obtained by the finite element method is a linear combination of basis functions that are nonzero only over small subdomains. Two classes of finite element methods which are widely used for hyperbolic problems are the discontinuous Galerkin method, [7], in which there is no continuity restriction on the interface of the elements, and the streamline diffusion method, [28], in which the basis functions are modified to produce a small amount of artificial diffusion in the direction of streamlines. These methods are particularly useful for problems with discontinuous solutions.

In spectral methods, [5], the solution is first written as its Fourier series. This

series is next substituted into the equation and a system of ordinary differential equation (ODE) is obtained. The ODEs are then solved using an ODE solver. The spectral method is similar to the finite element method in approximating the solution as a linear combination of basis functions. However, in contrast to the finite element method, the basis functions are continuous and nonzero over the whole domain. As a result of this, the spectral method usually works better when the solution is smooth. Moreover, it can only be applied to problems with simple computational domains, such as cubes.

In the boundary element method, [18], the PDE is rewritten as a boundary integral equation defined on the boundary of the domain using the Green's theorem. Therefore, only the boundary of the domain needs to be discretized, which in turn results in reducing the dimension of the problem at least by one. This is beneficial from computational complexity point of view. However, in contrast to the finite difference or the finite element method where the resulting system of linear equations has a sparse structure, here we get a dense system. Moreover, the boundary element method is applicable to problems for which Green functions can easily be calculated, for instance when the speed of wave propagation is constant.

It is beyond the scope of this thesis to study all these numerical methods for wave propagation problems. We only note that each numerical method is applicable for a certain type of problem. In order to choose a proper numerical algorithm, one usually considers the accuracy and efficiency of the method. A major part of the thesis focuses on developing accurate and efficient algorithms for some challenging wave propagation problems.

One challenging computational problem, which is treated in paper 5, [45], is when the system of equations consists of second order hyperbolic PDEs, involving mixed space-time derivatives. For instance, the harmonic formulation of Einstein's equations is a system of ten nonlinear second order hyperbolic equations with mixed space-time derivatives. After linearizing and reducing to first order form, we obtain a system of about sixty first order equations. This results in a notable increase in the computational complexity. As a better alternative, the problem can be treated in the second order form without any order reduction. Although the discretization of the second order system involves more subtle analysis because of numerical stability issues which are not present in first order formulations, this approach has advantages for both computational efficiency and accuracy over the first order formulation. The main difficulty in treating such second order systems is due to the presence of mixed space-time derivatives. For instance, if we simply use central difference approximations for both time and space derivatives, the scheme will not be stable for particular choices of the coefficients in the equations.

As a model problem with similar properties, we consider the initial value problem

for the second order hyperbolic system (2.7) with constant coefficients A_{jk} and $B_j \equiv C \equiv f \equiv 0$, and the initial conditions (2.8). In order to introduce mixed derivatives into the equations, we use a shifted coordinate

$$\boldsymbol{x} = \tilde{\boldsymbol{x}} + \beta\, t, \qquad \tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_n) \in \mathbb{R}^n, \qquad \beta = (\beta_1, \ldots, \beta_n) \in \mathbb{R}^n_+,$$

and obtain the shifted system

$$u_{tt} = 2P_1(\partial/\partial\tilde{\boldsymbol{x}})\, u_t - P_1^2(\partial/\partial\tilde{\boldsymbol{x}})\, u + P_0(\partial/\partial\tilde{\boldsymbol{x}})\, u, \tag{2.14}$$

with the operators

$$P_1(\partial/\partial\tilde{\boldsymbol{x}}) = \sum_{j=1}^{n} \beta_j \frac{\partial}{\partial\tilde{x}_j}, \qquad P_0(\partial/\partial\tilde{\boldsymbol{x}}) = \sum_{j,k=1}^{n} A_{jk} \frac{\partial}{\partial\tilde{x}_j} \frac{\partial}{\partial\tilde{x}_k}.$$

The shifted system (2.14) is an important model for describing black holes in numerical relativity. The study of this system also provides a firm basis for solving the harmonic Einstein system of equations, because of the existence of the mixed space-time derivatives which are essential features of the Einstein equations. This type of systems also arises in acoustic wave propagation in a medium with nonuniform macroscopic motion.

We use the method of lines and reduce the system of partial differential equations, in their second-order form, to a system of ordinary differential equations in time on a spatial grid. We then apply the energy method and Fourier-Laplace transformation to analyze and establish stable approximations.

Another challenging problem is numerical simulation of high frequency waves, which is the subject of papers 1-4. We discuss such numerical methods in the next chapter in more detail.

Chapter 3

High Frequency Waves

Simulation of high-frequency wave propagation is important in many engineering and science fields. Examples include radar and sonar technology, wireless communication, seismic tomography, medical imaging and non-destructive testing.

In this chapter, we study the numerical simulation of waves at high frequencies and the underlying mathematical models used. For simplicity we will mainly discuss the linear scalar wave equation,

$$u_{tt} - c(\boldsymbol{x})^2 \Delta u = 0, \qquad (\boldsymbol{x}, t) \in \mathbb{R}^n \times \mathbb{R}_+, \tag{3.1}$$

where $c(\boldsymbol{x})$ is the local speed of wave propagation of the medium. We complement (3.1) with initial data that generate high-frequency solutions. The exact form of the data will not be important here, but a typical example would be $u(\boldsymbol{x}, 0) = A(\boldsymbol{x}) \exp(i\omega \boldsymbol{k} \cdot \boldsymbol{x})$ where $|\boldsymbol{k}| = 1$ and the frequency $\omega \gg 1$. With slight modifications, the techniques we describe will also carry over to systems of wave equations, like the Maxwell equations and the elastic wave equation. See, for instance, [24] where the linear Schrödinger equation is treated. We also define the *index of refraction* as $\eta(\boldsymbol{x}) = c_0/c(\boldsymbol{x})$ with the reference velocity c_0 (*e.g.* the speed of light in vacuum). For simplicity we will henceforth let $c_0 = 1$.

3.1 Time-harmonic Helmholtz equation

We consider time harmonic waves of type $u(\boldsymbol{x}, t) = v(\boldsymbol{x}) \exp(i\omega t)$ with ω fixed. Inserting it into the time-dependent wave equation (3.1), we get the Helmholtz equation

$$c(\boldsymbol{x})^2 \Delta v + \omega^2 v = 0. \tag{3.2}$$

When the wave frequency is high and the wavelength is short compared to the size of the computational domain, we encounter a multiscale problem with a highly oscillatory solution. Direct simulations based on the standard wave equations are very expensive, since a large number of grid points is required to resolve the wave oscillations. It is, therefore, a difficult computational problem, and computations are a major challenge.

In general, numerical methods for high-frequency wave problems can be classified into three categories:

- **Direct methods**: One class of direct methods is based on the standard wave equations. The accuracy of the solution is then determined by the number of grid points or elements per wavelength, and the computational cost for a fixed accuracy increases with increasing frequency. The computational complexity is at least $\mathcal{O}(\omega^n)$. Another class is based on integral equations. Given a boundary condition, and a constant speed of propagation, the problem can be formulated as an integral equation on the boundary. Therefore only the boundary needs to be discretized instead of the whole domain, and the effective dimension is $n-1$. Standard methods for solving the boundary integral equations include the method of moments [21] and finite element methods [73, 55]. Using a fast iterative solver such as the fast multilevel multipole technique [9, 72], the complexity of these methods will be almost $\mathcal{O}(\omega^{n-1})$. There are, however, efforts to find robust algorithms of complexity $\mathcal{O}(1)$, [17].

- **Asymptotic methods**: These methods are based on constructing asymptotic expansions of the solution which are valid when $\omega \to \infty$. The accuracy increases with increasing frequency for a fixed computational cost. Most asymptotic techniques rely on geometrical optics equations with frequency independent unknowns. Among other asymptotic methods are wave optical methods (physical optics and physical theory of diffraction) and Gaussian beam methods.

- **Hybrid methods**: They combine direct and asymptotic techniques [43, 20]. Direct methods are applied on the regions where the geometric variations or the variations in $c(\boldsymbol{x})$ are of the same scale as the wavelength, and asymptotic methods are applied elsewhere. In some cases a linear combination of both methods are used.

In what follows, we will briefly review variants of geometrical optics approximations. Instead of the oscillating wave field the unknowns in standard geometrical optics are the phase and the amplitude, which typically vary on a much coarser scale than the full solution. Hence, they should in principle be easier to compute numerically. The main drawbacks of the infinite frequency approximation of geometrical

optics are that diffraction effects at boundaries are lost, and that the approximation breaks down at caustics, where the predicted amplitude is unbounded. For these situations more detailed models are needed, such as the geometrical theory of diffraction [31], which adds diffraction phenomena by explicitly taking into account the geometry of Ω and boundary conditions. The solution's asymptotic behavior close to caustics can also be derived, and a correct amplitude for finite frequency can be computed [36, 42, 22]. Numerically this can for instance be done with Gaussian beams [53, 2].

3.2 Geometrical Optics

In order to solve the Helmholtz equation (3.2) for large values of ω, we seek solutions of the form

$$v(\boldsymbol{x}) = a(\boldsymbol{x}, \omega) e^{i\omega\phi(\boldsymbol{x})}, \qquad \boldsymbol{x} \in \mathbb{R}^3. \tag{3.3}$$

The real-valued phase function $\phi(\boldsymbol{x})$ is independent of ω, and the amplitude function $a(\boldsymbol{x}, \omega)$ is assumed to be expanded in inverse powers of ω,

$$a(\boldsymbol{x}, \omega) \approx \sum_{k=0}^{\infty} a_k(\boldsymbol{x})(i\omega)^{-k} = \sum_{k=0}^{n} a_k(\boldsymbol{x})(i\omega)^{-k} + \mathcal{O}(\omega^{-n}). \tag{3.4}$$

It means that the series is an asymptotic expansion of a as $\omega \to \infty$. It is known as the asymptotic WKBJ expansion, [22]. Geometrical optics (GO) only considers the leading term of the series ($k = 0$), which is called the the geometrical optics term. Putting (3.3) with the leading term of (3.4) into (3.2) and canceling the phase factor $e^{i\omega\phi}$, we get

$$|\nabla\phi|^2 = \eta(\boldsymbol{x})^2, \tag{3.5}$$

$$2\nabla\phi \cdot \nabla a_0 + a_0 \Delta\phi = 0. \tag{3.6}$$

Equation (3.5) is the *eikonal* equation, which is a first order non-linear PDE for $\phi(\mathbf{x})$. Equation (3.6) is the *transport* equation, which is a linear PDE with variable coefficients for a_0, once ϕ is known.

GO can also be formulated in terms of ODEs. We first note that the eikonal equation is a nonlinear Hamilton-Jacobi equation with Hamiltonian $H(\boldsymbol{x}, \boldsymbol{p}) = |\boldsymbol{p}|/\eta(\boldsymbol{x}) \equiv 1$, where $\boldsymbol{p} = \nabla\phi$ is the *slowness* vector. We let $(\boldsymbol{x}(t), \boldsymbol{p}(t))$ be a bi-characteristic related to this Hamiltonian. Since H is constant along them, $H(\boldsymbol{x}(t), \boldsymbol{p}(t)) = H(\boldsymbol{x}_0, \boldsymbol{p}_0)$, we get the so called *ray* equations,

$$\frac{d\boldsymbol{x}}{dt} = \nabla_{\boldsymbol{p}} H = \frac{1}{\eta^2}\boldsymbol{p}, \qquad \frac{d\boldsymbol{p}}{dt} = -\nabla_{\boldsymbol{x}} H = \frac{\nabla\eta}{\eta}. \tag{3.7}$$

There are also ODEs for the amplitude, [11].

There is yet another formulation for GO based on a kinetic viewpoint. Considering rays as trajectories of particles (photons) and introducing the phase space $(t, \boldsymbol{x}, \boldsymbol{p})$, we note that the evolution of these particles in the phase space is given by the ray equations (3.7). We let $f(t, \boldsymbol{x}, \boldsymbol{p})$ be a particle density function. It will then satisfy the *Liouville* equation,

$$f_t + \nabla_{\boldsymbol{p}} H \cdot \nabla_{\boldsymbol{x}} f - \nabla_{\boldsymbol{x}} H \cdot \nabla_{\boldsymbol{p}} f = 0, \tag{3.8}$$

where $\nabla_{\boldsymbol{p}} H$ and $\nabla_{\boldsymbol{x}} H$ are given by (3.7).

There are different numerical techniques based on the three different mathematical models of GO:

1. Numerical methods based on the *ray* equations (3.7) include *ray tracing* [8, 29, 38]. In this method the ODEs (3.7) together with the ODEs for the amplitude are solved with standard ODE solvers such as 2nd or 4th order Runge-Kutta methods, giving the phase and amplitude along the rays. The solution at a desired point is then interpolated from the solutions along the rays. This can be rather difficult in the regions where ray tracing produces diverging or crossing rays. Moreover, ray tracing is only of interest for problems involving a small number of source points. For problems with many source points, ray tracing may be computationally expensive.

2. Numerical methods based on the *eikonal* equation (3.5) are Hamilton-Jacobi methods. They solve the eikonal and transport equations on a uniform Eulerian grid to control the error everywhere. Different types of numerical techniques have been proposed to compute the unique viscosity solution of the eikonal equation, including upwind methods of ENO or WENO type [69, 68, 52], fast marching method [66, 58, 59, 54], group marching method [32] and sweeping method [57, 33, 65]. However, since the eikonal equation is a nonlinear equation for which the superposition principle does not hold, these methods fail to capture multivalued solutions corresponding to crossing rays. Among the methods proposed for computing multivalued solutions are a domain decomposition based method by detecting kinks [15], big ray tracing [3, 1] and slowness matching method [62, 63]. The multivalued solutions, in these methods, are constructed by putting together the solutions of several eikonal equations. Nevertheless, finding a robust technique to compute multivalued solutions is still a computational challenge.

3. Numerical methods based on the *kinetic* equation (3.8) are so called phase space methods. The Liouville equation, like ray equations, benefits from the

linear superposition principle. Moreover, its solution can be computed on a fixed Eulerian grid. There is, however, a drawback with directly solving the Liouville equation. Because of introducing the phase space and increasing the number of independent variables, a direct simulation will computationally be very expensive. There are two different approaches to overcome this drawback; wave front methods and moment-based methods. In the former, special wave front solutions are computed, and the later is based on transforming the Liouville equation to a system of conservation law equations for moments of f in the reduced space $(t, \boldsymbol{x})$. See, for instance, [40, 6, 10, 56]. The classical wave front methods include Lagrangian front tracking, wave front construction [70], the segment projection method [12, 64] and level set method [51, 41, 27, 26, 25]. Related methods are the fast phase space method [16] and the phase flow method [76].

See [11, 4] for a survey of geometrical optics approximations.

3.3 Geometrical Theory of Diffraction

There are two deficiencies in the GO solution described above. First, it does not include diffraction effects. Secondly, it breaks down at caustics, where a_0 is unbounded. To overcome the first deficiency, in addition to the incident and reflected rays of GO, new classes of rays, namely *diffracted rays*, should be introduced to construct the full asymptotic expansion of the solution.

Geometrical theory of diffraction (GTD), developed by J. Keller [31], provides a technique for adding diffraction effects to the geometrical optics approximation. GTD is often used in scattering problems in computational electromagnetics, where boundary effects are of major importance, for example in radar cross section calculations and in the optimization of base station locations for cell phones in a city.

There are various kinds of diffracted rays. One type of diffracted rays is generated when there is a discontinuity in the scatterer surface, such as edges, tips or changes in material properties. At these singular points an infinite set of diffracted rays are produced which obey the usual geometrical optics equations. The amplitude of each diffracted ray is proportional to the amplitude of the ray hitting the corner and a diffraction coefficient D. The coefficient D depends on the directions of the inducing and diffracted rays, the frequency, the local boundary geometry and the shape of the incident wave front. In Figure 3.1 (left), the incident ray hitting the tip of a wedge generates a reflected ray, another ray that continues past the tip, and infinitely many diffracted rays in all directions.

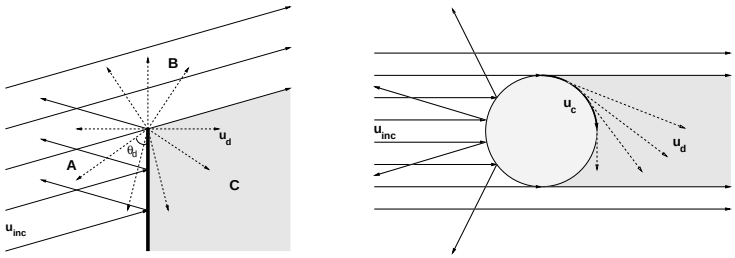


Figure 3.1: Diffraction by discontinuous and smooth scatterers. Left figure shows diffraction of an incident field u_{inc} by a wedge. The incident ray hitting the tip of the wedge generates a reflected ray, another ray that continues past the tip, and infinitely many diffracted rays u_d in all directions. Right figure shows a creeping ray u_c induced by the incident field u_{inc} at the north pole of a perfectly conducting cylinder, where the incident direction is orthogonal to the surface normal. As the creeping ray propagates on the boundary, it continuously emits surface-diffracted rays u_d with exponentially decreasing initial amplitude.

One typical improved expansion adds diffracted rays to GO by adding extra correction terms to the asymptotic solution (3.3-3.4),

$$v(\boldsymbol{x}) = a(\boldsymbol{x},\omega)e^{i\omega\phi(\boldsymbol{x})} + b(\boldsymbol{x},\omega)e^{i\omega\phi_d(\boldsymbol{x})}, \qquad b(\boldsymbol{x},\omega) \approx \sum_{k=0}^{\infty} b_k(\boldsymbol{x})(i\omega)^{-k-\frac{1}{2}}, \tag{3.9}$$

where $\phi_d(\boldsymbol{x})$ and $b_k(\boldsymbol{x})$ are the phase and amplitudes associated with diffracted rays. More elaborate expansions must sometimes be used, such as those given by the *uniform theory of diffraction* (UTD), [35].

Another type of diffraction is generated even for smooth scatterers. When an incident field hits a smooth body there will be a shadow zone behind it and the geometrical optics solution will again be discontinuous. There is a curve (point in 2D) dividing the shadow part and the illuminated part of the body. Along this *shadow line* (shadow point in 2D) the incident rays are tangent to the body surface. The shadow line will act as a source for *creeping rays*, that propagate along geodesics on the scatterer surface, if the surrounding medium is homogeneous, $\eta \equiv 1$. The creeping ray carries an amplitude proportional to the amplitude of the inducing ray. At each point on a convex surface with perfectly conducting material, the creeping ray sheds surface-diffracted rays in the tangential direction, with its current amplitude. The amplitude decays exponentially along the creeping ray's trajectory. In three di-

mensions, the amplitude also changes through geometrical spreading on the surface. The diffracted rays follow the usual geometrical optics laws. A 2D example is shown in Figure 3.1 (right). The incident ray hitting the north pole of a perfectly conducting circular cylinder generates a creeping ray propagating on the cylinder boundary and shedding diffracted rays along its way. Note that another creeping ray will be generated by the incident ray hitting the south pole.

The diffracted rays generated by discontinuities and shed by creeping rays obey the usual geometrical optics equations. The main computational task is thus based on the standard GO approximation discussed in Section 3.2. However, computing creeping ray contribution to the field involves more technicalities, and one needs to find geodesics on the scatterer surface as well. We assume that the scatterer surface can be represented by a regular parameterization $\boldsymbol{x} = \bar{X}(\boldsymbol{u})$, where $\boldsymbol{x} = (x, y, z) \in \mathbb{R}^3$ is the coordinate in $3D$ physical space, and the parameters $\boldsymbol{u} = (u, v)$ belong to a set $\Omega \subset \mathbb{R}^2$. Let the scatterer be illuminated by incident rays in a certain direction, and assume that the shadow line $\boldsymbol{u}_0(s)$ is represented by a curve in parameter space, with s being the arc length parameterization. A wave field, associated to the creeping rays, is generated on the surface

$$v_s(\boldsymbol{u}) = a(\boldsymbol{u})e^{i\omega\phi(\boldsymbol{u})}, \tag{3.10}$$

where $\phi(\boldsymbol{u})$ and $a(\boldsymbol{u})$ are surface phase and amplitude. The creeping rays are related to (3.10) in the same way as the standard GO rays are related to the leading term of the series (3.3-3.4). Like in GO, the surface wave field can be formulated as a system of either ODEs or PDEs. In the ODE formulation, we obtain a system of equations known as *surface ray* equations. In the PDE formulation, we get *surface eikonal* and *surface transport* equations. See [30, 44]. Based on these two formulations, there are different numerical techniques for computing creeping rays. Lagrangian techniques are based on surface ray equations. The simplest and most common method is standard ray tracing which solves these ODEs on triangulated surfaces [23]. Assuming the geodesic paths are given by piecewise linear curves, it is possible to find the linear ray path on each triangle, analytically. This method gives the surface phase and amplitude solutions along creeping rays. Interpolation must then be applied to obtain the solution everywhere. But, in regions where rays cross or diverge this can be rather difficult. However, the interpolation can be simplified by using wave front methods [71, 20] in which, instead of individual rays, an interface representing a wave front is evolved. Nevertheless, for some problems, such as radar cross section (RCS) computations, where creeping rays from all illumination angles must be computed, Lagrangian methods can be computationally expensive. Eulerian techniques are based on surface eikonal and surface transport equations. These PDEs are discretized on fixed computational grids, and there is no problem with interpolation [34].

However, these equations only give the correct solution when it is a single wave. In the case of crossing waves, more elaborate schemes have been devised to capture multivalued solutions, [47, 75].

In paper 1, [47], we present an adaptation of the fast phase space method, [16], for standard geometrical optics to computation of creeping waves. This method is based on a new PDE formulation of creeping rays given by so called *escape equations*. The escape PDEs solutions contain information about all possible creeping rays in all directions. To extract properties like phase and amplitude for a ray family, post-processing of the solution is needed.

This method requires one fixed parameterization of the scatterer. It has however been modified in paper 2, [48], for more complex scatterer surfaces which cannot be represented by a single non-singular explicit parameterization. The surface is split into several simpler surfaces with explicit parameterizations. These multiple patches collectively cover the scatterer surface in a non-singular manner. The escape PDEs are solved in every patch, individually. The creeping rays on the scatterer are then computed by connecting all individual solutions through a fast post-processing. The inter-patch boundaries are treated by the continuity of creeping rays.

3.4 Gaussian Beams

Close to caustics the amplitude grows rapidly in the geometrical optics approximation and blows up at the caustic itself. In reality the amplitude remains bounded, but increases with the frequency ω. The error in the standard series expansion (3.3-3.4) is thus unbounded around caustics. To capture the actual solution behavior there are better expansions that have small errors uniformly in ω, derived *e.g.* by Ludwig [42] and Kravtsov [36]. The expansions are different for different types of caustics. For a fold caustic there are two ray families meeting at the caustic, with phases ϕ^+ and ϕ^-. Letting $\rho = \frac{3}{4}(\phi^+ - \phi^-)$ a more suitable description of the solution u in this case is

$$u(\boldsymbol{x}) = \omega^{1/6}\, e^{i\omega\phi(\boldsymbol{x})} \left(\mathrm{Ai}(-(\omega\rho(\boldsymbol{x}))^{2/3}) \sum_{k=0}^{\infty} A_k(\boldsymbol{x})(i\omega)^{-k} \right.$$
$$\left. + \; i\omega^{-1/3}\mathrm{Ai}'(-(\omega\rho(\boldsymbol{x}))^{2/3}) \sum_{k=0}^{\infty} B_k(\boldsymbol{x})(i\omega)^{-k} \right),$$

where Ai is the Airy function. The dominant term close to the caustic, $|\rho|\omega \ll 1$ is of the order $\mathcal{O}(\omega^{1/6})$ with an error of $\mathcal{O}(\omega^{-1/3})$. Away from the caustic, on the convex side where $\rho > 0$, we can use the fact that $|\mathrm{Ai}(-x)| \sim x^{-1/4}$ and $|\mathrm{Ai}'(-x)| \sim x^{1/4}$

for large x, to conclude that the dominant term is of the order $\mathcal{O}(1)$ with an error of $\mathcal{O}(\omega^{-1})$, *i.e.* the standard situation for geometrical optics.

We will now discuss the Gaussian beam method for computing the wave field at caustics. The Gaussian beam method is an asymptotic method for computing high-frequency wave fields in smoothly varying inhomogeneous media. It was proposed by Popov [53], based on an earlier work of Babic and Pankratova [2]. Gaussian beams are closely related to ray tracing, but instead of viewing rays just as characteristics of the eikonal equation, Gaussian beams are fatter rays: They are approximate high frequency solution to the wave equation or the Helmholtz equation which are concentrated on a standard ray. Contrary to standard GO rays, Gaussian beams accept complex valued phase functions. The main advantage of this construction is that Gaussian beams give the correct solution also at caustics where standard geometrical optics breaks down.

We now review the governing equations. We first note that because of the constraint $H(\boldsymbol{x}, \boldsymbol{p}) = 1$, or $|\boldsymbol{p}| = \eta(\boldsymbol{x})$, the dimension of the phase space $(\boldsymbol{x}, \boldsymbol{p})$ can actually be reduced by one. For example in two dimensions, with $\boldsymbol{x} = (x, y)$, by setting $\boldsymbol{p} = \eta(\cos\theta, \sin\theta)$ and using θ as a dependent variable in (3.7) instead of $\boldsymbol{p}$, we get the reduced equations,

$$\frac{dx}{dt} = c(x, y)\cos\theta, \tag{3.11a}$$

$$\frac{dy}{dt} = c(x, y)\sin\theta, \tag{3.11b}$$

$$\frac{d\theta}{dt} = \frac{\partial c}{\partial x}\sin\theta - \frac{\partial c}{\partial y}\cos\theta. \tag{3.11c}$$

We consider a ray in a two-dimensional Cartesian coordinate system x, y given by the ray tracing system (3.11). In orthogonal ray-centered coordinates (t, q), where q is the axis perpendicular to the ray at point t with the origin on the ray, the paraxial Gaussian beam solution closely concentrated about the central ray is given by

$$u(t, q, \omega) = A(t, q)\ \exp\{i\omega\phi(t, q)\}. \tag{3.12}$$

Here the complex-valued amplitude A and the phase ϕ are given by the eikonal and transport equations with complex initial data for ϕ of the type $\phi(0, q) \sim iq^2$ to give $u(0, q)$ a Gaussian profile. They are approximated by Taylor expansions. For first-order Gaussian beams, for instance, we have

$$A \approx A(t, 0) = \sqrt{c(x(t), y(t))/Q(t)}, \tag{3.13}$$

$$\phi \approx \phi(t, 0) + q\phi_q(t, 0) + \frac{q^2}{2}\phi_{qq}(t, 0) = t + \frac{q^2}{2}\,\frac{P(t)}{Q(t)}. \tag{3.14}$$

The complex-valued scalar functions P and Q satisfy the *dynamic ray tracing* system

$$\begin{aligned} \frac{dQ}{dt} &= c^2\, P, \\ \frac{dP}{dt} &= -\frac{1}{c}\left(c_{xx}\,\sin^2\theta - 2c_{xy}\,\sin\theta\,\cos\theta + c_{yy}\,\cos^2\theta\right) Q. \end{aligned} \tag{3.15}$$

As initial data for (3.15), we may choose

$$Q(0) = Q_0 > 0, \qquad P(0) = i.$$

One can show that this choice will guarantee that two important conditions are satisfied along the ray: $Q(t) \neq 0$ and $Im(P(t)/Q(t)) > 0$. The first condition guarantees the regularity of the Gaussian beam (with finite amplitudes at caustics). The second condition guarantees the concentration of the solution close to the ray. Note that for higher order Gaussian beams, we need to include more terms in the Taylor expansions and in the WKBJ expansion.

In the Gaussian beam summation method, the initial/boundary condition for the wave field is decomposed into initial conditions for Gaussian beams. Individual Gaussian beams are computed by solving the ray tracing and dynamic ray tracing systems (3.11,3.15). The contributions of the beams concentrated close to their central rays are determined by the approximations (3.13,3.14) entered in (3.12). The wave field at a receiver is then obtained by a superposition of the Gaussian beams situated close to the receiver, [67].

In paper 3, [49], we study the accuracy of Gaussian beam summation method and derive error estimates related to the Taylor expansions for beams of any order. For first-order beams, for example, we show that the error is of order $\mathcal{O}(\omega^{-1})$. In fact, because of error cancelation effects between the beams, the error is smaller than $\mathcal{O}(\omega^{-1/2})$ which a simple analysis would indicate. Moreover, we investigate the effect of beam widths on the accuracy when the speed of propagation is constant. It has been proposed that the optimal choice of the initial parameters, $Q(0)$ and $P(0)$, produce Gaussian beams of minimum width at a receiver point, see [67] for instance. The main motivation for this choice is that for wide beams the Taylor expansion error should be large. Moreover, from the computational point of view, it is more convenient to work with beams which are as narrow as possible, because in the case of variable speed of propagation, where the central rays can bend, at some distance from the rays the phase may become non-smooth and therefore the Gaussian beam approximation may break down. However, we show that this choice will not necessarily give the minimum error in the case of constant speed of propagation. The optimal choice of the parameters should minimize the error and is still an open question.

In paper 4, [50], we construct a wave front method based on Gaussian beams. The method tracks a front of Gaussian beams with only two particular initial values $(Q_1(0), P_1(0)) = (1, 0)$ and $(Q_2(0), P_2(0)) = (0, 1)$, where (Q_1, P_1) and (Q_2, P_2) solve (3.15). This allows direct recreation of any other beam propagating from the initial front into the computational domain at no extra cost. Therefore, optimization, based on the minimization of either the beam width or the error is possible in the algorithm.

Chapter 4

Summary of Papers

4.1 Paper I: A Fast Phase Space Method for Computing Creeping Rays

In this paper, we consider creeping ray contributions to high frequency scattering problems. We assume that the scatterer surface can be represented by a single parameterization and present a new Eulerian formulation for the problem. Following the discussions in Section 3, we derive a set of *escape* partial differential equations in a three-dimensional phase space. The equations are then solved on a fixed computational grid using a version of first-order accurate fast marching algorithm. The solution to the escape equations contain information about all possible creeping rays. This information includes the phase and amplitude of the ray field, which are extracted by a fast post-processing.

We consider an application to mono-static radar cross section problems where creeping rays from all illumination angles must be computed and present the numerical results of the fast phase space method.

This paper is published in Journal of Computational Physics and has entry [47] in the bibliography.

4.2 Paper II: A Multiple-patch Phase Space Method for Computing Trajectories on Manifolds with Applications to Wave Propagation Problems

In this paper, we present a multiple-patch phase space method for computing trajectories on two-dimensional manifolds possibly embedded in a higher-dimensional

space. The dynamics of trajectories are given by systems of ordinary differential equations (ODEs). We split the manifold into multiple patches where each patch has a well-defined regular parameterization. The ODEs are formulated as *escape* equations, which are hyperbolic partial differential equations (PDEs) in a three-dimensional phase space. The escape equations are solved in each patch, individually. The solutions of individual patches are then connected using suitable inter-patch boundary conditions. Properties for particular families of trajectories are obtained through a fast post-processing.

We apply the method to two different problems: the creeping ray contribution to mono-static radar cross section computations and the multivalued travel-time of seismic waves in multi-layered media. We present numerical examples to illustrate the accuracy and efficiency of the method.

This paper is published in Communications in Mathematical Sciences and has entry [48] in the bibliography.

4.3 Paper III: Taylor Expansion Errors in Gaussian Beam Summation

In this paper, we study the accuracy of Gaussian beam summation method and derive error estimates related to the Taylor expansion of the phase and amplitude off the center of the beam. Unlike standard geometrical optics, Gaussian beams compute the correct solution of the wave field also at caustics. We show that in the case of using odd order beams, the error is smaller than a simple analysis would indicate because of error cancellation effects between the beams. Since the cancellation happens only when odd order beams are used, there is no remarkable gain in using even order beams. Moreover, in the case of constant coefficient equations, i.e. when the speed of propagation is constant, the local beam width is not a good indicator of accuracy, and there is no direct relation between the error and the beams width. We present numerical examples to verify the error estimates.

This paper has entry [49] in the bibliography.

4.4 Paper IV: A Wave Front-based Gaussian Beam Method for Computing High Frequency Waves

In this paper, we present a wave front method based on Gaussian beams for computing high-frequency wave propagation problems. The method tracks a front of Gaussian beams with two particular initial values for width and curvature which allows the direct recreation of any other beam propagating from the initial front into

the medium. This is used to approximate the field with different, optimally chosen, beams in different points on the front. The performance of the method is illustrated with two numerical examples.

This paper has entry [50] in the bibliography.

4.5 Paper V: Finite Difference Schemes for Second Order Systems Describing Black Holes

In this paper, we construct stable finite difference algorithms for second order hyperbolic systems arising in numerical relativity. We treat equations in second-order differential form without reducing them to first-order form. We apply the algorithms to a model black hole space-time consisting of a spacelike inner boundary excising the singularity, a timelike outer boundary and a horizon in between. These algorithms are implemented as stable, convergent numerical codes and their performance is compared in a 2-dimensional excision problem.

This paper is published in Journal of Physical Review D and has entry [45] in the bibliography.

4.6 Paper VI: Hyperbolic Initial Boundary Value Problems which are not Boundary Stable

In this paper, we extend the theory of boundary stable hyperbolic problems to a model problem which is not boundary stable. The Kreiss symmetrizer technique gives sharp estimates of the solution of hyperbolic initial boundary value problems including estimates at the boundaries. In this case, the problem is called boundary stable. There are, however, problems which are not boundary stable but are well-posed in a weaker sense, i.e., we can obtain energy estimates in the interior of the domain. These types of problems are important in many applications, including seismic, optical and gravitational waves. We consider a model problem which may not be boundary stable depending on the choice of boundary conditions. We show that the general theory of hyperbolic systems can be extended to this case, and the symmetrizer technique can be used to derive estimates of the solution off the boundary.

This paper has entry [46] in the bibliography.

Bibliography

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

[2] V. M. Babic and T. F. Pankratova. On discontinuities of green's function of the wave equation with variable coefficient. *Problemy Matem. Fiziki*, 6, 1973. Leningrad University, Saint-Petersburg.

[3] J.-D. Benamou. Big ray tracing: Multivalued travel time field computation using viscosity solutions of the eikonal equation. *J. Comput. Phys.*, 128(4):463–474, 1996.

[4] J.-D. Benamou. An introduction to Eulerian geometrical optics (1992-2002). *J. Sci. Comput.*, 19(1-3):63–93, 2003.

[5] J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. Springer, 1989.

[6] 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.

[7] F. Brezzi, L. D. Marini, and E. Süli. Discontinuous Galerkin methods for first-order hyperbolic problems. *M3AS*, 14(12):1893–1903, 2004.

[8] V. Červený, I. A. Molotkov, and I. Psencik. *Ray Methods in Seismology*. Univ. Karlova Press, 1977.

[9] E. Darve and P. Haveé. A fast multipole method for Maxwell equations stable at all frequencies. *Philos. Trans. R. Soc. London A*, 362:603–628, 2004.

[10] B. Engquist and O. Runborg. Multiphase computations in geometrical optics. *J. Comput. Appl. Math.*, 74:175–192, 1996.

[11] B. Engquist and O. Runborg. Computational high frequency wave propagation. *Acta Numerica*, 12:181–266, 2003.

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

[13] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational differential equations*. Studentlitteratur, 1996.

[14] L. C. Evans. *Partial differential equations*. American Mathematical Society, 1998.

[15] E. Fatemi, B. Engquist, and S. J. Osher. Numerical solution of the high frequency asymptotic expansion for the scalar wave equation. *J. Comput. Phys.*, 120(1):145–155, 1995.

[16] S. Fomel and J. A. Sethian. Fast phase space computation of multiple arrivals. *Proc. Natl. Acad. Sci. USA*, 99(11):7329–7334 (electronic), 2002.

[17] C. Geuzaine, O Bruno, and F. Reitich. On the $\mathcal{O}(1)$ solution of multiple-scattering problems. *IEEE Transactions on Magnetics*, 41(5):1488–1491, 2005.

[18] W. Gibson. *The Method of Moments in Electromagnetics*. Chapman and Hall/CRC, 2008.

[19] B. Gustafsson, H. O. Kreiss, and J. Ologer. *Time dependent problems and difference methods*. Wiley-Interscience, 1995.

[20] S. Hagdahl. *Hybrid Methods for Computational Electromagnetics in Frequency Domain*. PhD thesis, NADA, KTH, Stockholm, 2005.

[21] R. F. Harrington. *Field Computation by Moment Methods*. MacMillan, New York, 1968.

[22] L. Hörmander. *The analysis of linear partial differential operators. I-IV*. Springer-Verlag, Berlin, 1983–1985.

[23] P. E. Hussar, V. Oliker, H. L. Riggins, E.M. Smith-Rowlan, W.R. Klocko, and L. Prussner. An implementation of the UTD on facetized CAD platform models. *IEEE Antennas Propag.*, 42(2):100–106, 2000.

[24] S. Jin. Recent computational methods for high frequency waves in heterogeneous media. Preprint, 2008.

[25] S. Jin, H. Liu, S. Osher, and R. Tsai. Computing multi-valued physical observables for the high frequency limit of symmetric hyperbolic systems. *J. Comput. Phys.*, 210(2):497–518, 2005.

[26] S. Jin and S. Osher. A level set method for the computation of multivalued solutions to quasi-linear hyperbolic PDEs and Hamilton-Jacobi equations. *Commun. Math. Sci.*, 1(3):575–591, 2003.

[27] S. Jin and X. Wen. Hamiltonian-preserving schemes for the Liouville equation with discontinuous potentials. *Commun. Math. Sci.*, 3(3):285–315, 2005.

[28] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic problems. *Comput. Methods Appl. Mech. Engrg.*, 45:285–312, 1985.

[29] B. R. Julian and D. Gubbins. Three-dimensional seismic ray tracing. *J. Geophys. Res.*, 43:95–114, 1977.

[30] J. Keller and R. M. Lewis. Asymptotic methods for partial differential equations: the reduced wave equation and Maxwell's equations. *Surveys Appl. Math.*, 1:1–82, 1995.

[31] J. B. Keller. The geometric theory of diffraction. In *Symposium on Microwave Optics*, Eaton Electronics Research Laboratory, McGill University, Montreal, Canada, June 1953.

[32] S. Kim. An $\mathcal{O}(N)$ level set method for eikonal equations. *SIAM J. Sci. Comput.*, 22(6):2178–2193, 2000.

[33] S. Kim and R. Cook. 3-D traveltime computation using second-order ENO scheme. *Geophysics*, 64(6):1867–1876, 1999.

[34] R. Kimmel and J. A. Sethian. Computing geodesic paths on manifolds. *Proc. Natl. Acad. Sci. USA*, 95(15):8431–8435 (electronic), 1998.

[35] 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.

[36] Yu. A. Kravtsov. On a modification of the geometrical optics method. *Izv. VUZ Radiofiz.*, 7(4):664–673, 1964.

[37] H.-O. Kreiss and J. Lorenz. *Initial-boundary value problems and the Navier-Stokes equations*. SIAM, 2004.

[38] R. T. Langan, I. Lerche, and R. T. Cutler. Tracing of rays through heterogeneous media: An accurate and efficient procedure. *Geophysics*, 50:1456–1465, 1985.

[39] R. J. Leveque. *Finite volume methods for hyperbolic problems*. Cambridge University Press, 2002.

[40] C. D. Levermore. Moment closure hierarchies for kinetic theories. *J. Statist. Phys.*, 83(5-6):1021–1065, 1996.

[41] H. Liu, S. Osher, and R. Tsai. Multi-valued solution and level set methods in computational high frequency wave propagation. *Commun. Comput. Phys.*, 1(5):765–804, 2006.

[42] D. Ludwig. Uniform asymptotic expansions at a caustic. *Comm. Pure Appl. Math.*, 19:215–250, 1966.

[43] L. N. Medgyesi-Mitschang and D.-S. Wang. Hybrid methods for analysis of complex scatterers. *P. IEEE*, 77(5):770–779, 1989.

[44] M. Motamed. Phase space methods for computing creeping rays. Licentiate thesis, KTH, 2006. ISBN 91-7178-467-5.

[45] M. Motamed, M. Babiuc, B. Szilagyi, H-O. Kreiss, and J. Winicour. Finite difference schemes for second order systems describing black holes. *Commun. Math. Sci.*, 73(12), 2006. Appended as paper V.

[46] M. Motamed and H-O. Kreiss. Hyperbolic initial boundary value problems which are not boundary stable. Preprint, 2008. Appended as paper VI.

[47] M. Motamed and O. Runborg. A fast phase space method for computing creeping rays. *J. Comput. Phys.*, 219(1):276–295, 2006. Appended as paper I.

[48] M. Motamed and O. Runborg. A multiple-patch phase space method for computing trajectories on manifolds with applications to wave propagation problems. *Commun. Math. Sci.*, 5(3):617–648, 2007. Appended as paper II.

[49] M. Motamed and O. Runborg. Taylor expansion errors in Gaussian beam summation. Preprint, 2008. Appended as paper III.

[50] M. Motamed and O. Runborg. A wave front-based Gaussian beam method for computing high frequency waves. Preprint, 2008. Appended as paper IV.

[51] S. J. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.*, 79(1):12–49, 1988.

[52] S. J. Osher and C.-W. Shu. High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 28(4):907–922, 1991.

[53] M. M. Popov. A new method of computation of wave fields using gaussian beams. *Wave Motion*, 4:85–97, 1982.

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

[55] P. A. Raviart and J. M. Thomas. *A mixed finite element method for 2nd order elliptic problems*, volume 606 of *Lecture Notes in Math.*, pages 292–315. Springer-Verlag, New York, 1977.

[56] O. Runborg. Some new results in multiphase geometrical optics. *M2AN Math. Model. Numer. Anal.*, 34:1203–1231, 2000.

[57] W. A. Schnedier, K. A. Ranzinger, A. H. Balch, and C. Kruse. A dynamic programming approach to first arrival traveltime computation in media with arbitrary distributed velocities. *Geophysics*, 57(1):39–50, 1992.

[58] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proc. Nat. Acad. Sci. U.S.A.*, 93(4):1591–1595, 1996.

[59] J. A. Sethian. *Level set methods and fast marching methods*. Cambridge University Press, Cambridge, second edition, 1999. Evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science.

[60] Chi-Wang Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. ICASE Report 97-65, Brown University, 1997. Prepared for NASA Langley Research Center.

[61] J. C. Strikwerda. *Finite difference schemes and partial differential equations*. Wadsworth and Brooks/Cole, 1989.

[62] W. W. Symes. A slowness matching finite difference method for traveltimes beyond transmission caustics. Preprint, Dept. of Computational and Applied Mathematics, Rice University, 1996.

[63] W. W. Symes and J. Qian. A slowness matching Eulerian method for multivalued solutions of eikonal equations. *J. Sci. Comput.*, 19(1-3):501–526, 2003.

[64] A.-K. Tornberg and B. Engquist. The segment projection method for interface tracking. *Comm. Pure Appl. Math.*, 56(1):47–79, 2003.

[65] Y. R. Tsai, L. T. Cheng, S. Osher, and H. K. Zhao. Fast sweeping algorithms for a class of Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 41(2):673–694 (electronic), 2003.

[66] J. N. Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE Trans. Automat. Control*, 40(9):1528–1538, 1995.

[67] M. M. Popov V. Cerveny and I. Psencik. Computation of wave fields in inhomogeneous media - gaussian beam approach. *Geophys. J. R. Astr. Soc.*, 70:109–128, 1982.

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

[69] J. Vidale. Finite-difference calculation of traveltimes. *B. Seismol. Soc. Am.*, 78(6):2062–2076, December 1988.

[70] V. Vinje, E. Iversen, and H. Gjøystdal. Traveltime and amplitude estimation using wavefront construction. In *Eur. Ass. Expl. Geophys.*, pages 504–505, 1992. Extended abstracts.

[71] V. Vinje, E. Iversen, and H. Gjøystdal. Traveltime and amplitude estimation using wavefront construction. *Geophysics*, 58(8):1157–1166, 1993.

[72] V.Rokhlin. Rapid solution of integral equations for scattering theory in two dimensions. *J. Comput. Phys.*, 86:414–439, 1990.

[73] D. R. Wilton, S. M. Rao, and A. W. Glisson. Electromagnetic scattering by surfaces of arbitrary shape. *IEEE T. Antenn. Propag.*, 30:409–418, 1982.

[74] K. S. Yee. Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media. *IEEE T. Antenn. Propag.*, 14(3):302–307, 1966.

[75] L. Ying and E. J. Candes. Fast geodesics computation with the phase flow method. *J. Comput. Phys.*, 220(1):6–18, 2006.

[76] L. Ying and E. J. Candes. The phase flow method. *J. Comput. Phys.*, 220(1):184–215, 2006.

Paper I

Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 219 (2006) 276–295
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

A fast phase space method for computing creeping rays

Mohammad Motamed *, Olof Runborg

Department of Numerical Analysis and Computer Science, Royal Institute of Technology (KTH), Lindstadsvagen 3, 10044 Stockholm, Sweden

Received 17 October 2005; received in revised form 20 March 2006; accepted 21 March 2006
Available online 22 May 2006

Abstract

Creeping rays can give an important contribution to the solution of medium to high frequency scattering problems. They are generated at the shadow lines of the illuminated scatterer by grazing incident rays and propagate along geodesics on the scatterer surface, continuously shedding diffracted rays in their tangential direction.

In this paper, we show how the ray propagation problem can be formulated as a partial differential equation (PDE) in a three-dimensional phase space. To solve the PDE we use a fast marching method. The PDE solution contains information about all possible creeping rays. This information includes the phase and amplitude of the field, which are extracted by a fast post-processing. Computationally, the cost of solving the PDE is less than tracing all rays individually by solving a system of ordinary differential equations.

We consider an application to mono-static radar cross section problems where creeping rays from all illumination angles must be computed. The numerical results of the fast phase space method and a comparison with the results of ray tracing are presented.

Keywords: Creeping rays; High frequency wave propagation; Scattering problems; Numerical methods; Geometrical theory of diffraction; Eikonal equation

1. Introduction

The general problem that we are interested in is the scattering of a time-harmonic incident field by a bounded scatterer D. If the total field is split into an incident and a scattered field, this can be formulated as a boundary value problem for the scattered field in the region outside D, consisting of the Helmholtz equation,

$$\Delta W + n(\mathbf{x})^2\omega^2 W = 0, \quad \mathbf{x} \in \mathbb{R}^3 \setminus \bar{D}, \tag{1}$$

augmented with Dirichlet, Neumann or Robin boundary conditions on the boundary of the scatterer ∂D, and the Sommerfeld radiation condition at infinity. Here $n(\mathbf{x})$ is the index of refraction, and ω is the angular frequency.

* Corresponding author.
E-mail addresses: mohamad@nada.kth.se (M. Motamed), olofr@nada.kth.se (O. Runborg).

0021-9991/$ - see front matter
doi:10.1016/j.jcp.2006.03.024

In direct numerical simulations of (1) the accuracy of the solution is determined by the number of grid points or elements per wave length. The computational cost to maintain constant accuracy grows algebraically with the frequency, and for sufficiently high frequencies, a direct numerical simulation is no longer feasible. Numerical methods based on approximations of (1) are needed.

Fortunately, there exist good such approximations precisely for the difficult case of high frequency solutions. In free space, a typical high frequency solution can be approximated by a simple wave,

$$W(\mathbf{x}) \approx a(\mathbf{x})\mathrm{e}^{\mathrm{i}\omega\phi(\mathbf{x})}, \quad \mathbf{x} \in \mathbb{R}^3, \tag{2}$$

where the amplitude $a(\mathbf{x})$ and the phase function $\phi(\mathbf{x})$ depend only mildly on the parameter ω and vary on a much coarser scale than $W(\mathbf{x})$ itself. Geometrical optics (GO) considers the case when $\omega \to \infty$. The frequency then disappears from the model and the equations can be solved at a computational cost independent of ω. GO can be formulated as the partial differential equations for ϕ and a. The phase function ϕ satisfies the *eikonal equation*,

$$|\nabla\phi| = n(x), \tag{3}$$

and the leading order amplitude term a satisfies the *transport equation*,

$$2\nabla\phi \cdot \nabla a + \Delta\phi a = 0. \tag{4}$$

GO can also be formulated in terms of ordinary differential equations (ODE). It corresponds to solving the eikonal equation (3) through the method of characteristics, i.e. solving the system of ODEs,

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \nabla_p H(\mathbf{x},\mathbf{p}), \quad \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\nabla_x H(\mathbf{x},\mathbf{p}), \quad H(\mathbf{x},\mathbf{p}) = \frac{|\mathbf{p}|}{n(\mathbf{x})}, \tag{5}$$

where t is time. As long as ϕ is smooth, the relationship between the models is given by $\phi(\mathbf{x}(t)) = \phi(\mathbf{x}(0)) + t$. There are also ODEs giving the amplitude $a(\mathbf{x}(t))$ along the characteristics.

The main drawbacks of the infinite frequency approximation of geometrical optics are that diffraction effects at boundaries are lost, and that the approximation breaks down at caustics, where the predicted amplitude a is unbounded. Geometrical theory of diffraction (GTD), pioneered by J. Keller in the 1950s [14], adds diffraction effects to the GO approximations. One type of diffracted rays are *creeping rays*, which are generated at the *shadow line* of the scatterer, i.e. where the incident ray strikes the surface of the scatterer at grazing angle. At this point the incident ray divides into two parts: one part continues straight on, and a second part propagates along geodesics on the surface, continuously shedding diffracted rays in its tangential direction. See Fig. 1. In analogy with (2), a wave field is generated on the surface

$$W_{\mathrm{s}}(\mathbf{u}) = a(\mathbf{u})\mathrm{e}^{\mathrm{i}\omega\phi(\mathbf{u})}, \tag{6}$$

where $\phi(\mathbf{u})$ and $a(\mathbf{u})$ are now the surface phase and amplitude and $\mathbf{u} \in \mathbb{R}^2$ is a parameterization of the surface. The creeping rays satisfy a system of ODEs similar to (5). They are related to (6) in the same way as the standard GO rays are related to (2).

Creeping rays can give an important contribution to the solution at medium to high frequencies, for instance in radar cross section (RCS) computations for low observable objects [3] and in antenna coupling problems [16]. We want to compute the creeping rays and the associated wave field in (6).

Various methods have been devised to compute the geometrical optics solution. They can be divided into Lagrangian and Eulerian methods.

Lagrangian methods are based on the ODE formulation (5). The simplest Lagrangian method is standard ray tracing where the ODEs in (5) together with ODEs for the amplitude are solved directly with numerical methods for ODEs. This approach is very common in standard free space GO, [4,19], but is also done for the creeping ray case, [12,22]. Ray tracing gives the phase and amplitude solution along a ray, and interpolation must be applied to obtain those quantities everywhere. This can be rather difficult, in particular in regions where rays cross. Another problem with ray tracing is that it may produce diverging rays that fail to cover the domain. Even for smooth $n(\mathbf{x})$ there may be shadow zones where the field is hard to resolve. The interpolation can be simplified by instead using so-called wave front methods [30,11]. They are related to ray tracing, but instead of individual rays, an interface representing a wave front is evolved according to the ray equations.

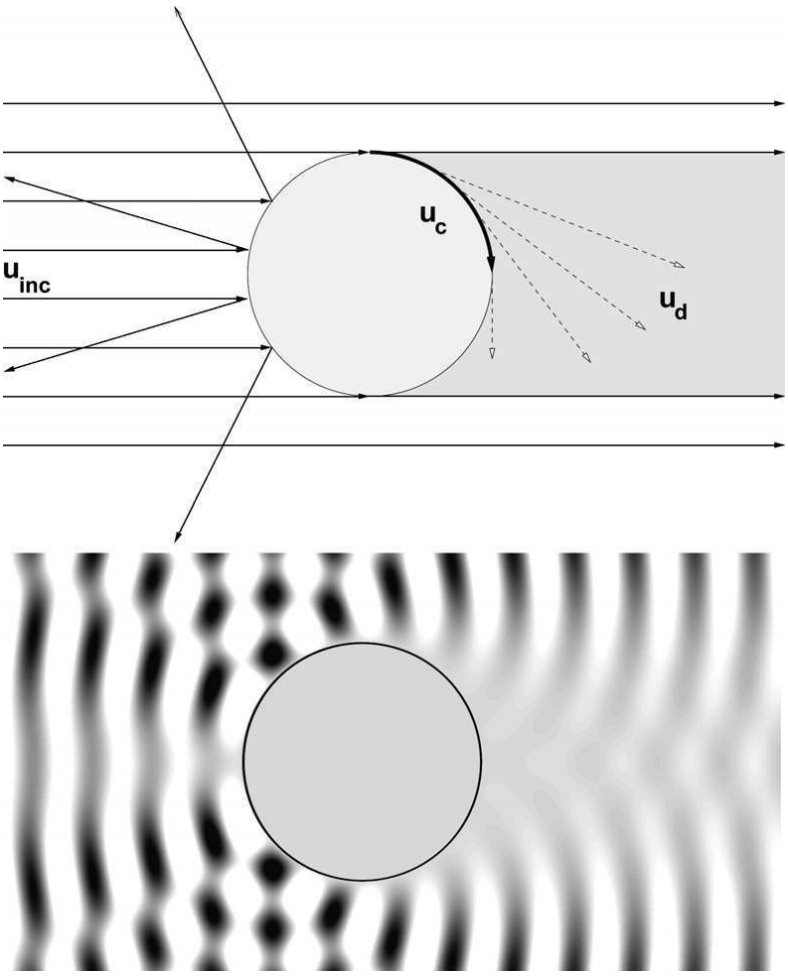


Fig. 1. Diffraction by a smooth cylinder. Top figure shows the solution schematically. The incident field u_{inc} induces a creeping ray u_{c} at the north (and south) pole of the cylinder, where the incident direction is orthogonal to the surface normal. As the creeping ray propagates along the surface, it continuously emits surface-diffracted rays u_{d} with exponentially decreasing initial amplitude. Bottom figure shows real part of a solution to the Helmholtz equation. The surface diffracted waves can be seen behind the cylinder.

More recently, Eulerian methods based on PDEs have been proposed to avoid some of the drawbacks of ray tracing. These methods discretize the PDEs on fixed computational grids to control accuracy everywhere and there is no need for interpolation. The simplest Eulerian methods solves the eikonal and transport equations (3,4). This technique has been used in standard GO, [29,28,7] and also in the surface case, [15]. However, the eikonal and transport equations only give the correct solution when it is a single wave of the form (2). When there are crossing waves, more elaborate schemes must be devised. In the free space GO case a number of methods have been developed in the last ten years using different approaches. Many of them are based on a third formulation of geometrical optics as a kinetic equation set in phase space. They include "big" ray tracing [1], patching together multiple eikonal solutions [2], moment methods [24,25,9], segment projection method [6], level set methods [21,23], slowness matching [26], the phase flow method [31] and fast phase space methods [8]. A survey of this research effort is given in [5].

These more advanced methods have so far not been used for the creeping ray case. In this paper we propose an adaptation of the fast phase space method of Fomel and Sethian [8] to this case. This method is computationally expensive if only a few solutions are computed. It becomes attractive when the solution is sought for many different sources but with the same index of refraction. In the creeping ray case this happens for instance when the solution for all illumination angles of a fixed scatterer is of interest. We consider one such example: computing the mono-static RCS.

Following [8] we formulate the ray propagation problem as a time-independent partial differential equation (PDE) in a three-dimensional phase space. We use a fast marching method to solve the PDE. The PDE solution contains information for all incidence angles. The phase and amplitude of the field are extracted by a fast post-processing. Computationally the cost of solving the PDE is less than tracing all rays individually. If the surface is discretized by N^2 points the complexity is $\mathcal{O}(N^3 \log N)$, while ray tracing would cost $\mathcal{O}(N^4)$ if a comparable number of incidence angles (N^2) and rays per angle (N) are considered.

In Section 2, we formulate the governing equations. The numerical method for solving the equations are discussed in Section 3. In Section 4, we show how to extract the information for a particular ray through post-processing. An application to a mono-static RCS problem is shown as an example in Section 5.

2. Governing equations

For simplicity we consider the case when the scatterer surface has an explicit parameterization. Let $\overline{X}$ be a regular hypersurface, representing a scatterer surface, with the parametric equations $\overline{X} = \overline{X}(\mathbf{u})$, where $\overline{X} = (x, y, z) \in \mathbb{R}^3$ is the coordinate in 3D physical space, and the parameters $\mathbf{u} = (u, v)$ belong to a bounded set $\Omega \subset \mathbb{R}^2$. Let the scatterer be illuminated by incident rays in a direction represented by a normalized vector $\widehat{I} = [\iota_1, \iota_2, \iota_3]$. The shadow line is then defined as the set of points where

$$\widehat{N}^\top \widehat{I} = 0, \tag{7}$$

where $\widehat{N}(\mathbf{u})$ is the surface normal at $\overline{X}(\mathbf{u})$,

$$\widehat{N} = \frac{\overline{X}_u \times \overline{X}_v}{|\overline{X}_u \times \overline{X}_v|}. \tag{8}$$

Here the subscripts denote differentiation with respect to u and v. We will assume that (7) defines a curve in parameter space, which we denote $\mathbf{u}_0(s)$, and s is the arc length parameterization.

2.1. Geodesics

We start by deriving the equations for creeping rays, which are indeed geodesics on the scatterer surface. According to Keller and Lewis [13], the surface phase satisfies the *surface eikonal equation*,

$$|\widetilde{\nabla}\phi| = n, \tag{9}$$

where $n(\mathbf{u})$ is the index of refraction at the surface, and $\widetilde{\nabla}$ is the surface gradient, defined as

$$\widetilde{\nabla}\phi := JG^{-1}\nabla\phi, \quad G = J^\top J,$$

with

$$J = [\overline{X}_u \overline{X}_v] \in \mathbb{R}^{3\times 2}.$$

We prescribe boundary conditions for (9) on the shadow line, which acts as the source for the creeping rays. The boundary condition is that the surface phase agrees with ϕ_{inc}, the phase of the incoming wave,

$$\phi(\mathbf{u}_0(s)) = \phi_0(\mathbf{u}_0) := \phi_{\text{inc}}(\overline{X}(\mathbf{u}_0(s))), \tag{10}$$

To avoid ambiguities as to which direction the surface waves propagate, we add the condition

$$\widetilde{\nabla}\phi(\mathbf{u}_0(s)) = \nabla\phi_{\text{inc}}(\overline{X}(\mathbf{u}_0(s))), \tag{11}$$

which is consistent with (9) since ϕ_{inc} satisfies the free space eikonal equation (3) and with (10) since

$$\frac{\mathrm{d}}{\mathrm{d}s}(\phi(\mathbf{u}_0(s)) - \phi_{\text{inc}}(\overline{X}(\mathbf{u}_0(s)))) = \nabla\phi^\top \mathbf{u}_0' - \nabla\phi_{\text{inc}}^\top \frac{\mathrm{d}\overline{X}}{\mathrm{d}s} = (J^\top \widetilde{\nabla}\phi)^\top \mathbf{u}_0' - \nabla\phi_{\text{inc}}^\top \frac{\mathrm{d}\overline{X}}{\mathrm{d}s} = (\widetilde{\nabla}\phi - \nabla\phi_{\text{inc}})^\top \frac{\mathrm{d}\overline{X}}{\mathrm{d}s}.$$

In the case when $n = 1$ and the incoming wave is a plane wave in direction $\widehat{I}$, we have $\phi_{\text{inc}}(\mathbf{x}) = \widehat{I}^\top \mathbf{x}$. Then (10), (11) reduce to

$$\phi_0(\mathbf{u}_0(s)) := \widehat{I}^\top \overline{X}(\mathbf{u}_0(s)), \quad \widetilde{\nabla}\phi(\mathbf{u}_0(s)) = \widehat{I}. \tag{12}$$

We can write (9) as a Hamilton–Jacobi equation $H(\mathbf{u},\nabla\phi) = 0$, with the Hamiltonian

$$H(\mathbf{u},\mathbf{p}) \equiv \frac{1}{2}\mathbf{p}^\top G^{-1}(\mathbf{u})\mathbf{p} - \frac{n^2(\mathbf{u})}{2}.$$

Note that in the case $n = $ constant, the geometrical rays associated with the eikonal equation (3) becomes straight lines. Analogously, for the surface eikonal equation (9), the creeping rays for constant n are geodesics, or shortest paths between two points on the surface. Henceforth, we will assume $n \equiv 1$ and a plane incoming wave.

Introducing a parameter τ, the bicharacteristics $(\mathbf{u}(\tau), \mathbf{p}(\tau))$ are determined by the solution of the following Hamiltonian equations

$$\dot{\mathbf{u}} = H_{\mathbf{p}} = G^{-1}\mathbf{p}, \tag{13a}$$

$$\dot{\mathbf{p}} = -H_{\mathbf{u}}. \tag{13b}$$

Here the dot denotes differentiation with respect to the parameter τ. At the shadow line, the initial direction of the geodesic should be parallel to the incident field. We demand that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\overline{X}(\mathbf{u}(\tau))\bigg|_{\tau=0} = \widehat{I}.$$

This implies that $\mathbf{p}(0) = G\dot{\mathbf{u}}(0) = J^\top J\dot{\mathbf{u}}(0) = J^\top\dot{\overline{X}}(0) = J^\top\widehat{I}$. The initial condition for the system (13) therefore reads,

$$\mathbf{u}(0) = \mathbf{u}_0(s), \tag{14a}$$

$$\mathbf{p}(0) = \mathbf{p}_0(s) := J^\top(\mathbf{u}_0(s))\widehat{I}. \tag{14b}$$

We note that by (12),

$$\mathbf{p}(0) = J^\top(\mathbf{u}_0(s))\widetilde{\nabla}\phi(\mathbf{u}_0(s)) = J^\top JG^{-1}\nabla\phi(\mathbf{u}_0(s)) = \nabla\phi(\mathbf{u}(0)).$$

As for any Hamiltonian system it therefore follows that

$$\mathbf{p}(\tau) = \nabla\phi(\mathbf{u}(\tau)), \tag{15}$$

for all $\tau \geqslant 0$, as long as ϕ is smooth. As a consequence, (13) and (15) give

$$|\dot{\overline{X}}| = \left|\frac{\mathrm{d}\overline{X}}{\mathrm{d}\tau}\right| = |J\dot{\mathbf{u}}| = |JH_{\mathbf{p}}| = |JG^{-1}\mathbf{p}| = 1, \tag{16}$$

and we can identify the parameter τ with arc length along the creeping rays $\overline{X}(\mathbf{u}(\tau))$. In this case, the system of four first-order ODEs (13) can be written as a system of two second-order equations [13],

$$\ddot{u} + \Gamma^1_{11}\dot{u}^2 + 2\Gamma^1_{12}\dot{u}\dot{v} + \Gamma^1_{22}\dot{v}^2 = 0, \tag{17a}$$

$$\ddot{v} + \Gamma^2_{11}\dot{u}^2 + 2\Gamma^2_{12}\dot{u}\dot{v} + \Gamma^2_{22}\dot{v}^2 = 0. \tag{17b}$$

Here $\Gamma^k_{ij}(\mathbf{u})$ are Christoffel symbols, defined by

$$\Gamma^k_{ij} = \sum_{m=1}^{2}\frac{1}{2}g^{km}[(g_{jm})_i + (g_{im})_j - (g_{ji})_m],$$

where $(g_{ij}) = G$ and $(g^{ij}) = G^{-1}$, and subscripts 1 and 2 denote differentiation with respect to u and v, respectively.

Now if we set $\dot{u} = \frac{\mathrm{d}u}{\mathrm{d}\tau} = \rho\cos\theta$ and $\dot{v} = \frac{\mathrm{d}v}{\mathrm{d}\tau} = \rho\sin\theta$, then $\dot{v} = \dot{u}\tan\theta$, and by differentiating with respect to τ,

$$\ddot{v} = \ddot{u}\tan\theta + \dot{u}\frac{1}{\cos^2\theta}\dot{\theta}. \tag{18}$$

Moreover by (16),

$$\rho = \rho(u, v, \theta) = \left| J \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \right|^{-1} = |\overline{X}_u \cos\theta + \overline{X}_v \sin\theta|^{-1}.$$

Let $\gamma := (u, v, \theta)$. Using (18), we get

$$\dot{\theta} = \rho(\gamma)\mathscr{V}(\gamma),$$

where

$$\mathscr{V}(\gamma) := (\Gamma^1_{11}\cos^2\theta + 2\Gamma^1_{12}\cos\theta\sin\theta + \Gamma^1_{22}\sin^2\theta)\sin\theta - (\Gamma^2_{11}\cos^2\theta + 2\Gamma^2_{12}\cos\theta\sin\theta + \Gamma^2_{22}\sin^2\theta)\cos\theta.$$

Therefore the system of ODEs (17), for geodesics, reduces to

$$\begin{pmatrix} \dot{u} \\ \dot{v} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} \rho(\gamma)\cos\theta \\ \rho(\gamma)\sin\theta \\ \rho(\gamma)\mathscr{V}(\gamma) \end{pmatrix} =: \mathbf{g}(\gamma). \tag{19}$$

2.2. Phase and amplitude

Let us now derive the ODEs for the surface phase ϕ and amplitude a. As before, we parametrize the creeping ray with the arc length τ in the physical space. In the surface field associated with the creeping ray (6), the phase function $\phi(\mathbf{u}(\tau))$ and the amplitude $a(\mathbf{u}(\tau))$ of the field vary with the distance τ along the ray.

From (13) and (15) it follows that the phase of the geodesic satisfies the ODE,

$$\frac{\mathrm{d}\phi(\mathbf{u}(\tau))}{\mathrm{d}\tau} = \nabla\phi \cdot \dot{\mathbf{u}} = \nabla\phi \cdot G^{-1}\nabla\phi = |\widetilde{\nabla}\phi|^2 = 1, \quad \phi(0) = \phi_0(\mathbf{u}_0). \tag{20}$$

Hence, the phase is the length of the ray.

Now consider a narrow strip of a creeping ray, starting at the incident point Q_0 on the shadow line and propagating along a geodesic on the scatterer surface. See Fig. 2.

To determine an equation for the amplitude, we apply the optical form of energy conservation principle in a small interval from τ to $\tau + \mathrm{d}\tau$, [18], and get

$$\frac{\mathrm{d}}{\mathrm{d}\tau}[a(\tau)^2 \mathrm{d}\sigma(\tau)] = -2\alpha(\tau)[a(\tau)^2 \mathrm{d}\sigma(\tau)], \tag{21}$$

where $\mathrm{d}\sigma(\tau)$ is the width of the strip at distance τ from Q_0, and $\alpha(\tau)$ is an attenuation factor. Solving (21) gives us

$$a(\tau) = a_0 \left(\frac{\mathrm{d}\sigma_0}{\mathrm{d}\sigma}\right)^{\frac{1}{2}} \exp\left(-\int_0^\tau \alpha(r)\mathrm{d}r\right), \tag{22}$$

where a_0 and $\mathrm{d}\sigma_0$ are the amplitude and strip width at Q_0, respectively. There are thus two parts in this equation which we can treat separately: the attenuation, represented by the exponential, and the geometrical spreading of the creeping ray, represented by $\frac{\mathrm{d}\sigma}{\mathrm{d}\sigma_0}$.

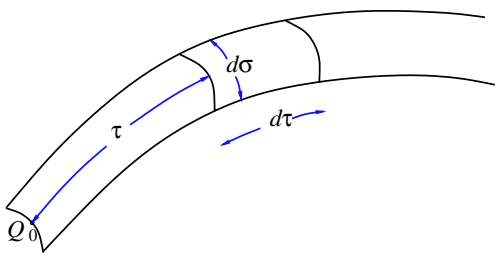


Fig. 2. A narrow strip of a creeping ray on the surface.

2.2.1. Attenuation

We will here show that the attenuation can be obtained by solving an ODE coupled to the geodesic system (19).

The attenuation factor α is given by [18,20],

$$\alpha = \frac{q_0}{\rho_g}\exp\left(\mathrm{i}\frac{\pi}{6}\right)\left(\frac{\omega\rho_g}{2}\right)^{1/3} := \omega^{1/3}\widetilde{\alpha}.$$

Here $q_0 \approx 2.33811$ is the smallest positive zero of the Airy function, and ρ_g is the radius of curvature of the surface with respect to arc length along the ray trajectory, given by [10],

$$\rho_g = \frac{1}{-\widehat{T}^\top D_{\mathbf{u}}\widehat{N}\dot{\mathbf{u}}}, \quad \widehat{T} = \frac{\mathrm{d}\overline{X}}{\mathrm{d}\tau}(\mathbf{u}(\tau)) = J\dot{\mathbf{u}}.$$

Here, $\widehat{T}$ is the tangent vector to the surface in the geodesics direction, and $D_{\mathbf{u}}\widehat{N} = [\widehat{N}_u\widehat{N}_v]$ is the Jacobian of the normal vector $\widehat{N}$. Note that $|\widehat{T}| = 1$ by (16). Since $\widehat{T}$, $\widehat{N}$ and $\dot{\mathbf{u}}$ are functions of (u, v, θ), so is $\widetilde{\alpha} = \widetilde{\alpha}(u, v, \theta)$. We can therefore add the ODE

$$\frac{\mathrm{d}\beta}{\mathrm{d}\tau} = \widetilde{\alpha}(u, v, \theta), \quad \beta(0) = 0, \tag{23}$$

to the geodesic system (19), and then express the attenuation as

$$\exp\left(-\int_0^\tau \alpha(r)\mathrm{d}r\right) = \exp(-\omega^{1/3}\beta(\tau)).$$

Note that β is independent of the frequency ω.

2.2.2. Geometrical spreading

To compute the amplitude of the creeping ray from (22), we also need to compute the geometrical spreading. We consider again a narrow strip of a geodesics, as in Fig. 2, and let $\mathrm{d}\sigma_0(s)$ and $\mathrm{d}\sigma(s,\tau)$ be the strip width at the shadow line and at the distance τ from the shadow line, respectively.

Set $\widetilde{\mathbf{u}}(s,\tau) := \mathbf{u}(\tau)$, where $(\mathbf{u}(\tau), \mathbf{p}(\tau))$ is a solution to (13) with the initial data (14) so that $\widetilde{\mathbf{u}}(s, 0) = \mathbf{u}_0(s)$. Moreover, let

$$\widetilde{X}(s,\tau) := \overline{X}(\widetilde{\mathbf{u}}(s,\tau)).$$

Then $\widetilde{X}$ is the point on the geodesic at the distance τ from the shadow line, and $\widetilde{X}_0(s) = \widetilde{X}(s, 0)$ is the starting point on the shadow line. Denote the geometrical spreading of the creeping ray at the point $\widetilde{X}(s,\tau)$ in the physical space by

$$\mathcal{Q}(s,\tau) := \frac{\mathrm{d}\sigma(s,\tau)}{\mathrm{d}\sigma_0(s)}.$$

Moreover, let $\mathrm{d}\sigma_0'$ and $\mathrm{d}\sigma'$ be the strip width in the direction of the shadow line, defined by $\mathrm{d}\sigma_0' = |\widetilde{X}_{0s}|\mathrm{d}s$ and $\mathrm{d}\sigma' = |\widetilde{X}_s|\mathrm{d}s$. See Fig. 3. Then we have

$$\cos\beta_0 = \frac{\mathrm{d}\sigma_0}{\mathrm{d}\sigma_0'} = \frac{\widetilde{X}_{0\tau}^\perp}{|\widetilde{X}_{0\tau}^\perp|}\cdot\frac{\widetilde{X}_{0s}}{|\widetilde{X}_{0s}|}, \tag{24}$$

$$\cos\beta = \frac{\mathrm{d}\sigma}{\mathrm{d}\sigma'} = \frac{\widetilde{X}_\tau^\perp}{|\widetilde{X}_\tau^\perp|}\cdot\frac{\widetilde{X}_s}{|\widetilde{X}_s|}, \tag{25}$$

where the τ- and s-subscripts denote differentiation along the ray and the shadow line, respectively, and $\widetilde{X}_\tau^\perp$ is orthogonal to $\widetilde{X}_\tau$ in the tangent plane to the surface. Since $|\widetilde{X}_{0\tau}^\perp| = |\widetilde{X}_\tau^\perp| = 1$ by (16), the geometrical spreading is then computed as,

$$\mathcal{Q}(s,\tau) = \frac{\widetilde{X}_\tau^\perp\cdot\widetilde{X}_s}{\widetilde{X}_{0\tau}^\perp\cdot\widetilde{X}_{0s}}. \tag{26}$$

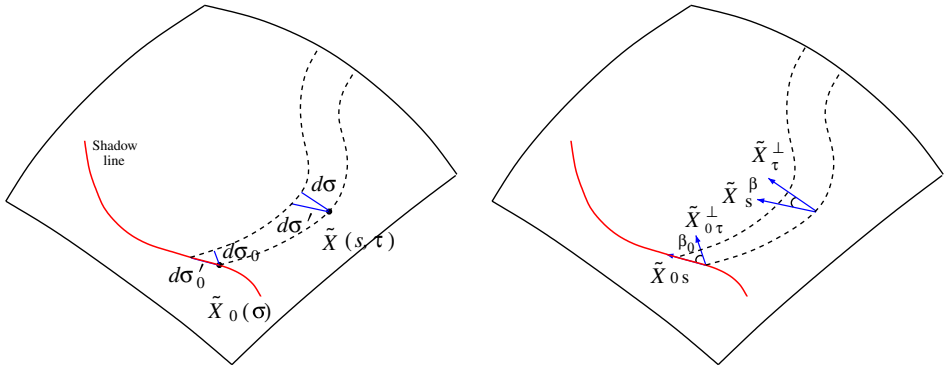


Fig. 3. Geometrical spreading of a creeping ray on the surface, starting at the shadow line and ending at the boundary.

We will show how to calculate the right hand side of (26) numerically, below.

2.3. Eulerian formulation

There are a number of drawbacks with Lagrangian methods based on solving the ODEs (19), (20) and (23). In particular, in the regions where rays diverge or cross, interpolation can be difficult. Instead, we use an Eulerian formulation and derive time-independent PDEs, which can be solved on a fixed computational grid.

We introduce the phase space $\mathbb{P} = \mathbb{R}^2 \times \mathbb{S}$, where $\mathbb{S}$ is the periodic sphere. We consider the triplet $\gamma = (u, v, \theta)$ as a point in this space. The geodesics on the scatterer are then confined to a subdomain $\Omega_p = \Omega \times \mathbb{S} \subset \mathbb{P}$ in phase space.

Let us now introduce an unknown function $F : \mathbb{P} \to \mathbb{P}$,

$$F(\gamma) = \begin{pmatrix} U(\gamma) \\ V(\gamma) \\ \Theta(\gamma) \end{pmatrix}, \tag{27}$$

which is the point where the geodesic starting at $\mathbf{u} = (u, v) \in \Omega$ with direction $\theta \in \mathbb{S}$ will cross the boundary of Ω_p. See Fig. 4. Since F is constant along a geodesic, we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}\tau} F(u(\tau), v(\tau), \theta(\tau)) = \frac{\mathrm{d}u}{\mathrm{d}\tau} F_u + \frac{\mathrm{d}v}{\mathrm{d}\tau} F_v + \frac{\mathrm{d}\theta}{\mathrm{d}\tau} F_\theta. \tag{28}$$

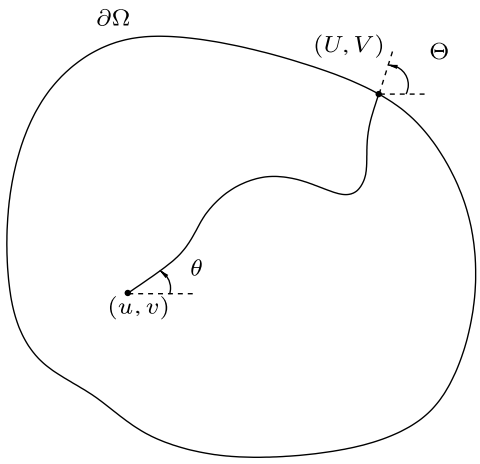


Fig. 4. A geodesic in the parameter space. The function F is defined as $F(u, v, \theta) = (U, V, \Theta)$, with the notation as in the figure.

Using (28) and (19), we can write the *escape* PDE for F as

$$\cos\theta F_u + \sin\theta F_v + \mathscr{V}(\gamma)F_\theta = 0, \quad \gamma \in \Omega_p, \tag{29}$$

with the boundary condition at inflow points, i.e., the points on $\partial\Omega_p$ at which geodesics are out-going,

$$F(\gamma) = \gamma, \quad \gamma \in \partial\Omega_p^{\text{inflow}}.$$

Note that inflowing characteristics correspond to out-going geodesics.

Now we define a surface phase $\Phi : \mathbb{P} \to \mathbb{R}$, such that $\Phi(\gamma)$ is the distance traveled by a geodesic starting at the point $\mathbf{u}$ with direction θ before it hits the boundary of Ω_p. Using (20), we can derive the PDE for Φ as

$$\cos\theta\Phi_u + \sin\theta\Phi_v + \mathscr{V}(\gamma)\Phi_\theta = \frac{1}{\rho(\gamma)}, \quad \gamma \in \Omega_p, \tag{30}$$

with the boundary condition at inflow points

$$\Phi(\gamma) = 0, \quad \gamma \in \partial\Omega_p^{\text{inflow}}.$$

In the same way we define a function $B : \mathbb{P} \to \mathbb{R}$ as the β-value of a geodesic starting at the point $\gamma \in \Omega_p$ when it hits the boundary of Ω_p. We then use (23) and derive the PDE for B as

$$\cos\theta B_u + \sin\theta B_v + \mathscr{V}(\gamma)B_\theta = \frac{\widetilde{\alpha}(\gamma)}{\rho(\gamma)}, \quad \gamma \in \Omega_p, \tag{31}$$

with the boundary condition at inflow points

$$B(\gamma) = 0, \quad \gamma \in \partial\Omega_p^{\text{inflow}}.$$

For the geometrical spreading we consider a fixed shadow line $\gamma_0(s) = (u_0(s), v_0(s), \theta_0(s))$ and like in Section 2.2.2 we define

$$\widetilde{u}(s,\tau) = u(\tau), \quad \widetilde{v}(s,\tau) = v(\tau), \quad \widetilde{\theta}(s,\tau) = \theta(\tau),$$

where (u, v, θ) solves (19) with initial data $(u_0(s), v_0(s), \theta_0(s))$. Setting $\widetilde{\gamma} = (\widetilde{u}, \widetilde{v}, \widetilde{\theta})$ we thus have

$$\widetilde{\gamma}_\tau = \mathbf{g}(\widetilde{\gamma}), \quad \widetilde{\gamma}(s,0) = \gamma_0(s),$$

with $\mathbf{g}$ defined in (19).

For a given shadow line, the creeping rays will lie on a submanifold of phase space $\mathbb{P}$ which we define as $\mathbb{L}(\gamma_0) = \{\widetilde{\gamma}(s,\tau) : \tau \geqslant 0\}$. We then introduce the function $Q : \mathbb{L}(\gamma_0) \to \mathbb{R}$ as

$$Q(\widetilde{\gamma}(s,\tau)) := \mathscr{Q}(s,\tau).$$

which is a Eulerian version of the geometrical spreading, restricted to $\mathbb{L}(\gamma_0)$. We will use the following simple Lemma.

Lemma 1. *The Jacobian $D_\gamma F(\gamma) \in \mathbb{R}^{3\times 3}$ has rank two for all $\gamma \in \Omega_p$ where it is well-defined. Its null space is spanned by $\mathbf{g}(\gamma)$.*

Proof 1. That $D_\gamma F(\gamma)\mathbf{g}(\gamma) = 0$ is just a restatement of (29). Suppose $D_\gamma F(\gamma)\mathbf{v} = 0$ and construct a curve $\gamma_0(s) \subset \mathbb{P}$ satisfying $\gamma_0(0) = \gamma$ and $\gamma_0'(0) = \mathbf{v}$. Let $\widetilde{\gamma}(s,\tau)$ be defined for this curve in the same way as above. Then $\frac{\mathrm{d}}{\mathrm{d}s}F(\gamma_0(s)) = 0$ for $s = 0$. Moreover, since $D_\gamma F(\gamma)$ is well-defined there is a differentiable function $\widehat{\tau}(s)$ such that $F(\gamma_0(s)) = \widetilde{\gamma}(s, \widehat{\tau}(s))$ in a neighborhood of $s = 0$. Together this means that

$$0 = \frac{\mathrm{d}}{\mathrm{d}s}\widetilde{\gamma}(s,\widehat{\tau}(\mathrm{s}))\bigg|_{s=0} = \widetilde{\gamma}_s(0,\widehat{\tau}(0)) + \widehat{\tau}'(0)\widetilde{\gamma}_\tau(0,\widehat{\tau}(0)). \tag{32}$$

Since $-\widehat{\tau}'(0)\widetilde{\gamma}_\tau(0,\tau)$ is a solution to the ODE $(\widetilde{\gamma}_s)_\tau = D_\gamma\mathbf{g}(\widetilde{\gamma})\widetilde{\gamma}_s$ for $s = 0$, uniqueness of ODE solutions implies that (32) holds for all $\tau \geqslant 0$, in particular

$$\widetilde{\gamma}_s(0,0) + \widehat{\tau}'(0)\widetilde{\gamma}_\tau(0,0) = 0 \quad \Longleftrightarrow \quad \mathbf{v} = -\widehat{\tau}'(0)\mathbf{g}(\gamma).$$

Hence, if $\mathbf{v}$ is in the nullspace, then it is parallel to $\mathbf{g}(\gamma)$, and the nullspace is thus one-dimensional. □

In order to compute Q we first find a solution $z = z(s,\tau)$ to

$$D_\gamma F(\widetilde{\gamma})z = \frac{\mathrm{d}}{\mathrm{d}s}F(\gamma_0(s)). \tag{33}$$

We note that $F(\widetilde{\gamma}(s,\tau)) = F(\gamma_0(s))$ for all $\tau \geqslant 0$, so this z satisfies

$$D_\gamma F(\widetilde{\gamma})z = D_\gamma F(\widetilde{\gamma})\widetilde{\gamma}_s.$$

By Lemma 1 we therefore get

$$z(s,\tau) = \widetilde{\gamma}_s + \alpha\mathbf{g}(\widetilde{\gamma}) = \widetilde{\gamma}_s + \alpha\widetilde{\gamma}_\tau,$$

for some α and since $\widetilde{X}_\tau = \widehat{T}(\widetilde{\gamma})$ by (16), we have

$$[\widehat{T}(\widetilde{\gamma}) \times \widehat{N}(\widetilde{u},\widetilde{v})]^\top J(\widetilde{u},\widetilde{v})\widetilde{z} = \widetilde{X}_\tau^\perp \cdot (\widetilde{X}_s + \alpha\widetilde{X}_\tau) = \widetilde{X}_\tau^\perp \cdot \widetilde{X}_s,$$

where $\widetilde{z} \in \mathbb{R}^2$ contains the first two components of z. Consequently, since $\widehat{T}(\gamma_0(s)) = \widehat{I}$,

$$Q(\widetilde{\gamma}) = \frac{[\widehat{T}(\widetilde{\gamma}) \times \widehat{N}(\widetilde{u},\widetilde{v})]^\top J(\widetilde{u},\widetilde{v})\widetilde{z}}{[\widehat{I} \times \widehat{N}(\mathbf{u_0}(s))]^\top \widetilde{X}_{0s}(s)}. \tag{34}$$

On the boundary, when $\widetilde{\gamma} \in \partial\Omega_p$ we can simplify the computation and avoid solving for z in (33). Let $\widehat{X} : \mathbb{R} \to \mathbb{R}^3$ be defined by $\widehat{X}(s) := \overline{X}(U(\gamma_0(s)), V(\gamma_0(s)))$ with U, V defined in (27). As in the proof of Lemma 1 there is a function $\widehat{\tau}(s)$ such that

$$\widehat{X}(s) = \widetilde{X}(s,\widehat{\tau}(s)). \tag{35}$$

After differentiating (35) with respect to s, we get

$$\widehat{X}_s(s) = \widetilde{X}_\tau\widehat{\tau}'(s) + \widetilde{X}_s.$$

Therefore, for $\widetilde{\gamma}$ on the boundary, i.e. $\widetilde{\gamma} = F(\gamma_0)$,

$$Q(\widetilde{\gamma}) = \frac{[\widehat{T}(\widetilde{\gamma}) \times \widehat{N}(\widetilde{u},\widetilde{v})]^\top \widehat{X}_s(s)}{[\widehat{I} \times \widehat{N}(\mathbf{u_0}(s))]^\top \widetilde{X}_{0s}}. \tag{36}$$

Note that $\widehat{X}_s(s)$ can easily be computed from the numerical solution to the PDE (29).

3. Numerical solution of the PDEs

All PDEs (29)–(31) are of the general form

$$af_u + bf_v + cf_\theta = d(u,v,\theta), \tag{37}$$

which are time-independent hyperbolic equations.

In the phase space $\mathbb{P}$, the direction of characteristics at the points on the boundary determines if boundary conditions are needed at that point. We assign boundary conditions at the points where a characteristic is in-going. For example a characteristic is in-going if $\dot{u} = \rho\cos\theta > 0$ on the left boundary and if $\dot{v} = \rho\sin\theta > 0$ on the lower boundary. More precisely, suppose Ω is the unit square and $-\pi < \theta \leqslant \pi$. Then we prescribe boundary condition on $\partial\Omega_p^{\text{inflow}}$ given by

$$\partial\Omega_p^{\text{inflow}} = \left\{u = 0,\ |\theta| < \frac{\pi}{2}\right\} \bigcup \left\{u = 1,\ |\theta - \pi| < \frac{\pi}{2}\right\} \bigcup \{v = 0,\ \theta > 0\} \bigcup \{v = 1,\ \theta < 0\}.$$

We always use periodic boundary conditions in the θ direction.

To solve these equations, we use a Fast Marching algorithm, given by Fomel and Sethian [8]. We let $f = (F, \Phi, B)$ and discretize the phase space domain $\Omega_p = \Omega \times \mathbb{S}$ uniformly, setting $u_i = i\Delta u$, $v_j = j\Delta v$ and $\theta_k = k\Delta\theta$, with the step sizes $\Delta u = \Delta v = \frac{1}{N}$ and $\Delta\theta = \frac{2\pi}{N}$. Then by solving the PDEs (37), we get the approximate solution

$$f_{ijk} = (F_{ijk}, \Phi_{ijk}, B_{ijk}) \approx (F(u_i, v_j, \theta_k), \Phi(u_i, v_j, \theta_k), B(u_i, v_j, \theta_k)).$$

The complexity is $\mathcal{O}(N^3 \log N)$. See [8] for more details.

4. Post-processing

To extract properties like phase and amplitude for a ray family, post-processing of the solution to the escape PDEs (37) is needed. It is based on the following simple observation. By the uniqueness of solutions of ODEs,

$$F(\gamma_1) = F(\gamma_2),$$

if and only if the points γ_1 and γ_2 lie on the same geodesic.

As an example, suppose we want to compute the surface phase at a point on the scatterer, when the scatterer is illuminated. We assume that the shadow line $\gamma_0(s) = (u_0(s), v_0(s), \theta_0(s))$ is known. For each point $(u, v) \in \Omega$ covered by the surface wave there is at least one creeping ray passing that point starting at the shadow line $\gamma_0(s)$. By the argument above, we can thus find $s = s^*(u, v)$ and phase angle $\theta = \theta^*(u, v)$, as the solution to

$$F(\gamma_0(s)) = F(u, v, \theta). \tag{38}$$

The phase at (u,v) is then given by

$$\phi(u, v) = \phi_0(\mathbf{u}_0(s^*)) + \Phi(\gamma_0(s^*)) - \Phi(\gamma^*), \quad \gamma^* = (u, v, \theta^*),$$

with ϕ_0 as in (12). Note that γ^* is now in the submanifold $\mathbb{L}(\gamma_0)$ which was defined in Section 2.3. There may be multiple solutions (s^*, θ^*) to (38), giving multiple phases.

We now introduce a function $A : \mathbb{L}(\gamma_0) \to \mathbb{R}$ as the amplitude at the point $\gamma \in \mathbb{L}(\gamma_0)$ on the geodesic starting at the shadow line $\gamma_0(s)$. By (22) we can write

$$A(\gamma^*) = A_0 Q(\gamma^*)^{\frac{-1}{2}} \exp\left(-\omega^{\frac{1}{3}}(B(\gamma_0(s^*)) - B(\gamma^*))\right),$$

where A_0 is the amplitude at the point $\gamma_0(s^*)$, and $Q(\gamma^*)$ is computed by (34).

The main difficulty here is to solve (38). We now show how to solve it. Since $F = (U, V, \Theta)$ is a point on the phase space boundary $\partial\Omega_p$, it can be reduced to a point (S, Θ) in $\mathbb{R}^2$. For example in a rectangular domain Ω, Fig. 5, we choose $S \in [0, 2\pi]$ along $\partial\Omega$ to be zero at the lower left corner, π at the upper right corner, and 2π again at the lower left corner. Now the left and right hand sides of (38) are curves in $\mathbb{R}^2$ parameterized by s and θ, and solving the algebraic equation (38) amounts to finding crossing points of these curves. See Fig. 5.

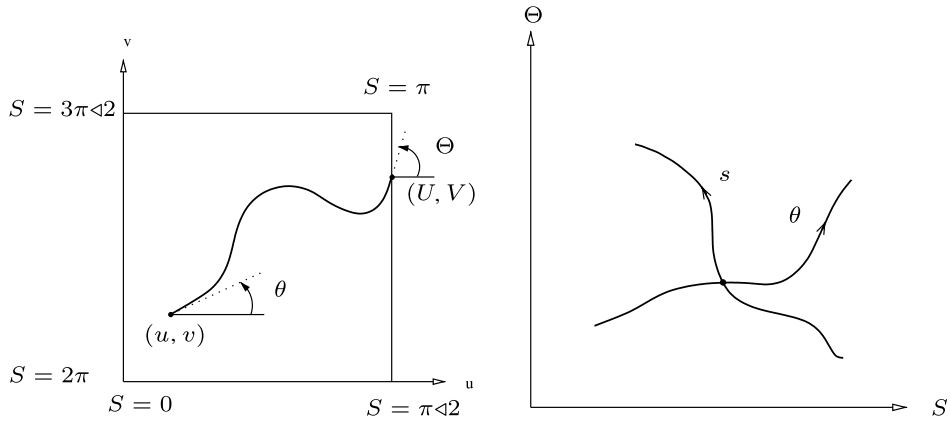


Fig. 5. Left figure shows a geodesic in a rectangular domain in the parameter space and the choice of S on the boundary. Right figure shows two crossing curves. One curve is for all points on the shadow line, parameterized by s. The other curve is for a single point in the parameter space with all directions, parameterized by θ.

Numerically, we discretize the parameterization of the shadow line in N grid points $\{s_m\}$, $m = 1, \ldots, N$. For each point $\{\mathbf{u}_0(s_m)\}$ on the parameter space shadow line, the ray direction $\theta_0(s_m)$ at the shadow line is computed using the fact that the tangential vector $\widehat{T}$ to the hypersurface at the point $\gamma_0(s_m)$ should be in the same direction as the incident angle $\widehat{I}$:

$$\widehat{T}(\gamma_0(s_m)) = \widehat{I}. \tag{39}$$

After obtaining the discretized phase space shadow line $\{\gamma_0(s_m)\}$, we then interpolate the approximate solution f_{ijk} (available on a regular grid) to find the approximate solution on the shadow line:

$$\widetilde{f}_{s_m} = (\widetilde{F}_{s_m}, \widetilde{\Phi}_{s_m}, \widetilde{B}_{s_m}) \approx (F(\gamma_0(s_m)), \Phi(\gamma_0(s_m)), B(\gamma_0(s_m))).$$

Having the discretized solution on the shadow line and at the point $(u, v) \in \Omega$ for all N directions $\theta \in [0, 2\pi]$, we then need to find crossing points of two complex lines of N straight line segments. These crossing points will then be the solutions to (38). The amount of work to do this is proportional to N, by using a monotonic sections algorithm; see e.g. [27]. For all N^2 points on the surface the computational cost for finding crossing points will then be $\mathcal{O}(N^3)$. The complexity to solve the PDEs using the Fast Marching method is $\mathcal{O}(N^3 \log N)$. Therefore the total complexity will be $\mathcal{O}(N^3 \log N)$.

If we only need to compute the field for one shadow line, it could be done faster. For example by using wave front tracking or solvers based on the surface eikonal equation, the complexity is $\mathcal{O}(N^2)$. But there are applications when we need the field for many shadow lines. In such cases, using the Fast Marching method can be much faster. We will show one such application in the next section.

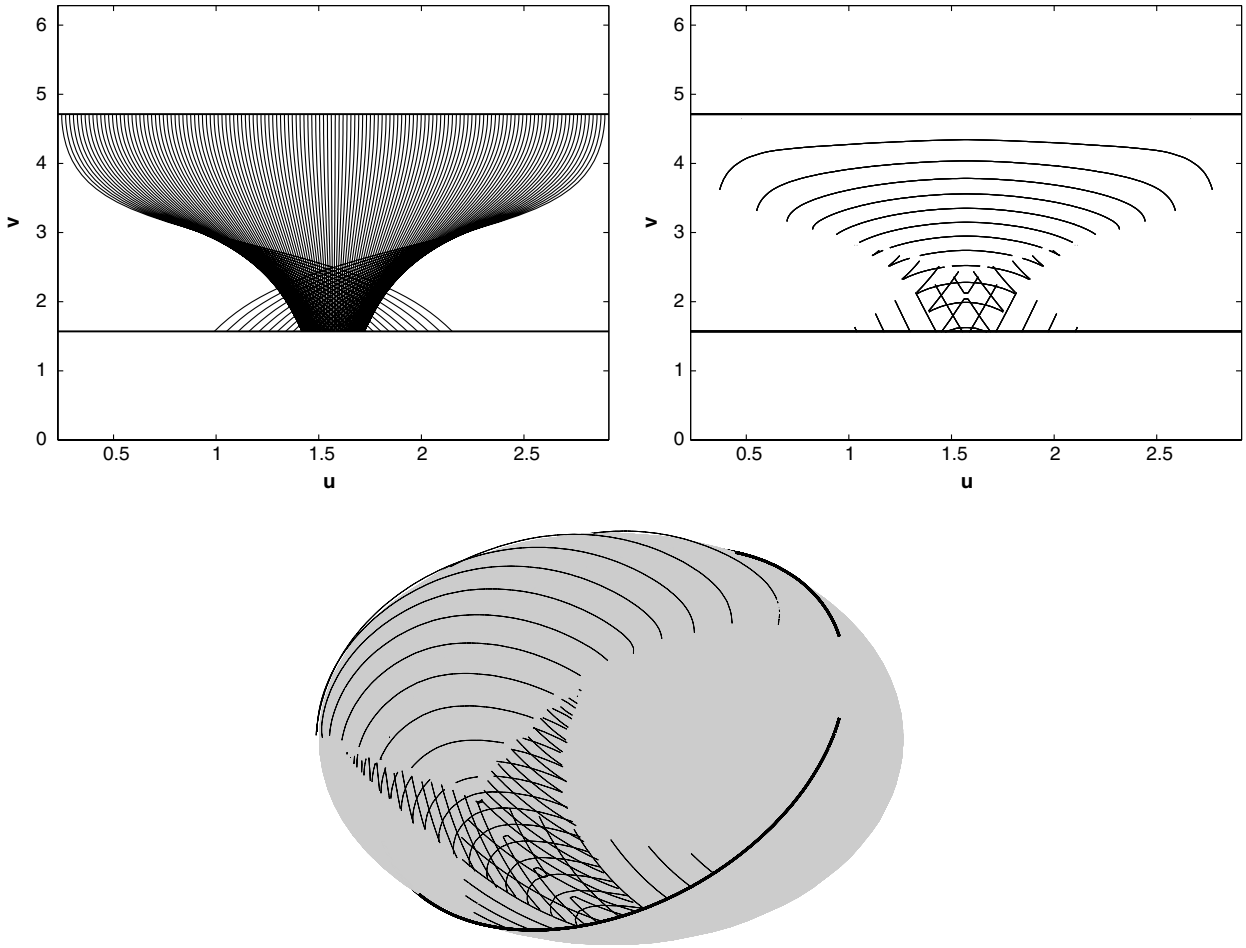


Fig. 6. Ray propagation on the shadow zone of an ellipsoid. Top figures show the creeping rays (left) and iso-phase curves (right) in the parameter space between two shadow lines. Bottom figure shows the iso-phase curves and the shadow line (bold) in the physical space.

As an example, in Fig. 6, the iso-phase curves are shown for an ellipsoid illuminated by incident rays in direction $\widehat{I} = [0, 1, 0]$. In the shadow zone between the two shadow lines, there are either one, two or three phases. As it can be seen, multiple phases can be captured. The solution here is computed by the Fast Marching method on a 120^3 grid and using the post-processing described above.

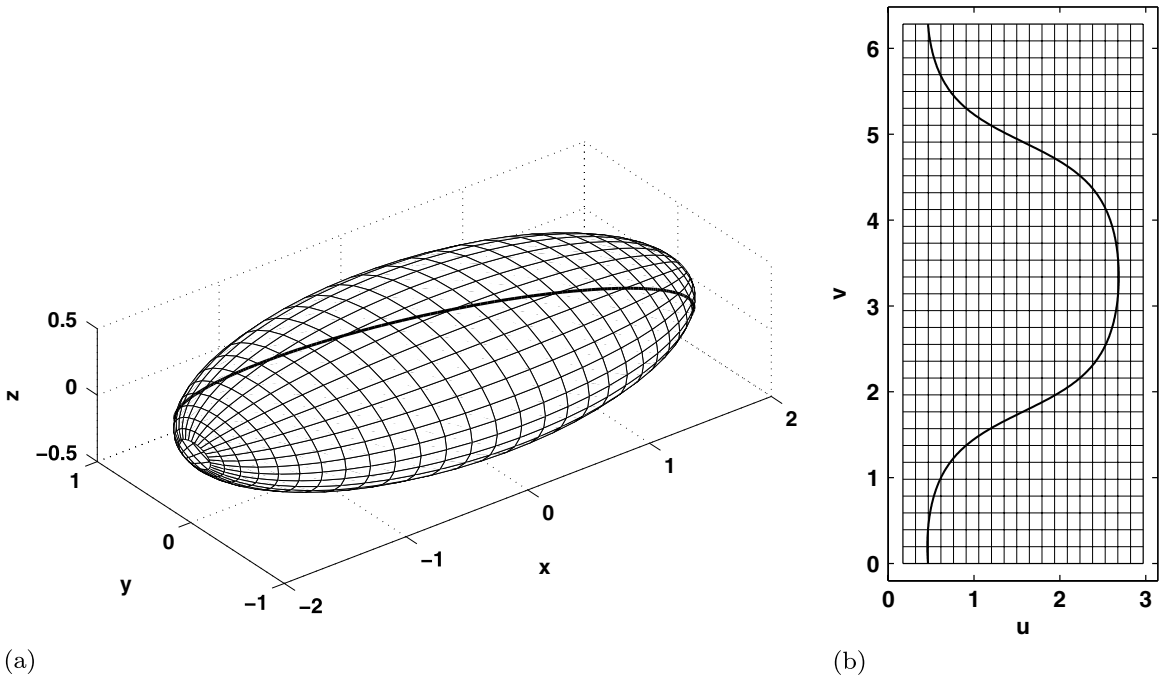


Fig. 7. Shadow line in the physical and parameter space: (a) shadow line in (x, y, z)-space; (b) shadow line in (u, v)-space.

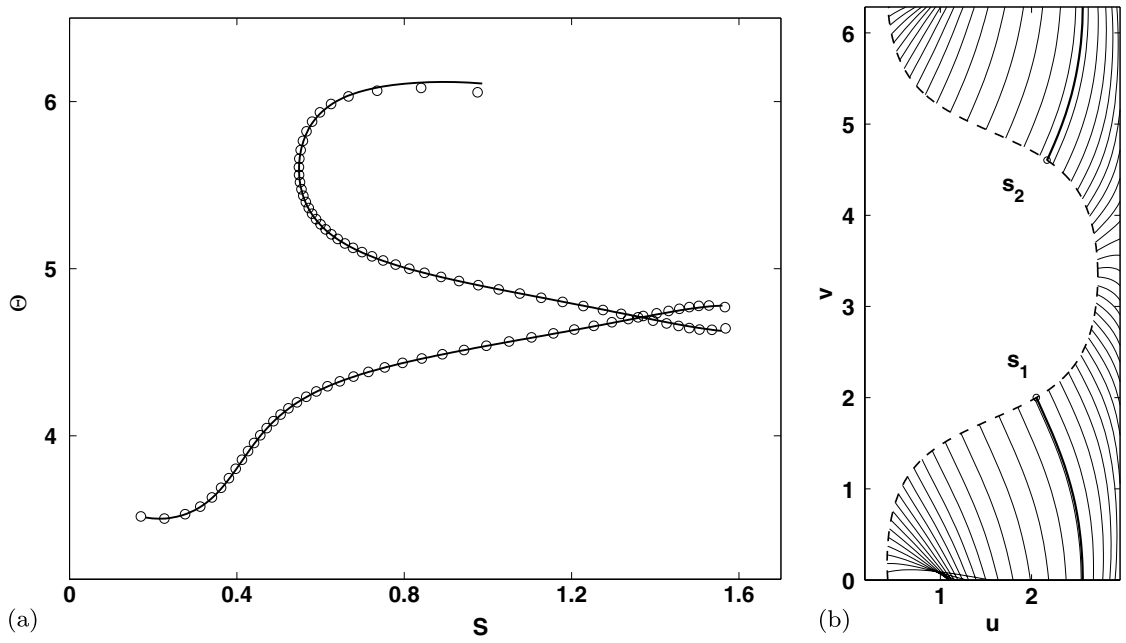


Fig. 8. Right figure shows all creeping rays starting at the shadow line (dashed) and ending at the boundary. The two bold curves are the backscattered ray. Left figure shows two curves corresponding to the rays hitting the top and bottom boundaries in the parameter space. Circles denote the values computed by the Fast Marching method and solid lines denote the values computed by a high order accurate ray tracing method. The crossing point corresponds to the backscattered ray. (a) $F(\gamma(s_1))$ and $F(\gamma(s_2)) + C$; (b) creeping rays in (u, v) space.

5. An application to mono-static RCS computations

Mono-static RCS is a measure of backscattered radiation in the direction of incident waves, when an object is irradiated. Normally most part of it consists of direct reflections, but for not too high frequencies there are situations where creeping rays can give important contribution [3]. The rays that propagate on the surface of the scatterer and return in the opposite direction of incident waves are called *backscattered creeping rays*.

In this section we apply the fast phase space method on a scattering problem and compute the contribution of the backscattered creeping rays to RCS. For simplicity we only consider the amplitude *on* the scatterer, ignoring the effect of diffraction coefficients and geometrical spreading outside the scatterer. We assume that the incoming amplitude is one on the shadow line and compute the backscattered amplitude on the shadow line before the ray leaves the scatterer. We compare the results with standard ray tracing.

5.1. Scattering problem

As a test case we consider a hypersurface $\overline{X} = \overline{X}(u, v)$ which is a patch of an ellipsoid with the following parametric equations:

$$x = -r_1 \cos u,$$
$$y = r_2 \sin u \cos v,$$
$$z = r_3 \sin u \sin v,$$

where $r_1 = 2$, $r_2 = 1$, and $r_3 = 0.5$ are the ellipsoid's semiaxes. Notice that in order to avoid the irregularity at the points $(\pm r_1, 0, 0)$, we cut off these points from the parameter space.

First, we need to compute the shadow lines on the scatterer. For this hypersurface we can find them analytically. By (7) and (8), the shadow line corresponding to the incident direction $\widehat{I} = [\iota_1, \iota_2, \iota_3]$ is given by

$$\iota_1 r_2 r_3 \cos u_0(s) - \iota_2 r_1 r_3 \sin u_0(s) \cos v_0(s) - \iota_3 r_1 r_2 \sin u_0(s) \sin v_0(s) = 0.$$

The ray directions $\theta_0(s)$ at the shadow line are then computed using (39). For example, in Fig. 7 the shadow line is shown for $\widehat{I} \| [0.9, 1, 0.1]$ in physical and parameter space, respectively.

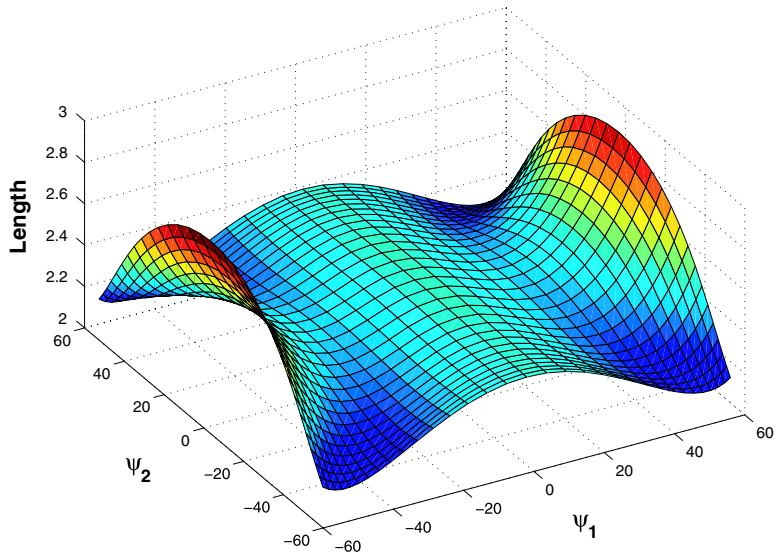


Fig. 9. Length of the backscattered creeping rays for many illumination angles.

5.2. Finding the backscattered rays

The goal is to find the length and amplitude of the backscattered creeping rays for different incident angles. In order to find the backscattered creeping rays, we use post-processing as before. A backscattered ray starting at point s_1 and ending at point s_2 on shadow line should satisfy

$$F(\gamma(s_1)) = F(\gamma(s_2)) + C, \tag{40}$$

where the constant C accounts for the fact that the upper and lower boundaries in the parameter space coincide on the hypersurface. It means that the points with $S = \pi, \ldots, 3\pi/2$ should be changed to $S = \pi/2, \ldots, 0$ and at the same time their Θ values should be added by π. The reason for adding by π is that we need to reverse the direction of the geodesic starting at s_2. Notice that we only consider the geodesics which hit the upper and lower boundaries, because the left and right boundaries are indeed artificial boundaries, introduced to avoid the irregularity.

As before, the right and left hand sides of (40) are curves in $\mathbb{R}^2$ parameterized by s, and to find the backscattered ray we need to find crossing points of these curves. Fig. 8(a) shows the intersecting curves in the (S, Θ)-plane for the points on the shadow line corresponding to geodesics hitting the lower and upper boundaries in parameter space, c.f. Fig. 5. Fig. 8(b) shows the creeping rays starting at all N points on the shadow line and the backscattered ray (bold line).

5.3. Length and amplitude of backscattered ray

The length and amplitude of the backscattered creeping rays are computed by a third order interpolation of the solution to the PDEs (37). For a given incident direction $\hat{I} = [\iota_1, \iota_2, \iota_3]$, the horizontal and vertical incident angles are calculated as

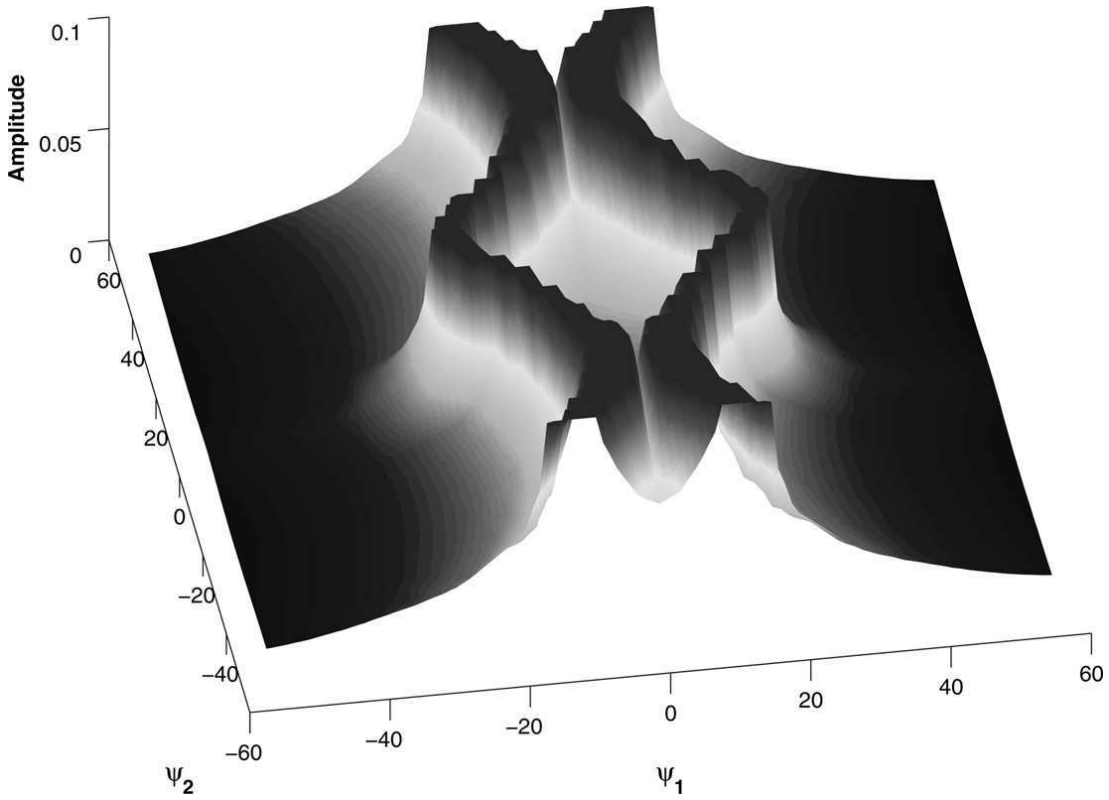


Fig. 10. Amplitude of the backscattered creeping rays for many illumination angles for $\omega = 1$.

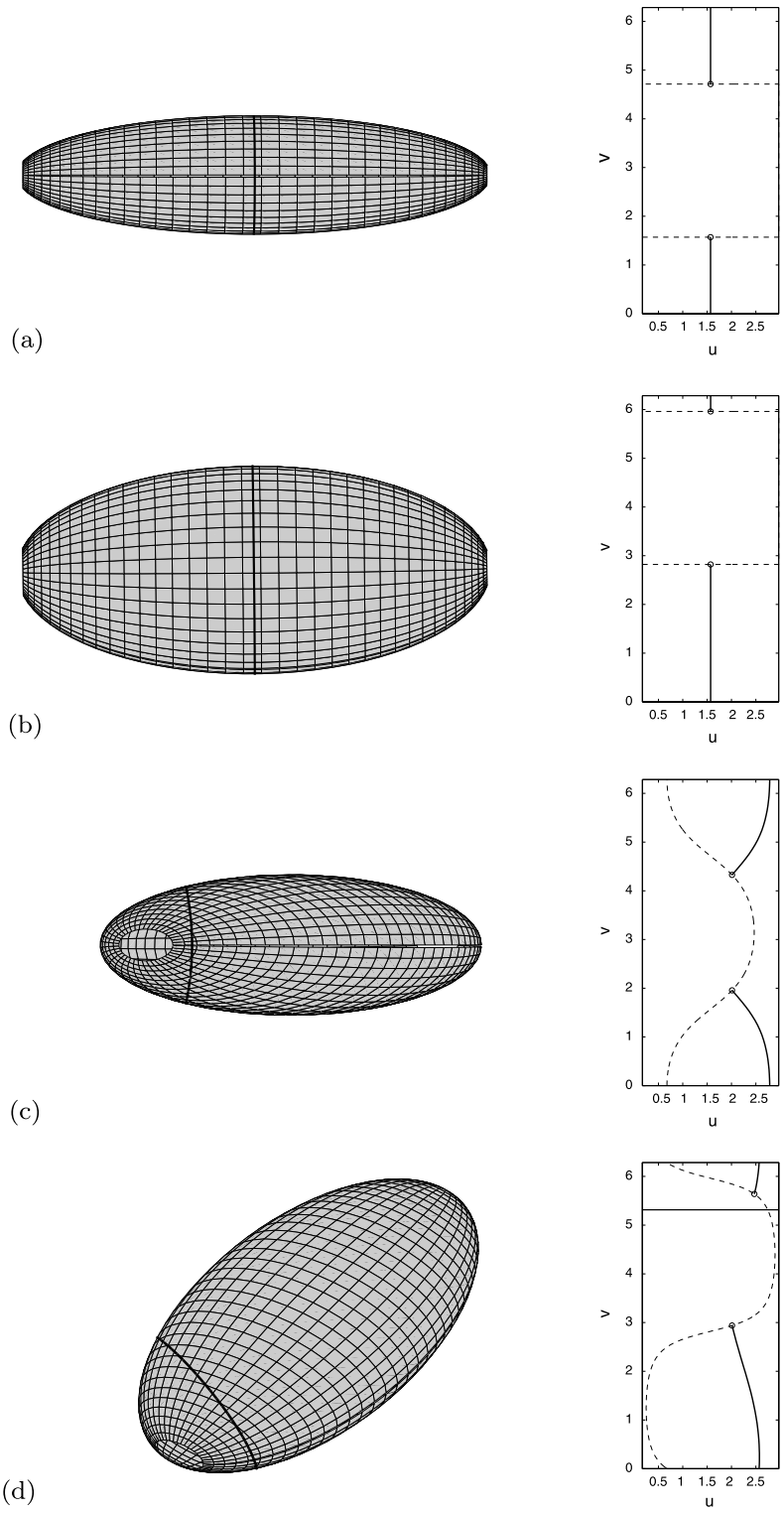


Fig. 11. The backscattered creeping rays for four different illumination angles and two different frequencies. Left figures show the backscattered rays in the physical space by bold solid lines. The view direction is in the illumination direction, so that the shadow line is the outer most curve around the ellipsoid. Right figures show the backscattered rays in the parameter space. Shadow lines here are shown by dashed lines. The amplitudes for $\omega = 1$ and $\omega = 20$ are denoted by a_1 and a_{20}, respectively. (a) $\psi_1 = 0$, $\psi_2 = 0$, length $= 2.44$, $a_1 = 0.022$; (b) $\psi_1 = 0$, $\psi_2 = 56$, length $= 2.43$, $a_1 = 0.044$; $a_{20} = 2.40 \times 10^{-5}$; (c) $\psi_1 = 58$, $\psi_2 = 0$, length $= 2.89$, $a_1 = 0.010$; $a_{20} = 6.84 \times 10^{-6}$; (d) $\psi_1 = 58$, $\psi_2 = 56$, length $= 2.16$, $a_1 = 0.012$; $a_{20} = 8.73 \times 10^{-6}$.

$$\psi_1 = \arctan\left(\frac{l_1}{l_2}\right), \quad \psi_2 = \arctan\left(\frac{l_3}{l_2}\right).$$

They vary from $-60°$ to $60°$. Fig. 9 shows the length for different incident angles.

For computing the geometrical spreading, we again use the fact that the upper and lower boundaries of the doman Ω in the parameter space coincide on the hypersurface. Therefore, one can consider a new domain $\widetilde{\Omega}$ consisting of two domains Ω on top of each other, connected by the boundary $v = 0$. The creeping ray starting at the point $\gamma(s_1)$ in the upper domain continues in the lower domain and hits the shadow line at the point $\widetilde{\gamma}(s_2) = \gamma(s_2) + C$, with $C = (0, -2\pi, \pi)$. Now, let $\widetilde{F}$ be the escape location and direction on $\partial\widetilde{\Omega}$ for the extended doman $\widetilde{\Omega}$. We will have $\widetilde{F}(\gamma(s_1)) = F(F(\gamma(s_1)) + \widetilde{C}) - \widetilde{C}$ and $\widetilde{F}(\widetilde{\gamma}(s_2)) = F(\widetilde{\gamma}(s_2) + \widetilde{C}) - \widetilde{C}$ where $\widetilde{C} = (0, 2\pi, 0)$. We can then use (34) to compute the geometrical spreading $Q(\widetilde{\gamma}(s_2))$ at the point $\widetilde{\gamma}(s_2)$ from the starting point $\gamma(s_1)$ The amplitude is computed by

$$A(\gamma(s_2)) = A(\gamma(s_1))(Q(\widetilde{\gamma}(s_2)))^{\frac{-1}{2}} \exp\left(-\omega^{\frac{1}{3}}(B(\gamma(s_1)) + B(\gamma(s_2)))\right).$$

Fig. 10 shows the amplitude for different incident angles. For some incident angles, the geometrical spreading of the creeping ray becomes zero, These rays are called *caustic backscattered creeping rays*, and their amplitude is infinite at the shawod line. However, away from the scatterer their contribution is bounded because of geometrical spreading outside the scatterer. Note that in Fig. 10 the amplitudes larger than a certain value are not shown.

Fig. 11 shows the backscattered creeping rays in the physical and parameter space for four different incident directions.

5.4. Convergence and complexity

We use a first order Fast Marching algorithm. Fig. 12 shows the length $\Phi(u, \pi, \pi/2)$ obtained using a coarse mesh of the size 60^3 and a fine mesh of the size 120^3. We compare the solution with a reference solution obtained by a high order accurate Ray tracing method. It confirms the first order accuracy of the Fast Marching algorithm.

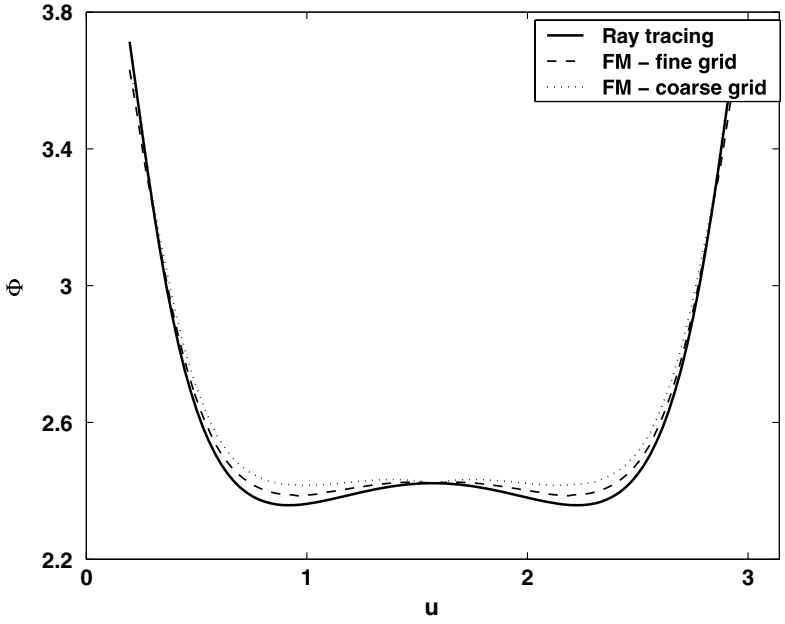


Fig. 12. The length $\Phi(u, \pi, \frac{\pi}{2})$ obtained using Fast Marching on a coarse and fine grid. They converge to a reference solution obtained by a high order solver using Ray tracing.

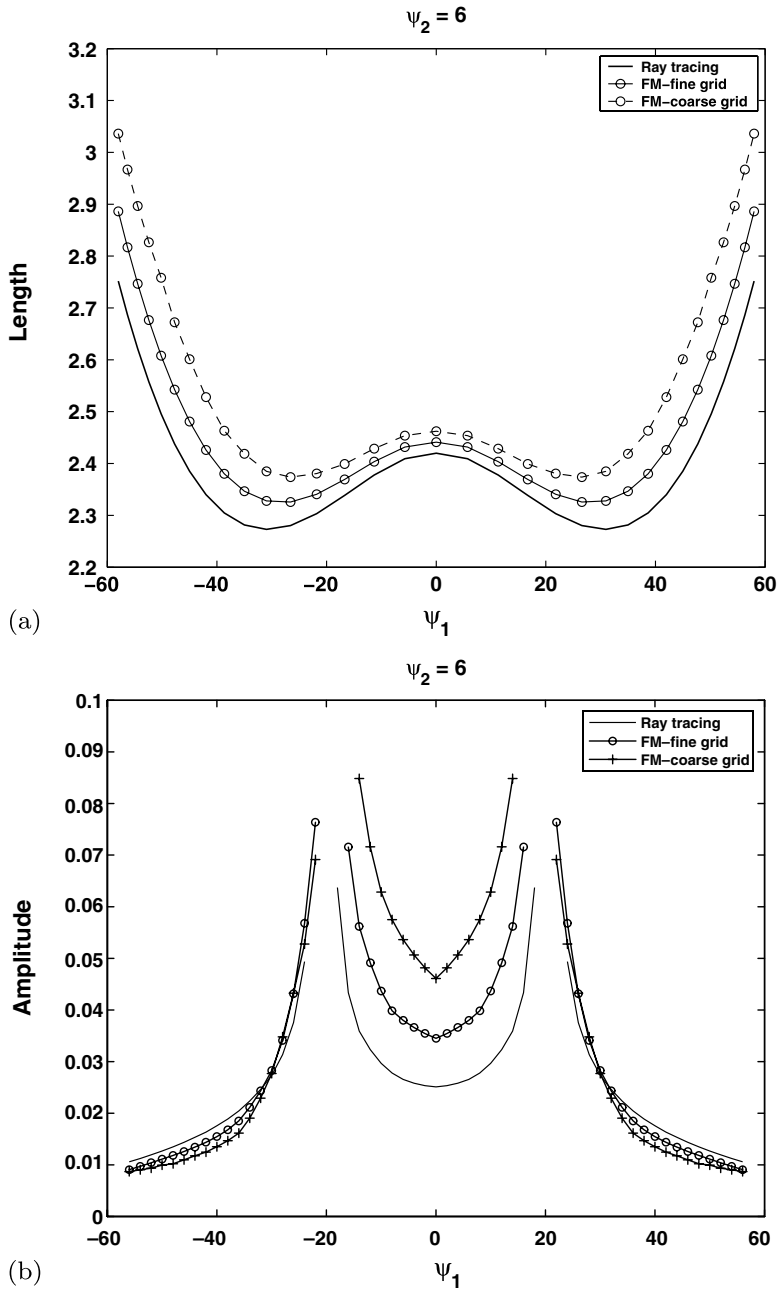


Fig. 13. Length and amplitude (at $\omega = 1$) of the backscattered ray for different horizontal incident angles ψ_1 and a fixed vertical incident angle $\psi_2 = 6$. Solutions of Fast Marching algorithm converge to a reference solution obtained by Ray tracing as we use a finer grid. (a) Length; (b) amplitude

The convergence of the length and amplitude (at $\omega = 1$) of the backscattered creeping ray is shown in Fig. 13 for a fixed vertical incident angle $\psi_2 = 6°$ and different horizontal incident angles ψ_1. Although the relative error is worse for the amplitude than for the phase, the rate of convergence confirms the first-order accuracy of the method. The accuracy of amplitude can be improved either by using a higher order fast marching method or by computing the geometrical spreading Q directly by using another ODE instead of numerically differentiating the functions U and V with respect to u, v and θ to compute $\widehat{X}_s(s)$ in (36) as done in [17,31].

The complexity of using the fast phase space method proposed here consists of two parts. First, the cost of solving the PDEs by the Fast Marching method is $\mathcal{O}(N^3 \log N)$. Second, the cost of finding the backscattered

rays for each shadow line is $\mathcal{O}(N)$. For all N^2 shadow lines, it is $\mathcal{O}(N^3)$. Therefore the total complexity will be $\mathcal{O}(N^3 \log N)$. The total cost by using other methods, like wave front tracking and solvers based on the surface eikonal equation, will be $\mathcal{O}(N^4)$, if the cost for each shadow line is $\mathcal{O}(N^2)$. In this case, using the Fast Marching method will then be much faster.

6. Conclusion

We have presented a new phase space method for computing creeping rays in an Eulerian framework. We have formulated the ray propagation problem as a set of time-independent PDEs in a three-dimensional phase space. To solve the PDEs we have used a first-order fast marching method. Properties like phase and amplitude for a ray family as well as wavefronts can be extracted through a fast post-processing. The method is computationally attractive when the solution is sought for many different sources but with the same index of refraction, for example in RCS computations.

In this paper, the surface is assumed to be represented by a single parameterization. In future work, we plan to extend the method to be applicable to more complicated and realistic geometries which can be represented by multiple parameterizations. The information can then be extracted by combining multi-patches through a post-processing. Moreover, we will use a higher order method in order to increase the accuracy.

References

[1] J.-D. Benamou, Big ray tracing: multivalued travel time field computation using viscosity solutions of the eikonal equation, J. Comput. Phys. 128 (4) (1996) 463–474.
[2] J.-D. Benamou, Direct computation of multivalued phase space solutions for Hamilton–Jacobi equations, Comm. Pure Appl. Math. 52 (11) (1999) 1443–1475.
[3] D.P. Bouche, J.-J. Bouquet, H. Manenc, R. Mittra, Asymptotic computation of the RCS of low observable axisymmetric objects at high frequency, IEEE T. Antenn. Propag. 40 (10) (1992) 1165–1174.
[4] V. Červený, I.A. Molotkov, I. Psencik, Ray Methods in Seismology, Univ. Karlova Press, 1977.
[5] B. Engquist, O. Runborg, Computational high frequency wave propagation, Acta Numer. 12 (2003) 181–266.
[6] B. Engquist, O. Runborg, A.-K. Tornberg, High frequency wave propagation by the segment projection method, J. Comput. Phys. 178 (2002) 373–390.
[7] E. Fatemi, B. Engquist, S.J. Osher, Numerical solution of the high frequency asymptotic expansion for the scalar wave equation, J. Comput. Phys. 120 (1) (1995) 145–155.
[8] S. Fomel, J.A. Sethian, Fast phase space computation of multiple arrivals, Proc. Natl. Acad. Sci. USA 99 (11) (2002) 7329–7334 (electronic).
[9] L. Gosse, S. Jin, X. Li, Two moment systems for computing multiphase semiclassical limits of the Schrödinger equation, Math. Models Methods Appl. Sci. 13 (12) (2003) 1689–1723.
[10] A. Gray, Modern Differential Geometry of Curves and Surfaces, CRC Press, 1993.
[11] S. Hagdahl, Hybrid Methods for Computational Electromagnetics in Frequency Domain, PhD thesis, NADA, KTH, Stockholm, 2005.
[12] P.E. Hussar, V. Oliker, H.L. Riggins, E.M. Smith-Rowlan, W.R. Klocko, L. Prussner, An implementation of the UTD on facetized CAD platform models, IEEE Antennas Propag. 42 (2) (2000) 100–106.
[13] J. Keller, R.M. Lewis, Asymptotic methods for partial differential equations: the reduced wave equation and Maxwell's equations, Surveys Appl. Math. 1 (1995) 1–82.
[14] J.B. Keller, The geometric theory of diffraction, in: Symposium on Microwave Optics, Eaton Electronics Research Laboratory, McGill University, Montreal, Canada, June 1953.
[15] R. Kimmel, J.A. Sethian, Computing geodesic paths on manifolds, Proc. Natl. Acad. Sci. USA 95 (15) (1998) 8431–8435 (electronic).
[16] E.M. Koper, W.D. Wood, S.W. Schneider, Aircraft antenna coupling minimization using genetic algorithms and approximations, IEEE T. Aero. Elec. Sys. 40 (2) (2004) 742–751.
[17] S. Leung, J. Qian, S. Osher, A level set method for three-dimensional paraxial geometrical optics with multiple point sources, Comm. Math. Sci. 2 (4) (2004) 657–686.
[18] B.R. Levy, J. Keller, Diffraction by a smooth object, Comm. Pure Appl. Math. 12 (1959).
[19] H. Ling, R. Chou, S.W. Lee, Shooting and bouncing rays: calculating the RCS of an arbitrarily shaped cavity, IEEE T. Antenn. Propag. 37 (1989) 194–205.
[20] R. MittraTopics in Applied Physics: Numerical and Asymptotic Techniques in Electromagnetics, vol. 3, Springer, 1975.
[21] S.J. Osher, L.-T. Cheng, M. Kang, H. Shim, Y.-H. Tsai, Geometric optics in a phase-space-based level set and Eulerian framework, J. Comput. Phys. 179 (2) (2002) 622–648.
[22] J. Perez, J.A. Saiz, O.M. Conde, R.P. Torre, M.F. Catedra, Analysis of antennas on board arbitrary structures modeled by NURBS surfaces, IEEE T. Antenn. Propag. 45 (6) (1997) 1045–1053.

[23] J. Qian, L.-T. Cheng, S. Osher, A level set-based Eulerian approach for anisotropic wave propagations, Wave Motion 37 (2003) 365–379.
[24] O. Runborg, Multiphase Computations in Geometrical Optics, Licentiate's thesis, NADA, KTH, Stockholm, 1996.
[25] O. Runborg, Some new results in multiphase geometrical optics, M2AN Math. Model. Numer. Anal. 34 (2000) 1203–1231.
[26] W.W. Symes, J. Qian, A slowness matching Eulerian method for multivalued solutions of eikonal equations, J. Sci. Comput. 19 (1–3) (2003) 501–526.
[27] G. Taylor, Another look at the line intersection problem, Int. J. Geogr. Inf. Syst. 3 (20) (1989) 192–193.
[28] J. van Trier, W.W. Symes, Upwind finite-difference calculation of traveltimes, Geophysics 56 (6) (1991) 812–821.
[29] J. Vidale, Finite-difference calculation of traveltimes, B. Seismol. Soc. Am. 78 (6) (1988) 2062–2076.
[30] V. Vinje, E. Iversen, H. Gjøystdal, Traveltime and amplitude estimation using wavefront construction, Geophysics 58 (8) (1993) 1157–1166.
[31] L. Ying, E.J. Candes, The Phase Flow Method, Preprint, 2005.

Paper II

A Multiple-Patch Phase Space Method for Computing Trajectories on Manifolds with Applications to Wave Propagation Problems

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

April 26, 2007

Abstract. We present a multiple-patch phase space method for computing trajectories on two-dimensional manifolds possibly embedded in a higher-dimensional space. The dynamics of trajectories are given by systems of ordinary differential equations (ODEs). We split the manifold into multiple patches where each patch has a well-defined regular parameterization. The ODEs are formulated as *escape* equations, which are hyperbolic partial differential equations (PDEs) in a three-dimensional phase space. The escape equations are solved in each patch, individually. The solutions of individual patches are then connected using suitable inter-patch boundary conditions. Properties for particular families of trajectories are obtained through a fast post-processing.

We apply the method to two different problems: the creeping ray contribution to mono-static radar cross section computations and the multivalued travel-time of seismic waves in multi-layered media. We present numerical examples to illustrate the accuracy and efficiency of the method.

Keywords. ODEs on a manifold; Phase space method; Escape equations; High frequency wave propagation; Geodesics; Creeping rays; Seismic waves; Travel-time

1 Introduction

We want to compute trajectories on two-dimensional compact manifolds possibly embedded in a higher-dimensional space. The dynamics of the trajectories we consider are given by systems of ODEs in a phase space. In many problems, we need to compute a large number of trajectories. In other words, the dynamical systems of ODEs need to be integrated for many different initial conditions. Examples include geodesics computation in computational geometry [11], robotics [2] and the theory of general relativity.

Our motivation for this comes from high frequency wave propagation problems. We consider the problem of scattering of a time-harmonic incident field by a bounded scatterer D. We split the total field into an incident and a scattered field. The scattered field in the region outside D is given by the Helmholtz equation,

$$\Delta W + n(\mathbf{x})^2\omega^2 W = 0, \qquad \mathbf{x} \in \mathbb{R}^3 \setminus \bar{D}, \tag{1}$$

where $n(\mathbf{x})$ is the index of refraction, and ω is the angular frequency. We can impose either a Dirichlet, Neumann or Robin boundary condition on the boundary of the scatterer ∂D and the Sommerfeld radiation condition at infinity.

The computational cost of direct numerical simulations of (1) grows algebraically with the frequency. Therefore, at high frequencies, numerical methods based on approximations of (1) are needed.

Geometrical optics (GO), for example, considers simple waves,

$$W(\mathbf{x}) \approx a(\mathbf{x})e^{i\omega\phi(\mathbf{x})}, \qquad \mathbf{x} \in \mathbb{R}^3, \tag{2}$$

when $\omega \to \infty$. The amplitude $a(\mathbf{x})$ and the phase function $\phi(\mathbf{x})$ depend only mildly on ω, and the computational cost will then be independent of ω. GO can be formulated either as PDEs for ϕ and a, known as *eikonal* and *transport* equations, respectively, or as a system of ordinary differential equations (ODEs).

Geometrical theory of diffraction (GTD), [18] is a correction to the GO approximations by adding diffraction effects. One type of diffracted rays are *creeping rays* which are generated at the *shadow line* of the scatterer and propagate along geodesics on the surface, continuously shedding diffracted rays in their tangential direction. A wave field, associated to a creeping ray, is generated on the surface

$$W_s(\mathbf{u}) = a(\mathbf{u})e^{i\omega\phi(\mathbf{u})}, \tag{3}$$

where $\phi(\mathbf{u})$ and $a(\mathbf{u})$ are surface phase and amplitude and $\mathbf{u} \in \mathbb{R}^2$ is a parameterization of the surface. The creeping rays are related to (3) in the same way as the standard GO rays are related to (2). Similar to GO rays, creeping rays can also be formulated either as PDEs or as a system of ODEs. There are two different approaches to compute the standard GO and creeping rays and the associated wave fields in (2) and (3); Lagrangian and Eulerian methods. Lagrangian methods are based on ODEs. The simplest Lagrangian method is standard ray tracing [6, 24, 13, 29] which gives the phase and amplitude solution along a ray. Interpolation must then be applied to obtain the solution everywhere. But, in regions where rays cross or diverge this can be rather difficult. The interpolation can be simplified by using wave front methods [38, 10]. In these methods, instead of individual rays, an interface representing a wave front is evolved. Eulerian methods, on the other hand, are based on PDEs. The PDEs are discretized on fixed computational grids to control accuracy everywhere, and there is no problem with interpolation. The simplest Eulerian methods solves the eikonal and transport equations [37, 36, 8, 20]. However, these equations only give the correct solution when it is a single wave. In the case of crossing waves, more elaborate schemes have been devised based on a third formulation of geometrical optics as a kinetic equation set in phase space. A survey of this research effort, in the free space GO case, is given in [7, 31, 25]. In the surface ray case, see [26, 39] for some recent works.

Fomel and Sethian [9] presented a fast phase space method for computing solutions of static Hamilton-Jacobi equations in phase space. Their method is based on *escape* equations which are time-independent PDEs in a three-dimensional phase space. The PDE solutions, computed by a fast marching method, give the information for all trajectories from all possible starting configurations.

Recently, the authors extended the fast phase space method [26] to efficiently computing all possible creeping rays on a hypersurface. The escape solutions contains information for all incident angles. The phase and amplitude of the field are then extracted by a fast post-processing. This method is computationally attractive when the solution is sought for many different sources but with the same index of refraction, for example for computing the mono-static radar cross section (RCS). The computational cost of solving the PDEs is less than tracing all rays individually. If the surface is discretized by N^2 points the complexity is $\mathcal{O}(N^3 \log N)$, which is close to optimal. In the mono-static RCS case, direct ray tracing would cost $\mathcal{O}(N^4)$ if a comparable number of incidence angles (N^2) and rays per angle (N) are considered.

However, it is only applicable for the scatterer surfaces with simple geometries. It assumes that the surface is represented by a single parameterization, and therefore surfaces with

coordinate singularities cannot be treated, and the singularity has to be excised. Most scatterer surfaces with complicated geometries, for example, cannot be represented by a single non-singular explicit parameterization. This problem can be resolved by splitting the scatterer surface into several simpler surfaces with explicit parameterizations. These multiple patches collectively cover the scatterer surface in a non-singular manner. Moreover, one can get other benefits by this way:

1. Smaller gradients in the solution by refining the patches with higher varying velocity coefficients.
2. Possibility to parallelize, since the patches can be handled independently.
3. Less internal memory needed.
4. Using the possible symmetry of the scatterer (for example for an ellipsoid).

In this paper, we consider a two-dimensional compact manifold M embedded in $\mathbb{R}^d$ and compute trajectories on the manifold. We first consider the case when the manifold is represented by a single regular parameterization and modify the fast phase space method [9, 26] to a more general class of problems. Second, we consider the case when the manifold is represented by an atlas of charts and modify the single-patch phase space method to this case. In both cases, dynamics of trajectories are given by systems of first-order ODEs.

Multiple-patch (or multi-block) finite difference schemes have long been used in computational science. They are a sub-class of domain decomposition methods for solving PDEs by iteratively solving sub-problems on smaller sub-domains [5]. However, the scheme presented here is not based on iterations. Another domain decomposition method related to the multiple-patch algorithm is the slowness matching Eulerian method [34], where local single-valued solutions of the eikonal equations are patched together by slowness matching to obtain a global, multi-valued traveltime field.

In Section 2, we give the governing equations describing the dynamics of trajectories on two classes of compact manifolds: the manifolds which can be represented by a single regular parameterization and the manifolds which are described by an atlas of charts. The construction of the single and multiple-patch schemes are described in Section 3 and 4, respectively. In Section 5 and 6, we present applications in computing creeping rays and seismic waves, together with sample numerical results from a prototype implementation of the scheme.

2 Governing Equations

Consider a two-dimensional compact manifold M embedded in $\mathbb{R}^d$. We want to compute trajectories on the manifold. Since we are interested in applications to wave propagation problems, it is natural to consider the trajectories as rays, and we will use this terminology henceforth.

We consider two cases: when the manifold is represented by a single regular parameterization, and when the manifold is represented by an atlas of charts. In both cases, dynamics of rays are given by systems of three first-order ODEs describing the rate of change of the rays' location and direction along the ray trajectories.

2.1 Single-Patch Manifolds

First, assume that the manifold can be represented by a regular parameterization $\mathbf{x} = \bar{X}(\mathbf{u})$, where $\mathbf{x} \in M$, and the parameters $\mathbf{u} = (u, v)$ belong to a set $\Omega \subset \mathbb{R}^2$. Note that if M is a hypersurface or a plane embedded in $\mathbb{R}^3$, then $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$, and if M is a plane in $\mathbb{R}^2$, then $\mathbf{x} = (x, y) \in \mathbb{R}^2$.

We introduce the phase space $\mathbb{P} = \mathbb{R}^2 \times \mathbb{S}$, where $\mathbb{S} = [0, 2\,\pi]$, and consider the triplet $\gamma = (u, v, \theta)$ as a point in this space. Let the rays be given by a system of three ODEs

$$\dot{\gamma} = \mathbf{g}(\gamma), \tag{4}$$

where the dot denotes differentiation with respect to the parameter τ being the arc length along the rays, and $\mathbf{g} = (g_1, g_2, g_3)$ is a given three-vector function which is periodic in $\theta \in \mathbb{S}$. The ray trajectories on M are then confined to a subdomain $\Omega_p = \Omega \times \mathbb{S} \subset \mathbb{P}$ in phase space. Note that the parameter values $\mathbf{u} = (u, v)$ represent the rays' location $\bar{X}(\mathbf{u})$ on M, and the angle θ represents the direction of the rays.

Remark 1. *A generic Hamiltonian system with Hamiltonian $H(\mathbf{u}, \mathbf{p})$ in four-dimensional space $\Omega \times \mathbb{R}^2$, with $\mathbf{p} \in \mathbb{R}^2$, can typically be reduced to the form (4). Here, θ can, for instance, be an angle representing the direction of vector $\mathbf{p}$. For example, if $H = |\mathbf{p}|^2 + V(\mathbf{u}) \equiv C$, one can reduce it by setting*

$$\mathbf{p} = (C - V(\mathbf{u}))^{1/2} \, (\cos\theta \;\; \sin\theta)^\top,$$

where C is determined by initial data. See also Section 5 and Section 6 for more examples.

Moreover, let any information transporting along the rays, represented by a (possibly vector-valued) function $\beta(\tau)$, be given by a more general system of the form

$$\dot{\beta} = \alpha(\gamma, \beta), \tag{5}$$

where $\alpha(\gamma, \beta)$ is some given function. For example, when β is the length of the ray, we have $\alpha \equiv 1$.

2.2 Multiple-Patch Manifolds

There are two main classes of problems for which representing the manifold by a single parameterization is not applicable: the manifolds which cannot be described by a single regular parameterization due to singularities, e.g. an airplane surface, and the manifolds with different (discontinuous) material properties, e.g. earth consisting of materials with different seismic velocities. The former is of topological and geometrical nature related to the underlying manifold, and the latter is more special application oriented.

We therefore, secondly, consider the more general case when M is described by an atlas of charts (M_j, w_j), with $j = 1, \ldots, P$, where the sets M_j collectively cover M, and the mapping $w_j : M_j \to \Omega$ is bijective. In particular, we assume that Ω is the unit square and M_j are patches with parametric equations

$$\mathbf{x} = \bar{X}_j(\mathbf{u}) : [0, 1]^2 \to M_j \subset \mathbb{R}^d,$$

and the mappings are $w_j = \bar{X}_j^{-1}$. Then $M = \bigcup_j w_j^{-1}([0, 1]^2)$. Note that although the sets are closed, we still consider M as an atlas. We assume further that the patches stick together

along their sides (patch boundaries) and denote the side between two connected patches M_j and $M_{j'}$ by $S_{jj'}$. Note that it is possible to have $j = j'$, for instance when M is a torus. When $j \neq j'$, we have $S_{jj'} = M_j \cap M_{j'}$. It is also possible that a patch does not share a side with another patch, for example, if the manifold has boundary (e.g. a finite cylinder). We denote such a side by S_{0j} which belongs only to M_j. Denote the set of all sides by $\mathcal{S}$.

For each patch with the id number j, let the rays be given by a system of three equations set in Ω_p,

$$\dot{\gamma} = \mathbf{g}_j(\gamma), \tag{6}$$

where $\mathbf{g}_j = (g_1^j, g_2^j, g_3^j)$ is a given three-vector function. Note that $\mathbf{g}_j$ may be different for different j. As before, the systems (6) are natural structures for Hamiltonian systems on four-dimensional spaces $\Omega \times \mathbb{R}^2$ with Hamiltonian $H_j(\mathbf{u}, \mathbf{p})$ whose order are reduced by one.

Correspondingly, let any information transporting along the rays, represented by a (possibly vector-valued) function $\beta(\tau)$, be given by a system of the form

$$\dot{\beta} = \alpha_j(\gamma, \beta), \tag{7}$$

where $\alpha_j(\gamma, \beta)$ is a given function.

A main difference between the numerical methods for the single patch representation of the manifold and the multiple-patch case is that in the latter we need to connect the solutions of adjacent patches and impose suitable conditions at the inter-patch boundaries. In order to treat this problem, we need to introduce a global space, which is bijective with the space $\mathbb{Z}_P \times \Omega_p$, and in which the boundary conditions are defined and can easily be handled. Here $\mathbb{Z}_P = \{1, 2, \ldots, P\}$. We first note that by our assumptions above, there is a bijective mapping between $(j, \mathbf{u}) \in \mathbb{Z}_P \times \Omega$ and $\mathbf{x} \in M$, except when $\mathbf{x}$ is at patch boundaries ($\mathbf{x} \in S_{jj'}$). Now, let $T_{\mathbf{x}}M$ be the tangent plane (the set of tangent vectors) to M at point $\mathbf{x} \in M$ and $TM = \bigcup_{\mathbf{x}\in M} T_{\mathbf{x}}M$ be the tangent bundle of M. The dimension of TM is twice the dimension of M. An element of TM is a pair $\Gamma := (\mathbf{x}, \mathbf{q})$ where $\mathbf{x} \in M$ and $\mathbf{q} \in T_{\mathbf{x}}M$. We consider the unit tangent bundle UTM of M which contains all unit-normed tangent vectors ($\|\mathbf{q}\| = 1$). Note that UTM is a three-dimensional manifold embedded in $\mathbb{R}^{2d}$.

We now want to prove that the unit tangent bundle UTM is in fact the global manifold which is bijective with the space $\mathbb{Z}_P \times \Omega_p$. But, before the proof, we notice that, by construction, for each point $\Gamma = (\mathbf{x}, \mathbf{q}) \in UTM$, there is a well-defined patch id number $j = \mathcal{J}(\Gamma)$, except when $\mathbf{x}$ is on patch boundaries. We extend this function also to the patch boundaries as follows:

- if $\mathbf{x} \in S_{jj'}$ and $\mathbf{q} \nparallel S_{jj'}$, then $\mathcal{J}(\Gamma) = \lim_{\epsilon\to 0} \arg\min_j \ dist(\mathbf{x} + \epsilon\mathbf{q}, M_j)$, which means that $\mathcal{J}(\Gamma)$ is the id of the patch into which the ray starting at Γ enters.
- if $\mathbf{x} \in S_{jj'}$ and $\mathbf{q} \parallel S_{jj'}$, then $\mathcal{J}(\Gamma) = \max(j, j')$.

Where by $\mathbf{q} \parallel S_{jj'}$, we mean that $\mathbf{q}$ is parallel to the patch boundary in an interval around $\mathbf{x} \in S_{jj'}$. Therefore in this case, Γ belongs to both UTM_j and $UTM_{j'}$, and we can choose either of j' and j' as the value of the function $\mathcal{J}(\Gamma)$. In order to have a well-defined function, we choose the larger one. Moreover, if $\mathbf{x}$ is at a corner sharing several patches $j, j', j'', \ldots,$ and $\mathbf{q}$ is parallel to $S_{jj'}$, we again choose $\mathcal{J}(\Gamma) = \max(j, j')$.

We now prove the following Lemma.

Lemma 1. *Suppose the Jacobian $J_j = D_{\mathbf{u}}\bar{X}_j \in \mathbb{R}^{d\times 2}$ has full rank for all $(j,\mathbf{u}) \in \mathbb{Z}_P \times \Omega$. For each j there is then a bijective mapping $W_j : UTM_j \to \Omega_p$ given by $W_j(\Gamma) = \gamma$, where*

$$\Gamma = (\mathbf{x}, \mathbf{q}), \quad \mathbf{q} = \frac{J_j(w_j(\mathbf{x}))\hat{s}(\theta)}{|J_j(w_j(\mathbf{x}))\hat{s}(\theta)|}, \quad \hat{s}(\theta) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \quad \gamma = (w_j(\mathbf{x}), \theta). \tag{8}$$

Moreover, there is a bijective mapping between $(j,\gamma) \in \mathbb{Z}_P \times \Omega_p$ and $\Gamma = (\mathbf{x},\mathbf{q}) \in UTM$.

Proof. First assume that $\mathbf{x} \in M_j$ and $\mathbf{q} \in UT_{\mathbf{x}}M_j$. Since the mapping $w_j = \bar{X}_j^{-1}$ is bijective, there is $\mathbf{u}$ such that $\bar{X}_j(\mathbf{u}) = \mathbf{x}$, given by $\mathbf{u} = w_j(\mathbf{x})$. Moreover, since the Jacobian $J_j(\mathbf{u})$ has full rank, its columns span the tangent plane at $\mathbf{x}$, and since $\mathbf{q}$ belongs to this plane, there exists a solution θ to

$$\frac{J_j(\mathbf{u})\hat{s}(\theta)}{|J_j(\mathbf{u})\hat{s}(\theta)|} = \mathbf{q}, \qquad \hat{s}(\theta) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}.$$

The second statement follows since $\mathcal{J}(\Gamma)$ is well-defined for all $\Gamma \in UTM$. This proves the lemma. □

Note that the atlas of charts (UTM_j, W_j) describe the space $UTM = \bigcup_j W_j^{-1}(\Omega_p)$. Figure 1 shows a schematic representation of the two-dimensional manifold M, the three-dimensional space UTM and the corresponding bijective mappings to the parameter space Ω and phase space Ω_p.

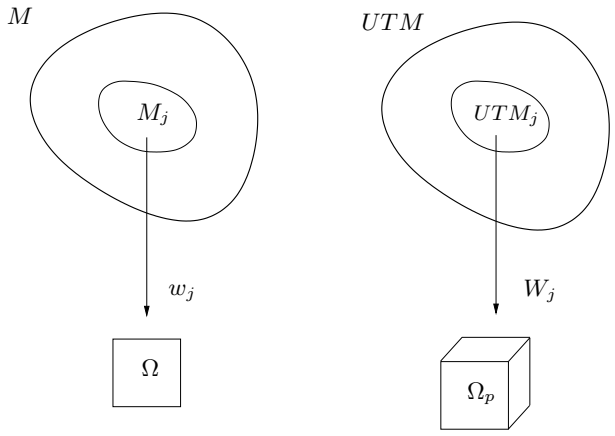


Figure 1: A schematic representation of the two-dimensional manifold M embedded in $\mathbb{R}^d$ and the three-dimensional space UTM embedded in $\mathbb{R}^{2d}$. The bijective mappings w_j and W_j map a chart j of these manifolds to the two-dimensional parameter space Ω and the three-dimensional phase space Ω_p, respectively.

2.2.1 Boundary Conditions

We may have different boundary conditions at the patch boundaries. In some problems, the rays are continuous at the patch boundaries. Such problems include geodesics and creeping rays computations on a hypersurafce with constant index of refraction. In these problems, the boundary conditions are determined easily by the continuity of rays. In some problems, the rays may not be continuous at the patch boundaries. For example, seismic propagation in a multi-layered media with different seismic velocities is such a problem, in which the boundary conditions are determined by Snell's law of refraction or the law of reflection.

As was mentioned before, the inter-patch boundary conditions are given in physical space in terms of $\Gamma \in UTM$, rather than in terms of $\gamma \in \Omega_p$. Let $\Gamma = (\mathbf{x}, \mathbf{q})$, where $\mathbf{x} \in S_{jj'}$ and $j' = \mathcal{J}(\Gamma) \neq j$, which means that the ray arrives at the side $S_{jj'}$ from patch M_j. The inter-patch boundary condition at $S_{jj'}$ is given by,

$$\tilde{\Gamma} = \mathcal{L}_{jj'}(\Gamma),$$

where $\mathcal{L}_{jj'}$ is some known function, and $\tilde{\Gamma} = (\tilde{\mathbf{x}}, \tilde{\mathbf{q}}) \in UTM_{\mathcal{J}(\tilde{\Gamma})}$. For example, depending on the ray arriving at the side $S_{jj'}$ from patch M_j, we may have the following boundary conditions:

- if the ray is continuous, then $\mathcal{L}_{jj'}$ is the identity function

$$\tilde{\mathbf{x}} = \mathbf{x}, \quad \tilde{\mathbf{q}} = \mathbf{q}.$$

- if the ray is refracted, then

$$\tilde{\mathbf{x}} = \mathbf{x}, \quad \tilde{\mathbf{q}} = \tilde{\mathcal{S}}(\mathbf{x}, \mathbf{q}).$$

- if the ray is reflected, then

$$\tilde{\mathbf{x}} = \mathbf{x}, \quad \tilde{\mathbf{q}} = \tilde{\mathcal{R}}(\mathbf{x}, \mathbf{q}).$$

Here, the functions $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{R}}$ are determined by Snell's law of refraction and the law of reflection, respectively. See Section 6.2 for more details.

In the next two sections, we present a patch-based phase space method for computing ray trajectories on manifolds. First, we consider the case when the manifold is represented by a single parameterization and construct a single-patch phase space method based on writing the systems (4-5) in a Eulerian framework. Next, we consider a wider class of manifolds which are represented by multiple parameterizations and introduce a multiple-patch phase space method based on solving the Eulerian version of systems (6-7) in each patch and connecting the solutions of individual patches using suitable inter-patch boundary conditions. In both methods, properties for particular ray families are obtained through a fast post-processing.

3 Single-Patch Phase Space Scheme

We consider the case when the two-dimensional manifold M embedded in $\mathbb{R}^d$ is represented by a single regular parameterization. The objective is to compute the ray trajectories together with the information transported along them on M. First, the system of ODEs (4) and (5), describing rays and other information, are formulated as time-independent Eulerian PDEs in phase space. These equations are then solved numerically on a fixed computational grid. The solution to the PDEs is post-processed to extract information for a particular family of rays.

3.1 Mathematical Formulation

We consider a ray $\bar{\gamma}(\tau)$ satisfying (4), starting at $\bar{\gamma}(0) = \gamma = (u, v, \theta) \in \Omega_p$ and ending at the boundary $\partial\Omega_p = \partial\Omega \times \mathbb{S}$. We call this end point $(U, V, \Theta) \in \partial\Omega_p$ the *escape point* of the ray. See Figure 2. We then define three types of unknown *escape functions* for this ray, as follows:

- $F : \mathbb{P} \to \mathbb{P}$, $F(\gamma) = (U, V, \Theta)$ is the escape point.
- $\Phi : \mathbb{P} \to \mathbb{R}$ is the length of the ray. We also refer to this as the travel-time of the ray.
- $B : \mathbb{P} \to \mathbb{R}$ is a function representing a relation between the β-values at the escape and starting points, where β satisfies (5).

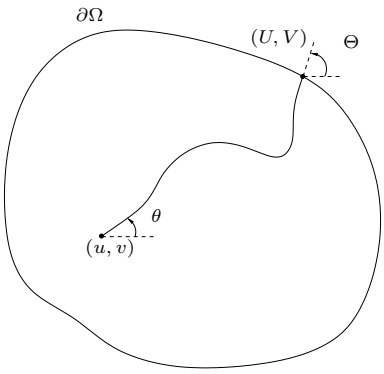


Figure 2: A ray trajectory in the parameter space, starting at $\gamma = (u, v, \theta) \in \Omega_p$ and ending at the escape point $F(\gamma) = (U, V, \Theta) \in \partial\Omega_p$.

Each escape function $f(\gamma)$ of the above types satisfies an ODE,

$$\frac{d}{d\tau} f(\gamma(\tau)) = h\big(\gamma(\tau), f(\gamma(\tau))\big), \tag{9}$$

where the forcing term h is 0, 1 and $\alpha(\gamma, f)$ for $f = F$, $f = \Phi$ and $f = B$, respectively.

Using the chain rule, the *escape* PDE for each escape function $f(\gamma)$ reads

$$g_1(\gamma)\, f_u + g_2(\gamma)\, f_v + g_3(\gamma)\, f_\theta = h(\gamma, f), \qquad \gamma \in \Omega_p, \tag{10}$$

with the boundary condition at inflow points of $\partial\Omega_p$,

$$f(\gamma) = b, \qquad \gamma \in \partial\Omega_p^{\text{inflow}}, \qquad \partial\Omega_p^{\text{inflow}} = \left\{\gamma \in \partial\Omega_p \,|\, \hat{n}(\gamma)^\top \mathbf{g}(\gamma) < 0\right\},$$

with $\hat{n}$ being the outward normal vector in the phase space.

Note that for the first two types of escape functions $f = F$ and $f = \Phi$, the boundary value b is γ and 0, respectively. For the third type $f = B$, if for instance B is the difference or ratio between β-values at the escape and starting points, the boundary value are $b = 0$ or $b = 1$, respectively.

The escape equation (10) is a linear hyperbolic equation, and the variable velocity coefficients $\mathbf{g} = (g_1, g_2, g_3)$ are known and determine the characteristic direction at every point $\gamma \in \Omega_p$.

One important property of the solutions to the escape PDEs is that they are in general discontinuous due to discontinuous boundary conditions. This happens, for example, when a characteristic touches a boundary tangentially, such that at some points on the plane the characteristic is in-going, and suddenly it becomes out-going.

3.2 Numerical Solution of the escape PDEs

We now want to solve (10) numerically. We discretize the phase space domain $\Omega_p = \Omega \times \mathbb{S}$ uniformly, setting $u_i = i\Delta u$, $v_j = j\Delta v$ and $\theta_k = k\Delta\theta$, with the step sizes $\Delta u = \Delta v = \frac{1}{N}$ and $\Delta\theta = \frac{2\pi}{N}$, assuming Ω is the unit square. Moreover, let f_{ijk} approximate an escape function $f(u_i, v_j, \theta_k)$.

In addition to the boundary condition at inflow points, since the function f is periodic in θ, we use periodic boundary conditions,

$$f(u, v, 0) = f(u, v, 2\,\pi),$$

as numerical boundary conditions.

There are different methods for solving the escape equation (10). One way is to discretize the PDEs in the phase space using a finite difference, finite volume or finite element approximation and arrive at a system of linear equations $A\bar{f} = \bar{b}$, where A is a $N^3 \times N^3$ matrix with a sparse structure and $\bar{b} \in \mathbb{R}^{N^3}$ represents the boundary conditions. This system can then be solved iteratively, and one can speed up the computations using suitable preconditioners [12, 4]. However, in the case that characteristics change direction many times in the phase space domain, it is difficult to find good preconditioners.

Another way to solve the escape equations is to write them as

$$f_t + g_1\, f_u + g_2\, f_v + g_3\, f_\theta = h,$$

and solve these time-dependent equations until the steady state $f_t = 0$. This method can be seen as an iterative method. Finding a fast algorithm which is not much restricted by the CFL condition is analogous to finding a good preconditioner in the iterative method.

Yet, another way to solve the equation (10) is to compute the approximate solution f_{ijk} using a ray tracing method, which traces back along the characteristic to the initial boundary from each grid point (i, j, k). The main drawback with this method is that it will be expensive, because one needs to trace back all N^3 points in the domain all the way to the boundary.

Instead, we use a Fast Marching algorithm, given by Fomel and Sethian [9]. A similar method in two-dimensional space was also proposed in [16]. The basic idea of the algorithm is to march the solution outwards from the boundary and use the characteristic directions to update grid values. Note that in the algorithm, we always also compute Φ_{ijk} besides f_{ijk}.

First, the grid points are divided into three classes:

- *Accepted*: the correct values of f_{ijk} and Φ_{ijk} have been computed.
- *Considered*: adjacent to *Accepted* for which f_{ijk} and Φ_{ijk} have already been computed, but may be corrected by a later computation.
- *Far*: the correct values of f_{ijk} and Φ_{ijk} are not known.

The major steps of the algorithm are then as follows:

0. Start with all nodes $(u_i, v_j, \theta_k) \in \Omega_p$ in *Far*, and assign Φ_{ijk} at these nodes a large value. This large value needs to be greater than the length (travel-time) of every possible ray in the computational domain. Put the boundary nodes $(u_i, v_j, \theta_k) \in \partial\Omega_p^{\text{inflow}}$ in *Accepted*, and assign f_{ijk} and Φ_{ijk} at these nodes the correct boundary values. Put all nodes

adjacent to *Accepted*, for which the characteristic[1] at that node points back to the boundary, in *Considered*. Each *Considered* node is then given a value by using a local cell characteristic method.

1. Take the *Considered* node with the smallest arrival time Φ_{ijk} as *Accepted*.
2. Find the octant toward which the characteristic going through that node points.
3. For each neighboring grid point in the octant which is not *Accepted* use the local cell characteristic method to (possibly) compute new values for f_{ijk} and Φ_{ijk}. In the case we can compute new values for a *Far* node, put it in *Considered*.
4. Loop to step 1 until all points are *Accepted*.

Since in [9] the local cell characteristic method, used in steps 0 and 3 of the algorithm, is not discussed, we will here describe a version of first and second order local cell-based ray tracing methods using a local linear and parabolic ray tracing and the Taylor expansion of the trajectory near the starting point.

Consider a grid cell in Ω_p, and assume we want to compute the value of f_{ijk} at a corner of this cell, knowing the correct values of f at some neighboring grid points. The output of the local ray tracing would be either a new value for f_{ijk} or no new value, depending on whether the neighboring points, to which the characteristic points back, are *Accepted* or not. See Figure 3.

Let τ be the arc length parameterization along the characteristic $\gamma(\tau)$. We start at $\gamma(0) = (u_i, v_j, \theta_k)$, where we want to compute a possibly new value, and trace backwards along the characteristic to intersect a cell face at $\gamma(\tau^*)$, $\tau^* < 0$. We Taylor expand f near the starting point,

$$f(\gamma(\tau^*)) = f(\gamma(0)) + \tau^* \frac{d}{d\tau} f(\gamma(0)) + \frac{\tau^{*2}}{2} \frac{d^2}{d\tau^2} f(\gamma(0)) + \mathcal{O}(\tau^{*3}), \tag{11}$$

with local truncation error $\mathcal{O}(\tau^{*3}) \approx \mathcal{O}(\Delta u^3)$. Note that $\frac{d}{d\tau} f(\gamma(0))$ and $\frac{d^2}{d\tau^2} f(\gamma(0))$ in (11) are given by:

$$\begin{aligned}
\frac{d}{d\tau} f(\gamma(0)) &= h(\gamma(0), f(\gamma(0))), \\
\frac{d^2}{d\tau^2} f(\gamma(0)) &= \frac{d}{d\tau} h\big(\gamma(0), f(\gamma(0))\big) \\
&= \mathbf{g}(\gamma(0)) \cdot \nabla_\gamma h\big(\gamma(0), f(\gamma(0))\big) + h_f\big(\gamma(0), f(\gamma(0))\big)\, h\big(\gamma(0), f(\gamma(0))\big).
\end{aligned}$$

Therefore, to find $f(\gamma(0))$, with accuracy of $\mathcal{O}(\tau^{*3})$, we need to know τ^* and $f(\gamma(\tau^*))$. Note that for $f = F$ and $f = \Phi$, since $\frac{d}{d\tau} F(\gamma(\tau)) = 0$ and $\frac{d}{d\tau} \Phi(\gamma(\tau)) = 1$, the expansion (11) reduces to

$$F(\gamma(\tau^*)) = F(\gamma(0)), \tag{12}$$

$$\Phi(\gamma(\tau^*)) = \Phi(\gamma(0)) + \tau^*. \tag{13}$$

[1] We approximate the characteristic by a piecewise linear curve for a first order method and piecewise parabolic for a second order method.

3.2.1 First Order Method

We assume that characteristics are linear in each cell. Therefore, we can write

$$\gamma(\tau) \approx \sigma_1 + \sigma_2\,\tau, \quad \sigma_1 = \gamma(0), \quad \sigma_2 = \dot{\gamma}(0) = \mathbf{g}(\gamma(0)).$$

Note that σ_1 and σ_2 are known. There are six possible planes, $u = u_{i\pm1}$, $v = v_{j\pm1}$ and $\theta = \theta_{k\pm1}$, which this line can intersect. We, therefore, get six crossing points $\tau_1, \ldots, \tau_6$, which are solutions of six linear equations. It is then clear that $\tau^* = \max_{\tau_j<0} \tau_j$. Knowing the crossing face and the crossing point $\gamma(\tau^*)$, we continue as follows:

a. If all four points of the cell face are *Accepted*, use these points to interpolate a value of $f(\gamma(\tau^*))$. Then use the first two terms of the Taylor expansion (11) to compute a new value for $f_{ijk} \approx f(\gamma(0))$. Note that we need to solve a (possibly) nonlinear algebraic equation, when h depends on f,

$$f(\gamma(0)) = f(\gamma(\tau^*)) - \tau^*\, h(\gamma(0), f(\gamma(0))).$$

Put this node in *Considered*. Since the method is first order, a two dimensional bilinear interpolation is used. See Figure 3.

b. If no points on the cell face are *Accepted*, do not update the value.

c. Else, continue tracing along the characteristic until either (a) or (b) occurs. Note that each time the characteristic enters a new cell, the new starting point needs to be updated.

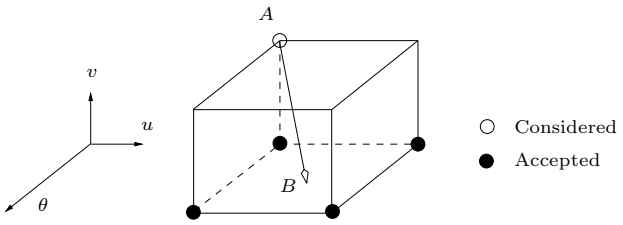


Figure 3: A grid cell in Ω_p. Point A is updated by tracing the characteristic back to point B and interpolating from the accepted values. Here, points A and B correspond to $\gamma(0)$ and $\gamma(\tau^*)$, respectively.

3.2.2 Second Order Method

We assume that characteristics are parabolic in each cell and write

$$\gamma(\tau) \approx \sigma_1 + \sigma_2\,\tau + \sigma_3\,\tau^2, \quad \sigma_1 = \gamma(0), \quad \sigma_2 = \dot{\gamma}(0), \quad \sigma_3 = \frac{1}{2}\ddot{\gamma}(0) = \frac{1}{2} D_\gamma \dot{\gamma}(0)\,\dot{\gamma}(0).$$

Note that σ_1, σ_2 and σ_3 are known. In this case, there are nine possible cell faces which can intersect this parabola; $u = u_i$, $v = v_j$, $\theta = \theta_k$ and the six faces in the linear case. By intersecting the parabola with the faces, we get nine crossing points $\tau_1, \ldots, \tau_9$, which are solutions of simple quadratic equations. We then get $\tau^* = \max_{\tau_j<0} \tau_j$ and continue in the following way:

a. Pick the crossing face and eight faces around it in the same plane, sharing sixteen grid points in total. If all sixteen points are *Accepted*, use these points to interpolate a value of $f(\gamma(\tau^*))$. Then use the first three terms of the Taylor expansion (11) to compute a new value for $f_{ijk} \approx f(\gamma(0))$. Note that, again, we need to solve a (possibly) nonlinear algebraic equation, when h depends on f,

$$f(\gamma(0)) = f(\gamma(\tau^*)) - \tau^* h(\gamma(0), f(\gamma(0))) - \frac{\tau^{*2}}{2} \frac{d}{d\tau} h\big(\gamma(0), f(\gamma(0))\big).$$

Put this node in *Considered*. Because the solution can be discontinuous, we use a version of two dimensional essentially non-oscillatory (ENO) interpolation based on Newton divided differences and the Newton formulation of the interpolation polynomial. Among four points in each dimension, we pick up either the left three or the right three points which have a smaller divided difference and use a second order polynomial. See [33].

b. If no points on the cell face are *Accepted*, do not update the value.

c. Else, continue tracing along the characteristic until either (a) or (b) occurs. Note that each time the characteristic enters a new cell, the new starting point needs to be updated.

The algorithm is a one-pass algorithm and is of complexity $\mathcal{O}(N^3 \log N)$. Note that we use heap sort algorithm for extracting the smallest arrival time Φ_{ijk} of *Considered* nodes and for inserting new updated values of *Considered* nodes. There is however no proof of convergence for the method.

3.3 Post-Processing

Solutions of the escape PDEs (10) give the escape point, length and other information for rays with all possible starting points in the phase space. These solutions need to be post-processed to extract properties for a ray family.

As an example, suppose we want to compute the length of the ray between two points $\mathbf{u}_1$ and $\mathbf{u}_2$ in the parameter space Ω. We first observe that $F(\gamma_1) = F(\gamma_2)$, if and only if the points γ_1 and γ_2 lie on the same ray. We can thus find θ_1 and θ_2, as the solution to

$$F(\mathbf{u}_1, \theta_1) = F(\mathbf{u}_2, \theta_2). \tag{14}$$

The length is then given by $|\Phi(\mathbf{u}_1, \theta_1) - \Phi(\mathbf{u}_2, \theta_2)|$. Note that there may be multiple solutions to (14), giving multiple lengths. If $\mathbf{u}_2 \in \partial\Omega$, the expression simplifies to solving

$$\big(U(\mathbf{u}_1, \theta_1), V(\mathbf{u}_1, \theta_1)\big) = \mathbf{u}_2, \tag{15}$$

for θ_1 to get the length $\Phi(\mathbf{u}_1, \theta_1)$.

To solve (14), we note that since $F = (U, V, \Theta) \in \partial\Omega_p$ is a point on the phase space boundary, it can be reduced to a point (S, Θ) in $\mathbb{R}^2$, where S represents the escape location on the boundary $\partial\Omega$. For example if $\Omega = [0, 1]^2$, we can choose $S \in [0, 2\,\pi]$ along $\partial\Omega$ such that $S = 0$, $S = \pi$ and $S = 2\pi$ for $(U, V) = (0, 0)$, $(U, V) = (1, 1)$ and $(U, V) = (0, 0)$, respectively. The left and right hand sides of (14) are then curves in $\mathbb{R}^2$ parameterized by θ_1 and θ_2, and solving the algebraic equation (14) amounts to finding crossing points of these curves.

Having the discrete solutions at the points $\mathbf{u}_1$ and $\mathbf{u}_2$ for all N directions, we then need to find crossing points of two complex lines of N straight line segments as the solutions to (14). This can be done with a complexity of $\mathcal{O}(N)$; see e.g. [35]. We note that in the case that a second order method for solving the escape equations is used, the linear intersection algorithm will not affect second order accuracy of the method. In fact, the intersection algorithm is performed only to find the intersection's neighboring points. We use a higher order interpolation to compute the initial angles θ_1 and θ_2 and the escape functions corresponding to these angles. The complexity of finding the ray length between one fixed source point and all other N^2 points in Ω is then $\mathcal{O}(N^3)$, and the total complexity, including solving the escape PDEs, will therefore be $\mathcal{O}(N^3 \log N)$. This is expensive for computing this so called travel-time field for only one source point. For example by using wave front tracking or solvers based on the surface eikonal equation, the complexity is $\mathcal{O}(N^2)$. However, if the solutions are sought for many source points, the phase space method can be more efficient. See Section 5 for such an example.

4 Multiple-Patch Phase Space Scheme

We now consider the more complicated and realistic case when the manifold M cannot be represented by one regular parameterization. We let M be described by an atlas of charts or multiple patches and want to compute the ray trajectories together with the information transported along them on the manifold. First, the system of ODEs (6) and (7) in each chart (patch) are formulated as time-independent Eulerian PDEs and solved numerically on a fixed computational grid in phase space. The solutions to the PDEs in each chart are then connected using suitable inter-patch boundary conditions. Information for a particular family of rays are then extracted through a fast post-processing.

We describe the multiple-patch scheme and the key design choices in such a scheme, including the number and shape of patches, the treatment of inter-patch boundaries and the choice of escape boundary.

4.1 Multiple-Patch Construction

We first want to define a function $\mathbf{F}$ for the multiple patch case that corresponds to the single patch solution F described in Section 3. Let $\mathcal{R}$ be some curve in M, representing an escape boundary. We consider a ray starting at a point $\Gamma \in UTM$ and define $\mathbf{F}(\Gamma) : UTM \to UTM$ as mapping the point Γ to another point in the space UTM where the projection of the ray onto M first crosses $\mathcal{R}$ (assuming such a point exists).

If the compact manifold has a boundary (e.g. a finite cylinder), we let this be the escape boundary, similar to the case of a single-patch manifold. Hence, $\mathcal{R} = \bigcup_j S_{0j}$. However, for a compact boundaryless manifold (e.g. a sphere or a torus), there is no obvious escape boundary, as in the single patch case. In this case we will let

$$\mathcal{R} = \bigcup_{(j,j')\subset R} S_{jj'} \subset \mathcal{S} \tag{16}$$

be the escape boundary, where R is some index set, to be determined (see below).

To compute $\mathbf{F}(\Gamma)$, we first recall that, by construction, for each point $\Gamma = (\mathbf{x}, \mathbf{q}) \in UTM$, there is a well-defined patch id number $j = \mathcal{J}(\Gamma)$ and a well-defined mapping $W_j : UTM_j \to$

Ω_p.

Now, suppose $F_j(\gamma)$ are the solutions to the escape PDE (10), with $f = F$, in Ω_p corresponding to each patch with $j = 1, \ldots, P$. The function $\mathbf{F}(\Gamma)$ is then given recursively by

$$\tilde{\Gamma}_0 = \Gamma, \tag{17}$$

and while $\tilde{\mathbf{x}}_n \notin \mathcal{R}$, where $\tilde{\Gamma}_n = (\tilde{\mathbf{x}}_n, \tilde{\mathbf{q}}_n)$,

$$j = \mathcal{J}(\tilde{\Gamma}_n), \quad \Gamma_{n+1} = W_j^{-1} F_j(W_j(\tilde{\Gamma}_n)), \quad j' = \mathcal{J}(\Gamma_{n+1}), \quad \tilde{\Gamma}_{n+1} = \mathcal{L}_{jj'}(\Gamma_{n+1}), \tag{18}$$

where $\mathcal{L}_{jj'}$ is the operator representing the inter-patch boundary conditions between patches M_j and $M_{j'}$. Then $\mathbf{F}(\Gamma) = \Gamma_{n^*}$, where n^* is the smallest index for which $\mathbf{x}_{n^*} \in \mathcal{R}$.

Remark 2. *If the rays are continuous at the patch boundaries, $\mathcal{L}_{jj'}$ will be the identity function ($\tilde{\Gamma}_{n+1} = \Gamma_{n+1}$). From the above recursive formula, it is easy to see that, in order to compute the function $\mathbf{F}$ for all points in UTM it is enough to know the escape PDE solutions F_j in all patches and the patch transfer functions $\mathcal{T}_{jj'} = W_{j'} W_j^{-1}$ at all sides connecting two patches M_j and $M_{j'}$. Note that these transfer functions can be easily calculated from the mappings W_j. As an example, in Section 5, we will discuss the computation of creeping rays which are continuous at patch boundaries.*

If the rays are not continuous at the patch boundaries, each time they pass a boundary, the coordinates of Γ_{n+1} may change ($\tilde{\Gamma}_{n+1} \neq \Gamma_{n+1}$). It happens when, for example, the rays change their direction as they enter another patch with different properties. The patch transfer functions are then changed to $\mathcal{T}_{jj'} = W_{j'} \mathcal{L}_{jj'} W_j^{-1}$. Here, transfer functions are again easily calculated from the mappings W_j and the inter-patch boundary conditions. We will consider such examples in Section 6, where the rays change direction according to Snell's law of refraction and the law of reflection.

Similar to $\mathbf{F}(\Gamma)$, we can define the functions $\mathbf{\Phi}(\Gamma)$ and $\mathbf{B}(\Gamma)$ in UTM for the multiple patch case corresponding to the single patch functions Φ and B described in Section 3. Assuming $\Phi_j(\gamma)$ and $B_j(\gamma)$ are the solutions to the escape PDE (10), with $f = \Phi$ and $f = B$, respectively, in Ω_p corresponding to each patch with $j = 1, \ldots, P$, we can write

$$\mathbf{\Phi}(\Gamma) = \sum_{n=0}^{n^*-1} \Phi_j(W_j(\tilde{\Gamma}_n)),$$

with j and $\tilde{\Gamma}_n$ as in (17)-(18), and

$$\mathbf{B}(\Gamma) = \sum_{n=0}^{n^*-1} B_j(W_j(\tilde{\Gamma}_n)),$$

if B is, for example, the difference between β-values at Γ_{n^*} and Γ_0, and

$$\mathbf{B}(\Gamma) = \prod_{n=0}^{n^*-1} B_j(W_j(\tilde{\Gamma}_n)),$$

if B represents, for example, the ratio between β-values.

4.2 Post-Processing

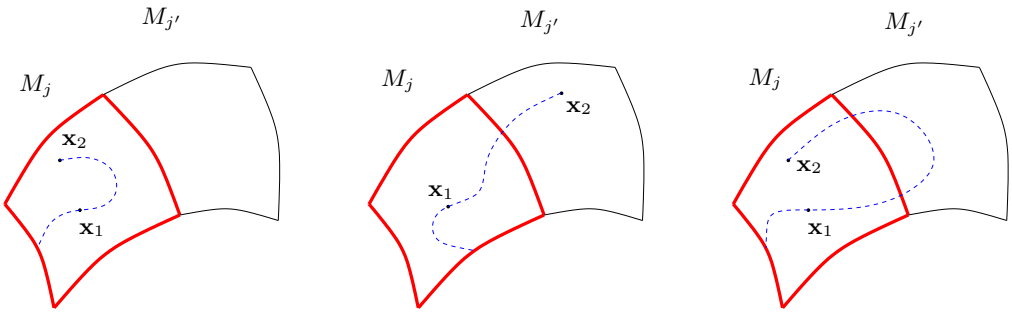
Suppose we want to compute the length of a ray connecting two points $\mathbf{x}_1 \in M_{j_1}$ and $\mathbf{x}_2 \in M_{j_2}$. In order to find this ray, if the manifold has a boundary, we let this be the escape boundary, and the post-processing is similar to the single patch case with F replaced by $\mathbf{F}$. In the case of a boundaryless manifold, we choose the boundaries of M_{j_1} as the escape boundary $\mathcal{R}$. We then find $\mathbf{F}\left(W_{j_1}^{-1}(w_{j_1}(\mathbf{x}_1), \theta_1)\right)$ for all directions $\theta_1 \in \mathbb{S}$.

We now modify the function $\mathbf{F}(\Gamma)$ by $\mathbf{F}_n(\Gamma)$, where n is the number of times which the ray starting at Γ hits the escape boundary. It is therefore obvious that $\mathbf{F}_1(\Gamma) = \mathbf{F}(\Gamma)$. In the case where the rays at the patch boundaries are continuous, we have,

$$\mathbf{F}_n(\Gamma) = \underbrace{\mathbf{F} \circ \mathbf{F} \cdots \circ \mathbf{F}}_{n \text{ times}}(\Gamma). \tag{19}$$

In general, the boundary function $\mathcal{L}_{jj'}$ must be applied in composition too. Analogously, we can define functions $\mathbf{\Phi}_n$ and $\mathbf{B}_n$.

For all directions $\theta_2 \in \mathbb{S}$ we then find $\mathbf{F}_n\left(W_{j_2}^{-1}(w_{j_2}(\mathbf{x}_2), \theta_2)\right)$. Since we do not know how many times the ray, which starts at $\mathbf{x}_2$ and passes through $\mathbf{x}_1$, hits $\mathcal{R}$, we need to find $\mathbf{F}_n$ for several values $n = 1, 2, \ldots$. See Figure 4 for three different cases where n is 1, 2 and 3.



(a) $n = 1$, $(j_1 = j_2 = j)$ (b) $n = 2$, $(j_1 = j,\ j_2 = j')$ (c) $n = 3$, $(j_1 = j_2 = j)$

Figure 4: Two neighboring patches M_j and $M_{j'}$. The ray (dashed curve) starting at $\mathbf{x}_2$ and passing through $\mathbf{x}_1$, hits the escape boundary $\mathcal{R}$ (thick curves) n times. Here, three different cases are shown where n is 1, 2 and 3.

We then find θ_1, θ_2 and n as the solutions to the algebraic equations

$$\mathbf{F}\left(W_{j_1}^{-1}(w_{j_1}(\mathbf{x}_1), \theta_1)\right) = \mathbf{F}_n\left(W_{j_2}^{-1}(w_{j_2}(\mathbf{x}_2), \theta_2)\right), \tag{20}$$

analogous to (14) in the single-patch case. There will be at most four systems of equations corresponding to four sides of patch M_{j_1}, for each value of n. The solutions to (20) can be computed by finding intersections of four sets of possibly crossing curves.

The length is then given by

$$\left|\mathbf{\Phi}\big(W_{j_1}^{-1}(\gamma_1)\big) - \mathbf{\Phi}_n\big(W_{j_2}^{-1}(\gamma_2)\big)\right|,$$

with $\gamma_1 = (w_{j_1}(\mathbf{x}_1), \theta_1)$ and $\gamma_2 = (w_{j_2}(\mathbf{x}_2), \theta_2)$.

4.3 Number and Shape of Patches and Parameterizations

One of the key design choices in such a multiple-patch scheme is the choice of patches and parameterizations. The important things are:

1. Patches should cover the physical domain with nonsingular parameterizations.
2. Parameterizations should have small coordinate distortions to make finite differencing accurate.
3. The right hand side $h(\gamma, f)$ in the escape PDEs should be well resolved by the patch discretization.

Remark 3. *Using overlapping patches, one can possibly reduce the number of patches. However, the objective in this work has not been optimizing the number of patches.*

4.4 Choosing Escape Boundary

Another key design choice is the choice of escape boundary. Two things are important about $\mathcal{R}$, and R:

1. The projection of each ray, which is of interest, onto M should cross $\mathcal{R}$ at some point. Otherwise $\mathbf{F}(\Gamma)$ is not well defined for all points. It is not obvious how to verify this rigorously. Having nonzero coefficients, $\mathbf{g}(\gamma) \neq 0$, everywhere is a necessary condition, but it is still possible to have rays that never reaches a given boundary, see e.g. [23].
2. If the compact manifold has a boundary, we can choose this as the escape boundary, similar to the single-patch manifold.

4.5 Limitations and Extra Problems

There are a couple of difficulties and problems:

1. In some cases, one cannot capture all rays of interest by only one choice of escape boundary. Different choices of escape boundary might be needed. A good implementation of the algorithm will then be the one which considers different combinations of patch boundaries as the escape boundary. Note that this is done in post-processing and does not require recomputation of the f_j solutions.
2. When a ray hits an inter-patch boundary, in order to find the escape solution at this point, we need to interpolate the discrete solutions computed on a fixed grid. The interpolation can be difficult if a ray is tangent to the inter-patch boundary. One possible way to overcome such a problem is to use overlapping patches. Another possibility is to choose another atlas of charts for the manifold.

5 Application to Creeping Ray Computations

Creeping rays are a type of diffracted rays which are generated at the *shadow line*[2] of the scatterer and propagate along geodesic paths on the scatterer surface. On a perfectly conducting convex body, they attenuate along their propagation path by tangentially shedding diffracted rays and losing energy. On a concave scatterer, they propagate on the surface and importantly, in the absence of dissipation, experience no attenuation.

The study of creeping rays is important in many high frequency problems, such as design of sophisticated and conformal antennas [19], antenna coupling problems [21], radar cross

[2]Shadow line or *horizon* is the locus of the points at which the incident rays are tangent to the scatterer surface.

section (RCS) computations [3, 19, 32, 26] and control of scattering properties of metallic structures coated with dielectric materials [28, 1, 22, 27].

In this section, we consider the application of the multiple-patch phase space method to computing creeping rays. Here, the computational domain is a scatterer surface which is a two-dimensional hypersurface embedded in $\mathbb{R}^3$. We split the surface into multiple patches represented by different parameterizations. The escape PDEs describing creeping rays are solved in each patch, individually. The creeping rays on the scatterer are then computed by connecting all individual solutions. The inter-patch boundaries are treated by the continuity of characteristics.

We first consider the case when the scatterer surface has a regular explicit parameterization and write the governing equations for computing creeping rays. We then discuss the multiple-patch scheme and give two numerical examples where the contribution of creeping rays to mono-static RCS is computed.

5.1 Governing Equations

We consider a scatterer surface with a regular explicit parameterization, represented by $\mathbf{x} = \bar{X}(\mathbf{u})$, where $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$, and the parameters $\mathbf{u} = (u, v)$ belong to a set $\Omega \subset \mathbb{R}^2$. Let the scatterer be illuminated by incident rays in a direction denoted by a normalized vector $\hat{I} = [\imath_1, \imath_2, \imath_3]$. We assume that the shadow line $\mathbf{u}_0(s)$ is represented by a curve in parameter space, with s being the arc length parameterization. The objective is to compute the geodesic paths on the scatterer surface together with the phase and amplitude of the wave field of creeping rays generated on the scatterer.

According to Keller and Lewis [17], the surface phase satisfies the *surface eikonal equation*,

$$|\tilde{\nabla}\phi| = n, \tag{21}$$

where $n(\mathbf{u})$ is the index of refraction at the surface, and $\tilde{\nabla}$ is the surface gradient, defined as

$$\tilde{\nabla}\phi := JG^{-1}\nabla\phi, \qquad G = J^\top J, \qquad J = [\bar{X}_u\, \bar{X}_v] \in \mathbb{R}^{3\times 2}.$$

We can write (21) as a Hamilton-Jacobi equation $H(\mathbf{u}, \nabla\phi) = 0$, with the Hamiltonian

$$H(\mathbf{u}, \mathbf{p}) \equiv \frac{1}{2}\,\mathbf{p}^\top\, G^{-1}(\mathbf{u})\,\mathbf{p} - \frac{n^2(\mathbf{u})}{2}. \tag{22}$$

Note that in the case n = constant, the rays associated with the surface eikonal equation (21) are geodesics, or shortest paths between two points on the surface. Henceforth, we will assume $n \equiv 1$.

We write (without derivation) the set of equations which are used in computing creeping rays and are obtained by reducing the order of the Hamiltonian system corresponding to (22) by one. For derivations see [26].

A geodesic on the surface is uniquely characterized by its location, (u, v), and direction, θ. Letting $\gamma := (u, v, \theta)$, the geodesics satisfy the system of ODEs (4) with

$$\mathbf{g}(\gamma) = \begin{pmatrix} \rho(\gamma)\,\cos\theta \\ \rho(\gamma)\,\sin\theta \\ \rho(\gamma)\,\mathcal{V}(\gamma) \end{pmatrix}. \tag{23}$$

The parameter τ is the arc length along the geodesic in the physical space, and

$$\rho = \rho(u, v, \theta) = \left|\bar{X}_u \cos\theta + \bar{X}_v \sin\theta\right|^{-1},$$

$$\begin{aligned}\mathcal{V}(\gamma) = &(\Gamma^1_{11}\cos^2\theta + 2\Gamma^1_{12}\cos\theta\sin\theta + \Gamma^1_{22}\sin^2\theta)\sin\theta - \\ &(\Gamma^2_{11}\cos^2\theta + 2\Gamma^2_{12}\cos\theta\sin\theta + \Gamma^2_{22}\sin^2\theta)\cos\theta,\end{aligned}$$

where $\Gamma^k_{ij}(\mathbf{u})$ are Christoffel symbols.

Moreover, we know that the phase ϕ is the length of the ray, given by (5) with $\beta = \phi$ and $\alpha \equiv 1$, and the amplitude a is computed by,

$$a(\tau) = a_0 \mathcal{Q}(s,\tau)^{\frac{-1}{2}} \exp\left(-\omega^{1/3}\beta(\tau)\right), \tag{24}$$

where a_0 is the amplitude at the starting point on the shadow line, $\mathcal{Q}(s,\tau)$ is the geometrical spreading at distance τ from the starting point, and $\beta(\tau)$ is a function representing the attenuation factor given by (5) with

$$\alpha(\gamma) = \frac{q_0}{\rho_g(\gamma)} \exp\left(i\frac{\pi}{6}\left(\frac{\rho_g(\gamma)}{2}\right)^{1/3}\right), \quad q_0 \approx 2.33811. \tag{25}$$

Here $\rho_g(\gamma)$ is the radius of curvature of the surface along the ray trajectory. We then let the escape function B be the difference between the β-values at the escape and starting points. All escape functions F, Φ and B satisfy equation (10), with the right hand side h being 0, 1 and α given by (25), respectively.

In order to compute the amplitude, in addition to β, we need also to compute geometrical spreading. We set $\tilde{\mathbf{u}}(s,\tau) := \mathbf{u}(\tau)$, with $\tilde{\mathbf{u}}(s,0) = \mathbf{u}_0(s)$ and let $\tilde{X}(s,\tau) := \bar{X}\big(\tilde{\mathbf{u}}(s,\tau)\big)$ be a point on the geodesic at the distance τ from the starting point $\tilde{X}_0(s) = \tilde{X}(s,0)$ on the shadow line. The geometrical spreading of the creeping ray at $\tilde{X}(s,\tau)$ in the physical space is given by, [26],

$$\mathcal{Q}(s,\tau) = \frac{\tilde{X}^{\perp}_{\tau} \cdot \tilde{X}_s}{\tilde{X}^{\perp}_{0\tau} \cdot \tilde{X}_{0s}}. \tag{26}$$

We consider a fixed shadow line $\gamma_0(s) = (u_0(s), v_0(s), \theta_0(s))$ and define $\tilde{\gamma}(s,\tau) := \gamma(\tau)$, where γ solves (4) with initial data $\gamma_0(s)$. Let $\mathbb{L}(\gamma_0) = \{\tilde{\gamma}(s,\tau) \ : \ \tau \geq 0\}$ be a sub-manifold of phase space $\mathbb{P}$ on which the creeping rays generated at $\gamma_0(s)$ lie. The Eulerian version of the geometrical spreading $Q : \mathbb{L}(\gamma_0) \to \mathbb{R}$, restricted to $\mathbb{L}(\gamma_0)$ and defined as $Q(\tilde{\gamma}(s,\tau)) := \mathcal{Q}(s,\tau)$ is then given by

$$Q(\tilde{\gamma}) = \frac{[\hat{T}(\tilde{\gamma}) \times \hat{N}(\tilde{u},\tilde{v})]^{\top} J(\tilde{u},\tilde{v}) z}{[\hat{I} \times \hat{N}(\mathbf{u_0}(s))]^{\top} \tilde{X}_{0s}(s)}, \quad \hat{T} = J\dot{\mathbf{u}}, \tag{27}$$

where $z = z(s,\tau)$ is a solution to

$$D_{\gamma} F(\tilde{\gamma}) z = \frac{d}{ds} F(\gamma_0(s)). \tag{28}$$

For $\tilde{\gamma}$ on the boundary, i.e. $\tilde{\gamma} = F(\gamma_0)$, the formula (27) can be simplified as,

$$Q(\tilde{\gamma}) = \frac{[\hat{T}(\tilde{\gamma}) \times \hat{N}(\tilde{u},\tilde{v})]^{\top} \hat{X}_s(s)}{[\hat{I} \times \hat{N}(\mathbf{u_0}(s))]^{\top} \tilde{X}_{0s}}, \tag{29}$$

where $\hat{X} : \mathbb{R} \to \mathbb{R}^3$ is defined by $\hat{X}(s) := X(U(\gamma_0(s)), V(\gamma_0(s)))$.

Note that $\hat{X}_s(s)$ in (29) and $D_\gamma F(\tilde{\gamma})$ and $F_s(\gamma_0(s))$ in (28) can be computed by numerically differentiating the solution to the PDEs in (10) with $f = F$, as was done in [26]. Instead, one can also directly compute $\tilde{X}_s$ in (26) by adding other ODEs to the geodesic system (4) as follows: First, we note that $\tilde{X}_s = J\tilde{\mathbf{u}}_s$. We then differentiate (4) with respect to s and derive the following ODE system

$$\dot{\tilde{\gamma}}_s = D_\gamma \mathbf{g}\, \tilde{\gamma}_s, \qquad \tilde{\gamma}_s(s,0) = \tilde{\gamma}_{0s}(s). \tag{30}$$

By solving this ODE, $\tilde{\mathbf{u}}_s$ and therefore $\tilde{X}_s$ can be computed. One can also write the escape PDE for (30) in the same way as before and post-process the phase space solution.

5.2 Multiple-Patch Scheme

We now split the scatterer surface M into several simpler surfaces with explicit regular parameterizations. As before, let M be given by an atlas of charts (M_j, w_j), where the patches $M_j \subset \mathbb{R}^3$ have the parametric equations $\mathbf{x} = \bar{X}_j(\mathbf{u}) : [0,1]^2 \to M_j$ and collectively cover M. Moreover, the mappings $w_j = \bar{X}_j^{-1} : M_j \to [0,1]^2$ are bijective.

Since on a geodesic $\hat{T}$ in (27) has unit length, we can consider the unit tangent bundle UTM of M as the global space. Note that UTM is a three-dimensional manifold embedded in $\mathbb{R}^6$. By Lemma 1, there is therefore a bijective mapping $W_j : UTM_j \to \Omega_p$ for each j, defined by $W_j(\Gamma) = \gamma$, with γ and Γ as in (8).

Knowing the bijective mappings w_j and W_j, and the solution to the escape PDEs in each patch, F_j, Φ_j and B_j, we can compute the multiple-patch escape functions $\mathbf{F}$, $\mathbf{\Phi}$ and $\mathbf{B}$ as described in Section 4.1.

5.3 Post-Processing

In order to compute phase and amplitude of a ray family, post-processing of the solutions to the escape PDEs (10) is needed.

For a given illumination direction, assume that the shadow line is known and given by $\Gamma_0(s)$ in the unit tangent bundle UTM. For each point $\mathbf{x} \in M_j$ covered by the surface wave, there is at least one creeping ray which starts at the shadow line and passes through that point. In order to find this ray, assuming the scatterer surface is boundaryless, we first choose the escape boundary $\mathcal{R}$ as the boundaries of M_j. Note that in the case of a surface with boundary, we choose its boundary as the escape boundary, and the post-processing will be similar to the single-patch case discussed in [26]. We then find $\mathbf{F}\left(W_j^{-1}(w_j(\mathbf{x}), \theta)\right)$ for all directions $\theta \in \mathbb{S}$. Moreover, for all points on the shadow line we find $\mathbf{F}_n\left(\Gamma_0(s)\right)$, defined by (19), with $n = 1, 2, \ldots$. We then find $s = s^*$, $\theta = \theta^*$ and $n = n^*$ as the solutions to the algebraic equations

$$\mathbf{F}\left(W_j^{-1}(w_j(\mathbf{x}), \theta)\right) = \mathbf{F}_n\left(\Gamma_0(s)\right), \tag{31}$$

analogous to (14) in the single-patch case. There will be at most four systems of equations corresponding to four sides of patch M_j, for each value of n. The solutions to (31) can be computed by finding intersections of four sets of possibly crossing curves.

Now we can use (28) to compute z with $\gamma_0 = W_{j_0}(\Gamma_0(s^*))$ and $\tilde{\gamma} = (w_j(\mathbf{x}), \theta^*)$ where $j_0 = \mathcal{J}(\Gamma_0(s^*))$. Note that $F(\tilde{\gamma})$ and $F(\gamma_0)$ in the left and right hand sides of (28) are

replaced by $W_j\big(\mathbf{F}(W_j^{-1}(\tilde{\gamma}))\big)$ and $W_j\big(\mathbf{F}_{n^*}(\Gamma_0(s^*))\big)$, respectively. The geometrical spreading $Q(\tilde{\gamma})$ at point $\mathbf{x}$ will be therefore computed by (27), and phase and amplitude are given by

$$\phi(w_j(\mathbf{x})) = \phi_0 + \mathbf{\Phi}\big(W_{j_0}^{-1}(\gamma_0)\big) - \mathbf{\Phi}\big(W_j^{-1}(\tilde{\gamma})\big),$$

$$A(\tilde{\gamma}) = A_0\, Q(\tilde{\gamma})^{\frac{-1}{2}} \exp\bigg(-\omega^{\frac{1}{3}}\Big(\mathbf{B}\big(W_{j_0}^{-1}(\gamma_0)\big) - \mathbf{B}\big(W_j^{-1}(\tilde{\gamma})\big)\Big)\bigg),$$

where ϕ_0 and A_0 are the phase and amplitude at the point γ_0, respectively.

5.4 Example 1 - A Scalene Ellipsoid

We consider the scatterer surfaces to be a scalene ellipsoid (an ellipsoid with different semi-axes) and apply the multiple-patch phase space method to compute the contribution of *backscattered* creeping rays to mono-static RCS, i.e., the rays that propagate on the surface of the scatterer and return in the opposite direction of incident waves. We assume that the incoming amplitudes are one at attachment points on the shadow line and compute the backscattered amplitude at detachment points on the shadow line. We also compute the length of the backscattered rays.

We consider an ellipsoid given by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

with $a = 2$, $b = 1$ and $c = 0.5$. Since there is no single non-singular parameterization for the ellipsoid, we split it into six patches with non-singular parameterizations (see Figure 5) and solve for $f(\gamma)$ in each patch, as described in Section 3.2.

In order to find the backscattered creeping ray by post-processing, we first choose the escape boundary consisting of six sides, as highlighted in Figure 6. We then continue as follows,

0. Given a pair of incident angles $(\Psi_1, \Psi_2) \in [0, 90]^2$, find the incident direction $\hat{I} = [\sin\Psi_1 \cos\Psi_2,\ \cos\Psi_1 \cos\Psi_2,\ \sin\Psi_2]$.

1. Find the shadow line $\gamma_0(s) = (\mathbf{u}_0(s), \theta_0(s))$ in the phase space Ω_p using the relations $\hat{N}^\top \hat{I} = 0$ and $\hat{T}(\gamma_0(s)) = \hat{I}$ in patch $j(s)$. Let the parameterization of the shadow line be discretized in N grid points $\{s_n\}$ with $n = 1, \ldots, N$.

2. For each point on the shadow line find $\mathbf{F}\left(W_{j(s_n)}^{-1}(\gamma_0(s_n))\right)$ as discussed in Section 4.1.

3. A backscattered ray starting at attachment point s_a and ending at detachment point s_d on the shadow line should satisfy

$$\mathbf{F}\left(W_{j_a}^{-1}(\gamma_0(s_a))\right) = \mathbf{F}\left(W_{j_d}^{-1}(\gamma_0(s_d))\right) + C,$$

where $j_a = j(s_a)$ and $j_d = j(s_d)$, and C is a constant accounting for the fact that the directions of creeping rays starting at s_a and s_d differ by a π on the escape boundary. The right and left hand sides of this equation can be represented as six sets of curves in $\mathbb{R}^2$ parameterized by s, corresponding to six sides of the escape boundary. To find the backscattered ray we need to find crossing points of these curves, as is done in the single patch case.

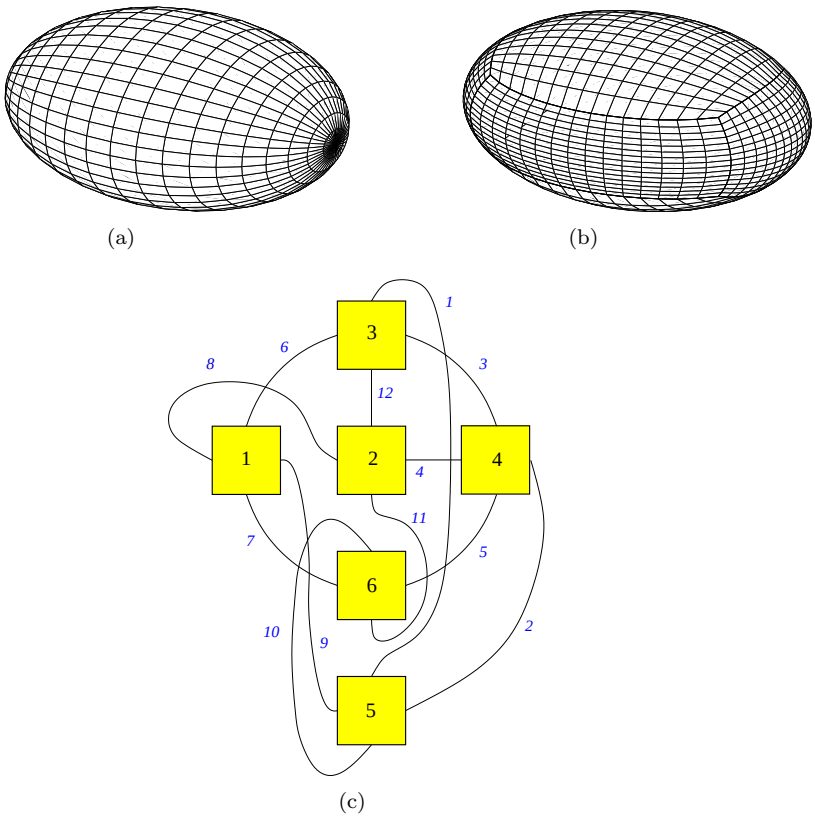


Figure 5: Upper left figure shows an ellipsoid with a single patch parameterization which is singular at two poles. Upper right figure shows the ellipsoid divided into 6 patches. Note that the singularities have been removed using non-singular multiple parameterizations. Lower figure shows the structure of patches and patch boundaries in parameter space. Patches $j = 1, \ldots, 6$ correspond to left, front, up, right, back and down patches, respectively. These 6 patches share 12 sides in total, shown with italic numbers.

4. For each crossing point, there is a pair of backscattered rays (two backscattered rays lying on top of each other); one starting at point s_a and ending at point s_d, the other starting at point s_d and ending at point s_a. Although these two rays have the same lengths, they do not have the same geometrical spreading and therefore not the same amplitude. Compute two geometrical spreadings as described in Section 5.3 with $\gamma_0 = \gamma(s_d)$ and $\tilde{\gamma} = \gamma(s_a)$ for the first backscattered ray and $\gamma_0 = \gamma(s_a)$ and $\tilde{\gamma} = \gamma(s_d)$ for the second one.

5. The length and amplitudes are then computed as,

$$\phi = \mathbf{\Phi}(\Gamma_{s_a}) + \mathbf{\Phi}(\Gamma_{s_d}), \quad \Gamma_{s_a} = W_{j_a}^{-1}(\gamma(s_a)), \quad \Gamma_{s_d} = W_{j_d}^{-1}(\gamma(s_d)),$$

$$A_1 = Q(\gamma(s_a))^{\frac{-1}{2}} \exp\left(-\omega^{\frac{1}{3}}\left(\mathbf{B}(\Gamma_{s_a}) + \mathbf{B}(\Gamma_{s_d})\right)\right),$$

$$A_2 = Q(\gamma(s_d))^{\frac{-1}{2}} \exp\left(-\omega^{\frac{1}{3}}\left(\mathbf{B}(\Gamma_{s_a}) + \mathbf{B}(\Gamma_{s_d})\right)\right).$$

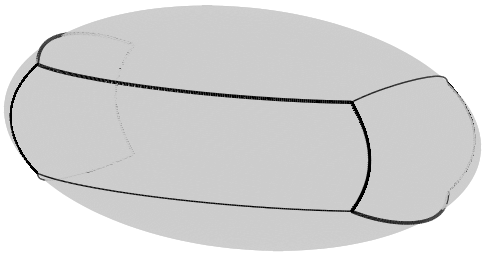


Figure 6: The ellipsoid with its patch boundaries. Thick lines show the escape boundary.

Figure 7 shows the backscattered rays for two different incident angles. There are three pairs of backscattered rays which can be detected by the algorithm. Every two rays of each pair lie on top of each other.

Figure 7: Left figure shows the backscattered creeping rays (thick curves) for $\Psi_1 = 30$ and $\Psi_2 = 0$. Right figure shows the backscattered creeping rays for $\Psi_1 = 30$ and $\Psi_2 = 10$. Thin curves represent the shadow lines.

We notice that in [26], because of using a single patch and excising the singularity at two poles, only the shortest backscattered ray could be captured. Figure 8 shows the length and amplitudes of the shortest backscattered ray for different incident angles, with $\omega = 1$. The peaks in the amplitude correspond to *caustic backscattered creeping rays* which have infinite amplitudes. Such rays are particularly important in near-field RCS computations. However, in far-field RCS, due to the the geometrical spreading outside the scatterer, their contribution may not be as important.

Figure 9 shows the convergence of length and amplitudes of the backscattered creeping ray for a fixed vertical angle $\Psi_2 = 70$ and different horizontal incident angles $\Psi_1 \in [-90, 90]$. We use a second order Fast Marching algorithm on a coarse grid of the size 50^3 and a fine grid of the size 100^3. We compare them with a reference solution obtained by a high order ray tracing method. The rate of convergence confirms the second order accuracy of the algorithm. We note that comparing to the results in [26], where a first order algorithm was used, the accuracy of amplitude has been improved dramatically. It shows that using a first order accurate method for computing the phase and amplitude results in a worse relative

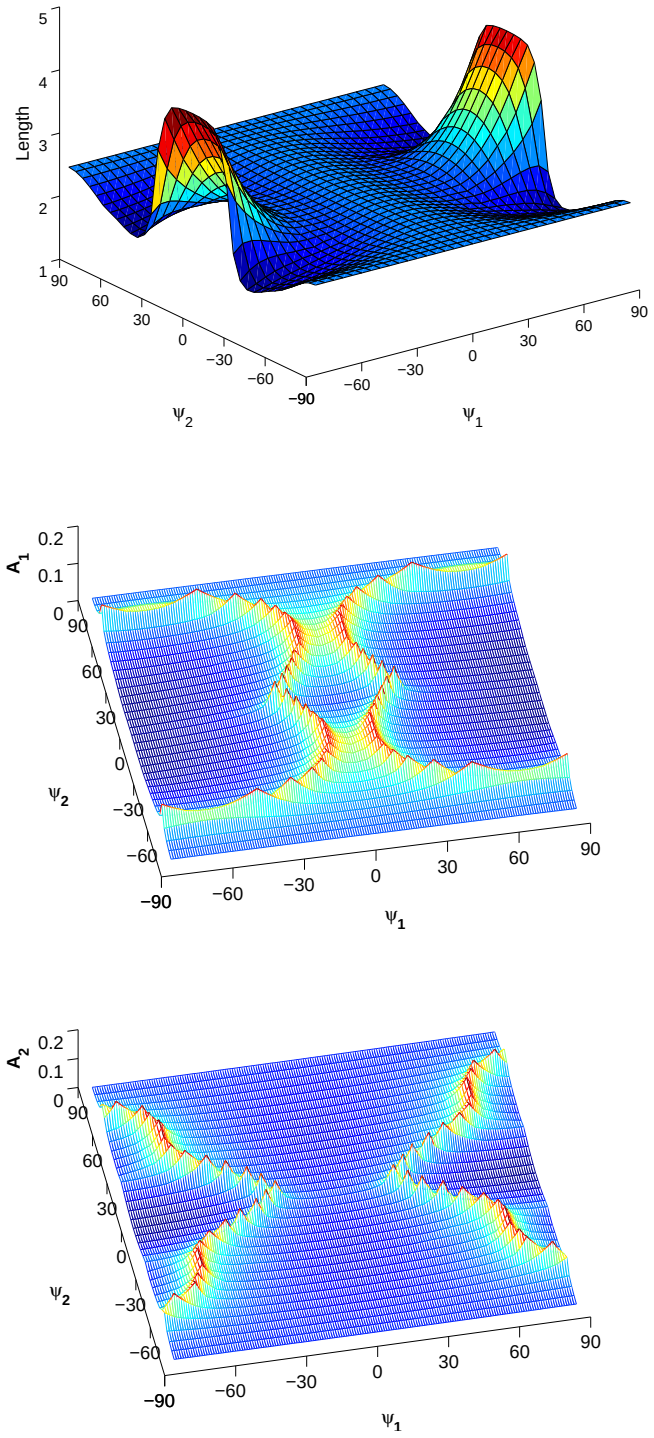


Figure 8: Length and amplitudes (with $\omega = 1$) of backscattered creeping rays for many illumination angles.

error for the amplitude than for the phase. Therefore, higher order algorithms are required to obtain low relative errors for the amplitude, as observed also in [30].

The complexity of using the fast phase space method proposed here consists of two parts. First, the cost of solving the PDEs by the Fast Marching method is $\mathcal{O}(N^3 \log N)$. Second, the cost of finding the backscattered rays for each shadow line is $\mathcal{O}(N)$. For all N^2 shadow lines, it is $\mathcal{O}(N^3)$. Therefore the total complexity will be $\mathcal{O}(N^3 \log N)$. The total cost by using other methods, like wave front tracking and solvers based on the surface eikonal equation, will be $\mathcal{O}(N^4)$, if the cost for each shadow line is $\mathcal{O}(N^2)$. In this case, using the phase space method will then be much faster.

Remark 4. *A graph structure can be useful for a general computer implementation. The topology of the surface can be described by a graph, in which each patch is a node and the edges go between connected patches. Figure 10a shows the graph corresponding to the ellipsoid divided into six patches which are connected through twelve sides (see Figure 5). The graph therefore has six nodes and twelve edges.*

We can also introduce another topology graph, in which the nodes are the sides of the patches and the edges correspond to the patches themselves. Each node (side) is therefore connected to six other nodes through two patches which are connected by that side. See Figure 10b. This structure can be useful for imposing inter-patch boundary conditions and computing $\mathbf{F}$.

5.5 Example 2 - A Balloon

We consider a balloon-shape surface consisting of a hemisphere in the positive side of the z-axis, centered at the origin and with radius r, and the surface created by rotating the part of parabola $z^2 = 2r(r-y)$ over the interval $-\sqrt{2}r \leq z \leq 0$ about the z-axis. This is a simple smooth version of the cone-hemisphere studied in [3] as a model for low-observable objects where creeping rays are important for RCS. We divide this surface into six patches, as shown in Figure 11; The hemisphere is split into five patches $j = 1, \ldots, 5$, and the parabolic part is represented by one patch $j = 6$. We excise the singularity at the vertex of the balloon by cutting it off. The lower boundary of patch $j = 6$ will therefore be an excision boundary and is not considered as a patch boundary. We also partition the upper boundary of patch $j = 6$ into four boundaries connecting to lower boundaries of patches $j = 1, \ldots, 4$. Note that the left and right boundaries of patch $j = 6$ are in fact the same. Therefore, there are in total thirteen sides connecting six patches. See Figure 11.

Since the surface is symmetric about the z-axis, we consider a fixed horizontal incident angle $\Psi_1 = 90$, and due to symmetry about the yz-plane, we consider the vertical angles $\Psi_2 \in [-90, 90]$. Figure 12 shows the backscattered rays for two different incident angles $\Psi_2 = 40$ and $\Psi_2 = -40$. For positive vertical incident angles, there are four pairs of backscattered rays which can be detected by the algorithm. Two of them are symmetric and have the same length and amplitudes. For negative vertical incident angles, only one backscattered ray can be captured. We notice that in the case $\Psi_2 = 90$, there will be infinitely many backscattered rays which results in high observability of the object in this incident direction. On the other hand, for $\Psi_2 = -90$, there will be no backscattered ray because we have excised the vertex. In fact even if we did not excise it, all creeping rays would go to the vertex and diffract in different directions.

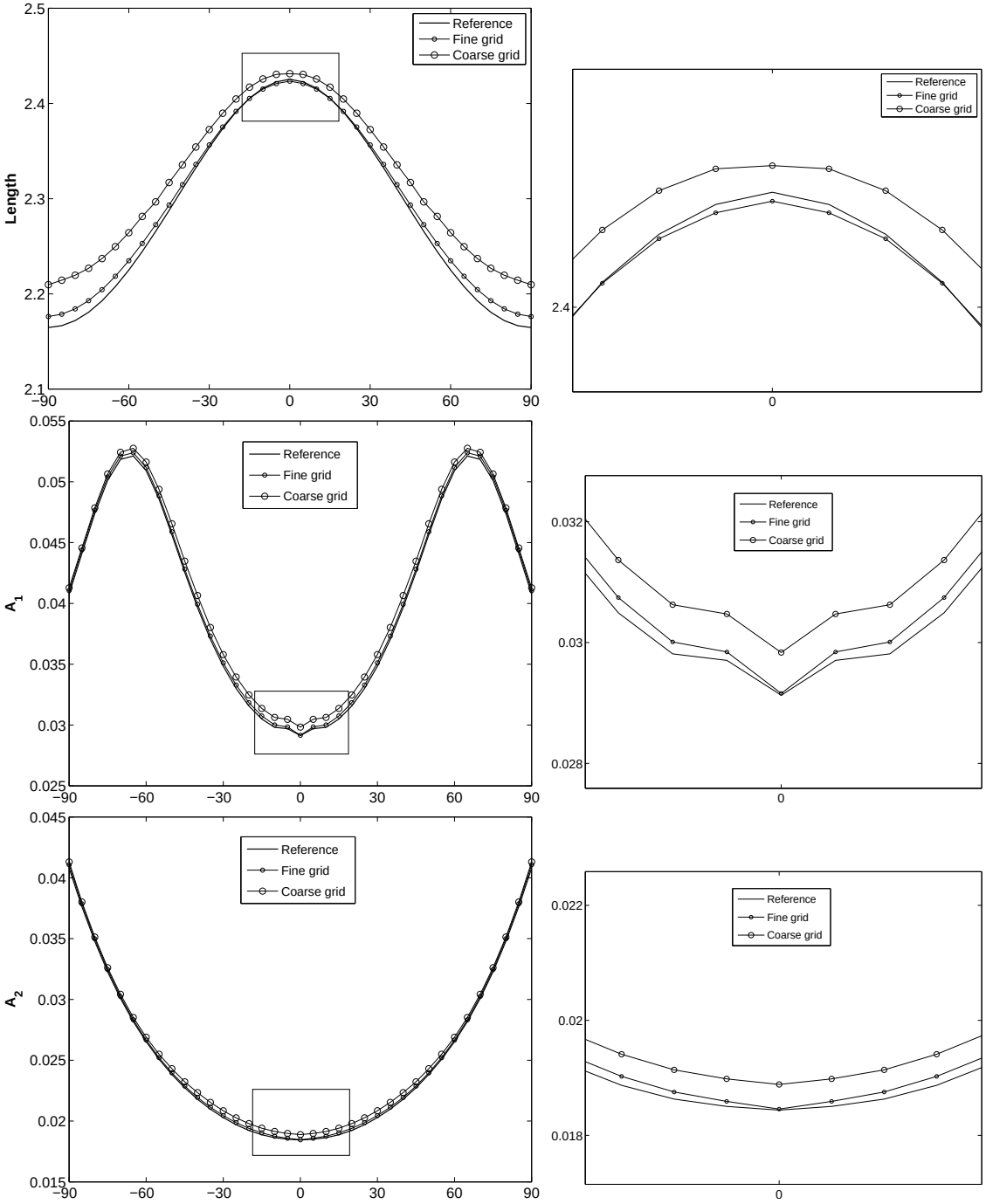


Figure 9: Length and amplitude (with $\omega = 1$) of the backscattered creeping rays for different horizontal incident angles and a fixed vertical angle $\Psi_2 = 70$. By refining the grid, solutions of the second order phase space algorithm converge to a reference solution obtained by a high order ray tracing method with a correct rate. Right figures show zoomed views of left figures.

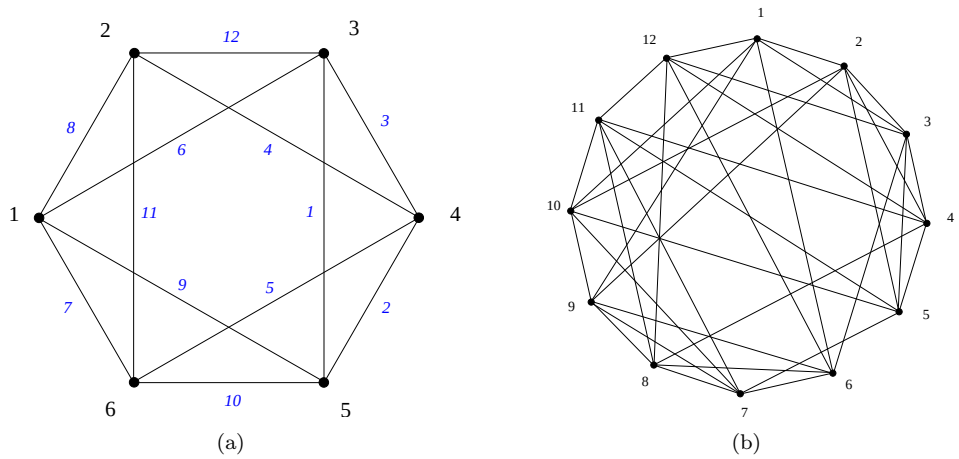


Figure 10: Representation of an ellipsoid divided into 6 patches by two different graph structures. Left figure shows the graph with 6 nodes and 12 edges. Here, the nodes 1 to 6 denotes the left, front, up, right, back and down patches, respectively. Right figure shows the graph with 12 nodes and 72 edges.

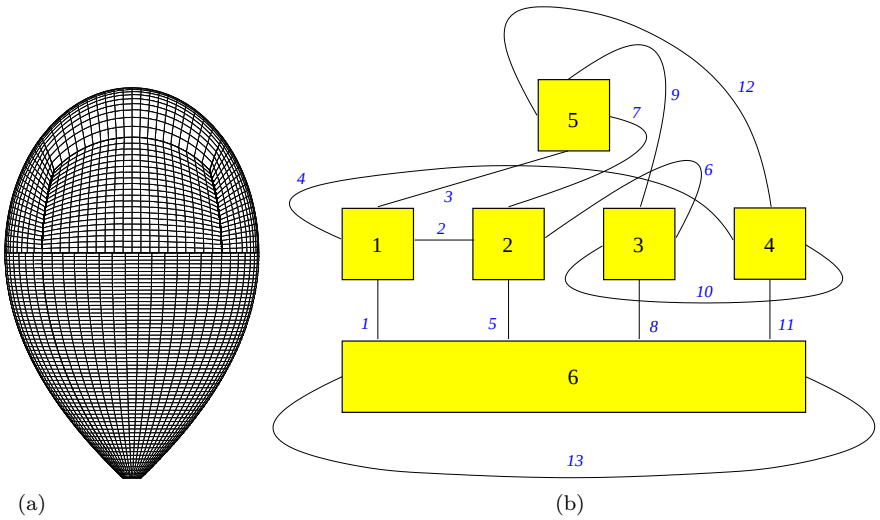


Figure 11: Left figure shows the balloon divided into 6 patches. Right figure shows the structure of patches and patch boundaries in parameter space. Patches $j = 1, \ldots, 6$ correspond to front, right, back, left, up and down patches, respectively. These 6 patches share 13 sides in total, shown with italic numbers.

Figure 13 shows the length and amplitude of backscattered rays in a polar coordinate system for all incident directions $\Psi \in [0, 360]$. The angles $\Psi \in [0, 90]$ in the polar system correspond to $\Psi_2 \in [0, -90]$, and the angles $\Psi \in [270, 360]$ correspond to $\Psi_2 \in [90, 0]$. The values for $\Psi \in [90, 270]$ are then calculated using the symmetry of the surface about the yz-plane.

Figure 12: Backscattered creeping rays (thick curves) for $\Psi_2 = 40$ (left figure) and $\Psi_2 = -40$ (right figure). Thin curves represent the shadow lines.

6 Application to Seismic Wave Computations

The inhomogeneity of earth causes deflection and reflection of seismic waves. The numerical study of seismic wave propagations, therefore, helps us to learn about the inhomogeneous structure of earth, which is important in direct and inverse problems of seismology and seismic exploration of oil.

In this section, we apply the multiple-patch phase space method to compute the travel-time of seismic rays. We consider a two-dimensional multi-layered medium whose different layers have different wave speeds. We split the medium into multiple patches corresponding to different layers. The escape PDEs describing seismic waves are solved in each patch, individually. The travel-time of the waves in the medium are then computed by connecting all individual solutions. The inter-patch boundaries are treated by Snell's law and the law of reflection.

We first consider the case when the medium has a regular explicit parameterization and derive the governing equations. We then discuss the multiple-patch scheme and give a numerical example for computing the travel-times.

6.1 Governing Equations

Consider a two-dimensional medium M represented by parametric equations $\mathbf{x} = \bar{X}(\mathbf{u})$, where $\mathbf{x} = (x, y) \in M \subset \mathbb{R}^2$ and $\mathbf{u} = (u, v) \in \Omega \subset \mathbb{R}^2$.

The phase ϕ of the wave satisfies the eikonal equation,

$$|\nabla \phi| = n(\mathbf{x}), \tag{32}$$

which is a Hamilton-Jacobi equation. The Hamiltonian for the eikonal equation can be written in the form

$$H(\mathbf{x}, \mathbf{p}) = c(\mathbf{x})|\mathbf{p}| \equiv 1, \tag{33}$$

where $c(\mathbf{x}) = 1/n(\mathbf{x})$ is the wave speed and $\mathbf{p} = \nabla \phi$. Introducing the arc length parameter

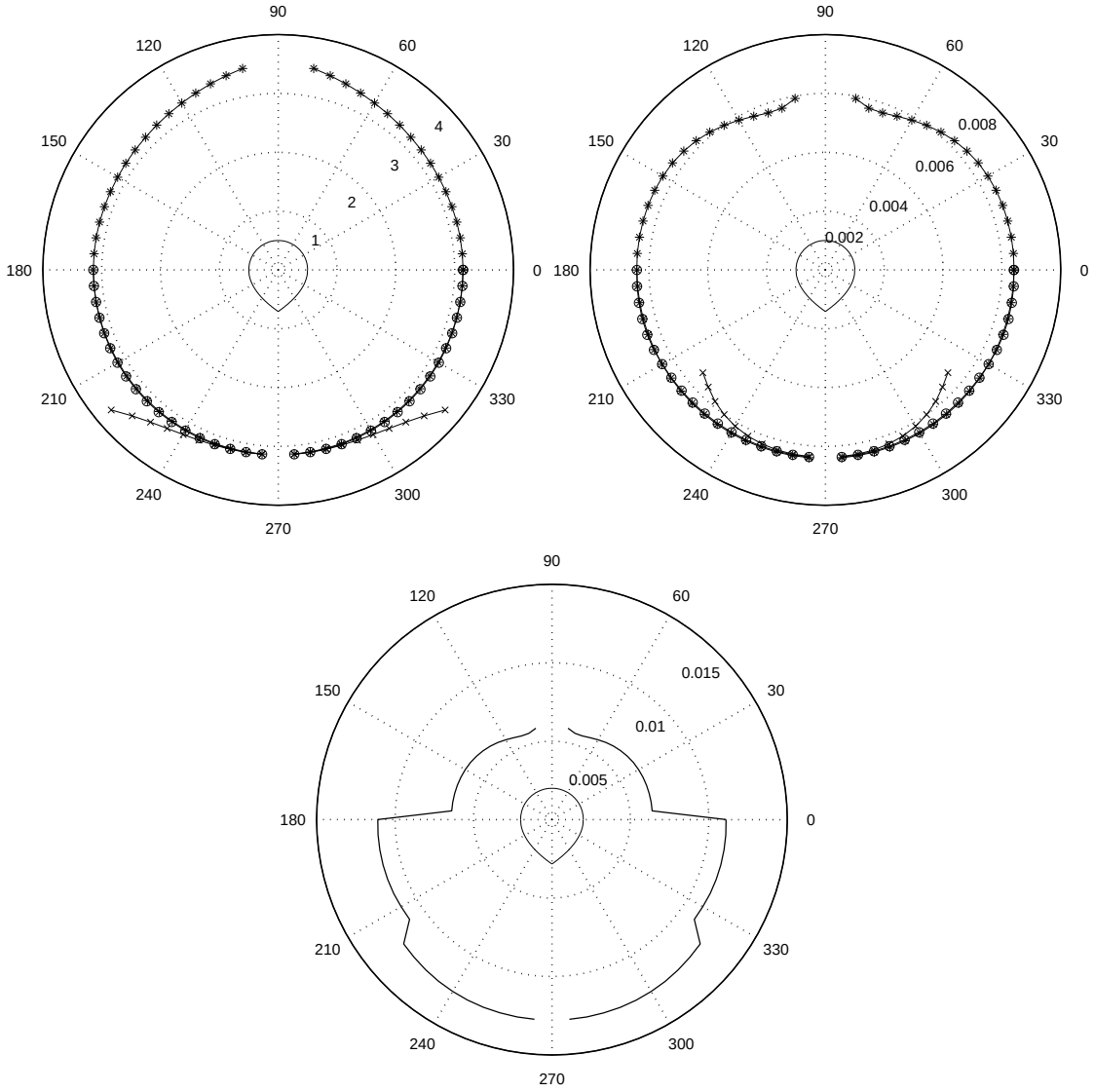


Figure 13: Length and amplitude (with $\omega = 1$) of the backscattered creeping rays for all illumination directions $\Psi \in [0, 360]$. Upper left and right figures show the length and amplitude of the backscattered rays, respectively. There are four pairs of rays among which two (illustrated by $\circ$) are symmetric. Note that at $\Psi = 90$ ($\Psi_2 = -90$), there will be no backscattered ray because all creeping rays go to the vertex and diffract in different directions. At $\Psi = 270$ ($\Psi_2 = 90$), however, there are infinitely many backscattered rays resulting in high observability of the object in this incident direction, and therefore the values are not shown. Because of the excision, the longest backscattered ray (illustrated by $\times$) can be captured only for $\Psi \in [220, 320]$ ($\Psi_2 \geq 40$). Bottom figure shows the total amplitude, $A_{tot} = \sqrt{A_1^2 + A_2^2 + A_3^2 + A_4^2}$, of all four backscattered creeping rays.

τ, a ray trajectory $(\mathbf{u}(\tau), \mathbf{p}(\tau))$ in $\Omega \times \mathbb{R}^2$ is given by the Hamiltonian system

$$\dot{\mathbf{x}} = c(\mathbf{x}) \frac{\mathbf{p}}{|\mathbf{p}|} = c^2(\mathbf{x})\, \mathbf{p}, \tag{34a}$$

$$\dot{\mathbf{p}} = -|\mathbf{p}|\, \nabla_{\mathbf{x}} c(\mathbf{x}) = -\frac{\nabla_{\mathbf{x}} c(\mathbf{x})}{c(\mathbf{x})}, \tag{34b}$$

where the dot denotes differentiation with respect τ.

Since $\dot{\mathbf{x}} = J(\mathbf{u})\,\dot{\mathbf{u}}$ with the Jacobian $J = [\bar{X}_u\, \bar{X}_v] \in \mathbb{R}^{2\times 2}$, we have

$$\dot{\mathbf{u}} = J^{-1}(\mathbf{u})\,\dot{\mathbf{x}} = c^2(\bar{X}(\mathbf{u}))\, J^{-1}(\mathbf{u})\,\mathbf{p}. \tag{35}$$

Moreover, inspired by $|\mathbf{p}| = \frac{1}{c(\mathbf{x})}$, we set $\mathbf{p} = (p_1, p_2)^\top = \frac{1}{c(\mathbf{x})}\,(\cos\theta, \sin\theta)^\top$. Differentiating $\mathbf{p}$ with respect to τ, we get

$$\dot{\mathbf{p}} = \begin{pmatrix} \nabla_{\mathbf{x}} \frac{1}{c(\mathbf{x})} \cdot \dot{\mathbf{x}}\, \cos\theta - \frac{1}{c(\mathbf{x})} \sin\theta\, \dot{\theta} \\ \nabla_{\mathbf{x}} \frac{1}{c(\mathbf{x})} \cdot \dot{\mathbf{x}}\, \sin\theta + \frac{1}{c(\mathbf{x})} \cos\theta\, \dot{\theta} \end{pmatrix}. \tag{36}$$

By (34) and (36), we get $\dot{\theta} = c_x(\mathbf{x}) \sin\theta - c_y(\mathbf{x}) \cos\theta$. Therefore, setting $\gamma := (u, v, \theta) \in \Omega_p$, the function $\mathbf{g}(\gamma)$ in (4) will be

$$\mathbf{g}(\gamma) = \begin{pmatrix} c(\bar{X}(\mathbf{u}))\,(g^{11}\,\cos\theta + g^{12}\,\sin\theta) \\ c(\bar{X}(\mathbf{u}))\,(g^{21}\,\cos\theta + g^{22}\,\sin\theta) \\ c_x(\bar{X}(\mathbf{u}))\sin\theta - c_y(\bar{X}(\mathbf{u}))\cos\theta \end{pmatrix}, \tag{37}$$

where $(g^{ij}) = J^{-1}(\mathbf{u})$. Note that since $\dot{\mathbf{x}} \parallel \mathbf{p}$ by (34), the angle θ represents the direction of the ray trajectory at $\mathbf{x}$ in the physical space. Moreover, with our choice of Hamiltonian,

$$\dot{\phi}(\mathbf{x}(\tau)) = \nabla\phi(\mathbf{x}(\tau)) \cdot \dot{\mathbf{x}}(\tau) = \mathbf{p} \cdot \mathbf{p}\frac{c(\mathbf{x}(\tau))}{|\mathbf{p}|} = |\mathbf{p}|\, c(\mathbf{x}(\tau)) = 1,$$

implying that ϕ corresponds to travel-time.

6.2 Multiple-Patch Scheme

We assume that the physical domain, representing a medium, is a two-dimensional compact manifold $M \subset \mathbb{R}^2$ with boundary. Since the wave speed distribution in a multi-layered inhomogeneous medium is not continuous, it is natural to split the medium to different patches with continuous wave speed distributions. We now let M be described by an atlas of charts (M_j, w_j) as before. The three-dimensional unit tangent bundle UTM is embedded in $\mathbb{R}^4$. In this case, there is an easier way to represent UTM by simplifying the mapping $W_j : UTM_j \to \Omega_p$ to be $W_j(\Gamma) = \gamma$, where

$$\Gamma = (\mathbf{x}, \mathbf{q}), \quad \mathbf{q} = \hat{s}(\theta) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \quad \gamma = (w_j(\mathbf{x}), \theta). \tag{38}$$

In the same way as before, we can define and compute multiple-patch escape functions $\mathbf{F}(\Gamma)$ and $\mathbf{\Phi}(\Gamma)$. However, here the rays are not continuous at the patch-boundaries due to the change of the wave speed at these points. When a ray passes the boundary between two layers (two neighboring patches) with different wave speeds, part of the ray is reflected (by the law of reflection), and part of it is refracted or transmitted into the second layer (by Snell's law of refraction). At each interface, therefore, the ray field splits into two new ray families, one reflected and one transmitted.

Figure 14a shows the reflection and refraction of a ray at the interface between two media of different wave speeds, with $c_L > c_R$. The law of reflection gives the relation between the angles of incidence (θ_{inc}) and of reflection (θ_{ref}) as

$$\theta_{inc} = \theta_{ref}. \tag{39}$$

The relation between the angles of incidence and of refraction (θ_{tr}) for a ray crossing a boundary between different media is given by Snell's law

$$\frac{\sin\theta_{inc}}{\sin\theta_{tr}} = \frac{c_L}{c_R}. \tag{40}$$

When a ray moves from a dense to a less dense medium ($c_L < c_R$), Snell's law cannot be used to calculate the refracted angle if $\sin\theta_{tr} = \sin\theta_{inc}\,(c_R/c_L) > 1$. At this point, the ray is reflected in the incident medium, known as *internal reflection*. There is therefore a critical angle (θ_{cr}) for which the ray travels directly along the surface between the two refractive media. The critical angle is found by Snell's law, putting in a transmitted angle of 90 degrees. This gives:

$$\theta_{cr} = \arcsin\frac{c_R}{c_L}. \tag{41}$$

For any angle of incidence larger than the critical angle ($\theta_{inc} > \theta_{cr}$), the ray is totally reflected off the interface, obeying the law of reflection. This phenomena is called *total internal reflection*. See Figure 14b.

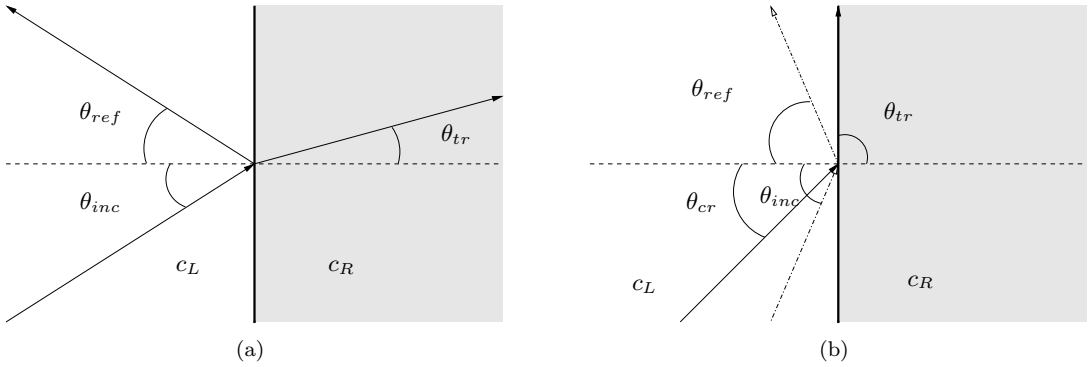


Figure 14: Reflection and refraction of a ray at the interface between two media of different wave speeds. Left figure shows the reflection and refraction when $c_L > c_R$. Right figure shows the internal reflection when $\theta_{inc} \geq \theta_{cr}$.

From (39)-(41), we can easily find the inter-patch boundary functions $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{R}}$ discussed in Section 2.2.1. Post-processing in this case is similar to that of the single-patch case, because the escape boundary that we chose coincides with the external boundary of the medium.

Note that the inter-patch boundary conditions above can be seen as a way to preserve the Hamiltonian (33) for a ray across the patch boundary. In cases where the discontinuity in $c(\mathbf{x})$ is not aligned with the patch boundary, the solution of the escape equations is not unique. Uniqueness can however be recovered by enforcing the extra condition that solutions should be continuous along constant Hamiltonian paths also inside the patches. This is the idea of so called Hamiltonian-preserving methods developed in [14, 15]. These methods capture the effect of a discontinuous $c(\mathbf{x})$ on uniform grids not aligned with the discontinuity.

6.3 Example 3 - A Multi-Layered Medium

We consider a multi-layered medium $M = [0, 6]^2$ consisting of three layers with different wave speeds (see Figure 15):

- Top layer: $c_1(x,y) = 1 + 0.05\,(x-3)^2 + 0.25\,y,$
- Middle layer: $c_2(x,y) = \frac{3}{1+e^{-((x-3)^2+(y-3)^2)}},$
- Bottom layer: $c_3(x,y) = 0.5 + 0.2\,x + 0.5\,y.$

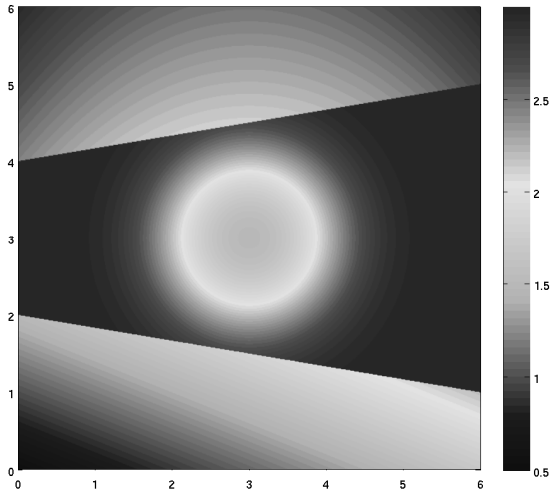


Figure 15: The medium consisting of three layers and grey scale plot of the wave speed field.

We want to compute multivalued travel-time of seismic rays in the medium from a given source point $\mathbf{x}_0$ on the boundary. We split the medium into three patches corresponding to the three layers, as shown in Figure 15. The escape equations for the escape point F and the travel-time Φ are derived and solved in each patch.

In order to find the travel-time with a given source point by post-processing, we first choose the four outermost boundaries of the entire physical domain as the escape boundary. We then continue as follows:

0. The source point $\mathbf{x}_0$ on the boundary is first reduced to a point $S_0 \in \mathbb{R}$.
1. For each point $\mathbf{x} \in M$, find $\mathbf{F}(\Gamma) = (\mathbf{U}, \mathbf{V}, \boldsymbol{\Theta})$ with $\Gamma = (\mathbf{x}, \mathbf{q}(\theta))$ for all $\theta \in \mathbb{S}$. Now $(\mathbf{U}, \mathbf{V})$ can again be reduced to points $S \in \mathbb{R}$, parameterized by θ.
2. Find $\theta = \theta^*$ such that $S_0 = S(\theta)$.
3. Travel-time at $\mathbf{x} \in M$ will then be $\boldsymbol{\Phi}(\Gamma^*)$ with $\Gamma^* = (\mathbf{x}, \mathbf{q}(\theta^*))$.

Figure 16 shows the distribution of transmitted seismic rays and equiarrival curves, i.e., the locus of all points in physical domain which have the same travel-time, from two different source points, $\mathbf{x}_0 = (3, 6)$ and $\mathbf{x}_0 = (3.5, 6)$. Note that we can track both reflected and transmitted ray families, but not at the same time. In order to get all rays, one needs to follow all ray families. Figure 17 shows the equiarrival curves of rays reflected from the top and bottom interfaces inside the top and the middle layers, respectively, for a source point at $\mathbf{x}_0 = (3, 6)$. If we repeat this procedure, we can also capture multiple rays reflected from the two interfaces that get trapped inside the middle layer and reverberate to infinity. Here, we do not consider reflections from the domain boundaries, as we have a truncated domain much smaller than the physical space in which the waves propagate.

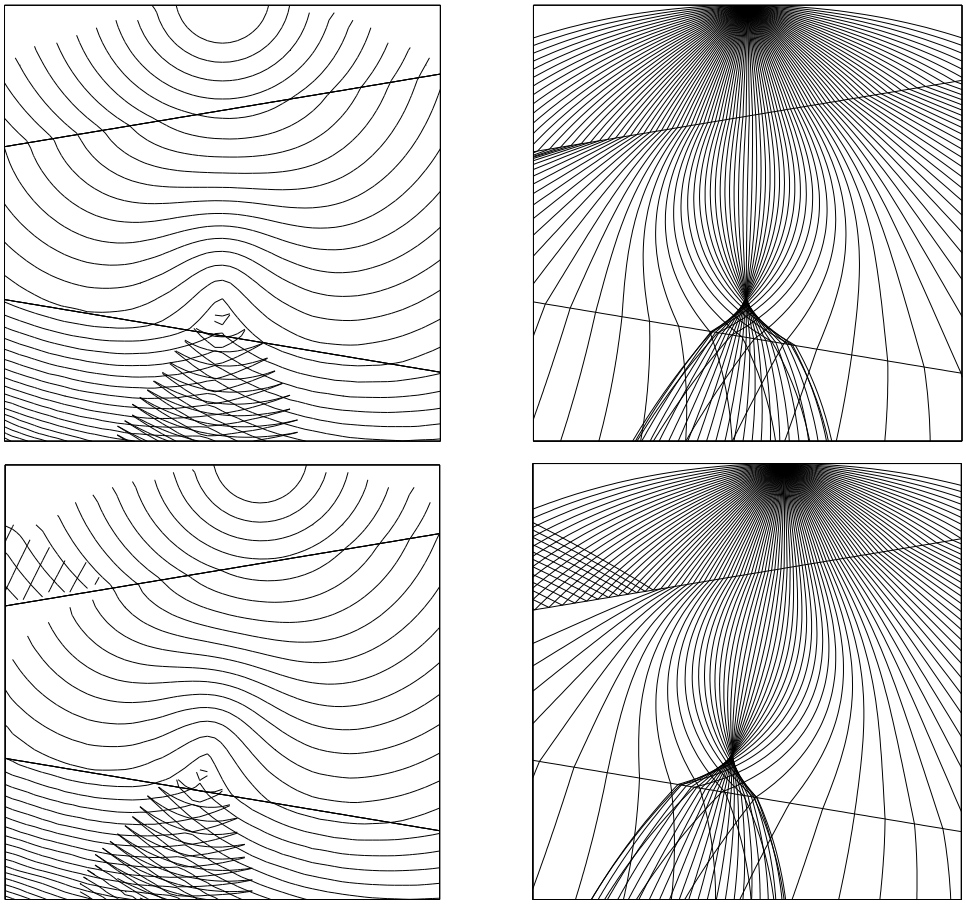


Figure 16: The equiarrival curves and the distribution of seismic rays for two different source points.

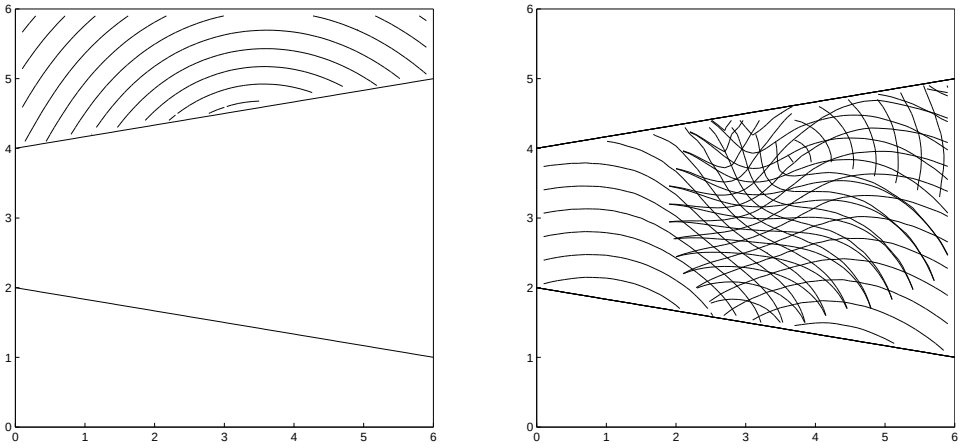


Figure 17: The equiarrival curves of reflected seismic rays from the top (left figure) and bottom (right figure) interfaces for a source point on the center of the top of the domain.

7 Conclusion

We have modified the single-patch phase space method for computing creeping rays to a multiple-patch method for computing trajectories on two-dimensional manifolds possibly

embedded in a higher-dimensional space. The dynamics of trajectories are given by systems of first-order ODEs in a phase space. We split the manifold into multiple patches where each patch has a well-defined regular parameterization. The *escape* equations, which are hyperbolic PDEs in a three-dimensional phase space, are derived and solved in each patch, individually, using a second-order version of the fast marching method. The solutions of individual patches are then connected using suitable inter-patch boundary conditions. Properties for particular families of trajectories are obtained through a fast post-processing. For some applications, the complexity of the method is attractive. Such applications include mono-static and bi-static RCS computations, antenna coupling problems, and travel-time computations of seismic waves when the solution is sought for many different sources.

References

[1] N. C. Albertsen. Creeping wave modes for a dielectric coated cylinder. *IEEE T. Antenn. Propag.*, 37(12):1642–1644, 1989.

[2] C. Belta and V. Kumar. Optimal motion generation for groups of robots: A geometric approach. *J. Mech. Des.*, 126:36–70, 2004.

[3] D. P. Bouche, J.-J. Bouquet, H. Manenc, and R. Mittra. Asymptotic computation of the RCS of low observable axisymmetric objects at high frequency. *IEEE T. Antenn. Propag.*, 40(10):1165–1174, 1992.

[4] H. Brandén and S. Holmgren. Convergence acceleration for hyperbolic systems using semicirculant approximations. *J. Sci. Comput.*, 14(4):357–393, 1999.

[5] T. F. Chan and T. P. Mathew. Domain decomposition algorithms. *Acta Numerica*, 3:61–143, 1994.

[6] V. Červený, I. A. Molotkov, and I. Psencik. *Ray Methods in Seismology*. Univ. Karlova Press, 1977.

[7] B. Engquist and O. Runborg. Computational high frequency wave propagation. *Acta Numerica*, 12:181–266, 2003.

[8] E. Fatemi, B. Engquist, and S. J. Osher. Numerical solution of the high frequency asymptotic expansion for the scalar wave equation. *J. Comput. Phys.*, 120(1):145–155, 1995.

[9] S. Fomel and J. A. Sethian. Fast phase space computation of multiple arrivals. *Proc. Natl. Acad. Sci. USA*, 99(11):7329–7334 (electronic), 2002.

[10] S. Hagdahl. *Hybrid Methods for Computational Electromagnetics in Frequency Domain*. PhD thesis, NADA, KTH, Stockholm, 2005.

[11] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii. Topology matching for fully automatic similarity estimation of 3D shapes. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 203–212, 2001. ACM Press, New York, NY, USA.

[12] S. Holmgren and K. Otto. Semicirculant preconditioners for first-order partial differential equations. *SIAM J. Sci. Comput.*, 15:385–407, 1994.

[13] P. E. Hussar, V. Oliker, H. L. Riggins, E.M. Smith-Rowlan, W.R. Klocko, and L. Prussner. An implementation of the UTD on facetized CAD platform models. *IEEE Antennas Propag.*, 42(2):100–106, 2000.

[14] S. Jin and X. Wen. Hamiltonian-preserving schemes for the Liouville equation with discontinuous potentials. *Comm. Math. Sci.*, 3:285–315, 2005.

[15] S. Jin and X. Wen. Hamiltonian-preserving schemes for the Liouville equation of geometrical optics with discontinuous local wave speeds. *J. Comput. Phys.*, 214(2):672–697, 2006.

[16] K. H. Karlsen, K. A. Lie, and N. H. Risebro. A fast marching method for reservoir simulation. *Computational Geosciences*, 4(2):185–206, 2000.

[17] J. Keller and R. M. Lewis. Asymptotic methods for partial differential equations: the reduced wave equation and Maxwell's equations. *Surveys Appl. Math.*, 1:1–82, 1995.

[18] J. B. Keller. The geometric theory of diffraction. In *Symposium on Microwave Optics*, Eaton Electronics Research Laboratory, McGill University, Montreal, Canada, June 1953.

[19] L. C. Kempel, J. L. Volakis, and S. Bindiganavale. Radiation and scattering from printed antennas on cylindrically conformal platforms. Technical report, Michigan Univ. Final Report, January 1994.

[20] R. Kimmel and J. A. Sethian. Computing geodesic paths on manifolds. *Proc. Natl. Acad. Sci. USA*, 95(15):8431–8435 (electronic), 1998.

[21] E. M. Koper, W. D. Wood, and S. W. Schneider. Aircraft antenna coupling minimization using genetic algorithms and approximations. *IEEE T. Aero. Elec. Sys.*, 40(2):742–751, 2004.

[22] A. P. Krasnojen. Features of creeping waves propagation on the dielectric-coated circular cylinder. *IEE Proc. Microw. Antennas Propag.*, 145(2):179–183, 1998.

[23] K. M. Kuperberg and C. S. Reed. A dynamical system on $\mathbb{R}^3$ with uniformly bounded trajectories and no compact trajectories. In *Proceedings of the American Mathematical Society*, volume 106, pages 1095–1097, 1989.

[24] H. Ling, R. Chou, and S. W. Lee. Shooting and bouncing rays: Calculating the RCS of an arbitrarily shaped cavity. *IEEE T. Antenn. Propag.*, 37:194–205, 1989.

[25] H. Liu, S. Osher, and R. Tsai. Multi-valued solution and level set methods in computational high frequency wave propagation. *Commun. Comput. Phys.*, 1:765–804, 2006.

[26] M. Motamed and O. Runborg. A fast phase space method for computing creeping rays. *J. Comput. Phys.*, 219(1):276–295, 2006.

[27] R. Paknys and D. R. Jackson. The relation between creeping waves, leaky waves, and surface waves. *IEEE T. Antenn. Propag.*, 53(3):898–907, 2005.

[28] R. Paknys and N. Wang. Creeping wave propagation constants and modal impedence for a dielectric coated cylinder. *IEEE T. Antenn. Propag.*, 34(5):674–680, 1986.

[29] J. Perez, J. A. Saiz, O. M. Conde, R. P. Torre, and M. F. Catedra. Analysis of antennas on board arbitrary structures modeled by NURBS surfaces. *IEEE T. Antenn. Propag.*, 45(6):1045–1053, 1997.

[30] J. Qian and W. W. Symes. An adaptive finite-difference method for traveltimes and amplitudes. *Geophysics*, 67(1):167–176, January-February 2002.

[31] O. Runborg. Mathematical models and numerical methods for high frequency waves. *Commun. Comput. Phys.*, 2:827–880, 2007.

[32] J. Shim and H. T. Kim. Dominance of creeping wave modes of backscattered field from a conducting sphere with dielectric coating. *Progress In Electromagnetic Research*, 21:293–306, 1999.

[33] Chi-Wang Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. ICASE Report 97-65, Brown University, 1997. Prepared for NASA Langley Research Center.

[34] W. W. Symes and J. Qian. A slowness matching Eulerian method for multivalued solutions of eikonal equations. *J. Sci. Comput.*, 19(1-3):501–526, 2003.

[35] G. Taylor. Another look at the line intersection problem. *Int. J. Geogr. Inf. Syst.*, 3(20):192–3, 1989.

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

[37] J. Vidale. Finite-difference calculation of traveltimes. *B. Seismol. Soc. Am.*, 78(6):2062–2076, December 1988.

[38] V. Vinje, E. Iversen, and H. Gjøystdal. Traveltime and amplitude estimation using wavefront construction. *Geophysics*, 58(8):1157–1166, 1993.

[39] L. Ying and E. J. Candes. Fast geodesics computation with the phase flow method. *J. Comput. Phys.*, 220(1):6–18, 2006.

Paper III

Taylor Expansion Errors in Gaussian Beam Summation

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

March 19, 2008

Abstract. Gaussian beam summation method is an asymptotic method for computing high frequency wave fields in smoothly varying inhomogeneous media. In this paper we study the accuracy of Gaussian beam summation method and derive error estimates related to the Taylor expansion of the phase and amplitude off the center of the beam. We show that in the case of using odd order beams, the error is smaller than a simple analysis would indicate because of error cancellation effects between the beams. Since the cancellation happens only when odd order beams are used, there is no remarkable gain in using even order beams. Moreover, in the case of constant coefficient equations, i.e. when the speed of propagation is constant, the local beam width is not a good indicator of accuracy, and there is no direct relation between the error and the beams width. We present numerical examples to verify the error estimates.

Keywords. wave propagation, high frequency, asymptotic approximation, summation of Gaussian beams, accuracy, error estimates

1 Introduction

Simulation of wave propagation is expensive when the frequency becomes high. In this case, a large number of grid points are needed to resolve the wave oscillations, and the computational cost to maintain constant accuracy grows algebraically with the frequency. At sufficiently high frequencies, therefore, direct simulations are no longer feasible.

Instead one can use high frequency asymptotic models for wave propagation. The most popular one is geometrical optics, which is obtained when the frequency tends to infinity. The unknowns in geometrical optics are phase and amplitude which are independent of the frequency and vary on a much coarser scale than the full wave solution. They can therefore be computed at a computational cost independent of the frequency. However, a main drawback of geometrical optics is that the model breaks down at caustics, where geometrical optics rays intersect and the predicted amplitude is unbounded.

Gaussian beams approximation is another high frequency asymptotic model which is valid also at caustics. It was introduced by Popov [1], based on an earlier work of Babic and Pankratova [2]. A Gaussian beam is an approximate high frequency solution to the linear wave equation which is concentrated close to a standard ray of geometrical optics, called the central ray of the beam. Although the phase function is real-valued along the central ray, Gaussian beams accept complex-valued phase functions off their central ray. The imaginary part of the phase is chosen such that the solution decays

exponentially away from the central ray, maintaining a Gaussian-shaped profile. The main advantage of this construction is that it gives the correct solution also at caustics. It has been proved to be beneficial in seismic imaging, [6, 7].

Numerical methods based on Gaussian beams use the superposition principle. Individual beams are computed via ray tracing like equations, where quantities such as the curvature and width of beams are calculated from ordinary differential equations (ODEs) along the central rays, and contribution of the beams concentrated close to their central rays are determined by Taylor expansion. The result is then summed to find the full wave field. See for example [3, 4, 5, 6, 7]. For a rigorous mathematical analysis of Gaussian beams we refer to [8].

In this paper we derive error estimates for the beam summation method. Some error estimates for this method have been derived earlier, [9, 10]. We aim to give a more complete picture of the error by also including the error due to the spreading of the beams, which is related to the Taylor expansion of the phase and amplitude off the center of the beam. This error is recognized as important in e.g. [9]. It turns out that, in the case of using odd order beams, the error is smaller than a simple analysis would indicate because of error cancellation effects between the beams. Since the cancellation happens only when odd order beams are used, there is no remarkable gain in using even order beams. Moreover, we show that in the case of constant coefficient equations, i.e. when the speed of propagation is constant, the local beam width is not a good indicator of accuracy, and there is no direct relation between the error and the beams width. However, this may not be true in the case of varying speed of propagation, where the beam width can be an important factor in Taylor expansion error.

In Section 2, we review the construction of Gaussian beams and the Gaussian beam summation method. Accuracy of Gaussian beam summation is studied in Section 3, where the main result is formulated together with numerical examples verifying the obtained error estimates. In Section 4, the proof of the main theorem is given in detail. Finally, in Section 5, we compute the errors analytically in the case of constant coefficient equations and give some remarks on how to select the Gaussian beam parameters.

2 Gaussian beam summation method

Gaussian beams are obtained when the linear wave equation is solved with initial or boundary data in the shape of a Gaussian bell. A Gaussian beam is an asymptotic solution concentrated on its central ray in the domain. By the superposition principle for linear equations, such solutions can be added to find the full wave field. The initial/boundary data for beams are obtained such that the wave data at the source is well approximated. In this section, we consider the Helmholtz equation (or the reduced wave equation) and review the construction of Gaussian beams and their summation.

2.1 Construction of Gaussian beams

Consider the Helmholtz equation

$$\Delta u(\boldsymbol{x}) + \frac{\omega^2}{c(\boldsymbol{x})^2}\, u(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \mathbb{R}^2,$$

where $\omega \gg 1$ and $c(\boldsymbol{x})$ are the frequency and speed of propagation, respectively. We substitute the WKBJ ansatz

$$u_{\mathrm{GB}}(\boldsymbol{x}) = e^{i\omega\phi(\boldsymbol{x})} \sum_{k=0}^{\infty} A_k(\boldsymbol{x})(i\omega)^{-k}, \tag{1}$$

into the Helmholtz equation. Here, the phase function ϕ and the amplitude functions A_k are assumed to be smooth and independent of ω. Equating coefficients of powers of ω to zero gives us the *eikonal equation* and the *transport equation* for the phase and the first amplitude term in the frequency domain,

$$|\nabla\phi| = 1/c(\boldsymbol{x}), \qquad 2\,\nabla A_0 \cdot \nabla\phi + A_0\,\Delta\phi = 0.$$

For the remaining amplitude terms, we get additional transport equations

$$2\,\nabla A_{k+1} \cdot \nabla\phi + A_{k+1}\,\Delta\phi + \Delta A_k = 0.$$

When ω is large, only the first term in the WKBJ expansion is significant. Therefore, in the standard Gaussian beam method only the first order term in the expansion is kept. In this paper, without loss of generality, we only consider the first term (with $k = 0$) and drop zero, writing A instead of A_0. The same result holds also with including higher order terms.

While in the standard geometrical optics, the phase is real-valued, in Gaussian beam construction, the phase is real-valued only on the central ray of the beam. Away from the central ray, it is complex-valued with *positive imaginary part*. The solution will then be exponentially decreasing away from the central ray, maintaining its Gaussian shape. Another difference between geometrical optics and Gaussian beams is that in the Gaussian beam construction, ϕ is constructed based on one specific ray (the beam's central ray), while in geometrical optics it is globally defined for all rays. Note that the Gaussian beam method can also be formulated globally. In this case we obtain a *complex eikonal equation*, [11, 12]. But unfortunately, the question of existence and uniqueness of the complex eikonal equation is to a certain extent still open. In particular what precise boundary conditions are well-posed for the above setting is not known.

The Gaussian beam approximation breaks down when $\phi(\boldsymbol{x})$ becomes non-smooth. This is also typical for complex eikonal equation. It happens in general some distance away from the central beam. On the other hand, away from the beam the solution rapidly goes to zero and the precise value of the phase is not important. One usually deals with this problem by multiplying the amplitude with a smooth cut-off function that is one close to the central ray, and zero for some fixed distance away from it.

$$u_{\mathrm{GB}}(\boldsymbol{x}) = \varphi(\boldsymbol{x})A(\boldsymbol{x})e^{i\omega\phi(\boldsymbol{x})}. \tag{2}$$

Here $\varphi(\boldsymbol{x})$ is smooth and compactly supported around the central ray.

For a beam starting at point $\boldsymbol{x}_0$ with direction $\boldsymbol{p}_0$, the corresponding central ray satisfies the ray tracing ODEs

$$\frac{d\boldsymbol{x}}{dt} = c^2(\boldsymbol{x})\boldsymbol{p}, \quad \frac{d\boldsymbol{p}}{dt} = -\frac{\nabla c(\boldsymbol{x})}{c(\boldsymbol{x})}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \quad \boldsymbol{p}(0) = \frac{\boldsymbol{p}_0}{|\boldsymbol{p}_0| c(\boldsymbol{x}_0)}, \tag{3}$$

with t being the real-valued travel time along the ray. Note that since $\boldsymbol{p}(t) = \nabla\phi(\boldsymbol{x}(t))$ satisfies the eikonal equation, we can set $\boldsymbol{p} = (\cos\theta \ \sin\theta)^\top / c(\boldsymbol{x})$ and reduce (3) to

$$\frac{dx}{dt} = c(\boldsymbol{x})\cos\theta, \qquad \frac{dy}{dt} = c(\boldsymbol{x})\sin\theta, \qquad \frac{d\theta}{dt} = c_x(\boldsymbol{x})\sin\theta - c_y(\boldsymbol{x})\cos\theta. \tag{4}$$

The complex-valued A and ϕ close to the central ray are then approximated by Taylor expansions around the ray,

$$A(\boldsymbol{x}) \approx A(\boldsymbol{x}^*) + (\boldsymbol{x} - \boldsymbol{x}^*) \cdot \nabla A(\boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top D^2 A(\boldsymbol{x}^*)\,(\boldsymbol{x} - \boldsymbol{x}^*) + \cdots, \tag{5}$$

$$\phi(\boldsymbol{x}) \approx \phi(\boldsymbol{x}^*) + (\boldsymbol{x} - \boldsymbol{x}^*) \cdot \nabla \phi(\boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top D^2 \phi(\boldsymbol{x}^*)\,(\boldsymbol{x} - \boldsymbol{x}^*) + \cdots, \tag{6}$$

where $\boldsymbol{x}^* = \boldsymbol{x}(t)$ for some t. The Taylor coefficients $\phi(\boldsymbol{x}(t))$, $\nabla\phi(\boldsymbol{x}(t))$, $A(\boldsymbol{x}(t))$, etc. on the central ray can be computed. The lowest ones are given directly,

$$\phi(\boldsymbol{x}(t)) = \phi(\boldsymbol{x}_0) + t, \qquad \nabla\phi(\boldsymbol{x}(t)) = \boldsymbol{p}(t),$$

and the higher ones can be obtained by solving ODEs similar to (3). The most common approximation by far is to approximate $A(\boldsymbol{x})$ to zeroth order and $\phi(\boldsymbol{x})$ to second order (a first order Gaussian beam). In this case we have, [5],

$$A(\boldsymbol{x}(t)) = A(\boldsymbol{x}_0)\left(\frac{c(\boldsymbol{x}(t))}{c(\boldsymbol{x}_0)}\frac{Q(0)}{Q(t)}\right)^{1/2}, \qquad D^2\phi(\boldsymbol{x}(t)) = HNH^{-1},$$

with

$$H = \begin{pmatrix} \sin\theta & \cos\theta \\ -\cos\theta & \sin\theta \end{pmatrix}, \quad N = \begin{pmatrix} P/Q & -c_1/c^2 \\ -c_1/c^2 & -c_2/c^2 \end{pmatrix}, \quad \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}\nabla c,$$

where the complex-valued scalar functions P and Q satisfy the dynamic ray tracing ODEs

$$\frac{dQ}{dt} = c^2(\boldsymbol{x})\,P, \qquad Q(0) = Q_0, \tag{7}$$

$$\frac{dP}{dt} = -\frac{c_{xx}\sin^2\theta - 2c_{xy}\sin\theta\cos\theta + c_{yy}\cos^2\theta}{c(\boldsymbol{x})}\,Q, \quad P(0) = P_0. \tag{8}$$

It can be shown that if $Q_0 \neq 0$ and $\Im(P_0/Q_0) > 0$, then $Q(t) \neq 0$ and $\Im(P(t)/Q(t)) > 0$ along the central ray, [1]. Therefore, by a proper choice of initial data Q_0 and P_0, each beam will be regular (with finite amplitude at caustics) and concentrate along the central ray. A common choice is $Q_0 > 0$ and $P_0 = i$. Note that the quantities P and Q determine the wavefront curvature and the beam width.

2.2 Beam summation

Let the source be a curve $\boldsymbol{x}_0(s)$ in $\mathbb{R}^2$ parameterized by s. We introduce the notation $A(\boldsymbol{x}, s)$, $\phi(\boldsymbol{x}, s)$ and $\varphi(\boldsymbol{x}, s)$ for the amplitude, phase and cut-off of a beam with initial position $\boldsymbol{x}_0(s)$. In the Gaussian beam summation method, the initial/boundary condition on $\boldsymbol{x}_0(s)$ for the wave field is decomposed into initial conditions for several beams with different initial positions $\boldsymbol{x}_0(s_j)$. Individual Gaussian beams are computed by solving the ODEs (3) and (7,8). The contributions of the beams concentrated close to their central rays are determined by the approximations (5,6) entered in (2). The wave field is then obtained by summing over the beams

$$u_s(\boldsymbol{x}) = \sum_{j\in\mathbb{Z}} \varphi(\boldsymbol{x}, s_j) A(\boldsymbol{x}, s_j) e^{i\omega\phi(\boldsymbol{x}, s_j)}. \tag{9}$$

The initial conditions for the Taylor coefficient ODEs are chosen such that u_s well approximates the exact initial/boundary data.

As an example, a plane wave on the y-axis, $\boldsymbol{x}_0(s) = (s, 0)$, can be approximated by a sum of beams, [6],

$$1 = \sum_j \frac{1}{\sqrt{\pi}} \frac{h}{w_0} e^{-(s-s_j)^2/w_0^2} + \mathcal{O}(e^{-(w_0/h)^2}), \quad s_j = jh, \tag{10}$$

with h and w_0 representing the initial spacing of the beams and the initial beam widths, See Figure 1. To properly choose the initial data, one must therefore take

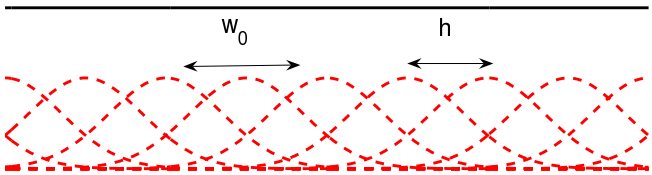


Figure 1: The sum of several Gaussian functions is almost constant. A plane wave can therefore be decomposed approximately to a sum of parallel Gaussian beams.

the parameters w_0 and h such that $w_0 > h$. Note that for computational efficiency, h should not be taken too small. Choosing a real-valued, positive Q_0, $P_0 = i$ and using first order beams, the wave field (9) produces a plane wave on $\boldsymbol{x}_0(s) = (s, 0)$, if

$$w_0 = \left(\frac{2Q_0}{\omega}\right)^{1/2}, \qquad A(\boldsymbol{x}_0, s_j) = \frac{1}{\sqrt{\pi}} \frac{h}{w_0}.$$

Motivated by this example, we write

$$u_s(\boldsymbol{x}) = \sum_{j\in\mathbb{Z}} \frac{\alpha h}{w_0} \varphi(\boldsymbol{x}, s_j) A(\boldsymbol{x}, s_j) e^{i\omega\phi(\boldsymbol{x}, s_j)}, \tag{11}$$

where α is a constant. For an initial plane wave, we have $\alpha = \sqrt{1/\pi}$ and $A(\boldsymbol{x}_0, s_j) = 1$.

Note that since $w_0 \propto \omega^{-1/2}$,

$$u_s(\boldsymbol{x}) \propto \omega^{1/2} h \sum_{j\in\mathbb{Z}} \varphi(\boldsymbol{x}, s_j) A(\boldsymbol{x}, s_j) e^{i\omega\phi(\boldsymbol{x},s_j)} \approx \omega^{1/2} \int \varphi(\boldsymbol{x}, s) A(\boldsymbol{x}, s) e^{i\omega\phi(\boldsymbol{x},s)} ds, \quad (12)$$

meaning that the summation (11) approximates the superposition integral.

In what follows, in order to simplify the calculations, we assume that all beams, originating from $\boldsymbol{x}_0(s)$, shoot out orthogonally. We denote by $\boldsymbol{X}(t, s)$ the location of the center ray originating in $\boldsymbol{x}_0(s)$ after time t. We further assume that $\phi(\boldsymbol{x}_0(s), s) = 0$.

We make one observation that will be used in the analysis below. It is well-known that $\boldsymbol{X}_t \parallel \nabla_x\phi$, $\boldsymbol{X}_t \cdot \nabla_x\phi = 1$ and $\boldsymbol{X}_s \perp \boldsymbol{X}_t$ under the assumptions made above. Therefore, since $\phi(\boldsymbol{X}(t, s), s) = t$,

$$0 = \frac{d}{ds}\phi(\boldsymbol{X}(t, s), s) = \boldsymbol{X}_s \cdot \nabla_x\phi + \phi_s. = \phi_s(\boldsymbol{X}(t, s), s) \quad (13)$$

Hence $\phi_s = 0$ everywhere on the central rays.

3 Accuracy of Gaussian beams summation

In this section we study the accuracy of summation of Gaussian beams. One can distinguish five different types of errors in the approximation:

1. High frequency approximation.
2. Error in initial data.
3. Taylor expansion error.
4. Cut-off error.
5. Error in numerical integrators for solving Taylor coefficient ODEs.

The first error depends on the number of terms used in the WKBJ approximation. For example, for standard beams it is of the order $O(1/\omega)$, meaning that each beam is a solution to the Helmholtz equation up to order $O(1/\omega)$. The second error represents how well the exact initial/boundary data is approximated by a sum of Gaussian beams. The third error is due to the fact that A and ϕ are not computed globally, and only their derivatives on the central beams are computed. One therefore needs to approximate their values around the central beams by Taylor expansions. The fourth error is caused by multiplying the solution by a smooth cut-off function in order to account for possible irregularities away from the central rays. Finally, the last error is the numerical error in solving the ODEs for computing Taylor coefficients. For example the global error in a fourth order Runge-Kutta method is $O(\Delta t^4)$, with Δt being the time-step.

Here, we will only concentrate on the Taylor expansion error.

3.1 Motivation and preliminaries

Let the source be a curve $\boldsymbol{x}_0(s)$ and assume that we look for the solution along a line $\boldsymbol{x} = (x, y^*)$. We simplify the notation by setting $A(\boldsymbol{x}, s) = A(x, s)$, $\phi(\boldsymbol{x}, s) = \phi(x, s)$, $\varphi(\boldsymbol{x}, s) = \varphi(x, s)$ and $s_j = jh$, with h representing the initial spacing of the beams,

$$u_s(x) = \sum_{j\in\mathbb{Z}} \frac{\alpha h}{w_0} \varphi(x, s_j) A(x, s_j) e^{i\omega\phi(x,s_j)}, \qquad s_j = jh.$$

To approximate this sum we let $X(s)$ denote the location of the center beam on the line (x, y^*) when the initial data is given at $\boldsymbol{x}_0(s)$. Hence, $X(s)$ is implicitly defined by

$$\boldsymbol{X}(t(s), s) = (X(s), y^*),$$

for some function $t(s)$. Figure 2 shows the setting for $\boldsymbol{x}_0(s) = (s, 0)$, as an example.

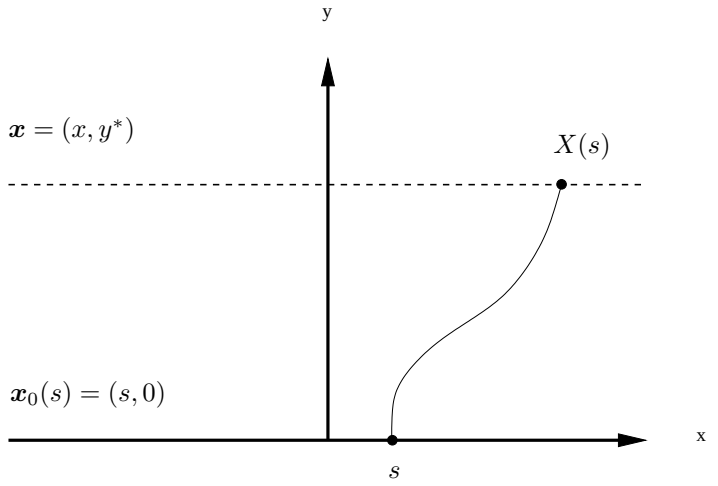


Figure 2: A schematic representation of the initial source and a beam central ray.

Then we approximate A up to level q and ϕ up to level $q + 2$,

$$A(x, s) \approx \tilde{A}_q(x, s) := A(X(s), s) + \cdots + \frac{(x - X(s))^q}{q!} \partial_x^q A(X(s), s), \tag{14}$$

$$\phi(x, s) \approx \tilde{\phi}_q(x, s) := \phi(X(s), s) + \cdots + \frac{(x - X(s))^{q+2}}{(q + 2)!} \partial_x^{q+2} \phi(X(s), s), \tag{15}$$

and assume that the approximate Gaussian beam solution is given by

$$\tilde{u}_s(x) = \sum_{j\in\mathbb{Z}} \frac{\alpha h}{w_0} \varphi_j(x) \tilde{A}_q(x, s_j) e^{i\omega\tilde{\phi}_q(x,s_j)}, \qquad \varphi_j(x) := \varphi(x, s_j).$$

We call this a $(q + 1)$-th order Gaussian beam solution. Note that for $q \geq 2$, one needs to include more terms in the WKBJ expansion in order to balance the high frequency approximation error and the Taylor expansion error. For example, for a third order

Gaussian beam with $q = 2$, the first amplitude term A_0 in (1) is approximated up to level 2, the second amplitude term A_1 in (1) is approximated up to level 0, and ϕ is approximated up to level 4.

Our motivation for considering the Taylor expansion error comes from the following observation. We define the width of the Gaussian beam passing through (x, y^*) as

$$w(x) := \frac{1}{\sqrt{\omega\, \Im\phi_{xx}(x, X^{-1}(x))}}.$$

Because of the term $e^{i\omega(x-X(s))^2\phi_{xx}/2}$ the solution will be close to zero for $|x - X(s)| > w(x)$. A simple error analysis would therefore give

$$u_s - \tilde{u}_s = (A - \tilde{A}_q)e^{i\omega\tilde{\phi}_q} + Ae^{i\omega\tilde{\phi}_q}(e^{i\omega(\phi-\tilde{\phi}_q)} - 1) = O(w^{q+1})e^{i\omega\tilde{\phi}_q} + Ae^{i\omega\tilde{\phi}_q}(e^{iO(\omega w^{q+3})} - 1).$$

Hence, the error would be $O(w^{q+1}(1+\omega w^2)) = O(w^{q+1}) = O(\omega^{-(q+1)/2})$. In particular, for first order beams with $q = 0$, the convergence rate in ω would be half order, i.e. proportional to $\omega^{-1/2}$.

We now consider two numerical examples and use first order Gaussian beams to verify this convergence rate. In the first example, a plane wave generated on the line $y = 0$ propagates orthogonally into the computational domain with a variable speed of propagation. Figure 3a shows the central rays of Gaussian beams, and Figure 3c shows the absolute value of the Gaussian beams and geometrical optics solutions along the line $y = 0.6$, shown in bold in Figure 3a. Figure 3e shows the logarithmic scale of the maximum error between the Gaussian beams solution and the geometrical optics solution as a reference solution. As it can be seen, the convergence rate of the error is surprisingly proportional to ω^{-1}, which is half order better than what we expected. Note that the geometrical optics error is ω^{-1}, and since a lower accuracy was expected for the Gaussian beam method, it is fine to compare the solution with the geometrical optics solution.

In the second example, a plane wave generated on the line $x = 0$ propagates with an angle of 45^o into the computational domain with a variable speed of propagation. The convergence rate of the error, shown in Figure 3f, is again proportional to ω^{-1}.

We will therefore study the Taylor expansion error more carefully to describe why it is smaller than what we expected.

3.2 Main result

For our results we make the following precise assumptions

(**A1**) *Smoothness of all coefficients.* We assume $A(x, s) \in C_b^{p+q+2}(\mathbb{R}^2)$, the space of functions with $p + q + 2$ continuous and bounded derivatives. Similarly $\phi(x, s) \in C^{p+q+4}$ and $\boldsymbol{X}(t, s) \in C^{p+1}$, with $p \geq 2$.

(**A2**) *Algebraic growth of phase off center beam.* For $p_1, p_2 \leq p$, we have

$$\partial_x^{p_1}\partial_s^{p_2}\phi(x, s) \leq C(1 + |x - X(s)|^k).$$

In particular, all derivatives are bounded on the center beam, $x = X(s)$.

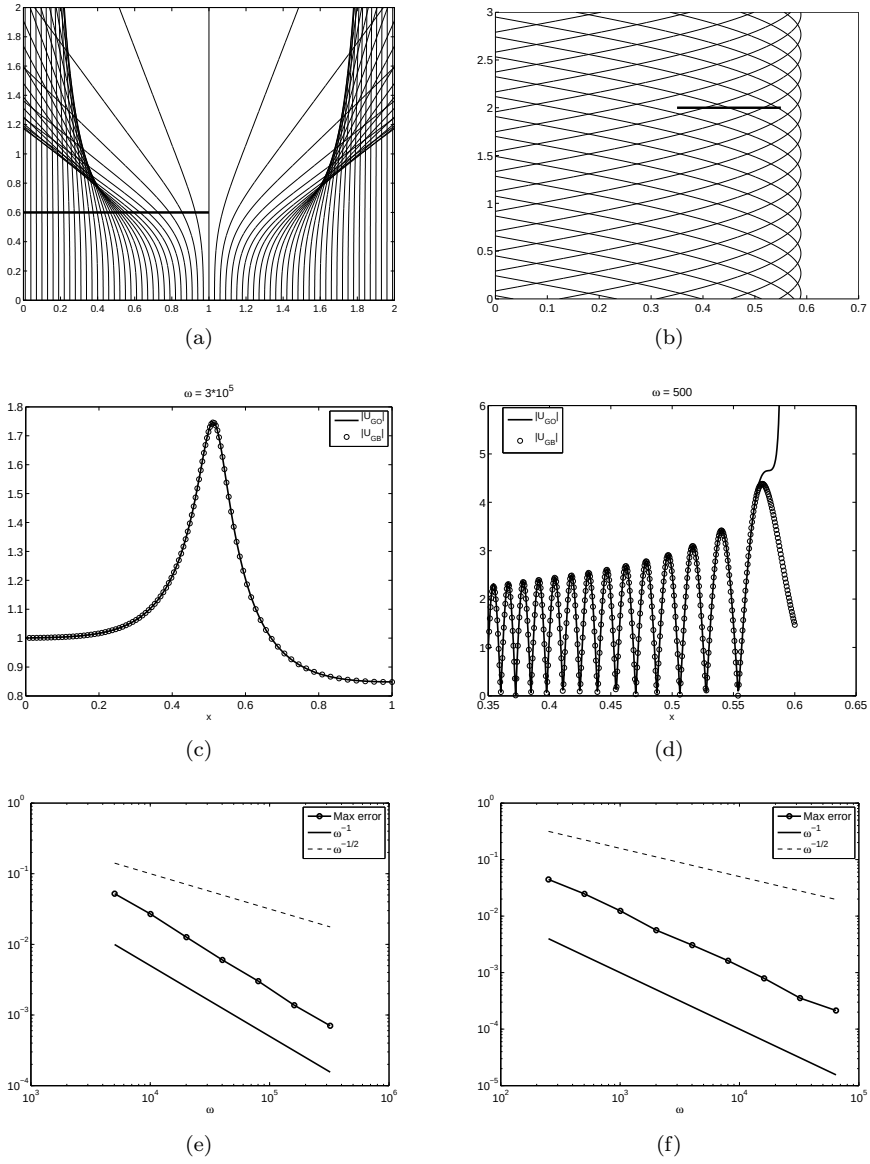


Figure 3: Left and right top figures show the central rays of Gaussian beams by an initial plane wave on x- and y-axis, respectively. Middle figures show the absolute value of the Gaussian beams and geometrical optics solutions along the lines $y = 0.6$ and $y = 2$. Bottom figures show the logarithmic scale of the maximum error between the Gaussian beams solution and the geometrical optics solution. The convergence rate of the maximum error is ω^{-1}.

(**A3**) *No caustics.* The derivative $X'(s)$ is bounded away from zero, $0 < c_0 \leq X'(s) \leq c_1 < \infty$.

(**A4**) *Non-degeneracy of each beam.* The imaginary part of ϕ_{xx} is strictly positive

$$0 < c_0 \leq \Im\phi_{xx}(X(s), s) \leq c_1 < \infty. \tag{16}$$

This means that the approximate beams will have a fast decay off the central beam for high frequencies, and also that the beam width never vanishes. The last point is an important feature of Gaussian beams, related to the fact that Gaussian beams can approximate the exact field at caustics.

(**A5**) *Cut-off of fixed size.* We use $\varphi(x,s) = \varphi(x - X(s))$ with $\varphi \in C^\infty$ such that $\varphi(x) = 1$ for $|x| \leq \alpha/2$ and $\varphi(x) = 0$ for $|x| > \alpha$. The size of α will be chosen "small enough" depending on ϕ but independent of ω. Moreover, we assume $\varphi'(0) = 0$.

Then we can show

Theorem 1. *(Main Theorem) For the $(q+1)$-th order Gaussian beams, we have*

$$|u_s(x) - \tilde{u}_s(x)| = |E_{non-osc} + E_{osc}| \leq C\left(\omega^{-\frac{q^*}{2}} + \omega^{\frac{q+1}{2}}\beta^{-p}\right), \tag{17}$$

where

$$\beta := w(x)/h, \qquad q^* = \begin{cases} q+2, & q \text{ even}, \\ q+1, & q \text{ odd}. \end{cases}.$$

The constant C depends on the initial data, P_0 and Q_0, for the beams but does not depend on x, ω or h.

For the first part of the error we have

$$\left| E_{\text{non}-\text{osc}} - \alpha\, C^*(x)\, \frac{w^{q^*+1}}{w_0} \right| \leq C' \frac{w^{q^*+2}}{w_0},$$

meaning that the leading order term of the error $E_{non-osc}$ in ω is $\alpha\, C^*(x)\, w^{q^*+1}/w_0$, with α being a constant and C^* given by (26).

Remark 1. *As the theorem shows, although the formal convergence rate, when $\beta = w(x)/h$ is held fixed, is just half order in ω for first order beams, but typically the second term in (17) is smaller because of the "fast" decay of β^{-p}. In practice therefore the convergence rate is full first order, which is the same as geometrical optics.*

In order to make the second term of the error small, one should therefore take $h < w(x)$. However, h should not be chosen too small for computational complexity reasons. It is also important to note that to balance the error with the error in initial data, h should also relate to the initial beam width w_0. For wider beams the first term dominates.

Remark 2. *As the estimate (17) suggests, there is no remarkable gain in using even order beams (with an odd q). However, one should note that this is only true in the case of the summation of beams. If we only have one beam, this does not hold.*

Remark 3. *The first term in (17) corresponds to the leading order error in the superposition integrals. The second term corresponds to the truncation error of the trapezoidal rule applied to the superposition integrals. This latter error is also investigated in [9], where it is called the discretization error.*

4 Proof of main result

Before going on to the proof of Theorem 1, we prove the following utility lemma concerning estimates for the composition of two functions.

Lemma 1. *Suppose $f(z)$ and $g_\delta(z)$ belongs to $C^p(\mathbb{R})$ for each value of the parameter δ. If*

$$|g_\delta^{(n)}(z)| \le C_g(1+|z|^q), \qquad 1 \le n \le p, \tag{18}$$

where C_g and $q \ge 0$ are constants independent of z and δ, then, for $1 \le n \le p$, there are functions $h_{m,n} \in C^{p-n}(\mathbb{R})$ and constants $C_{m,n}$ independent of z and δ, such that

$$\frac{d^n}{dz^n} f(g_\delta(z)) = \sum_{m=1}^{n} h_{m,n}(z) f^{(m)}(g_\delta(z)), \qquad \max_{0\le k\le p-n} |h_{m,n}^{(k)}(z)| \le C_{m,n}(1+|z|^{qn}). \tag{19}$$

If $q = 0$ and $|g_\delta(z)| \le C_g$, then

$$\left| \frac{d^n}{dz^n} e^{z^r g_\delta(z)} \right| \le C(1+|z|^{rn}) e^{z^r g_\delta(z)}, \qquad 0 \le n \le p, \tag{20}$$

for some constant C.

Proof. We show (19) by induction. For $n = 1$ we have $h_{1,1} = g_\delta'(z) \in C^{p-1}$ and the statement clearly holds. Suppose (19) is true for $1 \le n' \le n < p$. Then

$$\frac{d^{n+1}}{dz^{n+1}} f(g_\delta(z)) = \sum_{m=1}^{n} h_{m,n}' f^{(m)}(g_\delta) + g_\delta' h_{m,n}(z) f^{(m+1)}(g_\delta).$$

Thus

$$h_{m,n+1}(z) = \begin{cases} h_{m,n}', & m = 1, \\ h_{m,n}' + g_\delta' h_{m-1,n}, & 1 < m \le n, \\ g_\delta' h_{m-1,n}, & m = n+1. \end{cases}$$

Using the induction hypothesis, we immediately get that $h_{m,n+1}(z) \in C^{p-n-1}(\mathbb{R})$. Moreover,

$$\max_{0\le k\le p-n-1} |h_{m,n+1}^{(k)}(z)| \le \max_{0\le k\le p-n-1} |h_{m,n}^{(k+1)}(z)| + \max_{0\le k\le p-n-1} \sum_{j=0}^{k} c_{jk} |h_{m-1,n}^{(j)}(z)| |g_\delta^{(k+1-j)}(z)|$$

The first term is bounded by $C_{1,1}(1+|z|^{qn})$ by assumption and for the second term we can estimate

$$|h_{m-1,n}^{(j)}(z)| |g_\delta^{(k+1-j)}(z)| \le C_{m-1,n}(1+|z|^{qn})\, C_g(1+|z|^q) \le C'(1+|z|^{q(n+1)}),$$

which proves (19). When $q = 0$, we have for $1 \le n \le p$,

$$\frac{d^n}{dz^n} z^r g_\delta(z) = \sum_{j=0}^{\min(r,n)} c_{j,n} g_\delta^{(n-j)}(z) \frac{d^j}{dz^j} z^r \le \sum_{j=0}^{\min(r,n)} C_g |z|^{r-j} \le C'(1+|z|^r).$$

By taking $f(z) = e^z$ in (19) the result (20) follows. ∎

We can now start with the main proof. The error that we need to bound is given by

$$E(x) = u_s(x) - \tilde{u}_s(x) = \frac{\alpha h}{w_0} \sum_{j\in\mathbb{Z}} \varphi(X(s_j) - x) \left(A(x, s_j)\, e^{i\omega\phi(x,s_j)} - \tilde{A}_q(x, s_j)\, e^{i\omega\tilde{\phi}_q(x,s_j)} \right).$$

For a fixed x we set

$$f(j) = \varphi(X(s_j) - x) \left(A(x, s_j)\, e^{i\omega\phi(x,s_j)} - \tilde{A}_q(x, s_j)\, e^{i\omega\tilde{\phi}_q(x,s_j)} \right).$$

Then the Poisson summation formula gives

$$E = \frac{\alpha h}{w_0} \sum_{j\in\mathbb{Z}} f(j) = \frac{\alpha h}{w_0} \sum_{k\in\mathbb{Z}} \hat{f}(k),$$

where

$$\begin{aligned}
\hat{f}(k) &= \int f(s) e^{-2\pi i s k} ds \\
&= \int \varphi(X(sh) - x) \left(A(x, sh)\, e^{i\omega\phi(x,sh)} - \tilde{A}_q(x, sh)\, e^{i\omega\tilde{\phi}_q(x,sh)} \right) e^{-2\pi i s k} ds \\
&= \frac{1}{h} \int \varphi(X(s) - x) \left(A(x, s)\, e^{i\omega\phi(x,s)} - \tilde{A}_q(x, s)\, e^{i\omega\tilde{\phi}_q(x,s)} \right) e^{-2\pi i s k/h} ds.
\end{aligned}$$

Let us denote $X^{-1}(x)$ by $m(x)$ and then, since $X'(s)$ is bounded away from zero we can use the change of variables

$$z = \frac{X(s) - x}{w(x)} \quad \Rightarrow \quad s = m(x + w(x)z).$$

We obtain

$$\begin{aligned}
\hat{f}(k) = \frac{w}{h} \int \varphi(wz) \Big[& A(x, m(x + wz))\, e^{i\omega\phi(x,m(x+wz))} - \\
& \tilde{A}_q(x, m(x + wz))\, e^{i\omega\tilde{\phi}_q(x,m(x+wz))} \Big] e^{-2\pi i m(x+wz)k/h}\, m'(x + wz)\, dz.
\end{aligned}$$

Finally, letting

$$D_A(x, s) := A(x, s) - \tilde{A}_q(x, s), \qquad D_\phi(x, s) := \phi(x, s) - \tilde{\phi}_q(x, s).$$

and recalling that $\operatorname{supp}\varphi \subset [-\alpha, \alpha]$, we can write the integral as

$$\hat{f}(k) = \frac{w}{h} \int_{|z|\le \frac{\alpha}{w}} \varphi \left(D_A + A(e^{i\omega D_\phi} - 1) \right) e^{i\omega\tilde{\phi}_q} e^{-2\pi i m k/h} m' dz. \tag{21}$$

We consider the non-oscillatory case $k = 0$ and the oscillatory case $k \neq 0$ separately and write

$$E = \frac{\alpha h}{w_0}\, \hat{f}(0) + \frac{\alpha h}{w_0} \sum_{k\neq 0} \hat{f}(k) =: E_{\text{non-osc}} + E_{\text{osc}}.$$

The non-oscillatory case will normally be the dominant contribution. The oscillatory case corresponds to the truncation error when the superposition integral (12) is approximated by a trapezoidal rule discretization.

4.1 Non-oscillatory case

We will now consider the case $k = 0$ and approximate the terms in the integral (21) by their Taylor expansion. Let us use the shorthand

$$\tilde{a}_p(x) = \frac{(-1)^p}{p!}\partial_x^p A(x, m(x)), \qquad \tilde{b}_p(x) = \frac{1}{p!}\frac{d^p}{dz^p}A(x, m(x+z))\bigg|_{z=0}.$$

and

$$\tilde{p}_p(x) = \frac{(-1)^p}{p!}\partial_x^p \phi(x, m(x)), \qquad \tilde{r}_p(x) = \frac{1}{p!}\frac{d^p}{dz^p}\phi(x, m(x+z))\bigg|_{z=0}.$$

We note that, in this notation

$$\tilde{A}_q(x, m(x+z)) = \tilde{a}_0(x+z) + z\tilde{a}_1(x+z) + \cdots + z^q\tilde{a}_q(x+z),$$

$$\tilde{\phi}_q(x, m(x+z)) = \tilde{p}_0(x+z) + z\tilde{p}_1(x+z) + \cdots + z^{q+2}\tilde{p}_{q+2}(x+z).$$

Let

$$a_1(x) = \tilde{a}_{q+1}(x), \qquad a_2(x) = \tilde{a}_{q+2}(x) + \tilde{a}'_{q+1}(x),$$

$$b_1(x) = i\frac{\tilde{p}_{q+3}(x)}{\Im\phi_{xx}(x, m(x))}, \qquad b_2(x) = i\frac{\tilde{p}_{q+4}(x) + \tilde{p}'_{q+3}(x)}{\Im\phi_{xx}(x, m(x))},$$

$$c_1(x) = \Re\frac{\tilde{r}_2(x)}{\Im\phi_{xx}(x, m(x))}, \qquad c_2(x) = i\frac{\tilde{r}_3(x) - \sigma\tilde{p}_3(x)}{\Im\phi_{xx}(x, m(x))}.$$

where $\sigma = 1$ for $q = 0$ and $\sigma = 0$ for $q > 0$. We then approximate

$$D_A(x, m(x+wz)) \approx w^{q+1}\tilde{D}_A(x, z) := (wz)^{q+1}a_1(x) + (wz)^{q+2}a_2(x),$$
$$e^{i\omega D_\phi(x, m(x+wz))} - 1 \approx w^{q+1}\tilde{B}(x, z) := w^{q+1}b_1(x)z^{q+3} + w^{q+2}(b_2(x)z^{q+4} + \sigma b_1^2(x)z^{2q+6}/2),$$
$$e^{i\omega\tilde{\phi}_q(x, m(x+wz))} \approx \tilde{C}(x, z) =: e^{i\omega\phi(x, m(x)) + iz^2c_1(x) - z^2/2}(1 + c_2(x)wz^3).$$

The residual terms are denoted

$$D_A(x, m(x+wz)) - w^{q+1}\tilde{D}_A(x, z) =: w^{q+3}R_A(x, z),$$
$$e^{i\omega D_\phi(x, m(x+wz))} - 1 - w^{q+1}\tilde{B}(x, z) =: w^{q+3}R_B(x, z),$$
$$e^{i\omega\tilde{\phi}_q(x, m(x+wz))} - \tilde{C}(x, z) =: w^2R_C(x, z).$$

Then we have

Lemma 2. *Let the residual terms R_A, R_B and R_C be defined as above. Under assumptions (A1) and (A2), for small enough α,*

$$|R_A| \le C|z|^{q+3}, \qquad |R_B| \le Ce^{z^2/7}, \qquad |R_C| \le Ce^{-z^2/4}, \quad \forall|z| \le \alpha/w,$$

where the constant C is independent of x, ω and z.

Proof. We note that $\tilde{a}_q(x+z)$ are the first q coefficients in the Taylor expansion of $A(x+z-x', m(x+z))$ around $x'=0$. Therefore, by Taylor's theorem and assumption (A1)

$$\left|D_A(x, m(x+z)) - z^{q+1}\tilde{a}_{q+1}(x+z) - z^{q+2}\tilde{a}_{q+2}(x+z)\right| \le C|z|^{q+3}.$$

Expanding the second and third terms around $z=0$ gives the bound for R_A.

We now estimate ωD_ϕ in two different ways. By Taylor's theorem as above,

$$|\omega D_\phi(x, m(x+wz))| \le C\omega|wz|^{q+3}|(1+|wz|^{q^*}),$$

where we used the growth condition (A2) for ϕ to bound the error term. Then, for $|z| \le \alpha/w$, and small enough α,

$$|\omega D_\phi| \le Cw^{q+1}|z|^{q+3}, \qquad |\omega D_\phi| \le Cz^2\alpha^{q+1}(1+\alpha^{q^*}) \le \frac{z^2}{8},$$

implying

$$\left|e^{i\omega D_\phi} - 1 - i\omega D_\phi - \frac{(i\omega D_\phi)^2}{2}\right| \le \frac{1}{6}|\omega D_\phi|^3 e^{|\omega D_\phi|} \le Cw^{3q+3}|z|^{3q+9}\exp\left(\frac{z^2}{8}\right).$$

Moreover, the same steps as for D_A together with (A2) gives

$$\left|D_\phi(x, m(x+z)) - z^{q+3}\tilde{p}_{q+3}(x) - z^{q+4}(\tilde{p}_{q+4}(x) + \tilde{p}'_{q+3}(x))\right| \le C|z|^{q+5}(1+|z|^{q^*}), \quad (22)$$

and since $\omega = 1/w^2\Im\phi_{xx}$, when $|z| \le \alpha/w$,

$$\left|i\omega D_\phi(x, m(x+wz)) - w^{q+1}z^{q+3}b_1(x) - w^{q+2}z^{q+4}b_2(x)\right| \le Cw^{q+3}|z|^{q+5}.$$

Finally, for $q>0$, clearly $|\omega D_\phi|^2 \le Cw^{2q+2}|z|^{2q+6} \le Cw^{q+3}|z|^{2q+6}$ and for $q=0$ we get

$$|(i\omega D_\phi)^2 - w^2b_1^2z^6| = \frac{|D_\phi^2 - (w^3z^3\tilde{p}_3)^2|}{w^4\Im\phi_{xx}^2} = \frac{|D_\phi - w^3z^3\tilde{p}_3||D_\phi + w^3z^3\tilde{p}_3|}{w^4\Im\phi_{xx}^2} \le Cw^3|z|^7.$$

Thus,

$$|R_B| \le Cw^{2q}|z|^{3q+9}\exp\left(\frac{z^2}{8}\right) + C|z|^{q+5} + (1-\sigma)C|z|^{2q+6} + \sigma C|z|^{2q+7} \le C'\exp\left(\frac{z^2}{7}\right).$$

To show the third inequality, we note that since $\phi_s(x, m(x)) \equiv 0$ by (13), we have $\tilde{r}_1(x) = 0$. Therefore by Taylor's theorem and assumption (A2), for $q' \ge 2$,

$$\left|\phi(x, m(x+z)) - \phi(x, m(x)) - \sum_{p=2}^{q'} z^p\tilde{r}_p(x)\right| \le C|z|^{q'+1}(1+|z|^{q^*}). \quad (23)$$

Let $v(x,z) = \tilde{\phi}_q(x, m(x+z)) - \phi(x, m(x)) - z^2\tilde{r}_2(x)$. Then, by (22) and (23),

$$|v(x,z)| = |\phi(x, m(x+z)) - \phi(x, m(x)) - D_\phi(x, m(x+z)) - z^2\tilde{r}_2(x)| \le C|z|^3(1+|z|^{q^*}).$$

Moreover,

$$|v(x,z) - z^3(\tilde{r}_3(x) + \sigma\tilde{p}_3(x))| \le C|z|^4(1+|z|^{q^*}).$$

As above, if $|z| \le \alpha/w$,

$$|e^{i\omega v(x,wz)} - 1 - wz^3c_2(x)| \le |i\omega v(x,wz) - wz^3c_2(x)| + \frac{1}{2}|\omega v|^2 e^{|\omega v|}$$

$$\le C\omega|wz|^4(1+|wz|^{q^*}) + \frac{1}{2}|C\omega|wz|^3(1+|wz|^{q^*})|^2 e^{C\omega|wz|^3(1+|wz|^{q^*})}$$

$$\le Cw^2|z|^4(1+\alpha^{q^*}) + Cw^2|z|^6(1+\alpha^{q^*})^2 e^{Cz^2\alpha(1+\alpha^{q^*})} \le Cw^2 e^{z^2/4},$$

for small enough α. The estimate for $|R_C|$ follows. It remains to note that, since $\phi_s(x,m(x)) = \Im\phi_x(x,m(x)) \equiv 0$,

$$\Im\tilde{r}_2 = \frac{1}{2}m'(x)^2\Im\phi_{ss} = \frac{1}{2}m'(x)\Im\left(\frac{d}{dx}\phi_s - \phi_{sx}\right) = -\frac{1}{2}\Im\left(\frac{d}{dx}\phi_x - \phi_{xx}\right) = \frac{1}{2}\Im\phi_{xx},$$

which shows that $i\omega w^2 z^2\tilde{r}_2 = iz^2c_1 - z^2/2$. □

We Taylor expand the remaining quantities in (21) and use the assumption (A5) to get

$$\begin{aligned}\varphi(wz) &\approx 1,\\ A(x,m(x+wz)) &\approx \tilde{A}(x,z) := A(x,m(x)) + wz\tilde{b}_1(x),\\ m'(x+wz) &\approx \tilde{m}(x,z) := m'(x) + wzm''(x).\end{aligned}$$

It is easy to see that the residual terms for these approximations can all be bounded by Cw^2z^2. Since these residual terms as well as R_A and R_B above all grow slower than $\exp(z^2/4)$, we can replace the terms in the integral in (21) by their approximations and control the error by $O(w^{q+3})$,

$$\left|\hat{f}(0) - \frac{w}{h}w^{q+1}\int_{|z|\le\frac{\alpha}{w}}(\tilde{D}_A + \tilde{A}\tilde{B})\tilde{C}\tilde{m}dz\right| \le C\frac{w}{h}w^{q+3}.$$

Introduce the functions

$$d_p(x) = \int z^p e^{iz^2c_1(x) - z^2/2}dz = \begin{cases} N_p(1-2ic_1(x))^{-(p+1)/2}, & p \text{ even},\\ 0, & p \text{ odd},\end{cases} \tag{24}$$

with N_p being a constant. We note that $d_p(x) \equiv 0$ when p is odd and it is bounded in x when p is even. Therefore, since the leading order z in $\tilde{D}_A$ and $\tilde{B}$ is $q+1$ and $q+3$ respectively, when q is even the leading order term vanishes. Thus, when q is even we get

$$\left|\hat{f}(0) - \frac{w}{h}w^{q+2}e_{\text{even}}(x)\right| \le C\frac{w}{h}w^{q+3},$$

where

$$e_{\text{even}}(x) = a_1 m'' d_{q+2} + a_1 c_2 m' d_{q+4} + a_2 m' d_{q+2} + b_1 A m'' d_{q+4} + b_1 A c_2 m' d_{q+6} + \quad (25)$$
$$b_1 \tilde{b}_1 m' d_{q+4} + b_2 A m' d_{q+4} + \sigma \frac{b_1^2}{2} A m' d_{2q+6}.$$

When q is odd,

$$\left| \hat{f}(0) - \frac{w}{h} w^{q+1} e_{\text{odd}}(x) \right| \le C \frac{w}{h} w^{q+2},$$

where

$$e_{\text{odd}}(x) = (a_1 + A b_1) m' d_{q+3}.$$

We therefore write

$$\left| \hat{f}(0) - C^*(x) \frac{w^{q^*+1}}{h} \right| \le C \frac{w^{q^*+2}}{h}, \quad q^* = \begin{cases} q+2, & q \text{ even}, \\ q+1, & q \text{ odd}, \end{cases} \quad C^*(x) = \begin{cases} e_{\text{even}}, & q \text{ even}, \\ e_{\text{odd}}, & q \text{ odd}. \end{cases} \quad (26)$$

Note that $C^*(x)$ is independent of ω and h and can be bounded by a constant independent of x. We then have

$$\left| E_{\text{non-osc}} - \alpha C^*(x) \frac{w^{q^*+1}}{w_0} \right| \le C \frac{w^{q^*+2}}{w_0}. \quad (27)$$

Therefore, the leading order term of the error $E_{\text{non-osc}}$ in ω is $\alpha C^*(x)\, w^{q^*+1}/w_0$.

4.2 Oscillatory case

In the oscillatory case we need to show that the functions in Lemma 2 also are smooth, with bounded derivatives. Then the non-stationary phase lemma can be used to bound $\hat{f}(k)$ since the phase derivative $m'(x)$ never vanishes.

We need

Lemma 3. *Under assumptions (A1) and (A2), for $0 \le k \le p$ and $|z| \le \alpha/w$ with small enough α,*

$$\left| \frac{d^k}{dz^k} D_A(x, m(x+wz)) \right| \le C w^{q+1} (1 + |z|^{q+1}), \quad (28)$$

$$\left| \frac{d^k}{dz^k} (e^{i\omega D_\phi(x, m(x+wz))} - 1) \right| \le C' w^{q+1} e^{z^2/7}, \quad (29)$$

$$\left| \frac{d^k}{dz^k} e^{i\omega \tilde{\phi}_q(x, m(x+wz))} \right| \le C'' e^{-z^2/5}. \quad (30)$$

The constants C, C' and C'' are independent of of k, x, ω and z.

Proof. For $k = 0$ the inequalities follow in the same way as in the proof of Lemma 2. Now consider $1 \le k \le p$. Since $\tilde{A}_q$ is q-th order Taylor expansion of $A(x+z-x', m(x+z))$ around $x' = 0$, we can write

$$D_A(x, m(x+z)) = \frac{(-1)^{q+1}}{q!} \int_0^z \partial_x^{q+1} A(x+t, m(x+z)) t^q dt.$$

Consequently, for $k > 0$,

$$\frac{d^k}{dz^k} D_A(x, m(x+z)) = \frac{(-1)^{q+1}}{q!} \int_0^z \partial_x^{q+1} \frac{d^k}{dz^k} A(x+t, m(x+z)) t^q dt + \frac{d^{k-1}}{dz^{k-1}} \left(\tilde{a}_{q+1}(x+z) z^q \right),$$

and therefore, by (A1), for $k \geq 0$,

$$\left| \frac{d^k}{dz^k} D_A(x, m(x+z)) \right| \leq C(|z|^{q+1} + |z|^{\max(q+1-k,0)}),$$

and (28) follows since $w^k(|wz|^{q+1} + |wz|^{\max(q+1-k,0)}) \leq w^{q+1}(1 + |z|^{q+1})$.

In the same way as for D_A, and using (A2),

$$\begin{aligned} \left| i\omega \frac{d^k}{dz^k} D_\phi(x, m(x+wz)) \right| &\leq C\omega w^k (|wz|^{q+3} + |wz|^{\max(q+3-k,0)})(1 + |wz|^{q^*}) \\ &\leq C w^{q+1}(1 + |z|^{q+3}), \end{aligned}$$

if $|z| \leq \alpha/w$. By Lemma 1 with $f(g_w) = e^{g_w}$ and $g_w = i\omega D_\phi(x, m(x+wz))$, using (19)

$$\left| \frac{d^k}{dz^k} e^{i\omega D_\phi(x, m(x+wz))} \right| \leq C w^{q+1}(1 + |z|^{k(q+3)}) |e^{i\omega D_\phi}| \leq C' w^{q+1} e^{z^2/7}.$$

Therefore (29) follows for $1 \leq k \leq p$.

We now write

$$\tilde{\phi}_q(x, m(x+wz)) = \sum_{j=0}^{q+2} \tilde{p}_j(x+wz) \, (wz)^j.$$

Then, since $\tilde{p}_0' + \tilde{p}_1 \equiv 0$ by (13), we have

$$\frac{d}{dz} \tilde{\phi}_q(x, m(x+wz)) = w^2 \tilde{p}_1' z + \sum_{j=2}^{q+2} \left(w^{j+1} z^j \tilde{p}_j'(x+wz) + j w^j z^{j-1} \tilde{p}_j(x+wz) \right),$$

and therefore

$$\frac{d}{dz} \left(i\omega \tilde{\phi}_q(x, m(x+wz)) \right) = \frac{i}{\Im \phi_{xx}(x, m(x))} \sum_{j=1}^{q+2} \gamma_j(x+wz) \, w^{j-1} z^j,$$

where

$$\gamma_j := \tilde{p}_j' + (j+1)\tilde{p}_{j+1}, \quad 1 \leq j \leq q+1, \qquad \gamma_{q+2} := \tilde{p}_{q+2}'.$$

Since the phase derivatives are evaluated on a center beam, $\gamma_j \in C_b^p$ are bounded, for $0 \leq k \leq p$, uniformly in x and we therefore have

$$\left| \frac{d^k}{dz^k} \left(i\omega \tilde{\phi}_q(x, m(x+wz)) \right) \right| \leq C_k(1 + |z|^{q+2}), \qquad 1 \leq k \leq p.$$

Thus, by Lemma 1 with $f(g_w) = e^{g_w}$ and $g_w = i\omega \tilde{\phi}_q(x, m(x+wz))$, using (19), the inequality (30) follows for $1 \leq k \leq p$. This completes the proof. □

The remaining terms in (21), i.e. $A(x, m(x+wz))$, $\varphi(wz)$ and $m'(x+wx)$, are all assumed to be smooth with derivatives of order up to p bounded uniformly in x by the assumptions (A1) and (A5). Since the growth in (28) and (29) is offset by the rapid decay in (30), the above Lemma shows that all z-derivatives of the integrand,

$$g(x,z) := \varphi \left(D_A + A(e^{i\omega D_\phi} - 1)\right) e^{i\omega\tilde{\phi}_q} m',$$

up to order p belongs to L_1 and $||\partial_z^k g(x,\cdot)|| \leq C_k w^{q+1}$ for $0 \leq k \leq p$. The constants C_k are independent of x and ω. We can then use the following version of the non-stationary phase lemma.

Lemma 4. *Suppose $\psi(z) \in C^{p+1}(\mathbb{R})$ with $\psi'(z) \in C_b^p(\mathbb{R})$ and $\psi'(z) \geq c_0 > 0$. Moreover, let $\epsilon < \delta < 1$ and suppose $g(z) \in W^{p,1}$. Then*

$$\left| \int g(z) e^{-i\psi(\delta z)/\varepsilon} dz \right| \leq C||g||_{W^{p,1}} \left(\frac{\varepsilon}{\delta}\right)^p, \tag{31}$$

where C depends on $\psi(x)$ and p, but not on $g(z)$, δ and ε.

Proof. For the proof we refer to [13]. It is an easy adaptation of that proof of theorem 7.7.1. □

Taking ψ as $2\pi m(x+\cdot)$, δ as w and ε as h/k we can apply this to (21),

$$|\hat{f}(k)| = \frac{w}{h} \left| \int g(x,z) e^{-2\pi i m(x+wz)k/h} dz \right| \leq C \frac{w}{h} ||g(x,\cdot)||_{W^{p,1}} \left(\frac{h}{kw}\right)^p.$$

Consequently,

$$\left| \sum_{k\neq 0} \hat{f}(k) \right| \leq C \frac{w}{h} ||g(x,\cdot)||_{W^{p,1}} \sum_{k\neq 0} \left(\frac{h}{kw}\right)^p \leq C \frac{w}{h} w^{q+1} \left(\frac{h}{w}\right)^p \sum_{k=1}^{\infty} k^{-p} \leq C' \frac{w}{h} w^{q+1} \left(\frac{h}{w}\right)^p.$$

Thus since by the assumptions (A1) $p \leq 2$,

$$|E_{\text{osc}}| = \frac{\alpha h}{w_0} |\sum_{k\neq 0} \hat{f}(k)| \leq C' \frac{w^{q+2}}{w_0} \left(\frac{h}{w}\right)^p.$$

Together with (27) this shows the theorem.

5 Constant coefficient equations

It is often claimed that the beam width is important in the accuracy of Gaussian beams, because for wide beams the Taylor expansion error should be large. See for example [4, 6]. We therefore in this section consider the constant coefficient Helmholtz equation, with the speed of propagation $c(\boldsymbol{x}) \equiv 1$, for which exact Gaussian beam solutions and the dominant part of Taylor expansion error $|E_{\text{non-osc}}|$ can be computed. We investigate the importance of the beam width on Taylor error in this particular

case. Our conclusion is that the local beam width is not a good indicator of accuracy, and there is no direct relation between the error and the beams width.

We let $q = 0$ and consider the source $\boldsymbol{x}_0(s) = (s, y_0(s))$ and assume all beams originating from $\boldsymbol{x}_0$ shoot out orthogonally. Therefore $\theta_0(s) = \frac{\pi}{2} + \tan^{-1}(y_0'(s))$. In the constant coefficient case $c \equiv 1$, for a central ray Ω with $x(0) = x_0(s) = s$, $y(0) = y_0(s)$ and $\theta(0) = \theta_0(s)$, we get from (4) at $y = y^*$,

$$\theta(t(s)) = \frac{\pi}{2} + \tan^{-1}(y_0'(s)), \tag{32}$$

$$x(t(s)) = X(s) = s - y_0'(s)\,(y^* - y_0(s)), \tag{33}$$

$$t(s) = \left((X(s) - s)^2 + (y^* - y_0(s))^2\right)^{1/2}, \tag{34}$$

which implies that the central ray is a straight line.

Here, we will only compute the error at $\boldsymbol{x} = (0, y^*)$. For this point, let $s^* := m(0) = X^{-1}(0)$. To simplify the calculations, and without loss of generality, we assume $y_0(s^*) = y_0'(s^*) = 0$. Therefore, by (32)-(34), the central ray starting at $\boldsymbol{x}_0(s^*)$ will lie on the y-axis, and we have $s^* = X(s^*) = 0$ and $t(s^*) = y^*$. See Figure 4.

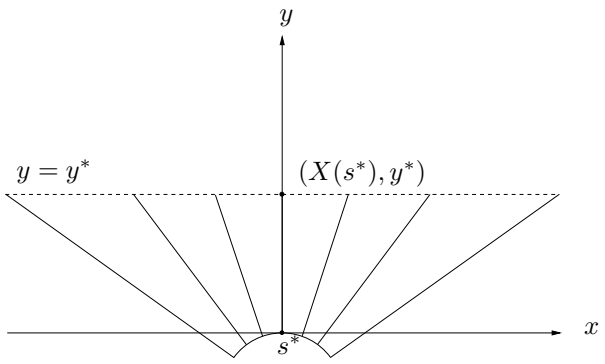


Figure 4: A schematic representation of the initial source and central beam rays which are straight lines.

Assuming the initial phase on $\boldsymbol{x}_0(s)$ to be zero, $\phi(\boldsymbol{x}_0) = 0$, we also get

$$\phi(X(s), s) = t(s). \tag{35}$$

To obtain ODEs for higher order Taylor coefficients, we introduce the orthogonal ray-centered coordinates t, n, where n is the axis perpendicular to the ray at point t with the origin on the ray. In this coordinate system, $\phi(t, n = 0)$ and $A(t, n = 0)$ correspond to $\phi(X(s), s)$ and $A(X(s), s)$ in the Cartesian coordinate, respectively. The eikonal equation and transport equation in the ray-centered coordinates read

$$\phi_t^2 + \phi_n^2 = 1, \tag{36}$$

$$2\nabla A \cdot \nabla\phi + A\Delta\phi = 0, \qquad \nabla\phi = (\phi_t \ \phi_n)^\top \tag{37}$$

Set $\phi^{(j)}(t) := \partial_n^j \phi(t, n=0)$ and $A^{(j)}(t) := \partial_n^j A(t, n=0)$, with $j = 0, 1, 2, \ldots$. We first note that by (35)

$$\phi^{(0)}(t) = t, \qquad \partial_t \phi(t, n=0) = 1, \qquad \partial_t^j \phi(t, n=0) = 0, \quad j = 2, 3, \ldots .$$

Moreover, by (36) and (37) and taking several of their derivatives with respect to t and n,

$$\begin{gathered}
\phi^{(1)}(t) = 0, \quad \partial_t \partial_n \phi(t, n=0) = 0, \quad \partial_t \partial_n^2 \phi(t, n=0) = -{\phi^{(2)}}^2(t), \\
\partial_t \partial_n^3 \phi(t, n=0) = 0, \quad \partial_t^2 \partial_n \phi(t, n=0) = 0, \quad \partial_t^3 \partial_n \phi(t, n=0) = 0, \\
\partial_t^2 \partial_n^2 \phi(t, n=0) = 2{\phi^{(2)}}^3(t), \quad \partial_t A(t, n=0) = -\frac{1}{2} A^{(0)}(t)\, \phi^{(2)}(t), \\
\partial_t^2 A(t, n=0) = \frac{3}{4} A^{(0)}(t)\, {\phi^{(2)}}^2(t), \quad \partial_t \partial_n A(t, n=0) = 0.
\end{gathered}$$

Now, let

$$\phi(t, n) \approx t + \frac{n^2}{2}\phi^{(2)}(t) + \frac{n^3}{6}\phi^{(3)}(t) + \frac{n^4}{24}\phi^{(4)}(t), \tag{38}$$

and

$$A(t, n) \approx A^{(0)}(t) + nA^{(1)}(t) + \frac{n^2}{2}A^{(2)}(t). \tag{39}$$

Putting (38) and (39) into (36) and (37), we obtain the following ODEs for Taylor coefficients,

$$\frac{d}{dt}\phi^{(2)} + {\phi^{(2)}}^2 = 0, \tag{40}$$

$$\frac{d}{dt}\phi^{(3)} + 3\phi^{(2)}\phi^{(3)} = 0, \tag{41}$$

$$\frac{d}{dt}\phi^{(4)} + 4\phi^{(2)}\phi^{(4)} + 3{\phi^{(2)}}^4 + 3{\phi^{(3)}}^2 = 0, \tag{42}$$

$$\frac{d}{dt}A^{(0)} + \frac{1}{2}\phi^{(2)}A^{(0)} = 0, \tag{43}$$

$$\frac{d}{dt}A^{(1)} + \frac{3}{2}\phi^{(2)}A^{(1)} + \frac{1}{2}\phi^{(3)}A^{(0)} = 0, \tag{44}$$

$$\frac{d}{dt}A^{(2)} + \frac{5}{2}\phi^{(2)}A^{(2)} + 2\phi^{(3)}A^{(1)} + \frac{1}{2}\phi^{(4)}A^{(0)} + \frac{3}{2}{\phi^{(2)}}^3 A^{(0)} = 0 \tag{45}$$

Setting $\phi^{(2)}(t) = P(t)/Q(t)$, the nonlinear Ricatti equation (40) can be reduced to the system of linear ODEs (7-8),

$$\frac{d}{dt}Q = P, \qquad \frac{d}{dt}P = 0.$$

Therefore, with $P(0) = P_0$ and $Q(0) = Q_0$, we obtain

$$\phi^{(2)}(t) = \frac{P_0}{Q_0 + P_0 t}.$$

Moreover, the equation (43) gives us

$$A^{(0)}(t) = \frac{Q_0^{1/2}}{(Q_0 + P_0 t)^{1/2}},$$

Note that we set $A^{(0)}(0) = 1$ motivated by (10).

The rest of ODEs are linear first order equations. With zero initial conditions, we get

$$\begin{aligned}
\phi^{(3)}(t) &= 0, \\
\phi^{(4)}(t) &= -3\frac{P_0^4 t}{(Q_0 + P_0 t)^4}, \\
A^{(1)}(t) &= 0, \\
A^{(2)}(t) &= -\frac{3}{2}\,\frac{Q_0^{1/2}\, P_0^3 t}{(Q_0 + P_0 t)^{7/2}}.
\end{aligned}$$

For a function $f(x, s) = F(t, n)$ in two different coordinates, we have

$$\partial_x^j f = \sum_{i=0}^{j} c_{i,j}\, \partial_t^{j-i} \partial_n^i F\, \sin^{j-i}\theta\, \cos^i\theta, \tag{46}$$

where θ is the angle between x-axis and t-axis, and $c_{i,j}$ are binomial coefficients. Therefore

$$\partial_x^1 \phi(X(s), s) = \cos\theta(t(s)), \tag{47}$$

$$\partial_x^2 \phi(X(s), s) = \phi^{(2)}(t(s))\, \sin^2\theta(t(s)), \tag{48}$$

$$\partial_x^3 \phi(X(s), s) = -3\, {\phi^{(2)}}^2(t(s))\, \sin^2\theta(t(s))\, \cos\theta(t(s)), \tag{49}$$

$$\partial_x^4 \phi(X(s), s) = \phi^{(4)}(t(s))\, \sin^4\theta(t(s)) + 12\, {\phi^{(2)}}^3(t(s))\, \sin^2\theta(t(s))\, \cos^2\theta(t(s)), \tag{50}$$

$$\partial_x^1 A(X(s), s) = -\frac{1}{2}\, A^{(0)}(t(s))\, \phi^{(2)}(t(s))\, \cos\theta(t(s)), \tag{51}$$

$$\partial_x^2 A(X(s), s) = A^{(2)}(t(s))\, \sin^2\theta(t(s)) + \frac{3}{4}\, A^{(0)}(t(s))\, {\phi^{(2)}}^2(t(s))\, \cos^2\theta(t(s)). \tag{52}$$

At the point $\boldsymbol{x} = (0, y^*)$, where $s = s^* = 0$, we have $\cos\theta(y^*) = 0$ and $\sin\theta(y^*) = 1$. In fact, at this point the n-axis is parallel to the x-axis, and therefore $\partial_x^j = \partial_n^j$.

Therefore,

$$a_1(0) = -\partial_x^1 A(0, 0) = 0, \quad b_1(0) = -\frac{i}{6}\frac{\partial_x^3 \phi(0, 0)}{\Im \partial_x^2 \phi(0, 0)} = 0.$$

Thus, e_{even} in (25) simplifies to

$$e_{\text{even}}(0) = m'(0)\, a_2(0)\, d_2(0) + m'(0)\, A(0, s^*)\, b_2(0)\, d_4(0),$$

and we therefore need only to calculate a_2, b_2 and c_1.

Differentiating (13), (47), (49) and (51) with respect to s, we obtain

$$\phi_{ss}(0,0) = X'^2(0)\phi_{xx}(0,0) + X'(0)\,\frac{d}{ds}\theta(y^*),$$
$$\phi_{xxxs}(0,0) = -X'(0)\phi_{xxxx}(0,0) + 3\frac{P_0^2}{(Q_0+P_0y^*)^2}\,\frac{d}{ds}\theta(y^*),$$
$$A_{xs}(0,0) = -X'(0)A_{xx}(0,0) + \frac{1}{2}\frac{Q_0^{1/2}\,P_0}{(Q_0+P_0y^*)^{3/2}}\,\frac{d}{ds}\theta(y^*).$$

Moreover, by (13),

$$\tilde{r}_2(x) = \frac{1}{2}\,\frac{d^2}{dz^2}\phi(x, m(x+z))\Big|_{z=0} = \frac{1}{2}m''\phi_s + \frac{1}{2}m'^2\phi_{ss} = \frac{1}{2}m'^2\phi_{ss}.$$

Therefore, after some algebraic manipulation, we obtain

$$a_2(0) = -\frac{3Q_0^{1/2}\,P_0^3y^* + 2Q_0^{1/2}\,P_0(Q_0+P_0y^*)^2m'(0)\frac{d}{ds}\theta(y^*)}{4(Q_0+P_0y^*)^{7/2}},$$
$$b_2(0) = -i\frac{P_0^4y^* + 4P_0^2(Q_0+P_0y^*)^2m'(0)\frac{d}{ds}\theta(y^*)}{8\Im(\frac{P_0}{Q_0+P_0y^*})\,(Q_0+P_0y^*)^4},$$
$$c_1(0) = \frac{\Re(\frac{P_0}{Q_0+P_0y^*}) + m'(0)\frac{d}{ds}\theta(y^*)}{2\Im(\frac{P_0}{Q_0+P_0y^*})}.$$

Assuming $P_0 = i$, $\Im Q_0 = 0$, $\Re Q_0 > 0$, we have

$$a_2(0) = i\frac{3Q_0^{1/2}\,y^* - 2Q_0^{1/2}\,(Q_0+iy^*)^2m'(0)\frac{d}{ds}\theta(y^*)}{4(Q_0+iy^*)^{7/2}}, \tag{53}$$
$$b_2(0) = i\frac{(Q_0^2+y^{*2})\left(-y^* + 4(Q_0+iy^*)^2m'(0)\frac{d}{ds}\theta(y^*)\right)}{8Q_0(Q_0+iy^*)^4}, \tag{54}$$
$$c_1(0) = \frac{y^* + (Q_0^2+y^{*2})m'(0)\frac{d}{ds}\theta(y^*)}{2Q_0}, \tag{55}$$

and

$$A(0,0) = \frac{Q_0^{1/2}}{(Q_0+iy^*)^{1/2}}. \tag{56}$$

Note that by (32-34),

$$\frac{d}{ds}\theta(y^*) = y_0''(0), \quad m'(0) = (X^{-1})'(0) = (1-y^*y_0''(0))^{-1}. \tag{57}$$

Therefore, knowing $y_0(s)$ and by (53-57) and (24), we can calculate $e_{\text{even}}(0)$. Note that $e_{\text{even}}(0)$ only depends on Q_0, y^* and $y_0''(0)$.

We consider the following two cases:

(1) $y_0''(0) = 0$,

(2) $y_0''(0) = -1$.

The first case corresponds to a line $y_0 = 0$. The second case corresponds to a circle $y_0(s) = -1 + \sqrt{1 - s^2}$ or a parabola $y_0(s) = -s^2/2$. Note that with an initial curve with positive second derivative, the rays will intersect and form caustic, and then our theory does not hold.

For the first case, we obtain the simple expression

$$e^{\mathrm{l}}_{\mathrm{even}}(0) = \frac{n_0\, y^*\, Q_0^2}{(Q_0 + iy^*)^2\, (Q_0^2 + y^{*2})^{3/2}}, \tag{58}$$

and for the second case, assuming $1 + y^* \approx y^*$, i.e. for large distances from the source, we have

$$e^{\mathrm{c}}_{\mathrm{even}}(0) \approx \frac{n_1\, Q_0^3 + n_2\, Q_0^2\, y^* + n_3\, Q_0\, y^{*2} + n_4\, y^{*3}}{Q_0^{1/2}\, y^{*1/2}\, (Q_0 + iy^*)^2\, (Q_0^2 + y^{*2})^{3/2}}, \tag{59}$$

where n_j, with $j = 0, 1, \ldots 4$, are constant complex numbers.

Now since

$$w(0) = \left(\frac{Q_0^2 + y^{*2}}{\omega\, Q_0}\right)^{1/2}, \qquad w_0(0) = \left(\frac{Q_0}{\omega}\right)^{1/2},$$

and the amplitude of the geometrical optics solution is proportional to $|1 - y^*\, y_0''(0)|^{-1/2}$, by (27), the relative error will be

$$|E_{\mathrm{rel}}| = |E_{\mathrm{non-osc}}|\, |1 - y^*\, y_0''(0)|^{1/2} = \frac{w^3(0)}{w_0(0)}\, |e_{\mathrm{even}}(0)|\, |1 - y^*\, y_0''(0)|^{1/2}.$$

We therefore obtain

$$|E^{\mathrm{l}}_{\mathrm{rel}}| = \left|\frac{n_0\, y^*}{\omega\, (Q_0 + iy^*)^2}\right|,$$

and

$$|E^{\mathrm{c}}_{\mathrm{rel}}| \approx \left|\frac{n_1\, Q_0^3 + n_2\, Q_0^2\, y^* + n_3\, Q_0\, y^{*2} + n_4\, y^{*3}}{\omega\, Q_0^{3/2}\, (Q_0 + iy^*)^2}\right|,$$

corresponding to (58) and (59), respectively.

Figure 5 shows the absolute values of the relative errors at $y^* = 3$. Note that here, $|E^{\mathrm{c}}_{\mathrm{rel}}|$ is calculated exactly, without the assumption $1 + y^* \approx y^*$.

As it can be seen from the formulas and figures, the relative error has a direct relation with Q_0, but not with the beam width w. It decreases as Q_0 increases. This result has also been noticed in [9] for the oscillatory part of the error (or the discretization error).

In many approximations, the optimal Q_0, corresponding to the minimum beam width at a receiver point, is chosen for computations, see [4] for instance. Although using this Q_0, we do not obtain the minimum error, but importantly the error does

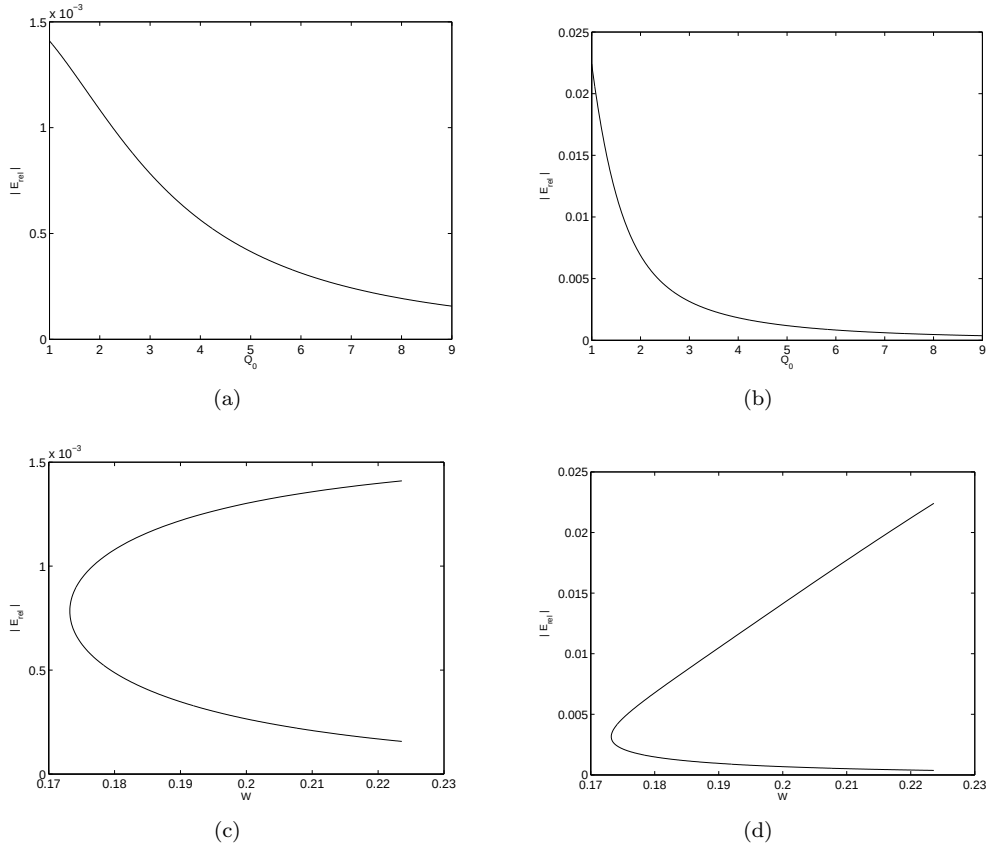


Figure 5: Upper and lower left figures show the absolute value of relative error as a function of Q_0 and the width w, in the case when the initial source is a line, respectively. Upper and lower right figures show the same variables when the initial source is a circle.

not increase as the distance from the source increases. Note that the minimum width is attained at $Q_0 = y^*$, and therefore

$$|E^{\mathrm{l}}_{\mathrm{rel}}| = \frac{N}{\omega\, y^*}, \qquad |E^{\mathrm{c}}_{\mathrm{rel}}| \approx \frac{N'}{\omega\, y^{*1/2}},$$

with N and N' being constant numbers. Moreover, in the general case of non-constant coefficient equations, where the rays can bend, it may not be possible to have very wide beams, since as was noted before, the Gaussian beam approximation may break down when the phase becomes non-smooth, and this happens at some distance away from the central ray (outside *the regularity region*). Also, in the presence of a varying speed of propagation where the properties may change dramatically as we get farther to the central rays, the Taylor expansion error can be large for wide beams. In this case therefore, Q_0 corresponding to the minimum beam width may be a proper choice.

Figure 6 shows the beam width as a function of Q_0. Note that for $Q_0 > y^*$, the width will increases, and therefore selecting a very large Q_0 results in having a very wide beam.

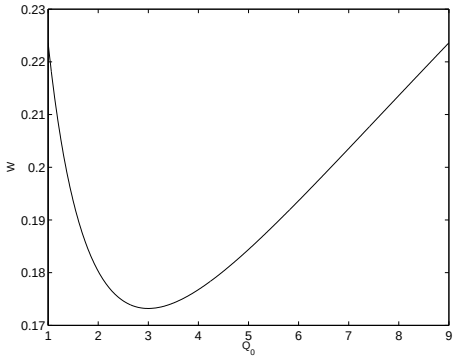


Figure 6: The beam width as a function of Q_0 at $y^* = 3$.

References

[1] M. M. Popov, *A New Method of Computation of Wave Fields Using Gaussian Beams*, Wave Motion, vol. 4, 1982, 85-97.

[2] V. M. Babic and T. F. Pankratova, *On Discontinuities of Green's Function of the Wave Equation with Variable Coefficient*, Problemy Matem. Fiziki, Leningrad University, Saint-Petersburg, vol. 6, 1973.

[3] A. P. Katchalov and M. M. Popov, *Application of the Method of Summation of Gaussian Beams for Calculation of High-frequency Wave Fields*, Sov. Phys. Dokl., vol. 26, 1981, 604-606.

[4] V. Cerveny, M. M. Popov and I. Psencik, *Computation of Wave Fields in Inhomogeneous Media - Gaussian Beam Approach*, Geophys. J. R. Astr. Soc., vol. 70, 1982, 109-128.

[5] L. Klimes, *Expansion of a High-frequency Time-harmonic Wavefield Given on an Initial Surface into Gaussian Beams*, Geophys. J. R. astr. Soc., vol. 79, 1984, 105-118.

[6] N. R. Hill, *Gaussian Beam Migration*, Geophysics, vol. 55, no. 11, 1990, 1416-1428.

[7] N. R. Hill, *Prestack Gaussian-Beam Depth Migration*, Geophysics, vol. 66, no. 4, 2001, 1240-1250.

[8] J. Ralston, *Gaussian Beams and the Propagation of Singularities*, In Studies in partial differential equations, Math. Assoc. America, Washington, DC, vol. 23, 1982, 206-248.

[9] L. Klimes, *Discretization Error for the Superposition of Gaussian Beams*, Geophys. J. R. astr. Soc., vol. 86, 1986, 531-551.

[10] N. M. Tanushev, *Superpositions and Higher Order Gaussian Beams*, Preprint, 2008.

[11] R. Magnanini and G. Talenti, *On Complex-Valued Solutions to a 2-D Eikonal Equation. I. Qualitative Properties*, Contemporary Mathematics, vol. 238, 1999, 203-229

[12] R. Magnanini and G. Talenti, *On Complex-Valued Solutions to a Two-Dimensional Eikonal Equation. II. Existence Theorems*, SIAM Journal on Mathematical Analysis, vol. 34, no. 4, 2003, 805-835.

[13] L. Hörmander, *The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*, Springer-Verlag, Berlin Heidelberg New York, 1983.

Paper IV

A Wave Front-Based Gaussian Beam Method for Computing High Frequency Waves

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

April 19, 2008

Abstract. We present a wave front method based on Gaussian beams for computing high-frequency wave propagation problems. Unlike standard geometrical optics, Gaussian beams compute the correct solution of the wave field also at caustics. The method tracks a front of Gaussian beams with two particular initial values for width and curvature which allows the direct recreation of any other beam propagating from the initial front into the medium. This is used to approximate the field with different, optimally chosen, beams in different points on the front. The performance of the method is illustrated with two numerical examples.

Keywords. wave propagation, high frequency, asymptotic approximation, summation of Gaussian beams, wavefront methods

1 Introduction

The Gaussian beam method is an asymptotic method for computing high-frequency wave fields in smoothly varying inhomogeneous media. It was proposed by Popov [1], based on an earlier work of Babic and Pankratova [2]. The method was first applied by Katchalov and Popov [3] and Cerveny et al. [4] to describe high-frequency seismic wave fields by the summation of paraxial Gaussian beams. It was later applied to seismic migration by Hill [5, 6]. For a rigorous mathematical analysis of Gaussian beams we refer to [7]. The main advantage of this method is that Gaussian beams give the correct solution also at caustics where standard geometrical optics breaks down.

In the Gaussian beam method, the initial/boundary condition for the wave field is decomposed into initial conditions for Gaussian beams. Individual Gaussian beams are computed by ray tracing, where quantities such as the curvature and width of beams are calculated from ordinary differential equations (ODEs) along the central ray of the beams. The contributions of the beams concentrated close to their central rays are determined by Taylor expansion. The wave field at a receiver is then obtained as a weighted superposition of the Gaussian beams situated close to the receiver.

It is also possible to design an Eulerian Gaussian beam summation method. In [8], the problem is formulated by Liouville-type equations in phase space giving uniformly distributed Eulerian traveltimes and amplitudes for multiple sources.

In this paper, we consider the Lagrangian formulation and present a wave front method for computing Gaussian beams. Wave front methods have been very successful

for standard geometrical optics as they provide a simple mechanism for controlling the resolution and accuracy of the numerical approximation [9]. Using them with Gaussian beams is not as straightforward since the beam method strongly depends on the distribution and width of the beams at the initial front. We show how one can use two canonical beams in the wave front method and from these afterward recreate any other beam. This is used to approximate the field with different, optimally chosen, beams in different points on the front. We present numerical examples to verify the efficiency and accuracy of the algorithm.

2 Gaussian beam equations

Gaussian beams are asymptotic solutions of linear wave equations. In the beam summation method, the initial/boundary data are decomposed into Gaussian beams. Individual beams are computed from ordinary differential equations along their central rays. The contribution of each beam close to its central ray is calculated by Taylor expansion. The wave field is then obtained by summing over the beams.

In this section, we review the governing equations for computing Gaussian beams and formulate the beam summation method. We consider the reduced wave equation in the frequency domain (Helmholtz equation) in a two-dimensional space,

$$\Delta u(\boldsymbol{x}) + \frac{\omega^2}{c(\boldsymbol{x})^2}\, u(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} = (x, y) \in \mathbb{R}^2, \tag{1}$$

where $\omega \gg 1$ is the angular frequency and $c(\boldsymbol{x})$ is the speed of propagation. A Gaussian beam as an approximate high frequency solution to (1) is written in the form,

$$u_{\mathrm{GB}}(\boldsymbol{x}) = A(\boldsymbol{x})\, e^{i\omega\phi(\boldsymbol{x})}, \tag{2}$$

where the amplitude function A and phase function ϕ are independent of ω. Note that (2) is in fact the first term of the WKBJ expansion, known as *geometrical optics term* and is of order $\mathcal{O}(\omega^{-1})$. However, unlike the geometrical optics, where these functions are real-valued, in Gaussian beam method they are complex-valued.

The beam central ray Ω is given by the *ray tracing system*

$$\begin{aligned}
\frac{dx}{dt} &= c(\boldsymbol{x})\, \cos\theta, \quad x(0) = x_0, \\
\frac{dy}{dt} &= c(\boldsymbol{x})\, \sin\theta, \quad y(0) = y_0, \\
\frac{d\theta}{dt} &= \frac{\partial c}{\partial x}(\boldsymbol{x})\, \sin\theta - \frac{\partial c}{\partial y}(\boldsymbol{x})\, \cos\theta, \quad \theta(0) = \theta_0,
\end{aligned} \tag{3}$$

where t is the real-valued travel-time (or the arc-length) along the ray, and θ is the angle between the tangent of the ray and the positive x-axis.

The complex-valued functions A and ϕ close to the central ray are approximated by

Taylor expansions around the ray,

$$A(\boldsymbol{x}) \approx A(\boldsymbol{x}^*), \tag{4}$$

$$\phi(\boldsymbol{x}) \approx \phi(\boldsymbol{x}^*) + (\boldsymbol{x}-\boldsymbol{x}^*)\cdot\nabla\phi(\boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^*)^\top D^2\phi(\boldsymbol{x}^*)\,(\boldsymbol{x}-\boldsymbol{x}^*), \tag{5}$$

where $\boldsymbol{x}^* = \boldsymbol{x}(t)$ for some t. The Taylor coefficients $\phi(\boldsymbol{x}^*)$, $\nabla\phi(\boldsymbol{x}^*)$, $D^2\phi(\boldsymbol{x}^*)$ and $A(\boldsymbol{x}^*)$ are computed only on the central ray,

$$\phi(\boldsymbol{x}^*) = \phi(\boldsymbol{x}(0)) + t, \qquad \nabla\phi(\boldsymbol{x}^*) = (\cos\theta\ \ \sin\theta)^\top / c(\boldsymbol{x}^*),$$

$$D^2\phi(\boldsymbol{x}(t)) = HNH^{-1}, \qquad A(\boldsymbol{x}^*) = (c(\boldsymbol{x}^*)/Q)^{1/2},$$

where

$$H = \begin{pmatrix} \sin\theta & \cos\theta \\ -\cos\theta & \sin\theta \end{pmatrix}, \quad N = \begin{pmatrix} P/Q & -c_1/c^2 \\ -c_1/c^2 & -c_2/c^2 \end{pmatrix}, \quad \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}\nabla c.$$

The complex-valued scalar functions P and Q satisfy the *dynamic ray tracing system*

$$\begin{aligned} \frac{dQ}{d\tau} &= c^2(\boldsymbol{x})\,P \qquad Q(0) = Q_0 \\ \frac{dP}{d\tau} &= -\frac{c_{xx}\,\sin^2\theta - 2c_{xy}\,\sin\theta\,\cos\theta + c_{yy}\,\cos^2\theta}{c(\boldsymbol{x})}\,Q, \quad P(0) = P_0. \end{aligned} \tag{6}$$

The quantities P and Q determine the wavefront curvature and the beam width.

A main difficulty in the Gaussian beam method is the choice of initial data P_0 and Q_0. These parameters are arbitrary and therefore the Gaussian beam solution is non-unique. However, despite this non-uniqueness, the summation of Gaussian beams is a high frequency asymptotic expansion of the wave field for all admissible values of these parameters. It can be shown that if we choose the admissible parameters satisfying

$$Q_0 \neq 0, \qquad \Im(P_0/Q_0) > 0, \tag{7}$$

then $Q(t) \neq 0$ and $\Im(P(t)/Q(t)) > 0$ along the central ray, [1]. The former guarantees the *regularity* of the Gaussian beam (with finite amplitudes at caustics), and the latter guarantees the *non-degeneracy* of the beam (concentration of the solution close to the ray).

It has been proposed that the optimal choice of the parameters produce Gaussian beams of minimum width at a receiver point, see [4, 10] for instance. The half-width of the Gaussian beam is given by

$$w(t) = \left(\frac{1}{2}\omega\, Im(P(t)/Q(t))\right)^{-1/2}.$$

The main motivation for this choice is that for wide beams the Taylor expansion error should be large. Moreover, from the computational point of view, it is more convenient to work with beams which are as narrow as possible, because in the case of variable

speed of propagation, where the central rays can bend, at some distance from the rays the phase may become non-smooth and therefore the Gaussian beam approximation may break down. However, it was shown in [11] that this choice will not necessarily give the minimum error. The optimal choice of the parameters should minimize the error and is still an open question.

Since the optimal parameters are different for different points along the beam central ray, we may need to solve (6) for different initial conditions P_0 and Q_0. This can computationally be very expensive. However, we take the advantage of linearity of (6) and make the following observation. We specify two real-valued canonical solutions (Q_1, P_1) and (Q_2, P_2) with two different sets of initial data

$$(Q_1, P_1)\,(0) = (1, 0), \quad (Q_2, P_2)\,(0) = (0, 1). \tag{8}$$

Then

$$Q = Q_0\,Q_1 + P_0\,Q_2, \qquad P = Q_0\,P_1 + P_0\,P_2, \tag{9}$$

is a complex-valued solution of (6) with the initial data P_0 and Q_0, [4]. Hence from two basis solutions, beams with all possible initial data can be computed by taking linear combinations at no extra cost. In particular, the geometrical optics solution can be obtained from (Q_1, P_1),

$$\phi_{GO}(\boldsymbol{x}(t)) = \phi(\boldsymbol{x}(0)) + t, \qquad A_{GO}(\boldsymbol{x}(t)) = A(\boldsymbol{x}(0)) \left(\frac{1}{Q_1(t)}\,\frac{c(\boldsymbol{x}(t))}{c(\boldsymbol{x}(0))}\right)^{1/2},$$

which corresponds to an infinitely wide beam.

Based on the two canonical solutions, a typical choice of initial parameters is

$$Q_0 = Q_0^{opt}(t) = \left|\frac{Q_2(t)}{Q_1(t)}\right|, \qquad P_0 = i, \tag{10}$$

where the real-valued Q_0 is optimally chosen to give the minimum half-width of the Gaussian beam

$$W^{min}(t) = 2\left(\frac{|Q_1(t)\,Q_2(t)|}{\omega}\right)^{1/2}.$$

In addition, letting Q_0 to be complex-valued with a positive real part will allow us to make the width arbitrarily small and give more control over the beam parameters.

Now, let the wave source be a curve $\boldsymbol{x}_0(s)$ in $\mathbb{R}^2$ parameterized by s. We introduce the notation $A(\boldsymbol{x}, s)$ and $\phi(\boldsymbol{x}, s)$ for the amplitude and phase of a beam with initial position $\boldsymbol{x}_0(s)$. We first decompose the initial/boundary condition for the wave field on $\boldsymbol{x}_0(s)$ into initial conditions for several beams with different initial positions $\boldsymbol{x}_0(s_j)$, see [12] for example. Individual Gaussian beams are computed by solving the above ODEs. The contributions of the beams concentrated close to their central rays are determined by the approximations (4,5) entered in (2). The wave field at a fixed receiver point $\boldsymbol{x}_R$ is then calculated by summing over the beams

$$u(\boldsymbol{x}_R) = \sum_{j\in\mathbb{Z}} \psi(s_j)\,A(\boldsymbol{x}_R, s_j)\,e^{i\omega\phi(\boldsymbol{x}_R, s_j)}. \tag{11}$$

The weights $\psi(s_j)$ and the initial conditions for the ODEs (3) and (6) are chosen such that u at $\boldsymbol{x}_0$ well approximates the exact initial/boundary data.

As an example, we show how to find the weights and the initial conditions when the wave field is generated by a plane wave at $\boldsymbol{x}_0(s) = (0, s)$ propagating into the domain orthogonally, i.e. $\theta_0(s) = 0$. We first note that a plane wave can be approximated by a sum of beams, [5],

$$u(0,s) = 1 = \sum_j \frac{1}{\sqrt{\pi}} \frac{h}{w_0} e^{-(s-s_j)^2/w_0^2} + \mathcal{O}(e^{-(w_0/h)^2}), \quad s_j = jh, \tag{12}$$

with h and w_0 representing the initial spacing of the beams and the initial beam half-widths, See Figure 1.

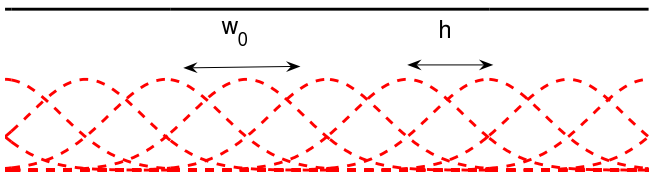


Figure 1: The sum of several Gaussian functions is almost constant. A plane wave can therefore be decomposed approximately to a sum of parallel Gaussian beams.

To properly choose the initial data, one must therefore take the parameters w_0 and h such that $w_0 > h$. Note that for computational efficiency, h should not be taken too small. The wave field (11) at $\boldsymbol{x}_0(s) = (0, s)$ is

$$u(0,s) = \sum_{j\in\mathbb{Z}} \psi_j \left(\frac{c(0,s_j)}{Q_0}\right)^{1/2} e^{\frac{i}{2}\omega(s-s_j)^2 \frac{P_0}{Q_0}}, \quad \psi_j := \psi(s_j). \tag{13}$$

Now, we choose a real-valued, positive Q_0 and $P_0 = i$ which satisfies the condition (7). Comparing to (12), the Gaussian beam solution (13) produces a plane wave if

$$w_0 = \left(\frac{2Q_0}{\omega}\right)^{1/2}, \qquad \psi_j = h\left(\frac{\omega}{2\pi c(0,s_j)}\right)^{1/2}.$$

3 Wavefront method

The usual way to compute high frequency wave fields by Gaussian beam summation is based on standard ray tracing, where the central rays of the beams are traced individually by solving the ODE systems (3) and (6). The main problem with ray tracing is that it may produce diverging rays that fail to cover the computational domain. In this case, one needs to increase the number of rays, which in turn increases the computational cost.

In standard geometrical optics, the problem of diverging rays can be overcomed by instead using so-called wave front methods, [13, 14]. They are related to ray tracing, but instead of tracing a sequence of individual rays, a wave front is evolved in physical

or phase space according to the ODE formulations. In physical space, a wave front at a travel-time $t \geq 0$ is a curve $\{\boldsymbol{x}(t,s) | \phi(\boldsymbol{x}(t,s),s) - \phi(\boldsymbol{x}_0(s),s) - t = 0\}$. Wave front methods provide a simple mechanism for controlling the resolution and accuracy of the numerical approximation. There are also Eulerian wave front methods based on PDE formulations of the problem, see [9] for example.

Using wave front methods with Gaussian beams is not as straightforward, since the beam method strongly depends on the distribution and width of the beams at the initial front. We introduce a Lagarangian wave front-based Gaussian beam method, in which a wave front is evolved in phase space $(\boldsymbol{x}, \theta)$ by solving the ODE systems (3) and (6). In order to overcome the problem of diverging rays in the ray tracing method, we use an automatic refinement criterion to keep the fronts uniformly sampled.

In this section we will show how to construct the wave front Gaussian beam method. Let the initial phase space wave front be $(\boldsymbol{x}_0(s), \theta_0(s))$ parameterized by s and assume that the exact phase space wave front at travel-time t is described by $(\boldsymbol{x}(t,s), \theta(t,s))$. Now let

$$\boldsymbol{x}_j^n \approx \boldsymbol{x}(n\Delta t, j\Delta s), \qquad \theta_j^n \approx \theta(n\Delta t, j\Delta s),$$

where (j,n) represents a marker (grid point) on a front at $t = n\Delta t$. We initialize the markers on the initial front at $t = 0$ as $(\boldsymbol{x}_j^0, \theta_j^0) = (\boldsymbol{x}_0(j\Delta s), \theta_0(j\Delta s))$. Each marker is then updated by a standard ODE-solver, applied to the ray tracing system (3). See Figure 2 (left).

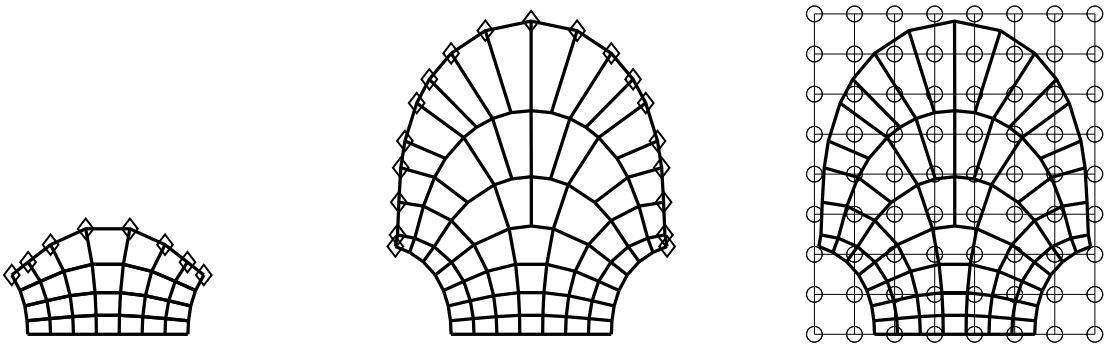


Figure 2: Wave front construction. Markers ($\diamond$) on the wave front are propagated as ordinary rays (left). When markers fail to acurately resolve the wave front, new markers are inserted via interpolation from the old markers (middle). The information carried by markers on the wave fronts are interpolated onto a regular grid (right).

When the resolution of the wave front deteriorates, new markers are inserted and computed by interpolation from the old markers. We add a new marker $(j + 1/2, n)$ between markers (j,n) and $(j+1,n)$ if

$$|\boldsymbol{x}_{j+1}^n - \boldsymbol{x}_j^n| \geq \delta_{\boldsymbol{x}} \qquad or \qquad |\theta_{j+1}^n - \theta_j^n| \geq \delta_\theta,$$

for some tolerances $\delta_{\boldsymbol{x}}$ and δ_θ. See Figure 2 (middle).

Figure 3 shows the central rays and the computed function θ versus y along the front $t = 3$ obtained by the wave front method without and with refinement. As it can

be seen, in the case of no refinement, the solution is poorly resolved in places where the rays diverge. It is well resolved if refinement is performed.

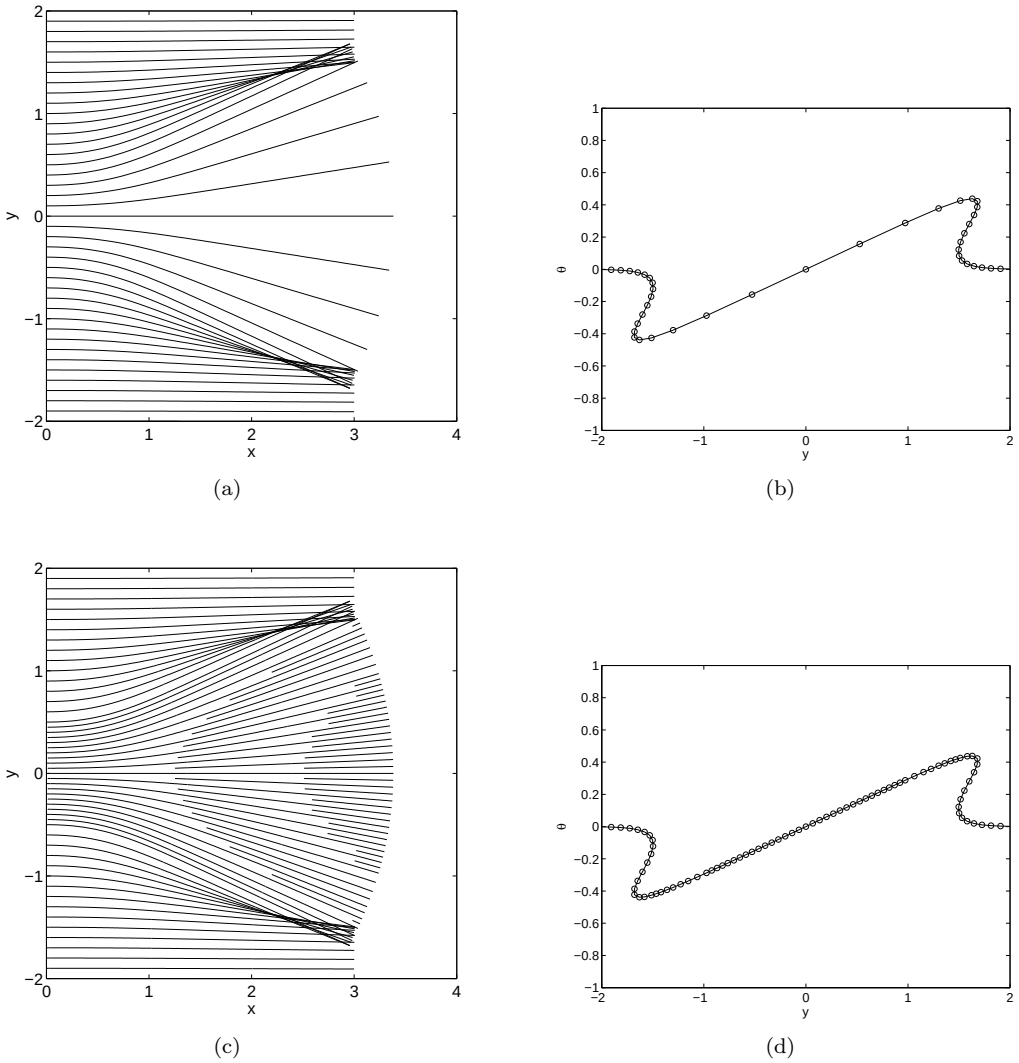


Figure 3: Figure (a) shows the beam central rays generated by a plane wave which propagates into the computational domain from the left boundary. Figure (b) shows the function θ versus y along the front $t = 3$. In places where the rays diverge, the solution, computed by the wave front method without refinement, is poorly resolved. Bottom figures show the rays and solution computed by the wave front method with refinement. In this case, the solution is uniformly resolved.

Note that inserting new markers on the fronts in the wave front method is analogous to inserting new rays in the ray tracing method. However, here, the rays are inserted only in places where the resolution deteriorates. These rays are then traced afterward, and there is no need to compute them from the source, as is done in the ray tracing method. Therefore, in Gaussian beam summation, the wave front method is computationally faster than the ray tracing method, while keeping the same order

of accuracy.

In parallel with computing $(\boldsymbol{x}_j^n, \theta_j^n)$, we also compute the real-valued canonical functions (Q_{1j}^n, P_{1j}^n) and (Q_{2j}^n, P_{2j}^n) by solving the dynamic ray tracing system (6) with the initial conditions (8). We note that via (9) we can recreate beams with *any* initial data P_0 and Q_0 from these two canonical solutions. We save $V_j^n := (x_j^n, y_j^n, \theta_j^n, Q_{1j}^n, P_{1j}^n, Q_{2j}^n, P_{2j}^n)^\top$ for each grid point (j, n).

Now assume we want to compute the wave field at a marker (j^*, n^*), as a receiver point, on the front at $t^* = n^*\Delta t$. We first select the beam parameters P_0, Q_0 and the initial spacing h such that the initial/boundary data is well approximated on the initial front. We discretize the initial front into M equi-distant grid points $s_m = m\,h$, with $m = 1, \ldots, M$. Each grid point on the initial front represents the initial point of a beam central ray Ω_m. We then find $V_m^{n^*}$ by interpolating the already computed values $V_j^{n^*}$. The complex-valued functions $P_m^{n^*}, Q_m^{n^*}$ on the front at t^* are obtained by (9). The total wave field at the marker (j^*, n^*) is then calculated by (11).

As an alternative way, if we need the wave field on a regular grid, we can first interpolate V_j^n values down on a regular Cartesian grid. See Figure 2 (right). We then use the same precedure as above, but instead of a wavefront, we consider a line passing the receiver point.

A main advantage of using the basis solutions (Q_1, P_1) and (Q_2, P_2) in the algorithm is that at different wavefronts we can use different initial data Q_0, P_0 to evaluate the solution at no extra cost. Therefore, optimization, based on the minimization of either the beam width or the error, is possible, and we can approximate the field with different, optimally chosen, beams in different points on the fronts. Moreover, since the geometrical optics solutions can be obtained by (Q_1, P_1), we can construct a hybrid algorithm and use Gaussian beam solutions only around caustics.

The cost of the wave front tracking is independent of ω and is typically $\mathcal{O}(1)$ per grid point.

4 Numerical example

In this section we consider two numerical examples and use the wave front method described in Section 3 to compute the wave field.

4.1 Example 1

We consider a rectangular domain $\mathcal{D} = [0, 4] \times [-2, 2]$ and the speed of propagation

$$c(x, y) = \frac{1}{1 + e^{-y^2}}, \qquad (x, y) \in \mathcal{D}.$$

The boundary data is given on the y-axis, $\boldsymbol{x}_0(s) = (0, s)$, by a plane wave propagating into the computational domain orthogonally, i.e. $\theta_0(s) = 0$. The plane wave is refracted as it propagates through the domain, and a cusp caustic is formed. Figure 4 shows the central rays of the Gaussian beams and the corresponding wave fronts.

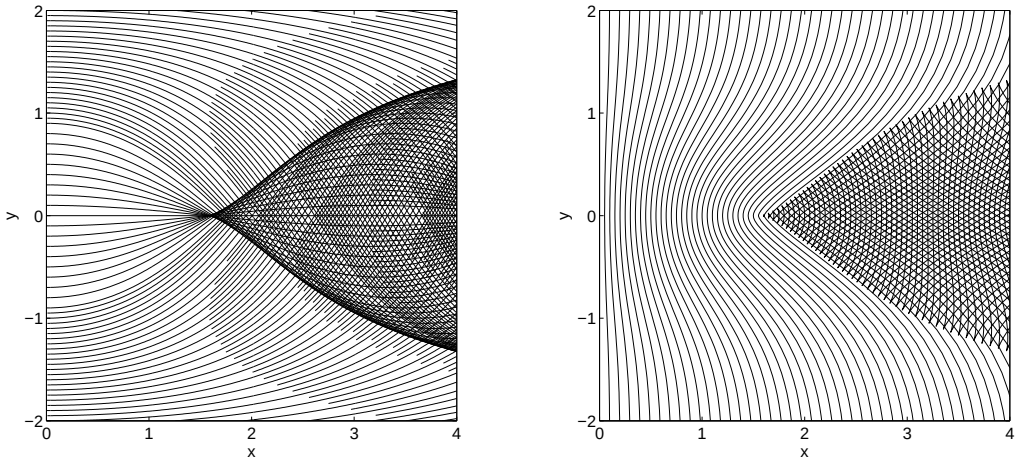


Figure 4: The central rays and the corresponding wave fronts generated by a plane wave propagating into the domain from the left boundary $x = 0$. The wave field is refracted inside the domain and forms a cusp caustic.

The total wave field along the line $x = 1$ is shown in Figure 5a for different frequencies. As it can be seen the solution obtained by the Gaussian beam method converges to the solution obtained by geometrical optics. Figure 5b shows the maximum pointwise error between the Gaussian beam solution and the geometrical optics solution. The error is proportional to ω^{-1} and agrees with the convergence rate obtained in [11].

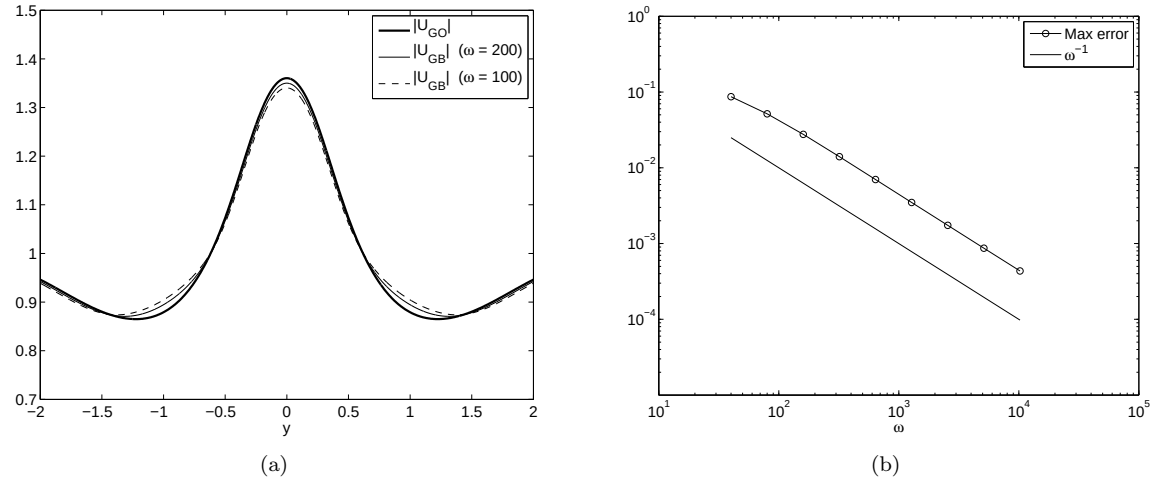


Figure 5: Figure (a) shows the magnitude of the solution obtained by wave front Gaussian beam method with different frequencies and geometrical optics at $x = 1$. Figure (b) shows the logarithmic scale of the maximum pointwise error between the Gaussian beam solutions and the geometrical optics solution. The error is of order $\mathcal{O}(\omega^{-1})$.

Figure 6 shows the total wave field along the line $x = 1.572$ where there is a cusp caustic at $y = 0$. A zoomed view at the caustic is shown in Figure 7a. Unlike the amplitude of the geometrical optics solution which is unbounded at the caustic, the amplitude of Gaussian beam solution is bounded and increases as the frequency increases. The rate of increase is shown in Figure 7b and agrees with the Maslov theory saying that at a cusp caustic, $|u| = \mathcal{O}(\omega^{1/4})$. See, for example, [15].

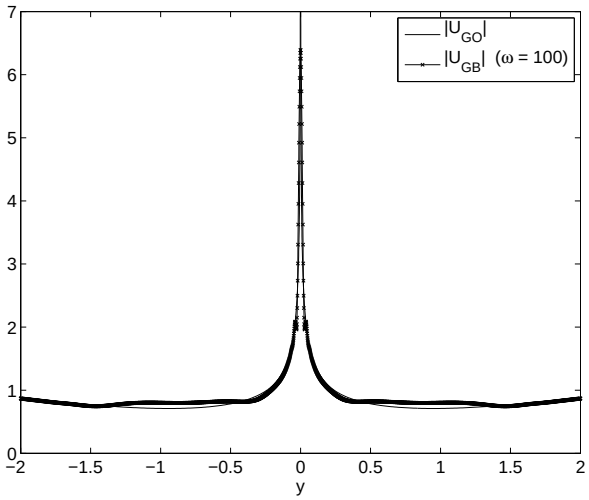


Figure 6: Absolute value of the wave field $|u|$ along the line $x = 1.572$. A cusp caustic is formed at $y = 0$ along this line. The amplitude of the geometrical optics solution is unbounded at the caustic, but Gaussian beam method gives a bounded amplitude.

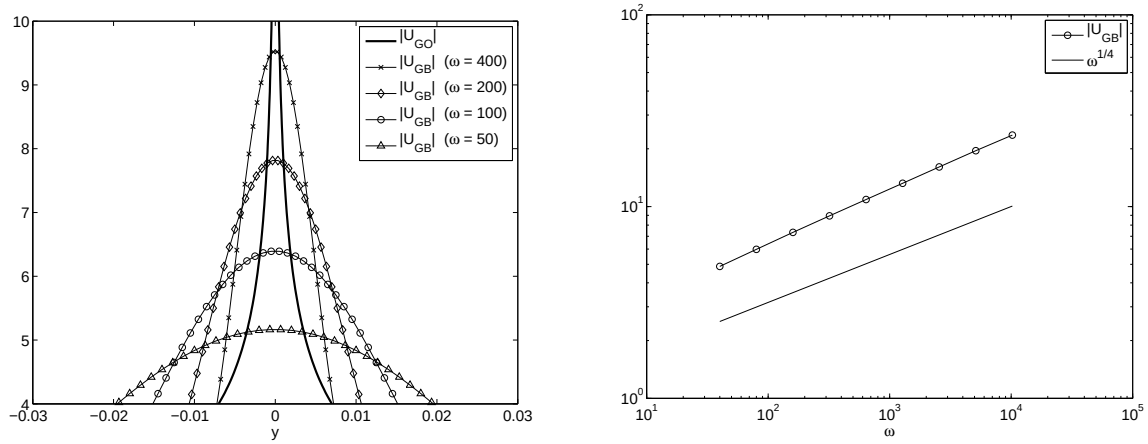
Figure 8 shows the total wave field after the caustic along the line $x = 2.5$ for two different frequencies. Note that in between the caustics, there are multiple arrival times, and the amplitude of the wave field is very oscillatory.

As it was mentioned in Section 3, a main advantage of this algorithm is that by using the basis solutions (Q_1, P_1) and (Q_2, P_2) we can use different initial data Q_0, P_0 for different points of the domain at no extra cost. It provides a simple and fast way of optimizing the solution. In order to verify this, we plot the magnitude of the solution along two different lines $x = 1$ and $x = 1.572$. First, we use a fixed value for the initial data, $Q_0 = 1, P_0 = i$ for both cases, see Figure 9 (top). Next, we use different values for the initial data, $Q_0 = 1 - 0.2i, P_0 = i$ along $x = 1$ and $Q_0 = 1.2 - 0.2i, P_0 = i$ along $x = 1.572$, see Figure 9 (bottom). As the figures show, by choosing different initial data at different points, it is possible to improve the solution.

4.2 Example 2

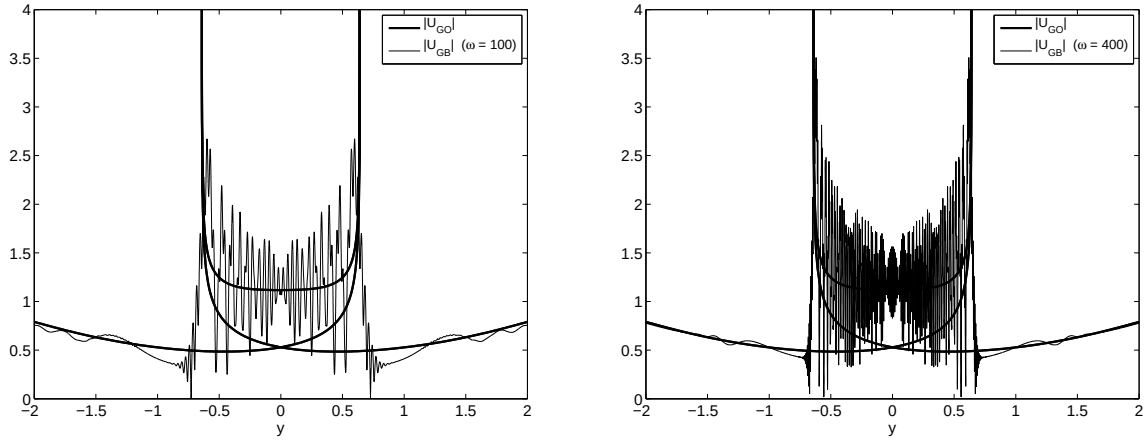
As the second example, we consider the speed of propagation

$$c(x, y) = 1 + 0.5\, e^{-2\,((x-0.5)^2+y^2)}, \qquad (x, y) \in \mathcal{D}.$$



(a) (b)

Figure 7: Figure (a) shows a zoomed view of the solution magnitude close to the cusp caustic at $(x, y) = (1.572, 0)$. While the amplitude of the geometrical optics solution is unbounded at the caustic, the Gaussian beam solutions are bounded and increase as the frequency increases. Figure (b) shows that the rate of increase is $\mathcal{O}(\omega^{1/4})$ as the Maslov theory predicts.

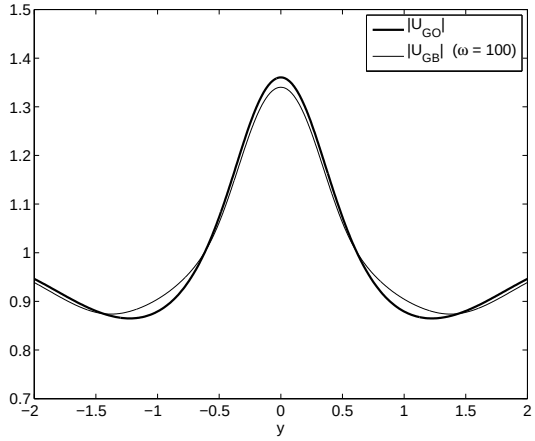


(a) (b)

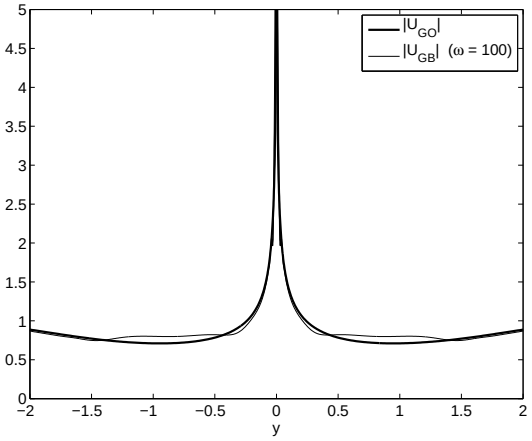
Figure 8: Amplitude of the wave field along the line $x = 2.5$ for two frequencies $\omega = 100$ and $\omega = 400$. In the region inside the caustic, there are three arrival times.

Similar to the first example, the wave field is generated by a plane wave propagating from the left boundary into the computational domain orthogonally. Two cusp caustics are formed. Figure 10 shows the central rays of the Gaussian beams and the corresponding wave fronts.

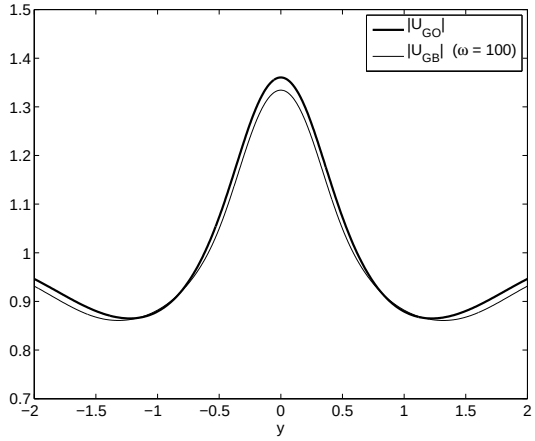
The total wave field for different frequencies along the line $x = 1$ and the maximum pointwise error between the Gaussian beam solution and the geometrical optics solution are shown in Figure 11. The error is proportional to ω^{-1}, as expected.



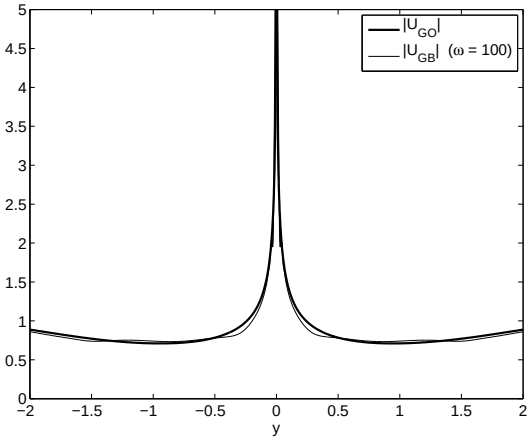
(a) $x = 1$, $Q_0 = 1$



(b) $x = 1.572$, $Q_0 = 1$



(c) $x = 1$, $Q_0 = 1 - 0.2i$



(d) $x = 1.572$, $Q_0 = 1.2 - 0.2i$

Figure 9: Figures (a) and (b) show the magnitude of the solution with a fixed $Q_0 = 1$ along the lines $x = 1$ and $x = 1.572$, respectively. Figures (c) and (c) show the magnitude of the solution along the same lines, but with different initial data $Q_0 = 1 - 0.2i$ and $Q_0 = 1.2 - 0.2i$, respectively.

Figure 12 shows the total wave field along the line $x = 2.125$ where there are two cusp caustic at $y = \pm 1.352$. A zoomed view at the caustic is shown in Figure 13a, and the rate of the increase of Gaussian beam solutions as the frequency increases is shown in Figure 13b. As it can be seen, $|u| = \mathcal{O}(\omega^{1/4})$ in agreement with the Maslov theory.

References

[1] M. M. Popov, "A New Method of Computation of Wave Fields Using Gaussian Beams", Wave Motion, vol. 4, pp. 85-97, 1982.

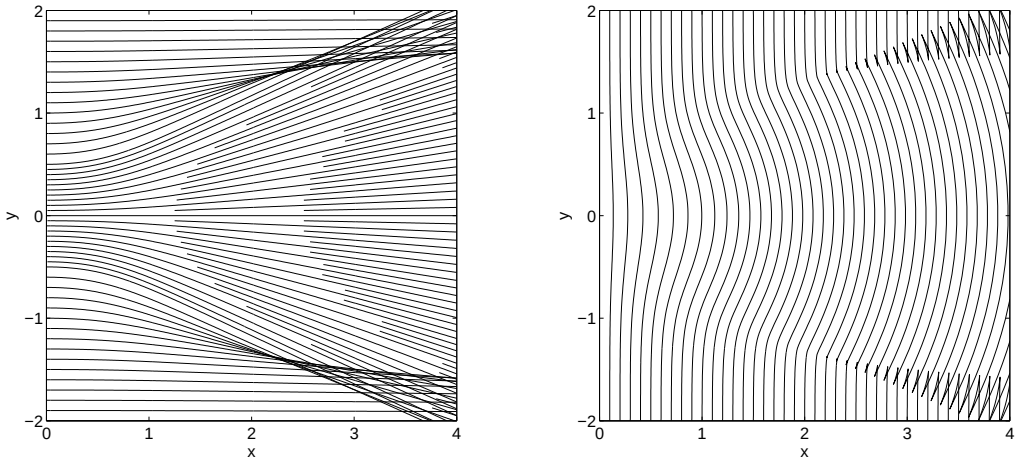


Figure 10: The central rays and the corresponding wave fronts generated by a plane wave propagating into the domain from the left boundary $x = 0$. The wave field is refracted inside the domain and form two cusp caustics.

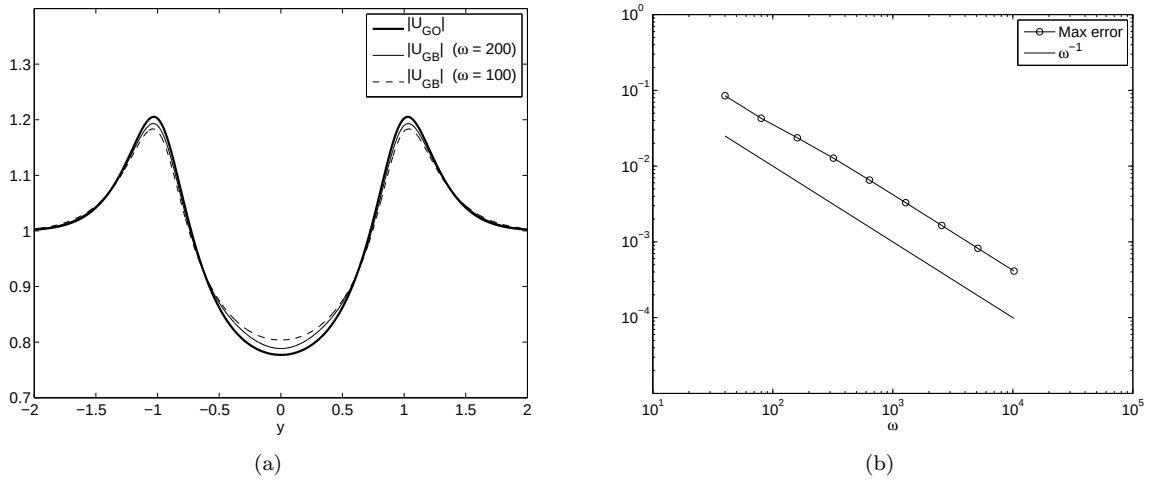


Figure 11: Figure (a) shows the magnitude of the solution obtained by wave front Gaussian beam method with different frequencies and geometrical optics at $x = 1$. Figure (b) shows the logarithmic scale of the maximum pointwise error between the Gaussian beam solutions and the geometrical optics solution. The error is of order $\mathcal{O}(\omega^{-1})$.

[2] V. M. Babic and T. F. Pankratova, "On Discontinuities of Green's Function of the Wave Equation with Variable Coefficient", Problemy Matem. Fiziki, vol. 6, Leningrad University, Saint-Petersburg, 1973.

[3] A. P. Katchalov and M. M. Popov, "Application of the Method of Summation of Gaussian Beams for Calculation of High-frequency Wave Fields", Sov. Phys. Dokl. ,

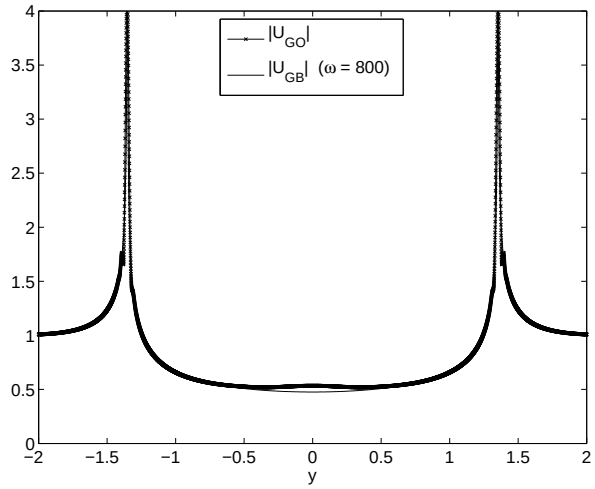
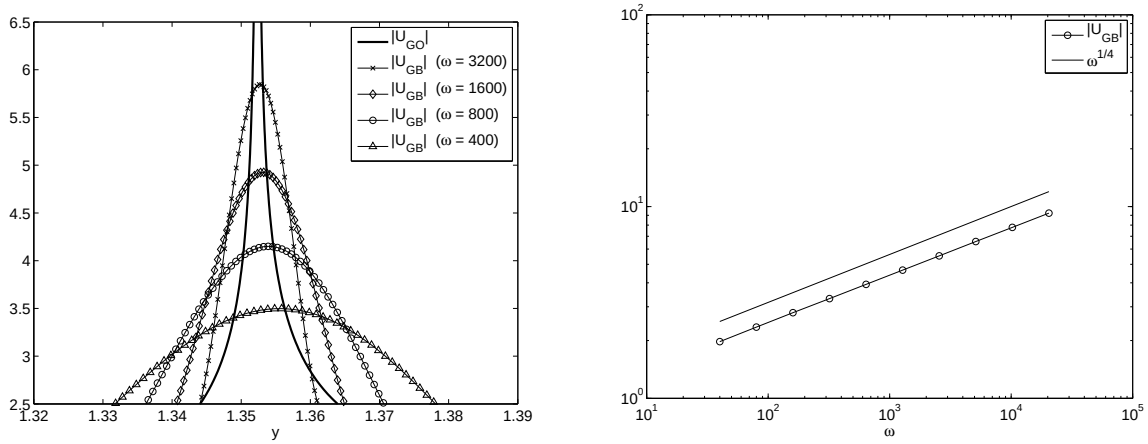


Figure 12: The absolute value of the wave field $|u|$ along the line $x = 2.125$. Two cusp caustics are formed at $y = \pm 1.352$ along this line.



(a) (b)

Figure 13: Figure (a) shows a zoomed view of the solution magnitude close to the right cusp caustic at $(x, y) = (2.125, 1.352)$. While the amplitude of the geometrical optics solution is unbounded at the caustic, the Gaussian beam solutions are bounded and increase as the frequency increases. Figure (b) shows that the rate of increase is $\mathcal{O}(\omega^{1/4})$ as the Maslov theory predicts.

vol. 26, pp. 604-606, 1981.

[4] V. Cerveny, M. M. Popov and I. Psencik, "Computation of Wave Fields in Inhomogeneous Media - Gaussian Beam Approach", Geophys. J. R. Astr. Soc., vol. 70, pp. 109-128, 1982.

[5] N. R. Hill, "Gaussian Beam Migration", Geophysics, vol. 55, No. 11, pp. 1416-1428, 1990.

[6] N. R. Hill, "Prestack Gaussian-Beam Depth Migration", Geophysics, vol. 66, No. 4, pp. 1240-1250, 2001.

[7] J. Ralston, "Gaussian Beams and the Propagation of Singularities", In Studies in partial differential equations, vol. 23 of MAA Stud. Math., pp. 206-248, Math. Assoc. America, Washington, DC, 1982.

[8] S. Leung, J. Qian and R. Burridge, "Eulerian Gaussian beams for high frequency wave propagation", Geophysics vol. 72, No. 5, pp. SM61-SM76, 2007.

[9] B. Engquist and O. Runborg. Computational high frequency wave propagation. *Acta Numerica*, 12:181–266, 2003.

[10] A. P. Katchalov and M. M. Popov, "Application of the Gaussian Beam Method to Elasticity Theory", Geophys. J. R. Astr. Soc., vol. 81, pp. 205-214, 1985.

[11] M. Motamed and O. Runborg, "Taylor Expansion Error in Gaussian Beam Summation", Preprint, 2008.

[12] L. Klimes, "Expansion of a High-frequency Time-harmonic Wavefield Given on an Initial Surface into Gaussian Beams", Geophys. J. R. astr. Soc., vol. 79, pp. 105-118, 1984.

[13] V. Vinje, E. Iversen and H. Gjoystdal, "Traveltime and amplitude estimation using wavefront construction", In Eur. Ass. Expl. Geophys., pp. 504-505, 1992. Extended abstracts.

[14] V. Vinje, E. Iversen and H. Gjoystdal, "Traveltime and amplitude estimation using wavefront construction", Geophysics, vol. 58, No. 8, pp. 1157-1166, 1993.

[15] M. E. Taylor, "Partial Differential Equations I: Basic Theory", volume 115 of Applied Mathematical Sciences. Springer-Verlag, 1996.

Paper V

Finite difference schemes for second order systems describing black holes

Mohammad Motamed,[1,2] M. Babiuc,[2,3] B. Szilágyi,[2] H-O. Kreiss,[1,2] and J. Winicour[2,3]

[1]*NADA, Royal Institute of Technology, 10044 Stockholm, Sweden*

[2]*Albert Einstein Institute, Max Planck Gesellschaft, Am Mühlenberg 1, D-14476 Golm, Germany*

[3]*Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*

(Received 19 March 2006; published 6 June 2006)

In the harmonic description of general relativity, the principal part of Einstein's equations reduces to 10 curved space wave equations for the components of the space-time metric. We present theorems regarding the stability of several evolution-boundary algorithms for such equations when treated in second order differential form. The theorems apply to a model black hole space-time consisting of a spacelike inner boundary excising the singularity, a timelike outer boundary and a horizon in between. These algorithms are implemented as stable, convergent numerical codes and their performance is compared in a 2-dimensional excision problem.

DOI: 10.1103/PhysRevD.73.124008 PACS numbers: 04.25.Dm, 04.20.Ex, 04.30.Db, 95.30.Lz

I. INTRODUCTION

A primary goal of numerical relativity is the computation of gravitational radiation waveforms from binary black holes. Radiation produced in the inspiral and merger of binary black holes is expected to provide a strong signal for gravitational wave observatories. However, the simulation of black holes has proved to be a difficult computational problem. The importance of this challenging problem has recently spurred a fertile interaction between numerical relativity and computational mathematics. The classic computational treatment of hyperbolic systems has been directed at fluid dynamics and has been based upon first differential order systems. Certain formulations of Einstein's equations take a more natural second order form, notably the harmonic formulation [1,2] for which well-posedness of the Cauchy problem was first established [3]. Here we present theorems regarding the stability of several evolution-boundary algorithms for such second order systems which have direct application to the black hole problem.

Harmonic coordinates $x^\alpha = (t, x^i) = (t, x, y, z)$ have only recently been used in designing numerical codes [4–12]. They satisfy the curved space wave equation

$$\Box_g x^\mu := \frac{1}{\sqrt{-g}}\partial_\alpha(\sqrt{-g}g^{\alpha\beta}\partial_\beta x^\mu) = 0. \tag{1.1}$$

In harmonic coordinates, Einstein's equations reduce to 10 quasilinear wave equations for the components of the metric,

$$\Box_g g^{\mu\nu} = S^{\mu\nu}, \tag{1.2}$$

where $S^{\mu\nu}$ are nonlinear terms which do not enter the principal part. Thus the scalar wave equation

$$g^{\alpha\beta}\partial_\alpha\partial_\beta u = 0, \tag{1.3}$$

which has the same principal part, provides a fundamental testing ground for designing algorithms to treat the nonlinear gravitational problem (1.2). In a previous study [13], we used this scalar equation to develop evolution and boundary algorithms for a model one dimensional black hole excision problem. Here we extend these results to two dimensions. While the extension to 2D involves substantial new features, the generalization from 2D to 3D is quite straightforward. Thus our results are immediately applicable to algorithms for the harmonic gravitational Eqs. (1.2), as well as their generalization to include harmonic gauge forcing terms [14] and other related generalizations such as the $Z4$ formulation [15].

We treat (1.3) in the second order differential form, which has advantages for both computational efficiency and accuracy over first order formulations [16,17]. Although the system can be reduced to first order symmetric hyperbolic form [18], this has the disadvantage of introducing auxiliary variables with their associated constraints and boundary conditions. The second order form is also best suited to the analogous wave equations governing elasticity and acoustics. Elasticity theory is governed by a coupled system of wave equations which for simple cases is similar to (1.3), in which the spatial components g^{ij} are determined by the elastic moduli. In fact, some of the techniques utilized here have been developed in a recent computational study of the wave equations governing an elastic body [19]. The new ingredient introduced in the wave Eq. (1.3) arises from the nonvanishing mixed space-time derivatives arising from the components g^{it}. Such terms do not ordinarily appear in the wave equations governing elasticity theory because they are treated in the rest frame of the body but they would necessarily arise in treating acoustic waves propagating in a medium with nonuniform macroscopic motion. In general relativity, these mixed space-time components of the metric correspond to a nonvanishing "shift," which is an essential feature of the black hole problem. In the standard 3 + 1 description of space-time [20], the Cauchy hypersurfaces $t =$ constant are required to be spacelike so that they have a length element with Euclidean signature

$$d\ell^2 = h_{ij}dx^i dx^j. \tag{1.4}$$

The inverse spatial metric h^{ij}, satisfying $h^{ij}h_{jk} = \delta^i_k$, is related to the spatial components of the 4-metric determining the wave operator by

$$h^{ij} = g^{ij} - \frac{1}{g^{tt}}\beta^i\beta^j, \tag{1.5}$$

where β^i are the components of the shift. Here, the spacelike character of the Cauchy hypersurfaces requires that $g^{tt} < 0$.

The wave equation with shift has not received a great deal of attention outside of recent work in general relativity [6,10,13,21–23], although the important causal effect of the shift in the black hole excision problem has been recognized. This causal effect has been incorporated in a first order computational treatment of Einstein's equations by using upwind differencing for the shift terms, see e.g. [24]. Even before the computational problem is attempted, new mathematical features introduced by the shift must be dealt with in formulating a well-posed initial-boundary value problem. The operator $h^{ij}\partial_i\partial_j$ is by construction an elliptic operator defined by the spatial metric of the Cauchy hypersurfaces. However, the operator $g^{ij}\partial_i\partial_j$ is elliptic only when the shift is sufficiently small. The elliptic case arises when the operator ∂_t is timelike, i.e. when the evolution proceeds in a timelike (subluminal) space-time direction.

Without loss of generality, we set $g^{tt} = -1$ and write the 2D version of (1.3) as

$$(\partial_t^2 - 2(\beta^x\partial_x + \beta^y\partial_y)\partial_t - (a_1 - \beta^x\beta^x)\partial_x^2 - (c_1 - \beta^y\beta^y)\partial_y^2 - 2(b_1 - \beta^x\beta^y)\partial_x\partial_y)u = 0, \tag{1.6}$$

where $h^{xx} = a_1$, $h^{xy} = b_1$ and $h^{yy} = c_1$. The Euclidean property of h^{ij} requires

$$a_1 > 0, \qquad c_1 > 0, \qquad a_1c_1 - b_1^2 > 0. \tag{1.7}$$

The components of g^{ij} are $g^{xx} = a = a_1 - \beta^x\beta^x$, $g^{yy} = c = c_1 - \beta^y\beta^y$ and $g^{xy} = b = b_1 - \beta^x\beta^y$. In the subluminal case when $g^{ij}\partial_i\partial_j$ is an elliptic operator, the simplest second order accurate difference approximation to (1.6) is

$$W := (\partial_t^2 - 2(\beta^x D_{0x} + \beta^y D_{0y})\partial_t - aD_{+x}D_{-x} - cD_{+y}D_{-y} - 2bD_{0x}D_{0y})u = 0. \tag{1.8}$$

(Here D_{0i}, D_{+i} and D_{-i} are, respectively, the centered, forward and backward difference operators in the x^i-direction defined in Sec. III). This leads to stable evolution-boundary algorithms for Dirichlet, Neumann, Sommerfeld or other dissipative boundary conditions. Stability was established for the 1D case using a semi-discrete energy norm in [13], and this was generalized using the discrete energy method to the full 3D case in [10,12].

The W-algorithm (1.8) is unstable when the shift is sufficiently large so that $g^{ii} \leq 0$ (for any diagonal component). This occurs in one of the strategies for avoiding problems with the singularity which ultimately forms inside a black hole. In this strategy, the singularity is "excised" by surrounding it with a spacelike inner boundary. The evolution direction which is adopted to this spacelike boundary is superluminal, so that g^{ij} no longer has Euclidean signature. In Sec. III, we establish the stability of several different second order, evolution-boundary algorithms for this superluminal case. For the Cauchy problem, we establish stability for a general system of wave equations in s spatial dimensions so that the results may be immediately applied to other second order systems such as elasticity theory or acoustics. For the boundary, we specialize our treatment to scalar equations in 2D in order to simplify the notation, but the extension to general systems in sD is straightforward. In particular, our results apply directly to the 3D harmonic evolution of black holes.

The analysis of the initial-boundary problem for (1.6) in Sec. III makes evident that the above geometric properties of the wave equation have a mathematical analogue which results independently from a consideration of the well-posedness of the problem. The geometrical and analytical approaches are complementary and provide a good meeting ground for the ideas of numerical relativity and computational mathematics. While the main concern of numerical relativity is the black hole problem, the stability theorems for the finite difference algorithms developed for the model problems considered here provide a firm basis for attacking this problem with the harmonic Einstein system (1.2).

In Sec. IV, we compare the performance of the algorithms for the superluminal case in a problem without boundaries. In Sec. V, we simulate a simple 2D model of the excision problem in which the inner boundary $\mathcal{S}$ is spacelike and the outer boundary $\mathcal{T}$ is timelike. Between the boundaries the operator $g^{ij}\partial_i\partial_j$ goes from nonelliptic to elliptic along a curve $\mathcal{H}$ where $\det(g^{ij}) = 0$. The metric is chosen so that no characteristics can leave the inner region between $\mathcal{S}$ and $\mathcal{H}$, so that $\mathcal{H}$ mimics the role of a horizon. The global simulation of (1.6) in the region bounded by $\mathcal{T}$ and $\mathcal{S}$ is achieved with a blended evolution algorithm. A stable superluminal algorithm is used in an inner region between $\mathcal{S}$ and $\mathcal{H}$. In the exterior region, this superluminal algorithm is blended to the W-algorithm (1.8), so that the W-algorithm is used to treat the outer boundary $\mathcal{T}$. This model excision problem involves many of the mathematical difficulties in the full gravitational case. We begin in Sec. II with some simple examples which illustrate the problem, its potential pitfalls and how to avoid them.

II. SOME SUBTLETIES ASSOCIATED WITH THE WAVE EQUATION WITH SHIFT

In an inertial coordinate system $\hat{x}^\alpha = (\hat{t}, \hat{x}^i)$ (in units where the velocity of light $c = 1$), the wave equation which governs special relativistic physics,

$$(\partial_{\hat{t}}^2 - \delta^{ij}\partial_{\hat{i}}\partial_{\hat{j}})u = 0, \tag{2.1}$$

does not contain a shift. The invariance of the velocity of light results from the property that this wave equation retains the same form (2.1) under a Lorentz transformation,

$$t' = \frac{1}{\sqrt{1-\beta^2}}(\hat{t} - \delta_{ij}\beta^i\hat{x}^j) \qquad x'^i = \frac{1}{\sqrt{1-\beta^2}}(\hat{x}^i - \beta^i\hat{t})$$
$$\beta^2 = \delta_{kl}\beta^k\beta^l, \tag{2.2}$$

to another inertial coordinate system with relative motion. In this way special relativity resolves the dilemma with experiment that under a Galilean transformation,

$$t = \hat{t} \qquad x^i = (\hat{x}^i - \beta^i\hat{t}), \tag{2.3}$$

(2.1) gives rise to the shifted wave equation

$$(\partial_t^2 - 2\beta^i\partial_i - (\delta^{ij} - \beta^i\beta^j)\partial_i\partial_j)u = 0 \tag{2.4}$$

whose solutions propagate with coordinate speeds in the range $|1 \pm \beta|$ (where $\beta^2 = \delta_{ij}\beta^i\beta^j$). This raises the question: why does the wave equation with shift arise in general relativity?

In fact, although there are no preferred inertial coordinates in general relativity, in any sufficiently small spacetime region it is always possible to introduce Gaussian coordinates in which the wave Eq. (1.3) reduces to the shift-free form

$$(\partial_t^2 - h^{ij}\partial_i\partial_j)u = 0. \tag{2.5}$$

The problem here is that in Gaussian coordinates the worldlines $x^i = \text{const}$ are geodesics, i.e. the worldlines of freely falling observers, which can be focused by the attractive nature of gravity to produce coordinate singularities. This can occur on a short time scale in a strong gravitational field.

Another reason for introducing a shift is the simplicity of harmonic coordinates in reducing Einstein's equations into the hyperbolic form (1.2). Since the shift components g^{it} satisfy a coupled system of nonlinear wave equations, even if they were initialized with vanishing Cauchy data they would in general evolve to be nonzero. This cannot be avoided by introducing a harmonic gauge forcing term, of the form $\Box x^\alpha = F^\alpha$, without choosing the forcing term F^α to depend upon the derivatives of the metric $\partial_\mu g^{\alpha\beta}$. This in turn jeopardizes the hyperbolic form of the reduced Einstein equations and the well-posedness of the Cauchy problem [14].

Yet another reason for introducing a shift arises in the simulation of black holes. Once a black hole of mass M has formed there is at most a proper time of order M (in gravitational units) until a physical singularity is encountered. On the other hand, a simulation which provides gravitational waveforms of physical interest typically requires an evolution for a proper time of more than $100M$ in the exterior region. One strategy for accomplishing this is to excise the singularity by surrounding it with a spacelike inner boundary for the simulation domain, i.e. an inner boundary which moves at superluminal speed. If the evolution tracks the inner boundary then a superluminal shift must be used.

This can be illustrated by a spherically symmetric Schwarzschild black hole for which the wave Eq. (1.3) becomes

$$\left(\left(1 + \frac{2M}{r}\right)\partial_t^2 - \frac{4M}{r}\partial_t\partial_r - \left(1 - \frac{2M}{r}\right)\partial_r^2 - \frac{1}{r^2}\left(\partial_\theta^2 + \frac{1}{\sin^2\theta}\partial_\phi^2\right)\right)u = 0, \tag{2.6}$$

in ingoing Eddington-Finkelstein coordinates. Here the evolution takes place on the spacelike Cauchy hypersurfaces $t = \text{const}$ which are nonsingular for $r > 0$. The black hole is located at $r = 2M$, which is a characteristic hypersurface with the horizon property that no characteristics leave the region $r \le 2M$. The singularity is excised by evolving in a domain $R_1 \le r \le R_2$, where $0 < R_1 < 2M$ and $R_2 \gg 2M$. The shift has the radial component

$$\beta^r = \frac{1}{1 + \frac{r}{2M}} > 0. \tag{2.7}$$

The change in sign of the coefficient of ∂_r^2 in passing inside the horizon does not change the hyperbolicity of the wave equation but it changes its mathematical properties. Outside the horizon, the curves of constant (r, θ, ϕ) are timelike, as well as the outer boundary $r = R_2$. In the outer region $2M < r \le R_2$, the W-algorithm (1.8) provides a stable second order evolution-boundary algorithm for the wave equation [10,12,13].

Inside the horizon, the t-direction, as well as the inner boundary $r = R_1$ is spacelike, i.e. evolution on a grid with constant (r, θ, ϕ) proceeds outside the light cone. This effects the mathematical properties of the wave equation. As a result, in this domain, the W-algorithm is unstable. The alternative algorithms presented in Sec. III are stable inside the horizon. But the W-algorithm has better accuracy than these algorithms in the exterior region [10]. In the simulation of the model excision problem in Sec. V, a stable algorithm for the superluminal regime is blended to the W-algorithm in the exterior.

The Schwarzschild horizon has the property that characteristics can not exit from inside, but can enter from the outside. Near the horizon, the radial part of Schwarzschild wave Eq. (2.6) has the same qualitative features as the wave equation

$$(\partial_t - \partial_x)(\partial_t + x\partial_x)u = 0, \tag{2.8}$$

which has a horizon $x = 0$. One set of characteristics of (2.8) cross the horizon at $x = 0$ in the negative x-direction. The other set of characteristics are tangent to the horizon and *diverge* away on either side. An observer at $x > 0$ cannot see beyond the horizon at $x = 0$. This is the situ-

ation which we dealt with in a model 1D excision problem [13] whose treatment we generalize to 2D in Sec. V. However, it should be emphasized that the related equation

$$(\partial_t - \partial_x)(\partial_t - x\partial_x)u = 0 \tag{2.9}$$

has a different mathematical character. Although (2.9) is also hyperbolic and has a well-posed Cauchy problem, one set of characteristics *converge* toward the horizon at $x = 0$. These characteristics approach each other exponentially fast and, in general, the gradients become exponentially large near $x = 0$. This would lead to the focusing of a wave into formation of a shock. Although we do not treat this case in this paper, it is important to bear in mind that it would require different methods.

Boundaries introduce additional subtleties. First consider a timelike boundary, similar to the outer boundary $r = R_2 > 2M$ for the Schwarzschild wave Eq. (2.6). Since the evolution is timelike in the neighborhood of the boundary, the W-algorithm can be used. The stability of dissipative boundary conditions for the W-algorithm was established for 1D in [13] and extended to 3D in [12] by means of a semidiscrete energy method. However, such an energy estimate does not preclude exponential growth of a wave traveling between two boundaries. A simple example [7] arises from the repetitive blue shifting of a wave packet in special relativity reflecting back and forth between two plane boundaries, whose velocities $\pm v$ are controlled to be always toward the packet during reflection. After many reflections the wave packet shrinks in size and its energy grows by a factor $e^{4\alpha T}$, where T is measured in units of the crossing time between reflections and $v = \tanh\alpha$. Dissipation must be used to control such growth of short wavelength error.

It is instructive to interpret the boundary conditions on a wave in special relativity in the shifted coordinate system (2.3) where the boundary has fixed location but moves relative to the $t =$ const Cauchy hypersurfaces. In the 1D case, this gives rise to the half-plane problem

$$(\partial_t^2 - 2\beta\partial_x\partial_t - (1 - \beta^2)\partial_x^2)u = 0, \tag{2.10}$$

in the region $x \leq 0$ (where we now write $\beta^x = \beta$). There are two different frames in which the energy of the wave can be considered - the rest frame of the boundary and the rest frame intrinsic to the Cauchy hypersurfaces. In the rest frame of the boundary, the energy is

$$E = \frac{1}{2}\int_{-\infty}^{0} dx((\partial_t u)^2 + (1 - \beta^2)(\partial_x u)^2) \tag{2.11}$$

and satisfies

$$\partial_t E = \partial_t u((1 - \beta^2)\partial_x + \beta\partial_t)u|_{x=0}. \tag{2.12}$$

In the case $\beta^2 < 1$, this energy provides a norm and the semidiscrete version of the flux-conservation law (2.12) provides the basis for establishing stable evolution-boundary algorithms for the W-algorithm (1.8). Note the sign of β is important here in formulating a stable Neumann boundary condition. A homogeneous Neumann boundary condition takes the dissipative form

$$((1 - \beta^2)\partial_x + \beta\partial_t)u = 0. \tag{2.13}$$

The familiar form $\partial_x u = 0$ implies $\partial_t E \leq 0$ and thus guarantees a well-posed problem only when $\beta > 0$, i.e. only when the motion of the boundary is outward relative to the Cauchy hypersurfaces.

The energy intrinsic to the Cauchy hypersurfaces,

$$E_0 = \frac{1}{2}\int_{-\infty}^{0} dx((\partial_t u - \beta\partial_x u)^2 + (\partial_x u)^2), \tag{2.14}$$

provides a norm even in the superluminal case when $\beta^2 > 1$. It satisfies

$$\partial_t E_0 = \left(\frac{\beta}{2}(\partial_t u - \beta\partial_x u)^2 + \frac{\beta}{2}(\partial_x u)^2 + (\partial_t u - \beta\partial_x u)\partial_x u\right)\Big|_{x=0}. \tag{2.15}$$

Thus, in the absence of a boundary, (2.15) would reduce to $\partial_t E_0 = 0$ so that the Cauchy problem is well-posed for any β. The energy analogous to E_0 is used in Sec. III to establish well-posedness of the Cauchy problem and the stability of superluminal algorithms in the general multi-dimensional case.

When $\beta < -1$, i.e. when the motion of the boundary is superluminal and directed toward the Cauchy hypersurfaces, it is easy to verify that (2.15) implies $\partial_t E_0 < 0$ so that there is always a loss of energy through the boundary. This is the case of a spacelike boundary through which all the characteristics leave, i.e. a pure ''outflow'' boundary. Stable algorithms for such a boundary are also given in Sec. III for the higher dimensional case. Note that for $\beta > 1$ the boundary is also spacelike but now (2.15) implies $\partial_t E_0 > 0$. This is the pure ''inflow'' case, in which all the characteristics enter the boundary. This should not be considered in the context of an initial-boundary value problem, but as a pure Cauchy problem where the boundary represents a nonsmooth extension of the Cauchy hypersurface.

Further subtleties arise in treating co-orbiting, binary black holes. One strategy for the binary problem is to use a rotating coordinate system which co-orbits with the black holes. In the Schwarzschild case, the use of a coordinate $\varphi = \phi - \omega t$ rotating with angular velocity ω transforms the wave Eq. (2.6) into

$$\left(\left(1 + \frac{2M}{r}\right)(\partial_t + \omega\partial_\varphi)^2 - \frac{4M}{r}\partial_t\partial_r - \left(1 - \frac{2M}{r}\right)\partial_r^2 - \frac{1}{r^2}\left(\partial_\theta^2 + \frac{1}{\sin^2\theta}\partial_\varphi^2\right)\right)u = 0. \tag{2.16}$$

Now the t-direction becomes spacelike in the region

$$\left(1+\frac{2M}{r}\right)r^2\omega^2\sin^2\theta > 1, \tag{2.17}$$

which intersects the outer boundary $r = R_2$ if R_2 is sufficiently large. In that case, although the boundary remains timelike the evolution is superluminal so that the W-algorithm is no longer stable. A stable algorithm for such a boundary problem has been established in the 1D case [23]. We will not consider the 2D version of this problem here.

A common strategy for treating the binary black hole problem is to use a grid based upon Cartesian coordinates. This poses a problem in dealing with inner and outer boundaries with the spherical shapes natural to the problem. In other second order wave problems, such curved boundaries have been successfully treated by the embedded boundary method [19,25]. Another approach being explored in general relativity is to use multiblock grids [26–29]. This is another problem which we defer to future work and do not consider here.

III. ALGORITHMS FOR THE 2D SUPERLUMINAL PROBLEM

In this section, we study a class of second order hyperbolic systems with shift which we will use in Sec. V to construct stable algorithms for a model 2D black hole excision problem. The excision problem is a strip problem with spacelike and timelike boundaries and a horizon in between. In the region where the shift is superluminal, the boundary is spacelike and where the shift is subluminal, the boundary is timelike. We replace this problem by Cauchy and half-space problems. The strip problem is well-posed if the corresponding Cauchy and half-space problems are well-posed [30].

For the Cauchy problem, we consider general systems of equations in s space dimensions to demonstrate that the results have applicability beyond numerical relativity. For the half-space problems, we only consider scalar equations in 2D to simplify the notation. The generalization from scalar equations in 2D to systems in sD is quite straightforward.

Here, we consider systems with constant coefficients. Systems with variable coefficients can be reduced to systems with constant coefficients by freezing the coefficients at all points. The problem with variable coefficients is strongly well-posed if the Kreiss condition holds uniformly for all problems with constant coefficients [31].

In order to analyze and establish stable approximations we use the method of lines and reduce the system of partial differential equations to a system of ordinary differential equations in time on a spatial grid. We then apply two standard techniques: the energy method and mode analysis. The stability of the semidiscrete approximation implies the stability of the totally discretized method for most standard methods of lines [32], e.g. with the use of a Runge-Kutta time integrator.

A. The Cauchy problem

We consider the Cauchy problem for a second order system with constant (possibly complex) coefficients in s space dimensions,

$$\begin{aligned}\mathbf{u}_{tt} &= \sum_{j,k=1}^{s} A_{jk}\frac{\partial}{\partial\tilde{x}_j}\frac{\partial}{\partial\tilde{x}_k}\mathbf{u} := P_0(\partial/\partial\tilde{x})\mathbf{u},\\ \tilde{\mathbf{x}} &= (\tilde{x}_1,\ldots,\tilde{x}_s)\in\mathbb{R}^s, \qquad t\ge 0,\end{aligned} \tag{3.1}$$

with the initial conditions

$$\mathbf{u}(\tilde{\mathbf{x}}, t=0) = \mathbf{f}(\tilde{\mathbf{x}}), \quad \mathbf{u}_t(\tilde{\mathbf{x}}, t=0) = \mathbf{g}(\tilde{\mathbf{x}}), \quad \mathbf{u},\mathbf{f},\mathbf{g}\in\mathbb{C}^n. \tag{3.2}$$

(We abbreviate $\partial_\alpha u = u_\alpha$ where confusion does not arise.) Here, for each (j,k), A_{jk} are constant Hermitian matrices $\in\mathbb{C}^{n,n}$, and the data $\mathbf{f} = \mathbf{f}(\tilde{\mathbf{x}})$ and $\mathbf{g} = \mathbf{g}(\tilde{\mathbf{x}})$ are smooth and 1-periodic in each $\tilde{x}_j$, $j = 1,\ldots,s$. The solution $\mathbf{u} = \mathbf{u}(\tilde{\mathbf{x}}, t)$ is then smooth and 1-periodic in each $\tilde{x}_j$. Moreover, we consider solutions with $\int_{\mathbb{R}^s}\mathbf{u}\,d\tilde{\mathbf{x}} = 0$.

We assume that the Hermitian operator P_0 in (3.1) is elliptic, i. e. there exists a positive constant δ such that

$$\sum_{j,k=1}^{s} A_{jk}\xi_j\xi_k \ge \delta|\xi|^2 I \tag{3.3}$$

for all vectors $\xi\in\mathbb{R}^s$. Here I is the $n\times n$ identity matrix.

We introduce a shift by

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{x} + \bar{\beta}t, \qquad \mathbf{x} = (x_1,\ldots,x_s)\in\mathbb{R}^s,\\ \bar{\beta} &= (\beta^1,\ldots,\beta^s)\in\mathbb{R}^s, \qquad \beta^j > 0,\end{aligned}$$

and obtain the shifted system

$$\mathbf{u}_{tt} = 2P_1(\partial/\partial x)\mathbf{u}_t - P_1^2(\partial/\partial x)\mathbf{u} + P_0(\partial/\partial x)\mathbf{u}. \tag{3.4}$$

Here P_1 is a scalar operator,

$$P_1(\partial/\partial x) = \sum_{j=1}^{s}\beta^j\frac{\partial}{\partial x_j}.$$

Theorem 1.—The Cauchy problem for (3.4) is well-posed.

Proof.—If we set $\mathbf{v} = \mathbf{u}_t - P_1(\partial/\partial x)\mathbf{u}$, we get the first order system

$$\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}_t = P_1(\partial/\partial x)\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} + \begin{pmatrix}0 & I\\ P_0(\partial/\partial x) & 0\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}. \tag{3.5}$$

We Fourier transform (3.5) and get

$$\begin{pmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{pmatrix}_t = \hat{P}_1(i\omega)\begin{pmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{pmatrix} + \begin{pmatrix}0 & I\\ -\hat{P}_0(\omega) & 0\end{pmatrix}\begin{pmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{pmatrix}, \qquad \omega\ne\mathbf{0}, \tag{3.6}$$

where $\omega = (\omega_1,\ldots,\omega_s)\in\mathbb{R}^s$ and $\hat{P}_1(i\omega) = i\sum_{j=1}^{s}\beta^j\omega_j$

and $\hat{P}_0(\omega) = \sum_{j,k=1}^{s} A_{jk}\omega_j\omega_k$. Since P_0 is elliptic, we have $\hat{P}_0 = \hat{P}_0^* \geq \delta_0|\omega|^2 I$, for some $\delta_0 > 0$. We can then introduce new variables

$$\hat{\mathbf{w}} = T\begin{pmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{pmatrix}, \qquad T = \begin{pmatrix} I & 0 \\ 0 & \hat{P}_0^{-1/2}\end{pmatrix} \tag{3.7}$$

and, since T and $\hat{P}_1$ commute, we obtain from (3.6)

$$\hat{\mathbf{w}}_t = \hat{P}_1(i\omega)\hat{\mathbf{w}} + \begin{pmatrix} 0 & \hat{P}_0^{1/2}(\omega) \\ -\hat{P}_0^{1/2}(\omega) & 0\end{pmatrix}\hat{\mathbf{w}} := S\hat{\mathbf{w}}. \tag{3.8}$$

Since the matrix S is skew Hermitian, $S^* = -S$, we obtain

$$\frac{\partial}{\partial t}||\hat{\mathbf{w}}||^2 = (S\hat{\mathbf{w}}, \hat{\mathbf{w}}) + (\hat{\mathbf{w}}, S\hat{\mathbf{w}}) = (\hat{\mathbf{w}}, (S^* + S)\hat{\mathbf{w}}) = 0.$$

Therefore, by Parseval's relation, there is an energy estimate and the Cauchy problem is well-posed [30].

Now we show how to construct stable finite difference approximations to (3.4). We leave time continuous and use the method of lines. For brevity, we treat the case $\beta^j > 0$.

Let $h_j = 1/N_j$, $j = 1, \ldots, s$, denote spatial gridlengths, where N_j are natural numbers. For any multi-index $\nu = (\nu_1, \ldots, \nu_s) \in \mathbb{Z}^s$, let $\mathbf{x}_\nu = (h_1\nu_1, \ldots, h_s\nu_s)$ denote the corresponding gridpoint. We consider gridfunctions $\mathbf{u}_\nu := \mathbf{u}_\nu(\mathbf{x}_\nu, t)$ approximating $\mathbf{u}(\mathbf{x}_\nu, t)$ and introduce a translation operator E_j in the j-th coordinate by

$$E_j^p\mathbf{u}_\nu = \mathbf{u}_\nu(\mathbf{x}_\nu + ph_j\mathbf{e}_j, t), \qquad p \in \mathbb{Z},$$

where $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the vector containing a 1 in the j-th position and zeros elsewhere. We then define the forward, backward, and the central difference operators in the j-th coordinate direction by

$$h_jD_{+j} = E_j^1 - E_j^0, \qquad h_jD_{-j} = E_j^0 - E_j^{-1},$$
$$2D_{0j} = D_{+j} + D_{-j}.$$

We approximate (3.4) by

$$\mathbf{u}_{\nu tt} = 2p_1(D)\mathbf{u}_{\nu t} - p_1^2(D)\mathbf{u}_\nu + p_0(D)\mathbf{u}_\nu, \tag{3.9}$$

where $p_0(D)$ is the centered approximation

$$p_0(D) = \sum_{j=1}^{s} A_{jj}D_{+j}D_{-j} + \sum_{j\neq k=1}^{s} A_{jk}D_{0j}D_{0k}, \tag{3.10}$$

and $p_1(D)$ is any one of the following approximations:

(1) Centered approximation,

$$p_1(D) = \sum_{j=1}^{s} \beta^j D_{0j}, \tag{3.11}$$

(2) First order accurate one-sided approximation,

$$p_1(D) = \sum_{j=1}^{s} \beta^j D_{+j}, \tag{3.12}$$

(3) Second order accurate one-sided approximation,

$$p_1(D) = \sum_{j=1}^{s} \beta^j D_{pj}, \tag{3.13}$$

where

$$D_{pj} = D_{+j} - \frac{h_j}{2}D_{+j}^2. \tag{3.14}$$

Remark.—It is not necessary to assume that $\beta^j > 0$ in (3.11), (3.12), and (3.13). In general, we can use $\frac{\beta^j+|\beta^j|}{2}D_{+j} + \frac{\beta^j-|\beta^j|}{2}D_{-j}$ in (3.12). For the second order one-sided approximation (3.13), we replace D_{+j} and D_{-j} by D_{pj} and $D_{mj} = D_{-j} + \frac{h_j}{2}D_{-j}^2$, respectively.

Theorem 2.—The approximation (3.9) is stable.

Proof.—As in the continuum case, we write (3.9) as a first order system and Fourier transform to get

$$\begin{pmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{pmatrix}_t = \hat{p}_1\begin{pmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{pmatrix} + \begin{pmatrix} 0 & I \\ -\hat{p}_0 & 0\end{pmatrix}\begin{pmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{pmatrix}, \tag{3.15}$$

where

$$\hat{p}_0 = \sum_{j=1}^{s} A_{jj}\frac{4}{h_j^2}\sin^2\frac{\xi_j}{2} + \sum_{j\neq k=1}^{s} A_{jk}\frac{1}{h_jh_k}\sin\xi_j\sin\xi_k,$$
$$\xi_j = \omega_jh_j, \qquad |\xi_j| \leq \pi, \tag{3.16}$$

and $\hat{p}_1$ is one of the following:

$$\hat{p}_1 = \sum_{j=1}^{s} \beta^j\frac{1}{h_j}i\sin\xi_j, \tag{3.17}$$

$$\hat{p}_1 = \sum_{j=1}^{s} \beta^j\frac{1}{h_j}\left(i\sin\xi_j - 2\sin^2\frac{\xi_j}{2}\right), \tag{3.18}$$

$$\hat{p}_1 = \sum_{j=1}^{s} \beta^j\frac{1}{h_j}\left(i\sin\xi_j\left(1 - 2\sin^2\frac{\xi_j}{2}\right) - 4\sin^4\frac{\xi_j}{2}\right), \tag{3.19}$$

corresponding to (3.11), (3.12), and (3.13).

Since

$$\sin^2\xi = 4\sin^2\frac{\xi}{2}\cos^2\frac{\xi}{2} = 4\sin^2\frac{\xi}{2} - 4\sin^4\frac{\xi}{2}, \tag{3.20}$$

we have

$$\hat{p}_0 = \sum_{j,k=1}^{s} A_{jk} \frac{1}{h_j h_k} \sin\xi_j \sin\xi_k + \sum_{j=1}^{s} A_{jj} \frac{4}{h_j^2} \sin^4 \frac{\xi_j}{2}. \tag{3.21}$$

From the ellipticity condition (3.3) it follows that $\hat{p}_0$ is positive. As $\xi_j \to 0$ we have $\sin\xi_j/h_j \to \omega_j$. Therefore, the first sum in (3.21) is strictly positive. When $|\xi_j| = \pi$, the first sum in (3.21) is zero but the second sum is not because $\sin\frac{\xi_j}{2} \neq 0$. Therefore $\hat{p}_0$ is positive definite, and we can use the same transformation as in (3.7) and write (3.15) as

$$\hat{\mathbf{w}}_t = \hat{p}_1 \hat{\mathbf{w}} + \begin{pmatrix} 0 & \hat{p}_0^{1/2}(\omega) \\ -\hat{p}_0^{1/2}(\omega) & 0 \end{pmatrix} \hat{\mathbf{w}}. \tag{3.22}$$

The second term on the right hand side of (3.22) is again skew Hermitian and has no influence on the stability. Thus, we need only consider

$$\hat{\mathbf{w}}_t = \hat{p}_1 \hat{\mathbf{w}},$$

which consists of difference approximations of scalar equations of the above type. To show that the approximations (3.11), (3.12), and (3.13) are stable, we set $\hat{u} = e^{\lambda t}\hat{u}_0$ and get $\lambda = \hat{p}_1$. By (3.17), (3.18), and (3.19), we have $\Re\lambda \leq 0$ and there are no exponentially growing modes.

The approximation (3.9) involves a wide stencil. Therefore extra boundary conditions (ghost points) are required and the resulting accuracy is less than with a more compact stencil. In order to investigate other approximations with a more compact stencil, we write (3.4) as

$$\begin{aligned} \mathbf{u}_{tt} &= 2P_1(\partial/\partial x)\mathbf{u}_t + P(\partial/\partial x)\mathbf{u}, \\ P(\partial/\partial x) &= P_0(\partial/\partial x) - P_1^2(\partial/\partial x) \end{aligned} \tag{3.23}$$

and approximate it by

$$\mathbf{u}_{\nu tt} = 2p_1(D)\mathbf{u}_{\nu t} + p(D)\mathbf{u}_\nu, \tag{3.24}$$

where $p_1(D)$ is given by (3.11) and $p(D)$ is the centered approximation

$$p(D) = \sum_{j=1}^{s} (A_{jj} - \beta^{j2}) D_{+j} D_{-j} + \sum_{j\neq k=1}^{s} (A_{jk} - \beta^j \beta^k) D_{0j} D_{0k}. \tag{3.25}$$

Theorem 3.—The approximation (3.24) is stable if $A_{jj} - \beta^{j2} > 0$.

Proof.—We write (3.24) as

$$\begin{aligned} \mathbf{u}_{\nu tt} &= 2p_1(D)\mathbf{u}_{\nu t} - p_1^2(D)\mathbf{u}_\nu + q(D)\mathbf{u}_\nu, \\ q(D) &= p(D) + p_1^2(D). \end{aligned} \tag{3.26}$$

We use the relation $D_{+j}D_{-j} = D_{0j}^2 - \frac{h_j^2}{4} D_{+j}^2 D_{-j}^2$ and write

$$q(D) = \sum_{j,k=1}^{s} A_{jk} D_{0j} D_{0k} - \frac{1}{4} \sum_{j=1}^{s} (A_{jj} - \beta^{j2}) h_j^2 D_{+j}^2 D_{-j}^2.$$

In the same way as in the continuum case, we write (3.26) as a first order system and Fourier transform to get

$$\begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{pmatrix}_t = \hat{p}_1 \begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{pmatrix} + \begin{pmatrix} 0 & I \\ -\hat{q} & 0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{pmatrix}, \tag{3.27}$$

where

$$\begin{aligned} \hat{q} &= \sum_{j,k=1}^{s} A_{jk} \frac{1}{h_j h_k} \sin\xi_j \sin\xi_k + \sum_{j=1}^{s} (A_{jj} - \beta^{j2}) \frac{4}{h_j^2} \sin^4 \frac{\xi_j}{2}, \\ \xi_j &= \omega_j h_j, \qquad |\xi_j| \leq \pi \end{aligned} \tag{3.28}$$

and $\hat{p}_1$ is given by (3.17). By the ellipticity condition (3.3), it is clear that $\hat{q}$ is a positive definite matrix if $A_{jj} - \beta^{j2} > 0$. Therefore, we can use the same transformation as in (3.7) and write (3.27) as

$$\hat{\mathbf{w}}_t = \hat{p}_1 \hat{\mathbf{w}} + \begin{pmatrix} 0 & \hat{q}^{1/2}(\omega) \\ -\hat{q}^{1/2}(\omega) & 0 \end{pmatrix} \hat{\mathbf{w}}. \tag{3.29}$$

The second term on the right hand side of (3.29) is again skew Hermitian and has no influence on the stability. Thus, we need only to consider

$$\hat{\mathbf{w}}_t = \hat{p}_1 \hat{\mathbf{w}},$$

and the stability follows in the same way as in Theorem 2.

Remark.—If the operator P is elliptic, we have $A_{jj} - \beta^{j2} > 0$, and by Theorem 3 the approximation (3.24) is stable. However, it is possible to have $A_{jj} - \beta^{j2} > 0$ while P is nonelliptic. In this case, the approximation (3.24) remains stable when P is nonelliptic. In other words, the stability of (3.24) does not depend upon the coefficients of mixed derivatives A_{jk}, $j \neq k$.

Remark.—In the scalar case, (3.24) reduces to the W-algorithm (1.8).

In the excision problem, we use the subluminal algorithm (3.24) in the subluminal region where $A_{jj} - \beta^{j2} > 0$. In the superluminal region where the shift β^j is large so that $A_{jj} - \beta^{j2} \leq 0$, we use the superluminal algorithm (3.9) instead. We need then a prescription for switching from one algorithm to the other. There are two distinct ways to do this. One is to make a sharp switch between the algorithms where the transition from superluminal to subluminal region takes place. The other, used in [13], is to introduce a smooth, monotonic blending function and use a blended algorithm, which turns into the superluminal algorithm inside the superluminal region and reduces monotonically to the subluminal algorithm in the outside. For this purpose, note that the superluminal algorithm remains stable in the subluminal region.

As a further alternative to the above approximations, we can approximate (3.23) by adding a fourth differential order term

$$\mathbf{u}_{\nu tt} = 2p_1(D)\mathbf{u}_{\nu t} + p(D)\mathbf{u}_\nu - Q(D)\mathbf{u}_\nu, \qquad (3.30)$$

where $p_1(D)$ is given by (3.11) and

$$Q(D) = \frac{1}{4}\sum_{j=1}^{s} \alpha_j h_j^2 D_{+j}^2 D_{-j}^2, \qquad \alpha_j \geq 0. \qquad (3.31)$$

The motivation for adding such a fourth order term is to modify the matrix $\hat{q}$ in (3.28) so that it becomes positive definite even if $A_{jj} - \beta^{j2} \leq 0$. When $A_{jj} - \beta^{j2} > 0$, the matrix $\hat{q}$ is positive definite and this added term is unnecessary. We can take advantage of this by embedding the switch or blending function in the choice of α_j, with $\alpha_j = 0$ in the outer region.

Theorem 4.—The approximation (3.30) is stable if $A_{jj} + \alpha_j I \geq \beta^{j2} I$.

Proof.—We use the relation $D_{0j}^2 = D_{+j}D_{-j} + \frac{h_j^2}{4} D_{+j}^2 D_{-j}^2$ and write

$$\begin{aligned} p(D) &= \sum_{j,k=1}^{s} (A_{jk} - \beta^j \beta^k) D_{0j} D_{0k} \\ &\quad - \frac{1}{4}\sum_{j=1}^{s} (A_{jj} - \beta^{j2}) h_j^2 D_{+j}^2 D_{-j}^2 \\ &= -p_1^2(D) + \sum_{j,k=1}^{s} A_{jk} D_{0j} D_{0k} \\ &\quad - \frac{1}{4}\sum_{j=1}^{s} (A_{jj} - \beta^{j2}) h_j^2 D_{+j}^2 D_{-j}^2. \end{aligned}$$

We can then write (3.30) as

$$\mathbf{u}_{\nu tt} = 2p_1(D)\mathbf{u}_{\nu t} - p_1^2(D)\mathbf{u}_\nu + q(D)\mathbf{u}_\nu, \qquad (3.32)$$

where

$$\begin{aligned} q(D) &= \sum_{j,k=1}^{s} A_{jk} D_{0j} D_{0k} \\ &\quad - \frac{1}{4}\sum_{j=1}^{s} (A_{jj} - \beta^{j2} + \alpha_j) h_j^2 D_{+j}^2 D_{-j}^2. \end{aligned} \qquad (3.33)$$

In the same way as before, we write (3.33) as a first order system and Fourier transform to get

$$\begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{pmatrix}_t = \hat{p}_1 \begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{pmatrix} + \begin{pmatrix} 0 & I \\ -\hat{q} & 0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{pmatrix}, \qquad (3.34)$$

where

$$\begin{aligned} \hat{q} &= \sum_{j,k=1}^{s} A_{jk} \frac{1}{h_j h_k} \sin\xi_j \sin\xi_k \\ &\quad + \sum_{j=1}^{s} (A_{jj} - \beta^{j2} + \alpha_j) \frac{4}{h_j^2} \sin^4 \frac{\xi_j}{2}. \end{aligned}$$

If $A_{jj} + \alpha_j I \geq \beta^{j2} I$ then, because of ellipticity, $\hat{q}$ is positive definite and stability follows in the same way as before.

B. Half-plane problems

We consider the scalar wave equation with constant coefficients in two space dimensions,

$$u_{\tilde{t}\tilde{t}} = a_1 u_{\tilde{x}\tilde{x}} + 2b_1 u_{\tilde{x}\tilde{y}} + c_1 u_{\tilde{y}\tilde{y}} := P_0 u. \qquad (3.35)$$

In the moving coordinate system, $t = \tilde{t}$, $x = \tilde{x} - \beta^x \tilde{t}$, $y = \tilde{y} - \beta^y \tilde{t}$, with $\beta^x, \beta^y > 0$, we get the shifted wave equation,

$$\begin{aligned} u_{tt} &= 2(\beta^x u_{xt} + \beta^y u_{yt}) + a u_{xx} + 2b u_{xy} + c u_{yy} \\ &:= 2P_1 u_t + Pu. \end{aligned} \qquad (3.36)$$

Here the coefficients $a = a_1 - \beta^{x2}$, $b = b_1 - \beta^x \beta^y$, and $c = c_1 - \beta^{x2}$ are assumed to be constant. Moreover, we assume that the space operator P_0 in (3.36) is elliptic, namely $a_1 > 0$ and $c_1 > 0$ and $b_1^2 < a_1 c_1$. Therefore, by Theorem 1, the Cauchy problem for (3.36) is well-posed.

We consider (3.36) in the half-space

$$0 \leq x < \infty, \qquad -\infty < y < \infty, \qquad t \geq 0$$

and we assume that u is 1-periodic in y. The number of boundary conditions needed at $x = 0$ is equal to the number of outgoing characteristics of the equation $u_{tt} = 2\beta^x u_{xt} + a u_{xx}$. We consider two distinct half-plane problems determined by the coefficients of the operator P.

Half-plane problem I: If $a > 0$ and $b^2 < ac$, then the operator P is elliptic and one boundary condition is needed at $x = 0$. In the excision problem, this is the case of subluminal shift with a timelike boundary.

Half-plane problem II: If $a < 0$, then the operator P is nonelliptic. In the excision problem, this is the case of a superluminal shift with a spacelike boundary.

1. Half-plane problem I (subluminal case)

This is the problem treated in [12] by the energy method. In the present context of (3.36), the energy is given by

$$E = \|u_t\|^2 + a\|u_x\|^2 + 2b(u_x, u_y) + c\|u_y\|^2 \qquad (3.37)$$

in terms of the L_2 scalar product and the corresponding norm

$$(v, w) = \int_0^1 \int_0^\infty v w dx dy, \qquad \|v\|^2 = (v, v). \qquad (3.38)$$

If u solves (3.36), then integration by parts gives

$$\partial_t E = -2u_t(\beta^x u_t + a u_x + b u_y)|_{x=0}. \tag{3.39}$$

Any boundary condition satisfying the dissipative condition $\partial_t E \le 0$ gives an energy estimate sufficient to establish the well-posedness of the Cauchy problem, including the Dirichlet condition

$$u_t(0, y, t) = 0 \tag{3.40}$$

and the Neumann condition

$$\beta^x u_t(0, y, t) + a u_x(0, y, t) + b u_y(0, y, t) = 0 \tag{3.41}$$

for which energy is conserved.

As difference approximation for the half-plane problem, we use (3.24), which in the present case reduces to the W-algorithm (1.8). By introducing a discrete energy norm and using summation by parts, a discrete version of (3.39) has been used to establish stability of the finite difference problem. For details we refer to [12].

2. Half-plane problem II (superluminal case)

To investigate the well-posedness of the continuum problem, we use mode analysis. We apply a Laplace transformation in t and Fourier transformation in y.

Theorem 5.—The half-plane problem (3.36) with $a < 0$ is well-posed.

Proof.—By substituting $u = \hat{u}(x)e^{st+i\omega y}$, $s \in \mathbb{C}$, $\omega \in \mathbb{R}$, into (3.36) we obtain

$$a\hat{u}_{xx} + (2ib\omega + 2\beta^x s)\hat{u}_x + (2i\beta^y \omega s - s^2 - c\omega^2)\hat{u} = 0. \tag{3.42}$$

The general solution to the ordinary differential Eq. (3.42) is of the form $\hat{u}(x) = \sigma_1 e^{\kappa_1 x} + \sigma_2 e^{\kappa_2 x}$, where κ_1 and κ_2 are the solutions of the characteristic equation

$$a\kappa^2 + (2bi\omega + 2\beta^x s)\kappa + 2i\beta^y \omega s - s^2 - c\omega^2 = 0. \tag{3.43}$$

Without restriction we can assume $a = -1$. Moreover, since the sign of $\Re\kappa$ does not depend on ω, we set $\omega = 0$. We then obtain

$$\kappa_{1,2} = \beta^x s \pm \sqrt{(\beta^{x2} - 1)s^2}.$$

For $\Re s > 0$, we have $\Re\kappa_{1,2} > 0$ and there is no bounded solution $\hat{u}$. Therefore no boundary condition is needed and the problem is well-posed.

As difference approximation for the half-plane problem, we can use either (3.9) or (3.30). We study the stability of the approximations by mode analysis. Below we show that (3.9) is stable with $p_1(D)$ in (3.12). The stability of the other approximations with $p_1(D)$ in (3.11) and (3.13) can be shown in the same way.

On a uniform spatial grid $\Omega_h = (\nu h, \mu h)$, $\nu = 0, 1, 2, \ldots, \mu = 1, 2, \ldots, N$, with spacing h, let $v(t) := u_{\nu\mu}(t)$ be the gridfunction approximating $u(x_\nu, y_\mu, t)$. We consider the shifted wave Eq. (3.36) and approximate it by

$$v_{tt} = 2(\beta^x D_{+x} + \beta^y D_{+y})v_t - (\beta^x D_{+x} + \beta^y D_{+y})^2 v + (a_1 D_{+x} D_{-x} + 2b_1 D_{0x} D_{0y} + c_1 D_{+y} D_{-y})v, \tag{3.44}$$

for $\nu = 1, 2, \ldots$. For every fixed μ, we need one extra boundary condition to determine $u_{0\mu}$. We use a third order extrapolation

$$h^3 D_{+x}^3 u_{0\mu} = 0. \tag{3.45}$$

We consider bounded solutions of type

$$u_{\nu\mu}(t) = e^{st+i\omega\mu h}\varphi_\nu, \qquad \|\varphi\|_h < \infty. \tag{3.46}$$

Putting (3.46) into (3.44), we get the eigenvalue problem

$$\begin{aligned}
&\varphi_\nu s^2 - 2\frac{\beta^x}{h}(\varphi_{\nu+1} - \varphi_\nu)s - 2\frac{\beta^y}{h}\left(i\sin\xi - 2\sin^2\frac{\xi}{2}\right)\varphi_\nu s \\
&+ \frac{\beta^{x^2}}{h^2}(\varphi_{\nu+2} - 2\varphi_{\nu+1} + \varphi_\nu) + 2\frac{\beta^x\beta^y}{h^2}(\varphi_{\nu+1} - \varphi_\nu) \\
&\times\left(i\sin\xi - 2\sin^2\frac{\xi}{2}\right) + \frac{\beta^{y^2}}{h^2}\left(i\sin^2\xi - 2\sin^2\frac{\xi}{2}\right)^2\varphi_\nu \\
&- \frac{a_1}{h^2}(\varphi_{\nu+1} - 2\varphi_\nu + \varphi_{\nu-1}) - \frac{b_1}{h^2}i\sin\xi(\varphi_{\nu+1} - \varphi_{\nu-1}) \\
&+ 4\frac{c_1}{h^2}\sin^2\frac{\xi}{2}\varphi_\nu = 0, \qquad \xi = \omega h.
\end{aligned} \tag{3.47}$$

The approximation (3.44) and (3.45) is stable if and only if the Kreiss condition is satisfied, or equivalently if (3.47) has no eigenvalue s with $\Re s \ge 0$ [31]. The constant-coefficient ordinary difference Eq. (3.47) has solution of the form

$$\varphi_\nu = \sum_{j=1}^{3} \sigma_j \kappa_j^\nu,$$

where κ_j are the three solutions of the characteristic equation

$$\begin{aligned}
&s^2 - 2\left(\frac{\beta^x}{h}(\kappa - 1) + \frac{\beta^y}{h}\left(i\sin\xi - 2\sin^2\frac{\xi}{2}\right)\right)s \\
&+ \left(\frac{\beta^x}{h}(\kappa - 1) + \frac{\beta^y}{h}\left(i\sin\xi - 2\sin^2\frac{\xi}{2}\right)\right)^2 \\
&- \frac{a_1}{h^2}\frac{(\kappa - 1)^2}{\kappa} - \frac{b_1}{h^2}\left(\kappa - \frac{1}{\kappa}\right)i\sin\xi + 4\frac{c_1}{h^2}\sin^2\frac{\xi}{2} = 0.
\end{aligned} \tag{3.48}$$

By Lemma 12.1.6 of [31], for $\Re s > 0$ the characteristic Eq. (3.48) has no solutions with $|\kappa| = 1$ and there is exactly one solution with $|\kappa| < 1$. Roughly speaking, the number of left points in the difference stencil determines the number of solutions to the characteristic equation with $|\kappa| < 1$. We call this solution κ_1 and write the bounded solution as

$$u_\nu(t) = e^{st+i\omega\mu h}\sigma_1\kappa_1^\nu. \tag{3.49}$$

By substituting (3.49) into the boundary condition (3.46), we get

$$\sigma_1(\kappa_1 - 1)^3 e^{st+i\omega\mu h} = 0. \tag{3.50}$$

Since $\kappa_1 \neq 1$ for $\Re s > 0$, (3.50) has only the trivial solution $\sigma_1 = 0$. Now, we let $\kappa \to 1$ and investigate if there is any sequence $\{s\}$ such that $\Re s \to 0$ with $\Re s > 0$. We then get from (3.48)

$$\tilde{s}^2 - 2\beta^y\left(i\sin\xi - 2\sin^2\frac{\xi}{2}\right)\tilde{s} + \beta^{y2}\left(i\sin\xi - 2\sin^2\frac{\xi}{2}\right)^2 + 4c_1\sin^2\frac{\xi}{2} = 0, \qquad \tilde{s} = sh, \tag{3.51}$$

and therefore

$$\tilde{s} = \beta^y\left(i\sin\xi - 2\sin^2\frac{\xi}{2}\right) \pm \sqrt{-4c_1\sin^2\frac{\xi}{2}}. \tag{3.52}$$

Since $\beta^y > 0$ and $c_1 > 0$, we have $\Re s < 0$ if $\xi \not\to 0$. In the case where $\xi \to 0$, we get from (3.48)

$$s^2 - 2\frac{\beta^x}{h}s(\kappa - 1) + \frac{\beta^{x2}}{h^2}(\kappa - 1)^2 - \frac{a_1}{h^2}\frac{(\kappa - 1)^2}{\kappa} = 0. \tag{3.53}$$

Letting $s \to 0$, we then get from (3.53) that $\kappa_{1,2} = 1$ and $\kappa_3 = a_1/\beta^{x2} < 1$. Since for $\Re s > 0$ there is no solution with $|\kappa| = 1$, the only solution is κ_3 which is strictly less than 1 and does not converge to 1. Therefore there is no positive sequence $\{s\}$ such that $\Re s \to 0$ for $|\xi| \leq \pi$. Now, we can prove the following theorem:

Theorem 6.—The approximation (3.44) and (3.45) is stable.

Proof.—Since there is no eigenvalue s with $\Re s \geq 0$ to the eigenvalue problem (3.47) giving bounded solutions (3.46), the Kreiss condition is satisfied and stability follows.

IV. TESTS OF THE SUPERLUMINAL ALGORITHMS

In the subluminal case where the evolution proceeds in a timelike direction, the W-algorithm (1.8) provides an accurate, flux-conservative, second order treatment of the IBVP. This was proved for a 1D quasilinear wave equation in [13] using the discrete energy method. In [10,12], the results were extended to the 3D case and applied to the harmonic Einstein system (1.2). The semidiscrete conservation laws extend to the principal part of the harmonic Einstein system and contribute to excellent long term performance in test problems. We use this W-algorithm to treat the outer region of the model excision problem considered in Sec. V.

In this model problem, the inner boundary is chosen to be spacelike, corresponding to the strategy for excising an interior singularity. The evolution near the inner boundary proceeds in a spacelike direction (superluminal shift) so that the spatial grid tracks the boundary. For this superluminal case, the W-algorithm is unstable and one of the algorithms considered in Sec. III must be used. These algorithms are either given by (3.9), with $p_1(D)$ given by one of the approximations (3.11), (3.12), and (3.13), or by (3.30).

In the case of the 2D shifted wave Eq. (1.6), the choice (3.11) reduces to the centered algorithm

$$V := ((\partial_t - \beta^x D_{0x} - \beta^y D_{0y})^2 - a_1 D_{+x}D_{-x} - c_1 D_{+y}D_{-y} - 2b_1 D_{0x}D_{0y})u = 0; \tag{4.1}$$

the choice (3.12) reduces to

$$V_+ := ((\partial_t - \beta^x D_{+x} - \beta^y D_{+y})^2 - a_1 D_{+x}D_{-x} - c_1 D_{+y}D_{-y} - 2b_1 D_{0x}D_{0y})u = 0, \tag{4.2}$$

in which the shift terms are treated by first order accurate one-sided difference operators; the choice (3.13) reduces to

$$V_p := ((\partial_t - \beta^x D_{px} - \beta^y D_{py})^2 - a_1 D_{+x}D_{-x} - c_1 D_{+y}D_{-y} - 2b_1 D_{0x}D_{0y})u = 0, \tag{4.3}$$

in which the shift terms are treated by second order accurate one-sided difference operators (3.14); and (3.30) is related to the subluminal W-algorithm (1.8) by

$$V_\alpha := W + \frac{h^2}{4}(\alpha_1(D_{+x}D_{-x})^2 + \alpha_2(D_{+y}D_{-y})^2)u = 0, \tag{4.4}$$

where Theorem 4 guarantees stability provided the inequalities

$$\alpha_1 \geq \beta^{x^2} - a_1 = -a, \qquad \alpha_1 \geq 0, \tag{4.5}$$

$$\alpha_2 \geq \beta^{y^2} - c_1 = -c, \qquad \alpha_2 \geq 0, \tag{4.6}$$

are satisfied.

The V-algorithm is related to the W-algorithm by the second order accurate modification

$$V = W + \frac{h^2}{4}(\beta^{x^2}(D_{+x}D_{-x})^2 + \beta^{y^2}(D_{+y}D_{-y})^2)u = 0. \tag{4.7}$$

In the subluminal case where the W and V algorithms can be compared, tests show that the W-algorithm has considerably better accuracy due to its more compact stencil [10]. Here we carry out a set of 2D superluminal tests to compare the performance of the superluminal algorithms in a periodic test problem (smooth toroidal boundary conditions) where the effect of the boundary is eliminated. The first order accurate V_+-algorithm (4.2) is highly dissipative and much less accurate than the second order accurate V_p version (4.3). For these reasons, we restrict our test comparisons to the V, V_p and V_α algorithms.

The V-algorithm is a special case of the V_α-algorithm (4.4) where $\alpha_1 = \beta^{x^2}$ and $\alpha_2 = \beta^{y^2}$. The accuracy of the V_α-algorithm might be expected to depend on the relative

weight of the higher order terms responsible for the stretched stencil in (4.4). For example, the V_α-algorithm might be expected to be most accurate for the minimum values, $\alpha_1 = (|a| - a)/2$ and $\alpha_2 = (|c| - c)/2$, which are allowed by (4.5) and (4.6). This is true for the case when a and c are positive, for which $\alpha_1 = \alpha_2 = 0$ and the V_α-algorithm reduces to the W-algorithm.

When $a < 0$ is negative and $c > 0$, the optimal value for α_2 remains 0 but the optimal value for α_1 is not necessarily the minimum allowed value $\alpha_1 = -a$. This value would result in approximating the $a\partial_x^2$ term in the wave operator by aD_{0x}^2, which decouples the even and odd grid points. Although the optimal choice of α_1 in this case is not obvious, the combination $\alpha_1 = \beta^{x^2}$ and $\alpha_2 = 0$ would give better accuracy than the V-algorithm. No general guidelines are suggested by examining the truncation error in the V_α-algorithm, which to order h^2 is given by

$$\tau = \frac{h^2}{12}((a - 3\alpha_1)\partial_x^4 + (c - 3\alpha_2)\partial_y^4 + 4b(\partial_x^3\partial_y + \partial_x\partial_y^3) + 4\beta^x\partial_x^3\partial_t + 4\beta^x\partial_y^3\partial_t)u. \quad (4.8)$$

Note that the values $\alpha_1 = a/3$ and $\alpha_2 = c/3$ correspond to the fourth order accurate approximations to the terms $a\partial_x^2$ and $c\partial_y^2$ in the wave operator. However, these choices are not allowed in the superluminal regime, where stability requires $\alpha_i \geq 0$.

As a test problem for comparing the accuracy of these evolution algorithms in the superluminal regime we pick a case where both a and c are negative. We consider the wave equation

$$(-\partial_t^2 + 4(\partial_x + \partial_y)\partial_t - 3\partial_x^2 - 3\partial_y^2 - 8\partial_x\partial_y)u = 0. \quad (4.9)$$

which arises from a 2D version of (2.4) with shift $\beta^x = \beta^y = 2$. With this superluminal choice of shift, there are no characteristics in the $(x > 0, y > 0)$ directions. Waves propagating along the diagonal have the form

$$u = F[x + y + (4 + \sqrt{2})t] + G[x + y + (4 - \sqrt{2})t]. \quad (4.10)$$

In our test, we simulate the solution

$$u = \sin(2\pi[x + y + (4 + \sqrt{2})t]) \quad (4.11)$$

in the domain $-.5 \leq (x, y) \leq .5$, on a grid with $N = 200$ points, with periodic boundary conditions. For this particular solution, the symmetries $\partial_x u = \partial_y u = \partial_t u/(4 + \sqrt{2})$ imply that the truncation error (4.8) has a minimum at $\alpha_1 = \alpha_2 = \alpha_m$, where

$$\alpha_m = \frac{13 + 8\sqrt{2}}{3} \approx 8.104\,569\,5. \quad (4.12)$$

Figure 1 plots the ℓ_∞ norm of the numerical error in the scalar field obtained in the simulation of (4.11) by evolving the wave Eq. (4.9) with the V_α-algorithm, for various values of $\alpha_1 = \alpha_2 = \alpha$. The error for $\alpha = 8.104\,569\,5$ is extremely small and the plots confirm that α_m is indeed the optimal value. The value $\alpha = 4$ corresponds to the V-algorithm, which gives significantly larger error. The value $\alpha = 3$, which is the smallest value allowed by stability, gives even larger error.

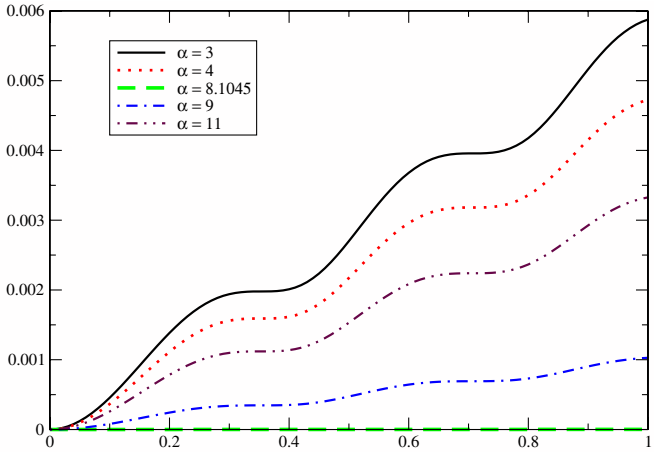


FIG. 1 (color online). The ℓ_∞ norm of the error of the scalar field obtained with V_α algorithm on a grid of 200 points is plotted vs time, in the interval $0 \leq t \leq 1$. For the value $\alpha_m \approx 8.104\,569\,5$, the error is barely discernible and the plots clearly indicate that α_m is the optimal value. The value $\alpha = 4$ corresponds to the V-algorithm, which has significantly larger error.

The error in Fig. 1 is predominantly phase error. Figure 2 shows snapshots of $u(t = 100, x)$ (100 crossing times) for the simulation of (4.11) using the V, V_p and V_α algorithms, with $\alpha = 8$. The simulations are compared with the analytical solution at $t = 100$. The solution with the V_p-algorithm leads in phase while that with the V-algorithm lags in phase and it has slightly better accuracy. As ex-

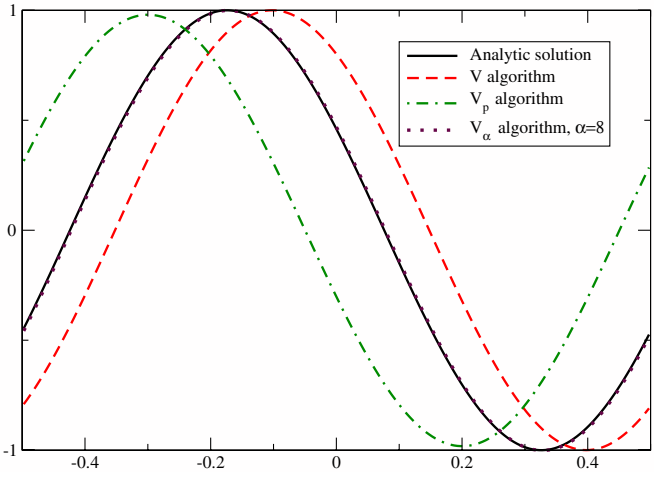


FIG. 2 (color online). Snapshots of the scalar field $u(t = 100, x)$ obtained with the V, V_p and V_α algorithms, compared with the analytic solution. The phase error with the V_p-algorithm is larger than with the V-algorithm. The V_α-algorithm, for $\alpha = 8$, is extremely accurate and barely distinguishable from the analytic solution.

pected from the above error analysis, the V_α-algorithm, with $\alpha = 8$, is extremely accurate.

V. SIMULATION OF A MODEL 2D EXCISION PROBLEM

In this section, we simulate a simple 2D model of the excision problem in which the inner boundary $\mathcal{S}$ is spacelike and the outer boundary $\mathcal{T}$ is timelike, with a horizon $\mathcal{H}$ in between. In the inner region between $\mathcal{S}$ and $\mathcal{H}$, since the shift is superluminal, the operator P in (3.23) is nonelliptic and both characteristics leave the inner boundary. In the outer region between $\mathcal{H}$ and $\mathcal{T}$, since the shift is subluminal, the operator P is elliptic and one characteristic leaves $\mathcal{T}$ and the other enters $\mathcal{T}$.

To model a wave pulse propagating into a horizon, we consider the shifted wave equation with a source term F,

$$u_{tt} = 2(\beta^x u_{xt} + \beta^y u_{yt}) + a u_{xx} + 2b u_{xy} + c u_{yy} + F(x, y, t), \tag{5.1}$$

on the spatial domain $(x, y) \in \Omega = [-2, 2] \times [-2, 2]$, and $t \geq 0$. We set the coefficients $\beta^x = \beta^y = 2$, $a = 0.5(x - \sin\frac{\pi y}{2})$, $b = 0.5$ and $c = 5$, for which the problem is well-posed. The spacelike boundary $\mathcal{S}$ at $x = -2$, the timelike boundary $\mathcal{T}$ at $x = 2$ and the horizon $\mathcal{H}$ are shown in Fig. 3. The horizon satisfies $ac - b^2 = 0$, which determines the curve

$$x = 0.1 + \sin\frac{\pi y}{2}. \tag{5.2}$$

For the smooth function

$$F(x, y, t) = \frac{2}{\sigma^2}(-(1 - 2\beta^x - a)(\sigma - 2(t + x - x_0)^2) - 4(\beta^y + b)(t + x - x_0)y + c(\sigma - 2y^2))e^{-((t+x-x_0)^2+y^2)/\sigma}, \tag{5.3}$$

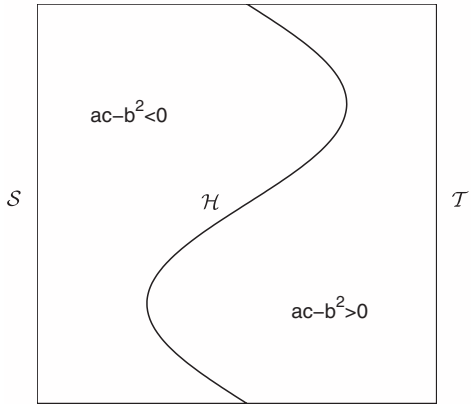


FIG. 3. Computational domain with the spacelike boundary $\mathcal{S}$ on the left, the timelike boundary $\mathcal{T}$ on the right and the sinusoidal shaped horizon $\mathcal{H}$ in between. The solution is periodic in the vertical y-direction.

the Eq. (5.1) has the solution

$$u(x, y, t) = e^{-((t+x-x_0)^2+y^2)/\sigma}, \tag{5.4}$$

which is a left-traveling wave packet, initially centered about $x_0 = 0.5$ outside the horizon and propagating towards the spacelike boundary. Here we set $\sigma = 0.05$. We uniformly discretize the spatial domain as $x_\nu = \nu h$ and $y_\mu = \mu h$ with $\nu, \mu = 0, \pm 1, \ldots, \pm N$ with the grid size $h = \frac{2}{N}$.

The global simulation of the model problem in the region between $\mathcal{S}$ and $\mathcal{T}$ is carried out by combining the superluminal V-algorithms established in Sec. III with the subluminal W-algorithm. The spacelike boundary and the superluminal region are treated with one of the V-algorithms. A region containing the timelike boundary is treated by the W-algorithm.

We consider the following three global algorithms:

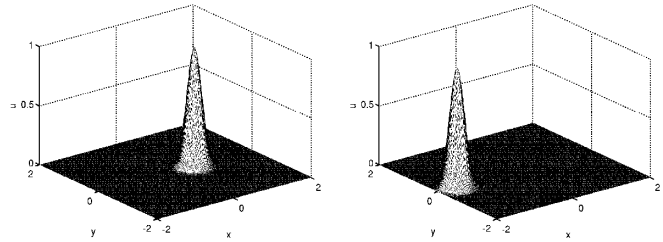
(i) *Algorithm 1.*—The superluminal region is treated by the V-algorithm (4.3). In the subluminal region we use the W-algorithm. We introduce a cutoff function ϕ which is 0 when $a > 0$ and $c > 0$ and is 1 when $a \leq 0$ or $c \leq 0$. Then we use the following approximation

$$\phi V + (1 - \phi)W = 0.$$

(ii) *Algorithm 2.*—This is similar to **1**, except the superluminal region is treated by the V_p-algorithm (4.3), which is then blended to the W-algorithm in the same way as in **1**.

(iii) *Algorithm 3.*—We use the V_α-algorithm (4.4) with $\alpha_1 = (|a| - a)/2$ and $\alpha_2 = (|c| - c)/2$.

The initial data and boundary condition at $x = 2$ are chosen according to the exact solution (5.4). In the first and third algorithms, we need two extra boundary conditions at $\nu = -N, -N + 1$. In the second algorithm, we need only one boundary condition at $\nu = -N$. We use third order extrapolations as the extra boundary conditions. In the y-direction we use periodic boundary conditions. For the integration in time, we use the standard 4th order Runge-Kutta method.



(a) The initial wave pulse at $t = 0$. (b) The wave pulse at $t = 2$.

FIG. 4. A pulse propagating across the horizon. The left figure shows the initial pulse in the region outside the horizon. The right figure shows the pulse at a later time, after it has crossed the horizon and is incident on the inner spacelike boundary.

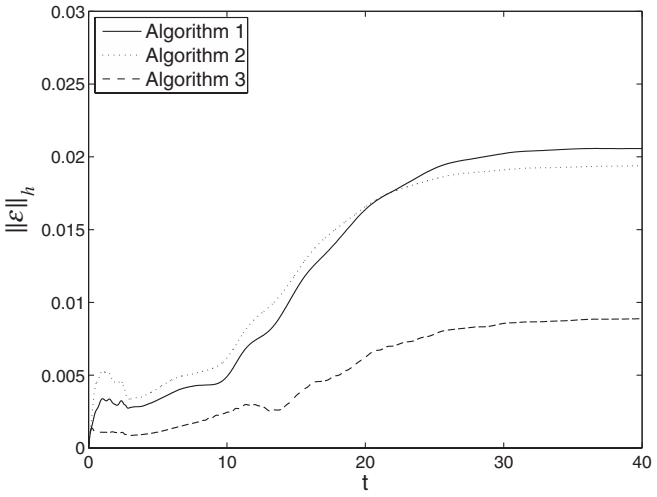


FIG. 5. Norm of the error $\|\mathcal{E}\|_h$ versus time, with $h = 0.02$.

Figure 4 shows the initial wave pulse and the pulse at a later time $t = 2$ computed by the third algorithm.

For the gridfunction $u_{\nu\mu}(t)$ approximating $u(x_\nu, y_\mu, t)$, we define the discrete norm as

$$\|u_{\nu\mu}\|_h^2 = \sum_{\nu,\mu=-N}^{N} u_{\nu\mu} h^2, \tag{5.5}$$

where $h = \Delta x = \Delta y$ is the gridlength. We then define the convergence factor by

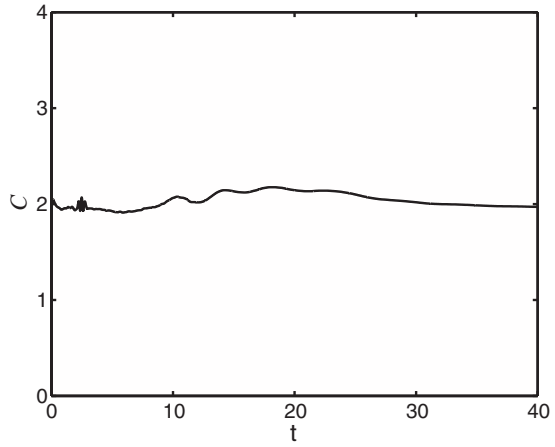
$$\mathcal{C}(t) = \log_2\left(\frac{\|\mathcal{E}(t)\|_h}{\|\mathcal{E}(t)\|_{h/2}}\right), \qquad \mathcal{E}(t) = u(x_\nu, y_\mu, t) - u_{\nu\mu}(t), \tag{5.6}$$

where $\mathcal{E}(t)$ is the error at time t, and $u(x_\nu, y_\mu, t)$ is the exact solution computed by (5.4).

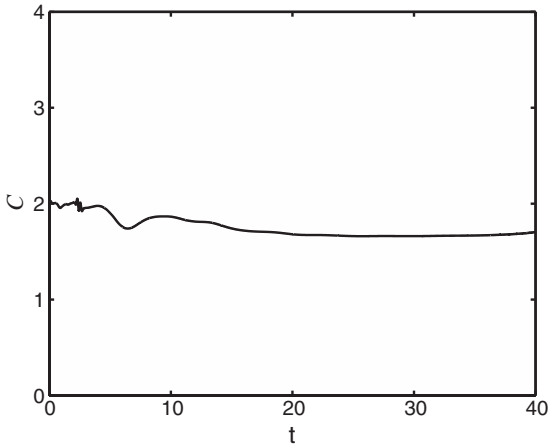
Figure 5 shows the norm of the error versus time for the three algorithms with $h = 0.02$ and $\Delta t = 0.001$.

Figure 6 shows the convergence factor as a function of time for the three algorithms with $h = 0.04$ and $\Delta t = 0.001$. It confirms the second order accuracy of the algorithms in space. The jumps in the convergence factor at about $t = 2$ is a result of using third order extrapolations at the spacelike inner boundary, while we use second order evolution algorithms. At this time the pulse reaches the spacelike boundary and an increase in the order of accuracy, from 2 to 3, is expected.

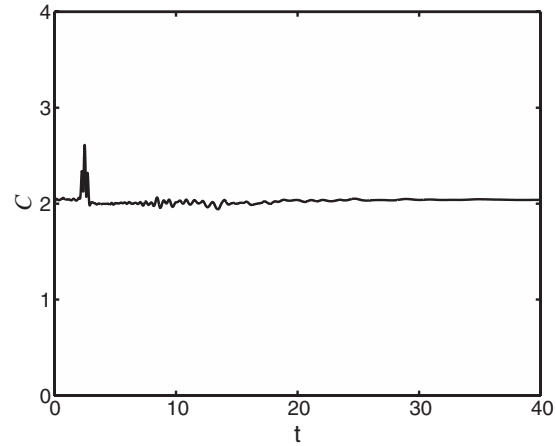
The third algorithm gives a better accuracy than the other two. The second algorithm, in which we use the second order one-sided stencil with extrapolation in one ghost point, gives a slightly smaller error than the first algorithm, in which the second order centered stencil with extrapolation in two ghost points is used.



(a) Algorithm 1.



(b) Algorithm 2.



(c) Algorithm 3.

FIG. 6. Convergence factor $\mathcal{C}$ versus time indicates that the algorithms are second order accurate in space.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant No. PH-0244673 to the University of Pittsburgh. We have used computing resources of the Pittsburgh Supercomputing Center and of the National Energy Research Scientific Computing Center; and we have benefited from the use of the Cactus Computational Toolkit (http://www.cactuscode.org).

[1] T. de Donder, *La Gravifique Einsteinienne* (Gauthiers-Villars, Paris, 1921).
[2] V. Fock, *The Theory of Space, Time and Gravitation* (MacMillan, New York, 1964).
[3] Y. Foures-Bruhat, Acta Math. **88**, 141 (1952).
[4] D. Garfinkle, Phys. Rev. D **65**, 044029 (2002).
[5] B. Szilágyi, B. Schmidt, and J. Winicour, Phys. Rev. D **65**, 064015 (2002).
[6] B. Szilágyi and J. Winicour, Phys. Rev. D **68**, 041501 (2003).
[7] M. C. Babiuc, B. Szilágyi, and J. Winicour, Lect. Notes Phys. **692**, 251 (2006).
[8] F. Pretorius, Classical Quantum Gravity **22**, 425 (2005).
[9] F. Pretorius, Phys. Rev. Lett. **95**, 121101 (2005).
[10] M. C. Babiuc, B. Szilágyi, and J. Winicour, gr-qc/0511154.
[11] L. Lindblom, M. A. Scheel, L. E. Kidder, R. Owen, and O. Rinne, gr-qc/0512093.
[12] M. C. Babiuc, B. Szilágyi, and J. Winicour, Phys. Rev. D **73**, 064017 (2006).
[13] B. Szilágyi, H-O. Kreiss, and J. Winicour, Phys. Rev. D **71**, 104035 (2005).
[14] H. Friedrich, Classical Quantum Gravity **13**, 1451 (1996).
[15] C. Bona, T. Ledvinka, C. Palenzuela, and M. Zácek, Phys. Rev. D **67**, 104005 (2003).
[16] H.-O. Kreiss, N. A. Peterson, and J. Yström, SIAM J. Numer. Anal. **40**, 1940 (2002).
[17] H.-O. Kreiss and O. E. Ortiz, Lect. Notes Phys. **604**, 359 (2002).
[18] A. E. Fisher and J. E. Marsden, Commun. Math. Phys. **28**, 1 (1972).
[19] H.-O. Kreiss, N. A. Petersson, and J. Yström, SIAM J. Numer. Anal. **42**, 1292 (2004).
[20] R. Arnowitt, S. Deser, and C. Misner, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[21] M. Alcubierre and B. Schutz, J. Comput. Phys. **112**, 44 (1994).
[22] G. Calabrese, Phys. Rev. D **71**, 027501 (2005).
[23] G. Calabrese and C. Gundlach, gr-qc/0509119.
[24] M. Alcubierre and B. Brugmann, Phys. Rev. D **63**, 104006 (2001).
[25] H.-O. Kreiss and N. A. Petersson, SIAM J. Sci. Comput. **27**, 1141 (2006).
[26] H. P. Pfeiffer, L. E. Kidder, M. A. Scheel, and S. A. Teukolsky, Comput. Phys. Commun. **152**, 253 (2003).
[27] M. A. Scheel, A. L. Erickcek, L. M. Burko, L. E. Kidder, H. P. Pfeiffer, and S. A. Teukolsky, Phys. Rev. D **69**, 104006 (2004).
[28] G. Calabrese, L. Lehner, O. Reula, O. Sarbach, and M. Tiglio, Classical Quantum Gravity **21**, 5735 (2004).
[29] L. Lehner, D. Neilsen, O. Reula, and M. Tiglio, Classical Quantum Gravity **22**, 5283 (2005).
[30] H.-O. Kreiss and J. Lorenz, *Initial-boundary Value Problems and the Navier-Stokes Equations*, reprint of the 1989 edition (SIAM, Philadelphia, PA, 2004).
[31] B. Gustafsson, H.-O. Kreiss, and J. Oliger, *Time Dependent Problems and Difference Methods* (A Wiley-Interscience Publication, New York, 1995).
[32] H.-O. Kreiss and L. Wu, Applied Numerical Mathematics **12**, 213 (1993).

Paper VI

Hyperbolic Initial Boundary Value Problems which are not Boundary Stable

Mohammad Motamed, Heinz-Otto Kreiss
Department of Numerical Analysis,
Royal Institute of Technology (KTH),
10044 Stockholm, Sweden
E-mail: mohamad@nada.kth.se, hokreiss@nada.kth.se

April 2, 2008

Abstract. The Kreiss symmetrizer technique gives sharp estimates of the solution of hyperbolic initial boundary value problems including estimates at the boundaries resulting in strongly well-posedness in the generalized sense. In this case, the problem is called boundary stable. There are, however, problems which are not boundary stable but are well-posed in a weaker sense, i.e., we can obtain energy estimates in the interior of the domain. We call these problems well-posed in the generalized sense. These types of problems are important in many applications, including seismic, optical and gravitational waves. Examples include surface waves and glancing waves in electromagnetic and elastic wave propagation problems. Unfortunately, there is no general theory for such problems.

In this paper, we consider a model problem which may not be boundary stable depending on the choice of boundary conditions. We show that the general theory of hyperbolic systems can be extended to this case, and the symmetrizer technique can be used to derive estimates of the solution off the boundary and verify well-posedness in the generalized sense.

Keywords. Partial differential equations; Hyperbolic systems; Boundary stable problems; Kreiss symmetrizers; Pseudo-differential operators.

1 Introduction

The theory of linear hyperbolic initial-boundary value problems is well developed for two classes of problems; the Friedrichs theory for symmetric systems with maximally dissipative boundary conditions and the Kreiss theory for hyperbolic systems with boundary conditions satisfying the uniform Kreiss eigenvalue condition.

For first-order symmetric hyperbolic systems with maximally dissipative boundary conditions, an energy estimate can be derived using integration by parts, [3, 4, 14, 5]. However, if the system is not symmetric or the boundary conditions are not maximally dissipative, other techniques are needed.

A rather comprehensive theory has been developed based on the principle of frozen coefficients, Fourier and Laplace transformation, construction of Kreiss-type symmetrizers and the theory of pseudo-differential operators, [9, 18, 2, 20, 19, 15, 11, 10, 13, 16]. This theory gives a necessary and sufficient algebraic condition , known as Kreiss eigenvalue condition, for *strongly well-posedness in the generalized sense*. The theory can also be applied to second-order hyperbolic systems, [12]. For problems with constant coefficients, one can directly derive an estimate for the solution of the

problem which does not rely on the construction of Kreiss symmetrizers and the theory of pseudo-differential operators. However, the importance of symmetrizers is that we can use the theory of pseudo-differential operators and treat systems with variable coefficients.

The Kreiss symmetrizer technique was first introduced by Kreiss for Strictly hyperbolic systems, [9], and was extended to systems with constant multiplicity, [2], and to a special class of systems with variable multiplicity, [16]. It gives sharp estimates including the estimate of the solution at boundaries which result in strong well-posedness in the generalized sense. In this case, the problem is called boundary stable, see [13]. There are, however, problems which are not boundary stable but are well-posed in a weaker sense. We call these problems *well-posed in the generalized sense*. The main purpose of this paper is to extend the theory and construction of symmetrizers for such problems by relaxing the strong eigenvalue condition and deriving estimates of the solution at the interior.

In Section 2, we shortly review the Kreiss theory for boundary stable hyperbolic systems. We introduce another concepts of well-posedness in Section 3, which is desirable for systems which are not boundary stable. We then consider a model problem which may not be boundary stable and discuss different choices of boundary conditions resulting in different types of well-posedness. We show that it is possible to extend the general theory of Hyperbolic systems to the problems which are not boundary stable. A number of auxiliary lemmas are collected in the appendix.

2 Boundary Stable Hyperbolic systems

In this section we give a short review of the Kreiss theory for first order systems which are boundary stable. We note that since the theory is based on the theory of pseudo-differential operators, it can also be applied to second order systems. In fact we can always write a second order system of differential equations as a first order system of pseudo-differential operators, [12].

Consider a first order system of partial differential equations

$$\frac{\partial u}{\partial t} = P(\frac{\partial}{\partial x})u + F(x,t), \qquad P(\frac{\partial}{\partial x}) = A\frac{\partial}{\partial x_1} + \sum_{j=2}^{m} B_j \frac{\partial}{\partial x_j}, \tag{1}$$

where $u(x,t) = \big(u_1(x,t), \ldots, u_n(x,t)\big)^\top$ is a vector-valued function of the real variables $(x,t) = (x_1, \ldots, x_m, t)$, and the coefficient matrices $A, B_j \in \mathbb{C}^{n\times n}$ are constant.

2.1 The Cauchy Problem

We first consider the Cauchy problem for (1) with initial conditions,

$$u(x,0) = f(x), \qquad x \in R, \tag{2}$$

in the space $R: -\infty < x_j < \infty$, $j = 1, \ldots, m$. We denote the L_2 scalar product and the corresponding norm in this space by

$$(u, v)_R = \int_R u^* v dx, \qquad ||u||_R^2 = (u, u)_R,$$

where u^* is the adjoint of u.

For numerical analysis and computations, there is a satisfactory way to define well-posedness as follows.

Definition 1. *The Cauchy problem is called well-posed in the semigroup sense, if*

1. *for a dense set of smooth data, there is a smooth solution,*
2. *the solutions of homogeneous equations ($F \equiv 0$) satisfy the energy estimate*

$$||u(., t)||_R \leq K e^{\alpha(t-t_0)} ||u(., t_0)||_R, \tag{3}$$

where K and α are constants.

The solutions of inhomogeneous systems can be determined and estimated using Duhamel's principle.

Theorem 1. *The Cauchy problem is well-posed in the semigroup sense, if and only if for every real $\omega = (\omega_1, \omega_-)$, $\omega_- = (\omega_2, \ldots, \omega_m)$ with $|\omega| = 1$, the symbol*

$$P(i\omega) = iA\omega_1 + iB(\omega_-), \qquad B(\omega_-) = \sum_{j=2}^{m} B_j \omega_j, \tag{4}$$

has purely imaginary eigenvalues and can be transformed to diagonal form by a transformation $S(\omega)$ with $|S|\,|S^{-1}| \leq K$, where K is a constant independ of ω.

Note that the theorem is true only for problems with constant coefficients. For variable coefficients problems, the symbol should in addition be smoothly symmetrizable in order for the Cauchy problem to be well-posed, [7].

Definition 1 does not only require properties of the eigenvalues of the symbol, but also properties of the eigenvectors. There is a weaker definition used by Hadamard [6] and Petrovskii [17] as follows.

Definition 2. *The Cauchy problem is well-posed in the sense of Hadamard if the estimate (3) is replaced by*

$$||u(., t)||_R \leq K e^{\alpha(t-t_0)} H_p^2(t_0), \quad H_p^2(t) = \sum_{|j| \leq p} ||\frac{\partial^{|j|} u(., t)}{\partial x_1^{j_1} \cdots \partial x_m^{j_m}}||_R^2, \quad p \leq n. \tag{5}$$

One can show that the Cauchy problem is well-posed in the sense of Hadamard if and only if the eigenvalues of the symbol are purely imaginary. However, if the solution of the homogeneous system do not satisfy (3) but only the weaker estimate (5), the well-posedness can be destroyed by lower order terms, [21]. The weaker definition 2, therefore, is not stable against lower order perturbations.

Henceforth, we assume that the system (1) is strictly hyperbolic, i.e., for all real ω with $|\omega| = 1$, the eigenvalues of the symbol (4) are purely imaginary and distinct. By Theorem 1, therefore, the Cauchy problem for (1) is well-posed in the semigroup sense.

We further assume, for simplicity, that A is nonsingular and without restriction assume it has the form

$$A = \begin{pmatrix} -\Lambda^I & \\ & \Lambda^{II} \end{pmatrix}, \tag{6}$$

where Λ^I and Λ^{II} are real positive definite diagonal matrices of order r and $n - r$, respectively. For the singular case see [15].

2.2 The Initial Boundary Value Problem

We now consider the initial boundary value problem (IBVP) for (1) with initial conditions,

$$u(x, 0) = f(x), \qquad x \in R_0, \tag{7}$$

in the half-space $R_0 : x_1 \geq 0$, $-\infty < x_j < \infty$, $j = 2, \ldots, m$, and boundary conditions at $x_1 = 0$,

$$u^I(0, x_-, t) = S\, u^{II}(0, x_-, t) + g(x_-, t). \tag{8}$$

Here $x_- = (x_2, \ldots, x_m)$ denotes a point in the $(m-1)$-dimensional space $R_- : -\infty < x_j < \infty$, $j = 2, \ldots, m$, and $u^I = (u_1, \ldots, u_r)^\top$ and $u^{II} = (u_{r+1}, \ldots, u_n)^\top$ correspond to the partitions Λ^I and Λ^{II}, respectively, and $S \in \mathbb{C}^{r\times(n-r)}$ is a rectangular matrix. All data are smooth, compatible and have compact support.

By a suitable change of variables we can make the boundary conditions homogeneous. We can therefore use Definition 1 also in this case ($F \equiv g \equiv 0$), with $||.||_R$ replaced by $||.||_{R_0}$.

This definition is satisfactory if the system is symmetric hyperbolic and the boundary conditions are of Friedrichs' type. In this case, integration by parts give the energy estimate. But if the system is not symmetric hyperbolic or the boundary conditions are not of Friedrichs' type, another approach needs to be devised.

We consider the IBVP (1), (7), (8) with homogeneous initial data ($f \equiv 0$) and discuss a concept of well-posedness for which we obtain necessary and sufficient conditions.

Definition 3. *Let $f(x) \equiv 0$. We call the IBVP (1), (7), (8) strongly well-posed in the generalized sense if for all smooth compatible data, F and g, there is a unique*

solution u, and in each time interval $0 \le t \le T$, there is a constant K_T independent of the data such that

$$\int_0^t ||u(x,\tau)||^2_{R_0} d\tau + \int_0^t ||u(0,x_-,\tau)||^2_{R_-} d\tau \le K_T \Big(\int_0^t ||F(x,\tau)||^2_{R_0} d\tau + \int_0^t ||g(0,x_-,\tau)||^2_{R_-} d\tau \Big). \tag{9}$$

Here $||.||_{R_0}$ and $||.||_{R_-}$ denote the L_2-norm over the half-space R_0 and the boundary space R_-, respectively.

2.2.1 A Necessary Condition for Strongly Well-posedness in the Generalized Sense

We start with a simple test to derive a necessary condition for the problem to be well-posed. For the IBVP (1), (7), (8) with $F \equiv f \equiv g \equiv 0$, we construct simple wave solutions

$$u(x,t) = e^{st + i\langle \omega_-, x_- \rangle} \, \phi(x_1), \qquad \langle \omega_-, x_- \rangle = \sum_{j=2}^{m} \omega_j x_j, \tag{10}$$

satisfying the boundary conditions

$$\phi^I(0) = S \, \phi^{II}(0), \qquad |\phi|_\infty < \infty. \tag{11}$$

We have

Lemma 1. *(Lopatinsky condition) The half-space problem with $F \equiv f \equiv g \equiv 0$ is not well-posed if for some $\omega_0 \in \mathbb{R}^{m-1}$ and $s_0 \in \mathbb{C}$ with $\Re s_0 > 0$ there is a solution (10) which satisfies (11).*

We can also express the condition as an eigenvalue condition, as follows. We Fourier transform the problem with respect to the tangential variables x_- and get

$$\frac{\partial \hat{u}}{\partial t} = A \frac{\partial \hat{u}}{\partial x_1} + iB(\omega_-)\hat{u} + \hat{F} \qquad \text{for } x_1 \ge 0, \tag{12}$$

$$\hat{u}^I = S \, \hat{u}^{II} + \hat{g} \qquad \text{for } x_1 = 0, \tag{13}$$

where $\hat{u} = \hat{u}(x_1, \omega_-, t) = \mathcal{F}u(x,t) = \int_{R_-} e^{-i\langle \omega_-, x_- \rangle} \, u \, dx_-$, $\hat{F} = \hat{F}(x_1, \omega_-, t) = \mathcal{F}F(x,t)$ and $\hat{g} = \hat{g}(\omega_-, t) = \mathcal{F}g(x_-, t)$ are the Fourier transforms of u, F and g with respect to x_-, respectively. We then Laplace transform it with respect to t and obtain the resolvent equation

$$s\tilde{u} = A \frac{d\tilde{u}}{dx_1} + iB(\omega_-)\tilde{u} + \tilde{F} \qquad \text{for } x_1 \ge 0, \tag{14}$$

$$\tilde{u}^I = S \, \tilde{u}^{II} + \tilde{g} \qquad \text{for } x_1 = 0, \tag{15}$$

with $\tilde{u} = \tilde{u}(x_1, \omega_-, s) = \mathcal{L}\hat{u}(x_1, \omega_-, t) = \int_0^\infty \hat{u}e^{-st}\,dt$, $\tilde{F} = \tilde{F}(x_1, \omega_-, s) = \mathcal{L}\hat{F}(x_1, \omega_-, t)$ and $\tilde{g} = \tilde{g}(\omega_-, s) = \mathcal{L}\hat{g}(\omega_-, t)$ being the Laplace transforms of $\hat{u}$, $\hat{F}$ and $\hat{g}$ with respect to t, respectively. Here $\omega \in \mathbb{R}$ and $s \in \mathbb{C}$.

Let $L_2(0 \le x_1 < \infty)$ be the space of all functions which are quadratically integrable for $0 \le x_1 < \infty$ and denote by

$$(u, v)_0 = \int_0^\infty u^* v dx_1, \qquad ||u||_0^2 = (u, u)_0,$$

the usual scaler product and the corresponding norm in this space. Then $\phi \in L_2(0 \le x_1 < \infty)$ is an eigenfunction of (14), (15) corresponding to an eigenvalue s, if ϕ is the solution of the eigenvalue problem

$$s\phi = A\frac{d\phi}{dx_1} + iB(\omega_-)\phi \qquad \text{for } x_1 \ge 0, \tag{16}$$

$$\phi^I = S\,\phi^{II}, \quad ||\phi||_0^2 < \infty \qquad \text{for } x_1 = 0. \tag{17}$$

We can now formulate the eigenvalue condition equivalent to the Lopatinsky condition,

Lemma 2. *(Eigenvalue condition, Agmon [1]) There are no solution of type (10) which satisfies (11) if and only if the eigenvalue problem (16), (17) has no nontrivial solution with $\Re s > 0$.*

By Lemma 2, the eigenvalue condition is a necessary condition for well-posedness of the IVBP. Hersch [8] has shown that this condition is also sufficient for the problem to be weakly well-posed in the sense of Hadamard. In fact, if there is no solution of type (10) which satisfies (11), the IBVP can be solved by Laplace-Fourier transform. Let $f(x) \equiv 0$. if u is the solution of the IBPV, then $\tilde{u} = \mathcal{L}\mathcal{F}u$ satisfies the resolvent equation (14-15). Conversely, if the eigenvalue condition is satisfied, then one can solve the resolvent equation for $\Re s > 0$. Inverting the Laplace-Fourier transform gives us the solution of the IBVP. However, in general, the eigenvalue condition is not stable against lower order perturbations, and therefore other stable conditions are needed.

We now derive algebraic conditions which determine whether s with $\Re s > 0$ is an eigenvalue.

Let κ be the solutions of the characteristic equation

$$\mathrm{Det}|A\kappa - (sI - iB(\omega_-))| = 0. \tag{18}$$

One can then prove the following lemma, see [9],

Lemma 3. *For solutions , κ, of the characteristic equation (18), we have:*

1. *for $\Re s > 0$, there are no κ with $\Re\kappa = 0$,*

2. *there are precisely r solutions with $\Re\kappa < 0$ and $n - r$ solutions with $\Re\kappa > 0$,*

3. *there exists a constant $\delta > 0$ such that $|\Re\kappa| > \delta\eta$ for all $s = i\xi + \eta$, $\eta > 0$ and all ω_-.*

Assuming all eigenvalues κ_j are distinct, we can write the solution of (16), (17) as

$$\phi = \sum_{\Re\kappa_j<0} \sigma_j e^{\kappa_j x_1}\varphi_j + \sum_{\Re\kappa_j>0} \sigma_j e^{\kappa_j x_1}\varphi_j, \tag{19}$$

where φ_j are the corresponding eigenvectors. Note that if the eigenvalues κ_j are not distinct, the usual modifications apply. Since we are only interested in bounded solutions, we set σ_j in the second term of (19) equal to zero. Introducing ϕ into the boundary conditions (17), we get a linear system of r equations for r unknowns $\sigma = (\sigma_1, \ldots, \sigma_r)$,

$$\tilde{S}(s, \omega_-)\,\sigma = 0.$$

Therefore s with $\Re s > 0$ is an eigenvalue if and only if

$$\mathrm{Det}|\tilde{S}(s, \omega_-)| = 0. \tag{20}$$

Assuming $\mathrm{Det}|\tilde{S}| \neq 0$ for $\Re s > 0$, we know that (14),(15) has a unique solution. Inverting the Fourier and Laplace transforms we obtain the solution to the IBVP.

2.2.2 A Sufficient Condition for Strongly Well-posedness in the Generalized Sense

We now introduce the concept of generalized eigenvalues and derive a sufficient condition such that the problem is strongly well-posed in the generalized sense.

We first introduce normalized variables

$$s' = \frac{s}{\sqrt{|s|^2 + |\omega_-|^2}} = i\xi' + \eta', \qquad \omega'_- = \frac{\omega_-}{\sqrt{|s|^2 + |\omega_-|^2}}, \tag{21}$$

and write the eigenvalue problem (16), (17) in terms of these variables.

Definition 4. *Let $(i\xi'_0, \omega'_{0-})$ be a fixed point, and consider the eigenvalue problem (16), (17) for $s' = i\xi'_0 + \eta'$, $\omega'_- = \omega'_{0-}$, $\eta' > 0$. Then, $(i\xi'_0, \omega'_{0-})$ is a generalized eigenvalue for a boundary condition if in the limit $\eta' \to 0$ the boundary condition is satisfied, or equivalently*

$$\lim_{\eta'\to 0} Det|\tilde{S}(i\xi'_0 + \eta', \omega'_{0-})| = 0.$$

Definition 5. *Let $F \equiv f \equiv 0$. We call the IBVP (1), (7), (8) boundary stable if for all smooth boundary data, g, there is a unique solution u, and in each time interval $0 \le t \le T$, there is a constant K_T independent of the data such that*

$$\int_0^t ||u(0, x_-, \tau)||^2_{R_-} d\tau \le K_T \int_0^t ||g(0, x_-, \tau)||^2_{R_-} d\tau. \tag{22}$$

One can also phrase the boundary stability condition as an eigenvalue condition, [13].

Definition 6. *(Kreiss eigenvalue condition) The eigenvalue problem (16), (17) has no eigenvalue or generalized eigenvalue for* $\Re s \geq 0$.

The boundary estimate (22) is crucial in the theory. It allows us to construct a symmetrizer to obtain an energy estimate in the generalized sense for the full problem.

Using the normalized variables (21), we write (14),(15) as

$$-A\frac{d\tilde{u}}{dx_1} + \sqrt{|s|^2+|\omega_-|^2}\big(s'I - iB(\omega'_-)\big)\tilde{u} = \tilde{F} \qquad \text{for } x_1 \geq 0, \tag{23}$$

$$\tilde{u}^I - S\,\tilde{u}^{II} = \tilde{g} \qquad \text{for } x_1 = 0. \tag{24}$$

We now formulate the main result.

Theorem 2. *Assume that the half-space problem is boundary stable. Then there exists a symmetrizer* $\hat{R} = \hat{R}(s', \omega'_-)$ *with the following properties:*

1. $\hat{R}$ *is uniformly bounded and a smooth function of* s' *and* ω'_- *and of the coefficients* A, B_j *and* S.
2. $\hat{R}A$ *is Hermitian.*
3. *For all vectors* y *satisfying the boundary conditions,*
$$y^*\,\hat{R}A\,y \geq \delta_1|y|^2 - C|\tilde{g}|^2.$$
4. $\sqrt{|s|^2+|\omega_-|^2}\Re\{\hat{R}\big(s'I - iB(\omega'_-)\big)\} \geq \delta_2\eta I$.

We can now prove

Theorem 3. *Assume that the half-space problem is boundary stable (or equivalently the Kreiss eigenvalue condition holds). Then it is strongly well-posed in the generalized sense.*

Proof. Multiplying (23),(24) by $\hat{R}$, we obtain

$$\begin{aligned}
\Re(\tilde{u}, \hat{R}\tilde{F})_0 &= \Re\{-(\tilde{u}, \hat{R}A\frac{d\tilde{u}}{dx_1})_0 + \Big(\tilde{u}, \sqrt{|s|^2+|\omega_-|^2}\hat{R}\big(s'I - iB(\omega'_-)\big)\tilde{u}\Big)_0\} \\
&= \Re\{-\frac{1}{2}\tilde{u}^*\hat{R}A\tilde{u}|_0^\infty + \Big(\tilde{u}, \hat{R}\big(sI - iB(\omega_-)\big)\tilde{u}\Big)_0\} \\
&\geq \frac{1}{2}\delta_1|\tilde{u}(0,\omega_-,s)|^2 + \delta_2\eta||\tilde{u}(x_1,\omega_-,s)||_0^2 - C|\tilde{g}|^2.
\end{aligned}$$

Thus we obtain

$$\eta||\tilde{u}(x_1,\omega_-,s)||_0^2 + |\tilde{u}(0,\omega_-,s)|^2 \leq \text{const.}\big(\frac{1}{\eta}||\tilde{F}||_0^2 + C|\tilde{g}|^2\big).$$

Inverting the Fourier and Laplace transforms proves the theorem. □

Definition 3 has many good properties. It is stable against lower order terms. Moreover, one can construct symmetrizers, and using the theory of pseudo-differential operators, one can treat variable coefficients in a general smooth domain. Rauch [20] has also shown

Theorem 4. *If the problem is strongly well-posed in the generalized sense, then it is well-posed in the semigroup sense.*

3 Well-posed Problems in the Generalized Sense

In the last section, we reviewed the general theory of hyperbolic systems for boundary stable problems. The theory gives necessary and sufficient conditions for the problems to be strongly well-posed in the generalized sense. There are, however, problems which are not boundary stable but are well-posed in a weaker sense. Examples include surface waves and glancing waves in electromagnetic and elastic wave propagation problems described by Maxwell's equations and elastic wave equations with certain types of boundary conditions. It is therefore necessary to develop a theory for such types of problems.

In this section, We first introduce another concept of well-posedness which is again stable against lower order perturbations and is desirable for problems which are not boundary stable. We then consider a model problem which may not be boundary stable and show that it is possible to extend the general theory to this case. We derive necessary and sufficient conditions for the model problem to be well-posed.

We consider the IBVP (1), (7), (8) in the half-space R_0 with homogeneous initial data and boundary conditions ($f \equiv g \equiv 0$). If the problem is well-posed in any of the above senses, then we can solve it by Laplace-Fourier transform. The transformed solution satisfies the resolvent equation (14-15) with $\hat{h} \equiv 0$. If the problem is well-posed in the semigroup sense, then the solution of the resolvent equation satisfy the estimate

$$||\tilde{u}||_0 \leq \frac{K}{\eta - \alpha}||\tilde{F}||_0, \qquad \eta = \Re s > \alpha, \tag{25}$$

i.e.,

$$||(sI - P)^{-1}||_0 \leq \frac{K}{\eta - \alpha}, \qquad P = A\frac{\partial}{\partial x_1} + iB(\omega_-) \tag{26}$$

We call (26) the resolvent condition. By Parseval's relation, (26) is equivalent with

$$\int_0^\infty e^{-2\eta t}||u(.,t)||^2_{R_0}dt \leq \frac{K^2}{(\eta - \alpha)^2}\int_0^\infty e^{-2\eta t}||F(.,t)||^2_{R_0}dt. \tag{27}$$

We now use estimate (27) to define a new definition for well-posedness.

Definition 7. *Let $f(x) \equiv g(x_-, t) \equiv 0$. We call the IBVP (1), (7), (8) well-posed in the generalized sense if estimate (27) holds.*

We shall now consider the following hyperbolic system as a model problem,

$$\frac{\partial u}{\partial t} = A\,\frac{\partial u}{\partial x} + B\,\frac{\partial u}{\partial y} + F(x,y,t), \qquad A = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{28}$$

in the half-space

$$(x,y) \in R_0 = \{(x,y)\,|\,x \geq 0,\, -\infty < y < \infty\}, \qquad t \geq 0.$$

We augment (28) with the initial condition

$$u(x,y,0) = f(x,y), \tag{29}$$

and the boundary condition at $x = 0$,

$$u_1(0,y,t) = a\,u_2(0,y,t) + g(y,t), \qquad a \in \mathbb{C}. \tag{30}$$

Here $u(x,y,t) = (u_1,u_2)^\top$ is a vector-valued function. The data F, f, g are assumed to be compatible smooth functions with compact support. Moreover, we are only interested in solutions with bounded L_2-norm, and, therefore, we assume $||u||_{R_0} < \infty$ for every fixed t.

Fourier transformation in y and Laplace transformation in t gives us

$$\frac{d\tilde{u}}{dx} = M\,\tilde{u} + H, \qquad M = \begin{pmatrix} -s & i\omega \\ -i\omega & s \end{pmatrix}, \qquad H = -A^{-1}\,\tilde{F}, \tag{31}$$

with the boundary condition

$$\tilde{u}_1(0,\omega,s) = a\,\tilde{u}_2(0,\omega,s) + \tilde{g}(\omega,s). \tag{32}$$

We consider bounded solutions $||\tilde{u}||_0^2 = \int_0^\infty |\tilde{u}(x)|^2 dx < \infty$. It can be considered as the boundary condition at infinity.

In order to investigate well-posedness of the IBVP (28)-(30), we first let $F \equiv g \equiv 0$, and construct simple wave solutions of type

$$u(x,y,t) = e^{st+i\omega y}\,\phi(x), \qquad \Re s > 0. \tag{33}$$

We then arrive at the following eigenvalue problem

$$s\phi = A\frac{d\phi}{dx} + i\omega B\phi, \qquad \phi = (\phi_1,\phi_2)^\top, \qquad x \geq 0, \tag{34}$$

$$\phi_1(0) = a\,\phi_2(0), \qquad ||\phi||_0^2 < \infty. \tag{35}$$

Lemma 4. *There is no eigenvalue of (34)-(35) with $\Re s > 0$ only for $a \in \mathbb{R}$ or $|a| \leq 1$.*

Proof. For the proof see Lemma 8.4.2 in [11]. □

As a result of the above lemma, if $|a| > 1$ and $a \notin \mathbb{R}$, there exisit eigenvalues of (34)-(35) with $\Re s > 0$ and therefore by the general theory, the problem is ill-posed. Moreover, if $a \in \mathbb{R}$ or $|a| \leq 1$, we can solve (31)-(32) with $F, g \neq 0$ for $\Re s > 0$. Inverting the Laplace-Fourier transform gives us the solution of the IBVP (28)-(30). However, the problem is well-posed only if we can derive proper estimates of the solution in terms of the data. In order to investigate the well-posedness, we first investigate if there is any generalized eigenvalue.

Lemma 5. *There is no eigenvalue and generalized eigenvalue of (34)-(35) with* $\Re s \geq 0$ *only in the case* $|a| < 1$.

Proof. For the proof see Section 8.4.3 in [11]. □

This lemma shows that the condition $|a| < 1$ is necessary and sufficient for the problem to be boundary stable and therefore strongly well-posed in the generalized sense. There are however two cases for which there exist generalized eigenvalues and therefore by the general theory the problem is not boundary stable:

1. $|a| > 1$, $a \in \mathbb{R}$,
2. $|a| = 1$.

These two cases are fundamentally different, and the corresponding boundary conditions have different types of generalized eigenvalues. As we will show in the following, in the first case the problem is not well-posed, while in the second case the problem is well-posed in the generalized sense.

We first note that the matrix M in (31) has two eigenvalues

$$\kappa_1 = -\kappa, \qquad \kappa_2 = \kappa, \qquad \kappa = \sqrt{s^2 + \omega^2}, \tag{36}$$

and the corresponding eigenvectors are

$$v_1 = \begin{pmatrix} s + \kappa \\ i\omega \end{pmatrix}, \qquad v_2 = \begin{pmatrix} s - \kappa \\ i\omega \end{pmatrix}. \tag{37}$$

For a complex number $z \in \mathbb{C}$, we define the argument of $\sqrt{z}$ by

$$\arg \sqrt{z} = \frac{1}{2} \arg z, \qquad -\pi < \arg z \leq \pi. \tag{38}$$

If the two eigenvalues (36) are distinct ($\kappa \neq 0$), we can diagonalize the matrix M,

$$\Lambda = V^{-1} M V = \begin{pmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{pmatrix}, \qquad V = \begin{pmatrix} v_1 & v_2 \end{pmatrix}, \tag{39}$$

and write (31) and (32) as

$$\frac{d\tilde{v}}{dx} = \Lambda \tilde{v} + G, \qquad \tilde{v} = V^{-1} \tilde{u}, \qquad G = V^{-1} H, \tag{40}$$

$$L_1\,\tilde{v}_1(0,\omega,s)+L_2\,\tilde{v}_2(0,\omega,s)=\tilde{g}(\omega,s),\qquad L_j=s-\kappa_j-i\,a\,\omega,\quad j=1,2. \tag{41}$$

If the two eigenvalues (36) are multiple ($\kappa=0$), we use Schur decomposition, and by a unitary transformation matrix

$$Q=\left((s+\kappa)\,(\bar{s}+\bar{\kappa})+\omega^2\right)^{-1/2}\begin{pmatrix} s+\kappa & i\,\omega \\ i\,\omega & \bar{s}+\bar{\kappa}\end{pmatrix}, \tag{42}$$

we transform the matrix M to an upper triangular matrix

$$T=Q^*MQ=\begin{pmatrix}\kappa_1 & p\\ 0 & \kappa_2\end{pmatrix},\qquad p=-i\,\omega\,\frac{(2s-\bar{s}-\bar{\kappa})\,(\bar{s}+\bar{\kappa})+\omega^2}{(s+\kappa)\,(\bar{s}+\bar{\kappa})+\omega^2}. \tag{43}$$

The equations (31) and (32) are then written as

$$\frac{d\tilde{v}}{dx}=T\,\tilde{v}+G,\qquad \tilde{v}=Q^*\tilde{u},\qquad G=Q^*H, \tag{44}$$

$$L_1\,\tilde{v}_1(0,\omega,s)+L_2\,\tilde{v}_2(0,\omega,s)=\tilde{g}(\omega,s),\quad L_1=s+\kappa-i\,a\,\omega,\quad L_2=-a(\bar{s}+\bar{\kappa})+i\,\omega. \tag{45}$$

We now consider the two cases $|a|>1$, $a\in\mathbb{R}$ and $|a|=1$, separately.

3.1 The case $|a|>1$, $a\in\mathbb{R}$

As was discussed above, the problem is not boundary stable in this case. Here, we will show that the problem is also not well-posed by employing Fourier and Laplace transformations and directly solving the resulting family of ordinary boundary value problems.

We only consider the case when $a\in\mathbb{R}$ and $a>1$. The other case when $a<-1$ is similar. We will first find the generalized eigenvalue. Let $s=i\,\frac{1+a^2}{2a}\omega+\eta$, where $0<\eta\ll|\omega|$. Then

$$\begin{aligned}\kappa=\sqrt{s^2+\omega^2}&\approx\sqrt{\frac{-(a^2-1)^2}{4\,a^2}\,\omega^2+2i\,\eta\frac{1+a^2}{2\,a}\omega}\\ &=i\,\frac{a^2-1}{2\,a}\,|\omega|\sqrt{1-2i\,\eta\frac{2a(1+a^2)}{(a^2-1)^2}\,\frac{1}{\omega}}\\ &\approx i\,\frac{a^2-1}{2\,a}\,|\omega|\left(1-i\,\eta\,\frac{2a(1+a^2)}{(a^2-1)^2}\,\frac{1}{\omega}\right).\end{aligned}$$

Since, by Lemma A3 of the Appendix, $\Re\kappa>0$ for $\Re s>0$, we only consider $\omega>0$ and therefore

$$\kappa\approx i\,\frac{a^2-1}{2\,a}\,\omega+\frac{1+a^2}{a^2-1}\,\eta,\qquad \omega>0. \tag{46}$$

Moreover, since $\kappa \neq 0$, we can diagonalize the system and use (40) and (41) with

$$L_1 = s + \kappa - i\,a\,\omega \approx \eta(1 + \frac{1+a^2}{a^2-1}) + i\,\omega(\frac{1+a^2}{2\,a} + \frac{a^2-1}{2\,a} - a) = \frac{2a^2}{a^2-1}\eta,$$
$$L_2 = s - \kappa - i\,a\,\omega \approx \frac{-2}{a^2-1}\eta - i\,\frac{a^2-1}{a}\omega.$$

We therefore have

$$|L_1| \approx \frac{2a^2}{a^2-1}\eta, \qquad |L_2| \approx \frac{a^2-1}{a}\omega. \tag{47}$$

By Definition 4, for a generalized eigenvalue, we have $L_1 = 0$ in the limit $\eta \to 0$. We have thus proved,

Theorem 5. *The generalized eigenvalue of (40), (41) with $a \in \mathbb{R}$ and $a > 1$ is*

$$s_0 = i\,\xi_0, \qquad \xi_0 = \frac{1+a^2}{2\,a}\,\omega_0. \tag{48}$$

The corresponding eigenfunction is

$$u = e^{i\omega_0(\frac{1+a^2}{2a}t - \frac{a^2-1}{2a}x + y)}. \tag{49}$$

We only need to discuss the estimates of the solution close to the generalized eigenvalue, since by the general theory, away from this eigenvalue the solution is benign. We therefore consider a neighborhood of the generalized eigenvalue (48),

$$s = i\xi_0 + \eta, \qquad \omega = \omega_0, \qquad 0 < \eta \ll 1. \tag{50}$$

Let $f \equiv 0$ in (29). In order to construct estimates for (40)-(41) with $a \in \mathbb{R}$ and $a > 1$, we split the solution into two parts; one solving the equation (40) with $G = 0$ and obeying inhomogeneous boundary condition (41), and the other satisfying the full equation but with homogeneous boundary condition $\tilde{g} = 0$.

We first assume $G = 0$ and $\tilde{g} \neq 0$. From the second equation of (40), we get

$$\tilde{v}_2(x, \omega, s) = 0. \tag{51}$$

Because, otherwise, the solution is not bounded (since $\Re\kappa > 0$). The boundary condition (41) and the relation (47) give us

$$|\tilde{v}_1(0, \omega, s)|^2 \approx \left(\frac{a^2-1}{2a^2}\right)^2 \frac{1}{\eta^2}\,|\tilde{g}|^2. \tag{52}$$

From the first equation of (40),

$$||\tilde{v}_1(x, \omega, s)||_0^2 = \int_0^\infty |\tilde{v}_1(0, \omega, s)\,e^{-\Re\kappa}|^2\,dx = \frac{1}{2\Re\kappa}\,|\tilde{v}_1(0, \omega, s)|^2,$$

and therefore

$$||\tilde{v}_1(x,\omega,s)||_0^2 \approx \frac{(a^2-1)^3}{8a^4\,(1+a^2)}\,\frac{1}{\eta^3}\,|\tilde{g}|^2. \tag{53}$$

We now assume $G=(G_1,G_2)^\top \neq 0$ and $\tilde{g}=0$. By Lemma A1 of the Appendix for the second equation of (40), we obtain

$$|\tilde{v}_2(0,\omega,s)|^2 \leq \frac{2}{\Re\kappa}\,||G_2||_0^2, \qquad ||\tilde{v}_2(x,\omega,s)||_0^2 \leq \frac{1}{(\Re\kappa)^2}\,||G_2||_0^2. \tag{54}$$

For the first equation of (40) we use Lemma A2 and write

$$||\tilde{v}_1(x,\omega,s)||_0^2 \leq \frac{1}{(\Re\kappa)^2}\,||G_1||_0^2 + \frac{1}{2\,\Re\kappa}\,|\tilde{v}_1(0,\omega,s)|^2.$$

Moreover, from the boundary condition (41) and the relation (47) we have

$$|\tilde{v}_1(0,\omega,s)| \approx \frac{(a^2-1)^2}{2\,a^3}\,\frac{\omega}{\eta}\,|\tilde{v}_2(0,\omega,s)|.$$

We therefore obtain

$$|\tilde{v}_1(0,\omega,s)|^2 \leq \frac{(a^2-1)^5}{2a^6\,(1+a^2)}\,\frac{\omega^2}{\eta^3}\,||G_2||_0^2, \tag{55}$$

$$||\tilde{v}_1(x,\omega,s)||_0^2 \leq \left(\frac{a^2-1}{1+a^2}\right)^2\,\frac{1}{\eta^2}\,||G_1||_0^2 + \frac{(a^2-1)^6}{4a^6\,(1+a^2)^2}\,\frac{\omega^2}{\eta^4}\,||G_2||_0^2. \tag{56}$$

By the estimates (51)-(56), we conclude

Theorem 6. *For the solution of (40), (41) with $a \in \mathbb{R}$ and $a > 1$ we have the following estimates near the generalized eigenvalue*

$$\begin{aligned}
|\tilde{v}_1(0,\omega,s)|^2 &\leq C_1\,\left(\frac{\omega^2}{\eta^3}\,||G_2||_0^2 + \frac{1}{\eta^2}\,|\tilde{g}|^2\right),\\
|\tilde{v}_2(0,\omega,s)|^2 &\leq C_2\frac{1}{\eta}\,||G_2||_0^2,\\
||\tilde{v}_1(x,\omega,s)||_0^2 &\leq C_3\,\left(\frac{1}{\eta^2}\,||G_1||_0^2 + \frac{\omega^2}{\eta^4}\,||G_2||_0^2 + \frac{1}{\eta^3}\,|\tilde{g}|^2\right),\\
||\tilde{v}_2(x,\omega,s)||_0^2 &\leq C_4\,\frac{1}{\eta^2}\,||G_2||_0^2,
\end{aligned}$$

where the coefficients $C_1,\ldots,C_4$ are constant. The same estimates follow for $\tilde{u}$.

After inverse Fourier-Laplace transformation, we obtain estimates for the solution to the IBVP (28)-(30). The solution looses one derivative at each reflection from the

boundary. Therefore, if we consider the problem in the strip $0 \le x \le 1$, $-\infty < y < \infty$ and add another boundary condition

$$u_1(1, y, t) = b\, u_2(1, y, t), \qquad |b| > 1,\ b \in \mathbb{R},$$

the solution looses many derivatives as the time goes by. We call the problem *illposed in the asymptotic sense.*

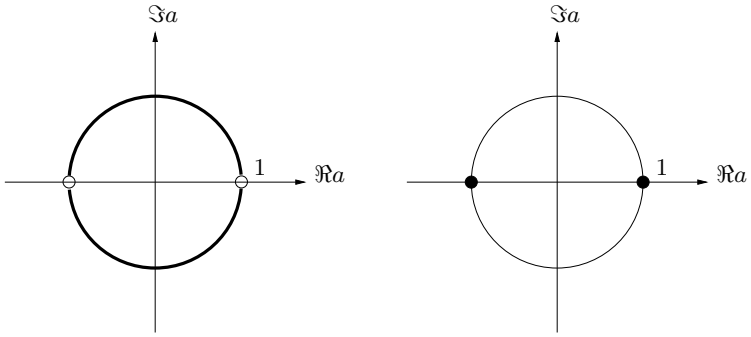
3.2 The case $|a| = 1$

In this case, the hyperbolic IBVP is not boundary stable and no theory exists for investigating the well-posedness. We will show that the Kreiss theory can be applied for this problem. We will prove that the problem is well-posed in the generalized sense by constructing the Kreiss symmetrizers and deriving estimates of type (27) for the solution inside the domain.

There are two different cases as depicted in Figure 1:

i) $|a| = 1$ with $\Im a \neq 0$,

ii) $a = \mp 1$.



(a) $|a| = 1$ and $\Im a \neq 0$ (b) $a = \mp 1$

Figure 1: Two different cases of $|a| = 1$

As we will show, these two different cases correspond to two different types of waves: surface waves and glancing waves which are important phenomena in Elastic wave equations and Maxwell's equations.

3.2.1 Surface Waves, $|a| = 1$ with $\Im a \neq 0$

Let $a = e^{i\theta}$ with $\theta \neq n\pi$, $n = 0, \pm 1, \pm 2, \ldots$. We first find the generalized eigenvalue. By Lemma A4, for $s_0 = i \cos\theta\, \omega_0$ and $\sin\theta\, \omega_0 < 0$, we have $L_1(s_0, \omega_0) = 0$.

Now let $s = \eta + i\,\cos\theta\,\omega$ and $\sin\theta\,\omega < 0$, where $0 < \eta \ll |\omega|$. Then

$$\begin{aligned}\kappa = \sqrt{s^2+\omega^2} &= \sqrt{\sin^2\theta\,\omega^2 + 2i\,\cos\theta\,\eta\,\omega + \eta^2} \\ &\approx |\sin\theta\,\omega|\left(1 + i\,\eta\,\frac{\cos\theta}{\sin^2\theta}\,\frac{1}{\omega}\right).\end{aligned}$$

Since $\sin\theta\,\omega < 0$, we have

$$\kappa \approx -\sin\theta\,\omega - i\,\frac{\cos\theta}{\sin\theta}\,\eta, \qquad \sin\theta\,\omega < 0. \tag{57}$$

We can therefore use (40) and (41) with

$$\begin{aligned}L_1 &= s + \kappa - i\,a\,\omega \approx (1 - i\,\frac{\cos\theta}{\sin\theta})\,\eta, \\ L_2 &= s - \kappa - i\,a\,\omega \approx (1 + i\,\frac{\cos\theta}{\sin\theta})\,\eta + 2\sin\theta\,\omega,\end{aligned}$$

and therefore

$$|L_1| \approx \frac{1}{|\sin\theta|}\eta, \qquad |L_2| \approx 2|\sin\theta|\,|\omega|. \tag{58}$$

We thus have

Theorem 7. *The generalized eigenvalue of (40), (41) with $|a| = 1$ and $\Im a \neq 0$ is*

$$s_0 = i\,\xi_0, \qquad \xi_0 = \cos\theta\,\omega_0, \qquad \sin\theta\,\omega_0 < 0. \tag{59}$$

The corresponding eigenfunction is

$$u = e^{i\omega_0(\cos\theta\,t+y) - |\sin\theta\,\omega_0|\,x}. \tag{60}$$

These eigenfunctions represent *surface waves* which decay exponentially normal to the boundary at $x = 0$. These type of waves are important in many applications (Elastic wave equations).

We will now construct the symmetrizer in a neighborhood of the generalized eigenvalue (59). By the general theory, away from these eigenvalues, there exist smooth symmetrizers and the solution is benign.

We use the normalized variables (21) and write the system (31) in the form

$$\frac{d\tilde{u}}{dx} = \sqrt{|s|^2+|\omega|^2}\,M'\,\tilde{u} + H, \qquad M' = \begin{pmatrix} -s' & i\omega' \\ -i\omega' & s' \end{pmatrix}. \tag{61}$$

Augmented with this system of ODEs, we consider the homogeneous boundary condition (32),

$$\tilde{u}_1(0,\omega,s) = a\,\tilde{u}_2(0,\omega,s). \tag{62}$$

We consider a neighborhood of the generalized eigenvalue,

$$s' = i\cos\theta\,\omega_0' + \eta', \qquad \sin\theta\,\omega_0' < 0, \qquad 0 < \eta' \ll 1.$$

We use the transformation matrix (39) at the generalized eigenvalue

$$V_0 = \omega_0 \begin{pmatrix} i\cos\theta - \sin\theta & i\cos\theta + \sin\theta \\ i & i \end{pmatrix},$$

and transform the system (61) to

$$\frac{d\tilde{v}}{dx} = \sqrt{|s|^2 + |\omega|^2}\,\Lambda'\,\tilde{v} + G, \qquad \tilde{v} = V_0^{-1}\tilde{u}, \qquad G = V_0^{-1}H, \tag{63}$$

where

$$\Lambda' = V_0^{-1}M'V_0 = \begin{pmatrix} i\cot\theta\,\eta' + \omega_0'\sin\theta & (1 + i\cot\theta)\,\eta' \\ (1 - i\cot\theta)\,\eta' & -i\cot\theta\,\eta' - \omega_0'\sin\theta \end{pmatrix}. \tag{64}$$

From the boundary condition (62) we obtain

$$\tilde{v}_2(0, \omega, s) = 0. \tag{65}$$

Following [9], we consider a symmetrizer of the form

$$\tilde{R} = \begin{pmatrix} b & d_1 \\ d_1 & d_2 \end{pmatrix} - i\eta' \begin{pmatrix} 0 & -c \\ c & 0 \end{pmatrix}. \tag{66}$$

Clearly, $\tilde{R}$ is Hermitian. Moreover, we have

$$2\Re(\tilde{R}\,\Lambda') = \begin{pmatrix} 2d_1\eta' + 2b\sin\theta\,\omega_0' & -i\cot\theta\,\eta'(2d_1 - d_2 - b) + (b + d_2)\eta' \\ i\cot\theta\,\eta'(2d_1 - d_2 - b) + (b + d_2)\eta' & 2d_1\eta' - 2d_2\sin\theta\,\omega_0' \end{pmatrix} + \mathcal{O}(\eta'^2).$$

We therefore need to have $2d_1 - d_2 - b = 0$. If we set $b = 0$, $d_2 = 2d_1 > 0$ and $c = 0$, we obtain

$$\Re(\tilde{R}\,\Lambda') = d_1 \begin{pmatrix} \eta' & \eta' \\ \eta' & \eta' - 2\sin\theta\,\omega_0' \end{pmatrix} \geq d_1\,\eta'\,I. \tag{67}$$

For the boundary term, we have by (65)

$$\langle \tilde{v}, \tilde{R}\,\tilde{v}\rangle_{x=0} = d1\,(\tilde{v}_1\,\bar{\tilde{v}}_2 + \bar{\tilde{v}}_1\,\tilde{v}_2) + 2d_1\,|\tilde{v}_2|^2 = 0. \tag{68}$$

We can therefore write

$$\begin{aligned} \Re(\tilde{v}, -\tilde{R}G)_0 &= \Re\{-(\tilde{v}, \tilde{R}\frac{d\tilde{v}}{dx})_0 + \Big(\tilde{v}, \sqrt{|s|^2 + |\omega|^2}\tilde{R}\Lambda'\tilde{v}\Big)_0\} \\ &= \Re\{\frac{1}{2}\langle\tilde{v}, \tilde{R}\tilde{v}\rangle_{x=0} + \Big(\tilde{v}, \sqrt{|s|^2 + |\omega|^2}\tilde{R}\Lambda'\tilde{v}\Big)_0\} \\ &\geq d_1\eta'\sqrt{|s|^2 + |\omega|^2}\,||\tilde{v}(x, \omega, s)||_0^2, \end{aligned}$$

i.e.,

$$\eta' \sqrt{|s|^2 + |\omega|^2}\, ||\tilde{v}||_0^2 \leq C ||\tilde{v}||_0\, ||G||_0.$$

We then obtain the estimate in a neighborhood of the generalized eigenvalue

$$||\tilde{v}||_0 \leq C \frac{1}{\eta}\, ||G||_0. \tag{69}$$

The problem is therefore well-posed in the generalized sense.

3.2.2 Glancing Waves, $a = \pm 1$

We now consider the case when $a = 1$. In order to find the generalized eigenvalue, we let $L_1 = s + \kappa - i\,\omega = 0$ and get $s = i\,\omega$, $\kappa = 0$. Since there are multiple eigenvalues ($\kappa_1 = \kappa_2 = 0$), we consider the system (44) with the boundary condition (45).

Now let $s = i\omega + \eta$ where $0 < \eta \ll |\omega|$. Then

$$\kappa \approx (2i\,\eta\,\omega)^{1/2} = (1+i)|\eta\,\omega|^{1/2}, \qquad \omega > 0. \tag{70}$$

Note that since $\Re\kappa > 0$, we only consider $\omega > 0$. Moreover, we have

$$\begin{aligned} L_1 &= s + \kappa - i\,\omega = \eta + \kappa, \\ L_2 &= -\bar{s} - \bar{\kappa} + i\,\omega = -\eta + 2i\,\omega - \bar{\kappa}, \end{aligned}$$

and therefore

$$|L_1|^2 \approx 2\eta\,\omega, \qquad |L_2|^2 \approx 4\omega^2. \tag{71}$$

We thus have

Theorem 8. *The generalized eigenvalue of (44), (45) with $a = 1$ is*

$$s_0 = i\,\xi_0, \qquad \xi_0 = \omega_0, \qquad \omega_0 > 0. \tag{72}$$

The corresponding eigenfunction is

$$u = e^{i\omega_0(t+y)}. \tag{73}$$

These eigenfunctions represent *glancing waves* which are constant normal to the boundary. These type of waves are important in many applications (Maxwell's wave equations).

We now derive estimates of the solution in a neighborhood of the generalized eigenvalue (72) by constructing symmetrizers in this neighborhood. Let

$$s' = i\omega_0' + \eta', \qquad \omega_0' > 0, \qquad 0 < \eta' \ll 1.$$

We use the transformation matrix (42) at the eigenvalue $s_0 = i\omega_0$

$$Q_0 = \frac{i}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

and transform the system (61) to

$$\frac{d\tilde{v}}{dx} = \sqrt{|s|^2 + |\omega|^2}\, T'\, \tilde{v} + G, \qquad \tilde{v} = Q_0^* \tilde{u}, \qquad G = Q_0^* H, \tag{74}$$

where

$$T' = Q_0^* M' Q_0 = \begin{pmatrix} 0 & -2i\,\omega_0' - \eta' \\ -\eta' & 0 \end{pmatrix}. \tag{75}$$

The boundary condition (62) gives us

$$\tilde{v}_2(0, \omega, s) = 0. \tag{76}$$

Considering a symmetrizer of the form (66), we obtain

$$2\Re(\tilde{R}\, T') = \begin{pmatrix} -2d_1\eta' & -2ib\omega_0' - (b + d_2)\eta' \\ 2ib\omega_0' - (b + d_2)\eta' & -2d_1\eta' - 4c\omega_0'\,\eta' \end{pmatrix} + \mathcal{O}(\eta'^2).$$

We therefore choose $b = 0$. Moreover, choosing $c = 0$, we obtain for the boundary terms,

$$\langle \tilde{v}, \tilde{R}\, \tilde{v} \rangle_{x=0} = 0.$$

Choosing $d_2 = 0$, we therefore obtain

$$\Re(\tilde{R}\, T') = -d_1 \eta' I. \tag{77}$$

We then obtain the same estimate as (69) if we choose $d_1 < 0$.

We summarize the results for the IBVP (28)-(30):

1) if $|a| < 1$, then the problem is strongly well-posed in the generalized sense.

2) if $|a| = 1$, then the problem is well-posed in the generalized sense.

3) if $|a| > 1, a \in \mathbb{R}$, then the problem is ill-posed in the asymptotic sense.

4) if $|a| > 1, a \notin \mathbb{R}$, then the problem is ill-posed in the sense that there are solutions which grow exponentially, arbitrarily fast.

We now formulate the main result.

Theorem 9. *(Main Theorem) Consider the hyperbolic initial boundary value problem (28)-(30), and assume that there is no eigenvalue with $\Re s > 0$ to the correponding eigenvalue problem. Then*

i) if there is no generalized eigenvalue, the problem is strongly well-posed in the generalized sense.

ii) if there exist generalized eigenvalues of either surface-wave modes or glancing-wave modes, the problem is well-posed in the generalized sense.

We conjecture that the theory holds also for general hyperbolic initial boundary value problems (1), (7), (8), which is the topic of future work.

Appendix

In this appendix we collect a number of auxiliary lemmas.

Lemma A1. *Consider the ordinary differential equation $u_x = \lambda\, u + F$ with $\Re\lambda > 0$, $0 \le x < \infty$. Then if the solution $u(x)$ vanishes at infinity, it satisfies the estimate*

$$|u(0)|^2 \le \frac{2}{\Re\lambda}\, ||F||_0^2, \qquad ||u||_0^2 \le \frac{1}{(\Re\lambda)^2}\, ||F||_0^2.$$

Proof. Integration by parts gives us

$$(u, u_x) = -|u(0)|^2 - (u_x, u),$$

i.e.,

$$2\,\Re(u, u_x) = -|u(0)|^2.$$

Therefore

$$\frac{1}{2}\,|u(0)|^2 + \Re\lambda\, ||u||^2 \le ||u||\, ||F||,$$

and the lemma follows. □

Lemma A2. *Consider $u_x = -\lambda\, u + F$ with $\Re\lambda > 0$, $0 \le x < \infty$. Then if the solution $u(x)$ vanishes at infinity, it satisfies the estimate*

$$||u||_0^2 \le \frac{1}{(\Re\lambda)^2}\, ||F||_0^2 + \frac{1}{2\Re\lambda}\, |u(0)|^2.$$

Proof. For $u(0) = 0$, we use integration by parts, and for $F = 0$, we can explicitly calculate the solution. The lemma follows after simple manupulations. □

Lemma A3. *There is a constant $\delta > 0$ such that for all $\omega \in \mathbb{R}$,*

$$\Re\kappa = \Re\sqrt{s^2 + \omega^2} \ge \delta\, \eta, \qquad \eta = \Re s > 0.$$

Proof. For the proof see Lemma 2 of [13]. □

Lemma A4. *Let $a = e^{i\theta}$. For $s_0 = i \cos\theta\, \omega_0$ and $\sin\theta\, \omega_0 < 0$, we have*

$$L_1(s_0, \omega_0) = s_0 + \sqrt{s_0^2 + \omega_0^2} - ia\, \omega_0 = 0.$$

Proof. The lemma follows after simple algebraic manupulations. □

References

[1] S. Agmon. Report. In *Paris Conference on Partial Differential Equations*, 1962.

[2] M.-S. Agranovich. Theorems on matrices depending on parameters and its applications to hyperbolic systems. *Functional Anal. Appl.*, 6(2):85–93, 1972.

[3] K. O. Friedrichs. Symmetric hyperbolic linear differential equations. *Comm. Pure Appl. Math.*, 7:345–392, 1954.

[4] K. O. Friedrichs. Symmetric positive linear differential equations. *Comm. Pure Appl. Math.*, 11:333–410, 1958.

[5] K. O. Friedrichs and P. D. Lax. On symmetrizable differential operators. In *Proc. Symposia in Pure Math., Amer. Math. Soc., Providence, R.I.*, volume 10, 1967.

[6] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. Silliman lectures series. Yale University Publications, 1921.

[7] D. C. Hernquist. Smoothly symmetrizable hyperbolic systems of partial differential equations. *Math. Scand.*, 61:262–275, 1987.

[8] R. Hersch. Mixed problems in several variables. *J. Math. Mech.*, 12:317–334, 1963.

[9] H.-O. Kreiss. Initial boundary value problems for hyperbolic systems. *Comm. Pure Appl. Math.*, 23:277–298, 1970.

[10] H.-O. Kreiss, B. Gustafsson, and J. Oliger. *Time dependent problems and difference methods*. A Wiley-Interscience Publication, 1995.

[11] H.-O. Kreiss and J. Lorenz. *Initial-boundary value problems and the Navier-Stokes equations*. SIAM, 2004.

[12] H.-O. Kreiss and O. E. Ortiz. Some mathematical and numerical questions connected with first and second order time-dependent systems of partial differential equations. *Lect. Notes Phys.*, 604:359–370, 2002.

[13] H.-O. Kreiss and J. Winicour. Problems which are well posed in a generalized sense with applications to the Einstein equations. *Class. Quantum Grav.*, 23:405–420, 2006.

[14] P. D. Lax and R. S. Phillips. Local boundary conditions for dissipative symmetric linear differential operators. *Comm. Pure Appl. Math.*, 13:427–455, 1960.

[15] A. Majda and S. Osher. Initial-boundary value problems for hyperbolic equations with uniformly characteristic boundary. *Comm. Pure Appl. Math.*, 28:607–675, 1975.

[16] G. Métivier and K. Zumbrun. Hyperbolic boundary value problems for symmetric systems with variable multiplicities. *J. Diff. Eqns*, 211:61–134, 2005.

[17] I. G. Petrovskii. On Cauchy's problem for a system of linear partial differential equations in a domain of non-analytical functions. *Moscow Univ. Bull. (A)*, 1(7):1–72, 1938. (In Russian).

[18] J. V. Ralston. Notes on a paper of Kreiss. *Comm. Pure Appl. Math.*, 24:759–762, 1971.

[19] J. Rauch. Energy and resolvent inequalities for hyperbolic mixed problems. *J. Diff. Eqns*, 11:528–450, 1972.

[20] J. Rauch. L_2 is a continuable condition for Kreiss' mixed problems. *Comm. Pure Appl. Math.*, 25:265–285, 1972.

[21] M.. Yamaguti and K. Kasahara. Sur le systeme hyperbolique a coefficients constants. In *Proc. Japan Acad.*, volume 35, pages 547–550, 1959.