# Numerical Solution of Partial Differential Equations

Sorin M. Mitran

(S. M. Mitran) Department of Mathematics, Applied Mathematics Program, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-3250

*E-mail address*, S. M. Mitran: mitran@amath.unc.edu

*URL*: http://www.amath.unc.edu/~mitran

2000 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18

Abstract.

# Contents

# Preface

These are lecture notes for course Math 221, Numerical Solution of Partial Differential Equations given by Sorin M. Mitran at the University of North Carolina at Chapel Hill during the Fall Semester of 2002.

# Part 1

# Basic Methods

CHAPTER 1

# Overview of frequently encountered PDE's

## 1. PDE's in the natural sciences

Ordinary and partial differential equations (ODE's and PDE's henceforth) are frequently encountered in numerous areas of study. A knowledge of the basic scientific background is necessary to write down equations of interest. Perhaps one should not be surprised that the same background knowledge is useful in devising solution methods. The typical procedures by which PDE's are derived should be known to researchers working on solution methods.

**1.1. Conservation laws.** A large number of PDE's arise from the physical principle of *conservation*. Physicists have always been interested in describing changes in the world surrounding us. By observation, theory and experiment certain concepts have been arrived at, among which the concept that one can define physical quantities that remain the same during some process. These quantities are said to be *conserved*. Typically a quantity is conserved in a hypothesized *isolated* system. In reality no system is truly isolated and the most interesting applications come about when we study the interaction of two or more systems. This leads to the question of how one can follow the changes in physical quantities of the separate systems. An extremely useful procedure is to set up an accounting procedure. To start with a mundane example, consider the physical quantity of interest to be the quantity of currency $Q$ in a building $B$. If the building is a commonplace one, it is to be expected that when completely isolated, the amount of currency in the building is fixed

$$Q = Q_0 \ . \tag{1.1}$$

$Q_0$ is some constant. Eq. (1.1) is self-evident but not particularly illuminating – of course the amount of money is constant if nothing goes in or out! Similarly in physics, statements such as "the total mass-energy of the universe is constant" are not terribly useful, though one should note this particular statement is not obviously true.

Things get more interesting when we consider a more realistic scenario in which the system is not isolated. People might be coming and going from building $B$ and some might actually have money in their pockets. In more leisurely economic times, one might be interested just in the amount of money in the building at the end of the day. Just a bit of thought leads to

$$Q_n = Q_{n-1} + \Delta Q_{n-1,n}$$

where $Q_n$ is the amount of money at the end of day $n$, $Q_{n-1}$ that from the previous day and $\Delta Q_{n-1,n}$ the difference between money received and that paid in the

building during day $n$

$$\Delta Q_{n-1,n} = R_{n-1,n} - P_{n-1,n} \ .$$

Keeping track of $R_{n-1,n}$ and $P_{n-1,n}$ separately, for instance in two distinct columns on a ledger, seems easier to people more inclined towards addition than subtraction, and this leads to double entry accounting, an important discovery of Renaissance Italy (see http://www.acaus.org/history/hs_pac.html).

As economic activity picks up and we take building $B$ to mean "bank" it becomes important to keep track of the money in the bank at all times, not just at the end of the day. It then makes sense to think of the rate at which money is moving in or out of the building so we can not only track the amount of currency at any given time, but also be able to make some future predictions. Since some time has passed in order for economic activity to pick up, we can assume that addition and subtraction have become much more familiar and are actively taught to small children. We'll therefore use a single quantity $F$ to denote the amount of money leaving or entering building $B$ during time interval $\Delta t$ with the understanding that positive values of $F$ represent incomes and negative ones expenditures. Such understandings go by the name of sign conventions. They're not especially meaningful but it aids communication if we all stick to the same ones. The amount of currency in the building then changes in accordance to

$$Q(t + \Delta t) = Q(t) + F \ \Delta t \ . \tag{1.2}$$

By the time such equations were being written out fluid flow was a scientific frontier investigated by the Bernoullis (see http://www.maths.tcd.ie/pub/ HistMath/People/Bernoullis/RouseBall/RB_Bernoullis.html) and $F$ got to be referred to as a *flux*, the Latin term for flow.

It is readily apparent that (1.2) is a good approximation for small intervals, but probably a bad one if $\Delta t$ is large since economic activity might change from hour to hour. In order to better keep track of things one might think of $F$ as being defined at any given time $t$ so we have $F(t)$ the instantaneous flux of currency at time $t$. By the time people were thinking along these lines Newton and Leibniz had introduced calculus and sufficient time has since passed that the notions of calculus are widely known at least among college students if not small children. We can therefore write

$$Q(t + \Delta t) = Q(t) + \int_t^{t+\Delta t} F(\tau) d\tau \tag{1.3}$$

and get a suitably impressive statement which, form nonewithstanding, carries the same significance as (1.2).

On the verge of the modern era economic activity might really expand and buildings become so large that it makes sense to keep track of the amount of money in individual rooms and also track inflows and outflows through individual doors. We can identify a room or door by its spatial position denoted by $\mathbf{x} = (x_1, x_2, x_3)$ but we encounter a problem in that position vectors such as $\mathbf{x}$ refer to a single point and no matter how small we make the currency it still has to occupy some space. This conceptual difficulty is overcome by introducing a fictitious "density of currency" at time $t$ which we shall denote by $q(\mathbf{x}, t)$. The only real meaning we associate with this density is that if we sum up the values of $q(\mathbf{x}, t)$ in some volume

$\omega$ we obtain the amount of currency in that volume

$$Q(\omega, t) = \int_{\omega} q(\mathbf{x}, t) d\mathbf{x} \ .$$

On afterthought, we might observe that the same sort of question should have arisen when we defined $Q(t)$ as being defined at one instant in time. Ingrained psychological perspectives make $Q(t)$ more plausible, but were we to live our lives such that quantum fluctuations are observable, $Q(t)$ would be much more questionable.

If we have a spatial density for $Q$ it seems natural to do the same for $F$ and we define $\mathbf{f}(\mathbf{x}, t)$ as being the instantaneous flux of currency in a small region around $(\mathbf{x}, t)$. A bit of thought suggests that the flux should be a vector quantity since we have three directions along which a density can be defined. Along any given direction a scalar flux is obtained by a scalar product; in particular along the direction normal to a boundary $\mathbf{n}(\mathbf{x})$ the scalar flux is given by $\mathbf{f}(\mathbf{x}, \tau) \cdot \mathbf{n}(\mathbf{x})$. The relation between the total flux and the flux densities is given by

$$F(\tau) = \int_{\partial B} \mathbf{f}(\mathbf{x}, \tau) \cdot \mathbf{n}(\mathbf{x}) \ d\mathbf{x} \ . \tag{1.4}$$

Careful observers will notice the appearance of $\partial B$ as defining the integration domain. By this we mean that the integration is to be taken over the boundary of the domain $B$, or in everyday terms, the exterior walls and doors of building $B$. There is a bit of inconsistency in the sciences as to what we mean by "flux". Sometimes it means the amount of some quantity passing through a finite region such as $F$ above. Other times it actually means "flux density" such as $\mathbf{f}$. This possibility of confusion shows the value of using the same conventions. Imposing such conventions is however a social activity and subject to historical iteration. In this course the convention "flux"=$\mathbf{f}$ shall be imposed by instructor fiat. Gathering together all the above we can write a much more sophisticated-looking statement

(1.5)

$$Q(B, t+\Delta t) = \int_B q(\mathbf{x}, t+\Delta t) d\mathbf{x} = \underbrace{\int_B q(\mathbf{x}, t) d\mathbf{x}}_{Q(B,t)} + \int_t^{t+\Delta t} \int_{\partial B} \mathbf{f}(\mathbf{x}, \tau) \cdot \mathbf{n}(\mathbf{x}) \ d\mathbf{x} \ d\tau$$

which nonetheless is essentially the same as (1.2) or (1.3).

There are some special cases in which additional events affecting the balance of $Q$ can occur. For instance if by $B$ we mean a reserve bank, money might be (legally) printed and destroyed in the building. Again by analogy with fluid dynamics when such an event occurs we say that there exist *sources* of $Q$ within $B$, much like a spring is a source of surface water. Let $\Sigma(t)$ be the total sources at time $t$. By now we know what to expect; $\Sigma(t)$ might actually be obtained by summing over several sources placed in a number of positions, for instance the separate printing presses and furnaces that exist in $B$. It is useful to introduce a spatial density of sources $\sigma(\mathbf{x}, t)$. Our conservation statement now becomes

$$\int_B q(\mathbf{x}, t+\Delta t) d\mathbf{x} - \int_B q(\mathbf{x}, t) d\mathbf{x} = \tag{1.6}$$

$$\int_t^{t+\Delta t} \int_{\partial B} \mathbf{f}(\mathbf{x}, \tau) \cdot \mathbf{n}(\mathbf{x}) \ d\mathbf{x} \ d\tau + \int_t^{t+\Delta t} \int_B \sigma(\mathbf{x}, \tau) d\mathbf{x} \ d\tau \tag{1.7}$$

The statement above encompasses all physical conservation laws. It is however quite straightforward in interpretation:

change in money in $B$ = net money coming in or going out of $B$ + net money produced or destroyed in $B$.

It should be emphasized that the above statement has true physical meaning and is referred to as an *integral formulation of a conservation law*. The key term is "integral" and refers to the fact that we are summing over some spatial domain. Remember that the densities were artificial constructs that we introduced.

Eq. (1.6) is useful and often applied directly in the analysis of physical systems. From an operational point of view it does have some inconveniences though. These have mainly to do with pesky integration domains $B$ which typically are difficult to describe and over which it is difficult to perform integrations. To avoid this, mathematicians and physicists have gone one further step and imagined $\mathbf{f}(\mathbf{x}, t)$ as being defined everywhere not only on $\partial B$ (the doors and windows of $B$). These internal fluxes can be shown to have a proper physical interpretation to which we shall come back later. For now let's see the implications of this extension. If we not only assume that $\mathbf{f}(\mathbf{x}, t)$ is defined everywhere, but also that it has nice properties such like enough smoothness to ensure differentiability then we can apply the Gauss theorem and transform the integral over $\partial B$ into one over $B$

$$\int_{\partial B} \mathbf{f}(\mathbf{x}, \tau) \cdot \mathbf{n}(\mathbf{x}) \; d\mathbf{x} = \int_B \nabla \cdot \mathbf{f}(\mathbf{x}, \tau) \; d\mathbf{x} \; . \tag{1.8}$$

Here we encounter another convention problem in that some disciplines use outward pointing normals in which case (1.8) holds while other disciplines use an inward pointing normal in which case we have

$$\int_{\partial B} \mathbf{f}(\mathbf{x}, \tau) \cdot \mathbf{n}(\mathbf{x}) \; d\mathbf{x} = - \int_B \nabla \cdot \mathbf{f}(\mathbf{x}, \tau) \; d\mathbf{x} \; . \tag{1.9}$$

Fluid dynamics uses the second convention which leads to (1.9) and this is the one we'll adopt since so many developments in numerical methods for PDE's initially arose from fluid dynamics problems. Applying (1.9) to (1.6) leads to

$$\int_B \left[ q(\mathbf{x}, t + \Delta t) - q(\mathbf{x}, t) + \int_t^{t+\Delta t} \nabla \cdot \mathbf{f}(\mathbf{x}, \tau) \; d\tau \right] d\mathbf{x} = \tag{1.10}$$

$$\int_t^{t+\Delta t} \int_B \sigma(\mathbf{x}, \tau) d\mathbf{x} \; d\tau \; . \tag{1.11}$$

There was nothing special about the shape of the building $B$ or the length of the time interval $\Delta t$ that we used in deriving (1.10). We can therefore consider special, infinitesimal domains and intervals and obtain a differential form

$$\frac{\partial q}{\partial t} + \nabla \cdot \mathbf{f} = \sigma \; , \tag{1.12}$$

where, as is customary, the dependence of $q, \mathbf{f}, \sigma$ on space and time is understood but not written out explicitly. Eq. (1.12) is known as the *local* or *differential form* of the conservation law for $E$. It is often easier to work with since there are no complications arising from the domain shape that appear directly in the statement of conservation.

In physics the above scenario is encountered many times. Physicists have arrived at certain quantities which obey (1.6). In many situation it is permissible to speak of local quantities and use (1.12). Classical physics arrived at mass, momentum, energy and electrical charge as physical concepts that lead to quantities that

satisfy conservation laws. Contemporary physics unified momentum and energy in the theory of relativity and also gave new, microscopic quantities that satisfy conservation such as lepton number.

**1.2. Special forms of conservation laws.**

1.2.1. *Newton's law.* The full general form (1.12) often arises in real-world applications. Many times it is possible to carry out certain simplifications that lead to equations that are easier to solve. As a simple example, consider the classic problem of dynamics of studying the motion of a point mass $m$. It has no internal structure and its motion is characterized by the second law of dynamics which is a statement of conservation of momentum

$$\frac{d}{dt}(m\mathbf{v}) = \sum \mathbf{F} \ . \tag{1.13}$$

Here we have the correspondence $q \longleftrightarrow (m\mathbf{v})$, $\sigma \longleftrightarrow \sum \mathbf{F}$ with (1.12), hence the statement: "external forces are sources of momentum". Instead of a PDE, the lack of internal structure has led to an ODE.

1.2.2. *Advection equations.* Other special forms of (1.12) are not quite so trivial. Often $\mathbf{f}, \sigma$ depend on $q$, that is we have $\mathbf{f}(q), \sigma(q)$. The specific form of this dependence is given by physical analysis typically. But accounting for all physical effects is so difficult that simple approximations are often used. For instance we can assume that $\mathbf{f}(q)$ is sufficiently smooth to have a Taylor expansion

$$\mathbf{f}(q) = \mathbf{f}_0 + \mathbf{f}'(q_0)(q - q_0) + \ \ldots = \tag{1.14}$$

and consider what happens when we use various truncations of the Taylor expansion.

Typically we can take $\mathbf{f}_0 = 0$ since it doesn't affect the PDE (1.12) anyway. Choosing a system of units such that $q_0 = 0$, the simplest truncation is

$$\mathbf{f}(q) = \mathbf{f}'(0)q = \mathbf{u}\ q \tag{1.15}$$

and the $\sigma = 0$ form of (1.12) is

$$\frac{\partial q}{\partial t} + \nabla \cdot (\mathbf{u}\ q) = 0 \ . \tag{1.16}$$

If we consider that $\mathbf{u}$ does not depend on the spatial coordinates we obtain

$$\frac{\partial q}{\partial t} + \mathbf{u} \cdot \nabla\ q = 0 \tag{1.17}$$

which goes by the name of the *constant velocity advection equation.* The name comes from its use in modeling the transport of some substance by a flow; this process is known as *advection.* Its one-dimensional form is the basis of much development in numerical methods for PDE's

$$\frac{\partial q}{\partial t} + u\frac{\partial q}{\partial x} = 0 \ , \tag{1.18}$$

and we shall study it in detail.

If $\mathbf{u}$ does depend on $\mathbf{x}$ we have

$$\frac{\partial q}{\partial t} + \nabla \cdot (\mathbf{u}\ q) = \frac{\partial q}{\partial t} + q\ \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla\ q = 0 \tag{1.19}$$

or

$$\frac{\partial q}{\partial t} + \mathbf{u} \cdot \nabla\ q = -q\ \nabla \cdot \mathbf{u} \tag{1.20}$$

known as the *variable velocity advection equation*. In very many cases the advection velocity field $\mathbf{u}$ is divergence free

$$\nabla \cdot \mathbf{u} \ = 0 \ , \tag{1.21}$$

so we have the simpler form

$$\frac{\partial q}{\partial t} + \mathbf{u} \cdot \nabla \ q = 0 \ . \tag{1.22}$$

1.2.3. *Diffusion equations.* Another widely encountered dependence of $\mathbf{f}$ on $q$ is of the form

$$\mathbf{f}(q) = -\alpha \nabla q \tag{1.23}$$

and this leads to

$$\frac{\partial q}{\partial t} - \nabla \cdot (\alpha \nabla q) = \sigma(q) \ . \tag{1.24}$$

This is known as the *heat equation* or the *diffusion equation*. If $\alpha$ (the thermal diffusivity) is a constant we have

$$\frac{\partial q}{\partial t} = \alpha \nabla^2 q + \sigma(q) \ , \tag{1.25}$$

a widely encountered form of the heat equation. For many problems time evolution is so slow that the $\partial q/\partial t$ derivative is negligible and (1.25) becomes

$$\nabla^2 q = -\sigma/\alpha \tag{1.26}$$

known as the *Poisson equation*. If $\sigma = 0$ we obtain the special form

$$\nabla^2 q = 0 \tag{1.27}$$

known as the *Laplace* or *harmonic equation*.

1.2.4. *Advection-diffusion equations.* As might be expected, the physical flux dependence might combine the two forms (1.15), (1.23) encountered above

$$\mathbf{f}(q) = \mathbf{u} \ q - \alpha \nabla q \ , \tag{1.28}$$

from which we obtain

$$\frac{\partial q}{\partial t} + \mathbf{u} \cdot \nabla \ q = \sigma(q) + \nabla \cdot (\alpha \nabla q) - q \ \nabla \cdot \mathbf{u} \ , \tag{1.29}$$

known, naturally enough, as the *advection-diffusion equation*. Again, in most applications $\mathbf{u}$ is divergence-free so (1.29) becomes

$$\frac{\partial q}{\partial t} + \mathbf{u} \cdot \nabla \ q = \sigma(q) + \nabla \cdot (\alpha \nabla q) \ . \tag{1.30}$$

1.2.5. *Vector valued conservation laws.* Up to now we have considered that the conserved quantity is a scalar $q$. Often it is more convenient to group scalars together as a vector, for instance when thinking of the momentum of a body. The generalization of the conservation law (1.12) is immediate

$$\frac{\partial \mathbf{q}(\mathbf{x},t)}{\partial t} + \nabla \cdot \mathbf{f}(\mathbf{q}(\mathbf{x},t),(\mathbf{x},t)) = \sigma(\mathbf{q}(\mathbf{x},t),(\mathbf{x},t)) \ . \tag{1.31}$$

Here the explicit dependence on space $\mathbf{x}$ and time $t$ of $\mathbf{q}$ has been pointed out, as well as the possible dependence of the fluxes $\mathbf{f}$ and sources $\sigma$ on both space and time and the conserved variables $\mathbf{q}(\mathbf{x},t)$. Note that $\nabla \cdot \mathbf{f}$ has a different meaning in the present context. As a result of taking the divergence we should still obtain

a vector quantity for (1.31) to be consistent. This means that $\mathbf{f}$ is now a *tensor* of dimension $n \times n$ where $n$ is the number of components of $\mathbf{e}$ (and $\sigma$).

1.2.6. *Convection-diffusion equations.* In fluid flow, among other applications, the velocity field $\mathbf{u}$ is related to the conserved quantities

$$\mathbf{u} = \mathbf{u}\ (\mathbf{x}, t, \mathbf{q})\ . \tag{1.32}$$

This particular situation goes by the name of *convection*. Similar to (1.29) we can write a convection-diffusion equation

$$\frac{\partial q}{\partial t} + \mathbf{u}(q) \cdot \nabla\ q = \sigma(q) + \nabla \cdot (\alpha \nabla q) - q\ \nabla \cdot \mathbf{u} \tag{1.33}$$

and its vector valued generalization

$$\frac{\partial \mathbf{q}}{\partial t} + \mathbf{u}(\mathbf{q}) \cdot \nabla\ \mathbf{e} = \sigma(\mathbf{q}) + \nabla \cdot (\alpha \nabla \mathbf{q}) - \mathbf{q}\ \nabla \cdot \mathbf{u}\ . \tag{1.34}$$

**1.3. Conservative and non-conservative forms.** We have seen that a large class of differential equations are derived from conservation laws. The basic form of a conservation law is:

time change = -(difference in outward fluxes) + (sources).

In mathematical terms we have the local, differential formulation

$$\frac{\partial q}{\partial t} = -\nabla \cdot \mathbf{f}(q) + \sigma\ . \tag{1.35}$$

This is known as the *conservative form* of the law of conservation of $q$. The same principle of conservation might be stated differently if $\nabla \cdot \mathbf{f}(q)$ is expanded. For instance, when $\mathbf{f} = \mathbf{u}\ q$ we can derive from the conservative form

$$\frac{\partial q}{\partial t} = -\nabla \cdot (\mathbf{u}\ q) + \sigma \tag{1.36}$$

the mathematically equivalent form

$$\frac{\partial q}{\partial t} = -q\ \nabla \cdot \mathbf{u} - \mathbf{u} \cdot \nabla q + \sigma\ . \tag{1.37}$$

Eq. (1.37) is known as the *non-conservative form* of the conservation law for $q$. Though equivalent from the analytical point of view, the numerical solution procedures for the two forms show different characteristics as we shall see later on.

## 2. PDE's in other disciplines

Historically, most of the effort in studying PDE's has been directed at those suggested by mathematical physics and that somehow arise from a conservation law. Recently PDE's have been introduced in a number of other fields of study such as mathematical finance or ecology. For instance an important development in mathematical finance is the Black-Scholes equation describing the trading of European options

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 V}{\partial s^2} + r S_t \frac{\partial V}{\partial s} - rV = 0 \tag{2.1}$$

with $V$ the value of an option, $t$ time and $s$ an asset allocation. Though such equations arise from different fundamental principles it is striking that they have the same form as those arising in mathematical physics. The Black-Scholes equation can be described as a mixed diffusion advection equation with a source term for example. We shall concentrate on a mathematical physics background in discussing

numerical solution of PDE's but keep in mind that the same methods are widely applicable.

## 3. Typical problems involving ODE's and PDE's

Now that we have arrived at the general form of PDE's which are of interest in many applications we can turn to actually finding solutions. An important first observation is that specifying the equation to be solved does not allow a unique solution. We must also specify additional *boundary* and/or *initial conditions*. Just as we have important special equations such as the advection or the Laplace equation, there are a number of important, archetypal problems involving ODE's and PDE's.

**3.1. Initial value problem for ODE's.** The simplest problem is the *initial value problem* for a first order system of ODE's

$$\begin{cases} \mathbf{q}' = \mathbf{f}(t, \mathbf{q}) \\ \mathbf{q}(t = 0) = \mathbf{q}_0 \end{cases} . \tag{3.1}$$

This also encompasses initial value problems for ODE's of higher order since an ODE of order $p$ can always be rewritten as a system of $p$ ODE's of order 1. To exemplify, consider

$$q''' = g(t, q, q', q'') \, . \tag{3.2}$$

By introducing the auxilliary functions

$$r = q', \; s = q'' \tag{3.3}$$

we obtain the system

$$\frac{d}{dt} \begin{bmatrix} q \\ r \\ s \end{bmatrix} = \begin{bmatrix} r \\ s \\ g(t, q, r, s) \end{bmatrix} \tag{3.4}$$

which is of the form (3.1).

**3.2. Boundary value problem for ODE's.** For ODE's of order 2 or greater or for systems of two or more ODE's one can meaningfully impose boundary conditions at distinct points within the computational domain. The archetypal ODE boundary value problem is for a second order ODE with conditions at the end points of the computational domain

$$\begin{cases} q'' = f(t, q, q') \\ q(a) = q_1 \\ q(b) = q_2 \end{cases} . \tag{3.5}$$

Instead of the function values, its derivatives might be specified such as in

$$\begin{cases} q'' = f(t, q, q') \\ q'(a) = r_1 \\ q'(b) = r_2 \end{cases} . \tag{3.6}$$

**3.3. Initial value problems for PDE's.** We can pose boundary and/or initial value conditions for PDE's also. It should be noted that not all combinations of PDE's and boundary conditions are compatible. For a large class of phenomena modeled by differential equations we have a reasonable expectation that small changes in the boundary conditions should lead to small changes in the solution. We would also expect the solution to exist and be unique. This means that the solution should depend continuously on the boundary data. Problems for which this holds are said to be *well posed in the sense of Hadamard* and we shall concentrate almost exclusively on this type of problems. Note that not all phenomena modeled need to behave this way as shown by the sensitive dependence on initial data shown in chaotic behavior.

The PDE initial value problem (IVP) most closely related to (3.1) is that posed for the scalar advection equation

$$\begin{cases} \dfrac{\partial q}{\partial t} + u\dfrac{\partial q}{\partial x} = \sigma(x,t,q) \\ q(x,t=0) = q_0(x),\ -\infty < x < \infty \end{cases} .$$

This problem is well posed and straight forward to solve as we shall see later on. The advection equation is the simplest example of the class of *hyperbolic* PDE's. The name is a result of historical accident; the first time PDE's were actively studied scientists were interested in second order PDE's the classification of which can be related to that of quadratic curves.

We can also consider PDE's of similar form for vector variables

$$\begin{cases} \dfrac{\partial \mathbf{q}}{\partial t} + \mathbf{A}\dfrac{\partial \mathbf{q}}{\partial x} = \sigma(x,t,\mathbf{q}) \\ \mathbf{q}(x,t=0) = \mathbf{q}_0(x),\ -\infty < x < \infty \end{cases} . \tag{3.7}$$

This IVP is well posed if the eigenvectors of the matrix $\mathbf{A}$ form a complete set.

**3.4. Boundary value problems for PDE's.** The archetypal boundary value problems are posed for the Poisson equation. Here are the most commonly encountered problems exemplified for the 2D Poisson equation.

(1) *Dirichlet problem*, in which the values of the unknown function are given along the solution domain's boundary

$$\begin{cases} \dfrac{\partial^2 q}{\partial x^2} + \dfrac{\partial^2 q}{\partial y^2} = \sigma(x,y,q),\ (x,y) \in \Omega \\ q(x,y) = F(x,y),\ (x,y) \in \partial\Omega \end{cases} . \tag{3.8}$$

Here, and in the following, $\Omega$ is the domain over which the problem is defined and $\partial\Omega$ is its boundary.

(2) *Neumann problem*, in which the values of the normal derivative of the unknown function are given along the solution domain's boundary

$$\begin{cases} \dfrac{\partial^2 q}{\partial x^2} + \dfrac{\partial^2 q}{\partial y^2} = \sigma(x,y,q),\ (x,y) \in \Omega \\ \dfrac{\partial q}{\partial n}(x,y) = F(x,y),\ (x,y) \in \partial\Omega \end{cases} . \tag{3.9}$$

(3) *Robin problem*, in which a linear combination of the function and its normal derivative are given on the boundary

$$\begin{cases} \dfrac{\partial^2 q}{\partial x^2} + \dfrac{\partial^2 q}{\partial y^2} = \sigma(x, y, q),\ (x, y) \in \Omega \\ q(x, y) + k(x, y)\dfrac{\partial q}{\partial n}(x, y) = F(x, y),\ (x, y) \in \partial\Omega \end{cases} . \tag{3.10}$$

**3.5. Mixed-type problems for PDE's.** A number of PDE's require both initial and boundary value conditions. The typical case is given by the problem of solving the heat equation on a finite strip $a \leq x \leq b$

$$\begin{cases} \dfrac{\partial q}{\partial t} = \alpha \dfrac{\partial^2 q}{\partial x^2} + \sigma(x, t, q), \\ q(a, t) = F_1(t),\ q(b, t) = F_2(t) \\ q(x, t = 0) = q_0(x) \end{cases} \tag{3.11}$$

CHAPTER 2

# Numerical approaches to solving PDE's

## 1. A general framework for numerical solution of PDE's

Now that we have an idea of the differential equations of interest in applications, let us turn to the problem of finding solutions. Analytical techniques such as separation of variables or Fourier analysis are very useful but for a limited class of problems, generally linear PDE's on domains of simple geometry. Most practical interest arises from non-linear PDE's over domains of complicated shape. The fundamental problem facing us is to determine a function $\tilde{q}$ that approximates the solution of the PDE of interest. A large numer of methods have been devised to solve this problem numerically and we shall study individual methods extensively in later chapters. It is instructive to see that basically all methods can be expressed in a general framework that allows comparison of the strengths and weaknesses of individual methods. Say we are faced with the following general problem:

PROBLEM 1. *Find $q : \Omega \to R^n$, $q \in \mathcal{F}$ that satisfies $Lq = 0$ on $\Omega$ and $Bq = 0$ on $\partial\Omega$ where $L, B$ are operators defined on the normed linear space $\mathcal{F}$.*

Let us see how this abstract statement maps onto one of the typical PDE problems, the Neumann problem:

$$\begin{cases} \dfrac{\partial^2 q}{\partial x^2} + \dfrac{\partial^2 q}{\partial y^2} = \sigma(x, y, q),\ (x, y) \in \Omega \\ \dfrac{\partial q}{\partial n}(x, y) = F(x, y),\ (x, y) \in \partial\Omega \end{cases} . \tag{1.1}$$

We have

$$L = \frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y} - \sigma(x, y, \cdot) \tag{1.2}$$

$$B = \frac{\partial}{\partial n} - F(x, y) \tag{1.3}$$

with $L : \Omega,\ B : \partial\Omega$ and $\mathcal{F}$ some reasonable space of functions, for instance the space of functions continuous up to second order defined on $\Omega$, $\mathcal{C}^{(2)}(\Omega)$.

The advantage of the abstract formulation is that it allows us to concentrate on the typical steps followed when building a numerical procedure without spending too much time on the individual application at hand. We assume that the exact solution $q$ is difficult to find and therefore concentrate on constructing an approximation $\tilde{q} \cong q$. Applying the operators $L, B$ to the approximation $\tilde{q}$ leads to an error, called a *residual*

$$L\tilde{q} = r,\ B\tilde{q} = s\ . \tag{1.4}$$

It is our objective that the residual be as small as possible. In order to quantify the error made, we need a mapping from the function space to which $r$ belongs to

a real number. Such a mapping is called a *functional* and the typical example is the norm functional, e.g. $\|r\|$ and $\|s\|$ . In numerical approximations we often work in the space of $p$-integrable functions $\mathcal{F} = \mathcal{L}^{(p)}(\Omega)$ for which the norm is defined by

$$\|f\| = \left( \int |f(x)|^p \, dx \right)^{1/p} \quad . \tag{1.5}$$

The norm is a good candidate for evaluating the quality of our approximation $\tilde{q}$ since if $\|r\| = 0$ then we know that $r = 0$ and therefore $\tilde{q} = q$. Unfortunately, the presence of the absolute value operation limits the operations that can be carried out on the norm and it is typical to use other functionals in constructing numerical procedures. Instead of using the norm consider the functionals

$$I(r) = \int_\Omega r \; w \; dx, \; J(s) = \int_{\partial\Omega} s \; w' \; dx \; . \tag{1.6}$$

Here we have introduced weight functions $w, w'$ that assign differing importance to errors made in various parts of the domain of integration. In adopting $I(r), \; J(s)$ as measures of the quality of our approximation $\tilde{q}$ we have abandoned the certainty of knowing that $\tilde{q} = q$ when $I(r) = 0$, since a zero value for $I(r)$ could be obtained by cancellation of positive and negative errors thorughout the domain. Appropriately chosen weight functions alleviate this concern somewhat and this is a tradeoff we accept for now with a view to simplicity of the ensuing algorithm. It is possible to eliminate this drawback using the square of the residual as we shall see later on.

To simplify the presentation let us concentrate on $I(r)$ only. The addition of boundary conditions is usually a straightforward matter. Now that we have established a means by which to quantify the quality of an approximation we have to decide on how to build the approximation itself. Since $\mathcal{F}$ is a normed linear space we can express any element $g$ of $\mathcal{F}$ as a linear combination over a set of basis functions

$$g = \sum_{i=1}^{\infty} a_i l_i \; . \tag{1.7}$$

For example the set $\{1, \sin x, \cos x, \sin 2x, \cos 2x, \ldots\}$ is a basis for the square-integrable functions defined on $[0, 2\pi]$, $\mathcal{L}^{(2)}([0, 2\pi])$, and the set $\{1, x, x^2, \ldots\}$ is a basis for the infinitely differentiable functions defined over the reals $C^{(\infty)}(\mathbb{R})$. In practical computations we cannot use an infinite sum, so we construct our approximation $\tilde{q}$ using only $N$ terms

$$\tilde{q} = \sum_{i=1}^{N} c_i l_i \; . \tag{1.8}$$

Using (1.8) $I(r)$ becomes

$$I(r) = \int_\Omega L \left( \sum_{i=1}^{N} c_i l_i \right) \; w \; dx \; . \tag{1.9}$$

For arbitrary coefficients $c_i$ we shall have $I(r) \neq 0$. Since we wish $\tilde{q}$ to be a good approximation we can reasonably impose $I(r) = 0$ and thus obtain an equation to

be used in determining the coefficients $c_i$

$$\int_\Omega L\left(\sum_{i=1}^{N} c_i l_i\right)\ w\ dx = 0\ . \tag{1.10}$$

Now this is just one equation but we have $N$ unknown coefficients. However the weight function is arbitrary so obtaining as many equations as we need is easy: we just choose $w$ from some set of functions $w \in \{w_1, w_2, \ldots, w_N\}$ and obtain a system of $N$ equations for $N$ unknowns

$$\int_\Omega L\left(\sum_{i=1}^{N} c_i l_i\right)\ w_j\ dx = 0,\ j = 1, 2, \ldots, N\ . \tag{1.11}$$

Up to this point we have concentrated on the approximation of $q$ itself. There still remains the question of how to approximate the presumably complex shape of $\Omega$. Generally this is done by approximating $\Omega$ by a set of simple-shaped subdomains $\omega_k$ such that the measure $\rho$ of the difference between the two sets goes to zero as we increase the number of subdomains

$$\rho\left(\Omega - \bigcup_{m=1}^{M} \omega_m\right) \to 0\ . \tag{1.12}$$

Using this approximation of the domain $\Omega$ (1.11) becomes

$$\sum_{m=1}^{M} \int_{\omega_m} L\left(\sum_{i=1}^{N} c_i l_i\right)\ w_j\ dx = 0,\ j = 1, 2, \ldots, N\ . \tag{1.13}$$

This is known as a *weighted residual formulation* and a large number of methods for numerically solving PDE's can be thus expressed. We shall turn to some examples shortly, but let us summarize the basic aspects:

(1) A function space from which we construct approximations is chosen along with a subset of a basis of this space $\{l_1, l_2, \ldots, l_N\}$;
(2) A set of weight functions $\{w_1, w_2, \ldots, w_N\}$ is chosen;
(3) A discretization of the domain $\{\omega_1, \omega_2, \ldots, \omega_M\}$ is chosen.

## 2. Basic numerical methods

### 2.1. Finite difference methods.

2.1.1. *Finite difference derivation.* In a finite difference method (FDM) the derivatives appearing in an ODE or PDE are approximated using finite differences. For example the IVP

$$\begin{cases} q' = f(t, q) \\ q(t = 0) = q_0 \end{cases} \tag{2.1}$$

can be solved over the domain $[0, T]$ by finite differences using the following procedure. We construct an approximation $\tilde{q}$ by a set of point values $Q^n = \tilde{q}(t^n)$ with $t^n = nk$, $n = 0, 1, \ldots, N$ and $k$ a step size $k = T/N$ and assume that $\tilde{q}$ varies linearly between the point values. From the point values we can construct myriad approximations of the value of the derivative of $\tilde{q}$, for example

$$\tilde{q}'(t^n) \cong \frac{Q^{n+1} - Q^{n-1}}{2k},$$

FIGURE 1. The piecewise linear form functions.

and obtain practical algorithms to solve (2.1) such as

$$Q^{n+1} = Q^{n-1} + 2k\ f(t^n, Q^n)\ , \tag{2.2}$$

known as the *midpoint rule*.

2.1.2. *Weighted residual derivation.* We shall analyze (2.1) and similar algorithms extensively later on, but let us now see how the same method can be obtained via the weighted residual formulation and what insights we can thereby gain. We have chosen $\tilde{q}$ as being a piecewise linear approximation defined at the points $t^n = nk$. In the general language of the weighted residual formulation we have $\Omega = [0, T]$, $\omega_m = [t^{m-1}, t^m]$. A basis for the piecewise linear functions defined on this partition of $\Omega$ is given by

$$l_n(t) = \begin{cases} 0 & t < t^{n-1} \\ \frac{1}{k}\left(t - t^{n-1}\right) & t^{n-1} \le t < t^n \\ \frac{1}{k}\left(t^{n+1} - t\right) & t^n \le t < t^{n+1} \\ 0 & t^{n+1} \le t \end{cases} . \tag{2.3}$$

See Fig. (1).

Any piecewise linear function over $\{\omega_m\}$ can be defined as a linear combination of $\{l_n\}$, for instance

$$\tilde{q}(t) = \sum_{i=0}^{N} c_i l_i(t)\ . \tag{2.4}$$

It is apparent from (2.3) that

$$l_m(t^n) = \delta_{mn} = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases} , \tag{2.5}$$

so imposing the conditions $\tilde{q}(t^n) = Q^n$ leads to

$$\tilde{q}(t^n) = \sum_{i=0}^{N} c_i l_i(t^n) = \sum_{i=0}^{N} c_i \delta_{in} = c_n = Q^n\ , \tag{2.6}$$

i.e. the coefficients of the expansion (2.4) are the nodal values $Q^n$. A similar expansion is made to approximate the values of the r.h.s. term

$$f(t, q) \cong \tilde{f}(t, q) = \sum_{i=0}^{N} F^i l_i(t) \tag{2.7}$$

with $F^n = \tilde{f}(t^n, Q^n)$.

We must now choose appropriate weight functions. Since the approximation we are building depends only on nodal values, a reasonable choice would be a set of weight functions that give importance to the residual obtained at the nodes. Such a set is given by

$$w_j = \delta(t - t^j) \tag{2.8}$$

where $\delta(t-t^j)$ is the Dirac delta functional centered on $t^j$ defined by its integral property

$$\int_0^T g(t)\delta(t-t^j)\ dt = g(t^j) \tag{2.9}$$

for any function $g(t)$.

Having chosen the function space on which to base our aproximation $\tilde{q}$, a basis in this space $\{l_m(t)\}$, the weight functions $\{\delta(t-t^j)\}$ and a discretization of the domain $[0,T]$ we can work through the weighted residual formulation (1.13) to obtain

$$\sum_{m=1}^{M}\int_{t^{m-1}}^{t^m}\left[\frac{d}{dt}\left(\sum_{n=1}^{N}Q^n l_n(t)\right) - \left(\sum_{n=1}^{N}F^n l_n(t)\right)\right]\delta(t-t^j)\ dt = 0\ . \tag{2.10}$$

It is easiest to use the properties of the Dirac-$\delta$ function to work through the above expression. We have

$$\int_0^T \frac{d}{dt}\left(\sum_{n=1}^{N}Q^n l_n(t)\right)\ \delta(t-t^j)\ dt = \int_0^T\left(\sum_{n=1}^{N}F^n l_n(t)\right)\ \delta(t-t^j)\ dt \tag{2.11}$$

$$\sum_{n=1}^{N}Q^n\int_0^T l_n'(t)\ \delta(t-t^j)\ dt = \sum_{n=1}^{N}F^n\int_0^T\ l_n(t)\ \delta(t-t^j)\ dt \tag{2.12}$$

$$\sum_{n=1}^{N}Q^n\ l_n'(t_j) = \sum_{n=1}^{N}F^n l_n(t^j) = \sum_{n=1}^{N}F^n\delta_{nj} = F^j \tag{2.13}$$

We should be careful in evaluating $l_n'(t_j)$ since $l_n(t)$ is not differentiable at the nodal points $t=t^n$. If we interpret $l_n'(t_j)$ in principal value as the average of the limits to the left and the right we have

$$l_n'(t_j) = \begin{cases} -\frac{1}{2k} & n = j-1 \\ 0 & n \neq j\pm 1 \\ \frac{1}{2k} & n = j+1 \end{cases}\ . \tag{2.14}$$

Eq. (2.13) becomes

$$\frac{Q^{j+1}-Q^{j-1}}{2k} = F^j \tag{2.15}$$

which is exactly the same expression we had obtained previously, Eq. (2.2).

2.1.3. *Comparison of the two derivations.* It is satisfying to see a method derived in two ways, but an immediate question is what is to be gained by more complicated weighted residual procedure in comparison to the straightforward finite difference derivation. Generally the benefit arises in theoretical considerations of the behavior of the method. For instance let us consider the following theorem from approximation theory.

THEOREM 1. *Let $\mathcal{V}$ be a metric linear space with a metric induced by the scalar product $(\cdot,\cdot)$ on $\mathcal{V}$ and let $\mathcal{S}$ be a subspace of $\mathcal{V}$. Let $v \in \mathcal{V}$ and $u \in \mathcal{S}$. If $v-u$ is orthognal to any $w \in \mathcal{S}$, then $u$ is the best approximation of $v$ within $\mathcal{S}$.*

PROOF. Let $d(u,v)$ be the distance induced by the scalar product between $u$ and $v$. Ask whether any other $w \in \mathcal{S}$ gives a smaller distance to $v$

$$[d(v,w)]^2 = (v-w, v-w) = (v-u+u-w, v-u+u-w) \tag{2.16}$$

$$= (v-u, v-u) + 2(u-w, v-u) + (u-w, u-w) \tag{2.17}$$

$$= \|v-u\|^2 + \|u-w\|^2 + 2(u-w, v-u)\ . \tag{2.18}$$

Since $v-u$ is orthogonal to any element in $\mathcal{S}$, it is orthogonal to $u-w$ so $(u-w, v-u) = 0$ and

$$[d(v,w)]^2 = \|v-u\|^2 + \|u-w\|^2 \geq \|v-u\|^2 = [d(v,u)]^2 \tag{2.19}$$

and we conclude that $u$ is the best approximation of $v$ within $\mathcal{S}$. □

We can apply such theorems to weighted residual derivations to infer the behavior of the numerical approximation. Applied to the above example, $\mathcal{V}$ would be the space of differentiable functions to which $q$ belongs. $\mathcal{S}$ would be the subspace of piecewise continuous functions where we defined our approximation $\tilde{q}$. The best approximation we could obtain would be orthogonal to the complement $\mathcal{S}$ of within $\mathcal{V}$. This subspace would contain functions that are not expressible as an expansion along the set (2.3), for instance functions that vary more rapidly than the time step chosen $k$. Predictions such as this are typically more difficult to obtain from the simpler finite difference derivation.

Let us consider an application of the above theorem. Let $\mathcal{V}$ be the space of differentiable functions defined on the interval $[0,T]$. We introduce the scalar product

$$(u,v) = \int_0^T u(t)v(t)\ dt\ ,$$

with $u, v \in \mathcal{V}$. The metric induced by the scalar product is

$$d(u,v) = (u-v, u-v)^{1/2} = \left( \int_0^T [u(t)-v(t)]^2\ dt \right)^{1/2}\ .$$

Now let us consider the problem of approximating elements of $\mathcal{V}$ by piecewise linear functions defined by their point values on a partition of the interval $[0,T]$. Let $\{0 = t^0, t^1, \ldots, t^{N-1}, t^N = T\}$ be the partition of this interval. To keep things simple we'll use an uniform partition with $t^n = t^0 + nk$, $k = T/N$. The subspace of piecewise linear functions is

$$\mathcal{S} = \left\{ (Q^n)\ \mid n = 0, 1, \ldots, N,\ \tilde{q}(t) = \frac{\left(Q^n - Q^{n-1}\right)}{k}(t - t^{n-1}) + Q^{n-1} \text{ for } t^{n-1} \leq t \leq t^n \right\}\ .$$

A piecewise linear function is obviously continuous but is not differentiable at the partition points $t^n$ so at present we cannot affirm that $\mathcal{S}$ is a subspace of $\mathcal{V}$. This drawback is easily eliminated by imposing a principal value definition of the derivative at the partition points, i.e. for $\tilde{q} \in \mathcal{S}$ we define

$$\tilde{q}'(t^n) = \frac{1}{2}\left[\lim_{\substack{t\to t^n \\ t<t^n}} \tilde{q}'(t) + \lim_{\substack{t\to t^n \\ t>t^n}} \tilde{q}'(t)\right] = \frac{Q^{n+1} - Q^{n-1}}{2k} \ .$$

With this definition we can now state that $\mathcal{S}$ is a subspace of $\mathcal{V}$. Let's determine the scalar product in $\mathcal{S}$ induced by that defined for $\mathcal{V}$. With $\tilde{q}, \tilde{r} \in \mathcal{S}$ we have

$$(\tilde{q}, \tilde{r}) = \int_a^b \tilde{q}(t)\tilde{r}(t)\ dt = \sum_{n=1}^{N} \int_{t^{n-1}}^{t^n} \left[\frac{\left(Q^n - Q^{n-1}\right)}{k}(t - t^{n-1}) + Q^{n-1}\right]\left[\frac{\left(R^n - R^{n-1}\right)}{k}(t - t^{n-1}) + R^{n-1}\right] dt$$
$$= \frac{k}{6}\left(2Q^nR^n + 2Q^{n-1}R^{n-1} + Q^nR^{n-1} + Q^{n-1}R^n\right) \ .$$

Let us now look for the best approximation of an element $q \in \mathcal{V}$ by an element $\tilde{q} \in \mathcal{S}$. By the above theorem, the best approximation possible satisfies the condition

$$(q - \tilde{q}, \tilde{r}) = 0$$

for all $\tilde{r} \in \mathcal{S}$. This leads to

$$\int_0^T [q(t) - \tilde{q}(t)]\ \tilde{r}(t)\ dt = 0$$

or

$$\sum_{n=1}^{N} \int_{t^{n-1}}^{t^n} \left[q(t) - \frac{\left(Q^n - Q^{n-1}\right)}{k}(t - t^{n-1}) - Q^{n-1}\right]\ \tilde{r}(t) = 0 \ .$$

The unknowns of the above problem are the nodal values $\{Q^n\}$ and we can obtain the relation

(2.20)

$$\sum_{n=1}^{N}\left\{\frac{\left(Q^n - Q^{n-1}\right)}{k}\int_{t^{n-1}}^{t^n}(t - t^{n-1})\ \tilde{r}(t)\ dt + Q^{n-1}\int_{t^{n-1}}^{t^n}\tilde{r}(t)\ dt\right\} = \int_0^T q(t)\,\tilde{r}(t)\,dt \ .$$

This is a linear relation in the unknowns. If this relation is valid for *all* $\tilde{r} \in \mathcal{S}$ we have determined the best approximation of $q$ by elements within $\mathcal{S}$. Note that $\mathcal{S}$ is of finite dimension $N+1$ so we need only choose $N+1$ functions $\tilde{r}_j \in \mathcal{S}$, $j = 0, 1, \ldots, N$ that form a basis of $\mathcal{S}$. If (2.20) is verified for the elements of the basis then it is verified for all elements within $\mathcal{S}$ (why?). We obtain the system of equations

$$\sum_{n=1}^{N}\left\{\frac{\left(Q^n - Q^{n-1}\right)}{k}\int_{t^{n-1}}^{t^n}(t - t^{n-1})\ \tilde{r}_j(t)\ dt + Q^{n-1}\int_{t^{n-1}}^{t^n}\tilde{r}_j(t)\ dt\right\} = \int_a^b q(t)\,\tilde{r}_j(t)\,dt\ ,\ j = 0, 1, \ldots, N,$$

which is linear system with $N+1$ equations for the $N+1$ unknowns $Q^n$.

The same technique can be applied to determine best approximations to elements $v \in \mathcal{V}$ which are not specified directly. In the above initial value problem (2.1) we are not given $q$ but rather its derivative $q'(t) = Lq(t) = f(t, q)$. We can impose a condition

$$(Lq - L\tilde{q}, \tilde{r}) = (f - L\tilde{q}, \tilde{r}) = 0$$

namely that the residual be orthogonal to the subspace $\mathcal{S}$. This leads to the system

$$\sum_{n=1}^{N} \left\{ \frac{\left(Q^n - Q^{n-1}\right)}{k} \int_{t^{n-1}}^{t^n} \tilde{r}_j(t)\ dt \right\} = \int_0^T f(t,q)\ \tilde{r}_j(t)\ dt\ ,\ j = 0, 1, \ldots, N,$$

again with $\{r_j\}$, $j = 0, 1, \ldots, N$ a basis of $\mathcal{S}$. Note that if we choose $r_j(t) = l_j(t)$, the linear form functions from (2.3), we obtain

$$\sum_{n=1}^{N} \left\{ \frac{\left(Q^n - Q^{n-1}\right)}{k} \int_{t^{n-1}}^{t^n} l_j(t)\ dt \right\} = \frac{Q^{j+1} - Q^{j-1}}{2} = \int_{t^{j-1}}^{t^{j+1}} f(t,q)\, l_j(t)\, dt\ ,\ j = 0, 1, \ldots, N\ .$$

If we additionally introduce a piecewise linear approximation for $f(t,q)$ by

$$f(t,q) = \sum_{n=0}^{N} F^n l_n(t)$$

with $F^n = f(t^n, Q^n)$ we obtain

$$\frac{Q^{j+1} - Q^{j-1}}{2} = \sum_{n=0}^{N} F^n \int_{t^{j-1}}^{t^{j+1}} l_n(t)\ l_j(t)\ dt\ ,\ j = 0, 1, \ldots, N\ .$$

Only three of the integrals give non-zero values,

$$\int_{t^{j-1}}^{t^{j+1}} l_{j-1}(t)\ l_j(t)\ dt = \frac{k}{6}\ ,$$
$$\int_{t^{j-1}}^{t^{j+1}} l_j(t)\ l_j(t)\ dt = \frac{2k}{3}\ ,$$
$$\int_{t^{j-1}}^{t^{j+1}} l_{j+1}(t)\ l_j(t)\ dt = \frac{k}{6}\ ,$$

leading to the final formula

$$\frac{Q^{j+1} - Q^{j-1}}{2} = kF^j\ ,$$

the same as that obtained above, (2.2). We can now however state that the numerical solution obtained by this formula is the *best* possible approximation of the initial value problem (2.1) by piecewise linear functions when the right-hand-side term $f(t,q)$ is also approximated by piecewise linear functions. This statement could not have been made from the simple finite difference derivation.

### 2.2. Finite volume methods.

2.2.1. *Derivation by integrating over a finite volume.* Finite volume methods for PDE's are derived by integrating the equation of interest over a finite region of the solution domain and introducing an average value for the unknown function over each finite region. Let us exemplify for an elementary PDE, the one-dimensional, constant-velocity advection equation

$$q_t + uq_x = 0\ .$$

Say we wish to determine the solution for $(x,t) \in [0,1] \times [0,T]$ starting from the initial condition $q(x, t=0) = \sin(\pi x)$. In a finite volume method we first introduce a partition of the definition domain into finite volumes $V_i^n = [x_{i-1}, x_i] \times [t^{n-1}, t^n]$

with $x_i = ih$, $h = 1/M$, $t^n = nk$, $k = T/N$. We then integrate the PDE over a finite volume

$$\int_{t^{n-1}}^{t^n} \int_{x_{i-1}}^{x_i} (q_t + uq_x)\ dx\ dt = 0\ ,$$

and obtain

$$\int_{x_{i-1}}^{x_i} \left[q(x,t^n) - q(x,t^{n-1})\right]\ dx + \int_{t^{n-1}}^{t^n} \left[uq(x_i,t) - uq(x_{i-1},t)\right] dt = 0\ .$$

An average value for the field variable $q$ over the interval $[x_{i-1}, x_i]$ is now introduced at each time

$$Q_i^n = \frac{1}{h} \int_{x_{i-1}}^{x_i} q(x,t^n)\ dx\ .$$

Similarily an average value for the flux $f(q) = uq$ is introduced over the time interval $[t^{n-1}, t^n]$ for each $x_i$

$$F_i^{n-1} = \frac{1}{k} \int_{t^{n-1}}^{t^n} uq(x_i,t)\ dt\ .$$

Such averages are known to exist by virtue of the mean value theorem. We obtain an update formula

$$Q_i^n = Q_i^{n-1} - \frac{k}{h}\left(F_i^{n-1} - F_{i-1}^{n-1}\right)\ .$$

At first glance this looks just like a backward in time, backward in space finite difference approximation. The interpretation is a bit different though; whereas in a finite difference approximation we would have taken $Q_i^n$ to be the value at $x = x_i$, $t = t^n$ in a finite volume method it is understood as the average value at $t^n$ over the cell $[x_{i-1}, x_i]$. This difference is slight though and indeed we shall see that most finite volume methods have equivalent finite difference formulations.

2.2.2. *Weighted residual derivation.* We can easily recognize the appropriate weights and approximation spaces that lead to a finite volume method. Since we're using average spatial values over a computational cell the appropriate approximation space is that of piecewise constant functions

$$c_i(x) = \begin{cases} 0 & x \le x_{i-1} \\ 1 & x_{i-1} < x \le x_i \\ 0 & x > x_i \end{cases}\ .$$

The time dependence of $\tilde{q}$ is not specified in the finite volume derivation. We can assume a linear variation given by (2.3) such that

$$\tilde{q}(x,t) = \sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t)$$

is our approximation of $q(x,t)$. The weight functions are unity within a cell and zero outside

$$w_j^n(x,t) = \begin{array}{ll} 1 & x_{i-1} \le x \le x_i \text{ and } t^{n-1} \le t \le t^n \\ 0 & \text{otherwise} \end{array}$$

The weighted residual formulation (1.13) becomes

$$\int_0^T \int_0^1 L\left(\sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t)\right) w_j^p(x,t)\ dx\ dt = 0$$

with $L = \partial_t + u\partial_x$. The properties of the weight functions leads to

$$\int_{t^{p-1}}^{t^p}\int_{x_{j-1}}^{x_j} L\left(\sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t)\right)\, dx\, dt = 0\ .$$

We can carry out some of the integrations analytically to obtain

$$\int_{x_{j-1}}^{x_j}\sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t^p)\, dx - \int_{x_{i-1}}^{x_i}\sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t^{p-1})\, dx +$$
$$\int_{t^{p-1}}^{t^p}\sum_{i=1}^{M}\sum_{n=0}^{N} u Q_i^n c_i(x_j) l_n(t)\, dt - \int_{t^{p-1}}^{t^p}\sum_{i=1}^{M}\sum_{n=0}^{N} u Q_i^n c_i(x_{j-1}) l_n(t)\, dt = 0$$

and we can use the identities $c_i(x_j) = \delta_{ij}$, $l_n(t^p) = \delta_{np}$ to obtain

$$\int_{x_{j-1}}^{x_j}\sum_{i=1}^{M} Q_i^p c_i(x)\, dx - \int_{x_{i-1}}^{x_i}\sum_{i=1}^{M} Q_i^{p-1} c_i(x)\, dx +$$
$$\int_{t^{p-1}}^{t^p}\sum_{n=0}^{N} u Q_j^n l_n(t)\, dt - \int_{t^{p-1}}^{t^p}\sum_{n=0}^{N} u Q_{j-1}^n l_n(t)\, dt = 0\ .$$

Another application of the properties of the form functions easily leads to

$$hQ_j^p - hQ_j^{p-1} + kF_j^{p-1} - kF_{j-1}^{p-1} = 0 \tag{2.21}$$

the same formula as was obtained previously. Note that because we have assumed a certain variation in time of $\tilde{q}(x,t)$ we actually can establish a formula for the time-averaged fluxes

$$F_j^{p-1} = \frac{u}{k}\int_{t^{p-1}}^{t^p}\left[Q_j^{p-1} l_{p-1}(t) + Q_j^p l_p(t)\right]\, dt \tag{2.22}$$
$$= \frac{u}{2}\left(Q_j^{p-1} + Q_j^p\right)\ . \tag{2.23}$$

**2.3. Finite element methods.** Finite element methods have traditionally been developed directly from the weighted residual formulation. There various ways in which weights and approximating functions are chosen that shall be studied in some detail later on. As an initial example let us consider the popular Galerkin procedure in which the weight functions are chosen identical to the approximating form functions. A simple example is to build a finite element method to solve the Poisson equation with Dirichlet boundary conditions

$$q_{xx} + q_{yy} = 0 \text{ on } \Omega \tag{2.24}$$
$$q(\Sigma) = f \text{ on } \Sigma = \partial\Omega. \tag{2.25}$$

Let us assume a simple rectangular domain $\Omega = [0,1]\times[0,1]$ and introduce an uniform partition $[x_{i-1}, x_i]\times[y_{j-1}, y_j]$, $x_i = ih$, $y_j = jh$, $h = 1/M$. There are many variants of the finite element method. The basic common thread is that a local approximation valid over a single finite element is used. Consider the so-called

*form functions*

$$N_1(\xi,\eta) = \frac{1}{4}(\xi-1)(\eta-1) \tag{2.26}$$

$$N_2(\xi,\eta) = \frac{1}{4}(\xi-1)(\eta+1) \tag{2.27}$$

$$N_3(\xi,\eta) = \frac{1}{4}(\xi+1)(\eta-1) \tag{2.28}$$

$$N_4(\xi,\eta) = \frac{1}{4}(\xi+1)(\eta+1) \tag{2.29}$$

for $-1 \le \xi, \eta \le 1$ and $N_1 = N_2 = N_3 = N_4 = 0$ otherwise. A local approximation valid over the element $(i,j)$ which has its lower left corner at $(x_i, y_j)$ is given by

$$\tilde{q}(\xi,\eta) = \sum_{k=1}^{4} Q_k^{(i,j)} N_k(\xi,\eta) \tag{2.30}$$

Here $k$ is the local numbering of the unknowns associated with element $(i,j)$. The weighted residual formulation leads to

$$\int_{-1}^{+1}\int_{-1}^{+1} L\left(\sum_{k=1}^{4} Q_k^{(i,j)} N_k(\xi,\eta)\right) N_k(\xi,\eta)\, d\xi d\eta = 0,\ 1 \le k \le 4,\ 1 \le i,j \le M\ . \tag{2.31}$$

This leads to a linear system for the unknown values $Q_k^{(i,j)}$. There intervene various practical complications involving numbering of unknowns and the fact that unknowns are based at the nodes and hence shared by more than one element. The numbering problem is readily solved while the issue of shared unknown values leads to the process of matrix assembly which shall be studied in detail when we analyze finite element methods.

**2.4. Spectral methods.** Up to now we have predominantly used *local*, *polynomial* approximating functions. There are good reasons to make such a choice for very many problems, especially those that contain jumps (discontinuities) in the unknown function or its derivatives. However if it is known that the function is smooth other basis functions can be used, for example global basis functions defined over an entire domain. When these basis functions have a special relation with the operator of the PDE being solved the resulting method is known as a spectral method. For instance for the operator

$$L = -\frac{d^2}{dx^2} \tag{2.32}$$

the functions $\sin(rx), \cos(rx)$ play a special role since they satisfy the eigenfunction relation

$$L\sin(rx) = r^2\sin(rx),\ L\cos(rx) = r^2\cos(rx)\ . \tag{2.33}$$

Such eigenfunctions have remarkable advantages as basis functions; in a very precise sense they offer the most compact representation of any given function in the space which they span. When they are used as basis functions in a numerical approximation the resulting method is known as a *spectral method*. Weight functions must also be chosen and there exist various procedures to do this. Again the Galerkin procedure is widely used in which the basis functions are also used as weight functions.

CHAPTER 3

# Initial Value Problems for Ordinary Differential Equations

## 1. Motivation

ODE numerical methods are directly applicable to solving PDE's. A simple example is the class of semi-discrete methods. Consider for instance the problem of finding $q(x,t)$ defined on $[0,1]\times[0,T]$ that satisfies

$$\begin{cases} q_t = q_{xx} \\ q(x,t=0) = f(x) \\ q(x=0,t) = g_0(t),\ q(x=1,t) = g_1(t) \end{cases} . \tag{1.1}$$

One approach to this PDE problem is to transform it into a system of ODE's by introducing a discretization in space

$$x_i = ih,\ h = 1/M \tag{1.2}$$

and the functions $Q_i(t)$ that approximate $q(x_i,t)$

$$Q_i(t) \cong q(x_i,t),\ i = 0,\ldots,M \tag{1.3}$$

at the nodes $x_i$. A finite difference approximation of the spatial derivative in (1.1) leads to

$$\frac{dQ_i(t)}{dt} = \frac{Q_{i+1}(t) - 2Q_i(t) + Q_{i-1}(t)}{h^2},\ i = 1,\ldots,M-1 \tag{1.4}$$

a system of $M-1$ ODE's. Initial conditions are given by $Q_i(0) = f(x_i)$ and we can apply the boundary conditions through $Q_0(t) = g_0(t)$, $Q_M(t) = g_1(t)$. Thus the IVBP for the heat equation given by Eq. (1.1) has been approximated by the ODE system (1.4).

Not only are the methods that we devise for ODE's applicable to PDE's but also much of the theoretical tools used in analyzing ODE's can be applied to the more difficult case of PDE's.

## 2. Existence of solutions

One of the basic ODE problems is the Cauchy or initial value problem given in general by (3.1). Let us consider the scalar form

$$\begin{cases} q' = f(t,q) \\ q(t=0) = q_0 \end{cases} , \tag{2.1}$$

with $f : \mathbb{R}^2 \to \mathbb{R}$ and establish the conditions under which solutions exist. The natural language translation of the above ODE is: *given the instantaneous rate of change for $q$ and a starting value $q_0$ at $t = 0$ find $q(t)$ for later times $t > 0$.* We suspect that $f$ should be smooth in order for a solution to exist. If $f$ were discontinuous at some $q^*$ we would be hard put to choose the proper rate of change.

It could conceivably be anywhere from the left limit $f(t, q^* - 0)$ to the right limit $f(t, q^* + 0)$. In fact mere continuity of the function $f$ is not sufficient to establish a unique solution as shown by the following counter-example.

EXAMPLE 1. *Consider the IVP $q' = f(q) = 3q^{2/3}$, $q(t = 0) = 0$. Both $q(t) = 0$ and $q(t) = t^3$ are solutions of the problem, so the IVP does not specify a unique solution even though $f(q)$ is continuous.*

Looking at the counter-example we might presume that the non-uniqueness is somehow associated with $f$ not being differentiable at $q = 0$. Indeed, if we impose that $f$ is differentiable everywhere then a unique solution to the IVP (2.1) does exist. From ODE theory it is known that the necessary and sufficient condition is stronger than continuity but weaker than differentiability. Continuity in $q$ at $q^*$would require $|f(t, q) - f(t, q^*)| \to 0$ as $q \to q^*$ while differentiability would require that there exist a limit value of the ratio

$$\frac{f(t, q) - f(t, q^*)}{q - q^*} \tag{2.2}$$

as $q \to q^*$ which we denote by

$$f_q(t, q^*) = \frac{\partial f}{\partial q}(t, q^*) \ . \tag{2.3}$$

An intermediate condition would be that the ratio (2.2) be bounded over some neighborhood of $q^*$, $|q - q^*| < \varepsilon$ for all $t$. This is known as Lipschitz continuity.

DEFINITION 1. *The function $f(t, q)$ is **Lipschitz continuous** in $q$ over $S = \{(t, q) \mid 0 \le t \le T,\ q \in \mathbb{R}\}$ if there exists $L > 0$ such that*

$$|f(t, q_1) - f(t, q_2)| \le L\,|q_1 - q_2| \ . \tag{2.4}$$

This is equivalent with stating that the function difference is of the same order as the difference in the $q$ arguments: $|f(t, q_1) - f(t, q_2)| = O(|q_1 - q_2|)$ at any given time $t$. It is easy to see that if $f$ is differentiable and $f_q = \partial f/\partial q$ is bounded, we can take $L$ as the maximum value of the differential over the neighborhood of $q$

$$L = \max_q \ |f_q(t, q)| \ . \tag{2.5}$$

Whenever we can establish a single $L$ value that holds over the entire strip $S$ we say that $f$ is *uniformly Lipschitz continuous* over $S$. There is a natural extension to vector valued functions $\mathbf{f} : \mathbb{R}^{n+1} \to \mathbb{R}^n$ in which instead of the absolute value we use some norm $\|\|$, saying that $\mathbf{f}$ is Lipschitz continuous if there exists $L > 0$ such that

$$\|\mathbf{f}(t, \mathbf{q}_1) - \mathbf{f}(t, \mathbf{q}_2)\| \le L\,\|\mathbf{q}_1 - \mathbf{q}_2\| \tag{2.6}$$

over a strip $\mathbf{S} = \{(t, \mathbf{q}) \mid 0 \le t \le T,\ \mathbf{q} \in \mathbb{R}^n\}$

With this preparation we can state the basic theorem for existence of solutions to the IVP.

THEOREM 2. *If $f$ in (2.1) is Lipschitz continuous in $q$ over $S$ then there exists a unique solution $q(t)$ to (2.1) for some interval $t \in [0, \tau]$. The solution $q(t)$ is continuous and continuously differentiable. If $f$ is uniformly Lipschitz continuous over $S$ then the solution $q(t)$ exists over the entire interval $t \in [0, T]$.*

Furthermore, the solution to (2.1) depends continuously on the initial data. Suppose we have two IVP's which have the same $f$ but different initial conditions $q_0^1, q_0^2$ leading to two different solutions $q\left(t; q_0^1\right)$, $q\left(t; q_0^2\right)$. Under the same conditions as the above theorem we can show that

$$\left|q\left(t; q_0^1\right) - q\left(t; q_0^2\right)\right| \leq e^{Lt}\left|q_0^1 - q_0^2\right| \ . \tag{2.7}$$

This states that later differences in the function values remain bounded by $e^{Lt}$ times the difference in the initial conditions. We see the role played by the Lipschitz constant $L$. It describes how quickly the two different solutions could diverge from one another.

## 3. Finite difference approximations

We shall first concentrate on solving the IVP (3.1) using a finite difference approach. A discretization of the interval $[0, T]$ over which the solution is sought is introduced by

$$t^{(n)} = nk, \ n = 0, 1, \ldots, N, \ k = T/N, \tag{3.1}$$

with $k$ the step size. This is known as an *uniform discretization* of the interval $[0, T]$. Non-uniform discretizations are also widely used but we shall restrict our attention to uniform discretizations for now. The notation $Q^{(n)}$ shall be used to denote approximations of the exact value of the unknown function at $t^{(n)}$

$$Q^{(n)} \cong q(t^{(n)}) \ . \tag{3.2}$$

Superscripts shall be used to denote discretization in time with a view to using subscripts for spatial discretization later on. To simplify the notation the parentheses around superscripts shall be omitted as in $t^n$, $Q^n$, when there is no risk of confusing superscripts with exponentiation.

Simple finite difference approximations of the derivative can be deduced by geometric reasoning. Thus

$$\frac{Q^{n+1} - Q^n}{k}, \ \frac{Q^n - Q^{n-1}}{k}, \ \frac{Q^{n+1} - Q^{n-1}}{2k}$$

can be interpreted as approximations of the true slope of the function $q(t)$ at $t^n$, $q'(t^n)$ obtained by various secants passing through points on the graph of $q(t)$. Such geometric constructions are difficult to extend to arbitrary accuracy though. One analytical procedure for achieving arbitrary accuracy is the *method of undetermined coefficients* which makes use of Taylor series expansions of $q(t^n + mk)$.

EXAMPLE 2. *A fourth-order accurate finite difference approximation of $q'(t)$ can be obtained by*

$$\begin{cases} a\ q(t-2k) = a\left[q - 2kq' + \frac{(2k)^2}{2!}q'' - \frac{(2k)^3}{3!}q''' + \frac{(2k)^4}{4!}q^{(iv)} + O(k^5)\right] \\ b\ q(t-k) = b\left[q - kq' + \frac{k^2}{2!}q'' - \frac{k^3}{3!}q''' + \frac{k^4}{4!}q^{(iv)} + O(k^5)\right] \\ c\ q(t) = c\ q(t) \\ d\ q(t+k) = d\left[q + kq' + \frac{k^2}{2!}q'' + \frac{k^3}{3!}q''' + \frac{k^4}{4!}q^{(iv)} + O(k^5)\right] \\ e\ q(t+2k) = e\left[q + 2kq' + \frac{(2k)^2}{2!}q'' + \frac{(2k)^3}{3!}q''' + \frac{(2k)^4}{4!}q^{(iv)} + O(k^5)\right] \end{cases} . \tag{3.3}$$

*Suppressed arguments denote function evaluation at $t$. Adding the above equations gives*

(3.4)

$$aq(t-2k)+bq(t-k)+cq(t)+dq(t+k)+eq(t+2k) = (a+b+c+d+e)\,q+$$

(3.5)
$$k(-2a-b+d+2e)q'+$$

(3.6)
$$k^2\left(2a+\frac{b}{2}+\frac{d}{2}+2e\right)q''+$$

(3.7)
$$k^3\left(-\frac{4a}{3}-\frac{b}{6}+\frac{d}{6}+\frac{4e}{3}\right)q'''+$$

(3.8)
$$k^4\left(\frac{2a}{3}+\frac{b}{24}+\frac{d}{24}+\frac{2e}{3}\right)q^{(iv)}+$$

(3.9)
$$O(k^5)$$

*Setting the coefficient of $kq'$ to 1 and those of $q, q'', q''', q^{(iv)}$ to zero leads to the formula*

$$q'(t) = \frac{-q(t+2k)+8q(t+k)-8q(t-k)+q(t-2k)}{12k}+O(k^4) \tag{3.10}$$

**3.1. Finite difference operators.** Finite difference approximations of arbitrarily high order of accuracy can be obtained using operator methods. Let $E$ be the translation operator defined by

$$Eq(t) = q(t+k)\ . \tag{3.11}$$

Repeated applications of $E$ lead to

$$E^n q(t) = q(t+nk) \tag{3.12}$$

with $n$ any integer. $E^0$ is the identity operator

$$E^0 q(t) = Iq(t) = q(t) \tag{3.13}$$

Define the forward and backward difference operators by

$$\Delta_+ = E - E^0 \tag{3.14}$$

$$\Delta_- = E^0 - E^{-1}\ . \tag{3.15}$$

We can also define a central difference operator by

$$\delta = E^{1/2} - E^{-1/2} \tag{3.16}$$

whose action is given by

$$\delta q(t) = q(t+k/2) - q(t-k/2)\ . \tag{3.17}$$

Say we're given a set of values of the function $q$ at nodes $t^n = nk$, $n = 0, \ldots, N$ that we denote by $Q^n$. One way we can determine values of the derivative $q'$ would be to construct the interpolating polynomial passing through the points $(t^n, Q^n)$ and then differentiate the polynomial. The Newton form of the interpolating polynomial $p_N(t)$ is given by

$$p_N(t) = Q^0 + \left[Q^1, Q^0\right](t-t^0) + \left[Q^2, Q^1, Q^0\right](t-t^1)(t-t^0) + \ldots + \tag{3.18}$$

$$\left[Q^N, Q^{N-1}, \ldots, Q^0\right](t-t^{N-1})(t-t^{N-2})\ldots(t-t^0) \tag{3.19}$$

with the divided differences $\left[Q^n, Q^{n-1}, \ldots, Q^0\right]$ defined recursively by

$$\left[Q^1, Q^0\right] = \frac{Q^1 - Q^0}{t^1 - t^0} \tag{3.20}$$

$$\left[Q^n, Q^{n-1}, \ldots, Q^0\right] = \frac{\left[Q^n, Q^{n-1}, \ldots, Q^1\right] - \left[Q^{n-1}, Q^{n-2}, \ldots, Q^0\right]}{t^n - t^0} \tag{3.21}$$

One could just differentiate (3.18), replace $t$ by the point of interest and thus obtain a finite difference approximation to the derivative. The resulting formula is unwieldy though since it involves divided differences evaluated recursively. We can however express the divided differences in terms of the finite difference operators $\Delta_+$, $\Delta_-$. The first divided difference is

$$\left[Q^1, Q^0\right] = \frac{\Delta_+ q(t_0)}{k} = \frac{\Delta_+ Q^0}{k} \tag{3.22}$$

and one can verify by induction that

$$\left[Q^n, Q^{n-1}, \ldots, Q^0\right] = \frac{\Delta_+^n Q^0}{n!\, k^n} = \frac{\Delta_-^n Q^n}{n!\, k^n}\,. \tag{3.23}$$

Proof. Formula (3.23) is true for $n = 1$ by Eq. (3.22). Assuming the above statement to be true for $n$, we must show that

$$\left[Q^{n+1}, Q^n, \ldots, Q^0\right] = \frac{\Delta_+^{n+1} Q^0}{(n+1)!\, k^{n+1}} = \frac{\Delta_-^{n+1} Q^{n+1}}{(n+1)!\, k^{n+1}}\,. \tag{3.24}$$

By the recursive definition of divided differences we have

$$\left[Q^{n+1}, Q^n, \ldots, Q^0\right] = \frac{\left[Q^{n+1}, Q^n, \ldots, Q^1\right] - \left[Q^n, Q^{n-1}, \ldots, Q^0\right]}{t^{n+1} - t^0}\,. \tag{3.25}$$

Using (3.23) we have

$$\left[Q^{n+1}, Q^n, \ldots, Q^0\right] = \frac{\dfrac{\Delta_+^n Q^1}{n!\, k^n} - \dfrac{\Delta_+^n Q^0}{n!\, k^n}}{(n+1)k} = \frac{\dfrac{\Delta_-^n Q^{n+1}}{n!\, k^n} - \dfrac{\Delta_-^n Q^n}{n!\, k^n}}{(n+1)k} \tag{3.26}$$

or

$$\left[Q^{n+1}, Q^n, \ldots, Q^0\right] = \frac{\Delta_+^n}{(n+1)!k^{n+1}}\left(Q^1 - Q^0\right) = \frac{\Delta_+^{n+1} Q^0}{(n+1)!k^{n+1}} \tag{3.27}$$

$$= \frac{\Delta_-^n}{(n+1)!k^{n+1}}(Q^{n+1} - Q^0) = \frac{\Delta_-^n Q^{n+1}}{(n+1)!k^{n+1}}\,. \tag{3.28}$$

□

□

Using these and replacing $t$ by $t = t_0 + \alpha k$ the Newton interpolating polynomial becomes

$$p_N(t) = Q^0 + \alpha\Delta_+ Q^0 + \frac{\alpha(\alpha-1)}{2}\Delta_+^2 Q^0 + \ldots + C_\alpha^n \Delta_+^n Q^0 \tag{3.29}$$

with $\alpha \in [0, n]$. There is a corresponding formula using backward differences

$$p_N(t) = Q^N + \beta\Delta_- Q^N + \frac{\beta(\beta+1)}{2}\Delta_-^2 Q^N + \ldots + (-1)^n C_{-\beta}^n \Delta_-^n Q^N \tag{3.30}$$

with $t = t_n + \beta k$ and $\beta \in [-n, 0]$. The above formulas may be succintly written as

$$p_N(t) = \left(E^0 + \Delta_+\right)^\alpha Q^0 = \left(E^0 - \Delta_-\right)^{-\beta} Q^N \tag{3.31}$$

using the binomial expansion. Differentiation of the interpolating polynomial with respect to $t$ may be rewritten as

$$\frac{dp_N(t)}{dt} = \frac{dp_N(t^0+\alpha h)}{d\alpha}\frac{d\alpha}{dt} = \frac{1}{h}\left[\left\{\ln(E^0+\Delta_+)\right\}(E^0+\Delta_+)^\alpha\right]Q^0 \tag{3.32}$$

where the quantity between the brackets is an operator. But the first operator to act on $Q_0$ gives just the evaluation of the Newton interpolant

$$\left[(E^0+\Delta_+)^\alpha\right]Q^0 = p_N(t) \tag{3.33}$$

so

$$\frac{dp_N(t)}{dt} = \frac{1}{k}\left[\ln(E^0+\Delta_+)\right]p_N(t) \tag{3.34}$$

Identifying the operators on the two sides leads to the identity

$$\frac{d}{dt} = \frac{1}{k}\left[\ln(E^0+\Delta_+)\right] \tag{3.35}$$

Analogously, if we use backward differences we have

$$\frac{d}{dt} = -\frac{1}{k}\left[\ln(E^0-\Delta_-)\right] \tag{3.36}$$

The logarithm of an operator is defined by its series representation so we obtain

$$\frac{d}{dt} = \frac{1}{k}\left(\Delta_+ - \frac{1}{2}\Delta_+^2 + \frac{1}{3}\Delta_+^3 - \frac{1}{4}\Delta_+^4 + \dots\right) \tag{3.37}$$

$$= \frac{1}{k}\left(\Delta_- + \frac{1}{2}\Delta_-^2 + \frac{1}{3}\Delta_-^3 + \frac{1}{4}\Delta_-^4 + \dots\right) \tag{3.38}$$

Truncation of the operator series at term $n$ leads to a finite difference formula of $O(k^n)$. The procedure may be extended to higher order derivatives. The result for the second order derivative is

$$\frac{d^2}{dt^2} = \frac{1}{k^2}\left(\Delta_+^2 - \Delta_+^3 + \frac{11}{12}\Delta_+^4 - \dots\right) \tag{3.39}$$

$$= \frac{1}{k^2}\left(\Delta_-^2 + \Delta_-^3 + \frac{11}{12}\Delta_-^4 + \dots\right) \tag{3.40}$$

EXAMPLE 3. *A second order forward difference formula is obtained by truncating the series (3.37) to the first two terms*

$$\frac{dq(t)}{dt} \cong \frac{1}{k}\left(\Delta_+ - \frac{1}{2}\Delta_+^2\right)q(t) \tag{3.41}$$

$$= \frac{1}{k}\left\{q(t+k) - q(t) - \frac{q(t+2k) - 2q(t+k) + q(t)}{2}\right\} \tag{3.42}$$

$$= \frac{-q(t+2k) + 4q(t+k) - 3q(t)}{2k} \tag{3.43}$$

$$= q'(t) + O(k^2) \tag{3.44}$$

From the above left and right formulas we can also construct a centered finite difference representation of the derivative operator. Let $D$ denote the time differentiation operator, i.e. $D = d/dt$. From (3.35) and (3.36) we can write

$$\Delta_+ = e^{kD} - 1,\ \Delta_- = 1 - e^{-kD}\ . \tag{3.45}$$

The average of the forward and backward difference operators is

$$\frac{1}{2}\left(\Delta_+ + \Delta_-\right) = 2\sinh kD \ . \tag{3.46}$$

Note that the centered difference operator $\delta$ can be expressed as the average of the forward and backward operators for a half step size, so we obtain

$$\delta = 2\sinh\frac{kD}{2} \ , \tag{3.47}$$

from where

$$D = \frac{2}{k}\operatorname{arcsinh}\frac{\delta}{2} \ . \tag{3.48}$$

The Taylor series expansion of $\operatorname{arcsinh} x$ around $x = 0$ is

$$\operatorname{arcsinh} x = x - \frac{x^3}{6} + \frac{3x^5}{40} - \frac{5x^7}{112} + \frac{35x^9}{1152} - \tag{3.49}$$

so the first few terms from (3.48) are

$$D = \frac{1}{k}\left(\delta - \frac{1}{24}\delta^3 + \frac{3}{640}\delta^5 - \frac{5}{7168}\delta^7 + \frac{35}{294912}\delta^9 - \ldots\right). \tag{3.50}$$

Notice that the series only contains odd powers of $\delta$. We can show by induction that truncation of the series at $\delta^{2n-1}$ gives a formula of $O(k^{2n})$ accuracy.

## 4. Common finite difference methods

Let us now consider how numerical algorithms may be built to solve the IVP (2.1) using finite differences.

**4.1. Taylor series.** The basic task of a finite difference method that solves (2.1) numerically is to furnish a value at the next time step $q(t+k)$ in terms of known values at previous time steps. An immediate way of doing this is by use of a Taylor series expansion

$$q(t+k) = q + kq' + \frac{k^2}{2!}q'' + \frac{k^3}{3!}q''' + \ldots \tag{4.1}$$

To ease notation, function arguments not explicitly written out are considered to be $t$, i.e. $q'$ means $q'(t)$. The ODE specifies that $q' = f(t,q)$. From this we can compute higher derivatives

$$q'' = \frac{d}{dt}f(t,q(t)) = f_t + f_q q' = \tag{4.2}$$

$$q''' = \frac{d}{dt}q'' = \frac{d}{dt}\left(f_t + f_q f\right) = f_{tt} + f_{tq}f + \left(f_{qt} + f_{qq}f\right)f + f_q\left(f_t + f_q f\right) \tag{4.3}$$

Truncation of the Taylor series to various orders leads to the algorithms:

(1) $O(k)$, Euler's method

$$Q^{n+1} = Q^n + kf(Q^n) \tag{4.4}$$

(2) $O(k^2)$

$$Q^{n+1} = Q^n + kf(Q^n) + \frac{k^2}{2}\left[f_t(Q^n) + f_q(Q^n)f(Q^n)\right] \tag{4.5}$$

(3) $O(k^3)$

(4.6)
$$Q^{n+1} = Q^n + kf + \frac{k^2}{2}\left[f_t + f_q f\right] + \frac{k^3}{6}\left[f_{tt} + f_{tq}f + \left(f_{qt} + f_{qq}f\right)f + f_q\left(f_t + f_q f\right)\right]$$

with $f$ and its derivatives evaluated at $Q^n$.

Formulas of arbitrarily high order can be built up through this procedure but evaluation of all the derivatives of $f$ is tedious. In recent years the symbolic computational capabilities of computers has led to renewed interest in the Taylor series method.

**4.2. Runge-Kutta methods.** The Runge-Kutta class of methods aims to achieve arbitrary accuracy not by evaluating the function $f$ and its derivatives at $t$ but just by evaluations of the function $f$ at various points in the interval $[t, t+k]$. The general form of a Runge-Kutta method is

$$q(t+k) = q(t) + \sum_{l=0}^{s} \gamma_l f_l \tag{4.7}$$

where $f_l$ denotes evaluation of the function $f$ at

$$t^l = t + \alpha_l k \tag{4.8}$$

$$q^l = q(t) + k \sum_{m=0}^{s} \beta_{l,m} f_m \tag{4.9}$$

Essentially this amounts to replacing the Taylor series expansion (4.1) with a weighted average of the values of $f$ along the integration step. Two common Runge-Kutta methods should be known to all practitioners of numerical methods.

(1) A common two-stage Runge-Kutta method

$$K_1 = k\ f(t^n, Q^n) \tag{4.10}$$

$$K_2 = k\ f(t^n + k/2, Q^n + K_1/2) \tag{4.11}$$

$$Q^{n+1} = Q^n + K_2 \tag{4.12}$$

(2) A common four-stage Runge-Kutta method

$$K_1 = k\ f(t^n, Q^n) \tag{4.13}$$

$$K_2 = k\ f(t^n + k/2, Q^n + K_1/2) \tag{4.14}$$

$$K_3 = k\ f(t^n + k/2, Q^n + K_2/2) \tag{4.15}$$

$$K_4 = k\ f(t^n + k, Q^n + K_3) \tag{4.16}$$

$$Q^{n+1} = Q^n + \frac{1}{6}\left(K_1 + 2K_2 + 2K_3 + K_4\right) \tag{4.17}$$

**4.3. Linear multi-step methods.** Linear multi-step methods (LMM) carry out the weighted average idea one step further and postulate a relationship of the form

$$\sum_{j=0}^{r} a_j Q^{(n+j)} = k \sum_{j=0}^{r} b_j f(t^{(n+j)}, Q^{(n+j)}) \tag{4.18}$$

so we not only have a weighted average of the $f$ values but also of the function values $Q$. Note that $r$ starting values $\left\{Q^{(0)}, Q^{(1)}, \ldots, Q^{(r-1)}\right\}$ are needed to apply (4.18),

but only one initial condition $Q^{(0)} = q_0$ is apparent from the IVP. The additional starting values are determined by some other procedure, for example by applying a Runge-Kutta method to determine $\{Q^{(1)}, Q^{(2)}, \ldots, Q^{(r-1)}\}$. It is convenient to define a LMM by the polynomials

$$\rho(\zeta) = \sum_{j=0}^{r} a_j \zeta^j, \ \sigma(\zeta) = \sum_{j=0}^{r} b_j \zeta^j \ . \tag{4.19}$$

Some common LMM methods are:

(1) $r$-step Adams-Bashforth methods

$$\rho(\zeta) = \zeta^r - \zeta^{r-1}, \ \sigma(\zeta) = \tag{4.20}$$

## 5. Linear difference equations

Relations such as (4.18) arise frequently in both ODE and PDE methods and it is important to be able to analyze the induced behavior. Consider the finite difference equation

$$\sum_{j=0}^{r} a_j Q^{(n+j)} = 0 \ . \tag{5.1}$$

We wish to find solutions to this equation given $r$ initial conditions $Q^{(0)}, Q^{(1)}, \ldots, Q^{(r-1)}$, that is to find the sequence $\{Q^{(l)}\}_{l \geq r}$ . Parantheses around superscripts are useful here in order to more easily distinguish between exponents and superscripts. We try to find solutions of the form

$$Q^{(l)} = \zeta^l \tag{5.2}$$

where $\zeta$ is some non-zero constant ($\zeta = 0$ leads to uninteresting, trivial solutions). We see that $\zeta$ must satisfy the polynomial equation

$$\rho(\zeta) = \sum_{j=0}^{r} a_j \zeta^j = 0 \ . \tag{5.3}$$

called the *characteristic equation* of the linear difference equation (5.1). Similarly $\rho(\zeta)$ is called the *characteristic polynomial*. Let $\zeta_1, \zeta_2, \ldots, \zeta_r$ be the roots of the characteristic polynomial

$$\rho(\zeta_m) = 0, \ m = 1, 2, \ldots, r \ . \tag{5.4}$$

Since (5.1) is linear, a linear combination of the roots is also a solution. The general form of the solution is

$$Q^{(n)} = \sum_{j=0}^{r} c_j \zeta_j^n \ . \tag{5.5}$$

The constants $c_j$ are determined from the initial conditions

$$\sum_{j=0}^{r} c_j \zeta_j^l = Q^{(l)}, \ l = 0, 1, \ldots, r-1 \tag{5.6}$$

a linear system of $r$ equations and $r$ unknowns.

## 6. Analysis of Convergence

The procedures outlined above can be used to derive various finite difference algorithms suited to ODE's. There arises naturally the question of which algorithms are better than others or whether they work for all cases. We would like an algorithm to give an approximation to the exact solution of the problem and to get better results as we increase the resolution of the algorithm by decreasing the step size $k$. Suppose we are seeking the solution to (2.1) for $t \in [0, T]$ through an algorithm $\mathcal{A}$ that produces the approximations $\{Q^n \cong q(t^n)\}_{n=0,1,\ldots,N}$ $t^n = nk$, $k = T/N$. At each step we expect there to be an error $E^n = Q^n - q(t^n)$. If we use exact arithmetic then $E^0 = 0$ since we just apply the initial condition, i.e. $Q^0 = q_0 \implies E^0 = 0$. If we use computer arithmetic however there might be an unavoidable initial error, e.g. when $q_0$ is irrational.

DEFINITION 2. *An algorithm $\mathcal{A}$ to solve the IVP (2.1) is* ***convergent*** *if*

$$\lim_{\substack{k\to 0 \\ Nk=T}} E^n = 0 \tag{6.1}$$

This just states that the error should go to zero as we decrease the step size but keep the product of step size and number of steps constant, i.e. $nk = T$.

We now ask whether commonly used algorithms are convergent. Since it is easier to work on simple cases first we start with the simplest algorithms and problems first. We know from ODE theory that linear equations can be solved analytically in closed form and one would expect that any numerical algorithm should be able to approximate the solution to any desired degree of precision also. Therefore we first study algorithms that solve the linear IVP

$$\begin{cases} q' = \lambda q + \sigma \\ q(t=0) = q_0 \end{cases}, \tag{6.2}$$

where $\sigma(t)$ is a source function. We shall call (6.2) the *model problem*. In plain language (6.2) states that the change in $q$ is given by two effects. The first effect is proportional to $q$ with a constant of proportionality $\lambda$ while the second is specified by the source function $\sigma$. The analytical solution to this problem is given by

$$q(t) = e^{\lambda t} q_0 + \int_0^t e^{\lambda(t-\tau)} \sigma(\tau) d\tau \tag{6.3}$$

which is known as the *Duhamel principle*. Its interpretation is quite educational. The first term shows that the initial condition gets amplified by $e^{\lambda t}$. The function $e^{\lambda t}$ is known as the *propagator function* since it transfers the initial condition forward to time $t$. Source terms also get propagated forward in time by the same function but over a time interval reflecting the difference between the time $t$ and the time $\tau$ when the source was applied. There is a cumulation of the effect of source terms as shown by the integral above.

**6.1. Convergence of the forward Euler method for the model problem.** Having selected a suitably simple IVP, we now select a simple numerical method, namely Euler's method

$$Q^{n+1} = Q^n + k\ f(Q^n) \tag{6.4}$$

and study its convergence properties. Euler's method applied to the model problem gives

$$Q^{n+1} = (1 + k\lambda)\, Q^n + k\ \sigma^n. \tag{6.5}$$

We must now construct an expression for the error $E^n$. Equation (6.5) suggests that we try to establish a relation recurrence between $E^{n+1}$ and $E^n$. A Taylor series expansion of $q(t^{n+1})$ gives

$$q(t^{n+1}) = q(t^n) + kq'(t^n) + \frac{k^2 q''(\xi^n)}{2} \tag{6.6}$$

with $\xi^n \in (t^n, t^{n+1})$ and $q''$ assumed to be bounded. Since $q'(t^n) = f(t^n, q(t^n))$ we obtain

$$q(t^{n+1}) = (1 + k\lambda)q(t^n) + k\ \sigma^n + \frac{k^2 q''(\xi^n)}{2}\ . \tag{6.7}$$

Subtraction of (6.7) from (6.5) leads to the desired recurrence relation between errors

$$E^{n+1} = (1 + k\lambda)\, E^n - \frac{k^2 q''(\xi^n)}{2}\ . \tag{6.8}$$

This shows that the error at the previous time step gets amplified by $1 + k\lambda$ and a new error proportional to $k^2$ is introduced during this step of the algorithm. The additional error introduced during step $n$ of the algorithm shall be called the *one-step error*

$$\omega^n = \frac{k^2 q''(\xi^n)}{2}\ . \tag{6.9}$$

Note that Euler's method is obtained by truncating the forward finite difference series (3.37) to the first term. This approximation leads to an error in the approximation of the derivative at time step $n$ which shall be called *the truncation error*

$$\tau^n = \left(\frac{1}{k}\Delta_+ - \frac{d}{dt}\right) q(t^n) \tag{6.10}$$

$$= \frac{q(t^{n+1}) - q(t^n)}{k} - q'(t^n)\ . \tag{6.11}$$

Taylor series expansion gives

$$\tau^n = \frac{k\ q''(\xi^n)}{2} \tag{6.12}$$

and we note that $\omega^n = k\ \tau^n$, i.e. the one-step error is one order higher than the truncation error.

We can repeatedly apply the recurrence relation (6.8)

$$E^n = (1 + k\lambda)\, E^{n-1} - \omega^{n-1} \tag{6.13}$$

$$= (1 + k\lambda)\left[(1 + k\lambda)\, E^{n-2} - \omega^{n-2}\right] - \omega^{n-1} \tag{6.14}$$

$$= (1 + k\lambda)^2\left[(1 + k\lambda)\, E^{n-3} - \omega^{n-3}\right] - (1 + k\lambda)\,\omega^{n-2} - \omega^{n-1} \tag{6.15}$$

$$= \ldots \tag{6.16}$$

and arrive at

$$E^{(N)} = (1+k\lambda)^N E^{(0)} - \sum_{j=0}^{N-1} (1+k\lambda)^{N-1-j} \omega^{(j)} \ . \tag{6.17}$$

We explicitly show superscripts by parantheses to avoid confusion with the exponentiation operations also present. Relation (6.17) is also known as the *discrete Duhamel principle* because of the similarity with (6.3). The initial error is amplified by $(1+k\lambda)^N$ and the additional error $\omega^j$ introduced during subsequent time steps is amplified by $(1+k\lambda)^{N-1-j}$. We see that $(1+k\lambda)^N$ plays the role of $e^{\lambda T}$ and is hence called the discrete propagator. Note that the discrete propagator is a first order truncation of the corresponding continuous propagation operator as shown by

$$e^{\lambda T} = e^{\lambda k N} = \left(e^{\lambda k}\right)^N = \left(1 + k\lambda + \frac{k^2\lambda^2}{2} + \ldots\right)^N \cong (1+k\lambda)^N \ . \tag{6.18}$$

Assume that we're using exact arithmetic and therefore $E^0 = 0$. If we can show that $(1+k\lambda)^N$ remains bounded as $k \to 0$ and $kN = T$ then the first term in (6.17) would have a zero limit. Recall that $e$ can be expressed as a limit

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e \ . \tag{6.19}$$

The limit arising from the first term in (6.17) is

$$\lim_{\substack{k\to 0 \\ kN=T}} (1+k\lambda)^N = \lim_{N\to\infty} \left(1 + \frac{\lambda T}{N}\right)^N = \lim_{N\to\infty} \left[\left(1 + \frac{\lambda T}{N}\right)^{\frac{N}{\lambda T}}\right]^{\lambda T} = e^{\lambda T} \tag{6.20}$$

so the first term in (6.17) does indeed have a zero limit since it leads to the multiplication of a finite quantity $e^{\lambda T}$ by $E^{(0)} = 0$. The second term can be bounded by

$$\left|\sum_{j=0}^{N-1} (1+k\lambda)^{N-1-j} \omega^{(j)}\right| \le \sum_{j=0}^{N-1} |1+k\lambda|^{N-1-j} \left|\omega^{(j)}\right| \le \sum_{j=0}^{N-1} e^{k|\lambda|(N-1-j)} \left|\omega^{(j)}\right| \tag{6.21}$$

$$\le e^{k|\lambda|N} \sum_{j=0}^{N-1} \left|\omega^{(j)}\right| = e^{|\lambda|T} \sum_{j=0}^{N-1} \left|\omega^{(j)}\right| \ . \tag{6.22}$$

There are $N$ terms in the sum with $N = T/k$. Let $\|\omega\|_\infty$ be the largest of the $\left|\omega^{(j)}\right|$ encountered in the sum, this being the one-step error at some $\xi$

$$\|\omega\|_\infty = \frac{k^2 \ |f''(\xi)|}{2} \tag{6.23}$$

We then obtain

$$\left|\sum_{j=0}^{n-1} (1+k\lambda)^{n-1-j} \omega^{(j)}\right| \le e^{|\lambda|T} \ N \ \|\omega\|_\infty = e^{|\lambda|T} \ T \ \frac{k \ |q''(\xi)|}{2} \tag{6.24}$$

This last quantity goes to zero as $k \to 0$ since $e^{|\lambda|T}$, $q''$, $T$ are bounded. We have therefore established that

$$\lim_{\substack{k\to 0 \\ kN=T}} E^N = 0 \tag{6.25}$$

and that the forward Euler algorithm is convergent in exact arithmetic.

**6.2. The effect of inexact arithmetic and truncation error - stability.** The forward Euler method is convergent in exact arithmetic but this does not necessarily hold true when inexact, computer arithmetic is used. We have shown that the first term in (6.17) goes to zero if $E^0 = 0$

$$\lim_{\substack{k\to 0 \\ kN=T}} (1+k\lambda)^N E^{(0)} = 0 \tag{6.26}$$

but if $E^0 \neq 0$ then $(1+k\lambda)^N E^{(0)}$ could lead to an error bounded by $e^{|\lambda|T}E^0$. Clearly, just establishing convergence for exact arithmetic is not sufficient for practical purposes. We must establish conditions such that inherent computer arithmetic errors can be controlled. We say that algorithms that permit us to control error growth are *stable* while those in which errors grow without bound are *unstable*.

These intuitive definitions of stability must be made more precise. We take our cue from the continuous dependence of the solution to an IVP on the initial conditions. Suppose we could exactly compute the solution to the linear ODE $q' = \lambda q + \sigma$ with slightly different initial conditions $\tilde{q}_0 = q_0 + E^0$. Here $E^0 \neq 0$ is intended to represent any initial error due to inexact computer arithmetic. From (**??**) we know that

$$|q(t;q_0) - q(t;\tilde{q}_0)| \leq e^{Lt}\,|q_0 - \tilde{q}_0| \ . \tag{6.27}$$

For the model problem this result can be sharpened since we know the exact solution from (6.3), and we have

$$|q(t;q_0) - q(t;\tilde{q}_0)| = e^{\lambda t}E^0 \ . \tag{6.28}$$

This means that if an initial error $E^0$ is unavoidable, and all subsequent computations are done exactly, we can expect an error $E(T) = e^{\lambda T}E^0$ at time $T$. Clearly we cannot hope to better this using approximate computations, so we should recognize that algorithms that are capable of reproducing this sort of behavior are good candidates for further study. Such an algorithm should reproduce the error growth of the exact solutions in the limit of $k \to 0$ though it might have different behavior when $k \neq 0$. Note that this discussion in terms of propagation of an initial error is also applicable to later stages of the algorithm. We can envisage that the truncation error of the algorithm introduces a certain error at each step. We would like this error to be kept under control as the algorithm progresses.

Definition 3 (intuitive). *An algorithm $\mathcal{A}$ producing the approximations $\{Q^n\}_{n=0,1,\ldots}$ to the exact solution of $q' = f(q)$, $q(0) = q_0$ is said to be **zero-stable** if $|Q^n - q(t^n)|$ is bounded in the limit of $k \to 0$, $n = 1, 2, \ldots, N = T/k$.*

Definition 4 (exact). *Let $S$ be the vector with $r$ components of initial errors in an $r$-step algorithm that produce the approximations $\{Q^n\}_{n=r,r+1,\ldots}$ to the exact solution of the IVP $q' = f(q)$, $q(0) = q_0$*

$$S = \left[Q^0 - q(t^0), Q^1 - q(t^1), \ldots, Q^{r-1} - q(t^{r-1})\right] \ . \tag{6.29}$$

*The algorithm $\mathcal{A}$ is said to be* ***zero-stable*** *if for any $\varepsilon > 0$ there exist two positive constants $\delta, c$ such that when $\|S\|_\infty < \varepsilon$, $k < \delta$ we have*

$$\|Q^n - q(t^n)\|_\infty < c\ \varepsilon \tag{6.30}$$

*with $n = r, r+1, \ldots, N = T/k$.*

Note that in the above definition we require bounded errors at later times when two conditions are met:

(1) The initial errors are small, i.e. $\|S\|_\infty < \varepsilon$,
(2) The step size is small $k < \delta$.

By this definition the forward Euler method is zero-stable. It is a one step method requiring just one starting value $Q^0$ so $S$ is a scalar $S = E^0 = Q^0 - q_0$. Suppose we know that error in approximating the starting value (due to computer imprecise arithmetic) is less than $\varepsilon$, $\left|E^0\right| < \varepsilon$. By the argument from the previous section we know that

$$|E^n| \leq A\ \left|E^0\right| + B\ k \tag{6.31}$$

with $A = e^{|\lambda| t^n}$, $B = e^{|\lambda| t^n} t^n\, |q''(\xi)|\, /2$. Both $A$ and $B$ are positive finite quantities for any finite time $t^n$. We can write

$$|E^n| \leq A\ \left|E^0\right| + B\ k < A\varepsilon + B\delta = \left(A + B\frac{\delta}{\varepsilon}\right)\varepsilon \tag{6.32}$$

and choosing $\delta = \varepsilon$ and $c = A + B$ we verify that

$$|E^n| < c\varepsilon \tag{6.33}$$

so Euler's method is zero stable for the model problem. This means that even if an initial error is introduced, the algorithm produces an approximation in which the error has not grown faster than we would have expected from solving the exact equation with perturbed initial conditions provided we use a small enough step size $k < \delta$. The above is also a proof of convergence for Euler's method, i.e. for sufficiently small initial error $\left|E^0\right| < \varepsilon$ and sufficiently small step sizes $k < \delta$ the error at time step $n$ can be bounded by $|E^n| < c\varepsilon$ and therefore be made as small as we'd like. Thus zero-stability is a necessary and sufficient condition for convergence of Euler's method.

An immediate question that arises is the effect of using a finite step size $k > 0$. In the above proof of zero-stability for Euler's method we used $\delta = \varepsilon$ for any $\varepsilon > 0$. This essentially means that we considered that we can take as small a step size as we'd like in order to establish the bound (6.33). Practical considerations might dictate otherwise though since decreasing the step size increases the number of steps that must be computed. Of course even when using finite step sizes we'd still want to obtain suitable approximations, in which errors have not grown so much that they completely mask the true solution. Let's look again at (6.17) which we repeat here for some intermediate time $t^{(n)}$

$$E^{(n)} = (1 + k\lambda)^n\, E^{(0)} - \sum_{j=0}^{n-1} (1 + k\lambda)^{n-1-j} \omega^{(j)}\ . \tag{6.34}$$

When we take a large number of steps $n$ the initial error gets amplified by $(1+k\lambda)^n$ and that introduced in the previous step $j$ is amplified by $(1+k\lambda)^{n-1-j}$. An immediate condition that suggests itself is to impose

$$|1+k\lambda| \leq 1 \tag{6.35}$$

so that the error $E^{(n)}$ does not grow as $n$ increases. The quantity $A(z) = 1+z$, $z = k\lambda$ is known as the *amplification factor* for Euler's method, and (6.35) says that the absolute value of the amplification factor should be less than one to have a stable computation.

DEFINITION 5. *An algorithm $\mathcal{A}$ using step size $k$ whose amplification factor $A(z)$ satisfies $|A(z)| \leq 1$ for the model problem $q' = \lambda q$ is said to be* ***absolutely stable****. The region in $z = k\lambda$ over which $|A(z)| \leq 1$ is called the region of* ***absolute stability*** *of the algorithm.*

In the above definition the "absolute" epithet is suggested by the taking of the absolute value of $A(z)$. The inequality (6.35) defines the region of absolute stability for Euler's method. Within the region of absolute stability Euler's method is convergent in the sense that

$$\lim_{n\to\infty} E^n = 0 \tag{6.36}$$

which means that as we integrate towards ever larger times $t^n$ the error between the approximation and the exact solution $E^n = Q^n - q(t^n)$ goes to zero. Note the rather subtle difference between zero-stability and absolute stability for Euler's method. Zero-stability establishes that the error of Euler's method at any given, finite time $T$ that is attained after $N = T/k$ steps goes to zero as the step size $k$ goes to zero. Absolute stability shows that the asymptotic error $E^n$ at $t^n = nk$ goes to zero as $n \to \infty$ with $k$ finite.

**6.3. Convergence on the model problem for other algorithms.** From the study of Euler's algorithm we have been able to establish the concepts of convergence, zero-stability, absolute stability and amplification factor. For these to be useful they must be applicable to other algorithms. We have seen that zero-stability ensures convergence at any fixed time as the step size goes to zero for Euler's method. Similarily, absolute stability ensures asymptotic convergence as $t \to \infty$ for Euler's method. We would like to know whether this holds for all algorithms or whether additional conditions are required. Also, it would be useful to find a quicker way to establish convergence than the bounding, "$\delta - \varepsilon$" proofs used above. We now turn to these tasks, still concentrating on the model problem $q' = f(q) = \lambda q$, $q(0) = q_0$.

6.3.1. *Consistency.* Suppose that a LMM is proposed for the model problem which leads to the recurrence relation

$$\frac{1}{k}\sum_{j=0}^{r} a_j Q^{n+j} = \sum_{j=0}^{r} b_j f(Q^{n+j}) = \lambda \sum_{j=0}^{r} b_j Q^{n+j} \ . \tag{6.37}$$

in which some linear combination of values on the lhs is intended to evaluate an average of the derivative on the rhs. In operator form this can be written as

$$\tilde{D}Q^n = DQ^n \tag{6.38}$$

$$\tilde{D} = \frac{1}{k}\sum_{j=0}^{r} a_j E^j,\ D = \sum_{j=0}^{r} b_j f(E^j)\ . \tag{6.39}$$

Here we interpret $\tilde{D}$ as the truncation of some infinite series that would give as its exact sum the exact operator $D$. This relation is intended to approximate the initial ODE so we expect that

$$\frac{1}{k}\sum_{j=0}^{r} a_j q(t^{n+j}) \cong \sum_{j=0}^{r} b_j q'(t^{n+j}) \tag{6.40}$$

By analogy with the definition of truncation error for Euler's method we define the truncation error for the LMM (6.37) as

$$\tau^n = \left(\tilde{D} - D\right) q(t^n) \tag{6.41}$$

Taylor series expansion around $t^n$ leads to

$$\tau^n = \frac{1}{k}\left\{\sum_{j=0}^{r} a_j\right\} q(t^n) + \left\{\sum_{j=0}^{r} (ja_j - b_j)\right\} q'(t^n) + \dots \tag{6.42}$$

$$+ k^{p-1}\left\{\sum_{j=0}^{r}\left[\frac{j^p}{p!}a_j - \frac{j^{p-1}}{(p-1)!}b_j\right]\right\} q^{(p)}(t^n) + \dots \tag{6.43}$$

It is clear that for the method to converge a necessary condition is that

$$\lim_{k\to 0} \tau^n = 0. \tag{6.44}$$

Definition 6. *An algorithm is said to be consistent if its truncation error goes to zero as the step size goes to zero.*

Euler's method has a truncation error of

$$\tau^n = \frac{k\ q''(\xi)}{2} \tag{6.45}$$

and is consistent. A general LMM is consistent if

$$\sum_{j=0}^{r} a_j = 0,\ \sum_{j=0}^{r} (ja_j - b_j) = 0 \tag{6.46}$$

which may be concisely stated in terms of the characteristic polynomials $\rho(\zeta)$, $\sigma(\zeta)$ as

$$\rho(1) = 0,\ \rho'(1) - \sigma(1) = 0\ . \tag{6.47}$$

The definition of consistency can be applied to other classes of methods also. Taylor's series methods (4.4-4.6) are obviously consistent and consistency conditions are imposed in order to determine the coefficients in Runge-Kutta methods.

6.3.2. *Operational definitions of stability.* The true practical value of the characteristic polynomials is that they allow us to quickly establish whether a LMM is stable. Here are the principal results of the theory.

Definition 7. *The roots $\zeta_j$ of a polynomial are said to satisfy the* ***stability condition*** *if either $|\zeta_j| < 1$ or when $|\zeta_j| = 1$, $\zeta_j$ is a simple root.*

Proposition 1. *The LMM (6.37) is zero-stable if the roots of the lhs characteristic polynomial $\rho(\zeta)$ satisfy the stability condition.*

PROPOSITION 2. *The LMM (6.37) is absolutely stable if the roots of* $\pi(\zeta; z) = \rho(\zeta) - z\sigma(\zeta)$ *satisfy the stability condition. The region of absolute stability is the region of values of* $z = k\lambda$ *for which the roots of* $\pi(\zeta; z)$ *satisfy the stability condition.*

These results are proved using the properties of finite difference equations to establish "$\delta - \varepsilon$" proofs as was done above for Euler's method. The proofs shall be omitted here but are given by Henrici. There is an analogy which may be made with standard calculus. The definition of a derivative $f'(x_0)$ is

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \tag{6.48}$$

but we rarely use that definition in practice to compute the derivative of $\sin(\cos x)$ say. Rather we establish differentiation rules and tables to more quickly evaluate derivatives. Likewise the definitions of stability involve bounding arguments which are incovenient to redo for every new method. Finding the roots of characteristic polynomials is more straightforward.

6.3.3. *Stability of common LMM's.*

Forward Euler.

$$Q^{n+1} = Q^n + \lambda k Q^n \tag{6.49}$$

$$\rho(\zeta) = \zeta - 1,\ \sigma(\zeta) = 1 \tag{6.50}$$

$$\pi(\zeta; z) = \zeta - 1 - z \tag{6.51}$$

The root of $\rho(\zeta)$ is $\zeta_1 = 1$ which satisfies the stability condition since it is a simple root of absolute value 1. The method is zero stable. The root of $\pi(\zeta)$ is

$$\zeta = 1 + z \tag{6.52}$$

so the region of absolute stability is

$$|1 + k\lambda| \leq 1 \tag{6.53}$$

Backward Euler.

$$Q^{n+1} = Q^n + \lambda k Q^{n+1} \tag{6.54}$$

$$\rho(\zeta) = \zeta - 1,\ \sigma(\zeta) = \zeta \tag{6.55}$$

$$\pi(\zeta; z) = \zeta - 1 - z\zeta \tag{6.56}$$

The root of $\rho(\zeta)$ is $\zeta_1 = 1$ so the method is zero stable. The region of absolute stability is

$$|\zeta_1| = \frac{1}{|1 - z|} \leq 1 \Rightarrow |1 - k\lambda| \geq 1 \tag{6.57}$$

Trapezoidal method.

$$Q^{n+1} = Q^n + \frac{k\lambda}{2}\left(Q^n + Q^{n+1}\right) \tag{6.58}$$

$$\rho(\zeta) = \zeta - 1,\ \sigma(\zeta) = \frac{1}{2}(\zeta + 1) \tag{6.59}$$

The method is zero stable. The region of absolute stability is defined by

$$\left|\frac{2 + z}{2 - z}\right| \leq 1\ . \tag{6.60}$$

Leapfrog method.

$$Q^{n+2} = Q^n + 2k\lambda Q^{n+1} \tag{6.61}$$

$$\rho(\zeta) = \zeta^2 - 1,\ \sigma(\zeta) = 2\zeta \tag{6.62}$$

The roots of $\rho(\zeta)$ are $\zeta_1 = 1$, $\zeta_2 = -1$ so the method is zero-stable. The absolute stability region is defined by

$$\left| z \pm \sqrt{z^2 + 1} \right| \leq 1 \tag{6.63}$$

6.3.4. *Boundary locus method.* Regions of absolute stability defined by conditions such as (6.60) or (6.63) are difficult to analyze. A useful tool to simplify the analysis is the *boundary locus method*. We are interested in regions of the values of $z$ for which the roots of $\pi(\zeta; z)$ satisfy the stability condition. In general $z$ can take complex values. As $z$ varies a particular root of $\pi$ takes different values; the interesting event is when the absolute value of a root is unity. We can therefore consider the problem of determining the curve in the complex $z$-plane where one of the roots of $\pi(\zeta; z)$ has absolute value unity. If $|\zeta| = 1$ then we can write $\zeta = e^{i\theta}$ and

$$\pi(e^{i\theta}; z) = \rho(e^{i\theta}) - z\sigma(e^{i\theta}) = 0 \tag{6.64}$$

from whence

$$z(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}\ . \tag{6.65}$$

The curve defined by (6.65) is known as the *boundary locus*. In the boundary locus method we trace $z(\theta)$. This delimits the complex $z$-plane into regions. We then take an arbitrary point $z^*$ within each region and find the roots of $\pi(\zeta; z^*) = 0$. If the roots satisfy the stability condition then that region of the complex $z$-plane is a region of absolute stability; if not it is a region of absolute instability.

Example 4. *For the trapezoidal method we have*

$$z(\theta) = 2\frac{e^{i\theta} - 1}{e^{i\theta} + 1} = 2i \tan\frac{\theta}{2} \tag{6.66}$$

*so the boundary locus is the imaginary axis. We have two regions. To the left of the imaginary axis the root of* $\pi(\zeta; z = -1)$ *is* $\zeta_1 = 1/3$ *which satisfies the stability condition. To the right the root of* $\pi(\zeta; 1)$ *is* $\zeta_1 = 3$ *which does not satisfy the stability condition. We conclude that* $\operatorname{Re} z \leq 0$ *is the region of absolute stability for the trapezoidal method.*

Example 5. *For the leapfrog method we have*

$$z(\theta) = \frac{e^{2i\theta} - 1}{e^{i\theta}} = 2i \sin\theta \tag{6.67}$$

*so the boudary locus is the segment from* $-i$ *to* $i$ *along the imaginary axis. There are two regions. One is outside of this segment. The roots of* $\pi(\zeta; z = 1)$ *are* $\zeta_{1,2} = 1 \pm \sqrt{2}$ *that do not satisfy the stability condition. The other region is the* $-i$ *to* $i$ *segment. The roots of* $\pi(\zeta; z = 0)$ *are* $\zeta_{1,2} = \pm 1$ *which satisfy the stability condition. The region of absolute stability for the leapfrog method is the segment from* $-i$ *to* $i$ *along the imaginary axis.*

6.3.5. *Conditions for convergence.* With the above preparation we are able to state the main result of the theory.

THEOREM 3. *A LMM is convergent if it is stable and consistent.*

Note that the general characterization "stable" is used in the above theorem. We can use any of the definitions of stability introduced above; we obtain different types of convergence. When "zero-stability" is used we obtain that the LMM converges to the solution of the IVP at a fixed time $T$ as $k \to 0$ but $nk = T$. When "absolute stability" is used we obtain asymptotic convergence as $n \to \infty$.

CHAPTER 4

# Fourier Analysis of Common Linear Partial Differential Equations

## 1. Fourier Series

Fourier transform techniques are useful in the study of PDE's in many ways. Linear equations can often be directly solved by Fourier transforms. Such solutions are useful in their own right and also as test cases for validation of numerical algorithms intended for more complicated, non-linear problems. Thinking about the properties of an analytical or numerical solution both in real space and in Fourier space brings many insights that guide algorithmic development and analysis. Here we'll briefly go over the basic results from Fourier analysis that are especially useful in numerical work. Most results shall be presented in brief. Further detail will be presented when spectral methods are studied later.

Consider a real function $f$ defined on $[-L, L]$. The function may be represented by a trigonometric or Fourier series

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos \xi_k x + b_k \sin \xi_k x) \tag{1.1}$$

with $\xi_k = k\pi/L$. Note that the representation (1.1) implicitly prolongs $f$ by periodicity outside of the original definition interval $[-L, L]$. Multiplying (1.1) by $\cos \xi_k x$, $\sin \xi_k x$ and integrating the result over $[-L, L]$ gives the Fourier coefficients

$$a_k = \frac{1}{L} \int_{-L}^{L} f(y) \cos \xi_k \ y \ dy, \ b_k = \frac{1}{L} \int_{-L}^{L} f(y) \sin \xi_k \ y \ dy \ . \tag{1.2}$$

Replacing the coefficients (1.2) back into (1.1) leads to

$$f(x) = \frac{1}{2L} \int_{-L}^{L} f(y) \ dy + \sum_{k=1}^{\infty} \frac{1}{L} \int_{-L}^{L} f(y) \cos \xi_k (y - x) dy \tag{1.3}$$

from which the Fourier transform can be defined in the limit $L \to \infty$.

Fourier series are useful in determining solutions to linear PDE's in conjunction with separation of variables.

Example 6. *Determine the temperature distribution in a bar given the initial temperatures $f(x)$ and the endpoint temperatures at all times $t$, $g_0(t)$, $g_1(t)$. The problem is modeled by the heat equation with initial and boundary conditions*

$$\begin{cases} q_t = q_{xx} \\ q(x, t = 0) = f(x) \\ q(x = 0, t) = g_0(t), \ q(x = 1, t) = g_1(t) \end{cases} . \tag{1.4}$$

*Physical units have been chosen so that the bar length and diffusion coefficient are equal to one. We can solve the problem by separation of variables by assuming that*

$$q(x,t) = X(x)T(t) \tag{1.5}$$

*which leads to*

$$\frac{T'}{T} = \frac{X''}{X} \ . \tag{1.6}$$

*The lhs depends only on $t$, the rhs only $x$. The relation has to be true for all $x, t$ which are independent variables so the only possiblity is that both $T'/T$ and $X''/X$ are constant.*

$$\frac{T'}{T} = \frac{X''}{X} = C \ . \tag{1.7}$$

*Solving for $T$ leads to $T(t) = T_0 e^{Ct}$. Physical reasoning leads to $C < 0$ since the temperature in the bar cannot increase without bound. Let $C = -a^2$. Solving for $X$ then leads to $X(x) = A_a \cos ax + B_a \sin ax$. Note that the heat equation is satisfied for any $a$. Since the equation is linear, any linear combination of solutions is also a solution, so the most general form of the solution is*

$$q(x,t) = \sum_a \left(A'_a \cos ax + B'_a \sin ax\right) e^{-a^2 t} \tag{1.8}$$

*with $A'_a = T_0 A_a$, $B'_a = T_0 B_a$ and the sum being taken over all possible values of $a$. The specific values of the coefficients shall be determined by the initial and boundary conditions*

$$q(x,0) = f(x) = \sum_a \left(A'_a \cos ax + B'_a \sin ax\right) \tag{1.9}$$

$$q(0,t) = g_0(t) = \sum_a A'_a e^{-a^2 t}, \ q(1,t) = g_1(t) = \sum_a \left(A'_a \cos a + B'_a \sin a\right) e^{-a^2 t} \tag{1.10}$$

*The details are problem dependent. If a single series is insufficient to represent all initial and boundary conditions the original problem may be separated into two parts*

$$\left\{ \begin{array}{c} q^{(1)}_t = q^{(1)}_{xx} \\ q^{(1)}(x,0) = f(x) \\ q^{(1)}(0,t) = 0, \ q^{(1)}(1,t) = 0 \end{array} \right. \quad \left\{ \begin{array}{c} q^{(2)}_t = q^{(2)}_{xx} \\ q^{(2)}(x,0) = 0 \\ q^{(2)}(0,t) = g_0(t), \ q^{(2)}(1,t) = g_1(t) \end{array} \right. \tag{1.11}$$

*The first problem has simple boundary conditions and the true initial condition and the second has a simple initial condition and the true boundary conditions. The solution to the initial problem is then given by $q = q^{(1)} + q^{(2)}$.*

## 2. Fourier Transform

Taking the $L \to \infty$ limit of (1.3) leads to

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} f(y) \cos \xi (y - x) \ dy \tag{2.1}$$

We can write $\cos \xi(y-x) = \left(\exp\left[i\xi(y-x)\right] + \exp\left[-i\xi(y-x)\right]\right)/2$ and obtain

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} f(y) e^{i\xi(x-y)} \ dy \tag{2.2}$$

whenever the integration operations above are well defined (e.g. when $f$ is absolutely integrable on $\mathbb{R}$). After the $y$ integration a function of $\xi, x$ is obtained which is again integrated to give $f(x)$. This suggests looking at the functions obtained in the intermediate step

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(y) e^{-i\xi y} dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\xi x} dx \ . \tag{2.3}$$

This is called the *Fourier transform* of $f$. We also have from (2.2)

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{i\xi x} d\xi \ . \tag{2.4}$$

We say that $f$ is the *inverse Fourier transform* of $\hat{f}$. The direct and inverse Fourier transforms are integral operators usually denoted by $\mathcal{F}$, $\mathcal{F}^{-1}$ respectively

$$\hat{f} = \mathcal{F}\ f,\ f = \mathcal{F}^{-1}\ \hat{f}\ . \tag{2.5}$$

The transformation is linear

$$f = a\ g + b\ h,\ \hat{f} = a\ \mathcal{F}g + b\ \mathcal{F}h \tag{2.6}$$

with $a, b$ scalars.

An important class of functions for which the Fourier transform is well defined is the $L^2$ functions $f$ for which

$$\|f\|_2 = \int_{-\infty}^{\infty} |f(x)|^2 \, dx \tag{2.7}$$

is finite. An important property of the Fourier transform is that it preserves the $L^2$ norm of a function, i.e.

$$\|f\|_2 = \left\|\hat{f}\right\|_2 \ . \tag{2.8}$$

An illuminating analogy is with vectors $v$ in $\mathbb{R}^n$. Linear transformations of vetors in $\mathbb{R}^n$ are given by matrices $A$. Most matrices do not preserve the norm of a vector, i.e. in general $\|v\| \neq \|Av\|$ . But some special matrices do preserve the norm; examples are rotation and reflection matrices. These basically allow us to look at vectors from various angles and in some orientations the vector becomes especially simple. For example by a succession of rotations we can transform a vector to be parallel to one of the axes in $\mathbb{R}^n$. In an analogous manner, Fourier transforms also allow us to look at functions from various viewpoints from which further insight is possible. We often speak of looking at a function in *real space* or in *wavenumber* or *Fourier space* referring to $f$ or $\hat{f}$, respectively.

An immediate benefit is when we consider PDE's. Note that the Fourier transform is a linear combination of function values amplified by $e^{i\xi x}$. The reason why this particular weighting is useful for PDE's is that it is an eigenfunction of the differentiation operator

$$\partial_x e^{i\xi x} = i\xi e^{i\xi x} \ . \tag{2.9}$$

We can read the above relation as saying that of all the functions we might conceive, for one particular function $e^{i\xi x}$ the differentiation operator reduces to a simple multiplication with a scalar. The benefit of using Fourier transform techniques is that the reduction of differentiation to multiplication carries over to all functions

when we look at them in Fourier space. Taking the differential of the Fourier representation of $f$ we obtain

$$f_x = \partial_x f = \frac{1}{\sqrt{2\pi}} \frac{\partial}{\partial x} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{i\xi x} d\xi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (i\xi)\, \hat{f}(\xi) e^{i\xi x} d\xi \tag{2.10}$$

From this we can deduce that the Fourier transform of $f_x$ is given by

$$\hat{f}_x = i\xi\ \hat{f}. \tag{2.11}$$

## 3. Fourier solution of common linear PDE's

Using the properties of the Fourier transform it is easy to solve linear PDE's. Let us do this for our basic model problems.

**3.1. Advection equation.** Taking the Fourier transform with respect to $x$ of

$$q_t + u q_x = 0 \tag{3.1}$$

gives

$$\hat{q}_t + i\xi u \hat{q} = 0 \tag{3.2}$$

which is an ODE in $t$ for $\hat{q}$. The solution is

$$\hat{q}(\xi, t) = \hat{q}(\xi, 0) e^{-i\xi u t} \tag{3.3}$$

which immediately gives $q(x, t0$

$$q(x,t) = \mathcal{F}^{-1}\ \hat{q} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{q}(\xi, t) e^{i\xi x} d\xi \tag{3.4}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{q}(\xi, 0) e^{i\xi(x-ut)} d\xi \tag{3.5}$$

The initial condition $q(x, t = 0) = f(x)$ gives $\hat{q}(\xi, 0) = \mathcal{F}\ f = \hat{f}$. If we write the Fourier representation of $f$ at $x - ut$ we obtain

$$f(x - ut) = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{i\xi(x-ut)} d\xi \tag{3.6}$$

so the solution to the advection equation is

$$q(x, t) = f(x - ut)\ . \tag{3.7}$$

**3.2. Heat equation.** We apply the same procedure to

$$q_t = q_{xx} \tag{3.8}$$

to obtain

$$\hat{q}_t = -\xi^2 \hat{q} \tag{3.9}$$

which can be integrated to give

$$\hat{q}(\xi, t) = \hat{q}(\xi, 0) e^{-\xi^2 t}\ . \tag{3.10}$$

Using this Fourier transform we obtain

$$q(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{q}(\xi, 0) e^{-\xi^2 t} e^{i\xi x} d\xi \tag{3.11}$$

It is possible to explicitly evaluate this integral, but we focus instead on interpretation of the behavior of the solution. The formula shows that as time $t$ increases

each Fourier mode decays exponentially. The rate of decay is proportional to the square of the wavenumber. Hence, rapid variations in the initial conditions are more quickly attenuated by the heat equation.

## 4. Von Neumann stability analysis

An important application of Fourier analysis to numerical solutions of PDE's is the ability to quickly determine stability criteria for finite difference schemes. After a finite difference discretization one works with the values of the unknown at grid points at various time levels. Consider the one-dimensional case with a the grid points $x_j = jh$, $h = 1/(M+1)$. Outside of the definition domain $[0,1]$ the grid function is prolonged by periodicity. We shall call the values $\left\{Q_j^n\right\}_{j=0,1,\ldots M}$ a *grid function*. The $\left\{Q_j^n\right\}$ can be represented by a Fourier integral

$$Q_j = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \hat{Q}^n(\xi)\, e^{i\xi jh} d\xi \tag{4.1}$$

Note that the wavenumber integral does not go over all possible values, but is restricted to the wavenumbers that can be resolved by the grid. The Fourier transform of the grid function is given by

$$\hat{Q}^n(\xi) = \frac{h}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} Q_j^n e^{-i\xi jh} \ . \tag{4.2}$$

Parseval's relation

$$\left\|\left\{Q_j^n\right\}\right\|_2 = \left\|\hat{Q}^n(\xi)\right\| \tag{4.3}$$

assures us that we can look at the magnitude of the grid function either in Fourier space or in real space. It is typically quite easy to determine a relation between the coefficients $\hat{Q}^n$ in Fourier space for finite difference discretizations of linear PDE's. Once this is done we can compute the amplification ratio

$$G(\xi) = \frac{\hat{Q}^{n+1}(\xi)}{\hat{Q}^n(\xi)} \ . \tag{4.4}$$

This ratio shows the growth of various solution components including the truncation or arithmetic errors. We want these to be kept under control so we shall impose

$$|G(\xi)| \leq 1 \tag{4.5}$$

CHAPTER 5

# Finite difference methods for the heat equation

## 1. One space dimension

**1.1. Semi-discretized system.** We start our analysis of numerical methods for PDE's with finite difference methods for the heat equation. The heat equation defined on the entire real line only requires initial conditions. Typically we solve the heat equation on a finite domain with boundary conditions that can be of Dirichlet, Neumann, or mixed type. We shall choose a model problem with Dirichlet boundary conditions

$$\left\{ \begin{array}{c} q_t = q_{xx} \\ q(x, t = 0) = f(x) \\ q(x = 0, t) = g_0(t),\ q(x = 2\pi, t) = g_1(t) \end{array} \right. . \tag{1.1}$$

Physical units have been chosen to obtain the simple form shown above. A time honored scientific method is to reduce a complicated unknown problem to one whose solution we already know. We therefore ask whether it is possible to reduce (1.1) to a system of ODE's. The idea is to carry out the discretization of just one of the differential operators. We choose to do this for the spatial derivatives and we shall approximate $\partial_x^2$ by a finite difference expression.

First we define a computational grid $x_j = jh$, $h = 2\pi/(M+1)$, $t^n = nk$ with step size $h, k$ in space and time. Define $Q_j(t)$ to be the restriction of $q(x,t)$ to $x = x_j$

$$Q_j(t) = q(x_j, t),\ j = 0, 1, \ldots, M+1 \ . \tag{1.2}$$

The indices $j = 1, 2, \ldots, M$ correspond to points in the interior of the computation domain. The indices $j = 0$ and $j = M+1$ correspond to points on the boundary of the computation domain where we can apply the Dirichlet boundary conditions to get

$$Q_0(t) = g_0(t),\ Q_{M+1}(t) = g_1(t) \ . \tag{1.3}$$

There are many possible finite difference approximations of $\partial_{xx}$. To show the general issues involved consider a simple centered approximation which is second order accurate

$$q_{xx}(x_j, t) \cong \frac{\delta_x^2}{h^2} Q_j(t) = \frac{Q_{j+1}(t) - 2Q_j(t) + Q_{j-1}(t)}{h^2} \ . \tag{1.4}$$

Using this in the heat equation at $x = x_j$ gives an ODE

$$\frac{dQ_j(t)}{dt} = \frac{Q_{j+1}(t) - 2Q_j(t) + Q_{j-1}(t)}{h^2} \ . \tag{1.5}$$

If we consider all interior points we obtain a system of ODE's which can be written in matrix form as

$$\frac{d}{dt}\mathbf{Q} = \frac{1}{h^2}\left(A\mathbf{Q} + \mathbf{b}\right) \tag{1.6}$$

with

$$\mathbf{Q} = \begin{bmatrix} Q_1 & Q_2 & \cdots & Q_M \end{bmatrix}^T,\ \mathbf{b} = \begin{bmatrix} Q_0 & 0 & \cdots & Q_{M+1} \end{bmatrix}^T . \tag{1.7}$$

$$A = \begin{matrix} -2 & 1 & 0 & \ddots & 0 & 0 \\ 1 & -2 & 1 & 0 & \ddots & 0 \\ 0 & 1 & -2 & 1 & 0 & \ddots \\ \ddots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & 0 & 1 & -2 & 1 \\ 0 & 0 & \ddots & 0 & 1 & -2 \end{matrix} \tag{1.8}$$

We have carried our program of reducing the initial PDE to a system of ODE's. This method is also known as the *method of lines*, since we're solving the initial PDE on the $x = x_j$ lines. Of course there is an approximation involved. We have used a finite spatial step size $h$ to obtain a finite sized ODE system whereas the initial PDE is equivalent to an infinite system of ODE's. We can now apply what we know from ODE's to solve (1.6). The boundary condition term does not play an essential role in the analysis of algorithms so we'll consider the special case $g_0(t) = g_1(t) = 0$ which leads to $\mathbf{b} = 0$.

1.1.1. *Forward Euler method.* Using the forward Euler method leads to

$$\mathbf{Q}^{n+1} = \mathbf{Q}^n + \frac{k}{h^2}A\mathbf{Q}^n = \left(I + \frac{k}{h^2}A\right)\mathbf{Q}^n . \tag{1.9}$$

A typical component from this equation reads

$$Q_j^{n+1} = Q_j^n + \sigma\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) \tag{1.10}$$

with $\sigma = k/h^2$. The same procedure could have been obtained by a Forward in Time, Centered in Space finite difference approximation of both derivatives appearing in the heat equation hence the name FTCS scheme. It is of $O(k, h^2)$, i.e. first order in time, second order in space. We need to choose a time step size $k$ that will ensure stability of the method. We're looking at finite step sizes so the appropriate concept to use is absolute stability. From Euler's method for systems we know that the region of stability is

$$|1 + z| \leq 1 \tag{1.11}$$

where $z = k\lambda$ and $\lambda$ is an eigenvalue of $\left(1/h^2\right)A$. The eigenvalues of an arbitrary matrix are difficult to determine in general. For the particular matrix here however we can use the fruitful analogy between discrete and continuum operators. The matrix $A$ is an approximation of the second derivative operator $\partial_x^2$. The eigenvalues and eigenfunctions of the continuum operator are given by

$$\partial_x^2 e^{i\xi x} = -\xi^2 e^{i\xi x} . \tag{1.12}$$

We suspect therefore that the eigenvectors of $A$ are discretizations of the continuum eigenfunction $e^{i\xi x}$. We therefore try a discretization of the eigenfunction $e^{i\xi x}$

$$\mathbf{W}_p = \left[ \begin{array}{ccccccc} e^{iph} & e^{ip2h} & \cdots & e^{ipjh} & \cdots & e^{ipMh} \end{array} \right] \tag{1.13}$$

Here $p$ is a discrete wavenumber (an integer in this case because we've chosen the definition domain as $[0, 2\pi]$). We compute $A\mathbf{W}_p$ hoping to obtain an eigenrelationship. The $j^{th}$ component of the result is

$$\left(A\mathbf{W}_p\right)_j = e^{ip(j-1)h} - 2e^{ipjh} + e^{ip(j+1)h} = 2\left[\cos ph - 1\right]\ e^{ipjh} \tag{1.14}$$

and we see that we obtain an unique scalar for an arbitrary component $j$. Our initial guess was correct and the eigenvalues of $(1/h^2)A$ are

$$\lambda_p = -\left(\frac{2}{h}\sin\frac{ph}{2}\right)^2 . \tag{1.15}$$

The stability region is given by

$$-1 \leq 1 - k\left(\frac{2}{h}\sin\frac{ph}{2}\right)^2 \leq 1 \tag{1.16}$$

which reduces to

$$k \leq \frac{h^2}{2}\frac{1}{\sin^2(ph/2)} \tag{1.17}$$

The most restrictive condition arises for high values of $p$, the fast Fourier modes. A condition that includes all possbile $p$ values is

$$k \leq \frac{h^2}{2} \tag{1.18}$$

This condition is quite restrictive in practice since a halving of the spatial step size (to get better accuracy) imposes a time step four times smaller.

Let us derive stability bounds using Von Neumann analysis. We replace the grid function values in (1.10) with their Fourier representation to obtain

$$\int_{-\pi/h}^{\pi/h} \hat{Q}^{n+1}(\xi)e^{ijh\xi}d\xi = \int_{-\pi/h}^{\pi/h} \hat{Q}^{n}(\xi)e^{ijh\xi}d\xi + \sigma\left(\int_{-\pi/h}^{\pi/h} \hat{Q}^{n}(\xi)e^{i(j+1)h\xi}d\xi - \right. \tag{1.19}$$

$$\left. 2\int_{-\pi/h}^{\pi/h} \hat{Q}^{n}(\xi)e^{ijh\xi}d\xi + \int_{-\pi/h}^{\pi/h} \hat{Q}^{n}(\xi)e^{i(j-1)h\xi}d\xi\right) \tag{1.20}$$

Grouping together terms this can be rewritten as

$$\int_{-\pi/h}^{\pi/h} \left[G(\xi) - 1 - 2\sigma\left(\cos h\xi - 1\right)\right]\hat{Q}^{n}(\xi)e^{ijh\xi}d\xi = 0\ . \tag{1.21}$$

Since the relation has to be true for all $\xi$ the coefficient of the exponential functions (which form a basis) must be zero. The only interesting possibility is

$$G = 1 - 4\frac{k}{h^2}\sin^2\frac{h\xi}{2} \tag{1.22}$$

Imposing the condition $|G| \leq 1$ leads to the same stability criterion as above, $k \leq h^2/2$. This application of von Neumann analysis was carried out in full detail. Typically we know that we shall arrive at the stage where a relation between Fourier coefficients is obtained and a number of intermediate steps can be short-circuited.

1.1.2. *Trapezoidal method.* Applying the trapezoidal method to (**??**) leads to the following relation between components

$$Q_j^{n+1} = Q_j^n + \frac{\sigma}{2}\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) + \frac{\sigma}{2}\left(Q_{j+1}^{n+1} - 2Q_j^{n+1} + Q_{j-1}^{n+1}\right) \ . \tag{1.23}$$

The method is known as the *Crank-Nicolson scheme*. This is now an implicit relation which requires the solution of a tridiagonal system at each time step. This is more work than the explicit FTCS method but not prohibitively so. If the method allows large step sizes the overall work for attaining a specific precision could be less than for FTCS. The stability criterion for the trapezoidal method is

$$|1 - z| \geq 1 \tag{1.24}$$

which leads to

$$\left|1 + k\left(\frac{2}{h}\sin\frac{ph}{2}\right)^2\right| \geq 1 \ . \tag{1.25}$$

This relation is satisfied for all $k > 0$ so the Crank-Nicolson method is unconditionally stable. This means that the only restriction on step size comes from the accuracy that we wish to attain.

Von Neumann stability analysis leads to

$$G = 1 + \sigma\left(\cos\xi h - 1\right) + G\sigma\left(\cos\xi h - 1\right) \tag{1.26}$$

or

$$G = \frac{1 - 2\sin^2\frac{\xi h}{2}}{1 + 2\sin^2\frac{\xi h}{2}} \tag{1.27}$$

from which we see that $|G| \leq 1$ always so we again obtain that the Crank-Nicolson scheme is unconditionally stable.

## 2. Two space dimensions

A typical 2D (two spatial dimensions) heat equation problem is to solve the IVBP

$$\begin{cases} q_t = q_{xx} + q_{yy}, \text{ for } (x,y) \in D,\ T > t > 0 \\ q(x,y,t=0) = f(x,y) \text{ for } (x,y) \in D \\ q(x,y,t) = g(x,y,t) \text{ for } (x,y) \in \partial D \end{cases} \ . \tag{2.1}$$

where $D$ is a region of the $(x,y)$ plane and $\partial D$ is its boundary. We shall first consider the simple case of a square region $D = [0,2\pi]\times[0,2\pi]$. Dirichlet boundary conditions are given here; we shall consider Neumann and mixed boundary conditions later. As in the 1D case we introduce an uniform discretization of the computational domain

$$x_j = jh,\ y_l = lh,\ t^n = nk \tag{2.2}$$

with $h = 2\pi/(M+1)$ the space step size and $k = T/N$ the time step. The exact value of the unknown function at $x_j, y_l, t^n$ is approximated by

$$Q_{jl}^n \cong q(x_j, y_l, t^n) \ . \tag{2.3}$$

Again the problem can be reduced to that of integrating a system of ODE's if we use the method of lines and introduce the functions

$$Q_{jl}(t) = q(x_j, y_l, t) \ , \tag{2.4}$$

and some procedure for approximating the spatial derivatives with finite differences. The second-order accurate approximation of the Laplace operator is the natural first thing to try

$$\nabla^2 q(x_j, y_l, t) = q_{xx}(x_j, y_l, t) + q_{yy}(x_j, y_l, t) \cong \nabla_h^2 Q_{j,l}(t) \tag{2.5}$$

$$\frac{1}{h^2}\left[Q_{j+1,l}(t) + Q_{j-1,l}(t) + Q_{j,l+1}(t) + Q_{j,l-1}(t) - 4Q_{j,l}(t)\right] \tag{2.6}$$

and this leads to an ODE system similar in form to that from the 1D case

$$\frac{d}{dt}\mathbf{Q} = \frac{1}{h^2}\left(A\mathbf{Q} + \mathbf{b}\right) \tag{2.7}$$

but with different definitions of $\mathbf{Q}$,$A, \mathbf{b}$

(2.8)

$$\mathbf{Q} = \begin{bmatrix} Q_{11} & Q_{1M} & \cdots & Q_{21} & Q_{22} & \cdots & Q_{MM} \end{bmatrix}^T,\ \mathbf{b} = \begin{bmatrix} Q_{01}+Q_{10} & Q_{02} & \cdots & Q_{M,M+1}+Q_{M+1,M} \end{bmatrix}^T.$$

$$A = \begin{bmatrix} -4 & 1 & 0 & \ddots & 1 & \ddots & 0 \\ 1 & -4 & 1 & 0 & \ddots & 1 & \ddots \\ 0 & 1 & -4 & 1 & 0 & \ddots & 1 \\ \ddots & 0 & \ddots & \ddots & \ddots & 0 & \ddots \\ 1 & \ddots & 0 & 1 & -4 & 1 & 0 \\ \ddots & 1 & \ddots & 0 & 1 & -4 & 1 \\ 0 & \ddots & 1 & \ddots & 0 & 1 & -4 \end{bmatrix} \tag{2.9}$$

We now have system matrix which is pentadiagonal, arising from the fact that the 2D field $Q_{jl}(t)$ is rendered as a one-dimensional vector $\mathbf{Q}$. The link between the $(j, l)$ indices and the index $m$ within the vector $\mathbf{Q}$ is $m = (j-1)M + l$.

Again, the ODE system can be solved by various algorithms such as forward Euler or trapezoidal. The update we obtain from applying forward Euler is

$$Q_{jl}^{n+1} = Q_{jl}^n + \frac{k}{h^2}\left(Q_{j+1,l}^n + Q_{j-1,l}^n + Q_{j,l+1}^n + Q_{j,l-1}^n - 4Q_{jl}^n\right) \tag{2.10}$$

$$= Q_{jl}^n + k\nabla_h^2 Q_{jl}^n \tag{2.11}$$

for $1 \le j, l \le M$. We can determine the time step restriction required for absolute stability if we know the eigenvalues of the matrix $A$. Since $\exp\left[i(\xi x + \eta y)\right]$ is an eigenfunction of the continuous $\nabla^2$ operator and the matrix $A$ discretizes this operator, we guess that the eigenvector should be

$$\mathbf{W}_{p,r} = \begin{bmatrix} e^{i(ph+rh)} & e^{i(p2h+rh)} & \cdots & e^{i(pjh+rlh)} & \cdots & e^{ip(Mh+Mh)} \end{bmatrix}, \tag{2.12}$$

the corresponding discretization of the continuous eigenfunction. Straightforward if somewhat tedious computation leads to

(2.13)

$$\left(A\mathbf{W}_{p,r}\right)_{j,l} = e^{i[p(j-1)h+rlh]} + e^{i[p(j+1)h+rlh]} + e^{i[ph+r(l-1)h]} + e^{i[pjh+r(l+1)h]} - 4e^{i[pjh+rlh]}$$

(2.14)

$$= 2\left[\cos ph + \cos rh - 2\right]\ e^{i(pjh+rlh)}$$

so the eigenvalues are indeed

$$\lambda_{p,r} = 2\left[\cos ph + \cos rh - 2\right] = -4\left(\sin^2\frac{ph}{2} + \sin^2\frac{rh}{2}\right) , \tag{2.15}$$

since these values do not exhibit any dependence on the $j, l$ indices. The forward Euler absolute stability region is

$$\left|1 + \frac{k\lambda}{h^2}\right| \leq 1 \tag{2.16}$$

which leads to

$$-1 \leq 1 - \frac{4k}{h^2}\left(\sin^2\frac{ph}{2} + \sin^2\frac{rh}{2}\right) \leq 1 \tag{2.17}$$

or

$$k \leq \frac{h^2}{2}\left(\sin^2\frac{ph}{2} + \sin^2\frac{rh}{2}\right)^{-1} . \tag{2.18}$$

The most restrictive condition arises for the $ph = rh = \pi$ wavenumbers; to cover any possible error growth at all modes we must limit the time step to

$$k \leq \frac{h^2}{4} . \tag{2.19}$$

As in the 1D case this is a severe limitation in practice since the heat equation leads to an ODE system which is stiff.

Faced with a stiff system, implicit methods for ODE's are good algorithms to consider. The trapezoidal method leads to

$$Q_{jl}^{n+1} = Q_{jl}^n + \frac{k}{2}\left(\nabla_h^2 Q_{jl}^n + \nabla_h^2 Q_{jl}^{n+1}\right) \tag{2.20}$$

an update which is unconditionally stable absolutely for the heat equation since all of the eigenvalues of $A$ are in the left half complex plane. This is known as the 2D Crank-Nicolson algorithm and is widely used in practice. A diificulty which appears now is that the linear system of equations resulting from (2.20) is harder to solve than the simple tridiagonal system obtained in the 1D case. Instead of a tridiagonal matrix we now have a band matrix with a half-bandwidth of $M$. The linear system to be solved is

$$(I - \frac{k}{2}\nabla_h^2)Q_{jl}^{n+1} = (I + \frac{k}{2}\nabla_h^2)Q_{jl}^n \tag{2.21}$$

or

$$M\mathbf{Q}^{n+1} = \mathbf{c}^n = N\mathbf{Q}^n \tag{2.22}$$

Direct solvers (Gaussian elimination, $LR$ factorization) would lead to what is known as zero fill-in. The system matrix $M$ has very few non-zero elements initially; during the course of Gaussian elimination these zero elements would be replaced by non-zero quantitities resulting from the elimination of the lower triangle of $M$. Furthermore, the matrix $M$ has such a simple structure that it usually is not stored as such in a computer program and only the non-zero values which are repeated along the diagonals are stored (two scalar variables). Zero fill-in would lead to the necessity of actually allocating storage for the matrix $M$, or at least for its non-zero band.

Given the above disadvantages of direct methods, iterative solvers are widely used to solve system (2.21). Let $\mathbf{Q}^{[0]}, \mathbf{Q}^{[1]}, \mathbf{Q}^{[2]}, \ldots$ be the succession of iterates

generated by some iterative method to solve linear systems (Gauss-Seidels successive over relaxation). We expect the iteration procedure to converge quicky since we're always starting with a good initial approximation given by $\mathbf{Q}$ at the previous time steps. We could try:

(1) $\mathbf{Q}^{[0]} = \mathbf{Q}^{[n]}$, i.e. setting the initial approximation for the iterative linear system solver to the field from the previous time level;
(2) $\mathbf{Q}^{[0]} = 2\mathbf{Q}^{[n]} - \mathbf{Q}^{[n-1]}$, i.e. setting the initial approximation by extrapolation of the change over the previous two time levels;
(3) $Q^{[0]}_{j,l} = Q^n_{j,l} + k\nabla^2_h Q^n_{jl}$, using forward Euler to give us an initial guess. Typically this works quite well even though the step size is usually much larger than that which would be admitted by the absolute stability restriction for forward Euler ($k \leq h^2/4$). But since we're only applying the forward Euler procedure once, to furnish an initial guess for the iterative linear solver, stability considerations are not a factor now.

CHAPTER 6

# Finite difference methods for hyperbolic equations

## 1. Scalar equations

**1.1. Constant velocity advection in one dimension.** The simplest example of a hyperbolic equation is the constant velocity advection equation

$$q_t + u\, q_x = 0 \tag{1.1}$$

with some initial condition $q(x, t = 0) = q_0(x)$. The equation can be solved along the entire real axis in $x$ or some portion thereof. In numerical work we always have a finite subdomain which we shall conveniently choose as $[0, 2\pi]$ with a view to applying Fourier analysis later on. When using a finite subdomain the question of boundary conditions arises which we shall postpone by considering periodic boundary conditions $q(x + 2\pi, t) = q(x, t)$.

1.1.1. *Exact solution by characteristics.* A first attack on finding the solution to (1.1) is to try to reduce it to a simpler problem. One can ask whether there is any subdomain over which the equation can be cast in a simpler form. For instance we can inquire whether there are any particular curves within the $(x, t)$ plane over which the equation simplifies. A general curve $\Gamma$ of curvilinear parameter is given by

$$\Gamma : \; x = x(s), \; t = t(s) \tag{1.2}$$

and the infinitesimal change in $q$ when going along $\Gamma$ is

$$\frac{dq}{ds} = \frac{\partial q}{\partial t}\frac{dt}{ds} + \frac{\partial q}{\partial x}\frac{dx}{ds} \tag{1.3}$$

Comparing (1.3) with (1.1) we see that if we impose

$$\frac{dt}{ds} = 1, \; \frac{dx}{ds} = u \tag{1.4}$$

then by (1.1) we must have that

$$\frac{dq}{ds} = 0 \; . \tag{1.5}$$

This means that $q$ is constant along the curves $\Gamma$ defined by (1.4) which are $x = ut + C$. The curves are shown in Fig. (1)

1.1.2. *Finite difference methods.* We can construct numerical methods for (1.1) by the same approaches used for the heat equation.

FIGURE 1. Characteristic curves for $q_t + q_x = 0$.

Semi-discretization. Define a computational grid $x_j = jh$, $h = 2\pi/(M+1)$, $t^n = nk$ with step size $h, k$ in space and time. Define $Q_j(t)$ to be the restriction of $q(x,t)$ to $x = x_j$

$$Q_j(t) = q(x_j, t),\ j = 0, 1, \ldots, M+1\ . \tag{1.6}$$

We can choose some approximation of the $x$ derivative. For instance approximating

$$\frac{dq(x_j,t)}{dx} \cong \frac{\delta}{h} Q_j = \frac{Q_{j+1}(t) - Q_{j-1}(t)}{2h} \tag{1.7}$$

leads to the ODE system

$$\frac{d}{dt}\mathbf{Q} = -\frac{u}{2h} B\mathbf{Q} \tag{1.8}$$

with

$$\mathbf{Q} = \begin{bmatrix} Q_1 & Q_2 & \cdots & Q_M \end{bmatrix}^T \tag{1.9}$$

$$B = \begin{bmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ 1 & & & & -1 & 0 \end{bmatrix} \tag{1.10}$$

We can now try various ODE schemes to solve (1.8). Using Euler's method would lead to a FTCS scheme

$$\mathbf{Q}^{n+1} = \mathbf{Q}^n - \frac{uk}{2h} B\mathbf{Q}^n = \left(I - \frac{uk}{2h} B\right)\mathbf{Q}^n\ . \tag{1.11}$$

If instead of Euler's scheme we use the midpoint method we obtain the update formula

$$\mathbf{Q}^{n+1} = \mathbf{Q}^{n-1} - \frac{uk}{h} B\mathbf{Q}^n \tag{1.12}$$

known as the *leap-frog* or *Dufort-Frankel* method.

Full discretization. Instead of the semi-discretization or method of lines approach we can also directly discretize both the space and time derivatives appearing in (1.1). A first-order forward in time, second-order centered in space discretization would lead to the FTCS scheme

$$Q_j^{n+1} = Q_j^n - \frac{uk}{2h}\left(Q_{j+1}^n - Q_{j-1}^n\right) \tag{1.13}$$

A modification of (1.13) of historical relevance is the Lax-Friedrichs scheme

$$Q_j^{n+1} = \frac{1}{2}\left(Q_{j+1}^n + Q_{j-1}^n\right) - \frac{uk}{2h}\left(Q_{j+1}^n - Q_{j-1}^n\right)\ . \tag{1.14}$$

This scheme is obtained by replacing $Q_j^n$ with its arithmetic average using values to the left and to the right. Since the formula has not been derived from discretizations of the derivative which we know to be consistent with the original equation it is useful to determine the truncation error. The exact advection operator is

$$D = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} \tag{1.15}$$

and we have $Dq = 0$ according to (1.1). Our approximation of this operator is

$$\tilde{D}q(x_j, t^n) = \frac{q_j^{n+1} - \frac{1}{2}\left(q_{j+1}^n + q_{j-1}^n\right)}{k} + \frac{u}{2h}\left(q_{j+1}^n - q_{j-1}^n\right) \tag{1.16}$$

with $q_j^n = q(x_j, t^n)$. The truncation error is therefore

$$\tau_j^n = \left(\tilde{D} - D\right) q(x_j, t^n) = \frac{1}{k} q_j^{n+1} - \frac{1}{2k}\left(q_{j+1}^n + q_{j-1}^n\right) + \frac{u}{2h}\left(q_{j+1}^n - q_{j-1}^n\right) \tag{1.17}$$

We can now carry out a Taylor's series expansion around $(x_j, t^n)$. To simplify notation $q$ and its derivatives will be understood to be evaluated at $(x_j, t^n)$ if not explicitly shown otherwise

(1.18)

$$\tau_j^n = \frac{1}{k}\left(q + kq_t + \frac{k^2}{2}q_{tt} + \ldots\right) - \frac{1}{2k}\left(q + hq_x + \frac{h^2}{2}q_{xx} + \ldots + q - hq_x + \frac{h^2}{2}q_{xx} + \ldots\right)$$

(1.19)

$$+ \frac{u}{2h}\left(q + hq_x + \frac{h^2}{2}q_{xx} + \frac{h^3}{6}q_{xxx} + \ldots - q + hq_x - \frac{h^2}{2}q_{xx} + \frac{h^3}{6}q_{xxx} + \ldots\right)$$

which gives

$$\tau_j^n = q_t + \frac{k}{2}q_{tt} + \ldots - \frac{h^2}{2k}q_{xx} + \ldots + uq_x + \frac{uh^2}{12}q_{xxx} + \ldots \tag{1.20}$$

Using (1.1) leads to the leading order error

$$\tau_j^n \cong \frac{k}{2}q_{tt} + \frac{uh^3}{12}q_{xxx} = O(k, h^2) \tag{1.21}$$

The analysis shows that the scheme is first order in time and second order in space. The Lax-Friedrichs scheme is consistent, i.e.

$$\lim_{k,h\to 0} \tau_j^n = 0\ . \tag{1.22}$$

Instead of centered finite differences, other approximations may be introduced. A good choice would be to use one-sided finite differences. This would take into account what we know from the exact solution to the advection equation: information travels along the characteristic lines. It would make sense to use finite differences which have a stencil that mimics this behavior. If $u > 0$ we would use left-sided differences and for $u < 0$ we would use right-sided differences. A first order approximation would be

$$Q_j^{n+1} = \begin{cases} Q_j^n - \dfrac{uk}{h}\left(Q_j^n - Q_{j-1}^n\right) & u \geq 0 \\ Q_j^n - \dfrac{uk}{h}\left(Q_{j+1}^n - Q_j^n\right) & u \leq 0 \end{cases} \tag{1.23}$$

This is known, naturally enough, as the *upwind scheme*.

Taylor series approach. A procedure useful in deriving higher order finite difference approximations for (1.1) is the Taylor series approach. In practical work it is economical to only store two time levels at any given stage in the computation. The general prescription for attaining higher order for the time derivative would involve keeping more terms from the operator series

$$\frac{\partial}{\partial t} = \frac{1}{k}\left(\Delta_+ - \frac{\Delta_+^2}{2} + \frac{\Delta_+^3}{3} - \ldots\right)\ . \tag{1.24}$$

This would be incovenient since the time stencil of the scheme would become wider and we would need to store more than two time levels. We can however use (1.1) to convert time derivatives into spatial derivatives

$$q_t = -uq_x \tag{1.25}$$

$$q_{tt} = \frac{\partial}{\partial t}(q_t) = -\frac{\partial}{\partial t}(uq_x) = -u\frac{\partial}{\partial x}(q_t) = u\frac{\partial}{\partial x}(uq_x) = u^2 q_{xx} \tag{1.26}$$

The Taylor series approach can now be applied to obtain as high an order of approximation in time as needed

$$q(t+k) = q + kq_t + \frac{k^2}{2}q_{tt} + \frac{k^3}{6}q_{ttt} + \dots \ . \tag{1.27}$$

As an example, let us construct a second order scheme by truncating

$$q(t+k) \cong q + kq_t + \frac{k^2}{2}q_{tt} \tag{1.28}$$

We now replace the time derivatives with spatial derivative

$$q(t+k) \cong q - ukq_x + \frac{u^2k^2}{2}q_{xx} \tag{1.29}$$

and use second order accurate, centered finite differences to approximate the spatial derivatives. The resulting scheme is

$$Q_j^{n+1} = Q_j^n - \frac{uk}{2h}\left(Q_{j+1}^n - Q_{j-1}^n\right) + \frac{u^2k^2}{2h^2}\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) \tag{1.30}$$

This is known as the *Lax-Wendroff scheme*.

Instead of centered finite differences we might want to use one-sided formulas to take into account the direction of the characteristics of (1.1). Let us construct a second order accurate one sided approximation using (**??**). One-sided differences can be obtained to arbitrary order of accuracy using the series

$$\frac{\partial}{\partial x} = \frac{1}{h}\left(\Delta_{x+} - \frac{\Delta_{x+}^2}{2} + \frac{\Delta_{x+}^3}{3} - \dots\right) \tag{1.31}$$

$$= \frac{1}{h}\left(\Delta_{x-} + \frac{\Delta_{x-}^2}{2} + \frac{\Delta_{x-}^3}{3} - \dots\right) \tag{1.32}$$

with the finite difference operators defined by

$$\Delta_{x+}q(x,t) = q(x+h,t) - q(x,t) \tag{1.33}$$

$$\Delta_{x-}q(x,t) = q(x,t) - q(x-h,t) \tag{1.34}$$

Let us assume that $u \geq 0$ and therefore that we will be using backward differences so that the computational stencil mimics the true domain of dependence. A second

order accurate approximation of $q_x$ is given by

(1.35)
$$q_x(x_j,t^n) = \frac{\partial q}{\partial x}(x_j,t^n) \cong \frac{1}{h}\left(\Delta_{x-} + \frac{\Delta_{x-}^2}{2}\right) q(x_j,t^n)$$

$$= \frac{1}{h}\left[q(x_j,t^n) - q(x_{j-1},t^n) + \frac{1}{2}\left(q(x_j,t^n) - 2q(x_{j-1},t^n) + q(x_{j-2},t^n)\right)\right] \tag{1.36}$$

$$\cong \frac{1}{h}\left[Q_j^n - Q_{j-1}^n + \frac{1}{2}\left(Q_j^n - 2Q_{j-1}^n + Q_{j-2}^n\right)\right] \tag{1.37}$$

$$= \frac{1}{2h}\left(3Q_j^n - 4Q_{j-1}^n + Q_{j-2}^n\right) \tag{1.38}$$

The second derivative is obtained from

(1.39)
$$\frac{\partial^2}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial}{\partial x} = \frac{1}{h}\left(\Delta_{x-} + \frac{\Delta_{x-}^2}{2} + \frac{\Delta_{x-}^3}{3} - \dots\right)\frac{1}{h}\left(\Delta_{x-} + \frac{\Delta_{x-}^2}{2} + \frac{\Delta_{x-}^3}{3} - \dots\right)$$

(1.40)
$$= \frac{1}{h^2}\left(\Delta_{x-}^2 + \Delta_{x-}^3 + \frac{11}{12}\Delta_{x-}^3\right)$$

Note that in (**??**) we have neglected terms of $O(k^3)$. The exact solution of the advection equation is $q(x,t) = q_0(x-ut)$. The $x-ut$ argument sugests that similar step sizes should be used for $x$ and $t$. This will be confirmed by our stability analysis below. So let us assume that $k = O(h)$. Since $q_{xx}$ already has a $k^2$ factor in (**??**) we only need an $O(h)$ approximation of $q_{xx}$. The leading order error term from $k^2 q_{xx}$ will then be of $O(k^2 h) = O(k^3) = O(h^3)$. We can therefore truncate the series (1.40) to just the first term and approximate

$$\frac{\partial^2 q}{\partial x^2}(x_j,t^n) \cong \frac{1}{h^2}\Delta_{x-}^2 q(x_j,t^n) = \frac{1}{h^2}\left[q(x_j,t^n) - 2q(x_{j-1},t^n) + q(x_{j-2},t^n)\right] \tag{1.41}$$

$$\cong \frac{1}{h^2}\left(Q_j^n - 2Q_{j-1}^n + Q_{j-2}^n\right) \tag{1.42}$$

Note that this is different from the procedure used in deriving the Lax-Wendroff scheme where a second order accurate expression of $q_{xx}$ was used. The reason is that in the Lax-Wendroff scheme the computational stencil already included $Q_{j-1}^n$, $Q_j^n$, $Q_{j+1}^n$ from the approximation of the first derivative $q_x$. Since the second order accurate approximation of $q_{xx}$ does not widen the stencil there is no penalty in using the more accurate, second order approximation of $q_{xx}$. In the one-sided scheme we are deriving here however, using a second order accurate approximation of $q_{xx}$ would involve widening the computational stencil to include $Q_{j-3}^n$. This increases the arithmetic cost of applying the formula without noticeable gain so we choose to use an $O(h)$ approximation of $q_{xx}$. Combining the above results we obtain

$$Q_j^{n+1} = Q_j^n - \frac{uk}{2h}\left(3Q_j^n - 4Q_{j-1}^n + Q_{j-2}^n\right) + \frac{1}{2}\left(\frac{uk}{h}\right)^2\left(Q_j^n - 2Q_{j-1}^n + Q_{j-2}^n\right) \tag{1.43}$$

for $u > 0$, which is known as the *Beam-Warming scheme*.

1.1.3. *Stability analysis.* We now turn to the analysis of the stability of the various schemes introduced above. The analysis can be done using the techniques for systems of ODE's or using Von Neumann analysis. We shall carry out both procedures.

Semi-discretized system. The matrix $B$ arising in the semi-discretized approach is skew-symmetric and will have purely imaginary eigenvalues. We can check this by explicitly calculating the eigenvalues. As usual, we guess that

$$\mathbf{W}_p = \begin{bmatrix} e^{iph} & e^{ip2h} & \cdots & e^{ipjh} & \cdots & e^{ipMh} \end{bmatrix} \tag{1.44}$$

will be an eigenvector since $B$ discretizes a derivation operator. Computing the $j^{th}$ component of $B\mathbf{W}_p$ we get

$$\left(B\mathbf{W}_p\right)_j = e^{ip(j+1)h} - e^{ip(j-1)h} = 2i\sin ph\ e^{ipjh} = 2i\sin ph\ \left(\mathbf{W}_p\right)_j \tag{1.45}$$

so the eigenvalue associated with $\mathbf{W}_p$ is

$$\lambda_p = 2i\sin ph \tag{1.46}$$

and is indeed purely imaginary.

To establish the stability region for the FTCS method (**??**) we use the eigenvalues $\lambda$ of $B$ in the criterion

$$|1+z| \le 1 \tag{1.47}$$

with $z = k\lambda$. It is immediately apparent that the scheme will be unconditionally unstable because $\lambda$ is purely imaginary $\lambda = ai$ so

$$|1+z| = \sqrt{1+(ka)^2} > 1 \tag{1.48}$$

for all $a > 0$.

The interval of stability for the midpoint scheme is $\mathrm{Re}\, z = 0$, $|\mathrm{Im}(z)| \le 1$. Here we would have

$$z = -\frac{uk}{h} i\sin ph \tag{1.49}$$

and the method is stable for

$$\left|\frac{uk}{h}\right| \le 1\ . \tag{1.50}$$

Von Neumann analysis. Obtaining analytical expression for the matrices arising from the semi-discretized approach becomes increasingly difficult as we use more accurate approximations of the derivatives in the PDE or study PDE's more complex than the advection equation. Von Neumann analysis is typically simpler to apply. We start by determing the stability region for the FTCS scheme (1.13). Substituting a typical wavemode $Q^n = \hat{Q}^n e^{i\xi jh}$ we obtain

$$\hat{Q}^{n+1}e^{i\xi jh} = \hat{Q}^n e^{i\xi jh} - \frac{uk}{2h}\left(\hat{Q}^n e^{i\xi(j+1)h} - \hat{Q}^n e^{i\xi(j-1)h}\right)\ . \tag{1.51}$$

The amplification ratio is

$$G = \frac{\hat{Q}^{n+1}}{\hat{Q}^n} = 1 - \frac{uk}{h} i\sin \xi h \tag{1.52}$$

which is always greater than 1

$$|G| \ge 1\ . \tag{1.53}$$

Thus the scheme is unconditionally unstable as we expected from the semi-discretized stability analysis done above.

For the Lax-Friedrichs scheme we obtain

(1.54)
$$\hat{Q}^{n+1}e^{i\xi jh} = \frac{1}{2}\left(\hat{Q}^n e^{i\xi(j+1)h} + \hat{Q}^n e^{i\xi(j-1)h}\right) - \frac{uk}{2h}\left(\hat{Q}^n e^{i\xi(j+1)h} - \hat{Q}^n e^{i\xi(j-1)h}\right)$$

and the amplification factor is

$$G = \cos\xi h - \frac{uk}{h} i \sin\xi h \tag{1.55}$$

Let us introduce the notation

$$\nu = \frac{uk}{h},\ \theta = \xi h\ . \tag{1.56}$$

The stability condition is that

$$|G| = \cos^2\theta + \nu^2\sin^2\theta \leq 1 = \cos^2\theta + \sin^2\theta \tag{1.57}$$

from where we obtain

$$\left(\nu^2 - 1\right)\sin^2\theta \leq 0\ . \tag{1.58}$$

The inequality is satisfied for

$$|\nu| \leq 1\ . \tag{1.59}$$

The quantity $\nu$ that appears repeatedly in analysis of numerical schemes for the advection equation is known as the *Courant-Friedrichs-Lewy* number or more concisely as the *CFL number*. We say that the Lax-Friedrichs scheme is stable for CFL numbers up to 1, it being implicitly understood that we're considering the absolute value of the velocity $|u|$. From the stability criterion we obtain a bound on the time step that we can use in the Lax-Friedrichs scheme

$$k \leq \frac{h}{|u|}\ . \tag{1.60}$$

For the Lax-Wendroff scheme the amplification ratio is

$$G = 1 - \nu i \sin\theta + \nu^2\left(\cos\theta - 1\right) \tag{1.61}$$

We have

$$|G| = 1 + 2\nu^2\left(\cos\theta - 1\right) + \nu^4\left(\cos\theta - 1\right)^2 + \nu^2\sin^2\theta \tag{1.62}$$

$$= 1 - 4\nu^2\sin^2\frac{\theta}{2}\left(1 - \cos^2\frac{\theta}{2}\right) + 4\nu^4\sin^4\frac{\theta}{2} \tag{1.63}$$

The stability condition is $|G| \leq 1$ leads to

$$\left(\nu^2 - 1\right)\sin^2\frac{\theta}{2} \leq 0 \tag{1.64}$$

so again the domain of stability is

$$|\nu| \leq 1\ . \tag{1.65}$$

Lax-Wendroff is a more efficient scheme than Lax-Friedrichs since we obtain $O(h^2, k^2)$ precision as opposed to $O(h, k^2)$ under the same time step restriction $k \leq h/|u|$.

Turning now to the one-sided schemes, for upwind when $u > 0$ we have

$$G = 1 - \nu\left(1 - e^{-i\theta}\right) \tag{1.66}$$

$$|G| = 1 - 2\nu\left(1 - \cos\theta\right) + \nu^2\left(1 - \cos\theta\right)^2 + \nu^2\sin^2\theta \tag{1.67}$$

FIGURE 2. Amplification factor $|G(\nu,\theta)|$ for the Beam-Warming scheme evaluated at $\theta = m\pi/8$, $m = 0, 1, \ldots 16$.

The stability condition $|G| \leq 1$ leads to

$$-2\nu\left(1-\cos\theta\right)+\nu^2\left(1-\cos\theta\right)^2+\nu^2\sin^2\theta\leq 0 \tag{1.68}$$

which can be rewritten in terms of the half-angle $\theta/2$ to give

$$-4\nu\sin^2\frac{\theta}{2}+4\nu^2\sin^4\frac{\theta}{2}+4\nu^2\sin^2\frac{\theta}{2}\cos^2\frac{\theta}{2}\leq 0 \tag{1.69}$$

and finally

$$\nu(\nu-1)\leq 0 \tag{1.70}$$

so the stability region is again $\nu \leq 1$.

For the Beam-Warming scheme we have

$$G=1-\frac{\nu}{2}\left(3-4e^{-i\theta}+e^{-2i\theta}\right)+\frac{1}{2}\nu^2\left(1-2e^{-i\theta}+e^{-2i\theta}\right)\ . \tag{1.71}$$

Notice that as we look at more complicated schemes the amplification factors become increasingly difficult to evaluate analytically. We can however use a numerical evaluation of $G(\nu,\theta)$ to generate plots such as Fig. 2. From the plot we deduce that the stability region is $\nu \leq 2$.

1.1.4. *Lax equivalence theorem.* The importance of establishing consistency and stability for a finite difference scheme for the advection equation is that these two properties guarantee convergence by the Lax equivalence theorem.

THEOREM 4. *A finite difference scheme for a linear PDE is convergent if the scheme is consistent with the PDE and it is stable.*

Convergence means that

$$\lim_{k,h\to 0} Q_j^n = q(x_j, t^n) \tag{1.72}$$

where $k, h$ go to zero in accordance with the stability criterion for the scheme. Convergence is obtained when the scheme is consistent, i.e. the truncation error goes to zero

$$\lim_{k,h\to 0} \tau_j^n = 0 \tag{1.73}$$

and the step sizes satisfy the stability criterion.

1.1.5. *Modified equations.* We have established a number of methods for solving the advection equation (1.1). Up to now we have characterized the error of any one scheme by its truncation error. Though indicative of the overall quality of an approximation, the precise nature of the error in the scheme is not apparent. It has proved very fruitful in the development of better methods to more accurately describe how a numerical approximation differs from the exact solution. A question one can ask is whether a given numerical scheme is perhaps a more accurate discretization of another PDE than the one it was originally designed for. Let us exemplify using the upwind scheme for the advection equation with $u > 0$

$$Q_j^{n+1} = Q_j^n - \nu\left(Q_j^n - Q_{j-1}^n\right)\ . \tag{1.74}$$

We know that this scheme is $O(k,h)$ accurate for the equation $q_t+uq_x=0$. Suppose that the scheme is an exact discretization of some unknown PDE $Ls=0$ with $L$ an unknown differential operator and $s=s(x,t)$. Then we would have

$$s(x,t+k)=s(x,t)-\nu\left[s(x,t)-s(x-h,t)\right]\ , \tag{1.75}$$

exactly. Let us carry out Taylor series expansion of $s$ around $(x,t)$

$$s+ks_t+\frac{k^2}{2}s_{tt}+\frac{k^3}{6}s_{ttt}+\ldots=s-\frac{uk}{h}\left[hs_x-\frac{h^2}{2}s_{xx}+\frac{h^3}{6}s_{xxx}-\ldots\right]\ . \tag{1.76}$$

To obtain a more concise notation the function arguments have been dropped. We obtain

$$s_t+us_x=-\frac{k}{2}s_{tt}+\frac{uh}{2}s_{xx}-\frac{k^2}{6}s_{ttt}-\frac{uh^2}{6}s_{xxx}+\ldots \tag{1.77}$$

This is of the form $As=E_{(h,k)}s$ with $A$ the advection operator $A=\partial_t+u\partial_x$ and $E_{(h,k)}$ an operator giving the deviation of the modified equation from the advection equation. Note that if $k=h=0$ we obtain the advection equation for which the scheme (1.74) is $O(h,k)$ accurate. We can interpret (1.77) as stating that the scheme (1.74) is:

(1) first order accurate for

$$s_t+us_x=0 \tag{1.78}$$

(2) second order accurate for

$$s_t+us_x=-\frac{k}{2}s_{tt}+\frac{uh}{2}s_{xx} \tag{1.79}$$

(3) third order accurate for

$$s_t+us_x=-\frac{k}{2}s_{tt}+\frac{uh}{2}s_{xx}-\frac{k^2}{6}s_{ttt}-\frac{uh^2}{6}s_{xxx} \tag{1.80}$$

The equations obtained above are called *modified equations*. These statements can be verfied by explicit computation of the truncation error. For example let us compute the truncation error in applying (1.74) to (**??**)

$$\tau_j^n=\left(\tilde{D}-D\right)s(x_j,t^n) \tag{1.81}$$

The finite difference approximation operator is

$$\tilde{D}s(x_j,t^n)=\frac{s_j^{n+1}-s_j^n}{k}+\frac{u}{h}\left(s_j^n-s_{j-1}^n\right) \tag{1.82}$$

The exact operator for the modified equation (**??**) is

$$D=\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}+\frac{k}{2}\frac{\partial^2}{\partial t^2}-\frac{uh}{2}\frac{\partial^2}{\partial x^2} \tag{1.83}$$

and we have $Ds=0$. We now expand $s_j^{n+1}$, $s_{j-1}^n$, $s_{j+1}^n$ from (1.82) around $(x_j,t^n)$ and obtain

$$\tau_j^n=\frac{1}{k}\left(s+ks_t+\frac{k^2}{2}s_{tt}+\ldots\right)-\frac{1}{k}s+\frac{u}{h}\left(s-s+hs_x-\frac{h^2}{2}s_{xx}+\frac{h^3}{6}s_{xxx}+\ldots\right) \tag{1.84}$$

$$\tau_j^n=s_t+\frac{k}{2}s_{tt}+us_x+\frac{uh}{2}s_{xx}+O(k^2,h^2)=Ds+O(k^2,h^2)=O(k^2,h^2) \tag{1.85}$$

so the truncation error is indeed of second order.

Now let us show the benefits of looking at the modified equation by using (**??**) for which the upwind scheme (1.77) is second order accurate. First we recast (**??**) so as to eliminate higher order derivatives in time. We can rewrite (**??**) as

$$s_t = -us_x - \frac{k}{2}s_{tt} + \frac{uh}{2}s_{xx} \tag{1.86}$$

and differentiate with respect to $t$ to obtain

$$s_{tt} = -us_{xt} - \frac{k}{2}s_{ttt} + \frac{uh}{2}s_{xxt} \tag{1.87}$$

Replacing (1.87) in (**??**) gives

$$s_t + us_x = -\frac{k}{2}\left(-us_{xt} - \frac{k}{2}s_{ttt} + \frac{uh}{2}s_{xxt}\right) + \frac{uh}{2}s_{xx} \tag{1.88}$$

$$= \frac{uk}{2}s_{xt} + \frac{uh}{2}s_{xx} + O(k^2, h^2, kh) \tag{1.89}$$

We can neglect the higher order terms since this is consistent with the order of accuracy used in obtaining (**??**). Differentiating (**??**) with respect to $x$ yields

$$s_{tx} = -us_{xx} - \frac{k}{2}s_{ttx} + \frac{uh}{2}s_{xxx} \tag{1.90}$$

and replacing this in (**??**) gives

$$s_t + us_x = -\frac{u^2k}{2}s_{xx} + \frac{uh}{2}s_{xx} = \frac{uh}{2}(1-\nu)s_{xx} \tag{1.91}$$

Equation (1.91) is the usual way to express the modified equation for the upwind scheme applied to the advection equation to second order. It shows that the upwind scheme does indeed model the advection equation in the limit $h \to 0$. For finite step sizes however the upwind scheme more accurately models the equation (1.91). The difference between (1.91) and the advection equation is the term

$$\frac{uh}{2}(1-\nu)s_{xx} \tag{1.92}$$

Note that this is a diffusive term whose effect is to smooth out any variations in $s(x,t)$ as long as $|1-\nu| \geq 0$ as has been seen in the study of the heat equation. The condition $|1-\nu| \geq 0$ is exactly the stability criterion for the upwind scheme. Indeed if $\nu > 1$ then we would obtain a negative diffusion coefficient for which the initial value problem is ill-defined. We can see that at exactly $\nu = 1$ there is no diffusion indicating that for $\nu = 1$ the upwind scheme achieves higher order accuracy for the advection equation. When $\nu < 1$ the error in the upwind scheme with respect to the true solution $q(x,t)$ of the advection equation will be diffusive: gradients will be smoothed out instead of being simply advected.

Now that we have seen the nature of the error introduced by the upwind scheme applied to the advection equation, we can also use this information to derive better schemes. Since the error is known to be given by (1.92) we can change the upwind scheme

$$Q_j^{n+1} = Q_j^n - \nu\left(Q_j^n - Q_{j-1}^n\right) \tag{1.93}$$

to counteract the known error by including a discretization of (1.92)

$$Q_j^{n+1} = Q_j^n - \nu\left(Q_j^n - Q_{j-1}^n\right) - k\frac{uh}{2}(1-\nu)\frac{Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n}{h^2} \tag{1.94}$$

Working this through leads to the scheme

$$Q_j^{n+1} = Q_j^n - \frac{\nu}{2}\left(Q_{j+1}^n - Q_{j-1}^n\right) + \frac{\nu^2}{2}\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right). \tag{1.95}$$

Thus we have obtained the Lax-Wendroff scheme (1.30) via another route.

The procedure can be continued to higher orders. We can now ask what is the modified equation more accurately described by the Lax-Wendroff scheme. Repeating the same procedures as above we first write

(1.96)

$$s(x,t+k) = s(x,t) - \frac{\nu}{2}\left[s(x+h,t) - s(x-h,t)\right] + \frac{\nu^2}{2}\left[s(x+h,t) - 2s(x,t) + s(x-h,t)\right]$$

and then carry out Taylor series expansions around $(x,t)$ to obtain

$$s + ks_t + \frac{k^2}{2}s_{tt} + \frac{k^3}{6}s_{ttt} + \ldots = s - \frac{uk}{2h}\left[2hs_x + \frac{h^3}{3}s_{xxx} + \ldots\right] + \tag{1.97}$$

$$\frac{u^2k^2}{2h^2}\left[h^2 s_{xx} + \frac{h^4}{12}s_{xxx}\right] \tag{1.98}$$

from where

$$s_t + us_x = -\frac{k}{2}\left(s_{tt} - u^2 s_{xx}\right) - \frac{k^2}{6}s_{ttt} - \frac{uh^2}{6}s_{xxx} + O(k^3, h^3) \tag{1.99}$$

Note the appearance of the $O(k)$ term. Had we carried out the Taylor expansion for the advection equation this term would have been proportional to $q_{tt} - u^2 q_{xx}$ which is zero according to (1.26). Here we cannot assume that $s_{tt} - u^2 s_{xx}$ is zero a priori. We must carry the term in the ensuing computations, expecting that it will give a higher order correction. Let us neglect the $O(k^3, h^3)$ contributions and proceed with our technique of replacing higher order time derivatives with spatial derivatives using

$$s_{tt} = -us_{xt} - \frac{k}{2}\left(s_{ttt} - u^2 s_{xxt}\right) \tag{1.100}$$

$$s_{ttt} = -us_{xtt}. \tag{1.101}$$

Higher order terms have been dropped since they would lead to $O(k^3, h^3)$ contributions in (1.99). Our intermediate result is

$$s_t + us_x = -\frac{k}{2}\left[-us_{xt} - \frac{k}{2}\left(s_{ttt} - u^2 s_{xxt}\right) - u^2 s_{xx}\right] + \frac{k^2}{6}us_{xtt} - \frac{uh^2}{6}s_{xxx} \tag{1.102}$$

and we continue by eliminating mixed derivatives. In the above formula we wish to express $s_{xtt}$ in terms of $x$ derivatives to $O(1)$

$$s_{xtt} = s_{ttx} = (s_t)_{tx} = (-us_x)_{tx} = -u\,(s_t)_{xx} = u^2 s_{xxx}\ . \tag{1.103}$$

We also need to express $s_{xt}$ in terms of $x$ derivatives to $O(k,h)$

$$s_{xt} = s_{tx} = -us_{xx} - \frac{k}{2}\left(s_{ttx} - u^2 s_{xxx}\right) \tag{1.104}$$

and $s_{ttt}$, $s_{xxt}$ to $O(1)$

$$s_{ttt} = -u^3 s_{xxx},\ s_{xxt} = -us_{xxx}\ . \tag{1.105}$$

Replacing in (1.102) leads to

(1.106)
$$s_t+us_x=-\frac{k}{2}\left\{-u\left[-us_{xx}-\frac{k}{2}\left(s_{ttx}-u^2s_{xxx}\right)\right]-u^2s_{xx}-\frac{k}{2}\left(-u^3s_{xxx}+u^3s_{xxx}\right)\right\}+\frac{u}{6}\left(k^2u^2-h^2\right)s_{xxx}$$

which simplifies to

$$s_t+us_x=\frac{k^2}{4}\left(s_{ttx}-u^2s_{xxx}\right)+\frac{u}{6}\left(k^2u^2-h^2\right)s_{xxx} \tag{1.107}$$

Since $s_{ttx}=u^2s_{xxx}$ to $O(1)$ we obtain in final

$$s_t+us_x=-\frac{uh^2}{6}\left(1-\nu^2\right)s_{xxx}\ . \tag{1.108}$$

The third order derivative now obtained shows that the Lax-Wendroff scheme introduces a *dispersive error* with different wave numbers traveling at different speeds. As expected, the dispersive error is proportional to $h^2$ since the Lax-Wendroff scheme is second order. A scheme more accurate than Lax-Wendroff could be obtained by adding a correction term modeling the dispersive error. Since this involves a third-order derivative the stencil of the scheme would become wider by at least one unit thereby entailing more computational work.

**1.2. Non-linear scalar equations.** We have introduced a number of finite difference methods for the simple constant-velocity advection equation (1.1). Of course, there is hardly much need for a numerical method in order to solve (1.1). Rather we have used (1.1) as a model problem to study the properties of numerical schemes on a simple case. We now proceed to consider more complicated problems and investigate how the methods already derived apply to these problems.

A general first order, hyperbolic scalar equation is given by

$$q_t+u(x,t,q)q_x=\sigma(x,t,q) \tag{1.109}$$

where $u$ may be interpretred as local advection velocity that depends in general upon $x,t$ and $q$. In a wide range of problems equations of the form

$$q_t+f(q)_x=\sigma(x,t,q) \tag{1.110}$$

arise where $f$ is known as the *flux function*. If $f$ is differentiable we can write

$$q_t+f_qq_x=\sigma(x,t,q) \tag{1.111}$$

so $f_q$ plays the role of the local advection velocity. Generally $u,f$ depend on $q$ so that the equations become non-linear in $q$. Equation (1.110) is said to be in *conservative form* as opposed to (1.109) which is said to be in *non-conservative form*. Generally we say that a PDE is in conservative form when it can be expressed as the space-time divergence of a vector field. For equation (1.110) the vector field would be $(q,f(q))$ and the space-time divergence is $\nabla_{(t,x)}\cdot\ =(\partial_t,\partial_x)\cdot$ so another way of writing (1.110) is

$$\nabla_{(t,x)}\cdot(q,f(q))=0\ . \tag{1.112}$$

An initial value problem is defined by specifying a solution domain along $x$ and an initial condition $q_0(x)$.

FIGURE 3. Characteristic curves for $q_t + e^{x+t} q_x = -\beta q$ .

1.2.1. *Solution by characteristics.* We can solve (1.109) using the method of characteristics. We again ask whether there are any special curves $\Gamma$ within the $(x,t)$ plane on which (1.109) reduces to a simpler form. Along the curves specified by the differential system

$$\frac{dt}{ds} = 1, \ \frac{dx}{ds} = u(x,t,q) \tag{1.113}$$

we do indeed obtain the simpler form

$$\frac{dq}{ds} = \sigma. \tag{1.114}$$

The essential difference with respect to the constant-velocity case is that the curves are no longer simple straight lines but depend on $x, t$ and $q$. Let us consider some examples in order to see the complications involved.

Variable-velocity advection. Consider the equation

$$q_t + u(x,t) q_x = \sigma \tag{1.115}$$

which describes the advection of the unknown field variable $q$ by an imposed velocity field $u(x,t)$. The velocity field is not influenced by $q$ itself; $q$ is said to be a *passive tracer*. The characteristic curves are given by the ODE

$$\frac{dx}{dt} = u(x,t) \ . \tag{1.116}$$

Note that we are no longer guaranteed that the characteristics exist for all times as they did for the constant-velocity advection equation. This is the case only if $u$ is uniformly Lipschitz.

EXAMPLE 7. *Consider the velocity field*

$$u(x,t) = x + t, \tag{1.117}$$

*the initial condition*

$$q_0(x) = \sin x, \tag{1.118}$$

*and the source term*

$$\sigma = -\beta q \ . \tag{1.119}$$

*The characteristic curves are*

$$x(t) = Ce^t - t - 1 \tag{1.120}$$

*which are shown in Fig. (3). At $t = 0$ the characteristic labeled by $C$ passes through the $x$ coordinate $x_0 = C - 1$. Along each characteristic the variable-velocity advection equation reduces to the ODE*

$$\frac{dq}{dt} = -\beta q \tag{1.121}$$

*which has the solution $q(x,t) = Ae^{-\beta t}$. We have to determine the constant $A$ from the initial conditions. Through any given point $(x,t)$ there passes the characteristic curve labeled by $C = e^{-t}(x+t+1)$. This particular characteristic curve will intersect*

FIGURE 4. Solution of $q_t + (x+t)q_x = -\beta q$, $q(x, t=0) = \sin x$ for $\beta = 0.1$ at $t = 0, 0.2, \ldots, 1$.

FIGURE 5. Crossing characteristics for inviscid Burgers equation with initial condition $q_0(x) = \sin x$ (shown in thick line).

*the x-axis at $x_0 = C - 1$ and this is the position from which we must take the initial value for q*

$$q(x,t) = q_0(e^{-t}(x+t+1) - 1)e^{-\beta t} = \sin(e^{-t}(x+t+1) - 1)e^{-\beta t} \ . \tag{1.122}$$

*We have found the solution to the PDE using the simpler expression of the PDE along the characteristics. The solution can be verified by direct substitution in (1.115) and is depicted in Fig. (4). The initial condition is spread out due to the spreading out of the characteristic curves and attenuated due to the source term $\sigma$.*

Burgers equation. A model equation used extensively in the study of non-linear equations is

$$q_t + qq_x = 0 \tag{1.123}$$

known as the *inviscid Burgers equation*. It is given in non-conservative form above. In conservative form it becomes

$$q_t + \left(\frac{q^2}{2}\right)_x = 0 \tag{1.124}$$

so the flux function is

$$f(q) = q^2/2 \ . \tag{1.125}$$

The characteristic curves are given by

$$\frac{dx}{dt} = q(x,t) \tag{1.126}$$

and along a characteristic curve $\Gamma$ equation (1.123) reduces to

$$\left(\frac{dq}{ds}\right)_\Gamma = 0 \ , \tag{1.127}$$

i.e. there is no variation in $q$ along the characteristic. This implies that the slope of each characteristic curve is constant and specified by the initial condition $q(x, t = 0) = q_0(x)$.

The type of difficulties that arise for non-linear equations is immediately apparent from the consideration of simple initial conditions. Consider $q_0(x) = \sin x$. The characteristics are sketched in Fig. 5. The problem is that the characteristic curves cross one another. At such a crossing point it is not apparent what the correct value of $q$ should be since different values are being transported along each of the crossing characteristics.

To get a better idea of what is happening it is useful to simplify the initial condition as much as possible. This leads to the so-called *Riemann problem*

$$q_0(x) = \begin{cases} q_l & x < 0 \\ q_r & x > 0 \end{cases} \tag{1.128}$$

Let us try to solve Burgers equation for this initial condition.

If $q_l > q_r$ characteristics from $x < 0$ will overtake those from $x > 0$. This will occur on some ray from the origin of equation $x = st$. To the left of this separating ray we will observe the value $q_l$ while to the right we will observe the value $q_r$. The solution is therefore

$$q(x,t) = \begin{cases} q_l & x < st \\ q_r & x > st \end{cases} \tag{1.129}$$

The initial discontinuity propagates at a velocity $s$. The discontinuity is called a *shock* using the language of compressible gas dynamics and $s$ is the *shock velocity*. The shock velocity can be determined by using the integrating Burgers equation over a domain having the shock as its diagonal $[st_1, st_2] \times [t_1, t_2]$

$$\int_{st_1}^{st_2} \int_{t_1}^{t_2} [q_t + f(q)_x]\, dtdx = 0 \tag{1.130}$$

from where

$$s = \frac{f(q_r) - f(q_l)}{q_r - q_l}\ . \tag{1.131}$$

If $q_l < q_r$ two solutions are possible. We can again have the shock solution (1.129) but also the solution

$$q(x,t) = \begin{cases} q_l & x < q_l t \\ x/t & q_l t \le x \le q_r t \\ q_r & x > q_r t \end{cases} \tag{1.132}$$

called a *rarefaction solution*, again using terms from gas dynamics. This an even worse conundrum, not only can discontinuities arise which invalidate the differentiation operations but multiple solutions seem to be possible. Clearly something is wrong and a way to correct the model that led to equation (**??**) must be found. From the physical point of view certain effects have been neglected, namely the viscosity of the fluid and we might be led to studying the viscous Burgers equation

$$q_t + qq_x = \nu q_{xx} \tag{1.133}$$

as a remedy to the difficulties encountered. This can be done and leads to smooth solutions with very large gradients in the regions where shocks would have formed for the inviscid Burgers equation. These large gradients are difficult to resolve properly requiring very fine grids, much finer than needed elsewhere in the solution domain. So a way that enables us to still work with the inviscid equation is quite useful.

1.2.2. *Weak solutions.* The possibility of crossing characteristic curves is indicative with a breakdown of the modeling assumptions that led to a certain hyperbolic PDE. In this situation one must revisit the method by which a certain PDE is derived and consider the validity of all intermediate hypotheses used in the derivation. Burgers equation serves as a useful example. The PDE

$$q_t + f(q)_x = 0 \tag{1.134}$$

with $f = q^2/2$ was proposed as a model for fluid flow in which the quantity $q$ is conserved but being advected by itself. The correct formulation of a conservation principle is through the integral statement

$$\int_{x_1}^{x_2} [q(x,t_2) - q(x,t_1)]\, dx = -\int_{t_1}^{t_2} [f(q(x_2,t)) - f(q(x_1,t))]\, dt \tag{1.135}$$

the one-dimensional expression of (1.10). In this form one can replace

$$q(x,t_2) - q(x,t_1) = \int_{t_1}^{t_2} \frac{\partial q}{\partial t}(x,t)dt \tag{1.136}$$

$$f(q(x_2,t)) - f(q(x_1,t)) = \int_{x_1}^{x_2} \frac{\partial f}{\partial x}dx \tag{1.137}$$

and obtain Burgers equation by going to the limits $t_2 \to t_1$, $x_2 \to x_1$ if the derivatives $\partial q/\partial t$, $\partial f/\partial x$ exist. However one cannot do this if $q$ is discontinuous. In this case only the integral form (1.135) is valid.

Nonetheless it is typically much more convenient to work with differential equations instead of integral equations. Therefore it is useful to extend the meaning we associate to "$q$ is a solution of a PDE" to cover the case where $q$ might be discontinuous at a few points. This is done through the techniques of the theory of distributions by requiring that $q$ satisfy a certain integral condition. Namely we consider the integral

$$I = \int_0^\infty \int_{-\infty}^{+\infty} \phi\, [q_t + f(q)_x]\, dxdt \tag{1.138}$$

with $\phi$ a smooth function of finite support and impose $I = 0$. Typically we require that $\phi$ be at least differentiable. We can integrate by parts to obtain

$$\int_0^\infty \int_{-\infty}^{+\infty} [\phi_t q + \phi_x f(q)]\, dxdt = -\int_{-\infty}^{+\infty} \phi(x,0)q(x,0)dx\ . \tag{1.139}$$

By this technique all differentiation operations on $q$ have been removed. We say that $q$ is a *weak solution* of (1.134) if (2.28) is satisfied for all $\phi$ from some space of test functions such as $\phi \in C^1(\mathbb{R}\times\mathbb{R})$.

1.2.3. *Difficulties of finite difference methods for non-linear hyperbolic equations.* The possibility of shocks for non-linear hyperbolic equations should alert us to possible difficulties with the finite difference methods we have introduced for the linear advection equation. Since these are based upon Taylor series expansions of $q(x,t)$ and $q$ can be discontinuous, the expansions will break down and not be valid near the discontinuities. Nevertheless, we would expect the methods to be adequate in regions where $q$ is smooth.

Let us see how we would apply the methods to a non-linear equation, taking Burgers equation as an example. One possibility is to interpret $q$ as the local advection velocity $u$. The upwind method for

$$q_t + qq_x = 0 \tag{1.140}$$

then becomes

$$Q_j^{n+1} = Q_j^n - \begin{cases} Q_j^n \left(Q_j^n - Q_{j-1}^n\right) & \text{if } Q_j^n \geq 0 \\ Q_j^n \left(Q_{j+1}^n - Q_j^n\right) & \text{if } Q_j^n < 0 \end{cases} \tag{1.141}$$

and the Lax-Wendroff methods reads

$$Q_j^{n+1} = Q_j^n - \frac{Q_j^n k}{2h}\left(Q_{j+1}^n - Q_{j-1}^n\right) + \frac{\left(Q_j^n\right)^2 k^2}{2h^2}\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) \ . \tag{1.142}$$

Applying this for a Riemann problem leads to a numerical solution similar to the exact shock solution but with oscillations near the shock (Fig. 1.2.3). There is also a smearing of the shock, instead of sharp discontinuity we have a smoothing of $q$ in the vicinity of the shock. Far from the shock the numerical solution is quite good however. This therefore leads to the search for so-called *high-resolution algorithms* that are able to preserve a high order of accuracy away from discontinuities and also sharply capture discontinuities.

## 2. Systems of hyperbolic equations

### 2.1. Linear systems.

2.1.1. *Classification of linear systems.* Consider now that we are interested in the simultaneous time evolution of a number of quantities

$$q = \left[\begin{array}{cccc} q_1 & q_2 & \dots & q_m \end{array}\right]^T \tag{2.1}$$

that satisfy

$$q_t + \mathbf{A}q_x = 0 \tag{2.2}$$

with $\mathbf{A}$ a constant $m \times m$ matrix of real numbers. Such a system is said to be hyperbolic if the eigenvectors of $\mathbf{A}$ form a basis for real $m$-vectors.

Example 8. *The second order wave equation is given in canonical form as*

$$\phi_{tt} - c^2\phi_{xx} = 0 \ . \tag{2.3}$$

*It can be reduced to a system of two first-order equations by introducing*

$$u = \phi_t, \ v = \phi_x \tag{2.4}$$

*We have*

$$u_t - c^2 v_x = 0 \tag{2.5}$$

*and since* $\phi_{xt} = \phi_{tx}$

$$v_t - u_x = 0 \tag{2.6}$$

*In vector form we obtain*

$$\frac{\partial}{\partial t}\left[\begin{array}{c} u \\ v \end{array}\right] + \left[\begin{array}{cc} 0 & -c^2 \\ -1 & 0 \end{array}\right]\frac{\partial}{\partial x}\left[\begin{array}{c} u \\ v \end{array}\right] = 0 \tag{2.7}$$

*or*

$$q_t + \mathbf{A}q_x = 0 \tag{2.8}$$

*with*

$$q = \left[\begin{array}{c} u \\ v \end{array}\right], \ \mathbf{A} = \left[\begin{array}{cc} 0 & -c^2 \\ -1 & 0 \end{array}\right] \tag{2.9}$$

*The eigenvalues of* $\mathbf{A}$ *are* $\lambda_{1,2} = \pm c$ *and the eigenvectors are*

$$r_1 = \left[\begin{array}{c} c \\ 1 \end{array}\right], \ r_2 = \left[\begin{array}{c} -c \\ 1 \end{array}\right] \tag{2.10}$$

*The eigenvectors are independent for $c \neq 0$ and therefore they form a basis for the space of real 2-vectors. The system (2.8) is hyperbolic.*

EXAMPLE 9. *Applying the same procedure to the Laplace equation*

$$\phi_{tt} + \phi_{xx} = 0 \tag{2.11}$$

*leads to the matrix*

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{2.12}$$

*whose eigenvalues are $\lambda_{1,2} = \pm i$ and eigenvectors are*

$$\begin{bmatrix} i \\ 1 \end{bmatrix}, \begin{bmatrix} -i \\ 1 \end{bmatrix} \tag{2.13}$$

*These have complex values and are not a basis for real 2-vectors. The system (2.12) is not hyperbolic, it is elliptic.*

2.1.2. *Solution by method of characteristics and reduction to diagonal form.* For hyperbolic systems we can apply a procedure similar to that used for systems of ODE's. We can write

$$\mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1} \tag{2.14}$$

with

$$\mathbf{\Lambda} = diag\,\{\lambda_1, \lambda_2, \ldots, \lambda_m\} \tag{2.15}$$

$$\mathbf{T} = [r_1\ r_2\ \cdots\ r_m] \tag{2.16}$$

$$\mathbf{A}r_j = \lambda_j r_j,\ j = 1, 2, \ldots, m\ . \tag{2.17}$$

and write

$$q_t + \mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1}q_x = 0\ . \tag{2.18}$$

Introducing the notation

$$w = \mathbf{T}^{-1}q \tag{2.19}$$

we obtain

$$w_t + \mathbf{\Lambda}w_x = 0\ . \tag{2.20}$$

Since $\mathbf{\Lambda}$ is a diagonal matrix, the equations of the original system have been decoupled and we can write the scalar $j^{\text{th}}$component equation

$$w_t^{(j)} + \lambda_j w_x^{(j)} = 0\ , \tag{2.21}$$

for $j = 1, 2, \ldots, m$. These are now simple constant-velocity advection equations for which we know the solution

$$w^{(j)}(x,t) = w_0^{(j)}(x - \lambda_j t) \tag{2.22}$$

with $w_0^{(j)}$ given by the initial conditions on $q$

$$w_0 = \mathbf{T}^{-1}q_0\ . \tag{2.23}$$

The value of each individual component of $w^{(j)}$ is constant along the family of characteristics $x - \lambda_j t = C_j$. Therefore $w$ are known as the *conservative variables.*

From a knowledge of the conservative variable solution we can recover the solution for the original variables

$$q = \mathbf{T}w \ . \tag{2.24}$$

2.1.3. *Finite difference methods.* The finite difference methods derived for the constant-velocity advection equation can be applied formally to hyperbolic systems also. For example, the Lax-Wendroff scheme is

$$Q_j^{n+1} = Q_j^n - \frac{k}{2h}\mathbf{A}\left(Q_{j+1}^n - Q_{j-1}^n\right) + \frac{k^2}{2h^2}\mathbf{A}^2\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) \tag{2.25}$$

There are some new features though due to the fact that there is no longer just a single "advection" or characteristic velocity. Let us try to apply the upwind method to the system (2.8). It is not apparent what the upwind direction should be for $q$. We can ascertain this for the conservative variables though. We have $\lambda_1 = -c$, $\lambda_2 = c$,

$$r_1 = \begin{bmatrix} c \\ 1 \end{bmatrix},\ r_1 = \begin{bmatrix} -c \\ 1 \end{bmatrix} \tag{2.26}$$

$$T = \begin{bmatrix} c & -c \\ 1 & 1 \end{bmatrix},\ T^{-1} = \begin{bmatrix} \frac{1}{2c} & \frac{1}{2} \\ -\frac{1}{2c} & \frac{1}{2} \end{bmatrix},\ \Lambda = \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix} \tag{2.27}$$

and the conservative variable system is

$$\frac{\partial}{\partial t}\begin{bmatrix} w^{(1)} \\ w^{(2)} \end{bmatrix} + \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} w^{(1)} \\ w^{(2)} \end{bmatrix} = 0 \ . \tag{2.28}$$

This system can be discretized in an upwind manner and we obtain the scheme

$$\left(W^{(1)}\right)_j^{n+1} = \left(W^{(1)}\right)_j^n + \frac{ck}{h}\left[\left(W^{(1)}\right)_{j+1}^n - \left(W^{(1)}\right)_j^n\right] \tag{2.29}$$

$$\left(W^{(2)}\right)_j^{n+1} = \left(W^{(2)}\right)_j^n - \frac{ck}{h}\left[\left(W^{(2)}\right)_j^n - \left(W^{(2)}\right)_{j-1}^n\right] \tag{2.30}$$

In matrix form this reads

$$W_j^{n+1} = (1-\nu)\,W_j^n + \mathbf{C}W_{j-1}^n + \mathbf{D}W_{j+1}^n \tag{2.31}$$

with

$$\nu = \frac{ck}{h},\ \mathbf{C} = \begin{bmatrix} 0 & 0 \\ 0 & \nu \end{bmatrix},\ \mathbf{D} = \begin{bmatrix} \nu & 0 \\ 0 & 0 \end{bmatrix} \tag{2.32}$$

Multiplying by $T$ leads to

$$Q_j^{n+1} = (1-\nu)\,Q_j^n + \mathbf{TCT}^{-1}Q_{j-1}^n + \mathbf{TDT}^{-1}Q_{j+1}^n \tag{2.33}$$

$$\mathbf{TCT}^{-1} = \frac{\nu}{2}\begin{bmatrix} 1 & -c \\ -\frac{1}{c} & 1 \end{bmatrix},\ \mathbf{TDT}^{-1} = \frac{\nu}{2}\begin{bmatrix} 1 & c \\ \frac{1}{c} & 1 \end{bmatrix} \tag{2.34}$$

This is the upwind scheme for the system (2.8).

**2.2. Non-linear systems.**

2.2.1. *Classification.* Non-linear systems are written as

$$q_t + f(q)_x = 0 \tag{2.35}$$

with

$$q = \begin{bmatrix} q_1 & q_2 & \dots & q_m \end{bmatrix}^T,\ f = \begin{bmatrix} f_1 & f_2 & \dots & f_m \end{bmatrix}^T \tag{2.36}$$

The classification of non-linear systems is made in accordance with the properties of the Jacobian of $f$ with respect to $q$

$$f_q = \begin{bmatrix} \frac{\partial f_1}{\partial q_1} & \frac{\partial f_1}{\partial q_2} & \dots & \frac{\partial f_1}{\partial q_m} \\ \frac{\partial f_2}{\partial q_1} & \frac{\partial f_2}{\partial q_2} & \dots & \frac{\partial f_2}{\partial q_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial q_1} & \frac{\partial f_m}{\partial q_2} & \dots & \frac{\partial f_m}{\partial q_m} \end{bmatrix} \tag{2.37}$$

If the eigenvectors of $f_q$ form a basis for $q$-vectors the system is said to be hyperbolic, otherwise it is elliptic or parabolic. Note that in this case the eigenvectors typically depend on the variables $q$ themselves so that the same system of equations may be hyperbolic in some regions and elliptic in others. The classification of PDE's as hyperbolic, parabolic and elliptic may be more familiar from the classification of second order equations. Let us show the equivalence of the two usages.

The canonical elliptic second order PDE is the Laplace equation

$$\phi_{tt} + \phi_{xx} = 0\ . \tag{2.38}$$

We reduce it to a system of first-order PDE's by introducing $u = \phi_t,\ v = \phi_x$. The Laplace equation states $u_t + v_x = 0$ and we also have $u_x = v_t$ by the equality of mixed derivatives. These two relations can be written in matrix form as

$$q_t + \mathbf{A} q_x = 0 \tag{2.39}$$

$$q = \begin{bmatrix} u \\ v \end{bmatrix},\ \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = 0 \tag{2.40}$$

The matrix $\mathbf{A}$ has the eigenvalues $\lambda_{1,2} = \pm i$ and eigenvectors $r_{1,2} = \begin{bmatrix} \pm i & 1 \end{bmatrix}$. The eigenvectors $r_{1,2}$ do not form a basis for two-component real vectors such as $q$ so the system is classified as elliptic in accord with the second-order Laplace equation's classification.

The canonic hyperbolic second order PDE is the wave equation

$$\phi_{tt} - \phi_{xx} = 0\ . \tag{2.41}$$

Following the same procedure we arrive at the study of the eigensystem of

$$\mathbf{B} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \tag{2.42}$$

which is given by $\lambda_{1,2} = \pm 1$, $r_{1,2} = \begin{bmatrix} \pm 1 & 1 \end{bmatrix}$. The eigenvectors now do form a basis for two-component real vectors and the system is classified as hyperbolic as expected from the wave equation.

Finally, the typical parabolic equation is

$$\phi_x = \phi_{tt} \tag{2.43}$$

for which we denote $u = \phi_t$ to obtain

$$q_t + Cq_x = \sigma \tag{2.44}$$

with

$$q = \begin{bmatrix} \phi \\ u \end{bmatrix}, \ \mathbf{C} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \ \sigma = \begin{bmatrix} u \\ 0 \end{bmatrix} . \tag{2.45}$$

The eigensystem of $\mathbf{C}$ is $\lambda_{1,2} = 0$, $r_1 = \begin{bmatrix} 0 & 1 \end{bmatrix}$, $r_2 = \begin{bmatrix} 0 & 0 \end{bmatrix}$ which does not form a basis for two-component real vectors. Note that in this case the rank of $\mathbf{C}$ is less than the dimension of the system; this is characteristic of parabolic equations.

2.2.2. *Solution by characteristics.* Let $\mathbf{A}$ be the Jacobian matrix for a non-linear hyperbolic system

$$q_t + \mathbf{A}(q)q_x = 0, \tag{2.46}$$

with $q$ a vector with $m$ components. By the definition of a hyperbolic system we know that $\mathbf{A}$ can be represented as

$$\mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1} . \tag{2.47}$$

The difference with respect to the linear system case is that the matrices $\mathbf{T}$, $\mathbf{\Lambda}$ are no longer constant but depend on $q$ and hence on $(x, t)$. Nevertheless, we can follow the same procedure of reduction to characteristic form locally for some neighborhood of a point $(x_0, t_0)$ where $q(x, t) = q_0$. We can write

$$q(x, t) = q_0 + \tilde{q}(x, t) \tag{2.48}$$

where $\tilde{q}$ is the perturbation from the value $q_0$. System (2.46) can now be written

$$\tilde{q}_t + \mathbf{A}_0\tilde{q}_x = 0 , \tag{2.49}$$

from where we obtain

$$\tilde{w}_t + \mathbf{\Lambda}_0\tilde{w}_x = 0 \tag{2.50}$$

with the perturbation characteristic variables given by

$$\tilde{w} = \mathbf{T}^{-1}\tilde{q}_0 . \tag{2.51}$$

The characteristic system (2.50) leads to the ODE's

$$\frac{d\tilde{w}^{(i)}}{ds_i} = 0, i = 1, \ldots, m . \tag{2.52}$$

where $d/ds_i$ indicates the derivative along the $i^{th}$ characteristic direction whose slope is given by the $\lambda_{0i}$ eigenvalue of $\mathbf{A}_0$

$$\frac{d}{ds_i} = \frac{\partial}{\partial t} + \lambda_{0i}\frac{\partial}{\partial x} . \tag{2.53}$$

A solution to (2.46) can be found by locally solving the ODE's (2.52). This is the *method of characteristics* for non-linear hyperbolic systems.

CHAPTER 7

# Finite volume methods for hyperbolic equations

## 1. Basic aspects

We have seen that the appearance of discontinuities even when starting from smooth initial data is a generic situation for non-linear hyperbolic PDE's. To define what is meant by a solution in such cases the concept of weak solutions was introduced which involved integrating the discontinuous solution over some domain. This suggests that it might be advantageous to construct a numerical method which involves an integration step. It is also the case that a large class of PDE's of practical interest are derived from conservation laws in which a direct expression of the quantity being conserved might prove useful in a numerical algorithm. To evaluate a conserved quantity an integration step is again required.

Let us construct a numerical method based upon the above observations. Consider a scalar hyperbolic equation in conservation form

$$q_t + f(q)_x = 0 \tag{1.1}$$

for which initial data

$$q\left(x, t=0\right) = q_0(x) \tag{1.2}$$

is given over some domain, say $0 \leq x \leq 1$. We divide the domain into a number of cells

$$C_i = \left[x_{i-1}, x_i\right],\ i = 1, \ldots, n \tag{1.3}$$

with $0 = x_0 < x_1 < \ldots < x_n = 1$. From our knowledge of the properties of equations of form (1.1), specifically the existence of characteristics, we expect an explicit time marching scheme to be an efficient numerical procedure. The basic problem faced in constructing such a procedure is to advance from time level $t^n$ to time level $t^{n+1}$. Since we wish to involve conserved quantities in our formulation we are led to integrating (1.1) over $[x_{i-1}, x_i] \times [t^n, t^{n+1}]$

$$\int_{x_{i-1}}^{x_i} \int_{t^n}^{t^{n+1}} \left[q_t + f(q)_x\right] dt dx = 0\ . \tag{1.4}$$

This leads to

$$\int_{x_{i-1}}^{x_i} \left[q(x, t^{n+1}) - q(x, t^n)\right] dx + \int_{t^n}^{t^{n+1}} \left[f(q(x_i, t)) - f(q(x_{i-1}, t))\right] = 0\ . \tag{1.5}$$

We introduce the quantities

$$Q_i^n = \frac{1}{h_i} \int_{x_{i-1}}^{x_i} q(x, t^n) dx,\ i = 1, \ldots, n \tag{1.6}$$

with $h_i = x_i - x_{i-1}$. Each $Q_i^n$ expresses the total amount of the physical quantity $q$ within cell $C_i$ at time $t^n$. We also introduce

$$F_i^n = \frac{1}{k}\int_{t^n}^{t^{n+1}} f(q(x_i,t))dt \tag{1.7}$$

with $k = t^{n+1} - t^n$ which denotes the flux of quantity $q$ through the interface between two cells at $x = x_i$ over the time interval $[t^n, t^{n+1}]$. We thus obtain the scheme

$$Q_i^{n+1} = Q_i^n - \frac{h_i}{k}\left(F_i^n - F_{i-1}^n\right),\ i = 1,\ldots,n\ . \tag{1.8}$$

Such a scheme is known as a *finite volume scheme* and it satisfies our requirements that conserved physical quantities are used and that the derivation of the method allows for discontinuous functions $q(x,t)$. The quantities $Q_i^n$ are known as *cell averages*, and they are taken to represent the value of $q(x,t^n)$ somewhere within cell $C_i$, typicall at the midpoint $(x_{i-1}+x_i)/2$. Our representation of the solution can be interpreted simply as a piecewise constant approximation

$$q(x,t^n) \cong Q_i^n,\ x_{i-1} < x < x_i\ . \tag{1.9}$$

One ingredient is still required to obtain a complete method: a procedure to compute the fluxes $F_i^n$ must be given. We know that hyperbolic equations have finite domains of depedence. If we choose a small enough time step the fluxes should only depend on the cell averages to the left and the right of the interface

$$F_i^n \cong \mathcal{F}(Q_i, Q_{i+1}),\ i = 1,\ldots,n-1\ . \tag{1.10}$$

The function $\mathcal{F}(Q_i, Q_{i+1})$ defines an approximation of the true flux and is known as the *numerical flux* at interface $x_i$. The basic task faced in the construction of a finite volume method is to specify how the numerical fluxes are computed starting from knowledge of the cell average data.

## 2. Godunov methods

One important aspect of finite volume methods is that their construction follows quite closely the physical behavior of the problem solution. This can be exploited further in constructing numerical fluxes. Consider that we have the cell average data $\{Q_i^n\}$. If we interpret this as specifying a piecewise-constant approximation of $q(x,t)$ we note that at each interface $x = x_i$ a Riemann problem has been specified. If the solution $q^\curlyvee(x_i,t)$ to the Riemann problem can be determined then an exact evaluation of the fluxes is possible by evaluating the physical flux function $f$ at $q^\curlyvee(x_i,t)$.

$$F_i^n = f(q^\curlyvee(x_i,t^n)). \tag{2.1}$$

Such methods that use the true physical solution in construction of the numerical fluxes are known as Godunov methods and were introduced to study gas dynamics problems.

A very simple example is given by the scalar, constant-velocity, advection equation

$$q_t + uq_x = 0 \tag{2.2}$$

solved on a uniform grid $x_i - x_{i-1} = h$. The solution to the Riemann problem is immediate

$$q^{\curlyvee}(x_i, t^n) = \begin{cases} Q_i^n & u > 0 \\ Q_{i+1}^n & u < 0 \end{cases} \tag{2.3}$$

leading to the numerical flux

$$f(q^{\curlyvee}(x_i, t^n)) = \begin{cases} uQ_i^n & u > 0 \\ uQ_{i+1}^n & u < 0 \end{cases} \tag{2.4}$$

and the scheme

$$Q_i^{n+1} = \begin{cases} Q_i^n - \frac{uh}{k}\left(Q_i^n - Q_{i-1}^n\right) & u > 0 \\ Q_i^n - \frac{uh}{k}\left(Q_{i+1}^n - Q_i^n\right) & u < 0 \end{cases} . \tag{2.5}$$

We recognize this as the upwind scheme derived in our study of finite difference methods. The interpretation is however a bit different. Whereas $Q_i^n$ in the upwind finite difference scheme denoted the value of $q$ at cell nodes $x_i$, here $Q_i^n$ is the value at cell centers. The difference is however non-essential and we find that a large number of finite volume schemes have a close finite difference equivalent. This is especially useful in determining stability restrictions since we can apply the theory derived for finite difference methods.

In a Godunov method we use the exact physical flux $f(q)$ to evaluate the numerical flux $\mathcal{F}(Q_i, Q_{i+1})$. This still involves an approximation, namely the piecewise constant approximation of $q(x, t^n)$, thus limiting the basic Godunov method to first order accuracy, $\tau_i^n = O(h, k)$. To obtain better accuracy we can introduce more accurate representations of $q(x, t^n)$. An obvious idea is to use a piecewise linear approximation

$$q(x, t^n) = Q_i^n + \sigma_i^n \left(x - x_{i-1/2}\right), \; x_{i-1} < x < x_i \tag{2.6}$$

with $x_{i-1/2} = (x_{i-1} + x_i)/2$. Here we have assumed that $Q_i^n$ represents the value of $q(x, t^n)$ at the midpoint $x_{i-1/2}$. The slope $\sigma_i^n$ may be constructed by interpolation between adjoining cell average values. A number of possibilities exist:

(1) *Downwind* or *Lax-Wendroff* slope

$$\sigma_i^n = \frac{Q_{i+1}^n - Q_i^n}{h} \tag{2.7}$$

(2) *Upwind* or *Beam-Warming* slope

$$\sigma_i^n = \frac{Q_i^n - Q_{i-1}^n}{h} \tag{2.8}$$

(3) *Centered* or *Fromm* slope

$$\sigma_i^n = \frac{Q_{i+1}^n - Q_{i-1}^n}{2h} \tag{2.9}$$

The names used for the slopes refer to the fact that when applied to the constant velocity advection equation each choice of slope leads to the corresponding finite difference scheme.

One problem associated with the desire for higher accuracy is that discontinuities can lead to non-physical oscillations in the numerical approximation. The

FIGURE 1. Non-physical oscillations introduced by piecewise-linear reconstruction. At $t = t^n$ a shock profile at $x = x_i$ is reconstructed using centered slopes. The profile is then advected to the new time level $t^{n+1}$ and used to construct new cell averages, some of which are in error (open circles).

difficulty is easily understood if we consider the Riemann problem for constant-velocity advection

$$q_t + uq_x = 0 \tag{2.10}$$

$$q(x,0) = \begin{cases} q_l & x > x_i \\ q_r & x < x_i \end{cases} \tag{2.11}$$

Reconstruction of the sharp discontinuity leads to overshoots not present in the initial condition. These are then advected downstream and contaminate the numerical solution as shown in Fig. 1.

Non-physical oscillations can lead to a breakdown of the entire computation in some applications. A common example is encountred when $q$ is some positive quantity physically but the numerical scheme oscillations lead to negative values. Typically the physical hypotheses used in constructing the algorithm are no longer valid and a runtime error results. This has led to the search for *high-resolution* algorithms that exhibit higher order accuracy (typically $O(h^2, k^2)$ ) away from discontinuities and capture discontinuities without oscillations. Typically the accuracy of the algorithm is only $O(h, k)$ near a discontinuity but this is not a problem since first-order accuracy is all we could expect. The procedure used in constructing such algorithms rests upon the identification of discontinuities in the initial data. If a discontinuity is identified a low-order, piecewise constant reconstruction of $q(x, t)$ is used. Otherwise a higher-order, say piecewise-linear, reconstruction is used. The technique is known as a *slope-limiter method*, since it attempts to maintain zero slope near discontinuities. Various slope-limiters have been proposed and analyzed:

(1) *minmod limiter*

$$\sigma_i^n = \text{minmod}\left(\frac{Q_i^n - Q_{i-1}^n}{h}, \frac{Q_{i+1}^n - Q_i^n}{h}\right) \tag{2.12}$$

where the minmod function is defined by

$$\text{minmod}(a,b) = \left\{ \begin{array}{ll} a & \text{if } |a| < |b|,\ ab > 0 \\ b & \text{if } |b| < |a|,\ ab > 0 \\ 0 & ab \le 0 \end{array} \right\} \tag{2.13}$$

(2) *monotonized central-difference*

$$\sigma_i^n = \text{minmod}\left(\frac{Q_{i+1}^n - Q_{i-1}^n}{2h}, 2\frac{Q_i^n - Q_{i-1}^n}{h}, 2\frac{Q_{i+1}^n - Q_i^n}{h}\right) \tag{2.14}$$

CHAPTER 8

# Equations of mixed type

## 1. Splitting methods

We have determined numerical methods suitable for various types of PDE's such as diffusion modeled by a parabolic equation or advection modeled by a hyperbolic equation. In very many applications multiple effects are present. A typical example would be the advection-diffusion equation

$$q_t + uq_x + vq_y = \alpha \left( q_{xx} + q_{yy} \right) \ . \tag{1.1}$$

The question naturally arises what to do when faced with the solution of such mixed equations. We could develop methods for each new class encountered or combine the methods already developed for simpler equations. The development of new methods for a certain problem class follows the general procedure introduced so far: a discretization is proposed and then stability and accuracy issues are studied.

We shall show however that it is also possible to combine methods developed for simple PDE's in order to obtain schemes for more complicated problems. A general abstract framework is quite useful in this context. Let $\mathcal{A}, \mathcal{B}$ be any two operators that act upon the unknown function $q$. We assume $q$ satisfies sufficient regularity hypothesis for the problem at hand. We are interested in whether a solution to the combined problem

$$\partial_t q = (\mathcal{A} + \mathcal{B})q \tag{1.2}$$

can be obtained by schemes suited to the simpler problems

$$\partial_t q = \mathcal{A}q \tag{1.3}$$

$$\partial_t q = \mathcal{B}q \tag{1.4}$$

Formally we can solve each of these equations using operator series. The solution to (1.2) is

$$q(t+k) = e^{(\mathcal{A}+\mathcal{B})k} q(t) \tag{1.5}$$

where the exponential of the operators is defined by its series representation

$$e^{(\mathcal{A}+\mathcal{B})k} = I + k(\mathcal{A}+\mathcal{B}) + \frac{k^2}{2!}(\mathcal{A}+\mathcal{B})^2 + \frac{k^3}{3!}(\mathcal{A}+\mathcal{B})^3 + \dots \tag{1.6}$$

We assume that the series above converge.

The procedure proposed to solve the original problem by breaking it down into two steps is:

(1) Solve $\partial_t q = \mathcal{A}q$ to obtain an intermediate value $q^*(t+k)$
(2) Use the intermediate value as an initial condition to the second step $\partial_t q = \mathcal{B}q$.

This can be expressed formally as

$$q^S(t+k) = e^{\mathcal{B}k} e^{\mathcal{A}k} q(t) \tag{1.7}$$

where the $S$ superscript denotes this procedure as a splitting approximation. The question is what is the error introduced by the split method. We therefore evaluate

$$E(t+k) = \mathcal{E}q(t) = \left( e^{(\mathcal{A}+\mathcal{B})k} - e^{\mathcal{B}k} e^{\mathcal{A}k} \right) q(t) \ . \tag{1.8}$$

The error operator $\mathcal{E}$ can be evaluated by taking the series expansions of all terms involved. In doing this we must be careful in the algebraic manipulations since $\mathcal{A}, \mathcal{B}$ do not necessarily commute. For instance if $\mathcal{A} = \partial_x$ and $\mathcal{B} = \alpha(x)\partial_x$ we have

$$\mathcal{A}\mathcal{B} = a'(x)\partial_x + \alpha(x)\partial_{xx} \tag{1.9}$$

but

$$\mathcal{B}\mathcal{A} = \alpha(x)\partial_{xx} \ . \tag{1.10}$$

We have

$$e^{(\mathcal{A}+\mathcal{B})k} = I + k(\mathcal{A}+\mathcal{B}) + \frac{k^2}{2}(\mathcal{A}^2 + \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A} + \mathcal{B}^2) + \ldots \tag{1.11}$$

$$e^{\mathcal{B}k} e^{\mathcal{A}k} = \left( I + k\mathcal{B} + \frac{k^2}{2}\mathcal{B}^2 + \ldots \right) \left( I + k\mathcal{A} + \frac{k^2}{2}\mathcal{A}^2 + \ldots \right) \tag{1.12}$$

$$= I + k(\mathcal{A}+\mathcal{B}) + \frac{k^2}{2}(\mathcal{A}^2 + 2\mathcal{B}\mathcal{A} + \mathcal{B}^2) + \ldots \tag{1.13}$$

The error operator is therefore

$$\mathcal{E} = (\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A}) \frac{k^2}{2} + \ldots = [\mathcal{A}, \mathcal{B}] \frac{k^2}{2} + \ldots \ . \tag{1.14}$$

The quantity $[\mathcal{A}, \mathcal{B}] = \mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A}$ is known as the *commutator* of $\mathcal{A}$, $\mathcal{B}$. We see that the splitting procedure introduces an $O(k^2)$ error in each time step and we take $O(1/k)$ steps so overall it reduces the numerical accuracy to first order irrespective of the accuracy employed in the individual steps.

A way to avoid the degradation of accuracy is to use a slightly more complicated splitting approach known as Strang splitting

$$q^{SS}(t+k) = e^{\mathcal{A}k/2} e^{\mathcal{B}k} e^{\mathcal{A}k/2} q(t) \ . \tag{1.15}$$

A Taylor series expansion of $e^{\mathcal{A}k/2} e^{\mathcal{B}k} e^{\mathcal{A}k/2}$ shows that the leading order error introduced in each time step is $O(k^3)$ for Strang splitting and $O(k^2)$ overall.

CHAPTER 9

# Spectral methods

## 1. Preliminaries

We have so far used Fourier methods in the theoretical analysis of numerical algorithms. However Fourier methods are also very useful in the *construction* of numerical methods for PDE's. By way of an introduction to spectral methods we shall concentrate on solving time dependent problems with periodic boundary conditions over a finite domain which we take to be $[-\pi, \pi]$, $f(x + 2\pi) = f(x)$. If $\|f\|_1 < \infty$ the function $f(x)$ can be expressed as a *Fourier series*

$$f(x) = \sum_{k=-\infty}^{\infty} \hat{f}_k e^{ikx} \tag{1.1}$$

with the Fourier coefficients given by

$$\hat{f}_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx \ . \tag{1.2}$$

This is the discrete wavenumber equivalent of the continuum Fourier transform introduced previously

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(k) e^{i\xi x} d\xi, \ \hat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(x) e^{-i\xi x} d\xi. \tag{1.3}$$

Typically we are interested in finding solutions that are smooth, say of class $C^1$ over $[-\pi, \pi]$ except a finite number of points of discontinuity. The important result from Fourier analysis relevant to this class of problems is:

THEOREM 5. *If $f$ is piecewise smooth with period $2\pi$ then the Fourier series of converges to $f(x)$*

$$\lim_{K \to \infty} \sum_{k=-K}^{K} \hat{f}_k e^{ikx} = f(x) \ . \tag{1.4}$$

*At a point of discontinuity we extend $f$ by the definition $f(x) = (f(x_-) + f(x_+))/2$.*

## 2. Evaluation of derivatives

Assume that $f \in \mathcal{C}^1$ and periodic over $[-\pi, \pi]$ so that both $f$ and $f'$ have convergent Fourier series. The series expansion of $f'$ is

$$f'(x) = \sum_{k=-\infty}^{\infty} d_k e^{ikx} \ . \tag{2.1}$$

Derivation of the Fourier series of $f$ leads to

$$f'(x) = i \sum_{k=-\infty}^{\infty} k \hat{f}_k e^{ikx} \tag{2.2}$$

so we conclude that

$$d_k = ik\hat{f}_k \ . \tag{2.3}$$

If we know the Fourier coefficients of $f$, the Fourier coefficients of the derivatives of $f$ are obtained by multiplication with $ik$. This again is essentially a consequence of the exponential being an eigenfunction of the differentiation operator

$$\partial_x e^{ikx} = ike^{ikx} \ . \tag{2.4}$$

## 3. Discrete Fourier transform

In practical work we can only use a finite number of wave modes and a finite number of function values. The finite version of the Fourier transform is known as the discrete Fourier transform and is given by

$$f_j = \sum_{k=-N/2+1}^{N/2} \hat{f}_k e^{ikx_j} \tag{3.1}$$

$$\hat{f}_k = \frac{1}{N} \sum_{j=0}^{N-1} f_j e^{-ikx_j} \tag{3.2}$$

where the function values are known at $x_j = jh - \pi$, $h = 2\pi/N$, $j = 0, 1, \cdots, N-1$ and the wavenumbers run from $-N/2+1$ to $N/2$. Introducing $\omega_N = e^{2\pi i/N}$, the $N^{th}$ root of unity the transformation formulas can be written

$$f_j = \sum_{k=-N/2+1}^{N/2} \hat{f}_k \omega_N^{jk}, \ \hat{f}_k = \frac{1}{N} \sum_{j=0}^{N-1} f_j \omega_N^{-jk} \ . \tag{3.3}$$

These specify two matrix-vector multiplication operations

$$f = F^{-1}\hat{f}, \ \hat{f} = Ff \tag{3.4}$$

and would seem to require $O(N^2)$ arithmetic operations to carry out. The operation count can be drastically reduced by use of the fast Fourier transform (FFT) algorithm. If $N$ is even, $N = 2P$ we can separate the sequence $\{f_0, f_1, \ldots, f_{2P-1}\}$ into two parts, one containing even indices and the other the odd ones

$$u_j = f_{2j}, \ v_j = f_{2j+1}, \ j = 0, \ldots, P-1 \ . \tag{3.5}$$

The Fourier coeffcients $\hat{f}_k$ are then computed by

$$\hat{f}_k = \frac{1}{2P} \sum_{j=0}^{P-1} \left( u_j \omega_N^{-2jk} + v_j \omega_N^{-(2j+1)k} \right) \tag{3.6}$$

$$= \frac{1}{2P} \left( \sum_{j=0}^{P-1} u_j \omega_P^{-jk} + \omega_N^k \sum_{j=0}^{P-1} v_j \omega_P^{-jk} \right) \tag{3.7}$$

and instead of one matrix-vector multiply costing $O(N^2)$ we obtain two matrix-vector multiplies each costing $O(N^2/4)$ for a total computational effort of $O(N^2/2)$.

FIGURE 1. An example of aliasing. $\sin x$, $\sin 5x$, $\sin 9x$ are the same when sampled on a coarse grid (circles).

The beauty of this approach is that if $N = 2^p$ we can continue the procedure and reduce the operation count down to $O(N \log_2 N)$ a significant improvement over $O(N^2)$. The FFT can be also implemented for $N$ as some composite of other powers of small prime numbers. Through application of the FFT we see that derivatives are economically evaluated using (2.3).

If we do not use a sufficient number of Fourier modes to completely capture all variations in the function, an error known as *aliasing* can occur. The minimum number of points needed to capture all variations in a function is given by the *Nyquist criterion*. If $K$ is the highest wavenumber present in function $f$ then we need at least $N = 2K$ points to completely represent $f$ through a discrete Fourier series. Aliasing is easily understood as an artefact of too coarse a sampling of a function. Consider for instance the case when $N = 4$ so the grid nodes are $x_j = -\pi + j\pi/2$, $j = 0, \ldots, 3$. Note that $\sin x$, $\sin 5x$, $\sin 9x$ are indistinguishable on this grid, Fig. 1. Aliasing leads to contamination of low wavenumber modes from higher wavenumbers present in $f$ that are not resolved by the chosen grid resolution. The errors thus introduced can be quite significant since the low wavenumber modes govern the coarse features of $f$. There exist a number of techniques to eliminate aliasing. Besides the obvious one of ensuring the Nyquist criterion is met filtering the high wavenumbers or using higher resolution in certain stages of a computation are also used.

## 4. Applications to PDE's

Fourier methods are especially useful in solving problems for which we know the solution is smooth. This is a result of a number of results from Fourier analysis. If $f$ is $L^2$ then we can state:

THEOREM 6. *If $f$ has $p-1$ continuous derivatives in $L^2$ for $p \geq 0$ and a $p^{th}$ derivative of bounded variation then*

$$\hat{f}(\xi) = O(|\xi|^{-p-1}) \text{ as } |\xi| \to \infty \ . \tag{4.1}$$

This statement tells us how the Fourier coefficients decay at high wavenumbers. In particular for infinitely differentiable functions the Fourier coefficents decay faster than any polynomial power. This behavior implies that once enough Fourier modes have been introduced to capture the characteristic scales of $f$ the amplitude of higher modes is essentially zero. For numerical work we expect that a finite number of modes will essentially reproduce the exact behavior of the smooth function and its derivatives. It is the combination of accuracy and ease of evaluation of derivatrives that make Fourier methods valuable in solving PDE's numerically. Note however that this is generally the case only when we have smooth functions and for very special boundary conditions such as the periodic boundary conditions considered here.

Let us now sketch how various common PDE's can be solved through Fourier methods. In all cases we shall consider periodic boundary conditions over the interval $[-\pi, \pi]$ along each spatial direction.

4.0.3. *Advection equation.* For the problem

$$q_t + uq_x = 0 \tag{4.2}$$

$$q(x, t = 0) = q_0(x) \tag{4.3}$$

with $u$ constant, we introduce the Fourier series

$$q_j(t) = \sum_{k=-N/2+1}^{N/2} \hat{q}_k(t)\omega_N^{jk} \tag{4.4}$$

giving the values of $q$ at the points $x_j = -\pi + jh$, $h = 2\pi/N$. This is a semi-discretization formulation in which time is kept as a continuous variable at this stage. The values of the $x$-derivative at $x_j$ are given by

$$\left(\frac{\partial q}{\partial x}\right)_j (t) = \sum_{k=-N/2+1}^{N/2} ik\hat{q}_k(t)\omega_N^{jk} \tag{4.5}$$

The powers of $\omega_N$ form a basis for grid functions so replacing the series into (4.2) leads to

$$\frac{d}{dt}\hat{q}_k + iuk\hat{q}_k = 0 \tag{4.6}$$

$$\hat{q}_k(t = 0) = \hat{q}_{0,k} \tag{4.7}$$

for $k = -N/2 + 1, \ldots, N/2$ with $\hat{q}_{0,k}$ the Fourier coefficents of the initial condition. For real $q$ we have $\hat{q}_k = (\hat{q}_{-k})^*$ so the complete Fourier coefficients can be obtained from a knowledge of the positive wavenumbers only. The system of ODE's is thus reduced to $k = 0, 1, \ldots, N/2$. Each equation can be solved analytically

$$\hat{q}_k(t) = \hat{q}_{0,k} \exp(-iukt) \tag{4.8}$$

and the problem is solved. Note that if $q$ is $C^\infty$ and $N$ large enough to capture all the modes present in $q_0(x)$ the solution is essentially exact and we would expect to observe errors on the order of machine zero when carrying out this computation in practice.

For the 2D adevction equation

$$q_t + uq_x + vq_y = 0 \tag{4.9}$$

$$q(x, y, t = 0) = q_0(x, y) \tag{4.10}$$

the procedure is quite similar. We shall employ a double Fourier series representation

$$q_{m,n} = \sum_{k=-M/2+1}^{M/2} \sum_{l=-N/2+1}^{N/2} \hat{q}_{k,l}(t)\omega_M^{mk}\omega_N^{nl} \tag{4.11}$$

to obtain the system

$$\frac{d}{dt}\hat{q}_{k,l} + i(uk + vl)\hat{q}_{k,l} = 0 \tag{4.12}$$

$$\hat{q}_{k,l}(t = 0) = \hat{q}_{0,k,l} \tag{4.13}$$

again easily solvable.

4.0.4. *Diffusion equation.* For the diffusion equation

$$q_t = \alpha q_{xx} \tag{4.14}$$

$$q(x, t = 0) = q_0(x) \tag{4.15}$$

the same procedure leads to the system of ODE's

$$\frac{d}{dt}\hat{q}_k = -\alpha k^2 \hat{q}_k \tag{4.16}$$

$$\hat{q}_k(t = 0) = \hat{q}_{0,k} \tag{4.17}$$

In the 2D case

$$q_t = \alpha(q_{xx} + q_{yy}) \tag{4.18}$$

$$q(x, y, t = 0) = q_0(x, y) \tag{4.19}$$

we obtain

$$\frac{d}{dt}\hat{q}_{k,l} = -\alpha(k^2 + l^2)\hat{q}_{k,l} \tag{4.20}$$

$$\hat{q}_{k,l}(t = 0) = \hat{q}_{0,k,l} \tag{4.21}$$

Note that solving this system of ODE's is quite easy. Compare with the necessity of solving an implicit system of $MN$ equations that would be obtained if we use a standard finite difference formulation such as Crank-Nicolson.

4.0.5. *Variable velocity advection.* Let us now consider

$$q_t + u(x)q_x = 0 \tag{4.22}$$

$$q(x, t = 0) = q_0(x) \tag{4.23}$$

Here things get more complicated since we have to introduce a Fourier series for $u(x)$ also to mimic the procedure followed above. This however would lead to a convolution product in Fourier space and the system of ODE's

$$\frac{d}{dt}\hat{q}_k + i \sum_{l+m=k} \hat{u}_l \hat{q}_m = 0 \tag{4.24}$$

and the solution of this system is no longer immediate; we need to also solve a dense linear system. This costs $O((N/2)^3/3)$ much more than the Fourier transforms or the $O(N)$ cost we would expect from a finite difference method. Instead of adopting this procedure we can carry out the following operations to advance our numerical solution from $\left\{Q_j^n\right\}$ to $\{Q_j^{n+1}\}$ (we have reverted to the $Q$ notation since the method is now fully discretized and we no longer will be able to solve the ODE systems that arise analytically):

(1) Compute $\left\{\hat{Q}_k^n\right\}$ from $\left\{Q_j^n\right\}$
(2) Compute the Fourier coefficients of the derivative $q_x$, $\left\{ik\hat{Q}_k^n\right\}$
(3) Carry out the inverse Fourier transform to find $\left\{(Q_x)_j^n\right\}$. We have at this stage completed the evaluation of the derivatives, $q_x$ through a process known as numerical spectral differentiation.
(4) Compute $c_j = u(x_j)\,(Q_x)_j^n$ for $j = 0, \ldots, N-1$
(5) Find the Fourier coefficients of the $\{c_j\}$ grid function, $\{\hat{c}_k\}$

(6) Solve the system of ODE's

$$\frac{d}{dt}\hat{q}_k + ik\hat{c}_k = 0 \tag{4.25}$$

$$\hat{q}_k(0) = \hat{Q}_k^n \tag{4.26}$$

over a time step $\Delta t$

$$\hat{q}_k(t^{n+1}) = \hat{Q}_k^n \exp\left(-ik\hat{c}_k\Delta t\right) \tag{4.27}$$

This is known as a pseudo-spectral method since we work both in spectral space to evaluate derivatives and in real space to evaluate products. There arises the problem that the product $u_j(Q_x)_j^n$ might be affected by aliasing errors. This is avoided typically by using a higher resolution at this stage of the algorithm, $3N/2$ instead of $N$ points being used to sample $c_j$.

4.0.6. *2D incompressible Navier-Stokes equations in $\omega - \psi$ formulation.* Let us conclude with a realistic practical example. The 2D incompressible Navier-Stokes equations

$$u_x + v_y = 0 \tag{4.28}$$

$$u_t + uu_x + vu_y = -p_x + \alpha\left(u_{xx} + u_{yy}\right) \tag{4.29}$$

$$v_t + uv_x + vv_y = -p_y + \alpha(v_{xx} + v_{yy}) \tag{4.30}$$

describe viscous fluid flow with velocity $(u, v)$ and pressure $p$. They are a widely used model in weather prediction in which periodic boundary conditions apply. The system of 3 PDE's can be reduced to 2 equations through use of the vorticity ($\omega$) stream function ($\psi$) formulation. The vorticity is defined as the curl of the velocity field. For a 2D flow there is only one non-zero component, perpendicular to the plane of flow

$$\omega = v_x - u_y \tag{4.31}$$

The streamfunction is defined by

$$\psi_y = u,\ \psi_x = -v \tag{4.32}$$

and is constant along a streamline of flow. Taking $\partial_y$ of (4.29) and -$\partial_x$ of (4.30) and adding the result leads to the vorticity transport equation

$$\omega_t + u\omega_x + v\omega_y = \alpha(\omega_{xx} + \omega_{yy})\ . \tag{4.33}$$

The vorticity can be expressed in terms of the stream function

$$\omega = (-\psi_x)_x - (\psi_y)_y \tag{4.34}$$

or

$$\nabla^2\psi = -\omega \tag{4.35}$$

Note that velocities obtained from a stream function automatically satisfy the continuity equation (4.28)

$$u_x + v_y = \psi_{yx} - \psi_{xy} = 0\ . \tag{4.36}$$

We can solve (4.33) and (4.35) by the following algorithm:

(1) From the current approximation of the vorticity field $\{\Omega_{ij}^n\}$ compute the Fourier transform $\{\hat{\Omega}_{kl}^n\}$

(2) Solve the Poisson equation for the stream function to find

$$\hat{\Psi}^n_{kl} = \frac{1}{k^2 + l^2}\hat{\Omega}^n_{kl} \tag{4.37}$$

(3) Evaluate the derivatives of the stream function needed to compute the velocities

$$\hat{U}^n_{kl} = il\hat{\Psi}^n_{kl},\ \hat{V}^n_{kl} = -ik\hat{\Psi}^n_{kl} \tag{4.38}$$

(4) Apply the inverse Fourier transform to find the velocity field in real space $\{U^n_{ij}\}$, $\{V^n_{ij}\}$

(5) Compute the derivatives of the vorticity in Fourier space

$$ik\hat{\Omega}^n_{kl},\ il\hat{\Omega}^n_{kl} \tag{4.39}$$

(6) Use the inverse Fourier transform to real space and obtain $\{(\Omega_x)^n_{ij}\}$, $\{(\Omega_y)^n_{ij}\}$

(7) Compute the convection term in real space

$$C^n_{ij} = U^n_{ij}\,(\Omega_x)^n_{ij} + V^n_{ij}\,(\Omega_y)^n_{ij} \tag{4.40}$$

(8) Fourier transform the convection term $\{\hat{C}^n_{kl}\}$

(9) Apply an ODE solver to advance the vorticity forward in time by solving

$$\frac{d}{dt}\hat{\Omega}_{kl} + \hat{C}_{kl} = -\alpha(k^2 + l^2)\hat{\Omega}_{kl} \tag{4.41}$$

The evaluation of the convective term can potentially introduce aliasing errors so this is carried out either with filtering or on an extended grid.

CHAPTER 10

# Finite difference methods revisited

## 1. Compact finite difference schemes

**1.1. Construction by Taylor series expansions.** Spectral methods have been shown to be very accurate methods in the computation of smooth solutions to PDE's. For a wide class of problems we are interested in high order approximation of the solution to a PDE but are not able to use a spectral method. Typical difficulties involve the appearance of complicated boundary conditions or some limited region of rapid variation in the solution which should be treated separately from the entire computational domain. Spectral methods are global methods and hence difficult to adapt locally. This difficulty has become overcome in recent years through the spectral element method but significant complications of method arise.

From spectral methods we can notice that the accurate evaluation of derivatives is based upon knowledge of the entire grid function not just a few adjoining values. This leads to the idea of extending a finite difference stencil to include more points. As we do this however the stencil becomes unwieldy encompessing perhaps a significant portion of the computational domain. We are interested in more accurate evaluations of the derivative but would like to keep the stencil compact. This has led to the application of Hermite polynomial interpolation ideas in which we seek the derivatives of a function $f(x)$ at the nodes of a uniform grid $\{x_j\}$ through formulas of the type

(1.1)
$$\frac{a}{2h}\left(f_{j+1}-f_{j-1}\right)+\frac{b}{4h}\left(f_{j+2}-f_{j-2}\right)+\frac{c}{6h}\left(f_{j+3}-f_{j-3}\right)=\beta f'_{j-2}+\alpha f'_{j-1}+f'_j+\alpha f'_{j+1}+\beta f'_{j+2}$$

Note that the computational stencil is widened and we have to solve a banded system to find the approximations of the derivative.

Taylor series expansions lead to relations among the coefficients introduced above

$$O(1): a+b+c=2\beta+2\alpha+1 \tag{1.2}$$

$$O(h^2): \frac{a}{6}+\frac{2b}{3}+\frac{3c}{2}=4\beta+\alpha \tag{1.3}$$

$$O(h^4): \frac{a}{120}+\frac{2b}{15}+\frac{27c}{40}=\frac{4\beta}{3}+\frac{\alpha}{12} \tag{1.4}$$

$$\cdots \tag{1.5}$$

The powers of $h$ corresponding to each relation are indicated in (1.2-1.4). We obtain second order accuracy if we apply just the first relation (1.2), $O(h^4)$ accuracy if we apply the first 2 formulas and so on. Each relation adds another restriction to the values of available coefficients. There are 5 unknown coefficients in the set of equations above. For example, choosing just the first three relations leads to a two

parameter family

$$a = \frac{1}{6}\alpha - \frac{10}{3}\beta + \frac{3}{2}, b = \frac{32}{15}\alpha + \frac{62}{15}\beta - \frac{3}{5}, c = \frac{6}{5}\beta - \frac{3}{10}\alpha + \frac{1}{10} \ . \tag{1.6}$$

In such a family of solutions we can impose additional constraints. Should we wish to solve a tridiagonal system for the derivatives we would impose $\beta = 0$ and obtain

$$a = \frac{1}{6}\alpha + \frac{3}{2}, \ b = \frac{32}{15}\alpha - \frac{3}{5}, \ c = -\frac{3}{10}\alpha + \frac{1}{10} \ . \tag{1.7}$$

Conversely, we might wish to obtain the smallest possible stencil for a given order of accuracy. This would lead to imposing $c = 0$ and the family

$$a = \frac{16}{9} - \frac{2}{3}\alpha, \ b = \frac{19}{6}\alpha - \frac{17}{18} \tag{1.8}$$

with $\beta = (3\alpha - 1)/12$. Thus we obtain a famil

Second derivatives are constructed throough a similar procedure

(1.9)

$$\frac{a}{h^2}\left(f_{j+1} - 2f_j + f_{j-1}\right) + \frac{b}{4h^2}\left(f_{j+2} - 2f_j + f_{j-2}\right) + \frac{c}{9h^2}\left(f_{j+3} - 2f_j + f_{j-3}\right) = \beta f''_{j-2} + \alpha f''_{j-1} + f''_j + \alpha f''_{j+1} + \beta f''_{j+2}$$

Taylor series expansions lead to the system

$$a + b + c = 2\alpha + 2\beta + 1 \tag{1.10}$$

$$\frac{a}{12} + \frac{b}{3} + \frac{3c}{4} = \alpha + 4\beta \tag{1.11}$$

$$\frac{a}{360} + \frac{2b}{45} + \frac{9c}{40} = \frac{\alpha}{12} + \frac{4\beta}{3} \tag{1.12}$$

with the two-parameter solution family

$$a = \frac{3}{2} - 3\beta - \frac{9}{4}\alpha, b = \frac{24}{5}\alpha - \frac{6}{5}\beta - \frac{3}{5}, c = \frac{31}{5}\beta - \frac{11}{20}\alpha + \frac{1}{10} \ . \tag{1.13}$$

**1.2. Fourier analysis of compact finite difference schemes.** The benefits of the procedure outlined above are shown by Fourier analysis of the ensuing finite difference formulas. Consider a periodic function $f(x + 2\pi) = f(x)$ and a uniform discretization $x_j = -\pi + jh$, $h = 2\pi/N$. The finite wavenumber, discrete Fourier series representing $f$ on this grid is

$$f_j = \sum_{k=-N/2+1}^{N/2} \hat{f}_k \omega_N^{jk} \tag{1.14}$$

with $\omega_N = exp(2\pi i/N)$, the $N^{th}$ root of unity. The full Fourier series representing the function is

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} \tag{1.15}$$

and that representing its derivative is

$$f'(x) = \sum_{k=-\infty}^{\infty} ikc_k e^{ikx} \ . \tag{1.16}$$

Assume that $N$ is large enough so all Fourier components of $f$ and $f'$ are captured on the finite representation and there is no aliasing error, i. e. $\hat{f}_k = c_k$ for $k = -N/2+1, \ldots, N/2$ and the derivative can be expressed as

$$f'_j = \sum_{k=-N/2+1}^{N/2} ik\hat{f}_k \omega_N^{jk} \tag{1.17}$$

We pose the problem of how well the Fourier representation of the derivative is approximated by various finite difference formulas. For instance the standard $O(h^2)$ centered formula leads to

$$f'_j \cong \frac{f_{j+1} - f_{j-1}}{2h} = \sum_{k=-N/2+1}^{N/2} \frac{i \sin kh}{h} \hat{f}_k \omega_N^{jk} \tag{1.18}$$

so the factor multiplying each Fourier coefficent is $i\left(\sin kh\right)/h$ instead of the exact factor $ik$. We can form a relative error

$$e(\kappa) = \left| \frac{\sin\kappa - \kappa}{\kappa} \right| = \left| \frac{A(\kappa) - \kappa}{\kappa} \right| \tag{1.19}$$

with $\kappa = kh$. A number of such expressions shall arise which will differ by $A(\kappa)$, the particular finite difference approximation of the correct Fourier differencing coefficient. For the family of schemes that result from imposing (1.1) we obtain

$$A(\kappa) = \frac{a\sin\kappa + (b/2)\sin 2\kappa + (c/3)\sin 3\kappa}{1 + 2\alpha\cos\kappa + 2\beta\cos 2\kappa} \ . \tag{1.20}$$

For a formula to reproduce as accurately as possible the Fourier representation of the derivative we would want $A(\kappa) = \kappa$. This may be used to impose additional constraints on the coefficients $a, b, c, \alpha, \beta$. For instance we could construct the sum over the represented wavenumbers

$$S = \sum_{k=-N/2+1}^{N/2} \left[ \frac{a\sin kh + (b/2)\sin 2kh + (c/3)\sin 3kh}{1 + 2\alpha\cos kh + 2\beta\cos 2kh} - kh \right]^2 \tag{1.21}$$

and minimize the this sum with respect to some of the coefficients to obtain additional restrictions.

A graphical representation of $A(\kappa)$ versus $\kappa$ is most useful in interpretation of the quality of various finite difference approximations in Fourier space. Let us write down some particular $A(\kappa)$ expressions:

(1) Standard, $O(h^2)$ centered finite difference formula

$$A_1(\kappa) = \sin\kappa \tag{1.22}$$

(2) $O(h^4)$, compact stencil, tridiagonal system for $f'$

$$a = \alpha = -1, \ b = c = \beta = 0 \tag{1.23}$$

$$A_2(\kappa) = \frac{3\sin\kappa}{2 + \cos\kappa} \tag{1.24}$$

(3) $O(h^6)$, tridiagonal system for $f'$

$$A_3(\kappa) = \frac{1}{3}\left(4 - \cos\kappa\right)\sin\kappa \tag{1.25}$$

FIGURE 1. Accuracy in capture of wavemodes present in $f'$ using various finite difference formulas. $A_1$ - circles, $A_2$ - x, $A_3$ - diamond.

This is shown in Fig. 1. The behavior of the exact Fourier coefficients of the derivative is shown by the $x = y$ line. We see that the standard $O(h^2)$ centered formula leads to large errors starting at $\kappa \cong \pi/6$. The full range of wave numbers resolvable by the grid is from $\kappa = 0$ to $\kappa = \pi$. The largest wavenumber corresponds to the Nyquist criterion. We can interpret these findings as suggesting that when using the standard centered formula we are able to accurately resolve Fourier components if they are sampled with 6 times the number of points suggested by the Nyquist criterion, 12 points instead of 2. The formula corresponding to $A_2$ shows much better behavior, requiring only 6 sample points per wavelength, still more than the theoretical minimum given by the Nyquist criterion. The higher order compact finite difference formulas, optimized to reproduce the derivatives spectral behavior do reach the Nyquist limit, thus reproducing the desirable behavior of the spectral schemes.

CHAPTER 11

# Finite element methods

## 1. Preliminaries

For a number of applications the restrictions imposed by finite difference or spectral methods with respect to the computational grid are too severe. This is especially the case in structural engineering where the elasticity equations are solved in domains of complicated geometry such as the interior of an automobile engine. A review of the finite difference and spectral methods would show that the reason relatively simple grids are required is that the differential form of the equation is used. Finite volume methods had no such restriction since they used an integral formulation, and indeed complicated geometries may be treated by finite volume methods. Another class of methods which are based upon an integral formulation are the finite element and closely related boundary element methods. We shall concentrate on finite element methods for now.

The basic idea behind the finite element methods is to employ a piecewise local approximation $\tilde{q}$ of the unknown function $q$ that satisfies some PDE of interest. The piecewise local approximation is defined over some general discretization of the domain of definition of $q$ denoted by $\Omega$. Instead of directly using the piecewise local approximation in the PDE we employ a weighted residual formulation. There arises the significant question of how to best relate the integral formulation to the PDE of interest. Once the discretization, piecewise local approximation and integral formulation are determined a system of equations is obtained whose solution gives the complete approximation to the problem of interest. We shall look at each of these components in detail.

**1.1. Spatial discretizations.** A domain $\Omega$ may be discretized into simple elements in very many ways. Nonetheless only a few are typically used in practice. General affine geometry furnishes some guidance for general discretization techniques. We know for instance that any $d-$dimensional domain may be expressed as a reunion of simplicia

$$\Omega = \cup_k S_k \ . \tag{1.1}$$

Simplicia are the simplest continuum geometric entitites one can construct in a space of dimension $d$. For 1D spaces the simplicia are line segments. In 2D they are triangles and in 3D they are tetrahedra. The measure of each of these elements is easily determined by the formulas:

(1) line segment in 1D of nodes $\{x_1, x_2\}$

$$l = x_2 - x_1 = \begin{vmatrix} 1 & 1 \\ x_1 & x_2 \end{vmatrix} \tag{1.2}$$

FIGURE 1. Example of the discretization into triangles of the domain between a circle and a NACA-0012 airfoil.

(2) triangle in 2D with nodes $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$

$$A = \frac{1}{2}\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} \tag{1.3}$$

(3) tetrahedron in 3D with nodes $\{(x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3)\}$

$$V = \frac{1}{6}\begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_3 & z_3 & z_4 \end{vmatrix} \tag{1.4}$$

In the above formulas the element measures are given with sign, the sign corresponding to orientation of the nodes. We understand that the positive value is to be taken whenever a true geometric measure (length, area, volume) is required. Simplicia have many attractive theoretical properties, in particular there exists a definition of what an optimal discretization is for a number of PDE's of interest, especially elliptic PDE's such as the Poisson equation. Fig. 1 shows an example of such a discretization.

Another widely used discretization is into generalized polyhedra having $2d$ sides, i.e. line segments in 1D, quadrilaterals in 2D, hexahedra in 3D. These have the advantage of enabling easier organization of programs since there is a natural ordering of the indices identifying each element. Thus discretizations which use these types of elements give rise to *structured computational grids*, similar to those encountered in finite difference methods whereas discretizations using simplicia lead to *unstructured computational grids*.

**1.2. Piecewise interpolations.** Once a discretization scheme for the geometric domain has been established the next step is to define a local approximation of $q$ over the element $E$. Typically the approximation is an interpolation based upon values $Q_j$ defined somewhere within the element $E$, but this is not obligatory and other approximations (spectral elements, Chebyshev elements) may be used. The position where the values $Q_j$ are to be defined must be established. A simple choice is the element nodes but again this is not obligatory and the values may be positioned at other points within $E$. Finally an interpolation scheme must be established such as polynomial interpolation. Let us give some typical examples:

1.2.1. *Linear elements in 1D.* The element $E$ has two nodes $\{x_1, x_2\}$, $x_2 > x_1$. Values representing $q(x)$ are defined at the nodes $\{Q_1, Q_2\}$. These define a linear polynomial approximation valid over $E$

$$\tilde{q}(x) = \frac{(x - x_1)Q_2 + (x_2 - x)Q_1}{x_2 - x_1} = N_1(x)Q_1 + N_2(x)Q_2 \tag{1.5}$$

The functions $N_1(x)$, $N_2(x)$ have properties reminiscent of the Dirac delta

$$N_1(x_1) = 1,\ N_1(x_2) = 0 \tag{1.6}$$

$$N_2(x_1) = 0,\ N_2(x_2) = 1 \tag{1.7}$$

FIGURE 2. Linear 1D form functions.

FIGURE 3. Quadratic element form functions in 1D.

and are called *form functions*. The particular ones used here are called the 1D linear form functions and are depicted in Fig. (2)

1.2.2. *Quadratic elements in 1D.* The element $E$ has three nodes $\{x_1, x_2, x_2\}$ and the local approximation is

$$\tilde{q}(x) = N_1(x)Q_1 + N_2(x)Q_2 + N_3(x)Q_3 \tag{1.8}$$

with the form functions

$$N_1(x) = \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)} \tag{1.9}$$

$$N_2(x) = \frac{(x-x_3)(x-x_1)}{(x_2-x_3)(x_2-x_1)} \tag{1.10}$$

$$N_3(x) = \frac{(x-x_1)(x-x_3)}{(x_3-x_1)(x_3-x_2)} \tag{1.11}$$

1.2.3. *Linear elements on triangles in 2D.* The element $E$ has 3 nodes of coordinates $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ at which the values $Q_1, Q_2, Q_3$ are defined. The local approximation of $q$ is given by

$$\tilde{q}(x, y) = N_1(x, y)Q_1 + N_2(x, y)Q_2 + N_3(x, y)Q_3 \tag{1.12}$$

with the form functions

$$N_1(x,y) = \frac{1}{2A}\begin{vmatrix} 1 & 1 & 1 \\ x & x_2 & x_3 \\ y & y_2 & y_3 \end{vmatrix} = \frac{1}{2A}\left(xy_2 - yx_2 - xy_3 + yx_3 + x_2y_3 - x_3y_2\right) \tag{1.13}$$

$$N_2(x,y) = \frac{1}{2A}\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x & x_3 \\ y_1 & y & y_3 \end{vmatrix} = \frac{1}{2A}\left(yx_1 - xy_1 + xy_3 - yx_3 - x_1y_3 + y_1x_3\right) \tag{1.14}$$

$$N_3(x,y) = \frac{1}{2A}\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x \\ y_1 & y_2 & y \end{vmatrix} = \frac{1}{2A}\left(xy_1 - yx_1 - xy_2 + yx_2 + x_1y_2 - x_2y_1\right) \tag{1.15}$$

Notice how the properties of simplicia enable the form functions to be easily determined.

1.2.4. *Linear along each direction elements on quadrilaterals in 2D.* The element $E$ has 4 nodes $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4\}$. It is convenient to introduce a local coordinate system $(\xi, \eta)$ so that the nodes correspond to the local coordinates $(\pm 1, \pm 1)$. The local approximation is then given in the local coordinates

by

$$\tilde{q}(\xi,\eta) = \sum_{k=1}^{4} N_k(\xi,\eta) Q_k \tag{1.16}$$

with the form functions

$$N_1(\xi,\eta) = \frac{1}{4}(1+\xi)(1+\eta) \tag{1.17}$$

$$N_2(\xi,\eta) = \frac{1}{4}(1-\xi)(1+\eta) \tag{1.18}$$

$$N_3(\xi,\eta) = \frac{1}{4}(1-\xi)(1-\eta) \tag{1.19}$$

$$N_4(\xi,\eta) = \frac{1}{4}(1+\xi)(1-\eta) \tag{1.20}$$

The local transformation of coordinates can also be written in terms of the form functions

$$x(\xi,\eta) = \sum_{k=1}^{4} N_k(\xi,\eta) x_k,\ y(\xi,\eta) = \sum_{k=1}^{4} N_k(\xi,\eta) y_k \tag{1.21}$$

## 2. Variational derivation of weighted residual formulations

We now turn to the problem of how to obtain a measure of the error introduced in approximating the exact solution $q$ to the PDE of interest with its piecewise approximation $\tilde{q}$. Some techniques were presented in the general presentation of weighted residual methods carried out in Chapter 2. For a wide class of problems of interest, especially elliptic problems there exist alternative formulations that lead to more efficient numerical algorithms. These are based upon variational and functional analysis and we shall consider the basics of the theory here.

**2.1. Variational calculus.** Consider the problem of determining the extremum of the integral

$$I(q) = \int_a^b f(x,q,q')dx \tag{2.1}$$

over all functions $q : \mathbb{R} \to \mathbb{R}$ that belong to some class, for example piecewise continuous functions and that satisfy the boundary conditions $q(x = a) = q_a$, $q(x = b) = q_b$. $I(q)$ is called a functional in that it associates a scalar value to each element from a space of functions. We can consider small perturbations of the function $q$ that we denote by $\delta\dot{q}$ The perturbations maintain the boundary conditions, i.e.

$$\delta q(x = a) = 0,\ \delta q(x = b) = 0\ . \tag{2.2}$$

The change in $I$ is

$$\delta I = I(q+\delta q) - I(q) = \int_a^b f(x, q+\delta q, q'+\delta q')dx - \int_a^b f(x,q,q')dx\ . \tag{2.3}$$

We shall consider $q$, $q'$ as independent variables in $f$ and carry out series expansions to obtain

$$\delta I = \int_a^b \left[ \left( \frac{\partial f}{\partial q} \right) \delta q + \left( \frac{\partial f}{\partial q'} \right) \delta q' \right] dx\ . \tag{2.4}$$

We can interchange the $\delta$ and $d/dx$ operators in the second term and then integrate by parts

$$\int_a^b \left(\frac{\partial f}{\partial q'}\right) \delta q' dx = \int_a^b \left(\frac{\partial f}{\partial q'}\right) \delta \frac{dq}{dx} dx = \int_a^b \left(\frac{\partial f}{\partial q'}\right) \frac{d}{dx} (\delta q)\, dx = \tag{2.5}$$

$$= \left(\frac{\partial f}{\partial q'}\right) (\delta q)\Big|_{x=a}^{x=b} - \int_a^b \frac{d}{dx}\left(\frac{\partial f}{\partial q'}\right) (\delta q)\, dx \tag{2.6}$$

Applying the boundary conditions and then replacing the above result in (2.4) leads to

$$\delta I = \int_a^b \left[\left(\frac{\partial f}{\partial q}\right) - \frac{d}{dx}\left(\frac{\partial f}{\partial q'}\right)\right] \delta q dx\ . \tag{2.7}$$

For $I$ to be at an extremum $\delta I$ must maintain the same sign under any perturbation of the extremum. This is only possible if the factor multiplying $\delta q$ in the above integral is zero everywhere. If it were not then $\delta q_1$ would give some value $\delta I_1$ and $-\delta q_1$ would lead to the opposite value $-\delta I_1$ and $I$ would not be at an extremum. We therefore have

$$\left(\frac{\partial f}{\partial q}\right) - \frac{d}{dx}\left(\frac{\partial f}{\partial q'}\right) = 0 \tag{2.8}$$

as the condition for $I$ to be at an extremum. This is known as the *Euler variational principle*. At the extremum we obviously have $\delta I = 0$.

The importance of the Euler variational principle for numerical solution of PDE's rests upon the link it furnishes between an integral formulation $I(q)$ and a differential equation (2.8). We can write down specific forms of $f$ that lead to PDE's of great practical interest. For example replacing

$$f(x, q, q') = \frac{1}{2}\left(\frac{dq}{dx}\right)^2 - gq \tag{2.9}$$

in (2.8) leads to the differential equation

$$q'' = g \tag{2.10}$$

with the boundary conditions $q(x = a) = q_a$, $q(x = b) = q_b$. This is the standard 2 point boundary problem for a second order ODE. Recall that this can be solved by either direct discretization leading to the linear system of equations

$$Q_{j-1} - 2Q_j + Q_{j+1} = h^2 g_j,\ j = 1, \ldots, N-1 \tag{2.11}$$

or by using a shooting method combined with an initial value solve in which we seek $z = q'(x = a)$ that leads to $q(x = b; z) = q_b$. The variational formulation above suggests a third approach. Instead of directly solving the ODE we can seek $q$ that minimizes $I(q)$ with $f$ given by (2.9). This is extremely useful in constructing finite element approximations as we will see below.

Other important expressions of the Euler variational principle can be derived for various situations. Let us consider the ones most often encountered.

(1) Functional of two functions in 1D. The functional is

$$I(p, q) = \int_a^b f(x, p, p', q, q') dx \tag{2.12}$$

and the Euler variational principle leads to

$$\left(\frac{\partial f}{\partial p}\right) + \left(\frac{\partial f}{\partial q}\right) - \frac{d}{dx}\left(\frac{\partial f}{\partial p'}\right) - \frac{d}{dx}\left(\frac{\partial f}{\partial q'}\right) = 0 \tag{2.13}$$

(2) Functional of a 2D function.

$$I(q) = \int_c^d \int_a^b f(x, y, q, q_x, q_y) dx dy \tag{2.14}$$

$$\left(\frac{\partial f}{\partial q}\right) - \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial q_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial q_y}\right) = 0 \tag{2.15}$$

(3) Functional involving second order derivatives in 1D.

$$I(p, q) = \int_a^b f(x, p, p', p'') dx \tag{2.16}$$

$$\left(\frac{\partial f}{\partial q}\right) - \frac{d}{dx}\left(\frac{\partial f}{\partial q'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial f}{\partial q''}\right) = 0 \tag{2.17}$$

**2.2. Ritz methods.** In the Ritz formulation of the finite element method we seek a piecewise approximation that minimizes the functional associated with the PDE of interest. The piecewise local approximation can be expressed as

$$\tilde{q}(x) = \sum_e \sum_k Q_k^e N_k^e(x) \tag{2.18}$$

where the $e$ sum is over all elements and the $k$ sum is over all nodes within an element. The unknowns of the problem are the nodal values $Q_k^e$. The form functions $N_k^e(x)$ correspond to some chosen approximation scheme. Let $f$ be associated with the PDE we are interested in solving. The problem reduces to finding $\{Q_k^e\}$ that minimizes

$$I(\tilde{q}) = \int_a^b f(x, \tilde{q}, \tilde{q}_x) dx \ . \tag{2.19}$$

This can be solved by finding the solution to the system of equations

$$\frac{\partial}{\partial Q_k^e} I(\tilde{q}) = 0 \tag{2.20}$$

with $e$ going over all elements and $k$ over all element nodes.

Note that the entire procedure rests upon the ability to determine a function $f$ that corresponds to a PDE of practical interest. In many situations we have physical guidance that such a variational principle formulation exists. The basic underpinning is furnished by analytical mechanics and the physical principle of least action which finds various expressions in different disciplines. The principle of least action asserts that of all the generalized trajectories $(p, q) = \{q_k(t), p_k(t) \mid k = 1, \ldots, 3N\}$ of a system of $N$ particles, the one actually followed minizes the action $S$

$$S = \int_{t_0}^{t_1} \mathcal{L}(t, p, q) dt \tag{2.21}$$

with $\mathcal{L}$ being the Lagrangean of the system. Here $q_k$ denote generalized coordinates and $p_k$ generalized momenta. Though not always immediately apparent this leads to other expressions typically called *minimum energy functionals*. These can be

written for systems with no dissipative effects. Here are some examples of functions $f$ linked to important PDE's:

(1) Poisson equation in 2D

$$f = \frac{1}{2}\left(q_x^2 + q_y^2\right) - gq \tag{2.22}$$

for which (2.15) gives

$$q_{xx} + q_{yy} = g \tag{2.23}$$

(2) Poisson equation in 3D

$$f = \frac{1}{2}\left(q_x^2 + q_y^2 + q_z^2\right) - gq \tag{2.24}$$

for which the Euler variational principle

$$\left(\frac{\partial f}{\partial q}\right) - \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial q_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial q_y}\right) - \frac{\partial}{\partial z}\left(\frac{\partial f}{\partial q_z}\right) = 0 \tag{2.25}$$

gives

$$q_{xx} + q_{yy} + q_{zz} = g \tag{2.26}$$

**2.3. Galerkin methods.** The Ritz formulation typically leads to a system of equations which has nice numerical properties. However there are many systems for which a variational formulation is not possible typically because the system has dissipative behavior. In such situations we can again use an integral reformulation of the PDE of interest based upon the concept of a weak solution already introduced in the study of hyperbolic problems. Suppose we're looking for a solution to the problem

$$\mathcal{A}q = g \tag{2.27}$$

with $\mathcal{A}$ some differential operator. A function $q$ that directly satisfies (2.27) is called a *classical solution*. Consider now some space of test functions $v$ and a scalar product defined for the functions $q$ and $v$. From (2.27) we can derive

$$(\mathcal{A}q, v) = (g, v) \tag{2.28}$$

where $(\cdot, \cdot)$ denotes the scalar product, e.g.

$$(u, v) = \int_a^b u(x)v(x)dx \ . \tag{2.29}$$

In (2.28) we can apply integration by parts to obtain

$$(q, \mathcal{A}^* v) = (g, v) \tag{2.30}$$

where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$. This typically enables us to avoid differentiating functions $q$ that might be discontinuous. We can now use (2.30) to determine the unknown coefficients of a finite element approximation

$$\tilde{q}(x) = \sum_e \sum_k Q_k^e N_k^e(x) \tag{2.31}$$

by requiring

$$\sum_e \sum_k Q_k^e \left(N_k^e(x), \mathcal{A}^* v\right) = (g, v) \ . \tag{2.32}$$

The only piece missing is how we choose the test functions $v$. In a Galerkin formulation these are chosen to be the form functions themselves leading to

$$\sum_e \sum_k Q_k^e \left( N_k^e(x), \mathcal{A}^* N_j^e(x) \right) = \left( g, N_j^e(x) \right) , \tag{2.33}$$

thus defining a linear system

$$\mathbf{A}Q = b \tag{2.34}$$

$$A_{jk} = \left( N_k^e(x), \mathcal{A}^* N_j^e(x) \right) \ . \tag{2.35}$$

**2.4. A detailed example.** Let us now carry out the steps involved in solving a Poisson equation in 2D using a Ritz formulation and quadrilateral elements. The mathematical statement of the problem is

$$\begin{cases} q_{xx} + q_{yy} = g & (x,y) \in \Omega \\ q = b & (x,y) \in \partial\Omega \end{cases} \tag{2.36}$$

with the domain $\Omega = [a,b] \times [c,d]$ and $\partial\Omega$ denoting the boundary of $\Omega$ on which Dirichlet conditions are given. The element form functions are given by (1.17)-(1.20) and the function $f$ is given by (2.22). The function $I(\tilde{q})$ is

$$I(\tilde{q}) = \int_c^d \int_a^b f(x,y,\tilde{q},\tilde{q}_x,\tilde{q}_y) dxdy = \int_c^d \int_a^b \left[ \frac{1}{2} \left( \tilde{q}_x^2 + \tilde{q}_y^2 \right) - g\tilde{q} \right] dxdy \ . \tag{2.37}$$

The finite element approximation is determined by the chosen form functions and the nodal values $\{Q_k^e\}$. The extremum of $I(\tilde{q})$ is attained when

$$\frac{\partial}{\partial Q_k^e} I(\tilde{q}) = 0 \tag{2.38}$$

which leads to

$$\int_c^d \int_a^b \left[ \left( \tilde{q}_x \frac{\partial \tilde{q}_x}{\partial Q_k^e} + \tilde{q}_y \frac{\partial \tilde{q}_y}{\partial Q_k^e} \right) - g \frac{\partial \tilde{q}}{\partial Q_k^e} \right] dxdy = 0 \tag{2.39}$$

Note that

$$\frac{\partial \tilde{q}}{\partial Q_k^e} = N_k^e, \ \frac{\partial \tilde{q}_x}{\partial Q_k^e} = \frac{\partial N_k^e}{\partial x}, \ \frac{\partial \tilde{q}_y}{\partial Q_k^e} = \frac{\partial N_k^e}{\partial y} \tag{2.40}$$

so these derivatives no longer contain the unknowns $\{Q_k^e\}$. We thus obtain

$$\sum_e \sum_j \left[ \iint \left( \frac{\partial N_j^e}{\partial x} \frac{\partial N_k^e}{\partial x} + \frac{\partial N_j^e}{\partial y} \frac{\partial N_k^e}{\partial y} \right) dxdy \right] Q_k^e = \sum_e \iint g N_k^e \ dxdy \tag{2.41}$$

with $k$ going over all the element nodes. The sum over the elements is typically known as an assembly operation, leading to the computation of the matrix elements

$$A_{jk} = \sum_e \iint \left( \frac{\partial N_j^e}{\partial x} \frac{\partial N_k^e}{\partial x} + \frac{\partial N_j^e}{\partial y} \frac{\partial N_k^e}{\partial y} \right) dxdy \tag{2.42}$$

known as the system *stiffness matrix*. We can easily compute the elements of this matrix. Analytical computation is possible as in

$$\frac{\partial N_k^e}{\partial x} = \frac{\partial N_k^e}{\partial \xi} \frac{\partial \xi}{\partial x} + \frac{\partial N_k^e}{\partial \eta} \frac{\partial \eta}{\partial x} \tag{2.43}$$

$$\frac{\partial \xi}{\partial x} = \frac{\frac{D(\xi,y)}{D(\xi,\eta)}}{\frac{D(x,y)}{D(\xi,\eta)}} = \frac{1}{J}\begin{vmatrix} 1 & 0 \\ y_\xi & y_\eta \end{vmatrix} = \frac{y_\eta}{J} \tag{2.44}$$

$$\frac{\partial \eta}{\partial x} = \frac{\frac{D(\eta,y)}{D(\xi,\eta)}}{\frac{D(x,y)}{D(\xi,\eta)}} = \frac{1}{J}\begin{vmatrix} 0 & 1 \\ y_\xi & y_\eta \end{vmatrix} = -\frac{y_\xi}{J} \tag{2.45}$$

$$J = \begin{vmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{vmatrix} = x_\xi y_\eta - x_\eta y_\xi \tag{2.46}$$

The $x(\xi,\eta)$ and $y(\xi,\eta)$ dependencies are given by (1.21) so we obtain

$$\frac{\partial x}{\partial \xi} = \sum_{k=1}^{4} \frac{\partial N_k}{\partial \xi} x_k, \quad \frac{\partial y}{\partial \xi} = \sum_{k=1}^{4} \frac{\partial N_k}{\partial \xi} y_k \tag{2.47}$$

$$\frac{\partial x}{\partial \eta} = \sum_{k=1}^{4} \frac{\partial N_k}{\partial \eta} x_k, \quad \frac{\partial y}{\partial \eta} = \sum_{k=1}^{4} \frac{\partial N_k}{\partial \eta} y_k \tag{2.48}$$

But analytical evaluations are not really required in this case. We can recognize that the integrand in (2.42) is quadratic in $(x,y)$ and that a 4-point Gauss-Legendre quadrature leads to an exact evaluation.