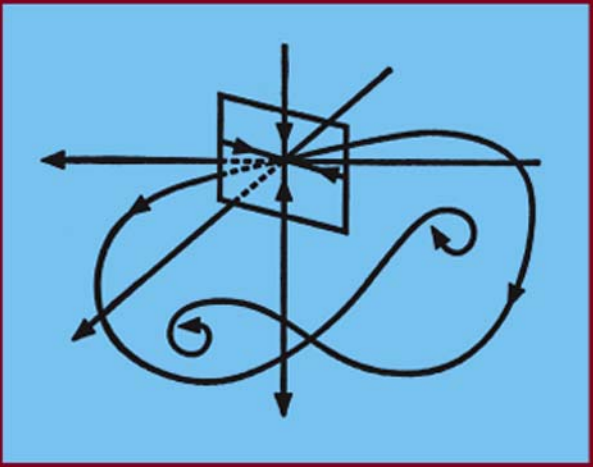
ELSEVIER

Mathematics in Science and Engineering *Volume 213*

series editor: C.K. Chui

Numerical Time-Dependent Partial Differential Equations for Scientists and Engineers



Moysey Brio

Aramais R. Zakharian

Gary M. Webb

Numerical Time-Dependent Partial Differential Equations for Scientists and Engineers

This is volume 213 in
MATHEMATICS IN SCIENCE AND ENGINEERING
Edited by C.K. Chui, *Stanford University*

A list of recent titles in this series appears at the end of this volume.

Numerical Time-Dependent Partial Differential Equations for Scientists and Engineers

Moysey Brio
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ARIZONA
TUCSON, ARIZONA
USA

Aramais R. Zakharian
CORNING INCORPORATED
SP TD 01-1
CORNING, NY 14831
USA

Gary M. Webb
CENTER FOR SPACE PLASMA AND AERONOMIC RESEARCH (CSPAR)
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE
HUNTSVILLE, AL 35805,
USA

ELSEVIER

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
84 Theobald's Road, London WC1X 8RR, UK

First edition 2010

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-133981-4

ISSN: 0076-5392

For information on all Elsevier publications visit our web site at www.elsevierdirect.com

Printed and bound in USA

10 11 12 9 8 7 6 5 4 3 2 1

Preface

Physical experiments and numerical simulations are the primary tools for validation of scientific theories and engineering designs. The aim of this book is to provide the reader with solid understanding of the universal principles that are necessary for successful application and development of numerical methods for partial differential equations. The material presented is based on a year-long graduate course taught over the last two decades at the University of Arizona to the students who are beginning to work on their PhD dissertations. The prerequisites assume familiarity with numerical analysis and partial differential equations on the senior undergraduate or first year graduate level, as currently taught in US universities. In addition to this course, the students in the computational sciences track would normally also take at least one more course covering numerical methods specific to their discipline, e.g. computational electrodynamics, finite element methods in hydrology, direct numerical simulation of turbulent flows, etc.

Our goal is to instill in students the following facts of numerical analysis that are well-known to every practitioner:

– Even though numerical analysis is a very diverse subject, there are universal principles that form the foundation of the field, and lead to time-tested methods that perform well on particular types of problems. Each method has its region of validity, advantages, disadvantages, limitations, difficulties, etc.

– On the physical level, one has to clearly understand what physical scales should be resolved to achieve the convergence regime and which ones will remain unresolved and will be treated phenomenologically or otherwise. Numerical method should treat the unresolved scales/singularities consistently with the appropriate physical laws.

– The total error in the numerical solution will have contributions from all approximations involved: geometry, boundaries and interfaces, initial conditions, sources, material properties, etc. The choice of the numerical method depends on the choice of the key properties that need to be approximated more accurately than the others.

– The stability of a numerical method is usually known only a posteriori, unless the method enforces symmetries, conservation laws, etc., that are sufficient to deduce the stability beforehand.

The question of whether the results of the code make physical sense must be asked. Usually results that defy “common sense” are suspect. However, sometimes surprising results can be correct. As Ortega y Gasset has observed: “If we examine more closely our ordinary notion of reality, perhaps we should find that we do not consider real what actually happens but a certain manner of happening that is familiar. In this vague sense, then, the real is not so much what is seen as forseen; not so much what we see as what (we think) we know”.

The first three chapters of the book are suitable for a one-semester course allocating approximately 4 weeks per chapter. In Chapter One, we focus on basic properties of partial differential equations, including analysis of the dispersion relation, symmetries, particular solutions and linear instabilities. The sections on modulational instabilities and resonant wave interaction are intended as additional reading for the interested students. In Chapter Two we discuss various discretization methods, including finite differences, compact finite differences, finite elements, finite volume and spectral methods. In Chapter Three we present Lax-Richtmyer convergence theory for the initial value problems. After each chapter we provide a sample of the project-like homework assignments allowing students about 3-4 weeks per assignment. The problems are designed to progress the students from the simple task of collecting facts and observations of the properties of the partial differential equations and corresponding numerical methods to the analysis and validation of numerical methods, as well as control and management techniques of the numerical artifacts. Specifically, we focus on such properties as diffusion, damping, dispersion, anisotropies, symmetries, conservation, etc.

The remaining chapters of the book are suitable for a second semester of the course. Here the available theoretical results are not as systematic as for the material covered in the first three chapters. Nevertheless, the understanding of these topics is vital for many numerical applications. In this part of the course we adopt a more heuristic and practical approach. In particular, the following questions are addressed: implementation of transparent and absorbing boundary conditions; practical stability analysis in the presence of the boundaries and interfaces; treatment of problems with different temporal/spatial scales; preservation of symmetries and additional constraints; physical regularization of singularities; resolution enhancement using adaptive mesh refinement; and moving meshes.

We suggest to structure the lectures so as to leave approximately 10 hours of the class time for monthly student presentations. The first two 15-minute presentations should be devoted to the statement of the physical problem and its mathematical model. The choice of the numerical algorithm and its performance on the simplest examples possible, e.g. using advection equation instead of Euler equations of fluid dynamics to illustrate dispersive/diffusive properties of a particular method, are the subject of the second round of student presentations. In the final 20-minute talks the students present the results and conclusion of their study. The papers for the projects that students choose with the consent of the instructor should be on the level of the papers published in Journal of Computational Physics, SIAM Journal on Scientific Computing, or similar high quality computational journals. The only restriction imposed is that the numerical methods in articles have to be related to the material of the course and the proposed work (reproducing the paper or part of it) can realistically be accomplished within a 3-month framework. The oral presentations are accompanied by written reports for each presentation. In the end, the final project reports accumulate to about 15-20 pages total. After each chapter we are providing abbreviated versions of sample projects. Our choice was strongly influenced by our research in space physics, electrical and optical engineering, applied mathematics, numerical analysis and professional software development. These sample projects are intended to serve only as a guide for the students and the instructors.

M.B. would like to acknowledge his family Alina, Ariel and Ashley for their patience and support. We would like to express our gratitude to Prof. Jerry Moloney, director of the Arizona Center of Mathematical Sciences (ACMS), for many years of fruitful collaboration that have presented us with ample opportunities to practice what we preach. We would like to thank Prof. Masud Mansuripur and ACMS members Krishna Mohan Gundu, Colm Dineen, Jorg Hader, Miroslav Kolesik, Patrick Kano, and Hongbo Li for many useful discussions of various numerical analysis issues. We are grateful to Jinjie Liu, who contributed the last two sections of the book on level set and front tracking grid generation methods. A.Z. would also like to acknowledge Prof. Charles Sonett for kindly providing support and mentoring during those first years of his graduate studies at the University of Arizona. G.M.W. acknowledges both stimulating and/or practical discussions (not necessarily both) with colleagues at the University of Arizona (Chuck Sonett, Randy Jokipii, Jozsef Kota, Joe Giacalone) and at the University of California Riverside (Gary Zank, Kobus le Roux, Jim McKenzie, Roma Ratkiewicz, Ming Ko, Nick Pogorelov, Vladimir Florinski, Gang Li, Jacob Heerikhuisen, Dastgeer Shaikh, Brahma Dasgupta, Qiang Hu, Haru Washimi, Olga Verkhoglyadova).

This page intentionally left blank

Contents

Chapter 1

Overview of Partial Differential Equations

1.1 Examples of Partial Differential Equations

Partial differential equations (PDEs) are equations that involve partial derivatives of the unknown quantities. In general, a system of PDEs can be written in the form:

$$G(D^p\mathbf{u}(x), D^{p-1}\mathbf{u}(x), \ldots, \mathbf{u}(x), x) = 0,$$

where $\mathbf{u} = (u^1, u^2, \ldots, u^m)$ denotes the dependent variables; $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denotes the independent variables which vary over some subset of R^n; and G describes functional relationships between the partial derivatives $D^p\mathbf{u}$, where we use the multi-index notation $p = (p_1, p_2, \ldots, p_n)$, and

$$D^p = \frac{\partial^{p_1+p_2+\cdots+p_n}}{\partial x_1^{p_1}\partial x_2^{p_2}, \ldots, \partial x_n^{p_n}}.$$

The order of the highest derivative involved defines the order of the system of the PDEs.

Below we describe common ways PDEs are derived and illustrate them with several examples. In applications, PDEs arise from first principles, which are experimental laws that hold for a wide variety of physical phenomena, like conservation of mass, momentum and energy, Newton's laws of motion, variational principles, etc. Phenomenological principles, are laws that are not as universal as first principles, for example, "the amount of business between two cities is inversely proportional to the geographical distance between them". Balance laws in population dynamics, which express the rates of change of population size and age distributions (dependent variables) as functions of the limited resources, species competition, cannibalism, etc., are further examples of phenomenological laws. Similarly, equations of state in gas dynamics and constitutive relations in elasticity are some other examples of such laws. However, some phenomenological principles can, in many cases, be thought of as a consequence of a physical theory; for example, the ideal gas law can be derived from the kinetic theory of gases.

Asymptotic reduction of PDEs in order to single out particular mechanisms, such as wave interactions, instabilities, and propagation, often leads to canonical PDEs such as Burgers equation, the nonlinear Schrödinger (NLS) equation, the Korteweg deVries (KdV) equation, the Benjamin Ono equation, the Boussinesq equation, three-wave interaction systems, etc. These asymptotic equations are, in general, more amenable to analysis, and are of interest in elucidating the behavior of the system.

In the development of a theory, one can postulate equations that have properties that account for the observations. For example, Schrödinger wanted to obtain an equation that would be dispersive and not diffusive like the heat equation, and incorporate advection like a uni-directional wave equation. Einstein postulated that the field equations of general relativity, namely that the stress energy tensor $T^{\mu\nu}$ of the cosmological fluid is proportional to the Einstein tensor $G^{\mu\nu}$, would be analogous to the familiar relation in mechanics that the stress due tension in a string is proportional to the curvature of the string. In the context of general relativity, the Einstein tensor is formed by contraction of the Riemann curvature tensor, which in turn can be expressed in terms of the metric tensor, and describes the gravitational field (see e.g. [131,188,197]).

Example. Heat Balance Law

Let the scalar quantity $T(x, y, z, t)$ be the temperature in some fixed smooth arbitrary region $V \in R^3$ called the control volume. The rate of change of the heat content or energy at time t is due to the heat flux $\mathbf{F}$ through the boundary ∂V of the region, and to sources or sinks of heat Q present within the region. Thus the conservation of heat energy, assuming there is no mechanical work done on the system, can be written in the form:

$$\frac{\partial}{\partial t}\int_V \rho c T dV = \int_{\partial V} \mathbf{F}\cdot(-\mathbf{n})dS + \int_V Q dV,$$

where ρ and c represent the density and specific heat of the material under consideration, and $\mathbf{n}$ is the outward normal to the surface. To close the system we assume Fourier's constitutive relation between the heat flux and the temperature, namely

$$\mathbf{F} = -\kappa \nabla T,$$

where κ is the heat conduction coefficient. Fourier's law can, in principle, be derived using kinetic theory (e.g. [150]). Other models for the heat flux through the boundary surfaces are possible, e.g. advection and reaction diffusion equations can occur in chemically reacting systems, which are described later in this section. Advection terms also appear in models with a moving control volume $V(t)$.

Using Gauss's divergence theorem on the surface integral above gives the integral form of the heat diffusion equation:

$$\frac{\partial}{\partial t} \int_V \rho c T dV = \int_V \nabla \cdot (\kappa \nabla T) dV + \int_V Q \; dV.$$

Additional assumptions on the smoothness of T, for example, $T \in C^2$, allow one to reduce the above equation locally to the heat conduction equation:

$$(\rho c T)_t = \nabla \cdot (\kappa \nabla T) + Q.$$

Note that the heat conduction coefficient κ may depend in some fashion on x, y, z, t and T, and in some instances should be replaced by a diffusion tensor. For example in plasma physics, the heat conduction coefficient due to the electrons in a fully ionized hydrogen plasma is predominantly along the magnetic field, and is proportional to $T^{5/2}$ [175].

Both the integral and differential forms of the heat conduction equation are used in numerical solutions modeling heat conduction. While the differential form is seemingly simpler, the integral form enforces the original finite volume balance exactly. The integral form only requires the functions to be integrable, and this allows one to seek a solution decomposition for T in terms of, say, piecewise linear functions that provide an easy approximation of the curved boundaries of V as well as simple computation of the weak derivatives, as described later when the finite element and finite volume approximations are discussed.

In the case of a time-independent source, the temperature distribution will, in general, approach a steady state equilibrium as $t \to \infty$. In this case the temperature distribution is found by solving the steady-state heat conduction equation:

$$-\nabla \cdot (\kappa \nabla T) = Q$$

with appropriate boundary conditions at the edge of the control volume.

Example. Advection-reaction-diffusion Equation

If in the previous example we drop the condition that the control volume V is fixed and instead assume that the control volume, $V(t)$, depends on t, and is advected by a known velocity field $\mathbf{v}(x, y, z, t)$, then the left-hand side of the integral equation can be expressed in an alternative form using the Reynolds transport theorem in fluid mechanics (e.g. [68,90]) to obtain the equation:

$$\frac{D}{Dt}\int_V \rho c T dV = \int_{V(t)} \frac{\partial(\rho c T)}{\partial t}\, dV + \int_{\partial V(t)} (\rho c T)(\mathbf{v}\cdot\mathbf{n}) dS,$$

where D/Dt denotes the total time derivative, due to both local time variations, $\partial/\partial t$, keeping $\mathbf{x}$ fixed, and the fact that the control volume $V(t)$ changes in time. Application of the divergence theorem to the last integral yields the advection diffusion equation:

$$(\rho c T)_t + \mathbf{v}\cdot\nabla(\rho c T) + (\rho c T)(\nabla\cdot\mathbf{v}) = \nabla\cdot(\kappa\nabla T) + Q.$$

A polynomial source function Q in the above equation would correspond to an advection-reaction-diffusion equation. The choices $Q \propto (1-T)T$ (logistic growth reaction) and $Q \propto T(T-1/2)(T-1)$ (cubic reaction) are some of the popular modeling choices.

In the case $\nabla \cdot \mathbf{v} = 0$, the control volume does not change as it is advected and the corresponding velocity field is incompressible, and leads to a simpler form of the heat transfer equation.

Example. Maxwell-Cattaneo Heat Flux [153]

Consider the one-dimensional (1D) case of the heat equation:

$$T_t + F_x = 0,$$
$$F = -\kappa T_x.$$

Replacing the last equation (Fourier heat flux) by an assumption that it takes some time τ (the relaxation time) for the gas to adjust to a new thermodynamic equilibrium, gives the Maxwell-Cattaneo heat flux:

$$F(t+\tau) = -\kappa T_x.$$

Assuming τ to be small and keeping only linear terms in the Taylor expansion modifies the heat equation to

$$\begin{aligned} &T_t + F_x = 0, \\ &\tau F_t + \kappa T_x = -F. \end{aligned}$$

In the case where τ and κ are constants, the consistency of the above two equations requires that the temperature T satisfies the telegrapher equation:

$$\tau T_{tt} + T_t - \kappa T_{xx} = 0.$$

The latter equation exhibits wave-like behavior at early times ($t \ll \tau$), and diffusive type behavior at late times ($t \gg \tau$). The telegrapher equation arises in wave propagation along transmission lines in electrical circuit theory, where the diffusive dissipation mechanism is due to electrical resistance, and the wave-like behavior is due to the inductance L and capacitance C [172]. The telegrapher equation has a finite propagation speed for disturbances and the dissipation of the wave depends on the wavelength. Telegrapher equations also describe energetic particle transport in the interplanetary medium (e.g. [65]). The diffusion equation, unlike the telegrapher equation, has an unphysical infinite speed for the propagation of disturbances (see next section). The telegrapher equation has two time scales present, dispersive and diffusive, and in the case where one is only interested in slow diffusive phenomena, the idealized model is much easier to solve numerically than the complete physical model [153].

Example. Discrete Motion on a Lattice [177]

Let u_i^n be the probability that at time $t = \Delta t$ a particle is at site x_i and p_i^n (q_i^n) is the probability of its jumping to the right (left) at time t from location x_i.

Suppose that evolution of the particle satisfies the Chapman-Kolmogorov equation:

$$u_i^{n+1} = p_{i-1}^n u_{i-1}^n + q_{i+1}^n u_{i+1}^n + (1 - p_i^n - q_i^n) u_i^n,$$

which gives the change in probability from one time to the next, due to a particle jumping from the left, right or staying in place.

Introducing the probability density function $u(x,t)$ and transport coefficients $c(x,t)$ and $D(x,t)$,

$$\begin{aligned}
&\lim_{\Delta x\to 0}\sum_{a\le x_i\le b} u_i^n \Delta x = \int_a^b u(x,t)dx,\\
&c(x,t) = \lim_{\Delta t,\Delta x\to 0}\frac{\Delta x}{\Delta t}(p_i^n - q_i^n),\\
&D(x,t) = \lim_{\Delta t,(\Delta x)^2\to 0}\frac{(\Delta x)^2}{\Delta t}\frac{p_i^n + q_i^n}{2},
\end{aligned}$$

and assuming that such a limit exists, we get a PDE called the Fokker-Planck equation:

$$u_t + (c(x,t)u)_x = (D(x,t)u)_{xx}.$$

Note that now we can go backwards and use the Chapman-Kolmogorov model as a discretization of the continuous model, noting again that the continuous model operates on the averaged (smoothed out) quantities. We will investigate in the following chapter the merits of such a discretization.

It is interesting to note that the Fokker-Planck equation can be obtained (e.g. [100]) from the Ito stochastic differential equation, in which u is the probability distribution function for the stochastic process. In the stochastic differential equation approach, x is taken to be a solution of a stochastic differential equation, with both non-random and random components. In general, the drift term $c = \langle \Delta x/\Delta t\rangle$ is an average velocity in the system, which can arise in part from deterministic processes, and in part as an average of stochastic processes. For example, in the motion of a charged particle in a magnetic field in a moving plasma, the particle drift c is due to particles being advected with the background plasma, plus a term due to particles drifting in the non-homogeneous, background magnetic field, as well as a further component due to the interaction of the particle with the random magnetic field (e.g. [91,92]). The second moment $\langle (\Delta x)^2/(2\Delta t)\rangle = D(x,t)$ is due to random or stochastic processes. The diffusion equation is obtained if $c = D_x$. Stochastic differential equations can be related to Monte Carlo methods. Both the Fokker-Planck equation and the Schrödinger equation can be studied using Feynman path integral formulations.

From the definition of the coefficients it follows that, depending on the dominance of the $c(x,t)$ or $D(x,t)$, the equation has drift or diffusion (averaging) as its dominant effect.

Example. Boundary Conditions

To set up the complete model ready for numerical simulation, an initial condition together with the boundary conditions are usually provided. The initial conditions describe the unknown functions and their derivatives of appropriate order. For example, a Gaussian distribution given at $t = 0$ by $T(x, y, z, 0) = \exp[-(x^2 + y^2 + z^2)/2\sigma]$ would be an example of an initial condition. In the numerical solution, a variety of boundary conditions that are physical in nature, or purely numerical due to truncation of the infinite domain of the physical problem, are used. Some common physical boundary conditions for the heat equations are:

(1) $T = g(\mathbf{x}_b, t)$, for $\mathbf{x}_b \in \partial V$ (temperature prescribed on ∂V);
(2) $\frac{\partial T}{\partial \mathbf{n}}|_{\partial V} = g(\mathbf{x}_b, t)$ (prescribed boundary flux on ∂V);
(3) $\kappa_1 \frac{\partial T_1}{\partial \mathbf{n}}|_{\partial V_1} - \kappa_2 \frac{\partial T_2}{\partial \mathbf{n}}|_{\partial V_2} = g(\mathbf{x}_b, t)$ (interface boundary with a source).

In the second example, Newton's cooling law with $g = \alpha(T - T_s)$ is a familiar example, where T_s denotes the surrounding temperature. In the last case, both the flux matching boundary condition with $g = 0$, or the flux matching boundary condition with a source with $g \neq 0$, are encountered in practice.

1.2 Linearization and Dispersion Relation

A useful first step in the analysis of a complicated PDE, or a system of PDEs, is linearization about a known solution. This allows one to obtain information on the local behavior of the small amplitude perturbations with respect to the background solution and to classify the equations accordingly. The procedure consists of substituting the following expansion

$$\mathbf{u}(x) = \mathbf{u}_0 + \epsilon \mathbf{u}_1(x) + \cdots, \quad \epsilon \ll 1,$$

into the original nonlinear PDE(s). Keeping terms of $O(\epsilon)$ results in a linear system of equations for the perturbed quantity $\mathbf{u}_1$. Dropping the subscript 1 on the perturbed quantities $\mathbf{u}_1$, we can write the linearized system in the form:

$$L(\mathbf{x}, D)[\mathbf{u}] = \sum_{|p| \leq m} a_p(\mathbf{x}) D^p(\mathbf{u}) = \mathbf{f}(\mathbf{x}).$$

In the constant coefficient case (assuming no space dependence of the coefficients and zero forcing function), we split the independent variables explicitly into time and space variables, and use the notation $L(\mathbf{x}, D) = \mathcal{L}(\partial_t, \partial_{x_i})$ for the differential operator L.

Next, consider the initial value problem with initial data consisting of a single harmonic

$$\mathbf{u}(x,0)=e^{i\mathbf{k}\cdot\mathbf{x}}.$$

The solution to the linear constant coefficient problem will be a plane wave solution

$$\mathbf{u}(x,t)=e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)},$$

as long as $\omega(\mathbf{k})$ satisfies the dispersion relation

$$\mathcal{L}(-i\omega,i\mathbf{k})=0,$$

which is the familiar heuristic rule in Fourier analysis of replacing time and space derivatives by a multiplication by $-i\omega$ and $i\mathbf{k}$, respectively.

The general initial data may be decomposed into Fourier harmonics and the solution due to linear superposition is given by the Fourier integral:

$$\mathbf{u}(\mathbf{x},t)=\int_{R^n}\hat{\mathbf{u}}_0(\mathbf{k})e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}\ d^n\mathbf{k}. \tag{1.1}$$

In the Fourier representation, ω and $\mathbf{k}$ denote the frequency and wavenumber vector, and $\theta=\mathbf{k}\cdot\mathbf{x}-\omega t$ is the phase of the plane wave component with amplitude $\hat{\mathbf{u}}_0(\mathbf{k})$. A plane wave with constant phase θ moves in the direction $\mathbf{k}$ with the phase velocity $\mathbf{c}=\frac{\omega}{|\mathbf{k}|}\hat{\mathbf{k}}$, where $\hat{\mathbf{k}}$ denotes the unit vector in the $\mathbf{k}$ direction. In the 1D case, the integration path in the Fourier integral (1.1), is over the real k-axis, i.e. $-\infty<k<\infty$, or a path parallel to the Re(k)-axis in the complex k-plane, on which $\hat{u}_0(k)$ is analytic. In addition, the integration path in the complex k-plane, needs to be chosen, so that causality constraints are satisfied. The usual proof of Fourier's theorem assumes that $\hat{\mathbf{u}}(\mathbf{k})$ is square integrable so that Parseval's theorem applies (for further discussion of the conditions for the Fourier inversion theorem to apply, in problems in mathematical physics, see e.g. [136]). Fourier transforms and generalized functions (e.g. Dirac delta distributions and their derivatives) are discussed by Lighthill [115] and Gelfand and Shilov [72].

A wave is called dispersive if the function $\omega(\mathbf{k})$ is a real nonlinear function of the real wavenumber vector $\mathbf{k}$. Diffraction is a phenomena associated with an anisotropic dependence of ω on $\mathbf{k}$. In the examples below, we illustrate that the wave group velocity $\mathbf{V}_g=\nabla_{\mathbf{k}}\omega$ describes the propagation velocity of wave packets as well as the characteristic velocity

Table 1.1: Dispersion relations and plane wave solutions for representative wave equations: 1. advection, 2. Airy, 3. diffraction, 4. wave, 5. Euler-Bernoulli, 6. Klein-Gordon, 7. heat

Equation	Dispersion relation	Plane wave solution
1. $u_t + \mathbf{a} \cdot \nabla u = 0, \mathbf{a} = \text{const}$	$\omega = \mathbf{a} \cdot \mathbf{k}$	$e^{i[\mathbf{k}\cdot(\mathbf{x}-\mathbf{a}t)]}$
2. $u_t + u_{xxx} = 0$	$\omega = -k^3$	$e^{ik(x+k^2t)}$
3. $u_{tx} + u_{yy} = 0$	$\omega = k_y^2/k_x$	$e^{i(k_x x+k_y y-k_y^2 t/k_x)}$
4. $u_{tt} - \Delta u = 0$	$\omega = \pm\lvert\mathbf{k}\rvert$	$e^{ik(\hat{\mathbf{k}}\cdot\mathbf{x}\mp t)}$
5. $u_{tt} + a^2 u_{xxxx} = 0, a \in C$	$\omega = \pm\lvert a\rvert k^2$	$e^{ik(x\mp\lvert a\rvert kt)}$
6. $u_{tt} - u_{xx} + \omega_0^2 u = 0$	$\omega = \pm(k^2 + \omega_0^2)^{\frac{1}{2}}$	$e^{i[kx\mp(k^2+\omega_0^2)^{\frac{1}{2}}t]}$
7. $u_t = \pm\Delta u$	$\omega = \mp i\lvert\mathbf{k}\rvert^2$	$e^{\mp\lvert\mathbf{k}\rvert^2 t}e^{i\mathbf{k}\cdot\mathbf{x}}$

for the transport of wave energy. Note that for dispersive and diffractive waves both phase and group velocities will depend on the wavenumber $\mathbf{k}$, and the phase velocity $\mathbf{V}_p = (\omega/|\mathbf{k}|)\hat{\mathbf{k}}$ and the group velocity $\mathbf{V}_g$ are generally different.

Table 1.1 provides examples of dispersion relations and the corresponding plane wave solutions for scalar wave equations. Several examples of systems of equations are discussed in the examples after the table.

The above table illustrates key wave propagation properties of the equations. For example, the first equation in the table exhibits pure advection; the forward (backward) heat equations (Case 7) show exponential decay (growth) of the high frequencies, respectively. The other examples illustrate different kinds of wave dispersion. In the Klein-Gordon case (Case 6), dispersion vanishes as $\omega_0 \to 0$.

Example. The Telegrapher Equation

The telegrapher equation discussed in Section 1.1, has dispersion equation:

$$\tau\omega^2 + i\omega - \kappa k^2 = 0,$$

with solutions

$$\omega = \frac{i}{2\tau}\left[-1 \pm (1 - 4\kappa\tau k^2)^{\frac{1}{2}}\right].$$

At long wavelengths $k \to 0$, the dispersion equation has approximate solutions

$$\omega_+ = -i\kappa k^2, \quad \omega_- = -\frac{i}{\tau}.$$

The root $\omega_+ = -i\kappa k^2$, corresponds to the dispersion equation for the heat equation; and the other root corresponds to an exponentially damped mode $\propto \exp(-t/\tau)$. At short wavelengths, the dispersion equation has roots $\omega_\pm = \pm k(\kappa/\tau)^{\frac{1}{2}}$, which are the characteristics of the backward and forward solutions of the wave equation with wave speed $c = (\kappa/\tau)^{\frac{1}{2}}$. The latter example illustrates that dispersion equations can exhibit different physical phenomena at different wavelengths.

Example. Hyperbolic Systems

Consider the 1D system

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0,$$

where the constant $n \times n$ matrix $\mathbf{A}$ has real eigenvalues. If the eigenvalues are distinct, the system is called strictly, or strongly, hyperbolic. If the system cannot be diagonalized it is classified as weakly hyperbolic (see next section). The system has solutions for $\mathbf{u}(x,t)$ proportional to a single Fourier harmonic $\mathbf{u}(x,t) = \mathbf{w}e^{i(kx-\omega t)}$, where

$$(\omega\mathbf{I} - k\mathbf{A})\mathbf{w} = 0,$$

$\mathbf{I}$ is the unit $n \times n$ matrix and ω satisfies the dispersion equation:

$$\det(\omega\mathbf{I} - k\mathbf{A}) = 0.$$

In the strongly hyperbolic case, the system can be diagonalized by using the orthonormal left and right eigenvectors $\{\mathbf{l}_i\}$ and $\{\mathbf{r}_i\}$ of the matrix $\mathbf{A}$:

$$(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{r}_i = 0, \quad \mathbf{l}_i(\mathbf{A} - \lambda_i\mathbf{I}) = 0, \quad \mathbf{l}_i \cdot \mathbf{r}_j = \delta_{ij},$$

where δ_{ij} is the Kronecker delta symbol, and the eigenvalues $\{\lambda_i : 1 \le i \le n\}$ satisfy the determinant equation:

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0.$$

After multiplying the original equations by the left eigenvectors $\mathbf{l}_i$, the system decouples into n scalar equations

$$\partial_t v_i(x,t) + \lambda_i \partial_x v_i(x,t) = 0, \quad i = 1, 2, \ldots, n,$$

where the characteristic variables $v_i(x,t)$ are defined by $v_i(x,t) = \mathbf{l}_i \cdot \mathbf{u}(x,t)$.

This implies that the general solution for $\mathbf{u}$ can be written in the form:

$$\mathbf{u}(x,t) = \sum_{i=1}^{n} v_i(x - \lambda_i t)\mathbf{r}_i.$$

Note that the different modes propagate without any interaction, and the amplitudes v_i are just advected with velocity λ_i along the ith characteristic, without any change. Thus if only one of the amplitudes is non-zero initially, one obtains only a single propagating mode. This is in contrast to wave propagation through an inhomogeneous medium. For example, for waves propagating on an inhomogeneous string, the backward and forward propagating modes interact with each other due to the inhomogeneous background medium, leading to reflection, transmission, and wave interaction phenomena.

In the multi-dimensional case,

$$\partial_t \mathbf{u}(\mathbf{x},t) + \sum_{i=1}^{n} \mathbf{A}_i \partial_{x_i} \mathbf{u}(\mathbf{x},t) = 0,$$

and the dispersion relation becomes

$$\det\left(\omega I - \sum_{i=1}^{n} k_i \mathbf{A}_i\right) = 0.$$

Example. Group Velocity and Wave Packets

Dispersive systems have an important property that the group velocity describes the velocity of a quasi-monochromatic wave packet. It is also the characteristic velocity for the transport of wave energy. Using the Fourier integral solution (1.1), one finds for a quasi-monochromatic wave packet, centered on $k = k_0$, and with a spread in k-space of δk about $k = k_0$ (where $|\delta k| \ll k_0$), that for x and t both large with $x/t = \text{const.}$, expanding $\omega(k)$ about $k = k_0$ in truncated Taylor series of the form:

$$\omega(k) = \omega_0 + \omega'(k_0)(k - k_0) + \frac{1}{2}\omega''(k_0)(k - k_0)^2,$$

yields the approximate solution for $u(x,t)$ of the form:

$$u \simeq \hat{u}_0(k_0)\left(\frac{2\pi}{|\omega_0'' t|}\right)^{\frac{1}{2}} \exp\left[i(k_0 x - \omega_0 t) - i\ \mathrm{sgn}\left(\omega_0''\right)\frac{\pi}{4}\right] \\ \exp\left(-i\frac{(x - \omega_0' t)^2}{2\omega_0'' t}\right) + c.c.,$$

where $\mathrm{sgn}(x)$ denotes the sign of x. The above approximate solution for

$u(x,t)$ shows that the wave envelope moves with the group velocity $V_g = \omega_0'$, and is damped like $t^{-\frac{1}{2}}$ at late times, and that there is a characteristic phase shift of $\pm\pi/4$ for the carrier wave with frequency ω_0 and wavenumber k_0. A similar result can be obtained by using the method of stationary phase or the saddle point method (e.g. [198, Equation (11.28), Section 11.4]), but the result in Whitham's book does not have the second exponential term in the above equation since the stationary phase analysis applies to a point moving at the group velocity with $x = \omega_0' t$. Also note that the analysis assumes that $\omega_0'' \neq 0$, and higher derivatives of ω are needed if $\omega_0'' = 0$.

The dispersion is called anomalous if the phase and group velocities have the opposite sign. Note that the envelope is moving with the group velocity, while $\omega''(k_0)$ (group velocity dispersion) influences the spreading of the envelope. In the nonlinear case the canonical equation describing the balance between nonlinear steepening and quadratic dispersion is the NLS equation. The NLS equation has *traveling wave* solutions and *soliton* solutions, which arise due to the balance between nonlinearity and dispersion (Section 1.4).

Another example in which the phase and group velocities have distinctly different behavior is that of short wavelength internal gravity waves in a stably stratified fluid (e.g. [198, Section 12.7, p. 421]; [9, Section 10.4, p. 161]). The effect of compressibility on short wavelength internal waves is small [117]. By eliminating compressibility effects associated with sound waves, one obtains the dispersion equation for internal gravity waves as:

$$\omega^2 = \frac{\omega_0^2 k_x^2}{k_x^2 + k_y^2 + \alpha^2/4}, \tag{1.2}$$

where

$$\omega_0^2 = \alpha g \quad \text{and} \quad \alpha = -\frac{1}{\rho_0(y)} \frac{d\rho_0(y)}{dy},$$

defines the Brunt-Väisälä frequency ω_0, in terms of the gravitational acceleration g and the vertical density gradient α (it is assumed that $\alpha > 0$). The Brunt-Väisälä frequency, or buoyancy frequency, is the frequency of vertical oscillations in the stratified fluid. The above dispersion equation shows that wave propagation is possible only for waves with frequencies below the Brunt-Väisälä frequency (i.e. for $|\omega| < \omega_0$). For waves with wavenumbers $k \gg \alpha$ (i.e. for waves with wavelength that is short compared to the stratification scale height $h = 1/\alpha$), the group velocity $\mathbf{V}_g = \partial\omega/\partial\mathbf{k}$

has components:

$$V_{gx} \approx \frac{\omega_0 k_y^2}{k^3}, \quad V_{gy} \approx -\frac{\omega_0 k_y k_x}{k^3}.$$

Thus for short wavelength waves ($k \gg \alpha$), the group velocity is essentially perpendicular to the phase velocity (i.e. $\mathbf{k} \cdot \mathbf{V}_g = 0$).

Example. Recovering a PDE from its Dispersion Relation

Using the correspondence $\partial/\partial t = -i\omega$, and $\partial/\partial x = ik$, allows one to reconstruct a linear PDE from a rational dispersion relation. For example, $\omega = k^2$ leads to the Schrödinger equation: $iu_t = -u_{xx}$. If we saturate the growth at large k by modifying the dispersion relation to $\omega = \frac{k^2}{(1+\epsilon k^4)}$, $\epsilon \ll 1$, then the Schrödinger equation correspondingly becomes modified to

$$i(\partial_t + \epsilon \partial_{xxxxt})u = -\partial_{xx} u.$$

Example. Wave Diffraction

Consider the two-dimensional (2D) wave equation:

$$u_{tt} = u_{xx} + u_{yy}$$

with dispersion relation

$$\omega^2 = k_1^2 + k_2^2,$$

where we use the notation $(x, y) = (x_1, x_2)$ and $(k_x, k_y) = (k_1, k_2)$. If the wave is almost a plane wave propagating in the x-direction, and is weakly curved in the y-direction, with $|k_2| \ll |k_1|$, then the dispersion relation may be approximated by

$$\omega = k_1 \sqrt{1 + \frac{k_2^2}{k_1^2}} \approx k_1 + \frac{1}{2}\frac{k_2^2}{k_1}.$$

The resulting PDE is the linearized 2D Burgers equation:

$$(u_t + u_x)_x + \frac{1}{2} u_{yy} = 0,$$

where the term $u_t + u_x$, describes the fact that the wave propagates as a uni-directional wave along the x-axis, and the u_{yy} term describes wave diffraction associated with the weakly curved wavefront. Looking for time harmonic solutions

$$u = A(x,y)e^{i(x-t)},$$

we obtain the Schrödinger equation:

$$iA_x + A_{xx} + \frac{1}{2}A_{yy} = 0.$$

This "parabolic approximation" is frequently used in wave propagation problems, where it provides substantial savings in computational effort (e.g. [184]).

Example. The 2D Schrödinger Equation

Consider a dispersive, weakly 2D, diffracting wave packet, in which the carrier wave has frequency ω_0 and wavenumber $\mathbf{k}_0$. The Taylor series expansion of the dispersion equation $\omega = \omega(k_1, k_2)$ about $\mathbf{k} = \mathbf{k}_0$ has the form:

$$\begin{aligned}\omega = {} & \omega_0 + \Delta k_1\omega_1 + \Delta k_2\omega_2 \\ & + \frac{1}{2}\left[\omega_{11}(\Delta k_1)^2 + \omega_{22}(\Delta k_2)^2 + 2\omega_{12}\Delta k_1\Delta k_2\right] + O\left[(\Delta k)^3\right].\end{aligned}$$

Here we use the notation $\Delta k_j = k_j - k_{j0}$, and $\omega_j = \partial\omega/\partial k_j$, $\omega_{ij} = \partial^2\omega/\partial k_i\partial k_j$ $(i,j = 1,2)$ denote the partial derivatives of ω with respect to $\mathbf{k}$, evaluated at $\mathbf{k} = \mathbf{k}_0$. It is assumed that $|\Delta k_j/k_0| \ll 1$ and $|\Delta k_1| \sim |\Delta k_2|$. The above expansion corresponds to a diffracting wave packet:

$$\psi = e^{i(\mathbf{k}_0\cdot\mathbf{x} - \omega_0 t)}u(x,t) + \text{c.c.},$$

where the wave envelope $u(x,t)$ satisfies the 2D Schrödinger equation:

$$i(u_t + \mathbf{V}_g\cdot\nabla u) + \frac{1}{2}\left(\omega_{11}u_{xx} + \omega_{22}u_{yy} + 2\omega_{12}u_{xy}\right) = 0,$$

and $\mathbf{V}_g = \omega_1\mathbf{e}_x + \omega_2\mathbf{e}_y$ is the group velocity.

The above generalized Schrödinger equation can be simplified by using the rotated coordinates

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix},$$

where

$$\tan 2\alpha = \frac{2\omega_{12}}{\omega_{11} - \omega_{22}}.$$

In the new coordinates we obtain the 2D Schrödinger equation:

$$i\left(u_t+\mathbf{V}_g'\cdot\nabla' u\right)+\frac{1}{2}\left(\lambda_1 u_{x'x'}+\lambda_2 u_{y'y'}\right)=0.$$

In the above equation, λ_1 and λ_2 are the eigenvalues of the Hermitian matrix

$$\mathbf{H}=\begin{pmatrix}\omega_{11} & \omega_{12}\\ \omega_{12} & \omega_{22}\end{pmatrix}.$$

The eigenvalues are given by the formulae:

$$\lambda_1=\omega_{11}+\omega_{12}\tan\alpha,\quad \lambda_2=\omega_{11}-\omega_{12}\cot\alpha.$$

Related wave diffraction phenomena, associated with the 1+3D inviscid Burgers equation, are discussed in Section 1.5, where the role of the group velocity wave action and wave energy equations in the WKB approximation are discussed.

1.3 Well-posedness, Regularity and the Solution Operator

A problem is called well-posed if it has a unique solution that depends continuously on the data, i.e. initial, boundary, and forcing data, or any other data. For example, the continuous dependence on the initial data with respect to a norm $\|.\|$, means that for $0\le t\le T$, where T is a fixed time, there is a constant $C(T)$, independent of the initial data, such that

$$\|w(t)-v(t)\|\le C(T)\|w(0)-v(0)\|,$$

where w and v are two solutions of the same problem.

Definition 1. *The Lebesgue-norm $\|\ \|_p$ of a Lebesgue measurable function u, on the interval $[a,b]$, is defined by*

$$\|u\|_p=\left(\int_a^b |u|^p\,dx\right)^{1/p},$$

where p is a positive integer. If the integral exists and is finite, the function u is said to be a member of the Lebesgue function class $L_p[a,b]$.

Comment

One of the most common norms used is the L_2 norm of square integrable functions. Another norm that is sometimes used is the L_∞ norm. When the essential-sup norm $\|.\|_\infty$ exists, it is equivalent to the supremum, or *sup* norm, defined as

$$\|u\|_{sup} = \sup_{a \le x \le b} |u(x)|,$$

where sup denotes the supremum of the function on the interval $[a, b]$.

For linear equations, the difference of two solutions is a solution and the above inequality can be stated as

$$\|u(t)\| \le C(T)\|u(0)\|.$$

When Fourier transform methods are applicable, the L_2 norm might be suitable since Parseval's identity $\|u\|_2 = \|\hat{u}\|_2$ applies to square integrable functions, where $\hat{u}$ is the Fourier transform of $u(\mathbf{x})$.

Example

An initial value problem for the Schrödinger equation $iu_t = -u_{xx}$ on a periodic domain has the property that

$$\frac{d}{dt}\|u(t)\|^2 = \left(\frac{d}{dt}u, u^*\right) + \left(u, \frac{d}{dt}u^*\right) = 0,$$

where the inner product (f, g) is defined by

$$(f, g) = \int_a^b f(x)g^*(x)\, dx,$$

and the interval (a, b) can be infinite or finite. Thus the initial value problem is well-posed in the L_2 norm.

The solution,

$$u = \sqrt{\frac{T}{T-t}} \exp\left(\frac{ix^2}{4(T-t)}\right), \quad \text{where } 0 \le t < T,$$

for arbitrary T is an exact solution of Schrödinger's equation with initial data such that $|u(x, 0)| = 1$, while the L_∞ norm of the solution is unbounded. This result is not surprising, given the probabilistic meaning of the L_2 norm, as the probability to find a particle somewhere on the line is one. The L_∞ norm or *sup* norm gives some indication of the size of the wave function u, but may well be unbounded if u is unbounded at a particular point in $(-\infty, \infty)$.

Example

Definition 2. *The function* $\mathbf{f}(\mathbf{u}, t)$ *is Lipschitz continuous with respect to* $\mathbf{u}$*, if*

$$\|\mathbf{f}(\mathbf{u}_2, t) - \mathbf{f}(\mathbf{u}_1, t)\| < M\|\mathbf{u}_2 - \mathbf{u}_1\|,$$

where M *is a constant.*

An initial value problem for an ordinary differential equation (ODE) system

$$\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}, t),$$

satisfying the initial condition $\mathbf{u}(t_0) = \mathbf{u}_0$, is well-posed in some open set near the initial data $(\mathbf{u}_0, t_0)$ if $\mathbf{f}(\mathbf{u}, t)$ is Lipschitz continuous with respect to $\mathbf{u}$ and continuous with respect to t. This result is known as Picard's theorem.

For example, $y'(t) = y^2(t)$ has a unique solution $y(t) = \frac{1}{1-t}$ that blows up at time $t = 1$, but does not contradict the well-posedness theorem.

Example

An initial value problem for the hyperbolic system described in the previous section is well-posed, since for a strictly hyperbolic system the eigenvalues $\{\lambda_i\}$ are real and distinct.

However, if the eigenvalues of the constant matrix $\mathbf{A}$ are complex, the system is not well-posed. Similarly, the backward heat equation solution discussed in the previous section does not depend continuously on the initial data. For example, the solution of the backward heat equation of the form $e^{k^2 t} e^{ikx}$ cannot be bounded independently of the initial data if k is an arbitrary wavenumber, but the solution exists and is unique! (See [63]). Such problems often arise in practice and require that the problem be converted into a well-posed problem by providing the extra information and/or smoothing. Even if the Fourier transform of the initial data has k in a bounded interval, numerical problems can still arise, unless high frequencies, inevitably generated by a numerical scheme, are periodically filtered out.

If the matrix in equation $\mathbf{u}_t + A\mathbf{u}_x = 0$ has a Jordan block J of size m, then

$$e^{tJ} = \sum_{i=1}^{m-1} \frac{(\lambda t J)^i}{i!},$$

where λ is the corresponding eigenvalue. In this case the problem is ill-posed because the solution has polynomial growth with respect to the initial frequency. This growth is not as fast as the exponential growth in the backward heat equation. Well-posedness for boundary value problems will be described in the chapter on numerical boundary conditions.

The solution operator $S(t, t_0)$ is a map of the initial data onto the solution at a later time t, i.e. $S(t, t_0) : \mathbf{u}_0 \to \mathbf{u}(t)$. For initial value problems with constant coefficients, it can be written explicitly in the exponential form.

Example

Consider first the matrix ODE system

$$\mathbf{u}'(t) = \mathbf{A}\mathbf{u}, \quad \mathbf{u}(t_0) = \mathbf{u}_0,$$

where $\mathbf{A}$ is a constant matrix. The solution can be obtained by use of the Taylor series expansion:

$$\begin{aligned}\mathbf{u}(t) &= \left(1 + (t - t_0)\partial_t + (t - t_0)^2 \frac{(\partial_t)^2}{2!} + \cdots\right) \mathbf{u}(t_0)\\ &= \exp\left((t - t_0)\partial_t\right) \mathbf{u}(t_0).\end{aligned}$$

In Fourier space this expresses the familiar property

$$F[\mathbf{u}(t + \Delta t)] = \exp(i\omega\Delta t)F[\mathbf{u}(t)],$$

of a shift in physical space becoming a multiplication in Fourier space (recall that $F[u(t)] = \int_{-\infty}^{\infty} \exp(-i\omega t)u(t)dt$). Because the operator ∂_t in the ODE system is equivalent to left multiplication by the matrix $\mathbf{A}$, it follows that $(\partial_t)^n \equiv \mathbf{A}^n$. Thus the solution for $\mathbf{u}(t)$ is given by

$$\mathbf{u}(t) = e^{(t-t_0)\mathbf{A}}\mathbf{u}(t_0),$$

where the exponential matrix operator $\exp((t - t_0)\mathbf{A})$ is defined by its Taylor expansion.

The above expression gives the explicit form of the solution operator as

$$S(t, t_0) = e^{(t-t_0)\mathbf{A}},$$

where $\mathbf{u}(t) = S(t, t_0)\mathbf{u}(t_0)$ is the solution of the matrix ODE system.

A similar argument shows that the matrix PDE system:

$$\mathbf{u}'(t) = (\mathbf{A}\partial_x + \mathbf{B}\partial_y + \mathbf{C}\partial_z + \mathbf{D})\mathbf{u}(t),$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are constant coefficient matrices, has the solution operator:

$$S(t,t_0) = e^{(t-t_0)(\mathbf{A}\partial_x+\mathbf{B}\partial_y+\mathbf{C}\partial_z+\mathbf{D})}.$$

Note that in the case where the constant matrices $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}$ commute,

$$e^{(t-t_0)(\mathbf{A}\partial_x+\mathbf{B}\partial_y+\mathbf{C}\partial_z+\mathbf{D})} = e^{(t-t_0)\mathbf{A}\partial_x}e^{(t-t_0)\mathbf{B}\partial_y}e^{(t-t_0)\mathbf{C}\partial_z}e^{(t-t_0)\mathbf{D}}.$$

This shows that the exact solution can be obtained by solving a sequence of 1D problems

$$\begin{aligned}
&\mathbf{u}'(t) = (\mathbf{A}\partial_x)\mathbf{u}(t),\\
&\mathbf{u}'(t) = (\mathbf{B}\partial_y)\mathbf{u}(t),\\
&\mathbf{u}'(t) = (\mathbf{C}\partial_z)\mathbf{u}(t),
\end{aligned}$$

by using the initial data at time t_0 for advancing the solution in z, and using the resultant solution as initial data to advance the solution in y, followed by an advancement of the solution in x, and then advance the solution due to the growth/decay or oscillations operator, $\exp((t-t_0)\mathbf{D})$. Note that the order in which the different operators are applied does not matter, since the exponential matrix operators commute. Physically, this says that we could advect the initial solution and than apply growth, decay or oscillations due to the $D\mathbf{u}$ term or apply the different exponential operators in any order we like. In general, commutators appear in the expansion of the exponential of a sum of matrices. For example, the Campbell-Baker-Hausdorff formula (e.g. [88, Section 5.5, p. 174]; [83, Section 16.3, p. 228]; [129, Ch. 5, Equation (1.19)]) states that if the matrices $\mathbf{A}$ and $\mathbf{B}$ are analytic functions of their matrix elements, then

$$e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{C}},$$

where the matrix $\mathbf{C}$ is given by

$$\mathbf{C} = \mathbf{B} + \int_0^1 g[\exp(t\mathrm{Ad}\mathbf{A})\exp(\mathrm{Ad}\mathbf{B})]A dt, \tag{1.3}$$

where

$$\mathrm{Ad}\mathbf{A}\ \mathbf{B} = [\mathbf{A},\mathbf{B}] = \mathbf{AB} - \mathbf{BA}$$

defines the adjoint representation operator Ad $\mathbf{A}$ associated with the matrix $\mathbf{A}$, and $[\mathbf{A},\mathbf{B}] = \mathbf{AB} - \mathbf{BA}$ is the matrix commutator (note if $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ matrices, then Ad $\mathbf{A}$ is an $n^2 \times n^2$ matrix). The function

$$g(z) = \frac{\ln z}{z-1} \equiv \sum_{j=0}^{\infty} \frac{(1-z)^j}{j+1},$$

is analytic for $|1-z| < 1$. Expanding the integrand in (1.3) as a power series in t and integrating it term by term, yields the expansion:

$$\mathbf{C} = \mathbf{A} + \mathbf{B} + \frac{1}{2}[\mathbf{A},\mathbf{B}] + \frac{1}{12}[\mathbf{A},[\mathbf{A},\mathbf{B}]] + \frac{1}{12}[\mathbf{B},[\mathbf{B},\mathbf{A}]] + \cdots,$$

for $\mathbf{C}$.

1.4 Physical Instabilities

In this section we provide a sketch of the different types of physical instabilities that can occur in model equations. Our purpose is not to provide a complete or encyclopedic description, but to emphasize that an analysis of the instabilities in the model equations is an essential element in designing a robust numerical scheme to treat the problem at hand. Instabilities that occur in the numerical solution of a system of equations can be due to: (a) the numerical scheme being unstable, or (b) a physical instability is present in the system. A numerical difference scheme may possibly not converge due to the presence of a linear numerical instability. These instabilities can sometimes be analyzed by using Von Neumann stability analysis, which is essentially Fourier analysis of the difference scheme, which will be discussed later in Chapter 3. However, nonlinear instabilities cannot be understood by using Von Neumann stability analysis. Thus it is important to be aware of both linear and nonlinear instabilities present in the equation system, and to distinguish whether the instability is of physical or numerical origin.

There is a huge literature on linear wave instabilities in fluids, plasmas, nonlinear optics, biological systems and other disciplines, in which initially small amplitude perturbations evolve from a given equilibrium, or background state (see, e.g. [42] for linear stability methods in hydrodynamics and magnetohydrodynamics; [84] for a review of nonlinear stability in fluids and plasmas).

In Section 1.4.1 we sketch the important distinction between linear convective and absolute instabilities depending on the form of the dispersion equation $D(\mathbf{k},\omega) = 0$ of the system under consideration (e.g. [18,31,178]). Section 1.4.2 discusses some of the subtleties that can arise in linear stability problems depending on the perturbation variables used the analysis. In particular, we discuss more fully the problem of gravity wave propagation in a stratified atmosphere discussed briefly in Section 1.2 and the Rayleigh-Taylor instability. We also discuss briefly the Kelvin-Helmholtz instability for stratified shear flows.

This is followed by discussions of nonlinear instabilities, including: wave breaking and shock formation in nonlinear hyperbolic waves (Section 1.4.3), the modulational instability or Benjamin Feir instability (Section 1.4.4) and resonant wave interactions and instabilities (Section 1.4.5).

1.4.1 Convective and Absolute Instabilities

One important distinction for linear wave instabilities is whether the instability is a convective or absolute instability [18,23,31,178,179].

For 1D systems in which the perturbations are proportional to $\exp[i(kx-\omega t)]$ and the dispersion equation has the form:

$$D(k,\omega)=0, \tag{1.4}$$

one can solve the dispersion equation for ω as a function of k in the form $\omega=\omega^{(\alpha)}(k)$, where the index α labels the different possible wave modes. Alternatively, one can solve the dispersion equation for k as a function of ω in the form $k=k^{(\beta)}(\omega)$ (note that the number of solution branches of the form $\omega=\omega(k)$, is not necessarily equal to the number of solution branches $k=k(\omega)$). If the solution $\omega=\omega(k)$ is complex for real k, or if the solution $k=k(\omega)$ is complex for real ω, the wave may be described as *unstable* or *stable* depending on whether the wave grows or decays in space and time. However, the distinction between stable and unstable waves is not always simple for there exist cases where the waves may grow in space but decay in time at a fixed point.

Definition 3. *A wave is said to be* unstable*, if for some real k, $\omega=\omega_r(k)+i\omega_i(k)$ has an imaginary part with $\omega_i>0$.*

Definition 4. *A wave with complex $k=k_r(\omega)+ik_i(\omega)$ is said to be an amplifying wave if the wave grows in the direction of energy flow, and* evanescent *if it decays in the direction of energy flow.*

In an unstable medium, a localized pulse can evolve in two distinct ways: (a) the pulse can grow or propagate away from its origin, so that eventually, at a fixed point in space, the disturbance decays in time (this is a *convective* instability); or (b) the growing pulse can encompass more of space, so that eventually, at each point, the disturbance grows with time (this is a *non-convective* or *absolute* instability).

Sturrock [178], investigated Fourier integral representations for waves, in which: (i) $k = k(\omega)$ (ω real) where the disturbance $\phi(x, t)$ is such that $|\phi| \to 0$ as $t \to \pm\infty$ (a time-like wave packet) and other representations in which (ii) $\omega = \omega(k)$, with k real, where $|\phi| \to 0$ as $|x| \to \infty$ (a space-like wave packet), and asked when are the two representations equivalent? He came to the conclusion that if the two representations are equivalent (i.e. the first Fourier integral over ω could be expressed in the form of a Fourier integral over k), then the wave packet is both a space-like and a time-like packet, and the unstable wave is an amplifying wave or a convectively unstable wave.

Briggs [31], Bers [18] and Brevdo [23] considered initial value problems and ways in which to distinguish between convective and absolute instabilities based on the complex and multiple roots of the dispersion equation. Bers [18] studied the late time asymptotics of the Green function solution $G(x, t)$, satisfying the wave equation:

$$\mathcal{L}(\partial_x, \partial_t) G(x, t) = \delta(x)\delta(t), \tag{1.5}$$

where the linear partial differential operator $\mathcal{L}$ is related to the dispersion function $D(k, \omega)$ by the usual Fourier analysis correspondence:

$$\mathcal{L}(\partial_x, \partial_t) \equiv D(-i\partial_x, i\partial_t). \tag{1.6}$$

The Green function solution is given by

$$G(x, t) = \int_L \frac{d\omega}{2\pi} \int_F \frac{dk}{2\pi} \frac{\exp[i(kx - \omega t)]}{D(k, \omega)}. \tag{1.7}$$

In (1.7), the contour F must lie in a strip of the complex k-plane, including the $\mathrm{Re}(k)$-axis (this allows the $x > 0$ eigenmodes to be separated off from the $x < 0$ eigenmodes). The integrand will be analytic on F for an appropriate choice of the contour L in the complex ω-plane. The contour L is a straight line parallel to the real ω-axis of the form $\omega = \omega_r + iL$, where the constant L is chosen so that there are no singularities of the integrand in $\mathrm{Im}(\omega) > L$ (this ensures that $G(x, t) = 0$ for $t < 0$).

Example

Consider the Green function solution of the Klein-Gordon equation:

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} - \gamma^2\right) G = \delta(x)\delta(t), \tag{1.8}$$

in which $|G| \to 0$ as $|x| \to \infty$. By Fourier analysis

$$G(x,t) = \int_L \frac{d\omega}{2\pi} \int_F \frac{dk}{2\pi} \frac{\exp[i(kx - \omega t)]}{k^2 - \omega^2 - \gamma^2}. \tag{1.9}$$

The integrand has poles at $k = \pm(\omega^2 + \gamma^2)^{\frac{1}{2}}$. Using the residue theorem to carry out the k-integral (F can be chosen as the $\mathrm{Re}(k)$ axis), one obtains

$$G(x,t) = \int_{c-i\infty}^{c+i\infty} \frac{ds}{2\pi i} \frac{\exp\left(st - (s^2 - \gamma^2)^{\frac{1}{2}}|x|\right)}{2(s^2 - \gamma^2)^{12}}, \tag{1.10}$$

for the Green function G where $s = -i\omega$. From [62, Vol. 1, p. 249, formula 36], the inverse Laplace transform (1.10) can be obtained in closed form to yield

$$G(x,t) = \frac{1}{2} I_0 \left(\gamma(t^2 - x^2)^{\frac{1}{2}}\right) H(t - |x|), \tag{1.11}$$

where $I_0(z)$ is the modified Bessel function of the first kind of order zero, and $H(z)$ is the Heaviside step function. Using the asymptotic form for $I_0(z)$ as $|z| \to \infty$ [6, formula 9.7.1, p. 377], we find as $|t| \to \infty$, with x finite, that

$$G \simeq \frac{\exp(\gamma t)}{(8\pi t)^{12}}. \tag{1.12}$$

Thus the Klein-Gordon Green function exhibits an *absolute* instability.

Comment

One can also obtain the same asymptotic result (1.12) by pinch point analysis. In pinch point analysis, one looks for multiple root solutions of the dispersion equation $D(k,\omega) = 0$. In the present case, the dispersion function $D(k,\omega) = k^2 - \omega^2 - \gamma^2$ has a double zero for k at $k = 0$ and $\omega = i\gamma$. At the double root $D(k,\omega) = D_k(k,\omega) = 0$. This double root solution of the dispersion equation is a pinch point, since it corresponds to the merging of two roots of the dispersion equation, one root coming from the upper half k-plane, with $k_i > 0$, representing an $x > 0$ eigenmode, and one root coming from the lower half k-plane, with $k_i < 0$, and representing an $x < 0$ eigenmode. A detailed discussion of pinch point analysis is given by Bers [18] and Brevdo [23].

Example

Consider the system of coupled uni-directional wave equations:

$$(\partial_t + 2\partial_x)\,\psi = \gamma\phi, \quad (\partial_t + \partial_x)\,\phi = \gamma\psi. \tag{1.13}$$

These equations [18] can be combined to yield the second order wave equation:

$$[(\partial_t + \partial_x)(\partial_t + 2\partial_x) - \gamma^2]\psi = 0. \tag{1.14}$$

Using the Fourier association $k = -i\partial_x$, $\omega = i\partial_t$, we obtain the dispersion equation for the system as:

$$D(k,\omega) = (\omega - 2k)(\omega - k) + \gamma^2 = 0. \tag{1.15}$$

The Green function (1.7) can be obtained by transform methods to yield the solution

$$G(x,t) = I_0\left(\gamma\theta^{\frac{1}{2}}\right) \quad \text{for } t < x < 2t, \tag{1.16}$$

and $G = 0$ otherwise, where

$$\begin{aligned}\theta &= 8\left(t - \frac{x}{2}\right)(x - t)\\ &\equiv t^2 - 4\left(x - \frac{3}{2}t\right)^2 \equiv \frac{x^2}{2} - 8\left(t - \frac{3}{4}x\right)^2.\end{aligned} \tag{1.17}$$

Using the asymptotic form for the Bessel function $I_0(z)$ for large $|z|$, it follows that if one moves convectively with the peak at the speed $V_c = 3/2$, then the Green function $G \sim \exp(\gamma t)/(2\pi\gamma t)^{\frac{1}{2}}$, indicating the presence of a convective instability. On the other hand, if one travels at the speed $V_s = 4/3$, the Green function behaves like an amplifying wave that grows spatially like $G \sim \exp(\gamma|x|/2^{\frac{1}{2}})/(2^{\frac{1}{2}}\pi\gamma|x|)^{\frac{1}{2}}$. These features of the solution can also be obtained by analysis of the dispersion function $D(k,\omega)$ (see [18]).

1.4.2 Rayleigh-Taylor and Kelvin-Helmholtz Instabilities

The Rayleigh-Taylor instability occurs in a stratified heavy fluid of variable density in a gravitational field in which heavier fluid lies on top of lighter fluid. Because of buoyancy forces the lighter fluid tends to rise to the top, which leads to the instability. An important special case is that of two uniform fluids, separated by a plane interface, in which the heavier fluid lies on top of the lighter fluid.

A second type of instability arises when different layers of a stratified fluid are in relative horizontal motion. In particular, the special case of two superposed, uniform fluids separated by a horizontal plane interface, in which the fluids have a non-zero relative horizontal velocity $\mathbf{V}$, is called the Kelvin-Helmholtz instability.

The simplest versions of these instabilities are obtained in the case where the fluid(s) obey the Euler equations of fluid dynamics in a gravitational field $\mathbf{g}$, in the nearly incompressible limit known as the Boussinesq approximation (e.g. [42]). In the perturbation analysis, the fluid density ρ, pressure p and velocity $\mathbf{u}$ are written as:

$$\rho = \rho_0 + \rho', \quad p = p_0 + p', \quad \mathbf{u} = \mathbf{u}_0 + \mathbf{u}', \tag{1.18}$$

where $\Psi_0 = (\rho_0, \mathbf{u}_0^T, p_0)^T$ denotes the background equilibrium state and $\Psi' = (\rho', \mathbf{u}'^T, p')^T$ denotes the perturbed state.

In the Boussinesq approximation, the perturbed mass continuity equation reduces to the two equations:

$$\nabla \cdot \mathbf{u}' = 0, \tag{1.19}$$

$$\frac{\partial \rho'}{\partial t} + \mathbf{u}_0 \cdot \nabla \rho' = 0. \tag{1.20}$$

The first of these equations expresses the fact that the velocity perturbations are incompressible, and the second equation expresses the fact that the density perturbations ρ' are advected with the background flow. The remaining equation in the model consists of the perturbed Euler equation:

$$\rho_0 \left(\frac{\partial \mathbf{u}'}{\partial t} + \mathbf{u}_0 \cdot \nabla \mathbf{u}' \right) = -\nabla p' + \rho' \mathbf{g}, \tag{1.21}$$

where $\mathbf{g} = -g(0,1,0)^T$ is the constant acceleration due to gravity. In the classical Rayleigh-Taylor instability, the background fluid velocity $\mathbf{u}_0 = 0$.

The Rayleigh-Taylor Instability

In the analysis of (1.19)–(1.21), for the case of the Rayleigh-Taylor instability it is assumed that $\rho_0 = \rho_0(y)$, $p = p_0(y)$, $\mathbf{u}_0 = 0$ define the background state. It is assumed that all perturbation variables depend only on two space coordinates x and y, and on the time t. The incompressibility constraint (1.19) is satisfied automatically by introducing the stream function $\psi(x,y,t)$ for the velocity perturbations, where $\mathbf{u}'=(u',v',0)^T=(\psi_y,-\psi_x,0)^T$. Whitham [198] shows that (1.19)–(1.21) may be reduced to the equation:

$$(\rho_0\psi_{ty})_{ty}+\rho_0\psi_{xxtt}-g\left(\frac{d\rho_0}{dy}\right)\psi_{xx}=0, \tag{1.22}$$

for the stream function ψ. Differentiating (1.22) with respect to x, we find that the vertical velocity component v' also satisfies the same wave equation as ψ, namely:

$$\left(\rho_0 v'_{ty}\right)_{ty}+\rho_0 v'_{xxtt}-g\frac{d\rho_0}{dy}v'_{xx}=0. \tag{1.23}$$

To analyze the perturbation equations it is useful to assume the solution ansatz of a harmonic wave of the form:

$$\Psi'=\left(\rho',\mathbf{u}'^T,p'\right)^T=\left(\tilde{\rho},\tilde{\mathbf{u}}^T,\tilde{p}\right)^T\exp[i(k_x x-\omega t)], \tag{1.24}$$

where ω is the frequency and k_x is the horizontal wavenumber. In this case, (1.23) reduces to the ODE:

$$\frac{d}{dy}\left(\rho_0\frac{d\tilde{v}}{dy}\right)+\rho_0 k_x^2\left(\frac{\omega_0^2}{\omega^2}-1\right)\tilde{v}=0, \tag{1.25}$$

where

$$\omega_0^2=-\frac{g}{\rho_0}\frac{d\rho_0}{dy} \tag{1.26}$$

defines the Brunt-Väisälä frequency or buoyancy frequency ω_0 of vertical oscillations in the fluid (cf. (1.2) et seq.).

The differential equation (1.25) is a Sturm-Liouville type equation, in which the allowed values of the eigen-frequency ω depends on the boundary conditions imposed on the differential equation at $y=y_0$ and at $y=y_1$, say. If one imposes homogeneous boundary conditions $\tilde{v}=0$ at $y=y_0$ and at $y=y_1$, then the required solution of (1.25) is obtained by finding the stationary point of the action:

$$J=\int_{y_0}^{y_1}\left[\frac{1}{2}\rho_0\left(\frac{d\tilde{v}}{dy}\right)^2+\frac{\rho_0 k_x^2\tilde{v}^2}{2}\left(1-\frac{\omega_0^2}{\omega^2}\right)\right]dy, \tag{1.27}$$

subject to the boundary conditions $\tilde{v} = 0$ at $y = y_0$ and at $y = y_1$. Alternatively, the variational principle (1.27) can be expressed in the form of a variational principle for the eigen-frequencies $\omega = i\nu$. In this latter approach, eigen-frequencies ω and eigenfunctions $\tilde{v}$ are sought for which

$$-\frac{gk_x^2}{\omega^2} = \frac{gk_x^2}{\nu^2} = \frac{\int_{y_0}^{y_1} \frac{1}{2}\rho_0 \left[\tilde{v}_y^2 + k_x^2\tilde{v}^2\right] dy}{\int_{y_0}^{y_1} 12\tilde{v}^2\rho_{0y} dy} = \frac{I_1}{I_2} \tag{1.28}$$

is stationary, where $J = I_1 + (gk_x^2/\omega^2)I_2$ is the action (1.27). Once the eigenvalues $\{\omega_n\}$ and eigenfunctions $\{\tilde{v}_n\}$ (n, integer) are determined from (1.28) ([136, Vol. 2, p. 1117], using for example the Rayleigh-Ritz method), the required solution of the initial value problem for v' can be constructed by a suitable linear sum over the eigen-functions.

It is obvious from (1.28) that $\nu^2 > 0$ (i.e. $\omega^2 < 0$) implying instability if $d\rho_0/dy > 0$, whereas $\omega^2 > 0$, implying stability if $d\rho_0/dy < 0$. Thus the variational principle (1.28) reveals the basic result that the perturbations are stable (i.e. they don't grow with increasing time) for a stably stratified fluid with $d\rho_0/dy < 0$, in which the lighter fluid lies on top of the heavier fluid. However, instability results in the opposite case in which the heavy fluid lies on top of the lighter fluid (the $d\rho_0/dy > 0$ case).

Equation (1.25) can be solved exactly for the case of a stably stratified background, with a constant scale height, with $\rho_0 = \rho_{00}\exp(-y/h)$. By using the transformation:

$$\tilde{v} = \frac{w}{\rho_0^{1/2}}, \tag{1.29}$$

we obtain the differential equation:

$$\frac{d^2w}{dy^2} + k_y^2 w = 0, \tag{1.30}$$

for w where

$$k_y^2 = k_x^2\left(\frac{\omega_0^2 - \omega^2}{\omega^2}\right) - \frac{1}{4h^2}. \tag{1.31}$$

Equation (1.31) is equivalent to the dispersion equation:

$$\omega^2 = \frac{\omega_0^2 k_x^2}{k_x^2 + k_y^2 + 1/(4h^2)} \tag{1.32}$$

for the internal gravity waves described in (1.2) et seq.

Since we assume $k_y^2 > 0$, (1.30) has solutions in terms of $\sin(k_y y)$ and $\cos(k_y y)$. Thus the general solution for $\tilde{v}$ is of the form:

$$\tilde{v} = \exp[y/(2h)]\left[A\cos(k_y y) + B\sin(k_y y)\right], \tag{1.33}$$

where A and B are constants determined by the boundary conditions. In particular, if $\tilde{v} = 0$ at $y = 0$ and $y = L$, the eigenvalues for k_y are $k_{ym} = 2\pi m/L$ (m, integer) and the eigenfunctions are

$$\tilde{v}_m = A_m \exp\left(\frac{y}{2h}\right)\sin\left(\frac{2m\pi y}{L}\right), \tag{1.34}$$

where the A_m are normalization constants. Thus the solution for $\tilde{v}$ consists of exponentially amplifying oscillations with increasing height y.

From the Fourier space perturbation equations, it is straightforward to show that the velocity field components $\tilde{u}, \tilde{v} \sim \mathrm{O}\left(\exp(y/2h)\right)$, but the density and pressure perturbations $\tilde{\rho}, \tilde{p} \sim \mathrm{O}\left(\exp(-y/2h)\right)$. Thus in the numerical solution of gravity wave propagation in a stable, exponentially stratified atmosphere, it is necessary to take into account the exponential behavior of the perturbations $\tilde{u}$, $\tilde{v}$, $\tilde{\rho}$ and $\tilde{p}$ to avoid exponential growth of small numerical errors.

In the classical version of the Rayleigh-Taylor instability, the background fluid density profile has the form:

$$\rho_0(y) = \rho_1[1 - H(y)] + \rho_2 H(y), \tag{1.35}$$

where $H(y)$ is the Heaviside step function. Thus $\rho_0 = \rho_1$ for $y < 0$ and $\rho_0 = \rho_2$ for $y > 0$ and $\mathbf{u}_0 = 0$ for the background state. In this case one obtains the dispersion equation:

$$\omega^2 = \frac{gk_\perp(\rho_1 - \rho_2)}{\rho_2 + \rho_1}, \tag{1.36}$$

from matching the perturbations at the interface [42]. Here $k_\perp = (k_x^2 + k_z^2)^{1/2}$. Thus the system is unstable (i.e. $\omega^2 < 0$) if $\rho_2 > \rho_1$.

The Kelvin-Helmholtz Instability

If in the classical Rayleigh-Taylor instability, described by (1.35) and (1.36), the upper medium has uniform velocity $\mathbf{u}_2 = (U_2, 0, 0)^T$ and the lower medium has uniform velocity $\mathbf{u}_1 = (U_1, 0, 0)^T$, then the matching conditions on the perturbations at the interface $y = 0$ give a dispersion equation for the perturbed interface which is a quadratic in ω with roots

$$\omega = k_x(\alpha_1 U_1 + \alpha_2 U_2) \pm [gk_\perp(\alpha_1 - \alpha_2) - k_x^2\alpha_1\alpha_2(U_1 - U_2)^2]^{1/2}, \tag{1.37}$$

[42, Ch. 11, Equation (30)], where $k_\perp = (k_x^2 + k_z^2)^{1/2}$, and

$$\alpha_1 = \frac{\rho_1}{\rho_1 + \rho_2}, \quad \alpha_2 = \frac{\rho_2}{\rho_1 + \rho_2}. \tag{1.38}$$

Thus if $\rho_2 > \rho_1$, the roots (1.37) are complex and the fluid is unstable. However, if $\rho_1 > \rho_2$ (the Rayleigh-Taylor stable case), instability can occur if

$$k_x^2\alpha_1\alpha_2(U_1 - U_2)^2 > gk_\perp(\alpha_1 - \alpha_2). \tag{1.39}$$

Thus if

$$k_\perp > \frac{g(\alpha_1 - \alpha_2)}{[\alpha_1\alpha_2(U_1 - U_2)^2 \cos^2\vartheta]}, \tag{1.40}$$

where $k_x = k_\perp \cos\vartheta$, there is always instability no matter how small $(U_1 - U_2)$ may be. Equation (1.40) shows that at sufficiently large wavenumbers, there is instability. There is an extensive literature on the Kelvin-Helmholtz instability (see e.g. [42]). Cairns [38] identified the Kelvin-Helmholtz instability with the coalescence of positive and negative energy waves. Later work by Balmforth and Morrison [13,14] showed that for plane-parallel, incompressible shear flows, there are both continuous and discrete eigenmodes, and that instability in flows with an inflection point are associated with negative energy modes. Hamiltonian description of shear flows is given by Balmforth and Morrison [14].

1.4.3 Wave Breaking and Gradient Catastrophe

Wave breaking of nonlinear hyperbolic waves in one Cartesian space dimension can be demonstrated by considering the conservation equation:

$$\rho_t + [Q(\rho)]_x = 0. \tag{1.41}$$

The properties and solutions of (1.41) are detailed in [198, Ch. 2], where they are used in discussions of nonlinear hyperbolic waves in a variety of physical contexts (e.g. flood waves, traffic flow and simple sound waves in gas dynamics). The quantity ρ is a density and $Q(\rho)$ is the flux. An alternative form of (1.41) is

$$\rho_t + \lambda(\rho)\rho_x = 0, \tag{1.42}$$

where

$$\lambda(\rho) = Q'(\rho) \tag{1.43}$$

is the characteristic speed of the wave.

Equation (1.42) is a first order PDE for ρ and can be solved by the method of characteristics [51,171]. The characteristics of (1.42) are

$$\frac{dx}{d\tau} = \lambda, \quad \frac{d\rho}{d\tau} = 0, \quad \frac{dt}{d\tau} = 1, \tag{1.44}$$

where τ is a parameter along the characteristics. Consider the initial value problem for (1.42) in which

$$\rho(x, 0) = f(x) \tag{1.45}$$

at time $t = 0$. The characteristics (1.44) may be integrated to yield the solution for $\rho(x, t)$ in the form:

$$\rho(x, t) = f(\xi), \tag{1.46}$$

where ξ satisfies the implicit equation:

$$G(x, t, \xi) = x - F(\xi)t - \xi = 0, \tag{1.47}$$

and $F(\xi) \equiv \lambda(\rho) = \lambda(f(\xi))$ is the eigen-velocity on the characteristics. For $\xi = \text{const.}$, (1.47) is a straight line in the (x, t)-plane, where $(x, t) = (\xi, 0)$ labels the point where the characteristic cuts the x-axis. The complete family of characteristics obtained by varying the parameter ξ is needed to generate the solution for $\rho(x, t)$.

We now demonstrate that the derivatives of ρ can become unbounded on the envelope of the family of plane waves with phase fronts (1.47). By implicit differentiation of (1.46) and (1.47) we find

$$\rho_t = -\frac{f'(\xi)F(\xi)}{1 + tF'(\xi)}, \quad \rho_x = \frac{f'(\xi)}{1 + tF'(\xi)}, \tag{1.48}$$

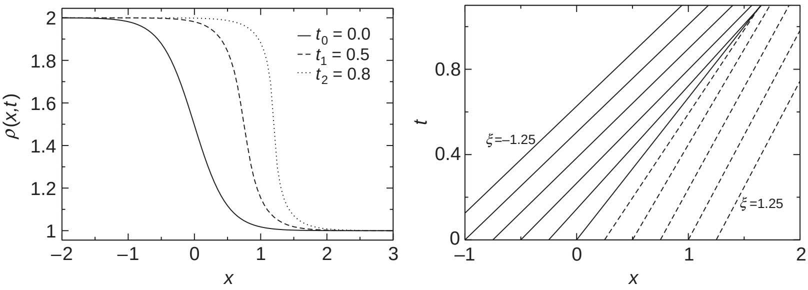


Figure 1.1: Left: solutions of the Equation (1.42) with $\lambda(\rho) = \rho$ at $t = t_1$, t_2, with initial condition $\rho(x, 0) = 1 + (1 + \tanh(-2x))/2$ at $t_0 = 0$. Right: corresponding characteristics (1.47) for $\xi \leq 0$ (solid lines) and $\xi > 0$ (dashed lines).

for the derivatives of ρ. Thus the derivatives of ρ diverge when

$$1 + tF'(\xi) = 0 \quad \text{or} \quad t = -\frac{1}{F'(\xi)}. \tag{1.49}$$

However, the envelope of the family of plane wave fronts (1.47) obtained by solving the simultaneous equations $G = G_\xi = 0$, also yields (1.49). In particular, from (1.47)

$$G_\xi = -[F'(\xi)t + 1] = 0, \tag{1.50}$$

which is precisely the condition (1.49) for the derivatives of ρ to diverge.

For the case of a forward propagating wave, with $\lambda'(\rho) > 0$, $f'(\xi) < 0$ one finds that $F'(\xi) = \lambda'(\rho)f'(\xi) < 0$, and the minimum break time $t = t_B$ for the wave to break is given by

$$t_B = -\frac{1}{F'(\xi_B)}, \tag{1.51}$$

where $|F'(\xi)|$ is maximal at $\xi = \xi_B$. From a physical standpoint, the wave breaks because the points on the wave profile $\rho(x, t)$ with the larger values of ρ travel at a faster speed than the points with smaller ρ (this assumes $\lambda'(\rho) > 0$ and that $\lambda(\rho) > 0$). The characteristics and the breaking of the wave are illustrated in Figure 1.1. At times $t > t_B$, the solution becomes multi-valued, and it is then necessary to insert a shock in the flow in order to obtain a consistent weak solution of the equation. The multi-valued character of the solution for $t > t_B$, can also be seen from the fact that more than one characteristic, with different ξ's, cover the same patch of the (x, t)-plane where the solution is multi-valued.

Similar arguments show that multi-dimensional simple waves, constructed using plane wave fronts (e.g. [20,192]), also break on the envelope of the family of plane wave fronts defining the simple wave solution. Explicit examples of vortex simple waves and simple sound waves in gas dynamics in two Cartesian space dimensions are obtained in [192]. In the case of simple Alfvén waves in magnetohydrodynamics the electric current diverges on the wave envelope [191].

1.4.4 Modulational Instabilities

The Benjamin Feir instability, or modulational instability, occurs in a variety of physical situations (e.g. water waves, nonlinear optics, and plasma physics). If dispersion and nonlinearity act against each other, monochromatic wave trains do not tend to remain monochromatic. The sidebands of the carrier wave draw on its energy by resonant wave-wave interactions, resulting in a modulated envelope. For soliton equations in one space dimension, the envelope modulation continues to grow until a soliton shape is formed, at which point there is exact balance between nonlinearity and dispersion, and no further change in the waveform occurs.

The canonical nonlinear wave equation that is used to describe wave modulation of a carrier wave, due to nonlinearity and dispersion, is the NLS equation (e.g. [57,138]). An alternative approach to modulational theory based on variational principles was developed by Whitham [198, Chs 14 and 15]. Modulational theory of wave trains assumes that the wave train has a fast varying phase and a slowly varying envelope, so that the method of multiple scales can be applied to the system of interest. In Whitham's approach to wave modulation theory, an ansatz for the waves is first specified, consisting of a slowly varying envelope and a fast varying phase. This ansatz is then substituted into the Lagrangian, and the Lagrangian is then averaged over the fast variations to determine the averaged Lagrangian density $\langle L\rangle$. Wave modulation equations are then obtained by requiring that the variations of the averaged Lagrangian functional $\langle\mathcal{L}\rangle$ be stationary.

There are some subtle differences between wave envelope evolution in the NLS equation and Whitham's theory [138, p. 49]. In particular, Whitham's theory applies to finite amplitude waves, but it breaks down for small amplitude waves due to a change in the solvability conditions at order ϵ (ϵ is a measure of the wave amplitude). In Whitham theory, one obtains an algebraic equation relating the slow wave frequency to the wavenumber and wave amplitude as the solvability condition, whereas in the NLS case the algebraic relation is replaced by a differential equation, which corresponds to the real part of the NLS equation. Below, we illustrate the occurrence of the modulational instability in the NLS equation, and also illustrate the use of Whitham's theory using the derivative nonlinear Schrödinger (DNLS) equation.

NLS Equation

The NLS equation in one space dimension for the complex wave amplitude a (the wave envelope) is

$$ia_T + \frac{\omega_0''}{2} a_{\xi\xi} + \beta |a|^2 a = 0, \tag{1.52}$$

where $\omega_0(k)$ describes the dispersion function for the carrier wave, with wavenumber k,

$$\xi = \epsilon(x - \omega_0' t), \quad T = \epsilon^2 t, \tag{1.53}$$

are the long space (ξ) and time (T) variables, ω_0' is the group velocity of the carrier wave, and ϵ measures the wave amplitude.

To illustrate the modulational instability, consider perturbations of the NLS equation about the exact, space independent solution of the NLS equation:

$$a = a_0 \exp(i\beta |a_0|^2 T) \equiv A_0(T), \tag{1.54}$$

[57,138]. Using the ansatz

$$a = A_0(T)(1 + B(\xi, T)), \tag{1.55}$$

for the perturbed solution, the linearized NLS equation (1.52) becomes

$$iB_T + \frac{\omega_0''}{2} B_{\xi\xi} + \beta |a_0|^2 (B + B^*) = 0. \tag{1.56}$$

The perturbed NLS equation (1.56) has solutions of the form:

$$B = B_1 \exp[i(K\xi + \Omega T)] + B_2 \exp[-i(K\xi + \Omega^* T)], \tag{1.57}$$

where we have allowed for the possibility that $\mathrm{Im}(\Omega) \neq 0$, corresponding to the possibility of unstable solutions.

Substituting the ansatz (1.57) in (1.56) we obtain the linear algebraic equations:

$$\left(\beta|a_0|^2 - \Omega - \frac{\omega_0''}{2}K^2\right) B_1 + \beta|a_0|^2 B_2^* = 0, \tag{1.58}$$

$$\beta|a_0|^2 B_1^* + \left(\beta|a_0|^2 + \Omega^* - \frac{\omega_0''}{2}K^2\right) B_2 = 0. \tag{1.59}$$

Taking the complex conjugate of (1.59), together with (1.58), yields a homogeneous matrix equation for B_1 and B_2^*, which has a non-trivial solution for B_1 and B_2^* if the determinant of the matrix is zero. This latter requirement yields the dispersion equation

$$\Omega^2 = \frac{(\omega_0'' K)^2}{4}\left(K^2 - \frac{4\beta|a_0|^2}{\omega_0''}\right), \tag{1.60}$$

relating Ω and K. From (1.60), it follows that $\Omega^2 < 0$ if

$$\beta\omega_0'' > 0 \quad \text{and} \quad 0 < K^2 < \frac{4\beta|a_0|^2}{\omega_0''}. \tag{1.61}$$

The results (1.61) indicate that instability occurs when $\beta\omega_0'' > 0$. For $\beta\omega_0'' < 0$, the wave is stable. In the case $\beta\omega_0'' > 0$, analysis based on the inverse scattering transform shows that solitons solutions of the NLS equation can be obtained in this case, whereas no soliton solutions are possible if $\beta\omega_0'' < 0$. In the focusing case ($\beta\omega_0'' > 0$), bunching of the unstable wave envelope occurs, as the envelope tries to evolve to an equilibrium state. The soliton solutions are stable solutions of the NLS equation in which there is a perfect balance between nonlinearity and dispersion. In the defocussing case ($\beta\omega_0'' < 0$) the envelope is stable. The physical explanation of the instability is given in [138].

Modulation Theory

As an example of wave modulation theory, consider the DNLS equation, for weakly nonlinear, Alfvén waves propagating along a uniform background magnetic field $\mathbf{B} = B_0\mathbf{e}_x$ directed along the x-axis of a rectangular Cartesian coordinate system. The DNLS equation is (e.g. [98, 130,132,133,154,208])

$$\psi_t + (C|\psi|^2\psi + i\sigma D\psi_x)_x = 0, \tag{1.62}$$

where $\sigma = \pm 1$ describes the polarization of the wave,

$$\psi = \frac{\delta B_y + i\delta B_z}{B_0}, \tag{1.63}$$

is the normalized, complex wave amplitude and $\delta \mathbf{B} = (0, \delta B_y, \delta B_z)$ is the magnetic field perturbation describing the wave. The nonlinear coefficient C and dispersion coefficient D are given by

$$C = \frac{V_A}{4(1-\beta)}, \quad D = \frac{\chi V_A}{2}, \quad \beta = \frac{a^2}{V_A^2}. \tag{1.64}$$

In (1.64), V_A is the Alfvén speed, and χ is the ion inertial length, and β is proportional to the plasma beta. The case $\sigma = -1$ corresponds to the forward propagating, right-hand polarized, whistler branch of the dispersion equation, and $\sigma = 1$ corresponds to the forward propagating, left-hand polarized ion-cyclotron wave. The DNLS equation is an integrable Hamiltonian system that can be integrated using the inverse scattering method [98]. The DNLS equation can be obtained from the variational principle $\delta L = 0$, where the variational functional $\mathcal{L}$ has the form:

$$\mathcal{L} = \int dx \int dt L(u_t, u_x, u_{xx}, u_t^*, u_x^*, u_{xx}^*), \quad \psi = u_x, \tag{1.65}$$

where u is a potential for ψ, and

$$L = -\frac{1}{2}\left(u_x u_t^* + u_x^* u_t\right) - i\sigma \frac{D}{2}\left(u_{xx} u_x^* - u_x u_{xx}^*\right) - \frac{1}{2} C \left(u_x u_x^*\right)^2, \tag{1.66}$$

is the Lagrangian density. Variations of $\mathcal{L}$ with respect to u^* yield the DNLS equation (1.62), whereas variations with respect to u yield the complex conjugate of the DNLS equation. The density perturbation in the wave is given by the eigen-relation $\delta\rho/\rho_0 = |\psi|^2/(2(1-\beta))$, and hence there is a density enhancement if $\beta < 1$, but a density depletion if $\beta > 1$.

Using the plane wave ansatz

$$u = \frac{\phi}{k}\exp(i\theta), \quad \psi = \phi \exp(i\theta), \quad \theta = kx - \omega t, \tag{1.67}$$

where θ is the fast varying phase, and averaging L over a period in θ, yields the averaged Lagrangian density

$$\langle L \rangle = \frac{\omega}{k}|\phi|^2 + \sigma D k |\phi|^2 - \frac{1}{2} C |\phi|^4. \tag{1.68}$$

The corresponding averaged Lagrangian functional $\langle\mathcal{L}\rangle$ is given by

$$\langle\mathcal{L}\rangle = \int dt \int dx \langle L\rangle. \tag{1.69}$$

The Lagrangian density $\langle L\rangle$ is a function only of the long time and space variables associated with the envelope, and $k = \theta_x$ and $\omega = -\theta_t$ also depend only on the long space and time variations (see [198], for a detailed description and justification of the method).

Taking variations with respect to the wave amplitude variable ϕ or ϕ^*, yields the amplitude dependent dispersion equation for the waves:

$$\omega = C|\phi|^2 k - \sigma D k^2. \tag{1.70}$$

Similarly, taking variations with respect to θ yields the equation:

$$\frac{\partial}{\partial t}\left(\frac{|\phi|^2}{k}\right) + \frac{\partial}{\partial x}\left(\frac{|\phi|^2}{k}(C|\phi|^2 - 2\sigma D k)\right) = 0, \tag{1.71}$$

where we have used the dispersion equation (1.70) to eliminate reference to ω. It is interesting to note that the group velocity of the waves from (1.70) is

$$V_g = \omega_k = C|\phi|^2 - 2\sigma D k. \tag{1.72}$$

Thus (1.71) may be identified as the wave action equation for the system where $A = |\phi|^2/k$ is proportional to the wave action ([132]; see also Section 1.5 for further discussion of wave action). Using the equations $k = \theta_x$, $\omega = -\theta_t$, the compatibility condition $\theta_{xt} = \theta_{tx}$ yields the wavenumber conservation equation:

$$k_t + \omega_x = 0. \tag{1.73}$$

Using the dispersion equation (1.70), the ray equation (1.73) becomes

$$\frac{\partial k}{\partial t} + \frac{\partial}{\partial x}\left(Ck|\phi|^2 - \sigma D k^2\right) = 0. \tag{1.74}$$

Thus both the wavenumber k and the wave energy $E = |\phi|^2$ vary on the long space and time scales, and are governed by the hydrodynamic type equation system (1.71) and (1.74).

The system of equations (1.71) and (1.74) can be written in the matrix form:

$$\frac{\partial}{\partial t}\begin{pmatrix} E \\ k \end{pmatrix} + \begin{pmatrix} 3CE - 2\sigma Dk & -2\sigma DE \\ Ck & CE - 2\sigma Dk \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} E \\ k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (1.75)$$

Looking for simple wave solutions of the above system in which E and k are functions of a single phase variable $\varphi(x, t)$, in which $K = \varphi_x$ and $\Omega = -\varphi_t$, yields the dispersion equation:

$$\Omega^2 + \alpha\Omega K + \gamma K^2 = 0, \quad (1.76)$$

where

$$\alpha = 4(\sigma Dk - CE), \quad \gamma = 3C^2E^2 - 6C\sigma DEk + 4D^2k^2. \quad (1.77)$$

It is straight forward to solve the quadratic dispersion equation (1.76) for the envelope. An inspection of the roots of the quadratic reveals that the solutions for Ω for a given K are unstable if

$$|\phi^2| - 4\sigma\chi k(1 - \beta) < 0. \quad (1.78)$$

Thus the left-hand polarized wave ($\sigma = 1$) is modulationally unstable for $\beta < 1$, but is stable if $\beta > 1$. Similarly, the right-hand polarized wave ($\sigma = -1$) is modulationally unstable for $\beta > 1$ and stable for $\beta < 1$.

1.4.5 Resonant Interactions

In this subsection we present an elementary discussion of resonant interaction phenomena that, under some circumstances, can lead to instabilities. These ideas have application to resonant wave interactions, involving multiple waves (e.g. three- and four-wave interactions in fluid mechanics and plasma physics).

Resonant Interaction of three Oscillators

This example is discussed in [158], where they use it to illustrate the physical phenomena underlying three-wave resonant interactions in plasmas. Consider three coupled, simple harmonic oscillators, described by the Hamiltonian:

$$H = \sum_{i=1}^{3} \frac{1}{2}\left(p_i^2 + \omega_i^2 x_i^2\right) + V x_1 x_2 x_3, \quad (1.79)$$

where the coupling constant V describes how the three oscillators, with canonical coordinates $\{(x_i, p_i) : \ i = 1, 2, 3\}$, and frequencies $\{\omega_i : \ i = 1, 2, 3\}$, are coupled to each other. The oscillators are independent of each other if $V = 0$.

Hamilton's equations for the system:

$$\dot{x}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial x_i}, \quad i = 1, 2, 3, \tag{1.80}$$

where $\dot{x}_i = dx_i/dt$ and $\dot{p}_i = dp_i/dt$, reduce to the equation system:

$$\begin{aligned} &\dot{x}_i = p_i, \quad i = 1, 2, 3, \\ &\dot{p}_1 = -\left(\omega_1^2 x_1 + V x_2 x_3\right), \\ &\dot{p}_2 = -\left(\omega_2^2 x_2 + V x_1 x_3\right), \\ &\dot{p}_3 = -\left(\omega_3^2 x_3 + V x_1 x_2\right). \end{aligned} \tag{1.81}$$

Elimination of reference to the canonical momenta $\{p_i\}$ in (1.81) yields the coupled harmonic oscillator system:

$$\begin{aligned} &\ddot{x}_1 + \omega_1^2 x_1 = -V x_2 x_3, \\ &\ddot{x}_2 + \omega_2^2 x_2 = -V x_1 x_3, \\ &\ddot{x}_3 + \omega_3^2 x_3 = -V x_1 x_2. \end{aligned} \tag{1.82}$$

In order to illustrate resonant interaction phenomena, it is useful to consider perturbation solutions of the equations.

The perturbation analysis of (1.82) can be carried out by using the method of normal forms (e.g. [137]). In the method of normal forms, one first notices that the homogeneous system of equations (1.82) has solutions $x_j(t) = A_j \exp(i\omega_j t) + B_j \exp(-i\omega_j t)$ for $j = 1, 2, 3$. One of the key ingredients of the method is to reduce the system to a first order system, by introducing new variables ζ_j and ζ_j^* (here $*$ denotes complex conjugate), so that the homogeneous system, with no source terms (i.e. the equations (1.82) with $V = 0$) has solutions proportional to $\exp(\pm i\omega_j t)$ for $j = 1, 2, 3$. This can be achieved in the present case by using the normal variables

$$z_j = \frac{1}{2}\left(x_j - i\frac{\dot{x}_j}{\omega_j}\right) \quad \text{and} \quad z_j^*, \quad j = 1, 2, 3, \tag{1.83}$$

where $\dot{x}_j = dx_j/dt$. From (1.83), the original variables $\{x_j\}$ and $\{\dot{x}_j\}$ are given by

$$x_j = z_j + z_j^*, \quad \dot{x}_j = i\omega_j(z_j - z_j^*), \quad j = 1, 2, 3. \tag{1.84}$$

The mathematical procedure to systematically obtain the normal variables for the first oscillator equation in (1.82) is to introduce new variables $y_1 = x_1$ and $y_2 = \dot{x}_1$ to reduce the equation to a matrix system of two coupled first order equations for y_1 and y_2 of the form $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{S}$, where $\mathbf{y} = (y_1, y_2)^T$ and $\mathbf{S}$ is the source term. The normal variables correspond to expanding the solution for $\mathbf{y}$ in terms of the right eigenvectors $\mathbf{R}_j$ of the matrix $\mathbf{A}$ corresponding to the eigenvalue λ_j. The normal variables are then given by $\zeta_j = \mathbf{L}_j \cdot \mathbf{y}$, where $\mathbf{L}_j$ are the orthonormal left eigenvectors of $\mathbf{A}$.

In terms of the normal variables $\{z_j\}$ the oscillator equation system (1.82) reduces to the diagonalized system

$$\dot{z}_j - i\omega_j z_j = \frac{iV}{2\omega_j} \prod_{\substack{s=1 \\ s\neq j}}^{3} (z_s + z_s^*), \quad j = 1, 2, 3. \tag{1.85}$$

An alternative set of normal variables, which is related to the Hamiltonian structure of the equations, is the set of re-scaled normal variables:

$$\hat{z}_j = i\,(2|\omega_j|)^{1/2}\, z_j. \tag{1.86}$$

In terms of the re-scaled variables (1.86), the normal equations (1.85) take the form:

$$\dot{\hat{z}}_j = i\omega_j \hat{z}_j + \sigma_j K \prod_{\substack{s=1 \\ s\neq j}}^{3} \left(\hat{z}_s - \hat{z}_s^*\right), \quad j = 1, 2, 3, \tag{1.87}$$

where

$$K = \frac{V}{2^{3/2}|\omega_1\omega_2\omega_3|^{1/2}} \tag{1.88}$$

is the resonant coupling coefficient and

$$\sigma_j = \mathrm{sgn}(\omega_j), \quad j = 1, 2, 3, \tag{1.89}$$

denote the signs of the oscillator frequencies ω_j. The equations (1.87) can be written in the form:

$$\frac{d\hat{z}_j}{dt} = i\sigma_j \frac{\partial H}{\partial \hat{z}_j^*}, \quad \frac{d\hat{z}_j^*}{dt} = -i\sigma_j \frac{\partial H}{\partial \hat{z}_j}, \tag{1.90}$$

where

$$H = \sum_{j=1}^{3} |\omega_j| \hat{z}_j \hat{z}_j^* + iK \prod_{s=1}^{3} (\hat{z}_s - \hat{z}_s^*) \tag{1.91}$$

is the Hamiltonian (1.79), expressed in terms of the normal variables $\hat{z}_j$. Strictly speaking, (1.90) are not in canonical Hamiltonian form. However, if one uses the variables

$$\tilde{z}_j = \frac{1}{2}[(1+\sigma_j)\hat{z}_j + (1-\sigma_j)\hat{z}_j^*], \tag{1.92}$$

(i.e. $\tilde{z}_j = \hat{z}_j$ if $\sigma_j > 0$ but $\tilde{z}_j = \hat{z}_j^*$ if $\sigma_j < 0$) then the $\tilde{z}_j$ and $\tilde{z}_j^*$ are canonically conjugate variables, satisfying Hamilton's equations:

$$\frac{d\tilde{z}_j}{dt} = i\frac{\partial H}{\partial \tilde{z}_j^*}, \quad \frac{d\tilde{z}_j^*}{dt} = -i\frac{\partial H}{\partial \tilde{z}_j}, \tag{1.93}$$

with Hamiltonian $\tilde{H} = iH$.

Resonant Interaction Equations

To analyze the long-time behavior of the oscillator equations (1.82), or the equivalent normal variable equations (1.87), we look for solutions of (1.87) of the form:

$$\hat{z}_j = \hat{C}_j(t)e^{i\theta_j}, \quad \theta_j = \omega_j t, \quad j = 1, 2, 3, \tag{1.94}$$

where the $\hat{C}_j(t)$ are slowly varying functions of t (i.e. $\omega_j \gg |\dot{\hat{C}}_j/\hat{C}_j|$). If the fast varying oscillator phases $\{\theta_s\}$ or the eigen-frequencies $\{\omega_s\}$ satisfy a resonance condition of the form:

$$\theta_j = \theta_l + \theta_m \quad \text{or} \quad \omega_j = \omega_l + \omega_m, \tag{1.95}$$

where (j, l, m) is a permutation of $(1, 2, 3)$, then the resonant interactions result in a systematic exchange of energy between the oscillators on the long time scale.

Substituting the solution ansatz (1.94) in (1.87) we obtain the equation:

$$\frac{d\hat{C}_j}{dt} = \sigma_j K \left(\hat{C}_l \hat{C}_m e^{i(\theta_m+\theta_l-\theta_j)} + \hat{C}_l^* \hat{C}_m^* e^{-i(\theta_m+\theta_l+\theta_j)} - \hat{C}_l^* \hat{C}_m e^{i(\theta_m-\theta_l-\theta_j)} - \hat{C}_l \hat{C}_m^* e^{i(\theta_l-\theta_m-\theta_j)} \right). \tag{1.96}$$

Averaging the latter equation over the fast time scale, and assuming the resonance condition (1.95) holds, we obtain the equation:

$$\frac{d\hat{C}_j}{dt} = \sigma_j K \hat{C}_l \hat{C}_m, \tag{1.97}$$

for the evolution of the wave amplitude $\hat{C}_j(t)$ on the long time scale. Carrying out a similar averaging procedure for the equations for $\hat{z}_l$ and $\hat{z}_m$, we obtain the resonant, three oscillator equation system:

$$\begin{aligned} \frac{d\hat{C}_j}{dt} &= \sigma_j K \hat{C}_l \hat{C}_m, \\ \frac{d\hat{C}_l}{dt} &= -\sigma_l K \hat{C}_j \hat{C}_m^*, \\ \frac{d\hat{C}_m}{dt} &= -\sigma_m K \hat{C}_j \hat{C}_l^*, \end{aligned} \tag{1.98}$$

where the oscillators are assumed to satisfy the resonance conditions (1.95). The equations (1.98) may be written in the Hamiltonian-like form:

$$\frac{d\hat{C}_s}{dt} = \sigma_s \frac{\partial H_c}{\partial \hat{C}_s^*}, \quad \frac{d\hat{C}_s^*}{dt} = -\sigma_s \frac{\partial H_c}{\partial \hat{C}_s}, \quad s = j, l, m, \tag{1.99}$$

where the Hamiltonian H_c has the form:

$$H_c = K \left(\hat{C}_l \hat{C}_m \hat{C}_j^* - \hat{C}_l^* \hat{C}_m^* \hat{C}_j \right). \tag{1.100}$$

As in the discussion following (1.90) et seq., the equations (1.99) can be cast in proper Hamiltonian form by using the variables $\tilde{C}_s = \hat{C}_s$ if $\sigma_s > 0$, and $\tilde{C}_s = \hat{C}_s^*$ if $\sigma_s < 0$.

The integrals:

$$\sigma_l |\hat{C}_l|^2 + \sigma_j |\hat{C}_j|^2 = \text{const.}, \quad \sigma_m |\hat{C}_m|^2 + \sigma_j |\hat{C}_j|^2 = \text{const.}, \tag{1.101}$$

of the system (1.98) are known as the Manley-Rowe relations. In a quantum mechanical interpretation, the Manley Rowe equations (1.101) correspond to conservation laws for the number of wave quanta, due to resonant interaction of the three oscillators.

The resonant interaction equations (1.99) have exact analytic solutions in terms of Jacobian elliptic functions (e.g. [10,158]). However, it is also instructive to consider perturbative solutions of (1.99) as discussed below.

Parametric Instabilities

Consider solutions of the oscillator system (1.82), for the special case where the coupling constant V is small, and one of the oscillators, say x_1, has much larger amplitude than the other two oscillators. More precisely, we assume

$$|x_2| \ll |x_1|, \quad |x_3| \ll |x_1|, \quad |x_1 V| \ll \omega_i^2, \quad i = 1, 2, 3. \tag{1.102}$$

In this case (1.82) may be approximated by the linear system

$$\begin{aligned} &\ddot{x}_1 + \omega_1^2 x_1 = 0, \\ &\ddot{x}_2 + \omega_2^2 x_2 = -V x_1 x_3, \\ &\ddot{x}_3 + \omega_3^2 x_3 = -V x_1 x_2. \end{aligned} \tag{1.103}$$

The question now arises, under what conditions on the parameters $\{\omega_i\}$ and the coupling parameter V, are the solutions for x_2 and x_3 driven unstable by the large amplitude oscillator x_1.

We assume the oscillators satisfy the resonance conditions:

$$\theta_2 = \theta_1 + \theta_3 \quad \text{or} \quad \omega_2 = \omega_1 + \omega_3. \tag{1.104}$$

Thus $(j, l, m) = (2, 1, 3)$ in the resonance conditions (1.95), and in the resonant three-oscillator equations (1.98). The resonant interaction equations (1.98) corresponding to the approximate system (1.103) are:

$$\frac{d\hat{C}_1}{dt} \approx 0, \quad \frac{d\hat{C}_2}{dt} = \sigma_2 K \hat{C}_1 \hat{C}_3, \quad \frac{d\hat{C}_3}{dt} = -\sigma_3 K \hat{C}_1^* \hat{C}_2. \tag{1.105}$$

From (1.105), it follows that $\hat{C}_1 = \text{const.}$ is the approximate solution for $\hat{C}_1$. The remaining two equations (1.105), with $\hat{C}_1 = \text{const.}$, are linear equations that can readily be integrated in terms of exponential or trigonometric functions. Thus setting $\hat{C}_j = r_j \exp(i\nu t)$ $(j = 2, 3)$, (1.105) have non-trivial solutions provided

$$\nu^2 = \sigma_2\sigma_3 K^2 |\hat{C}_1|^2 \equiv \frac{V^2}{4\omega_2\omega_3}|C_1|^2. \tag{1.106}$$

The second form of ν^2 in (1.106) follows by using the expression (1.88) for the resonant coupling coefficient K, the relation (1.86) between $\hat{z}_j$ and z_j, and noting that $z_j = C_j \exp(i\theta_j)$.

The equation (1.106) has pure imaginary solutions for ν indicating instability if $\omega_2\omega_3 < 0$. Note that the resonance conditions (1.104) imply that

$$|\omega_1| > |\omega_2| \quad \text{and} \quad |\omega_1| > |\omega_3| \quad \text{if} \quad \omega_2\omega_3 < 0. \tag{1.107}$$

In quantum theory, the energy of an oscillator with frequency ω is $E = \hbar\omega$, where $\hbar = h/2\pi$, and h is Planck's constant. Thus in a quantum interpretation, the resonance condition (1.104) corresponds to the conservation of energy $E_2 = E_1 + E_3$. Note that for instability $\omega_2\omega_3 < 0$ and $|E_1| > |E_2|, |E_3|$. One can think of the interaction as a three-oscillator interaction, in which oscillator 1 interacts with oscillator 3 resulting in the excitation and energy transfer to oscillator 2. It should also be kept in mind, that in the above analysis the amplitudes of oscillators 2 and 3 and the interaction strength V are assumed to be so small that the amplitude and energy of oscillator 1 remains unaffected by the interaction. It is important to emphasize that the above results are based on linear perturbation theory, and the initial exponential growth of x_2 and x_3 will eventually saturate when nonlinear effects come into play. It is straightforward to see from (1.79) that the Hamiltonian, or total energy of the system, does not depend explicitly on t, and hence is a constant of the motion. The fully nonlinear system is conservative, and the exact solution of the initial value problem for (1.81) would show a continuous sloshing of energy between the different components of the system.

Three-Wave Resonant Interaction Equations

An early review of nonlinear three-wave resonant interaction equations in nonlinear optics, parametric amplifiers and water waves has been given by Kaup et al. [99]. Methods of deriving the equations from first principles are described in the monograph by Anile et al. [9], whereas the solution of the equations by the inverse scattering method is described by Ablowitz and Segur [3]. The equations in a lossless, homogeneous medium, in one

Cartesian space dimension have the form:

$$\begin{aligned}
&\left(\frac{\partial}{\partial t}+v_i\frac{\partial}{\partial x}\right)a_i=p_iKa_ja_k,\\
&\left(\frac{\partial}{\partial t}+v_j\frac{\partial}{\partial x}\right)a_j=-p_jK^*a_ia_k^*,\\
&\left(\frac{\partial}{\partial t}+v_k\frac{\partial}{\partial x}\right)a_k=-p_kK^*a_ia_j^*,
\end{aligned}\tag{1.108}$$

where the a's are the complex wave packet amplitudes, the v's are the group velocities, the p's are the signs of the wave energies and K is the complex coupling coefficient. The three interacting waves must satisfy the resonance conditions

$$\omega_j+\omega_k=\omega_i,\tag{1.109}$$

$$k_j+k_k=k_i.\tag{1.110}$$

The frequency resonance condition (1.109) is exactly that obtained in the case of three coupled oscillators. This resonance condition corresponds to the conservation of energy of the waves in an elementary wave interaction. The second resonance condition on the wavenumbers (1.110) corresponds to conservation of wave momentum of the waves during the interaction. The space independent versions of these equations are in fact the resonant three-oscillator equations (1.98), with an appropriate change of nomenclature. The equations have solutions in terms of Jacobian elliptic functions [10, 158]. For positive energy waves ($p_i=p_j=p_k$), the solutions are periodic in time. The corresponding solution for $p_j=p_k=-p_i$ was derived by Coppi et al. [48]. These latter solutions become singular at a finite time, corresponding to an explosive instability.

A simple solution of (1.108) that shows blow-up at a finite time, can be obtained by assuming that K and the a's are real, and that $p_i=1$, $p_j=p_k=-1$; and by searching for solutions of (1.108) which depend only on t. It turns out that a solution of the equations exists, in which $a_i=a_j=a_k=a$, where a satisfies the ODE

$$\frac{da}{dt}=Ka^2.\tag{1.111}$$

The latter equation has the simple solution

$$a=\frac{1}{K(t_\infty-t)},\quad t_\infty=\frac{1}{a_0K},\tag{1.112}$$

where a_0 is the value of a at time $t = 0$. The above solution exhibits blow-up at time $t = t_\infty$. This example and other simple parametric and nonlinear differential equation modeling instabilities are discussed by Weiland and Wilhelmsson [196, Ch. 1].

Related resonant wave interaction equations for hyperbolic systems of equations were derived by Majda and Rosales [123], who obtained equations describing the resonant interaction of sound waves and entropy waves, in one Cartesian space dimension. The latter equations also contained Burgers self-wave steepening terms for the sound waves, which are not present in the above resonant interaction equations (1.108). The generalization of the Majda-Rosales equations, to describe three-wave resonant interactions in magnetohydrodynamics, were derived by Ali and Hunter [8] and Webb et al. [193,194].

1.5 Group Velocity, Wave Action and Wave Energy Equations

In this section we discuss the general notion of group velocity and its role in wave conservation laws, and the wave action and wave energy equations that arise in WKB analysis. We first discuss the method of stationary phase and Lighthill's method of constructing the group velocity surface from the wavenumber surface. This is followed by a discussion and examples of alternative ways of viewing the group velocity, using the method of characteristics for first order PDEs, and the role of the group velocity in the WKB approximation for short wavelength waves.

Group Velocity Surface and Stationary Phase

Consider a plane wave with phase

$$S = \mathbf{k} \cdot \mathbf{r} - \omega(\mathbf{k})t, \tag{1.113}$$

where $\mathbf{k}$ and ω satisfy the dispersion equation

$$D(\mathbf{k}, \omega) = 0. \tag{1.114}$$

In the asymptotic evaluation of Fourier integrals by the method of stationary phase (e.g. [198, Section 11.4]), it is argued that the main contribution to the integral comes from the region of $\mathbf{k}$-space, in the vicinity of the point where the phase is stationary. From (1.113), one obtains

$$\delta S = \delta\mathbf{k} \cdot (\mathbf{r} - \omega_{\mathbf{k}} t) = 0, \tag{1.115}$$

as the condition for the phase to be stationary.

By differentiation of (1.114), one obtains

$$\mathbf{V}_g = \frac{\partial\omega}{\partial\mathbf{k}} = -\frac{D_{\mathbf{k}}(\mathbf{k},\omega)}{D_\omega(\mathbf{k},\omega)}, \tag{1.116}$$

for the group velocity $\mathbf{V}_g$ of waves with wavevector $\mathbf{k}$.

Thus the group velocity surface

$$\mathbf{r} = \mathbf{V}_g t \tag{1.117}$$

is defined as the surface on which the phase is stationary.

Lighthill's method of stationary phase (e.g. [116]) for constructing the group velocity surface from the wavenumber surface is described below. The wavenumber surface $D(\mathbf{k},\omega) = 0$ with ω fixed, is the surface in $\mathbf{k}$-space, with normal $\mathbf{n} = D_{\mathbf{k}}/|D_{\mathbf{k}}|$. Since $\mathbf{V}_g = V_g\mathbf{n}$, it follows that the group velocity surface can be written in the form:

$$\mathbf{r} = \frac{(\mathbf{k}\cdot\mathbf{r})\mathbf{n}}{\mathbf{k}\cdot\mathbf{n}} = \frac{\phi\mathbf{n}}{\mathbf{k}\cdot\mathbf{n}}, \tag{1.118}$$

where

$$\phi = \mathbf{k}\cdot\mathbf{r} \equiv S + \omega t. \tag{1.119}$$

Since S is stationary, and ω is fixed, the phase $\phi = S + \omega t$ is a constant for fixed t. This allows one to geometrically construct the wave group velocity surface from the wavenumber surface $D(\mathbf{k},\omega) = 0$ as follows. First, construct the wavenumber surface. At a generic point P on the wavenumber surface ($\mathbf{k} = \mathbf{OP}$ is the wavevector, where O is the origin in $\mathbf{k}$-space), determine the wavenormal $\mathbf{n} = D_{\mathbf{k}}/|D_{\mathbf{k}}|$. Draw the tangent plane to the wavenumber surface at P, and determine the perpendicular distance, $OT = \mathbf{k}\cdot\mathbf{n}$ between the tangent plane and the origin. The group velocity surface (1.118) is then given by

$$\mathbf{r} = \frac{\phi\mathbf{n}}{OT}. \tag{1.120}$$

Since ϕ is a constant, the shape of the group velocity surface can be obtained by setting $\phi = 1$. The resulting group velocity surface $\mathbf{r} = \mathbf{n}/OT$ is the reciprocal polar of the wavenumber surface.

This method of constructing the group velocity surface is widespread in wave theory. Examples include the wavenumber and group velocity surfaces for magnetohydrodynamic waves [52,116,127,163,199] and ship waves for deep water gravity waves (e.g. [198]).

Below, we illustrate the role of the group velocity in WKB theory, associated wave conservation laws, and the role of the theory of characteristics in defining the ray equations of geometrical optics.

Example

Consider the linear wave dispersion equation:

$$D(\mathbf{k},\omega) = k_x\omega - \lambda_a k_y^2 - \lambda_b k_z^2 = 0. \tag{1.121}$$

Using the Fourier analysis correspondence $\omega \rightarrow i\partial/\partial t$ and $\mathbf{k} \rightarrow -i\nabla$, the above dispersion equation results in the linear wave equation:

$$u_{tx} + \lambda_a u_{yy} + \lambda_b u_{zz} = 0. \tag{1.122}$$

This equation arises, for example, in the description of weakly diffracting sound waves and magnetoacoustic waves, where the diffraction coefficients λ_a and λ_b govern the curvature of the wave front. The equation is the linearized version of the inviscid Burgers equation:

$$(u_t + uu_x)_x + \lambda_a u_{yy} + \lambda_b u_{zz} = 0. \tag{1.123}$$

In magnetohydrodynamics, the diffraction coefficients λ_a and λ_b are both positive for the fast magnetoacoustic wave, but λ_a and λ_b can have opposite signs for the slow magnetoacoustic wave (e.g. [190]).

The wave group velocity $\mathbf{V}_g = \nabla_{\mathbf{k}}\omega$ can be obtained by differentiation of the dispersion equation, and is given by the formula:

$$\mathbf{V}_g = -\left(\lambda_a \frac{k_y^2}{k_x^2} + \lambda_b \frac{k_z^2}{k_x^2}\right)\mathbf{e}_x + 2\lambda_a \frac{k_y}{k_x}\mathbf{e}_y + 2\lambda_b \frac{k_z}{k_x}\mathbf{e}_z, \tag{1.124}$$

where $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_z$ are unit vectors along the x, y and z axes. The equations $\mathbf{r} \equiv \mathbf{r}_g = \mathbf{V}_g t$ for a fixed t define the group velocity surface.

In particular, the equations:

$$x = -\left(\lambda_a \frac{k_y^2}{k_x^2} + \lambda_b \frac{k_z^2}{k_x^2}\right)t, \quad y = 2\lambda_a \frac{k_y}{k_x}t, \quad z = 2\lambda_b \frac{k_z}{k_x}t, \tag{1.125}$$

can be used to parametrically define the group velocity surface in terms of the two parameters k_y/k_x and k_z/k_x. Alternatively, eliminating reference to k_y/k_x and k_z/k_x in the above equations yields the equation:

$$s(x,y,z,t) \equiv x + \frac{y^2}{4\lambda_a t} + \frac{z^2}{4\lambda_b t} = 0, \tag{1.126}$$

for the group velocity surface. For $\lambda_a\lambda_b > 0$, the group velocity surface is a paraboloid of one sheet, but for $\lambda_a\lambda_b < 0$, the surface is a hyperboloid.

WKB Analysis

Next consider the WKBJ ansatz [198, Ch. 7], :

$$u = \sum_{n=0}^{\infty} \Phi_n(x,y,z,t) f_n(S), \tag{1.127}$$

for solutions of the diffractive wave equation (1.122), where the functions $f_n(S)$ are chosen so that

$$f_n'(S) = f_{n-1}(S), \tag{1.128}$$

with S defining the wave phase. The choice

$$f_n(S) = \exp(i\omega S)(i\omega)^{-n}, \quad S = t - \sigma(\mathbf{x}), \tag{1.129}$$

corresponds to the usual WKB expansion for high frequency waves (this choice is not essential, and is not used in the present analysis).

Substituting the expansion (1.127) into the wave equation (1.122), and setting the coefficients of the $f_n(S)$ equal to zero, we obtain a sequence of equations for S and the Φ_n. At the lowest order we set the coefficient of $f_{-2}(S)$ equal to zero [note $f_{-2}(S) = f_0''(S)$] and obtain the wave eikonal equation:

$$S_x S_t + \lambda_a S_y^2 + \lambda_b S_z^2 = 0. \tag{1.130}$$

Using the identifications

$$\omega = -S_t, \quad \mathbf{k} = (S_x, S_y, S_z), \tag{1.131}$$

for the frequency ω and wavevector $\mathbf{k}$, the wave eikonal equation (1.130), becomes

$$D(\mathbf{k},\omega) = \omega k_1 - \lambda_a k_2^2 - \lambda_b k_3^2 \equiv -(S_x S_t + \lambda_a S_y^2 + \lambda_b S_z^2) = 0, \tag{1.132}$$

which is the dispersion equation (1.121).

At the next highest order, we require the coefficient of $f_{-1}(S)$ to be zero, yielding a linear first order PDE for the evolution of the wave amplitude, namely

$$\begin{aligned}&\Phi_{0,x}S_t + S_x\Phi_{0,t} + \Phi_0 S_{xt}\\ &\qquad + \lambda_a(\Phi_0 S_{yy} + 2\Phi_{0,y}S_y) + \lambda_b(\Phi_0 S_{zz} + 2\Phi_{0,z}S_z) = 0. \qquad (1.133)\end{aligned}$$

Similarly, equations for the higher order wave amplitudes $\{\Phi_n : \ n \geq 1\}$ can be obtained.

WKB Conservation Laws

Equation (1.133) can be re-arranged to yield the conservation equation:

$$\frac{\partial}{\partial t}(\Phi_0^2 S_x) + \frac{\partial}{\partial x}(\Phi_0^2 S_t) + \frac{\partial}{\partial y}(2\lambda_a S_y \Phi_0^2) + \frac{\partial}{\partial z}(2\lambda_b S_z \Phi_0^2) = 0. \quad (1.134)$$

By noting that $V_{gx} = S_t/S_x$ and the results (1.124) for the group velocity, the above equation can be reduced to the wave energy equation:

$$\frac{\partial}{\partial t}(E_w) + \nabla \cdot (\mathbf{V}_g E_w) = 0, \qquad (1.135)$$

where $E_w = \Phi_0^2$ is the wave energy density.

Further conservation laws can be obtained by manipulating the wave amplitude equation (1.133), and by using the results

$$\frac{dk_i}{dt} = \frac{\partial k_i}{\partial t} + \mathbf{V}_g \cdot \nabla k_i = 0, \quad \frac{d\omega}{dt} = \frac{\partial \omega}{\partial t} + \mathbf{V}_g \cdot \nabla \omega = 0. \qquad (1.136)$$

The first of these equations follows from using the continuity of the partial derivatives of S, namely $S_{x_i t} = S_{t x_i}$, whereas the second equation is obtained by differentiating the wave eikonal equation (1.130). These latter results can also be derived from the characteristic equations for the wave eikonal equation (see below).

Introducing the wave action

$$A = -E_w/S_t, \quad E_w = \Phi_0^2, \qquad (1.137)$$

where E_w is the wave energy density, equation (1.135) yields the wave action conservation equation:

$$\frac{\partial A}{\partial t} + \nabla \cdot (\mathbf{V}_g A) = 0. \qquad (1.138)$$

In general, the wave action conservation equation is different from the wave energy equation. For WKB wave propagation in an inhomogeneous background flow (e.g. the propagation of WKB Alfvén waves in the Solar Wind, [55,89]) the wave action is conserved, but the physical wave energy is not conserved as the wind does work on the waves.

Using the fact that $\mathbf{k}$ is constant in the group velocity frame, as described by (1.136), it is straightforward to derive the canonical wave momentum equation:

$$\frac{\partial(A\mathbf{k})}{\partial t}+\nabla\cdot(\mathbf{V}_g A\mathbf{k})=0, \tag{1.139}$$

for the system. The above conservation laws could be derived by Lagrangian formulations of the equations, coupled with an application of Noether's theorem. These results can be related to physical and canonical wave stress energy tensors and conservation laws in Lagrangian field theories (see, e.g. [55,56]).

Characteristic Equations

Since the wave eikonal equation (1.132) is a first order PDE for S, its solutions can be obtained by the method of characteristics (e.g. [171, Ch. 2]). The characteristics are:

$$\begin{aligned}&\frac{dx_i}{d\tau}=D_{k_i},\quad \frac{dt}{d\tau}=-D_\omega,\quad \frac{dS}{d\tau}=\omega D_\omega+\mathbf{k}\cdot D_{\mathbf{k}},\\ &\frac{dk_i}{d\tau}=-\left(D_{x_i}+D_S S_{x_i}\right),\quad \frac{d\omega}{d\tau}=D_t+D_S S_t,\end{aligned} \tag{1.140}$$

where τ is a parameter along the characteristics. The characteristics, in WKB theory, are also known as the ray equations and are used for ray tracing in geometrical optics in inhomogeneous media. The use of the ray equations in this case assumes that the wavelength of the wave λ is much less than the characteristic scale of the inhomogeneous background medium L (i.e. $\lambda \ll L$). The first two characteristics in (1.140) can be combined to yield the equations:

$$\frac{d\mathbf{x}}{dt}=\mathbf{V}_g=-\frac{D_{\mathbf{k}}}{D_\omega}. \tag{1.141}$$

Thus on the characteristics, the derivative

$$\frac{d}{d\tau}=-D_\omega\frac{d}{dt}\equiv -D_\omega\left(\frac{\partial}{\partial t}+\mathbf{V}_g\cdot\nabla\right), \tag{1.142}$$

corresponds to the comoving derivative in the group velocity frame.

In the case where $D(\mathbf{k},\omega)=\omega-W(\mathbf{k},\mathbf{x},t)$, the ray equations (1.140) can be written in the Hamiltonian form:

$$\frac{d\mathbf{x}}{dt}=-W_{\mathbf{k}},\quad \frac{d\mathbf{k}}{dt}=W_{\mathbf{x}}, \tag{1.143}$$

where W is the Hamiltonian, and $\mathbf{x}$ and $\mathbf{k}$ are the canonically conjugate variables. The dispersion equation:

$$D(\mathbf{k},\omega)=\omega-W(\mathbf{k},\mathbf{x},t)\equiv-\left(S_t+W(\nabla S,\mathbf{x},t)\right)=0 \tag{1.144}$$

is then equivalent to the Hamilton-Jacobi equation.

Solution Example

It is straightforward to verify that the function $s(x,y,z,t)$ used in (1.126) to define the group velocity surface is a solution of the wave eikonal equation (1.130). Below we show that this solution $S=s$ can be derived by integrating the characteristics. The solution $S=s$ can also be regarded as an envelope of plane wave solutions which are tangent to the group velocity surface.

The characteristics (1.140) for the eikonal equation (1.132), reduce to:

$$\begin{aligned}&\frac{dx}{d\tau}=\omega,\quad \frac{dy}{d\tau}=-2\lambda_a k_2,\quad \frac{dz}{d\tau}=-2\lambda_b k_3,\quad \frac{dt}{d\tau}=-k_1,\\ &\frac{d\omega}{d\tau}=\frac{dk_1}{d\tau}=\frac{dk_2}{d\tau}=\frac{dk_3}{d\tau}=0,\quad \frac{dS}{d\tau}=2D(\mathbf{k},\omega)\equiv 0.\end{aligned} \tag{1.145}$$

Thus ω, k_1, k_2, k_3 and S are all constant on the characteristics. The characteristic equations for x, y, and z can be integrated to yield the integrals:

$$x+\frac{\omega}{k_1}t=c_1,\quad y-2\lambda_a\frac{k_2}{k_1}t=c_2,\quad z-2\lambda_b\frac{k_3}{k_1}t=c_3, \tag{1.146}$$

where c_1, c_2 and c_3 are integration constants. Using the method of characteristics, it follows that a general solution for S is of the form:

$$S=g(c_1,c_2,c_3)\equiv g\left(x+\frac{\omega}{k_1}t,y-2\lambda_a\frac{k_2}{k_1}t,z-2\lambda_b\frac{k_3}{k_1}t\right), \tag{1.147}$$

where g is an arbitrary, differentiable function of its arguments.

If we choose $c_2 = c_3 = 0$ and $g = c_1$, then the integrals (1.146), and the eikonal equation (1.130) yield

$$\frac{k_2}{k_1} = \frac{y}{2\lambda_a t}, \quad \frac{k_3}{k_1} = \frac{z}{2\lambda_b t}, \quad S = x + \left(\lambda_a \frac{k_2^2}{k_1^2} + \lambda_b \frac{k_3^2}{k_1^2}\right) t. \tag{1.148}$$

Eliminating k_2/k_1 and k_3/k_1 from (1.148) yields the solution

$$S = x + \frac{y^2}{4\lambda_a t} + \frac{z^2}{4\lambda_b t}, \tag{1.149}$$

of the wave eikonal equation (1.130). The equation $S = 0$ for the above solution is the group velocity surface (1.126).

Since $\mathbf{k}$ and ω are constant on the characteristics, the wave eikonal possesses a complete integral in the form of a plane wave, namely

$$S = k_1 x + k_2 y + k_3 z - \omega t, \tag{1.150}$$

where k_1, k_2, k_3, and ω are constants satisfying (1.132). The general solution for S consists of an envelope of plane waves. In particular, the envelope of the family of plane wave solutions

$$S = x + k_2 y + k_3 z - (\lambda_a k_2^2 + \lambda_b k_3^2) t, \tag{1.151}$$

obtained by regarding k_2 and k_3 as parameters defining the normals to the plane waves, yields the group velocity solution surface $S = s$. The envelope is obtained by setting $S_{k_2} = S_{k_3} = 0$, and subsequently solving these equations for k_2 and k_3, and substituting for k_2 and k_3 in the plane wave solution (1.151).

It is of interest to note that the original PDE (1.122) for u has the Green function solution

$$u = \frac{\delta(S)}{4\pi |\lambda_a \lambda_b|^{12} t}, \tag{1.152}$$

where $\delta(S)$ is the Dirac delta distribution [190].

1.6 Project Assignment

For your choice of a PDE, or a system of equations, provide: derivation, description of physical scales and units, well-posedness results known for the full or reduced system, classification of the linearized equations, symmetries, invariants and special solutions of the full system.

1.7 Project Sample

Simulations of semiconductor devices are often employed for systematic design and analysis of performance limiting processes, allowing optimization of structures with complex geometry, material composition and realistic physical properties. In this section we discuss the derivation of the system of PDEs that form the basis of many semiconductor device simulation models.

Physical models of semiconductor devices account for two coupled principal phenomena: charged carrier (electron and hole) transport and electromagnetic field generation and distribution due to the electric charges and currents. An *ab initio* quantum mechanical approach will fully characterize the system, but, due to the relatively large device volumes currently used, this method is not practical for most situations. Thus a hierarchy of semiconductor simulation models is considered in which the physical method is based either on the statistical description of the system, such as the Boltzmann equation and Monte Carlo particle methods, or on the momentum averaged description, such as the hydrodynamic transport and drift-diffusion models. The rest of this section focuses on the derivation of the semiconductor device simulation equations in the drift-diffusion approximation.

The electromagnetic field distribution inside and outside of the device volume is described by Maxwell's equations:

$$\begin{aligned}
&\frac{\partial \mathbf{D}}{\partial t} = \nabla \times \mathbf{H} - \mathbf{J}, \quad \nabla \cdot \mathbf{D} = \rho, \\
&\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E}, \quad \nabla \cdot \mathbf{B} = 0
\end{aligned}$$

where $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields, $\mathbf{D} = \epsilon\mathbf{E}$, $\mathbf{B} = \mu\mathbf{H}$ and ϵ, μ are the material permittivity and permeability. ρ and $\mathbf{J}$ are the total electric charge and current density, related by the charge continuity equation, that can be derived by applying the divergence operator to Maxwell's equation for the displacement current $\partial\mathbf{D}/\partial t = \nabla \times \mathbf{H} - \mathbf{J}$:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0. \tag{1.153}$$

Many practically important devices operate on time-scales and have spatial dimensions that result in a small contribution from the time-derivative of the magnetic field, $\partial \mathbf{B}/\partial t \approx 0$. Thus $\nabla \times \mathbf{E} = 0$ and a quasi-static approximation can be used, with the electric-field given by the gradient of the potential, $\mathbf{E} = -\nabla\phi$. The spatial distribution of the potential is therefore determined by Poisson's equation:

$$\nabla \cdot (\epsilon \mathbf{E}) = -\nabla \cdot (\epsilon \nabla \phi) = \rho. \tag{1.154}$$

The charge density in the equation (1.154) can be written as $\rho = e(p - n) + N_D - N_A$, where e is the elementary charge, n and p are the electron and hole number densities, and N_D, N_A are the charge densities due to the donor and acceptor doping, respectively. Similarly, the current density can be represented by the sum of the electron and hole currents: $\mathbf{J} = \mathbf{J}_n + \mathbf{J}_p$. Substituting these equations for ρ and $\mathbf{J}$ into the continuity equation (1.153) and into equation (1.154), one obtains a coupled system of charge number-density conservation equations for n, p and the electrostatic potential ϕ:

$$e\frac{\partial n}{\partial t} = \nabla \cdot \mathbf{J}_n + e(G - R), \tag{1.155}$$

$$e\frac{\partial p}{\partial t} = -\nabla \cdot \mathbf{J}_p + e(G - R), \tag{1.156}$$

$$-\nabla \cdot (\epsilon \nabla \phi) = e(p - n + N_D - N_A), \tag{1.157}$$

where G and R represent the electron-hole pair generation and recombination rates,

$$\begin{aligned} R = &\frac{np - n_i^2}{\tau_p(n + n_i) + \tau_n(p + n_i)} \\ &+ (np - n_i^2)B + (np - n_i^2)(C_n n + C_p p), \end{aligned} \tag{1.158}$$

and $n_i, \tau_{n,p}, B, C_{n,p}$ are material dependent constants. Carrier recombination takes place through three main processes, represented by three terms in the equation above: the Shockley-Read-Hall (SRH) mechanism, electron-hole spontaneous recombination and Auger recombination. The SRH processes take place through the mediation of trap levels localized in the semiconductor band gap, a mechanism that results in a recombination rate proportional to the carrier concentrations for $n \gg n_i$. The spontaneous recombination rate results in the emission of a photon, and is proportional to the product of electron-hole number densities; and for the high carrier densities, the Auger recombination, occurring when carriers perform direct transitions between the conduction and valence bands, with the energy transferred to another particle, also becomes important, and is proportional to the third power of the carrier density. Equations (1.155)–(1.158), along with the expressions for the current densities $\mathbf{J}_n, \mathbf{J}_p$ derived below as functions of n, p, ϕ, form the basis of the device simulation model.

The current densities $\mathbf{J}_{n,p}(n, p, \mathbf{E})$ can be derived by starting from the Boltzmann kinetic transport equation for electrons and holes ($\nu = n, p$) [164]:

$$\frac{\partial f_\nu}{\partial t} + \frac{\mathbf{F}_{\nu e}}{\hbar}\nabla_k f_\nu + \mathbf{u}_\nu \nabla_x f_\nu = -\int f_\nu(\mathbf{k})[1 - f_\nu(\mathbf{k}')]S(\mathbf{k}, \mathbf{k}') \\ - f_\nu(\mathbf{k}')[1 - f_\nu(\mathbf{k})]S(\mathbf{k}', \mathbf{k})d\mathbf{k}', \quad (1.159)$$

where $\hbar$ is the Plank constant normalized by 2π, $\mathbf{F}_{\nu e} = e\mathbf{E}$ is the force exerted on the charge, $\mathbf{p} = \hbar\mathbf{k}$ is the particle momentum, and $S(\mathbf{k}, \mathbf{k}')$ is the rate of scattering of particles with momentum $\mathbf{k}$ to $\mathbf{k}'$. The distribution function $f_\nu(\mathbf{x}, \mathbf{k}, t) = f_{\nu_0}(\mathbf{x}, \mathbf{k}, t) + f_{\nu_1}(\mathbf{x}, \mathbf{k}, t)$ is assumed to be close to the equilibrium Fermi-Dirac statistics at temperature T:

$$f_{\nu_0} = \frac{1}{1 + \exp[\pm(E_\xi(\mathbf{x}, \mathbf{k}) - E_{f\nu}(\mathbf{k}))/k_B T]}, \quad \xi = c, v,$$

where k_B is the Boltzmann constant, E_c and E_v are the conduction and valence band energies, and $E_{f\nu}$ is the Fermi energy level for electrons and holes. The physical origin of the terms contributing to the current density can be most readily seen by simplifying the equation (1.159) to one space dimension and using the relaxation-time approximation:

$$-\frac{eE}{m_e^*}\frac{\partial f}{\partial u} + u\frac{\partial f}{\partial x} = -\frac{f - f_0}{\tau}, \quad (1.160)$$

where, for simplicity, only the equation for electrons is considered, $u = p/m_e^*$ and m_e^* are the carrier velocity and its effective mass, and τ is the characteristic relaxation time. By taking the first moment of the Boltzmann equation (multiplying equation (1.160) by the charged particle velocity and integrating over the u) and defining $J = -e\int_u uf du$, one obtains:

$$-\frac{eE}{m_e^*}\int_u \frac{\partial f}{\partial u}du + \int_u u\frac{\partial f}{\partial x}du = \frac{J}{e\tau}.$$

Defining electron density $n(x) = \int_u f du$, the first term, after integration by parts, can be seen to be equal to $n(x)eE/m_e^*$. Then, re-writing the second integral in terms of $\langle u^2 \rangle = [\int_u u^2 f du]/n(x)$, the current density J can be expressed as:

$$J = e\tau \left[n(x)eE/m_e^* + \langle u^2 \rangle \frac{\partial n(x)}{\partial x} + n(x) \frac{\partial \langle u^2 \rangle}{\partial x} \right].$$

Finally, neglecting the energy drift term proportional to $\partial\langle u^2\rangle/\partial x$, and defining electron mobility μ_n and diffusion D_n constants, the current density becomes:

$$J = \mu_n n(x)eE + eD_n \frac{\partial n(x)}{\partial x},$$

where the first term corresponds to the current induced by the carrier drift in the electric field, while the second term represents the carrier concentration gradient dependent transport via diffusion. Similar derivations can be carried out for the holes and the result can be generalized to multiple space dimensions by defining $\mathbf{J}_\nu = \mp \frac{e}{4\pi^3} \int_{V_k} \mathbf{u}_\nu f_\nu d\mathbf{k}$, which results in the following expressions for the current densities:

$$\mathbf{J}_n = en\mu_n \mathbf{E} + eD_n \nabla n,$$
$$\mathbf{J}_p = ep\mu_p \mathbf{E} - eD_p \nabla p.$$

The above equations, along with the system (1.155)–(1.157) and appropriate carrier recombination-generation models, constitute the basic components of the semiconductor device simulation drift-diffusion model. For application in a proper physical context, the approximations made during the derivation of the model have to be considered. These include: parabolic energy bands $E = p^2/2m^*$, field independent mobility and diffusion coefficients, simple analytic models for the carrier recombination rates in equation (1.158), assumption that the typical time-scales involved in the problem are much larger than the relaxation time τ, small spatial gradients in the fields, small contribution from thermal effects and negligible energy drift. Additional terms, or a quantum mechanical description, must be used in carrier transport equations to account for these effects.

While for practical problems a finite-volume or finite-element discretization [164] is typically used to obtain numerical solutions of the semiconductor device model equations, the 1D carrier drift-diffusion equation can be solved exactly for a problem of carrier transport in a bulk semiconductor with a simple n/τ approximation for the recombination rate [181]. Consider a p-type semiconductor material slab, across which an external electric field of magnitude E_a is applied, Figure 1.2 (inset). If an incident light pulse generates excess carriers, with the number of electrons

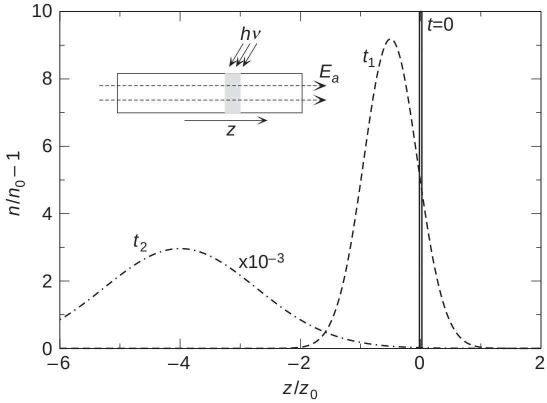


Figure 1.2: Time-snapshots at $t = t_1$,t_2 of electron distribution in a slab of bulk semiconductor in an external electric field. The slab is illuminated at $t = 0$ by a pulse of light, with generated excess carriers undergoing recombination, diffusion and drift.

or holes generated per unit area given by C^{gen}, then the subsequent evolution of the carrier distribution can be approximately described by the carrier transport equation:

$$\frac{\partial n}{\partial t} = -\frac{n - n_0}{\tau_n} + \mu_n E_a \frac{\partial n}{\partial z} + D_n \frac{\partial^2 n}{\partial z^2}.$$

Integrating this equation with an initial condition corresponding to the instantaneous carrier generation, one obtains:

$$n(t, z) = n_0 + \frac{C^g}{\sqrt{4\pi D_n t}} \exp\left[-\frac{(z - E_a \mu_n t)^2}{4 D_n t} - \frac{t}{\tau_n}\right].$$

As the solution shows in Figure 1.2, the excess carriers recombine, diffuse away from the point of injection and are advected with the drift velocity $E_a \mu_n$.

This page intentionally left blank

Chapter 2

Discretization Methods

Discretization of partial differential equations (PDEs) is based on the theory of function approximation. Several choices have to be made. The form of the equations needs to be chosen: for example, can the problem be written as a differential equation system in terms of an integral equation formulation, or in terms of an approximate solution operator? The type of space-time discretization, defined by the function subspace in which we approximate the solution, needs to be decided, as well as the choice of grids: e.g. regular or irregular grids to fit the geometry, or static versus solution adaptive grids.

In this chapter we explore some of the common approaches to the choice of form of the PDE and the space-time discretization, while leaving discussion of adaptive and moving grids for a later chapter. The goal is to introduce the reader to various forms of discretization and to illustrate the numerical performance of the different methods in order to build up experience. The questions raised by the analysis will be discussed in the following chapter. In particular, we will address how to choose a method that is accurate, robust and efficient for the problem at hand.

In general, the approximation of the solution may be written in the form:

$$u(\mathbf{x},t) \approx \sum_{i=1}^{N} c_i(t)\phi_i(\mathbf{x}).$$

Popular choices of the basis functions $\phi_i(\mathbf{x})$ include translations of a single function, $\phi_i(\mathbf{x}) = \phi(\mathbf{x} - \mathbf{x}_i)$, with $\phi(\mathbf{x})$ being top-hat, B-spline, Gaussian, sech, polynomial and rational approximations. For example, the choice

$$\phi_i(x) = H(x - x_{i-1}) - H(x - x_i),$$

where $H(x)$ is the Heaviside step function, corresponds to the finite difference approximation in which $u(x_i,t) = c_i(t)$ if $x_{i-1} < x < x_i$. The compact finite difference schemes use the Padé approximation involving rational functions. Spline methods involve piecewise polynomial fits, in which both derivatives and the function values are matched at the grid points. Finite element methods (FEMs) for elliptic problems are usually based on a variational formulation of the problem, use an appropriate interpolation between the nodes and with a choice of the parameters used in the approximation that assures the variational functional is stationary. Particle methods give rise to another class of approximations and discretization. The use of global orthogonal polynomials (e.g. eigenfunctions of a self-adjoint operator suggested by symmetry of the problem or eigenfunctions chosen due to accuracy considerations for smooth solutions) leads to the discretizations that go under names of spectral, spectral finite element and spectral elements (also known as h-p finite elements).

2.1 Polynomial Interpolation and Finite Differences

Consider a one-dimensional (1D) situation, and take polynomial basis functions that pass through a few neighboring points with respect to the nodes x_i. In this case, the expansion coefficients may be thought of as approximations to the exact solution $u(x,t)$, where $c_i(t) \approx u(x_i,t)$.

Lagrangian Interpolation

In Lagrangian interpolation, $u(x,t) = u_i(t)$ at the grid points $x = x_i$, where $i_1 \le i \le i_2$, and the approximate polynomial fit to $u(x,t)$ is of the form:

$$u(x,t) = \sum_{i=i_1}^{i_2} u_i(t) l_i(x),$$

where

$$l_i(x) = \frac{(x - x_{i_1})(x - x_{i_1+1}) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_{i_2})}{(x_i - x_{i_1})(x_i - x_{i_1+1}) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_{i_2})}$$

is the i^{th} Lagrangian interpolation polynomial. Note that $l_i(x_j) = \delta^i_j$ where δ^i_j is the Kronecker delta symbol, and hence, $u(x_i,t) = u_i(t)$. Derivatives with respect to x are simply obtained by differentiating each Lagrangian interpolation polynomial, term by term, in the series.

Example

Consider, for example, a parabolic approximation using Lagrangian interpolation, in which the parabola passes through the three-points (x_{i-1}, u_{i-1}), (x_i, u_i), (x_{i+1}, u_{i+1}), and use it to compute derivative approximations. The approximation used is

$$u(x) = u_{i-1}l_{i-1}(x) + u_i l_i(x) + u_{i+1}l_{i+1}(x).$$

The Lagrangian polynomial $l_{i-1}(x)$ in this case is given by

$$l_{i-1}(x) = \frac{(x - x_i)(x - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})},$$

and similar formulae define $l_i(x)$ and $l_{i+1}(x)$. Since the approximation involves a quadratic polynomial in x, the second derivative is constant, and the higher order derivatives are zero. Evaluating the first derivative of the parabola at x_i we obtain:

$$u'(x_i) \approx -\frac{u_{i-1}h_i}{h_{i-1}(h_{i-1} + h_i)} + \frac{u_i(h_i - h_{i-1})}{h_i h_{i-1}} + \frac{u_{i+1}h_{i-1}}{h_i(h_{i-1} + h_i)},$$

where

$$h_i = x_{i+1} - x_i.$$

Similarly, the constant second derivative is:

$$u''(x_i) \approx \frac{2}{h_i + h_{i-1}} \left(\frac{u_{i+1} - u_i}{h_i} - \frac{u_i - u_{i-1}}{h_{i-1}} \right). \tag{2.1}$$

Evaluating the first derivative at the end point x_{i-1} on a uniform grid with $h_i = h_{i-1} = h$ results in the formula:

$$u'(x_{i-1}) \approx \frac{1}{2h}(-3u_{i-1} + 4u_i - u_{i+1}).$$

Note that the derivative of the parabola is a linear function and the above formula can be obtained by using linear interpolation between derivative approximations at points $x_{i\pm1/2}$ to approximate the derivative at point x_i. This remark is useful in multi-dimensions, where instead of writing down a quadratic interpolant, one may use the fact that each 1D cut is a parabola and the derivative is a linear function of the independent variables.

Third and higher order derivatives may be computed similarly, for example on a uniform grid,

$$u_{xxx}(x_i) \approx D_+D_-D_+u(x_i) = \frac{u_{i+2} - 3u_{i+1} + 3u_i - u_{i-1}}{h^3},$$

or

$$\begin{aligned} u_{xxx}(x_i) &\approx \frac{1}{2}(D_+D_-D_+ + D_-D_+D_-)u(x_i) \\ &= \frac{u_{i+2} - 2u_{i+1} + 2u_{i-1} - u_{i-2}}{2h^3}, \end{aligned}$$

where

$$D_+[u(x)] = \frac{u(x+h) - u(x)}{h}, \quad D_-[u(x)] = \frac{u(x) - u(x-h)}{h},$$

are Euler's forward and backward difference operators.

Thus there are different difference operators representing the same derivative, resulting in different orders of accuracy, types of the error (dispersive, diffusive) and mode dynamics (exponentially growing modes, oscillatory modes, resonances, etc.).

Definition 5. *A finite difference operator P_h is consistent of order p in a variable x with a continuous operator P if $(P_h - P)v(x) = O(h^p)$ as $h \to 0$, where $v(x)$ is an arbitrary smooth function, $h = (h_1, h_2, \ldots, h_n)$ and $p = (p_1, p_2, \ldots, p_n)$ is a multi-index notation.*

Often $v(x)$ is assumed to be an exact solution of the discrete or the continuous problem. In the case when $v(x)$ is a solution of the discrete problem, we assume that it was obtained from discrete values by "connecting the dots" to form a function defined on the whole space, called the grid function. This may be achieved by using a piecewise continuous polynomial or global interpolation. However, it must be kept in mind that there will be no increase in accuracy if we use a higher order approximation than that used to evaluate the derivatives. The definition of consistency suggests that we use the Taylor series expansion in order to determine the order of consistency.

For example, expanding near $x_{i+1/2}$ and near x_i gives

$$\frac{u_{i+1} - u_i}{h} = u'(x_{i+1/2}) + O(h^2),$$

as a second-order approximation of the derivative, while

$$\frac{u_{i+1} - u_i}{h} = u'(x_i) + O(h),$$

is only a first-order approximation.

Similarly, the approximation of the second derivative near x_i on a non-uniform grid described in example (2.1) gives

$$\frac{2}{h_i + h_{i-1}} \left(\frac{u_{i+1} - u_i}{h_i} - \frac{u_i - u_{i-1}}{h_{i-1}} \right)$$
$$= u''(x_i) + \frac{1}{3}(h_i - h_{i-1})u'''(x_i) + O(h^2).$$

If we assume that the non-uniform grid is obtained by a smooth mapping $x(\xi)$, where ξ varies smoothly, we see that the change in the grid step, $(x_{i+1} - x_i) - (x_i - x_{i-1}) \approx x''(\xi)\Delta\xi^2$, is large if the grid derivative $x''(\xi)$ is large. This, in fact, is the key difficulty that arises at the grid interfaces when the grids are refined and have either steep gradients or vary discontinuously. Extra care should be taken to prevent reflections, oscillations and instabilities that may arise. This will be discussed in the sections on grid refinement and interface boundary conditions.

On uniform grids we may explore the difference approximations for derivatives using Fourier analysis, assuming the grid-function $u(x)$ has a period of N with respect to the node number of the grid point. Suppose $u(x)$ is given on $[a, b]$, then we can introduce a normalized variable $\theta = 2\pi(x - a)/(b - a)$, so that $0 \leq \theta \leq 2\pi$, where $\theta = 0$ corresponds to $x = a$ and $\theta = 2\pi$ corresponds to $x = b$. At a general point on the grid $\theta = kh$ and is known as the grid wavenumber; it corresponds to the k^{th} grid point where $h = 2\pi/N$ is the uniform grid spacing and N is the number of grid points. Setting $u(x) = w(\theta)$, the derivatives of $u(x)$ can be obtained from the derivatives of $w(\theta)$ (e.g. $u_x = w_\theta \theta_x$ and $u_{xx} = w_{\theta\theta}\theta_x^2$, where $\theta_x = 2\pi/(b - a)$). If we use the Discrete Fourier Transform (DFT) (e.g. [32, 39]) to represent the periodic continuation of the data $(x_j, u(x_j))$ on the interval $[a, b]$, then the analogue of Fourier's theorem is:

$$\hat{w}_k = \frac{1}{N} \sum_{j=0}^{N-1} w(\theta_j) \exp(-ik\theta_j), \tag{2.2}$$

$$w(\theta_j) = \sum_{k=-N/2}^{(N/2-1)} \hat{w}_k \exp(ik\theta_j), \tag{2.3}$$

where it is assumed that N is an even integer and

$$\theta_j = jh, \quad h = \frac{2\pi}{N}.$$

The discrete wavenumber k is an integer with $-N/2 \leq k \leq N/2 - 1$. Using (2.2), it follows that

$$\begin{aligned} F\left(\frac{1}{2}(D_+ + D_-)w(\theta_j)\right) &= \frac{\hat{w}_k(e^{ikh} - e^{-ikh})}{2h} \equiv i\frac{\hat{w}_k \sin(kh)}{h} \\ &\approx ik\hat{w}_k[1 + O(k^2h^2)], \end{aligned}$$

where $\frac{1}{2}(D_+ + D_-)w(\theta_j)$ is the central difference approximation for $w'(\theta_j)$ and $F(w(\theta)) \equiv \hat{w}_k$, is the DFT of w. Similarly, the discrete second derivative has a DFT of:

$$\begin{aligned} F(D_+D_-w(\theta_j)) &= \frac{\hat{w}_k(e^{ikh} - 2 + e^{-ikh})}{h^2} \equiv -\frac{4\hat{w}_k \sin^2(kh/2)}{h^2} \\ &\approx -k^2\hat{w}_k[1 + O[(kh)^4]]. \end{aligned}$$

Note the importance of the function $\operatorname{sinc}(kh) = \sin(kh)/(kh)$ in the above analysis. At the lowest order, the difference approximations have the same Fourier correspondence $\partial/\partial\theta \to ik$ and $\partial^2/\partial\theta^2 \to -k^2$ as that for the exact derivatives.

The above expansions may be viewed as expansions in $\theta = kh$. The Fourier transform of the unit step function $S(x) = H(x-h) - H(x+h)$ is $2h\operatorname{sinc}(kh)$. The function $\operatorname{sinc}(x)$ also plays an important role in the sampling theorem (e.g. [32]), which effectively states that if the Fourier transform of $g(t)$ has no frequency components with $f > f_c$, then $g(t)$ can be reconstructed, by sampling the values of $g(t)$ at intervals of $T = 1/(2f_c)$ (here $\omega = 2\pi f$ and $\exp(\pm i\omega t)$ are the exponential factors that occur in the Fourier inversion theorem). Further discussion of the sampling theorem and the Nyquist frequency is given in Section 2.3. The above expressions are related to the Taylor series expansions of $z\log(z)$ and $z^2\log^2(z)$ about $z = 1$, where $z = \exp(ikh)$. In particular,

$$z\log(z) = \frac{z^2 - 1}{2} + O\left[(z-1)^3\right] \equiv \frac{e^{ikh}(e^{ikh} - e^{-ikh})}{2},$$

and

$$z^2\log^2(z) = 1 - 2z + z^2 + O\left[(z-1)^3\right] \equiv e^{ikh}\left(e^{ikh} - 2 + e^{-ikh}\right).$$

These results also play a role in compact finite differences and Padé approximations for discrete derivatives, discussed in the next section.

The method of undetermined coefficients is another approach that is often used to derive finite difference approximations. The method assumes an approximation in a particular form, for example,

$$u'(x) \approx \alpha u(x-h) + \beta u(x) + \gamma u(x+h). \tag{2.4}$$

The unknown coefficients α, β and γ are obtained by using Taylor's theorem to expand the right-hand side of (2.4) about x, and match the coefficients at the lowest order. This yields the equations:

$$\alpha + \beta + \gamma = 0, \quad (\gamma - \alpha)h = 1, \quad (\alpha + \gamma)h^2/2 = 0.$$

Solving for α, β and γ gives the central difference approximation

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h},$$

for $u'(x)$. The *truncation* error for the above central difference formula for $u'(x)$ is $O(h^2)$, as can be verified by using the Taylor series expansion for the terms on the right-hand side (the leading term in the truncation error is of the order of $h^2 u_{xxx}/6$). The method of undetermined coefficients in other cases is more general, since other criteria can be used to determine the unknown coefficients. For example, in dispersion preserving approximations discussed in the next section, some coefficients are tuned to have at least second order accuracy (second-order tangency near $\theta = 0$ in Fourier space) while others are chosen to have the best least squares fit to the θ-line on the interval $[0, \pi/2]$ or part thereof. The discussion of the grid frequency range is described in greater detail in the next section on spectral differentiation in the context of aliasing errors, resolution, Nyquist wavenumber and the Shannon sampling theorem.

Next, we consider time dependent problems. Discretizing one independent variable at a time will lead to a full discretization. Often, for analysis purposes, semi-discretization is a very useful intermediate step in investigating the behavior of the system and for making a choice for subsequent discretization. When space is discretized and the time is left continuous, the corresponding semi-discretization is called the method of lines, which results in a system of $N-1$ ordinary differential equations (ODEs), where $N+1$ is number of grid points after the space discretization is done.

Example. The Diffusion Equation

Consider the diffusion equation in one Cartesian space dimension, with a source f, discretized in the form:

$$\frac{du_i(t)}{dt} = \frac{u_{i+1}(t) - 2u_i(t) + u_{i-1}(t)}{h^2} + f_i, \quad i = 1, 2, \ldots, N-1, \quad (2.5)$$

where $u_i(t)$ approximates the solution on a uniformly distributed grid $x_i = ih, i = 0, 2, \ldots, N$ and f_i denotes $f(x_i, t, u_i(t))$.

Introducing the array $\mathbf{u}(t) = (u_1(t), u_2(t), \ldots, u_{N-1}(t))^t$ of unknowns, the above system of ODEs can be written in matrix form as

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{D}\mathbf{u}(t) + \mathbf{f},$$

where the first and last rows incorporate the boundary conditions and the differentiation matrix $\mathbf{D}$ is the tri-diagonal matrix $\mathbf{D} = \frac{1}{h^2}\text{tridiag}(1, -2, 1)$, with appropriate boundary condition modifications to the first and last rows.

The diffusion equation (2.5) can be discretized using either an explicit or an implicit scheme.

Definition 6. *A difference approximation for a derivative is called an* explicit *formula if it can be calculated from known values of the solution for the u_i at the nodes. If the formula involves unknown values of the u_i, which have not been calculated, it is called an implicit formula. For example, u_i^n is an explicit term, but u_i^{n+1} is an implicit term, in a difference scheme, if the solution is only known at times t_0, $t_1, \ldots, t = t_n$, and is not known at time $t = t_{n+1}$.*

Discretizing (2.5) by the explicit, forward Euler method with uniform time step Δt gives

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} + f_i^n, \quad i = 1, 2, \ldots, N-1,$$

with boundary updates for u_0^{n+1} and u_N^{n+1} provided by physical or numerical boundary conditions. More concisely:

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} = \mathbf{D}\mathbf{u}^n + \mathbf{f}^n,$$

is the matrix form of the difference scheme.

Crank-Nicolson Scheme

For the 1D heat equation $u_t - u_{xx} = 0$, the Crank-Nicolson scheme consists in using a semi-implicit difference for the spatial derivative, and a forward Euler difference for the time derivative. The difference equation is:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \frac{1}{2}\left(L_{xx}u_j^{n+1} + L_{xx}u_j^n\right),$$

where

$$L_{xx}u_j^n = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2},$$

defines the second-order central difference operator L_{xx}. Note that $L_{xx}(u_j^n)$ is an explicit difference, but $L_{xx}(u_j^{n+1})$ is an implicit difference. The scheme is unconditionally stable. The Crank-Nicolson scheme can also be applied to the advection diffusion equation $u_t + vu_x - \kappa u_{xx} = 0$.

Example. The Nonlinear Schrödinger Equation

The two-dimensional (2D) nonlinear Schrödinger (NLS) equation:

$$iu_t + u_{xx} + u_{yy} + |u|^2 u = 0$$

on a unit square may be converted to a system of ODEs

$$\frac{idu_{ij}(t)}{dt} + (u_{xx}(t))_{ij} + (u_{yy}(t))_{ij} + |u(t)_{ij}|^2 u(t)_{ij} = 0,$$

or, introducing $P(u(t)_{i,j})$, the system of equations may be written symbolically in the form:

$$\frac{du_{ij}(t)}{dt} = P(u(t)_{i,j}),$$

where $(u_{xx}(t))_{ij}$ and $(u_{yy}(t))_{ij}$ are finite difference or other types of discretization (e.g. compact differences) at location (x_i, y_j) of the non-uniform rectangular mesh.

The resulting system of ODEs may be approximated by a Runge-Kutta scheme with fixed or adaptive time stepping, Adams-Bashforth, backward or forward Euler differences, midpoint or trapezoidal rule, backward schemes, fractional step (split-step methods), stiff solvers, symplectic solvers, etc.

Introducing the array $(u^n)_{ij} \approx u(x_i, y_j, t_n)$, where $u(x_i, y_j, t_n)$ denotes an exact solution at the discretization point (x_i, y_j, t_n) and applying forward, backward, midpoint and trapezoidal rules, third-order TVD Runge-Kutta, and second order Adams-Bashforth with uniform time stepping for simplicity, results in the following algebraic systems of equations:

$$\frac{u_{ij}^{n+1} - u_{ij}^{n}}{\Delta t} = P(u_{ij}^{n}),$$

$$\frac{u_{ij}^{n+1} - u_{ij}^{n}}{\Delta t} = P(u_{ij}^{n+1}),$$

$$\frac{u_{ij}^{n+1} - u_{ij}^{n}}{\Delta t} = P\left(\frac{u_{ij}^{n} + u_{ij}^{n+1}}{2}\right),$$

$$\frac{u_{ij}^{n+1} - u_{ij}^{n}}{\Delta t} = \frac{1}{2}[P(u_{ij}^{n}) + P(u_{ij}^{n+1})],$$

$$u1 = u_{ij}^{n} + \Delta t P(u_{ij}^{n}),$$

$$u2 = \frac{3}{4}u_{ij}^{n} + \frac{1}{4}u1 + \frac{1}{4}\Delta t P(u1),$$

$$u_{ij}^{n+1} = \frac{1}{3}u_{ij}^{n} + \frac{2}{3}u2 + \frac{2}{3}\Delta t P(u2),$$

$$\frac{u_{ij}^{n+1} - u_{ij}^{n}}{\Delta t} = \frac{3}{2}\left[P(u_{ij}^{n}) - \frac{1}{2}P(u_{ij}^{n-1})\right].$$

The above systems are more difficult to arrange into a matrix form as we have some freedom in the way we pack a double array u_{ij} into a single array. We can do it directly row by row, $\mathbf{u}^n = (u_{11}, u_{12}, u_{13}, \ldots, u_{N-1,1}, \ldots u_{N-1,N-2}, u_{N-1,N-1})$, or column by column, or along the diagonal, etc. After the double array is packed into a 1D array, it is necessary to create the corresponding sparse differentiation matrix.

Fractional Step and Alternating-direction-implicit Methods

The fractional step and Alternating-Direction-Implicit (ADI) method splits the solution of the matrix difference equations into two or more steps. For example, in using the ADI method to discretize the 2D NLS equation above, the solution is split into two steps. For the first step at time $t = t_n$, one discretizes the equations for the time step $t = t_n$ to $t = t_n^*$, using an implicit formula for u_{xx}, but an an explicit formula for u_{yy}. For the time step from $t = t_n^*$ to $t = t_{n+1}$, one uses an explicit formula for u_{xx} and an implicit formula for u_{yy}. This leads to a $(N-1) \times (N-1)$ tridiagonal matrix system for the step from $t = t_n$ to $t = t_n^*$, where there are $N - 1$ internal grid points in the x direction, and $M - 1$ internal grid points in the y-direction. This step corresponds to moving across the grid in the x direction, keeping the y index j constant. The tridiagonal system must be solved for each j, leading to the solution for the $(N-1)\times(M-1)$ variables at the intermediate time step $t = t_n^*$. The following step in time from $t = t_n^*$ to $t = t_{n+1}$ yields a similar $(M - 1) \times (M - 1)$ tridiagonal matrix system, which corresponds to moving across the grid in the y-direction, with x fixed. This system must be solved $N - 1$ times to yield the solution at time $t = t_{n+1}$. The original problem involving the above large sparse matrices, thus, reduces to a sequence of problems involving the inversion of tridiagonal matrix systems, which is much simpler, and involves less computation, than solving the original system. The original system involves $(N - 1)^2 \times (M - 1)^2$ matrices, whereas only $(N - 1) \times (N - 1)$ matrices and $(M-1)\times(M-1)$ tridiagonal matrices are used in the fractional step and ADI methods. These methods preserve the accuracy and robustness of the computation, and involve much less computation.

For the Schrödinger equation, the trapezoidal rule (Crank-Nicolson method in the PDE context) is the best with respect to stability, accuracy and long-time behavior, as it preserves the discrete energy $\sum_{ij} |u_{ij}^n|^2$. The method is implicit and requires solving a nonlinear system of algebraic equations for u_{ij}^{n+1}. A simpler version, involving the linearized Crank-Nicolson scheme, is often used and is equivalent to using the one-step version of Newton's method of solving the algebraic system of equations. It consists of replacing $P(u_{ij}^{n+1})$ by its Taylor expansion around u_{ij}^n,

$$P(u_{ij}^{n+1}) = P(u_{ij}^n) + (u_{ij}^{n+1} - u_{ij}^n)P_u(u_{ij}^n),$$

and solving the resulting linear system for u_{ij}^{n+1}. The success of this method is based on the fact that Newton's method often requires only one or two applications in practice to get within the accuracy of the finite difference approximation. If a large number of iterations is required to obtain a convergent solution, this usually indicates that a divergent, physically inconsistent and irrelevant solution has been obtained.

The above example illustrates the choices to be made with respect to robustness, accuracy and efficiency of the method. Systematic analysis of time stepping will be done in the next chapter.

We conclude this section with a method that is based on solution operator expansion and is popular (with appropriate modifications to control oscillations leading to nonlinear instabilities) for advection problems with discontinuities.

Example. Advection Equation

The advection equation:

$$T_t + uT_x = 0, \tag{2.6}$$

where u is a constant, is the canonical equation used to illustrate the Lax-Wendroff scheme (e.g. [66, Section 9.1.3]).

Forward Time, Centered Space Scheme

If one uses the forward time, centered space scheme (FTCS), to difference (2.6) one obtains the difference equation:

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + u\left(\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x}\right) = 0.$$

This equation can be written as

$$T_j^{n+1} = T_j^n - 0.5C(T_{j+1}^n - T_{j-1}^n), \tag{2.7}$$

where

$$C = u\frac{\Delta t}{\Delta x}, \tag{2.8}$$

is the Courant-Friedrichs-Lewy (CFL) number. A Von-Neumann stability analysis shows that (2.7) is unconditionally unstable.

Upwind Scheme

Assuming $u > 0$, and using upwind differencing, (i.e. a backward difference formula for $\partial T/\partial x$, this corresponds to the propagation of information in the direction of the advection, and hence the nomenclature 'upwind'), one obtains the discretized equation:

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + u\frac{T_j^n - T_{j-1}^n}{\Delta x} = 0, \tag{2.9}$$

which can also be written in the form:

$$T_j^{n+1} = (1 - |C|)T_j^n + |C|T_{j-1}^n.$$

Von Neumann stability analysis shows that this scheme is stable if $|C| < 1$ (the condition $|C| < 1$, also takes into account the case $u < 0$). This condition is known as the CFL condition, and states that a particle of fluid should not travel more than one space-step Δx in one time step Δt.

Lax-Wendroff Scheme

In the Lax-Wendroff scheme, one uses the Taylor series expansion of T with respect to time t out to order $(\Delta t)^2$ to obtain the approximation:

$$\begin{aligned} T(x,t+\Delta t) &= e^{\Delta t\partial_t}T(x,t) \equiv T(x,t) + \Delta t T_t + \frac{(\Delta t)^2}{2}T_{tt} + \cdots \\ &= T(x,t) - u\Delta t T_x + \frac{u^2(\Delta t)^2}{2}T_{xx} + O\left[(\Delta t)^3\right]. \end{aligned} \tag{2.10}$$

In (2.10) the operator $e^{\Delta t\partial_t}$ is the solution operator. The advection equation (2.6) implies that $T_t = -uT_x$ and $T_{tt} = u^2 T_{xx}$ etc.. These results have been used to replace time derivatives by space derivatives in (2.10). Equation (2.10) can also be written in the form:

$$T(x,t+\Delta t) - T(x,t) + u\Delta t T_x - \frac{u^2(\Delta t)^2}{2}T_{xx} = 0.$$

If we use central differences for the space discretization, the above equation becomes

$$T_j^{n+1} - T_j^n = -\frac{1}{2}C\left(T_{j+1}^n - T_{j-1}^n\right) + \frac{1}{2}C^2\left(T_{j+1}^n - 2T_j^n + T_{j-1}^n\right), \tag{2.11}$$

which is the Lax-Wendroff scheme. The Lax-Wendroff scheme is consistent with the advection equation (2.6) with truncation error $O[(\Delta t)^2, (\Delta x)^2]$, and is stable if $|C| < 1$.

Multi-dimensional Lax-Wendroff Scheme

Consider

$$\mathbf{u}_t = A\mathbf{u}_x + B\mathbf{u}_y,$$

with constant matrices A and B and expand the solution operator

$$\begin{aligned} \mathbf{u}(t+\Delta t) &= e^{\Delta t(A\partial_x + B\partial_y)}\mathbf{u}(t) \\ &= \left[I + \Delta t(A\partial_x + B\partial_y) + \frac{(\Delta t)^2}{2}(A\partial_x + B\partial_y)^2 + \cdots\right]\mathbf{u}(t), \end{aligned}$$

as a power series in Δt. If we keep $O[(\Delta t)^2]$ terms we obtain a scheme that is second-order accurate in time, and if we replace the spatial partial derivatives by an at least second-order difference approximation, then we obtain a high order and accurate one-level scheme. Thus noting that

$$(A\partial_x + B\partial_y)^2 = A^2(\partial_{xx})_{ij} + (AB + BA)(\partial_{xy})_{ij} + B^2(\partial_{yy})_{ij},$$

and using central differences for the space derivatives at point (x_i, y_j), we get

$$\begin{aligned}\mathbf{u}_{ij}^{n+1} = \mathbf{u}_{ij}^n + A(\mathbf{u}_x)_{ij}^n + B(\mathbf{u}_y)_{ij}^n + \frac{1}{2}A^2(\mathbf{u}_{xx})_{ij}^n \\ + (AB + BA)(\mathbf{u}_{xy})_{ij}^n + \frac{1}{2}B^2(\mathbf{u}_{yy})_{ij}^n.\end{aligned}$$

For example, the mixed derivative $(\mathbf{u}_{xy})_{ij}^n$ on a uniform grid is:

$$(\mathbf{u}_{xy}^n)_{ij} = \frac{1}{2\Delta y}\left(\frac{\mathbf{u}_{i+1,j+1}^n - \mathbf{u}_{i-1,j+1}^n}{2\Delta x} - \frac{\mathbf{u}_{i+1,j-1}^n - \mathbf{u}_{i-1,j-1}^n}{2\Delta x}\right).$$

Similar central difference formulae are also used for the other spatial derivatives. This leads to the multi-dimensional Lax-Wendroff scheme.

Solution Operator and Difference Schemes

It is instructive to view the different difference schemes used for differencing the advection equation (2.6) in terms of the solution operator $e^{\Delta t\partial_t}$. It follows from (2.10), that

$$T(x, t + \Delta t) = e^{\Delta t\partial_t}T(x, t) \equiv e^z T(x, t), \tag{2.12}$$

where

$$z = -u\Delta t\partial_x.$$

If we regard z as a small order quantity, we can expand the solution operator e^z in a variety of ways. For example,

$$e^z \approx 1 + z + O(z^2), \tag{2.13}$$

$$e^z \approx 1 + z + \frac{z^2}{2} + O(z^3), \tag{2.14}$$

$$e^z \approx \frac{1}{1 - z} + O(z^2), \tag{2.15}$$

$$e^z \approx \frac{1+z/2}{1-z/2} + O(z^2). \tag{2.16}$$

The expansion (2.13)–(2.14) are Taylor expansions of e^z; (2.15) is the $(0,1)$-Padé approximation of e^z; and (2.16) is the $(1,1)$-Padé approximation of e^z.

In general the (m,k)-Padé approximation of a function $f(x)$ is given by the ratio of two polynomials in the form:

$$r_{mk} = \frac{P_m(x)}{Q_k(x)}, \tag{2.17}$$

where $P_m(x)$ and $Q_k(x)$ are polynomials of degree at most of m and k respectively. One form of Padé approximation matches the derivatives of the Taylor series about $x = 0$ for as many derivatives as possible, assuming that the Taylor series about $x = 0$ is well defined (one can change the independent variable so that the new independent variable now contains $x = 0$ in the new interval). Padé approximations that minimize the minimum, maximum error over the interval are called the Chebyshev or *minmax* approximations, and involve the use of Chebyshev polynomials.

Using the Taylor series expansion (2.13) as the approximate solution operator in (2.12), yields the equation:

$$T(x,t+\Delta t) = (1 - u\Delta t\partial_x)T(x,t). \tag{2.18}$$

Using upwind spatial differencing for T_x in (2.18) we get the upwind scheme (2.9). Similarly, using the approximate solution operator (2.14), and central differences for T_x and T_{xx} we get the Lax-Wendroff scheme (2.11).

The $(0,1)$-Padé approximation (2.15), for the solution operator yields the equation:

$$T(x,t+\Delta t) = \frac{1}{1 - u\Delta t\partial_x}T(x,t). \tag{2.19}$$

Multiplying (2.19) by $(1 - u\Delta t\partial_x)$, and using upwind differencing for T_x, we get the implicit scheme

$$T_j^{n+1} - C\left(T_j^{n+1} - T_{j-1}^{n+1}\right) = T_j^n,$$

where C is the CFL number (2.8). Similarly, the $(1,1)$-Padé approximation (2.16) yields the equation:

$$T(x,t+\Delta t) = \frac{1 - \frac{1}{2}u\Delta t\partial_x}{1 + \frac{1}{2}u\Delta t\partial_x}T(x,t).$$

Multiplying this equation by $(1 + \frac{1}{2}u\Delta t\partial_x)$, and using upwind spatial differencing, we get the Crank-Nicolson scheme:

$$T_j^{n+1} = T_j^n - \frac{1}{2}C\left(T_j^{n+1} - T_{j-1}^{n+1} + T_j^n - T_{j-1}^n\right). \tag{2.20}$$

Note that the denominator in the Padé approximation schemes (2.15) and (2.16) gives rise to implicit difference derivatives.

Padé approximations for the solution operator e^z, are an essential ingredient of the compact finite difference schemes, which are discussed in the next section.

2.2 Compact Finite Differences and Dispersion Preserving Schemes

Consider the approximation of the derivative in Fourier space, as in the previous section, as an approximation of ikh or $\ln z$ near $k = 0$ or $z = 1$, where $z = e^{ikh}$. In this section, instead of a Taylor expansion we will use local rational approximations (Padé approximations) to develop compact finite difference formulae for the derivatives of u_i' at the grid points.

The essential idea of compact finite differences can be illustrated by a consideration of the central difference approximation:

$$D_0(h)u = \frac{u(x+h) - u(x-h)}{2h} \equiv \frac{\sinh(h\partial_x)}{h}u, \tag{2.21}$$

where $\sinh x$ is the hyperbolic sine function. Equation (2.21) can be inverted to give

$$h\partial_x = \operatorname{arcsinh}(hD_0) \equiv \operatorname{arcsinh}(Z), \tag{2.22}$$

for the exact derivative operator, where

$$Z = hD_0. \tag{2.23}$$

One can also think of the transformation (2.22) as a transformation in Fourier space. In this latter case, we have

$$F(hu_x) = ikh\hat{u}_k, \quad F[u(x+h)] = e^{ikh}\hat{u}_k, \quad F[u(x-h)] = e^{-ikh}\hat{u}_k, \tag{2.24}$$

where $F[u(x)] = \hat{u}_k$ denotes the Fourier transform of $u(x)$. Setting $z = e^{ikh}$, one can also think of compact differences as arising from Padé approximations for $ikh \equiv \ln(z)$ in terms of z.

A natural question that arises from (2.22) can be posed as follows: is it possible to obtain a more accurate difference approximation for u_x than the central difference approximation (2.21), by using a Padé approximation for $\operatorname{arcsinh}(Z)$? Below, we show that by using a (1,2)-Padé approximation for $\operatorname{arcsinh}(Z)$, it is possible to obtain a difference approximation that has an $O(h^4)$ truncation error. This improved difference approximation is related to three-point compact difference schemes for u_x. More general Padé approximations, with higher order truncation errors can also be constructed.

By setting

$$\operatorname{arcsinh}(Z) \approx \frac{a_0 + a_1 Z}{1 + b_1 Z + b_2 Z^2}, \tag{2.25}$$

using the Taylor series expansion for $\operatorname{arcsinh}(Z)$,

$$\operatorname{arcsinh}(Z) = Z - \frac{1}{2}\frac{Z^3}{3} + \frac{1.3}{2.4}\frac{Z^5}{5} + \cdots,$$

[6, p. 88, formula 4.6.31], clearing fractions in (2.25), and equating powers of Z, we get simple algebraic equations for a_0, a_1, b_1 and b_2. Solving these latter equations results in the Padé approximation

$$\operatorname{arcsinh}(Z) = \frac{Z}{1 + Z^2/6} + O(Z^5). \tag{2.26}$$

Thus, (2.22) and (2.26) yield the equation:

$$h\partial_x \approx \frac{hD_0}{1 + h^2 D_0^2/6}, \tag{2.27}$$

as the $(1,2)$-Padé approximation for $h\partial_x$. If we work to an accuracy of $O(h^4)$, we may replace the operator D_0^2 by D_+D_-. In fact,

$$D_0^2 u - D_+D_- u = \frac{h^2}{12} u_{4x} + O(h^4).$$

Replacing D_0^2 by D_+D_- in (2.27) yields the equation:

$$\left(1+\frac{h^2D_+D_-}{6}\right)u_x = D_0u.$$

Using the definitions of D_0, D_+ and D_- in this latter equation we get the implicit system of equations:

$$u'_{i-1}+4u'_i+u'_{i+1}=\frac{3}{h}(u_{i+1}-u_{i-1}) \tag{2.28}$$

for the derivatives of u at the mesh nodes for $0\leq i\leq N$. Equation (2.28) defines a three-point compact difference scheme, which may be solved for the u'_i at the inner grid points with appropriate corrections at the endpoints to accommodate the boundary conditions. The compact difference formula for derivatives requires one to solve a tridiagonal system in order to approximate the unknown derivatives at all grid points at once with fourth-order accuracy. Compared with the finite difference approximation for the derivative of the same order, the compact finite difference has a three-point stencil instead of a five-point stencil.

Note that (2.28) can also be obtained, by using the Fourier space map:

$$hu_x\approx\frac{hD_0}{1+h^2D_+D_-/6}u\to ikh\hat{u}_k\equiv\ln z\hat{u}_k\approx\frac{3(z^2-1)\hat{u}_k}{1+4z+z^2}, \tag{2.29}$$

which is equivalent to the $(2,2)$-Padé expansion of $\ln z$ about $z=1$. Thus the compact difference scheme (2.28), can be thought of as a Padé expansion of $\ln z$. Alternatively, using the grid wavenumber $\theta=kh$, the Padé expansion (2.29) can be written in the form:

$$\theta=\tilde{\theta}(\theta)+O(\theta^4)\quad\text{where }\tilde{\theta}(\theta)=\frac{3\sin\theta}{2+\cos\theta}. \tag{2.30}$$

Thus one can visualize the approximation by plotting $\tilde{\theta}$ versus θ. The exact derivative, in this case, is represented by a straight line $\tilde{\theta}=\theta$, in the $(\theta,\tilde{\theta})$-plane. The deviation of the curve $\tilde{\theta}=\tilde{\theta}(\theta)$ from $\tilde{\theta}=\theta$ can be used to visualize how good the approximate compact difference is as the grid wavenumber increases. If the stencil is wider than a three-point stencil then the usual practice is to switch gradually to one-sided formulae near the boundary. The introduction of gradients in the system will, in general, produce wave reflection, transmission and interaction of different wave modes, analogous to similar phenomena for wave propagation on a non-uniform string. Similar three-point and five-point compact difference schemes are described by Lelé [108].

An approximate local solution formula for the u_i' that is accurate to $O(h^4)$ can be obtained from (2.26) by using the fact that $D_0^2 \approx D_+D_-$, and using the Binomial expansion to obtain the approximate formula

$$u_x \approx D_0 \left(1 - \frac{h^2 D_+ D_-}{6}\right) u,$$

for u_x. Using this latter formula, and the definitions of D_0, D_+ and D_-, yields the explicit formula

$$u_i' = \frac{4}{3}\frac{u_{i+1} - u_{i-1}}{2h} - \frac{u_{i+2} - u_{i-2}}{12h} \equiv \frac{4D_0(h) - D_0(2h)}{3} u_i, \qquad (2.31)$$

for u_i', where $D_0(h)$ and $D_0(2h)$ are the central difference operators with step length h and $2h$, respectively. The truncation error in (2.31) is $O(h^4)$.

Padé approximations allow one to achieve similar or slightly better accuracy than the same order finite difference schemes, but the width of the stencil (nearby points) is smaller, at the expense of being implicit. This may be a desirable property for an easier treatment of the boundary conditions, both physical and numerical.

Spectral methods are the ultimate finite difference approximations when applied to smooth functions, meaning that all the scales are resolved, on a periodic domain. In fact, as the number of nodes in the stencil of the uniform centered finite difference tends to infinity, the resulting derivative approximation becomes spectral differentiation. Again, periodicity, smoothness and uniform grids are the usual price to pay. The first and last limitations mentioned above have been overcome by using spectral element or h-p finite element methods where orthogonal polynomials other than trigonometric functions are used as a basis. This will be discussed in a later section of this chapter.

One can also derive compact difference formulae using the method of undetermined coefficients [108]. In this approach, one takes averages of derivatives at about the i^{th} node, in order to increase the accuracy of approximation at point x_i. For example,

$$\begin{aligned}
&\beta u_{i-2}' + \alpha u_{i-1}' + u_i' + \alpha u_{i+1}' + \beta u_{i+2}' \\
&\quad = a\frac{u_{i+1} - u_{i-1}}{2h} + b\frac{u_{i+2} - u_{i-2}}{4h} + c\frac{u_{i+3} - u_{i-3}}{6h}
\end{aligned}$$

or in Fourier space

$$\theta\left(2\beta\cos(2\theta)+2\alpha\cos\theta+1\right)=a\sin\theta+\frac{b}{2}\sin(2\theta)+\frac{c}{3}\sin(3\theta). \qquad (2.32)$$

Matching the Taylor expansion about $\theta=0$ gives

$$\begin{aligned}
&a+b+c=1+2\alpha+2\beta,\\
&a+2^2b+3^2c=2\cdot 3(\alpha+2^2\beta),\\
&a+2^4b+3^4c=2\cdot 5(\alpha+2^4\beta),\\
&\dots\\
&a+2^{2n}b+3^{2n}c=2\cdot(2n+1)(\alpha+2^{2n}\beta),
\end{aligned}$$

to obtain a $2n+2$-order of approximation. Note that the three-point compact difference scheme (2.28) corresponds to the case $n=1$, $\alpha=\frac{1}{4}$, $a=\frac{3}{2}$, $b=c=\beta=0$. The function $\tilde{\theta}(\theta)$ analogous to (2.30) in this case is:

$$\tilde{\theta}(\theta)=\frac{a\sin\theta+\frac{1}{2}b\sin(2\theta)+\frac{1}{3}c\sin(3\theta)}{1+2\alpha\cos\theta+2\beta\cos(2\theta)},$$

where $\theta\approx\tilde{\theta}(\theta)$ is the approximate grid wavenumber relation for the evaluation of first derivatives.

The compact difference scheme for second-order derivatives has the form

$$\begin{aligned}
\beta u''_{i-2}+\alpha u''_{i-1}+u''_i+\alpha u''_{i+1}+\beta u''_{i+2}&=a\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}\\
+b\frac{u_{i+2}-2u_i+u_{i-2}}{4h^2}&+c\frac{u_{i+3}-2u_i+u_{i-3}}{9h^2}. \qquad (2.33)
\end{aligned}$$

The Fourier space version of (2.33) in this case is:

$$\begin{aligned}
&\theta^2\left(2\beta\cos(2\theta)+2\alpha\cos\theta+1\right)\\
&\quad=2a(1-\cos\theta)+\frac{b}{2}(1-\cos(2\theta))+\frac{2c}{9}(1-\cos(3\theta)). \qquad (2.34)
\end{aligned}$$

Equating powers of θ in the Taylor series expansion about $\theta=0$ in (2.34) yields the equations:

$$\begin{aligned}
&a+b+c=1+2\alpha+2\beta,\\
&a+2^{2n}b+3^{2n}c=2(n+1)(2n+1)(\alpha+2^{2n}\beta), \quad n=1,2,\dots,
\end{aligned}$$

to get a $2n{+}2$-order approximation of the second derivative. These formulae should be compared to spline approximations that give only second order accuracy for the second-order derivative and fourth-order accuracy for the first-order derivative (Section 2.4).

Spectral differentiation results if the α and β are set to zero, while the right-hand side is taken as a full Fourier series. See next section.

Example. Dispersion Preserving Scheme [183]

Set α and β equal to zero in the above approximation for the first derivative compact differences. The Fourier space equation (2.32) becomes:

$$\theta \approx \arcsin\theta + \frac{b}{2}\sin(2\theta) + \frac{c}{3}\cos(3\theta).$$

Using this result one determines two of the unknown coefficients a, b and c, by requiring fourth-order accuracy, while the third coefficient is obtained by minimizing

$$\left\| \theta - \left(a\sin\theta + \frac{b}{2}\sin(2\theta) + \frac{c}{3}\cos(3\theta) \right) \right\|_2$$

for $\theta \in [-\pi/2, \pi/2]$; or by symmetry for $\theta \in [0, \pi/2]$, we minimize

$$\int_0^{\pi/2} d\theta \, \left[\theta - \left(a\sin\theta + \frac{b}{2}\sin(2\theta) + \frac{c}{3}\cos(3\theta) \right) \right]^2 .$$

These constraints make the above expression a quadratic function with respect to one unknown variable and gives

$$\begin{aligned} a &= 1.59853286, \\ b/2 &= -0.37882628, \\ c/3 &= 0.0530399. \end{aligned}$$

Similar expressions may be obtained for other intervals, say for $[0, \pi/3]$ or $[0, \pi/4]$. In [67] different schemes were applied on different intervals of the grid-wavenumber $\theta = kh$.

2.3 Spectral Differentiation

In this section we consider spectral differentiation by using the DFT (e.g. [32,39]; as well as [67,189]). The DFT was used in Section 2.1 and

Section 2.2 (Equation (2.2) et seq.), to discuss the accuracy of finite difference and compact finite difference approximations for derivatives.

One approach to the DFT [32] involves the sampling of a function $h(t)$ in the time domain at regularly spaced intervals $\Delta t = T$, with windowing of the data over a time interval $T_0 = NT$, and sampling of the data in the frequency domain over a frequency period of $1/T_0$. The convolution theorem for Fourier transforms, coupled with a Fourier series representation of the Fourier extended data in the time domain, then leads to the DFT inversion formulae. We use a simpler, algebraic approach to the DFT that is sufficient for our purposes. Brigham [32] provides a physical, intuitive approach to the subject. A more rigorous, but readable account is given by Weaver [189].

Consider a periodic function $u(x)$, where $x \in [0, 2\pi]$, that can be represented by a Fourier series as:

$$u(x) = \sum_{k=-\infty}^{\infty} \hat{c}_k z^k \quad \text{where } z = e^{ix}.$$

We interpolate $u(x)$ by a trigonometric polynomial (resulting from truncation of the above series)

$$p(x) = \sum_{k=0}^{N-1} c_k z^k, \quad z = e^{ix}, \tag{2.35}$$

at uniformly spaced points $x_j = hj = \frac{2\pi}{N} j, \quad j = 0, 1, 2, \ldots, N-1$, where h denotes the grid spacing.

Since we choose the function to match the approximation at the grid points $x_j = jh$ (i.e., $u_j \equiv u(x_j) = p(x_j)$), it follows that

$$u_j = \sum_{k=0}^{N-1} c_k e^{ikx_j}, \quad j = 0, 1, 2, \ldots, N-1. \tag{2.36}$$

These equations can also be written in the matrix form $\mathbf{u} = \mathbf{Oc}$, where $O_{jk} = \exp(ikx_j)$. The matrix $\mathbf{O}$ is a symmetric orthogonal matrix, since $kx_j = jx_k$, and the rows and columns are orthogonal in the usual definition of the dot product for complex vectors. Thus

$$(\mathbf{OO}^*)_{km} = \sum_{j=0}^{N-1} e^{ijx_k} e^{-ijx_m} \equiv \sum_{j=0}^{N-1} e^{ikx_j} e^{-imx_j} = N\delta_{km}, \tag{2.37}$$

where δ_{km} is the Kronecker delta symbol ($\delta_{km} = 1$ if $k = m$ and $\delta_{km} = 0$ otherwise). The proof of (2.37) depends on noting that each term in the series (2.37) is unity for $k = m$, whereas for $k \neq m$, the sum is a geometric series with common ratio $w = \exp[2\pi i(k - m)/N]$.

From (2.37), $\mathbf{O}^{-1} = \mathbf{O}^*/N$, and $\mathbf{c} = (1/N)\mathbf{O}^*\mathbf{u}$ is the required solution for $\mathbf{c}$. Thus we obtain the inverse transform formulae:

$$c_k = \frac{1}{N} \sum_{j=0}^{N-1} u_j e^{-ikx_j}, \quad k = 0, 1, 2, \ldots, N - 1, \tag{2.38}$$

for the $\{c_k\}$. Equations (2.36) and (2.38) define the inversion formulae for the DFT. The Equations (2.36) and (2.38) are analogous to the corresponding formulae for Fourier series and the Fourier transform.

Fourier Series, Discrete Fourier Transform and Dirichlet Kernel

From the theory of Fourier series, a 2π-periodic, Lebesgue square integrable function $u(x)$ can be expanded in a Fourier series of the form:

$$u(x) = \lim_{N \to \infty} S_N[u(x)], \tag{2.39}$$

where

$$S_N[u(x)] = \sum_{k=-m}^{m} \hat{c}_k e^{ikx}, \tag{2.40}$$

and $N = 2m + 1$. The Fourier coefficients $\{\hat{c}_k\}$ in (2.40) are given by the formulae:

$$\hat{c}_k = \frac{1}{2\pi} \int_0^{2\pi} u(x) e^{-ikx} dx, \quad -m \leq k \leq m. \tag{2.41}$$

The formulae (2.41) for the Fourier coefficients can be formally obtained by multiplication of (2.39) by $\exp(-ikx)$, integration from $x = 0$ to $x = 2\pi$, interchange of the order of summation and integration, and use of the orthogonality of the complex exponential functions $\{\exp(ikx) : k \text{ integer}\}$, with respect to the inner product

$$(u, v) = \int_0^{2\pi} u(x) v^*(x)\, dx. \tag{2.42}$$

For square integrable functions $u(x) \in L_2[0, 2\pi]$, the Fourier series converges in the mean square (i.e. with respect to the L_2 norm, for details refer to e.g. [189]). It is also of interest to investigate the pointwise convergence of the Fourier series and, in particular, the manner in which the Fourier series converges at points of discontinuity of the function $u(x)$ (i.e., Gibbs phenomena).

It is of interest to note that a version of the DFT inversion formulae (2.36) and (2.38) can be obtained formally by using the trapezoidal rule to approximate the integral for the Fourier coefficients $\hat{c}_k$ in (2.41) at the nodes $x_j = 2\pi j/N$, where $|j| \leq m$ and $N = 2m+1$ is the number of nodes. The only difference between (2.38) and (2.41) is that the Fourier index k runs from $k = -m$ to $k = m$ in (2.41).

Substituting (2.41) in (2.40) and interchanging the order of integration and summation, we get the equation:

$$\begin{aligned} S_N[u(x)] &= \sum_{k=-m}^{m} \frac{1}{2\pi} \left(\int_0^{2\pi} u(y) e^{-iky}\, dy \right) e^{ikx} \\ &= \frac{1}{2\pi} \int_0^{2\pi} D_m(x-y) u(y)\, dy, \end{aligned} \tag{2.43}$$

where

$$D_m(x-y) = \sum_{k=-m}^{m} e^{ik(x-y)} = \frac{\sin\left[\left(m + \frac{1}{2}\right)(x-y)\right]}{\sin\left[\frac{1}{2}(x-y)\right]}, \tag{2.44}$$

is the Dirichlet kernel. The sum in (2.44) may be summed by noting that for $x \neq y$ the series is a geometric series with common ratio $w = \exp[i(x-y)]$ and first term w^{-m}. For $x = y$ each term in the series is equal to unity, in which case we get $D_m(0) = N = 2m+1$. One can formally get the same result for $D_m(0)$ by letting $x \to 0$, and using Hôpital's rule. The formulae (2.43)–(2.44) for the truncated Fourier series sum $S_N[u(x)]$ and the Dirichlet kernel (2.44) are important in determining the pointwise convergence of the Fourier series (e.g. [39,189]). Similar formulae also apply in questions of convergence of the trigonometric polynomial expansion (2.35). We return to these issues later on in this section.

The Sampling Theorem

In the Fourier analysis of data, and in using the Fourier transform or the DFT, it is important to sample the data at sufficiently small intervals in order to resolve the fine scale features in the data. This problem is also important in spectral differentiation based on any complete, orthogonal system of eigenfunctions. The main idea in the sampling theorem, is that to resolve the finest time scale T_c in a time series, one must sample the data on scales $T \leq \frac{1}{2}T_c$, in order to prevent aliasing of the data in Fourier space. Thus it is necessary to sample the data at a frequency $f = 1/T \geq 2f_c$, where $f_c = 1/T_c$. The frequency $2f_c$ is known as the Nyquist frequency. The sampling theorem also obviously applies for data which are functions of x.

Following the analysis of Brigham [32], we introduce the sampling distribution (or comb distribution):

$$\Delta(x) = \sum_{j=-\infty}^{\infty} \delta(x - jL), \tag{2.45}$$

where L is the sampling interval, and $\delta(x)$ is the Dirac delta distribution. The Fourier transform

$$\hat{u}(k) = \int_{-\infty}^{\infty} u(x)e^{-ikx}\,dx, \tag{2.46}$$

and the Fourier inversion formula:

$$u(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{u}(k)e^{ikx}\,dk, \tag{2.47}$$

also play an important role in the analysis. The Fourier transform (2.46) and its inverse (2.47), are well defined for $u(x) \in L_2[-\infty, \infty]$. Fourier theory can be generalized to take into account more pathological functions, such as generalized functions (e.g. the Dirac delta distribution, see [72,115,189]).

The sampling function $\Delta(x)$ can be used to generate the periodic extension $u_E(x)$ of a function $u(x)$ defined for $x \in [0, L)$ by convolving $u(x)$ with $\Delta(x)$ to obtain:

$$u_E(x) = u(x) * \Delta(x) = \int_{-\infty}^{\infty} u(x - x')\Delta(x')\,dx'. \tag{2.48}$$

Here the operation $*$ in (2.48) denotes convolution. From (2.45) and (2.48) it follows that $u_E(x) = u(x - jL)$ for $jL \leq x < (j + 1)L$, where j is an integer ($j \in Z$).

The Fourier transform of $\Delta(x)$ is given by:

$$\hat{\Delta}(k)=\lim_{m\to\infty}\sum_{j=-m}^{m} e^{-ijkL}=\lim_{m\to\infty} D_m(kL), \tag{2.49}$$

where $D_m(z)$ is the Dirichlet kernel (2.44) with argument $z=kL$. An alternative expression for $\hat{\Delta}(k)$ can be obtained by using the Fourier series representation for $\Delta(x)$, namely:

$$\Delta(x)=\frac{1}{L}\sum_{n=-\infty}^{\infty}\exp\left(\frac{2\pi inx}{L}\right)\equiv\frac{1}{L}\lim_{m\to\infty} D_m\left(\frac{2\pi x}{L}\right). \tag{2.50}$$

Taking the Fourier transform of (2.50) yields the alternative formula:

$$\hat{\Delta}(k)=\lim_{m\to\infty}\frac{2\pi}{L}\sum_{j=-m}^{m}\delta\left(k-\frac{2\pi j}{L}\right), \tag{2.51}$$

for $\hat{\Delta}(k)$.

It is interesting to note, that by setting $z=2\pi x/L$ in (2.50) and using the definition of $\Delta(x)$ in (2.45),

$$\lim_{m\to\infty} D_m(z)=2\pi\sum_{j=-\infty}^{\infty}\delta(z-2\pi j). \tag{2.52}$$

Thus as $m\to\infty$ the Dirichlet kernel $D_m(z)$ behaves like a periodic array of Dirac delta distributions located at $z=2\pi j$, where j is an integer. The sampling theorem is given below.

Definition 7. *If $u(x)$ and its Fourier transform $\hat{u}(k)$ satisfy Fourier's theorem, and $\hat{u}(k)=0$ for $|k|>k_c$, then the Fourier transform $\hat{u}_s(k)$ of the sampled distribution:*

$$u_s(x)=u(x)\Delta(x)\equiv\sum_{j=-\infty}^{\infty} u(jL)\delta(x-jL), \tag{2.53}$$

exhibits aliasing if the sampling frequency $\nu=1/L$, with $k=2\pi/L$, is such that $\nu<2\nu_c$, where $2\nu_c$ is the Nyquist frequency, and $\nu_c=1/L_c$. If $\nu\geq 2\nu_c$ there is no aliasing of $\hat{u}_s(k)$. If $\nu=2\nu_c$ then $u(x)$ can be reconstructed *uniquely from the sampled data* (2.53)*, and is given by the formula:*

$$u(x) = \sum_{j=-\infty}^{\infty} u(jL)\mathrm{sinc}[k_c(x - jL)], \tag{2.54}$$

where $\mathrm{sinc}(x) = \sin(x)/x$.

To establish the above results, we use the convolution theorem for Fourier transforms, namely

$$\mathcal{F}[f(x) * g(x)] = \hat{f}(k)\hat{g}(k) \quad \text{and} \quad \mathcal{F}[f(x)g(x)] = \frac{1}{2\pi}(\hat{f}(k) * \hat{g}(k)),$$

where $\mathcal{F}[f(x)] \equiv \hat{f}(k)$, and the convolution integral is defined in (2.48).

Using the convolution theorem for Fourier transforms, on (2.53), yields $\hat{u}_s(k) = \hat{u}(k) * \hat{\Delta}(k)/(2\pi)$, for the Fourier transform of $u_s(x)$. Since $\hat{u}(k) = 0$ for $|k| > k_c$, and using the result (2.51) for $\hat{\Delta}(k)$ we find that

$$\hat{u}_s(k) = \frac{1}{2\pi} \int_{-k_c}^{k_c} \hat{u}(k') \frac{2\pi}{L} \sum_{j=-\infty}^{\infty} \delta\left(k - k' - \frac{2\pi j}{L}\right) dk'. \tag{2.55}$$

A sketch of the resonance curves $k' = k - 2\pi j/L$ in the (k, k')-plane, coupled with the fact that $|k'| < k_c$, in (2.55), shows that if $L < \frac{1}{2}L_c$ (i.e., $\nu \geq 2\nu_c$) the transform $\hat{u}_s(k)$ does not suffer from aliasing, and has the form:

$$\hat{u}_s(k) = \begin{cases} \dfrac{1}{L}\hat{u}\left(k - \dfrac{2\pi j}{L}\right) & \text{if } \dfrac{(2j-1)\pi}{L} - k_c < k < \dfrac{(2j+1)\pi}{L} + k_c, \\ 0 & \text{otherwise.} \end{cases} \tag{2.56}$$

Aliasing occurs if $\nu < 2\nu_c$.

To derive formula (2.54) for the re-constructed function $u(x)$ for the case $\nu = 2\nu_c$, introduce the window function:

$$\hat{q}(k) = L \quad \text{if } |k| < k_c. \tag{2.57}$$

Since the support for $\hat{u}(k)$ is for $|k| < k_c$, it follows that

$$\hat{u}(k) = \hat{q}(k)\hat{u}_s(k) \quad \text{and} \quad u(x) = q(x) * u_s(x). \tag{2.58}$$

The inverse Fourier transform of $\hat{q}(k)$ is:

$$q(x) = \operatorname{sinc}(k_c x). \tag{2.59}$$

The formula (2.54) now follows from evaluating the convolution $u(x) = q(x) * u_s(x)$. This completes the proof. A similar form of the sampling theorem can also be written down if the data is sampled in k-space (see [32]).

Spectral Differentiation, Discrete Fourier Transform and Fast Fourier Transform

To approximate the derivatives of the function $u(x)$ one uses appropriate derivatives of $p(x)$, for example

$$u'(x) \approx p'(x) = \sum_{k=0}^{N-1} ikc_k e^{ikx}. \tag{2.60}$$

This expression can be evaluated at any desired position x, which does not have to be an interpolation node. Formula (2.60) is analogous to the derivative of a Fourier integral, i.e. differentiation with respect to x is equivalent to multiplication by ik in Fourier space. Thus once the Fourier coefficients $\{c_k\}$ of the function are known, the Fourier coefficients of the derivative are given by ikc_k. A note of caution: in a particular software package, one has to check carefully the definition of the Fourier transform adopted, and how the Fourier coefficients are stored in an array, to find which indices in the array have to be multiplied by the appropriate factors.

The computation of the Fourier coefficients may be accomplished by a Fast Fourier Transform (FFT) algorithm that involves order $N \log N$ operations instead of direct matrix multiplication that uses order N^2 operations. The algorithm can be thought of as a decomposition of the matrix $\mathbf{O}$ into the product of sparse matrices, for which the evaluation of $\mathbf{O}$ by multiplying the sparse matrices is computationally more efficient than direct evaluation of $\mathbf{O}$ (e.g. [32]). The algorithm exploits the periodicity and additive property of the complex exponential function. This allows one to cut the computational effort involving an even number of nodal points, N, by half in each subsequent split of the data into smaller samples. Therefore the best results are achieved for $N = 2^m$, where m is an integer. In general, the performance of the algorithm can be significantly modified by changing the length of the array. In practice, it usually pays to use FFT on arrays whose length is of the order of a few hundred. Multi-dimensional transforms are implemented as a sequence of 1D transforms.

Several useful properties follow from the definition of the DFT, which are analogous to the known properties of the usual Fourier transform (e.g. [21,32,39]).

The DFT can be interpreted as changing the basis in C^N from the standard orthogonal basis to another orthogonal basis formed by the discrete Fourier harmonics $\{\exp(ikx) : 0 \leq k \leq N-1\}$. The analog of the Parseval identity expresses the invariance of the dot product (modulo a multiplicative factor of N) in both representations, which can be written in the form:

$$\sum_{k=0}^{N-1} u_k v_k^* = N \sum_{k=0}^{N-1} c_k d_k^*,$$

where c_k and d_k represent the DFT of the arrays u_k and v_k, respectively. In particular, it implies the conservation of the 2-norm of the vector (up to a normalization constant) under the orthogonal transformation (rotation). Note, that when the data is real, sampling at N points has, in fact, double the amount of information than one might think at first glance, since $c_j^* = c_{-j}$ for real data. Also, since $\exp(ik2\pi/N)$ is an N-periodic function with respect to k, $c_k = c_{k+mN}, m = 0, \pm 1, \pm 2, \ldots$, which allows one to shift indices by any multiple of N (for example $c_{-1} = c_{N-1}$). Rewriting the DFT inversion formulae (2.36) and (2.38) in which the summation index takes the integer values from $-N/2$ to $N/2-1$ (we assume N is an even number), instead of from 0 to $N-1$, shows that the highest physical wavenumber k represented is not $N-1$ but $k = N/2$, which is called the Nyquist wavenumber. This alternative way of writing the DFT transform pair is also discussed in Equations (2.2) et seq., where it was used to discuss the accuracy of finite difference approximations for numerical derivatives.

The Nyquist wavenumber corresponds to the wavelength of two grid spacings, $\lambda = \frac{2\pi}{N/2} = 2h$. The corresponding Nyquist grid wavenumber is defined as $\theta = kh = \pi$. The grid wavenumbers smaller than $\pi/2$ are called smooth or low wavenumbers, while grid wavenumbers $\pi/2 < |\theta| < \pi$ are called oscillatory or high wavenumbers. An important observation with theoretical and practical implications is that the grid wavenumber changes from high to low as Δx tends to zero. In other words, resolution of a particular physical mode with a sufficient number of points converts it from an oscillatory to a smooth numerical wavenumber.

The wavenumbers higher than the Nyquist wavenumber are indistinguishable on the grid from the lower frequencies as described above, and lead to the aliasing property of the DFT described below. In other words, the wavenumbers higher than the Nyquist wavenumber are not represented properly in the DFT. The wavenumbers that need to be resolved have to be in the smooth part of the spectrum. Only 2-4 points per wavelength yields a very poor representation. In practice, the wavenumbers that need to be resolved in a continuous problem require a minimum of 6-10 points per wavelength, and 20-60 points per wavelength in second-order finite difference schemes are typical.

For practical applications it is important to relate the physical wavenumber K and the node number k computed using the DFT where it was assumed that the sampling was taken on the interval $[0, 2\pi]$, If the sampling interval $L = N\Delta x$, then DFT node number k has to be multiplied by factor $2\pi/L$ to obtain the physical wavenumber K. Thus, increasing the sampling interval L will increase the resolution in Fourier space by making $\Delta k = \frac{2\pi}{L}$ smaller.

The aliasing formula is a particular case of a Poisson's summation formula [136, Section 4.8], which says that a coefficient of the DFT is equal to the sum of all the continuous Fourier coefficients that have the same grid wavenumbers modulo N (aliases of a particular grid wavenumber):

$$\begin{aligned} c_k &= \frac{1}{N}\sum_{j=0}^{N-1} u_j e^{-ikx_j} = \frac{1}{N}\sum_{j=0}^{N-1}\left(\sum_{m=-\infty}^{\infty} \hat{c}_m e^{imx_j}\right) e^{-ikx_j} \\ &= \sum_{m=-\infty}^{\infty} \hat{c}_m \left(\frac{1}{N}\sum_{j=0}^{N-1} e^{imx_j} e^{-ikx_j}\right) = \sum_{m=k \ mod(N)} \hat{c}_m. \end{aligned} \tag{2.61}$$

Thus the amplitude of the m^{th} mode of the DFT is equal to the m^{th} mode plus the sum of the amplitudes of all other continuous modes that are indistinguishable from the m^{th} mode on the grid.

In Fourier transform theory

$$u(x) = \lim_{n\to\infty} \int_{-\infty}^{\infty} \mathcal{D}_n(x-y)u(y)\,dy = \int_{-\infty}^{\infty} \delta(x-y)u(y)\,dy, \tag{2.62}$$

where

$$\mathcal{D}_n(x-y) = \frac{1}{2\pi}\int_{-n}^{n} e^{-ik(x-y)}\,dy = \frac{\sin[n(x-y)]}{\pi(x-y)}, \tag{2.63}$$

is analogous to the Dirichlet kernel in (2.44). The analog of (2.43) and (2.44) in DFT theory is:

$$\begin{aligned} p(x) &= \sum_{k=-m}^{m} c_k e^{ikx} = \sum_{k=-m}^{m} \left(\frac{1}{N} \sum_{j=-m}^{m} u_j e^{-ikx_j} \right) e^{ikx} \\ &= \sum_{j=-m}^{m} \left(\frac{1}{N} \sum_{k=-m}^{m} e^{ikx} e^{-ikx_j} \right) u_j = \frac{1}{N} \sum_{j=-m}^{m} D_m(x - x_j) u_j, \end{aligned} \tag{2.64}$$

where $D_m(x)$ is the Dirichlet kernel in (2.44) and $N = 2m + 1$. The summation over k in (2.64) can be carried out by using the formula for the sum of a geometric progression with common ratio $w = \exp[i(x - x_j)]$. On the grid $(1/N)D_m(x_i - x_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta symbol. The above expression is a statement of a particular case of the Shannon's sampling theorem for periodic functions. It says that a band-limited function (a function having a Fourier transform confined to a finite interval $[-N/2, N/2]$) can be reconstructed explicitly by sampling at a uniform rate N. One can also differentiate (2.64) to obtain a formula for the derivative $p'(x)$, in terms of the derivative of the Dirichlet kernel, $D'_m(x - x_j)$ at any point x.

Formula (2.64) is analogous to the Lagrange interpolation formula:

$$p(x) = \sum_{k=-m}^{m} l_j(x) u_j,$$

where $l_j(x)$ is the Lagrangian interpolation polynomial discussed in Section 2.1, for which $l_j(x_i) = \delta_{ij}$.

Another basis that is sometimes used consists of Gaussians, $\exp[(x - x_j)^2/(2\sigma_j^2)]$. This basis is not orthogonal and the support is infinite for each element.

Smoothness of Fourier Approximations

We now overview the relation between the smoothness of the function in the physical space and the decay of the Fourier coefficients. The example below illustrates the Fourier space behavior of splines. Splines are piecewise smooth polynomial approximations [53].

Example

Consider a B-spline of order 0,

$$B_0(x) = \begin{cases} 1, & \text{in } [-1/2, 1/2]; \\ 0, & \text{otherwise.} \end{cases}$$

Its Fourier transform is: $\hat{B}_0(k) = \operatorname{sinc}\left(\frac{1}{2}k\right)$. Now introduce a spline of order one,

$$B_1(x) = B_0(x) * B_0(x) = \begin{cases} 1 - x, & \text{in } [0, 1]; \\ 1 + x, & \text{in } [-1, 0]; \\ 0, & \text{otherwise.} \end{cases}$$

The Fourier transform of $B_1(x)$ is $\hat{B}_1(k) = \operatorname{sinc}^2\left(\frac{1}{2}k\right)$.

Continuing in the same fashion, we get B-splines of higher and higher order. By induction, one can prove that the support of $B_p(x)$ is increasing and is contained in the interval $[-(p+1)/2, (p+1)/2]$. The smoothness increases as well, B_p has piecewise continuous derivatives of order p and continuous derivatives of the lower order. The corresponding decay of the transform $B_p(k) \sim \operatorname{sinc}^p\left(\frac{1}{2}k\right)$ is $O(\frac{1}{k^p})$. The pointwise error is intimately related to the decay of the Fourier coefficients, and decays like $\sim O(\frac{1}{N^p})$ near the singularity for a function having a piecewise continuous p^{th} derivative and continuous derivatives of lower order. The order of the pointwise convergence is $\sim O\left(\frac{1}{N^{p+1}}\right)$ away from the singularity.

For C^∞ functions, the order of convergence is faster than any power of n. This is called spectral convergence. For example, for a discontinuous function, an $O(1)$ error manifests itself in the Gibbs phenomenon, which is an oscillatory overshoot near the discontinuity.

The Gibbs Phenomenon

Consider the periodic extension of the step function:

$$\phi(x) = \begin{cases} 1, & \text{if } 0 \le x < \pi; \\ 0, & \text{if } \pi \le x < 2\pi. \end{cases} \tag{2.65}$$

The complex Fourier series for the step function $S_N[\phi(x)]$ consisting of $N = 2m+1$ terms obtained from (2.39), with $u = \phi(x)$, exhibits oscillations near the points of discontinuity at $x = 0$ and $x = \pi$ (see also [39, Ch. 2]). A discussion of the maximum overshoot of the Fourier series at $x = 0_+$ for large N is given below.

From (2.43)–(2.44),

$$S_N[\phi(x)] = \frac{1}{2\pi}\int_0^{2\pi} D_m(x-y)\phi(y)\,dy = \frac{1}{2\pi}\int_0^{\pi} D_m(x-y)\,dy,$$

where the Dirichlet kernel is given by (2.44). Using the integration variable $\xi = x - y$ in the above equation, we obtain

$$\begin{aligned} S_N[\phi(x)] &= \frac{1}{2\pi}\int_{x-\pi}^{x} D_m(\xi)\,d\xi = \frac{1}{2\pi}\int_0^{x} D_m(\xi)\,d\xi \\ &\quad + \frac{1}{2\pi}\left(\int_{-\pi}^{0} D_m(\xi)\,d\xi + \int_{x-\pi}^{-\pi} D_m(\xi)\,d\xi\right). \end{aligned} \tag{2.66}$$

If x is not in the neighborhood of $x = \pi$, and m is large, then the last integral on the right-hand side of (2.66) can be neglected (note, for large m, $D_m(z)$ behaves like a sequence of delta functions located at $z = 2\pi j$, where j is integral). Also from (2.44),

$$\int_{-\pi}^{0} D_m(\xi)\,d\xi = \int_0^{\pi} D_m(\xi)\,d\xi = \pi.$$

For x in the vicinity of $x = 0$, the first integral on the right-hand side of (2.66) has extrema at the points $x = \xi_j$ where $D_m(\xi_j) = 0$, i.e. at $\xi_j = 2\pi j/(2m+1)$. This behavior leads to oscillations of $S_N[\phi(x)]$ near $x = 0$, which are responsible for the Gibbs phenomenon. The maximum excursion of the oscillations near $x = 0$, in the region $x > 0$, occurs at $x = \xi_1 = 2\pi/N$. Thus the maximum amplitude near $x = 0_+$ is approximately given by

$$\begin{aligned} S_N[\phi(\xi_1)] &\approx \frac{1}{2\pi}\left(\pi + \int_0^{2\pi/N} D_m(\xi)\,d\xi\right) \\ &\approx \frac{1}{2} + \frac{1}{\pi}\int_0^{2\pi/N} \frac{\sin\left[\left(m+\frac{1}{2}\right)x\right]}{x}\,dx \\ &= \frac{1}{2} + \frac{1}{\pi}\int_0^{\pi} \frac{\sin(z)}{z}\,dz = 1.08949, \end{aligned} \tag{2.67}$$

which is approximately 9% of the jump. A similar estimate can be carried out to give the overshoot at $x = 0_-$.

It is desirable to reduce the Gibbs phenomenon near discontinuity points, while producing an accurate approximation of the function elsewhere. This smoothing process can be achieved by replacing the Fourier coefficients c_k by $\sigma_k c_k$ in the truncated Fourier series $S_N[u]$, to obtain the smoothed, truncated Fourier series $\mathcal{S}_N[u]$, with Fourier coefficients $\sigma_k c_k$, with increased decay rate. Typically the σ_k are non-negative numbers, with $\sigma_0 = 1$, $\sigma_k = \sigma_{-k}$, and $\sigma_{|k|}$ a decreasing function of k. Smoothing by Cesáro sums is one method of smoothing in which one takes the arithmetic mean of the $\{S_k[u] : 0 \leq k \leq m\}$ as:

$$\mathcal{S}_N[u] = \frac{1}{m+1} \sum_{k=0}^{m} S_k[u] = \sum_{j=-m}^{m} \left(1 - \frac{|j|}{m+1}\right) \hat{c}_j e^{ijx},$$

where $N = 2m+1$. The smoothing factors in this case are $\sigma_j = 1-|j|/(m+1)$. Other possible choices are Lanczos smoothing ($\sigma_j = \operatorname{sinc}(j\pi/m)$) or raised cosines ($\sigma_j = \cos^2(j\pi/m)$).

Similar ideas will be used in a later section to smooth out the solution of the numerical scheme, where popular choices for low pass filters (eliminating oscillatory modes) are $\exp(-k^{2\sigma})$, (super-Gaussian), and $\frac{1}{(1+k^{2\sigma})}$. Numerous other filters are often used in digital signal processing [141]. Application of DFT to pseudo-spectral methods and spectral methods of approximations will be described in later chapters. The DFT is also crucial in the linear stability analysis of finite difference schemes, which is called von Neumann analysis.

2.4 Method of Weighted Residuals, Finite Element and Finite Volume Methods

In this section, we describe the main ideas of the method of weighted residuals (MWR). We start with an illustration of the method on a 1D boundary value problem for the advection-diffusion equation

$$u_t + au_x = \epsilon u_{xx}, \quad 0 < x < 1, \quad t > 0, \tag{2.68}$$

subject to the initial and boundary value conditions:

$$u(0,x) = x, \quad u(t,0) = 0, \quad u(t,1) = 1. \tag{2.69}$$

The coefficients a and ϵ are positive constants.

Consider an approximation to the solution $u_N(x)$ represented in some basis as

$$u_N(t,x) = \sum_{j=0}^{N} c_j(t)\phi_j(x). \tag{2.70}$$

This approximation, in general, will not satisfy (2.68)–(2.69) exactly. The functions $\{\phi_j(x)\}$ in (2.70) are known as trial functions. The amount by which the quantity

$$R(u_N) = u_{Nt} + au_{Nx} - \epsilon u_{Nxx} \tag{2.71}$$

deviates from zero is called the residual.

The approximation (2.70) is substituted into a weak form of the differential equation that is obtained by multiplication of the equation with the test functions $\{\psi_i(x) : i = 0, 1, \dots, N\}$, which vanish at $x = 0$ and $x = 1$. After integrating the last term by parts once, we obtain the equation:

$$\int_0^1 R(u_N)\psi_i\,dx = \int_0^1 u_{Nt}\psi_i + au_{Nx}\psi_i + \epsilon u_{Nx}\psi_{ix}\,dx = 0, \qquad i = 0, 1, \dots, N. \tag{2.72}$$

The test functions $\{\psi_i(x)\}$ have been chosen to have compact support on $[0,1]$, meaning that the functions $\psi_i(x)$ are bounded on $[0,1]$ and vanish for $x \notin (0,1)$. The integration by parts is done in order to reduce the smoothness requirement on the basis (for example existence of piecewise continuous derivatives is sufficient to integrate this term).

There are $N+1$ unknown coefficients $c_j(t)$ to be determined from the above set of equations. The choice of the test functions $\psi_i(x)$ gives rise to various weighted residual methods. When $\psi_i(x) = \phi_i(x)$, we get the Galerkin method. The Petrov-Galerkin method corresponds to $\psi_i(x) \neq \phi_i(x)$ and is often used to better approximate the residual rather than the solution. Setting the residual to zero at a fixed set of x values, or equivalently choosing $\psi_i(x) = \delta(x - x_i), i = 0, 1, \dots, N$, we get the collocation method.

The choice of basis functions with compact support (usually piecewise polynomials expanded in terms of the B-splines) gives the finite element method, while the choice of an orthogonal basis (usually eigenfunctions to some self-adjoint problem like complex exponential functions, or trigonometric functions used in Fourier methods, or orthogonal polynomials with respect to some weight functions $w(x)$) gives a spectral weighted residual method, such as the spectral Galerkin method. A mix between the two, when the choice of the nodes for piecewise polynomials consists of the roots of the orthogonal polynomials, with Gaussian nodes, results in the spectral element method. It is known from the theory of quadrature rules for numerical integration, that the Gaussian quadrature rules based on approximating the integrand by a system of orthogonal polynomials $\{p_n(x) : 0 \leq n \leq N\}$ with respect to a weight function $w(x)$, and with integration nodes chosen to be the roots of the $(N+1)^{st}$ polynomial $p_{N+1}(x)$, yields approximations with the best least squares approximation (e.g. [148]).

We continue our example (2.72) by choosing the Galerkin approximation with a piecewise linear approximation space generated by use of the hat functions or $B_1(x - x_i)$ splines as trial functions. The first and last hat functions are chosen to contain only half of the "hat". Thus, we use the trial functions:

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h_{i-1} & \text{if } x \in [x_{i-1}, x_i], \\ -(x - x_{i+1})/h_i & \text{if } x \in [x_i, x_{i+1}], \\ 0 & \text{otherwise} \end{cases} \tag{2.73}$$

where $h_i = x_{i+1} - x_i$. The boundary conditions (2.69) at $x = 0$ and $x = 1$ fix the values of $c_0(t)$ and $c_N(t)$. In the present case $u_N(t,0) = c_0(t) = 0$ (note only $\phi_0(x)$ can contribute, and that $\phi_0(0) = 1$), and similarly, $u_N(t,1) = c_N(t)\phi_N(1) = c_N(t) = 1$. Thus the boundary conditions (2.69) are satisfied by setting $c_0(t) = 0$ and $c_N(t) = 1$. Using the Galerkin method in (2.72) in which the test functions $\psi_i(x) = \phi_i(x)$, we find that (2.72) reduces to:

$$\begin{aligned} &\int_0^1 \sum_{j=1}^{N-1} \left[c_{jt}\ \phi_j\phi_i + c_j(t)(a\phi_{jx}\phi_i + \epsilon\phi_{jx}\phi_{ix})\right] dx \\ &\qquad + \int_0^1 \phi_{Nx}(a\phi_i + \epsilon\phi_{ix})\, dx = 0, \quad i = 1, 2, \ldots, N-1. \end{aligned} \tag{2.74}$$

Rewriting (2.74) in matrix form, and switching the integration and summation order, we get the matrix differential equation system:

$$\mathbf{M}\frac{d}{dt}\mathbf{c}(t) + \mathbf{A}\mathbf{c}(t) = \mathbf{b}. \tag{2.75}$$

In (2.75) the arrays **M** and **A** and the boundary vector **b** have components:

$$M_{ij} = (\phi_i, \phi_j), \quad A_{ij} = (a\phi_i, \phi_{jx}) + (\epsilon\phi_{ix}, \phi_{jx}),$$
$$b_i = -(\phi_{Nx}, a\phi_i + \epsilon\phi_{ix}). \tag{2.76}$$

Here, the matrix elements are written in terms of the usual L_2 inner product and the indices $i, j = 1, 2, \ldots (N-1)$. The matrix **M** is called the mass matrix and the matrix **A** is called the stiffness matrix. The initial values for the column vector $\mathbf{c}(t) = (c_1(t), c_2(t), \ldots, c_{N-1}(t))^T$ at time $t = 0$, can be determined from the initial data (2.69) by noting that the spline basis functions satisfy the equations $\phi_i(x_j) = \delta_{ij}$ $(i, j = 1, 2, \ldots, N-1)$, and hence the equations:

$$u_N(0, x_i) = c_i(0) = x_i, \tag{2.77}$$

determine the initial values $\{c_i(0)\}$. The ODE system (2.75) can be integrated as an initial value problem, with initial data (2.77) by a standard method, like the Adams-Bashforth or Runge-Kutta. These methods are described in the next section.

For a given basis, the book-keeping of inner products is a matter of integration of piecewise smooth functions (constants, linear or quadratic) on the interval $[0, 1]$. The non-zero inner products are:

$$(\phi_i, \phi_i) = \frac{1}{3}(h_{i-1} + h_i), \quad (\phi_i, \phi_{i-1}) = \frac{1}{6}h_{i-1},$$
$$(\phi_{(i-1)x}, \phi_i) = -\frac{1}{2}, \quad (\phi_{i-1}, \phi_{ix}) = \frac{1}{2},$$
$$(\phi_{ix}, \phi_{ix}) = \frac{1}{h_{i-1}} + \frac{1}{h_i}, \quad (\phi_{ix}, \phi_{(i-1)x}) = -\frac{1}{h_{i-1}}. \tag{2.78}$$

For a uniform grid, we find:

$$M_{ij} = \frac{2h}{3}\delta_{i,j} + \frac{h}{6}\left(\delta_{i,j+1} + \delta_{i,j-1}\right),$$
$$A_{ij} = \frac{a}{2}\left(\delta_{i,j-1} - \delta_{i,j+1}\right) + \epsilon\left(\frac{2}{h}\delta_{i,j} - \frac{1}{h}\left(\delta_{i,j-1} + \delta_{i,j+1}\right)\right),$$
$$b_i = \left(\frac{\epsilon}{h} - \frac{a}{2}\right)\delta_{i,N-1}, \tag{2.79}$$

where $\delta_{i,j}$ is the Kronecker delta symbol. The matrix equation (2.75) reduces to:

$$\frac{2h}{3}\dot{c}_i + \frac{h}{6}\left(\dot{c}_{i-1} + \dot{c}_{i+1}\right) + ahD_0c_i - \epsilon hD_+D_-c_i = \left(\frac{\epsilon}{h} - \frac{a}{2}\right)\delta_{i,N-1}, \tag{2.80}$$

where $D_0c_i = (c_{i+1} - c_{i-1})/(2h)$ and $D_+D_-c_i = (c_{i+1} - 2c_i + c_{i-1})/h^2$ are the central difference approximations for the first and second spatial derivatives, and $\dot{c}_i = dc_i/dt$ is the time derivative of c_i. Thus on a uniform grid the discretization of the advection and diffusion terms reduces to central finite differences, but the matrix M is a weighted average of the nearby values of the time derivatives similar to compact finite differences. This indicates that there will be similarities in the behavior of these methods, like numerical dispersion, diffusion, possible nonlinear instabilities, etc.

In some cases, FEMs are known to preserve some invariants of the continuous problem, giving another rationale for the usage of the FEM, especially to study long-time behavior of the solution. Other methods are often designed to do the same as well; for example, to preserve a conserved density such as the energy, a symplectic property or other important invariants as described in later chapters.

In the case of self-adjoint elliptic problems, a weak formulation leads to the symmetric bilinear form that allows one to define a new inner product space ("energy" space) and interpret the whole procedure as a projection of the exact solution onto a subspace defined by the chosen basis as well as a minimization procedure of the quadratic (or convex) functional that has a unique minimum (variational formulation), as illustrated in the examples below. In the case of time-dependent problems, the variational formulation is lost and the finite element procedure is usually used to accurately approximate complex geometries by choosing an appropriate basis like triangulation with adaptive mesh to fit the geometry or the solution, or quadrilaterals that may be mapped onto curved boundaries.

In summary, the finite element method consists of the following four steps: generation of adaptive grids to accommodate complex geometry (element construction); choice of polynomials over each element and setting up inner product matrices; discretization of the ODE system; and, finally, efficient solution of the large and sparse linear matrix system where the usual choices are direct or iterative methods with preconditioning depending on the memory requirements and rates of convergence desired. Each step is quite involved, and testimony to this are the numerous software companies that have developed algorithms for the different steps.

Example. Minimization Methods for Elliptic Problems

Consider Poisson's equation:

$$-\Delta u = f, \quad \text{in} \quad \Omega, \tag{2.81}$$

satisfying the Dirichlet boundary conditions:

$$u = 0 \quad \text{on} \quad \Gamma, \tag{2.82}$$

where Ω is a bounded open domain in R^n with smooth boundary $\Gamma = \partial\Omega$.

Multiplying Poisson's equation by a continuous test function v with piecewise continuous first derivatives and zero boundary values, and integrating by parts, we obtain the following weak form

$$(u, v)_s = (f, v), \tag{2.83}$$

where (f, v) denotes the usual scalar product while $(u, v)_s$ is equivalent to the scalar product in Sobolev space H^1 and is defined as $(u, v)_s = \int_\Omega \nabla u \cdot \nabla v dx$. Writing

$$F_1(u, v) = (u, v)_s - (f, v) \equiv \int_\Omega d^3x \left(\nabla u \cdot \nabla v - fv\right), \tag{2.84}$$

we have $\delta F_1(u, v)/\delta v = -(\nabla^2 u + f)$ (e.g. [71]). Thus Poisson's equation (2.81) is recovered by requiring that the functional $F_1(u, v)$ be stationary with respect to variations of v.

Similarly, the functional:

$$F_2(v) = \frac{1}{2}(v, v)_s - (f, v) \equiv \int_\Omega d^3x \left(\frac{1}{2}\nabla v \cdot \nabla v - fv\right), \tag{2.85}$$

has variational derivative $\delta F_2(v)/\delta v = -[\nabla^2 v + f]$. Thus the requirement that the functional $F_2(v)$ is stationary with respect to v, meaning that $\delta F_2(v)/\delta v = 0$, yields Poisson's equation (2.81). The above weak form of the equations (2.83) corresponds to the Euler-Lagrange equation of the functional $F_2(v)$ with respect to variations of v, and the critical solution turns out to be the unique minimizer of this functional.

When the solution u is expanded in some basis, this can be interpreted as solving the weak form of the equation in the corresponding subspace of the Hilbert space in which the original solution is sought. Thus the approximate solution $\tilde{u}$ satisfies the equation:

$$(\tilde{u}, v)_s = (f, v),$$

for all $v \in \tilde{H}$ which is a subspace of H. Subtracting from the original weak form gives

$$(u - \tilde{u}, v)_s = 0.$$

In other words, the approximate solution is an orthogonal projection of the exact solution onto a subspace of the original Hilbert space of weak solutions.

The typical interpolation error of the function in Hilbert space, say by piecewise linear functions, is

$$||u - \tilde{u}|| \leq Ch^2,$$

where h represents the discretization parameter, and the constant C depends on the derivatives of u. This implies that to minimize the error, one can manipulate the discretization parameter in regions of large gradients and geometric singularities like corners, as well as avoid small angles in triangulation that will lead to large cross derivative terms. The gradients include gradients due to grid variations as well. Therefore, smoothly varying grids are the best, while sharp interfaces are potential sources of instabilities, numerical reflections, etc. that might be difficult to control, as examples in the later sections will illustrate.

Example. Spectral Element Methods

These methods attempt to combine the grid adaptivity of the finite elements with the spectral accuracy of spectral methods for smooth solutions, if over each element one introduces Gaussian nodes (zeroes of orthogonal polynomials), and inner products are computed using Gaussian quadratures to achieve spectral accuracy (exact for polynomials of highest degree possible due to the choice of the nodes and the weights). For example, over each element a Lagrangian interpolation polynomial is introduced with Gaussian nodes, corresponding to Gauss-Lobatto or Gauss-Legendre-Lobatto quadrature rules, etc., depending on the weight function $w(x)$ used for the quadrature rule.

In several space dimensions, tensor products of the basis functions are used to generate the basis. For example in 3D, the solution is approximated as:

$$u(x,y,z) \approx \sum_{i,j,k} u(\xi_i,\xi_j,\xi_k) l_i(x) l_j(y) l_k(z),$$

where ξ_i, ξ_j, ξ_k represent the nodes and each dot product is computed using the corresponding Gaussian quadrature,

$$(f,g) = \sum_{i,j,k} w_{ijk} f(\xi_i,\xi_j,\xi_k) g(\xi_i,\xi_j,\xi_k),$$

and w_{ijk} are the Gaussian weights at the nodes.

As the Gaussian nodes are the inner nodes, the interface between the elements is left open. The interface is treated differently for various types of problems. For example, continuity of the variables or its derivatives is often enforced for elliptic or parabolic problems, while discontinuous elements are often used for hyperbolic problems with computation of the flux at the interface via some wave propagation algorithm like a Riemann solver. Note that both strategies may clash for advection-diffusion problems. Stability of the high order time-stepping that goes with spectral elements is often difficult to control as well.

Example. Boundary Element Method

Consider the solution of Laplace's equation:

$$\Delta u = 0, \quad \text{in } \Omega,$$

satisfying the Dirichlet boundary conditions:

$$u = \phi \quad \text{on } \Gamma,$$

where Ω is a bounded open domain in R^n with smooth boundary Γ.

Assume that the solution is a linear combination of the free space Green's functions that are weighted by an unknown function $g(\mathbf{x})$ to satisfy the boundary condition:

$$u(\mathbf{x}) = \int_\Gamma \frac{\partial G(\mathbf{x}-\mathbf{y})}{\partial n} g(\mathbf{y})\, dS(\mathbf{y}), \tag{2.86}$$

where $G(\mathbf{x}-\mathbf{y}) = 1/(4\pi|\mathbf{x}-\mathbf{y}|)$ is the fundamental Green function solution of the Laplace equation in free space, $\mathbf{n}$ denotes the outward normal to the boundary Γ, $dS(\mathbf{y})$ is the differential area element on the boundary surface; $\partial/\partial n$ denotes the normal derivative on the boundary, and $g(\mathbf{x})$ is an unknown function. The solution form (2.86) applies to a point inside the boundary $\partial\Omega$. Since G satisfies Laplace's equation, the function $g(\mathbf{x})$ is determined by enforcing the boundary condition. Taking the limit towards the boundary, with a small deformation about the boundary point $\mathbf{x}$, such that the deformed boundary now includes $\mathbf{x}$ inside the deformed domain, yields the following integral equation along the boundary for the unknown function $g(\mathbf{x})$:

$$\phi(\mathbf{x}) = \frac{1}{2}g(\mathbf{x}) + \text{CPV}\left(\int_\Gamma \frac{\partial G(\mathbf{x}-\mathbf{y})}{\partial n} g(\mathbf{y})dS(\mathbf{y})\right), \tag{2.87}$$

where CPV in (2.87) denotes the Cauchy principal value integral. Thus the boundary value problem for the whole region Ω, reduces to the solution of an integral equation on the boundary Γ of the region.

Example. Finite Volume Method

Consider the test function that is a characteristic function that equals one or zero over each element. An example of the finite volume method for hexagonal elements generated by a regular Delaunay-Voronoi dual mesh, is shown in Figure 5.2. Consider an integral form of a system of hyperbolic conservation laws

$$\frac{d}{dt}\int_A u\,dS + \int_\Gamma \mathbf{f}\cdot\mathbf{n}\,dl = 0, \tag{2.88}$$

where u is one of the conserved variables, A and Γ represent the volume and the boundary of the control region, and $\mathbf{n}$ is the outward normal to the cell. Integrating in time and over the computational cell, shown in Figure 5.2, we get the following finite volume approximation,

$$u_{ij}^{n+1} = u_{ij}^n - \frac{1}{A}\sum_k \int_0^{\Delta t} dt \int_{\Gamma_k} \mathbf{f}\cdot\mathbf{n}\,dl, \tag{2.89}$$

where u_{ij}^n represents the cell average at time t_n and Γ_k is the length of the edge of the computational cell. The cell averages are assumed to be given, while the fluxes are approximated using the values on the edge computed as solutions to the linear Riemann problem, that is described in a later chapter.

In the case of Maxwell's equations, written for simplicity for the case of a vacuum,

$$\begin{aligned}
\mathbf{E}_t - \nabla\times\mathbf{B} &= 0,\\
\mathbf{B}_t + \nabla\times\mathbf{E} &= 0,
\end{aligned}$$

or in integral form over a control volume A is

$$\frac{d}{dt}\int_A \mathbf{E}\,dS - \int_\Gamma \mathbf{B}\,dl = 0,$$
$$\frac{d}{dt}\int_A \mathbf{B}\,dS + \int_\Gamma \mathbf{E}\,dl = 0.$$

The Yee scheme [201] exploits the fact that one variable defines the circulation for the other, and staggers the variables in space and time as

$$\frac{\mathbf{E}^{n+1} - \mathbf{E}^n}{\Delta t} - \frac{1}{A}\sum_{\Gamma_k} \Delta l_k \mathbf{B}_k^{n+1/2} = 0,$$
$$\frac{\mathbf{B}^{n+1/2} - \mathbf{B}^{n-1/2}}{\Delta t} - \frac{1}{A}\sum_{\Gamma_j} \Delta l_j \mathbf{E}_j^n = 0.$$

Here the electric field is defined as an average over the computational cells and the magnetic field is defined at the middle of the cell interfaces at intermediate time levels.

2.5 Project Assignment

Choose two canonical PDEs (or similar equations) from Chapter One, for example, advection and NLS equations. For each of these equations select two spatial discretization methods. You may choose the same two methods for both or they might be different for each equation. The resulting semi-discrete approximations may be advanced in time using any standard ODE solver as long as it results in a stable algorithm. Before implementation, discuss the choices of both spatial and temporal discretizations with the instructor. For all four algorithms, the computational domain is $[-5, 5]$ in suitably selected units, and the initial conditions are zero outside of the interval $[-1, 1]$, and top-hat, hat, Gaussian and Gaussian pulse functions specified in the interval $[-1, 1]$, respectively, are:

1. $u(x, 0) = \frac{1}{2}(H(x + 1) + H(1 - x))$,
2. $u(x, 0) = 1 - |x|$,
3. $u(x, 0) = \exp(-x^2/(2 * \sigma^2))$, $\sigma = 0.2$,
4. $u(x, 0) = \exp(-x^2/(2 * \sigma^2)) \cos(40x)$, $\sigma = 0.2$.

Note that the last two initial conditions are discontinuous at $x = \pm 1$, albeit the discontinuity is of the order of 10^{-6}. For the computational domain boundary conditions at $x = \pm 5$ you may choose Dirichlet, Neumann, periodic or any other suitable boundary condition. State clearly which boundary condition was implemented for each algorithm. For example, if a simple “no update” numerical boundary condition is used and all of the functions are set initially to zero at the boundary, state explicitly that this boundary is equivalent to a homogeneous Dirichlet boundary condition.

For the above initial data, run each of the four algorithms up to some appropriate dimensionless time $t = 1$, showing four time snapshots every 0.25 units of time. Compare properties of the computed numerical solutions with the known properties of each canonical equation in terms of numerical damping, dispersion, propagation speed, etc., when changing the spatial resolution (number of spatial points per “peak” of the initial distribution, or per wavelength, for the last initial condition), while keeping the numerical accuracy of the time integrator constant. State clearly the values of all numerical discretization parameters used in each computation.

2.6 Project Sample

Finite-difference, finite-volume, finite-element and spectral collocation methods can be considered as particular cases of a general approach based on the MWR. The basic principles of such formulation can be described by considering spatial discretization for a generic PDE given by:

$$\begin{cases} \mathcal{L}v = g, & \mathbf{r} \in \Omega \\ v = v_{\partial\Omega}, & \mathbf{r} \in \partial\Omega, \end{cases} \tag{2.90}$$

where $\mathcal{L}$ is a differential operator continuous in spatial coordinates $\mathbf{r}$, $g(\mathbf{r})$ is a given function and Ω, $\partial\Omega$ are the solution domain and its boundary. Taking a space of trial functions $v \in V$ and a space of test functions $w \in W$, the MWR requires the projection of the function $\mathcal{L}v - g$ (the continuous residual), on a space of test functions, to be zero for each $w \in W$:

$$\int_\Omega (\mathcal{L}v - g)w \, d\Omega = 0.$$

If finite-dimensional sub-spaces $v^h \subset V$ and $w^h \subset W$ are chosen for the trial and test functions, then the discrete weighted residual formulation can be written as:

$$\int_\Omega (\mathcal{L}^h v^h - g)w^h \, d\Omega = 0,$$

where $\mathcal{L}^h$ is the discrete approximation of the differential operator and, due to the approximation, the discrete residual $\mathcal{L}^h v^h - g$ of the equation is not zero. By choosing appropriate basis functions ϕ_i and ψ_i, the approximate solution v^h and the test functions can be represented by the truncated series expansions:

$$v^h = \sum_{i=0}^{N} a_i \phi_i, \quad w^h = \sum_{i=0}^{N} b_i \psi_i. \tag{2.91}$$

Then the discrete weighted residual method becomes:

$$\sum_{i=0}^{N} a_i \int_\Omega (\mathcal{L}^h \phi_i)\psi_j \, d\Omega = \int_\Omega g\psi_j \, d\Omega, \quad j = 0, \ldots, N, \tag{2.92}$$

which is a system of equations for coefficients a_i, from which the approximate solution v^h of the differential equation (2.90) can be constructed using expansion (2.91).

The choice of the test functions ψ_j leads to different numerical discretization methods. For example, in the finite difference methods a discrete set of collocation points $\mathbf{r}_j$ is defined in the solution domain Ω, and delta functions are chosen for the test functions:

$$\psi_j(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_j), \quad j = 0, \ldots, N.$$

The discrete residual is required to be zero at all collocation points $\mathbf{r}_j$, $j = 0, \ldots, N$. Using a truncated Taylor series expansion for $\mathcal{L}^h$ around these collocation points, and without actual approximation of the solution v in terms of ϕ_i, results in a finite difference scheme:

$$\mathcal{L}^h v|_j = f(\mathbf{r}_j), \quad j = 0, \ldots, N.$$

Thus the numerical accuracy of the finite difference scheme depends both on the truncation order of approximation for $\mathcal{L}^h$, and on the number of collocation points N.

Spectral collocation methods can be derived by using delta functions for ψ_j and the values v_i, defined at the collocation points $\mathbf{r}_j$, in place of the coefficients a_i in the approximate solution given by the truncated series expansion (2.91),

$$v^h = \sum_{i=0}^{N} v_i \phi_i,$$

and by choosing the approximating functions to be orthogonal polynomials in V, i.e. $(\phi_i, \phi_j) = 0$. Then, since ϕ_i are polynomial functions, the functions $\mathcal{L}\phi_i$ can either be evaluated without approximating the continuous operator $\mathcal{L}$, or, a discrete approximation $\mathcal{L}^h$ can be derived by computing the derivatives in terms of the coefficients of the approximate solution.

The finite-volume discretization uses partitioning of the domain Ω into a set of non-overlapping sub-domains Ω_j, with the test functions ψ_j defined to be equal to unity inside of the domain Ω_j, and zero outside of it:

$$\begin{cases} \psi_i = 1, & \mathbf{r} \in \Omega_j \\ \psi_i = 0, & \text{otherwise.} \end{cases}$$

Similar to the finite difference method, the finite-volume approach does not use an explicit approximation for the solution v^h, so the Equation (2.92) can be written as:

$$\int_{\Omega_j} (\mathcal{L}^h v - g)\, d\Omega_j = 0, \quad j = 0, \ldots, N.$$

If the volume integral over Ω_j can be evaluated in terms of the surface integrals using Green's theorem, then the resulting discrete set of equations can be solved numerically, with the accuracy of the approximation determined by the accuracy of the integration and by the number of sub-domains N.

As an example of a finite-volume method, we consider discretization of the electron-hole conservation equations used in semiconductor device modeling [164]. We start with an integral form of the steady-state current density equation for electrons (the derivation for holes is analogous):

$$\int_{\Omega} \nabla \cdot \mathbf{J}\, d\Omega = \int_{\partial\Omega} \mathbf{J}_{i,j}\, d\mathbf{s} = \int_{\Omega} (G - R)_{i,j}\, d\Omega, \tag{2.93}$$

where $\nabla \cdot = \mathcal{L}$, $G - R = g$ and Ω is the volume of the computational cell centered around a grid point i, j. Using a second-order accurate central finite-difference approximation for $\mathcal{L}^h$, and taking the generation-recombination rate $G - R$ to be constant in the cell volume and the current density $\mathbf{J}$ to be constant on each surface area, the integrals (2.93) in the 2D case can be approximated by:

$$\begin{aligned} &(J_{x\,i+1/2,j} - J_{x\,i-1/2,j})\Delta y_i + (J_{y\,i,j+1/2} - J_{y\,i,j-1/2})\Delta x_i \\ &\qquad = (G - R)_{i,j} \Delta x_i \Delta y_i, \end{aligned} \tag{2.94}$$

with the current density defined on the surface of the rectangular computational cell in Cartesian coordinates. Considering the definition of the current density in terms of the number density n and the electrostatic potential ϕ, e.g.

$$J_{x\,i+1/2,j} = \mu_{i+1/2} n(x) \partial_x \phi(x) - D_{i+1/2} \partial_x n(x), \tag{2.95}$$

the variation of the carrier density $n(x)$ between the grid points x_i and x_{i+1} can be obtained by assuming a constant electric field (linear potential ϕ) in the interval $[x_i, x_{i+1}]$, and integrating Equation (2.95) with respect to x:

$$\begin{aligned} n(x) = A \exp\left(\frac{\phi(x)}{\phi_T}\right) \\ + (\partial_x \phi(x))^{-1} \frac{J_{x\,i+1/2,j}}{\mu_{i+1/2}} \left(1 - \exp\left(\frac{\phi(x)}{\phi_T}\right)\right), \end{aligned} \tag{2.96}$$

where A is an integration constant, and we used the definition of the thermal potential $\phi_T = D_{i+1/2}/\mu_{i+1/2} = kT/e$, which depends only on temperature. Applying the boundary conditions $n(x_i, j) = n_i$, $n(x_{i+1}, j) = n_{i+1}$ and using the approximation $\partial_x \phi(x) = (\phi_{i+1} - \phi_i)/\Delta x_i$, one gets

$$\begin{aligned} n_{i+1} \exp\left(\frac{\phi_i - \phi_{i+1}}{\phi_T}\right) = n_i + \frac{J_{x\,i+1/2}}{\mu_{i+1/2}} \frac{x_{i+1} - x_i}{\phi_{i+1} - \phi_i} \\ \times \left[\exp\left(\frac{\phi_i - \phi_{i+1}}{\phi_T}\right) - 1\right]. \end{aligned} \tag{2.97}$$

From this solution for n_{i+1}, the current density can be written in the following form:

$$J_{x\,i+1/2} = \frac{\mu_{i+1/2}}{\Delta x_i} \left[B\left(\frac{\phi_i - \phi_{i+1}}{\phi_T}\right) n_i - B\left(\frac{\phi_{i+1} - \phi_i}{\phi_T}\right) n_{i+1} \right], \tag{2.98}$$

where $B(u) = u/[\exp(u) - 1]$ is the Bernoulli function. Equation (2.98) explicitly accounts for the exponential dependence of the carrier density on the potential difference across the grid cell. Substituting Equation (2.98) and the corresponding expression for the $J_{y\,i,i+1/2}$, etc. into Equation (2.94) results, for a given potential distribution, in a set of linear algebraic equations for the unknowns $n_{i,j}$.

As an example of spectral collocation discretization [39] we consider numerical solution of the linear wave mixing equations:

$$\frac{\partial a_p}{\partial t} + \lambda_p \frac{\partial a_p}{\partial x} = -\sum_{s=1}^{7} \left(\Lambda_{ps} + \frac{\partial \lambda_s}{\partial x} \delta_{ps} \right) a_s, \qquad p = 1(1)7, \tag{2.99}$$

where a_p is the amplitude of the p-th wave mode, λ_p is its characteristic speed, and Λ_{ps} are the wave coupling coefficients that describe the interaction of backward and forward magnetohydrodynamics waves in an energetic-particle modified large scale background flow [193]. The eigenspeeds are ordered such that $\lambda_{p+1} > \lambda_p$. λ_1 and λ_7 correspond to the backward and forward fast magnetoacoustic waves, λ_2 and λ_6 correspond to the backward and forward Alfvén waves, λ_3 and λ_5 correspond to the backward and forward slow magnetoacoustic waves and $\lambda_4 \equiv u_x$ corresponds to the entropy wave. Using normalized coordinates, we look for a solution of (2.99) which is periodic in space on the interval $x \in [0, 2\pi]$, with an initial condition

$$a_p(x, 0) = a_{p0}(x), \quad p = 1(1)7. \tag{2.100}$$

To simplify the notation, we consider only one of the functions a_p, which we denote by v. For the set of N points (Fourier nodes) $x_j = 2\pi j/N$, $j = 0, \ldots, N-1$, the solution v can be written in terms of its discrete Fourier coefficients as:

$$v(x_j) = \sum_{k=-N/2}^{N/2-1} \tilde{v}_k e^{ikx_j}, \quad j = 0, \ldots, N-1, \tag{2.101}$$

where coefficients $\tilde{v}_k$ are given by the DFT:

$$\tilde{v}_k = \frac{1}{N} \sum_{j=0}^{N-1} v(x_j) e^{-ikx_j}, \quad -N/2 \le k \le N/2 - 1. \tag{2.102}$$

The Fourier collocation discretization requires that Equation (2.99) be satisfied at the points x_j. Denoting the numerical approximation to the solution by v^h, from Equation (2.99) we have conditions:

$$\begin{aligned} &\left. \frac{\partial v^h}{\partial t} + \lambda \frac{\partial v^h}{\partial x} + S(v^h) \right|_{x_j} = 0, \\ &v^h(x_j) = v_0(x_j), \quad j = 0, \ldots, N-1, \end{aligned} \tag{2.103}$$

where S corresponds to the right-hand side of Equation (2.99). At the

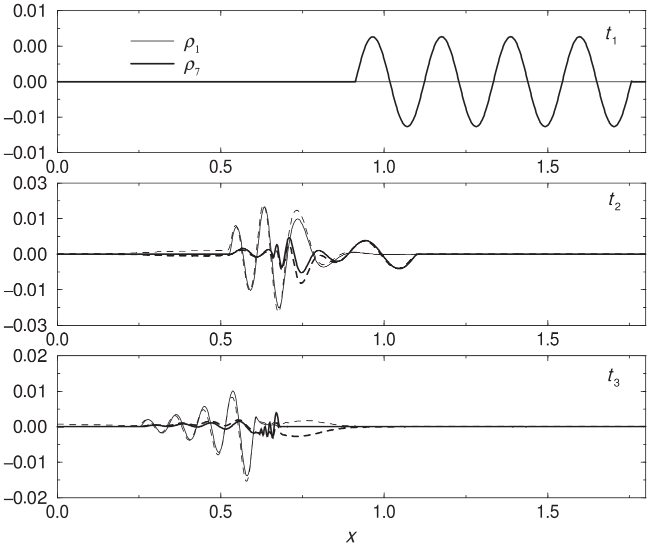


Figure 2.1: Comparison of the numerical solution of the wave mixing equations (2.99) using a Fourier collocation method (solid lines) with the solution of the equations for charged-particle modified fluid dynamics [193] using a second-order accurate finite-volume method (dashed lines).

Fourier nodes the spatial derivative in Equation (2.103) can be evaluated in the transform space using the Fourier collocation derivative $\mathcal{D}_N$ operator:

$$(\mathcal{D}_N v)_l = \sum_{k=-N/2}^{N/2-1} b_k e^{2ikl\pi/N},$$

$$b_k = \frac{ik}{N} \sum_{j=0}^{N-1} v_j e^{-2ikj\pi/N}, \quad j = 0, \ldots, N-1. \tag{2.104}$$

The system (2.103) then represents N ODEs that can be integrated in time using any suitable method, e.g. Runge-Kutta method.

Figure 2.1 shows numerical solutions of the wave-wave interaction equations, for an initial-value problem in which a forward fast wave mode a_7 with density perturbation ρ_7 is specified at $t = t_1 = 0$, and the backward wave amplitude a_1 is set to zero. Wave coupling is shown at $t = t_2$ and $t = t_3$ as the density perturbations ρ_1 and ρ_7 propagate through a slowly-varying background flow due to a charged particle modified shock wave. In Figure 2.1 the Fourier collocation solution is compared with the solution of the nonlinear charged-particle modified fluid flow equations [193], which are integrated using a second-order accurate finite-volume upwind scheme. As expected for the small amplitude of the initial density perturbation, there is a good agreement between the solution of the linear wave-wave interaction and nonlinear fluid flow equations.

Chapter 3

Convergence Theory for Initial Value Problems

3.1 Introduction to Convergence Theory

The convergence theory for numerical methods approximating time-dependent problems parallels the theory of ordinary differential equations (ODEs) where two types of behavior are studied, namely (1) the finite time solution and (2) the long-time asymptotic behavior where the solution either passes through an initial transient state and sets into a steady state, evolves into a periodic or chaotic motion or escapes to infinity.

For ODEs, these two questions lead to the study of dynamics and asymptotic states involving critical points, attractors and fast and slow evolution etcetera. The characterization of what is transient depends on the time scale of interest. For example, when a light is turned on, the onset of the current may be considered as a transient as far as the long-time behavior is concerned, whereas on a shorter time scale it is a dynamical process that is of interest in its own right.

Similarly, when discussing the behavior of a numerical scheme in time-dependent problems, we consider the dynamical behavior up to any fixed time T for a small but fixed discretization parameter Δt or a limiting case, assuming that Δt tends to zero. In this section we consider the nature of the convergence of the solution in the limit as $\Delta t \to 0$. In practice, one is often in a situation that is between the two theories, when Δt and space discretization steps are small, but not small enough to make numerical artifacts negligible. In such situations we will rely on the properties of the scheme revealed by the dispersion relation (phase and group velocities, amplification or damping of certain waves), symmetries and conservation properties.

There are two different types of convergence that are of interest. In the first type of convergence, the limit in which $\Delta t \to 0$ is studied. In the second type of convergence, Δt is a small but fixed parameter, and the convergence as an iteration parameter $N \to \infty$ is studied. The difference between these two notions of convergence may be illustrated by a simple example. The limit $\lim_{\Delta t \to \infty}(1+\Delta t)^{\frac{1}{\Delta t}} = e$ is a convergent limit. However, $(1 + \Delta t)^N \to \infty$ when we let the number of iterations $N \to \infty$ and keep Δt small but fixed.

We start the discussion of convergence theory by recalling the definition of consistency.

Definition 8. *A discrete operator is* consistent *of order $O(\Delta t^p, \Delta x^q)$ if*

$$(P - P_\Delta)v = O(\Delta t^p, \Delta x^q),$$

in the limit of vanishing discretization parameters.

Here the operator P is taken to be a continuous operator and P_Δ is its discrete approximation, and v is an arbitrary smooth function often taken as an exact solution of the continuous or discrete problem.

Convergence theory is concerned with the question of showing that solutions of the exact and approximate problems will be close to each other once the equations (the operators and mappings) are close to each other. Continuity or continuous dependence of the solutions on the small discretization parameters is the topic of study.

Definition 9. *A discrete approximation is called* convergent *(accurate) to order $O(\Delta t^p, \Delta x^q)$ if*

$$U - U_\Delta = O(\Delta t^p, \Delta x^q),$$

in the limit of vanishing discretization parameter for any time t, $0 \le t \le T$, where T is a fixed but arbitrary finite time; and U and U_Δ represent the exact and approximate solutions, respectively.

If the discrete solution depends on an integer parameter N, then one can talk about the rate of convergence of the corresponding approximating sequences, such as linear, quadratic or exponential.

To check the consistency of a finite difference scheme, one may use the Taylor expansion of the different terms in the equation system since the definition of consistency implicitly assumes the solutions are smooth. For linear problems with constant coefficients, the consistency of the scheme can be studied by taking the Fourier transform of the difference equations. In the case of finite or spectral elements, approximation theory can be used to assess how well the continuous operators are approximated. For example, for cubic splines, an error of $O(\Delta x^4)$ holds for the function approximation. The first derivative approximation has an error of $O(\Delta x^3)$, and the error is $O(\Delta x^{4-n})$ for the n^{th} derivative $(n < 4)$.

Definition 10. *Suppose v is an exact solution of the continuous problem $Pu = 0$. The quantity $\tau = (P - P_\Delta)v$ is called the* local truncation error *or* discretization error*. It measures how well the discrete operator approximates the exact operator P.*

The truncation error $\tau = O(\Delta t^p, \Delta x^q)$ $(p > 0,\ q > 0)$ indicates that the error is small for small Δt and Δx. Let U_Δ be a solution of the discrete problem $P_\Delta(U_\Delta) = 0$, then the amount by which $R(U_\Delta) = P(U_\Delta)$ deviates from zero is called the *residual.* The residual $R(U_\Delta) = O(\Delta t^p, \Delta x^q)$ measures how well the approximate solution U_Δ satisfies the exact equation $Pu = 0$.

Example

Consider the heat equation:

$$P[u] = u_t - \kappa u_{xx} = 0, \tag{3.1}$$

and the explicit finite difference scheme:

$$P_\Delta[u] = \frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{\kappa}{(\Delta x)^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) = 0, \tag{3.2}$$

Here κ is assumed to be constant and u_j^n is the solution of the difference equation at the grid point $x_j = j\Delta x$ at time $t_n = n\Delta t$ (we assume $0 \leq x \leq L$ is the domain for x). In general the solution of the difference equation (3.2) will differ from the exact solution $u = U(x,t)$, by the error $e_j^n = U_j^n - u_j^n$ where $U_j^n = U(j\Delta x, n\Delta t)$ is the exact solution at the grid point. The truncation error $\tau = (P - P_\Delta)[U]$ is obtained by replacing u_j^n by U_j^n in (3.2) and expanding $P_\Delta[U]$ in Taylor series about (x_j, t_n). This gives the truncation error τ as:

$$\tau = (P - P_\Delta)[U] = \kappa\frac{(\Delta x)^2}{12}U_{4x} - \frac{\Delta t}{2}U_{tt} + O\left[(\Delta t)^2, (\Delta x)^4\right].$$

Since $U_t = \kappa U_{xx}$, the truncation error can also be written as:

$$\tau = \kappa\frac{(\Delta x)^2}{12}\left(1 - \frac{6\kappa\Delta t}{(\Delta x)^2}\right)U_{4x} + O\left[(\Delta t)^2, (\Delta x)^4\right]. \tag{3.3}$$

Thus the truncation error $\tau = O[(\Delta x)^2, \Delta t]$ in this case.

Stability

The essence of stability is that there should be some limit to the extent to which any component of the initial data can be amplified in the numerical solution of $P_\Delta[u] = 0$. Consider for example the explicit finite difference scheme (3.2) for the heat equation. The requirement that the solution scheme be stable requires that $|u^n|$ should be bounded in some sense. Writing (3.2) in the form $u^{n+1} = C(\Delta t)u^n$, the formal solution of the difference scheme is $u^n = C(\Delta t)^n u^0$. The approximation operator $C(\Delta t)$ is said to be *stable* if for some fixed but arbitrary τ with $0 < \Delta t < \tau < T$ ($N\Delta t = T$, is the largest time considered in the solution), the infinite set of operators $\{C(\Delta t)^n\}$ is uniformly bounded (i.e. $\|C^n\| < K(T)$ where $K(T)$ is independent of n) (see e.g. [151, Section 3.4, p. 45]).

A simple Fourier method (von Neumann analysis) that can be used to determine the stability of (3.2) in the present case, is to note that the difference scheme possesses solutions of the form $u_j^n = A_m \xi^n \exp(imj\Delta x)$. Substituting this solution ansatz in (3.2) we find

$$\xi = \xi(m) = 1 - \frac{4\kappa\Delta t}{(\Delta x)^2}\sin^2\left(\frac{m\Delta x}{2}\right). \tag{3.4}$$

For stability, we require $|\xi| \leq 1$ (if $|\xi| > 1$ then $|\xi|^n \to \infty$ as $n \to \infty$: note $N\Delta t = T$, implies $\Delta t \to 0$ as $N \to \infty$). From (3.4), $\xi(m) \leq 1$. Hence we require $\xi(m) \geq -1$ for stability. This latter constraint is satisfied by choosing the grid spacing so that $0 < \sigma = \kappa\Delta t/(\Delta x)^2 \leq \frac{1}{2}$. If $\sigma > \frac{1}{2}$, the numerical difference scheme (3.2) is unstable.

The condition $\sigma < \frac{1}{2}$, i.e. $\kappa\Delta t/(\Delta x)^2 < \frac{1}{2}$, imposes severe constraints on the time step Δt if Δx is small. This can be overcome in the present case by using implicit, or partially implicit, differencing for the spatial derivative in (3.1). Consider the semi-implicit finite difference scheme:

$$\begin{aligned} u_j^{n+1} - u_j^n = {} & \frac{\kappa\Delta t}{(\Delta x)^2}\left[\Theta\left(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}\right)\right] \\ & + \frac{\kappa\Delta t}{(\Delta x)^2}\left[(1-\Theta)\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right)\right], \end{aligned} \tag{3.5}$$

where $0 \leq \Theta \leq 1$. Setting $u_j^n = A\xi^n \exp(imj\Delta x)$ as the solution ansatz in (3.5) we find

$$\xi = \xi(m) = \frac{1 - 4\sigma(1-\Theta)\sin^2(\frac{1}{2}m\Delta x)}{1 + 4\sigma\Theta\sin^2(\frac{1}{2}m\Delta x)}. \tag{3.6}$$

From (3.6) we find that for $0 \leq \Theta < \frac{1}{2}$, the stability constraint $|\xi| \leq 1$ is satisfied if $\sigma \leq 1/[2(1-2\Theta)]$. If $\Theta \geq \frac{1}{2}$ the scheme is stable for all $\sigma > 0$. For $\Theta = 0$, the scheme (3.5) reduces to the fully explicit scheme (3.2); for $\Theta = 1$ the scheme is fully implicit; and for $\Theta = \frac{1}{2}$ one obtains the Crank-Nicolson scheme.

Consider the linear problem:

$$P(U) = U_t - A(U, U_x, U_{xx}, \ldots) = 0, \tag{3.7}$$

subject to appropriate (i.e. well-posed) initial and boundary conditions, and the corresponding difference equation approximation:

$$P_\Delta(u) = \frac{u_j^{n+1} - u_j^n}{\Delta t} - A_\Delta(u) = 0, \tag{3.8}$$

where the operator A_Δ is the discretized form of the linear operator A. In general A_Δ will be a matrix since the difference scheme will involve the values of u_j^n at neighboring grid points.

The convergence of the difference equation solution u to the exact solution U depends on the discretization steps Δt and Δx, and can be monitored by following the evolution of the discretization error $e = U - u$. Since the operators are linear $P_\Delta(u) = P_\Delta(U - e) = P_\Delta(U) - P_\Delta(e)$. Using this fact and subtracting (3.8) from (3.7) then gives the equation:

$$P_\Delta(e) + \tau_\Delta \equiv \frac{e^{n+1} - e^n}{\Delta t} - A_\Delta(e) + \tau_\Delta = 0, \tag{3.9}$$

for the error e where $\tau_\Delta = P(U) - P_\Delta(U)$ is the truncation error. Note that $\tau_\Delta = -P_\Delta(U)$ if U is an exact solution of the continuous equation $P(U) = 0$. Equation (3.9) is a difference equation relating e^{n+1} and e^n. Alternatively, using $U = u + e$ and the linearity of P, we find $P(U) - P_\Delta(u) = P(u) + P(e) - P_\Delta(u) = 0$, which can be re-written as a continuous evolution equation for the error e:

$$P(e) \equiv e_t - A(e) = \tau, \tag{3.10}$$

where $\tau = P_\Delta(u) - P(u)$ is the truncation error. Note that $\tau = -P(u)$ if u is an exact solution of the difference equation $P_\Delta(u) = 0$. The truncation terms τ and τ_Δ in the above equations are in general functions of space and time. The formal solution of (3.10) for e is

$$e(t) = \exp\left(\int_0^t A\ dt\right) \int_0^t \exp\left(-\int_0^{t'} A\ dt''\right) \tau(t')dt' \\ + \exp\left(\int_0^t A\ dt\right) e(0). \tag{3.11}$$

Assume $u = u_\Delta = U - e$ is uniformly bounded in some norm, in a Banach space $\mathcal{B}$, i.e.

$$\|u_\Delta\| \le C(T)\|u_\Delta^0\|,$$

with constant $C(T)$, independent of the initial data and discretization parameters for any t, $0 \le t \le T$. This is called the stability, or well-posedness, of the discrete problem. Taking the norm of (3.11) leads to the key estimate for linear convergence:

$$\|e(t)\| \le e^{T\|A\|}\|e(0)\| + \frac{e^{T\|A\|} - 1}{\|A\|}\|\tau\|,$$

where $\|\cdot\|$ represents the L_2 or L_∞ norm, or some other appropriate norm, in which we can study the stability of the initial-boundary value problem.

This statement is a part of the Lax-Richtmyer equivalence theorem for linear partial differential equations (PDEs) and the Dahlquist theorem for linear multistep schemes for linear and nonlinear ordinary differential equations (ODEs), to be discussed in the later sections of this chapter. These theorems are related to a theorem due to [97] stating the equivalence of stability and convergence. In this theorem, a sequence of linear operators $\{T_m\}$ is used to approximate the linear operator equation $Tu = f$, where $T :\ \mathcal{B}' \to \mathcal{B}''$ and $T_m :\ \mathcal{B}' \to \mathcal{B}''$ are linear operators mapping the Banach space $\mathcal{B}'$ with norm $\|.\|'$ to the Banach space $\mathcal{B}''$, with norm $\|.\|''$ [151, Section 2.6, p. 37]. In the present context, the sequence of operators $\{T_m\}$ can be thought of as a sequence of difference operators, with finer and finer discretization steps, as m increases. If (1) $Tu = f$ has a unique solution, and (2) if T_m has a bounded inverse T_m^{-1} and $\|Tu - T_m u\| \to 0$ as $m \to \infty$, then the sequence of solutions $\{u_m\}$ of $T_m u_m = f$ converges to the solution u of $Tu = f$ *if and only if* the approximations are stable (i.e. the inverse operators $\{T_m^{-1}\}$ are uniformly bounded). These theorems effectively state that for a consistent approximation, the stability (well-posedness) of the discrete problem is equivalent to the convergence of the approximate solutions to the exact solution. This is a very useful piece of information in practice. Similar results for nonlinear hyperbolic and parabolic equations are described by Strang's theorem for smooth solutions. For other types of nonlinear equations, the notion of the equivalence of stability and convergence is often more in the realm of folklore than a rigorous theory. In practice, many other additional considerations, like fixed discretization, interfaces, boundaries, singularities, resonances, different-time scales, etc. come into play. These issues will be addressed in later chapters.

Examples

Model equation and local truncation error notions. Consider the advection equation:

$$u_t + au_x = 0, \tag{3.12}$$

where a is the constant advection velocity.

Leap-frog Scheme

The leap-frog scheme for (3.12) is:

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0. \tag{3.13}$$

The difference equation (3.13) can also be written in the form:

$$u_j^{n+1} = u_j^{n-1} - C\left(u_{j+1}^n - u_{j-1}^n\right), \tag{3.14}$$

where $C = a\Delta t/\Delta x$ is the Courant-Friedrichs-Levy (CFL) number. This non-dimensional number is the ratio of the physical and numerical propagation speeds and always appears in the stability restriction on the size of the time step in explicit schemes for hyperbolic PDEs. Substituting the exact solution $u(x,t)$ of (3.12) into the difference equation (3.12) we obtain the modified advection equation:

$$u_t + au_x = \tau, \tag{3.15}$$

where

$$\begin{aligned} \tau &= -\frac{1}{3!}\left(\Delta t^2 u_{ttt} + a\Delta x^2 u_{xxx}\right) + O\left(\Delta t^4, \Delta x^4\right), \\ &\equiv -\frac{a\Delta x^2}{3!}(1 - C^2)u_{xxx} + \cdots \end{aligned} \tag{3.16}$$

is the truncation error. Using von Neumann analysis (Section 3.3) one can show that the leap-frog scheme is stable if $|C| < 1$.

The Lax-Friedrichs Scheme

The finite difference approximation:

$$P_\Delta(u) = \frac{u_j^{n+1} - \frac{1}{2}(u_{j+1}^n + u_{j-1}^n)}{\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0, \tag{3.17}$$

for the advection equation (3.12) is known as the Lax-Friedrichs scheme. It can also be written in the form:

$$u_j^{n+1} = \frac{1}{2}\left(u_{j+1}^n + u_{j-1}^n\right) - \frac{C}{2}\left(u_{j+1}^n - u_{j-1}^n\right), \tag{3.18}$$

where $C = a\Delta t/\Delta x$ is the CFL number. Expanding (3.17) in Taylor series about $(x,t) = (x_j, t_n)$ gives the modified advection equation (3.15) with truncation error:

$$\tau = \frac{\Delta x^2}{2\Delta t}(1 - C^2)u_{xx} + \frac{a(C^2 - 1)}{3!}\Delta x^2 u_{xxx} + \cdots. \tag{3.19}$$

Thus the Lax-Friedrichs scheme is only consistent with the advection equation if $\Delta x^2/\Delta t \to 0$ as both $\Delta x \to 0$ and $\Delta t \to 0$. If $\Delta x^2/\Delta t \to$ *const.* as $\Delta x, \Delta t \to 0$, then the scheme is inconsistent, since the difference equation then converges to the advection-diffusion equation, with diffusion coefficient $\kappa = \Delta x^2/(2\Delta t)$, rather than to the advection equation (3.12). However, if $\Delta t \propto \Delta x$ (i.e. the CFL number $C = a\Delta t/\Delta x =$ *const.*), then the scheme is consistent with the original equation (3.12), and the scheme converges to the original equation if $|C| < 1$.

The Central Difference Scheme

The difference scheme:

$$P_\Delta(u) = \frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0, \tag{3.20}$$

for the advection equation (3.12) is known as the central difference scheme. Expanding (3.20) in Taylor series about $(x,t) = (x_j, t_n)$ gives the modified advection equation (3.15) with truncation term:

$$\tau = -\frac{a^2\Delta t}{2}u_{xx} - \frac{a\Delta x^2}{3!}(1 - C^2)u_{xxx} + O(\Delta t^3, \Delta x^4). \tag{3.21}$$

Thus as $\Delta t \to 0$ and $\Delta x \to 0$, the central difference scheme (3.20) is consistent with the advection equation (3.12). However, the negative diffusion term $-(a^2\Delta t/2)u_{xx}$ in (3.21) hints that the scheme is unstable, since the backward heat equation is ill-posed (see Chapter 1). The rigorous proof of the instability will be discussed in the next chapter and follows from the dispersion relation. The dispersion relation arises in the context of Fourier analysis or plane wave decomposition of the solutions of the linear problem. In the context of difference schemes it is called von Neumann analysis and will be discussed in the next section.

In practice, the check of consistency involves the Taylor expansion of the difference equation to check that it is consistent with the original equation in the limit of $\Delta t \to 0$ and $\Delta x \to 0$, as well as a check of the program code of the scheme. Errors in the coefficients, boundary conditions and sources often produce convergent solutions to the equations with wrong wave speeds, profiles, etc. Thus the validation of the code with a few special analytical solutions may help to weed out such errors. Finally, for most problems the convergence proofs for the corresponding schemes are not available in practice. Monitoring known invariants, like energy, momentum and symmetries, becomes a necessary check of accuracy and convergence. However, these checks might not still be sufficient to guarantee that the correct solution has been obtained.

3.2 Lax-Richtmyer Equivalence Theorem

The theorem states that for a linear well-posed problem approximated by a linear consistent discrete approximation, stability (well-posedness of the discrete problem) is equivalent to convergence. Thus for linear problems the verification of convergence consists of checking that the continuous and discrete problems are consistent and well-posed.

Consider the linear initial value problem:

$$U_t = LU, \tag{3.22}$$

where L is a linear spatial differential operator that is approximated by a discrete operator L_Δ together with a time discretization that results in a discrete problem:

$$U^{n+1} = \mathbf{Q}U^n = \mathbf{Q}^n U^0, \tag{3.23}$$

where $\mathbf{Q}$ is a matrix representation of the discrete approximation.

Some of the main ideas implicit in the formulation of the equivalence theorem (e.g. [151, Ch. 3]) are discussed below.

A *genuine* solution of the initial value problem (3.22) is a one parameter family $U = U(t)$ in the parameter t, such that:

(1) $U(t) \in \mathcal{D}_L$ where $\mathcal{D}_L$ is the domain of the operator L for each $t \in [0, T]$ where T is a fixed constant value of $t > 0$.
(2) Using the usual definition of the derivative of $U(t)$ we require:

$$\left\| \frac{U(t+\Delta t) - U(t)}{\Delta t} - LU(t) \right\| \to 0, \tag{3.24}$$

as $\Delta t \to 0$ $(0 \le t \le T)$.

The problem (3.22) is *well-posed* if (a) the family of *genuine* solutions is sufficiently large and (b) the solutions depend *uniquely and continuously on the initial data.*

Not all initial data U_0 will give rise to a *genuine* solution. Let $\mathcal{D}$ be the set of U_0 in the Banach space $\mathcal{B}$ in which the problem is posed, for which there is a unique genuine solution $U(t)$ such that $U(0) = U_0$ (a Banach space is a complete, normed linear space: see, e.g. [167]). The correspondence or map $U_0 \to U(t)$ is linear. We write

$$U(t) = E_0(t)U_0 \quad \forall\ U_0 \in \mathcal{D}, \tag{3.25}$$

where $E_0(t)$ is the solution operator.

The initial value problem (3.22) is called *properly posed* (a notion originally due to Hadamard) if

(1) The domain $\mathcal{D}$ of $E_0(t)$ is *dense* in $\mathcal{B}$. This effectively means that any element $u \in \mathcal{B}$ can be approximated as the limit of a convergent sequence $\{u_n\}$ where $u_n \in \mathcal{D}$.
(2) The family of operators $E_0(t)$ for different t is *uniformly bounded* such that $\|E_0(t)\| < K$ for $0 \le t \le T$.

Point (2) in the above definition of a properly posed problem effectively says that the solution depends continuously on the initial data. Thus if $U(t)$ and $V(t)$ are *genuine* solutions corresponding to initial data U_0 and V_0, then $U(t) = E_0(t)U_0$ and $V(t) = E_0(t)V_0$ are such that

$$\|U(t) - V(t)\| = \|E_0(t)(U_0 - V_0)\| \le K\|U_0 - V_0\|. \tag{3.26}$$

Thus if U_0 and V_0 are nearly equal, then the solutions $U(t)$ and $V(t)$ at time t are also very nearly equal where $0 \le t \le T$. As noted in Section 3.1, the concept of a *uniformly bounded* operator $E_0(t)$ for $0 \le t \le T$ is equivalent to that of a *stable* operator.

The gist of point (1) in the above definition of a properly posed problem is that even although a *genuine solution* might not exist for given initial data U_0, we can still approximate U_0 as closely as we wish by initial data for which a *genuine solution* does exist. For example, in the solution of the heat equation $T_t = T_{xx}$, the initial temperature distribution $T(x, 0)$ can be discontinuous, but we can approximate the initial data by a sequence of continuous, twice differentiable functions $T_n(x, 0)$ that approach the discontinuous initial data in the limit as $n \to \infty$ (this is effectively what is done in the Fourier series solution of the heat equation). The corresponding solution $T_n(x, t)$ approaches a function $T(x, t)$ as $n \to \infty$, which is interpreted as the solution of the heat equation for discontinuous initial data.

For an initial boundary value problem in a bounded domain to be *properly posed* the boundary conditions must be appropriate, since the domain of L is defined in part by the boundary conditions. For example, if too many boundary conditions are specified in the solution of the heat equation $T_t = T_{xx}$, e.g. $T = T_x = 0$ at $x = a$ and $x = b$ for all t, then there are no *genuine* solutions at all, unless the initial data $T(x, 0) \equiv 0$. In this case the domain $\mathcal{D}$ of the operator ∂_{xx} is not *dense* in $\mathcal{B}$. Similarly, if there are too few boundary conditions (e.g. none), then there is no unique solution of (3.22) for any initial data $U(x, 0) = U_0(x)$, and again the domain $\mathcal{D}$ of L is not dense in $\mathcal{B}$.

By using the *extension theorem* [151, Section 2.4], if $E_0(t)$ is a bounded linear operator with domain $\mathcal{D}$ *dense* in $\mathcal{B}$, then it has a unique *extension* $E(t)$ whose domain is the whole Banach space $\mathcal{B}$, such that the action of $E(t)$ coincides with that of $E_0(t)$ on the domain $\mathcal{D}$, and $\|E(t)\| = \|E_0(t)\|$. Thus $U(t) = E(t)U_0$ is interpreted as the *generalized solution* corresponding to the initial data $U_0 \in \mathcal{B}$. If L does not explicitly depend on t, then $E(s+t) = E(s)E(t)$ for $s \geq 0, t \geq 0$. This is called the *semi-group property*; it may be used to simplify proofs when it is applicable. For cases where L depends on t, E depends on both t and the initial time t_0, and the analog of the semi-group property is then $E(t; t_0) = E(t; t_1)E(t_1; t_0)$. In the case where the semi-group property holds one can prove that $E(t)$ and L commute, i.e. $E(t)LU_0 = LE(t)U_0$. Using these results, it follows that for a *properly posed* problem a *genuine* solution is continuous, and that the convergence in (3.24) is *uniform*.

The concepts of *stability* and *convergence* for the difference equation (3.23) implicitly assume that we can carry out an infinite sequence of calculations with increasingly finer meshes in which the discretization steps $\Delta x_i = g_i(t)$, $i = 1, \ldots, d(t, x_1, \ldots, x_d)$ are the independent variables) are specified as functions of t, such that $\Delta x_i \to 0$ as $\Delta t \to 0$. Effectively, one introduces a sequence of time intervals $\{\Delta_j t\}$ and integers n_j such that $\Delta_j t \to 0$ as $j \to \infty$ and $n_j \Delta_j t \to t$ as $j \to \infty$.

Proof of the Lax-Richtmyer Equivalence Theorem

The proof proceeds by first establishing the theorem is true for *genuine* solutions $U = U(t)$ of (3.22), for which U_0 and $U(t) \in \mathcal{D}_L$, where $\mathcal{D}_L$ is the domain of the operator L, and then extending the proof to cases where $U_0, U(t) \in \mathcal{B}$, but $U_0, U(t) \notin \mathcal{D}_L$.

To establish that *stability* implies *convergence*, first note that (3.22) and (3.23) can be written as:

$$P(U) \equiv U_t - LU = 0 \tag{3.27}$$

and

$$P_\Delta(u)^n \equiv \frac{u^{n+1} - Qu^n}{\Delta t} = 0. \tag{3.28}$$

The *truncation error* τ_n, obtained by substituting the exact solution $U = U(t,x)$ of (3.27) in the difference equation (3.28), is given by $\tau_n = P(U^n) - P_\Delta(U^n) \equiv -P_\Delta(U^n)$. From this latter equation, it follows that

$$U^{n+1} = \mathbf{Q}U^n - \tau_n \Delta t. \tag{3.29}$$

For a *consistent* solution of the difference equation (3.28) we require that the truncation error $\tau_n \to 0$ as $\Delta t \to 0$.

The solution of the difference equation (3.28) will be a *convergent* solution if $u^n \to U$ as $n \to \infty$ and $\Delta t \to 0$ such that $n\Delta t \to t \in [0, T]$. Alternatively, a convergent solution is obtained if the norm of the error $\|e^n\| = \|U^n - u^n\| \to 0$ as $n \to \infty$, $\Delta t \to 0$ and $n\Delta t \to t$. Using (3.28) and (3.29) it follows that the error e^n satisfies the difference equation:

$$e^{n+1} = \mathbf{Q}e^n - \tau_n \Delta t. \tag{3.30}$$

Iterating (3.30) gives the equation:

$$e^{n+1} = \mathbf{Q}^{n+1} e^0 - \Delta t \left(\tau_n + \mathbf{Q}\tau_{n-1} + \mathbf{Q}^2 \tau_{n-2} \ldots + \mathbf{Q}^n \tau_0 \right), \tag{3.31}$$

for the error at time $t_{n+1} = (n+1)\Delta t$ in terms of the error e^0 at time $t = 0$ and in terms of the truncation errors $\{\tau_j : \ 0 \leq j \leq n\}$.

For a *consistent* solution of the difference equation (3.28), the truncation errors $\tau_n \to 0$ in the limit as $\Delta t \to 0$. Thus for a given $\tau > 0$, the truncation errors will have norm $\|\tau_j\| < \tau$ provided Δt is small enough. Using the *stability* of the numerical approximation, it follows that $\|U^n\| \leq C(T)\|U^0\|$ uniformly with respect to n (i.e. $C(T)$ is independent of n) for all $t \in [0, T]$. Alternatively, the matrices Q^n are uniformly bounded, meaning $\|\mathbf{Q}^n U\| \leq C(T)\|U\|$ for all n.

Taking the norm of (3.31) and using the triangle inequality we obtain the equation:

$$\|e^{n+1}\| \leq \|\mathbf{Q}^n\| \, \|e^0\| + \Delta t \, \tau \, \left(\|\mathbf{Q}^0\| + \|\mathbf{Q}^1\| + \cdots + \|\mathbf{Q}^n\|\right), \tag{3.32}$$

as an upper bound for $\|e^{n+1}\|$.

Using stability and noting $t_{n+1} = (n+1)\Delta t \leq T$, we obtain the estimate

$$\|e^{n+1}\| \leq C(T) \, \|e^0\| + T \, C(T) \, \tau. \tag{3.33}$$

The key estimate (3.33) shows that $\|e^{n+1}\| \to 0$ in the ideal limit, that the initial error $\|e^0\| = \|U_0 - u_0\|$ is made arbitrarily small, and the upper bound to the truncation error, τ, is made arbitrarily small. For a consistent approximation, the truncation error bound $\tau \to 0$ as $\Delta t \to 0$. Strictly speaking, the above argument only applies to the error at the discrete time $t_n = n\Delta t$. However, the error at time t, $e(t) = U(t) - u^n$, has a norm $\|e(t)\|$ which can be bounded by using the triangle inequality:

$$\|e(t)\| \equiv \|U(t) - U(n\Delta t) + U^n - u^n\| \leq \|U(t) - U(n\Delta t)\| + \|e^n\|, \tag{3.34}$$

where $n\Delta t \approx t$ ($n\Delta t \to t$ as $\Delta t \to 0$ and $n \to \infty$). For a *properly posed* problem $U(t)$ is continuous so that $\|U(t) - U(n\Delta t)\| \to 0$ in the above limit. Similarly, from (3.33) $\|e^n\| \to 0$ in the same limit. This proves that *stability and consistency* of the difference approximation implies *convergence* for the class of initial data in the domain of L (i.e. $U_0 \in \mathcal{D}_L$).

For problems in which the initial data $U_0 \notin \mathcal{D}_L$, one uses the assumption that $\mathcal{D}_L$ is dense in $\mathcal{B}$, which means that we can approximate U_0 arbitrarily closely by a sequence $\{U_0^{(j)}\}$ where $U_0^{(j)} \in \mathcal{D}_L$ (i.e. $U_0^{(j)} \to U_0$ as $j \to \infty$). In this case,

$$\begin{aligned}\|e(t)\| = \|&U(t) - U(n\Delta t) + U^n - U^{(j)n} + U^{(j)n} \\ &- u^{(j)n} + u^{(j)n} - u^n\| \leq \|U(t) - U(n\Delta t)\| \\ &+ \|U^n - U^{(j)n}\| + \|U^{(j)n} - u^{(j)n}\| + \|u^{(j)n} - u^n\|.\end{aligned} \tag{3.35}$$

Noting that $\|U(t) - U(n\Delta t)\| \to 0$ in the limit as $\Delta t \to 0$ and $n\Delta t \to t$, and using the estimates:

$$\begin{aligned}
&\|U^n - U^{(j)n}\| = \|E(n\Delta t)(U^0 - U^{(j)0})\| \leq \|E\| \, \|U^0 - U^{(j)0}\|,\\
&\|u^{(j)n} - u^n\| = \left\|\mathbf{Q}^n \left(u^{(j)0} - u^0\right)\right\| \leq C(T)\|u^{(j)0} - u^0\|,\\
&\|U^{(j)n} - u^{(j)n}\| = \|e^n\|, \qquad (3.36)
\end{aligned}$$

it follows that $\|e(t)\| \to 0$ in the limit as $j \to \infty$, $\Delta t \to 0$ and $n\Delta t \to t$.

This completes the proof that *stability and consistency* of the difference approximation for a well-posed problem implies *convergence.*

The norm in the global error estimate (3.33) is the same as the norm in which it is possible to establish the stability and consistency of the numerical method. Then convergence is assured if the initial error e^0 and the truncation error bound τ tend to zero in the vanishing limit of space-time discretization. The order of accuracy turns out to be the same as the order of consistency. In other words the error in the solution is the same as the size of the perturbation due to discretization in the original equation. Note that convergence is still guaranteed even if $\|\mathbf{Q}^n\|$ is weakly unstable (algebraic growth with respect to Δt and Δx), as long as the truncation error decays faster. The estimate (3.33) also shows that any error introduced due to under-resolution, singularities, interfaces, etc., at any time in the computation, may be viewed as an error in the initial data and will contribute to the global error.

To prove that *convergence* implies *stability*, we assume, on the contrary, that *convergence* implies *instability*, and obtain a contradiction, which implies that the original proposition was correct. If the numerical scheme is unstable then there is a mode for which $\|\mathbf{Q}^n u_0\|$ grows without bound as $n \to \infty$, $\Delta t \to 0$ and $n\Delta t \to t$. On the other hand, since the numerical approximation is convergent, the norm of the error $\|e^n\| = \|U^n - u^n\| \to 0$ as $n \to \infty$, i.e. given $\epsilon > 0$ there exists an integer N and positive number δ such that $\|U^n - u^n\| < \epsilon$ whenever $n > N$ and $\Delta t < \delta$. Since $\|U^n - u^n\| \geq |\|U^n\| - \|u^n\||$ it follows that $\|u^n\| - \epsilon < \|U^n\| < \|u^n\| + \epsilon$, whenever $n > N$. Since the numerical scheme is unstable, this implies that $\|u^n\| \to \infty$ as $n \to \infty$, and hence $\|U^n\| \to \infty$ also in the same limit. But this is a contradiction, since the initial value problem for $U(t, x)$ is well-posed, i.e. $\|U^n\| < C(T)\|U^0\|$. Hence the original assumption was incorrect. Thus *convergence* implies *stability.* This completes the proof of the theorem.

Useful Norms

In the analysis of stability, the l_p-norms (e.g. [167, pp. 88-90, 218-219]) are a useful class of norms which apply to a finite dimensional vector space.

These norms are defined as:

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^{m} |x_j|^p \right)^{1/p}, \quad (p \geq 1), \tag{3.37}$$

where the x_j denote the components of the vector $\mathbf{x}$ (note that it is necessary to choose $p \geq 1$ in order that $\|\mathbf{x}\|_p$ satisfies the definition of a norm). Some useful inequalities associated with this norm are given below. The inequality

$$\sum_{j=1}^{m} |x_j|\ |y_j| \leq \|\mathbf{x}\|_p\ \|\mathbf{y}\|_q \quad \text{where} \quad \frac{1}{p} + \frac{1}{q} = 1, \tag{3.38}$$

is known as Hölder's inequality. In the case $p = q = 2$ Hölder's inequality is known as Cauchy's inequality. The inequality

$$|(\mathbf{x}, \mathbf{y})| \leq \|\mathbf{x}\|\ \|\mathbf{y}\|, \tag{3.39}$$

is the Schwarz inequality, where $(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{m} x_j y_j^*$ defines the inner product of $\mathbf{x}$ and $\mathbf{y}$. The triangle inequality,

$$\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p, \tag{3.40}$$

(also known as Minkowski's inequality) is proved by using Hölder's inequality. The sup-norm

$$\|\mathbf{x}\| = \sup_{1 \leq j \leq m} |x_j| \tag{3.41}$$

has the same effect as taking the l_∞-norm obtained by letting $p \to \infty$ in (3.37).

The above norms have continuous analogs. For example,

$$\|f\|_p = \int_a^b |f(x)|^p dx, \tag{3.42}$$

defines the L_p-norm for f on $[a, b]$ where $f(x) \in L_p[a, b]$ and $L_p[a, b]$ denotes the p^{th} class of Lebesgue integrable functions, for which the integrals (3.42) are well defined. Similarly,

$$\|f\| = \sup\{|f(x)| : \ x \in [a, b]\} \tag{3.43}$$

defines the *sup*-norm for bounded functions $f(x)$ on $[a, b]$.

Next, we examine several examples of stability/instability that are established directly from the definition of stability.

Example

Consider the advection equation:

$$u_t + au_x = 0, \quad (a > 0), \tag{3.44}$$

approximated by an upwind difference scheme:

$$u_j^{n+1} = (1-\lambda)u_j^n + \lambda u_{j-1}^n, \quad (0 \leq \lambda \leq 1), \tag{3.45}$$

where $\lambda = a\Delta t/\Delta x$ is the CFL number (see also (2.9) in Section 2.1). Using the l_∞-norm, and the triangle inequality we find

$$\|u^{n+1}\| \leq (1-\lambda)\|u^n\| + \lambda\|u^n\| \equiv \|u^n\|. \tag{3.46}$$

Thus $\|u^{n+1}\| \leq \|u^n\|$, which implies $\|u^n\| \leq \|u^0\|$. Hence, the scheme is stable in the l_∞-norm for $0 < \lambda \leq 1$. The Lax-Richtmyer equivalence theorem then implies the scheme is convergent for $0 < \lambda \leq 1$.

Example

Consider the advection-diffusion equation:

$$u_t + au_x = \epsilon u_{xx}, \quad a > 0,\ \epsilon > 0, \tag{3.47}$$

approximated by the difference scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_j^n - u_{j-1}^n}{\Delta x} = \epsilon\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}. \tag{3.48}$$

This is a three-point scheme, with explicit upwind differencing for the advection term and a centered difference approximation for the diffusion term. The scheme can also be written in the form:

$$u_j^{n+1} = (\sigma+\lambda)u_{j-1}^n + (1-(2\sigma+\lambda))u_j^n + \sigma u_{j+1}^n, \tag{3.49}$$

where

$$\lambda = a\frac{\Delta t}{\Delta x}, \quad \sigma = \epsilon\frac{\Delta t}{\Delta x^2}, \tag{3.50}$$

and we restrict λ to lie in the range $0 < \lambda \leq 1$. Note that the updated value in (3.49) is a convex combination of the values at the previous time level.

As in the previous example, use of the triangle inequality for $0 < \lambda \leq 1$ and $0 < 2\sigma + \lambda \leq 1$, gives the estimate:

$$\|u^{n+1}\| \leq (\sigma + \lambda)\|u^n\| + (1 - \lambda - 2\sigma)\|u^n\| + \sigma\|u^n\| \equiv \|u^n\|. \tag{3.51}$$

Thus again $\|u^{n+1}\| \leq \|u^n\|$ and hence $\|u^n\| \leq \|u^0\|$, and the scheme is stable for $0 < \lambda \leq 1$ and $0 < 2\sigma + \lambda \leq 1$. In the next section we use Fourier analysis to establish that this condition is also necessary for l_2 stability.

Example

Consider the diffusion equation [174]:

$$u_t = \epsilon u_{xx}, \quad (\epsilon > 0), \tag{3.52}$$

approximated by the explicit, centered finite difference scheme:

$$u_j^{n+1} = \sigma u_{j-1}^n + (1 - 2\sigma)u_j^n + \sigma u_{j+1}^n, \tag{3.53}$$

where $\sigma = \epsilon\Delta t/\Delta x^2$, subject to the initial condition:

$$u_j^0 = \delta_{j,j'} \tag{3.54}$$

where $\delta_{j,j'}$ is the Kronecker delta symbol. Consider the case where $\sigma > \frac{1}{2}$. Previous discussion of the difference scheme (3.52) based on von Neumann analysis (Fourier modal analysis) in (3.4) et seq. showed that the scheme (3.53) was unstable for $\sigma > \frac{1}{2}$. Below we study more exactly the manner in which the scheme develops an instability for the initial data (3.54).

Using the initial data (3.54) in (3.53) we find

$$u_j^1 = \sigma\left(\delta_{j,j'+1} + \delta_{j,j'-1}\right) + (1 - 2\sigma)\delta_{j,j'}. \tag{3.55}$$

Thus $u_j^1 \neq 0$ for $j = j' - 1$, j' and $j' + 1$. Since $\sigma > \frac{1}{2}$, u_j^1 oscillates in sign as j increases from $j = j' - 1$ to $j = j' + 1$. Similarly, we find

$$\begin{aligned} u_j^2 = {} & \sigma^2\left(\delta_{j,j'-2} + \delta_{j,j'+2}\right) + [2\sigma^2 + (1 - 2\sigma)^2]\delta_{j,j'} \\ & + 2\sigma(1 - 2\sigma)\left(\delta_{j,j'-1} + \delta_{j,j'+1}\right). \end{aligned} \tag{3.56}$$

Thus $u_j^2 \neq 0$ for $j' - 2 \leq j \leq j' + 2$, and oscillates in sign as j increases from $j = j' - 2$ to $j = j' + 2$. Generalizing these results (by induction), we find that u_j^n is non-zero for the $2n+1$ values of j in the range $j' - n \leq j \leq j' + n$, and that u_j^n oscillates in sign as j increases from $j = j' - n$ to $j = j' + n$.

Taking into account the oscillations in the u_j^n, we find that for $\sigma > \frac{1}{2}$:

$$\begin{aligned}|u_j^{n+1}| &= |\sigma u_{j-1}^n + (1-2\sigma)u_j^n + \sigma u_{j+1}^n| \\ &= \sigma|u_{j-1}^n| + (2\sigma-1)|u_j^n| + \sigma|u_{j+1}^n|.\end{aligned} \tag{3.57}$$

Summing over j in (3.57) we obtain

$$\|u^{n+1}\|_1 = (4\sigma-1)\|u^n\|_1 = (4\sigma-1)^{n+1}\|u^0\|_1, \tag{3.58}$$

for the l_1-norm $\|u^{n+1}\|_1$. Since $4\sigma - 1 > 1$, it follows that $\|u^n\|_1 \to \infty$ as $n \to \infty$, which implies instability with respect to the l_1-norm. Noting that the sum $S^n = \sum_{j=j'-n}^{j'+n} |u_j^n| \equiv \|u^n\|_1$ consists of a sum of $2n+1$ terms, then at least one of the terms in the sum must be greater than $S^n/(2n+1)$. Thus $\|u^n\|_\infty \equiv \sup |u_j^n| \geq \|u^n\|_1/(2n+1)$, which implies that $\|u^n\|_\infty \to \infty$ also as $n \to \infty$, and hence the scheme is also unstable in the sup-norm or l_∞-norm.

Example

Consider the Schrödinger equation:

$$iu_t = Hu = V(x)u - u_{xx}, \tag{3.59}$$

where $H = V(x) - \partial_x^2$ is the Hamiltonian operator. Using the Crank-Nicolson scheme (i.e. semi-implicit differencing for Hu), we obtain the difference scheme:

$$\left(1 + i\frac{\Delta t}{2}[V(x_j) - L_{xx}]\right) u_j^{n+1} = \left(1 - i\frac{\Delta t}{2}[V(x_j) - L_{xx}]\right) u_j^n, \tag{3.60}$$

where

$$L_{xx}(u_j^n) = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}, \tag{3.61}$$

is the central difference approximation for u_{xx} at $x_j = j\Delta x$ and $t_n = n\Delta t$. Equation (3.60) can be written in the matrix form:

$$\mathbf{P}u^{n+1} = \mathbf{P}^* u^n \quad \text{or} \quad u^{n+1} = \mathbf{Q}u^n, \tag{3.62}$$

where

$$\mathbf{Q} = \mathbf{P}^{-1}\mathbf{P}^*, \quad \mathbf{P} = \mathbf{I} + i\mathbf{A}, \tag{3.63}$$

and the matrix $\mathbf{A}$ has the symmetric, tridiagonal form:

$$\mathbf{A} = \begin{pmatrix} \alpha_1 & -\frac{\sigma}{2} & 0 & \dots & \dots & 0 \\ -\frac{\sigma}{2} & \alpha_2 & -\frac{\sigma}{2} & \dots & \dots & 0 \\ 0 & -\frac{\sigma}{2} & \alpha_3 & -\frac{\sigma}{2} & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & \dots & \dots & \dots & -\frac{\sigma}{2} & \alpha_{m-1} \end{pmatrix}. \tag{3.64}$$

In (3.62) it is assumed that $u_0^n = u_m^n = 0$. In (3.64), $\sigma = \Delta t/\Delta x^2$ and $\alpha_j = \frac{1}{2}\Delta t V(x_j) + \sigma$, $1 \leq j \leq m$, are the diagonal elements of $\mathbf{A}$.

In order to investigate the stability of the difference scheme $u^{n+1} = \mathbf{Q}u^n$, it suffices to determine the eigenvalues $\{\lambda_s : 1 \leq s \leq m-1\}$, or at least obtain a bound on the eigenvalues of the matrix $\mathbf{Q}$. Noting that $\mathbf{P}$ and $\mathbf{P}^{-1}$ are symmetric matrices (e.g. $\mathbf{P} = \mathbf{P}^T$), and that $\mathbf{P}^*\mathbf{P} = \mathbf{P}\mathbf{P}^* = \mathbf{I} + \mathbf{A}^2$, it is straightforward to verify that $\mathbf{Q}$ is a unitary matrix (i.e. $\mathbf{Q}\mathbf{Q}^{*T} = \mathbf{Q}^{*T}\mathbf{Q} = \mathbf{I}$). Since unitary matrices have complex eigenvalues with magnitude $|\lambda_s| = 1$ (e.g. [128, p. 319]) it follows that the eigenvalues of $\mathbf{Q}$ are complex and have unit magnitude, and hence $\|\mathbf{Q}\|_2 = 1$. Thus if we use the l_2-norm, we find:

$$\begin{aligned} \|u^{n+1}\|_2 &= (u^{n+1}, u^{n+1}) = (\mathbf{Q}u^n, \mathbf{Q}u^n) = (\mathbf{Q}^{*T}\mathbf{Q}u^n, u^n) \\ &= (u^n, u^n), \end{aligned} \tag{3.65}$$

where $(x, y) = \sum_{j=1}^m x_j y_j^*$ is the complex inner product. Thus $\|u^{n+1}\|_2 = \|u^n\|_2$, $\|u^n\|_2 = \|u^0\|_2$ and hence the scheme is stable with respect to the l_2-norm, which is similar to the result that the solutions of the continuous Schrödinger equation (3.59) are preserved with respect to the L_2-norm.

Example

Consider the diffusion equation:

$$u_t = u_{xx} \tag{3.66}$$

approximated by the Crank-Nicolson scheme:

$$\left(1 - \frac{\Delta t}{2} L_{xx}\right) u_j^{n+1} = \left(1 + \frac{\Delta t}{2} L_{xx}\right) u_j^n, \tag{3.67}$$

where $L_{xx}(u_j^n)$ is the central difference approximation (3.61) for u_{xx}. The difference scheme (3.67) can be written in the matrix form:

$$u^{n+1} = \mathbf{Q}u^n, \tag{3.68}$$

where

$$\mathbf{Q} = (2\mathbf{I} - \sigma\mathbf{T}_{m-1})^{-1}(2\mathbf{I} + \sigma\mathbf{T}_{m-1}), \tag{3.69}$$

and $\mathbf{T}_{m-1}$ is the $(m-1)\times(m-1)$ tridiagonal matrix:

$$\mathbf{T}_{m-1} = \begin{pmatrix} -2 & 1 & \dots & \dots & \dots & 0 \\ 1 & -2 & 1 & \dots & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & \dots & \dots & \dots & 1 & -2 \end{pmatrix}, \tag{3.70}$$

and $\sigma = \Delta t/\Delta x^2$ (see e.g. [169, p. 64-65]).

The matrix $\mathbf{T}_{m-1}$ has eigenvalues

$$\lambda_s = -4\sin^2\left(\frac{s\pi}{2m}\right), \quad (s = 1, 2, \dots, m-1). \tag{3.71}$$

Using the result from matrix theory that if a matrix $\mathbf{T}$ has eigenvalues λ_s, then the matrix $f(\mathbf{T})$ has eigenvalues $f(\lambda_s)$, it follows that $\mathbf{Q}$ has eigenvalues:

$$\mu_s = \frac{2+\sigma\lambda_s}{2-\sigma\lambda_s} \equiv \frac{2-4\sigma\sin^2[s\pi/(2m)]}{2+4\sigma\sin^2[s\pi/(2m)]}. \tag{3.72}$$

These eigenvalues are clearly less than one for all positive σ and, as shown below, this implies that the Crank-Nicolson scheme is stable.

To show that condition $|\mu_s| \le 1$ implies stability, we expand the initial data u^0 in terms of the orthonormal eigenvectors $\{\mathbf{w}_s\}$ of the matrix $\mathbf{Q}$ in the form:

$$u^0 = \sum_{s=1}^{m-1} a_s\mathbf{w}_s, \tag{3.73}$$

where $\mathbf{Q}\mathbf{w}_s = \mu_s\mathbf{w}_s$ and $(\mathbf{w}_p, \mathbf{w}_q) = \delta^p_q$. It then follows that

$$\|u^{n+1}\|_2 = \|\mathbf{Q}^n u^0\|_2 = \left\|\sum_{s=1}^{m-1} a_s\mu_s^n\mathbf{w}_s\right\|_2 \equiv \sum_{s=1}^{m-1} |a_s|^2|\mu_s|^{2n}. \tag{3.74}$$

If $|\mu_s| < 1$ for all s, then (3.74) implies $\|u^{n+1}\|_2 \leq \|u^0\|_2$. Defining the spectral radius of the matrix $\mathbf{Q}$ as $\rho(\mathbf{Q}) = \max_s|\mu_s|$, it follows that the condition for stability may be expressed as $\rho(\mathbf{Q}) \leq 1$.

Thus l_2-stability is equivalent to the condition that the spectral radius $\rho(\mathbf{Q}) \leq 1$. Note that matrix stability analysis is also applicable in the case of initial boundary value problems, as will be illustrated in the next chapter.

Example

If $\|\mathbf{Q}\| \leq 1 + C(T)\Delta t$ for some constant $C(T)$ as $\Delta t \to 0$, then the scheme is stable as follows from the simple estimate:

$$\|\mathbf{Q}^n\| \leq \|\mathbf{Q}\|^n \leq (1 + C(T)\Delta t)^{\frac{T}{\Delta t}} \leq e^{CT}, \tag{3.75}$$

where $T = n\Delta t$. Similarly, the condition $\rho(\mathbf{Q}) \leq 1 + C(T)\Delta t$ is necessary for stability. Suppose that $\rho(\mathbf{Q}) > 1 + C(T)\Delta t$ for every constant $C(T)$, then for sufficiently large n, $\rho(\mathbf{Q})^n$ approximates e^{CT} for arbitrary C and therefore tends to infinity as $n \to \infty$. Since $\rho(\mathbf{Q})^n \leq \|\mathbf{Q}^n\|$ in any norm, it follows that the condition $\rho(\mathbf{Q}) > 1 + C(T)\Delta t$ for every constant $C(T)$ is a necessary condition for instability in any norm.

An example of such a case occurs when $\mathbf{Q}$ has a Jordan block form. For example, such a $\mathbf{Q}$ arises from the leap-frog scheme for the advection equation, for which $\mathbf{Q}$ has polynomial growth as already mentioned in Section 3.1. Similar results can be obtained from eigenvalue analysis (see next section). For the leap-frog scheme, the matrix $\mathbf{Q}$ has a double eigenvalue $\lambda = -i$.

The condition $\rho(\mathbf{Q}) \leq 1 + C(T)\Delta t$ is also sufficient for l_2-stability in the case that the matrix $\mathbf{Q}$ is symmetrizable [75]. In the scalar case in particular, and for cases involving symmetric, diagonalizable, normal and orthogonal matrices, there are examples where $\mathbf{Q}$ can be symmetrized. Analysis of l_2-stability is easier to carry out in Fourier space in the case of a constant coefficient matrix $\mathbf{Q}$ and a uniformly spaced mesh. Then in Fourier space, $\tilde{U}^{n+1} = \tilde{\mathbf{Q}}\tilde{U}^n$ (the tilde accent denotes the solution in Fourier space) and $\|\mathbf{Q}\|_2 = \|\tilde{\mathbf{Q}}\|_2$. As already noted, such an analysis will produce necessary and in many cases sufficient conditions for linear stability in the l_2-norm for a uniform mesh in space and time.

Example. Practical Implications of the Convergence Theory

Convergence theory applies for the limiting case of vanishing discretization parameters. In practice that means that all features of the solution and geometry are resolved as discretization steps are refined. Suppose

$$\|u^n - U(x,t)\| \sim O(\Delta t^p, \Delta x^p). \tag{3.76}$$

Assuming a constant relation between Δt and Δx then

$$u^n_j(\Delta x) \sim U(\Delta x)^n_j + C\Delta t^p, \tag{3.77}$$

where the notation $u^n_j(\Delta x)$ and $U(\Delta x)$ means that Δx is the spatial discretization step used and U denotes the exact solution. Assuming that the solution converges as the discretization step Δx is decreased, then the order of accuracy p can be estimated from the ratio:

$$p \sim \log_2 \frac{\|U(\frac{\Delta x}{2}) - U(\frac{\Delta x}{4})\|}{\|U(\Delta x) - U(\frac{\Delta x}{2})\|}, \tag{3.78}$$

in the desired norm.

However, the exact solution is often not available and the approximate solution on the finest grid is taken as the "exact" solution. In such a situation one can only claim, but not prove, that the rate of convergence is similar to (3.78). Whether the limiting solution is in fact the desired solution of the differential equation requires that the consistency of the solution be checked. This is usually done by validating the computer code on a sequence of examples to see if the the code has been properly coded (i.e. the formulae in the code are checked for correctness, as well as a proper treatment of the boundary conditions). There are many examples where approximations are fast converging, but to the wrong solution.

3.3 Von Neumann Analysis and Courant-Friedrichs-Levy Necessary Stability Condition

Von Neumann analysis is a discrete Fourier analysis of the linear constant coefficient difference equation on a periodic domain with a uniform mesh size. The analysis shows that for L_2 stability, the spectral radius of the Fourier-space matrix $\hat{Q}$, (called the amplification matrix) needs to be of the order $1 + O(\Delta t)$ uniformly with respect to the number of iterations n. The Discrete Fourier Transform (DFT) corresponds to a rotation in the complex plane represented by a unitary matrix, which preserves the L_2 norm, $\|\hat{\mathbf{Q}}\|_{L_2} = \|\mathbf{Q}\|_{L_2}$. The above von Neumann restriction is a necessary condition for L_2 stability. It is sufficient in cases when power boundness of the amplification matrix follows from the above restrictions on spectral radius. This is the case when the amplification matrix is scalar, or similar via a uniformly bounded transformation to a diagonal, symmetric or normal matrix (a scalar matrix is $c\mathbf{I}$ where c is a scalar and $\mathbf{I}$ is the unit matrix). It is not sufficient in the case of a Jordan form of the amplification matrix, as happens for example for the leap-frog scheme applied to the advection equation.

The von Neumann analysis is equivalent to computing the dispersion relation of the discrete system by assuming the usual plane wave ansatz,

$$u_j^n = \hat{u}e^{i(kx-\omega t)} = \hat{u}e^{i(kj\Delta x-\omega n\Delta t)}. \tag{3.79}$$

We introduce the following notation for normalized wavenumber $\theta = k\Delta x$ and normalized frequency $\psi = \omega\Delta t$, which are called the grid wavenumber and grid frequency, respectively. It is also useful to introduce the amplification factor:

$$r = e^{-i\omega\Delta t}, \tag{3.80}$$

and to write the plane wave ansatz as $u_j^n = \hat{u}r^n e^{ij\theta}$, where u_j^n represents the dependent variables of the system, with $\theta \in [-\pi, \pi]$. The Nyquist wavenumber corresponds to $\theta = \pi$ or $k = \pi/\Delta x$. Similar considerations hold for the frequency as well.

The relation between $r = |r|e^{i\phi}$, where $\tan\phi = \mathrm{Im}(r)/\mathrm{Re}(r)$, and ω implies that $|r| = e^{-\mathrm{Im}(\omega)\Delta t}$ and $\Delta t\mathrm{Re}(\omega) = -\arctan[\mathrm{Im}(r)/\mathrm{Re}(r)]$. Thus we may compute the dispersion relation and numerical phase and group velocities in terms of the amplification factor r instead of ω.

For example, the phase speed $c = \mathrm{Re}(\omega)/k$ can be written in the form:

$$c(\theta) = \frac{\mathrm{Re}(\omega)}{k} = -\frac{1}{\lambda\theta}\arctan\left(\frac{\mathrm{Im}(r)}{\mathrm{Re}(r)}\right), \tag{3.81}$$

where $\lambda = \frac{\Delta t}{\Delta x}$. The group velocity can be computed similarly in terms of θ as:

$$g(\theta) = \frac{d\mathrm{Re}(\omega)}{dk} = \frac{d(\theta c(\theta))}{d\theta} = c(\theta) + \theta\frac{dc}{d\theta}. \tag{3.82}$$

Note that both $c(\theta)$ and $g(\theta)$ depend on the real part of the frequency $\mathrm{Re}(\omega)$, whereas the amplification factor $|r| = \exp[-\mathrm{Im}(\omega)\Delta t]$ depends on the imaginary part of the frequency $\mathrm{Im}(\omega)$. It is also clear from the discussion of convective and absolute instabilities in Section 1.4.1, that the dispersion equation $D(\omega, k) = 0$ can be thought of as an equation for $\omega = \omega(k)$ where k is real, or as an equation for $k = k(\omega)$ for real ω. More generally ω and k can both be complex. This latter viewpoint is that adopted by Bers [18] in his discussion of absolute and convective instabilities, amplifying and evanescent waves and pinch point analysis.

For numerical schemes with leading diffusion error, the numerical dispersion usually does not play an important role as high frequencies are damped at the highest rate. Similarly, for dispersive schemes, the phase errors and parasitic wave packets are more pronounced before any damping due numerical diffusion significantly affects the solution.

Example

Consider the advection equation:

$$u_t + au_x = 0,\ a > 0, \tag{3.83}$$

discretized by a variety of different difference schemes, including: central finite difference in space and forward/backward finite difference in time, leap-frog, Lax-Wendroff, upwind and Lax-Friedrichs schemes.

Substituting the von Neumann ansatz $u_j^n = \hat{u} r^n \exp(ij\theta)$ into the explicit forward time, central space (FTCS) scheme:

$$u_j^{n+1} - u_j^n + \frac{\mu}{2}\left(u_{j+1}^n - u_{j-1}^n\right) = 0, \tag{3.84}$$

where $\mu = a\Delta t/\Delta x$ is the Courant number, yields the amplification factor solution:

$$r = 1 + i\mu \sin\theta, \quad |r| = \left(1 + \mu^2 \sin^2\theta\right)^{1/2} \geq 1. \tag{3.85}$$

Since $|r| > 1$, the method is unstable.

The implicit FTCS scheme:

$$u_j^{n+1} - u_j^n + \frac{\mu}{2}\left(u_{j+1}^{n+1} - u_{j-1}^{n+1}\right) = 0, \tag{3.86}$$

has amplification factor solution:

$$r = \frac{1}{1 + i\mu\sin\theta}, \quad |r| = \left(1 + \mu^2 \sin^2\theta\right)^{-1/2} \leq 1. \tag{3.87}$$

This method is unconditionally stable with damping factor $|r(\theta)|$.

The leap-frog scheme:

$$u_j^{n+1} - u_j^{n-1} + \mu \left(u_{j+1}^n - u_{j-1}^n \right) = 0, \tag{3.88}$$

has amplification factor $r(\theta)$ satisfying the quadratic equation:

$$r^2 + 2i\mu r \sin\theta - 1 = 0, \tag{3.89}$$

with roots:

$$r = -i\mu \sin\theta \pm \left(1 - \mu^2 \sin^2\theta\right)^{1/2}. \tag{3.90}$$

If $|\mu| < 1$ then $|r(\theta)| = 1$ for all θ. Thus the leap-frog scheme is non-dissipative if $|\mu| < 1$.

If $\mu \sin\theta = 1$, the quadratic equation (3.89) has two equal roots $r = -i$. This case is more subtle to analyze than cases where the two roots of (3.89) are unequal. The leap frog scheme is a three-point scheme, since it involves the three time levels: t_{n+1}, t_n and t_{n-1}. For solutions involving the von Neumann ansatz, the scheme can be written as a matrix recurrence relation:

$$\mathbf{w}^{n+1} = \mathbf{G}(\theta)\mathbf{w}^n, \tag{3.91}$$

where $\mathbf{w}^n = (u_j^n, u_j^{n-1})^T$. The amplification matrix $\mathbf{G}(\theta)$, in (3.91) is given by:

$$\mathbf{G}(\theta) = \begin{pmatrix} -2i\mu\sin\theta & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.92}$$

If $|\mu \sin\theta| \neq 1$, then the matrix $\mathbf{G}$ has two distinct eigenvalues $\lambda \equiv r$, satisfying the eigenvalue equation (3.89) (i.e. λ satisfies the determinantal equation $\det(\mathbf{G} - \lambda\mathbf{I}) = 0$). In this case, the matrix $\mathbf{G}$ is similar to the diagonal matrix $\mathbf{G}'$:

$$\mathbf{G}' = \mathbf{PGP}^{-1} = \begin{pmatrix} \lambda_- & 0 \\ 0 & \lambda_+ \end{pmatrix}, \tag{3.93}$$

where

$$\mathbf{P} = \begin{pmatrix} \lambda_- & 1 \\ \lambda_+ & 1 \end{pmatrix}, \tag{3.94}$$

is the transformation matrix. The transformation (3.93) corresponds to the transformation $\hat{\mathbf{w}} = \mathbf{P}\mathbf{w}$ in which the eigenvectors of $\mathbf{G}$ are used as the new coordinate base vectors. The eigenvalues $\lambda_{\pm} \equiv r_{\pm}$ are given by (3.90).

In the case $\mu \sin\theta = 1$, the amplification matrix (3.92) has the form:

$$\mathbf{G} = \begin{pmatrix} -2i & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.95}$$

Because $\lambda_- = \lambda_+ = -i$, the matrix $\mathbf{P}$ in (3.94) is singular and the similarity transformation (3.93) does not apply. However, the complex matrix (3.95) is equivalent to a Jordan canonical matrix, $\hat{\mathbf{G}}$, via the similarity transformation:

$$\hat{\mathbf{G}} = \hat{\mathbf{P}}\mathbf{G}\hat{\mathbf{P}}^{-1} = \begin{pmatrix} -i & 1 \\ 0 & -i \end{pmatrix}, \tag{3.96}$$

where

$$\hat{\mathbf{P}} = \begin{pmatrix} 0 & 1 \\ 1 & i \end{pmatrix}. \tag{3.97}$$

Introducing the state vector:

$$\hat{\mathbf{w}}^n = \hat{\mathbf{P}}\mathbf{w}^n = \left(u_j^n, u_j^n + iu_j^{n-1}\right)^T, \tag{3.98}$$

the recurrence relation (3.91) may be written in the form:

$$\hat{\mathbf{w}}^{n+1} = \hat{\mathbf{G}}\hat{\mathbf{w}}^n = \hat{\mathbf{G}}^n\hat{\mathbf{w}}^1 = (-i)^{n-1}\begin{pmatrix} -i & n \\ 0 & -i \end{pmatrix}\hat{\mathbf{w}}^1. \tag{3.99}$$

In terms of the original variables, (3.99) can be written in the form:

$$u_j^n = (-i)^{n-1}\left((n-1)iu_j^0 + nu_j^1\right), \tag{3.100}$$

$$u_j^{n+1} + iu_j^n = (-i)^n\left(u_j^1 + iu_j^0\right). \tag{3.101}$$

It is straightforward to show that (3.101) is a consequence of (3.100).

The solution (3.100) shows that $\mathrm{Re}(u_j^n)$ has a component that grows linearly with n. Note that the condition $\mu\sin\theta = 1$ is satisfied if $\mu = 1$ and $\theta = \frac{\pi}{2}$. The condition $\theta = \pi/2$ corresponds to a grid wavenumber that is half the Nyquist frequency. A further interesting solution of the leap-frog finite difference equation (3.88) occurs for $\theta = \pi$, which corresponds to the

Nyquist wavenumber. In this case, the amplification factor r in (3.90) has solutions $r = \pm 1$. In addition, the difference equation possesses oscillatory solutions, behaving like $c(-1)^n$, where c is a constant.

The Lax-Wendroff, upwind and Lax-Friedrichs schemes are analyzed below. The Lax-Wendroff scheme:

$$u_j^{n+1} - u_j^n + \frac{1}{2}\mu \left(u_{j+1}^n - u_{j-1}^n\right) - \frac{1}{2}\mu^2 \left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) = 0, \tag{3.102}$$

has amplification factor:

$$\begin{aligned} r &= 1 - i\mu \sin\theta - 2\mu^2 \sin^2(\theta/2), \\ |r| &= \left(1 - 4\mu^2(1-\mu^2)\sin^4(\theta/2)\right)^{1/2}. \end{aligned} \tag{3.103}$$

The upwind scheme:

$$u_j^{n+1} - u_j^n + \mu \left(u_j^n - u_{j-1}^n\right) = 0, \tag{3.104}$$

has amplification factor:

$$r = (1-\mu) + \mu e^{-i\theta}, \quad |r| = \left(1 - 4\mu(1-\mu)\sin^4(\theta/2)\right)^{1/2}. \tag{3.105}$$

The Lax-Friedrichs scheme:

$$u_j^{n+1} - \frac{1}{2}\left(u_{j+1}^n + u_{j-1}^n\right) + \frac{\mu}{2}\left(u_{j+1}^n - u_{j-1}^n\right) = 0, \tag{3.106}$$

has amplification factor:

$$r = \cos\theta - i\mu\sin\theta, \quad |r| = [1 - (1-\mu^2)\sin^2\theta]^{1/2}. \tag{3.107}$$

The Lax-Wendroff scheme is second order accurate, while the other two schemes are only first order accurate both in space and time. All three schemes are stable only if $|\mu| \leq 1$.

The leap-frog scheme applied to the advection equation $u_t + au_x = 0$ has phase velocity $c(\theta)$ and group velocity $g(\theta)$ given by the equations:

$$c(\theta) = \frac{\mathrm{Re}(\omega)}{k} = \frac{\arcsin(\mu\sin\theta)}{\lambda\theta}, \tag{3.108}$$

$$g(\theta) = \frac{\mu\cos\theta}{\lambda\sqrt{1-\mu^2\sin^2\theta}}, \tag{3.109}$$

where $\lambda = \Delta t/\Delta x$. The graph or Taylor expansion (for smooth well-resolved modes, θ near zero), shows that the numerical phase and group velocities are lagging the physical ones that are constant and equal to a. For high frequency components, the behavior of the unresolved modes is dramatically different from the physical behavior. For example, the group velocity may become negative leading to contra-propagating parasitic wave packets. Similar conclusions hold for the Lax-Wendroff scheme as well. The leap-frog scheme has another peculiar property. The dispersion relation $\sin\phi + \mu\sin\theta = 0$ (which may also be written in the form: $\sin(\omega\Delta t) = \mu\sin(k\Delta x)$), with stability restriction $\mu < 1$, implies that $k(\omega)$ will become complex and the wave will decay exponentially in space, while the group velocity becomes zero, giving rise to so-called non-propagating modes. More detailed discussion of this phenomenon is given later, in the section on interface boundary conditions for adaptive mesh refinement (AMR) algorithms for Maxwell equations.

Example

Consider the diffusion equation:

$$u_t = \kappa u_{xx}. \tag{3.110}$$

It can be approximated by the semi-implicit scheme (3.5). Von Neumann analysis of (3.5) for the explicit ($\Theta = 0$), implicit ($\Theta = 1$), and Crank-Nicolson ($\Theta = \frac{1}{2}$) schemes, gives, respectively:

$$r = 1 - 4\sigma^2\sin^2(\theta/2),\ \frac{1}{1 + 4\sigma\sin^2(\theta/2)},\ \frac{1 - 2\sigma\sin^2(\theta/2)}{1 + 2\sigma\sin^2(\theta/2)}, \tag{3.111}$$

for the amplification factor r, where $\sigma = \kappa\Delta t/\Delta x^2$.

The explicit scheme is stable if $\sigma \le \frac{1}{2}$, while the implicit and Crank-Nicolson schemes are unconditionally stable. Note in the Crank-Nicolson scheme that if a large time step is chosen ($\sigma \gg 1$), and if the wavenumber is set equal to the Nyquist wavenumber $\theta = \pi$, we find $r = (1-2\sigma)/(1+2\sigma) \to -1$ as $\sigma \to \infty$ (i.e. $\Delta t \gg \Delta x^2/\kappa$). Thus in this case $r \approx -1$, ($|r| \le 1$), $u_j^n = r^n \exp(ij\theta)$ will be oscillatory for each consecutive value of n. Also, even although stability conditions may not impose time step restrictions, the accuracy requirements do, as the error for transient solutions will be $O(\Delta t, \Delta x^2)$ in the first two cases, and $O(\Delta t^2, \Delta x^2)$ in the Crank-Nicolson case. If one requires a solution of the diffusion equation at late times where the solution converges to the steady state solution, then a first order implicit scheme may be advantageous. These types of problems will be discussed in the next section, where we discuss the stability of long-time integration.

Also, note that taking $\kappa = i$ shows that Crank-Nicolson is non-dissipative when applied to Schrödinger's equation as $|r| = 1$ in this case.

Example. Finite Element Method

The stability of the finite element method and the method of weighted residuals discussed in Section 2.4 can also be analyzed by using the DFT. As in von Neumann analysis, we assume a plane wave ansatz as a solution of the discrete equations. In Section 2.4, we discussed the solution of the advection-diffusion equation (2.68) by using the method of weighted residuals. In this method, a solution of the advection diffusion equation was sought in the form:

$$u(x,t) = \sum_{m=1}^{N} c_m(t)\phi_m(x), \tag{3.112}$$

where the $\{\phi_m(x)\}$ are the trial functions. By substituting the solution ansatz (3.112) into a weak form of the equation (i.e. taking moments of the advection-diffusion equation with appropriate test functions $\{\psi_i(x) : 1 \leq i \leq N\}$), results in the matrix equation system (2.75). We consider the stability of the homogeneous version of (2.75), namely:

$$M\frac{d\mathbf{c}(t)}{dt} + A\mathbf{c}(t) = 0, \tag{3.113}$$

where M is the mass matrix and A is the stiffness matrix, given by (2.76). Here $\mathbf{c}(t) = (c_1(t), c_2(t), \ldots, c_N(t))^T$. These modal values are linearly related to nodal values by the equations $u(x_j,t) = \sum_{i=1}^{N} c_i(t)\phi_i(x_j)$. For example, for piecewise linear basis functions $\{\phi_i(x)\}$ generated by the hat functions (2.73), we have $\phi_i(x_j) = \delta_{ij}$, and the modal and nodal values are the same in this case (i.e. $c_i(t) = u(x_i,t)$).

For the above matrix ODE system (3.113), we use a von Neumann ansatz for the solution that is continuous in time and discrete in space, of the form:

$$u(x_j,t) = e^{-i\omega t}e^{ij\theta}, \tag{3.114}$$

where $\theta = k\Delta x$ and $j = 0, 1, \ldots, N-1$. Since $u(x_j,t) = c_j(t) = e^{-i\omega t}e^{ij\theta}$, it follows that $\mathbf{c}(t) = \hat{u}e^{-i\omega t}$, where $\hat{u} = (1, e^{i\theta}, e^{i2\theta}, \ldots, e^{i(N-1)\theta})^T$. Substituting this solution ansatz for $\mathbf{c}(t)$ in the matrix differential equation (3.113) we obtain the matrix equation:

$$(-i\omega M + A)\hat{u} = 0, \tag{3.115}$$

for the eigenvector $\hat{u}$. For a non-trivial solution of (3.115) for $\hat{u}$ we require that ω satisfies the eigenvalue equation $\det(A - i\omega M) = 0$. The latter determinental equation defines the dispersion relation for the discretized system. To analyze the full discretization, one should check that all eigenvalues ω are such that $\mathrm{Im}(\omega) < 0$, corresponding to a stable discretization of the ODE system (3.113). These ideas will be considered in more detail in the next section.

Example. Courant-Friedrichs-Lewy Condition

Another simple necessary stability restriction for hyperbolic problems (linear or nonlinear) arises from the finite speed of propagation in such problems. The CFL restriction implies that all initial data should be taken into account by the numerical scheme to assure convergence. Alternatively, the numerical speed should be greater than or equal to the physical speed of propagation, or the numerical domain of dependence should include the physical domain of dependence. This is equivalent to requiring that, if the physical solution is influenced by some region in the initial data, then the numerical scheme has to propagate this information in a timely fashion, otherwise changes in the physical domain of dependence will not be accounted for by the numerics. In practice, the check that the CFL condition is satisfied is the simplest check on the stability and convergence of the code. Only the determination of the numerical speed requires some thought.

For example, implicit schemes for hyperbolic problems will automatically satisfy the CFL restriction as all information on the previous time levels goes into the update. The three-point scheme in one dimension will propagate one grid point per time step and therefore the CFL restriction will imply that the numerical grid speed $\Delta x/\Delta t \geq \lambda$, where λ is the largest possible local physical speed in the system. For example, for the Euler equations of gas dynamics, $\lambda = \max_x(|\mathbf{u}| + c)$, where $\mathbf{u}$ is the local fluid velocity and c is the local sound speed. In the case of the compressible Euler equations, the propagation of information depends on whether the flow is supersonic or subsonic. For the case of supersonic flow, the information can only propagate inside the local Mach cone, whereas for subsonic flow, the information can propagate in all directions. It should also be noted that the concept of supersonic and subsonic in general depends on the frame of reference in which the fluid velocity is measured. For example, for two-dimensional self-similar flow, in which the physical variables depend only on x/t and y/t, it is appropriate to measure the fluid velocity relative to the local, self-similar frame moving with velocity $\mathbf{w} = (x/t, y/t)^T$. That the condition $\Delta x/\Delta t \geq \lambda$ is only necessary, but not sufficient for stability, follows from the fact that a numerical scheme using explicit, central, spatial and forward time differences, may satisfy the CFL condition but is still unstable. Similarly, other three-point schemes may satisfy the CFL condition, but the schemes are unstable. For example, the explicit FTCS scheme (3.84) for the advection equation (3.83) satisfying the CFL condition $\mu = a\Delta t/\Delta x < 1$, is still unstable, presumably because downwind information is needed to compute the first order spatial derivative. On the other hand, the Lax-Wendroff scheme (3.102), which is the same as the FTCS scheme if the second order spatial derivatives are neglected, is stable if the CFL condition is satisfied. The second order spatial difference operator in the Lax-Wendroff scheme takes into account higher order derivative information associated with the solution operator $e^{\Delta t \partial t}$ and is stable if the CFL condition $|\mu| \leq 1$ is satisfied.

In the case of fluid flow in two-space dimensions, if we look separately at the propagation along each of the coordinate directions we get one-dimensional (1D) CFL restrictions. However, if we consider propagation along the diagonal of the rectangular mesh, it takes two time steps for the three-point scheme to affect the value across the diagonal. Thus in this case, the CFL restriction is $\frac{\sqrt{\Delta x^2+\Delta y^2}}{2\Delta t} \geq \lambda$. If $\Delta x = \Delta y$ then the CFL restriction is equivalent to $\lambda\Delta t/\Delta x \leq 1/\sqrt{2}$, which is a more stringent condition than the 1D CFL condition $\lambda\Delta t/\Delta x \leq 1$. Again, in the case of the Euler equations, $\lambda = \max_{x,y}(\sqrt{u^2+v^2}+c)$, where c is the sound speed and u and v are the two components of the fluid velocity in the x and y directions, respectively. Similar considerations apply to difference equations in any number of dimensions and for schemes having arbitrary domains of dependence, and not just for three-point schemes.

3.4 Project Assignment

The aim of this assignment is to study numerically the order of accuracy of spatial discretization and the overall efficiency of the four algorithms implemented in the Project Assignment from Chapter Two. For each algorithm provide a graph of the spatial error in L_2-norm versus Δx. The plots should be on the log-log or semi-log scale to demonstrate polynomial, $O(\Delta x^p)$, or spectral order of accuracy, $O(e^{-\gamma/\Delta x})$, respectively, as $\Delta x \to 0$. To compute the error use the exact solution, when available, or the numerical result computed with the finest grid, when the exact solution cannot be computed analytically. In the latter case you have to establish consistency of your numerical implementation, for example, by showing that each term and/or the reduced equation, obtained by dropping some terms, are approximated consistently. Note that if consistency is not established, the only claim that can be made about convergence is the

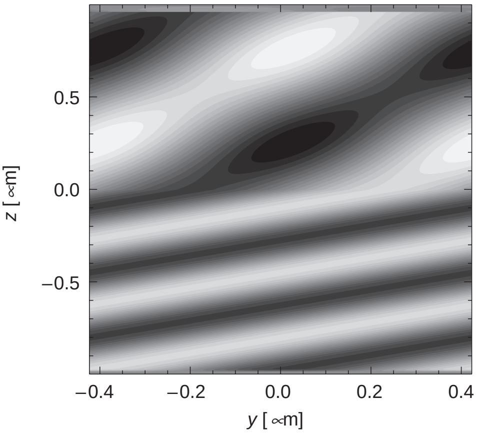


Figure 3.1: Time-snapshot of the E_y field of a planewave scattering from a a dielectric interface. The minimum (black) and the maximum (white) of the field distribution correspond to ∓ 0.75.

convergence “to itself”, without a claim that the numerical solution is even convergent, let alone the claim that it converges to the solution of the continuous problem.

The efficiency study should be illustrated by graphs of the CPU time, or the number of floating point operations, required to achieve an accuracy tolerance (both in space and time) of the order $10^{-n}, n = 1, 2, 3, \ldots$. As before, the axis should be scaled properly to detect polynomial or exponential behavior of the efficiency dependence on the desired tolerance.

3.5 Project Sample

The Finite-Difference Time-Domain (FDTD) method for Maxwell’s equations [201] uses a staggered grid for the **E** and **H** fields in both space and time. In addition to the physically motivated discrete representation of the integral form of the Maxwell equations, this staggering also corresponds to a central finite-difference operator applied to the space and time coordinates, and thus leads to a second-order accurate approximation, with the truncation error proportional to $O(\Delta t^2) + O(\Delta x^2)$. The discretization error of the numerical method can be computed by comparing the numerical and exact solutions.

Figure 3.1 shows the FDTD solution of the reflection-transmission problem, in which a TE_x (E_y, E_z, H_x) polarized planewave with wavelength $\lambda_0 = 650$ nm is incident at an angle of 50° from the air-side on a

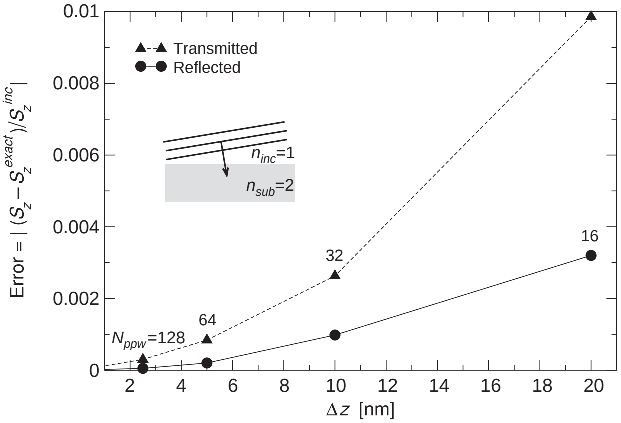


Figure 3.2: Rate of convergence for a problem of a planewave scattering from a a dielectric interface. The error in the numerical solution decreases as $O(\Delta z^2)$ with increasing number of points per wavelength, $N_{ppw} = \lambda_0/(n_{sub}\Delta z)$.

dielectric with relative permittivity $\epsilon = 4$. The fringes resulting from the superposition of the incident and reflected waves above the interface ($z > 0$), as well as the transmitted wave in the glass below the interface ($z < 0$), are evident in a time-snapshot of the distribution of the y-component of the $\mathbf{E}$ field.

The numerical error in the reflection and transmission coefficients can be computed by comparing the Poynting vector component along the z-axis for the exact and discrete solutions. The real valued components of the time averaged Poynting vector are computed from the complex valued $\mathbf{E}(\lambda_0)$ and $\mathbf{H}(\lambda_0)$ fields as $\langle \mathbf{S} \rangle = \frac{1}{2}\mathrm{Re}(\mathbf{E} \times \mathbf{H}^*)$. The $\mathbf{E}(\lambda_0)$ and $\mathbf{H}(\lambda_0)$ fields are evaluated by applying a DFT at the source frequency to the time-dependent data of $\mathbf{E}(t)$ and $\mathbf{H}(t)$ fields at the monitor points above and below the interface.

Figure 3.2 shows the computed error in the energy flux along the z-axis, given by the Poynting vector component S_z, as a function of the grid step-size. The difference between the computed and exact S_z, normalized by the incident energy flux, indicates second-order $O(\Delta z^2)$ convergence of the numerical solution to the exact solution, for which the reflectivity and transmissivity are given by $\mathrm{R} = 0.0268, \mathrm{T} = 0.9732$. In problems with material interfaces not aligned along the grid lines, the stair-cased approximation of the curved material interfaces on the finite-difference grid in general will reduce the order of convergence to the exact solution to $O(\Delta)$.

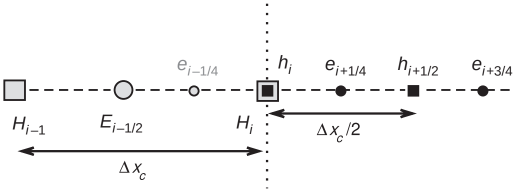


Figure 3.3: Interface between coarse $(E_{j-1/2}, H_j)$ and fine $(e_{j+1/4}, h_{j+1/2})$ FDTD spatial grids. Coarse grid magnetic field node H_j is collocated with the fine grid node h_j. The "ghost" electric field value $e_{j-1/4}$ is interpolated to provide a grid interface "boundary condition" for the fine grid.

Introduction of sources, boundary conditions or interfaces that are discretized using a lower order approximation than the order of the method itself, will also reduce the overall accuracy of the solution. One such example is the effect of the accuracy of discretization at the grid interfaces on the convergence rate of the FDTD method with grid refinement. Figure 3.3 shows an interface between coarse (Δx_c) and fine $(\Delta x_f = \Delta x_c/2)$ patches of a 1D FDTD grid, on which for every time step Δt_c on the coarse grid two time steps $\Delta t_c/2$ are applied to advance solution on the fine grid. Solution on the coarse grid can be advanced in time from $H^{n-1/2}$ to $H^{n+1/2}$ and E^n to E^{n+1} using the interface node value $H_j^{n+1/2}$, which can be computed from the fine grid values $h_i^{n+1/4}$ and $h_i^{n+3/4}$ collocated with it [203]. To compute the magnetic field values $h_i^{n+1/4}$ and $h_i^{n+3/4}$ from Maxwell's equations, the electric field "ghost" values $e_{i-1/4}^n$ and $e_{i-1/4}^{n+1/2}$ must be interpolated from the nearby coarse and fine E field nodes. Linear interpolation in space can be used to obtain a second-order accuracy for the interface condition, consistent with the FDTD scheme:

$$e_{i-1/4}^n = \frac{2}{3}E_{i-1}^n + \frac{1}{3}e_{i+1/4}^n. \tag{3.116}$$

For the second time step on the fine grid linear interpolation in time can also be applied to the coarse grid values:

$$e_{i-1/4}^{n+1/2} = \frac{2}{3}\hat{E}_{i-1}^{n+1/2} + \frac{1}{3}e_{i+1/4}^{n+1/2}, \tag{3.117}$$

where $\hat{E}_{i-1}^{n+1/2} = (E_{i-1}^n + E_{i-1}^{n+1})/2$. If a constant $\hat{E}_{i-1}^{n+1/2} = E_{i-1}^n$ interpolation is used instead, then the accuracy of the interface will reduce to first-order accuracy in time.

■

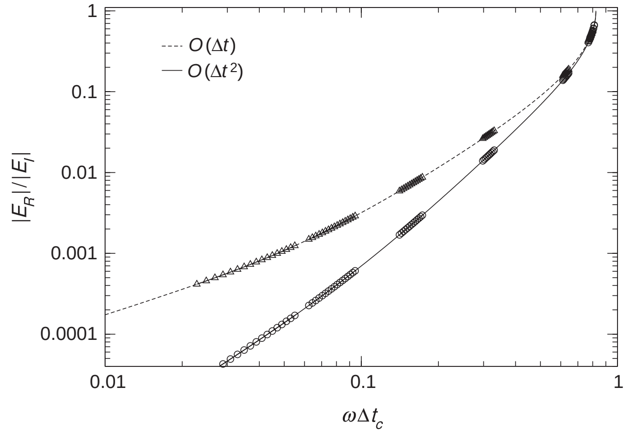


Figure 3.4: Rate of convergence for a problem of a pulse incident on the grid refinement interface. The reflection coefficient for the waves traveling from the coarse to the fine side of the interface is shown against the dimensionless angular frequency $\omega\Delta t_c$. Symbols represent numerical solution, while the lines correspond to the analytic result [203].

The amplitude of waves reflected at the grid refinement interface, for incidence on the interface from either the coarse or fine side of the grid, can be expected to be proportional to the order of accuracy of the discretization. Hence, the numerical reflection coefficient should approach zero (perfect transmission) in the limit of $\Delta t \to 0$, $\Delta x \to 0$.

Numerical reflection coefficients for the waves incident from the coarse grid side can be computed by taking a difference E_R between time series of a solution, sampled near the grid interface on the coarse side, and a reference solution sampled at the same point, but computed without the interface. Then the numerical reflection coefficient $R(f) = E_R(f)/E_I(f)$ is evaluated by taking the ratio of amplitudes of DFT of time samples of the two signals, where the incident pulse is given by

$$E_I = \sin(\omega t)\exp(-[(t-t_0)/\sigma]^2). \tag{3.118}$$

The coarse to fine (fine to coarse) transmission coefficient $T(f)$ can be similarly evaluated by taking the ratio of DFTs of the transmitted signal on the fine (coarse) grid and a reference solution computed on the fine (coarse) grid without the interface. The reflection coefficient for waves incident on the interface from the coarse grid side is shown in Figure 3.4. Linear interpolation in both space and time, used to compute a “ghost” field value at the interface, leads to the $O(\Delta t^2)$ convergence rate, while a linear in space and constant in time interpolation results in a slower convergence rate $O(\Delta t)$.

In cases when an exact solution is not known one may take for reference the numerical solution computed on the finest grid possible. This will only demonstrate "self-convergence", however, and is not sufficient to establish consistency. For example, if the numerical method is implemented inconsistently, say by shifting a grid node by half a cell, that would reduce the accuracy to first order, but "self-convergence" will not detect this, as the convergence is tested with respect to the shifted solution. Typically one needs to validate the implementation on a series of tests in order to establish the convergence rate to the exact solution, without actually knowing it.

Chapter 4

Numerical Boundary Conditions

4.1 Introduction to Numerical Boundary and Interface Conditions

Numerical boundary conditions arise in the process of numerical implementation of given physical boundary conditions on the original physical or truncated domain, due to restrictions imposed by computational resources. The latter situation is often encountered when an infinite domain is truncated or remapped onto a finite computational domain. Numerical implementation of the physical or numerical interface boundary conditions should preserve the accuracy and stability of the inner numerical method. The inaccuracies and instabilities created by numerical implementation of the interface and boundary conditions may be localized at the boundaries or interfaces, but more often they may propagate throughout the whole computational domain.

Example. Advection Equation

The advection equation:

$$u_t + u_x = 0, \quad 0 \leq x \leq 1, \tag{4.1}$$

describes a right-going wave. The artificial sources created by the numerical implementation of the boundary condition at $x = 0$ will generate waves propagating into the computational domain. On the other hand, a boundary condition at $x = 1$ may create a discontinuous boundary layer if the boundary condition is mismatched with the incoming solution. Since any perturbation of the left boundary, e.g. shifting the numerical boundary by a half grid point, will propagate throughout the computational domain, such perturbations may result in an overall lower order of accuracy. Similarly, abruptly turned on sources may cause the same problem, due to the propagation outwards of an initial jump discontinuity.

Example. Upwind Differencing for the Advection Equation

The advection equation, discretized by an upwind numerical method with uniform space-time stepping, gives the following explicit iteration,

$$u_j^{n+1} = u_j^n - \mu(u_j^n - u_{j-1}^n), \quad j = 1, 2, \ldots, n, \tag{4.2}$$

where $\mu = \Delta t / \Delta x$ is the Courant number. The method admits only right-going waves and may be used up to and including the right-hand boundary point. If a different boundary condition is applied for example at $x_n = 1$, e.g. by using a different numerical method to update the boundary point, the boundary condition information will not propagate back into the computational domain. On the other hand, any perturbations or inaccuracies at the left-hand boundary will propagate throughout the whole domain without creating instability, due to the dissipative nature of the upwind numerical method.

Example. Leap-frog Differencing for the Advection Equation

If the advection equation:

$$u_t - u_x = 0, \quad 0 \leq x \leq 1,$$

is discretized by the second-order leap-frog numerical method,

$$u_j^{n+1} = u_j^{n-1} + \mu(u_{j+1}^n - u_{j-1}^n), \quad j = 1, 2, \ldots, M-1, \tag{4.3}$$

then both boundary conditions must be provided as the method is based on centered differencing in space. The leap-frog scheme admits both the right and the left-going waves as can be seen from its dispersion relation. In addition, the leap-frog method is non-dissipative and therefore more susceptible to instabilities due to numerical boundary conditions. For example, the following implementations of transparent (non-reflective) boundary conditions at $x = 1$ have the following properties, [177],

$$u_0^{n+1} = u_0^{n+1}, \quad \text{(unstable)},$$

$$u_0^{n+1} = u_0^n, \quad \text{(stable, but only first-order accurate)},$$

$$u_0^{n+1} = u_0^{n-1} + 2\mu(u_1^n - u_0^n), \quad \text{(unstable)},$$

$$u_0^{n+1} = u_0^n + \mu(u_1^n - u_0^n), \quad \text{(stable and second-order accurate)}.$$

Stability analysis in the presence of the boundaries and interfaces will be discussed in the last section of this chapter. The above example illustrates a general observation that there is no a priori way to know whether the mathematically reasonable interpolation designed to preserve accuracy will also preserve the stability as well, unless some assurance principles are used in the design, e.g. symmetries leading to energy conservation or dissipation to damp instabilities. Therefore, in most cases, a posteriori analysis is required after the whole update algorithm is in place. Usually, the analysis cannot be done analytically, except for a few simple equations, domains and boundary conditions. In practice, a numerical evaluation of the stability is performed for several domain configurations with long runs to test the method on a variety of examples. Often, high frequency modes are the culprit, and numerical experiments with initial conditions on small computational domains and random noise initial conditions often reveal instabilities quite inexpensively.

4.2 Transparent Boundary Conditions for Hyperbolic and Dispersive Systems

The design of transparent boundary conditions is based on particular solutions of the system at hand, such as plane waves or far-field solutions obtained analytically or asymptotically. A similar approach using a prescribed form of the solution will be used in the next section to design an absorbing layer boundary condition. In contrast, transparent boundaries are applied at a single point, line or surface. Once the desired ansatz for the boundary behavior is chosen, we reconstruct the partial differential equation (PDE) to be satisfied at the boundary. Finally, these equations are discretized and the whole update scheme is analyzed for accuracy and stability.

Example. One-dimensional Wave Equation

Consider a transparent boundary condition design for the one-dimensional (1D) wave equation for sound waves:

$$u_{tt} - c^2(x) u_{xx} = 0, \tag{4.4}$$

with transparent boundary conditions at $x = L$ for waves impinging from the left. Assume that the solution is smooth and consists of outgoing waves with constant sound speed for $x \geq L$, with solution of the form $u(x,t) = f(x - c(L)\ t)$, for $x \geq L$, where f is some unknown function.

From the given form of the solution, after differentiation in both variables, it is easy to see that this solution satisfies the PDE $u_t + c(L)u_x = 0$, where $c(L)$ is the sound speed value at the boundary $x = L$. This equation needs to be further discretized to preserve accuracy and stability of the inner discretization scheme.

Example. Radiation Boundary Condition for Three-dimensional Wave Equation

Using a similar argument as we used for the 1D wave equation example above, consider the three-dimensional (3D) wave equation

$$u_{tt} = (u_{xx} + u_{yy} + u_{zz}), \quad \text{for } 0 < r < R. \tag{4.5}$$

Assume that the solution becomes radially symmetric as $r \to \infty$, i.e. $u(r,t) = f(r-t)/r + g(r+ct)/r$. Retaining only the outgoing solution implies that u satisfies the Sommerfeld radiation boundary condition, $u_t + u_r + \frac{1}{r}u = 0$, at $r = R$, [79].

Example. Radiation Boundary Condition for Two-dimensional Wave Equation

Instead of the exact analytical solution ansatz, one may use the following far-field asymptotic expansion of the outgoing solution [79]:

$$u(r,\theta,t) = \frac{f(r-t)}{\sqrt{r}}\left(a_0(\theta) + \frac{a_1(\theta)}{r} + \cdots\right). \tag{4.6}$$

Taking derivatives of the approximate solution ansatz (4.6) with respect to each variable and combining them into PDE form gives the following radiation boundary condition at $r = R$,

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial r} + \frac{1}{2r}\right)u = O\left(\frac{1}{r^{5/2}}\right). \tag{4.7}$$

Example. Three-dimensional Wave Equation in a Strip $0 < x < L$ Domain

Consider the 3D wave equation [78],

$$u_{tt} = u_{xx} + u_{yy} + u_{zz}.$$

After substituting a plane wave ansatz $u(x,y,z,t) = e^{i(k_1x+k_2y+k_3z-\omega t)}$ into the equation, we get the following dispersion relation:

$$\omega^2 = k_1^2 + k_2^2 + k_3^2.$$

Selecting the outgoing wave at $x = L$ according to the positive sign of the phase velocity, $\omega/k_1 > 0$, gives the relation:

$$k_1 = \omega\sqrt{1 - \frac{k_2^2 + k_3^2}{\omega^2}},$$

or

$$\tilde{k}_1 = \sqrt{1 - \tilde{k}_t^2},$$

where the tilde denotes the normal and transverse wave numbers, k_1 and $k_t^2 = k_2^2 + k_3^2$, scaled by the frequency ω. Since the right-hand side of the equation is not a rational function of $\tilde{k}_t$, this dispersion relation, when transformed back into physical domain, will result in a nonlocal equation that has to be discretized at $x = L$. In order to avoid this complication, the square root function is approximated by a rational or polynomial expression in terms of $\tilde{k}_t$. For example, taking for simplicity $\tilde{k}_t = \tilde{k}_2$ and applying various Padé approximations, we obtain the following transparent boundary equations after converting the above dispersion relation back into the physical domain,

$$\begin{aligned}
&\sqrt{1 - \tilde{k}_2^2} \approx 1, && u_t + u_x = 0,\\
&\sqrt{1 - \tilde{k}_2^2} \approx 1 - \frac{1}{2}\tilde{k}_2^2, && 2u_t + 2u_x + u_{yyt} = 0,\\
&\sqrt{1 - \tilde{k}_2^2} \approx \frac{1}{1 - \frac{1}{2}\tilde{k}_2^2}, && 2u_t + 2u_x - u_{yyx} = 0,\\
&\sqrt{1 - \tilde{k}_2^2} \approx \frac{1 - \frac{3}{4}\tilde{k}_2^2}{1 - \frac{1}{4}\tilde{k}_2^2}, && 4u_{ttt} + 4u_{ttx} - 3u_{tyy} - u_{xyy} = 0.
\end{aligned}$$

The ratio $\tilde{k}_2/\tilde{k}_1$ describes the angle of propagation for the plane wave. Since all of the above expansions were taken near $\tilde{k}_2 = 0$, they are called paraxial and wide-angle approximations, respectively. After the choice of the continuous boundary condition is made, the task is to determine whether the resulting initial boundary value problem is well-posed. After that the discretization step follows. Often, to maintain stability of the discrete approximation, an implicit Crank-Nicolson type method is applied. The semi-circle $\tilde{k}_1 = \sqrt{1 - \tilde{k}_t^2}$ in $\tilde{k}_1, \tilde{k}_t$ plane can be approximated by many other polynomial and rational interpolations. Unfortunately, none of them work well for all angles of incidence, $\tan\phi = \tilde{k}_1/\tilde{k}_t$. Some do very well at nearly normal incidence ($k_t = 0$), while others perform better at a nearly glancing angle ($\tilde{k}_1 = 0$).

Example. An Application to Fourier Optics

A similar uni-directional propagation approximation in optics is called the Beam Propagation Method (BPM). Consider a plane wave solution,

$$u(x, y, z, t) = e^{i(k_1 x + k_2 y + k_3 z - \omega t)},$$

for the Maxwell wave equation along a fiber or a waveguide that is homogeneous in the z-direction. The dispersion relation is the same as that described in the previous example,

$$\tilde{k}_3 = \sqrt{1 - \tilde{k}_t^2}.$$

Then, in the moving frame $\tilde{z} = z - \frac{\omega t}{\tilde{k}_3}$ the diffraction pattern at distance $\tilde{z}$ is given by:

$$u(x, y, \tilde{z}) = \iint \tilde{u}(\tilde{k}_1, \tilde{k}_2) e^{2\pi i(\tilde{k}_1 x + \tilde{k}_2 y + \tilde{k}_3 \tilde{z})} d\tilde{k}_1 d\tilde{k}_2,$$

with Fourier transform $\tilde{u}(\tilde{k}_1, \tilde{k}_2)$ defined as:

$$\tilde{u}(\tilde{k}_1, \tilde{k}_2) = \iint u(x, y, 0) e^{-2\pi i(\tilde{k}_1 x + \tilde{k}_2 y)} dx dy.$$

Various analytical solutions and asymptotic approximations are known for special cases of $\tilde{u}(\tilde{k}_1, \tilde{k}_2)$ and are associated with Airy, Fresnel, Poisson and Fraunhofer diffraction patterns, to name a few. Numerically, computation of $u(x, y, \tilde{z})$ consists of computing the Fourier transform of the initial distribution, $u(x, y, 0)$, multiplying it by the phase factor $e^{2\pi i \tilde{k}_3 \tilde{z}}$ and applying the inverse Fourier transform to the resulting function. A similar procedure was discussed for the pseudo-spectral approximation of the derivative, filtering and the evaluations of the convolution integrals using Fourier representation in Chapter 2.

In the next section we will study a sponge (absorbing) boundary layer that is non-reflecting for plane waves of all frequencies ω, wavenumbers k and for all angles of incidence, using a perfectly matched boundary layer (PML) method [16]. The reflection error for these boundary conditions is several orders of magnitude less than for the transparent boundary conditions described above. For example, errors of 10^{-4}-10^{-5} are routinely obtained in practice for the PML boundaries. The absorbing boundary layer is usually about 10 to 15 grid points wide, versus a single boundary point for the transparent boundary condition. In addition, the PML equations are twice the size of the original PDEs. As a result, up to half of the memory and CPU time may be spent in the PML regions. Therefore, if high accuracy is not required, the standard transparent boundary conditions might be more efficient, but in the majority of problems, PML boundaries are superior to the standard transparent boundary conditions.

Example. Linear Hyperbolic Equations

Consider the linear hyperbolic system of equations:

$$U_t + A_0 U_x = 0.$$

We write the solution in characteristic form (as was discussed in Chapter 1), as:

$$U(x,t) = \sum_{i=1}^{n} v_i(x - \lambda_i t) R_i.$$

The characteristic variables $v_i = L_i u$ satisfy the advection equations:

$$v_{it} + \lambda_i v_{ix} = 0, \quad i = 1, 2, \ldots, n,$$

with R_i and L_i denoting normalized right and left eigenvectors corresponding to the eigenvalue λ_i. Assuming that at the right boundary $x = L$ there are no incoming waves, we let $\partial_t v_i = 0$ or $L_i u = const$ in time for all indices i such that $\lambda_i < 0$.

The generalization of this method to a nonlinear hyperbolic system where the matrix $A(U)$ is dependent on the state vector U was carried out by Hedstrom under the assumption that solution is a simple wave (that is $U = U(\phi(x,t))$ for some function $\phi(x,t)$, [81]). Substituting this solution ansatz into the original system:

$$U_t + A(U)\, U_x = 0,$$

leads to the equation:

$$\left(I \frac{\phi_t}{\phi_x} + A(U) \right) \frac{dU}{d\phi} = 0.$$

Therefore, $\lambda = -\frac{\phi_t}{\phi_x}$ is an eigenvalue of the matrix A with right eigenvector $\frac{dU(\phi)}{d\phi} = R(U(\phi))$. The first integrals of this ordinary differential equation (ODE) system are called Riemann invariants (or simple wave integrals). Strictly speaking, Riemann invariants are defined as quantities that are constant on the characteristics (e.g. [44]). They consist of functions $\mathcal{R}(U)$ where $\partial\mathcal{R}/\partial U$ is a left eigenvector of the matrix $A(U)$. However, in the above gas dynamical example, the matrix A is symmetric, and the left and right eigenvectors are equivalent, but this is not always the case (e.g. in magnetohydrodynamics). The significance of simple wave integrals will be illustrated in the next example.

Example. Riemann Invariants for Euler Equations of Gas Dynamics

Consider the 1D Euler equations in the form:

$$\begin{aligned}
&\rho_t + (\rho u)_x = 0,\\
&u_t + uu_x + P_x/\rho = 0,\\
&s_t + us_x = 0,
\end{aligned}$$

where ρ, u, s, and P represent density, velocity, entropy and the pressure of the gas. The system is closed by assuming a polytropic gas with equation of state, $P = e^s\rho^\gamma$, with constant γ denoting the ratio of specific heats (here s is the normalized entropy $s = S/C_v$ and $\gamma = C_p/C_v$ where C_p and C_v are the specific heats of the gas at constant pressure and constant volume, respectively). If the system is written in the matrix form $U_t + AU_x = 0$, where $U = (\rho, u, s)^t$ is the state vector, one finds that the matrix A has eigenvalues $u, u \pm c$, where $c = \sqrt{\gamma P/\rho}$ is the gas sound speed. The eigenvalues $\lambda_1 = u - c$, $\lambda_2 = u$ and $\lambda_3 = u + c$ correspond to the backward sound wave, the contact discontinuity and the forward sound wave eigenmodes, respectively. The right eigenvector for the contact discontinuity is $R_u = (1, 0, -P_\rho/P_s)^t$ and $R_{u\pm c} = (1, \pm c/\rho, 0)^t$ are right eigenvectors for the sound waves, respectively. The Riemann invariants for the contact discontinuity are the first integrals of the ODE system:

$$\frac{d\rho}{1} = \frac{du}{0} = \frac{ds}{-P_\rho/P_s} = d\phi, \tag{4.8}$$

and the Riemann invariants for the sound modes satisfy the ODE system:

$$\frac{d\rho}{1} = \frac{du}{\pm c} = \frac{ds}{0} = d\phi. \tag{4.9}$$

The integrals of (4.8) for the contact discontinuity are $u = \text{const}$ and $p = \text{const.}$, and the integrals of (4.9) for the sound wave eigenmodes are $s = \text{const}$ and $\mathcal{R}_{\mp} = u \mp 2c/(\gamma - 1) = \text{const}$. The above analysis gives isentropic simple sound waves, which correspond to the sound wave Riemann invariants for waves with $s = \text{const}$. However, there exist generalized simple sound waves, in which the entropy is not constant, which can be derived, for example, by using the Monge-Ampere equation formulation of 1D gas dynamics (see e.g. [187]). These solutions involve so-called first integrals, or intermediate integrals of the Monge-Ampere equation and only apply for specific distributions of the entropy, which depend on the stream function.

The significance of the Riemann invariants is twofold. They represent smooth expansion wave solutions that, together with shocks, can be superimposed to generate the solution of the Riemann problem, an initial value problem with a piecewise constant initial condition (shock tube problem) [50]. Riemann solvers are at the heart of modern Godunov-type numerical methods for nonlinear hyperbolic systems that find numerous applications in fluid dynamics, magneto-hydrodynamics, astrophysics, elasticity, and other applications of continuum mechanics, [111]. In addition, Riemann invariants are nearly constant across the weak shocks, the variation is of the third-order of the strength of the shock measured in terms of the ratio $r = (p_2 - p_1)/p_1$, and even for a moderately strong shock, $r \in (0.5 - 10)$, the variation is small, [198].

The Hedstrom's boundary condition:

$$L_j(U)\frac{dU}{dt} = 0,$$

where j corresponds to the incoming waves is the first transparent boundary condition that explicitly uses nonlinear self-similar solutions. It works quite well in one dimension [81] even for shocks of moderate strength.

Example. Hadley's Transparent Boundary Condition for Dispersive Waves

Given any linear dispersive discrete system, assume that solution consists of a single plane wave $u(x, y, z, t) = e^{i(k_1 x + k_2 y + k_3 z - \omega t)}$, with x, y, z and t being uniformly discretized. Consider the boundary located at $x = 0$. Using the above plane wave ansatz, the solution satisfies the relation $u(0, y, z, t)u(2\Delta x, y, z, t) = u(\Delta x, y, z, t)^2$. Therefore, one can determine the boundary value at $x = 0$ as the ratio of the two nearby inner solution values [76]. Unfortunately, this approach will break down and generate noticeable reflections if the solution consists of more than one plane wave.

Example. Higdon's Transparent Boundary Condition for Dispersive Waves

Assume that a 1D linear dispersive system satisfies the dispersion relation $\omega = \omega(k^2)$. At the right boundary, $x = L$, Higdon's transparent boundary condition [82] consists of the product of 1D outgoing linear operators,

$$\prod_{j=1}^{n} \left(\frac{\partial}{\partial t} + c_j \frac{\partial}{\partial x} \right) u = 0, \tag{4.10}$$

where the number of operators n and the values of the constants c_j have to be chosen according to the following procedure. Assume that the solution u consists of the incident (right-going) and reflected (left-going) waves,

$$u(x,t) = e^{i(kx-\omega t)} + Re^{i(-kx-\omega t)}, \tag{4.11}$$

with ω/k being positive. The reflection coefficient R then can be explicitly computed by substituting the above ansatz (4.11) into the boundary condition (4.10). This gives the reflection coefficient as a product of the reflection coefficients for each operator,

$$R = -\prod_{j=1}^{n} \frac{(\omega - kc_j)}{(\omega + kc_j)} \exp(2ikL) \quad \text{and} \quad |R(\tilde{k})| = \prod_{j=1}^{n} \left| \frac{1 - c_j\tilde{k}}{1 + c_j\tilde{k}} \right|,$$

where $\tilde{k} = k/\omega$ is a normalized form of the wave number, known as the slowness. The determination of the free parameters of the problem, the coefficients c_j, as well as the number of equations, n, to be used in the boundary condition is usually done by numerically minimizing $|R(\tilde{k})|$ as a function of $\tilde{k}$. Note, that the above formula shows that $R(0) = 1$, which in turn implies that there are large reflections for nearly constant solutions. Typically the coefficients c_j are chosen as wave packet velocities corresponding to dominant wavenumbers present in the problem. The resulting boundary condition is often described by a PDE system of higher order than the original system. Finally, the resulting continuous boundary PDEs need to be discretized in an accurate and stable fashion. For an application to two coupled nonlinear Klein-Gordon equations, see [106].

4.3 Berenger's Perfectly Matched Layer Boundary Conditions

Perfectly Matched Layer [16] absorbing boundary conditions produce an absorbing boundary layer for linear wave-type equations. Theoretically, they are reflection-free for plane waves regardless of wavenumber, direction and frequency of the incident wave. This is in contrast to the transparent boundary conditions discussed in the previous section, or pre-Berenger boundary absorbers done via simple ansatz to include damping, or by applying filtering to the solution in the physical domain by multiplying it by a top hat function. In practice, Berenger's boundary conditions produce reflections that are several orders of magnitude smaller than the transparent boundary conditions. The performance of the PML boundary conditions is degraded for nonlinear systems and the implementation of Berenger's boundary conditions is usually much more involved than either the implementation of the inner scheme or the transparent boundary conditions. In addition, memory and CPU usage usually ranges from ten to as high as fifty percent of the inner iterations, whereas transparent boundary conditions require only a negligible part of the computational resources. Despite this, in the majority of practical applications we are familiar with, the PML boundary conditions easily outperform the standard transparent boundary conditions in terms of efficiency (CPU per given accuracy).

Example. One-dimensional Advection Equation

Consider the advection equation on interval $[0, L]$,

$$u_t + u_x = 0,$$

with the solution ansatz $u = e^{-\int_L^x \sigma(y)dy} e^{ik(x-t)}$, where the damping coefficient $\sigma(x) \geq 0$ varies gradually from its zero value in the computational domain in order to damp the solution in the region $x > L$ for an arbitrary wavenumber k. Taking partial derivatives one can derive the following equation satisfied by this ansatz,

$$u_t + u_x = -\sigma(x)u.$$

The damping coefficient is taken to be zero in the computational domain and then turned on smoothly from zero value to some limiting value $\sigma_{\max}$ in the absorbing layer $[L, \tilde{L}]$, e.g. $\sigma(x) = (x-L)^p, 1-e^{-(x-L)^{2p}}$, etc. Note that the absorber works for arbitrary wavenumber and frequency, therefore this is a PML boundary layer. Also, it is interesting to note that the solution is independent of the gradient of the damping coefficient $\sigma(x)$, which is not the case for the discrete problem, and experimentation is needed to determine the optimal shape of the damping coefficient. For the 1D case we found that a super-Gaussian with $p = 2$ was the best choice among those we considered and that it was better than any polynomial function.

A similar approach works for other wave equations. For example, consider the 2D wave equation for $x < 0$,

$$u_{tt} = u_{xx} + u_{yy}.$$

It admits the Fourier solution $u = e^{ik(x\cos(\theta)+y\sin(\theta)-t)}$, where the wavevector $\vec{k}$ is written in polar coordinates, $\vec{k} = (k\cos\theta, k\sin\theta)$. Outside the physical domain, in the region $x > 0$, we assume a solution of the form: $u = e^{-\alpha(x)}e^{ik(x\cos(\theta)+y\sin(\theta)-t)}$, where $\alpha(x) = \int_0^x \sigma(x')dx'$. Using this solution ansatz, and evaluating second derivatives, gives a wave equation with an absorbing layer in $0 < x < L$ of the form:

$$u_{tt} - 2\sigma(x)u_x - (\sigma^2 + \sigma_x)u = u_{xx} + u_{yy}.$$

Note, that in the absorption region the sign of the u_x term has to be reversed for the left-going waves, otherwise they will be amplified back to the original amplitude, unless the amplitude of the wave at the outer boundary is exactly zero. Such an absorber is called non-reciprocal. This non-PML approach may be feasible if numerical discretization is to be based on a wave decomposition like that used in Riemann solvers, which are discussed in the next chapter. In this case, damping can be applied to each wave separately regardless of the propagation direction. Such a method would trade off memory required by the reciprocal PML equations for extra computation involved in wave decomposition methods. Their stability and efficiency in comparison to PML boundary conditions need further investigation.

Note that in the corner region, repeating the procedure with an ansatz having both x and y damping factors, $\sigma_1(x)$ and $\sigma_2(y)$, gives a corner absorbing equation as:

$$u_{tt} - 2\sigma_1(x)u - 2\sigma_2(y)u_y - (\sigma_1^2 + \sigma_{1x} + \sigma_2^2 + \sigma_{2y})u = u_{xx} + u_{yy}.$$

4.3.1 Maxwell's Equations

In this section we consider application of Berenger's PML method for absorption of waves at the boundaries of the computational domain. The basis of the analysis are the Maxwell equations:

$$\begin{aligned}
&\nabla\times\mathbf{E}=-\frac{\partial\mathbf{B}}{\partial t},\quad \nabla\cdot\mathbf{B}=0,\\
&\nabla\times\mathbf{H}=\mathbf{J}+\frac{\partial\mathbf{D}}{\partial t},\quad \nabla\cdot\mathbf{D}=\rho_c,
\end{aligned}\tag{4.12}$$

where **E**, **D**, **B**, **H** are the electric field strength, electric field displacement, magnetic field induction and magnetic field strength, respectively; ρ_c is the free charge density and $\mathbf{D}=\epsilon\mathbf{E}$ and $\mathbf{B}=\mu\mathbf{H}$ are the constitutive relations between **D** and **E** and **B** and **H**, and ϵ and μ are the electric permittivity and magnetic permeability, respectively. In lossy media, **J** is related to **E** via Ohm's law $\mathbf{J}=\sigma\mathbf{E}$, where σ is the electric conductivity.

Example. Berenger's Perfectly Matched Layer Boundary Condition for Two-dimensional Maxwell Equations

Consider the solution of the two-dimensional (2D) version of Maxwell's equations (4.12) where the spatial coordinate z is ignorable (i.e. $\partial/\partial z=0$). In general, the solutions of the 2D Maxwell equations (4.12) can be split up into transverse electric (TE) and transverse magnetic (TM) modes. The TE modes involve only the variables (E_x,E_y,B_z), whereas the TM modes involve (B_x,B_y,E_z). Consider the problem of matching the TE mode equations across a boundary $x=0$ between two homogeneous, dielectric media. Assume that the region $x<0$ (region 1) is lossless, but region 2 $(x>0)$ is a lossy medium (i.e. $\sigma\neq 0$). In the lossless region, $x<0$, the TE mode equations are:

$$\begin{aligned}
&\epsilon_1\frac{\partial E_x}{\partial t}=\frac{\partial H_z}{\partial y},\\
&\epsilon_1\frac{\partial E_y}{\partial t}=-\frac{\partial H_z}{\partial x},\\
&\mu_1\frac{\partial H_z}{\partial t}=\frac{\partial E_x}{\partial y}-\frac{\partial E_y}{\partial x}.
\end{aligned}\tag{4.13}$$

The first two equations in (4.13) correspond to the x and y components of Ampére's law $\nabla\times\mathbf{H}=\mathbf{J}+\partial\mathbf{D}/\partial t$, with $\mathbf{J}=0$, and the z component of Faraday's law $\partial\mathbf{B}/\partial t=-\nabla\times\mathbf{E}$, respectively. Similarly, in the lossy medium (region 2, with $x>0$), the TE modes satisfy the equations:

$$\begin{aligned}
&\epsilon_2\frac{\partial E_x}{\partial t}+\sigma_2 E_x=\frac{\partial H_z}{\partial y},\\
&\epsilon_2\frac{\partial E_y}{\partial t}+\sigma_2 E_y=-\frac{\partial H_z}{\partial x},
\end{aligned}$$

$$\mu_2 \frac{\partial H_z}{\partial t} + \sigma_{m2} H_z = \frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x}. \tag{4.14}$$

In (4.14) an extra, non-physical term $\sigma_{m2} H_z$ has been added to the left-hand side of the equation, which is useful in producing an absorbing boundary in which the reflection coefficient $R = 0$ for normal incidence.

The term $J_m = \sigma_m H_z$ added to Faraday's law in (4.14) is analogous to a "magnetic current" that is similar to the electric current $\mathbf{J} = \sigma \mathbf{E}$ in Ampére's law. For waves impinging on the boundary from $x < 0$, it was shown by Berenger that at such absorbing boundary layers, the reflection coefficient $R = 0$ for arbitrary wavenumbers only at normal incidence [16]. In order to overcome this restriction, Berenger introduced a more refined strategy involving the splitting of the magnetic field H_z in the lossy medium 2, which we discuss later in this section. We note also for later reference, that the TE mode equations (4.14) can be combined to yield the PDE:

$$\left[\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + (\sigma\mu + \epsilon\sigma_m)\frac{\partial}{\partial t} + \sigma\sigma_m - \nabla_\perp^2\right] H_z = 0, \tag{4.15}$$

for the magnetic field H_z, where $\nabla_\perp^2 = \partial_x^2 + \partial_y^2$. If $\sigma_m = 0$, then (4.15) is a telegrapher equation for H_z. However, if $\sigma_m \neq 0$, (4.15) is a mixed telegrapher Klein-Gordon type equation. In the case that $\sigma = \sigma_m = 0$, (4.15) is a 2D wave equation for H_z.

Assume that an incident plane wave in $x < 0$ impinges onto the interface $x = 0$ at an angle θ. The solution ansatz in region 1 ($x < 0$) consists of an incident wave and a reflected wave of the form:

$$\begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix} = \mathbf{U}^{(i)} \exp\left(ik(x\cos\theta + y\sin\theta) - i\omega t\right) \\ + \mathbf{U}^{(r)} \exp\left(ik(-x\cos\theta + y\sin\theta) - i\omega t\right), \tag{4.16}$$

where $\mathbf{U} = (\hat{E}_x, \hat{E}_y, \hat{H}_z)^t$ is the state vector or eigenvector in Fourier space. Similarly, in region 2 ($x > 0$), the transmitted wave has the solution form:

$$\begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix} = \mathbf{U}^{(t)} \exp\left[i(\tilde{k}_x x + ky\sin\theta) - i\omega t\right]. \tag{4.17}$$

The superscripts i, r and t in (4.16)–(4.17) refer to the incident, reflected and transmitted waves, respectively.

In the solutions (4.16)–(4.17), the eigenvectors $\mathbf{U} = (\hat{E}_x, \hat{E}_y, \hat{H}_z)^t$ must satisfy the eigenvector equation of the form $\mathbf{AU} = 0$, where

$$\mathbf{A} = \begin{pmatrix} \sigma - i\omega\epsilon & 0 & -ik_y \\ 0 & \sigma - i\omega\epsilon & ik_x \\ -ik_y & ik_x & \sigma_m - i\mu\omega \end{pmatrix}, \tag{4.18}$$

and we have temporarily dropped the subscripts 1 and 2, in order to emphasize, that the same eigenequations formally apply in both regions, except that $\sigma = \sigma_m = 0$ in region 1. The dispersion equation for the waves is given by setting the determinant of the matrix A equal to zero in (4.18), i.e.

$$\det(\mathbf{A}) = (\sigma - i\omega\epsilon)\left[k_x^2 + k_y^2 + (\sigma - i\epsilon\omega)(\sigma_m - i\mu\omega)\right] = 0. \tag{4.19}$$

Dropping the root associated with the factor $(\sigma - i\omega\epsilon)$ as irrelevant in the present context (it represents a damped, non-propagating mode in region 2 and a zero frequency mode in region 1), we obtain the dispersion equation solution in region 2 as:

$$\tilde{k}_x^2 + k_y^2 = \omega^2\mu_2\epsilon_2\left(1 + \frac{i\sigma_2}{\epsilon_2\omega}\right)\left(1 + \frac{i\sigma_{m2}}{\mu_2\omega}\right), \tag{4.20}$$

where in general $\tilde{k}_x \neq k\cos\theta$ is the x-component of the wave vector $\mathbf{k}$ in region 2. Similarly, in region 1, the dispersion equation (4.19) reduces to:

$$k^2 \equiv k_x^2 + k_y^2 = \omega^2\mu_1\epsilon_1. \tag{4.21}$$

Note that in region 1, $k^2 = \omega^2/c_1^2$, where c_1 is the speed of light in region 1.

The eigenvectors $\mathbf{U} = (\hat{E}_x, \hat{E}_y, \hat{H}_z)^t$ for the incident, (i), reflected (r) and transmitted (t) waves are given by the equations:

$$\begin{aligned} \mathbf{U}^{(i)} &= \frac{H_{z1}^{(i)}}{\omega\epsilon_1}\left(-k_y, k_x, \omega\epsilon_1\right)^t, \\ \mathbf{U}^{(r)} &= \frac{H_{z1}^{(r)}}{\omega\epsilon_1}\left(-k_y, -k_x, \omega\epsilon_1\right)^t, \\ \mathbf{U}^{(t)} &= \frac{H_{z2}^{(r)}}{\omega\epsilon_2 + i\sigma_2}\left(-k_y, \tilde{k}_x, \omega\epsilon_2 + i\sigma_2\right)^t. \end{aligned} \tag{4.22}$$

The integral form of Maxwell's equations imply that the transverse (tangential) fields E_y and H_z are continuous across the discontinuous interface $x = 0$. In the above matching conditions, both the frequency ω and transverse wavenumber $k\sin\theta$ are preserved across the interface due to the fact that the continuity of transverse field components is valid at any time and at any place along the interface. In fact, derivatives of all orders with respect to t and y of the transverse variables are continuous across the interface. Thus the matching conditions at the interface $x = 0$ are:

$$E_{y1} = E_{y2}, \quad H_{z1} = H_{z2}, \tag{4.23}$$

and ω and $k\sin\theta$ are continuous across $x = 0$.

Using the eigenvector relations (4.22), the matching conditions (4.23) at the interface, may be written in the form:

$$\frac{k_x\left(\hat{H}_{z1}^{(i)} - \hat{H}_{z1}^{(r)}\right)}{\omega\epsilon_1} = \frac{\tilde{k}_x \hat{H}_{z2}^{(t)}}{\omega\epsilon_2 + i\sigma_2}, \tag{4.24}$$

$$\hat{H}_{z1}^{(i)} + \hat{H}_{z1}^{(r)} = \hat{H}_{z2}^{(t)}. \tag{4.25}$$

The transmission and reflection coefficients may be defined as:

$$R = \frac{E_{y1}^{(r)}}{E_{y1}^{(i)}}, \quad T = \frac{E_{y2}^{(t)}}{E_{y1}^{(i)}}. \tag{4.26}$$

Note, that relations (4.22) also imply that $R = -H_{z1}^{(r)}/H_{z1}^{(i)} = -E_{x1}^{(r)}/E_{x1}^{(i)}$. Using (4.22)–(4.26) we obtain:

$$R = \frac{\eta_2 - \eta_1}{\eta_2 + \eta_1}, \quad T = \frac{2\eta_2}{\eta_2 + \eta_1}, \tag{4.27}$$

where

$$\eta_1 = \frac{k\cos\theta}{\omega\epsilon_1}, \quad \eta_2 = \frac{\tilde{k}_x}{\omega\epsilon_2\left(1 + i\sigma_2/(\omega\epsilon_2)\right)}. \tag{4.28}$$

From (4.27) we note that $1 + R = T$.

To obtain some insight into the expressions (4.27)–(4.28) for R and T, consider the special case of normal incidence, with $\theta = 0$. Using the dispersion equations (4.20)–(4.21) in (4.28) we obtain for $\theta = 0$, the results:

$$\eta_1 = \sqrt{\frac{\mu_1}{\epsilon_1}}, \quad \eta_2 = \sqrt{\frac{\mu_2}{\epsilon_2}}\left(\frac{1 + i\sigma_{m2}/(\mu_2\omega)}{1 + i\sigma_2/(\epsilon_2\omega)}\right)^{1/2}, \tag{4.29}$$

for η_1 and η_2. The reflection coefficient $R = 0$ and transmission coefficient $T = 1$ in (4.27) if $\eta_1 = \eta_2$, and, since ω is arbitrary, this occurs in (4.29) if

$$\sqrt{\frac{\mu_1}{\epsilon_1}} = \sqrt{\frac{\mu_2}{\epsilon_2}} \quad \text{and} \quad \frac{\sigma_{m2}}{\mu_2} = \frac{\sigma_2}{\epsilon_2}. \tag{4.30}$$

Hence, for normal incidence, the reflection coefficient $R = 0$ if the lossless impedances $Z_1 = \sqrt{\mu_1/\epsilon_1}$ and $Z_2 = \sqrt{\mu_2/\epsilon_2}$ match, and if the artificial magnetic conductance $\sigma_{m2} = \sigma_2\mu_2/\epsilon_2$. This example shows that a reflectionless, absorbing boundary can indeed be constructed in the case of normal incidence if the impedance and dissipation matching conditions (4.30) are fulfilled and waves of all frequencies and wave numbers are absorbed in this case.

Berenger's Split-field Perfectly Matched Layer Equations

For oblique incidence, setting $R(\theta) = 0$ gives a non-trivial relation between k, ω and θ that, in general, will not be satisfied. In order to overcome this difficulty Berenger introduced extra parameters into the system by nonphysically splitting $H_z = H_{zx} + H_{zy}$, and by splitting Faraday's law with an unphysical magnetic current term, into two separate equations for H_{zx} and H_{zy}, to obtain the equation system:

$$\epsilon\frac{\partial E_x}{\partial t} + \sigma_y E_x = \frac{\partial H_z}{\partial y}, \tag{4.31}$$

$$\epsilon\frac{\partial E_y}{\partial t} + \sigma_x E_y = -\frac{\partial H_z}{\partial x}, \tag{4.32}$$

$$\mu\frac{\partial H_{zx}}{\partial t} + \sigma_{mx} H_{zx} = -\frac{\partial E_y}{\partial x}, \tag{4.33}$$

$$\mu\frac{\partial H_{zy}}{\partial t} + \sigma_{my} H_{zy} = \frac{\partial E_x}{\partial y}, \tag{4.34}$$

for the TE modes. Note that Berenger's PML equations generalize the previously considered system (4.14). If $\sigma_x = \sigma_y = \sigma$ and $\sigma_{mx} = \sigma_{my} = \sigma_m$, then the Berenger's PML equations system reduces to the system (4.14). The PML conductivities σ_j, σ_{mj} $(j = x, y)$ are introduced, not for each dependent variable, but for each independent spatial variable making the PML medium anisotropic and dispersive. Note that in $x < 0$, the medium is assumed lossless with $\sigma_j = \sigma_{mj} = 0$ $(j = x, y)$, but in $x > 0$, σ_j and σ_{mj} $(j = x, y)$ are in general non-zero.

For solutions of (4.34) in the form of plane waves: $\psi = \hat{\psi}\exp[i(k_x x + k_y y - \omega t)]$, where ψ denotes any of the dependent variables, the system (4.34) reduces to the matrix system:

$$\begin{pmatrix} \sigma_y - i\omega\epsilon & 0 & -ik_y & -ik_y \\ 0 & \sigma_x - i\omega\epsilon & ik_x & ik_x \\ 0 & ik_x & \sigma_{mx} - i\omega\mu & 0 \\ -ik_y & 0 & 0 & \sigma_{my} - i\omega\mu \end{pmatrix} \mathbf{U} = 0, \tag{4.35}$$

where $\mathbf{U} = (\hat{E}_x, \hat{E}_y, \hat{H}_{zx}, \hat{H}_{zy})^t$ is the state vector. Existence of non-trivial solutions for $\mathbf{U}$ requires that the determinant of the 4×4 matrix on the left-hand side of (4.35) is zero, which will be satisfied if $\omega(\mathbf{k})$ satisfies the dispersion equation:

$$\frac{k_x^2}{s_x^2} + \frac{k_y^2}{s_y^2} = 1, \tag{4.36}$$

where

$$s_x^2 = \omega^2\epsilon\mu\left(1 + \frac{i\sigma_x}{\omega\epsilon}\right)\left(1 + \frac{i\sigma_{mx}}{\omega\mu}\right),$$
$$s_y^2 = \omega^2\epsilon\mu\left(1 + \frac{i\sigma_y}{\omega\epsilon}\right)\left(1 + \frac{i\sigma_{my}}{\omega\mu}\right). \tag{4.37}$$

The factors $1/s_x$ and $1/s_y$ have the dimensions of length, and can be thought of as scaling or stretching parameters in the x and y directions (note however, that s_x and s_y are, in general, complex).

The eigenvector solution of (4.35), $\mathbf{U} = (\hat{E}_x, \hat{E}_y, \hat{H}_{zx}, \hat{H}_{zy})^t$ may be expressed in the form:

$$\mathbf{U} = H_z\left(\beta_1, \beta_2, \beta_3, \beta_4\right)^t,$$
$$\beta_1 = \frac{-k_y}{\omega\epsilon[1 + i\sigma_y/\omega\epsilon]}, \quad \beta_2 = \frac{k_x}{\omega\epsilon[1 + i\sigma_x/\omega\epsilon]},$$
$$\beta_3 = \frac{k_x^2}{s_x^2}, \quad \beta_4 = \frac{k_y^2}{s_y^2}, \tag{4.38}$$

where $H_z = H_{zx} + H_{zy}$ and $\mathbf{k}$ is related to ω via the dispersion equation (4.36).

The boundary conditions at the interface $x = 0$ require that E_y, H_z, ω and $k_y = k \sin\theta$ are continuous across $x = 0$. The reflection and transmission coefficients at the interface, R and T are defined in (4.26). Using the eigenvector relations (4.38), we find:

$$R = \frac{-H_{z1}^{(r)}}{H_{z2}^{(i)}} \quad \text{and} \quad T = \frac{\eta_2 H_{z2}^{(t)}}{\eta_1 H_{z1}^{(i)}}, \tag{4.39}$$

where

$$\eta_1 = \frac{k_x}{\omega\epsilon_1}, \quad \eta_2 = \frac{\tilde{k}_x}{\omega\epsilon_2[1 + i\sigma_{2x}/\omega\epsilon_2]}. \tag{4.40}$$

Using the boundary conditions (4.23) for the continuity of E_y and H_z at the boundary, (4.39) now gives the standard expressions for reflection and transmission coefficients R and T at the interface given in (4.27), except that now η_1 and η_2 are given by (4.40), in which ω and $\mathbf{k}$ must satisfy the modified dispersion equation (4.36). The choice of the four parameters σ_x, σ_y, σ_{mx} and σ_{my} in the PML region $x > 0$, leads to more possibilities for minimizing the reflection coefficient R in the PML layer for all angles of incidence of the waves impinging on the interface from $x < 0$. The condition for $R = 0$ is obtained by setting $\eta_1 = \eta_2$ in (4.27) and (4.40), and by requiring that the dispersion equation (4.36) is satisfied.

Using the dispersion equation (4.36) in region 2 ($x > 0$), and noting that $\omega = kc_1$ in region 1 ($x < 0$) ($c_j = 1/\sqrt{\mu_j\epsilon_j}$ for $j = 1, 2$ denotes the speed of light in regions 1 and 2, respectively) the parameters η_1 and η_2 in (4.40) can be written in the form:

$$\begin{aligned} \eta_1 &= Z_1 \cos\theta, \\ \eta_2 &= Z_2 \left(\frac{1 + i\sigma_{mx}/\omega\mu_2}{1 + \sigma_x/\omega_2\epsilon_2}\right)^{1/2} \\ &\quad \times \left(1 - \frac{c_2^2 \sin^2\theta}{c_1^2(1 + i\sigma_y/\omega\epsilon_2)(1 + i\sigma_{my}/\omega\mu_2)}\right)^{1/2}, \end{aligned} \tag{4.41}$$

where

$$Z_1 = \sqrt{\frac{\mu_1}{\epsilon_1}}, \quad \text{and} \quad Z_2 = \sqrt{\frac{\mu_2}{\epsilon_2}}, \tag{4.42}$$

are the lossless impedances in regions 1 and 2, respectively. From (4.41)–(4.42) the previous result that $R = 0$ for normal incidence if the matching conditions (4.30) are satisfied, is recovered. However, in general, to enforce $R = 0$ by requiring $\eta_1 = \eta_2$, requires that the imaginary part of $\eta_2 = 0$ and that the real part of η_1 must equal the real part of η_2 in (4.41). Taking into account Snell's law, i.e. $c_2/c_1 = \sin\theta_2/\sin\theta_1$, and taking the lossless limit in (4.41), one obtains $\eta_2 = Z_2\cos\theta_2$. In this case $R = 0$ if $Z_1\cos\theta_1 = Z_2\cos\theta_2$, where $\theta_1 \equiv \theta$, is the angle of incidence of the impinging plane wave (4.16).

Using the theoretical reflection coefficient worked out by Berenger,

$$R(\theta) = \exp\left[-2\alpha\sigma_{\max}\cos(\theta)\int_0^d \left(\frac{x}{d}\right)^m dx\right],$$

we observe the incident angle dependence of the reflection coefficient. In fact, PML acts as a perfect waveguide for waves entering at a glancing angle, while they will be maximally attenuated as they enter into the corner region at a normal angle of incidence.

Unsplit Formulation of the Perfectly Matched Layer Equations

Another approach, called uniaxial PML (UPML), that replaces the material properties of the original system by introducing anisotropic dispersive media was proposed in [160]. In this brief overview, we will skip the derivation that is analogous to the unsplit case discussed above. The interested reader may consult the original paper for the derivation of the reflection/transmission coefficients, while the implementation of the method is extensively discussed in the standard text on computational electromagnetics [182]. In [160] it was shown that the following modification of the original Maxwell equations in the frequency domain has zero reflection coefficient regardless of the frequency, wave number or angle of the incident plane wave. The UPML equations are

$$i\omega\epsilon_2 S\vec{E} = \vec{\nabla}\times\vec{H}, \quad i\omega\mu_2 S\vec{H} = -\vec{\nabla}\times\vec{E},$$

where the matrix S consists of a single diagonal matrix, e.g. $S_x = (1/s_x, s_x, s_x)$, or the product of such matrices in the overlapping PML regions. For example in the corner region $S = \text{diag}(\frac{s_y s_z}{s_x}, \frac{s_x s_z}{s_y}, \frac{s_x s_y}{s_z})$. To illustrate the conversion of the PML equations back into the time domain, consider the equation for the z-component of the electric field:

$$i\omega\epsilon_2\left(\frac{s_x s_y}{s_z}\right)E_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}.$$

Introducing new variable D_z via the following constitutive relation, $D_z = \epsilon \frac{s_x}{s_y} E_z$, implies:

$$i\omega\epsilon_2 s_y D_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y},$$

or in physical space:

$$\frac{\partial D_z}{\partial t} + \sigma_y D_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}.$$

The constitutive equation in the time domain becomes:

$$\frac{\partial D_z}{\partial t} + \sigma_z D_z = \epsilon_2 \frac{\partial E_z}{\partial t} + \epsilon_2 \sigma_x E_z.$$

The rest of the PML equations are derived similarly.

Implementation of either the split or the unsplit approach will result in doubling the size of the original system. The split-field PML equations may be unstable in certain situations [1]. In [200], it was shown that the unsplit and split approaches are connected by scaling normal field components by factors s_x and s_y, respectively. In addition, the paper contains a comparison of the performance of split, unsplit and Complex Frequency Shifted (CFS) PMLs. In CFS the scaling factors are modified as follows:

$$s_x = \kappa(x) + \frac{\sigma(x)}{\beta(x) + i\omega},$$

in order to improve the absorption of the evanescent (decaying) waves that allow for complex wavenumber k and frequency ω, whereas in the original plane wave ansatz (4.16)–(4.17) they were assumed to be real variables (propagating waves).

4.4 Matrix Stability Analysis in the Presence of Boundaries and Interfaces

Consider a linear boundary value problem:

$$\frac{\partial(A(x,t)U(x,t))}{\partial t} = B(x,t,\partial_x,\partial_{xx}\ldots)U + F(x,t),$$

where A and B are linear operators with respect to U. For example, operator A may represent a time convolution, time integration or a multiplication operator with a given kernel. Similarly, the operator B is a spatial linear operator that may depend on various spatial derivatives of U. The spatial variable x is an n-dimensional variable in a multi-index notation, $x = (x_1, x_2, \ldots, x_n)$, and the function F represents a given "soft" source. In addition, the initial and linear boundary conditions are:

$$U(x,0) = f(x), \quad L_B U(x,t) = g(x,t), \quad L_I U(x,t) = h(x,t).$$

The boundary conditions include internal boundaries and interfaces as well. For example, "hard" sources that prescribe the solution as a given function $h(x,t)$ at specified locations of the domain, are represented as an internal boundary condition, be that a point, line, surface or distributed volume source.

We assume that the continuous problem is well-posed, in particular, that there is a unique solution which is uniformly bounded in time in terms of the initial, boundary and interface data in some norm, defined appropriately at the boundaries and interfaces, so that:

$$\|U(x,t)\| \leq K(T)(\|f(x)\| + \|F(t,x,y)\| + \|g(t,y)\|_B + \|h(t,y)\|_B),$$

holds for $0 \leq t \leq T$, with T being an arbitrary but fixed time, while $K(T)$ is independent of the initial, boundary/interface or forcing data.

A similar uniform boundedness requirement for a consistent discrete approximation is called stability and it is a necessary and sufficient condition for the convergence of the discrete approximation. The proof is analogous to the proof of the Lax equivalence theorem discussed in Chapter 3, with the only distinction that the definition of the consistency and truncation error has to include boundary and interface discretizations [111]. For example, a discrete approximation of the form:

$$V^{(n+1)} = Q_n V^{(n)} + \Delta t_n G_n,$$

where $V^{(n)}$ is an unknown array approximating the continuous variable $U(x,t)$ at time $t = t_n$, after substitution of the exact solution, U becomes

$$U^{(n+1)} = Q_n U^{(n)} + \Delta t_n G_n + \Delta t_n \tau_n.$$

Matrix Q_n represents a discretization of a spatial linear operator L using any spatial discretization, e.g. finite differences, finite or spectral elements, multidomain methods, etc., and G_n represents a discrete source. According to the generalized definition of consistency, the truncation error τ_n accounts for the boundary, source and interface discretization errors, including singularities.

The practical difficulty is that the uniform boundness in the definition of stability for the product of the update operators or, when Q_n is n-independent, powers of the matrix Q^n should be uniformly bounded with respect to n, the size of the matrix (spatial grid size) and the structure of the domain in inhomogeneous media. For example, the size of the matrix will increase under vanishing spatial and temporal discretization parameters, while various inhomogeneous domains may manifest themselves in various matrix structures and corresponding spectra.

Only in a very few simple cases of homogeneous media and uniform space and time discretization, complemented by simple boundary conditions, e.g. zero or periodic, the resulting update matrix has banded and/or periodic (circular) structure that allows for analytic or semi-analytic investigation of the stability. For example, one may apply Laplace-Fourier transforms in space and time, respectively, [75]; if only one of these conditions holds, one may proceed semi-analytically. For example, if time is uniformly discretized, applying a discrete Laplace transform (z-transform) results in an eigenvalue problem for the spatial operators in the problem that may in turn be studied numerically.

In some cases stability follows from considerations of preservation of symmetries, energy estimates, etc., that are analogous to the invariants and conserved quantities of the continuous system under consideration. In the discrete case, that is reflected in the structure of the matrices that preserve the properties of the original operators. For example, the Schrödinger equation:

$$i\psi_t = \psi_{xx} + V(x)\psi,$$

preserves the L_2 norm of the solution. In [96] the discretization is done on a multi-domain using orthogonal polynomials on each sub-domain. The resulting submatrices are symmetric. When the interface conditions are implemented using space-symmetric finite differences and the time discretization is done via a Crank-Nicolson method, the resulting global update matrix is unitary and the preservation of the l_2 norm is guaranteed, which in turn implies stability.

In practice, for the general case of a linear boundary-value problem, the update matrix may be difficult to set up, but since modern matrix eigenproblem software utilizes matrix multiplication as a main step in its iterative procedure, it is sufficient to provide a routine that returns the product of the update matrix with an arbitrary vector, [107]. This is equivalent to a routine that does one step of the update algorithm even if it is written in index notation.

The matrix stability analysis provides a necessary and sufficient condition for stability on a given fixed grid and domain structure. It states that the method is stable if the update matrix has eigenvalues lying inside or on the unit circle. In addition, eigenvalues lying on the unit circle should have the same geometric and algebraic multiplicity. Implementing this numerically is equivalent to determining whether the canonical form of the matrix contains non-trivial Jordan blocks. This problem is ill-conditioned due to sensitivity to perturbations and requires an extrapolation procedure from a series of high precision computations.

In practice, the high frequencies are often the cause of the instabilities, and stability testing on small spatial domains while varying the discretization parameters allows one to determine stability regions numerically. In the project section of this chapter we will illustrate this approach to matrix stability analysis for adaptive mesh refinement algorithm, to be described in Chapter Six. The algorithm involves non-uniform space-time meshes together with numerical interface boundary conditions that have to be designed to preserve the stability of the original Yee algorithm applied to the Maxwell wave equations.

4.5 Project Sample

(a) Material Interfaces

In semiconductor device simulation models the discretization of the governing equations at the interface between two materials involves numerical solution of Poisson's equation in an inhomogeneous medium:

$$-\nabla \cdot (\epsilon(\mathbf{r})\nabla\phi) = e(p-n), \tag{4.43}$$

where e is the elementary charge, ϕ is the electrostatic potential, and p, n are the space-coordinate dependent number densities of positively and negatively charged carriers. In particular, at an interface between a semiconductor material and an insulator, the Gauss law in a differential form can be written as:

$$\epsilon_s \frac{\partial\phi}{\partial\mathbf{n}} - \epsilon_i \frac{\partial\phi}{\partial\mathbf{n}} = \sigma_{int}, \tag{4.44}$$

where $\mathbf{n}$ is the vector normal to the interface, and σ_{int} is the surface charge density [164]. Figure 4.1 shows a finite difference grid in two-space dimensions near the semiconductor-insulator interface. The boundary between the two materials is assumed to be at the position of the grid point j. In the semiconductor the electrostatic potential satisfies the Poisson

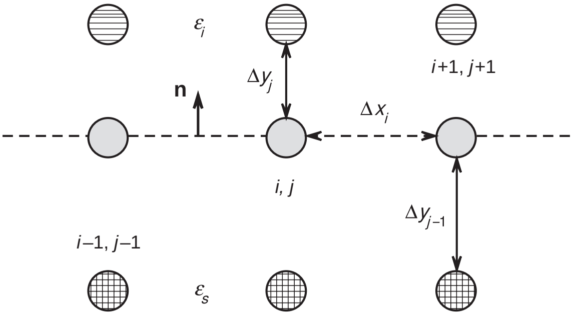


Figure 4.1: Schematic of the finite-difference grid and an interface (indicated by a dashed line) between a semiconductor and insulator materials.

equation, which at the boundary can be written in a semi-discrete form as:

$$\left.\frac{\partial^2\phi}{\partial^2 y}\right|_{i,j}^{s} + \frac{2}{\Delta x_i + \Delta x_{i+1}}\left(\frac{\phi_{i+1,j}-\phi_{i,j}}{\Delta x_i} - \frac{\phi_{i,j}-\phi_{i-1,j}}{\Delta x_{i-1}}\right) = \frac{1}{\epsilon_s}(n_{i,j}-p_{i,j}). \tag{4.45}$$

On the insulator side of the interface the Laplace equation is valid,

$$\left.\frac{\partial^2\phi}{\partial^2 y}\right|_{i,j}^{i} + \frac{2}{\Delta x_i + \Delta x_{i+1}}\left(\frac{\phi_{i+1,j}-\phi_{i,j}}{\Delta x_i} - \frac{\phi_{i,j}-\phi_{i-1,j}}{\Delta x_{i-1}}\right) = 0, \tag{4.46}$$

where a second-order accurate central finite-difference discretization is used to approximate the second derivative of ϕ. On either side of the interface a Taylor series expansion of the first-derivative of the potential along the y-direction gives:

$$\left.\frac{\partial\phi}{\partial y}\right|_{i,j}^{s} = \frac{\phi_{i,j}-\phi_{i,j-1}}{\Delta y_{j-1}} + \frac{\Delta y_{j-1}}{2}\left.\frac{\partial^2\phi}{\partial^2 y}\right|_{i,j}^{s} + O(\Delta^2), \tag{4.47}$$

$$\left.\frac{\partial\phi}{\partial y}\right|_{i,j}^{i} = \frac{\phi_{i,j+1}-\phi_{i,j}}{\Delta y_j} - \frac{\Delta y_j}{2}\left.\frac{\partial^2\phi}{\partial^2 y}\right|_{i,j}^{i} + O(\Delta^2). \tag{4.48}$$

Replacing the second-order derivatives in equations (4.47)–(4.48) by the corresponding expressions from the semi-discrete representations (4.45) and

(4.46), one obtains a second-order accurate approximation for the first derivative of ϕ on both sides of the interface. Substituting the resulting expressions for $[\partial\phi/\partial y]_{i,j}^{s,i}$ into the Neumann boundary condition given by the equation (4.44), results in a discrete interface condition:

$$\frac{2}{\Delta x_i + \Delta x_{i+1}}\left(\frac{\phi_{i+1,j}-\phi_{i,j}}{\Delta x_i} - \frac{\phi_{i,j}-\phi_{i-1,j}}{\Delta x_{i-1}}\right) + \tag{4.49}$$

$$\frac{2}{\epsilon_i \Delta y_j + \epsilon_s \Delta y_{j-1}}\left(\epsilon_i \frac{\phi_{i,j+1}-\phi_{i,j}}{\Delta y_j} - \epsilon_s \frac{\phi_{i,j}-\phi_{i,j-1}}{\Delta y_{j-1}} + \sigma_{i,j}\right) \tag{4.50}$$

$$= \frac{\Delta y_{j-1}}{\epsilon_i \Delta y_j + \epsilon_s \Delta y_{j-1}}(n_{i,j} - p_{i,j}). \tag{4.51}$$

Application of the above interface condition requires modification of the appropriate coefficients of the discretization matrix and does not introduce any additional nodes into the grid.

(b) Grid Interfaces

Boundary conditions of a different type must be applied at the grid interfaces where the numerical properties of the discrete equations, rather than the physical coefficients of the region or the solution itself, are discontinuous. One example of the grid interface is the non-conformal grid refinement, in which the grid lines are terminated before reaching the boundaries of the computation volume, Figure 4.2. The physical properties of the equations can be taken into account in order to construct a grid interface condition that is consistent with the scheme used away from the interface, and preserves the order of accuracy of the discretization. In the context of conservation laws, the physical quantity of interest is the discrete approximation of the flux function, with the boundary condition following from the requirement of the flux conservation at the grid interfaces [17]. In addition, the discrete boundary conditions applied at the grid interfaces, in general, will also change the stability properties of the method.

In this section we consider a non-conformal grid refinement applied to the Finite-Difference Time-Domain (FDTD) discretization of Maxwell's equations. In two-space dimensions, taking z-invariant solutions and uniform material properties, Maxwell's equations simplify to two uncoupled systems for the TM_x polarization (H_y, H_z, E_x) and the TE_x polarization (E_y, E_z, H_x). We consider below the TE_x subset. To advance the numerical solution in time, the FDTD method uses a grid that is staggered in both

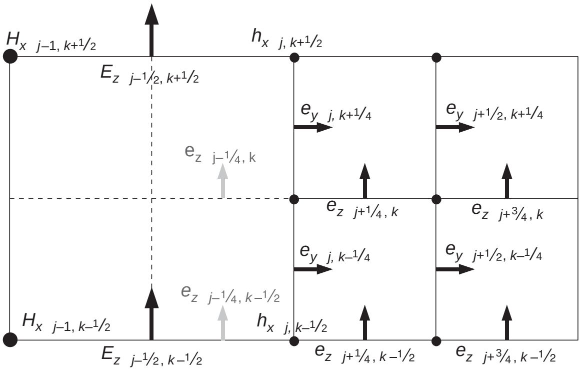


Figure 4.2: A Finite-Difference Time-Domain grid interface between coarse and fine grids, showing the location of the variables on the coarse grid ($\mathbf{E}$ and H_x) and on the fine grid ($\mathbf{e}$ and h_x). The space-time interpolated ghost boundary e_z values are shown in greyscale.

space (Figure 4.2) and time:

$$E^{n}_{y\,j,k} - E^{n-1}_{y\,j,k} = \nu_z \left(H^{n-1/2}_{x\,j,k+1/2} - H^{n-1/2}_{x\,j,k-1/2} \right), \tag{4.52}$$

$$\begin{aligned} &E^{n}_{z\,j-1/2,k-1/2} - E^{n-1}_{z\,j-1/2,k-1/2} \\ &\qquad = -\nu_y \left(H^{n-1/2}_{x\,j,k-1/2} - H^{n-1/2}_{x\,j-1,k-1/2} \right), \end{aligned} \tag{4.53}$$

$$\begin{aligned} H^{n+1/2}_{x\,j,k-1/2} - H^{n-1/2}_{x\,j,k-1/2} &= \nu_z \left(E^{n}_{y\,j,k+1/2} - E^{n}_{y\,j,k-1/2} \right) \\ &\quad - \nu_y \left(E^{n}_{z\,j+1/2,k-1/2} - E^{n}_{z\,j-1/2,k-1/2} \right), \end{aligned} \tag{4.54}$$

where ν_y, ν_z are the Courant (CFL) numbers defined as $\Delta t/\Delta y$ and $\Delta t/\Delta z$, respectively, and we use dimensionless equations in which the permeability, permittivity and the speed of light are all set to unity. Away from the interface, the above FDTD update can be applied on both the coarse and the fine grid. At the interface, the update of the fine (coarse) grid values requires a corresponding missing (or “ghost”) field value from the coarse (fine) grid, e.g. the $e_{z\,j-1/4,k-1/2}$ field is needed to update $h_{x\,j,k-1/2}$. Hence, some interpolation in space and time is required to compute the values at the “ghost” nodes, resulting in a grid interface boundary condition. For example, the value of the $e_{z\,j-1/4,k-1/2}$ field can be computed from the nearby coarse and fine grid values through a simple second-order accurate linear interpolation:

$$e^n_{z\,j-1/4,k-1/2} = \frac{2}{3}E^n_{z\,j-1/2,k-1/2} + \frac{1}{3}e^n_{z\,j+1/4,k-1/2}.$$

To illustrate the dependence of the stability properties of the scheme on the choice of the interpolation used at the grid interfaces, we apply a matrix stability analysis to a fully discrete problem on a finite domain with a single refinement patch [204]. This analysis provides a necessary and sufficient condition for stability on a given grid. It states that the method is stable if the update matrix has eigenvalues that fall within, or on, a unit circle. In addition, eigenvalues lying on the unit circle should be non-defective. Note that this analysis is more general than the spectral radii analysis, as the latter is inconclusive about the coinciding eigenvalues on the unit circle, which require consideration of the dimension of the corresponding eigenspace to rule out the mode resonances.

To compute the eigenvalues of the update matrix numerically, the solution state vector is represented by the **e**, **h** field values on the fine grid and the **E**, **H** field values in the regions of the coarse grid that do not overlap with the fine grid. Solution of a large scale eigenvalue problem using the Arnoldi process for approximating a few eigenvalues, such as the Implicitly Restarted Arnoldi Method [107], usually requires only the result of the multiplication of the matrix by a given vector, during the iterative solution process. Therefore it is possible to apply such eigenvalue solution methods without actually constructing explicitly the discrete update matrix, but instead using just the update of a solution vector by the discrete scheme.

We consider solutions of an eigenvalue problem resulting from a discretization based on a single fine grid patch of size $n_f \times n_f$ centered on a $n_c \times n_c$ coarse grid, and use two particular interpolation methods, labeled `TEQuHav` and `TEWeil`, that are described in detail in [204]. The Dirichlet boundary condition $\mathbf{E} = 0$ is applied on the outer edges of the computational domain covered by the coarse grid. This boundary condition corresponds physically to a non-dissipative reflecting boundary. For $CFL = 0.4$ and the `TEQuHav` algorithm, eigenvalues with $|\lambda| > 1$ exist for the case of $n_f = 4$ and a placement of the fine grid within one grid cell space of the coarse grid center. Additional modes with $|\lambda| > 1$ exist in the case of $n_f > 4$ and a symmetric arrangement of the fine grid with respect to the coarse grid center. Figure 4.3 shows eigenvalue distributions in the complex-plane for the case of a refined patch, $n_f = 8$, placed in the center of the coarse grid, $n_c = 10$.

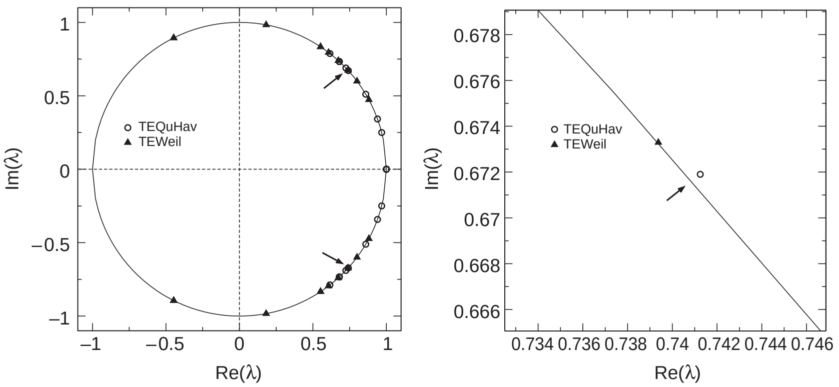


Figure 4.3: Left: Distribution of the 20 largest (in magnitude) eigenvalues on the complex plane for `TEQuHav`, `TEWeil` algorithms computed on a symmetric two-level grid with $n_c = 10$, $n_f = 8$ $CFL = 0.4$. For the `TEQuHav` algorithm arrows indicate the complex-conjugate eigenvalue pair that falls outside of the unit circle, $|\lambda| > 1$. Right: Close-up of the $|\lambda| > 1$ eigenvalue.

(c) Open Boundary Conditions

For many wave-propagation problems the PML method (see [182] and references therein) allows simulation of transparent, open boundary conditions with minimal reflections.

To demonstrate the dependence of the amount of reflection on the number of grid points used in the PML layer, we consider dimensionless Maxwell's equations for a TM_x mode in two-space dimensions for a dielectric medium with uniform permittivity and permeability:

$$\frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}, \quad \frac{\partial H_z}{\partial t} = \frac{\partial E_x}{\partial y}, \quad \frac{\partial H_y}{\partial t} = -\frac{\partial E_x}{\partial z}.$$

Figure 4.4 shows the time evolution of the electric field E_x for an initial-value problem corresponding to $H_y(y,z) = 0$, $H_z(y,z) = 0$ and $E_x(y,z) = \exp[-(y^2 + z^2)/w^2]$ at time $t = 0$. A cylindrical wave initiated in the center of the computational domain, propagates radially and becomes attenuated at the boundaries of the domain, where a UPML absorbing layer is applied. The magnitude of the reflection can be quantified by referring to Figure 4.5 that shows the time-evolution of a short optical pulse (amplitude full-width at half-maximum equal to 10 fs, carrier wavelength $\lambda_0 = 650$ nm) incident on a PML layer, and the reflected wave amplitude as a function of the PML layer width. A polynomial of order $m_{pml} = 3$ is used to compute

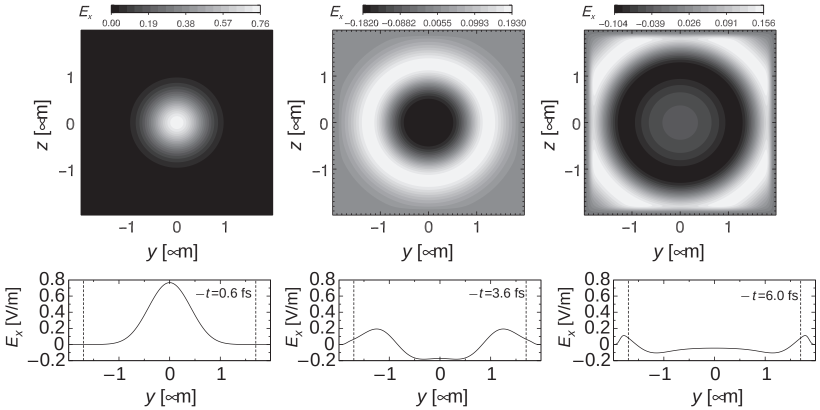


Figure 4.4: Time-snapshots of the distribution of the electric field $E_x(t, y, z)$ in a two-dimensionsal computational domain (top row) and along the line-cut at $z = 0$ (bottom row), for an initial-value problem corresponding to $H_y(0, y, z) = 0$, $H_z(0, y, z) = 0$ and $E_x(0, y, z) = \exp[-(y^2 + z^2)/w^2]$, $w = 0.5$ μm. The perfectly matched boundary layer is outlined by the dashed lines.

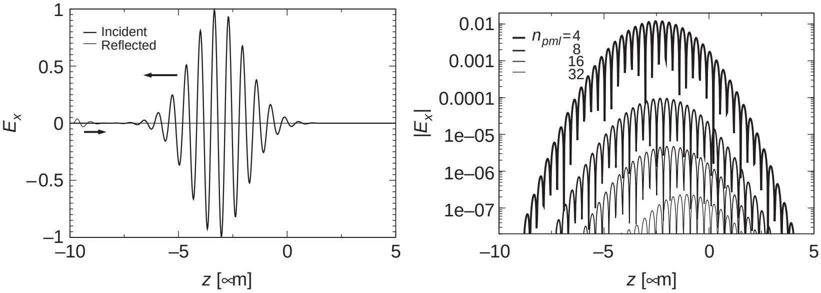


Figure 4.5: Left: Time-snapshot of a pulse of unit amplitude incident on a uniaxial perfectly matched boundary layer (UPML). Right: Distribution of the magnitude of the reflected wave for different values of the number of cells n_{pml} in a UPML layer.

the grading of the PML layers, with $\sigma_{\max} = 8 \times (m_{pml}+1)/(n_{pml}\Delta)$, where $\Delta = \lambda_0/32$ is the grid cell size in the PML region. The maximum of the amplitude of the reflected wave decreases about five orders of magnitude from 10^{-2} for $n_{pml} = 4$ to 2×10^{-7} for $n_{pml} = 32$.

Chapter 5

Problems with Multiple Temporal and Spatial Scales

5.1 Examples of Weakly and Strongly Interacting Multiple Scales

Problems with multiple disparate temporal and spatial scales are at the heart of most physical problems, as well as in mathematical modeling problems in economics and social sciences. Numerical treatment usually allows one to deal with two or, at most, three disparate scales due to limitations of computer resources. Problems with different scales arise when the interest is in the description of the problem on the larger scales, for example the motion of weather fronts in the presence of fast gravity waves, the average trajectory of a fast rotating charged particle in a magnetic field in which one averages over the fast gyro-period of the particle the behavior of society incorporating the dynamics of individuals, the role of singularities in solutions of differential equations or the role of geometry, etc. Often the proper physical description of the parameters and material properties on the microscopic level requires the solution of a large system of partial differential equations (PDEs) comparable to the number of equations needed to describe the macroscopic system. For example, Maxwell's equations of electrodynamics may require solving a many-body problem for each time iteration to determine material properties of the media.

The methods of dealing with such problems are based on either resolving (if ever possible) or not resolving the fine scales. In the latter case, asymptotic analysis of the dominant terms is often used to separate scales or compute the cumulative effect of the small scale motion on the large scale dynamics. For problems where scales are weakly coupled, numerical treatment may often produce physical results as long as the choice of the method has remnants of the smaller scale behavior such as: damping, diffusion, dispersion, etc. Under-resolved numerical computations usually never produce physical results as cumulative effects are not of the form of the truncation errors, in addition to the fact that the details of the small scale motion are important and have to be accounted for on either a theoretical or experimental basis.

Example. Stiff Ordinary Differential Equation

Consider the equation:

$$\epsilon\frac{dx(t)}{dt} = -x(t) + \cos(t), \quad x(0) = 0, \ \epsilon \ll 1.$$

There are two time scales present, the slow time scale described by t/T where $T = 1$, and the fast time scale T/ϵ. If interest is in the motion on the slow time scale and the solution evolves from an exponentially fast decay at early times to a slowly evolving solution at late times, then the problem is called *stiff.* Note, that when interest is in the transient fast motion, the problem is not stiff. Also, highly oscillatory problems, even if they may fall under the category of a multiple scales problem, are not stiff. To examine the behavior of the above system, we assume the following asymptotic expansion:

$$x(t) = x_0(t) + \epsilon x_1(t) + O(\epsilon^2).$$

Substituting this ansatz into the ordinary differential equation (ODE) and collecting zero and first order terms gives:

$$x_0(t) = \cos(t), \quad x_1(t) = -\frac{dx_0(t)}{dt} = \sin(t).$$

The exact solution of the original system is:

$$x(t) = \frac{\cos(t)}{1+\epsilon^2} + \epsilon\frac{\sin(t)}{1+\epsilon^2} - \frac{\exp(-t/\epsilon)}{1+\epsilon^2}.$$

It shows an exponentially fast decay of the solution to the motion on the slow attractor, within error $O(\epsilon)$, in the transition layer of width $O(\epsilon)$. Note that the approximate solution based on the long scale perturbation expansion to lowest order is:

$$x = x_0(t) + \epsilon x_1(t) + O(\epsilon^2) = \cos(t) + \epsilon \sin t + O(\epsilon^2),$$

which is consistent with the exact solution for large enough $t \sim O(1)$. Note however, that this approximate solution does not capture the fast initial decay due to the $-\exp(-t/\epsilon)$ term in the exact solution.

If explicit Euler or higher order Runge-Kutta methods are applied, the stability requirement would force the time stepping size to resolve the motion on the fast time scale. For example, for the explicit Euler method:

$$x_{n+1} = x_n(1 + \lambda\Delta t) - \lambda\Delta t \cos(n\Delta t), \quad \lambda = \frac{1}{\epsilon}.$$

The stability constraint is:

$$|1 + \lambda\Delta t| < 1, \quad \lambda = \frac{1}{\epsilon}.$$

For implicit Euler or higher order implicit stiff solvers, usually there are no restrictions on time step due to stability considerations, but the rapidly decaying exponential part of the solution is damped out without resolution in these methods. The only requirement for smallness of the time step is due to accuracy (resolution) requirements of the solution varying on the slow time scale. The theory for stiff ODEs states, [77]:

1. Absolutely stable linear multistep methods are implicit and first- or second-order accurate (e.g. implicit Euler and trapezoidal rule or mixture of the two, Gear's method).
2. There are implicit k-stage Runge-Kutta methods of order $2k$.
3. Backward difference formula methods are absolutely stable for systems with purely negative eigenvalues up to order 6.

A more delicate situation arises in the next problem for Burgers equation that is derived asymptotically from the full Navier-Stokes equation, [157].

Example. Shock Wave Solutions of Burgers Equation

Consider traveling waves for Burgers equation:

$$u_t + uu_x = \epsilon u_{xx}, \tag{5.1}$$

where

$$u \to u_2 \quad \text{as} \quad x \to -\infty \quad \text{and} \quad u \to u_1 \quad \text{as} \quad x \to \infty, \tag{5.2}$$

and $u_2 > u_1$, [157,198]. This solution represents a diffusively smoothed shock. The required solution is (e.g. [198, Chapter 4]):

$$u = u_1 + \frac{u_2 - u_1}{1 + \exp\theta} \equiv \frac{1}{2}(u_1 + u_2) - \frac{1}{2}(u_2 - u_1)\tanh\left(\frac{\theta}{2}\right), \tag{5.3}$$

where

$$\theta = \frac{(u_2 - u_1)}{2\epsilon}\,(x - st) \quad \text{and} \quad s = \frac{1}{2}(u_1 + u_2). \tag{5.4}$$

Here s can be identified with the shock speed, or the traveling wave speed.

Comment 1:

Whitham [198, Chapter 4, Section 4.3] also considers the initial value problem (IVP) for (5.1) with initial data:

$$u(x, 0) = u_1 H(x) + u_2 H(-x), \tag{5.5}$$

where $H(x)$ is the Heaviside step function. By using the Cole-Hopf transformation (see, e.g. [198]) the solution of the IVP (5.5) is:

$$u = u_1 + \frac{u_2 - u_1}{1 + h \exp\theta}, \tag{5.6}$$

where θ is given by (5.4):

$$h = \frac{\operatorname{erfc}[(u_1 t - x)/\sqrt{4\epsilon t}]}{\operatorname{erfc}[(x - u_2 t)/\sqrt{4\epsilon t}]}, \tag{5.7}$$

and $\operatorname{erfc}(x)$ is the complementary error function (e.g. [6]). For fixed x/t in the range $u_1 < x/t < u_2$, $h \to 1$ as $t \to \infty$ and the solution (5.6) converges to the traveling wave solution (5.3).

Comment 2:

If $u_2 = 1$ and $u_1 = -1$ then the traveling wave solution (5.3) reduces to the simple form $u(x, t) = -\tanh(x/2\epsilon)$. This solution has a characteristic length scale of $L = 2\epsilon$ for the shock transition. Thus the length of the shock transition $L \to 0$ as $\epsilon \to 0$. This observation has important consequences in numerical schemes. For example, assume that the numerical method introduces as leading order a truncation error due to numerical diffusion proportional to $\Delta x\ u_{xx}$, or a numerical dispersion proportional to $(\Delta x)^2\ u_{xxx}$, where $\epsilon \ll \Delta x$, so that we are dealing with an under-resolved computation. Clearly one should ignore the artificial width of the numerical layer proportional to $O(\Delta x)$, instead of the physically correct $O(\epsilon)$, and sharpen or reconstruct the image to a step function in the first instance if one is interested in the vanishing viscosity limit. On the other hand, if the numerical method is dispersive, the dispersive nature of the solution will generate Airy-type oscillations with Gibbs phenomena that also can be filtered out in a post-processing step to restore a visually more pleasing non-oscillatory solution. But, a zero dispersion limiting solution can be quite different from a vanishing viscosity limit, [105] and one has to be aware of the physically relevant limit.

In fact, in a great majority of practical problems, the under-resolved computations produce results that cannot be made physically correct in a simple post-processing step. A further investigation to determine whether a reasonable looking numerical solution reflects the correct physics or is just due to numerics is needed.

We conclude this section with an example of strongly interacting scales. In such examples, instead of simple additive scale separation, the cumulative effect of the fine scale dynamics defines the slow motion. This situation occurs quite often in applications. Examples are: gravity waves in meteorology determining the motion of the weather fronts, backward and forward cascades in turbulence, the fast gyrational motion of charged particles about the magnetic field in a plasma, and the slow drift of the particle guiding center across the field, due to large scale gradients in the magnetic field, as well as drifts due to electric fields and gravity. The usual remedies to deal with problems having multiple temporal and spatial scales are as follows:

1. work out asymptotically on physical or heuristic grounds reduced or slow motion equations eliminating the need to resolve the fine scale motion;
2. resolve all necessary physical scales;
3. under-resolve the smaller scales numerically, but treat them as singularities while enforcing correct physical boundary/interface relations around singularities.

The second approach is often impossible to implement due to limited computational resources. For example, to resolve the high wave numbers and short scales in the dissipation range in high Reynolds number fluid turbulence, the large scale energy containing range as well as the intermediate inertial range, is in general computationally prohibitive. For Maxwell's equations of electrodynamics, the physically correct computation of material properties may require the solution of a coupled system of three-dimensional (3D) Schrödinger equations to describe the multi-atom interactions at every time step. Currently, computation for a single atom takes several hours using parallel computer processors [96]. On the other hand, adaptive and moving grids may greatly assist in the resolution of small scales and reduction of errors near singularities. This latter approach will be pursued in the next chapter on numerical grid generation methods.

Adiabatic Motion of Charged Particles

A problem of considerable interest in plasma physics is the adiabatic motion of charged particles in slowly varying (both in space and time) electric ($\mathbf{E}$) and magnetic ($\mathbf{B}$) fields (e.g. [37,46,139,155]). This problem is naturally described as a system with multiple space and time scales. There is first of all, the small space scale r_g, the particle gyro-radius, which is assumed to be small compared to the length scale of variations of the magnetic and electric fields, L. Secondly it is assumed that the gyration time scale $T_g = 2\pi/\Omega$ of the particle about the magnetic field is small compared to the time scale T for the variations of $\mathbf{E}$ and $\mathbf{B}$, where Ω is the gyro-frequency for the particle about the magnetic field. The drift approximation effectively describes the slow time and space scale evolution of the particle guiding center drifts across and along the mean magnetic field, in which the equation of motion (the Lorentz force equation) has been averaged over the fast gyrational motion of the particle about the field, and in which the fast scale motions are coupled to the long time and space scales in the problem over many gyro-periods. The standard perturbation theory in most texts is carried out as a power series expansion in the parameter m/q, where m and q are the particle mass and electric charge, respectively, and uses dimensional units for the space, time and electromagnetic fields. This expansion is not formally a proper multiple scales perturbation expansion in which the equations have first been cast in dimensionless form, and the small dimensionless parameter is identified as $\epsilon = r_g/L$. However, it does give the correct description of the slow drift motions of the guiding center (note that the reason the dimensional expansion works is that the gyro-period is proportional to m/q).

Below, we give a brief synopsis of the formulae for the slow drift velocities $\mathbf{V}_D$ of the particle guiding centers, obtained in the standard analysis. We then give a simple example, in which we derive the drift velocity of the guiding center, in an inhomogeneous magnetic field, by using a multiple scales perturbation approach. Although more cumbersome than the usual textbook approach, using m/q as the small parameter, the multiple scales method identifies the long time and space scales in the problem, and how these scales are related to $\epsilon = r_g/L$.

Drift Velocity Formulae

The basis of the analysis is the Lorentz force equation (Newton's equation of motion) for the particle in the electric and magnetic fields, namely:

$$m\frac{d\mathbf{v}}{dt} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) + \mathbf{F}, \tag{5.8}$$

where m, q and $\mathbf{v} = d\mathbf{x}/dt$ denote the particle mass, charge, and velocity $\mathbf{x}(t)$ is the particle position, and $\mathbf{p} = m\mathbf{v}$ is the particle momentum. $\mathbf{E}$ and $\mathbf{B}$ denote the electric and magnetic fields and $\mathbf{F}$ denotes an external force (such as gravity, for which $\mathbf{F} = m\mathbf{g}$, where $\mathbf{g}$ is the acceleration due to gravity). We restrict our discussion to non-relativistically moving particles for which we can identify m as the particle rest mass and we assume that the external force $\mathbf{F}$ is negligible. Assuming r_g/L and $T_g/T = 2\pi/(\Omega T)$ are small parameters, where $\Omega = qB/m$ is the particle gyro-frequency and $r_g = p/(qB)$ is the particle gyro-radius, the guiding center drift velocity $\mathbf{V}_g$ can be written in the form:

$$\mathbf{V}_g = \mathbf{V}_{g\|} + \mathbf{V}_{g\perp}, \tag{5.9}$$

where the subscripts $\|$ and $\perp$ denote components parallel and perpendicular to the mean magnetic field. The guiding center drift velocity parallel to the magnetic field is given by:

$$\mathbf{V}_{g\|} = \left(v\cos\theta + vp\sin^2\theta\frac{\mathbf{B}\cdot\nabla\times\mathbf{B}}{2qB^3}\right)\mathbf{e}_B, \tag{5.10}$$

where $\mathbf{e}_B = \mathbf{B}/B$ is the unit vector along the field, and θ is the particle pitch angle ($\cos\theta = \mathbf{v}\cdot\mathbf{B}/vB$). The guiding center drift velocity perpendicular to the magnetic field is given by:

$$\mathbf{V}_g = \mathbf{V}_E + \mathbf{V}_{\nabla B} + \mathbf{V}_{AC}, \tag{5.11}$$

where

$$\begin{aligned}
&\mathbf{V}_E = \frac{\mathbf{E}\times\mathbf{B}}{B^2}, \quad \mathbf{V}_{\nabla B} = \frac{M}{q}\frac{\mathbf{B}\times\nabla B}{B^2},\\
&\mathbf{V}_{AC} = \frac{m\mathbf{B}}{qB^2}\times\frac{d}{dt}\left(v_\|\mathbf{e}_B + \mathbf{V}_E\right), \quad M = \frac{mv^2\sin^2\theta}{2B},\\
&\frac{d}{dt} = \frac{\partial}{\partial t} + \left(v_\|\mathbf{e}_B + \mathbf{V}_E\right)\cdot\nabla, \quad v_\| = v\cos\theta.
\end{aligned} \tag{5.12}$$

In (5.12), $\mathbf{V}_E$ is the electric field drift velocity, $\mathbf{V}_{\nabla B}$ is the grad B drift velocity, $\mathbf{V}_{AC}$ is the acceleration drift velocity, M is the adiabatically conserved magnetic moment of the particle, and d/dt is the time derivative moving with the guiding center at lowest order in ϵ. If $v_{\|} \gg V_E$, then the dominant term in the acceleration drift $\mathbf{V}_{AC}$ is the curvature drift term, $\mathbf{V}_C$, where

$$\mathbf{V}_{AC} \approx \mathbf{V}_C = \frac{m\mathbf{B}}{qB^2} \times \left(\frac{v_{\|}^2 \mathbf{B} \cdot \nabla \mathbf{B}}{B^2} \right). \tag{5.13}$$

Note that $\mathbf{V}_C$ depends on the curvature of the magnetic field. If there is a non-negligible external force $\mathbf{F}$ in the equation of motion (5.8), then the drift velocity $\mathbf{V}_F = \mathbf{F} \times \mathbf{B}/qB^2$ must be added to the right hand side of (5.11).

Example. Motion of a Charged Particle in a Slowly Varying Magnetic Field

Consider the motion of a charged particle in a slowly varying magnetic field of the form $\mathbf{B} = (0, 0, B(x, y))^t$, which depends only on two spatial coordinates x and y. The equations of motion for the particle are:

$$m\frac{d\mathbf{v}}{dt} = q\mathbf{v} \times \mathbf{B} = qB(v_y, -v_x, 0)^t, \tag{5.14}$$

where $\mathbf{v} = d\mathbf{x}/dt$ is the particle velocity. The first two equations in (5.14) may be combined to give the perpendicular energy integral:

$$\frac{1}{2}m(v_x^2 + v_y^2) = K \quad \text{or} \quad v_x^2 + v_y^2 = v_{\perp 0}^2, \tag{5.15}$$

where $v_\perp^2 = v^2 \sin^2\theta = v_x^2 + v_y^2$ is the square of the particle velocity perpendicular to $\mathbf{B}$ and θ is the particle pitch angle. Similarly, integration of the z-component of the Lorentz force equation (5.14) implies that:

$$v_z = v\cos\theta = v_{\|0}, \tag{5.16}$$

is also a constant of the motion. From (5.11)–(5.12) the guiding center drift velocity perpendicular to the magnetic field in the present example, is due solely to the grad $|B|$ drift, namely:

$$\frac{d\mathbf{x}_{g\perp}}{dt} = \mathbf{V}_{\nabla B} = \frac{pv\sin^2\theta}{2qB^3}\mathbf{B} \times \nabla B \equiv \frac{mv_\perp^2}{2qB_g^2}\left(-\frac{\partial B_g}{\partial y_g}, \frac{\partial B_g}{\partial x_g}, 0\right)^t, \tag{5.17}$$

where the subscript g denotes evaluation at the position of the guiding center. From (5.15)–(5.17) one can verify that the magnetic moment $M = mv_\perp^2/(2B_g)$ is an integral of the guiding center equations (5.17), i.e. $dM/dt = 0$. It is interesting to note that the x and y components of the drift velocity equations (5.17) can be cast in the Hamiltonian form:

$$\frac{dx_g}{dt} = \frac{\partial M}{\partial y_g}, \quad \frac{dy_g}{dt} = -\frac{\partial M}{\partial x_g}, \tag{5.18}$$

where the magnetic moment M is the conserved Hamiltonian, and (x_g, y_g) are the canonical coordinates. The parallel guiding center velocity component $v_z = v\cos\theta$ is also a constant of the motion.

Comment:

The guiding center drift equations (5.17) in the (x_g, y_g) plane are integrable. Since $v_\perp^2$, m are constants in the drift equations, then the integral $M = \text{const.}$ is equivalent to $B_g(x_g, y_g) = \text{const.}$ The first equation in (5.17) for dx_g/dt may then be integrated in the form:

$$t = -\frac{qB_g}{M}\int^{x_g} \frac{dx'_g}{\partial B_g(x'_g, y'_g)/\partial y'_g} + c_2, \tag{5.19}$$

where c_2 is an integration constant. Here $y_g = f(x_g, c_1)$ is a function of x_g, obtained by solving the first integral $B_g(x_g, y_g) = c_1$ for y_g, where we assume that the implicit function theorem applies.

Multiple Scales Approach

Although the drift equations (5.17) follow in a straightforward fashion from the standard guiding center drift formulae (5.9)–(5.13), they do not show explicitly the time scales involved. Below, we reconsider the example in (5.14) et seq., using a multiple scales perturbation analysis of the Lorentz force equation (5.14). The first step is to introduce dimensionless space and time variables and dimensionless physical variables as follows:

$$\begin{aligned} &\bar{\mathbf{x}} = \frac{\mathbf{x}}{L}, \quad \bar{t} = \Omega_0 t, \quad \Omega_0 = \frac{qB_0}{m}, \\ &\bar{\mathbf{B}} = \frac{\mathbf{B}}{B_0}, \quad \bar{\mathbf{v}} = \frac{\mathbf{v}}{U_0}, \quad \tau = \epsilon^2\bar{t}, \quad \epsilon = \frac{U_0}{\Omega_0 L_0} \equiv \frac{r_{g0}}{L_0}. \end{aligned} \tag{5.20}$$

In (5.20) L_0 is a characteristic length scale for the variation of the slowly varying background magnetic field, $\Omega_0 = qB_0/m$ is a characteristic frequency for particle gyration about the magnetic field. We set $U_0 = v_{\perp 0}$ as the perpendicular speed of the particle about the magnetic field given in (5.15). The parameter $\epsilon = r_g/L_0 \equiv U_0/(\Omega_0 L_0)$ is the small parameter in the problem; $\bar{t}$ is the fast time variable, associated with gyration of the particle about the field; and $\tau = \epsilon^2\bar{t} \sim O(1)$ defines the long time scale, on which the cumulative effects of the fast gyrational motion lead to a significant drift of particles relative to the field. The normalized particle velocity components normal to the field are of the form:

$$\bar{v}_x = \frac{v_x}{v_{\perp 0}} = \cos\phi, \quad \bar{v}_y = \frac{v_y}{v_{\perp 0}} = \sin\phi, \tag{5.21}$$

where ϕ is the gyration phase of the particle. We obtain:

$$\frac{d\bar{x}}{d\bar{t}} = \epsilon\cos\phi, \quad \frac{d\bar{y}}{d\bar{t}} = \epsilon\sin\phi, \tag{5.22}$$

$$\frac{d\phi}{d\bar{t}} = -\bar{B}. \tag{5.23}$$

In the following analysis we drop over-bars on dimensionless quantities in order to simplify the notation. The two equations (5.22) follow from the definitions of the normalized transverse velocities, whereas (5.23) is equivalent to the Lorentz force equation. We use complex, transverse spatial variables $x^{\pm} = x \pm iy$ to carry out the analysis and use expansions for $x^{\pm}$ of the form:

$$x^{\pm} = x_g^{\pm}(\tau) + \epsilon x_1^{\pm}(\bar{t}) + \epsilon^2 x_2^{\pm}(\bar{t}) + \cdots. \tag{5.24}$$

Note that the guiding center term $x_g^{\pm}(\tau)$ depends only on the slow time variable, whereas $x_1^{\pm}$ and $x_2^{\pm}$ depend on the fast time variable $\bar{t}$ associated with fast gyration of the particle about the field. The magnetic field is expanded as a Taylor series in the form:

$$\bar{\mathbf{B}} = \mathbf{B}_g[\mathbf{x}_g(\tau)] + \epsilon\left(x_1^{+}(\bar{t})\frac{\partial}{\partial x_g^{+}} + x_1^{-}(\bar{t})\frac{\partial}{\partial x_g^{-}}\right)B_g + O(\epsilon^2), \tag{5.25}$$

where B_g is the value of the magnetic induction at the location of the guiding center. Equations (5.22), written in terms of $x^{\pm}$ assume the form:

$$\frac{d\bar{x}^{\pm}}{d\bar{t}} = \epsilon\exp(\pm i\phi). \tag{5.26}$$

Integrating (5.23) using the expansion (5.25) for $\bar{B}$, we obtain:

$$\begin{aligned}\phi &= -B_g\bar{t} + \epsilon\phi_1(\bar{t}) + O(\epsilon^2),\\ \phi_1 &= -\left(\frac{\partial B_g}{\partial x_g^{+}}\int_0^{\bar{t}} x_1^{+}(t')dt' + \frac{\partial B_g}{\partial x_g^{-}}\int_0^{\bar{t}} x_1^{-}(t')dt'\right),\end{aligned} \tag{5.27}$$

as the solution for $\phi(t)$ accurate to $O(\epsilon^2)$. Also (5.26) can be expanded as a power series in ϵ to give balance equations at $O(\epsilon)$ and $O(\epsilon^2)$:

$$\frac{dx_1^+}{d\bar{t}} = \exp(-iB_g\bar{t}), \tag{5.28}$$

$$\frac{dx_g^+}{d\tau} + \frac{dx_2^+}{d\bar{t}} = i\phi_1(t)\exp(-iB_g\bar{t}). \tag{5.29}$$

Integration of (5.28) with respect to $\bar{t}$ gives the solution:

$$x_1^+(\bar{t}) = \frac{\exp(-iB_g\bar{t})}{-iB_g}, \tag{5.30}$$

for $x_1^+(\bar{t})$. Substituting (5.30) for $x_1^+(\bar{t})$ in (5.27) and carrying out the integrations over t' gives the expression:

$$\phi_1 = -\frac{1}{B_g^2}\left(\frac{\partial B_g}{\partial x_g^+}\left[1-\exp(-iB_g\bar{t})\right] + \frac{\partial B_g}{\partial x_g^-}\left[1-\exp(iB_g\bar{t})\right]\right), \tag{5.31}$$

for $\phi_1(\bar{t})$. Next, use (5.31) for ϕ_1 in (5.29) and integrate (5.29) over one period of $\bar{t}$, i.e. from $\bar{t}=0$ to $\bar{t}=T_g=2\pi/B_g$ and assuming that $x_2(\bar{t})$ has zero integral over one period, we obtain the compatibility condition:

$$\frac{dx_g^+}{d\tau} = \frac{i}{T_g}\int_0^{T_g}\phi_1(t')\exp(-iB_gt')dt', \tag{5.32}$$

for (5.29) to have a consistent perturbation solution (note that condition (5.32) ensures that the solution for x_2 does not grow linearly on the fast time scale $\bar{t}$). Substituting (5.31) for ϕ_1 in (5.32) and evaluating the source term, the compatibility condition (5.32) reduces to the equation:

$$\frac{dx_g^+}{d\tau} = \frac{i}{B_g}\frac{\partial B_g}{\partial x_g^-}, \tag{5.33}$$

for the long time evolution of $x_g^+(\tau)$.

Introducing again the over-bar notation for normalized variables, and using $\bar{x}_g$ and $\bar{y}_g$ as the dependent variables (recall $\bar{x}_g^\pm = \bar{x}_g \pm i\bar{y}_g$), (5.33) splits into the two equations:

$$\frac{d\bar{x}_g}{d\tau} = -\frac{1}{2\bar{B}_g^2}\frac{\partial\bar{B}_g}{\partial\bar{y}_g}, \quad \frac{d\bar{y}_g}{d\tau} = \frac{1}{2\bar{B}_g^2}\frac{\partial\bar{B}_g}{\partial\bar{x}_g}. \tag{5.34}$$

The system can also be written in the Hamiltonian form:

$$\frac{d\bar{x}_g}{d\tau} = \frac{\partial H}{\partial \bar{y}_g}, \quad \frac{d\bar{y}_g}{d\tau} = -\frac{\partial H}{\partial \bar{x}_g}, \quad H = \frac{1}{2\bar{B}_g}, \tag{5.35}$$

where $H = 1/(2\bar{B}_g)$ is the Hamiltonian.

Equations (5.34) and (5.35) are, in fact, the dimensionless form of the guiding center drift equations (5.17) and (5.18) associated with grad $|B|$ drifts. Using the transformations (5.20) to convert (5.34) to dimensional form gives the grad B drift equations (5.17). Similarly, using (5.20), (5.35) reduces to the dimensional Hamiltonian equations (5.18) governing the guiding center drifts. Note that the dimensionless Hamiltonian $H = 1/(2\bar{B}_g)$ corresponds to the magnetic moment $M = mv_\perp^2/(2B_g)$ used in the dimensional Hamiltonian drift equations (5.18).

The main points to emerge from the multiple scales derivation of the guiding center drift equations (5.34)–(5.35) are: (1) the demonstration that the particle drifts occur on the long time scale $T_L = T_g/\epsilon^2$, where T_g is the gyro-period and $\epsilon = r_g/L_0 = v_{\perp_0}/(\Omega_0 L_0)$ is the perturbation parameter; (2) the solution for the gyro-phase ϕ in (5.27) depends on both the fast ($\bar{t}$) and long time variables (τ); (3) the gradient drifts arise from the fact that the particle samples the inhomogeneous magnetic field during its gyro-orbit about the field, and that this field is different from that at the position of the gyro-center. It is the cumulative effects of the non-homogeneous magnetic fields on the particle orbit over many gyro-orbits that give rise to the drifts.

Note that explicit numerical methods on the full problem would lead to small time steps, while the stiff and symplectic time integrators discussed in the next sections would perform as follows. The stiff implicit solvers would damp the energy in a conservative system and eventually spiral the circular motion into a single point. The explicit symplectic integrators can be designed to preserve energy, momentum and symplectic structure of the motion, but that would not exempt them from the requirement to resolve the fast time scale motion. For example, in [168], the authors compared performance of the explicit, implicit midpoint-type method (involving several Newton's iterations per time step) and linearly implicit (involving only a single Newton's iteration per time step) symplectic integrators on a model problem mathematically similar to the problem of motion of the charged particle. They observed that the fully implicit method is as costly as the explicit method to achieve the same accuracy, but the linearly implicit method was several times less expensive. On the other hand, while all three methods preserved energy to the same extent, the position error of the linearly implicit method was considerably larger in comparison to the other two methods.

5.2 Stiff Ordinary Differential Equation Solvers

Recall that a problem is called stiff if in a multiple scales problem, the interest is in the motion on the slow scale and the solution decays exponentially fast toward the motion on the slow scale. Mathematically, the stiffness of the system usually manifests itself when the linearized problem has eigenvalues with large negative real parts. Neutral and slowly growing modes are allowed.

Example. Stiff Linear Conservative System

Consider a linear system of ODEs, [166],

$$\begin{aligned}\dot{x}_1 &= -k_1 x_1 + k_2 x_3,\\ \dot{x}_2 &= -k_4 x_2 + k_3 x_3,\\ \dot{x}_3 &= k_1 x_1 + k_4 x_2 - (k_2 + k_3) x_3.\end{aligned}$$

The system conserves total mass in time, $x_1+x_2+x_3 = \text{const}$, as can be seen by summing the above equations. For $(k_1, k_2, k_3, k_4) = (8.43 \times 10^{-10}, 2.90 \times 10^{11}, 2.46 \times 10^{10}, 8.76 \times 10^{-6})$ and initial condition $(x_1, x_2, x_3) = (0, 1, 0)$, the exact solution that can be explicitly computed contains a top hat function in the second components on the order of 10^{-17}, [166]. If the absolute error tolerance of the ODE solver is left at the default value of 10^{-6}, the ODE solvers "converge" to an incorrect but reasonable looking solution. Experimenting with the error tolerance, one observes that convergence to the correct solution near zero occurs when the absolute error tolerance is of the order of the solution itself or smaller. Without physical insight, an asymptotic estimate or a rigorous mathematical estimate, one cannot claim that the solution lies in the "convergence regime", which is defined heuristically as "everything (meaning physical scales of importance) that needs to be resolved is resolved".

Example. Stiff Nonlinear Conservative System

Consider the nonlinear system of ODEs, [166],

$$\begin{aligned}\dot{x}_1 &= -k_1 x_1 + k_2 x_2 x_3,\\ \dot{x}_2 &= k_1 x_1 - k_2 x_2 x_3 - k_3 x_2^2,\\ \dot{x}_3 &= k_3 x_2^2,\end{aligned}$$

with $(k_1, k_2, k_3) = (0.04, 10^4, 3 \times 10^7)$ and initial condition $(x_1, x_2, x_3) = (1, 0, 0)$. The system conserves total "mass", i.e. $x_1 + x_2 + x_3 = \text{const}$, as can be seen by summing the above equations. In addition, it can be shown that the solution remains nonnegative for all time. Setting the right hand side to zero gives $x_1 = x_2 = 0$, while x_3 is arbitrary. The mass conservation and initial condition imply that $x_3 = 1$. One of the numerical methods reported in [166] produces small negative values for one of the components that starts to grow and the solution blows-up. Note, that the conservation property by itself may not prevent the growth of large negative and positive components which may cancel each other.

The next two stiff variable coefficient and nonlinear stiff ODE examples show that numerical methods well-suited for stiff linear systems may or may not preserve their unconditional damping properties of the modes, depending on the damping coefficient λ, where $\text{Re}(\lambda) < 0$ in the linear damping case, when the solution evolves outside its original region of validity.

Example. An Ordinary Differential Equation with Variable Decay Rate

Consider the following differential equation, [7],

$$\dot{y} = \lambda(t) y,$$

where $\lambda(t) < 0$. It is easy to check that the implicit Euler and midpoint rule:

$$\begin{aligned} y_{n+1} &= y_n + \Delta t \lambda_{n+1} y_{n+1}, \\ y_{n+1} &= y_n + \frac{\Delta t}{2} \lambda_{n+1/2} \frac{y_{n+1} + y_n}{2}, \end{aligned}$$

will preserve the damping property, while the trapezoidal rule:

$$y_{n+1} = y_n + \frac{\Delta t}{2} (\lambda_{n+1} y_{n+1} + \lambda_n y_n),$$

will require additional constraints on the derivative of $\lambda(t)$. Let the amplification factor $\rho(t_n, t_{n+1})$ be defined as:

$$\rho(t_n, t_{n+1}) = \frac{1 + \frac{1}{2}\Delta t \lambda_n}{1 - \frac{1}{2}\Delta t \lambda_{n+1}}.$$

Then $|\rho(t_n, t_{n+1})| < 1$, implies that $\Delta t < \frac{4}{\lambda_{n+1} - \lambda_n}$ under the original assumption that $\lambda(t) < 0$ and $\lambda_{n+1} > \lambda_n$ (i.e. $(\lambda_{n+1} - \lambda_n)/\Delta t < 4$).

Example. Stiff Nonlinear System

Consider the nonlinear ODE, [7],

$$y'' + a(y)y' + (1 + b(y)\cos(2\pi x))\, y = 0,$$

linearized around constant state y_0:

$$y'' + a(y_0)y' + (1 + b(y_0)\cos(2\pi x))\, y = 0.$$

It was shown in [7], that the characteristic roots of the linearized equation have negative real parts if $a(y_0) > 0$ and $b(y_0) < 1$, but the original nonlinear system has unbounded solutions. An application of stiff solvers would have no advantage in this case.

Example. Method of Lines for the Heat Equation

Consider an initial-boundary value problem for the heat equation:

$$u_t = \epsilon u_{xx},$$

on the interval $[0, \pi]$ with zero boundary conditions at $x = 0$ and $x = \pi$. Discretizing the space variable x on a uniform mesh, $x_j = j\pi/N$, $j = 0, 1, \ldots, N$, and using centered differencing results in a linear system of ODEs:

$$\frac{dU(t)}{dt} = AU,$$

where $U = (u_1(t), u_2(t), \ldots, u_{N-1}(t))^t$ denotes the unknown values at given spatial locations, $u_j(t) \approx u(x_j, t)$. The matrix A is a symmetric tridiagonal matrix with diagonal elements $a = \frac{-2\epsilon}{(\Delta x)^2}$ and off diagonal elements $b = c = \frac{\epsilon}{(\Delta x)^2}$. Using an explicit formula for the eigenvalues of a tridiagonal matrix A with the following structure, [169],

$$A = \begin{pmatrix} a & b & 0 & \cdots & 0 \\ c & a & b & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & c & a & b \\ 0 & \cdots & 0 & c & a \end{pmatrix},$$

$$\lambda_j = a + 2\sqrt{bc}\cos\frac{\pi j}{N}, \quad 1 \le j \le N-1.$$

In our case, the formula gives, $\lambda_j = \epsilon \frac{-4}{(\Delta x)^2} \sin^2(\frac{j\pi}{2N})$. The eigenvalues are monotonically distributed and for small Δx large (N) the extreme eigenvalues can be approximated as $\lambda_{N-1} \sim -\frac{4\epsilon}{(\Delta x)^2}$ and $\lambda_1 \sim -\epsilon$. The large negative eigenvalues in the discrete problem are numerical artefacts, and carry no physical significance. For this example, application of stiff solvers will rapidly damp out the high wavenumber modes in analogy to the continuous case. In particular, applying the trapezoidal rule results in the popular Crank-Nicolson approximation. In this case, stiff solvers allow one to keep the time step according to the desired resolution and accuracy, e.g. $\Delta t \sim \Delta x$, while explicit numerical integrators would require unnecessarily small time stepping to satisfy the stability restrictions, $\Delta t \sim (\Delta x)^2$. For example, the explicit Euler method would require:

$$\left| 1 + \Delta t \frac{(-4\epsilon)}{(\Delta x)^2} \right| < 1,$$

or $\epsilon \frac{\Delta t}{(\Delta x)^2} < \frac{1}{2}$, in order for all numerical modes to decay to zero.

5.3 Long-Time Integrators for Hamiltonian Systems

Many PDEs possess symmetries and invariants due to underlying physical laws. In order to increase accuracy and efficiency of numerical approximations it is often advantageous to design numerical methods that preserve some of the key symmetries and invariants. This property might be the physical energy, amplitude, momentum, divergence, total charge, etc. Occasionally, the added benefit of such construction is the guaranteed stability of the method.

Example

Consider the complex ODE:

$$\frac{dz(t)}{dt} = -iz \quad \text{where } z = u + iv, \tag{5.36}$$

describing the motion of a particle moving in a circle. Alternatively, the equations of motion (5.36) can be written in terms of u and v in the form:

$$\frac{du(t)}{dt} = v, \quad \frac{dv(t)}{dt} = -u. \tag{5.37}$$

As already mentioned in Sections 5.1 and 5.2, explicit and implicit Euler methods will spiral the particle in or out of the correct physical circular orbit, while the trapezoidal rule:

$$\begin{aligned} u_{n+1} &= \frac{1}{2}(v_{n+1}+v_n)\Delta t + u_n, \\ v_{n+1} &= -\frac{1}{2}(u_{n+1}+u_n)\Delta t + v_n, \end{aligned}$$

besides being of second-order accuracy, also preserves the amplitude and the energy of the motion. The second-order error is due to the phase error of the motion. This example gives a simple illustration of an amplitude and energy preserving method, but if one is interested in accurate phase computation, then the standard higher order methods are more accurate and efficient.

In the simplest, finite dimensional case, Hamilton's equations are:

$$\frac{dq(t)}{dt} = \frac{\partial H(q,p)}{\partial p}, \quad \frac{dp(t)}{dt} = -\frac{\partial H(q,p)}{\partial q}, \tag{5.38}$$

where the Hamiltonian H is a scalar function of n-dimensional column vectors q and p. In the example of (5.37), the Hamiltonian is $H(u,v) = (u^2+v^2)/2$, and (u,v) are the canonical coordinates.

In this section we first give an overview of both finite dimensional and infinite dimensional Hamiltonian systems. Marsden and Ratiu [124] and Olver [140] give good introductions. Useful mathematical background material is given by Abraham et al. [5] and Sattinger and Weaver [161]. Holm et al. [85] review Euler-Poincaré and semi-direct product Lie algebras with applications to Lagrangian and Hamiltonian systems. Section 5.3.2 gives a description of multi-symplectic Hamiltonian systems. An overview of numerical integration techniques for Hamiltonian systems is given in Section 5.3.3. Our aim is not to be exhaustive, but to give an indication of the types of methods and applications.

5.3.1 Overview of Hamiltonian Systems

Finite dimensional Hamiltonian systems may be written in several different, but equivalent forms. The symplectic matrix form of Hamilton's equations is:

$$\frac{dz}{dt} = J\nabla H, \tag{5.39}$$

where

$$z = \begin{pmatrix} q \\ p \end{pmatrix}, \quad \text{and} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{5.40}$$

Here $z = (q, p)^t$ is a $(2n)$-dimensional vector describing the system, I is the $n \times n$ unit matrix, and J is an anti-symmetric or skew-symmetric matrix known as the symplectic operator, with the properties $J^t = -J = J^{-1}$.

Using the complex variable $w = q + ip$, Hamilton's equations (5.38) can be written in the form:

$$\frac{dw}{dt} = -2i\frac{\partial H}{\partial w^*}, \quad \frac{dw^*}{dt} = 2i\frac{\partial H}{\partial w}, \tag{5.41}$$

where the superscript $*$ in w^* denotes the complex conjugate of w.

The Poisson bracket for the system is defined as:

$$\{F, G\} = \sum_{k=1}^{n} \frac{\partial F}{\partial q^k}\frac{\partial G}{\partial p_k} - \frac{\partial F}{\partial p_k}\frac{\partial G}{\partial q^k} \equiv F_z^t J G_z. \tag{5.42}$$

Using the Poisson bracket (5.42), Hamilton's equations can be written in the form:

$$\dot{z} = \{z, H\}. \tag{5.43}$$

More generally, the time rate of change of any functional F of the canonical variables $(q, p)^t$ is given by the Poisson bracket equation:

$$\dot{F} = \{F, H\} = \hat{V}_H(F), \tag{5.44}$$

where the Hamiltonian evolution operator (or vector field) $\hat{V}_H$ is defined as:

$$\hat{V}_H = \frac{\partial H}{\partial p}\frac{\partial}{\partial q} - \frac{\partial H}{\partial q}\frac{\partial}{\partial p}. \tag{5.45}$$

In terms of $\hat{V}_H$ the formal solution of Hamilton's equations (5.43) for z is given by the equation:

$$z(t) = \exp\left(\int_0^t \hat{V}_H dt\right) z(0). \tag{5.46}$$

If H does not depend explictly on t, then $z(t) = \exp(t\hat{V}_H)z(0)$. If one uses the complex variables $w = q + ip$ to define the Hamiltonian system, then

$$\hat{V}_H = iL \quad \text{where } L = 2\left(\frac{\partial H}{\partial w}\frac{\partial}{\partial w^*} - \frac{\partial H}{\partial w^*}\frac{\partial}{\partial w}\right), \tag{5.47}$$

is known as the Liouville operator. One important question that can arise in this regard is whether or not the map $z(0) \to z(t)$ is one-to-one. Clearly, problems can arise if the Jacobian of the map $\partial z(t)/\partial z(0)$ vanishes, or becomes unbounded at singularities of the system.

A fundamental property of Hamiltonian systems is that the phase space velocity $(\dot{q}, \dot{p})$ is incompressible:

$$\frac{\partial \dot{q}}{\partial q} + \frac{\partial \dot{p}}{\partial p} = H_{pq} - H_{qp} = 0, \tag{5.48}$$

where $\dot{\psi} = d\psi/dt$. This result is related to Liouville's theorem in statistical mechanics. It is also related to the conservation of the phase space volume element $dp \wedge dq$ following the flow:

$$\begin{aligned} \frac{d}{dt}\left(dp \wedge dq\right) &= d\dot{p} \wedge dq + dp \wedge d\dot{q} \\ &= d(-H_q) \wedge dq + dp \wedge d(H_p) \\ &= (-H_{qp} + H_{pq})\, dp \wedge dq = 0, \end{aligned} \tag{5.49}$$

(note $dp \wedge dp = dq \wedge dq = 0$ in the algebra of exterior differential forms). The conservation of the volume element of the phase space volume element, can in turn be related to the invariance of the first Poincaré invariant:

$$\frac{d}{dt}\left(\oint_{C(t)} pdq\right) = 0, \tag{5.50}$$

where the integral is over a closed path $C(t)$ moving with the flow. The latter result is a consequence of Stokes theorem: $\int_A dp \wedge dq = \oint_{C(t)} p\, dq$, where A is the area element encompassing the closed curve $C(t)$ (note $d(pdq) = dp \wedge dq$). There are, in fact, higher order Poincaré invariants that play an important role in Hamiltonian systems, that have topological significance [11,12].

If we restrict our attention to smooth functions F, H on a manifold M such that $F, H : M \to R$ are maps onto the reals R, then the Poisson bracket must have the basic properties [140]:

(a) Bilinearity

$$\{cF + c'P, H\} = c\{F, H\} + c'\{P, H\}, \tag{5.51}$$

$$\{F, cH + c'P\} = c\{F, H\} + c'\{F, P\}, \tag{5.52}$$

(b) Skew symmetry

$$\{F,H\}=-\{H,F\}, \tag{5.53}$$

(c) Jacobi identity

$$\{\{F,H\},P\}+\{\{H,P\},F\}+\{\{P,F\},H\}=0, \tag{5.54}$$

(d) The Leibniz rule (i.e. differentiation rule)

$$\{F,HP\}=\{F,H\}P+H\{F,P\}. \tag{5.55}$$

Comment:

The Hamiltonian vector fields $\hat{V}_H$ defined in (5.44)–(5.45) such that $\hat{V}_H(F)=\{F,H\}$ have Lie bracket:

$$[\hat{V}_H,\hat{V}_F]=\hat{V}_H\hat{V}_F-\hat{V}_F\hat{V}_H, \tag{5.56}$$

satisfying the equation:

$$[\hat{V}_H,\hat{V}_F]=\hat{V}_{\{F,H\}}. \tag{5.57}$$

The detailed proof of (5.57) follows by proving that $[\hat{V}_H,\hat{V}_F]P=\hat{V}_{\{F,H\}}P$. The latter result follows by using the definition (5.56) of the Lie bracket and by using the Jacobi identity (5.54). Thus there is a map between the Lie bracket commutators $[\hat{V}_H,\hat{V}_F]$ and the corresponding Poisson bracket $\{F,H\}$ and a one-to-one correspondence between the Poisson bracket Lie algebra and the Lie algebra of Hamiltonian vector fields induced by the map (5.57).

Example

Consider Newton's equations of motion for N particles in a potential force field $\mathbf{F}=-\partial V/\partial\mathbf{x}^{(j)}$ where $V(\mathbf{x})$ is the potential, i.e.:

$$\begin{aligned}&\frac{d\mathbf{x}^{(j)}(t)}{dt}=\mathbf{v}^{(j)}(t),\\ &m_j\frac{d\mathbf{v}^{(j)}(t)}{dt}=\mathbf{F}^{(j)}=-\nabla_{\mathbf{x}^{(j)}}V(x),\quad 1\leq j\leq N,\end{aligned} \tag{5.58}$$

with molecular dynamics and N-body gravitational problems as particular cases. The Hamiltonian for the system is:

$$H=\sum_{j=1}^{N}\frac{1}{2}m_j|\mathbf{v}^{(j)}|^2+V(\mathbf{x})=\sum_{j=1}^{N}\frac{p_j^2}{2m_j}+V(\mathbf{x}), \tag{5.59}$$

where $p_j = m_j|\mathbf{v}|$ is the momentum of the $j^{(th)}$ particle. Hamilton's equations $d\mathbf{x}^{(j)}/dt = \partial H/\partial\mathbf{p}^{(j)}$ and $d\mathbf{p}^{(j)}/dt = -\partial H/\partial\mathbf{x}^{(j)}$ are equivalent to Newton's equations (5.58).

Generalizations

For continuous systems, such as ideal fluid mechanics, magnetohydrodynamics (MHD) and multi-fluid plasmas, the canonical Poisson bracket has the form:

$$\begin{aligned}\{F,G\}&=\int d^3x\sum_{j=1}^{N}\left(\frac{\delta F}{\delta q_j}\frac{\delta G}{\delta p_j}-\frac{\delta F}{\delta p_j}\frac{\delta G}{\delta q_j}\right)\\&\equiv\int d^3x\left(\frac{\delta F}{\delta z}\right)^t J\frac{\delta G}{\delta z},\end{aligned} \tag{5.60}$$

where the $\{q_j\}$ and $\{p_j\}$ are the canonical variables. In (5.60) $\delta F/\delta z$ is the variational derivative, or Frechet derivative of the functional F with respect to z (the variational derivative is also sometimes referred to as the Gateaux derivative; there are, in fact, technical differences between the Frechet and Gateaux derivative, which we will not go into here), $z = (q,p)^t$ and J is the symplectic matrix or operator (5.40). The symplectic inner product for the bracket (5.60) is defined by the equation:

$$\langle\mathbf{u},\mathbf{v}\rangle=\int d^3x\ \mathbf{u}^t\mathbf{J}\mathbf{v}, \tag{5.61}$$

where $\mathbf{u}$ and $\mathbf{v}$ are N-dimensional vectors.

Example. Ideal Gas Dynamics

For barotropic, potential flow of an ideal gas, the velocity of the fluid $\mathbf{v} = \nabla\Phi$ where Φ is the velocity potential, and the internal energy of the gas per unit volume $\varepsilon = \varepsilon(\rho)$ is independent of the entropy S (i.e. the entropy is constant). For an ideal polytropic gas with adiabatic index γ, $\varepsilon = p/(\gamma - 1)$ where the gas pressure $p \propto \rho^\gamma$. The basic equations for the system reduce to the mass continuity equation:

$$\frac{\partial \rho}{\partial t}+\nabla\cdot(\rho\nabla\Phi)=0, \tag{5.62}$$

and the Bernoulli equation:

$$\frac{\partial \Phi}{\partial t}+\frac{1}{2}|\nabla\Phi|^2+W=0, \tag{5.63}$$

where $W=\partial\varepsilon(\rho)/\partial\rho$ is the gas enthalpy (for a gas with a constant adiabatic index γ, $W=\gamma p/[(\gamma-1)\rho]$). Equations (5.62)–(5.63) are equivalent to the Hamiltonian equations:

$$\frac{\partial \rho}{\partial t}=\{\rho,H\}=\frac{\delta H}{\delta\Phi},\quad \frac{\partial \Phi}{\partial t}=\{\Phi,H\}=-\frac{\delta H}{\delta\rho}, \tag{5.64}$$

where

$$H=\int\left(\frac{1}{2}\rho|\nabla\Phi|^2+\varepsilon(\rho)\right)d^3x, \tag{5.65}$$

is the Hamiltonian functional and:

$$\{F,G\}=\int d^3x\left(\frac{\delta F}{\delta\rho}\frac{\delta G}{\delta\Phi}-\frac{\delta F}{\delta\Phi}\frac{\delta G}{\delta\rho}\right), \tag{5.66}$$

is the Poisson bracket. The velocity representation $\mathbf{v}=\nabla\Phi$ is known as a Clebsch representation, where Φ is the Clebsch potential. The above flow is irrotational. More general formulations for solenoidal and non-isentropic flow, using Clebsch potentials are discussed by Zakharov and Kuznetsov [207].

The Korteweg de Vries (KdV) Equation

There are other generalizations of the Poisson bracket, and the symplectic operator that describe nonlinear waves and soliton equations. The Korteweg de Vries (KdV) equation:

$$u_t+auu_x+u_{xxx}=0, \tag{5.67}$$

is an integrable, bi-Hamiltonian system [122]. It admits an infinite number of conservation laws and can be solved by the inverse scattering transform (e.g. [2]). Equation (5.67) can be written in the conservation form:

$$u_t + D_x\left(\frac{1}{2}au^2 + u_{xx}\right) = 0, \tag{5.68}$$

where $D_x \equiv \partial/\partial x$. Equation (5.68) can be written in the Hamiltonian form:

$$u_t + L_u\left(\frac{\delta H_2}{\delta u}\right) = 0, \tag{5.69}$$

where $L_u \equiv -D_x$ is the skew-symmetric, symplectic operator and:

$$H_2 = \int_{-\infty}^{\infty} dx\left(\frac{1}{2}u_x^2 - \frac{au^3}{6}\right), \tag{5.70}$$

is the Hamiltonian functional. In (5.69)–(5.70) we use the inner product:

$$\langle u, v\rangle = \int_{-\infty}^{\infty} dx\ uv. \tag{5.71}$$

Assuming both u and v vanish at infinity, one finds from (5.71) that $\langle u, D_x v\rangle = -\langle v, D_x u\rangle$. Thus the adjoint operator $D_x^\dagger = -D_x$, and D_x is a skew adjoint operator with respect to the inner product (5.71).

The Poisson bracket for the KdV equation, due to [70] is:

$$\{F, G\} = -\int_{-\infty}^{\infty} \frac{\delta F}{\delta u} L_u\left(\frac{\delta G}{\delta u}\right) dx. \tag{5.72}$$

In this formulation $u(x, t)$ is the functional defined by the equation:

$$u(x, t) = \int_{-\infty}^{\infty} u(x', t)\delta(x' - x)dx', \tag{5.73}$$

where $\delta(x' - x)$ is the Dirac delta distribution. Thus $\delta u/\delta u' = \delta(x' - x)$ and $u_t = \{u, H\} = -L_u(\delta H/\delta u)$ where $L_u = -D_x$.

A second Hamiltonian form for the KdV equation is:

$$u_t + M_u\left(\frac{\delta H_1}{\delta u}\right) = 0, \tag{5.74}$$

where

$$H_1 = \int_{-\infty}^{\infty} \frac{1}{2}u^2 dx, \tag{5.75}$$

is the Hamiltonian functional, and:

$$M_u = D_x^3 + \frac{2}{3}auD_x + \frac{1}{3}au_x, \tag{5.76}$$

is the skew-symmetric, symplectic operator. The Poisson bracket in this case is given by (5.72) except that the symplectic operator L_u is replaced by M_u. Magri [122] showed that the infinite sequence of conservation laws and Lie symmetries of the KdV equation are intimately connected to the above bi-Hamiltonian structure.

Example. One-dimensional (1D) Schrödinger Equation

Consider the generalized Schrödinger equation:

$$i\psi_t + \psi_{xx} - V(x)\psi + 2\nu\psi^2\psi^* = 0. \tag{5.77}$$

If $\nu = 0$, (5.77) reduces to the standard Schrödinger equation used in quantum mechanics. If $V(x) = 0$ and $\nu = 1$, then (5.77) reduces to the nonlinear Schrödinger (NLS) equation, used to describe modulated wave trains in nonlinear optics, fluid mechanics and plasma physics. The NLS equation is an exactly integrable evolution equation, which can be solved by the inverse scattering transform (e.g. see the book by Ablowitz and Segur [3]). Equation (5.77) can be written in the Hamiltonian form:

$$\psi_t = i\frac{\delta H}{\delta\psi^*}, \tag{5.78}$$

where the operator i acts as the symplectic operator, and:

$$H = \int_{-\infty}^{\infty} \left(\nu\psi^2\psi^{*2} - V(x)\psi\psi^* - \psi_x\psi_x^*\right) dx, \tag{5.79}$$

is the Hamiltonian functional. The inner product used in the Hamiltonian system (5.78)–(5.79) is:

$$\langle u, v\rangle = \int_{-\infty}^{\infty} (u^*v + uv^*)dx. \tag{5.80}$$

For this inner product, the operator $L = i$ satisfies $\langle Lu, v\rangle = -\langle u, Lv\rangle = -\langle L^\dagger u, v\rangle$, and hence L is skew-adjoint, i.e. $L^\dagger = -L$. The nonlinear Schrödinger equation obtained with $V(x) = 0$ and $\nu = 1$ is a bi-Hamiltonian system (see [122] for details).

5.3.2 Multi-symplectic Hamiltonian Systems

Multi-symplectic formulations of Hamiltonian systems with two or more independent variables x^α have been developed as a useful extension of Hamiltonian systems with one evolution variable t. This development has connections with dual variational formulations of traveling wave problems (e.g. [24]), and is useful in numerical schemes for Hamiltonian systems. Bridges and co-workers have used the multi-symplectic approach to study linear and nonlinear wave propagation, generalizations of wave action, wave modulation theory and wave stability problems [25,26]. Multi-symplectic Hamiltonian systems have been studied by Marsden and Shkoller [126] and Bridges et al. [28]. Reich [149] and Bridges and Reich [29] developed difference schemes, and Cotter et al. [49] developed a multi-symplectic, Euler-Poincaré formulation of fluid mechanics. The present discussion is based mainly on the work of Hydon [87].

It is first useful to note that Hamiltonian systems, with one evolution variable t, can in general be written in the form:

$$K_{ij}(z)\frac{dz^j}{dt} = \nabla_{z^i} H(z), \tag{5.81}$$

where the fundamental invariant phase space volume element:

$$\kappa = \frac{1}{2} K_{ij}(z) dz^i \wedge dz^j, \tag{5.82}$$

is required to be a closed two-form, i.e. $d\kappa = 0$. Here d denotes the exterior derivative and $\wedge$ denotes the anti-symmetric wedge product used in the algebra of exterior differential forms. The condition that κ be a closed two-form implies, by the Poincaré Lemma, that $\kappa = dg$ where $g = L_j dz^j$ is a one-form (note that $d\kappa = ddg = 0$ by antisymmetry of the wedge product). It turns out, that the condition that κ be a closed two-form implies that $K_{ij} = -K_{ji}$ is a skew symmetric operator (see [87,205] for further discussion of this approach). By taking the exterior derivative of the two-form (5.82) and setting the result equal to zero, we obtain the identity:

$$K_{ij,k} + K_{jk,i} + K_{ki,j} = 0, \tag{5.83}$$

which can be related to the Jacobi identity for the Poisson bracket. If the system (5.81) has an even dimension, and K_{ij} has non-zero determinant, then (5.81) can be written in the form:

$$\frac{dz^i}{dt} = R_{ij} \nabla_{z^j} H(z), \tag{5.84}$$

where R_{ij} is the inverse of the matrix K_{ij}. Here $R_{ij} = -R_{ji}$ is a skew-symmetric matrix. The closure relations (5.83) then are equivalent to the relations:

$$R_{im}\frac{\partial R_{jk}}{\partial z^m} + R_{km}\frac{\partial R_{ij}}{\partial z^m} + R_{jm}\frac{\partial R_{ki}}{\partial z^m} = 0, \tag{5.85}$$

(see e.g. [205,207]). The Poisson bracket for the system in the finite dimensional case is given by

$$\{A, B\} = \sum R_{ij}\frac{\partial A}{\partial z^i}\frac{\partial B}{\partial z^j}. \tag{5.86}$$

Equation (5.85) can then be shown to be equivalent to the Jacobi identity.

Example. Finite Dimensional Hamiltonian System

Consider a finite dimensional Hamiltonian system of dimension $2n$ with canonical variables $z = (q^1, q^2, \ldots, q^n, p_1, p_2, \ldots, p_n)^t$. It can be written in the form (5.81), where

$$\mathbf{K} = \mathbf{J}^t = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}. \tag{5.87}$$

Here the matrix $\mathbf{K}$ is the inverse of the symplectic matrix $\mathbf{J}$ and I_n is the unit $n \times n$ matrix. The invariant phase space element form (5.82) is:

$$\kappa = dp_j \wedge dq^j = d(p_j dq^j). \tag{5.88}$$

A natural generalization of the Hamiltonian system (5.81) to the multi-symplectic case with n independent variables x^α is the system:

$$K^\alpha_{ij} z^j_{,\alpha} = \nabla_{z^i} H(z), \tag{5.89}$$

where $z^j_\alpha = \partial z^j/\partial x^\alpha$. In this case the fundamental invariant two-forms are:

$$\kappa^\alpha = \frac{1}{2}K^\alpha_{ij} dz^i \wedge dz^j, \quad \alpha = 1(1)n. \tag{5.90}$$

The invariance of the phase space element $D_t(dp_j \wedge dq^j) = 0$ for the standard Hamiltonian formulation with evolution variable t, is replaced by the symplectic, or structural conservation law:

$$\kappa^\alpha_{,\alpha} = 0, \tag{5.91}$$

which is sometimes referred to as the symplecticity conservation law.

The closure of the two-forms κ^α means that the exterior derivative of $\kappa^\alpha = 0$. By the Poincaré Lemma, this implies that κ^α is the exterior derivative of a one-form, i.e.

$$\kappa^\alpha = d(L^\alpha_j dz^j) = d\omega^\alpha \quad \text{where } \omega^\alpha = L^\alpha_j dz^j. \tag{5.92}$$

Note that $d\kappa^\alpha = dd\omega^\alpha = 0$. Taking the exterior derivative of ω^α in (5.92) and using the anti-symmetry of the wedge product we obtain:

$$\kappa^\alpha = \frac{1}{2}\left(\frac{\partial L^\alpha_k}{\partial z^j} - \frac{\partial L^\alpha_j}{\partial z^k}\right) dz^j \wedge dz^k. \tag{5.93}$$

From (5.90) and (5.93) it follows that:

$$K^\alpha_{jk} = \frac{\partial L^\alpha_k}{\partial z^j} - \frac{\partial L^\alpha_j}{\partial z^k}, \tag{5.94}$$

which implies that the matrices K^α_{ij} are skew-symmetric, i.e. $K^\alpha_{ij} = -K^\alpha_{ji}$. The analogue of the Legendre transformation for multi-symplectic systems is the identity:

$$\left(L^\alpha_j dz^j\right)_{,\alpha} = d\left\{L^\alpha_j(z) z^j_{,\alpha} - H(z)\right\} \equiv dL, \tag{5.95}$$

where

$$L = L^\alpha_j(z) z^j_{,\alpha} - H(z), \tag{5.96}$$

is the Lagrangian density.

The proof of (5.95) proceeds by noting:

$$\left(L^\alpha_j dz^j\right)_{,\alpha} = \frac{\partial L^\alpha_j}{\partial z^i} z^i_{,\alpha} dz^j + L^\alpha_j(z) D_\alpha dz^j. \tag{5.97}$$

Then noting that the operators d and D_α commute and the fact that K^α_{ij} are skew symmetric in the indices i and j, we obtain:

$$\left(L^\alpha_j dz^j\right)_{,\alpha} = -K^\alpha_{ji} z^i_{,\alpha} dz^j + d\left(L^\alpha_j(z) z^j_{,\alpha}\right). \tag{5.98}$$

The identity (5.95) then follows by using the Hamiltonian evolution equations (5.89). The symplecticity or structural conservation law (5.91) now follows by taking the exterior derivative of (5.95) and using the results $ddL = 0$ and $dD_\alpha = D_\alpha d$. Other conservation laws are obtained by sectioning the forms in (5.95) (i.e. we impose the requirement that $z^j = z^j(\mathbf{x})$, which is also referred to as the pull-back to the base manifold). This gives the conservation law:

$$D_\alpha \left(L_j^\alpha(z) z^j_{,\beta} - L\delta^\alpha_\beta \right) = 0. \tag{5.99}$$

This conservation law is, in fact, the conservation law obtained due to the invariance of the action $A = \int L dx$ under translations in x^β which follows from Noether's first theorem (i.e. $x'^\alpha = x^\alpha + \epsilon\delta^\alpha_\beta$).

A further set of $n(n-1)/2$ conservation laws is obtained from pull-back of the structural conservation law (5.91) to the base manifold, namely:

$$D_\alpha \left(K^\alpha_{ij} z^i_{,\beta} z^j_{,\gamma} \right) = 0, \quad \beta < \gamma. \tag{5.100}$$

The conservation laws (5.100) can be obtained by cross-differentiation of the conservation laws (5.99), i.e. they are a consequence of the equations:

$$D_\gamma \left\{ D_\alpha \left(L_j^\alpha(z) z^j_{,\beta} \right) - D_\beta(L) \right\} - D_\beta \left\{ D_\alpha \left(L_j^\alpha(z) z^j_{,\gamma} \right) - D_\gamma(L) \right\} = 0.$$

Example. Nonlinear Wave Equation

Consider the following nonlinear wave equation:

$$u_{tt} - u_{xx} - V'(u) = 0, \tag{5.101}$$

(e.g. [29,149]). This equation may be obtained by extremizing the action:

$$A = \int dt \int dx \; L(u, u_t, u_x), \tag{5.102}$$

where

$$L = \frac{1}{2}\left(u_t^2 - u_x^2 \right) + V(u). \tag{5.103}$$

is the Lagrangian density. Writing

$$z = (u, v, w)^t, \quad v = u_t, \quad w = -u_x, \tag{5.104}$$

the wave equation (5.101) may be written as the multi-symplectic system:

$$\mathbf{K}^1 \mathbf{z}_t + \mathbf{K}^2 \mathbf{z}_x = \nabla_z S(z), \tag{5.105}$$

where

$$\mathbf{K}^1 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{K}^2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \tag{5.106}$$

and

$$S(z) = \frac{1}{2}\left(v^2 - w^2\right) - V(u) \tag{5.107}$$

is the multi-symplectic Hamiltonian.

Remark 1:

The multi-symplectic scheme (5.105)–(5.107) can be obtained intuitively by setting $p_1 = v = \partial L/\partial u_t$ and $p_2 = w = \partial L/\partial u_x = -u_x$ as canonical momenta and using the standard Legendre transformation to obtain $H = S(z)$. Computing $\nabla_z S(z)$ then gives (5.105). This method is not guaranteed to work in all cases.

It is interesting to use (5.92)–(5.94) to determine the two-forms κ^1 and κ^2 and the corresponding L^j_α matrices from the K^α_{ij} matrices ($\alpha = 1, 2$). We find:

$$\begin{aligned} &\kappa^1 = dz^2 \wedge dz^1 = d(z^2 dz^1), \quad \kappa^2 = dz^3 \wedge dz^1 = d(z^3 dz^1), \\ &L^1_j = \delta^1_j z^2, \quad L^2_j = \delta^1_j z^3. \end{aligned} \tag{5.108}$$

Using the Legendre transformation (5.96) with $H = S(z)$ we obtain the Lagrangian (5.103).

The multi-symplectic formalism also gives conservation laws. Using (5.99) with $\beta = 1$ and $\beta = 2$ gives the time (t) and x-translation symmetry conservation laws:

$$\frac{\partial}{\partial t}\left(\frac{1}{2}\left(u_t^2 + u_x^2\right) - V(u)\right) - \frac{\partial}{\partial x}(u_x u_t) = 0, \tag{5.109}$$

$$\frac{\partial}{\partial t}(u_t u_x) - \frac{\partial}{\partial x}\left(\frac{1}{2}\left(u_t^2 + u_x^2\right) + V(u)\right) = 0, \tag{5.110}$$

respectively.

Remark 2:

Equation (5.105) can also be obtained by extremizing the Cartan-Poincaré form:

$$I = \int \left(dS \wedge dx \wedge dt + \kappa^1 \wedge dx - \kappa^2 \wedge dt\right), \tag{5.111}$$

where $z'^i = z^i + \epsilon Z^i$ gives the variation Z^i in z^i, and in which the $z^i(x,t)$ are forced to be functions of x and t (i.e. one sections the forms: see [27,126]).

Remark 3:
The multi-symplectic formulation (5.105) of (5.101) is not unique. Bridges and Reich [29] show that (5.101) satisfies the system:

$$\mathbf{K}z_t + \mathbf{L}z_x = \nabla_z S, \tag{5.112}$$

where $z = (u, v, w, \phi)^t$ and

$$\mathbf{K} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{5.113}$$

The Hamiltonian $S(z)$ is given by (5.107) and $\phi(x,t)$ satisfies the wave equation $\phi_{tt} - \phi_{xx} = 0$.

5.3.3 Numerical Methods

Since we are interested in numerical methods, we will only consider finite dimensional Hamiltonian systems that result from finite dimensional approximations of the partial derivative operators in terms of the differentiation matrices. The origin of the differentiation matrix could be due to any spatial discretization, e.g. finite differences, finite elements, spectral methods, etc.

Example. One-dimensional Schrödinger Equation
Consider the solution of the 1D Schrödinger equation:

$$i\psi_t = -\psi_{xx} + V(x)\psi, \tag{5.114}$$

by the method of lines. Setting $\psi = u+iv$, the Schrödinger equation (5.114) splits into the two equations:

$$\begin{aligned} u_t &= -\left(v_{xx} - V(x)v\right), \\ v_t &= u_{xx} - V(x)u. \end{aligned} \tag{5.115}$$

Below, we use the standard method of lines to discretize the second order spatial derivatives in (5.115) to obtain a set of $N-1$ ODE for the vectors:

$$U = (u_1, u_2, \ldots, u_{N-1})^t \quad \text{and} \quad V = (v_1, v_2, \ldots, v_{N-1})^t \,, \tag{5.116}$$

where $u_j = u(x_j)$ and $v_j = v(x_j)$ are the discretized variables. The set of ODEs for U and V turns out to be a Hamiltonian system. This is expected, since the Schrödinger equation is a Hamiltonian system. The same set of differential equations for U and V are then obtained by discretizing the continuous, exact Hamiltonian for the system. This example shows that the difference equation system obtained from discretizing (5.115) is essentially the same as that obtained by discretizing the Hamiltonian functional of the continuous system. However, it is in general necessary to carefully consider the boundary values for the system at $j = 0$ and $j = N$ in the numerical implementation.

For the sake of concreteness, consider the discretized system, in which central differences are used for the second-order spatial derivatives in (5.115), namely:

$$\begin{aligned} \frac{dU}{dt} &= DV - a(v_0, 0, \ldots, 0, v_N)^t, \\ \frac{dV}{dt} &= -DU + a(u_0, 0, \ldots, 0, u_N)^t, \end{aligned} \tag{5.117}$$

where Δx is the discretization step in x and $a = 1/(\Delta x)^2$. The matrix D in (5.117), written in index notation is:

$$D_{jk} = (V_j + 2a)\, \delta_{jk} - a\, (\delta_{j,k-1} + \delta_{j,k+1})\,, \tag{5.118}$$

where $1 \leq j, k \leq (N - 1)$, $\delta_{j,k}$ is the Kronecker delta symbol and $V_j = V(x_j)$. Note that the matrix D_{jk} is a symmetric, tridiagonal matrix (D is the discrete analogue of the operator $V(x) - d^2/dx^2$).

The ODE system (5.117), with zero boundary data: $u_0 = u_N = 0$ and $v_0 = v_N = 0$ can be written in the Hamiltonian form:

$$\frac{dU}{dt} = \frac{\delta H_d}{\delta V}, \quad \frac{dV}{dt} = -\frac{\delta H_d}{\delta U}, \tag{5.119}$$

where

$$H_d = \frac{1}{2} \left(U^t D U + V^t D V \right), \tag{5.120}$$

is the Hamiltonian for the discretized system (5.117). In (5.119), $\delta H_d / \delta U = 1/2(D^t + D)U = DU$ because D is a symmetric matrix. Similarly,

$\delta H_d/\delta V = DV$. This establishes the basic Hamiltonian structure of the ODEs (5.117) used in the method of lines.

We now demonstrate how the discretized Hamiltonian H_d is related to the Hamiltonian for the exact equations (5.115) for u and v. These equations can be written in the Hamiltonian form:

$$u_t = \frac{\delta H}{\delta v}, \quad v_t = -\frac{\delta H}{\delta u}, \tag{5.121}$$

where

$$H = \int \frac{1}{2}\left[V(x)(u^2+v^2) + u_x^2 + v_x^2\right] dx, \tag{5.122}$$

is the Hamiltonian. Note that the Hamiltonian:

$$\tilde{H} = \frac{1}{2}\left(\sum_{j=1}^{N-1} V_j(u_j^2+v_j^2) + a\left(u_{j+1}-u_j\right)^2 + a\left(v_{j+1}-v_j\right)^2\right),$$

where $a = 1/(\Delta x)^2$ obtained by discretizing the Hamiltonian functional (5.122) is equivalent to the discretized Hamiltonian (5.120) where it is assumed that $u_j = v_j = 0$ for $j = 0, N$ for the boundary data.

Thus the discretization scheme used to approximate the Hamiltonian functional plays a central role in the solution of the discretized Schrödinger equation by the method of lines. Different discretizations of the Hamiltonian will clearly lead to different matrices D. In particular, one may wish to use forward differencing at the lower boundary $j = 0$, and backward differencing at the upper boundary at $j = N$, or some other form of differencing such as compact differences.

Example

In mechanics, [4], one often encounters a Hamiltonian of the form:

$$H = \frac{1}{2}p^t M(q)p + U(q),$$

representing for example the Hamiltonian for a collection of particles moving in a potential $U(q)$. This Hamiltonian is similar to the discretized Hamiltonian obtained in the previous example.

Example

Consider the normalized, linear and isotropic 3D Maxwell equations:

$$\mathbf{E}_t = \nabla \times \mathbf{H}, \quad \mathbf{H}_t = -\nabla \times \mathbf{E},$$

converted by the method of lines into an ODE system by discretizing *curl* operators on a staggered grid, [162]. The discretized system has the form:

$$U_t = CV, \quad V_t = -CU.$$

A continuous *curl* operator is skew-symmetric and hence the matrix C is also skew-symmetric. In particular, it is a skew-symmetric block matrix with skew-symmetric blocks,

$$C = \begin{pmatrix} 0 & -D_z & D_y \\ D_z & 0 & -D_x \\ -D_y & D_x & 0 \end{pmatrix},$$

where matrices D_x, D_y and D_z represent anti-symmetric differentiation matrices. The Hamiltonian for this system, written in terms of the variables U and V is the same as for the Schrödinger equation example above, namely $H = \frac{1}{2}U^tDU + \frac{1}{2}V^tDV$.

It is known that Hamiltonian systems admit at least n time invariants called Poincare invariants that can be written in terms of differential forms, [125]. Some Hamiltonian systems are integrable and have an infinite number of invariants, [206]. Here we only mention two time invariants. The first one is $H(q,p)$ and is often related to the physical energy of the system. Its invariance follows directly from the Hamiltonian form of equations:

$$H_t = H_q q_t + H_p p_t = H_q H_p - H_p H_q = 0.$$

The second invariant is called the symplecticity of the solution operator and is defined in terms of symplectic transformations, called canonical transformations in Hamiltonian dynamics. A smooth map $T : z \to \tilde{z}$ is called symplectic if and only if it preserves the Hamiltonian structure of the system, e.g. in terms of new variables:

$$\frac{d\tilde{z}}{dt} = J\nabla_{\tilde{z}}\tilde{H}(\tilde{z}).$$

Applying the chain rule it follows that a symplectic map satisfies the following identity:

$$(T')^t\, J\, T' = J,$$

where T' is the Jacobian matrix of this transformation, $T' = \frac{\partial \tilde{z}}{\partial z}$. It turns out that the solution operator, $S_t : z(0) \to z(t)$, that maps the initial conditions to a solution at arbitrary time t (see Chapter One for the definition of the solution operator) is symplectic. In fact, an ODE $\dot{z} = f(z)$ can be written in Hamiltonian form if and only if its solution operator is symplectic in some neighborhood of (t, z), [185]. Mathematically, symplecticity is related to volume-preserving property in the phase space, [125,185], while in physical terms it is related in some applications to vorticity, helicity and similar curl-like quantities due to Stokes (Kelvin, Afvén) circulation theorems, [28,159]. Discrete symplectic maps are defined identically to the continuous case by replacing continuous time with a discrete time, discrete map $T_n : z^n \to z^{n+1}$ is symplectic if and only if $(T'_n)^t\ J\ T'_n = J$, where T'_n is defined as before:

$$T'_n = \frac{\partial z^{n+1}}{\partial z^n}.$$

A very useful property of symplectic maps is that composition of symplectic maps is symplectic, [125]. This implies the simplecticity is preserved for each step or sub-step, like in multistep or time-split algorithms. We proceed with the illustrations of symplectic algorithms applied to several continuous Hamiltonian systems discussed before.

Consider an ODE system:

$$u_t = v, \quad v_t = -u,$$

discretized by an explicit Euler method:

$$u^{n+1} = u^n + \Delta t\ v^n, \quad v^{n+1} = v^n - \Delta t\ u^n.$$

Then the discrete map T'_n is:

$$T'_n = \begin{pmatrix} 1 & \Delta t \\ -\Delta t & 1 \end{pmatrix}.$$

It does not satisfy the symplectic property, $(T'_n)^t\ J\ T'_n = J$, while the following modification:

$$\begin{aligned} u^{n+1} &= u^n + \Delta t\ v^n, \\ v^{n+1} &= v^n - \Delta t\ u^{n+1} = v^n - \Delta t\ (u^n + \Delta t\ v^n), \end{aligned}$$

with:

$$T'_n = \begin{pmatrix} 1 & \Delta t \\ -\Delta t & 1 - (\Delta t)^2 \end{pmatrix},$$

renders a symplectic integrator. This method is equivalent to a leap-frog scheme if u^{n+1} is understood as approximating $u(t^{n+1/2})$. It can be rewritten in the familiar form:

$$u^{n+1/2} = u^{n-1/2} + \Delta t\, v^n,$$
$$v^{n+1} = v^n - \Delta t\, u^{n+1/2}.$$

For nonlinear Newton's law with potential force, the implicit midpoint rule:

$$x^{n+1} = x^n + \frac{\Delta t}{2}\,(v^n + v^{n+1}),$$
$$v^{n+1} = v^n + \Delta t\, F\left(\frac{x^n + x^{n+1}}{2}\right),$$

is symplectic, while the trapezoidal rule that treats forcing as $\frac{F(x^n)+F(x^{n+1})}{2}$ is not, [168]. Another symplectic algorithm is a leap-frog type Verlet algorithm [168]:

$$x^{n+1/2} = x^n + \frac{\Delta t}{2}\, v^n,$$
$$v^{n+1} = v^n + \Delta t\, F(x^{n+1/2}),$$
$$x^{n+1} = x^{n+1/2} + \frac{\Delta t}{2}\, v^{n+1}.$$

For the linear Schröedinger equation the trapezoidal rule results in the unitary update matrix and is called Cayley's method, [147]. For the NLS equation with cubic nonlinearity, $|\psi|^2\psi$, for a comparison of the variety of fully implicit, semi-implicit discretization of the nonlinear term, $|\psi^n|^2(\psi^n + \psi^{n+1})/2$, together with pseudo-spectral methods, the reader is referred to [43].

Physical systems, in addition to symplectic structure and conservation of energy, may have other invariants that are important to preserve for accurate computation. For example, for Maxwell equations application of the leap-frog scheme (Yee method) that is second order symplectic integrator preserves $\operatorname{div} B = 0$ constraint, while higher order symplectic integrators may not. For variational conservative mechanical systems one can achieve preservation of energy, momentum and symplectic structure using adaptive time stepping, [93], while with fixed time stepping one has to choose between symplectic-momentum or energy-momentum preserving methods, [125]. In some applications, like molecular dynamics, it is advantageous to preserve energy, while in others, like astrophysical computations, preservation of the angular momentum leads to better accuracy and efficiency of the computation.

Multi-symplectic Schemes

Bridges and Reich [29] give an extensive overview of discretization schemes for multi-symplectic Hamiltonian PDEs. A widely used strategy for developing discretizations of (5.105) or (5.112) representing the nonlinear wave equation (5.101) is to concatenate one-dimensional schemes. Thus applying the implicit midpoint rule to the system (5.112) in both space and time gives the scheme:

$$\mathbf{K}\delta_t^+ z_{i+1/2}^n + \mathbf{L}\delta_x^+ z_i^{n+1} = \nabla_z S\left(z_{i+1/2}^{n+1/2}\right). \tag{5.123}$$

Here the edge midpoint approximations are given by:

$$z_i^{n+1} = \frac{1}{2}\left(z_i^{n+1} + z_i^n\right), \quad z_{i+1/2}^n = \frac{1}{2}\left(z_{i+1}^n + z_i^n\right), \tag{5.124}$$

the cell center approximation is:

$$z_{i+1/2}^{n+1/2} = \frac{1}{4}\left(z_{i+1}^{n+1} + z_i^{n+1} + z_{i+1}^n + z_i^n\right), \tag{5.125}$$

and the forward finite difference operators δ_t^+ and δ_x^+ are defined by:

$$\delta_t^+ z_{i+1/2}^n = \frac{1}{\Delta t}\left(z_{i+1/2}^{n+1} - z_{i+1/2}^n\right), \tag{5.126}$$

$$\delta_x^+ z_i^{n+1/2} = \frac{1}{\Delta x}\left(z_{i+1}^{n+1/2} - z_i^{n+1/2}\right). \tag{5.127}$$

This scheme is an example of a “box scheme”, and is called the Preissman box scheme. Preissman introduced the scheme for shallow water equations. The Preissman scheme (5.123) exactly preserves a discrete multi-symplectic conservation law [27].

5.4 Hyperbolic Conservation Laws

In this section we discuss singular solutions for PDEs. We choose examples of hyperbolic conservation laws as the theory and understanding of discontinuous solutions for such systems reached its maturity over the last few decades. The singularities of PDEs are defined as “solutions” or geometrical features of the problem where the PDE model at hand breaks down. For example, in the formation of discontinuities, cusps or infinite quantities near discontinuities or corners in the geometry, the knowledge exists that beyond some time the model describes non-physical behavior even though nothing "blows-up". Often singularities arise from the desire to stay with a model that has only one or two scales that need to be resolved, the fine scales are not resolved numerically. To obtain accurate solutions that do not resolve the fine scale structure requires additional constraints to deal with the smaller scale effects, for example implementing them as singularities. In some cases, the implicit or explicit implementation of such extra conditions in numerical methods are known, while in others the additional constraints that will render a physically desirable solution are not known or are very complicated and one has to go through the expense of resolving extra physical scales. Often this approach may not be feasible due to limited computational resources, but when it is possible, various grid generation strategies discussed in the next chapter are an indispensable tool for this task.

The additional information required to define the singularity is called a regularization procedure, a terminology used in mechanics to describe the solution past the break time of the system. The validity of a particular regularization procedure needs to be established by comparison with the physical solution obtained by resolving necessary additional scales. That process might be quite costly or impossible due to limitation of the computational resources. For example, in engineering application of turbulence computation, resolution of the correct physical Kolmogorov scale by direct numerical simulation (DNS) can be prohibitive or impossible. The Kolmogorov scale is given by $\eta = (\nu^3/\varepsilon)^{1/4}$ where ν is the kinematic viscosity and ε is the rate of kinetic energy dissipation. The scales that need to be resolved in DNSs stretch from the integral scale L of the turbulence associated with large scale motions of the turbulent eddies down to the dissipation scale, or Kolmogorov scale η. N is the number of spatial mesh points needed to cover the entire range of scales, $\eta < x < L$, using a uniform grid with scale size η, in 3D computations; $N \approx \text{Re}^{9/4}$ where $\text{Re} = u'L/\nu$ is the Reynolds number of the flow and u' is the turbulent velocity fluctuation at the integral scale. The largest DNS of Navier Stokes turbulence, as of October 2007, by the Japanese Earth simulator supercomputer, used 4096^3 mesh points, which corresponds to a modest Reynolds number of $\text{Re} = (4096)^{4/3}$ (see Wikpedia, under Direct numerical simulation). Thus various turbulence closure models are compared to the experimental results to determine their region of applicability, [198].

Example. Advection Equation

Consider the advection equation:

$$u_t + u_x = 0,$$

with singular initial data having discontinuities in u or its derivatives. There are many mathematical ways to regularize the system, for example by smoothing the initial data, smoothing the solution by introducing numerical diffusion, dispersion, etc. and viewing the solution as a limiting case of the vanishing smoothing effects. Another approach would be to switch to an integral formulation obtained by integration with some weight function, $w(x)$. For example, taking $w(x) = 1$ and integrating over x from $x = a$ to $x = b$ in the above advection equation gives the integral formulation:

$$\frac{d}{dt}\int_b^a u(x,t)dx + u(b,t) - u(a,t) = 0.$$

Various mathematical regularization procedures may provide different singular solutions and correspond to different physical regularizations. For example, dispersive regularization would create a Gibbs-type spike near the discontinuity together with oscillations, while diffusive regularization will smear the solution with the transition region spreading according to solution of the heat equation with initial step function, $\mathrm{erf}(\sqrt{(t)})$, [198]. If the advection equation came from a physical problem modeling pure advection, conservation of some physical quantity might be the choice for the regularization procedure, and in some cases happen to be equivalent to a simple algebraic relation.

Example. Numerical Approximations of Singular Solutions of the Advection Equation

Consider the following numerical methods to approximate the above advection equation:

$$\begin{aligned}
u_j^{n+1} &= u_j^n - \nu(u_j^n - u_{j-1}^n), \quad \text{(upwind scheme)};\\
u_j^{n+1} &= u_j^n - \frac{\nu}{2}(u_{j+1}^n - u_{j-1}^n)\\
&\qquad + \frac{\nu^2}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \quad \text{(Lax-Wendroff)};\\
u_j^{n+1} &= u_j^{n-1} - \frac{\nu}{2}(u_{j+1}^n - u_{j-1}^n), \quad \text{(leap-frog)};\\
u_j^{n+1} &= \tilde{u}_j^n, \quad \text{(method of characteristics)};
\end{aligned}$$

where $\nu = \Delta t/\Delta x$ denotes the Courant number and the value of $\tilde{u}_j^n$ is defined using an interpolation at position $x_j - a\Delta t$, with advection speed $a = 1$ in our example. One may use local, piecewise linear interpolation or a non-oscillatory form of cubic spline (described later in this section) to determine the unknown value $\tilde{u}$ from nearby grid values.

The non-oscillatory interpolation would reduce accuracy to first-order near extrema. The first three methods described above require the Courant number $|\nu| < 1$ for stable solutions. The discretization (truncation) errors are diffusive for the upwind scheme and dispersive with no numerical damping for the leap-frog scheme. The Lax-Wendroff scheme has a second-order dispersive truncation error and a fourth-order diffusion as the leading terms of the modified (or truncation error) equation. The correct choice of the numerical method depends on the regularization appropriate for the physical problem at hand. For example the leap-frog type scheme, Yee algorithm, is the most popular numerical method in electrodynamic applications where, in addition to correct treatment of the material discontinuities, it preserves the additional constraint on the magnetic field to be void of magnetic charges, i.e. $\operatorname{div} B = 0$. For electromagnetic problems, higher order upwind methods perform well for short runs of a few thousand iterations where effects of numerical diffusion do not deteriorate the accuracy of the computation due to artificial damping or artificial magnetic charges.

In the following example, we will overview the material that is discussed in more detail in the subsequent subsections.

Example. Definition of Discontinuous Solutions for Nonlinear Hyperbolic Conservation Laws

Consider a 1D system:

$$U_t + F(U)_x = 0,$$

where a Jacobian matrix of F, $A(U) = \frac{\partial F}{\partial U}$ has real eigenvalues $\{\lambda_j\}$ and a complete set of right and left eigenvectors $\{R_j\}$ and $\{L_j\}$ for all values of U. For example, the inviscid Burgers equation, $u_t + uu_x = 0$, Euler's equations of gas dynamics, the ideal MHD equations, the equations of elasticity and Maxwell's equations of electrodynamics in nonlinear Kerr media, can all be cast in the above conservation form. The inviscid Burgers equation:

$$u_t + uu_x = 0,$$

with smooth initial data, say, $u(x, 0) = f(x) = \sin(x)$ on the interval $[0, 2\pi]$, will have a solution that eventually steepens up into a solution with an infinite slope due to the fact that the advection speed is larger at spatial positions where the initial function values are larger. The implicit solution of the inviscid Burgers equation before the break time is given by the formulae: $u(x, t) = f(\xi)$, where $x - ut = \xi$. The implicit formula for the solution up to break time, $u(x, t) = f(x - u(x, t)t)$, shows that both derivatives $u_x = f'/(1 + tf')$, $u_t = -uf'/(1 + tf')$ become infinite the first time the determinant becomes zero, $t_b = \min(-1/f')$, assuming there is at least one point ξ that $f'(\xi)$ is negative, [198]. The exact solution for $\sin(x)$ initial data may be written in terms of Fubini series, [34,156]:

$$u(x,t) = -2\sum_{k=1}^{\infty} \frac{J_k(kt)}{kt}\sin(kx),$$

and it can be used to illustrate performance of various numerical methods near the break up time. In the Lagrangian particle dynamics interpretation of the inviscid Burgers equation the Cauchy characteristics, $dx/dt = u(x,t)$ and $du/dt = 0$ may be integrated with initial data $u(x,0) = f(\xi)$ and $x = \xi$ at time $t = 0$. In this formulation, ξ is regarded as the Lagrangian fluid label, and the solution for x has the form $x = X(\xi,t) = u(\xi)t + \xi$. The inviscid Burgers equation then states that the particle acceleration is zero, i.e.:

$$\frac{d}{dt}u(x(t),t) = u_t + u_x\frac{d}{dt}x(t) = u_t + uu_x = 0,$$

where u represents the velocity particle. In other words, each particle moves with its own constant velocity given by the initial condition. Therefore if there are particles ahead that are moving more slowly than any particle behind it, $f' < 0$, the collision is inevitable and the model stating that particles are free streaming ceases to be valid. A regularization procedure is required to prescribe the collision and post-collision scenarios. It is quite clear that imposing mere conservation of particle number still leaves a variety of particle interactions possible, for example, reflection, pile-up, passing through, etc. The Gauss divergence theorem implies that across the discontinuous solution the conservation principle is equivalent to fluxes being constant on both sides of a moving discontinuity. Introducing the moving frame $\xi = x - st$, and changing variables, $\tilde{U}(\xi,t) = U(x,t)$, where s is the discontinuity speed, transforms the conservation law to the form:

$$\tilde{U}_t + (\tilde{F} - s\tilde{U})_\xi = 0.$$

The jump conditions across the discontinuity are $F_l - sU_l = F_r - sU_r$. For the inviscid Burgers equation it reduces to the relation $s = (u_l + u_r)/2$. In a multi-dimensional case the jump conditions are the same if the fluxes and the discontinuity speed are understood as being taken in the direction normal to the discontinuity surface.

For the inviscid Burgers equation the non-uniqueness of the solution satisfying the conservation principle can be illustrated for initial data $u(x,0) = sign(x)$. This function is a solution to the Burgers equation if we consider the integral conservative form of the equation. It satisfies the differential equation in smooth regions and satisfies the jump conditions across the discontinuity. But we could preserve the conservation principle if we introduce any number of intermediate discontinuities, e.g between states 1.0 and 0.9, moving with speed 0.95, between states 0.9 and 0.8 moving with speed 0.85, etc. In fact, taking the number of intermediate states to infinity we end up with a continuous, but discontinuous in the first derivative, weak solution that consists of a linear transition between the end states that separate with constant characteristic speeds ± away from the initial discontinuity. A physical regularization may consist of considering vanishing viscosity or dispersive solutions resulting in different limiting solutions, [105]. The former case can be shown to be the simple condition that discontinuity should be compressive, e.g. $u_l > u_r$, [111]. In other fields of application the situation is not that simple. For example, for polytropic gas dynamics, vanishing viscosity limit is equivalent to compressibility or increase of entropy across the discontinuity that follows from the second law of thermodynamics [118]. For a two-dimensional (2D) ideal MHD, the gas dynamic constraints are still not sufficient as the limiting solution depends discontinuously on magnetic viscosities, with various ratios producing different limiting results. An additional requirement on the solution to be the limiting case of a 3D problem, that allows for additional rotation of Alfven waves to be part of the solution, does lead to uniqueness, [15]. For Maxwell's equations with Lorentz dispersion and Kerr nonlinearity, the dispersion regularizes the solution if the nonlinearity is sufficiently weak, which happens to be the case in current applications, but a correct physical mechanism for stronger nonlinearities is not known [195]. In combustion, numerical methods that rely on Godunov-type methods together with time-splitting may add local diffusion and produce incorrect speeds of the discontinuities [111]. Therefore one has to be cautious of Godunov-type methods for instability-type, turbulence, MHD, electrodynamics, combustion and other problems where the solution depends on the details of the interactions on the smaller scales, and validation studies have to be performed before any claim of physical correctness of the obtained numerical solutions.

5.4.1 Overview of Godunov Methods

Below we give a broad overview of Godunov methods for conservation laws. This is followed by a more detailed exposition, including: a discussion of shocks and weak solutions, Riemann problems, Lax geometrical entropy conditions, linear Riemann problems, upwind difference schemes, Godunov, finite volume Riemann solvers and approximate Riemann solvers, including Roe solvers.

Example. Introduction to Godunov-type Methods for Nonlinear Conservation Laws

Historically, the Godunov method came about quite serendipitously. Godunov was considering the numerical approximation of the system of conservation laws in gas dynamics (Euler's equations) in the form of:

$$U_j^{n+1} = U_j^n - \nu(F_{j+1/2}^n - F_{j-1/2}^n).$$

After realizing that taking numerical flux as a simple arithmetic average of nearby fluxes, $F_{j+1/2}^n = (F_j + F_{j+1})/2$ renders an unstable method, Godunov worked out numerical flux values based on averaged values of the Riemann invariants, [74]. The resulting formulae turned out to be equivalent to solving the Riemann problem for the exact system of conservation laws with initial data consisting of two constant states U_j and U_{j+1}, and substituting the resulting solution $U_{j+1/2}$ into the exact physical flux formula, $F_{j+1/2} = F(U_{j+1/2})$. The solution of the exact nonlinear Riemann problem requires Newton's iterations to solve the resulting nonlinear algebraic equations [111], and increases the cost of the overall computation. Later, simplified versions of the Godunov method were introduced, such as the Roe scheme described in (5.203)–(5.216). The Roe-type Riemann solver has explicit analytic solutions in terms of the eigenvalues and the eigenvectors of the Jacobian matrix $\frac{\partial \hat{F}}{\partial U}$ of the associated simpler system (5.211). The numerical flux is computed as in the linear case, but the matrix A varies with spatial position, $A_{j+1/2}$. There are numerous ways to compute an averaged matrix, for example by the trapezoidal or midpoint rule, $A_{j+1/2} = (A_j + A_{j+1})/2$, $A_{j+1/2} = A(U_{j+1/2})$, respectively. Another approach would be to average other values in order to satisfy additional constants.

For example, Roe proposed an averaging that preserves the jump conditional across a stationary discontinuity. The averaging should also preserve the positivity of physical quantities such as density and pressure, [33,119]. In the linear case, solution to the Riemann problem can be written as:

$$U_{j+1/2}^n = \frac{U_l + U_r}{2} - \frac{1}{2}\sum_k sign(\lambda_k) c_{k,j+1/2}^n R_k,$$

or multiplying by constant matrix A as,

$$F_{j+1/2}^n = \frac{F_l + F_r}{2} - \frac{1}{2}\sum_k |\lambda_k| c_{k,j+1/2}^n R_k = \frac{F_l + F_r}{2} - \frac{1}{2}|A|(U_r - U_l).$$

For the nonlinear case, the $F(U_{j+1/2})$ and the above $F_{j+1/2}$ are not equivalent, but either can be postulated as numerical flux for the Roe-type Riemann solver. The $c^n_{k,j+1/2}$ determine the jump between the left and right states U_l and U_r in the Riemann problem (see equation (5.196) and Section 5.4.5). The λ_k are the eigenvalues of the matrix $\hat{A}$ used in the Roe solver in (5.211).

Example. Nonlinear Filtering via Slope and Flux Limiters

In this example we discuss solution-dependent nonlinear spatial filtering procedures that allow one to extend Godunov-type methods to higher spatial order of accuracy while preserving the non-oscillatory property of the first-order methods. In a series of papers, [186], van Leer showed that the standard Lax-Wendroff method, when applied to a scalar advection equation, may be viewed as a Riemann solver with interface values $u_{j+1/2}$ determined by linear interpolation between u_j and u_{j+1}. The modified equation for the Lax-Wendroff method has a leading truncation error term as a dispersive term proportional to $(\Delta x)^2 u_{xxx}$, and thus is oscillatory. To eliminate oscillations, van Leer introduced slope limiters $s_{j+1/2}$ into piecewise linear interpolation of the interface values:

$$u_{j+1/2} = u_j + \frac{1}{2}(1-\nu)(u_{j+1} - u_j)s_{j+1/2},$$

where the slope is set to zero in order to flatten the oscillations if nearby slopes are of opposite sign, and the slope is taken as a harmonic mean between the slopes in the upwind direction in cases where the slopes have the same sign, i.e.:

$$(u_{j+1} - u_j)s_{j+1/2} = \phi(u_{j+1} - u_j, u_j - u_{j-1}),$$

where $\nu = \Delta t/\Delta x$ for advection equation $u_t + u_x = 0$ and the harmonic mean function is defined as:

$$\phi(a,b) = \frac{1}{2(\frac{1}{a} + \frac{1}{b})} = \frac{2ab}{a+b}, \quad \text{for } ab > 0,$$

and zero otherwise. Sweeby showed, [180], that van Leer's limiter is a total variation diminishing (TVD) limiter with an amount of numerical diffusion between the most diffusive TVD limiter, the minmod limiter, that can be written in terms of the slopes or slope ratios as follows:

$$\phi(a,b) = \max(0, \min(a,b)) = a \max(0, \min(1,r)),$$

where $r = b/a$, and the least diffusive TVD limiter, the superbee limiter defined as:

$$\begin{aligned}\phi(a,b) &= \max(0, \min(a,2b), \min(b,2a)) \\ &= a\max(0, \min(1,2r), \min(r,2)).\end{aligned}$$

Similarly, van Leer's limiter can be cast into a single formula:

$$\phi(a,b) = a\phi(r) = a\frac{r+|r|}{1+|r|}.$$

It is known that the accuracy of the TVD approximations drops to the first-order near local extrema and that numerical diffusion decreases as one moves from lower to upper boundary of the TVD region, [111]. The van Leer limiter provides a limiter that passes in the middle of the TVD region. This non-oscillatory TVD property of the harmonic mean accounts for its popularity in other applications like Hermite splines, where the unknown slopes at data points are determined as the harmonic mean of nearby slopes or set to zero as described above [135]. This also reduces the original fourth-order approximation of the Hermite splines (when derivatives are provided) down to the first-order approximation. For applications to spectral Fourier and finite difference methods in optics to reduce oscillations near discontinuities in material boundaries see [94,112].

Second-order accuracy in time can be achieved simply by applying higher-order time integrators like second-order Runge-Kutta methods, [111]. This semi-discretization approach requires the Riemann solver to be applied at each sub-step. On the other hand, incorporation of the limiters as wave amplitude limiters into the modified Lax-Wendroff method provides second-order accuracy in time and only a single application of the Riemann solver, [111,119]. The numerical flux is a modification of the Roe flux, [119],

$$\begin{aligned}F^n_{j+1/2} &= \frac{F_l+F_r}{2} - \frac{1}{2}\sum_k |\lambda_k| c^n_{k,j+1/2} R_k \\ &= \frac{F_l+F_r}{2} - \frac{1}{2}|A|(U_r - U_l),\end{aligned}$$

by replacing $|\lambda_k|$ with an expression:

$$\alpha|\lambda_k|(1-\phi^0_k) + \beta\lambda_{max}(1-\phi^1_k) + \frac{\Delta t}{\Delta x}\lambda^2_k.$$

The last term in the new expression is a Lax-Wendroff-type term that makes the method second-order accurate in time. The middle term, with

λ_{max} denoting the spectral radius of the matrix $A_{j+1/2}$, is an entropy fix introduced independently by Rusanov and Einfeldt et al. [61,173], where λ_{max} is the spectral radius of the matrix $A_{j+1/2}$. The introduction of an entropy fix does not allow numerical diffusion of Roe schemes to drop to zero when the averaged eigenspeed $\lambda_{j+1/2} = 0$. Both the first term and the middle term are multiplied by flux limiters, that are chosen in the paper to be superbee and minmod, respectively. The weights α and β allow one to manipulate the amount of numerical diffusion and have to satisfy the constraints, $0 \leq \alpha \leq 1$, and $\alpha + \beta = 1$. In the paper $\alpha = 0.9$ and $\beta = 0.1$.

This formula gives an example of applying the flux limiters to numerical fluxes versus the application of the slope limiters to the initial conditions for the Riemann solver as discussed before. Also notice the possibility of application of different limiters to different waves, labeled by index k. More numerical diffusion is usually added to compressive fields, like shocks in gas dynamics, and less to linear or material discontinuities like contact and shear discontinuities. The reader may consult [45,111,119] for examples of code implementation as well as multi-dimensional extensions of the Lax-Wendroff method that incorporate cross derivative terms and improve on the usual 1D treatment of numerical fluxes in each spatial direction.

The mathematical ideas behind nonlinear data filtering used in TVD limiters methods were incorporated into the design of level set methods, that in turn found numerous and diverse applications including image processing and computer vision, [143].

5.4.2 Shocks and Discontinuous Solutions

Consider the scalar conservation law:

$$u_t + f(u)_x = 0, \tag{5.128}$$

in one Cartesian space dimension x. This equation was discussed in detail in Section 1.4.3, but with u replaced by ρ and $f(u)$ by $Q(\rho)$ (see also [44, 198]). For smooth solutions, (5.128) is equivalent to the nonlinear wave equation:

$$u_t + \lambda(u)u_x = 0, \tag{5.129}$$

where $\lambda(u) = f'(u)$ is the characteristic wave speed of the system. The general solution of (5.129) by the method of characteristics, satisfying the initial data $u(x, 0) = g(x)$ is given by the implicit formulae:

$$x - G(\xi)t = \xi, \quad u(x,t) = g(\xi), \quad \lambda(u) = G(\xi). \tag{5.130}$$

The analysis of Section 1.4.3 showed that the solution (5.130) has a gradient catastrophe (i.e. $|u_t| \to \infty$ and $|u_x| \to \infty$ at time $t = -1/G'(\xi)$ where the characteristics cross). The minimum break time, t_B is given by $t_B = -1/G'(\xi_B)$ where $|G'(\xi)|$ is maximal at $\xi = \xi_B$. Note that the Jacobian $\partial(\xi, t)/\partial(x, t) = \xi_x = 1/[1 + tG'(\xi)]$ blows up at time $t = -1/G'(\xi)$.

Beyond the break time t_B, the solution (5.130) becomes multi-valued, and it is necessary to regularize the flow in order to obtain a well-defined weak solution of the equation. A weak solution of (5.128) is defined by the integral condition:

$$I = -\iint_{\mathcal{R}} (\nabla \cdot \mathbf{F})\phi dx\, dt = 0, \tag{5.131}$$

where $\phi(x, t)$ is a smooth function with compact support over the region $\mathcal{R}$ of the (x, t) plane of interest. Here:

$$\nabla \cdot \mathbf{F} = \nabla \cdot [f(u), u] = u_t + f(u)_x = 0, \tag{5.132}$$

is an alternative way of writing the conservation law (5.128) in which $\mathbf{F} = [f(u), u]$ is the flux and $\nabla = (\partial_x, \partial_t)$ is the gradient in (x, t) space. The test function $\phi(x, t)$ is assumed to be continuous and first-order differentiable (i.e. a C^1 function) and to vanish outside the region $\mathcal{R}$ in the (x, t) plane. By applying Green's theorem in the plane, (5.131) can be written in the form:

$$\begin{aligned} I &= -\iint_{\mathcal{R}} [\nabla \cdot (\mathbf{F}\phi) - \nabla\phi \cdot \mathbf{F}]\, dx\, dt \\ &= \iint_{\mathcal{R}} \nabla\phi \cdot \mathbf{F}\, dx\, dt - \oint_{\partial\mathcal{R}} (\mathbf{F} \cdot \mathbf{n})\phi\, d\sigma = 0, \end{aligned} \tag{5.133}$$

where

$$\mathbf{n} = (y_\sigma, -x_\sigma) \quad \text{and} \quad \mathbf{t} = (x_\sigma, y_\sigma), \tag{5.134}$$

define the unit outward normal $\mathbf{n}$ and and tangent vector $\mathbf{t}$ to the boundary curve $\mathcal{C} = \partial\mathcal{R}$ of the domain $\mathcal{R}$ and σ denotes the arclength along $\mathcal{C}$. Assuming that ϕ vanishes s on $\mathcal{C}$, the weak solution condition (5.133) reduces to the equation:

$$I = \iint_{\mathcal{R}} \nabla\phi \cdot \mathbf{F}\, dx\, dt = 0. \tag{5.135}$$

Both formulations (5.133) and (5.135) are useful in deriving shock jump conditions for discontinuous weak solutions of (5.128).

Assuming that the solution was regularized by inserting a shock into the weak solution after the breaking time, we consider the application of the above formulas in the derivation of shock jump conditions for weak, discontinuous shock solutions of (5.128), in which the region $\mathcal{R}$ is split into two disjoint sets $\mathcal{R}_1$ and $\mathcal{R}_2$ in the $(x.t)$ plane: $\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2$, in which $\mathcal{R}_1$ is to the left of the discontinuity surface $\Sigma(x,t) = 0$ and $\mathcal{R}_2$ is to the right of $\Sigma(x,t) = 0$. Applying Green's formula to $\mathcal{R}_1$ we obtain:

$$\begin{aligned} I_1 &= \iint_{\mathcal{R}_1} \mathbf{F} \cdot \nabla\phi \, dx \, dt = \iint_{\mathcal{R}_1} \nabla \cdot (\mathbf{F}\phi) - \phi\nabla \cdot \mathbf{F} \, dx \, dt \\ &= \int_{\Sigma=0} \phi(\mathbf{F}_1 \cdot \mathbf{n}) \, d\sigma. \end{aligned} \tag{5.136}$$

where $\mathbf{F}_1$ is the value of $\mathbf{F}$ on the shock, just inside $\mathcal{R}_1$ and $\mathbf{n}$ is the outward normal to $\mathcal{R}_1$. In the derivation of (5.136) we use the fact that $\nabla \cdot \mathbf{F} = 0$ in $\mathcal{R}_1$ and that ϕ vanishes on that part of the boundary $\partial\mathcal{R}_1$ not including the shock. Similarly, for the region $\mathcal{R}_2$ we find:

$$I_2 = \iint_{\mathcal{R}_2} \mathbf{F} \cdot \nabla\phi \, dx \, dt = -\int_{\Sigma=0} \phi(\mathbf{F}_2 \cdot \mathbf{n}) \, d\sigma. \tag{5.137}$$

Noting that $I = I_1 + I_2$, the weak solution condition (5.135) reduces to:

$$\int_{\Sigma=0} \phi(\mathbf{F}_1 - \mathbf{F}_2) \cdot \mathbf{n} \, d\sigma = 0. \tag{5.138}$$

Since $\phi(x,t)$ is arbitrary, (5.138) implies:

$$(\mathbf{F}_2 - \mathbf{F}_1) \cdot \mathbf{n} = [\mathbf{F}] \cdot \mathbf{n} = 0, \tag{5.139}$$

where $[\mathbf{F}] \cdot \mathbf{n}$ is the jump in $\mathbf{F} \cdot \mathbf{n}$ across the shock. Finally, noting that $\mathbf{n} = (1, -s)/(1+s^2)^{1/2}$ is the outward normal to Σ in region $\mathcal{R}_1$ where $s = dx/dt$ is the shock speed and using $\mathbf{F} = [f(u), u]$, the jump condition (5.139) reduces to the usual Rankine Hugoniot shock jump conditions:

$$[f(u)] - s[u] = 0. \tag{5.140}$$

If $s \neq 0$, then (5.140) may be solved for the shock speed s as:

$$s = \frac{[f(u)]}{[u]}. \tag{5.141}$$

There are many examples and generalizations of the shock jump conditions (5.139)–(5.141). Some examples are:

(1) Consider solutions of the inviscid Burgers equation:

$$u_t + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0. \tag{5.142}$$

In this case $f(u) = u^2/2$ and (5.141) gives:

$$s = \frac{1}{2}\frac{u_r^2 - u_l^2}{u_r - u_l} = \frac{1}{2}\left(u_r + u_l\right), \tag{5.143}$$

for the shock speed, where u_l and u_r denote the values of u just left and right of the shock, respectively.

(2) The Rankine Hugoniot conditions for shocks in 1D gas dynamics may be described by shock jump conditions of the form (5.140). In this case, the mass, momentum and energy conservation equations give rise to the shock jump conditions:

$$\begin{aligned}&[\rho u] = s[\rho],\\ &[\rho u^2 + p] = s[\rho u],\\ &\left[u\rho\left(\frac{1}{2}u^2 + \varepsilon + \frac{p}{\rho}\right)\right] = s\left[\rho\varepsilon + \frac{1}{2}\rho u^2\right],\end{aligned}$$

where ρ, u, p, ε denote the gas density, velocity, pressure and internal energy density per unit mass, respectively. In physics and mathematics texts the shock jump conditions (5.144) are analyzed in the shock frame, where $s = 0$ (e.g. [44, Ch. 2], [102, Ch. 9]).

(3) In relativistic gas dynamics the jump conditions (5.139) can be easily generalized to describe shocks in three space and one time dimension. In this case, the shock jump conditions:

$$\left[n^\beta\right] w_\beta = 0, \quad \left[T^{\alpha\beta}\right] w_\beta = 0, \tag{5.144}$$

correspond to the number density conservation equation and the conservation of energy and momentum, respectively. Here n^β is the number density four current, $T^{\alpha\beta}$ is the stress energy tensor, w_α is the four-velocity of the shock, and $\alpha, \beta = 0, 1, 2, 3$ correspond to the time and space coordinates, respectively.

5.4.3 Riemann Problems

In this section we discuss some simple examples of Riemann problems. A typical Riemann problem in one space dimension consists of obtaining weak solutions of a given hyperbolic system, in which the left and right states U_l and U_r are specified constant states to the left and right of a discontinuity at $x = 0$ at time $t = 0$. As time t increases, a region develops between the left and right states, which may involve discontinuities of different types coupled by compound waves (e.g. as in the case of piston problems in 1D gas dynamics [102]). Our aim is not to be exhaustive (there are many good texts that treat the Riemann problem in greater detail: e.g. [44,111,170,177]). Our aim is to give some simple examples of Riemann problems for linear wave systems and for Burgers equation that illustrate the basic principles. In particular, the Riemann problem for the linear hyperbolic system:

$$U_t + (AU)_x = 0, \tag{5.145}$$

where A is a constant matrix with real, distinct, eigenvalues and a complete set of eigenvectors will be discussed in detail (see also [111]). This problem is essential for understanding the basis of finite volume Riemann solvers and linear Roe solvers, discussed in detail in Section 5.4.5.

Example

Consider the solution of the linear wave equation:

$$u_t + au_x = 0, \tag{5.146}$$

with initial data:

$$u(x, 0) \equiv g(x) = u_l[1 - H(x)] + u_r H(x), \tag{5.147}$$

where $H(x)$ is the Heaviside step function ($H(x) = 0$ if $x < 0$ and $H(x) = 1$ for $x > 0$). We assume that $u_l > u_r$ and that a is a positive constant.

The general solution of (5.146) by the method of characteristics, with initial data (5.146) in the regions not involving shocks is:

$$u(x, t) = g(x - at) = u_l[1 - H(x - at)] + u_r H(x - at), \tag{5.148}$$

In other words:

$$u(x, t) = \begin{cases} u_l & \text{if } x < at, \\ u_r & \text{if } x > at, \end{cases} \tag{5.149}$$

is the solution of the Riemann problem. The speed of the shock according to (5.141) is given by:

$$s = \frac{[au]}{[u]} = a. \tag{5.150}$$

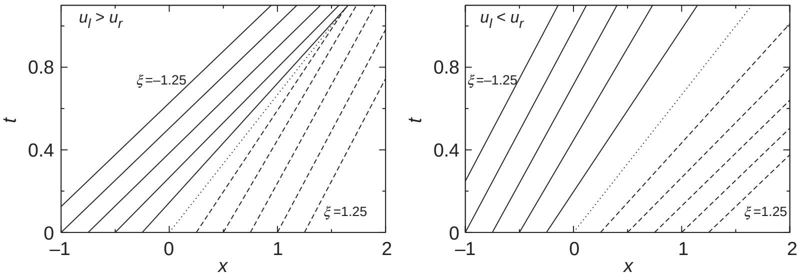


Figure 5.1: Left: characteristics of the equation (5.151) for the initial value problem with $u_l > u_r$. Right: corresponding characteristics for the case $u_l < u_r$.

Hence the speed of the shock is given by the linear wave speed a. It is instructive to plot the characteristics and shock locus in the (x, t)-plane, in which x is the abscissa and t is the ordinate, as illustrated Figure 5.1. The characteristics are given by $t = (x - \xi)/a$ where ξ is the point where the curve cuts the x-axis ($\xi < 0$ to the left of the shock where $u = u_l$ and $u = u_r$ on the right of the shock where $\xi > 0$). Note that there is a jump in the value of u across the shock $t = x/a$ between the left and right states.

Example

Consider the solution of the inviscid Burgers equation:

$$u_t + uu_x = 0, \tag{5.151}$$

subject to the initial conditions (5.147) (i.e. $u(x, 0) = u_l$ if $x < 0$ and $u(x, 0) = u_r$ for $x > 0$, where $u_l > u_r$). By the method of characteristics, the solution for $u(x, t)$ in regions not involving the shock, is given by:

$$u(x, t) = g(\xi) \quad \text{where } \xi = x - g(\xi)t. \tag{5.152}$$

In other words:

$$u(x, t) = \begin{cases} u_l & \text{if } x < st, \\ u_r & \text{if } x > st, \end{cases} \tag{5.153}$$

where

$$s = \frac{[f(u)]}{[u]} = \frac{1}{2}\frac{[u^2]}{[u]} = \frac{1}{2}(u_r + u_l), \tag{5.154}$$

is the speed of the shock (see also (5.143)). Notice that the solution of the Riemann problem for Burgers equation above is different from that for the linear wave equation in (5.146) with the same initial data (the shock speed is different for the two problems).

A plot of the characteristics in the (x, t) plane for the case $u_l > u_r$ is given in Figure 5.1, left. The slopes of the characteristics are $m_l = 1/u_l$, and $m_r = 1/u_r$ to the left and right of the shock, respectively, and the slope of the shock locus is $m_s = 1/s$, where s is given by (5.154), and $m_l < m_s < m_r$ (i.e. $1/u_l < 1/s < 1/u_r$). Note that the solution can be traced back to the initial data on the x-axis in an unambiguous fashion.

However, if $u_l < u_r$ (Figure 5.1, right), there is not a well-defined, unique solution of the Riemann problem between the characteristics $t = x/u_l$ and $t = x/u_r$. In this wedge shaped region, the characteristics cannot be traced back to the initial data at $x = 0$, but instead, they trace back to a point on the shock. In other words, the characteristics in this region originate at the shock and are directed forward in time.

For $u_l < u_r$ there is another weak solution of the inviscid Burgers equation (5.151), namely the rarefaction wave:

$$u(x,t) = \begin{cases} u_l & \text{if } x < u_l t, \\ x/t & \text{if } u_l t \leq x \leq u_r t, \\ u_r & \text{if } x > u_r t. \end{cases} \tag{5.155}$$

This solution is stable to perturbations and is the vanishing viscosity limit to the viscous Burgers equation.

Lax Geometrical Entropy Conditions

Lax [103,104] considered the solution of the Riemann problem for the hyperbolic system:

$$U_t + A(U)U_x = 0, \tag{5.156}$$

where $A(u)$ is an $m \times m$ matrix, with m real distinct eigenvalues: $\lambda_1 < \lambda_2 \ldots < \lambda_m$, and a complete set of corresponding right and left eigenvectors. A necessary condition that the Riemann problem:

$$U(x,0) = \begin{cases} U_l & \text{if } x < 0, \\ U_r & \text{if } x > 0, \end{cases} \tag{5.157}$$

has a unique solution, known as a k-shock, is if the eigenvalues for the left and right states are ordered so that:

$$\begin{aligned} &\lambda_{k-1}(U_l) < s < \lambda_k(U_l), \\ &\lambda_k(U_r) < s < \lambda_{k+1}(U_r), \end{aligned} \tag{5.158}$$

where s is the speed of the shock, separating the left and right states.

The Lax inequalities (5.158) may also be combined to give the equivalent inequalities:

$$\begin{aligned}&\lambda_k(U_r) < s < \lambda_k(U_l),\\ &\lambda_{k-1}(U_l) < s < \lambda_{k+1}(U_r),\end{aligned} \tag{5.159}$$

which are often easier to use in applications.

The proof of the Lax entropy conditions depends on the solution of the so-called quarter plane problem for the wave equation $u_t + au_x = 0$. For the quarter plane problem for the right state, one considers the solution of $u_t + au_x = 0$ in the quarter plane $x > 0$ and $t > 0$ with initial data $u = u_r$ at $t = 0$. If $a < 0$, the solution is completely specified by the initial data along the positive x-axis. However, if $a > 0$, one must also specify information along the t-axis (i.e. $u(0,t)$). A similar quarter plane problem for the left state $u = u_l$ for $x < 0$ and $t > 0$, shows that for $a < 0$, data must be supplied along both the $x < 0$-axis, as well as along the $t > 0$-axis, whereas for $a > 0$ one only needs to give initial data along the $x < 0$ axis.

Next, consider the extension of these ideas to a hyperbolic system with m distinct eigenvalues, $\lambda_1(U) < \lambda_2(U) < \cdots < \lambda_m(U)$, in which there is a k-shock $x = st$ in the solution of the Riemann problem, for which:

$$\lambda_1(U_r) < \lambda_2(U_r) < \cdots < \lambda_k(U_r) < s < \lambda_{k+1}(U_r) < \cdots < \lambda_m(U_r).$$

In other words, $\lambda_i > s$ for $i > k$. From consideration of the equivalent quarter plane problem, it follows that for $\lambda_i - s > 0$, we need to specify data on the shock, as well as on the positive x axis. Thus we need to specify $U(st,t)$ for the $m-k$ eigenvalues $\lambda_{k+1}(u_r), \ldots, \lambda_m(U_r)$. Next look at the left state boundary conditions, for which:

$$\lambda_1(U_l) < \lambda_2(U_l) < \cdots < \lambda_j(U_l) < s < \lambda_{j+1}(U_l) < \cdots < \lambda_m(U_l),$$

for some index j. Consideration of the equivalent quarter plane problem for the eigenvalues $\lambda_i(U_l)$ with $i < j$, reveals that it is necessary to specify $U(st,t)$ on the shock for those λ_i for which $\lambda_i - s < 0$ (i.e. for $i < j$). However, there are also m Rankine Hugoniot conditions $s[U] = [f(U)]$ relating the upstream and downstream states. Eliminating the shock speed from these relations implies that there are $m-1$ conditions relating the left and right states at the shock. For a unique solution, it is necessary that the number of conditions arising from the characteristic conditions, i.e. $m - k + j$ should be equal to $m - 1$. Thus we require $j = k-1$, which in turn implies the Lax inequalities (5.159). This completes the proof.

Remark 1:

The proof and further discussion of the above result may be found in [44, 103,104].

See also [210], for an application to particle injection in cosmic ray modified shocks, and [209], for an application to mass loading at cometary bow shocks. Chorin and Marsden [44] apply the Lax entropy conditions to discuss combustion shocks.

Remark 2:

Lax [104] has shown that for an ideal gas, a shock is compressive if and only if it satisfies the Lax inequalities (5.158). Thus for ideal gases, the Lax inequalities (Lax entropy conditions) coincide with the condition that the physical entropy S must increase in the transition between the upstream and downstream states.

Example

Consider the case of shocks in 1D gas dynamics. In this case, the characteristic speeds are:

$$\lambda_1 = u - c, \quad \lambda_2 = u, \quad \lambda_3 = u + c,$$

corresponding to the backward sound wave, the entropy wave and the forward sound wave, where u is the fluid velocity and c is the sound speed, (e.g. [44,50]). In the shock frame, the fluid speeds, upstream and downstream of the shock are:

$$v_1 = u_1 - s = -|v_1|, \quad \text{and} \quad v_2 = u_2 - s = -|v_2|,$$

where s is the shock velocity. Analysis of gas dynamical shocks shows [102] that the shock consists of a compressive, supersonic to subsonic transition (i.e. $|v_1| > c_1$, $|v_2| < c_2$, $\rho_2 > \rho_1$, where the subscripts 1 and 2 denote the upstream and downstream states, and ρ is the gas density), in which the gas entropy S increases across the shock (i.e. $S_2 > S_1$). Using the above results, it follows that:

$$\begin{aligned}
&\lambda_1(U_l) < \lambda_2(U_l) < s < \lambda_3(U_l),\\
&\lambda_1(U_r) < \lambda_2(U_r) < \lambda_3(U_r) < s,
\end{aligned}$$

where U_r denotes the right, or upstream state and U_l denotes the left or downstream state. Thus the shock satisfies the Lax entropy conditions (5.159) for a 3-shock.

Remark 3:

The Lax entropy conditions, turn out to be the same as the so-called "evolutionary conditions for shock stability". In the shock evolutionary conditions, the question addressed is whether an initial perturbation upstream of the shock will give rise to a stable shock. In some cases, the evolutionary conditions based on linear wave analysis do not give a definitive answer as the expansion of the solution around the constant state starts with quadratic terms and one has to consider the nonlinear model equations, [15,33,35].

Remark 4:

In 1D gas dynamics, the Riemann problem, in which two gases, at relative rest at $x = 0$ at time $t = 0$, but with different pressures, in which $p_r > p_l$, at time $t > 0$, develops into a shock wave moving to the left, separated by a contact discontinuity from a rarefaction wave moving to the right. The rarefaction wave propagates into the undisturbed right state (e.g. [102, Ch. 10, Section 100]). This example shows the importance of compound waves in the solution of Riemann problems.

The Linear Riemann Problem

The Riemann problem for the system:

$$U_t + (AU)_x = 0, \tag{5.160}$$

where A is a constant $m \times m$ matrix, with m distinct, real eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_m$, and a complete set of right and left eigenvectors $\{R_j\}$ and $\{L_j\}$ $(j = 1(1)m)$, normalized so that $L_j \cdot R_k = \delta_{jk}$, is important as a step in understanding the fully nonlinear Riemann problem in which $A = A(U)$ is nonlinear in U (e.g. [111]).

As discussed in Chapter 1, Section 1.1, the system (5.160) can be diagonalized by taking the scalar product, on the left with the left eigenvectors $\{L_p\}$ to obtain the diagonalized system:

$$\frac{\partial v_p}{\partial t} + \lambda_p \frac{\partial v_p}{\partial x} = 0, \quad \text{where } v_p = L_p \cdot U. \tag{5.161}$$

The general solution of (5.161) for v_p has the form $v_p(x,t) = v_p(x - \lambda_p t, 0)$. By using the orthogonality of the left and right eigenvectors: $L_j \cdot R_k = \delta_{jk}$, it follows that the general solution for U has the form:

$$U(x,t) = \sum_{p=1}^{m} v_p(x - \lambda_p t) R_p, \tag{5.162}$$

where, in an abuse of mathematical notation, we use $v_p(x-\lambda_p t)$ to mean $v_p(x-\lambda_p t, 0)$. In the Riemann problem, one can expand the left and right states U_l and U_r in right eigenfunction expansions of the form:

$$U_l = \sum_{p=1}^{m} \alpha_p R_p, \quad \text{if } x < 0, \quad U_r = \sum_{p=1}^{m} \beta_p R_p, \quad \text{if } x > 0, \tag{5.163}$$

where the α_p and β_p are constants. Note that

$$\begin{aligned} \alpha_p &= L_p \cdot U_l = v_p(x,0), \quad \text{if } x < 0, \\ \beta_p &= L_p \cdot U_r = v_p(x,0), \quad \text{if } x > 0. \end{aligned}$$

Hence at time t:

$$v_p(x,t) = \begin{cases} \alpha_p & \text{if } x - \lambda_p t < 0, \\ \beta_p & \text{if } x - \lambda_p t > 0. \end{cases} \tag{5.164}$$

Based on the above analysis, the solution of the linear Riemann problem is:

$$U(x,t) = \sum_{p=1}^{P(x,t)} \beta_p R_p + \sum_{p=P(x,t)+1}^{m} \alpha_p R_p, \tag{5.165}$$

where $P(x,t)$ is the maximum integer for which $x - \lambda_p t > 0$ [111]. A geometric picture of the solution (5.165) for the case $m = 3$ is described by LeVeque [111]. Given a fixed point (x,t) in the (x',t') plane, the equation for the p^{th} characteristic passing through (x,t) is given by the straight line $x' = x + \lambda_p(t'-t)$, which cuts the x'-axis at $x' = x - \lambda_p t$ when $t' = 0$. The location of shocks in the (x',t') plane occur along the characteristics $x' = \lambda_j t'$ $(1 \le j \le m)$. Thus, for example, in the case $m = 3$, if the point (x,t) lies between $x' = \lambda_1 t'$ and $x' = \lambda_2 t'$ $(t' > 0)$, then

$$U(x,t) = \beta_1 R_1 + \alpha_2 R_2 + \alpha_3 R_3. \tag{5.166}$$

Thus there is a jump of $[U] = U - U_l = (\beta_1 - \alpha_1)R_1$ as one passes from the left state in the region $x' < \lambda_1 t'$ across the shock $x' = x_1'$ into the region $x_1' < x' < x_2'$ where $x_j' = \lambda_j t'$ denotes the position of the j^{th} shock. Similarly, as one passes across the p^{th} shock, $x' = x_p'$, there is a jump in the value of U of:

$$\begin{aligned}[U] &= \left(\sum_{j=1}^{p} \beta_j R_j + \sum_{j=p+1}^{m} \alpha_j R_j\right) - \left(\sum_{j=1}^{p-1} \beta_j R_j + \sum_{j=p}^{m} \alpha_j R_j\right) \\ &= (\beta_p - \alpha_p) R_p. \end{aligned} \tag{5.167}$$

According to the shock jump conditions (5.140), the jump in the flux across the p^{th} shock is:

$$\begin{aligned}[f(U)] = [AU] = A[U] &= A(\beta_p - \alpha_p) R_p \\ &= \lambda_p (\beta_p - \alpha_p) R_p = \lambda_p [U], \end{aligned} \tag{5.168}$$

which demonstrates that the velocity of the p^{th} shock is $s = \lambda_p$.

From the solution (5.165) it also follows that:

$$\begin{aligned} U(x,t) - U_l &= \sum_{x - \lambda_p t > 0} \beta_p R_p + \sum_{x - \lambda_p t < 0} \alpha_p R_p - \sum_{p=1}^{m} \alpha_p R_p \\ &= \sum_{x_p - \lambda_p t > 0} (\beta_p - \alpha_p) R_p, \end{aligned} \tag{5.169}$$

and hence:

$$U(x,t) = U_l + \sum_{\lambda_p < \xi} (\beta_p - \alpha_p) R_p, \quad \text{where } \xi = x/t, \tag{5.170}$$

is the solution for $U(x,t)$ in terms of the left state U_l plus the jumps across the shocks in between the left state and the position (x,t). A similar calculation gives the solution for $U(x,t)$ in the form:

$$U(x,t) = U_r - \sum_{\lambda_p > \xi} (\beta_p - \alpha_p) R_p, \tag{5.171}$$

where U_r is the right state. Finally, using (5.161) or the above two equations, we obtain:

$$U_r - U_l = \sum_{p=1}^{m} (\beta_p - \alpha_p) R_p. \tag{5.172}$$

Thus the basic idea behind the solution of the linear Riemann problem is in how to split up the sum (5.172) to give the solution for $U(x,t)$.

The same idea also applies to the fully nonlinear Riemann problem, except that the jumps across the shocks, $[f(U)] \neq \lambda_p[U]$ because the eigenvalues are now nonlinear functions of U. Also for the nonlinear Riemann problem, rarefaction waves may be a necessary part of the solution for $U(x,t)$, connecting the left and right states.

Riemann Invariants

Riemann invariants, are physical quantities that are constant on the characteristics. For example, in isentropic flow in 1D gas dynamics, the Riemann invariants for the forward ($\lambda = u + c$) and backward ($\lambda = u - c$) sound waves have the form:

$$J_{\pm} = u \pm \int c(\rho)\frac{d\rho}{\rho} = u \pm \frac{2c}{\gamma - 1}. \tag{5.173}$$

Here γ is the adiabatic index of the gas, and we assume $\gamma \neq 1$. If $\gamma = 1$ (isothermal gas) $c =$ const. and the integral over ρ equals $c\ln(\rho)$ in that case.

5.4.4 Upwind Differencing Schemes

Example. Advection equation with variable speed.

Let

$$u_t + a(x)u_x = 0,$$

where $a(x)$ denotes the variable advection speed. Consider an upwind numerical method that applies for both forward or backward differentiation depending on whether the sign of the advection speed is positive or negative, respectively, i.e.

$$u_j^{n+1} = u_j^n - \nu_{j-1/2}\,(u_j^n - u_{j-1}^n), \quad \text{for } a_{j-1/2} > 0, \tag{5.174}$$

$$u_j^{n+1} = u_j^n - \nu_{j+1/2}\,(u_{j+1}^n - u_j^n), \quad \text{for } a_{j+1/2} < 0, \tag{5.175}$$

where $\nu_{j\pm1/2} = a_{j\pm1/2}\Delta t/\Delta x$. The modified equation describing the truncation error for this method has a leading order truncation error term consisting of first-order numerical diffusion. For constant a case, (5.174) and (5.175) can be combined in the formula:

$$u_j^{n+1} = u_j^n - \nu_{j-1/2}^{+}(u_j^n - u_{j-1}^n) - \nu_{j+1/2}^{-}(u_{j+1}^n - u_j^n), \tag{5.176}$$

where

$$\begin{aligned}&\nu^+_{j-1/2} = \max(\nu_{j-1/2}, 0), \quad \nu^-_{j+1/2} = \min(\nu_{j+1/2}, 0),\\ &\nu_j = \nu^+_{j-1/2} + \nu^-_{j+1/2}, \quad |\nu_j| = \nu^+_{j-1/2} - \nu^-_{j+1/2}.\end{aligned} \tag{5.177}$$

Noting that

$$\nu^+_{j-1/2} = \frac{1}{2}(\nu_j + |\nu_j|), \quad \nu^-_{j+1/2} = \frac{1}{2}(\nu_j - |\nu_j|),$$

(5.176) can be re-written in the form:

$$u^{n+1}_j = u^n_j - \frac{\nu_j}{2}(u^n_{j+1} - u^n_{j-1}) + \frac{|\nu_j|}{2}(u^n_{j+1} - 2u^n_j + u^n_{j-1}), \tag{5.178}$$

showing that upwinding introduces numerical diffusion proportional to $|a|$. Note that for $a > 0$, x_{j-1} is upstream of x_j, but for $a < 0$, x_{j+1} is upstream of x_j.

Example. Linear System of Hyperbolic Conservation Laws

Consider the system:

$$U_t + (AU)_x = 0, \tag{5.179}$$

where A is a constant matrix with real eigenvalues and a complete set of eigenvectors. As was discussed in Chapter 1, the solution can be decomposed into eigenmodes by multiplying both sides by the left eigenvector L_j:

$$\partial_t w_j + \lambda_j \partial_x w_j = 0, \tag{5.180}$$

where the j^{th} wave amplitude w_j is defined as $w_j = L_j \cdot U$. The solution can be reassembled as superposition of the eigenmodes in the form:

$$U = \sum_j w_j R_j, \tag{5.181}$$

with right and left eigenvectors normalized so that $L_j R_k = \delta_{jk}$. Such wave decomposition allows one to use the exact solution for each scalar advection equation or to use different numerical methods for various fields. For example, one can introduce numerical diffusion into some fields via upwind discretization while avoiding numerical diffusion for others. To obtain an upwind difference scheme analogous to (5.178) in the present example, first note that:

$$U = \sum_{k=1}^{m} w_k R_k = \sum_{s=1}^{m} U^s a_s, \tag{5.182}$$

where a_s is the s^{th} unit vector describing the state vector U in the original variables. From (5.182) we obtain $U^s = R^s_k w_k$ or in matrix notation:

$$U = RW, \tag{5.183}$$

describing the effect of the change of base vectors. In the $\{\mathbf{R}_k\}$ base system, we use the notation:

$$\begin{aligned} \lambda_p^+ &= \max(\lambda_p, 0), \quad \Lambda^+ = \text{diag}(\lambda_1^+, \lambda_2^+, \dots, \lambda_m^+), \\ \lambda_p^- &= \min(\lambda_p, 0), \quad \Lambda^- = \text{diag}(\lambda_1^-, \lambda_2^-, \dots, \lambda_m^-), \end{aligned} \tag{5.184}$$

to describe the right and left propagating waves in the diagonalized system (5.180). The upwind difference equations for (5.180) can then be written in the form:

$$W_j^{n+1} = W_j^n - \frac{\Delta t}{\Delta x}\Lambda^+ \left(W_j^n - W_{j-1}^n\right) - \frac{\Delta t}{\Delta x}\Lambda^- \left(W_{j+1}^n - W_j^n\right), \tag{5.185}$$

which is the analog of (5.176). Pre-multiplying (5.185) by the matrix R and converting (5.185) back to the original base gives the equation:

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x} A^+ \left(U_j^n - U_{j-1}^n\right) - \frac{\Delta t}{\Delta x} A^- \left(U_{j+1}^n - U_j^n\right), \tag{5.186}$$

where

$$A^+ = R\Lambda^+ R^{-1}, \quad \text{and} \quad A^- = R\Lambda^- R^{-1}. \tag{5.187}$$

Noting that

$$\begin{aligned} &A = A^+ + A^- = R\Lambda R^{-1}, \quad |A| = A^+ - A^- = R|\Lambda|R^{-1}, \\ &\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m), \quad |\Lambda| = \text{diag}(|\lambda_1|, |\lambda_2|, \dots, |\lambda_m|), \end{aligned} \tag{5.188}$$

(5.186) may be written in the form:

$$\begin{aligned} U_j^{n+1} = U_j^n &- \frac{1}{2}\frac{\Delta t}{\Delta x} A (U_{j+1}^n - U_{j-1}^n) \\ &+ \frac{1}{2}\frac{\Delta t}{\Delta x}|A|(U_{j+1}^n - 2U_j^n + U_{j-1}^n). \end{aligned} \tag{5.189}$$

Equation (5.189) is the analog of (5.178) for the linear advection equation for the scalar u. The last term in (5.189) involving $|A|$ represents numerical diffusion. The matrix R consists of the right eigenvectors of A written as columns and Λ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_m$. The solutions for U_j^{n+1} in (5.186) and (5.189) are obtained in the next section by using Godunov's method applied to the linear system (5.179). For nonlinear systems, analogous splitting of the equations can be introduced to incorporate upwinding. These methods are related to Riemann problems, with right and left going waves.

5.4.5 Godunov, Finite Volume Riemann Solvers

In Godunov's method, the numerical solution of the conservative system:

$$U_t + (F(U))_x = 0, \tag{5.190}$$

is conceived as consisting of piecewise constant values U_j^n on the j^{th} grid cell $x_{j-1/2} < x < x_{j+1/2}$. Integration of (5.190) over the region $t_n < t < t_{n+1}$ and $x_{j-1/2} < x < x_{j+1/2}$ gives the finite volume approximation:

$$U_j^{n+1} = U_j^n - \nu \left(F_{j+1/2}^n - F_{j-1/2}^n \right), \tag{5.191}$$

where $\nu = \Delta t / \Delta x$ and

$$U_j^n = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} U(x, t_n) dx, \tag{5.192}$$

$$F_{j+1/2}^n = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} F\left[U(x_{j+1/2}, t) \right] dt. \tag{5.193}$$

The numerical interface flux $F_{j\pm1/2}^n = F(U_{j\pm1/2}^n)$ where $U_{j\pm1/2}^n$ is defined by appropriate rules depending on the approximation used. A simple approach of arithmetic averaging of nearby cell values leads to the unstable centered difference scheme. The Riemann solver approach determines the interface value at position $x_{j+1/2}$ by solving exactly the initial value problem with discontinuous initial data consisting of two constant states U_l and U_r. In first-order methods $U_l = U_j^n$ and $U_r = U_{j+1}^n$, while in higher order methods U_l and U_r are computed by higher order non-oscillatory interpolations from several nearby values which are described at the end of this section.

The Linear Case

In the linear case:

$$F(U) = AU, \tag{5.194}$$

where A is a constant matrix, use is made of the linear Riemann problem described by (5.160)–(5.172). From (5.170)–(5.171), the interface value $U^n_{j+1/2}$ is given by:

$$U^n_{j+1/2} = U_l + \sum_{\lambda_k<0} c^n_{k,j+1/2} R_k = U_r - \sum_{\lambda_k>0} c^n_{k,j+1/2} R_k, \tag{5.195}$$

where $U_l = U^n_j$ and $U_r = U^n_{j+1}$ and $c^n_{k,j+1/2} \equiv \beta_k - \alpha_k$. The decomposition coefficients (wave amplitudes) $c^n_{k,j+1/2}$ are defined in terms of the mode decomposition of the initial jump:

$$U_r - U_l = \sum_k c^n_{k,j+1/2} R_k. \tag{5.196}$$

The corresponding flux $F^n_{j+1/2} = F(U^n_{j+1/2}) = AU^n_{j+1/2}$. If we use the expansion (5.195) for $U^n_{j+1/2}$, involving U_r, then we obtain:

$$F^n_{j+1/2} = AU^n_{j+1} - \sum_{\lambda_k>0} \lambda_k c^n_{k,j+1/2} R_k. \tag{5.197}$$

By using the matrices A^- and A^+ introduced in the upwind scheme in (5.187)–(5.188), (5.197) becomes:

$$F^n_{j+1/2} = AU^n_{j+1} - A^+ \left(U^n_{j+1} - U^n_j\right), \tag{5.198}$$

which gives the interface flux in terms of the left and right states $U_l = U^n_j$ and $U_r = U^n_{j+1}$. Alternatively, by noting that $A = A^- + A^+$, (5.198) can be written in the form:

$$F^n_{j+1/2} = AU^n_j + A^- \left(U^n_{j+1} - U^n_j\right). \tag{5.199}$$

This form for $F^n_{j+1/2}$ corresponds to the left state expansion in (5.195). Yet another expansion follows by noting that:

$$A^+ = \frac{1}{2}\left(A + |A|\right), \quad A^- = \frac{1}{2}\left(A - |A|\right), \tag{5.200}$$

in (5.198) to obtain:

$$\begin{aligned}F^n_{j+1/2} &= \frac{1}{2}A\left(U^n_{j+1}+U^n_j\right)-\frac{1}{2}|A|\left(U^n_{j+1}-U^n_j\right)\\ &\equiv \frac{1}{2}\left[F_r+F_l-|A|(U_r-U_l)\right],\end{aligned} \tag{5.201}$$

where $F_l = AU^n_j$ and $F_r = AU^n_{j+1}$ are the left and right fluxes.

The flux $F^n_{j-1/2}$ across the $x_{j-1/2}$ interface can be obtained from the above formulae (5.198)–(5.201) simply by replacing j by $j-1$. Then using (5.198) to compute $F^n_{j+1/2}$ and $F^n_{j-1/2}$ in the Godunov scheme (5.191) gives the required solution for U^{n+1}_j in the form:

$$U^{n+1}_j = U^n_j - \nu\left[A^+\left(U^n_j - U^n_{j-1}\right)+A^-\left(U^n_{j+1}-U^n_j\right)\right]. \tag{5.202}$$

The solution (5.202) can be recognized as the solution for U^{n+1}_j obtained in the upwind scheme (5.186) with $\nu = \Delta t/\Delta x$. Thus the upwind scheme for the linear hyperbolic system, in which the flux $F(U) = AU$, where A is a constant matrix, is equivalent to the Godunov scheme solution based on (5.191). The solution (5.202) can also be written in the alternative form (5.189) obtained by using the matrices A and $|A|$ instead of A^+ and A^-.

Approximate Riemann Solvers

In this section we give a brief synopsis of approximate Riemann solvers, and Roe solvers. A more complete analysis is given for example by LeVeque [111]. Godunov's method and related higher order methods require the solution of Riemann problems at every cell boundary. In the nonlinear case, these Riemann problems could be solved in principle, but in practice this is expensive and involves iteration of the equations. The resulting fine detail of the Riemann solver is not used in Godunov's method, in which the exact solution is averaged over each grid cell. Thus it is more practical and less expensive to obtain equally good numerical solutions with approximate Riemann solvers. In approximate Riemann solvers, the approximate solution $\hat{U}(x,t) = \hat{W}(\xi)$ is typically used in which $\xi = x/t$ and:

$$\hat{W}(\xi) = \begin{cases} U_l & \text{for } \xi < a_{\min},\\ U_r & \text{for } \xi > a_{\max},\end{cases}$$

where $a_{\min}$ and $a_{\max}$ are the minimum and maximum speeds based on the eigenvalues of the matrices $F'(U_l)$ and $F'(U_r)$, and $F(U)$ is the flux function in (5.190). In addition some version of the Courant condition, for example,

$C = a\Delta t/\Delta x < 1$, needs to be imposed in the numerical implementation to maintain the stability of the method.

Consider the solution of the Riemann problem in one Cartesian space dimension x in which $U = U_l$ for $x < 0$ and $U = U_r$ for $x > 0$ at time $t = 0$. Note that the linear Riemann solution in (5.160) et seq. can be expressed in the similarity solution form $U(x,t) = W(\xi)$, where $\xi = x/t$. More generally, we can write $U(x,t) = W(\xi,t)$, at least in regions where $J = \partial(\xi,t)/(x,t)$ is well defined. Changing variables from (x,t) to (ξ,t) in the conservative system (5.190) gives rise to the PDE system:

$$t\frac{\partial W}{\partial t} - \xi\frac{\partial W}{\partial \xi} + \frac{\partial F(W)}{\partial \xi} = 0. \tag{5.203}$$

Integrating the system (5.203) over the region $-M < \xi < M$ and $0 < t < T$ gives rise to the conservation equation:

$$\int_{-M}^{M} d\xi\ W(\xi,T) = M(U_r + U_l) + F(U_l) - F(U_r). \tag{5.204}$$

Because there are no length scales in the problem, it is reasonable to assume $W(\xi,T) = W(\xi)$ is independent of T.

In the approximate Riemann problem, one approximates $W(\xi)$ by $\hat{W}(\xi)$ where $\hat{W}(\xi)$ is the solution of the Riemann problem for the modified (and easier to solve) conservative system:

$$\frac{\partial \hat{U}}{\partial t} + \frac{\partial \hat{F}(\hat{U})}{\partial x} = 0. \tag{5.205}$$

In this case we obtain, analogous to (5.204), the conservation law:

$$\int_{-M}^{M} d\xi\ \hat{W}(\xi) = M(U_r + U_l) + \hat{F}(U_l) - \hat{F}(U_r). \tag{5.206}$$

Hence from (5.204) and (5.206) we require the solution of the approximate Riemann problem to have a flux function satisfying the equation:

$$\hat{F}(U_r) - \hat{F}(U_l) = F(U_r) - F(U_l), \tag{5.207}$$

where $F(U)$ is the flux for the original problem.

By integrating (5.203) over the region $(0,M) \times (0,T)$ and also over the region $(-M,0) \times (0,T)$ of the (ξ,t) plane, we obtain two equations for the interface flux function $F(\hat{W}(0)) = \mathcal{F}(U_l,U_r)$ (note for the j^{th} cell

$F_{j+1/2} = \mathcal{F}(U_j, U_{j+1})$ in (5.191)), namely:

$$\mathcal{F}(U_l, U_r) = F(U_r) - MU_r + \int_0^M \hat{W}(\xi)\, d\xi, \tag{5.208}$$

$$\mathcal{F}(U_l, U_r) = F(U_l) + MU_l - \int_{-M}^0 \hat{W}(\xi)\, d\xi. \tag{5.209}$$

By using the analog of (5.208) for the approximate Riemann problem, it follows that:

$$\mathcal{F}(U_l, U_r) = \hat{\mathcal{F}}(U_l, U_r) - \hat{F}(U_r) + F(U_r), \tag{5.210}$$

where $\hat{\mathcal{F}} = \hat{F}[\hat{W}(0)]$ is the interface flux for the approximate Riemann solver.

Roe's Approximate Riemann Solver

Roe suggested that instead of solving the exact Riemann problem for a nonlinear system (5.190) as is done in the Godunov method, one solves the constant matrix linear system:

$$\hat{U}_t + \left(\hat{A}(U_l, U_r)\hat{U}\right)_x = 0, \tag{5.211}$$

with flux function $\hat{F}(\hat{U}) = \hat{A}\hat{U}$, where $\hat{A}(U_l, U_r)$ is a constant matrix depending on U_l and U_r with a complete set of real eigenvalues, λ_p and right and left eigenvectors $\hat{R}_p$ and $\hat{L}_p$. The matrix $\hat{A}(U_l, U_r)$ is chosen to satisfy the conditions:

$$\begin{aligned}
&(1)\ \hat{A}(U_l, U_r)(U_r - U_l) = F(U_r) - F(U_l),\\
&(2)\ \hat{A}(U_l, U_r) \text{ is diagonalizable with real eigenvalues},\\
&(3)\ \hat{A}(U_l, U_r) \to F'(U) \quad \text{as } U_r, U_l \to \bar{U},
\end{aligned} \tag{5.212}$$

where $\bar{U}$ is the average state.

The solution of the linear Riemann problem for (5.211) using (5.170)–(5.172) is given by:

$$\hat{W}(\xi) = U(x,t) = U_l + \sum_{\lambda_p < \xi} c_p \hat{R}_p = U_r - \sum_{\lambda_p > \xi} c_p \hat{R}_p, \tag{5.213}$$

where

$$U_r - U_l = \sum_{p=1}^{m} c_p \hat{R}_p, \tag{5.214}$$

is the expansion for $U_r - U_l$ in terms of the eigenvectors $\hat{R}_p$ of the matrix $\hat{A}(U_l, U_r)$ and $\xi = x/t$. The condition (5.212)(1), is a restatement of the condition (5.207) that the difference of the fluxes for the approximate Riemann problem $\hat{F}(U_r) - \hat{F}(U_l)$ should equal $F(U_r) - F(U_l)$ for the exact problem. Condition (5.212)(1) also implies the conservation law (5.204) if $W(\xi, t)$ is replaced by $\hat{W}(\xi)$. It also ensures that if U_l and U_r are connected by a single shock or contact discontinuity, then the approximate solution agrees with the exact solution.

Noting that $\hat{\mathcal{F}}(U_l, U_r) = \hat{A}[\hat{W}(0)]$ and using (5.213) to determine $\hat{W}(0)$, (5.210) gives the expressions:

$$\mathcal{F}(U_l, U_r) = F(U_l) + \sum_{p=1}^{m} \lambda_p^- c_p \hat{R}_p = F(U_r) - \sum_{p=1}^{m} \lambda_p^+ c_p \hat{R}_p, \tag{5.215}$$

for the flux at the cell boundary between U_l and U_r, where $\lambda_p^+ = \max(\lambda_p, 0)$ and $\lambda_p^- = \min(\lambda_p, 0)$ for the forward and backward waves. The flux formulae (5.215) applied to the j^{th} cell $(x_{j-1/2}, x_{j+1/2})$ gives the flux $F_{j+1/2} = \mathcal{F}(U_j, U_{j+1})$ in the alternative forms:

$$\begin{aligned} F^n_{j+1/2} &= F\left(U^n_{j+1}\right) - \sum_p \lambda_p^+ c^n_{p,j+1/2} \hat{R}_p \\ &= F\left(U^n_j\right) + \sum_p \lambda_p^- c^n_{p,j+1/2} \hat{R}_p, \end{aligned} \tag{5.216}$$

for the flux through the cell boundary at $x_{j+1/2}$ (one can also use the average of the above two fluxes for $F^n_{j+1/2}$). To obtain the corresponding fluxes through the left boundary at $x = x_{j-1/2}$, simply replace j by $j-1$ in the above formulae. The required solution for U_j^{n+1} can now be obtained by using the above formulae to determine the fluxes $F^n_{j+1/2}$ and $F^n_{j-1/2}$ in the Godunov scheme (5.191).

Example: Roe Solver for Isothermal Gas Dynamics

For 1D, isothermal gas dynamics, the conservation laws for mass and momentum may be written in the conservative form:

$$U_t + F(U)_x = 0, \tag{5.217}$$

where

$$U=(\rho,m)^t,\quad F(U)=\left(m,\frac{m^2}{\rho}+a^2\rho\right)^t,\quad m=\rho v,\ p=a^2\rho. \tag{5.218}$$

Here, ρ is the gas density, $m=\rho v$ is the mass flux and v is the fluid velocity, $a=\sqrt{p/\rho}$ is the constant isothermal sound speed and p is the gas pressure. LeVeque [111] introduces the state vector $z=U/\sqrt{\rho}$ and shows that the matrix

$$\hat{A}=\begin{pmatrix}0 & 1\\ a^2-\bar{v}^2 & 2\bar{v}\end{pmatrix}, \tag{5.219}$$

where

$$\bar{v}=\frac{\sqrt{\rho_l}v_l+\sqrt{\rho_r}v_r}{\sqrt{\rho_l}+\sqrt{\rho_l}}, \tag{5.220}$$

satisfies the conditions (5.212) required for the Roe solver matrix $\hat{A}$.

The above analysis gives the basic idea used in the Roe scheme to obtain approximate solutions of the Riemann problem for (5.190). The form of the matrix $\hat{A}$ in (5.219) and its eigen-system for the case of isothermal gas dynamics is worked out in [111]. The original Roe scheme applied to Euler's equations was worked out in [152]. However, modifications of the above scheme are needed if there are rarefaction waves involved in the Riemann problem, in which case it is necessary to apply an entropy fix. One such approach due to [80] is described by LeVeque [111, Ch. 14].

Godunov showed that for numerical approximations in the Godunov method to be monotonicity preserving, the Riemann solver has to be first=order accurate, [80]. Later on, to overcome this restriction, Harten replaced the monotonicity constraint by a weaker notion of TVD that allowed increasing of the spatial order of accuracy away from the local extrema points, [111]. These methods utilize non-oscillatory interpolation, which will be described later in the section.

5.5 Methods of Fractional Steps, Time-Split and Approximate Factorization Algorithms

For modeling complicated physical processes or solving multi-dimensional problems, it is tempting to design a numerical method using reliable and efficient methods known separately for each physical process at hand, for example, to construct a multi-dimensional algorithm from 1D ones; to split advection and collision terms in Boltzmann's equation; to separate advection-diffusion from reaction terms in advection-diffusion-reaction equations; to approximate 3D Euler or Naiver-Stokes in fluid dynamics using 1D algorithms, etc.

Consider, an initial value problem for a system of PDEs (or integro-differential equations):

$$U_t + A_1(U) + A_2(U) + A_3(U) = 0,$$

where $A_i s$ are nonlinear operators that may depend on spatial derivatives and time, but not on time derivatives of U. The claim is that the solution of the full system can be approximated by solving a sequence of sub-systems:

$$U_t + A_1(U) = 0, \quad U_t + A_2(U) = 0, \quad U_t + A_3(U) = 0,$$

with solution of the previous system serving as an initial data (input) for the next equation. In terms of the solution operator, see Chapter One, it can be written concisely as, $U(t) = S(t)U(0) \approx S_3(t)S_2(t)S_1(t)U(0)$, where S, S_1, S_2 and S_3 are solution operators to the full problem and each sub-system, respectively. Advantages of this approach are:

1. Design accurate and efficient algorithms for many complicated problems using available methods and software for each sub-problem.
2. Increase of accuracy by trading-off splitting error for truncation error of a sub-system that can be solved very accurately (e.g. spectrally accurate), or explicitly for problems with desperate time scales.

The possible disadvantages and difficulties that may preclude the use of fractional step methods are:

1. A need for boundary conditions at intermediate steps.
2. Large splitting error for the first order splitting algorithms, while high order extensions may be costly and very cumbersome to design.
3. For problems with discontinuities or other singularities, the splitting may break the balance between the terms that defines correct singular solutions and cause convergence to nonphysical solutions.
4. Possible instabilities and numerical oscillation triggered by splitting.

Example. Approximate Factorization: Alternating Direction Implicit Method

Historically, the alternating direction implicit method (ADI) was first applied [59,144] to the heat equation. The motivation was to replace the multi-dimensional Crank-Nicolson method by a sequence of 1D tridiagonal Gaussian elimination solvers (Thomas algorithm) that are more efficient, especially in terms of memory requirements in comparison to Gaussian elimination solvers for banded matrices, [177]. Today's computer resources and availability of fast preconditioned sparse iterative solvers may weaken the original motivation, but ADI inspired development of other fractional step methods may be used in applications far beyond the original problem. Let

$$U_t + L_1 U + L_2 U = 0,$$

where $L_1 = \partial_{xx}$ and $L_2 = \partial_{yy}$. Apply a two step scheme where the first step is implicit in the x-direction while the second step is implicit in the y-direction:

$$\begin{aligned} \frac{U^{n+1/2} - U^n}{\Delta t/2} &= \tilde{L}_1 U^{n+1/2} + \tilde{L}_2 U^n, \\ \frac{U^{n+1} - U^{n+1/2}}{\Delta t/2} &= \tilde{L}_1 U^{n+1/2} + \tilde{L}_2 U^{n+1}, \end{aligned}$$

or

$$\begin{aligned} \left(1 - \frac{\Delta t}{2}\tilde{L}_1\right) U^{n+1/2} &= \left(1 + \frac{\Delta t}{2}\tilde{L}_2\right) U^n, \\ \left(1 - \frac{\Delta t}{2}\tilde{L}_2\right) U^{n+1} &= \left(1 + \frac{\Delta t}{2}\tilde{L}_1\right) U^{n+1/2}, \end{aligned}$$

where $\tilde{L}_1$ and $\tilde{L}_2$ are numerical approximations to continuous operators L_1 and L_2, respectively, and $U^{n+1/2}$ is an intermediate variable. Multiplying the last equation by $\left(1 - \frac{\Delta t}{2}\tilde{L}_1\right)$, switching the order of the factors on the right hand side (since they commute), and using the first equation to eliminate intermediate variable $U^{n+1/2}$, gives:

$$\left(1 - \frac{\Delta t}{2}\tilde{L}_1\right)\left(1 - \frac{\Delta t}{2}\tilde{L}_2\right) U^{n+1} = \left(1 + \frac{\Delta t}{2}\tilde{L}_1\right)\left(1 + \frac{\Delta t}{2}\tilde{L}_2\right) U^n.$$

Comparing this approximate factorization update operator with the Crank-Nicolson operator,

$$\left(1 - \frac{\Delta t}{2}\tilde{L}_1 - \frac{\Delta t}{2}\tilde{L}_2\right) V^{n+1} = \left(1 + \frac{\Delta t}{2}\tilde{L}_1 + \frac{\Delta t}{2}\tilde{L}_2\right) V^n,$$

gives a local splitting error of order $O(\Delta t)^3$, $\frac{(\Delta t)^2}{4}\tilde{L}_1\tilde{L}_2(V^{n+1} - V^n) = O(\Delta t)^3$, [177], thus implying a second order global splitting error.

Example. Time-split Algorithms: Strang-type Splitting

Consider a linear system of PDEs,

$$U_t + L_1 U + L_2 U + L_3 U = 0,$$

for example,

$$u_t + u_x = u_{xx} + u.$$

In this case, all three operators, ∂_x, ∂_{xx} and multiplication by one, are commuting. On the other hand, the following example originating from a 2D TE mode Maxwell, has two spatial linear operators that do not commute,

$$\frac{\partial}{\partial t}\begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix} = \left[\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\partial_x \\ 0 & \partial_x & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & \partial_y \\ 0 & 0 & 0 \\ -\partial_y & 0 & 0 \end{pmatrix}\right]\begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix}.$$

Now, going back to the general linear system, the solution operator $e^{\Delta t L}$ can be factored exactly $e^{\Delta t L} = e^{\Delta t L_3} e^{\Delta t L_2} e^{\Delta t L_1}$, in the case when operators L_1, L_2 and L_3 commute, or approximately when they do not:

$$e^{\Delta t L} = e^{\Delta t L_3} e^{\Delta t L_2} e^{\Delta t L_1} + O(\Delta t)^2,$$

or to second order accuracy via Strang-type splitting:

$$e^{\Delta t L} = e^{\frac{\Delta t}{2} L_1} e^{\frac{\Delta t}{2} L_2} e^{\frac{\Delta t}{2} L_3} e^{\frac{\Delta t}{2} L_3} e^{\frac{\Delta t}{2} L_2} e^{\frac{\Delta t}{2} L_1} + O(\Delta t)^3,$$

where the middle product can be combined, $e^{\frac{\Delta t}{2} L_3} e^{\frac{\Delta t}{2} L_3} = e^{\Delta t L_3}$. The motivation for Strang-type splitting may be seen from the Campbell-Hausdorff-Baker formula (see Chapter one), that hints of the possibility of eliminating the commutator by switching the order of operators. The proof of the identities comes from a Taylor expansion of the exponents on both sides, and is, in fact, still valid [109] for splitting of arbitrary smooth nonlinear operators. In the same work, LeVeque showed that the local truncation error for Strang-type splitting can be written as:

$$E_{lt}(\Delta t) = E_{split}(\Delta t) + E_3(\Delta t) + 2E_2\left(\frac{\Delta t}{2}\right) + 2E_2\left(\frac{\Delta t}{2}\right),$$

where the splitting error is independent of the local truncation errors for each subproblem, E_1, E_2 and E_3, and is defined as the difference between exact solution operators:

$$E_{split} = S(\Delta t) - S_1\left(\frac{\Delta t}{2}\right) S_2\left(\frac{\Delta t}{2}\right) S_3(\Delta t) S_2\left(\frac{\Delta t}{2}\right) S_1\left(\frac{\Delta t}{2}\right).$$

Linear stability of each subproblem in general is unrelated to the stability of the full problem, as the spectrum of the matrix product is unrelated to the spectrum of each matrix, unless additional constraints are imposed, like all update matrices can be normalized or simultaneously diagonalized by the same similarity transformation. In that case, indeed, stability of each subproblem does imply the overall stability, [109]. We illustrate the splitting technique using several examples.

Example. Counter-propagating Waves in Optics

Consider a system of coupled nonlinear Schrödinger equations, [64]:

$$2ik\left(\frac{1}{v_g}\frac{\partial}{\partial t} + \frac{\partial}{\partial z}\right)\psi_R + \nabla_t r\psi_R = ikg_R\psi_R,$$
$$2ik\left(\frac{1}{v_g}\frac{\partial}{\partial t} - \frac{\partial}{\partial z}\right)\psi_L + \nabla_t r\psi_L = ikg_L\psi_L,$$

where k and v_g are constants, the nonlinear coupling terms are as follows, $g_R = g_L = g_0/(1 + |\psi_R|^2 + |\psi_L|^2)$. In this case, weak diffraction effects are modeled by the 2D transverse Laplacian with respect to (x, y) variables. The domain is assumed periodic in the transverse direction and splitting the diffraction term allows us to apply spectrally accurate Fourier differentiation methods. The rest of the system is 1D with counter-propagating waves moving with constant speed v_g. Using the method of characteristics with a particular time step $\Delta t = \Delta x/v_g$ solves the advection part exactly. For a 1D NLS with cubic nonlinearity $|\psi|^2\psi$, splitting the nonlinear term, $i\psi_t = |\psi|^2\psi$, allows for the exact solution, $\psi(x, y, t) = \psi_0(x, y)e^{-i|\psi|^2 t}$, due to the fact that the amplitude of the solution is time invariant.

Example. Semi-analytic Approach to Highly Oscillatory System of Ordinary Differential Equations

Given a system, [54]:

$$u(t)_t + Au(t) = f(u(t), t),$$

where A is a constant matrix describing a highly oscillatory part of the system. Splitting of the linear part may result in the exact solution in some cases, or be reduced to numerical computation of the matrix exponential,

e^{-At}. The latter task, in most situations, can be done accuratly and reliably, [95,134].

The next two examples illustrate non-physical behavior arising in splitting problems with discontinuous solutions (shock waves) when modeling compressible Navier-Stokes equations for reactive mixtures, [47,110].

Example

Let

$$u_t + uu_x = \mu \left(u - \frac{1}{2} \right) (u - 1),$$

be split as

$$u + uu_x = 0, \quad u_t = \mu \left(u - \frac{1}{2} \right) (u - 1).$$

Also, consider the following system:

$$\begin{aligned} &u_t + uu_x - q_0 z_x = \mu u_{xx}, \\ &z_x = K\phi(u)z, \end{aligned}$$

that is split as

$$\begin{aligned} &u_t + uu_x = 0, \\ &z_x = K\phi(u)z, \\ &u_t - q_0 z_x = \mu u_{xx}. \end{aligned}$$

The function $\phi(u)$ is a Heaviside step function $H(u - u_{ign})$, where u_{ign}, K, q_0 and μ are given constants. In the cited papers, it was shown that when nonlinear advection is approximated by Godunov-type methods; an ODE is marched from the given boundary condition by one-sided spatial differences; and the linear advection-diffusion equation is discretized using an implicit Euler or Crank-Nicolson method, the resulting solutions have incorrect shock speeds. In order to fix the problem, special procedures that restore the correct balance between nonlinear advection, diffusion and nonlinear reaction terms have to be devised, [47,145,146].

5.6 Project Sample

A number of accurate, high resolution finite-volume methods are available for modeling compressible flow in multiple space dimensions based on the exact or approximate solution of the 1D Riemann problem: i.e. the solution of the hyperbolic system of equations with piecewise constant initial data separated by a discontinuity. Consider for simplicity a 2D hyperbolic system of conservation laws:

$$u_t + f(u)_x + g(u)_y = 0. \tag{5.221}$$

On a uniform Cartesian grid with spacing $\Delta x = \Delta y$, a finite volume discretization takes the form:

$$U_{i,j}^{n+1} = U_{i,j}^n - \frac{\Delta t}{\Delta x}(F_{i+1,j}^n - F_{i,j}^n + G_{i,j+1}^n - G_{i,j}^n), \tag{5.222}$$

where $U_{i,j}^n$ represents approximation to the cell averages:

$$U_{i,j}^n \cong \frac{1}{\Delta x^2}\int_{x_i}^{x_{i+1}}\int_{y_j}^{y_{j+1}} u(x,y,t_n), \tag{5.223}$$

and (F,G) is an approximation to the true flux. For example, flux through the left boundary of the cell is:

$$F_{i,j}^n \cong \frac{1}{\Delta x \Delta t}\int_{t_n}^{t_{n+1}}\int_{y_j}^{y_{j+1}} f(u(x_i,y,t))\,dy\,dt. \tag{5.224}$$

Riemann solvers use an approach to calculating flux as $F_{i,j}^n = f(\bar{u})$, where $\bar{u}$ is the solution to the Riemann problem with $U_{i-1,j}$ and $U_{i,j}$ as initial left and right states, evaluated at $t + \Delta t$ at the boundary between cells $i-1$ and i.

In two and three dimensions, solvers that are based on the Riemann problem employ some form of dimensional splitting, where the 1D operator of the equations projected on the normals to the control cell is used. In the multiplicative version of this approach, one applies a 1D operator successively to each coordinate direction:

$$U^{n+1} = \hat{S}_k^y \hat{S}_k^x U^n, \tag{5.225}$$

where U^{n+1} and U^n denote solutions at time $t + \Delta t$ and t, and $\hat{S}_k^i$ is a numerical approximation to the exact 1D solution operator $\hat{S}$ with time step $k = \Delta t$. For systems of linear equations and nonlinear equations, such as Euler's equations, the operators for each coordinate direction usually do not commute, and using a Taylor series expansion, the truncation error in the modified equation can be shown to be $O(\Delta t)$, leading to

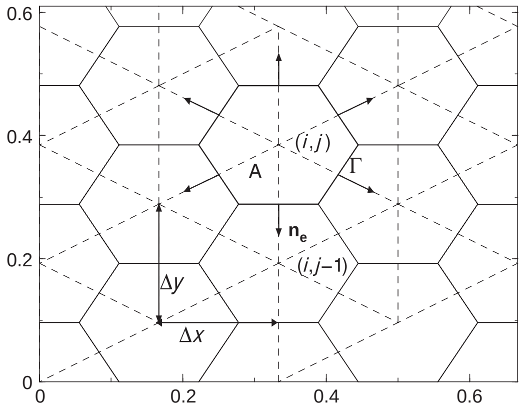


Figure 5.2: Control volume of a dual Delaunay-Voronoi hexagonal mesh.

the first-order accurate scheme. Using a different product of operators suggested by Strang [176]:

$$U^{n+1} = \hat{S}^x_{k/2}\hat{S}^y_k\hat{S}^x_{k/2}U^n, \tag{5.226}$$

gives a splitting which has truncation error $O(\Delta t^2)$. For smooth solutions, if each of the methods $\hat{S}^i_k$ is second-order accurate, then formally so is the split method.

In the case of additive splitting, the computation uses the same 1D operator as in the multiplicative version, but the value in the cell is updated simultaneously, that is, $\hat{S}^y_k$ and $\hat{S}^x_k$ are connected additively. The advantage of using direction splitting is in the relative simplicity of the underlying 1D solution operator and the existence of an analytic solution to the Riemann problem for the case of the Euler equations and some other nonlinear systems. But both of these methods do not account for physically relevant directions of propagation, which may lead to difficulties such as numerical diffusion for shocks traveling at the angle to the grid lines, anisotropic wave propagation and numerical instabilities, which will ultimately restrict the maximum allowed time step or Courant-Friedrichs-Lewy number. The main reason for this is that the multidimensional form of the equations, unlike 1D case, has infinitely many propagation directions.

Generalizing the two-state 1D Riemann problem to a three-state 2D problem, one can apply the difference scheme on a grid in which there are three neighboring cells. Figure 5.2 shows a regular Delaunay-Voronoi dual mesh on which a linear three-state 2D Riemann solver can be used in a cell-centered finite volume method on a hexagonal grid [36].

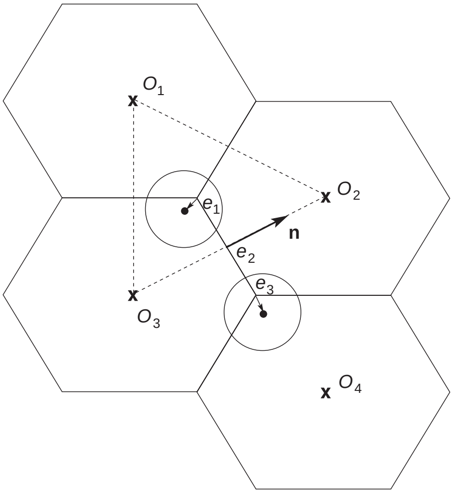


Figure 5.3: Flux computation using three-state Riemann problems.

Consider an integral form of a system of hyperbolic conservation laws

$$\frac{d}{dt}\int_A u\,dS + \int_\Gamma \mathbf{f}\cdot\mathbf{n}\,dl = 0, \tag{5.227}$$

where u is one of the conserved variables, $\mathbf{f}$ is a flux vector and A and Γ represent the volume and the boundary of the control region. Integrating in time and over the computational cell, shown in Figure 5.2, gives the following finite volume approximation,

$$U_{ij}^{n+1} = U_{ij}^{n} - \frac{1}{A}\sum_k \int_0^{\Delta t} dt \int_{\Gamma_k} \mathbf{f}\cdot\mathbf{n}\,dl, \tag{5.228}$$

where U_{ij} represents the cell average and Γ_k is the edge of the computational cell. Denoting approximation to the flux through the k-th edge by Φ_k we have:

$$U_{ij}^{n+1} = U_{ij}^{n} - \frac{\Delta t}{A}\sum_k \Gamma_k\Phi_k. \tag{5.229}$$

The cell averages are assumed to be given, while the fluxes Φ_k are approximated using the values on the edge computed as solutions to the three- and two-state linear Riemann problems.

The initial data needed to determine the values of the flux density $\mathbf{f} = (F, G)$ along one of the edges consists of four states, as shown in Figure 5.3.

The circles in this figure represent the position of a sonic wave front based on the average sound speed of the three surrounding states. The centers of the sonic circles are shifted by the position vector $-\bar{\mathbf{u}}\Delta t$ to account for the advection with average velocity $\bar{\mathbf{u}}$. Thus the numerical flux across the sections of the edge, denoted by e_1, e_3, results from multidimensional waves originating from the corners. These fluxes are approximated using solutions to the three-state linear Riemann problems [36]. For example, in Figure 5.3, the flux across the section e_1 is computed as $\mathbf{f}(\mathbf{u}^*(x, y, t))$, where $\mathbf{u}^*(x, y, t)$ is the solution of the three-state Riemann problem with the initial data from the cells O_1, O_2 and O_3. The flux across the section e_2 is determined only by the states O_2 and O_3 and can be computed using the 1D Riemann solver. The resulting numerical flux can be viewed as a 1D flux across the cell boundaries plus the corrections due to the waves emanating from the corners.

This page intentionally left blank

Chapter 6

Numerical Grid Generation

6.1 Non-uniform Static Grids, Stability and Accuracy Issues

In this chapter we overview the basic ideas behind numerical grid generation through analysis of several representative examples. Grid adaptation to geometrical features of the domain, its interfaces and boundaries, as well as adaptation to the solution features, such as high gradients, is often employed to improve the accuracy and efficiency of the computation. In some cases, due to limited computational resources, it is the only way to resolve the features in the geometry or solution needed to obtain an accurate numerical result. The trade-offs that have to be addressed in many sophisticated grid generation strategies typically involve implementation and stability issues. Several appealing ideas illustrated in examples below are still actively researched after decades of work, due to inability to overcome the numerical stability problems.

We will consider grid generation as a remapping problem that transforms a regular Cartesian region, called logical space, into a desired region in the physical space. The mapped domain may consist of a single computational cell, as in the finite element method applied on unstructured grids; a part of the domain (block mapping) singled out by its geometrical, material or solution properties; or one can remap the entire computational domain. For example, in the last two cases applied in two dimensions, one may use a logical space consisting of any regular polygons that tessellate the plane (e.g. triangles, squares or hexagons), while for the unstructured mesh a single computational cell in logical space could be an arbitrary regular polygon. The complexity, adaptation and flexibility varies with each approach. With respect to the time evolution, the remapping may be done once, to adapt to fixed geometrical features of the domain (called static remapping), or it may be done periodically as the solution evolves, or it may be done continuously (dynamically), adapting the mesh to the solution features.

Example. Linear, Algebraic and Exponential One-dimensional (1D) Maps

Linear map,

$$x(\xi) = (1-\xi)x_0 + \xi x_1,$$

transforms $[0,1]$ into an arbitrary interval $[x_0, x_1]$, for $\xi_j = j/N$, where N is the number of uniform intervals in the logical space $[0,1]$ and x_j is defined as $x_j = x(\xi_j)$. The exponential map is given by the formula,

$$x(\xi) = \frac{e^{\lambda\xi}-1}{e^{\lambda}-1},$$

where λ is a free parameter to be specified. It also can be written as,

$$x_{j+1} = x_j + r(x_j - x_{j-1}),$$

where $r = e^{\lambda\Delta\xi}$ and $\Delta\xi = 1/N$. Remapping of the infinite interval in physical space, $[0,\infty]$, into a finite logical interval can be done using exponential, algebraic or trigonometric functions, for example,

$$\xi = 1 - e^{-\frac{x}{L}}, \quad \xi = \frac{\frac{x}{L}}{\frac{x}{L}+1}, \quad \xi = \frac{2}{\pi}\arctan\left(\frac{x}{L}\right),$$

respectively, with logical variable ξ being discretized uniformly as before. The parameters in the mapping formulae are chosen to achieve desired compression or stretching ratios according to the relation,

$$\Delta x \sim x'(\xi)\,\Delta\xi, \quad \Delta\xi = const.$$

Note, when remapping of the infinite interval into a finite one is done to replace the boundary condition at infinity by the boundary condition at $\xi = 1$, the resulting problem is not necessarily simpler, since the gradient of the transformation now will affect both the stability and accuracy of the solution.

Example. Two-dimensional (2D) Mappings

Similarly to the 1D maps, the 2D uniform logical Cartesian grid, $(\xi,\eta) \in [0,1] \times [0,1]$ is transformed into physical $(x(\xi,\eta), y(\xi,\eta))$ coordinates [101]. For example a bilinear map, that preserves the counter-clockwise orientation of the logical unit square, is as follows,

$$\vec{r} = (1-\xi)(1-\eta)\vec{r}_{00} + (1-\xi)\eta\vec{r}_{01} + \xi(1-\eta)\vec{r}_{10} + \xi\eta\vec{r}_{11},$$

where $\vec{r}$ is defined as $\vec{r} = (x, y)$ and the other four vectors provide vertex coordinates in physical space. Similarly, one may remap a curvilinear quadrilateral region in physical space. For example, a bi-quadratic map that has five control points at the midpoints of the unit square and its center in logical space, in addition to the four vertices, is

$$\begin{aligned}\vec{r} = {} & (1-\xi)(1-\eta)(1-2\xi)(1-2\eta)\vec{r}_{00} \\ & + (1-\xi)\eta(1-2\xi)(-1+2\eta)\vec{r}_{01} \\ & + \xi(1-\eta)(-1+2\xi)(1-2\eta)\vec{r}_{10} + \xi\eta(-1+2\xi)(-1+2\eta)\vec{r}_{11} \\ & + 4\xi(1-\xi)(1-2\eta)(1-\eta)\vec{r}_{\frac{1}{2},0} + 4(1-\xi)(1-2\xi)\eta(1-2\eta)\vec{r}_{0,\frac{1}{2}} \\ & + 4\xi(-1+2\xi)\eta(1-\eta)\vec{r}_{1,\frac{1}{2}} + 4\xi(1-\xi)\eta(-1+2\eta)\vec{r}_{\frac{1}{2},1} \\ & + 16\xi\eta(1-\xi)(1-\eta)\vec{r}_{\frac{1}{2},\frac{1}{2}}.\end{aligned}$$

For three-dimensional (3D) generalizations of triangular and quadrilateral grids, as well as a discussion of mapping degeneracies, we refer the reader to [22] and references therein.

The polar coordinates are given by the familiar formulae,

$$\begin{aligned} x(\xi,\eta) &= r\cos(\theta), \\ y(\xi,\eta) &= r\sin(\theta), \end{aligned}$$

where $r \in [r_0, r_1]$ and $\theta \in [\theta_0, \theta_1]$ are linearly stretched into a unit logical space ξ and η,

$$\begin{aligned} \theta(\xi,\eta) &= \theta_0 + \xi(\theta_1 - \theta_0), \\ r(\xi,\eta) &= r_0 + \eta(r_1 - r_0). \end{aligned}$$

Other 2D parametric equations may be utilized in the same fashion. In particular, parabolic and elliptic grids may be generated by formulae,

$$\begin{aligned} x(\xi,\eta) &= (r^2 - s^2)/2, \quad r = 1+\xi, \\ y(\xi,\eta) &= rs, \quad s = 1+\eta, \end{aligned}$$

and

$$x(\xi,\eta) = a\cosh(r)\cos(s), \quad r = 1+\xi,$$

$$y(\xi,\eta) = a\sinh(r)\sin(s), \quad s = \pi\eta,$$

respectively.

Once the grid is generated, the corresponding partial differential equation (PDE) may be discretized in physical or logical space. The first approach would require one to differentiate/integrate appropriate interpolation functions over non-uniform grid and curvilinear regions, as is the case for finite element or finite volume computations. The second approach is equivalent to changing the variables and it is the way integrals over curvilinear regions are computed anyway. Therefore, in this section we will only concentrate on the change of variables technique.

Example. One-dimensional Heat Equation

Consider

$$u_t = u_{xx},$$

and $x = x(\xi)$, $\tilde{u}(\xi,t) = u(x(\xi),t)$. Therefore, $\tilde{u}_\xi = u_x x_\xi$, or in operator form, $\frac{\partial}{\partial x} = \frac{1}{x_\xi}\frac{\partial}{\partial \xi}$. The transformed heat equation becomes

$$\frac{\partial \tilde{u}}{\partial t} = \frac{1}{x_\xi}\frac{\partial}{\partial \xi}\left(\frac{1}{x_\xi}\frac{\partial \tilde{u}}{\partial \xi}\right).$$

The subsequent discretization is done on a uniform grid in the logical ξ-space, via finite difference, pseudo-spectral, finite element or any other suitable discretization method. Since both stability (for explicit time differentiation) and accuracy depend on a Δx that now involves the gradient of the transformation, $\Delta x = x_\xi \Delta\xi$, the gradient of the mapping will affect both the stability condition and the accuracy of the numerical solution. In particular, the steeper the gradient, the more stringent is the requirement on the size of $\Delta\xi$, with the "infinite" gradient representing the interface boundary between the two grids. It requires design of an appropriate interface boundary condition in order to preserve accuracy (and therefore correct "physics") and stability of the underlying numerical method. Boundaries between refinement patches in adaptive mesh refinement algorithms provide such an example and will be discussed later in this chapter.

Example. Two-dimensional Heat Equation

Consider the heat equation with space-dependent conductivity given by the symmetric positive definite matrix κ,

$$u_t = \nabla\cdot(\kappa\nabla u), \quad \kappa = \begin{pmatrix} \alpha(x,y) & \beta(x,y) \\ \beta(x,y) & \gamma(x,y) \end{pmatrix}.$$

Similar to the 1D case, applying the chain rule gives,

$$\tilde{u}_t = \frac{1}{J}\tilde{\nabla}\cdot\left(\tilde{\kappa}\tilde{\nabla}\tilde{u}\right), \quad \tilde{\kappa} = \begin{pmatrix}\tilde{\alpha}(\xi,\eta) & \tilde{\beta}(\xi,\eta)\\ \tilde{\beta}(\xi,\eta) & \tilde{\gamma}(\xi,\eta)\end{pmatrix},$$

[101], where

$$\begin{aligned}
J &= x_\xi y_\eta - x_\eta y_\xi,\\
\hat{\alpha} &= \frac{1}{J}\left(\tilde{\alpha} y_\eta^2 - 2\tilde{\beta} x_\eta y_\eta + \tilde{\gamma} x_\eta^2\right),\\
\hat{\beta} &= -\frac{1}{J}\left(\tilde{\alpha} y_\xi y_\eta - \tilde{\beta}(x_\xi y_\eta + x_\eta y_\xi) + \tilde{\gamma} x_\xi x_\eta\right),\\
\hat{\gamma} &= \frac{1}{J}\left(\tilde{\alpha} y_\xi^2 - 2\tilde{\beta} x_\xi y_\xi + \tilde{\gamma} x_\xi^2\right),
\end{aligned}$$

and “tilde” is defined as in the previous example, e.g. $\tilde{\alpha}(\xi,\eta) = \alpha(x(\xi,\eta), y(\xi,\eta))$.

Example

Consider 2D Maxwell equations for (TM) mode,

$$\begin{aligned}
\mu\frac{\partial H^x}{\partial t} &= -\frac{\partial E^z}{\partial y},\\
\mu\frac{\partial H^y}{\partial t} &= \frac{\partial E^z}{\partial x},\\
\epsilon\frac{\partial E^z}{\partial t} &= \frac{\partial H^y}{\partial x} - \frac{\partial H^x}{\partial y}.
\end{aligned}$$

Assuming $x(\xi,\eta)$ and $y(\xi,\eta)$, and applying the chain rule gives, [211],

$$\begin{aligned}
\mu J\frac{\partial \tilde{H}^\xi}{\partial t} &= -\frac{\partial \tilde{E}^z}{\partial \eta},\\
\mu J\frac{\partial \tilde{H}^\eta}{\partial t} &= \frac{\partial \tilde{E}^z}{\partial \xi},\\
\epsilon J\frac{\partial \tilde{E}^z}{\partial t} &= \alpha_1\frac{\partial \tilde{H}^\eta}{\partial \xi} - \alpha_2\frac{\partial \tilde{H}^\xi}{\partial \eta} - \alpha_3\left(\frac{\partial \tilde{H}^\eta}{\partial \eta} - \frac{\partial \tilde{H}^\xi}{\partial \xi}\right) + \alpha_4\tilde{H}^\xi + \alpha_5\tilde{H}^\eta,
\end{aligned}$$

where

$$J = x_\xi y_\eta - x_\eta y_\xi,$$

$$
\begin{aligned}
\tilde{H}^{\xi} &= \frac{1}{J}\left(\tilde{H}^{x} y_{\eta} - \tilde{H}^{y} x_{\eta}\right),\\
\tilde{H}^{\eta} &= \frac{1}{J}\left(\tilde{H}^{y} x_{\xi} - \tilde{H}^{x} y_{\xi}\right),\\
\alpha_1 &= x_{\eta}^2 + y_{\eta}^2,\\
\alpha_2 &= x_{\xi}^2 + y_{\xi}^2,\\
\alpha_3 &= x_{\xi} x_{\eta} + y_{\xi} y_{\eta},\\
\alpha_4 &= x_{\xi\xi} x_{\eta} - x_{\xi\eta} x_{\xi} + y_{\xi\xi} y_{\eta} - y_{\xi\eta} y_{\xi},\\
\alpha_5 &= x_{\eta\xi} x_{\eta} - x_{\eta\eta} x_{\xi} + y_{\eta\xi} y_{\eta} - y_{\eta\eta} y_{\xi}.
\end{aligned}
$$

For orthogonal curvilinear grids ($\gamma = 0$) and with the introduction of modified material properties, $\tilde{\mu} = J\mu$ and $\tilde{\epsilon} = J\epsilon$, the above equations require simple modification of the existing Cartesian Maxwell codes. In addition, orthogonal grids eliminate the problem of small angles and degenerate transformations, thus improving the accuracy and stability of the method.

6.2 Adaptive and Moving Grids Based on Equidistribution Principle

The equidistribution principle in its simplest form is described by equation,

$$M\Delta x = \Delta\xi,$$

where $M \geq 0$ is a solution and/or geometry-dependent monitor function that is proportional to the desired Δx, because $\Delta\xi = \mathit{const}$ large M will produce small Δx and vice versa. Taking the ξ-derivative of $M\Delta x/\Delta\xi = \mathit{const}$, motivates the following elliptic grid generation equation

$$\frac{\partial}{\partial\xi}\left(M\frac{\partial x(\xi)}{\partial\xi}\right) = 0,$$

and similarly in the 2D case,

$$
\begin{aligned}
&\nabla\cdot\left(M\nabla x(\xi,\eta)\right) = 0,\\
&\nabla\cdot\left(M\nabla y(\xi,\eta)\right) = 0,
\end{aligned}
$$

where the monitor function M may depend on the amplitude, gradient and curvature of the solution, arclength and orthogonality of the coordinate transformation. The derivatives involved in the monitor function may be taken with respect to the physical or logical space. Here, we follow the recommendation in [41] to utilize the logical space derivatives in the monitor function. When the monitor function depends on the solution u and its derivatives, but not on variables x and y or their derivatives, the above elliptic equations may be viewed as Euler-Lagrange equations of the corresponding variational problem [86,101],

$$J[x,y] = \frac{1}{2} \iint \big((\nabla x)^t M \nabla x + (\nabla y)^t M \nabla y \big)\, d\xi d\eta.$$

In applications, a typical monitor function is as follows,

$$M = \sqrt{1 + \alpha_0 |f|^2 + \alpha_1 |\nabla f|^2 + \alpha |\Delta f|^2},$$

where f represents a solution of the problem or a function that characterizes material/geometrical features of the domain, or a combination of both. We reiterate, the ∇ operator is taken with respect to the logical (ξ, η) space. The α_i weights are application dependent and have to be determined experimentally [41]. For use of other variational grid generation methods the reader may consult the text [101] and a recent article [40] containing a literature overview of the variational approaches.

Example. Moving Adaptive Meshes

To obtain a moving mesh equation we assume that x is also time dependent, $x(\xi, t)$ [40,41]. This approach is identical to Lagrangian or particle description in continuum dynamics. Introducing a new variable, $\tilde{u}(\xi, t) = u(x(\xi, t), t)$, and applying the chain rule to compute the total (also called the Lagrangian or material) derivative,

$$\tilde{u}_t = u_t + u_x x_t, \quad \tilde{u}_\xi = u_x x_\xi.$$

These formulae are then used to modify the original equation to the new (ξ, t) coordinate system. For example, the heat equation

$$u_t = u_{xx}$$

becomes

$$\frac{\partial \tilde{u}}{\partial t} - \frac{x_t}{x_\xi} \frac{\partial \tilde{u}}{\partial \xi} = \frac{1}{x_\xi} \frac{\partial}{\partial \xi} \left(\frac{1}{x_\xi} \frac{\partial \tilde{u}}{\partial \xi} \right).$$

To close the system, a dynamic equation for moving grids $x(\xi, t)$ is needed. Here we chose the parabolic equation that was suggested after experiments with six plausible hyperbolic/parabolic equations for time dynamics of the moving grids [86]. It generalizes the adaptive static grid generation equation, considering it as an equilibrium solution to the parabolic equation that is achieved as time tends to infinity. In the one-dimensional case,

$$x_t = (Mx_\xi)_\xi,$$

and similarly for the 2D case,

$$x_t = \nabla \cdot (M\nabla x), \quad y_t = \nabla \cdot (M\nabla y).$$

This nonlinear parabolic equation may have numerous equilibrium solutions, convergence to which will depend on the choice of the initial data. The discretization of the above combined system often runs into stability complications and various ad hoc remedies such as artificial viscosities and temporal and spatial filtering steps are introduced, [41,58,86,113].

6.3 Level Set Methods

In 1988, S. J. Osher and J. A. Sethian developed a numerical algorithm for following fronts/surfaces with a curvature-dependent speed, known as PSC methods (for Propagation of Surfaces under Curvature) [142,143,165]. In two Cartesian space dimensions, these methods are related to the so-called curve shortening algorithms [69]. In this section we describe the relation of the geometric heat equation and the affine heat equation to front propagation and level set methods.

A curve $C(\mathbf{x}(t,\tau),t)$, in the (x,y)-plane can be thought of as a level set of a function $u(x,y,t)$. For example,

$$C = \{x, y : u(x(t), y(t), t) = 0\}, \tag{6.1}$$

defines the curve $C(t)$. Differentiation of (6.1) with respect to t yields the equation:

$$u_t + \mathbf{x}_t \cdot \nabla u = 0. \tag{6.2}$$

If $C(t,\tau)$ is a moving closed curve, with parameter τ along the curve, the curve can be chosen to evolve so that it expands or contracts at a speed $F(\kappa)$ normal to itself, where F is a function of the curvature, κ, of the curve, i.e. the curve $\mathbf{x}(t,\tau)$ evolves as:

$$\mathbf{x}_t = F(\kappa)\mathbf{n}. \tag{6.3}$$

From (6.2)–(6.3), with $F(\kappa) = \kappa$, we obtain:

$$u_t + \kappa \mathbf{n} \cdot \nabla u = 0. \tag{6.4}$$

If we choose $\mathbf{n} = -\nabla u/|\nabla u|$ as the inward normal to the curve, then u evolves according to the equation:

$$u_t = \kappa|\nabla u|. \tag{6.5}$$

In the above development, the curve $C(t, \tau)$ evolves according to the equation:

$$u_x x_\tau + u_y y_\tau = 0 \quad \text{or} \quad \hat{\mathbf{t}} \cdot \nabla u = 0, \tag{6.6}$$

where

$$\hat{\mathbf{t}} = \frac{(x_\tau, y_\tau)}{\sqrt{x_\tau^2 + y_\tau^2}}, \quad \mathbf{n} = \frac{(y_\tau, -x_\tau)}{\sqrt{x_\tau^2 + y_\tau^2}}, \tag{6.7}$$

are the tangent vector and normal to the curve. The curvature κ of the curve $C(t, \tau)$, the tangent vector $\hat{\mathbf{t}}$ and bi-normal $\mathbf{n}$ to the curve, are related by the Serret-Frenet formula:

$$\frac{d\hat{\mathbf{t}}}{ds} = \kappa \mathbf{n}, \tag{6.8}$$

where $s \equiv \tau$ is the arc-length along the curve. From (6.7)–(6.8) it follows that

$$\kappa = \frac{y_\tau x_{\tau\tau} - x_\tau y_{\tau\tau}}{(x_\tau^2 + y_\tau^2)^{3/2}}. \tag{6.9}$$

From (6.9) and (6.6) we obtain:

$$\kappa = \frac{\operatorname{sgn}(u_y)\left(u_y^2 u_{xx} - 2u_x u_y u_{xy} + u_x^2 u_{yy}\right)}{(u_x^2 + u_y^2)^{3/2}}, \tag{6.10}$$

as an alternative formula for the curvature of the curve $C(t, \tau)$. Choosing $\operatorname{sgn}(u_y) = 1$ in (6.10) we obtain:

$$\kappa = \frac{u_y^2 u_{xx} - 2u_x u_y u_{xy} + u_x^2 u_{yy}}{(u_x^2 + u_y^2)^{3/2}} \equiv \nabla \cdot \left(\frac{\nabla u}{|\nabla u|}\right). \tag{6.11}$$

Using (6.11) in (6.5) with $F(\kappa) = \kappa$ we obtain the geometric heat equation:

$$u_t = \frac{u_y^2 u_{xx} - 2u_x u_y u_{xy} + u_x^2 u_{yy}}{u_x^2 + u_y^2}. \tag{6.12}$$

Thus, the boundary curve $[x(t,\tau), y(t,\tau)]$ is such that $u[x(t,\tau), y(t,\tau), t] = 0$, where $u(x,y,t)$ satisfies the geometric heat equation (6.12). Hence the level set $u(x,y,t) = 0$ describes the evolution of the front.

Other related PDEs can be obtained for other choices of the function $F(\kappa)$. The choice $F(\kappa) = \kappa^{1/3}$ in (6.5) results in the affine heat equation:

$$u_t = \left(u_y^2 u_x x - 2u_x u_y u_{xy} + u_x^2 u_{yy}\right)^{1/3}. \tag{6.13}$$

This completes our discussion of the relationship between level set methods, front evolution with a curvature-dependent speed and equations related to the geoemetric heat equation. The implementation of these methods, with examples and applications, may be found in the book by Sethian [165].

The level set technique can be used for mesh generation in two and three dimensions. Given a curve in two space dimensions or a surface in three space dimensions, the level set method allows one to generate a logically rectangular, body-fitted grid. For logically rectangular grids, each grid node has four neighbors in 2D and six neighbors in 3D. The body-fitted grid has one set of the grid lines that can match the body itself. Such grids are important in many areas, for example, when applying the FDTD technique on nonorthogonal grids, the logically rectangular grid makes the implementation easier and more efficient, and the staircasing error is eliminated by the body-fitted feature of the grid.

First, a given body is treated as a propagating front. By following the trajectories of the particles on the front, level set lines are obtained. For example, in two dimensions, assume a closed curve $C(\mathbf{x}(t,\tau), t)$ is moving normal to itself. The exterior and interior meshes can be obtained by moving the front outwards and inwards, respectively. For the exterior mesh generation, we can set the speed function $F(\kappa) = 1 - \epsilon\kappa$, where ϵ is used to control the decay of the curvature. For interior mesh generation, the speed function $F(\kappa) = -\kappa$ can be used for a convex body. In the case that the body is non-convex initially, a threshold value is introduced to ensure the front always moves inwards. In such cases, the speed function becomes

$$F(\kappa) = \min(-\kappa, F_{threshold}).$$

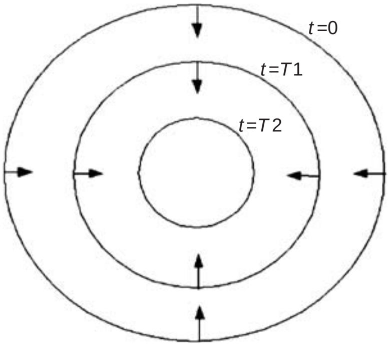


Figure 6.1: Level curve motion.

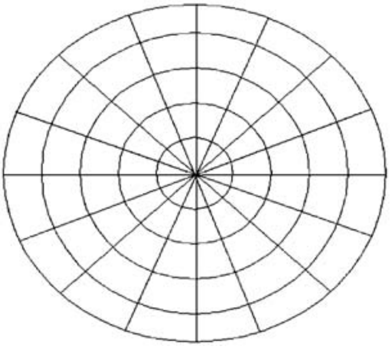


Figure 6.2: Two-dimensional body-fitted mesh in a circle.

The time evolution of the curve provides a set of level curves, as shown in Figure 6.1. By constructing the transverse lines (connecting corresponding nodes on these level curves), we obtain a body-fitted mesh, as shown in Figure 6.2. The transverse lines can be constructed in different ways. The basic idea is to follow the trajectories of the moving front. In the case that big curvature variation happens, some transversal adjustment may be needed. Several examples are shown in Figure 6.3. Besides the examples presented above, a variety of grids can be generated using the level set technique, such as the non-uniform adaptive and moving grids [165], and unstructured tetrahedral meshes [30].

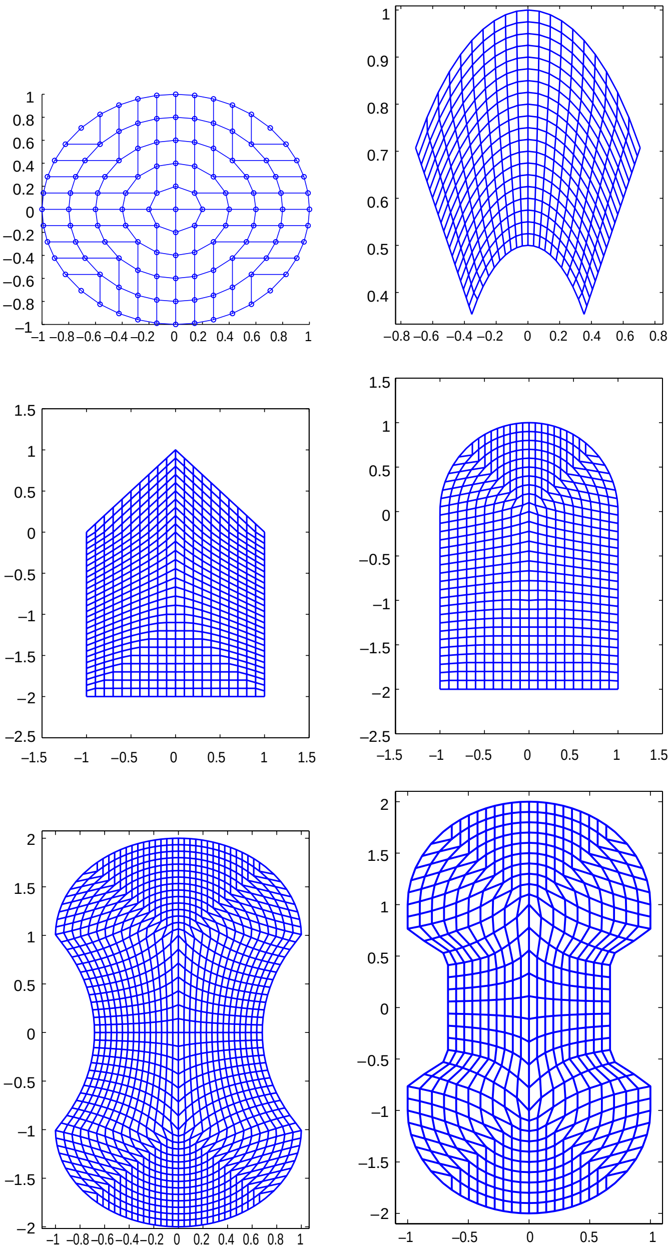


Figure 6.3: Two-dimensional meshes.

6.4 The Front Tracking Method

The front tracking method is a Lagrangian method for the propagation of a moving front. It was introduced by [151] and has been successfully applied in high accuracy aerodynamic computations. A thorough study of the front tracking method has been carried out by Glimm and his collaborators (see [73] for the references). In this section we describe the relation of the hyperbolic conservation laws, the front tracking method and unstructured higher dimensional mesh generation technique.

The front tracking method tracks the front by moving the particles on the interface. An interface is a collection of geometric objects, such as points, curves (piecewise linear segments) and surfaces (triangulated mesh), that correspond to zero-, one-, and two-dimensional meshes, respectively. The interfaces are represented explicitly as lower dimensional meshes. The particles on the interface are advanced via an ordinary differential equation

$$\frac{dx}{dt} = \mathbf{v}. \tag{6.14}$$

If a velocity field is given, $\mathbf{v}$ is a function of space and time only, so $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$. When the front tracking method is used for solving nonlinear hyperbolic conservation laws, the velocity function is determined by computing the solution to the local Riemann problem with initial states being those on either side of the interface point and using the method of characteristics.

There are two types of interface representations: grid free (Lagrangian) interface representation and grid based (Eulerian) interface representation. A grid free interface is independent of the underlying grid. The interface element size is specified by the user. For grid based interface, all the points lie on the cell edges. Thus, the elements are constructed from vertices which are the intersections between the interface and the grid lines. The grid based method is more robust, while the grid free method is more accurate. Two-dimensional bifurcations of interface topology in front tracking are resolved accurately through detection of interface intersections. In three dimensions, a local Eulerian reconstruction method is used. This method has the robustness of the Eulerian method while it maintains the high resolution and accuracy of the Lagrangian method. Figure 6.4 shows examples of three types of interface representations in 2D. For 3D grid based two-fluid interface reconstruction, there are in total 2^8 cases in each cell and they can be reduced to 14 cases, as shown in Figure 6.5. This technique is the marching cubes method [121].

For nonlinear hyperbolic conservation laws, the front tracking method tracks a discontinuity interface using analytical solutions of Riemann

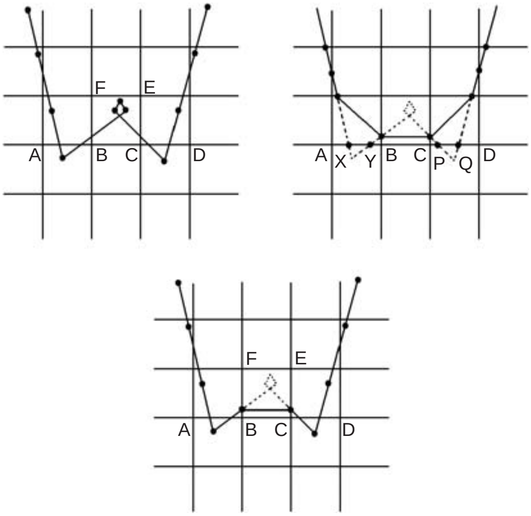


Figure 6.4: Example of three types of interface representations in 2D. Upper-left: grid free interface; Upper-right: grid based interface; Lower: locally Eulerian interface.

problems across the interface, and applies a finite difference scheme to solve equations on different sides of the discontinuity interface using the ghost cell extrapolation method. This ghost cell method was also used in tracking using the level set method. The propagating front represents the discontinuity interface as a zero level set curve or surface. The points on different sides of the interface are labeled as positive and negative, respectively. The ghost cell method does not preserve conservation at the cells cut by the interface and it has less accuracy at discontinuities. Conservative front tracking increases the order of convergence for the truncation error at a discontinuity and therefore converges faster than ordinary tracking or than untracked algorithms for the physical solution. A fully conservative front tracking algorithm for an N-dimensional system of nonlinear conservation laws has been developed [120].

The conservative front tracking method is based on the divergence form of the conservation law in a space-time domain. Consider a system of conservation laws in N spatial dimensions in differential form

$$\frac{\partial U}{\partial t} + \nabla \cdot F(U) = 0, \quad F = (f_1, f_2, \ldots, f_N), \tag{6.15}$$

where $U \in R^p$ and $f_j(U) = (f_{1j}(U), \ldots, f_{pj}(U))^T \in R^p$ are defined in a spatial domain $\Omega \subset R^N$.

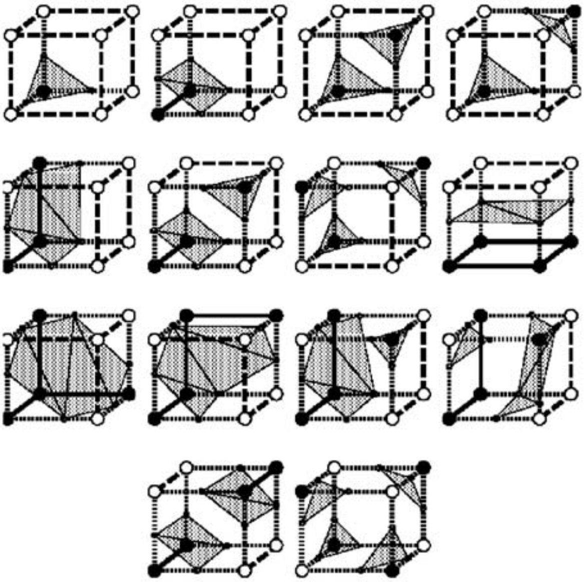


Figure 6.5: Grid based 3D interface. For a two-material interface within each grid cell, there are $2^8 = 256$ possible configurations for the crossings of the cell edge by the interface. They can be reduced to 14 isomorphically distinct configurations of block interface.

Integrating in a time-space domain $\mathcal{V} \subset R^{N+1}$, we obtain the integral form of (6.15),

$$\int_{\mathcal{V}} \left(\frac{\partial U}{\partial t} + \nabla \cdot F(U) \right) d\mathcal{V} = 0. \tag{6.16}$$

By the divergence theorem, we have

$$\int_{\partial \mathcal{V}} (U, F(U)) \cdot n dS = 0. \tag{6.17}$$

The finite difference method presented here is an explicit finite volume integration scheme based on the integral form (6.17).

Assume a space-time discretization $\{\mathcal{V}_i\}$ which conforms to the space-time interface as U evolves in one time step from t_n to t_{n+1}. We solve (6.17) explicitly in this region. We define each $\mathcal{V}_i$ as a space-time control volume, and $\partial \mathcal{V}_i = D_i^n \cup D_i^{n+1} \cup \hat{S}_i$ with D_i^n, D_i^{n+1}, and $\hat{S}_i$ meeting at most at

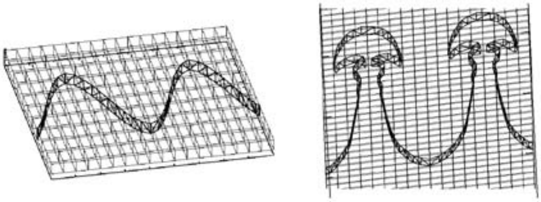


Figure 6.6: Examples of two 3D space-time interfaces at two sequential time steps.

their boundaries, where D_i^n and D_i^{n+1} are the boundary surfaces of $\mathcal{V}_i$ at time level n and $n+1$, respectively, and $\hat{S}_i$ is the complementary boundary surface of $\mathcal{V}_i$. Dividing the calculation of the integral (6.17) into three parts over D_i^n, D_i^{n+1} and $\hat{S}_i$, respectively, we have

$$|D_i^{n+1}|\overline{U}\,|_{t_{n+1}} = |D_i^n|\overline{U}\,|_{t_n} - \int_{\hat{S}_i} (U, F(U)) \cdot ndS. \tag{6.18}$$

$\overline{U}\,|_{t_m}$ is cell average at time level m $(m = n, n+1)$, and it is defined as

$$\overline{U}\,|_{t_m} = \frac{1}{|D_i^m|} \int_{|D_i^m|} U(x_1, \ldots, x_N, t_m) dx_1 ... dx_N,$$

where $|D_i^m|$ is the face area of D_i^m, and n is the outward normal of $\hat{S}_i$.

A key step in solving equation (6.18) is the construction of the space-time control volume $\{\mathcal{V}_i\}$. In three dimensions, $\mathcal{V}_i$ is the volume of a four-dimensional space-time cell. This requires construction of a four-dimensional space-time mesh. There are three major steps in this space-time mesh generation technique:

(1) Propagating the spatial interface at time level n to a new time level $n+1$;
(2) Constructing space-time interface by joining two spatial interfaces at time levels n and $n+1$;
(3) Reconstructing space-time cells by a convex hull algorithm.

In step 1, the front tracking method is applied to advance the spatial interface. A sample space-time interface generated after step 2 is shown in Figure 6.6. The convex hull algorithm in step 3 is based on an isosurface construction technique developed by Bhaniramka et al. [19]. Figure 6.7 shows the procedure to construct a 2D space-time interface. The cell corners labeled as positive and negative are in black and white, respectively.

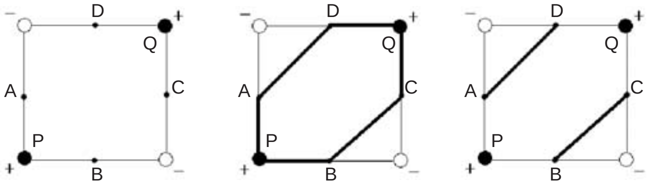


Figure 6.7: Space-time interface construction in two spatial dimensions.

A, B, C, D are cell edge crossings. Let $P^+(c)$ be the union of the corners with positive labels together with four crossings A, B, C, D. First, a convex hull $PBCQDA$ is constructed using the incremental construction method on $P^+(c)$. After deleting the four line segments AP, PB, CQ and QD that lie on the cell boundary, we obtain the space-time interface in this space-time cell: $BC \cup DA$. The space-time interface cuts the space-time cell into three fragments, so that we can obtain the space time control volume by computing the volumes of these fragments. The proof of the correctness for the convex hull reconstruction algorithm is given in [19].

There are two major interface tracking methods: the Lagrangian approach and the Eulerian approach. The front tracking method is a Lagrangian method and the level set method is an Eulerian method. Each method has its advantages and disadvantages. The level set method is efficient and easy to implement. The resolution of the level set method depends on the mesh size and the order of the numerical solver. For a highly curved interface, the error of the level set method can be amplified. As a result, the interface could be smeared and it may lead to inaccurate solutions. The Lagrangian front tracking method tracks the interface explicitly and resolves bifurcations through interface reconstruction. For multiphase flow problems, the front tracking method provides accurate and robust solutions. A triangulated mesh that represents the moving interface should be stored throughout the computation and reconstructed at every time step. The front evolution requires solving a Riemann problem locally on the interface. Both the interface reconstruction and the local Riemann solver are time-consuming. Figure 6.8 shows a comparison benchmark test for a rotating slotted disk using two tracking methods. The level set method shows edge smoothing after several rounds of rotation and eventually results in a topologically incorrect bifurcation. In the front tracking method, the disk maintains its shape after as many as 13 rounds of rotation. Figure 6.9 shows the 3D surface evolution of a sphere under a prescribed shear velocity field. More examples and applications of the front tracking method can be found in [60,73].

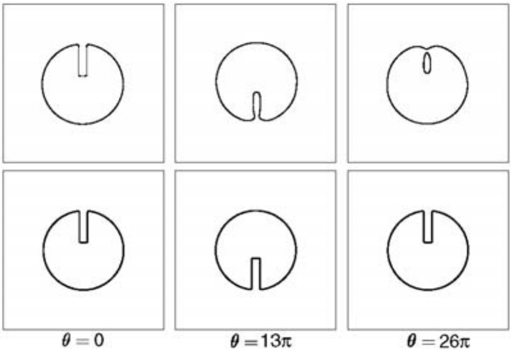


Figure 6.8: Comparison of slotted disk simulation. The upper sequence shows the result of the level set method using the fifth order WENO scheme and the lower sequence shows the result of front tracking using the fourth order Runge-Kutta method.
Source: [60]

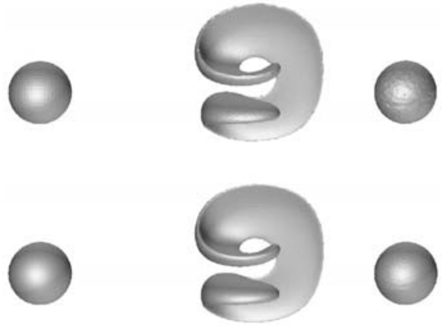


Figure 6.9: 3D surface evolution of a sphere under a prescribed shear velocity field. The upper sequence and the lower sequence have the mesh of 64^3 and 128^3, respectively. From left to right are the interfaces at three sequential time steps.
Source: [60]

6.5 Project Sample

The type of a coordinate system and the geometry of the associated discretization grid for the finite-difference and finite-volume methods are often determined by the physics of the problem and the properties of the numerical method. The weak shock diffraction problem in fluid dynamics

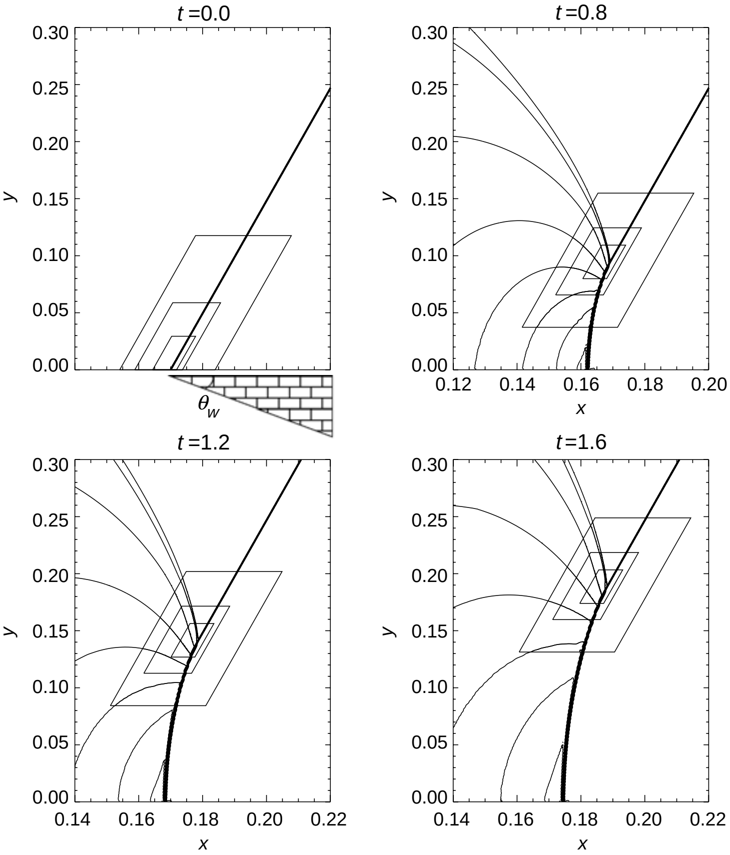


Figure 6.10: Fluid flow density contours on grid levels l = 0-3. The initial condition corresponds to a shock wave incident on a wedge of angle θ_w, shown at $t = 0.0$. The vertical Mach stem, the weak reflected wave and the incident shock wave meet at the triple point that propagates along the incident shock.

is one example, in which the choice of the grid in conjunction with the properties of the finite-volume scheme used to discretize the equations, plays an important role in obtaining accurate solutions [202].

When a weak vertical shock impinges on a wedge of small angle, a pattern resembling Mach reflection may arise [114]. Mach reflection contains three shocks coming together - the incident, reflected wave and a Mach stem, as depicted in Figure 6.10. There are two main parameters in the shock reflection problem, the Mach number M of the incident shock, and wedge angle θ_w. The transition from regular to Mach reflection for weak shocks occurs for thin wedges when the shock strength and the wedge angle are related by

$$M = 1 + O({\theta_w}^2), \qquad \text{as } \theta_w \to 0.$$

To solve the shock reflection problem numerically, a second order, accurate in space and time finite volume formulation, can be used to integrate the equations of ideal fluid dynamics. Evaluation of the fluxes at the interfaces between finite-volume cells can be accomplished using the solution to the 1D linearized Riemann problem ("Roe scheme"), in conjunction with the operator splitting for 2D computations. To remove numerical dispersion error at the shocks, a monotonized central-difference flux limiter can be applied to wave amplitudes throughout the domain.

It is convenient to change the coordinate system to the reference frame moving with the incident shock in the x direction, Figure 6.10. The choice of the reference frame, in addition to minimizing the required computational domain size, allows representation of the incident shock as a single point transition, since the "Roe scheme" admits stationary grid aligned shocks as its exact solution. This exact representation of the incident shock on the grid eliminates small initial waves as well as the limiter produced oscillations that propagate downstream and may contaminate the region of interest.

Alignment of the incident shock with the grid lines on a rectangular domain would require implementation of more complicated boundary conditions to simulate the surface of the wedge, since the wedge then would not be grid aligned. To address this problem the grid is slanted in the direction of the x axis by the angle θ_w. The computational cell is a parallelogram and Riemann problems are solved in the directions normal to the cell edges.

In order to resolve the small region around the triple point, the block adaptive grid refinement technique is used. It would be computationally expensive to refine regions near all three shocks throughout the domain. On the other hand, refining only the region around the triple point causes shocks to cross the grid interfaces, which act as boundaries between domains with different numerical viscosity. Special care is required to eliminate diffusion waves and oscillations at discontinuities emanating from these boundaries.

One can use the fact that the incident shock is grid aligned to represent it as an exact solution at all grid levels to avoid any disturbances at grid boundaries. The reflected shock is represented by a traveling wave that becomes a linear diffraction wave away from the triple point. Therefore it

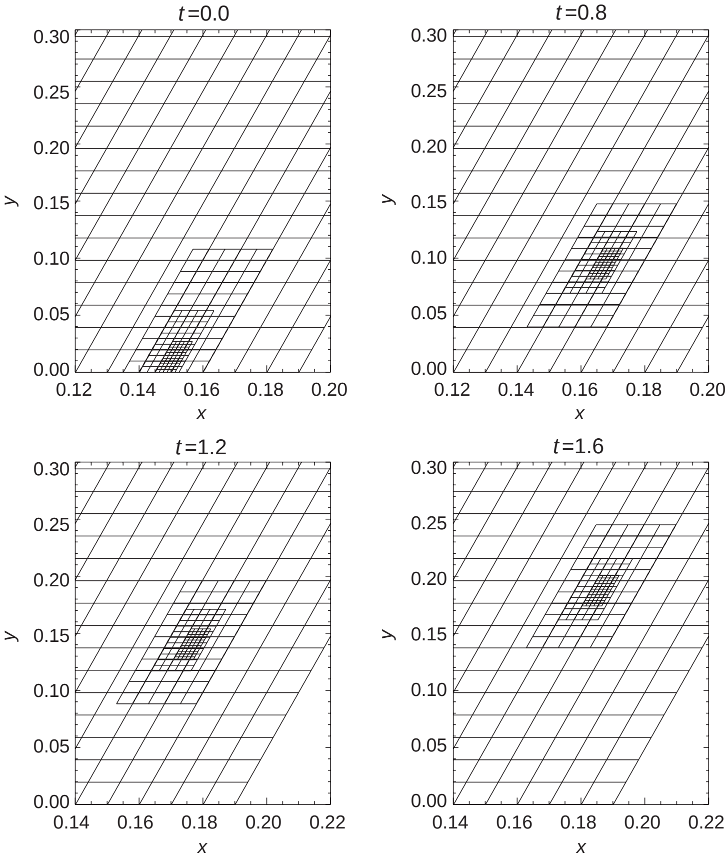


Figure 6.11: Evolution of the adaptive grid structure with time. Grid levels $l = 0\text{-}3$.

is sensitive to the numerical diffusion error that propagates from coarse to fine levels. The distance between boundaries of successively refined regions was taken large enough to reduce this source of errors to an acceptable level.

The Mach stem, on the other hand, becomes stronger as one moves away from the triple point towards the wedge. To remove wave reflections at the grid interfaces, a limiter at the lower grid boundary is used to reduce the second order space interpolation from coarse to fine grids to first order.

In addition, since the stem is a slowly moving shock, numerical diffusion is drastically decreased and small oscillations are generated by the limiter in the region where the stem is at a significant angle to the grid lines.

However, this does not present a problem, since it happens far from the triple point.

Initially, at $t = 0$, all grids have their lower boundary located at the tip of the wedge, which is located at $x \approx 0.15$, Figure 6.11. When the triple point forms and propagates up along the incident shock, the refined grids move independently to keep the triple point centered. Starting and keeping the triple point embedded into the finest refinement region at all times is crucial, since by specifying the initial condition on a coarse mesh and refining the grid later, one would just get a highly resolved coarse solution.

Bibliography

[1] S. Abarbanel, D. Gottlieb and J.S. Hesthaven, "Long time behavior of the perfectly matched layer equations in computational electromagnetics", *J. Sci. Comput.* **17** 405–422 (2002)

[2] M.J. Ablowitz, D.J. Kaup, A.C. Newell and H. Segur, "The inverse scattering transform-fourier analysis for nonlinear problems", *Stud. Appl. Math.* **53** 249–315 (1974)

[3] M.J. Ablowitz and H. Segur, Solitons and the Inverse Scattering Transform, SIAM: Philadelphia, PA (1981)

[4] R. Abraham and J.E. Marsden, Foundations of Mechanics, Addison-Wesley, 2nd ed. (1978)

[5] R. Abraham, J.E. Marsden and T. Ratiu, Manifolds, Tensor Analysis and Applications, Springer, 2nd ed. (1988)

[6] M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions, Dover: New York (1965)

[7] R.C. Aitken (Ed.), Stiff Computation, Oxford University Press (1985)

[8] G. Ali and J.K. Hunter, "Wave interactions in magnetohydrodynamics", *Wave Motion* **27** 257–277 (1998)

[9] A.M. Anile, J.K. Hunter, P. Pantano and G. Russo, Ray Methods for Nonlinear Waves in Fluids and Plasmas, Longmans: New York (1993)

[10] J.A. Armstrong, N. Bloembergen, J. Ducuing and P.S. Pershan, "Interaction between light waves in a nonlinear dielectric medium", *Phys. Rev. Lett.* **127** 1918 (1962)

[11] V.I. Arnold, Ordinary Differential Equations, The MIT Press (1978)

[12] V. Arnold and B. Khesin, Topological Methods in Hydrodynamics, Springer (1998)

[13] N.J. Balmforth and P.J. Morrison, "A necessary and sufficient instability condition for inviscid shear flow", *Stud. Appl. Math.* **102** 309–344 (1999)

[14] N.J. Balmforth and P.J. Morrison, Hamiltonian Description of Shear Flow, in Large Scale Atmosphere-Ocean Dynamics 2: Geometric Methods and Models (Eds. J. Norbury and I. Roulstone), Cambridge University Press: Cambridge, pp. 117–142 (2002)

[15] A.A. Barmin, A.G. Kulikovskiy and N.V. Pogorelov, "Shock-capturing approach and nonevolutionary solutions in magnetohydrodynamics", *J. Comp. Phys.* **126** 77–90 (1996)

[16] J.P. Berenger, "A perfectly matched layer for the absorption of electromagnetic waves", *J. Comp. Phys.* **114** 185–200 (1994)

[17] M.J. Berger and J. Oliger, "Adaptive mesh refinement for hyperbolic partial differential equations", *J. Comp. Phys.* **53** 484–512 (1984)

[18] A. Bers, Space-time Evolution of Plasma Instabilities: Absolute and Convective, in Handbook of Plasma Physics, Vol. 1 (Eds. M.N. Rosenbluth and R.Z. Sagdeev), Northholland: Amsterdam, p. 451 (1983)

[19] P. Bhaniramka, R. Wenge and R. Crawfis, "Isosurface construction in any dimension using convex hulls", *IEEE Tran. Visualiz. Comp. Graphics* **10** 130–141 (2004)

[20] G. Boillat, "Multi-dimensional simple waves in N-dimensional propagation", *J. Math. Phys.* **11** 1482–1483 (1970)

[21] R.N. Bracewell, The Fourier Transform and its Applications, McGraw Hill: New York, 2nd ed. (1978)

[22] L. Branets and G.F. Carey, "A local cell quality metric and variational grid smoothing algorithm", *Engineering with Computers* **21** 19–28 (2005)

[23] L. Brevdo, "A study of absolute and convective instabilities with an application to the Eady model", *Geophys. Astrophys. Fluid Dynamics* **40** 1 (1988)

[24] T.J. Bridges, "Spatial Hamiltonian structure, energy flux and the water wave problem", *Proc. Roy. Soc. Lond. A* **439** 297–315 (1992)

[25] T.J. Bridges, "Multi-symplectic structures and wave propagation", *Math. Proc. Camb. Philos. Soc.* **121** 147–190 (1997)

[26] T.J. Bridges, "A geometric formulation of the conservation of wave action and its implications for signature and classification of instabilities", *Proc. Roy. Soc. A* **453** 1365–1395 (1997)

[27] T.J. Bridges and S. Reich, "Multi-symplectic integrators: Numerical schemes for Hamiltonian PDEs that conserve symplecticity", *Phys. Lett. A* **284** 184–193 (2001)

[28] T.J. Bridges, P.E. Hydon and S. Reich, "Vorticity and symplecticity in Lagrangian fluid dynamics", *J. Phys. A: Math. Gen.* **38** 1403–1418 (2005)

[29] T.J. Bridges and S. Reich, "Numerical methods for Hamiltonian PDEs", *J. Phys. A, Math. Gen.* **39** 5287–5320 (2006)

[30] R. Bridson, J. Teran, N. Molino and R. Fedkiw, "Adaptive physics based tetrahedral mesh generation using level sets", *Eng. Comput.* **21** (1) 2–18 (2005)

[31] R.J. Briggs, Electron Stream Interaction with Plasmas, MIT Press: Cambridge, MA (1964)

[32] E.O. Brigham, The Fast Fourier Transform and its Applications, Prentice-Hall: New Jersey (1988)

[33] M. Brio and C.C. Wu, "An upwind differencing scheme for the equations of ideal magnetohydrodynamics", *J. Comp. Phys.* **75** 400–422 (1988)

[34] M. Brio, M. Temple-Raston, "Regularizations of the inviscid Burgers equation, in viscous profiles and numerical methods for shock waves" (Ed. M. Shearer), *SIAM Proc. Appl. Math.* **12** (1991)

[35] M. Brio and P. Rosenau, "Stability of shock waves for 3 × 3 system of model MHD equations", *Notes on Num. Fluid Mech.* **43** 77 (1993)

[36] M. Brio, A.R. Zakharian and G.M. Webb, "Two-dimensional Riemann solver for Euler equations of gas dynamics", *J. Comp. Phys.* **167** 177–195 (2001)

[37] A.J. Brizard and T.S. Hahm, "Foundations of nonlinear gyro-kinetic theory", *Rev. Mod. Phys.* **79** 421–467 (2007)

[38] R.A. Cairns, "The role of negative energy waves in some instabilities of parallel flows", *J. Fluid Mech.* **92** 1–14 (1979) (Part 1)

[39] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang, Spectral Methods in Fluid Dynamics, Springer-Verlag: New York (1988)

[40] W. Cao, W. Huang and R.D. Russell, "Approaches for generating moving adaptive meshes: Location versus velocity", *Appl. Numer. Math.* **47** 121–138 (2003)

[41] H. Ceniceros and T.Y. Hou, “An efficient dynamically adaptive mesh for potentially singular solutions”, *J. Comp. Phys.* **172** 1–31 (2001)

[42] S. Chandrasekhar, Hydrodynamic and Hydromagnetic Stability, Clarendon Press: Oxford (1961)

[43] Q. Chang, E. Jia and W. Sun, “Difference schemes for solving the generalized nonlinear Schrödinger equation”, *J. Comp. Phys.* **148** 397–415 (1999)

[44] A.J. Chorin and J.E. Marsden, A Mathematical Introduction to Fluid Mechanics, Springer-Verlag: New York (1979)

[45] CLAWPACK, Conservation LAWs software PACKage, www.amath.washington.edu/˜claw

[46] P.C. Clemmow and J.P. Dougherty, Electrodynamics of Particles and Plasmas, Addison Wesley, 2nd ed. (1990)

[47] P. Colella, A. Majda and V. Roytburd, “Theoretical and numerical structure for reacting shock waves”, *Siam J. Sci. Stat. Comput.* **7** 1059–1080 (1986)

[48] B. Coppi, M.N. Rosenbluth and R.N. Sudan, “Nonlinear interactions of positive and negative energy modes in a rarefied plasma, I”, *Ann. Phys.* **55** 207–247 (1969)

[49] C.J. Cotter, D.D. Holm and P.E. Hydon, “Multi-symplectic formulation of fluid dynamics using the inverse map”, *Proc. Roy. Soc. Lond. A* **463** 2617–2687 (2007)

[50] R. Courant and K.O. Friedrichs, Supersonic Flow and Shock Waves, Springer: New York (1976) (Reprint of 1948 edition)

[51] R. Courant and D. Hilbert, Methods of Mathematical Physics, Vol. 2, Wiley: New York (1989)

[52] G.D. Crapper, “Steady magnetohydrohydrodynamic flow past a source”, *J. Inst. Maths. Appl.* **1** 241 (1965)

[53] C. De Boor, A Practical Guide to Splines, Springer-Verlag: New York (1978)

[54] G. Denk and C. Penski, “Integration schemes for highly oscillatory DAEs with applications to circuit simulation”, *J. Comput. Appl. Math.* **82** 79–91 (1997)

[55] R.L. Dewar, “Interaction between hydromagnetic waves and a time dependent inhomogeneous medium”, *Phys. Fluids* **13** (11) 2710–2720 (1970)

[56] R.L. Dewar, “Energy momentum tensors for dispersive electromagnetic waves”, *Aust. J. Phys.* **30** 533–575 (1977)

[57] R.K. Dodd, J.C. Eilbeck, J.D. Gibbon and H.C. Morris, Solitons and Nonlinear Wave Equations, Academic Press: London, UK (1982)

[58] E.A. Dorfi and L.O’C. Drury, “Simple adaptive grids for 1-D initial value problems”, *J. Comp. Phys.* **69** 175–195 (1987)

[59] J. Douglas and H. Rachford, “On the numerical solution of heat conduction problems in two or three space variables”, *Trans. Amer. Math. Soc.* **82** 421–439 (1956)

[60] J. Du, B. Fix, J. Glimm, X.-C. Jia, X.-L. Li, Y.-H. Li and L.-L. Wu, “A simple package for front tracking”, *J. Comp. Phys.* **213** 613–628 (2006)

[61] B. Einfeldt, C.D. Munz, P.L. Roe and B. Sjögreen, “On Godunov-type methods near low densities”, *J. Comp. Phys.* **92** 273–295 (1991)

[62] A. Erdelyi, W. Magnus, F. Oberhettinger and F.G. Tricomi, Tables of Integral Transforms, Volume 1, (Bateman Manuscript Project), McGraw-Hill: New York (1954)

[63] L.C. Evans, Partial Differential Equations, AMS: Providence, RI (2002)

[64] D. Feit and J. Fleck, “Split-step Fourier methods applied to model nonlinear refractive effects in optically thick media”, *Appl. Optics* **17** 3990 (1978)

[65] L.A. Fisk and W.I. Axford, “Anisotropies of solar cosmic rays”, *Solar Phys.* **7** 486–498 (1969)

[66] C.A.J. Fletcher, Computational Techniques in Fluid Dynamics, Vol. 1, Springer-Verlag: Berlin, New York (1991)

[67] B. Fornberg, A Practical Guide to Pseudo-spectral Methods, Cambridge University Press: Cambridge, UK (1996)

[68] D. Frederick and T.S. Chang, Continuum Mechanics, Scientific Publ. Inc.: Boston, MA (1972)

[69] M. Gage and R.S. Hamilton, “The heat equation shrinking convex plane curves”, *J. Differential Geom.* **23** 69–96 (1986)

[70] C.S. Gardner, “Korteweg-de Vries equation and generalizations, IV. The Korteweg de-Vries equation as a Hamiltonian system”, *J. Math. Phys.* **12** 1548–1551 (1971)

[71] I.M. Gelfand and S.V. Fomin, Calculus of Variations, Prentice Hall: NJ (1963) (translated by R.A. Silverman)

[72] I.M. Gelfand and G.E. Shilov, Generalized Functions, (Properties and Operations), Vol. 1, Academic Press: NY (1964) (translated by Eugene Saletan)

[73] J. Glimm, J.W. Grove, X.-L. Li, K.-M. Shyue, Q. Zhang and Y. Zeng, "Three dimensional front tracking", *SIAM J. Sci. Comp.* **19** 703–727 (1998)

[74] S.K. Godunov, "Reminiscences about difference schemes", *J. Comp. Phys.* **153** 625 (1999)

[75] B. Gustafsson, H.O. Kreiss and J. Oliger, Time Dependent Problems and Difference Methods, Wiley (1995)

[76] G.R. Hadley, "Transparent boundary condition for the beam propagation method", *IEEE J. Quantum Electronics* **28** 363–370 (1992)

[77] E. Hairer and G. Wanner, Solving Ordinary Differential Equations, II: Stiff and Differential-Algebraic Problems, Springer, 2nd revised ed. (2004)

[78] L. Halperin and L.N. Trefethen, "Wide-angle one-way wave equations", *J. Acoust. Soc. Am.* **84** 1397–1404 (1988)

[79] S.I. Hariharan and T.H. Moulden (Eds.), Numerical Solution of Partial Differential Equations, Pitman/Longmans (1986)

[80] A. Harten, P.D. Lax and B. van Leer, "On upstream differencing and Godunov-type schemes for hyperbolic conservation laws", *SIAM Rev.* **25** 35–61 (1983)

[81] G.W. Hedstrom, "Nonreflecting boundary conditions for non-linear hyperbolic systems", *J. Comp. Phys.* **30** 222–237 (1979)

[82] R.L. Higdon, "Radiation boundary conditions for dispersive waves", *SIAM J. Num. Anal.* **31** 64–100 (1994)

[83] G.P. Hochschild, Basic Theory of Algebraic Groups and Lie Algebras, Springer-Verlag: New York, Heidelberg, Berlin (1981)

[84] D.D. Holm, J.E. Marsden, T.S. Ratiu and A. Weinstein, "Nonlinear stability of fluids and plasmas", *Phys. Reports* **123** 1–116 (1985)

[85] D.D. Holm, J.E. Marsden and T.S. Ratiu, "The Euler-Poincaré equations and semi-direct products with applications to continuum theories", *Adv. Math.* **137** 1–81 (1998)

[86] W. Huang, Y. Ren and R.D. Russell, "Moving mesh partial differential equations (MMPDEs) based on the equidistribution principle", *SIAM J. Numer. Anal.* **31** 709–730 (1994)

[87] P.E. Hydon, "Multi-symplectic conservation laws for differential and differential-difference equations", *Proc. Roy. Soc. A* **461** 1627–1637 (2005)

[88] N. Jacobson, Lie Algebras, Dover Publ.: NY, Dover edition (1979) (which is a corrected republication of the 1962 book published by Interscience, a division of J. Wiley and Sons)

[89] S.A. Jacques, "Momentum and energy transport by waves in the solar atmosphere and in the solar wind", *Astrophys. J.* **215** 942–951 (1977)

[90] A. Jeffrey, Quasilinear Hyperbolic Systems and Waves, Pitman: London (1976)

[91] J.R. Jokipii, "Cosmic ray propagation, I. Charged particles in a random magnetic field", *Astrophys. J.* **146** 480–487 (1966)

[92] J.R. Jokipii, "Propagation of cosmic rays in the solar wind", *Rev. Geophys. Space Phys.* **9** 27–87 (1971)

[93] C. Kane, J.E. Marsden and M. Ortiz, "Symplectic-energy-momentum preserving variational integrators", *J. Math. Phys.* **40** 3353–3371 (1999)

[94] N. Kaneda, B. Houshmand and T. Itoh, "FDTD analysis of dielectric resonators with curved surfaces", *IEEE Trans. Microwave Theory Tech.* **45** 1645–1649 (1997)

[95] P.O. Kano, M. Brio and J.V. Moloney, "Application of Weeks method for the numerical inversion of the Laplace transform to the matrix exponential", *Comm. Math. Sci.* **3** 335–372 (2005)

[96] P.O. Kano, M. Brio and J.V. Moloney, "Numerical analysis of the ab initio computation of the effects of ionization on the nonlinear susceptibility coefficients of the hydrogen atom", *Comm. Math. Sci.* **4** 5380 (2006)

[97] L.W. Kantorovitch, "Functional analysis and applied mathematics", *Uspekhi Mat. Nauk, USSR* **3** 89 (1948)

[98] D.J. Kaup and A.C. Newell, "An exact solution for a derivative nonlinear Schrödinger equation", *J. Math. Phys.* **19** 798 (1978)

[99] D.J. Kaup, A. Reiman and A. Bers, “Space time evolution of nonlinear three wave resonant interactions I, interaction in a homogeneous medium”, *Rev. Mod. Phys.* **51** (2) 275–309 (1979)

[100] P.E. Kloeden and E. Platen, Numerical Solution of Stochastic Differential Equations, Springer: New York (1999)

[101] P. Knupp and S. Steinberg, Fundamentals of Grid Generation, CRC Press: Florida (1997)

[102] L.D. Landau and E.M. Lifshitz, Fluid Mechanics, Course of Theoretical Physics, Vol. 6, Pergamon Press: Oxford, 2nd ed. (1987)

[103] P.D. Lax, “Hyperbolic systems of conservation laws II”, *Comm. Pure and Appl. Math.* **10** 105–119 (1957)

[104] P.D. Lax, Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves, Regional Conf. Series in Appl. Math., SIAM: Philadelphia, PA (1973)

[105] P.D. Lax and C.D. Levermore, “The small dispersion limit of the Korteweg-de Vries equation. I”, *Comm. Pure Appl. Math.* **36** 253–290 (1983)

[106] J. Lega and A. Goriely, “Pulse, fronts and oscillations of an elastic rod”, *Physica D* **132** 373–391 (1999)

[107] R.B. Lehoucq, D.C. Sorensen and C. Yang, ARPACK Users’ Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods, SIAM: Philadelphia, PA (1998)

[108] S.K. Lelé, “Compact finite difference schemes with spectral like resolution”, *J. Comp. Phys.* **103** 16–42 (1992)

[109] R.J. LeVeque, Time-Split Methods for Partial Differential Equations, Ph.D. Dissertation, Stanford University (1982)

[110] R.J. LeVeque and H.C. Yee, “A study of numerical methods for hyperbolic conservation laws with stiff source terms”, *J. Comp. Phys.* **86** 187–210 (1990)

[111] R.J. LeVeque, Numerical Methods for Conservation Laws, Birkhauser Verlag, 2nd ed. (1992)

[112] L. Li, “New formulation of the Fourier modal method for crossed surface-relief gratings”, *J. Opt. Soc. Am. A* **14** 2758–2767 (1997)

[113] S.T. Li, J.M. Hyman and L.R. Petzold, “An adaptive moving mesh method with static rezoning for partial differential equations”, *Comput. Math. Appl.* **46** 1511–1524 (2003)

[114] M.J. Lighthill, “The diffraction of a blast. I”, *Proc. Roy. Soc. Lond. A.* **198** 454–470 (1949)

[115] M.J. Lighthill, Introduction to Fourier Analysis and Generalised Functions, Cambridge University Press: Cambridge, UK (1958)

[116] M.J. Lighthill, “Studies on magneto-hydromagnetic waves and other anisotropic wave motions”, *Phil. Trans. Roy. Soc. London* **A** 398–427 (1960)

[117] M.J. Lighthill, Waves in Fluids, Cambridge University Press: Cambridge, UK (1978)

[118] T.P. Liu, “Admissible solutions of hyperbolic conservation laws”, *Amer. Math. Soc.* (1981)

[119] X.D. Liu and P.D. Lax, “Positive schemes for solving multi-dimensional hyperbolic systems of conservation laws”, *J. Comp. Fluid Dyn.* **5** 133–156 (1996)

[120] J.-J. Liu, H.-K. Lim, J. Glimm and X.-L. Li, “A conservative front tracking method in N-dimensions”, *J. Sci. Comput.* **31** (1–2) 213–236 (2007)

[121] W.E. Lorensen and H.E. Cline, “Marching cubes: A high resolution 3D surface construction algorithm”, *Computer Graphics* **21** (4) 163–169 (1987)

[122] F. Magri, “A simple model of the integrable Hamiltonian equation”, *J. Math. Phys.* **19** (5) 1156–1162 (1978)

[123] A. Majda and R.R. Rosales, “Resonantly interacting weakly nonlinear hyperbolic waves I - a single space variable”, *Stud. In Appl. Maths.* **71** 149 (1984)

[124] J.E. Marsden and T.S. Ratiu, Introduction to Mechanics and Symmetry, Springer: New York (1994)

[125] J.E. Marsden, G.W. Patrick and W.F. Shadwick, “Integration algorithms and classical mechanics”, *Am. Math. Soc.* (1996)

[126] J.E. Marsden and S. Shkoller, “Multisymplectic geometry, covariant Hamiltonians and water waves”, *Math. Proc. Camb. Phil. Soc.* **125** 553–575 (1999)

[127] J.F. McKenzie, “Stationary MHD waves generated by a source in a moving plasma”, *J. Geophys. Res.* **96** (A6) 9491–9501 (1991)

[128] E. Merzbacher, Quantum Mechanics, Wiley: NY (1963)

[129] W. Miller Jr., Symmetry Groups and their Applications, Academic Press: New York (1972) (Note there is a sign error in one of the terms in Miller's book)

[130] K. Mio, T. Ogino, K. Minami and S. Takeda, "Modified nonlinear Schrödinger equation for Alfvén waves propagating along the magnetic field in cold plasmas", *J. Phys. Soc. Japan* **41** 265 (1976)

[131] C.W. Misner, K.S. Thorne and J.A. Wheeler, Gravitation, W.H. Freeman and Co.: San Francisco (1973)

[132] E. Mjølhus, "On the modulational instability of hydromagnetic waves parallel to the magnetic field", *J. Plasma Phys.* **16** 321 (1976)

[133] E. Mjølhus and J. Wyller, "Nonlinear Alfven waves in a finite beta plasma", *J. Plasma Phys.* **40** 299 (1988)

[134] C.B. Moler and C.F. Van Loan, "Nineteen dubious ways to compute the exponential of a matrix", *SIAM Rev.* **20** 801–836 (1978)

[135] C.B. Moler, Numerical Computing with MATLAB, Cambridge University Press (2006)

[136] P.M. Morse and H. Feshbach, Methods of Theoretical Physics, Vols. 1 and 2, Academic Press: New York (1953)

[137] A. Nayfeh, Methods of Normal Forms, Wiley: New York (1993)

[138] A.C. Newell, Solitons in Mathematics and Physics, CBMS-NSF Regional Conf. Series in Appl. Math., Vol. 48, SIAM: Philadelphia, PA (1985)

[139] T.G. Northrop, The Adiabatic Motion of Charged Particles, Interscience: New York (1963)

[140] P.J. Olver, Applications of Lie Groups to Differential Equations, Springer: New York, 2nd ed. (1993)

[141] A.V. Oppenheim and R.W. Schafer, Discrete-Time Signal Processing, Prentice-Hall: Engelwood Cliffs, NJ (1989)

[142] S.J. Osher and J.A. Sethian, "Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations", *J. Comp. Phys.* **97** 12–49 (1988)

[143] S.J. Osher and R. Fedkiw, Level Set Methods and Dynamic Implicit Surfaces, Springer (2002)

[144] D.W. Peaceman and H. Rachford, “The numerical solution of parabolic and elliptic equations”, *J. Soc. Ind. Appl. Math.* **3** 2841 (1955)

[145] R.B. Pember, “Numerical methods for hyperbolic conservation laws with stiff relaxation, I: Spurious solutions”, *SIAM J. Appl. Math.* **53** 1293–1330 (1993)

[146] R.B. Pember, “Numerical methods for hyperbolic conservation laws with stiff relaxation, II. Higher-order Godunov methods”, *SIAM J. Sci. Comp.* **14** 824–859 (1993)

[147] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetter, Numerical Recipes-The Art of Scientific Computing, Cambridge University Press (1986)

[148] A. Ralston, A First Course in Numerical Analysis, McGraw Hill, Kogakusha, Ltd. (1965)

[149] S. Reich, “Multi-symplectic Runge-Kutta Collocation methods for Hamiltonian wave equations”, *J. Comp. Phys.* **157** 473 (2000)

[150] F. Reif, Fundamentals of Statistical and Thermal Physics, McGraw-Hill: New York (1965)

[151] R.D. Richtmyer and K.W. Morton, Difference Methods for Initial-value Problems, Interscience Publishers: New York (1967)

[152] P.L. Roe, “Approximate Riemann solvers, parameter vectors and difference schemes”, *J. Comp. Phys.* **43** 357–372 (1981)

[153] P.L. Roe and M. Arora, “Characteristic-based schemes for dispersive waves, I. The method of characteristics for smooth solutions”, *Numer. Methods Partial Differential Equations* **9** 459–505 (1993)

[154] A. Rogister, “Parallel propagation of low frequency waves in high beta plasmas”, *Phys. Fluids* **14** 2733 (1971)

[155] B. Rossi and S. Olbert, Introduction to the Physics of Space, McGraw Hill (1970)

[156] O.V. Rudenko and S.I. Soluyan, Theoretical Foundations of Nonlinear Acoustics, Consultants Bureau (Plenum): New York (1977) (English translation by R.T. Beyer)

[157] P.L. Sachdev, Nonlinear Diffusive Waves, Cambridge University Press (1987)

[158] R.Z. Sagdeev and A.A. Galeev, Nonlinear Plasma Theory (Eds. T.M. O'Neill and D.L. Book), W.A. Benjamin: New York (1969)

[159] I. Saitoh, Y. Suzuki and N. Takahashi, "The symplectic finite difference time domain method", *IEEE Trans. Magnetics* **37** 3251–3254 (2001)

[160] Z.S. Sacks, D.M. Kingsland, R. Lee and J.F. Lee, "A perfectly matched anisotropic absorber for use as an absorbing boundary condition", *IEEE Trans. Ant. Prop.* **43** 1460–1463 (1995)

[161] D.H. Sattinger and O.L. Weaver, Lie Groups and Lie Algebras with Applications to Physics, Geometry and Mechanics, Springer: New York (1986)

[162] R. Schuhmann and T. Weiland, Conservation of Discrete Energy and Related Laws in the Finite Integration Technique, PIER Monograph Series, Vol. 32, p. 301 (2001)

[163] W.R. Sears, "Some remarks about flow past bodies", *Rev. Mod. Phys.* **32** 701–705 (1960)

[164] S. Selberherr, Analysis and Simulation of Semiconductor Devices, Springer-Verlag: Wien (1984)

[165] J. Sethian, Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, and Computer Vision, Cambridge University Press (1996)

[166] L.F. Shampine, I. Gladwell and S. Thompson, Solving ODEs with MATLAB, Cambridge University Press (2003)

[167] G.F. Simmons, Introduction to Topology and Modern Analysis, McGraw-Hill: NY (1963)

[168] R.D. Skeel, G. Zhang and T. Schlick, "A family of symplectic integrators: Stability, accuracy and molecular dynamics applications", *SIAM J. Sci. Comput.* **18** 203–222 (1997)

[169] G.D. Smith, Numerical Solution of Partial Differential Equations: Finite Difference Methods, Oxford University Press, 3rd ed. (1985)

[170] J. Smøller, Shock Waves and Reaction Diffusion Equations, Springer: New York (1983)

[171] I.N. Sneddon, Elements of Partial Differential Equations, McGraw-Hill: New York (1957)

[172] I.S. Sokolnikoff and R.M. Redheffer, Mathematics of Physics and Modern Engineering, McGraw-Hill: New York (1966)

[173] G.A. Sod, “A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws”, *J. Comp. Phys.* **27** 1–31 (1978)

[174] G.A. Sod, Numerical Methods in Fluid Dynamics, Initial Boundary Value Problems, Cambridge University Press (1986)

[175] L. Spitzer Jr., The Physics of Fully Ionized Gases, Interscience: New York (1962)

[176] G. Strang, “On the construction and comparison of difference schemes”, *SIAM J. Numer. Anal.* **5** 506 (1968)

[177] J.C. Strikwerda, Finite Difference Schemes and Partial Differential Equations, Cambridge University Press, 2nd ed. (2004)

[178] P.A. Sturrock, “Kinematics of growing waves”, *Phys. Rev.* **112** 1488 (1958)

[179] D.G. Swanson, Plasma Waves, Academic Press: Boston, MA (1989)

[180] P.K. Sweeby, “High resolution schemes using flux limiters for hyperbolic conservation laws”, *SIAM J. Num. Anal.* **21** 995–1011 (1984)

[181] S.M. Sze, Physics of Semiconductor Devices, Wiley-Interscience, 2nd ed. (1981)

[182] A. Taflove and H.C. Hagness, Computational Electrodynamics: The Finite-difference Time-domain Method, Artech House: Boston, 3rd ed. (2005)

[183] C.K.W. Tam and J.C. Webb, “Dispersion relation preserving finite difference schemes for computational acoustics”, *J. Comp. Phys.* **107** 262–281 (1993)

[184] F.D. Tappert, The Parabolic Approximation Method, Wave Propagation and Under Water Acoustics, Lecture Notes in Physics, Vol. 70, pp. 224–287 (1977)

[185] M.E. Taylor, Partial Differential Equations, I: Basic Theory, Springer (1996)

[186] B. van Leer, “Towards the ultimate conservative difference scheme. V. – A second-order sequel to Godunov’s method (for ideal compressible flow)”, *J. Comp. Phys.* **32** 101–136 (1979)

[187] R. Von Mises, Mathematical Theory of Compressible Fluid Flows, Academic Press (1958) Arts. 15.7-15.8, pp. 231–233

[188] R.M. Wald, General Relativity, University of Chicago Press: Chicago (1984)

[189] H.J. Weaver, Theory of Discrete and Continuous Fourier Analysis, Wiley: New York (1989)

[190] G.M. Webb and G.P. Zank, "Wave diffraction in weak cosmic ray modified shocks", *Astrophys. J.* **396** 549–574 (1992)

[191] G.M. Webb, R. Ratkiewicz, M. Brio and G.P. Zank, Multi-dimensional MHD Simple Waves, Solar Wind 8 (Eds. D. Winterhalte, J.T. Gosling, S.R. Habbal, W.S. Kurth and M. Neugebauer), AIP Conf. Proc., Vol. 382, pp. 335–338 (1996)

[192] G.M. Webb, R. Ratkiewicz, M. Brio and G.P. Zank, "Multi-dimensional simple waves in gas dynamics", *J. Plasma Phys.* **59** 417–460 (1998)

[193] G.M. Webb, A.R. Zakharian, M. Brio and G.P. Zank, "Wave interactions in MHD and cosmic ray modified shocks", *J. Plasma Phys.* **61** 295–346 (1999)

[194] G.M. Webb, A.R. Zakharian, M. Brio and G.P. Zank, "Nonlinear and three wave resonant interactions in magnetohydrodynamics", *J. Plasma Phys.* **63** 393 (2000)

[195] G.M. Webb, M.P. Sørensen, M. Brio, A.R. Zakharian and J.V. Moloney, "Variational principles, Lie point symmetries, and similarity solutions of the vector Maxwell equations in non-linear optics", *J. Physica D* **191** 49–80 (2004)

[196] J. Weiland and H. Wilhelmsson, Coherent and Nonlinear Interaction of Waves in Plasmas, Pergamon Press: Oxford (1977)

[197] S. Weinberg, Gravitation and Cosmology, J. Wiley and Sons, Inc.: New York (1972)

[198] G.B. Whitham, Linear and Non-linear Waves, Wiley (1974)

[199] T.I. Woodward and J.F. McKenzie, "Stationary MHD structures", *Planet. Space Sci.* **41** (3) 217–228 (1993)

[200] J.-P. Wrenger, "Numerical reflection from FDTD-PMLs: A comparison of the split PML with the unsplit and CFS PMLs", *IEEE Trans. Ant. Prop.* **50** 258–265 (2002)

[201] K.S. Yee, “Numerical solutions of initial boundary value problems involving Maxwell’s equations in isotropic media”, *IEEE Trans. Anten. Prop.* **AP-14** 302–307 (1966)

[202] A.R. Zakharian, M. Brio, J.K. Hunter and G.M. Webb, “The von Neumann paradox in weak shock reflection”, *J. Fluid Mechanics* **422** 193–205 (2000)

[203] A.R. Zakharian, M. Brio and J.V. Moloney, “FDTD based second-order accurate local mesh refinement method for Maxwell’s equations in two space dimensions”, *Comm. Math. Sci.* **2** (3) 497–513 (2004)

[204] A.R. Zakharian, M. Brio, C. Dineen and J.V. Moloney, “Stability of 2D FDTD algorithms with local mesh refinement for Maxwell’s equations”, *Comm. Math. Sci.* **4** (2) 345–375 (2006)

[205] V.E. Zakharov and E.A. Kuznetsov, “Hamiltonian formalism for systems of hydrodynamic type”, *Soviet Sci. Rev., Sec. C: Math. Phys. Rev.* **4** 167–220 (1984) (Ed. S.P. Novikov (Chur: Harwood Academic))

[206] V.E. Zakharov (Ed.), What is Integrability, Springer (1991)

[207] V.E. Zakharov and E.A. Kuznetsov, “Hamiltonian formalism for nonlinear waves”, *Physics, Uspekhi* **40** 1087–1116 (1997)

[208] G.P. Zank, “Oscillatory cosmic ray shock structures”, *Ap. Space Sci.* **140** 301–324 (1988)

[209] G.P. Zank, S. Oughton, M. Neubauer and G.M. Webb, “Properties of mass loading shocks II, magnetohydrodynamics”, *J. Geophys. Res.* **A11** 17051–17074 (1992)

[210] G.P. Zank, G.M. Webb and D.J. Donohue, “Particle injection and the structure of energetic particle modified shocks”, *Astrophys. J.* **406** 67–91 (1993)

[211] Z. Xie, C.H. Chan and B. Zhang, “An explicit fourth-order orthogonal curvilinear staggered-grid FDTD method for Maxwell’s equations”, *J. Comp. Phys.* **175** 739–763 (2002)

This page intentionally left blank

Index

Mathematics in Science and Engineering
Edited by C.K. Chui, Stanford University

Recent titles:
C. De Coster and P. Habets, *Two-Point Boundary Value Problems: Lower and Upper Solutions*
Wei-Bin Zang, *Discrete Dynamical Systems, Bifurcations and Chaos in Economics*
I. Podlubny, *Fractional Differential Equations*
E. Castillo, A. Iglesias, R. Ruíz-Cobo, *Functional Equations in Applied Sciences*
V. Hutson, J.S. Pym, M.J. Cloud, *Applications of Functional Analysis and Operator Theory (Second Edition)*
V. Lakshmikantham and S.K. Sen, *Computational Error and Complexity in Science and Engineering*
T.A. Burton, *Volterra Integral and Differential Equations (Second Edition)*
E.N. Chukwu, *A Mathematical Treatment of Economic Cooperation and Competition Among Nations: with Nigeria, USA, UK, China and Middle East Examples*
V.V. Ivanov and N. Ivanova, *Mathematical Models of the Cell and Cell Associated Objects*
Z. Zong, *Information-Theoretic Methods for Estimating Complicated Probability for Distributions*
A.G. Ramm, *Dynamical Systems Method for Solving Operator Equations*
J. Mishra and S.N. Mishra, *L-System Fractals*
I.V. Konnov, *Equilibrium Models and Variational Inequalities*
H.G. Dehling, T. Gottschalk and A.C. Hoffmann, *Stochastic Modeling in Process Technology*
A. Torokhti and P. Howlett, *Computational Methods for Modelling of Nonlinear Systems*