LECTURES ON NUMERICAL METEOROLOGY

Roger K. Smith and Wolfgang Ulrich

December 5, 2008

Contents

Numerical Weather Prediction

The story goes that in 1904 Wilhelm Bjerknes was the first to point out that the future state of the atmosphere could be predicted by integrating the partial differential equations that govern the behaviour of the atmosphere, using as initial fields the observed state of the atmosphere at a particular time. However, the equations are too complicated for analytic solutions to be found and one must resort to numerical methods. We refer now to such integrations as *numerical weather prediction*, commonly abbreviated NWP.

The first attempt at NWP was carried out by Lewis Frey Richardson during the First World War. At this time the calculations had to be carried out by hand and were very tedious and time-consuming. The result was a spectacular failure, but the details were published in what has become one of the most famous books in meteorology (Richardson, 1922).

Richardson estimated that it would require a team of 64,000 persons to carry out a 24 hour forecast in 24 hours. This together with the unrealistic nature of his calculation, which predicted a surface pressure change of 145 mb in 6 hours, cast doubt on the practicality of the method! Several later developments changed this pessimistic view. Courant, Friedrichs and Lewy (1928) found that space and time increments chosen to discretize the differential equations have to meet a certain stability requirement. Later, mainly through the work of Rossby and his collaborators in the late 1930s, it was found that the large-scale motions in the atmosphere could be represented approximately by a rather simple equation expressing the conservation of absolute vorticity following the motion of air columns. Finally, at the end of World War II the first electronic computer ENIAC (Electronic Numerical Integrator and Computer) was constructed. This computer was used by Charney, Fjørtoft and von Neumann in the late 1940s for the first successful numerical forecast, based on integration of the absolute vorticity conservation equation (Charney *et al.* 1950).

In the last four decades, progress in the accuracy and sophistication of NWP models has been swift, partly through the development of improved numerical algorithms for solving the equations and also because of the as-

tonishing technological developments in computer engineering.

In these lectures we attempt to survey some of the basic results from the theory of partial differential equations (PDEs) before describing a range of numerical techniques for the solution of such equations. The emphasis will be on techniques that have application to numerical weather prediction.

The basic theory of PDEs is covered in Chapter 1. The various techniques for their solution are covered in subsequent chapters.

The main body of notes is concerned with grid point methods (Chapter 2-5). Later grid spectral methods and finite element methods are discussed.

Further reading

- G. J. Haltiner, 1971: Numerical weather prediction. John Wiley, New York, 317 pp.
- G. Stephenson, 1970: An introduction to partial differential equations for science students. Longmon, 152 pp.
- I. N. Sneddon, 1957: Elements of partial differential equations. McGraw-Hill, New York, 327 pp.

Chapter 1

Partial Differential Equations

1.1 Introduction

The differential equations describing atmospheric motion including those used for numerical weather prediction (NWP) are examples of partial differential equations. Partial differential equations (PDEs) arise in problems involving more than one independent variable and those with which we shall be concerned here are ones that arise in the representation of various types of physical processes such as diffusion, wave propagation, or equilibrium states of various physical systems.

An example is the equation for the temperature distribution, $T(x,t)$, of a heated rod (Fig. 1.1). In this case, the two independent variables are the distance along the rod, x, and the time t.

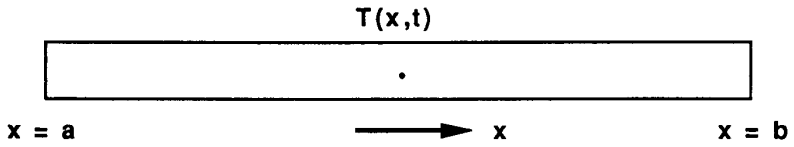

Figure 1.1: Temperature distribution in a heated rod.

Physical laws show that for a rod with a uniform diffusivity k, the temperature satisfies the equation

$$\frac{\partial T}{\partial t} = k\frac{\partial^2 T}{\partial x^2}. \tag{1.1}$$

Understanding the solution of this problem provides an understanding of a range of problems of interest to the atmospheric scientist. An example

is the temperature distribution in the ground due to the diurnal variation of heating at the earth's surface. Knowledge about the temperature of the earths surface as a function of time is needed for the complete solution of the equations for the atmosphere, itself.

It is useful to have means of picturing the solution to a PDE. In the case of Eq. (1.1), the solution can be expressed as a surface, $z = T(x,t)$, in a three-dimensional space (x,t,z), as shown in Fig 1.2. The domain of the solution, D, is the region $0 \leq t < \infty$ and $a \leq x \leq b$. The temperature distribution at some time $t_0 > 0$ is the curve $z = T(x,t_o)$, where the plane $t = t_0$ intersects the solution curve. The curve $z = T(x,0)$ is the initial temperature distribution.

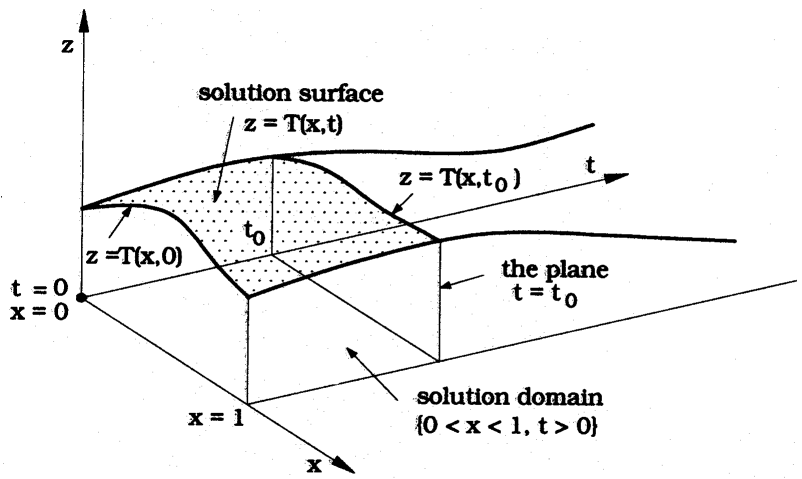

Figure 1.2: Solution surface and solution domain, D, for Eq. (1.1).

Equation (1.1) relates locally (i.e. at any point x,t) the slope of the surface in the t-direction to the rate-of-change of the slope in the x-direction. We shall see that in order to obtain a unique solution, we need to say something about the behaviour of the surface at the edges of the solution domain: at $t = 0$ and at $x = a, b$. It would seem physically reasonable, for example, that in order to predict the future evolution of the temperature, we need to know the initial state, i.e., the initial temperature distribution in the rod, $T(x,0)$. It would seem reasonable also that we need to know something about what is happening at the ends of the rod at any particular time. We shall return

to consider such problems later.

In contrast to the general solution of ordinary differential equations (ODEs), which involve arbitrary constants of integration, the general solution of PDEs involves arbitrary functions. Consider, for example, the equation

$$\frac{\partial^2 u}{\partial x \partial y} = 0 \tag{1.2}$$

Integrating partially with respect to x we obtain

$$\frac{\partial u}{\partial y} = F(y),$$

where $F(y)$ is an arbitrary function of y. We may integrate once again, this time with respect to y to obtain

$$u(x, y) = f(y) + g(x), \tag{1.3}$$

where $f(y) = \int F(y)\, dy$ and $g(x)$ are arbitrary functions. To determine $f(y)$ and $g(x)$ we need to have some additional information about the problem, for example, the *initial conditions* (if time is one of the independent variables) and/or *boundary conditions*. An example of the latter are the conditions that might be required to solve the heated rod problem.

To be specific, suppose that we wish to find $u(x, y)$ satisfying (1.2) in the region $x \geq 0$, $y \geq 0$ and that we are given that $u = x$ when $y = 0$ and $u = y$ when $x = 0$. Then the surface $u(x, y)$ must intersect the plane $x = 0$ in the line $u = y$ and the plane $y = 0$ in the line $u = x$ (Fig. 1.3).

Let us see how the functions $f(y)$ and $g(x)$ in (1.3) are determined. We are given that

$$u(x, 0) = f(0) + g(x) = x$$

and

$$u(0, y) = f(y) + g(0) = y$$

It follows that

$$u(x, y) = x + y - f(0) - g(0).$$

The only way this can satisfy the equation and the boundary conditions is if $f(0)$ and $g(0)$ are both zero, whereupon

$$u(x, y) = x + y.$$

It is easily verified that this equation satisfies the equation and the boundary condition. It may be shown also to be unique.

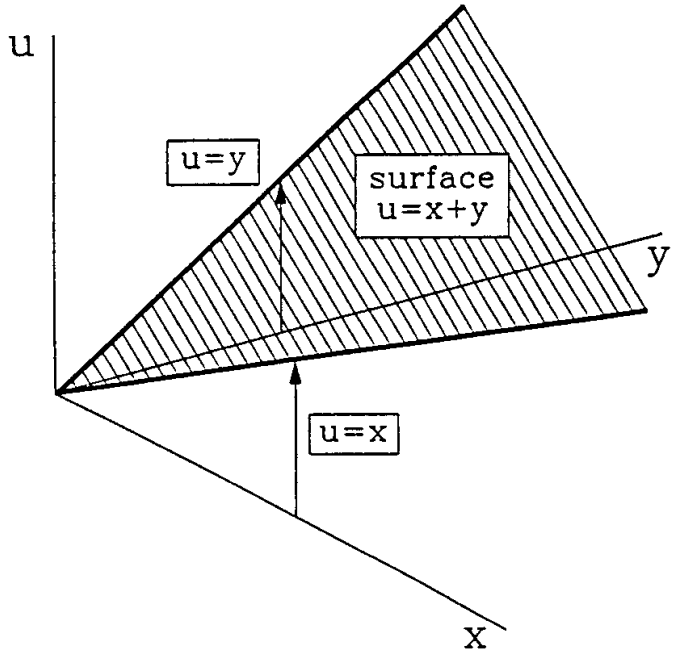

Figure 1.3: Solution surface $u(x,t)$ for Eq. (1.2) satisfying the conditions $u = x$ at $y = 0$ and $u = y$ at $x = 0$.

Clearly, boundary conditions are an essential part of the problem and in order to obtain a meaningful solution, we need sufficient boundary conditions but not too many. We know, for example, that for a third-order ODE, we require just three conditions to determine a unique solution; with only two conditions the problem is under-prescribed and, with four conditions it is over-prescribed.

The question is: what is a sufficient set of boundary conditions for a PDE? We shall see that this depends on the type of equation, which, in turn, reflects the nature of the physical problem. Often two different types of boundary condition lead to different types of solution of the same PDE. Therefore methods of solution are normally built around boundary conditions and on the physically expected type of solution.

A relevant problem in the atmosphere is that of numerical weather prediction (NWP), which is what this course is all about. With some approximations (see e.g. Smith, Lectures on Dynamical Meteorology, Chapter 8), one can formulate an equation for the evolution of atmospheric pressure as a function of space and time, $p(x,y,z,t)$. It is physically reasonable to expect that to be able to predict the pressure at some time in the future, we

would need to know the pressure distribution at some particular time, e.g., $p(x, y, z, 0)$. Furthermore, it is clear that if we are making the prediction for a limited region on the earth, for example in the European region, we shall need to know about pressure systems that are entering or leaving this region. That is, we shall need to specify the pressure, or some related quantity such as the horizontal pressure gradient, at the boundary of the computational domain for all time t. Again, as in the heated rod problem, we expect to require an initial condition and some boundary conditions.

1.2 Some definitions

To continue our discussion we need to introduce some basic definitions:

Operators Consider the expression

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2}.$$

This describes a set of manipulations which assigns to any sufficiently differentiable function $u(x, t)$ a new function of x and t. Such an assignment of one function to another is called an operator. We may denote it by L. We write the function that it assigns to u by $L[u]$ or Lu.

If $u(x, t)$ and $v(x, t)$ are any twice differentiable functions, so is an arbitrary linear combination of them, $au+bv$, where a and b are constants. Then $L[au + bv]$ is defined as

$$L[au + bv] = aL[u] + bL[v].$$

Definition An operator with these properties is called a *linear operator*. The order of an operator refers to the degree of the highest derivative (as in theory of ODEs). For example

$$L\,[u] = \frac{\partial u}{\partial x} - ku \quad \text{is a } \textit{first-order} \text{ operator ,}$$

$$L\,[T] = \frac{\partial T}{\partial t} - k\frac{\partial^2 T}{\partial x^2} \text{ and } L\,[u] = \frac{\partial^2 u}{\partial x \partial y} \quad \text{are } \textit{second-order} \text{ operators.}$$

An equation of the type

$$L[u] = f,$$

where L is a linear operator and f is a function of the independent variables only is called a *linear partial differential equation*. If $f \neq 0$ it is a *non homogeneous equation*; if $f = 0$ it is *homogeneous*.

Examples:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \frac{\partial u}{\partial x} = ku \quad \text{are } \textit{homogeneous} \text{ equations,}$$

$$\frac{\partial u}{\partial x} = kx, \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = kxy, \quad \text{are } \textit{nonhomogeneous} \text{ equations.}$$

In general, *homogeneous* means that every term includes the dependent variable to the same power. For example, the nonlinear equation

$$u\frac{\partial u}{\partial x} - \left(\frac{\partial^2 u}{\partial y^2}\right)^2 = 0$$

is homogeneous. Then if u is a solution, so is λu, for any constant λ.

Note 1.

The definition of a linear equation refers only to the operator and does not consider the boundary conditions. Thus, while the solutions of a linear equation may be added (or linearly superposed) to generate other solutions, different solutions may satisfy different boundary conditions. However, superposition may enable us to construct a solution satisfying given boundary conditions from ones which do not.

Note 2.

Superposition does not apply to nonlinear or to nonhomogeneous equations. For example, if u_1 nd u_2 are solutions of

$$L[u] = u\frac{\partial u}{\partial x} - \frac{\partial^2 u}{\partial x^2} = ky,$$

then

$$\frac{\partial}{\partial x}(u_1 + u_2) - \frac{\partial^2}{\partial x^2}(u_1 + u_2)$$
$$\frac{\partial u_2}{\partial x} + u_2\frac{\partial u_1}{\partial x} + 2ky \neq ky.$$

1.3 First order equations

One of the simplest first-order PDE's is the one-dimensional *advection equation*,

$$\frac{\partial T}{\partial t} + U\,\frac{\partial T}{\partial x} = 0, \tag{1.4}$$

where U is a constant. To understand the solution method it may be helpful to see how the equation might arise.

Suppose we fly horizontally in a research aircraft along a straight track with speed c in an atmosphere at rest. Suppose that there is horizontal temperature gradient in this atmosphere[1] in the direction of flight and that the background temperature varies with time. What will be the rate-of-change of temperature recorded in the aircraft? Suppose that in a short time interval Δt, the change in temperature is ΔT. Then by the chain rule for differentiation,

$$\Delta T = \frac{\partial T}{\partial t}\Delta t + \frac{\partial T}{\partial x}\Delta x,$$

so that the total rate-of-change of temperature is

$$\frac{\Delta T}{\Delta t} = \frac{\partial T}{\partial t} + \frac{\Delta x}{\Delta t}\frac{\partial T}{\partial x}.$$

Now $\Delta x = c\Delta t$ and in the limit as $\Delta t \to 0$, the total rate-of-change of T is

$$\frac{DT}{Dt} = \frac{\partial T}{\partial t} + c\frac{\partial T}{\partial x}. \tag{1.5}$$

The first term on the right hand side of this expression is the local rate-of-change of temperature associated with the time variation of the background field (it might be the normal diurnal variation of temperature, for example); the second term is associated with the motion of the aircraft in the presence of the horizontal temperature gradient. Suppose now that the air mass is moving in the x-direction with uniform speed U and that we drift with the air in a balloon, again measuring temperature. Then, the rate of temperature change we measure in the balloon is given by (1.5) with $c = U$. Since we are drifting with the air, we cannot know about the horizontal temperature gradient within the air mass and must measure the local rate-of-change of temperature at the point of the air parcel with which we are drifting, say $\partial T_0/\partial t$; i.e.

[1] In actual fact, such an atmosphere cannot be at rest (why?), but this is immaterial for the present illustration. The assumption can be rigorously justified if the flight speed of the aircraft is much larger than the wind speed.

$$\frac{DT}{Dt} = \frac{\partial T}{\partial t} + U\,\frac{\partial T}{\partial x} = \frac{\partial T_o}{\partial t}, \tag{1.6}$$

If the local rate-of-change of air temperature at this point is zero, then

$$\frac{\partial T}{\partial t} + U\,\frac{\partial T}{\partial x} = 0. \tag{1.7}$$

This is therefore the PDE representing the behaviour of temperature as a function of x and t in a uniformly translating air mass in which there is a horizontal temperature gradient. Equation (1.7) is simply a statement of the fact that the temperature of an air parcel being carried along with the air mass remains constant.

Knowing how equations such as (1.7) arise provides a clue to their solution. Suppose know that a certain quantity $q(x,t)$ satisfies the PDE

$$\frac{\partial q}{\partial t} + \lambda\frac{\partial q}{\partial x} = 0, \tag{1.8}$$

We recognize the operator $L \equiv \partial/\partial t + \lambda\partial/\partial x$, where λ is a constant, as representing a total rate-of-change following a point moving with speed $dx/dt = \lambda$, (cf. expression (1.5)). That is, if one moves with a point $x(t) = x_0 + \lambda t$, the operator $L[q]$ will tell us the rate-of-change of q. Equation (1.8) tells us that this rate-of-change is zero; in other words q is a constant along a line $x = x_0 + \lambda t$. Of course, it may vary from line to line and since each line is characterized by the constant of integration x_0, we can say that $q = q(x_0)$, or alternatively, $q = q(x - \lambda t)$. If we know the distribution of q at time $t = 0$, say $q = Q(x)$, then it is clear that the solution of (1.8) at time t is

$$q = Q(x - \lambda t). \tag{1.9}$$

All this becomes clearer when we sketch the solution surface in the (x,t,z)-space, analogous to Fig. 1.2. The lines $x = x_0 + \lambda t$ are a family of parallel lines in the plane $z = 0$ and intersect the plane $t = 0$ at $x = 0$. The equation says that the height of the solution surface is always the same along such a line. Then, it is not too difficult to see that the intersection of this surface with the plane $t =$ constant is a curve which is identical with the curve $z = Q(x)$ at $t = 0$, but displaced in the x-direction by a distance λt. Thus the solution represents a disturbance with arbitrary shape $Q(x)$ translating uniformly with speed λ in the positive x-direction if $\lambda > 0$, or in the negative x-direction if $\lambda < 0$.

It is clear that 'information' about the initial distribution of q 'propagates' or is 'carried along' the lines $x = x_0 + \lambda t$ in the plane $z = 0$. These lines are called the *characteristic curves*, or simply the *characteristics* of the equation.

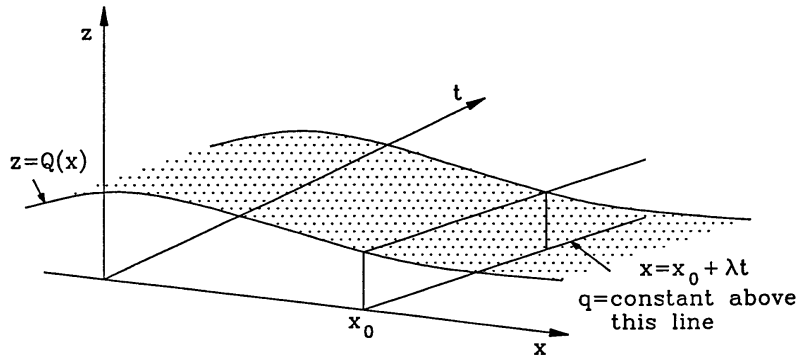

Figure 1.4: Solution surface for Eq. (1.8).

If λ were a function of t and/or x, the solution procedure would be the same. Essentially we would solve first for the characteristic curves by integrating the equation

$$\frac{dx}{dt} = \lambda(x,t). \tag{1.10}$$

The solution of this ODE involves an integration constant that determines where the characteristic intersects the x-axis. Then we need to construct the solution surface which has the same value along each characteristic in the $x-t$ plane as that in the initial plane, $t = 0$. Of course, for a general $\lambda(x,t)$, some characteristics may never cross the initial line, but in a physical problem, this would not be expected to be the case (why?).

If the equation is nonhomogeneous, for example Eq. (1.6) with $T_0(t)$ prescribed, the procedure is again to solve for the characteristics, but in this case, T is not constant along the characteristics, but varies at the rate $\partial T_0/\partial t$ (see Ex. 1.1 below).

Note that the characteristics of a linear equation depend on the coefficient(s) of the derivatives, but that is all. However, in a nonlinear equation such as

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0, \tag{1.11}$$

they depend upon the solution also. For example, the characteristics of (1.11) satisfy the equation

$$\frac{dx}{dt} = u. \tag{1.12}$$

Equation (1.11) is not much more difficult to treat than the linear equation (1.8), but the solution is markedly different and may develop discontinuities.

Again, (1.11) says that u is constant along a characteristic given by (1.12) and this being the case, (1.12) can be integrated to give

$$x = x_0 + ut \tag{1.13}$$

along a characteristic. Thus the characteristics are again straight lines in the $x - t$ plane, but their slope depends on the value of the solution u, which is equal to the initial value, say $u_0(x_0)$. Thus depending on the slope of the initial condition $u(x, 0) = u_0(x)$, the characteristics may diverge or converge. If they converge, there will be a finite time at which two characteristics meet at some point and the solution will become discontinuous there (see Ex. 1.2). This happens in the case of the nonlinear shallow-water equations without diffusion.

1.4 Linear second-order equations

The general linear homogenous equation of second-order is

$$a\frac{\partial^2 u}{\partial x^2} + 2b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + 2f\frac{\partial u}{\partial x} + 2g\frac{\partial u}{\partial y} + hu = 0, \tag{1.14}$$

where a, b, c, f, g, h are either constants or more generally functions of x and y. The diffusion equation (1.1) is an example for which $y = t$, $a = k$, $g - \frac{1}{2}$ and all other coefficients are zero. Another example is the so-called *wave equation*:

$$\frac{\partial^2 u}{\partial t^2} = c_0^2 \frac{\partial^2 u}{\partial x^2}, \tag{1.15}$$

where $y = t$, $a = 1$, $b = 0$, c_0^2 and $f = g = h = 0$. Note that the form of (1.14) resembles the equation for a conic section,

$$ax^2 + 2bxy + cy^2 + 2fx + 2gy + h = 0, \tag{1.16}$$

where a, b, c .... are constants. Equation (1.16) represents an ellipse, parabola, or hyperbola according as $ac - b^2 > 0$, $= 0$ or < 0. Thus we classify the PDE (1.14) as

$$\begin{aligned} \textit{elliptic} \quad &\text{when} \quad ac - b^2 > 0 \\ \textit{parabolic} \quad &\text{when} \quad ac - b^2 = 0 \\ \textit{hyperbolic} \quad &\text{when} \quad ac - b^2 < 0 \end{aligned}$$

It follows that the diffusion equation is parabolic and the wave equation is hyperbolic. We shall see that the type of boundary and/or initial conditions

required to determine a solution of (1.14) depend on the type of equation according to the above classification. We begin by studying a slightly simpler form of (1.14), namely

$$a\frac{\partial^2 u}{\partial x^2} + 2b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = 0, \tag{1.17}$$

where a, b, and c are constants. This is known as Euler's equation. To obtain the general solution we make a linear transformation of the independent variables to ξ and η given by

$$\xi = px + qy \quad \text{and} \quad \eta = rx + sy,$$

where the constants p, q , r and s will be chosen later. Now, using the chain rule for partial differentiation,

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial x} = p\frac{\partial u}{\partial \xi} + r\frac{\partial u}{\partial \eta},$$

and

$$\frac{\partial u}{\partial y} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial y} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial y} = q\frac{\partial u}{\partial \xi} + s\frac{\partial u}{\partial \eta}.$$

It follows that

$$\begin{aligned}
\frac{\partial^2 u}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) \\
&= \left(p\frac{\partial}{\partial \xi} + r\frac{\partial}{\partial \eta}\right)\left(p\frac{\partial u}{\partial \xi} + r\frac{\partial u}{\partial \eta}\right) \\
&= p^2\frac{\partial^2 u}{\partial \xi^2} + 2pr\frac{\partial^2 u}{\partial \xi \partial \eta} + r^2\frac{\partial^2 u}{\partial \eta^2},
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 u}{\partial y^2} &= \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) \\
&= \left(q\frac{\partial}{\partial \xi} + s\frac{\partial}{\partial \eta}\right)\left(q\frac{\partial u}{\partial \xi} + s\frac{\partial u}{\partial \eta}\right) \\
&= q^2\frac{\partial^2 u}{\partial \xi^2} + 2qs\frac{\partial^2 u}{\partial \xi \partial \eta} + s^2\frac{\partial^2 u}{\partial \eta^2},
\end{aligned}$$

Then, substitution into (1.17) gives

$$\begin{aligned}
(ap^2 + 2bpq + cq^2)\frac{\partial^2 u}{\partial \xi^2} &+ 2[apr + csq + b(rq + sp)] + \frac{\partial^2 u}{\partial \xi \partial \eta} \\
&+ (ar^2 + 2brs + cs^2) = 0.
\end{aligned} \tag{1.18}$$

Now we choose $p = 1$, $r = 1$, and q and s to be the two roots λ_1, λ_2 of the quadratic equation

$$a + 2b\lambda + c\lambda^2 = 0. \tag{1.19}$$

Then (1.18) becomes

$$[a + b(\lambda_1 + \lambda_2) + c\lambda_1\lambda_2]\frac{\partial^2 u}{\partial\xi\partial\eta} = 0. \tag{1.20}$$

Noting that $\lambda_1 + \lambda_2 = -2b/c$ and $\lambda_1\lambda_2 = a/c$, Eq. (1.20) reduces to

$$\frac{2}{c}(ac - b^2)\frac{\partial^2 u}{\partial\xi\partial\eta} = 0. \tag{1.21}$$

If $c \neq 0$ and if the equation is not parabolic, then

$$\frac{\partial^2 u}{\partial\xi\partial\eta} = 0, \tag{1.22}$$

which has the same form as (1.2), with solution (cf. Eq. 1.3)

$$u(\xi, \eta) = F(\xi) + G(\eta), \tag{1.23}$$

where F and G are arbitrary functions. In terms of x and y, it follows that

$$u(x, y) = F(x + \lambda_1 y) + G(x + \lambda_2 y). \tag{1.24}$$

The functions F and G will need to be determined from suitably-prescribed initial and/or boundary conditions.

In the case of the wave equation (1.15), $a = 1$, $b = 0$, $c = -c_0^{-2}$ and $y = t$. Then $ac - b^2 < 0$ and the equation is hyperbolic. Equation (1.19) reduces to $1 - \lambda^2/c_0^2 = 0$, whereupon $\lambda_1 = -c_0$, $\lambda_2 = c_0$ so that the general solution is

$$u(x, t) = F(x - c_0 t) + G(x + c_0 t). \tag{1.25}$$

In analogy with (1.9), the solution corresponds with the sum of two shapes $F(x)$ and $G(x)$, the former translating in the positive x-direction with speed c_0, the latter propagating with the same speed in the negative x-direction. In this case, there are two sets of characteristic curves: $\xi = x - c_0 t =$ constant and $\eta = x + c_0 t =$ constant. Part of the 'initial information' ($u(x, 0) = F(x) + G(x)$) propagates along one set of characteristics and part propagates along the other set (see Ex. 1.6). Clearly, to have enough information to determine F and G we need an extra initial condition. This is not surprising since the equation involves the second derivative with respect to time rather than just the first (note that the ODE $dx/dt = f(t)$ requires only one initial

derivative condition whereas the equation $d^2x/dt^2 = f(t)$ requires two). We might guess that an appropriate additional condition would be to prescribe $\partial u/\partial t = v_0(x)$ at $t = 0$, equivalent to specifying the initial slope of the solution surface at $t = 0$ as well as its height. From (1.25), this implies that

$$-c_0\frac{dF}{dx} + c_0\frac{dG}{dx} = v_0(x),$$

whereupon

$$-F + G = \frac{1}{c_o}\int_0^x v_o(x')dx' \tag{1.26}$$

Since

$$F + G = u_0(x), \tag{1.27}$$

the full solution of (1.15) follows readily in terms of $u_0(x)$ and $v_0(x)$ (see Ex. 1.6).

For *hyperbolic equations* in general, $ac - b^2 < 0$ and the roots of (1.19) are *real* and *distinct*. For elliptic equations, $ac - b^2 > 0$ and the roots are *complex conjugate* pairs. Thus elliptic equations have complex characteristics.

In the case of a *parabolic equation*, $ac = b^2$ and the roots of (1.19) are equal; in fact $\lambda_1 = \lambda_2 = -b/c$. In this case we choose $p = 1$ and $q = -b/c$, whereupon the first and second terms of (1.18) vanish and then provided that r and s are not both zero, it reduces to

$$\frac{\partial^2 u}{\partial \eta^2} = 0. \tag{1.28}$$

This equation may be integrated directly to give

$$u(x, y) = F(\xi) + \eta G(\xi), \tag{1.29}$$

where F and G are arbitrary functions of ξ. Now since $p = 1$ and $q = -b/c\lambda$, say, then

$$\xi = x + \lambda y \quad \text{and} \quad \eta = rx + sy,$$

where r and s are *arbitrary* and not both zero. For simplicity we may choose $r = 0$ and $s = 1$ to obtain

$$u(x, y) = F(x + \lambda y) + yG(x + \lambda y). \tag{1.30}$$

Note that a parabolic equation, unlike a hyperbolic equation, has only one set of characteristics given by $\xi = x + \lambda y =$ constant.

If one considers the solution surface $z = u(x, y)$ in (x, y, z)-space, it is clear that a knowledge of $u(x, 0)$ would determine the function F. However, it is not immediately clear what condition is suitable to determine G.

While the method of obtaining the general solution of Euler's equation shows some important features about the structure of solution, the general solutions are not usually easy to fit directly to any prescribed initial or boundary conditions.

Often the general solutions to PDE's are far too *general* to be practically useful and specific solutions require initial and/or boundary conditions to be prescribed. In trying to see what boundary conditions might be appropriate, it is enlightening to introduce three physical problems which lead to equations of the three basic types we have discussed: hyperbolic, parabolic and elliptic. We regard these problems as *prototype problems* for understanding the structure of solutions to the equations.

1.5 Parabolic equations

A prototype for studying parabolic systems is the heated rod problem discussed in section 1.1. Suppose that the initial temperature distribution in the rod is $T_0(x)$ and that for $T > 0$, the ends are held at uniform temperature T_a and T_b, respectively - in our physical 'thought' experiment we assume that the rod is insulated along its length so that no heat escapes and that the ends are held in contact with heat reservoirs, which maintain the temperature there at the required values. It is reasonable to expect that if the rod has a uniform conductivity k, the temperature distribution along it will eventually become linear in x, i.e.

$$T(x, \infty) = T_a + (T_b - T_a)(x - a)/(b - a). \tag{1.31}$$

It is easy to check that this is a steady solution of Eq. (1.1) satisfying the required boundary conditions on T at $x = a$ and b. Let us define the deviation temperature at time t, $\theta(x, t)$, to be the actual temperature $T(x, t)$ minus the steady state temperature given by (1.31). Then clearly $\theta(x, t)$ satisfies (1.1) because the latter is a linear equation; moreover $\theta(a, t) = 0$ and $\theta(b, t) = 0$ for all time, because $T(a, t) = T(a, \infty)$ and $T(b, t) = T(a, \infty)$.

We try now to obtain a solution for $\theta(x, t)$ by the method of separation of variables. We set $\theta(x, t) = X(x)\Theta(t)$ in the equation for θ and divide by θ to obtain

$$\frac{1}{\Theta}\frac{d\Theta}{dt} = \frac{k}{X}\frac{d^2X}{dx^2}, \tag{1.32}$$

Since the left-hand side is a function of t only and the right-hand side is function of x only, the only way that (1.32) can be satisfied as x and t vary independently is if both sides are the same constant, say μ. This observation

enables us to reduce the PDE to two ODE's, namely

$$\frac{d\Theta}{dt} = \mu\Theta, \tag{1.33}$$

and

$$\frac{d^2X}{dx^2} - \frac{\mu}{k}X = 0. \tag{1.34}$$

In order to satisfy the boundary conditions on θ for all time we must have $X(a) = X(b) = 0$. The solution of (1.33) has the form

$$\Theta = \Theta_o e^{\mu t} \tag{1.35}$$

and since $\theta(x,t)$ is the transient part of the solution for $T(x,t)$ that must decay to zero as $t \to \infty$, it is clear that μ must be taken to be negative. Then, writing $m^2 = -\mu/k > 0$ in (1.34), we see that solutions for X must have the form

$$X(x) = A\cos mx + B\sin mx \tag{1.36}$$

The algebra in implementing the boundary conditions on X is simplified if we relocate the origin of x so that $a = 0$. Then the condition $X(0) = 0$ requires that $A = 0$ and the condition $X(b) = 0$ requires that $B\sin mb = 0$. If we choose $B = 0$, the solution for X is trivial and of little use to us. If $B \neq 0$ then $\sin mb = 0$, whereupon mb must be an integer multiple n, say, of π. It follows that possible solutions for θ have the form

$$B_n \exp\left(-n^2\pi^2 kt/b^2\right) \sin\,(n\pi x/b)$$

and by the principle of superposition, the general solution for $\theta(x,t)$ satisfying $\theta(0,t) = 0$ and $\theta(b,t) = 0$ is the Fourier series

$$\theta(x,t) = \sum_{n=1}^{\infty} B_n \exp(-n^2\pi^2 kt/b^2) \sin\left(\frac{n\pi x}{b}\right). \tag{1.37}$$

Given the initial condition that $T(x,0) = T_0(x)$ we find that

$$T(x,0) - T_0(x,\infty) = \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{b}\right),$$

which relates the Fourier coefficients B_n to the initial temperature distribution in the rod, i.e.,

$$B_n = \frac{2}{b}\int_0^b \left[T_0(x) - T(x,\infty)\right] \sin\left(\frac{n\pi x}{b}\right) dx.$$

The solution of this problem brings out clearly the constraints on parabolic equations in general. It shows that one initial condition is required together with two boundary conditions. Again, this fits with the idea that the time derivative is only first order whereas the space derivative is second order.

1.6 Hyperbolic equations

The prototype problem for a hyperbolic equation considers the vibrations of a stretched string with the ends at $x = 0$ and $x = b$, say (Fig. 1.5). The lateral displacement of the string, $y(x,t)$, satisfies the wave equation (1.15), where the constant c_0^2 is proportional to the tension in the string and inversely proportional to its mass per unit length.

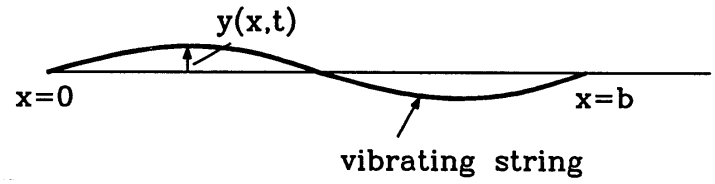

Figure 1.5: Schematic of a vibrating string.

The boundary conditions are clearly that $y(0,t)$ and $y(b,t) = 0$ and since the equation is derived by applying Newton's second law to each element of the string, we can guess that it will be necessary to know both the positions and velocity of each element, i.e. $y(x,0)$ and $\partial y/\partial t$ at $t = 0$. Substituting $y(x,t) = X(x)T(t)$ in this case and dividing through by $y(x,t)$ gives

$$\frac{1}{X}\frac{d^2X}{dx^2} = \frac{1}{c_o^2 T}\frac{d^2T}{dt^2} = -m^2,$$

where $m^2 =$ constant. We choose a negative separation constant in this case in order to be able to satisfy the boundary conditions at $x = 0$ and $x = b$; these imply that $X(0)$ and $X(b) = 0$. Then X satisfies

$$\frac{d^2X}{dx^2} + m^2 X = 0,$$

with appropriate solutions

$$X = B_n \sin\left(\frac{n\pi x}{b}\right) \quad (n = 1, 2, 3, ...),$$

as before (see section 1.5). The equation for T is similar in form and has solutions

$$T = A_n \cos\left(\frac{n\pi c_o t}{b}\right) + C_n \sin\left(\frac{n\pi c_o t}{b}\right).$$

Then the general solution for y satisfying $y(0,t) = 0$ and $y(b,t) = 0$ for all time is

$$y(x,t) = \sum_{n=1}^{\infty} \left[A_n \cos \left(\frac{n\pi c_o t}{b} \right) + C_n \sin \left(\frac{n\pi c_o t}{b} \right) \right] \sin \left(\frac{n\pi x}{b} \right), \quad (1.38)$$

where the constants A_n, C_n have been redefined to incorporate the constants B_n. Note that, in contrast to the solution of the diffusion equation, there are now two sets of Fourier coefficients to be determined. The first set is obtained by putting $t = 0$ in (1.38) so that

$$y(x,t) = \sum_{n=1}^{\infty} A_n \sin \left(\frac{n\pi x}{b} \right). \quad (1.39)$$

Then

$$A_n = \frac{2}{b} \int_0^b y(x,0) \sin \left(\frac{n\pi x}{b} \right) dx. \quad (1.40)$$

The other set is obtained by differentiating (1.38) with respect to time and the setting $t = 0$ so that

$$\left. \frac{\partial y}{\partial t} \right|_{t=0} = v_o(x) = \frac{\pi c_0}{b} \sum_{n=0}^{\infty} n C_n \sin \left(\frac{n\pi x}{b} \right), \quad (1.41)$$

or

$$C_n = \left(\frac{2}{n\pi c_o} \right) \int_0^b v_o(x) \sin \left(\frac{n\pi x}{b} \right) dx. \quad (1.42)$$

Again, the analysis shows rather well how the physical problem is precisely constrained by a certain combination of initial and boundary conditions. In particular it indicates how the formulation would fall apart if one or more of these conditions were unavailable. For example the physical problem would be clearly ill-conceived if we removed either boundary condition at the end of the string!

1.7 Elliptic equations

Elliptic equations tend to arise in problems involving equilibrium or balanced states. The prototype physical problem considers the equilibrium of a stretched membrane subject to a distribution of forces normal to it (Fig. 1.6). It can be shown that the lateral displacement of the membrane $z(x,y)$

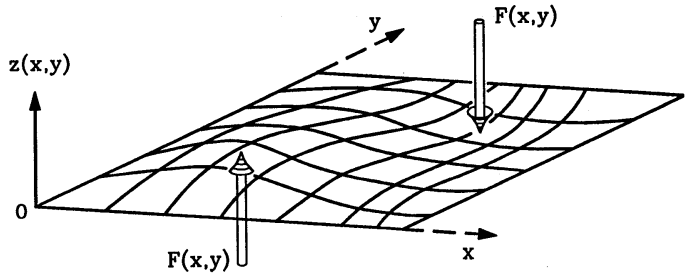

Figure 1.6: Equilibrium displacement of a stretched membrane.

subject to the force distribution $F(x, y)$ per unit area normal to it satisfies the elliptic PDE:

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = -F\,(x, y). \tag{1.43}$$

Equation (1.43) is called *Poisson's equation.* Typically the membrane would have zero displacement along its boundary (e.g. a drum skin), i.e. an appropriate boundary condition would be $z = 0$ along all of its closed sides. Suppose that the membrane is stretched across a closed piece of wire which does not lie exactly in a plane, but whose displacement is known and suppose that there is no force distribution across the membrane. Then the displacement satisfies the equation

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0 \tag{1.44}$$

subject to the displacement being prescribed along the entire closed boundary formed by the wire. Equation (1.44) is called Laplace's equation.

In section 1.4 we showed that the transformation $\xi = x + \lambda_1 y$, $\eta = x + \lambda_2 y$ reduces Euler's equation (1.17) to the form $\partial^2 u/\partial\xi\eta = 0$, where λ_1 and λ_2 are the roots of (1.19), assuming that $ac - b^2 \neq 0$. This transformation proves useful for a hyperbolic equation where λ_l and λ_2 are real, but is of limited use for elliptic equation where λ_l and λ_2 are complex conjugates and the characteristics are complex. Suppose that in the latter case,

$$\lambda_1 = p + iq \quad \text{and} \quad \lambda_2 = p - iq, \tag{1.45}$$

where

$$p = -b/c, \quad q = (ac - b^2)^{1/2}/c. \tag{1.46}$$

Instead of transforming to ξ and η, we transform to λ and μ where

$$\lambda = \tfrac{1}{2}(\xi + \eta) = x + py, \tag{1.47}$$

and

$$\mu = \tfrac{1}{2i}(\xi - \eta) = qy. \tag{1.48}$$

Then

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial \lambda}\frac{\partial \lambda}{\partial x} + \frac{\partial}{\partial \mu}\frac{\partial \mu}{\partial x}\frac{\partial}{\partial \lambda}$$

and

$$\frac{\partial}{\partial y} = \frac{\partial}{\partial \lambda}\frac{\partial \lambda}{\partial y} + \frac{\partial}{\partial \mu}\frac{\partial \mu}{\partial y}p\frac{\partial}{\partial \lambda} + q\frac{\partial}{\partial \mu},$$

whereupon Euler's equation reduces to

$$\frac{1}{c}(ac - b^2)\left(\frac{\partial^2 u}{\partial \lambda^2} + \frac{\partial^2 u}{\partial \mu^2}\right) = 0. \tag{1.49}$$

Similarly, the general elliptic equation can be transformed to the canonical form

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + e\frac{\partial u}{\partial x} + f\frac{\partial u}{\partial y} + gu = 0. \tag{1.50}$$

Other problems in which elliptic equations arise are:

(i) Steady heat flow in a medium with uniform conductivity The heat flux in such a medium is $\mathbf{q} = -k\nabla T$, where T is the temperature and k is the conductivity. If there are no heat sources or sinks in the medium, $\nabla \cdot \mathbf{q} = 0$, whereupon T satisfies $\nabla^2 T = 0$.

(ii) Steady irrotational incompressible flow of a homogeneous fluid Irrotational means that $\nabla \wedge \mathbf{v} = 0$, where $\mathbf{v}$ is the velocity vector. This implies the existence of a velocity potential ϕ such that $\mathbf{v} = \nabla\phi$. Steady, incompressible and homogeneous implies that $\nabla \cdot \mathbf{v} = 0$, whereupon ϕ satisfies $\nabla^2\phi = 0$.

Since elliptic problems arise naturally for steady problems, problems concerned essentially with states of equilibrium, initial values are not relevant. Thus typical problems are *boundary value problems* in which the boundary data are prescribed on given *closed* boundary curves (or surfaces/hypersurfaces for equations with more than two independent variables).

The two most commonly occurring problems are:

(a) The Dirichlet problem, e.g.

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

with u prescribed as a function of position along the whole boundary of the domain

(b) The Neumann problem, e.g.

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

with $\partial u/\partial n$ prescribed as a function of position along the whole boundary of the domain, where $\partial/\partial n$ is the normal derivative along the boundary. Note that in this case, the prescribed function must satisfy a compatibility condition (see below).

A third problem is the *mixed problem* where $au + b\partial u/\partial n$ is prescribed along the boundary, possibly with a or b zero along segments of the boundary.

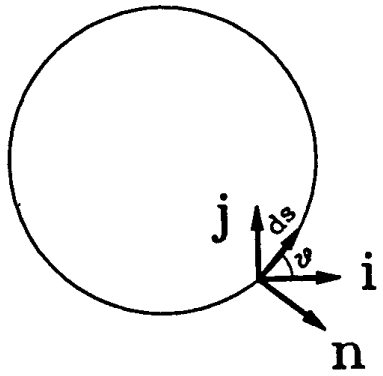

Figure 1.7:

Green's theorem If $P(x, y)$ and $Q(x, y)$ are functions defined inside and on the boundary C of a closed domain S, then

$$\int\int_s \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}\right) dx\, dy = \oint_C (P dy - Q dx), \tag{1.51}$$

where $\oint_C$ is taken in the positive (counter-clockwise) direction.

The proof may be found in many books on advanced calculus, e.g. R.P. Gillespie, Integration, Oliver and Boyd, p56, §20.

Let $\hat{\mathbf{n}}$ be the outward unit normal to C and let ds be an element of the curve C. Then

$$dx = ds \cos\theta = -\mathbf{j} \cdot \hat{\mathbf{n}}\ ds,$$

and (1.51) may be written as

$$\int\int_s \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}\right) dx\, dy = \oint_C (P\mathbf{i} - Q\mathbf{j}) \cdot \hat{\mathbf{n}}\, ds. \tag{1.52}$$

This is the two-dimensional form of the *divergence theorem*. In the case of Neumann's problem for Poisson's equation,

$$\oint_C \frac{\partial u}{\partial n}\, ds = \int\int_s \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) dx\ dy = -\int\int_s F\ dx\, dy, \tag{1.53}$$

the first equality being obtained by setting $P = \partial u/\partial x$ and $Q = \partial u/\partial y$ in (1.52). Noting that $\partial u/\partial n = \hat{\mathbf{n}} \cdot \nabla u$, it follows that $\partial u/\partial n$ can be prescribed only subject to the constraint on F implied by Eq. (1.53). Clearly, for Laplace's equation we must have

$$\oint_C \frac{\partial u}{\partial n} ds = 0. \tag{1.54}$$

An example of an incorrectly-posed boundary value problem for Laplace's equations was given by Hadamard. It is shown diagrammatically in Fig. 1.8.

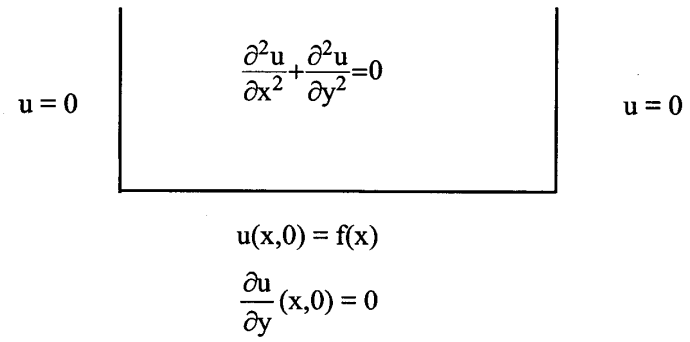

Figure 1.8: Specification of Hadamard's problem for Laplace's equation. These boundary conditions are the same as those that are appropriate for the wave equation.

It turns out that one can find a unique solution to this problem, but the solution is not continuous with respect to the data. By this we mean that an arbitrarily small change in $f(x)$ can change the solution by an arbitrarily large amount.

Example

Let $f(x) = e[\sin(4n+1)\pi x]$, where n is a positive integer. It is easy to verify that a solution which satisfies the boundary conditions is

$$u_n(x, y) = e^{-\sqrt{n}} \cosh\left[(4h+1)\pi y\right].$$

By taking n sufficiently large, one can make f and any number of its derivatives arbitrarily small, but

$$u_n(\tfrac{1}{2}, y) = e^{-\sqrt{n}} \cosh\left[(4n+1)\pi y\right]$$

can be made arbitrarily large for any fixed $y > 0$ by choosing n sufficiently large. Hence the solution is not continuous with respect to the data and the problem is *incorrectly posed*. It is easy to see why the problem is ill-posed by considering a physical analogy. Suppose that u represents the temperature distribution in a block of material and write $u = T - T_0$. The physical analogy is that of a block of material of unit width as shown in Fig. 1.8. If $T(x) > 0$, the temperature gradient between $y = 0$ and the side-boundaries implies a steady heat flux through $y = 0$ towards the side boundaries. However, the condition $\partial T/\partial y = 0$ at $y = 0$ is inconsistent with this as it implies zero heat flux. Note also that the absence of any condition as $y \to \infty$ allows an arbitrary heat flux through that boundary and therefore a multiplicity of solutions.

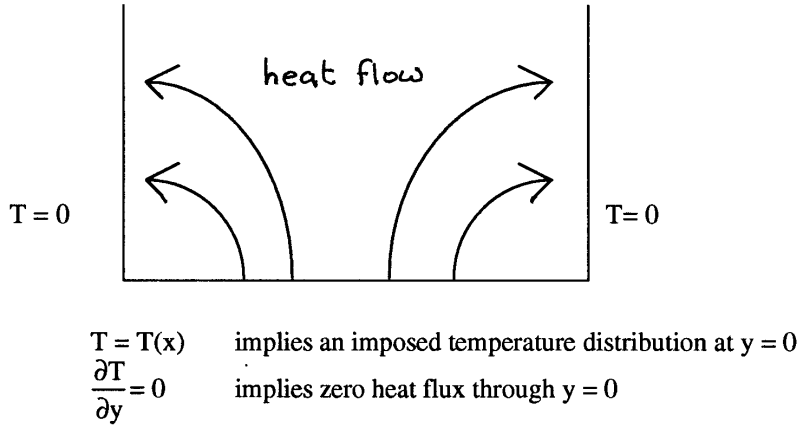

Figure 1.9: Physical analogue to Hadamard's problem. Arrows indicate direction of heat flow.

Exercises

1.1 Solve the PDE's

$$(a) \quad \frac{\partial T}{\partial t} + U\frac{\partial T}{\partial x} = q(1 - e^{-\mu t}),$$

$$(b) \qquad \frac{\partial T}{\partial t} + Ue^{-\mu t}\frac{\partial T}{\partial x} = 0.$$

in the domain $-\infty \leq x \leq \infty$, $t \geq 0$, given that $T = T_0(x)$ at $t = 0$. Assume that U, μ and q are constants.

1.2 Obtain the characteristics of the PDE

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$$

in the domain $-\infty \leq x \leq \infty$, $t \geq 0$, given that $u = 1 - e^{-x}$ at $t0$. Show that the characteristics converge in the region $x > 0$ and diverge in the region $x < 0$.

1.3 A model for the temperature variation $T(z, t)$ in the ground as a function of depth z and time t is to assume that $T(z, t)$ satisfies the diffusion equation

$$\frac{\partial T}{\partial t} = k\frac{\partial^2 T}{\partial z^2},$$

where k is a constant. The surface temperature, $T(0, t)$, and the temperature at great depth, $T(\infty, t)$, are prescribed. Taking $T(0, t) = T_s + \Delta\cos(\Omega t)$ to represent the diurnal temperature variation at the earth's surface resulting from insolation and nocturnal cooling and $T(\infty, t) = T_\infty$, and assuming that T_s, Δ and Ω are constants, calculate $T(z, t)$. Is it realistic to have T_s and T_∞ different? Indeed, how realistic is this model?

1.4 Determine the nature of each of the following equations (i.e. whether elliptic, parabolic or hyperbolic) and obtain the general solution in each case:

$$\text{(a)} \quad 3\frac{\partial^2 u}{\partial x^2} + 4\frac{\partial^2 u}{\partial x \partial y} - \frac{\partial^2 u}{\partial y^2} = 0, \quad \text{(b)} \quad \frac{\partial^2 u}{\partial x^2} - 2\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} = 0,$$

$$\text{(c)} \qquad 4\frac{\partial^2 u}{\partial y^2} + 2\frac{\partial^2 u}{\partial x^2} = 0, \qquad \text{(d)} \qquad \frac{\partial^2 u}{\partial y^2} + 2\frac{\partial^2 u}{\partial x^2}0,$$

(e) $$\frac{\partial^2 u}{\partial y^2} + 2\frac{\partial^2 u}{\partial x^2} = 0.$$

1.5 A function $u(r,t)$ satisfies the equation

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2},$$

where c is constant. By introducing the new dependent variable $v(r,t)$ and writing $\xi = r + ct$, $\eta = r - ct$, reduce this equation to

$$\frac{\partial^2 v}{\partial \xi \partial \eta} = 0.$$

Hence show that the general solution $u(r,t)$ has the form $u(r,t) = \frac{1}{r}[f(r+ct)+g(r-ct)]$, where f and g are arbitrary (twice differentiable) functions.

1.6 Using Eqs. (1.26) and (1.27), show that the solution of the wave equation satisfying the initial conditions $u = u_0(x)$ and $\partial u/\partial t = v_0(x)$ at $t = 0$ is

$$u(x,t) = \frac{1}{2}\left[u_0(x-ct) + u_0(x+ct)\right] + \frac{1}{2}c\int_{x-ct}^{x+ct} v_0(s)\,ds.$$

This is known as D'Alembert's solution.

Chapter 2

Finite Difference Methods

2.1 The basic ideas

The most direct way to solve PDEs numerically is through the use of finite-difference approximations to the derivatives. The domain of integration spanned by the independent variables is covered with a finite grid of points at discrete space and/or time intervals. The dependent variables are stored and evaluated only at these grid points and derivatives are approximated by values at adjacent grid points. In this way the differential equation is reduced to a system of algebraic equations for the function values at each grid point. These equations are solved numerically.

As an example consider the ordinary differential equation (ODE),

$$\frac{du}{dx} = f(u, x), \tag{2.1}$$

on the interval $a \leq x \leq b$. We divide the interval up into a set of $N+1$ grid points, $x_i = a + \Delta x, i = 1, \ldots, N+1$, where $\Delta x = (b-a)/N$ is the grid length. Let u_i denote the value of u at the grid point x; i.e., $u_i = u(x_i)$.

There are various ways to approximate the derivative at the grid point:

I) *A forward difference*,

$$\left(\frac{du}{dx}\right)_i = \frac{u_{i+1} - u_i}{\Delta x},$$

II) *A backward difference*,

$$\left(\frac{du}{dx}\right)_i = \frac{u_i - u_{i-1}}{\Delta x},$$

III) *A centred (or central) difference,*

$$\left(\frac{du}{dx}\right)_i = \frac{u_{i+1} - u_{i-1}}{2\Delta x}.$$

The names are clear when one considers the graphical representation of these quantities (Fig. 2.1). One gains the impression from the figure that the centred difference may be a more accurate approximation to the tangent curve at x_i than either of the other differences. We show now that this is the case. Suppose $u(x)$ is expanded as a Taylor series about the point x_i. Then

$$u_{i+1} = u_i + \left(\frac{du}{dx}\right)_i \Delta x + \frac{1}{2}\left(\frac{d^2u}{dx^2}\right)_i \Delta x^2 + \frac{1}{6}\left(\frac{d^3u}{dx^3}\right)_i \Delta x^3 \tag{2.2}$$

and

$$u_{i-1} = u_i - \left(\frac{du}{dx}\right)_i \Delta x + \frac{1}{2}\left(\frac{d^2u}{dx^2}\right)_i \Delta x^2 - \frac{1}{6}\left(\frac{d^3u}{dx^3}\right)_i \Delta x^3. \tag{2.3}$$

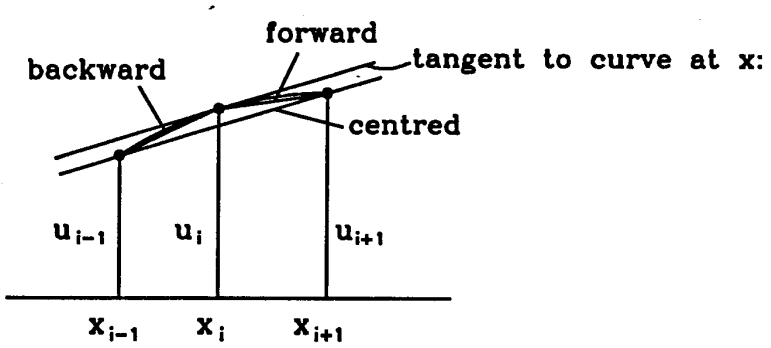

Figure 2.1: Graphical representation of forward, backward and centred differences.

It follows from (2.2 and 2.3) that

$$\left(\frac{du}{dx}\right)_i = \frac{u_{i+1} - u_i}{\Delta x} - \Delta x\left[\frac{1}{2}\left(\frac{d^2u}{dx^2}\right)_i + \ldots\right], \tag{2.4}$$

$$\left(\frac{du}{dx}\right)_i = \frac{u_i - u_{i-1}}{\Delta x} + \Delta x\left[\frac{1}{2}\left(\frac{d^2u}{dx^2}\right)_i - \ldots\right] \tag{2.5}$$

and by subtracting (2.3) from (2.2) that

$$\left(\frac{du}{dx}\right)_i = \frac{u_{i+1} - u_{i-1}}{2\Delta x} - \Delta x \left[\frac{1}{6}\left(\frac{d^3u}{dx^3}\right)_i + \ .....\right]. \tag{2.6}$$

Thus the error in replacing the derivative by a forward or backward difference is $0(\Delta x)$ while that for the central difference is $0(\Delta x^2)$. The error, characterized by the term in Δx or Δx^2 in the foregoing equations, is called the *truncation error* of the approximation. The leading power of Δx in the truncation error is the order of accuracy of the approximation. Clearly, for an approximation to be consistent it must be at least first-order accurate, and *consistency* in this sense *is a requirement of any finite-difference approximation*. It is clear that knowledge of a discrete set of values u_i, even if the approximations were perfect, gives less information than knowledge of the function $u(x)$. Suppose that the function $u(x)$ defined on the interval $(0, L)$ is represented by a Fourier series

$$u(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} [a_n \cos(2\pi nx/L) + b_n sin(2\pi nx/L)]. \tag{2.7}$$

If only $N+1$ values of $u(x)$ are known, we cannot expect to determine more than $N+1$ of the Fourier coefficients a_0, a_1, a_2, ..., b_1, b_2, ... It is intuitively obvious from Fig. 2.2 that to resolve a sine wave of a given wavelength, a minimum of three points are required. That is, $2\Delta \leq \lambda$.

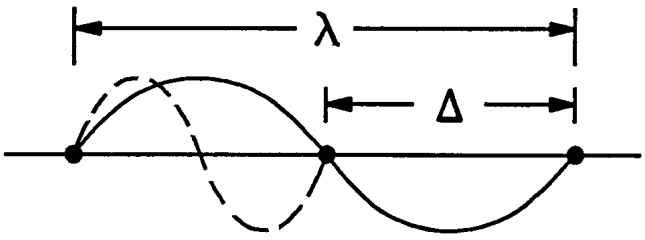

Figure 2.2: Representation of a sine wave of wavelength λ requires a minimum of three grid points.

It is clear also from Fig. 2.2 that a wave of wavelength Δ cannot be distinguished from one of wavelength 2Δ. In view of this, it is reasonable to expect that with only $N+1$ points, only waves longer than a certain

wavelength in (2.7) will be resolved, i.e. only the coefficients a_0 and a_1, a_2, ...$a_{N/2}$, b_1, b_2, ... $b_{N/2}$ can be determined (assuming that N is even). Note that the shortest wave has wavelength $L/(N/2) = 2L/N = 2\Delta x$, confirming the intuitive limit suggested by Fig. 2.2. This is an important result: a grid with a spacing Δx can not resolve wavelengths shorter than $2\Delta x$.

2.2 The advection equation

Many of the important ideas can be illustrated by reference to the advection equation (1.4) which we write in the form

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0. \tag{2.8}$$

We divide the (x, t)-plane into a series of discrete points ($i\Delta x$, $n\Delta t$) and denote the approximate solution for u at this point by u_i^n (see Fig. 2.3). We have seen that the true solution 'propagates' along the characteristics in the $x - t$ plane, which are given by $dx/dt = c$. This suggests replacing the time derivative by a *forward* difference and the space derivative by a *backward* difference, an idea that is explored further in Chapter 3: see Fig. 3.8. Then a possible finite-difference scheme for the equation is

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_i^n - u_{i-1}^n}{\Delta x} = 0. \tag{2.9}$$

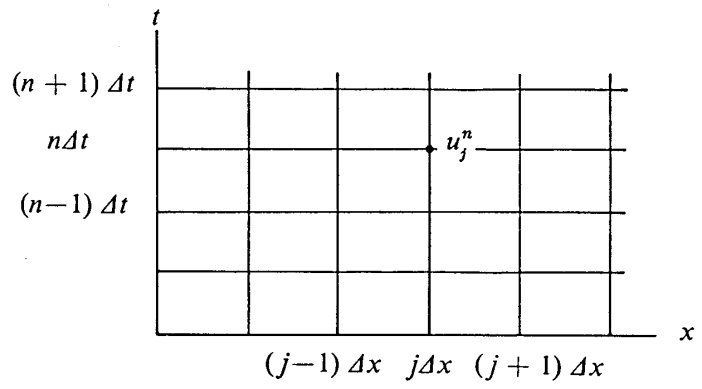

Figure 2.3: A finite-difference grid for finding an approximate solution to (2.3).

This scheme is called a *forward* and *upstream* scheme, upstream indicating the position of the point $i-1$ relative to the advection velocity c, assumed here to be positive.

Equation (2.9) is only one of many possible *consistent* finite-difference schemes for (2.8). There are many schemes that approach the PDE as Δx and Δt approach zero.

The difference between the numerical solution and the *true* solution is

$$u_i^n - u(i\Delta x,\ n\Delta t), \tag{2.10}$$

which represents the error of the numerical solution. Since $u(x,t)$ is normally unobtainable, we cannot usually determine this error. However, we can estimate the accuracy of a scheme in terms of the grid spacings Δx, Δt in the following way. For Eq. (2.9), for example, we substitute the true solution (assumed known) in the difference equation (2.9) to obtain

$$\frac{u(i\Delta x,(n+1)\Delta t) - u(i\Delta x,\ n\Delta t)}{\Delta t} + c\frac{u(i\Delta x,n\Delta t) - u((i-1)\Delta x,\ n\Delta t)}{\Delta x} = \varepsilon. \tag{2.11}$$

The quantitiy e is called the *truncation error* of the finite-difference scheme. It is a measure of how closely the true solution satisfies the equation for the scheme and therefore a measure of the accuracy of the scheme. Using expressions equivalent to (2.4) and (2.5) it follows at once that

$$\begin{aligned} \varepsilon &= \Delta t\left[\frac{1}{2}\frac{\partial^2 u}{\partial t^2} + \frac{1}{6}\frac{\partial^3 u}{\partial t^3}\Delta t + \ ....\right] \\ &- c\,\Delta x\left[\frac{1}{2}\frac{\partial^2 u}{\partial x^2} - \frac{1}{6}\frac{\partial^3 u}{\partial x^3}\Delta x\ ....\right]. \end{aligned} \tag{2.12}$$

As before, these are the terms that are 'truncated off' to make the PDE reduce to the finite difference scheme. The order of accuracy of a finite-difference scheme is the lowest power of Δx and Δt that appears in the truncation error. The scheme (2.9) is first order accurate, because

$$\varepsilon = O(\Delta t) + O(\Delta x) = O(\Delta t, \Delta x).$$

It is useful to make a distinction between orders of accuracy in space and in time, especially when the lowest powers of Δx and Δt are not the same. A necessary condition for the consistency of a scheme is that it be at least of the first order of accuracy.

2.3 Convergence

The truncation error of a consistent scheme can be made arbitrarily small by making Δt and Δx sufficiently small. However, we cannot be sure that this will reduce the error of the numerical solution given by (2.10). There are two questions that emerge:

(a) How does the error $u_i^n - u(i\Delta x, n\Delta t)$ behave when, for a fixed total time $n\Delta t$, the increments Δx, Δt approach zero?

(b) How does this error behave when, for fixed values of Δx and Δt, the number of time steps increases?

If in (a) the error approaches zero as the grid is progressively refined (i.e. as Δt, $\Delta x \to 0$), the solution is called *convergent*. If a *scheme* gives a convergent solution for any initial conditions, then the scheme itself is called convergent.

We may show that consistency of a scheme does not guarantee convergence. For example, we have noted that (2.9) is a consistent scheme for (2.8). However, additional conditions must be satisfied for it to be a convergent scheme.

Consider the numerical solution when the grid lines and characteristics are as shown in Fig. 2.4. The characteristic passing through the origin passes also through the grid point A, denoted by a square. Thus, according to our discussion in §1.3, the true value of u at A is equal to the value of u at the origin. However, in the numerical solution, the value of u at A is computed using the values at points denoted by circles. The domain spanned by these points is called the domain of dependence of the numerical scheme. For the grid shown, the grid point at the origin is outside this domain and therefore cannot affect the numerical solution at A. Therefore the error of the numerical solution at A can be arbitrarily large.

It is obvious that the situation is not improved if Δt and Δx are reduced in proportion to one another since the domain of dependence would remain the same. Refinement of the grid in this way cannot reduce the error of the numerical solution. Clearly, a necessary condition for the convergence of a scheme for Eq. (2.8) is that the characteristic defining the true solution at a grid point lies within the domain of dependence of the numerical solution at that point. This will happen when the slope of the characteristics is greater than slope of the line AB, i.e. when

$$c\Delta t \leq \Delta x. \tag{2.13}$$

This necessary condition for the convergence of (2.9) is the condition found by Courant *et al.* (1928) referred to in the Prologue.

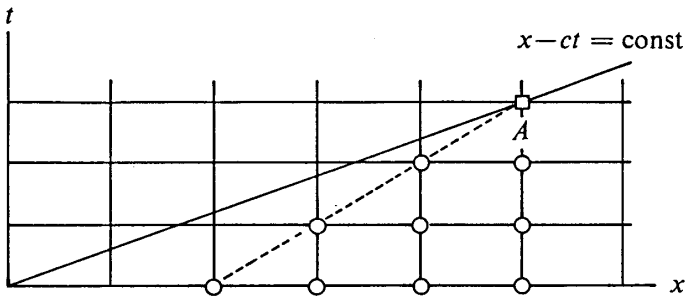

Figure 2.4: A possible grid for the solution of (2.9) showing the characteristics through the point A and the domain of dependence of this point, (i.e. the set of points, denoted by open circles, that affect the solution at A).

2.4 Stability

The answer to question (b) raised above depends on the *stability* of the numerical solution.

A difficulty in defining stability arises because the true solution does not have to be bounded (consider for example an unstable Eady-wave governed by the quasi-geostrophic potential vorticity equation). When we know that the true solution is bounded, as is the case of (2.8), we can construct a definition based on the error $u_i^n - u(i\Delta x, n\Delta t)$.

Definition: The solution u_i^n is *stable* if the error remains bounded as n increases. Moreover, *finite-difference scheme is stable* if it gives a stable solution for any initial conditions.

Stability of a scheme is a property of great importance. There exist consistent schemes of high order of accuracy that still give solutions diverging unacceptably far from the true solution. Thus, conditions for stability should be known. There are three methods for doing this which we illustrate using the forward and upstream scheme (2.9).

Direct method: We know that the true solution is bounded so it suffices to test the boundedness of the numerical solution. We may rewrite (2.9) as

$$u_i^{n+1} = (1-\mu)u_i^n + \mu u_{i-1}^n, \tag{2.14}$$

where

$$m = c\Delta t/\Delta x.$$

If $0 \leq \mu \leq 1$, which happens to be the necessary condition for convergence (2.13), we have

$$\left|u_i^{n+1}\right| \leq (1-\mu)\ \left|u_i^n\right| + \mu\left|u_{i-1}^n\right|. \tag{2.15}$$

We apply this at the point i where at time level $n+1$, $\left|u_i^{n+1}\right|$ is a maximum, $Max_{(i)}\left|u_i^{n+1}\right|$. Since the right-hand-side is increased only if each of the terms is replaced with $Max_{(i)}\left|u_i^n\right|$, we obtain from (2.15),

$$Max_{(i)}\left|u_i^{n+1}\right| \leq Max_{(i)}\left|u_i^n\right|.$$

This proves the boundedness of the numerical solution. Thus $0 \leq \mu \leq 1$ is a sufficient condition for the stability of (2.9).

While the direct method is simple, it is successful only for a rather limited number of schemes.

Energy method: This method is of much wider applicability and can be used as well for nonlinear equations.

If we know that the true solution is bounded, we test whether $\sum_i (u_i^{n+1})^2$ is bounded also. If it is, then every u_i^n must be bounded and the stability of the scheme has been proved. The method is called the energy method because in physical applications u^2 is often proportional to some form of energy.

Squaring (2.14) and summing over i gives

$$\sum_i (u_i^{n+1})^2 = \sum_i \left[(1-\mu)^2(u_i^n)^2 + 2\mu(1-\mu)u_i^n u_{i-1}^n + \mu^2(u_{i-1}^n)^2\right]. \tag{2.16}$$

We assume a *cyclic boundary condition*

$$u_{-1} = u_N$$

whereupon

$$\sum_i (u_{i-1}^n)^2 = \sum_i (u_i^n)^2, \tag{2.17}$$

Now we use Schwarz's inequality which states that

$$\sum ab \leq \sqrt{\sum a^2}\sqrt{\sum b^2}.$$

Equation (2.16) then shows that

$$\sum_i u_i^n u_{i-1}^n \leq \sqrt{\sum_i (u_i^n)^2}\sqrt{\sum_i (u_{i-1}^n)^2} = \sum_i (u_i^n)^2. \tag{2.18}$$

Using (2.17) and (2.18) it follows that if $1-\mu \geq 0$, (2.16) gives the inequality

$$\sum_i (u_i^{n+1})^2 \leq \left[(1-\mu)^2 + 2\mu(1-\mu) + \mu^2\right] \sum_i (u_i^n)^2,$$

or

$$\sum_i (u_i^{n+1})^2 \leq \sum_i (u_i^n)^2.$$

Therefore $1 \geq 1-\mu \geq 0$, together with the cyclic boundary condition, is a sufficient condition for the stability of (2.14).

Von Neumann's method: This method, otherwise called the Fourier series method is the most frequently used one for establishing the stability of a finite-difference scheme. However, we cannot use it for nonlinear equations, but must apply it to the linearized versions of these.

A solution to a linear equation can be expressed in the form of a Fourier series, each harmonic of which is a solution also. The idea is to test for the stability of a single harmonic solution. Then a necessary condition for the stability of the scheme is that it is stable to all admissible harmonics.

As an illustration of the method we consider again the advection equation (2.8), which has an analytic solution in the form of a single harmonic.

$$u(x,t) = Re[U(t)e^{ikx}]. \tag{2.19}$$

Here $U(t)$ is the wave amplitude and k the wavenumber. Substitution into (2.8) gives

$$\frac{dU}{dt} + ikcU = 0,$$

which has the solution

$$U(t) = U(0)e^{-ikct},$$

$U(0)$ being the initial amplitude. Hence

$$u(x,t) = \mathrm{Re}\left[U\,(0)\,e^{ik(x-ct)}\right], \tag{2.20}$$

as expected from the general solution (cf. 1.9).

In the von Neumann method we look for an analogous solution of the finite-difference equation (2.14). We substitute

$$u_j^n = \mathrm{Re}\left[U^{(n)} e^{ikj\Delta x}\right], \tag{2.21}$$

where $U(n)$ is the amplitude at level n. This is a solution provided that

$$U^{(n+1)} = (1-\mu)U^{(n)} + \mu U^{(n)} e^{-ik\Delta x}. \tag{2.22}$$

We can now study the behaviour of the amplitude $U(n)$ as n increases. We define an amplification factor $|\lambda|$ by

$$U^{(n+1)} = \lambda U^{(n)}. \tag{2.23}$$

Then

$$\left|\mathrm{U}^{(\mathrm{n})}\right| = |\lambda|^n \left|U^0 < B\right|,$$

where B is a finite number, i.e.

$$n \ln|\lambda| < \ln\left(B/|U^{(0)}|\right) = B', \quad \text{say.}$$

Since $n = t/\Delta t$, this necessary condition for stability becomes

$$ln < (B'\Delta t). \tag{2.24}$$

If we require boundedness of the solution for *finite* time t, condition (2.24) is that

$$ln|\lambda| \leq O(\Delta t).$$

Now define $|\lambda| = 1 + \delta$. Then since for $|\delta| << 1$, $ln(1+\delta) \approx \delta$, the stability criterion is equivalent to $\delta \leq 0(\Delta t)$, or

$$|\lambda| \leq 1 + O(\Delta t). \tag{2.25}$$

This is the von *Neumann necessary condition for stability.*

The von Neumann condition allows an exponential growth of the solution when $|\lambda| > 1$, but no faster. This is necessary when the true solution grows exponentially. However, when we know that the true solution doesn't grow, as in the problem at hand, it is customary to replace (2.25) by a sufficient condition

$$|\lambda| \leq 1. \tag{2.26}$$

In the present example, substitution of (2.23) into (2.22) gives

$$\lambda = 1 - \mu + \mu e^{-ik\Delta x}, \tag{2.27}$$

whereupon

$$|\lambda|^2 = 1 - 2\mu(1-\mu)(1 - cosk\Delta x). \tag{2.28}$$

It follows once again that $1 \geq 1 - \mu \geq 0$ is a sufficient condition for the stability of (2.14).

Equation (2.28) gives further information about the numerical solution. We consider $|\lambda|^2$ as a function of $\mu(= c\Delta t/\Delta x)$ for various fixed values of $k\Delta x$. These curves are all parabolas satisfying $|\lambda|^2 = 1$ at $\mu = 0$ and $\mu = 1$ and with a minimum value at $\mu = \frac{1}{2}$ that depends on $k\Delta x$ (Fig. 2.5). The absolute minimum $|\lambda|^2 = 0$ occurs when $\cos k\Delta = -1$, i.e. $k\Delta x = \pi$, corresponding with a wavelength $L = 2\mu/k = 2\Delta x$. Recall that this is the minimum resolvable wavelength. Figure 2.5 shows plots of $|\lambda|^2$ against μ for $L = 2\Delta x$, $4\Delta x$ and $8\Delta x$. Clearly, within the stable region $0 < \mu < 1$, the scheme is damping, but the degree of damping decreases as the wavelength increases. Since the true solution has a constant amplitude, the damping indicates an error due to finite differencing, At the shortest resolvable wavelength, the damping may be very large unless Δt is extremely small.

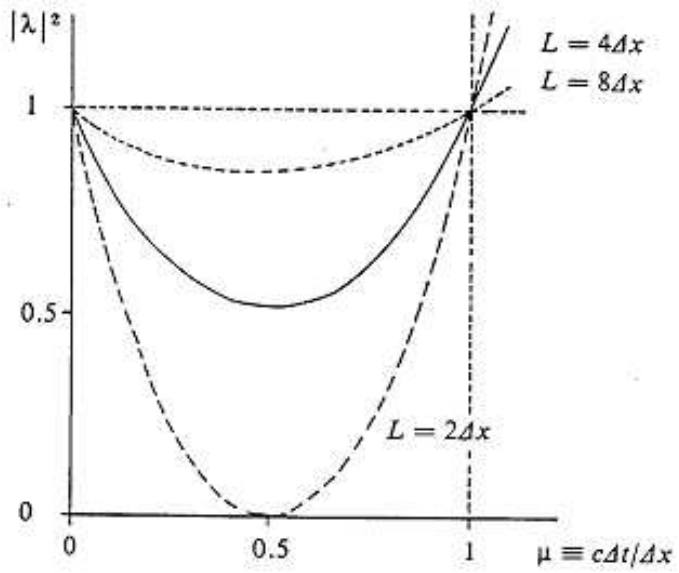

Figure 2.5: Plots of $|\lambda|^2$ versus μ for the finite-difference scheme (2.14) and for various values of L.

2.5 Time differencing schemes

We consider time differencing schemes first in the context of ODE's of the type

$$\frac{dU}{dt} = f(U,t), \quad U = U(t), \tag{2.29}$$

where t denotes the time. We divide the time axis into segments of equal length Δt and denote the approximate value of U at time $n\Delta t$ by $U^{(n)}$. We assume that $U^{(0)}$, $U^{(1)}$, ... , $U^{(n)}$ are known and seek to construct a scheme for computation of an approximate value $U^{(n+1)}$. There are many possibilities.

2.5.1 Two-level schemes

These relate $U^{(n+1)}$ to $U^{(n)}$ by approximating the exact formula

$$U^{(n+1)} = U^{(n)} + \int_{n\Delta t}^{(n+1)\Delta t} f(U,t)dt. \tag{2.30}$$

Euler (or forward) scheme: In this, $f(U,t)$ is assumed constant, equal to the value at time $n\Delta t$. Then

$$U^{(n+1)} = U^{(n)} + \Delta t\, f^{(n)},$$

where

$$f^{(n)} = f(U^{(n)}, n\Delta t). \tag{2.31}$$

The truncation error of this scheme is $O(\Delta t)$, i.e., it is a first-order accurate scheme. Because f is computed at one end of the interval, the scheme is said to be *uncentred* in time. Such schemes are invariably only first-order accurate.

Backward scheme: Now $f(U,t)$ is assumed constant, equal to $f^{(n+1)}$, i.e.

$$U^{(n+1)} = U^{(n)} + \Delta t\, f^{(n+1)}. \tag{2.32}$$

Since $f^{(n+1)}$ depends on $U^{(n+1)}$, the scheme is called *implicit*. For an ODE it may be simple to solve such a difference equation for $U^{(n+1)}$, but for PDE's this requires solving a set of simultaneous equations, one for each grid point of the computational region.

If the value of $f^{(n)}$ does not depend on $U^{(n+1)}$, such as in (2.31), the scheme is called *explicit*.

The truncation error of (2.32) is also $O(\Delta t)$.

Trapezoidal scheme: Here f is assumed to vary linearly between its values at $n\Delta t$ and $(n+1)\Delta t$ so that its mean value is used for the integral, i.e.

$$U^{(n+1)} = U^{(n)} + \tfrac{1}{2}\Delta t\,(f^{(n+1)} + f^{(n)}). \tag{2.33}$$

This is an implicit scheme also, but its truncation error is $O(\Delta t^2)$.

Next we consider a pair of iterative schemes:

Matsuno (or Euler-backward) schemes: First a step is made using the Euler scheme. The value of U_*^{n+1} so obtained is used to approximate $f^{(n+1)}$, say $f_*^{(n+1)}$, which is then used to make a backward step. Therefore

$$\left.\begin{aligned} U_*^{(n+1)} &= U^{(n)} + \Delta t\, f^{(n)}\,, \\ U^{(n+1)} &= U^{(n)} + \Delta t\, f_*^{(n+1)}\,, \end{aligned}\right\}, \tag{2.34}$$

where

$$f_*^{(n+1)} = f(U_*^{(n+1)}, (n+1)\Delta t).$$

This is an explicit scheme and is first-order accurate.

Heun scheme: This is similar to the previous one, it is explicit and second-order accurate, but the second step is made using the trapezoidal scheme, i.e.

$$\left.\begin{aligned} U_*^{(n+1)} &= U^{(n)} + \Delta t\, f^{(n)}\,, \\ U^{(n+1)} &= U^{(n)} + \tfrac{1}{2}\Delta t\,\left(f^{(n)} + f_*^{(n+1)}\right)\,. \end{aligned}\right\} \tag{2.35}$$

2.5.2 Three-level schemes

Except at the first time step, one can store the value $U^{(n-1)}$ and construct schemes using the 'time history' of U. These are three-level schemes. They approximate the formula

$$U^{(n+1)} = U^{(n-1)} + \int_{(n-1)\Delta t}^{(n+1)\Delta t} f(U,t)\,dt, \tag{2.36}$$

or they can use the additional value to improve the approximation to f in (2.30). Two examples are:

Leapfrog scheme: In this, f is taken to be constant, equal to the value at time $n\Delta t$, whereupon

$$U^{(n+1)} = U^{(n-1)} + 2\Delta t\, f^{(n)}. \tag{2.37}$$

This is probably the scheme most widely used in the atmospheric sciences. It is second-order accurate with truncation error $O(\Delta t^2)$.

Adams-Bashforth scheme: The scheme that is usually called the Adams-Bashforth scheme in the atmospheric sciences is, in fact, a simplified version of the original Adams-Bashforth scheme, which is fourth-order accurate. The simplified version is obtained when f in (2.30) is approximated by a value obtained at the centre of the interval Δt by a linear extrapolation using values $f^{(n-1)}$ and $f^{(n)}$. This gives

$$U^{(n+1)} = U^{(n)} + \Delta t\, (\tfrac{3}{2} f^{(n)} - \tfrac{1}{2} f^{(n-1)}). \tag{2.38}$$

This is an explicit, second-order accurate scheme.

Milne-Simpson scheme: In this, Simpson's rule is used to calculate the integral in (2.36), giving an implicit scheme.

2.6 Properties of schemes - the oscillation equation

The stability and other important properties of the foregoing time differencing schemes depend on the form of the function $f(U,t)$. In order to discuss these properties we need to prescribe this function. For applications to atmospheric models it is of particular interest to consider the case $f = i\omega U$, i.e., the ODE:

$$\frac{dU}{dt} = i\omega U, \qquad U = U(t). \tag{2.39}$$

We shall refer to this as the *oscillation equation*. We allow U to be complex (e.g. $u + iv$) so that, in general, the equation represents a system of two equations. The parameter ω is taken to be real and is called the *frequency*. Note that the time variation of a single harmonic of the advection equation satisfies this equation if $\omega = -kc$ (cf. Eq. (2.19) and the equation below it).

The general solution of (2.39) is $U(t) = U(0)e^{i\omega t}$, or, for discrete values $t = n\Delta t$,

$$U(n\Delta t) = U(0)e^{in\omega\Delta t}. \tag{2.40}$$

If we consider the solution in the complex plane, its argument rotates by $\omega\Delta t$ each time step and there is no change in amplitude. We use Neumann's method to analyse the properties of various schemes. Let

$$U^{(n+1)} = \lambda U^{(n)}, \tag{2.41}$$

and

$$\lambda = |\lambda|e^{i\theta}. \tag{2.42}$$

Then the numerical solution can be written

$$U^{(n)} = |\lambda|^n U^{(0)} e^{in\theta}. \tag{2.43}$$

Here θ represents the change in argument (or phase change) of the numerical solution during each time step. Since we know that the amplitude of the true solution does not change, we shall require $\lambda \leq 1$ for the stability.

Definition 1: We say that a scheme is *unstable, neutral or damping*, according as $|\lambda| > 1,\ = 1$ or < 1, respectively.

Definition 2: We say that a scheme is *accelerating*, has *no effect on the phase speed, or decelerating*, according as $\theta/(\omega\Delta t) > 1, = 1$ or < 1.

For accuracy it is desirable to have both the amplification and the relative speed, $(\theta/\Delta t)/\omega$, close to unity.

2.6.1 Two-level schemes

The three non-iterative two-level schemes in section 2.5.1 can be described by a single finite-difference equation

$$U^{(n+1)} = U^{(n)} + \Delta t\,(\alpha f^{(n)} + \beta f^{(n+1)}), \tag{2.44}$$

with a consistency requirement that $\alpha + \beta = 1$. The Euler scheme has $\alpha = 1$, $\beta = 0$; the backward scheme $\alpha = 0$, $\beta = 1$ and the trapezoidal scheme $\alpha = \frac{1}{2}$, $\beta = \frac{1}{2}$.

Applied to the oscillation equation, (2.44) gives

$$U^{(n+1)} = U^{(n)} + i\omega\Delta t\,(\alpha U^{(n)} + \beta U^{(n+1)}). \tag{2.45}$$

Then

$$U^{(n+1)} = \lambda U^{(n)}, \tag{2.46}$$

where

$$\lambda = \frac{1 + i\alpha p}{1 - i\beta p} = \frac{1 - \alpha\beta p^2 + i(\alpha + \beta)p}{1 + \beta^2 p^2}, \tag{2.47}$$

and

$$p = w\Delta t. \tag{2.48}$$

2.6.2 Euler scheme

$$\lambda = 1 + ip, \qquad |\lambda| = (1 + p^2)^{1/2}. \tag{2.49}$$

This scheme is always *unstable*. If $\Delta t << 1/|\omega|$, then $p << 1$ and $|\lambda| = 1 + \frac{1}{2}p^2 + \ldots$, i.e. $|\lambda| = 1 + O(\Delta t^2)$. Then $|\lambda| - 1$ is an order of magnitude less than the maximum allowed by the von Neumann stability condition. Nevertheless, an indiscriminate use of the Euler scheme for the solution of the atmospheric equations leads to amplification at a quite unacceptable rate.

2.6.3 Backward scheme

$$\lambda = (1 + \tfrac{1}{4}ip)/(1 + p^2), \qquad |\lambda| = (1 + p^2)^{-\frac{1}{2}}. \tag{2.50}$$

This scheme is stable irrespective of the size of Δt; it is an *unconditionally-stable* scheme. However, since $|\lambda| < 1$, it is damping *with the amount of damping increasing as the frequency* (and hence p) *increases*. The latter property is often considered desirable of a scheme, because it leads to the selective removal of undesirable high-frequency 'noise'.

2.6.4 Trapezoidal scheme

$$\lambda = \left(1 - \frac{1}{4}p^2 + ip\right) / \left(1 + \frac{1}{4}p^2\right), \qquad |\lambda| = 1. \tag{2.51}$$

This scheme is always neutral.

Note that both implicit schemes are stable irrespective of the size of Δt. The *iterative* two-level schemes can be described by a single equation also:

$$\left.\begin{aligned} &U_*^{(n+1)} = U^{(n)} + \Delta t\, f^{(n)}, \\ &U^{(n+1)} = U^{(n)} + \Delta t \left(\alpha f^{(n)} + \beta f_*^{(n+1)}\right) \\ &\alpha + \beta = 1. \end{aligned}\right\} \tag{2.52}$$

The Matsuno scheme has $\alpha = 0$, $\beta = 1$; the Heun scheme $\alpha = \frac{1}{2}$, $\beta = \frac{1}{2}$. When applied to the oscillation equation, (2.52) gives

$$\left.\begin{aligned} U_*^{(n+1)} &= \mathrm{U}^{(n)} + i\omega\Delta t\, U^{(n)} \\ \mathrm{U}^{(n+1)} &= \mathrm{U}^{(n)} + i\omega\Delta t \left(\alpha U^{(n)} + \beta U_*^{(n+1)}\right) \end{aligned}\right\} \tag{2.53}$$

whereupon

$$U^{(n+1)} = \lambda U^{(n)},$$

with

$$\lambda = 1 - \beta p^2 + ip. \tag{2.54}$$

2.6.5 Matsuno scheme

$$\lambda = 1 - p^2 + ip, \quad |\lambda| = (1 - p^2 + p^4)^{1/2}. \tag{2.55}$$

This scheme is stable if $|p| \leq 1$ so that to achieve stability we must ensure that

$$\Delta t \leq 1/|\omega|. \tag{2.56}$$

Accordingly, the Matsuno scheme is *conditionally stable*. The higher the frequency, the more restrictive the stability condition.

Differentiating (2.55) gives

$$\frac{d|\lambda|}{dp} = p(2p^2 - 1)/(1 - p^2 + p^4)^{1/2}.$$

Thus the amplification factor of this scheme has a minimum when $p = 1/\sqrt{2}$. Therefore, if we can choose Δt so that $0 < p < 1/\sqrt{2}$ for all the frequencies present, then the scheme will reduce the relative amplitudes of higher frequencies.

2.6.6 Heun scheme

$$\lambda = 1 - \frac{1}{2}p^2 + ip, \qquad |\lambda| = (1 + \frac{1}{4}p^4)^{1/2}.. \tag{2.57}$$

This is always > 1 so that the Heun scheme is always unstable. However, for small p,

$$|\lambda| = 1 + \frac{1}{8}p^4 + \ldots, \tag{2.58}$$

i.e. $|\lambda| = 1 + O(\Delta t^4)$. The instability is therefore quite weak. Experience shows that it can be tolerated when we can choose Δt sufficiently small.

The results of the five two-level schemes are summarized in Fig. 2.6. Since the amplification factors are all even functions of p, it is only necessary to show the curves for $p \geq 0$.

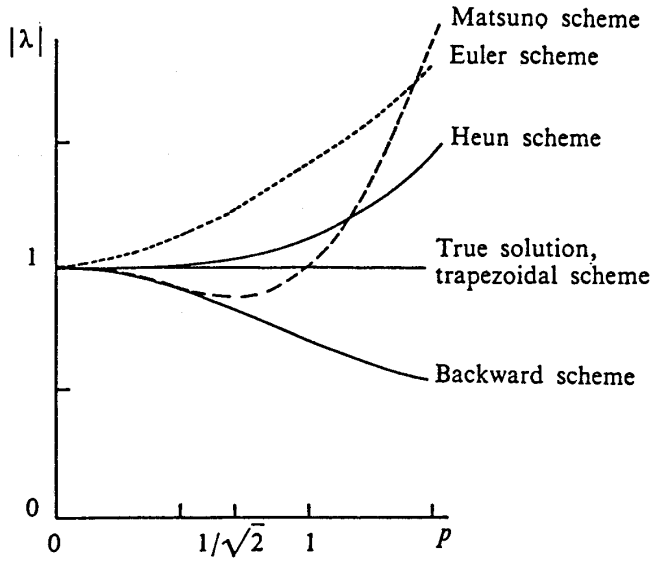

Figure 2.6: The amplification factors as a function of $p = \omega\Delta t$ for the five two level schemes and for the true solution.

2.6.7 Phase change of schemes

It is of interest to consider the phase change per time step, θ, together with the relative phase change per time step, θ/p, the former being defined by

$$\theta = \tan^{-1}[\mathrm{Im}(\lambda)/\mathrm{Re}(\lambda)]. \tag{2.59}$$

2.6.8 Euler and backward schemes

From (2.49) and (2.50)

$$\theta/p = (1/p)\tan^{-1} p, \tag{2.60}$$

and since $|(1/p)\tan^{-1} p| < 1$, $|\theta/p| < 1$ for a stable scheme. Thus both schemes are decelerating. For $p = 1$, $\theta/p = \pi/4$.

Matsuno scheme Equation (2.55) gives

$$\theta/p = (1/p)\tan^{-1}[p/(1-p^2)]. \tag{2.61}$$

For stability $|p| \leq 1$ and for relatively small frequencies ($p << 1$),

$$\theta/p = 1 + \frac{2}{3}p^2 + ...$$

Therefore, the Matsuno scheme is accelerating. When $p = 1$, $\theta/p = \pi/2$. The analysis of phase errors of schemes applied to the oscillation equation is not so important as an analysis of the amplification factor. Firstly, phase errors do not affect the stability of a scheme. Moreover, when the schemes are used to solve the PDEs of atmospheric motion, additional phase errors occur as a result of space differencing. These are usually the dominant source of error.

2.7 Three level schemes, computational modes

2.7.1 Leapfrog scheme

When applied to the oscillation equation, Eq. (2.37) gives

$$U^{(n+1)} = U^{(n-1)} + 2i\omega\Delta t\, U^{(n)}. \tag{2.62}$$

Three-level schemes such as this require two initial conditions to initialize the calculation. Physically, a single initial condition should have been adequate. It turns out that, besides this *physical initial condition*, three-level schemes require a *computational initial condition* $U^{(1)}$. This will have to be determined by using a two-level scheme for the first step. Now (2.23) gives

$$U^{(n+1)} = \lambda U^{(n-1)} \qquad U^{(n+1)} = \lambda^2 U^{(n-1)}. \tag{2.63}$$

whereupon (2.62) reduces to

$$\lambda^2 - 2ip\lambda - 1 = 0, \tag{2.64}$$

which is a quadratic equation for λ. The solutions are

$$\begin{aligned} \lambda_1 &= \sqrt{1-p^2} + ip \\ \lambda_2 &= -\sqrt{1-p^2} + ip. \end{aligned} \tag{2.65}$$

It follows that there are two solutions of the form $U^{(n+1)} = \lambda U^{(n)}$. In general, an m-level scheme gives $m-1$ solutions of this form. Each possible solution is called a mode. If such a solution represents an approximation to the true solution, it is clear that λ must tend to unity as $\Delta t \to 0$, i.e. as $p \to 0$. In (2.64), $\lambda_1 \to 1$, but $\lambda_2 \to -1$ as $\to 0$. Thus the solution with $\lambda\lambda_1$ is called the physical mode; that with $\lambda = \lambda_2$ is not an approximation to the solution and is called the *computational mode*.

A simple example. Consider the case $\omega = 0$. Then (2.39) becomes

$$\frac{dU}{dt} = 0 \quad \text{with solution} \quad U = \text{constant} U^{(0)}.$$

The leapfrog scheme gives $U^{(n+1)} = U^{(n-1)}$ and from (2.64) $\lambda = \pm 1$, or $\lambda_1 = 1$, $\lambda_2 = -1$.

Suppose the first step happens to give $U^{(1)} = U^{(0)}$. Then the solution with $\lambda\lambda_1$ gives the true solution, and consists of the physical mode only. In contrast, if the first step gives $U^{(1)} = -U^{(0)}$, then, for all n, $U^{(n+1)} = -U^{(n)}$; i.e. the solution would consist entirely of the computational mode. Apparently a good choice for the computational initial condition is crucial for a satisfactory numerical solution to the problem.

In general, since (2.62) is linear, its solution will be a linear combination of the two solutions

$$\begin{aligned} U_1^{(n)} &= \lambda_1^n U_1^{(0)} \quad \text{and} \quad U_2^{(n)} = \lambda_2^n U_2^{(0)}, \text{ i.e.,} \\ U^{(n)} &= a\lambda_1^n U_1^{(0)} + b\lambda_2^n U_2^{(0)}, \end{aligned} \tag{2.66}$$

where a and b are constants. To satisfy the two initial conditions ($U^{(0)}$, $U^{(1)}$ specified) requires that

$$\begin{aligned} U^{(0)} &= aU_1^{(0)} + bU_2^{(0)}, \\ U^{(0)} &= a\lambda_1 U_1^{(0)} + b\lambda_2 U_2^{(0)}, \end{aligned}$$

which can be solved for $aU_1^{(0)}$ and $bU_2^{(0)}$. Then (2.66) gives

$$U^{(n)} = [\lambda_1^n (U^{(1)} - \lambda_2 U^{(0)}) - \lambda_2^n (U^{(1)} - \lambda_1 U^{(0)})]/(\lambda_1 - \lambda_2). \tag{2.67}$$

Clearly, the amplitudes of the physical and computational modes are proportional to

$$\left|U^{(1)} - \lambda_2 U^{(0)}\right| \quad \text{and} \quad \left|U^{(1)} - \lambda_1 U^{(0)}\right|,$$

respectively, and these depend on $U^{(1)}$. If we were able to choose $U^{(1)} = \lambda_1 U^{(0)}$, the solution would consist of the physical mode only; if it happened that $U^{(1)}\lambda_2 U^{(0)}$, the solution would consist entirely of the computational mode. Unfortunately, for more complicated functions $f(U,t)$ than studied here, it is frequently not possible to determine λ_1 and λ_2 and we have to live with the presence of the computational mode.

Normally, $U^{(1)}$ is computed using one of the two-level schemes. The simplest method is to use the Euler scheme, but use of a higher-order scheme such as the Heun scheme gives a smaller amplitude of the computational mode.

2.7.2 Stability of the leapfrog scheme

Case $|p| < 1$ In (2.65), $1 - p^2 > 0$ and $|\lambda|_1 = |\lambda|_2 1$.

In this range of p, both the physical mode and the computational mode are stable and neutral. For the phase change, using (2.59) and (2.65),

$$\begin{aligned} \theta_1 &= \tan^{-1}\left[p/\sqrt{(1-p^2)}\right] \\ \theta_2 &= \tan^{-1}\left[-p/\sqrt{(1-p^2)}\right] \end{aligned} \tag{2.68}$$

and for $0 \leq p < 1$, $\theta_2 = \pi - \theta_1$. As $p \to 0$, $\theta_1 \to p$ while $\theta_2 \to \pi - p$ so that for small Δt, the physical mode approximates the true solution and the behaviour of the computational mode is quite different. For $p < 0$, $\theta_2 = -\pi - \theta_1$. The phase change is plotted in Fig. 2.7 as a function of $x = p/(1-p^2)^{1/2}$.

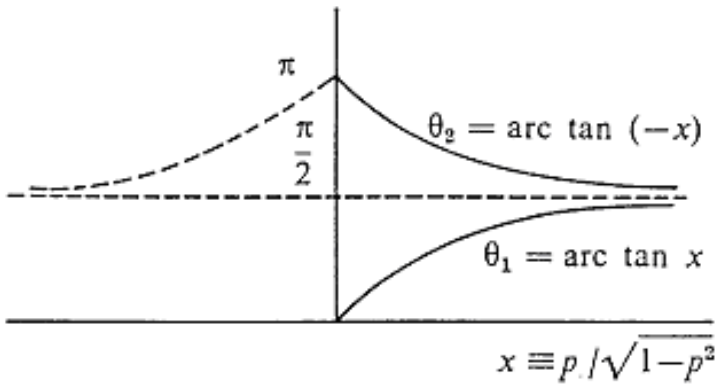

Figure 2.7: Phase change of the physical and of the computational mode for the leapfrog scheme.

For accuracy of the physical mode, θ_1 should closely approximate the phase change of the true solution, p. For $|p| << 1$, (2.68) gives

$$\theta_1 = p + \frac{1}{6}p^3 + \ldots,$$

i.e. the leapfrog scheme is accelerating. However, the acceleration is four times less than that of the Matsuno scheme.

One can illustrate the behaviour of the two modes in the complex plane. Note that

$$U_1^{(n)} = U_1^{(0)} e^{in\theta_1}, \qquad U_2^{(n)} = U_2^{(0)} e^{in(\pm\pi - \theta_1}. \tag{2.69}$$

For simplicity assume that $\theta_1 = \pi/8$ and that Im $(U^{(0)}) = 0$ at $t = 0$. As seen in the figure, the physical mode rotates in the positive sense by an angle θ_1 in each time step Δt, while in the case $p > 0$, the computational mode rotates by an angle $\pi - \theta_1$.

2.7.3 The Adams-Bashforth scheme

The stability analysis of this scheme when applied to the oscillation equation is left as an exercise (Ex. 2.1.). It turns out that the physical mode of the scheme is always unstable, but like the Heun scheme, the amplification is only by a fourth order term and it can be tolerated when Δt is sufficiently small. It has a useful property that the computational mode is damped.

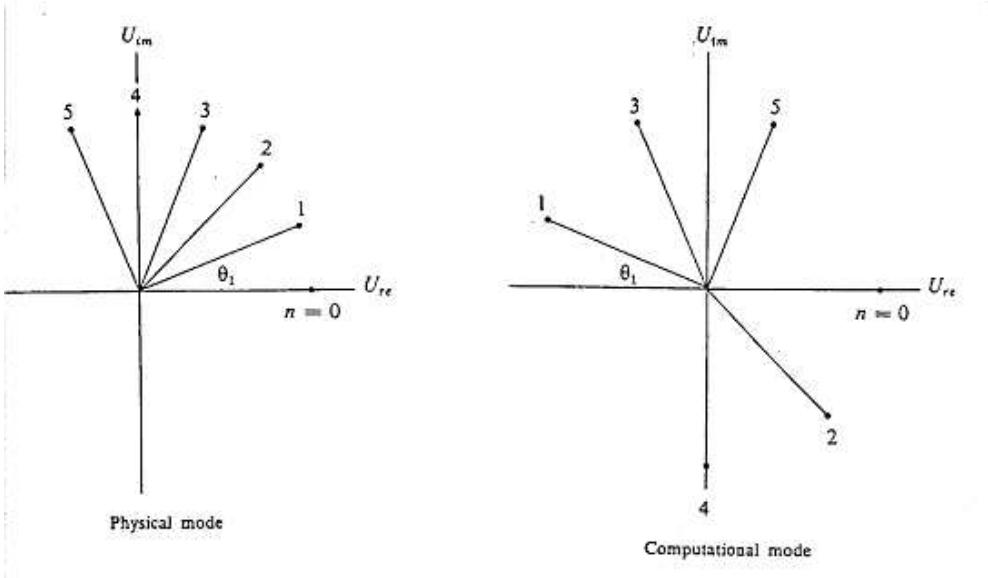

Figure 2.8: The physical and computational modes for the leapfrog scheme when $\theta = \frac{\pi}{8}$ and when the imaginary parts are zero at the initial moment. The first five time steps are shown.

2.8 Properties of schemes applied to the friction equation.

We consider now the properties of schemes when applied to the equation

$$\frac{dU}{dt} = -\kappa U, \qquad U = U(t), \qquad \kappa > 0. \tag{2.70}$$

Following Mesinger and Arakawa (1976), we refer to this as the friction equation. An equation of this type arises if we try to solve the diffusion equation [cf. Eq. (1.1)] .

$$\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2} \qquad \sigma > 0 \qquad \sigma > 0. \tag{2.71}$$

Then, the substitution $u(x,t) = \mathrm{Re}\left[U(t)e^{ikx}\right]$ gives

$$\frac{dU}{dt} = -\sigma k^2 U, \tag{2.72}$$

equivalent to Eq. (2.70) if $\kappa = \sigma\kappa^2$. Alternatively, if $U = u + iv$ represents the velocity vector in the complex plane, Eq. (2.70) describes the motion of a particle of unit mass subject to a frictional retardation proportional to the velocity. The general solution of (2.70) is

$$U(t) = U(0)e^{-\kappa t} \tag{2.73}$$

Thus both the real and imaginary parts of U (e.g. u and v) decrease exponentially with time.

The properties of schemes applied to Eq. (2.70) will be analyzed using the von Neumann method.

2.8.1 Noniterative two-level schemes - see Eq. (2.44)

Applied to the friction equation, (2.44) gives

$$U^{(n+1)} = U^{(n)} - \kappa\Delta t\,(\alpha U^{(n)} + \beta U^{(n+1)}). \tag{2.74}$$

Let

$$K = \kappa\Delta t. \tag{2.75}$$

Then (2.74) may be written

$$U^{(n+1)} = \frac{1 - \alpha K}{1 + \beta K} U^{(n)}. \tag{2.76}$$

Euler scheme

$$\alpha = 1, \qquad \beta = 0; \qquad \lambda = 1 - K.$$

Hence the stability requires that $|1 - k| \leq 1$, or $0 < K \leq 2$. We see at once that the stability criteria of particular schemes do not have to be the same when they are applied to different equations. In the case of (2.76), we would wish to be more demanding in the choice of Δt by choosing $K < 1$; this prevents the solution (2.76) from oscillating from time step to time step.

Backward scheme

$$\alpha = 0, \quad \beta = 1; \quad \lambda = 1/(1 + K).$$

Now stability requires only that $K > 0$. Furthermore, the solution does not oscillate in sign.

Trapezoidal scheme

$$\alpha = \frac{1}{2}, \quad \beta = \frac{1}{2}; \quad \lambda = (1 - K)/(1 + K).$$

Iterative two-level-schemes The scheme (2.52) when applied to (2.70) gives

$$U^{(n+1)} = (1 - K + \beta K^2)U^{(n)}. \tag{2.77}$$

Therefore both the Matsuno and Heun schemes are stable for sufficiently small values of K.

Leapfrog scheme One can show that when this scheme is applied to (2.71) the computational mode is unstable (see Ex. 2.2). The scheme is therefore unsuitable for solving the friction equation.

Adams-Bashforth scheme One can show that this scheme is stable for sufficiently small values of K and that the computational mode is damped (see Ex. 2.3).

2.9 A combination of schemes

The question arises: what do we if the equation contains both the oscillation and the friction term, i.e.

$$\frac{dU}{dt} = i\omega U - \kappa U. \tag{2.78}$$

For example, we may wish to use the leapfrog scheme because of the oscillation term $i\omega U$, but we have just seen that it cannot be used for the friction term κU. In this and in similar situations we can use different schemes for the different terms; e.g. we might use the leapfrog scheme for the oscillation term and the forward scheme for the friction term. This would give

$$U^{(n+1)} = U^{(n-1)} + 2\Delta t(i\omega\, U^{(n)} - \kappa U^{(n-1)}). \tag{2.79}$$

Other combinations would be possible also.

Exercises

(2.1) Show that when applied to the oscillation equation, the Adams-Bashforth scheme - Eq. (2.38) - gives

$$U^{(n+1)} = U^{(n)} + \Delta t\,(\tfrac{3}{2}U^{(n)} - \tfrac{1}{2}U^{(n-1)}). \tag{2.80}$$

Show further that the substitution (2.63) into this equation gives

$$\lambda^2 - \left(1 + \tfrac{3}{2}ip\right)\lambda + \tfrac{1}{2}ip = 0. \tag{2.81}$$

Deduce that for small p, the roots of this equation are approximately

$$\begin{aligned} \lambda_1 &= \left(1 - \tfrac{1}{2}p^2 - \tfrac{1}{8}p^4 - \ldots\right) + i\left(p + \tfrac{1}{4}p^3 + \ldots\right) \\ \lambda_2 &= \left(\tfrac{1}{2}p^2 + \tfrac{1}{8}p^4 - \ldots\right) + i\left(\tfrac{1}{2}p - \tfrac{1}{4}p^3 + \ldots\right) \end{aligned} \tag{2.82}$$

and that the amplification factors are: $|\lambda_1| = 1 + \frac{1}{4}p^2 + \ldots, \quad |\lambda_2| = \frac{1}{2}p + \ldots$. Note that the scheme is weakly unstable.

(2.2) Show that when applied to the friction equation (2.70), the leapfrog scheme gives

$$U^{(n+1)} = U^{(n-1)} - 2k\Delta t U^{(n)}, \tag{2.83}$$

and that the (complex) amplification factor has solutions

$$\begin{aligned} \lambda_1 &= -K + \sqrt{1 + K^2} \\ \lambda_2 &= -K - \sqrt{1 + K^2}, \end{aligned} \tag{2.84}$$

where $K = k\Delta t$. Show that the solution associated with λ_1 is the physical mode and that associated with λ_2 is the computational mode.

Deduce that the computational mode is always unstable. It follows that the leapfrog scheme is unsuitable for the solution of the friction equation.

(2.3) Show that when applied to the friction equation (2.70), the Adams-Bashforth scheme has the (complex) amplification factor

$$\lambda = \tfrac{1}{2}\left(1 - \tfrac{3}{2}K \pm \sqrt{1 - K + \tfrac{9}{4}K^2}\right). \tag{2.85}$$

Deduce that this scheme is stable for sufficiently small values of K and that the computational mode is damped.

(2.4) The linearized shallow-water equations for unidirectional waves on a homogeneous layer of nonrotating fluid of uniform undisturbed depth H are

$$\frac{\partial u}{\partial t} = -g\frac{\partial h}{\partial x}, \qquad \frac{\partial h}{\partial t} = -H\frac{\partial u}{\partial x}, \tag{2.86}$$

where $h(x, y)$ is the disturbed depth and $u(x, y)$ is the horizontal velocity perturbation. Show that solutions have the form

$$\left(u\left(x,t\right), h\left(x,t\right)\right) = \mathrm{Re}\left[\left(\hat{u}, \hat{h}\right) e^{ik(x-ct)}\right], \tag{2.87}$$

where $c = \pm\sqrt{gH}$. Taking central space differences, show that the corresponding differential-difference equations have solutions of the form

$$\left(u_j, h_j\right) = \mathrm{Re}\left[\left(\hat{u}, \hat{h}\right) e^{ik(j\Delta x - c^* t)}\right], \tag{2.88}$$

and that the corresponding dispersion relation is

$$c^* = \pm\sqrt{gH}\frac{\sin k\Delta x}{k\Delta x}. \tag{2.89}$$

Show that the dispersion equation in the analogous two-dimensional formulation of the continuous system, corresponding to (2.86), is $\omega^2 = gH\left(k^2 + l^2\right)$, where, for example,

$$h\left(x, y, t\right) = \mathrm{Re}\left[\hat{h} e^{i(kx+ly-\omega t)}\right].$$

Chapter 3

The advection equation

We return now to consider finite-difference schemes for the advection equation. This will set the stage for our later study of the equations governing atmospheric motion. We begin with the simplest one-dimensional linear form of the equation already discussed in section

3.1 Schemes with centred second-order space differencing

The one-dimensional linear advection equation is

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \qquad c = \text{constant}, \tag{3.1}$$

where $u = u(x,t)$, and its general solutionis $u = f(x - ct)$, where f is an arbitrary function. One of the finite-difference schemes for this equation is the forward and upstream scheme (2.9) or (2.14) which we showed in Chapter 2 to have excessive damping. If the space derivative in (3.1) is approximated by a central difference, we obtain

$$\frac{\partial u_j}{\partial t} = -c\frac{u_{j+1} - u_{j-1}}{2\Delta x}. \tag{3.2}$$

A number of schemes for (3.1) can be constructed by approximating the time derivative in (3.2) by one of the methods in section 2.2. For example, the leapfrog scheme gives

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} = -c\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}. \tag{3.3}$$

The properties of schemes constructed in this way can be inferred from the known properties of time-differencing schemes applied to the oscillation equation. We substitute in (3.2) a tentative solution in the form

$$uj = Re[U(t)e^{ikj\Delta x}] \tag{3.4}$$

whereupon

$$\frac{dU}{dt} = i\left(-\frac{c}{\Delta x}\sin k\Delta x\right)U. \tag{3.5}$$

This is a form of the oscillation equation studied in section 2.2 if

$$\omega = -\frac{c}{\Delta x}\sin k\Delta x. \tag{3.6}$$

Leapfrog scheme If we apply this to (3.5) we obtain

$$U^{(n+1)} = U^{(n-1)} + 2i\left(-c\frac{\Delta t}{\Delta x}\sin k\Delta x\right)U^{(n)}. \tag{3.7}$$

Let

$$p = -c\frac{\Delta t}{\Delta x}\sin k\Delta x. \tag{3.8}$$

Then, according to our earlier analysis we require $|p| \leq 1$ for stability, i.e.

$$\left|-\frac{c\Delta t}{\Delta x}\sin k\Delta x\right| \leq 1$$

for any possible k. But $|\sin k\Delta x| \leq 1$, so the criterion reduces to

$$\Delta t/\Delta x \leq 1, \tag{3.9}$$

which is, again, the CFL criterion.

Note that the maximum value of $|p|$, i.e. the minimum stability, is associated with the wave with $k\Delta x = \frac{\pi}{2}$. This is the component with wavelength $4\Delta x$, twice the shortest resolvable wavelength $2\Delta x$. There are two solutions for $U(0)$, the physical and the computational mode,

$$U_1^{(n)} = \lambda_1^n U_2^{(0)}, \qquad U_2^{(n)} = \lambda_2^n U_2^{(0)} \tag{3.10}$$

where λ_1 and λ_2 are given by (2.65). In the stable case, we have

$$\lambda_1 = e^{i\theta}, qquad\theta = \tan^{-1}\left[p/\sqrt{(1-p^2)}\right] \tag{3.11}$$

$$\lambda_2 = e^{i(\pm\pi-\theta)} = -e^{-i\theta}, \tag{3.12}$$

where the + or − sign is taken according as $p > 0$ or $p < 0$. Combining (3.10) with (3.4) we obtain the physical mode

$$u_j^n = \mathrm{Re}[U_1^{(0)} e^{ik(j\Delta x + \frac{\theta}{k\Delta t} n\Delta t)}], \tag{3.13}$$

and the computational mode

$$u_j^n = \mathrm{Re}[(-1)^n U_2^{(0)} e^{ik(j\Delta x + \frac{\theta}{k\Delta t} n\Delta t)}]. \tag{3.14}$$

These expressions, should be compared with the true solution (2.20), i.e.,

$$u(x,t) = \mathrm{Re}[U(0) e^{ik(x-ct)}]. \tag{3.15}$$

It is clear that the phase speed of the physical mode, c_1, is equal to $-\theta/k\Delta t$. Equation (3.12) shows that as $\Delta t \to 0$, $\theta \to p$ and Eq. (3.8) shows that as $\Delta x \to 0$, $p \to -ck\Delta t$. It follows that, as Δx, $\Delta t \to 0$, $c_1 \to c$; i.e. the phase speed of the physical mode approaches the phase speed of the solution, while $c_2 \to -c$. In addition, the computational mode changes sign at all grid points from time step to time step, because of the factor $(-1)^n$ in (3.14).

Matsuno scheme First the approximate values $u_j^{(n+1)}*$ are calculated using the forward scheme, i.e.,

$$\frac{u_j^{(n+1)^*} - u_j^n}{\Delta t} = -c\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}. \tag{3.16}$$

Then the are used in the backward scheme, that is

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -c\frac{u_{j+1}^{(n+1)^*} - u_{j-1}^{(n+1)^*}}{2\Delta x}. \tag{3.17}$$

We eliminate the starred quantities in (3.17) using (3.16) to obtain

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -c\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + c^2\Delta t\frac{u_{j+2}^n - 2u_j^n + u_{j-2}^n}{(2\Delta x)^2}. \tag{3.18}$$

Without the last term, this represents the finite-difference equation obtained when the forward scheme is used for the time derivative in Eq. (3.2). The third term approaches zero as Δx, $\Delta t \to 0$ and (3.18) is therefore a consistent scheme for the advection equation.

3.2 Energy method

We show now another example of the use of the energy method for testing stability. This method ca be applied also to non-linear equations, and to study the effect of boundary conditions on stability. The method can be used to test the stability of a group of schemes for solving (3.1). Many schemes of interest can be written as

$$u_j^{n+1} - u_j^n = -\tfrac{1}{2}\mu(u_{j+1}^* - u_{j-1}^*) \tag{3.19}$$

where

$$m = c\Delta t/\Delta x, \tag{3.20}$$

and $u_j^\star$ is a linear function of a number of values u_j^n.

Non-iterative two-level schemes take

$$u_j^* = \alpha u_j^n + \beta u_j^{n+1}. \tag{3.21}$$

Iterative two-level schemes, take

$$u_j^* = u_j^n - \tfrac{1}{2}\beta\mu(u_{j+1}^n - u_{j-1}^n). \tag{3.22}$$

Adam-Bashforth scheme take

$$u_j^* = \frac{2}{3}u_j^n - \frac{1}{2}u_{j-1}^n. \tag{3.23}$$

Here we analyse the stability of the non-iterative two-level schemes.

First we multiply (3.19) by u_j^* and sum over j to obtain

$$\sum_j u_j^*(u_j^{n+1} - u_j^n) = -\frac{1}{2}\mu\sum_j u_j^*(u_{j+1}^* - u_{j-1}^*).$$

If cyclic boundary conditions are assumed, the right-and side vanishes whereupon

$$\sum_j u_j^*(u_j^{n+1} - u_j^n) = 0. \tag{3.24}$$

Now

$$\sum_j \tfrac{1}{2}\left[(u_j^{n+1})^2 - (u_j^n)^2\right] \sum_j \tfrac{1}{2}(u_j^{n+1} + u_j^n)(u_j^{n+1} - u_j^n).$$

Substituting (3.21) into (3.24) and writing $\beta = 1 - \alpha$ gives

$$\sum_j \tfrac{1}{2}\left(\left(u_j^{n+1}\right)^2 - \left(u_j^n\right)^2\right) = \left(\alpha - \tfrac{1}{2}\right)\sum_j \left(u_j^{n+1} - u_j^n\right)^2. \tag{3.25}$$

Thus if $\alpha > \frac{1}{2}$ the scheme is unstable; if $\alpha = \frac{1}{2}$ it is neutral; if $\alpha < \frac{1}{2}$ the total energy, $\sum \frac{1}{2}\left(u_j^n\right)^2$, decreases monotonically with time.

3.3 Lax-Wendroff scheme

Unlike the schemes described so far, the two-step Lax-Wendroff scheme cannot be constructed by an independent choice of finite-difference approximations to the space and to the time derivative of the advection equation. Consider the stencil shown in Fig. 3.1. First, provisional values of u are calculated at the centre of the two rectangular meshes of the stencil, denoted by stars. This is done using centred space and forward time differencing.

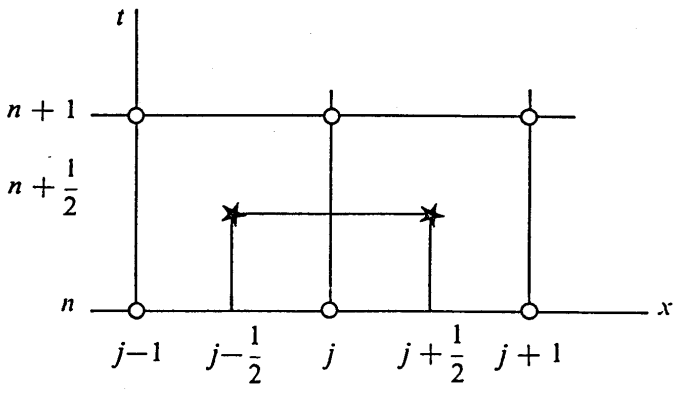

Figure 3.1: The space-time stencil for the Lax-Wendroff scheme.

The values of $u^n_{j+\frac{1}{2}}$ and $u^n_{-\frac{1}{2}}$ are obtained by taking arithmethic averages of the values u_j^n at the nearest grid points. Then

$$\frac{u^{n+\frac{1}{2}}_{j+\frac{1}{2}} - \frac{1}{2}(u^n_{j+1} + u^n_j)}{\frac{1}{2}\Delta t} = -c\frac{u^n_{j+1} - u^n_j}{\Delta x} \tag{3.26}$$

$$\frac{u_{j-\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}(u_j^n + u_{j-1}^n)}{\frac{1}{2}\Delta t} = -c\frac{u_j^n - u_{j-1}^n}{\Delta x}. \tag{3.27}$$

Using these provisional values a further step is made, this one is time-centred in both space and time, i.e.,

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -c\frac{u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - u_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x}. \tag{3.28}$$

Elimination of the provisional values from (3.28) using Eqs. (3.26) - (3.27) gives

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -c\Delta t\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \frac{1}{2}c^2\Delta t\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}. \tag{3.29}$$

This difference equation is rather similar to (3.18) obtained using simple centred space differencing and the Matsuno time difference, the only difference being in the last term. Again, this approaches zero as Δx, $\Delta t \to 0$; with Δt fixed it now approaches $\frac{1}{2}c^2\Delta t(\partial^2 u/\partial x^2)$, again a diffusion term, but with a coefficient half the previous one. Furthermore, this term is now calculated over an interval $2\Delta x$, and its damping effect will be a maximum at that wavelength. This kind of dependence on wavelength is often considered desirable of a finite difference scheme. This is because there are serious problems with finite-difference around $2\Delta x$. It is often possible to alleviate these problems by using a dissipative scheme which preferentially damps the two-grid interval waves.

While (3.18) was first-order accurate in time, (3.29) has truncation error $O[(\Delta x)^2 + (\Delta t)^2]$ and is therefore second-order accurate in both space and time.

Stability properties Substitute

$$u_j^n = Re[U(n)e^{ikj\Delta x}], \tag{3.30}$$

into (3.29) to obtain

$$U^{(n+1)} = [1 - \mu^2(1 - \cos k\Delta x) - i\mu \sin k\Delta x]U^{(n)}). \tag{3.31}$$

Therefore

$$\lambda = 1 - \mu^2(1 - \cos k\Delta x) - im \sin k\Delta x. \tag{3.32}$$

It is easy to show that

$$|\lambda| = \left[1 - 4\mu^2\left(1 - \mu^2\right)\sin^4 \tfrac{k\Delta x}{2}\right]^{\frac{1}{2}}. \tag{3.33}$$

Since $\left|\sin \frac{1}{2}k\Delta x\right| \leq 1$, the expression in brackets cannot fall below $1-4\mu^2(1-\mu^2)$, which is unity if $\mu = 1$ and has a minimum when $\mu^2 = \frac{1}{2}$. Hence $|\lambda| \leq 1$. Thus like the Lax-Wendroff scheme, the scheme is stable if

$$|c|\,\frac{\Delta t}{\Delta x} \leq 1,$$

again, the CFL condition. The scheme is damping for $|\mu| < 1$. For a given wavelength, the amplification factor has a minimum value

$$\lambda_{min} = (1 - \sin 4\frac{1}{2}k\Delta x)^{1/2}, \tag{3.34}$$

which equals zero for $k\Delta x = \frac{1}{2}\pi$, or $2\pi/k = 2\Delta x$ and approaches unity as $k \to 0$, i.e. as the wavelength tends to infinity. The amplification factors for waves of wavelength $2\Delta x$ and $4\Delta x$ are shown in Fig. 3.2. The amount of damping is generally quite large for shorter wavelengths. Unfortunately, it depends also on the time step and on the advection velocity (through m). This is a disadvantage of the Lax-Wendroff scheme because there is no reason why the amount of damping should depend on these quantities and it is either impractical or impossible to control the damping by changing them. For example, for small values of m, expansion of (3.33) gives

$$|\lambda| = 1 - 2\mu^2 \sin 4\frac{1}{2}k\Delta x + \ldots$$

showing that for a given amount of time (a fixed value of $n\Delta t$) the total damping will be approximately proportional to Δt. However, we would prefer to choose Δt to give the best accuracy and stability properties, not to give the optimum amount of damping. The Lax-Wendroff scheme has been widely used in atmospheric models on account of its reasonably good behaviour. It is second-order accurate, explicit, not unconditionally unstable, and has no computational mode. None of the schemes obtained by combining centred space differencing with one of the seven time differencing schemes studied in section 2.2 has all of these advantages. The dissipation of the scheme will not be harmful if its total effect is negligible compared with the physical dissipation, and it can be useful for controlling the shortest waves. If the physical dissipation is very small or non-existent, it is better to use a neutral scheme.

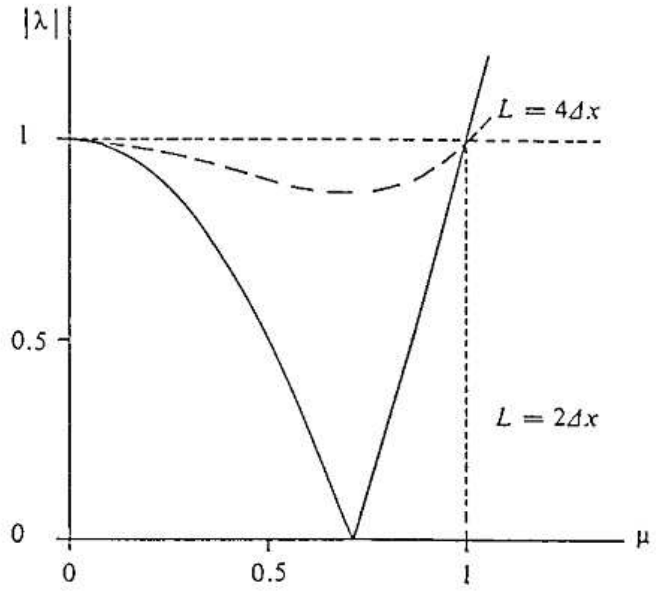

Figure 3.2: Amplification factor of the Lax-Wendroff scheme as a function of $\mu = c\Delta t/\Delta x$, for waves of wavelength $2\Delta x$ and $4\Delta x$.

3.4 Computational dispersion

The linear advection equation (3.1) has no dispersion in the sense that all wave modes have the same phase speed c. We shall show that this is generally not true of the finite-difference approximation to it.

Consider the equation

$$\frac{\partial u_j}{\partial t} + c\frac{u_{j+1} - u_{j-1}}{2\Delta x} = 0, \tag{3.35}$$

that is obtained by approximating $\partial u/\partial x$ by a centred space difference. An equation of this type is a differential-difference equation. Since the time derivative has retained its differential from, any error in (3.35) is due to the space differencing. Equation (3.35) has a solution in the form of a single harmonic component

$$u_j = \mathrm{Re}\left[U\left(t\right)e^{ikj\Delta x}\right], \tag{3.36}$$

provided that

$$\frac{dU}{dt} + ik\left(c\frac{\sin k\Delta x}{k\Delta x}\right)U = 0, \tag{3.37}$$

while (3.1) with the substitution gives

$$\frac{dU}{dt} + ikcU = 0. \tag{3.38}$$

Thus the solutions of (3.35) propagate with the phase speed $c^\star$ given by

$$c^* = c\frac{\sin\, k\Delta x}{k\Delta x}, \tag{3.39}$$

which is a function of the wavenumber k. Thus, the finite differencing in space gives rise to a dispersion of the waves. As $k\Delta x$ increases from zero, the phase speed $c^\star$ decreases monotonically from c and becomes zero for the shortest resolvable wavelength $2\Delta x$, when $k\Delta x = \pi$.

In summary, all waves propagate at a speed that is less than the true phase speed c, with this decelerating effect increasing as the wavelength decreases. The two-grid interval wave is stationary.

It is clear why the two-grid interval wave is stationary when we look at the plot of that wave, shown in Fig. 3.3. For such a wave, $u_{j+1} = u_{j-1}$ at all grid points, whereupon (3.35) gives a zero value for $\partial u_j/\partial t$.

The consequence of a reduction of the true advection speed is a general retardation of the advection process. Moreover, the dependence of this reduction on wavelength is particularly serious for the shortest waves.

If the pattern that is being advected represents a superposition of more than one wave, the false dispersion will result in a deformation of that pattern. In the atmosphere, small-scale patterns such as fronts, shear lines, etc. represent a superposition of many waves with a significant proportion of the energy in the shortest waves resolvable. For this reason, in NWP models, such patterns, if present in the initial fields, are rather rapidly deformed, until they acquire a form which is less sharp than in the beginning. Since such small-scale features are often associated with significant weather, the effect of computational dispersion require careful attention.

3.5 Group velocity

In linear dispersive wave theory, the group velocity is defined by

$$c_g = \frac{d\omega}{dk} = \frac{d}{dk}(kc), \tag{3.40}$$

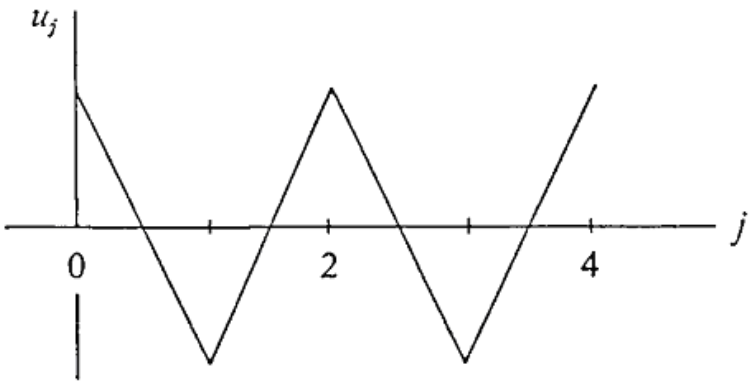

Figure 3.3: Plot of a two-grid interval wave, with a wavelength of $2\Delta x$.

where $\omega = \omega(k)$ is the dispersion relationship that relates frequency ω to wavenumber k. In the case of the linear advection equation (3.1), c is a constant, whereupon $c_g = c$. However, for the differential-difference equation (3.35),

$$c_g^\star = \frac{d}{dk}(kc^\star) = c\cos k\Delta x. \tag{3.41}$$

Thus as $k\Delta x$ increases from zero, the group velocity $c_g^\star$ decreases monotonically from c_g and becomes equal to $-c_g$ for the shortest resolvable wavelength $2\Delta x$. The results are summarized in Fig. 3.4. For the exact advection equation (3.1), both individual harmonic waves and wave packets (i.e. places where superposition of waves results in a maximum amplitude of a group of neighbouring wavenumbers) propagate at the same constant speed $c = c_g$. Introduction of the centred space finite difference in (3.35) makes both the phase speed and the group velocity decrease as the wavenumber increases. The error is especially great for the shortest resolvable wavelengths; waves with wavelengths less than $4\Delta x$ even have a negative group velocity. Thus wave packets made up of these waves propagate in the direction opposite to the advection velocity and opposite to the direction of individual waves.

As an example, consider the sinusoidal-like function $Y(x)$ that is slowly-varying in space on the scale of a grid length. Let us define

$$u_j = (-1)^j Y_j, \tag{3.42}$$

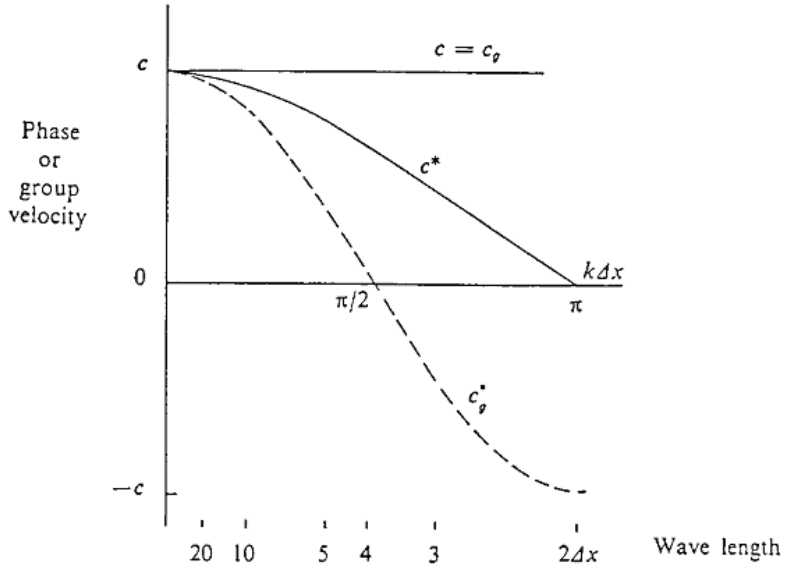

Figure 3.4: Phase speed and group velocity for the linear advection equation and the corresponding differential-difference equation with second-order space differencing.

shown in Fig. 3.5. Thus the function $\pm Y(x)$ is the *envelope* of the function u_j. Suppose that we calculate the advection of u_j using (3.35), i.e.

$$\frac{\partial Y_j}{\partial t} - c\frac{Y_{j+1} - Y_{j-1}}{2\Delta x} = 0,$$

which is obtained by substituting (3.42) into (3.35). It is clear that the advection of Y_j is governed by the same equation as for the advection of u_j, except that the advection velocity appears with an opposite sign! Therefore, as the individual short waves of the function u_j slowly propagate in the positive x direction, the envelope $\pm Y(x)$, which has much longer wavelength, propagates relatively fast in the opposite direction.

When $Y(x)$ is a sine function, so that it consists of only a single harmonic, it propagates with no change in shape. Then, because of (3.42), u_j must also be advected with no change in shape, from which we conclude that u_j consists also of a single harmonic component. If, on the other hand, the function $Y(x)$ consisted of a number of harmonic components, the shapes of both $Y(x)$ and u_j would change during the advection process as a result of the computational dispersion of these components.

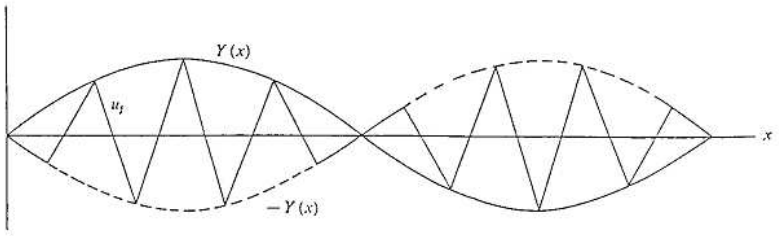

Figure 3.5: Functions u_j and $\pm Y(x)$, illustrating the propagation of an envelope of short as waves when their advection is calculated using second-order space differencing.

Analytic solution of (3.35). It is possible to solve Eq. (3.35) analytically. With the re-scaled time variable, $\tau = ct/\Delta x$, the equation can be written

$$2\frac{d}{d\tau}u_j(\tau) = u_{j-1}(\tau) - u_{j+1}(\tau). \tag{3.43}$$

This is the recurrence formula of the Bessel function of the first kind of order j, $J_j(\tau)$, i.e., a solution is

$$u_j(\tau) = J_j(\tau). \tag{3.44}$$

Several of these functions, of the order, are shown in Fig. 3.6. Note that $\forall j$, $(-1)^j J_j = J_{-j}$.

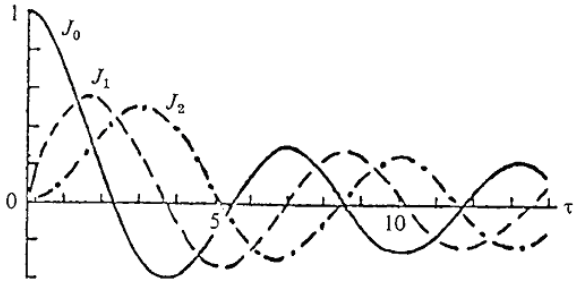

Figure 3.6: The Bessel functions $J_0(\tau)$, $J_1(\tau)$ and $J_2(\tau)$. Note that $J_{-1}(\tau) = -J_1(\tau)$, $J_{-2}(\tau) = J_2(\tau)$.

Now in (3.43), j can take any integer value, since the location for which we choose j = 0 is arbitrary. Thus, a solution that is more general than (3.44) is

$$u_j(\tau) = J_{j-p}(t),$$

where p is an arbitrary integer. Moreover, because Eq. (3.43) is linear, a still more general solution is

$$u_j(\tau) = \sum_{p=-\infty}^{\infty} a_p J_{j-p}(\tau), \tag{3.45}$$

where a_p are arbitrary constants. Now, $J_0(0) = 1$ and $J_n(0) = 0$ for $n \neq 0$. Hence,

$$p = j, \qquad u_j(0) = a_j.$$

It follows that the constants in (3.45) can be chosen to satisfy an arbitrary initial condition $u_j = u_j(0)$, $\forall j$. Therefore (3.45) is the general solution of (3.43) .

Suppose that the initial condition consists of a spike function

$$U_j(0) = \begin{cases} 1 & \text{for} \quad j = 0 \\ 0 & \text{for} \quad j \neq 0 \end{cases} \tag{3.46}$$

shown at the top of Fig. 3.7.

It follows from (3.43) that $du_j/d\tau = 0$ at $\tau = 0$, except at points $j - 1$ and $j = 1$, where it is equal to $-\frac{1}{2}$ and $\frac{1}{2}$, respectively.

If the spike function were an initial condition for the advection equation (3.1), it would simply translate to the right with uniform speed while remaining unchanged in shape. However, in the difference equation (3.35), the disturbance spreads out in both directions as series of waves; i.e. it disperses.

Figure 3.7 illustrates an example of computational dispersion associated with second-order space differencing. The main disturbance is advected at a speed only slightly less than the physical one. It mostly consists of the longer wave components which have an advection speed not much different from the physical advection velocity. One sees also the propagating of a group of short waves in the direction opposite to that of the physical advection. Since the appearance of these waves contradicts the physical properties of the advection equation, such waves are called *parasitic waves*.

For the spike function, the solution of the differential-difference equation is quite unsatisfactory as an approximation to the true solution. However, this is an extreme example, although it does illustrate one of the limitations involved with the numerical technique.

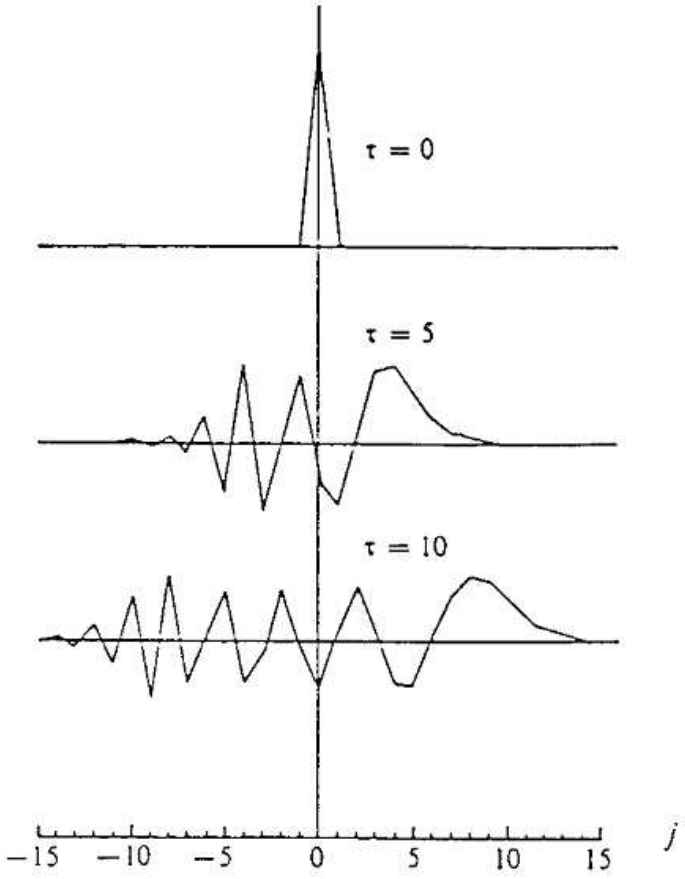

Figure 3.7: The analytic solution of (3.41) for the initial condition condition shown in the uppermost diagram for $\tau = 5$ and $\tau = 10$.

3.6 Schemes with uncentred differencing

The space differencing in the advection equation (3.1) can be approximated also using uncentred differencing. Two possibilities are to use a forward or backward scheme, according to whether c is negative or positive, i,e.,

$$\frac{\partial u_j}{\partial t} + c\frac{u_j - u_{j-1}}{\Delta x} = 0 \quad \text{for} \quad c > 0, \tag{3.47}$$

or

$$\frac{\partial u_j}{\partial t} + c\frac{u_{j+1} - u_j}{\Delta x} = 0 \quad \text{for} \quad c < 0, \tag{3.48}$$

These equations are again differential-difference equations. In both cases, the differences are calculated on the side from which the advection velocity reaches the centre point. For this reason we call Eqs. (3.47)-(3.48) an *upstream difference schemes.*

A series of schemes for the advection equation can be constructed by approximating the time derivatives in Eqs. (3.47)-(3.48) by one of the many possible consistent methods. The resulting schemes will be only first order accurate. However, they have one advantage over centred schemes in space when applied to the advection equation: with upstream differences, a disturbance cannot propagate in the direction opposite to the physical advection. Thus no parasitic waves will contaminate the numerical solution. Suppose that a forward difference is used for the time derivative in (3.1) and that $c > 0$; then (3.47) gives

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \tag{3.49}$$

This scheme was used in Chapter 2 where it was found to be damping, with the amount of damping depending on the wavelength. The maximum damping occurred for the $2\Delta x$ wave. The analytic solution of the difference equation (3.49) was discussed by Wurtele (1961).

The advantage that is achieved, at least in principle, by using upstream differencing as compared with centred or downstream differencing (e.g. 3.47 for $c > 0$), can be illustrated by considering the *domain of influence* of a grid point in different schemes.

We stay with the case $c > 0$. As shown in Chapter 1 (section 1.3), a grid point value can be said to propagate along the characteristics $x - ct =$ constant. Figure 3.8 shows a grid point marked by a circle with the associated characteristic passing through it. With upstream differencing as in (3.49), the value at the grid point will influence the values at points within the domain shaded by vertical lines. The figure shows also the domains of influence with centred and downstream differencing. One of the domains of influence, that given by upstream differencing, is clearly the best approximation to the characteristic line representing the domain of influence in the solution.

The foregoing discussion suggests constructing a scheme for Eq. (3.1) by tracing a characteristic back from a point $(j\Delta x, (n+1)\Delta t)$ to intersect the previous time level $t = n\Delta t$ and calculating the value at the point of intersection by interpolation (see Fig. 3.9). We then set $u_j^{n+1} = u^\star$. If we choose

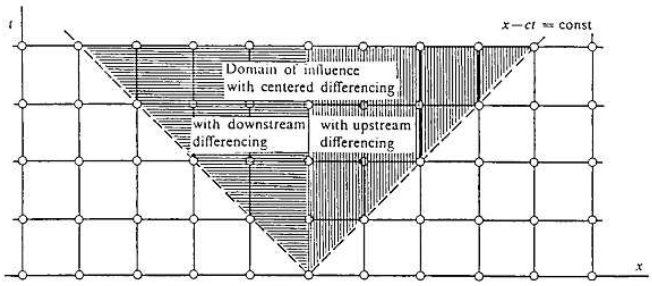

Figure 3.8: Domains of influence of a grid point for the scheme (3.49) with upstream differencing as well as for schemes with centred differencing, downstream differencing and the true solution.

a linear interpolation procedure that employs values at two neighbouring points at the time $n\Delta t$, we obtain

$$u_j^{n+1} = u_{j-1}^n + \frac{u_j^n - u_{j-1}^n}{\Delta x}(\Delta x - c\Delta t).$$

This is identical to the scheme (3.49), with upstream differencing. If a quadratic interpolation procedure is chosen using three neighbouring points, one obtains the Lax-Wendroff scheme (Ex. 3.1).

Analytic solution of 3.47 Further insight into the properties of schemes that can be obtained from (3.47) may be obtained from an analytic solution to this differential-difference equation. For small values of Δt this solution will approximate the solution obtained from the difference schemes. As before, we scale the time by writing $\tau = ct/Dx$. Then Eq. (3.47) becomes

$$\frac{d}{d\tau}u_j(\tau) + u_j(\tau) - u_{j-1}(\tau) = 0. \tag{3.50}$$

It may easily be verified that a solution of this equation is the Poisson frequency function.

$$u_j(\tau) = \begin{cases} \frac{e^{-\tau}\tau^{j-p}}{(j-p)!} for\ j \geq p \\ 0 for\ j < p \end{cases} \tag{3.51}$$

Again, p is an arbitrary integer which relates to the fact that the location of the point $j = 0$ is arbitrary. An example of the Poisson frequency function

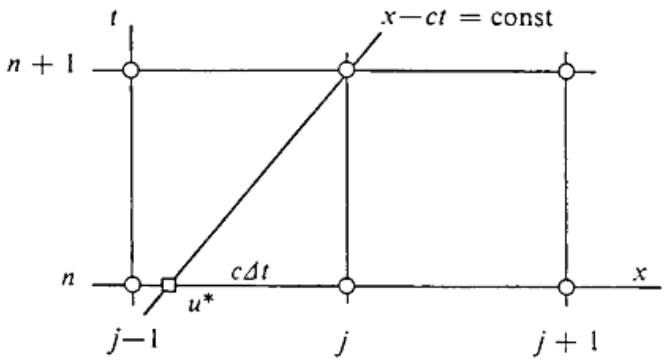

Figure 3.9: Sketch for the construction of schemes by calculation of a previous value on a characteristic passing through the point $(j\Delta x,\ (n+1)\Delta t)$.

is shown in Fig. 3.10 for the case when $\tau = 4$. The inclusion of a vertical scale is unnecessary because the area enclosed by the graph is always unity. In particular, when $\tau = 0$, the histogram consists of a single rectangle with its ordinate equal to unity. Thus it corresponds with the initial pulse-like disturbance (3.46) studied earlier. As τ increases, (3.51) transforms into a skew bell-shaped histogram of the type shown in Fig. (3.10). Its mean position on the x-axis is

$$\sum_{j-p=0}^{\infty} (j-p)\frac{e^{-\tau}\tau^{j-p}}{(j-p)!}\tau,$$

i.e., it moves at unit non-dimensional speed. Thus the mean position propagates at the physical advection velocity. However, the maximum point of the histogram logs behind as indicated by the skewed-shape of the histogram shown in Fig. 3.10. Negative values of u_j, which would be physically unrealistic, do not occur and there are no parasitic waves on the opposite side of zero from the direction of the physical advection. Finally, since the total amount of the advected quantity is exactly conserved, but the disturbance is damped quite rapidly. As before, a more general solution than (3.51) is the linear combination

$$\sum_{p=-\infty}^{j} a_p \frac{e^{-\tau}\tau^{j-p}}{(j-p)!}, \tag{3.52}$$

where a_p are arbitrary constants. Putting $\tau = 0$ in this formula gives

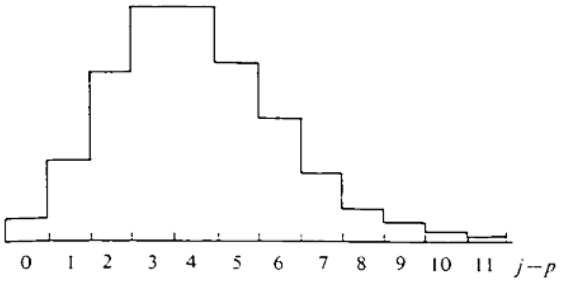

Figure 3.10: The Poisson frequency function (3.51) when $\tau = 4$.

$$u_j(0) = a_j,$$

so that the constants ap can be chosen to satisfy arbitrary initial conditions. The form of (3.52) indicates that the value $u_j(t)$ at a point j can be attributed to the superposition of the effect of the initial values at that point and of the initial values at all the points located upstream of it.

Figure 3.11 compares the solution (3.45) for centred differencing with that from (3.52) for upstream differencing for the initial disturbance

$$u_j(0) = \begin{cases} 1 & \text{for} \quad j = -1,\, 0\,,\, 1, \\ 0 & \text{otherwise,} \end{cases}$$

at three nondimensional times. If the grid distance is 300 km and $c = 15\ ms^{-1}$, then 5 units of nondimensional time corresponds with about 1 day in physical time. Therefore, the damping effect of upstream differencing is seen to be appreciable.

3.7 Schemes with centred fourth-order-space differencing

Many of the difficulties described above, in particular the phase speed error and the computational dispersion, result from the approximation used for

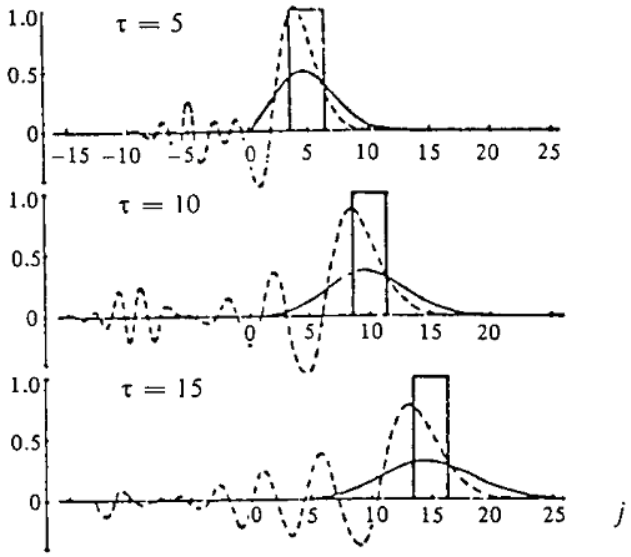

Figure 3.11: Analytic solutions of the exact advection equation (heavy solid line), of the equation using centred-differencing (dashed line), and of the equation using upstream differencing (thin solid line), for three different values of non-dimensional time.

space differencing. Therefore we are led to consider schemes using high-order space differencing.

Using the Taylor series expansion discussed in section 2.1, one can show that

$$\frac{\partial u}{\partial x} = \frac{4}{3}\frac{u_{j+1} - u_{j-1}}{2\Delta x} - \frac{1}{3}\frac{u_{j+2} - u_{j-2}}{4\Delta x} + O(\Delta x^4),$$

which is a fourth-order accurate approximation to $\partial u/\partial x$ (see Ex. 3.2). Then, in analogy with (3.2), the differential-difference approximation to (3.1) is

$$\frac{\partial u_j}{\partial t} + c\left(\frac{4}{3}\frac{u_{j+1} - u_{j-1}}{2\Delta x} - \frac{1}{3}\frac{u_{j+2} - u_{j-2}}{4\Delta x}\right) = 0. \tag{3.53}$$

Then, taking $u_j(t) = Re[U(t)e^{ikj\Delta x}]$ as before, we obtain the phase speed

$$c** = c\left(\frac{4}{3}\frac{\sin\, k\Delta x}{k\Delta x} - \frac{1}{3}\frac{\sin\, 2k\Delta x}{2k\Delta x}\right), \tag{3.54}$$

instead of $*c = c(\sin k\Delta x/k\Delta x)$ for the case of second-order differencing. For small $k\Delta x$, the latter formula gives

$$c^* = c\left(1 - \frac{1}{3!}(k\Delta x)^2 + \ldots\right),$$

whereas Eq. (3.54) gives

$$c^{**}c\left(1 - \frac{4}{5!}(k\Delta x)^4 + \ldots\right).$$

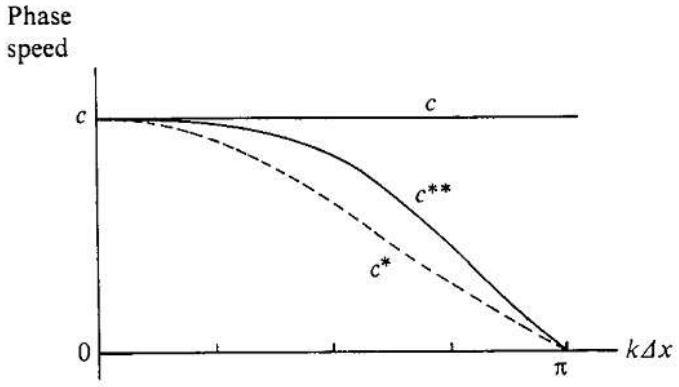

Figure 3.12: Phase speed for the linear advection equation, c, and for the corresponding differential-difference equations with second-order ($c^\star$) and with fourth-order ($c^{\star\star}$) centred-space differencing.

Thus, even though the decelerating effect is still present, the phase speed error is considerably reduced. The two phase speeds c^* and c^{**} are shown in Fig. 3.12 for all values of k. There is clearly a very significant increase in accuracy of the phase speed for large-scale and medium-scale waves. However, as the wavelength approaches its minimum value of $2\Delta x$, the increase in phase speed obtained by fourth order differencing diminishes, until, finally, the wave with wavelength $2\Delta x$ is again stationary. Moreover, for short waves the slope of the phase speed curve is greater than with second-order differencing implying that the computational dispersion of these waves is greater. Thus a short wave disturbance will be distorted more rapidly.

The use of additional grid points for higher order space differencing does introduce some side difficulties. As in the case of higher order time differencing, it leads to the appearance of computational modes, but in space. Moreover, the formulation of boundary conditions becomes more complicated. For small-scale disturbances, comparable to two grid intervals in space, no finite-difference scheme is very satisfactory

3.8 The two-dimensional advection equation

The two-dimensional linear advection equation is

$$\frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} + V\frac{\partial u}{\partial y} = 0, \tag{3.55}$$

where $u = u(x, y, t)$ is a function of two space variables x, y and of time t, and U, V are components of the advection velocity, assumed constants. The advection speed is

$$c = \sqrt{U^2 + V^2}. \tag{3.56}$$

The corresponding differential-difference equation is

$$\frac{\partial}{\partial t}u_{j,m} = -U\frac{u_{j+1,m} - u_{j-1,m}}{2\Delta x} - V\frac{u_{j,m+1} - u_{j,m-1}}{2\Delta y}, \tag{3.57}$$

where the space derivatives are approximated by second-order central differences. Here, denotes $u(j\Delta x, m\Delta y)$ and the grid-points are located at points $x = j\Delta x$, $y = m\Delta y$. The substitution

$$u_{j,m} = \mathrm{Re}\left[\hat{U}(t)e^{i(kx+ly)}\right], \tag{3.58}$$

into (3.57) gives the oscillation equation

$$\frac{d\hat{U}}{dt} = i\left(-\frac{U}{\Delta x}\sin\, k\Delta x - \frac{V}{\Delta y}\sin\, l\Delta y\right)\hat{U}. \tag{3.59}$$

If the leapfrog scheme is used for the time derivative, the stability criterion is

$$\left|\left(\frac{U}{\Delta x}\ \sin\, k\Delta x + \frac{V}{\Delta y}\ \sin\, l\Delta y\right)\Delta t\right| \leq 1 \tag{3.60}$$

This equation has to be satisfied for all possible values of the wavenumbers k and l. Let us consider the case $\Delta x = \Delta y = d$, say. As before, the minimum resolvable wavelength in the x-direction occurs when $kd = \pi$. Likewise, the minimum resolvable wavelength in the y-direction occurs when $ld = \pi$. Hence the admissible region of wavenumber space for the square grid is the shaded region shown in Fig. 3.12. Consider the function

$$\Lambda(k, l, U, V) = \frac{U}{\Delta x}\ \sin\, k\Delta x + \frac{V}{\Delta y}\ \sin\, l\Delta y \geq 0$$

in this region. This function has an extremum where $\partial\Lambda/\partial k = 0$ and $\partial\Lambda/\partial l = 0$, i.e., at $kd = \frac{\pi}{2}$, $ld\frac{\pi}{2}$, the centre of the region. At this point a wave has

wavelength components $2\pi/k = 4d$, $2\pi/l = 4d$. For a given value of the advection speed $c(U^2 + V^2)^{\frac{1}{2}\frac{1}{2}}$, the left-hand-side of (3.59) has a maximum value at this point if $\partial\Lambda/\partial U = 0$ and $\partial\Lambda/\partial V = 0$, i.e., if $1 - U/|c|^2 - U^2)^{\frac{1}{2}\frac{1}{2}} = 0$, which requires that $U = V$. In this case the advection velocity makes an angle of $\pi/4$ with the x-axis and $U = V = c/\sqrt{2}$. Thus the stability criterion (3.60) becomes

$$\sqrt{2}c\frac{\Delta t}{\Delta x} \leq 1. \tag{3.61}$$

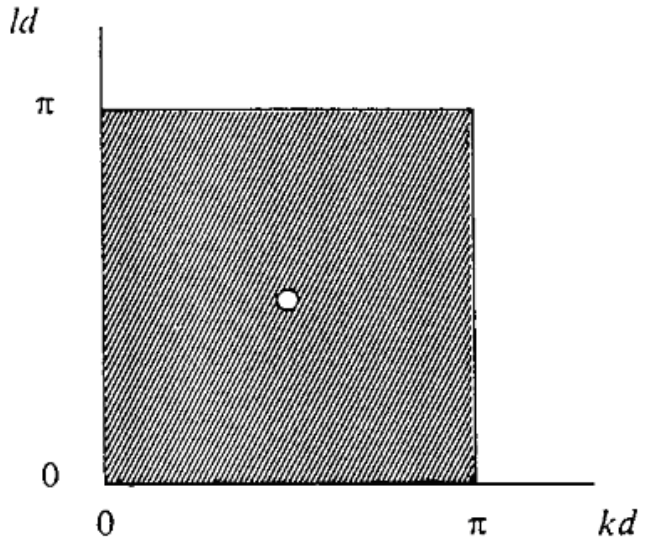

Figure 3.13: Admissible region of wavenumber for a square two-dimensional grid, with grid length d.

It follows that we must choose a smaller time step ($1/\sqrt{2}$ times smaller) than in the one-dimensional case to ensure stability. The minimum stability is associated with wavelengths in both the x and y directions twice as long as the shortest resolvable wavelength of $2d$, just as in the one-dimensional case. The two-dimensional wavenumber of this wave is $(k^2 + \lambda_2)^{1/2}$ which is greater by a factor of $\sqrt{2}$ than wavenumbers along the axes, and its wavelength is shorter by the same factor. This applies to all waves with $k = l$.

3.9 Aliasing error and nonlinear instability

We consider now the one-dimensional nonlinear advection equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0, \tag{3.62}$$

discussed in section 1.3. As shown there, the general solution of this equation is

$$u = F(x - ut), \tag{3.63}$$

where F is some arbitrary function. Here we consider the effect of the multiplication in (3.62). When performed in finite-differences, it results in an error related to the inability of the discrete grid to resolve wavelengths shorter than $2\Delta x$, i.e. wavenumbers greater than $k_{max} = \pi/\Delta x$. Consider, for example, the function

$$u = \sin kx, \tag{3.64}$$

where $k < k_{max}$. Substitution into the nonlinear term of (3.62) gives

$$u(\partial u/\partial x) = k \sin kx \cos kx = \frac{1}{2} k \sin 2kx.$$

Hence, if the wavenumber in (3.64) lies in the interval $\frac{1}{2}k_{max} < k \leq k_{max}$, the nonlinear term will give a wavenumber that is outside the range that can be resolved by the grid. Such a wave cannot be properly reproduced in a finite-difference calculation.

Consider a wave for which $k > k_{max}$, e.g., let $l = \Delta x$ as indicated by the continuous line in Fig. 3.16. Knowing only the values at grid points it is not possible to distinguish this wave from the one shown by the dashed line. Thus from the considerations explained in the section 2.1, where it is assumed that only the longest waves are present, there will be an error. This is called an *aliasing error*.

In a more general case, suppose that the function u consists of a number of harmonic components

$$u = \sum_n u_n.$$

Then the nonlinear term will contain products of harmonics of different wavelengths, such as

$$\sin k_1 x \cos k_2 x = [\sin(k_1 + k_2)x + \sin(k_1 - k_2)x].$$

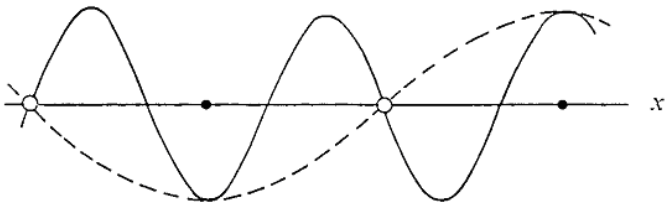

Figure 3.14: A wave of wavelength $\frac{4}{3}\Delta x$, misrepresented by the finite difference grid as a wave of wavelength $4\Delta x$.

Thus, even if a finite-difference calculation is begun with waves which all have $k \leq k_{max}$ nonlinear interaction will lead ultimately to wave components with $k > k_{max}$ and these waves will be mispresented.

Note that for all k,

$$\sin kx = sin[2k_{max} - (2k_{max} - k)]x,$$

and since $k_{max} = \pi/\Delta x$, this becomes

$$\sin\ kx = \sin \frac{2\pi}{\Delta x} x\ \cos\ \left(\frac{2\pi}{\Delta x} - k\right) x - \cos\ \frac{2\pi}{\Delta x} x\ \sin\ \left(\frac{2\pi}{\Delta x} - k\right) x.$$

In particular, at grid points $x = j\Delta x$,

$$\sin \frac{2\pi}{\Delta x} j\Delta x = 0,\ \cos\ \frac{2\pi}{\Delta x} j\Delta x = 1,$$

whereupon

$$sinkj\Delta x = -\sin(2k_{max} - k)j\Delta x. \tag{3.65}$$

It follows that, knowing only grid point values, we cannot distinguish the wavenumbers k from the wavenumbers $2k_{max} - k$. Further, if $k > k_{max}$,the wavenumber k will be misrepresented as the wavenumber

$$k^* = 2k_{max} - k. \tag{3.66}$$

Therefore, as shown in Fig. 3.14, the resulting wave has wavenumber k^* which is less that k_{max} by an amount equal to that by which k is greater than k_{max}. We can think of the wavenumber k^* as being an image obtained

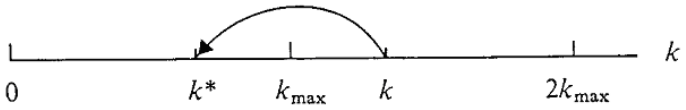

Figure 3.15: Misrepresentation of a wavenumber $k > k_{max}$ in accordance with (3.66).

by the reflection of k across the value k_{max} into the admissible range of wavenumber.

In the example considered above ($\lambda = \frac{4}{3}\Delta x$), $k = \frac{3\pi}{2}(\Delta x)^{-1}$ and $k^* = \frac{\pi}{2}(\Delta x)^{-1}$, the wavenumber "seen" by the finite difference grid. The latter wave has wavelength $4\Delta x$.

The consequence of aliasing error in a numerical integration of the equations of atmospheric motion is a "cascade" of energy into scales that cannot be resolved. After a time, this energy can be expected to grow beyond physically acceptable limits. Indeed, it is found that if no measures are taken to prevent or control this build up, the calculation suffers a catastrophic breakdown. The phenomenon is called *nonlinear instability* and was first discovered by Norman Phillips in 1956 in an early attempt to model the atmospheric general circulation.

Phillips started from an atmosphere at rest and sought to integrate the vorticity equation for a model time of 30 days. The calculation came to a catastrophic end before this time, a fact that Phillips attributed initially to excessive truncation error. Therefore he repeated the experiment with reduced space and time steps. However, the breakdown still occurred at about the same time. Later he repeated the experiment, but after every two hours of simulated time he carried out a harmonic analysis of the vorticity fields, eliminating all components with $k > \frac{1}{2}k_{max}$[1]. This solved the instability problem.

[1]Orszag (1971) showed that, in fact, it is sufficient to eliminate only the top one-third of wavenumbers, because if waves with $k > \frac{2}{3}k_{max}$ are filtered out all aliases satisfy $k > \frac{2}{3}k_{max}$ and will be removed.

3.10 Ways to prevent nonlinear instability

If we wish to carry out an integration for an extended period of time it is necessary to suppress nonlinear instability. Possible methods are:

I) To use a differencing scheme that has a built-in damping of the shortest waves, e.g. the Lax-Wendroff scheme. Indeed, Kasahara (1969) showed that it is sufficient to apply the scheme intermittently at relatively long time intervals.

II) To add a dissipative term to a scheme that is not dissipative, enabling the amount of dissipation to be better controlled.

III) To use a Lagrangian formulation of the advection terms instead of an Eulerian formulation. More about such schemes will be presented later.

IV) To use schemes that are free of the spurious inflow of energy into the shorter waves, instead of suppressing the amplitude of these waves. Such schemes were developed by Arakawa (1966, 1972).

3.11 Arakawa's conservative scheme

Consider a homogeneous, incompressible, inviscid, two-dimensional flow. The vorticity (ζ)-streamfunction (ψ) equations are

$$\frac{\partial \zeta}{\partial t} = -\mathbf{u} \cdot \nabla \zeta = -\nabla \cdot (\mathbf{u}\zeta), \tag{3.67}$$

$$\mathbf{u} = \mathbf{k} \wedge \nabla \psi, \tag{3.68}$$

$$\zeta = \nabla^2 \psi, \tag{3.69}$$

in standard notation. Suppose that the flow is contained in a closed domain D so that there is no flow across the domain boundary S. Then

$$\frac{\partial}{\partial t} \int_D \zeta dV = -\int_D \nabla \cdot (\mathbf{u}\zeta) dV = -\int_D \zeta \mathbf{u} \cdot \hat{\mathbf{n}} dS = 0,$$

where $\hat{\mathbf{n}}$ is a unit vector normal to S and pointing from D. It follows that the total (or mean) vorticity is conserved.

Multiplying (3.67) by ζ gives

$$\frac{\partial}{\partial t}\left(\tfrac{1}{2}\zeta^2\right) = -\tfrac{1}{2}\nabla\cdot\left(\zeta^2\mathbf{u}\right),$$

and application of the divergence theorem to this shows that the total or mean square vorticity is conserved also. Vorticity squared is called *enstrophy*.

Finally, multiplying (3.67) by ψ and using (3.69) gives

$$\psi\frac{\partial}{\partial t}\left(\tfrac{1}{2}\zeta^2\right) = -\tfrac{1}{2}\nabla\cdot\left(\zeta^2\mathbf{u}\right). \tag{3.70}$$

The left-hand-side of this can be written

$$\psi\frac{\partial}{\partial t}\left(\nabla\cdot\nabla\psi\right)\psi\nabla\cdot\nabla\left(\frac{\partial\psi}{\partial t}\right) = \nabla\cdot\left(\psi\nabla\frac{\partial\psi}{\partial t}\right) - \nabla\psi\cdot\nabla\left(\frac{\partial\psi}{\partial t}\right)$$

and the right-hand-side can be rearranged as

$$\psi\nabla\cdot(\zeta\mathbf{u}) = \nabla\cdot(\psi\zeta\mathbf{u}) - \zeta\mathbf{u}\cdot\nabla\psi.$$

Substitution of these into (3.70) gives

$$\nabla\psi\cdot\nabla\left(\frac{\partial\psi}{\partial t}\right) = \nabla\cdot\left(\psi\nabla\frac{\partial\psi}{\partial t}\right) + \nabla\cdot(\psi\zeta\mathbf{u}) - \zeta\mathbf{u}\cdot\nabla\psi.$$

The last term vanishes because of (3.68), while the first term is

$$\frac{\partial}{\partial t}\left[\frac{1}{2}\left(\nabla\psi\right)^2\right],$$

just the rate-of-change of kinetic energy. Again with the use of Gauss's theorem, it follows that

$$\frac{\partial}{\partial t}\int_D \tfrac{1}{2}u^2\,dV = 0,$$

i.e. the mean kinetic energy is conserved in D.

In summary, for any closed region, the model expressed by (3.67) - (3.69) conserves mean *kinetic energy, mean vorticity, and mean enstrophy*. Obviously it would be desirable and perhaps even necessary to retain these properties when the differential equation is replaced by a difference equation.

Suppose that D is a rectangular region of dimensions L_x and L_y over which the streamfunction can be represented as a double Fourier series. The implication is that the streamfunction is periodic in the x-direction with

period L_x and is zero along the upper and lower boundary, but kinetic energy, vorticity and enstrophy are still conserved.

$$\psi = \sum_{m=0}^{\infty}\sum_{n=1}^{\infty}\left[a_{m,n}(t)\cos\frac{2\pi mx}{L_x} + b_{m,n}(t)\sin\frac{2\pi mx}{L_x}\right]\sin\frac{n\pi y}{L_y}. \tag{3.71}$$

If we calculate the product $-\mathbf{u}\times\nabla\zeta$ in (3.67) and use standard trigonometric identities, it is found that if a term in u with wavenumbers m_1, n_1 is multiplied by a term in $\nabla\zeta$ with wavenumbers m_2 ,n_2, the resulting term involves the four pairs of wavenumbers

$$\begin{aligned} &m_1+m_2, \quad n_1+n_2 \qquad m_1-m_2, \quad n_1+n_2 \\ &m_1+m_2, \quad n_1-n_2 \qquad m_1-m_2, \quad n_1-n_2 \end{aligned} \tag{3.72}$$

It follows that there is a transfer of vorticity and energy between different wavenumbers associated with this term. Again, *nonlinear interaction* results in the exchange of energy within the total spectrum and may develop wavenumbers which previously did not exist.

Now the individual terms of (3.72) are orthogonal eigenfunctions of the Helmholtz equation

$$\nabla^2\psi_{m,n} + m_{m,n}^2\psi_{m,n} = 0, \tag{3.73}$$

where the eigenvalues are given by

$$\mu_{m,n}^2 = (2\mu\pi/L_x)^2 + (n\pi/L_y)^2. \tag{3.74}$$

Because the eigenfunctions $\psi_{m,n}$ are orthogonal, the mean kinetic energy and mean enstrophy are

$$\overline{K} = \tfrac{1}{2}\overline{\sum\sum(\nabla\psi_{m,n})^2} = \overline{\sum\sum K_{m,n}} = \tfrac{1}{2}\overline{\sum\sum\mu_{m,n}^2\psi_{m,n}^2} \tag{3.75}$$

$$\overline{\zeta^2} = \overline{\sum\sum(\nabla^2\psi_{m,n})^2}2\overline{\sum\sum\mu_{m,n}^2K_{m,n}} = \overline{\sum\sum\mu_{m,n}^4\psi_{m,n}^2}, \tag{3.76}$$

where the bar indicates a domain average[2]. Since $\overline{K}$ and $\overline{\zeta^2}$ are conserved, so also is their ratio $\overline{\zeta^2}/\overline{K} = 2\overline{\mu}^2$, where $\overline{\mu}$ is an average wavenumber. In

[2]The penultimate expression in (3.76) follows from the vector identity $\nabla\psi\cdot\nabla\phi = \nabla\cdot(\psi\nabla\phi) - \psi\nabla^2\phi$.

other words, *a systematic energy cascade towards higher wavenumbers is not possible* in this flow. Furthermore, to obtain the enstrophy the contributions $K_{m,n}$ are multiplied by the wavenumber squared. Therefore the *fraction of energy that can flow into high wavenumbers is clearly limited*, and the higher the wavenumber, the more it is limited. These results were obtained by Fjørtoft (1953).

Charney (1966) showed that the situation can be illustrated by a simple mechanical analogy. Equations (3.75)-(3.76) give

$$\overline{K}\overline{\mu}^2 = \overline{\sum\sum \mu_{m,n}^2 K_{m,n}}.$$

As shown in Fig. (3.16), we can imagine a semi-infinite weightless rod on which a weight $\overline{K}$ is suspended at a distance $\lambda = \overline{\mu}^2$ to the left of the point at which the rod is suspended, and weights $K_1 = K_{1,1}$, $K_2 = K_{1,2}$, $K_3 K_{1,3}$, are suspended at distances $\lambda_1 = \mu_{1,1}$, $\lambda_2 = \mu_{1,2}^2$, $\lambda_3 = \mu_{2,1}$, to the right of that point. The rod, as defined, would be in mechanical equilibrium. Its left-hand-side cannot change, while on the right-hand-side an interchange of mass between the weights is permitted, but only so as not to disturb the equilibrium, i.e., the total moment of the forces. Since the value of μ for a particular K is fixed, it follows that at least three components must always take part in an energy transfer (see Ex. 3.3). In particular, very little energy can be expected to accumulate in the highest wavenumbers through a cascade of energy from the lowest wavenumbers.

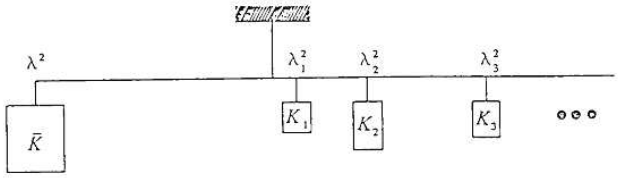

Figure 3.16: A mechanical analogy of the interchange of energy between harmonic components.

Consider now the numerical solution of Eqs. (3.67) - (3.69). These can be written as a single equation

$$\frac{\partial}{\partial t}(\nabla^2\psi) = -J(\psi, \nabla^2\psi), \tag{3.77}$$

using the Jacobian notation, $J(f,g) = (\partial f/\partial x)(g/y) - (\partial g/\partial x)(\partial f/\partial y)$. If a finite-difference scheme could be constructed so as to conserve the average

values of the kinetic energy and enstrophy, a systematic transfer of energy towards the highest wavenumbers could not occur and the nonlinear instability found by Phillips would be avoided. The method for doing this is due Arakawa op. cit.. The idea is to find a form of the Jacobian term in (3.77) that has the appropriate conservation properties.

Note that for any two scalar functions f and g,

$$J(f,g) = \mathbf{k}\cdot\nabla\wedge(f\nabla g) = -\mathbf{k}\cdot\nabla\wedge(g\wedge f).$$

Using Stoke' s circulation theorem it follows that the domain average of J is zero, i.e.,

$$\overline{J(f,g)} = 0, \tag{3.78}$$

if either f or g is a constant along the boundary of the domain. Moreover, under the same conditions,

$$\overline{fJ(f,g) = 0}\,,\ \ \overline{gJ(f,g) = 0}. \tag{3.79}$$

Now there are three alternative ways to write the Jacobian in (3.77), i.e.,

$$\begin{aligned} j(\psi,\zeta) &= \frac{\partial\psi}{\partial x}\frac{\partial\zeta}{\partial y} - \frac{\partial\psi}{\partial y}\frac{\partial\zeta}{\partial x} && (3.80)\\ &= \frac{\partial}{\partial x}\left(\psi\frac{\partial\zeta}{\partial y}\right) - \frac{\partial}{\partial y}\left(\psi\frac{\partial\zeta}{\partial x}\right) && (3.81)\\ &= \frac{\partial}{\partial y}\left(\zeta\frac{\partial\psi}{\partial x}\right) - \frac{\partial}{\partial x}\left(\zeta\frac{\partial\psi}{\partial y}\right). && (3.82) \end{aligned}$$

Based on these, we consider possible ways to construct second-order accurate finite-difference approximations to the Jacobian. With the simplest centred space differencing, we require values of ψ and ζ from a box of nine adjacent grid points to evaluate (3.80)-(3.82) - see Fig. 3.17. Let d be the grid size and ψ_n, ζ_n the values of ψ and ψ at the point n.

Then the corresponding approximations to (3.80-3.82) are

$$\begin{aligned} J_0^{++} &= [(\psi_1-\psi_3)(\zeta_2-\zeta_4) - (\psi_2-\psi_4)(\zeta_1-\zeta_3)]/4d^2 && (3.83)\\ J_0^{+x} &= [\psi_1(\zeta_5-\zeta_8) - \psi_3(\zeta_6-\zeta_7) - \psi_2(\zeta_5-\zeta_6) + \psi_4(\zeta_8-\zeta_7)]/4d^2 && (3.84)\\ J_0^{x+} &= = [\zeta_2(\psi_5-y6) - z4(\psi_8-\psi_7) - \zeta_1(\psi_5-\psi_8) + \zeta_3(\psi_6-\psi_7)]/4d^2 && (3.85) \end{aligned}$$

The superscripts + and x denote the position of the points from which values of ψ and ζ, respectively, are used to form the approximations. A more general approximation to the Jacobian is

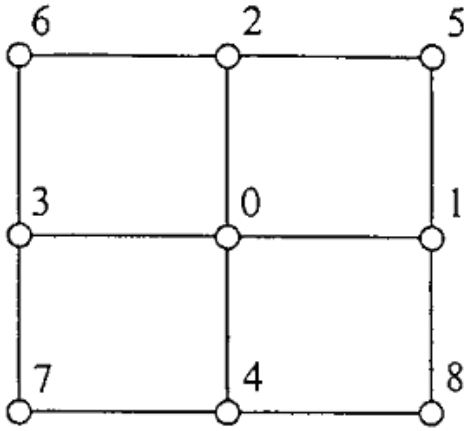

Figure 3.17: Stencil used for the finite-difference approximation to the Jacobian in (3.80)-(3.82).

$$J(\psi,\zeta) = \alpha J^{++} + \beta J^{x+} + \gamma J^{+x}, \tag{3.86}$$

with the consistency requirement that $\alpha+\beta+\gamma = 1$. We shall seek to choose α, β and γ to ensure conservation of mean kinetic energy and enstrophy. Consider the gain in enstrophy at point zero in fig. 3.17 due to the value at point 1 and vice versa. The results, excluding the factor $4d^2$, are (3.87)

$$\begin{aligned} \zeta_0 J_0^{++} \sim \; & -\zeta_0\zeta_1(\psi_2-\psi_4) \; + \; \text{other terms} \\ \zeta_1 J_1^{++} \sim \; & \;\; \zeta_1\zeta_0(\psi_5-\psi_8) \; + \; \text{other terms} \end{aligned} \tag{3.87}$$

Since these terms do not cancel and no other opportunity for the product $\zeta_0\zeta_1$ occurs over the grid, the sum $\overline{\zeta J^{++}}$ cannot vanish in general, and $\overline{\zeta^2}$ will not be conserved (i.e., $\overline{\partial\zeta^2}/\partial t \neq= 0$). Similarly

$$\begin{aligned} \zeta_0 J_0^{+x} &\sim \zeta_0\zeta_5(\psi_1-\psi_2)+\ldots \\ \zeta_5 J_5^{+x} &\sim \zeta_5\zeta_0(\psi_y 2-\psi_1)+\ldots \\ \zeta_0 J_0^{x+} &\sim -\zeta_0\zeta_1(\psi_5-\psi_8)+\ldots \\ \zeta_1 J_1^{x+} &\sim \zeta_1\zeta_0(\psi_2-\psi_4)+\ldots \end{aligned} \tag{3.88}$$

From (3.87) and (3.88) it is clear that both J^{+x} and $\frac{1}{2}(J^{++}+J^{x+})$ *conserve* $\overline{\zeta^2}$. A similar examination of the finite-difference approximation of $\psi J(\psi,z)$ shows that

$$J^{x+} \quad \text{and} \quad J^{+x}\tfrac{1}{2}(J^{++} + J^{+x}) \quad \text{conserve} \quad \overline{K} = \tfrac{1}{2}\mathbf{u}^2.$$

Combining the two results it follows that the

$$\tfrac{1}{3}(J^{++}J^{+x} + J^{x+}) \quad \text{conserves both} \quad \overline{\zeta^2} \quad \text{and} \quad \overline{K}. \tag{3.89}$$

Consequently it conserves also the *mean wavenumber*. These conservation properties prevent nonlinear instability. *Aliasing is still present* in the form of phase errors, but the latter results also from linear finite-difference truncation errors. The expression (3.89) is called the *Arakawa-Jacobian*.

Arakawa (1966) showed also how to construct an approximation of fourth-order accuracy to the Jacobian with the same conservation properties.

3.12 Conservative schemes for the primitive equations

Consider the nonlinear advection equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \tag{3.90}$$

and its corresponding kinetic energy equation

$$\frac{\partial}{\partial t}\left(\tfrac{1}{2}u^2\right) + \frac{\partial}{\partial x}\left(\tfrac{1}{3}u^3\right) = 0. \tag{3.91}$$

Integration of (3.91) with respect to x gives

$$\int_0^L \frac{\partial}{\partial t}\left(\tfrac{1}{2}u^2\right)\, dx = -\tfrac{1}{3}\left(u_L^3 - u_0^3\right). \tag{3.92}$$

If $u_0 = u_L$, including the case of zero flux at the boundaries when both are zero, then

$$\frac{\partial}{\partial t}\int_0^L \tfrac{1}{2}\,u^2 dx = 0, \text{ i.e. the 'kinetic energy' is conserved.}$$

If the interval $[0, L]$ is divided into n equal segments with end points at $x_j = jd$ where $d = L/n \quad (j = 0, 1, \ldots n)$, then the right-hand side of (3.92) can be written as the sum

$$-\tfrac{1}{3}[(u_1^3 - u_0^3) + (u_2^3 - u_1^3) + \ldots + (u_j^3 - u_{j-1}^3) + (u_{j+1}^3 - u_j^3) + \ldots + (u_n^3 - u_{n-1}^3)]$$

Note that when the variables are discretized, the integral becomes a sum over the domain. Moreover, for conservation to occur, the successive terms in the sum must be of the form $(A_{j+1} - A_j)$, so that all intermediate terms cancel as in the above expression. Suppose one were to use centred space differencing to approximate (3.90); i.e.,

$$\frac{\partial u_j}{\partial t} + u_j \frac{u_{j+1} - u_{j-1}}{2d} = 0. \tag{3.93}$$

Multiplying this by u_j and forming the kinetic energy integral as above gives

$$\frac{\partial}{\partial t} \int_0^L \tfrac{1}{2}\, u^2 dx = -\tfrac{1}{2} \sum_{j=0}^{n} (u_j^2 u_{j+1} - u_j^2 u_{j-1}). \tag{3.94}$$

It is clear that the terms in the summation are not of the form $(A_{j+1} - A_j)$, whereupon the scheme does not conserve kinetic energy. This is despite the fact that it is a consistent and linearly stable scheme for Δt sufficiently small. Another possibility would be to use the flux form of the equation, i.e.,

$$\frac{\partial u_j}{\partial t} = -\frac{\partial}{\partial x}(\tfrac{1}{2}u_j^2) = -(u_{j+1}^2 - u_{j-1}^2)/4d.$$

This would give for the right-hand side of (3.94),

$$-\tfrac{1}{4} \sum_{j=0}^{n} (u_{j+1}^2 u_j - u_j u_{j-1}^2),$$

in which again, the terms are not of the form $(A_{j+1} - A_j)$. Hence this does not lead to an energy conserving scheme either. A further possibility would be to use the difference form

$$\frac{\partial u_j}{\partial t} = -\frac{1}{6d}(u_{j+1} + u_j + u_{j-1})(u_{j+1} - u_{j-1}).$$

Multiplying this by u_j and summing leads to the right-hand side for (3.94):

$$-\frac{1}{6} \sum_{j=0}^{n} \left[(u_j^2 u_{j+1} + u_j u_{j+1}^2) - (u_{j-1}^2 u_j + u_{j-1} u_j^2)\right],$$

in which the terms are of the form $(A_j - A_{j-1})$. Accordingly the scheme is energy conserving for interior points. Although this finite-difference form is too simple for more complex systems of equations, it does suggest the use of some kind of averaging technique to achieve the conservation property.

Exercises

3.1. Show that by using a quadratic interpolation formula for tracing back the characteristic to the time level $t = n\Delta t$ in Fig. (3.9), one obtains the Lax-Wendroff scheme.

3.2. Derive the fourth-order accurate approximation

$$\frac{\partial u}{\partial x} = \frac{4}{3}\frac{u_{j+1} - u_{j-1}}{2\Delta x} - \frac{1}{3}\frac{u_{j+2} - u_{j-2}}{4\Delta x} + O(\Delta x^4),$$

on page 68 using a Taylor series.

3.3. Verify the statement on page 77 concerning Fig. 3.15 that "at least three components must always take part in an energy transfer".

Chapter 4

The gravity and inertia-gravity wave equations

Like advection, wave propagation is a particularly important process contained in the equations for atmospheric motion. Accordingly, we shall be concerned with investigating numerical methods for treating the terms in the equations that represent wave propagation. We begin with the simplest linearized system governing the horizontal propagation of gravity- or inertia-gravity waves.

4.1 One dimensional gravity waves

For perturbations to a layer of homogenous fluid of uniform undisturbed depth H, there are two equations:

$$\frac{\partial u}{\partial t} = -g\frac{\partial h}{\partial x}, \tag{4.1}$$

$$\frac{\partial h}{\partial t} = -H\frac{\partial u}{\partial x}, \tag{4.2}$$

where $u(x,t)$ and $h(x,t)$ are the perturbation horizontal velocity and fluid depth, respectively. Equation (4.1) is the linearized form of the momentum equation while (4.1b) is the depth-integrated form of the continuity equation (see e.g. DM, Chapter 11). We look for solutions of (4.1)-(4.2) of the form

$$[u(x,t), h(x,t)] = Re[(\hat{u}, \hat{h})e^{i(kx-\omega t)}] \tag{4.3}$$

where $\hat{u}$ and $\hat{h}$ are constants that satisfy the homogeneous algebraic equations

$$\omega\hat{u} = gk\hat{h}, \quad \omega\hat{h} = Hk\hat{u}.$$

These equations can be combined to give the dispersion relation

$$\omega^2 = gHk^2. \tag{4.4}$$

The phase speed of waves is therefore

$$c = \frac{\partial h}{\partial t} = -H\frac{\partial u}{\partial x}. \tag{4.5}$$

It follows that gravity waves can propagate along the x-axis in either direction with speed $\pm\sqrt{gH}$. In particular, they are nondispersive.

When the space derivatives are approximated by central differences, the differential-difference equations corresponding with Eqs. (4.1) - (4.2) are

$$\frac{\partial u_j}{\partial t} = -g\frac{h_{j+1} - h_{j-1}}{2\Delta x}, \tag{4.6}$$

$$\frac{\partial h_j}{\partial t} = -H\frac{u_{j+1} - u_{j-1}}{2\Delta x}. \tag{4.7}$$

The solution (4.3) then takes the form

$$(u_j, h_j) = \mathrm{Re}[(\hat{u},\, \hat{h})\, e^{i(kj\Delta x - \omega t)}], \tag{4.8}$$

and substitution of these into Eqs. (4.6) - (4.7) gives

$$\omega\hat{u} = g\frac{\sin\, k\Delta x}{\Delta x}\hat{h}\ , \quad \omega\hat{h}H\frac{\sin\, k\Delta x}{\Delta x}\hat{u}.$$

These combine to give the dispersion relation

$$\omega^2 = gH\left(\frac{\sin\, k\Delta x}{\Delta x}\right)^2. \tag{4.9}$$

Thus, with the finite-difference approximation to the space derivatives, the gravity waves propagate with the phase speed

$$c^* = \pm\sqrt{gH}\left(\frac{\sin\, k\Delta x}{k\Delta x}\right) \tag{4.10}$$

$$= c\frac{\sin\, k\Delta x}{k\Delta x}. \tag{4.11}$$

Again, *space differencing leads to computational dispersion.* The formula (4.11) is the same as that obtained when considering the advection equation (see e.g. 3.39). It follows that both the phase speed and group velocity depend on the wavenumber as shown in Fig. 3.4. The phase speed decreases

as the wavelength decreases, and the wave with wavelength $2\Delta x$ is stationary. Nevertheless, *there is an important difference between this problem and the advection problem in that there are now two dependent variables.*

We have assumed that u and h are carried at every grid point as shown in Fig. 4.1a. However, as far as the system (4.6) - (4.7) is concerned, u_j depends only on h at adjacent points and h_j depends only on u at adjacent points. Thus the grid contains two elementary 'subgrids', with the solution on one of these subgrids being completely decoupled from the other. It would appear to be sufficient to calculate only one of these solutions by using the grid shown in Fig. 4.1b. Such a grid, with variables carried at alternate points in space, is called a staggered grid. The computation time needed to solve (4.6) - (4.7) on this grid is halved, whereas the truncation error is the same. Moreover, the waves with $k\Delta x > \frac{1}{2}\pi$ have been eliminated and these are just the waves with large phase speed errors and negative group velocities. This amounts to a significant improvement in the properties of the scheme, because, when using the staggered grid, the phase and group velocity diagram in Fig. 3.4 is reduced to its left half, covering waves with wavelengths no shorter than $4\Delta x$.

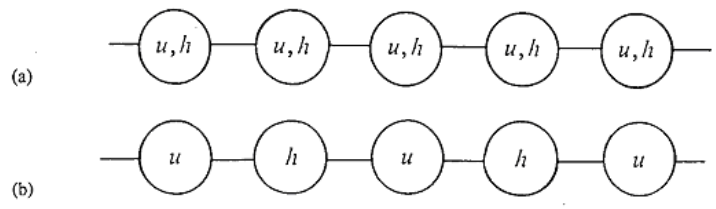

Figure 4.1: A grid with two dependent variables u and h. In (a) these are stored at every grid point; in (b) at alternate grid points.

If we wish to include waves with wavelengths between 4Dx and 2Dx in our calculation we can halve the grid length and will still require no more computation time than if we had used a nonstaggered grid.

4.2 Two-dimensional gravity waves

In two-dimensions, the linearized shallow-water equations are

$$\frac{\partial u}{\partial t} = -g\frac{\partial h}{\partial x}, \quad \frac{\partial v}{\partial t} = -g\frac{\partial h}{\partial y},$$

$$\frac{\partial h}{\partial t} = -H\nabla \cdot \mathbf{u}. \tag{4.12}$$

Substituting the wave solution

$$(u, v, h) = Re[(\hat{u}, \hat{v}, \hat{h})^{i(kx+ly-\omega x)}], \tag{4.13}$$

we obtain the dispersion relation

$$\omega^2 = gH(k^2 + l^2). \tag{4.14}$$

The phase speed, $c = w/|\mathbf{k}|$, where $\mathbf{k} = (k, l)$ is the wavenumber vector. It follows again that $c = \sqrt{gH}$.

The construction of a suitable grid, either non-staggered or staggered, presents a number of possibilities. Three of these, denoted (A), (E) and (C) are shown in Fig. 4.2.

Let $d^\star$ denote the shortest distance between the grid points. With the same value of $d^\star$, the lattice (E) will have half the number of variables per unit area than the lattice (A), while lattice (C) will have only a quarter of the number. The lattice (E) can be obtained by a superposition of two (C) lattices and the lattice (A) by a superposition of two (E) lattices, or of four (C) lattices.

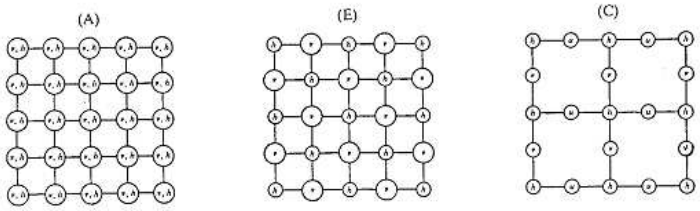

Figure 4.2: Three types of lattice for the finite-difference solution of 4.12.

The admissible range of wavenumbers in the wavenumber plane can be determined by considering the shortest resolvable wavelengths. In lattice (E), the lines joining the nearest points with the same variable makes an angle of $\pi/4$ with the grid lines while with the other two lattices these lines are along the grid lines. Fig. 4.3 shows the admissible range of wavenumbers. A halving of the number of variables is accompanied by a halving of the admissible region of the wavenumber plane.

The same standard finite-difference approximations can be used for the space derivatives in (4.12) for all three lattices. We define the difference operator

$$\delta_x h = [h(x + d^\star, y) - h(x - d^\star, y)]/2d^\star,$$

and similarly for $\delta_y h$. Then (4.12) can be approximated by

$$\frac{\partial h}{\partial t} = -H(\delta_x u + \delta_y v). \tag{4.15}$$

Substituting the appropriate wave solutions (analogous to 4.3) we obtain the dispersion relation

$$\omega^2 = gH\frac{\sin^2 kd^* + \sin^2 ld^*}{d^{*2}}. \tag{4.16}$$

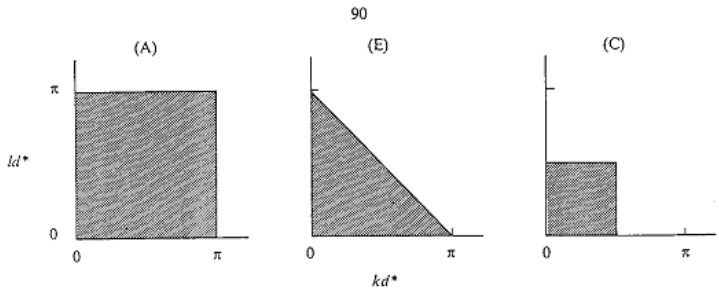

Figure 4.3: Admissible regions of wavenumbers for the three types of lattice shown in Fig. 4.2.

If we define $X = d^*k$, $Y = ld^*$, the ratio of the phase speed given by (4.16), c^*, to the true phase speed $\sqrt{gH}$, can be written as

$$\frac{c^*}{\sqrt{gH}} = \sqrt{\frac{\sin^2 X + \sin^2 Y}{X^2 + Y^2}}. \tag{4.17}$$

This formula reduces to (4.10) when applied to the one-dimensional case.

The values of the relative phase speed (4.17) on the wavenumber range admitted by lattice (E) are shown in Fig. 4.4. By symmetry about the line $l = k$, only half of the region needs to be shown. Fig. 4.3 shows that lattice (C) admits only the left half of the triangular region shown in Fig. 4.4. Clearly lattice (C) gives a more accurate phase speed for gravity waves than the other two lattices. Unfortunately, because it does not carry the two velocity components at the same points, there is some difficulty if the Coriolis terms have to be included. Of the other lattices, the staggered lattice (E) is

much superior to the non-staggered lattice (A). A result with the same truncation error can be achieved in about half of the computation time, (exactly half if the equations are linear), and a significant fraction of wavenumbers that are associated with large phase speed errors and computational dispersion are eliminated. The additional time needed for a calculation on an (A) lattice is spent on waves that can hardly be expected to improve the integration.

Despite its superiority over lattice (A), the lattice (E) is not free of computational problems. As with the non-staggered one-dimensional grid discussed in section 4.1, the solution of (4.15) on each of the two type (C) subgrids forming the (E) lattice are independent and can diverge from each other. This can lead to serious problems. For example, if the values of the dependent variables on one of these (C) lattices are constants, they will constitute a stationary solution on that lattice, no matter what the values of the variables on the other (C) lattices are. Two stationary solutions, with different constant values on each of these complementary lattices, will give a stationary wave represented by the right-hand corner of the triangular region in Fig. 4.4, with a zero phase speed. This wave is usually referred to as *the two-grid-interval* wave. In the same way, the (A) lattice admits four independent stationary solutions, with different constant values on each of its four type (C) subgrids.

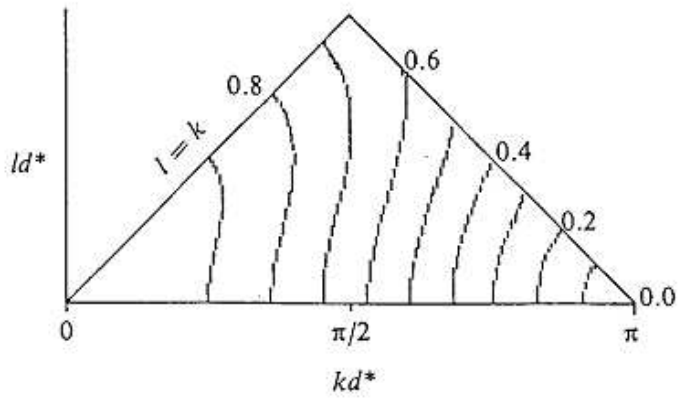

Figure 4.4: Relative phase speed of gravity waves when the space derivatives in (4.12) are approximated by space-centred finite-difference analogues.

The two-grid-interval wave can easily be generated when boundary conditions are artificially prescribed, and, with more complete equations, in cases when gravity waves are generated inside the computational region. These

can be caused by heating, for example through the release of latent heat, and also by the influence of mountains. When gravity waves are excited involving variables of one of the (C) subgrids only, for example by forcing at individual grid points or lines of points, the gravity wave will propagate through the variables (C) of this subgrid only. The variables of the other (C) subgrid will be influenced only through the Coriolis and advection terms on a much larger time scale. Thus physical effects which may excite relatively long waves in the atmosphere may excite spurious waves with wavelengths of approximately two grid intervals in a computation. When these reach an excessive amplitude, some remedial measures have to be taken - see later.

4.3 Inertia-gravity waves

We consider now the effect of centred space differencing on inertia-gravity waves. The linearized shallow-water equations with rotation present are

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\frac{\partial h}{\partial x} + fv, \quad \frac{\partial v}{\partial t} = -g\frac{\partial h}{\partial y} - fu, \\ \frac{\partial h}{\partial t} &= -H\nabla \cdot \mathbf{u}, . \end{aligned} \tag{4.18}$$

These differ from Eqs. (4.12) through the appearance of the two Coriolis terms. Since these do not involve derivatives, they are difficult to calculate on the (C) lattice, which was ideal for pure gravity waves. For this reason we need to reconsider the problem of the distribution of the variables.

Equations (4.18) admit two distinct types of motion: low-frequency, quasi-geostrophic and quasi-nondivergent flow; and high-frequency inertia-gravity waves (see e.g. DM, Chapter 11). Inertia-gravity waves are continually excited in the atmosphere, but as they are dispersive, a local accumulation of wave energy disperses with time. This process is known as geostrophic adjustment. The remaining motion is in approximate geostrophic balance and changes only slowly with time. Here we are concerned with the correct representation of this process, which is governed essentially by the inertia-gravity wave equations.

We are interested both in waves caused by physical effects, and in those caused by inadequacies of the initial data and of the numerical procedures. Accordingly, we shall investigate the effect of the space distribution on the dispersive properties of the inertia-gravity waves. We accomplish this using the simplest centred approximations for the space derivatives, leaving the time derivatives in their differential form. The discussion follows that of

Winnighoff and Arakawa, as presented by (Arakawa, 1972; Arakawa *et al.*, 1974).

We consider five ways of distributing the dependent variables in space, shown in Fig. 4.5. Let d denote the shortest distance between neighbouring points carrying the same dependent variable. In the panel, d is the same for each of the five lattices. Therefore all the lattices have the *same number of dependent variables per unit area* and the computation time required for an integration on each of the lattices will about the same. However, the properties of the solutions obtained will differ because of the effect of the space arrangement of variables.

Using the subscripts shown in the figure, we define the centred space-differencing operator by

$$(\delta_x a)_{i,j} \frac{1}{d'}(a_{i+\frac{1}{2},j}, -a_{i-\frac{1}{2},j}),$$

where d' is the distance separating the points between which the finite-difference is taken. Thus, for lattices (A) to (D), $d' = d$, whereas for lattice (E), $d' = \sqrt{2d}$. We define also an average taken over the two points by

$$(\overline{a}^x)_{i,j} = \tfrac{1}{2}(a_{i+\frac{1}{2},j} - \ i_{i-\frac{1}{2},j}).$$

The analogous quantities $(\delta_y a)_{i,j}$ and $(a^{-y})_{i,j}$ are defined similarly, but with respect to the y axis. Finally

$$(\overline{a}^x)_{i,j} = \overline{(a^x)}^y_{i,j}.$$

For each of the five lattices we use the simplest centred approximations for the space derivatives and Coriolis terms in (4.18) and obtain the five equation systems:

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\overline{\delta_x h}^x + fv, \quad \frac{\partial v}{\partial t} = -g\overline{\delta_y h}^y - fu, \\ \frac{\partial h}{\partial t} &= -H\left(\overline{\delta_x u}^x + \overline{\delta_y v}^y\right), \end{aligned} \tag{4.19}$$

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\overline{\delta_x h}^y + fv, \quad \frac{\partial v}{\partial t} = -g\overline{\delta_y h}^x - fu, \\ \frac{\partial h}{\partial t} &= -H\left(\overline{\delta_x u}^y + \overline{\delta_y v}^x\right), \end{aligned} \tag{4.20}$$

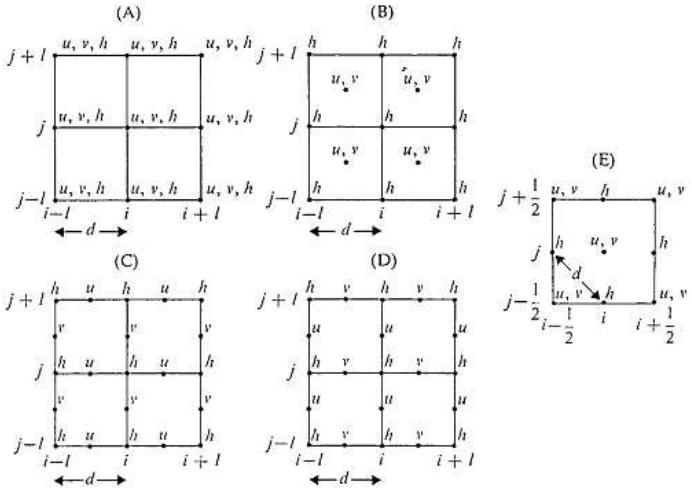

Figure 4.5: Five types of lattice considered for the finite-difference solution of Eqs. (4.1) - (4.2). Note that the grid-spacing in (E) is smaller than that in (A) - (D) by a factor $\sqrt{2}$, because d is the same.

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\delta_x h + f\overline{v}^{xy}, \quad \frac{\partial v}{\partial t} = -g\delta_y h - f\overline{u}^{xy}, \\ \frac{\partial h}{\partial t} &= -H\left(\delta_x u + \delta_y v\right), \end{aligned} \tag{4.21}$$

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\overline{\delta_x h}^{xy} + f\overline{v}^{xy}, \quad \frac{\partial v}{\partial t} = -g\overline{\delta_y h}^{xy} - f\overline{u}^{xy}, \\ \frac{\partial h}{\partial t} &= -H\left(\overline{\delta_x u}^{xy} + \overline{\delta_y v}^{xy}\right), \end{aligned} \tag{4.22}$$

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\delta_x h + fv, \quad \frac{\partial v}{\partial t} = -g\delta_y h - fu, \\ \frac{\partial h}{\partial t} &= -H\left(\delta_x u + \delta_y v\right). \end{aligned} \tag{4.23}$$

We analyze first a one-dimensional case, in which the variables u, v and h do not vary with y. Thus we have

$$u, v, h = u, v, h(x, t).$$

The system (4.18) then reduces to

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\frac{\partial h}{\partial x} + fv, \quad \frac{\partial v}{\partial t} = -fu, \\ \frac{\partial h}{\partial t} &= -H\,\frac{\partial u}{\partial x}. \end{aligned} \tag{4.24}$$

Substituting the wave solutions (4.13) we obtain the frequency equation which can be written as

$$\left(\frac{\omega}{f}\right)^2 = 1 + \frac{gH}{f^2}k^2. \tag{4.25}$$

Thus, as the *radius of deformation*

$$\lambda = \sqrt{gH/f},$$

is never equal to zero, the frequency of the inertia-gravity waves is a monotonically increasing function of k. Therefore, the group velocity $\partial\omega/\partial k$ is never equal to zero. This is very important for the geostrophic adjustment process, as it precludes a local accumulation of wave energy.

We now look at the effect of the finite differencing in space in this case. As the variables are assumed not to depend on y, the systems (4.19) - (4.23) reduce to

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\overline{\delta_x h}^x + fv, \quad \frac{\partial v}{\partial t} = -fu, \\ \frac{\partial h}{\partial t} &= -H\overline{\delta_x u}^x, \end{aligned} \tag{4.26}$$

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\delta_x h + fv, \quad \frac{\partial v}{\partial t} = -fu, \\ \frac{\partial h}{\partial t} &= -H\delta_x u, \end{aligned} \tag{4.27}$$

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\delta_x h + f\overline{v}^x, \quad \frac{\partial v}{\partial t} = -f\overline{u}^x, \\ \frac{\partial}{\partial t} &= -H\delta_x u, \end{aligned} \tag{4.28}$$

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\overline{\delta_x h}^x + f\overline{v}^x, \quad \frac{\partial v}{\partial t} = -f\overline{u}^x, \\ \frac{\partial}{\partial t} &= -H\overline{\delta_x u}^x, \end{aligned} \tag{4.29}$$

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\delta_x h + fv, \quad \frac{\partial v}{\partial t} = -fu, \\ \frac{\partial h}{\partial t} &= -H\delta_x u. \end{aligned} \tag{4.30}$$

Substitution of wave solutions into these systems gives the frequency equations

$$\left(\frac{\omega}{f}\right)^2 = 1 + \left(\frac{\lambda}{d}\right)^2 \sin^2 \ kd, \tag{4.31}$$

$$\left(\frac{\omega}{f}\right)^2 = 1 + 4\left(\frac{\lambda}{d}\right)^2 \sin^2 \frac{kd}{2}, \tag{4.32}$$

$$\left(\frac{\omega}{f}\right)^2 = \cos^2 \frac{kd}{2} + 4\left(\frac{\lambda}{d}\right)^2 \sin^2 \frac{kd}{2}, \tag{4.33}$$

$$\left(\frac{\omega}{f}\right)^2 = \cos^2 \frac{kd}{2} + \left(\frac{\lambda}{d}\right)^2 \sin^2 kd, \tag{4.34}$$

$$\left(\frac{\omega}{f}\right)^2 = 1 + 2\left(\frac{\lambda}{d}\right)^2 \sin^2 \frac{kd}{\sqrt{2}}. \tag{4.35}$$

The non-dimensional frequency ω/f is now seen to depend on two parameters, kd and λ/d.

We shall analyze the dispersion properties revealed by the expressions for each of the five lattices. The wave length of the shortest resolvable wave along the x-axis is $2d$ for lattices (A) through (D), and for the lattice (E). Thus, we have to consider the range $0 < kd \leq \sqrt{2\pi}$ for lattices (A) through (D), and the range $0 < kd \leq \sqrt{2\pi}$ for the lattice (E).

Lattice (A): The frequency reaches a maximum at $kd = \pi/2$. Thus, the group velocity is zero for k equal to $\pi/(2d)$. If inertia-gravity waves of approximately that wave number are excited near a point inside the computational region, for example by nonlinear effects or by forcing through heating or ground topography, the wave energy stays near the point. Beyond this maximum value, for $\pi/2 < kd < \pi$, the frequency decreases as the wave

number increases. Thus, for these waves the group velocity has the wrong sign. Finally, the two-grid-interval wave with $kd = \pi$ behaves like a pure inertia oscillation, and its group velocity is again zero.

Lattice (B): The frequency increases monotonically throughout the range $0 < kd < \pi$. However, it reaches a maximum at the end of the range, so that the group velocity is zero for the two-grid-interval wave with $kd = \pi$.

Lattice (C): The frequency increases monotonically with kd if $\lambda/d > {}^1/_2$ and decreases monotonically with kd if $\lambda/d < {}^1/_2$. It reaches an extreme value again at $kd = \pi$, associated with a zero group velocity. For $\lambda/d = {}^1/_2$ the group velocity is equal to zero for all k.

Lattice (D): The frequency reaches a maximum at $(\lambda/d)2coskd = {}^1/_4$. The two-grid-interval wave at $kd = \pi$ is stationary.

Lattice (E): The frequency reaches a maximum at $kd = \pi/\sqrt{2}$. The shortest resolvable wave with $kd = \sqrt{2}\pi$ behaves like a pure inertia oscillation, and its group velocity is again zero.

A summary of these results is shown in Fig. 3.2. It shows the functions $|\omega|/f$, in the case $\lambda/d = 2$.

The figure vividly illustrates the inadequacy of the lattices (D) and (A). The phase speed and dispersion properties of the remaining three lattices are much better: however, zero group velocities occur with every lattice. Thus, with any lattice there will be difficulties in the geostrophic adjustment process.

The difference between the results for lattices (B) and (E) is interesting because these two lattices can be obtained from one another by a rotation through an angle of $\pi/4$. If we consider the one-dimensional case in which the dependent variables are constant along the lines $y = x + c$, we obtain results for these two lattices that are exactly opposite to those in Fig. 4.6. In general, we define the coordinate system x', y' by rotating the system x, y in the positive direction through an angle of $\pi/4$, and then, using the relations change from variables u, v, h to new dependent variables u', v', h. We find that this transforms the system (4.20) into (4.23). Thus, the dispersion properties of the lattices (B) and (E) can be considered equivalent. A inertia-gravity wave in one of these lattices has phase speed and dispersion properties identical to those of the same wave with its front rotated through the angle of $\pi/4$ in the other lattice.

$$u' = \frac{\sqrt{2}}{2}(u + v), \qquad v' = \frac{\sqrt{2}}{2}(-u + v),$$

Obviously, we should consider also the two-dimensional case. The values of $|\omega|/f$ that are obtained in the two-dimensional case for the true solution and those using lattices (B) and (C) are shown in Fig. 4.7 with $\lambda/d' = 2$.

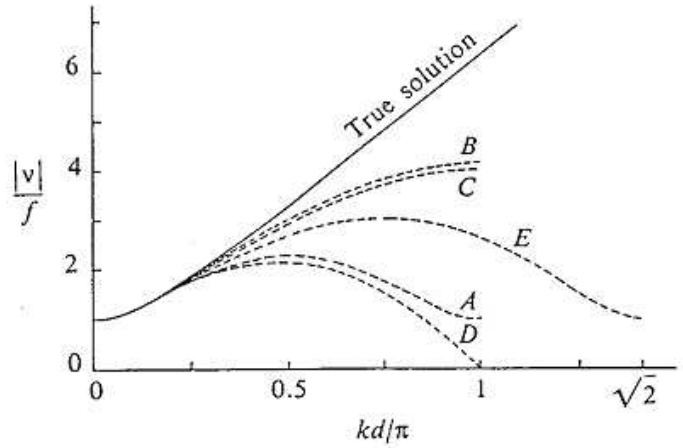

Figure 4.6: The functions $|\omega|/f$ given by (4.25) and (4.31)-(4.35), with $\lambda/d = 2$.

The diagram for lattice (E) can be obtained by a counter-clockwise rotation of the (B) lattice diagram. The diagram for lattice (C) in the two-dimensional case is seen to be a much better approximation to the exact solution than the (B) or (E) lattice diagram. In the (B) lattice diagram the dot-dashed line shows the maximum $|\omega|/f$ for a given ratio l/k; note that there is no such line in the (C) lattice diagram and the exact solution. Such a maximum occurs at only two corner points of the (C) lattice diagram. Thus, with the (C) lattice, no waves have a group velocity with the wrong sign. The situation, though, does depend on the parameter λ/d. Within a stratified atmosphere the radius of deformation λ depends on the stability; if the stability is so weak as to make l/d of the order 1 or less, the (C) lattice loses the advantages shown in Fig. 4.7. However, for typical grid sizes used in atmospheric models this is not the case and therefore Arakawa (Arakawa and Lamb, 1976) concludes that the lattice (C) is the best lattice to simulate the geostrophic adjustment process. Accordingly, it was used in the general circulation model at the University of California at Los Angeles, and also in the British operational model. The (B) or (E) lattices have a problem with false low frequencies of the shortest waves. The two-grid-interval wave, that was stationary as a pure gravity wave, now behaves like a pure inertia oscillation. The difficulty arises from decoupling of the gravity wave solutions on the two complementary (C) type subgrids. Methods of dealing with this will be discussed later.

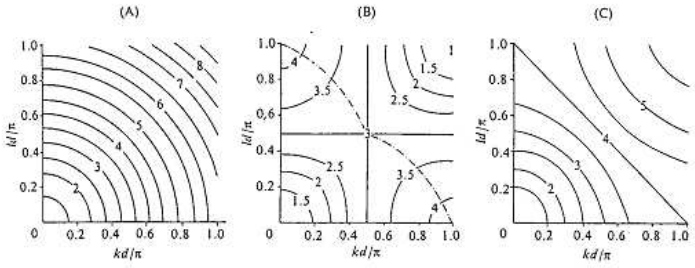

Figure 4.7: The functions $|\omega|/f$, for the true solution and for solutions of systems (4.20) and (4.21), with $\lambda/d = 2$.

4.4 The normal form of the gravity wave equations

We consider the one-dimensional equations

$$\begin{aligned} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} + g\frac{\partial h}{\partial x} &= 0 \\ \frac{\partial h}{\partial t} + c\frac{\partial}{\partial x} + H\frac{\partial u}{\partial x} &= 0. \end{aligned} \tag{4.36}$$

We multiply the second of these equations by an arbitrary parameter λ, and add the result to the first equation. We obtain

$$\frac{\partial}{\partial t}(u + \lambda h) + (c + \lambda H)\frac{\partial u}{\partial x} + (g + \lambda c)\frac{\partial h}{\partial x} = 0. \tag{4.37}$$

We wish to choose λ so that

$$\frac{g + \lambda c}{c + \lambda H} = \lambda \tag{4.38}$$

to obtain an equation with only one dependent variable, $u + \lambda h$. The two solutions are

$$\lambda = \pm\sqrt{\frac{g}{H}}. \tag{4.39}$$

Substituting these into (4.37) we obtain

$$\left[\frac{\partial}{\partial t} + \left(c + \sqrt{gH}\right)\frac{\partial}{\partial x}\right]\left(u + \sqrt{\frac{g}{H}}h\right) = 0$$

$$\left[\frac{\partial}{\partial t}+\left(c-\sqrt{gH}\right)\frac{\partial}{\partial x}\right]\left(u-\sqrt{\frac{g'H}{H}}h\right) \;=\; 0. \tag{4.40}$$

This is the *normal form* of (4.36). It shows that (4.36) is *equivalent to a system of two advection equations*. The quantity $u+\sqrt{g/H'h}$ is seen to be advected at a velocity $c+\sqrt{gH}$ in the direction of the x axis, while, at the same time, the quantity is advected in the same direction at a velocity $c-\sqrt{gH}$. Thus, if the leapfrog scheme is used for the time differencing and we choose a grid that carries both u and h at every grid point, we obtain the Courant-Friedrichs-Levy stability criterion

$$\left(c+\sqrt{gH}\right)\frac{\Delta t}{\Delta x}\leq 1. \tag{4.41}$$

The advection velocity in the atmosphere is normally about an order of magnitude less than the phase speed of external gravity waves, giving the approximate stability requirement

$$\sqrt{gH}\frac{\Delta t}{\Delta x}<1. \tag{4.42}$$

When using the *three-dimensional primitive* equations, external gravity waves are normally eliminated by permitting no vertical velocity at the upper boundary. The highest phase speed admitted by the system is then that of the Lamb waves, which for an isothermal atmosphere is

$$\sqrt{\left(\frac{f}{k}\right)^2+\frac{c_p}{c_v}RT}.$$

If we neglect the first term and recall that the scale height of an isothermal atmosphere is $H^*=RT/g$, we see that the phase speed of the Lamb waves is of the same order of magnitude as that of external gravity waves (300m/s). The CFL stability condition thus means that a large amount of computer time is required because of the presence of these fast moving waves. For this reason implicit time differencing schemes were developed, so that the choice of the time step can be based solely on accuracy and not on stability.

4.5 The leapfrog scheme and the Eliassen grid

We consider the leapfrog scheme with centred space differencing applied to the two-dimensional system of gravity wave equations

$$\begin{aligned} \frac{\partial u}{\partial t} + g\frac{\partial h}{\partial x} &= 0, \quad \frac{\partial v}{\partial t} + g\frac{\partial h}{\partial y} = 0, \\ \frac{\partial h}{\partial t} + H\nabla \cdot v &= 0. \end{aligned} \tag{4.43}$$

In finite-difference notation

$$\begin{aligned} u^{n+1} &= u^{n-1} - 2g\Delta t\delta_x h^n, \quad v^{n+1} = v^{n-1} - 2g\Delta t\delta_y h^n \\ h^{n+1} &= h^{n-1} - 2H\Delta t(\delta_x u + \delta_y v)^n. \end{aligned} \tag{4.44}$$

Substituting the wave solutions

$$\begin{aligned} u^n &= \mathrm{Re}\left[\lambda^n \hat{u}\, e^{i(kx+ly)}\right], \\ v^n &= \mathrm{Re}\left[\lambda^n \hat{v}\, e^{i(kx+ly)}\right], \\ h^n &= \mathrm{Re}\left[\lambda^n \hat{h}\, e^{i(kx+ly)}\right], \end{aligned} \tag{4.45}$$

we obtain the homogeneous system

$$\begin{aligned} (\lambda^2 - 1)\hat{u} + i\lambda 2\sqrt{2}g\mu \sin(X)\hat{h} &= 0 \\ (\lambda^2 1)\hat{v} + i\lambda 2\sqrt{2}g\mu \sin(Y)\hat{h} &= \\ 0i\lambda 2\sqrt{2}H\mu(\sin\,(X)\hat{u} + \sin(Y)\hat{v}) + (\lambda^2 - 1)\hat{h} &= 0 \end{aligned} \tag{4.46}$$

with $X = kd^*$, $Y = ld^*$, $\mu = \Delta t/(d^*)$. Note that $\mu = \Delta t/d$ when lattice (E) is chosen. The requirement that the determinant of (4.46) is equal to zero gives six solutions for λ. Two of these are

$$\lambda = 1 \tag{4.47}$$

and

$$\lambda = -1. \tag{4.48}$$

The remaining four are given by

$$\lambda^2 = 1 - 4A \pm 2\sqrt{2A(2A-1)} \tag{4.49}$$

where

$$A = gH\mu^2(\sin^2 X + \sin^2 Y).$$

We can now analyze the solution (4.45) associated with the values found for λ. Equation (4.47) gives a neutral and stationary solution. If either sin X or sin Y is non-zero in this case, then according to (4.46) we have $\hat{h} = 0$, and the solution represents a physically acceptable translatory motion. However, if $\sin X$ *and* $\sin Y$ are both equal to zero, the amplitudes of all three dependent variables can take arbitrary values. In addition to the physically acceptable solution where all the dependent variables are constant ($k = l = 0$), there is a solution with one or both of the wave numbers k and l equal to π/d^*. This is the two-grid-interval wave. Again it appears as a false computational solution; since it is stationary and it is not affected by the introduction of time differencing.

The second value $\lambda = -1$, represents a false computational mode in time, with a period of 2Dt. This computational mode results from using a three time-level scheme. To prove stability of the scheme the behaviour of the remaining solutions given by (4.49) has to be investigated. They will all be neutral for $2A < 1$. To obtain the condition in the form $B\Delta t \leq 1$ we write

$$\sqrt{2A} \leq 1.$$

Since this has to be satisfied for all the admissible waves, we find that the CFL criterion in the two-dimensional case is now

$$2\sqrt{gH}\mu \leq 1 \tag{4.50}$$

or

$$\sqrt{2gH}\frac{\Delta t}{\Delta x} \leq 1. \tag{4.51}$$

This is in agreement with the previous results (Exercise 4.1). The nondimensional constant on the left is sometimes called the *Courant number*.

With solutions like (4.45), the frequency is given by

$$\lambda = |\lambda| \ e^{-i\nu\Delta t}.$$

Thus, the expressions obtained for λ can be used to calculate the relative phase speed using the relation (Exercise 4.x)

$$\frac{c^*}{\sqrt{gH}} = \frac{1}{\Delta t\sqrt{gH}\sqrt{k^2+l^2}} \ \text{arc} \ \tan\frac{-\lambda_{im}}{\lambda_{re}}. \tag{4.52}$$

For a more explicit illustration of the effect of time differencing, we can perform a series expansion of (4.52). One obtains for $\sqrt{2A} < 1/\sqrt{2}$,

$$\frac{c^*}{\sqrt{gH}} = \sqrt{\frac{\sin^2 X + \sin^2 Y}{X^2 + Y^2}}(1 + \tfrac{1}{3}A + \tfrac{3}{10}A^2 + \ldots.$$

The factor multiplying the series in parenthesis describes the effect of space differencing as given by (4.17). The acceleration resulting from the leapfrog time differencing is beneficial, as it reduces the phase error due to space differencing. The values of the relative phase speed show very little difference compared with Fig. 4.4. The relative phase speed is still poor.

4.6 The Eliassen grid

Eliassen (1956) pointed out the advantages of a space-time grid staggered both in space and time as shown in Fig. 4.8. Such a grid is convenient for the leapfrog scheme associated with centred space-differencing which is suitable for the linearized system:

$$\begin{aligned} \frac{\partial u}{\partial t} + g\frac{\partial h}{\partial x} - fv &= 0, \quad \frac{\partial v}{\partial t} + g\frac{\partial h}{\partial x} + fu0, \\ \frac{\partial h}{\partial t} + H\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) &= 0. \end{aligned} \tag{4.53}$$

The second grid can be obtained by shifting the grid a distance $\sqrt{2}d^*$ along the line $y = x$. This grid is called *Eliassen grid*.

4.7 Economical explicit schemes

The fact that we are now solving *two* equations, the equation of motion and the continuity equation, suggests new ways of constructing time differencing schemes. The time step for gravity waves imposed by the CFL stability criterion is generally much less than that required for an accurate integration of the slower quasi-geostrophic motions. With these short timesteps, the errors due to space differencing are much greater than those due to time differencing. Two explicit schemes that are more economical than the standard leapfrog scheme will be given here. They both achieve economy by using a different integration procedure for the height gradient terms of the equation of motion and for the divergence term of the continuity equation. For brevity, we call these terms the *gravity wave terms* in the governing equations.

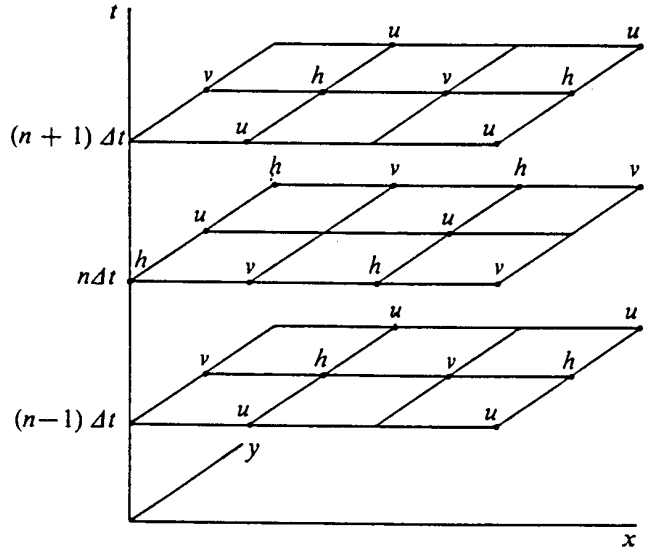

Figure 4.8: Eliassen grid.

4.7.1 The forward-backward scheme

This scheme is obtained by first integrating the gravity wave terms of *either* the equation of motion or of the continuity equation forward, and then those of the other equation backward in time. With centred space differencing (4.43) is approximated by

$$\begin{aligned} u^{n+1} &= u^n - g\Delta t\delta_x h^n, \quad vn+1 = vn - g\Delta t\delta_y h^n, \\ h^{n+1} &= h^n - H\Delta t(\delta_x u + \delta_y v)^{n+1}, \end{aligned} \tag{4.54}$$

or by an analogous system in which the order of integration is reversed. Substituting wave solution as before, we find three solutions for λ. One of these,

$$\lambda = 1, \tag{4.55}$$

gives again a neutral and stationary solution. The remaining two are

$$\lambda = 1 - A \pm \sqrt{A(A-2)} \tag{4.56}$$

where the quantity A is defined as in the preceding section. Solutions (4.55) and (4.56) are obtained for both versions of the scheme (Exercise .4.x.),

that is, no matter which of the two equations - the equation of motion or the continuity equation - is first integrated forward. Examination of the amplification factors given by (4.56) shows that the scheme is stable and neutral for $A \leq 2$, that is, for

$$\sqrt{A} \leq 2.$$

To satisfy this for all admissible waves, we must have

$$2\sqrt{gH}\mu \leq 2. \tag{4.57}$$

Thus the forward-backward scheme is stable and neutral with time steps twice those allowed by the CFL criterion for the leapfrog scheme. The amplification factors of the forward-backward and of the leapfrog scheme are equal within their regions of stability. We now compare their effect on the phase speed by comparing the expression (4.56) for the forward-backward scheme, with (4.49), for the leapfrog scheme. The right-hand side of (4.56), with A replaced by $4A$, is equal to the right-hand side of (4.49). Because of the definition of A, this means that λ for the forward-backward scheme is identical to λ^2 for the leapfrog scheme when the time steps used for the forward-backward scheme are twice as long as those for the leapfrog scheme! Thus, the forward-backward scheme gives the same result using only half the computation time needed for the leapfrog scheme. In addition, as a two-level scheme, it has no computational mode in time. To understand this advantage over the leapfrog scheme we compare the finite-difference analogues that these two schemes give for the wave equation, since the system of gravity wave equations is equivalent to a single wave equation. Consider the one-dimensional version of this system:

$$\frac{\partial u}{\partial t} + g\frac{\partial h}{\partial x} = 0, \quad \frac{\partial h}{\partial t} + H\frac{\partial u}{\partial x} = 0 \tag{4.58}$$

Eliminating one of the variables u, h we obtain a wave equation

$$\frac{\partial^2 h}{\partial t^2} - gH\frac{\partial^2 h}{\partial x^2} = 0. \tag{4.59}$$

We can perform the same elimination for each of the finite-difference schemes. The forward-backward and space-centred approximation to (4.58) is

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + g\frac{h_{j+1}^n - h_{j-1}^n}{2\Delta x} = 0, \quad \frac{h_j^{n+1} - h_j^n}{\Delta t} + H\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\Delta x} = 0. \tag{4.60}$$

We now subtract from the second of these equations an analogous equation for the time level $n-1$ instead of n, divide the resulting equation by Δt, and

finally, eliminate all u values from it using the first of Eqs. (4.60), written for space points $j+1$ and $j-1$ instead of j. We obtain

$$\frac{h_j^{n+1} - 2h_j^n + h_j^{n-1}}{(\Delta t)^2} - gH\frac{h_{j+2}^n - 2h_j^n + h_{j-2}^n}{(2\Delta x)^2} = 0, \tag{4.61}$$

This is a finite-difference analogue of the wave equation (4.59). Note that although each of the two equations (4.60) is only of first order of accuracy in time, the wave equation analogue equivalent to (4.60) is seen to be of the second order of accuracy. If we use a leapfrog and space-centered approximation to (4.58), and follow an elimination procedure like that used in deriving (4.61), we obtain

$$\frac{h_j^{n+1} - 2h_j^{n-1} + h_j^{n-3}}{(2\Delta t)^2} - gH\frac{h_{j+2}^{n-1} - 2h_j^{n-1} + h_{j-2}^{n-1}}{(2\Delta x)^2} = 0, \tag{4.62}$$

This is also an analogue to the wave equation (4.59) of second-order accuracy. However in (4.61) the second time derivative was approximated using values at three consecutive time levels; in (4.62) it is approximated by values at every second time level only, that is, at time intervals $2\Delta t$. Thus, while the time step required for linear stability with the leapfrog scheme was half that with the forward-backward scheme (4.62) shows that we can omit the variables at every second time step, and thus achieve the same computation time as using the forward-backward scheme with double the time step. This method was discussed in the previous section for the two-dimensional case, it is the Eliassen grid. Thus, comparing (4.61) and (4.62) shows that the economy accomplished by the forward-backward scheme is equivalent to that accomplished with leapfrog time differencing by the Eliassen grid. Both of these methods avoid calculating the false time computational mode, and thus save half of the computation time with no effect on the physical mode of the solution. A disadvantage of the forward-backward scheme is that it is not possible to use the leapfrog scheme for the advection terms. However the second order accurate Adams-Bashforth scheme can be used for these terms.

4.8 Implicit and semi-implicit schemes

The time step permitted by the economical explicit schemes, twice that prescribed by the CFL criterion, is still considerably shorter than that required for accurate integration of the quasi-geostrophic motions. Thus we consider implicit schemes which are stable for any choice of time step. We shall consider here only the simplest of the implicit schemes, the trapezoidal rule

applied to the system (4.43) of pure gravity waves. For brevity it will simply be called the *implicit scheme*.

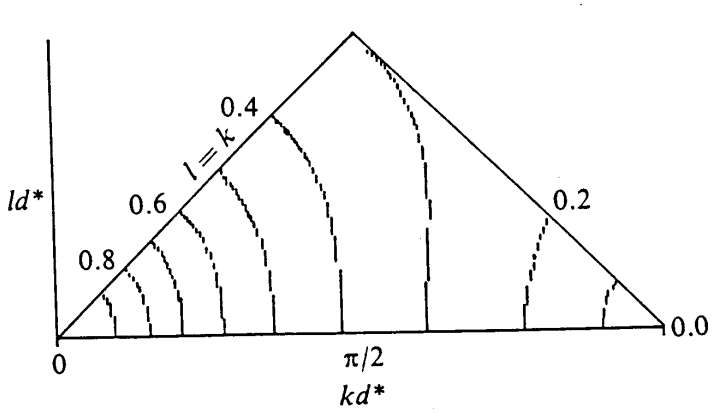

Figure 4.9: Relative speed of gravity waves, with implicit time and centred space differencing, and a Courant number equal to 5.

4.9 The implicit scheme (trapezoidal rule)

We consider the finite-difference scheme

$$\begin{aligned} u^{n+1} &= u^n - g\Delta t\frac{1}{2}(\delta_x h^n + \delta_x h^{n+1}), \\ v^{n+1} &= v^n - g\Delta t\frac{1}{2}(\delta_y h^n + \delta_y h^{n+1}), \\ h^{n+1} &= h^n - H\Delta t[(\delta_x u + \delta_y v)^n + (\delta_x u + \delta_y v)^{n+1}]. \end{aligned} \tag{4.63}$$

Substituting the wave solutions (4.45) we find three solutions for λ. One of these,

$$\lambda = 1, \tag{4.64}$$

is again that associated with neutral a stationary solution. The remaining two are

$$\lambda = \frac{1}{1+\frac{1}{2}A}\left(1 - \tfrac{1}{2}A \pm i\sqrt{2A}\right) \tag{4.65}$$

Examination of (4.65) shows that it always gives amplification factors satisfying

$$|\lambda| = 1, \tag{4.66}$$

and so the scheme is unconditionally stable and neutral. Using (4.65) and (4.52) we find for the relative phase speed of the nonstationary solutions,

$$\frac{c_*}{\sqrt{gH}} = \frac{1}{\mu\sqrt{2gH(X^2+y^2)}} \arctan\left(\pm\frac{\sqrt{2A}}{1-\frac{1}{2}A}\right). \tag{4.67}$$

The numerical values given by (4.67) for the physical mode with $2\sqrt{gH}\mu = 5$ are shown in Fig. 4.9. The wave number region is the same as in the earlier diagram, Fig. 4.4. Comparing the isolines of the present figure with those of Fig. 4.4, where the effect of the space differencing alone was considered, shows that the effect of time differencing on phase speed is now not negligible. Implicit time differencing is seen to result in a *considerable retardation of gravity waves* of the same order of the magnitude as that due to centred space differencing. The solution of the implicit scheme for new variables u^{n+1}, v^{n+1} is no longer trivial.

4.10 The semi-implicit method of Kwizak and Robert

There is no advantage in using an implicit method for advection, Coriolis, and other terms of governing equations in atmospheric models. They are associated with slower phase speeds, and should not require excessively small time steps for linear stability when calculated explicitly. Since the trapezoidal implicit scheme is a two-level scheme like the forward-backward scheme, it is convenient to use Adams-Bashforth scheme for this purpose. Kwizak and Robert (1971) chose, however, to use the leapfrog scheme. The usual procedure used for solving the semi-implicit difference system for variables at time level $n+1$ will be illustrated for the shallow water equations. These equations can be written in a compact form

$$\begin{aligned} \frac{\partial u}{\partial t} &= -g\frac{\partial h}{\partial x} + A_u \quad \frac{\partial v}{\partial t} = -g\frac{\partial h}{\partial y} + A_v, \\ \frac{\partial h}{\partial t} &= -H\nabla\cdot v + A_h, \end{aligned} \tag{4.68}$$

where A_u, A_v and A_h denote the terms that were omitted in the system (4.43) describing the propagation of pure gravity waves. These additional terms,

and implicit differencing over a time interval $2\Delta t$ for the gravity wave terms and centred space differencing (4.68) is replaced by

$$\begin{aligned} u^{n+1} &= u^{n-1} - g\Delta t(\delta_x h^{n-1} + \delta_x h^{n+1}) + 2\Delta t A_u^n, \\ v^{n+1} &= v^{n-1} - g\Delta t(\delta_y h^{n-1} + \delta_y h^{n+1}) + 2\Delta t A_v^n, \\ h^{n+1} &= h^{n-1} - H\Delta t[(\delta_x u + \delta_y v)^{n-1} + (\delta_x u + \delta_y v)^{n+1}] \qquad (4.69) \\ &+ 2\Delta t A_h^n. \qquad (4.70) \end{aligned}$$

We now apply the operator δ_x to the first and δ_y to the second of these equations, respectively and add the results. We introduce the notation

$$\delta_{xx}h = \delta_x(\delta_x h) and \delta_{yy}h = \delta_y(\delta_y h).$$

We obtain

$$\begin{aligned} (\delta_x u + \delta_y v)^{n+1} &= (\delta_x u + \delta_y v)^{n-1} - g\Delta t[(\delta_{xx} + \\ \delta_{yy})h^{n-1} + (\delta_{xx} &+ \delta_{yy})h^{n+1}] + 2\Delta t \cdot (\delta_x A_u + \delta_y A_v)^n \end{aligned}$$

Substituting the right-hand side into the third of Eqs. (4.69), and defining the 'finite-difference Laplacian' by

$$\nabla^2_{\oplus} h = (\delta_{xx} + \delta_{yy})h$$

we find that

$$\begin{aligned} h^{n+1} &= h^{n-1} - 2H\Delta t(\delta_x u + \delta_y v)^{n-1} + gH(\Delta t)^2(\Delta^2_{\oplus} h^{n-1} + h^{n+1}) \\ &+ 2\Delta^2_{\oplus} h^{n+1} + 2\Delta t[A_h - H\Delta t(\delta_x A_u + \delta_y A_v)]^n \end{aligned}$$

Using, in addition the definitions

$$F^{n-1} := h^{n-1} 2H\Delta t(\delta_x u + \delta_y v)^{n-1} + gH(\Delta t)^2 \Delta^2_{\oplus} h^{n-1},$$

$$G^n := 2\Delta t[Ah - H\Delta t(\delta_x A_u + \delta_y A_v)]^n$$

this can be written as

$$h^{n+1} - gH(Dt)^2 h^{n+1} = F^{n-1} + G^n. \qquad (4.71)$$

The terms have been arranged to show that at time level n the right hand side is known at all space grid points. Once the equation has been solved for the values h^{n+1}, u^{n+1} and v^{n+1} can be obtained directly from the first and the second of Eqs. (4.69). The algebraic system (4.71) is an approximation to an elliptic equation

$$\nabla^2 h + ah + b(x, y) = 0. \tag{4.72}$$

4.11 The splitting or Marchuk method

Since there are different factors in a complex system of hydrodynamic equations for weather prediction, we will normally wish to use different schemes for terms associated with them. Thus considering the linearized system with advection and gravity wave terms,

$$\begin{aligned} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} + g\frac{\partial h}{\partial x} &= 0 \\ \frac{\partial h}{\partial t} + c\frac{\partial h}{\partial x} + H\frac{\partial u}{\partial x} &= 0 \end{aligned} \tag{4.73}$$

we might wish to use one scheme for the advection terms, and another for the gravity wave terms - in much the same way as was done for the semi-implicit scheme. In such a situation, even though both of the schemes to be used are stable considered one at a time, we cannot be certain that the scheme obtained as a combination of the two will also be stable. These problems can be avoided using the splitting method. The idea of this method is to construct

schemes for a complex system of equations so that within each time step this system is split into a number of simpler subsystems, which are then solved consecutively one at a time. In case of (4.73), within a given time step, we could first solve the system of advection equations

$$\begin{aligned} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} &= 0 \\ \frac{\partial h}{\partial t} + c\frac{\partial h}{\partial x} &= 0. \end{aligned} \tag{4.74}$$

Denote the provisional values u^{n+1}, h^{n+1}, obtained in this way by u^*, h^*. Use these values at the beginning of the time step for solving the remaining subsystem

$$\begin{aligned} \frac{\partial u}{\partial t} + g\frac{\partial h}{\partial x} &= 0 \\ \frac{\partial h}{\partial t} + H\frac{\partial u}{\partial x} &= 0. \end{aligned} \tag{4.75}$$

The values u^{n+1}, h^{n+1}, obtained after solving this other subsystem, are now taken as actual approximate values of these variables at the level $n+1$. The procedure is repeated in each following time step. A solution obtained by the splitting method will represent a consistent approximation to the true solution. This can be proved easily for a particular choice of schemes for solving the subsystems. The approximate values of the dependent variables then have to approach the true values as the time step approaches zero. Stability of the splitting method To study the stability of schemes constructed by the splitting method, we consider the example above. Denote by la and lb the values of l of the schemes chosen for the numerical solution of subsystems (4.74) and (4.75), respectively. Then we have

$$u^* = Re(\lambda_a \lambda^n \hat{u} e^{ikx}), \quad h^* = Re(\lambda_a \lambda^n \hat{h} e^{ikx}) \tag{4.76}$$

and

$$u^{n+1} = Re(\lambda_b \lambda_a \lambda^n \hat{u} e^{ikx}), \quad h^{n+1} = Re(\lambda_b, \lambda_a \lambda^n \hat{h} e^{ikx}). \tag{4.77}$$

Therefore, we find,

$$\lambda = \lambda_b \lambda_a, \tag{4.78}$$

and

$$|\lambda| = |\lambda_b| \cdot |\lambda_a|. \tag{4.79}$$

Thus, if both of the schemes chosen for the solution of subsystems (4.74) and (4.76) - (4.79) are stable, the combined scheme constructed by the splitting method will also be stable. This conclusion can be generalized for an arbitrary system of equations and number of subsystems.

When applying the splitting method, we do not necessarily have to use equal time steps for each of the subsystems. This may be the main advantage of the splitting method: we can chose a relatively long time step for the subsystem governing a slow process, advection in the present example, and then use a number of smaller steps to calculate the faster process. Significant economies can be accomplished in this way. A disadvantage of the splitting method is that calculation of the effects of different physical factors one at

a time usually leads to an increase in the truncation error. The splitting method was first used in atmospheric models by Marchuk (1967); thus in meteorology it is known as the Marchuk method.

4.12 Two-grid-interval noise suppression

A number of methods have been used to cope with two-grid-interval noise. In many models dissipative schemes are used to give maximum damping for the two-grid-interval wave, or lateral diffusion is added with relatively large diffusion coefficients. The appearance of two-grid interval noise is thereby suppressed. However, instead of attacking the consequences of inadequacies in a simulation of a physical process, it is generally better to look for a method that would achieve a physical correct simulation of that process, and thus eliminate the cause of the difficulty. Mesinger (1973) showed how two-grid-interval wave noise could be prevented in some cases even by using centred differencing. We consider the system of linearized equations (4.43).

Mesinger proposed a grid in which a height perturbation at a single grid point is propagated by gravity waves to all the other height grid points. Therefore there can be no grid-splitting and two-grid-interval noise in the height field. Since a velocity perturbation can propagate as a gravity wave only by exciting height perturbations, the procedure will prevent false two-grid-interval noise in all the variables. We shall illustrate this procedure using the implicit scheme (4.63). The velocity components at regular velocity points are computed in the way as before, so the first two equations of that system remain unchanged. To calculate the velocity divergence in the continuity equation we define auxiliary velocity points midway between the neighbouring height points, as shown by the circled numbers 5, 6, 7 and 8 in Fig. 4.10.

Using the system x', y' shown in this figure, components u' are needed at points 5 and 7, and components v' at points 6 and 8. At the beginning of the time step Δt these components are obtained by space-averaging, that is

$$u'^n = \frac{\sqrt{2}}{2}\left(\overline{u}^{y'} + \overline{v}^{y'}\right)^n, \quad v' = \frac{\sqrt{2}}{2}\left(-\overline{u}^{x'} + \overline{v}^{x'}\right)^n \tag{4.80}$$

An overbar denotes a two-point average taken along the direction indicated following the bar sign. Acceleration contributions are added to these initial values to obtain values at the end of the time step

$$\begin{aligned} u'^{n+1} &= u'^n - g\Delta t \tfrac{1}{2}\left(\delta_{x'} h^n + \delta_{x'} h^{n+1}\right), \\ v' &= v'^n - g\Delta t \tfrac{1}{2}\left(\delta_{y'} h^n + \delta_{y'} h^{n+1}\right). \end{aligned} \tag{4.81}$$

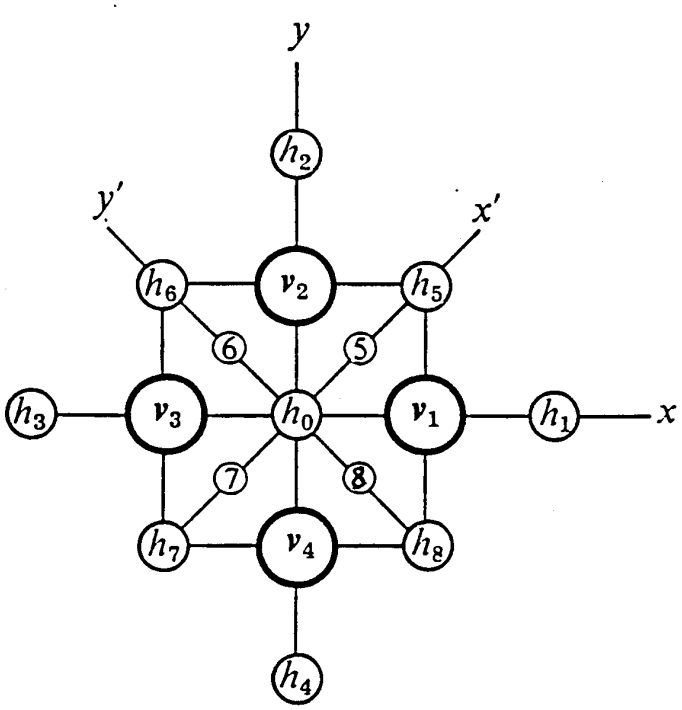

Figure 4.10: Stencil used to denote the height and velocity grid point values surrounding a height point.

The velocity divergence in the continuity equation can now be approximated by

$$\tfrac{1}{2}\left(\delta_x u + \delta_y v\right) + \tfrac{1}{2}\left(\delta_{x'} u' + \delta_{y'} v'\right), \tag{4.82}$$

giving weight to all eight directions of the lattice. In this way the implicit approximation to the continuity equation may be obtained as

$$h^{n+1} = h^n - H\Delta t\left(\delta_x u + \delta_y v\right)^n + \tfrac{1}{4} gH\left(\Delta t\right)^2\left(\nabla^2_{\otimes} h^n + \nabla^2_{\otimes} h^{n+1}\right). \tag{4.83}$$

Here the velocity components at level $n+1$ have already been eliminated using the first two of Equations (4.63), and

$$\nabla^2_{\otimes} = \frac{1}{4d^2}(h_1 + h_2 + h_3 + h_4 + 2(h_5 + h_6 + h_7 + h_8) - 12h_0). \tag{4.84}$$

This is again a finite-difference approximation to $\nabla^2 h$, but now it is calculated using the height values of nine neighbouring points. Comparing this scheme

with the standard implicit scheme (4.63), the only modification is that this nine-point Laplacian has replaced the five-point Lapalcian in the continuity equation. This allows propagation of gravity waves between all height points of the grid, thus admitting no false space noise in the height field. The modification has no effect on the unconditional stability of the implicit scheme; however, instead of being neutral for all waves, the scheme now damps shorter waves to some extent. The modified scheme has a smaller truncation error than the unmodified scheme. It is important to be aware that this method is *not* attempting to improve the calculation of short gravity waves of wave lengths close to two grid intervals. At this scale the finite-difference representation is very poor, and significant improvements in accuracy can hardly be expected. The problem is that gravity waves, with longer wave lengths can propagate independently on individual (C) type subgrids, and thus erroneously appear to have wave lengths close to two grid intervals. Thus we are confronted with a kind of aliasing error. The proposed method enables these waves to appear with wave length close to their physical value instead in the noise region with wave lengths close to two grid intervals.

Mesinger denotes the application of the grid structure in Fig. 4.10 as "cceleration-modified" versions of a particular scheme. The phase speed of gravity waves as a function of wave numbers k and l is shown in Fig. 4.11 for the implicit scheme, the Heun scheme and the leapfrog scheme. The shading close to the right corner corresponds with a region of zero phase speed. Since the solutions in these regions are damped, there is no danger that it can be excited to a significant amplitude.

4.13 Time noise and time filtering

In addition to the appearance of spurious short-wave noise in space, spurious short-wave noise in time, that is high frequency noise can appear in numerical models. As a result of initial conditions which are interpolated from observations, the numerical forecasts will contain spurious gravity waves of unrealistically large amplitudes. Experience shows that the noise generated by assimilation of the observed data typically dies out to an acceptable level in about 24 hours of simulated time due to geostrophic adjustment. However it may be desirable to accelerate this adjustment by appropriate numerical techniques. The Matsuno schemes can be used for this purpose.

Another method that can be used to increase the damping of high frequency noise in atmospheric models is *time filtering* originally proposed by Robert(1966). To apply this at least three consecutive values of the function to be filtered are needed. It suffices to consider one function only, which we

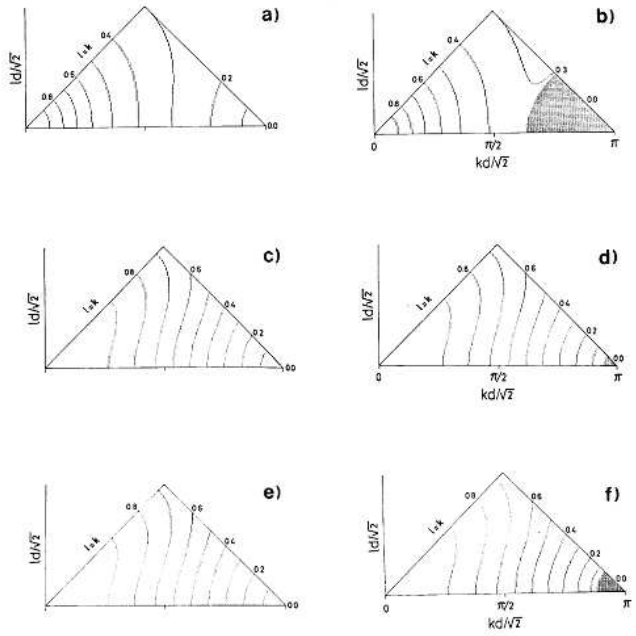

Figure 4.11: Relative phase speed of gravity waves as a function of wave numbers k and l. a) for the implicit scheme. b) for the acceleration-modified implicit scheme. c) the Heun-scheme. d) the acceleration-modified Heun-scheme. e) the leapfrog-scheme. f) the acceleration-modified leapfrog-scheme.

assume to be a solution of the oscillation equation. Thus we consider the function

$$U(t) = U(0)e^{i\omega t}, \tag{4.85}$$

where the values $U(t - \Delta t)$, $U(t)$ and $U(t + \Delta t)$ are known. We shall first examine the effect of changing only the middle of these three values using the relation

$$(t) = \overline{U}(t) + \tfrac{1}{2}S[U(t - \Delta t) - 2U(t) + U(t + \Delta t)] \tag{4.86}$$

known as the *centred filter*. The overbar now denotes the filtered value of a function, and S is the *filter parameter*. The expression within the square bracket in (4.86) is proportional to the simplest approximation to the second derivative in time; thus, for sufficiently small positive values of S application

of the filter (4.86) will decrease the curvature in a graph of the three values of $U(t)$. For quantitative analysis of the effect of the filter we define

$$\overline{U}(t) = R \cdot \times U(t), \tag{4.87}$$

where the complex factor R is called the *response* of the filter. When this is substituted in into (4.85) and we use (4.86), we obtain

$$R = 1 - S(1 - \cos \omega \Delta t). \tag{4.88}$$

It is convenient to define

$$R = |R| e^{i\delta}.$$

We can then say that the phase change δ resulting from the centred filter is zero, and that within the CFL stability criterion and for small positive values of S the amplitude factor $|R|$ exerts a damping effect increasing with increasing frequencies.

When, however, a filter is continually applied during a numerical integration, the value $U(t - \Delta t)$ has already been changed prior to changing $U(t)$. It is then appropriate to consider the filter

$$(t) = U(t) + S[(t - \Delta t) - 2U(t) + U(t + \Delta t)]. \tag{4.89}$$

Asselin (1972) calls this the *basic time filter*. A procedure like the one used in deriving (4.87) now gives

$$R = \frac{(2 - S)^2 + 2S^2(1 - \cos\ \omega \Delta t)}{(2 - S)^2 + 4S(1 - \cos\ \omega \Delta t)} e^{i\omega \Delta t}. \tag{4.90}$$

Thus, there is now a phase change that is different from zero; however it is small for small values of $\omega \Delta t$. The amplitude factor is not much different from that of the centred filter for small values of S.

Using an analogous approach one can analyze the effect of smoothing and filtering in space.

Extension of the centred filter to two space dimensions may be accomplished in two ways. First, one may smooth in each dimension, independently of the other dimension. It can be shown that the final result is independent of the dimension in which one first smooths, and also independent of the order in which one applies the smoothing elements.

4.13.1 Example of filtering

We consider the effect of the centred filter applied to a two-dimensional array. In Fig. 4.5a the example distribution is displayed. It consists of a high of the product of single waves $\sin(x) \cdot \sin(x)$ and a low, which has a Gaussian height distribution. The maximum value is 10 and the minimum value is -10. The

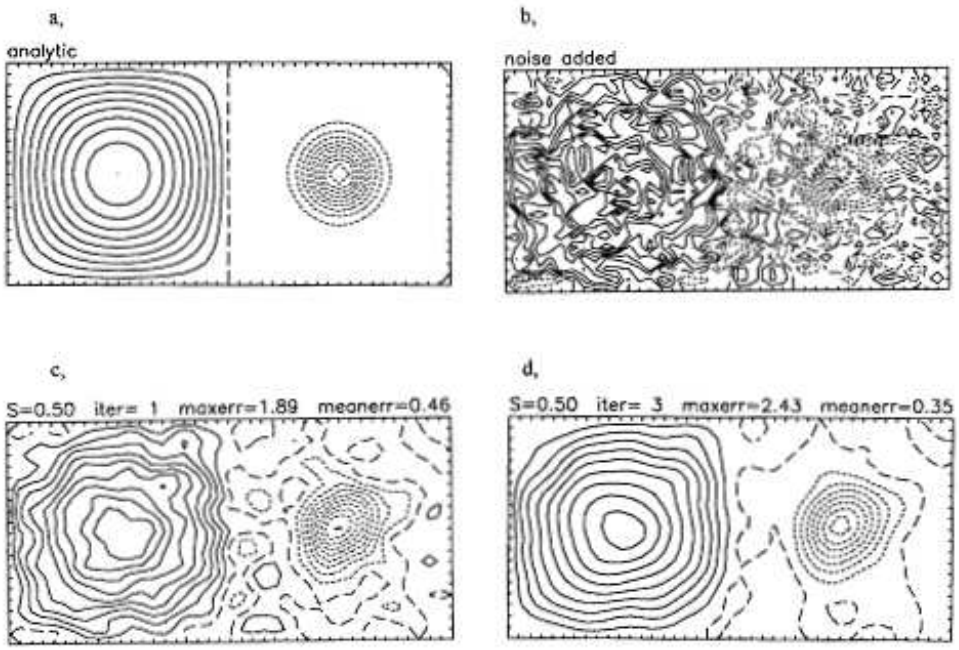

Figure 4.12: Example of space filtering with the centred filter applied to a two-dimensional array. a) undisturbed initial (ideal) distribution. Contour interval in all panels 1 height unit, solid contours represent positive values, dashed contours negative values. The extrema correspond with 10 and -10 height units. b) The initial distribution is disturbed by random numbers with amplitude ± 2.5 height units. c) Height distribution after one iteration with $S = 0.5$. The maximum error (maxerr) and mean error (meanerr) are displayed on top of this panel and the following panel. d) As in c) but for 3 iterations.

number of intervals is 40 by 20. This distribution is disturbed with random noise with a maximum amplitude of ± 2.5 at all inner grid points, shown in Fig. 4.5b. The following panels display the effect of the centred filter applied first in x-direction to the interior of the whole array

$$\overline{h}^x(x,y) = h(x,y) + \frac{1}{2}S\left[h(x+\Delta x,y) - 2h(x,y) + h(x-\Delta x,y)\right],$$

and then in y-direction to the interior of the whole array

$$\overline{h}^y(x,y) = h(x,y) + \frac{1}{2}S\left[h(x,y+\Delta y) - 2h(x,y) + h(x,y-\Delta y)\right].$$

Once the whole array is processed, the process is repeated, denoted as iterations. Fig. 4.5 shows that the original pattern can be retrieved this way. Best results are obtained with a few iterations. The filter parameter S and the maximum and the mean error, with respect to the original field Fig. 4.5 are displayed in a line above the panels.

Exercises

4.1 Write down the splitting method for the system (4.67) with the leapfrog method for advection and the forward-backward-scheme for the gravity wave terms. Derive a stability criterion.

4.2 Derive the elliptic finite-difference equation (4.84).

4.3 Show, using the acceleration modified Heun-scheme with time step 2Dt that the equation for the height-field can be written

$$h^{n+1} = h^{n-1} - 2H\Delta t(\delta_x u + \delta_y v)^{n-1} + 2gH(\Delta t)^2\nabla_{\otimes}^2 h^{n-1}$$

4.4 Show that the application of two smoothing elements, one in each dimension, the two elements may be combined into a single 9-point operator.

Chapter 5

Methods for Solving Elliptic Equations

In this chapter we consider methods for the solution of difference equations arising from discretization of elliptic partial differential equations. As a special case of elliptic equations we consider the Helmholtz-equation

$$\nabla^2\psi - \alpha\psi = \zeta. \tag{5.1}$$

This arises from the use of the semi-implicit method for shallow water equations as introduced in chapter 4 or as a Poisson-Equation ($\alpha = 0$). Another example is the inversion of pressure in nonhydrostatic Boussinesq equations. The solutions of elliptic equations are known as boundary value problems. The boundaries may be of periodic, of Dirichlet type (ψ prescribed), of Neumann type (normal gradient $\frac{\partial\psi}{\partial n}$ prescribed) or of Robin type (also known as mixed type, a combination of Dirichlet and Neumann type). As a paradigm of elliptic equations we chose the Poisson equation. For the iterative methods we present here the extension to general elliptic equations is straightforward.

5.1 The Poisson equation

The discretized form of the Poisson equation, using second-order differences is

$$\delta_{xx}\psi + \delta_{yy}\psi = \zeta \tag{5.2}$$

resulting in a five point formula

$$\frac{\psi_{i+1,j} - 2\psi_{i,j} + \psi_{i-1,j}}{\Delta x^2} + \frac{\psi_{i,j+1} - 2\psi_{i,j} + \psi_{i,j-1}}{\Delta y^2} = \zeta_{i,j} \tag{5.3}$$

where $\zeta_{i,j}$ is known.

5.2 General considerations

In a rectangular domain where Max(i) = I and Max(j) = J, equation (5.3) and the boundary equations give a system of $N = (I-2) \times (J-2)$ simultaneous linear algebraic equations. Except for the *source* or non-homogeneous term, $\zeta_{i,j}$, it is the same block-tridiagonal system which results from a fully implicit formulation of the two-dimensional diffusion equation and the economical tridiagonal algorithm is not applicable. The most elementary methods of solving such a system are Cramer's rule and the various forms of Gaussian elimination. For the problems of interest, N is a large number, and these methods are inadequate. Cramer's rule involves quite an unbelievable number of operations - approximately $(N+1)!$ multiplications. Even if the time were available ,the accuracy would be destroyed by round-off error. The multiplications in Gaussian methods vary as N^3, and the accuracy may be expected to deteriorate for N much greater than 50 depending on the details of the method and the computer word length.

In recent years, highly efficient direct methods have been developed. Dorr (1970) reviews the 'block' methods, cyclic reduction methods, tensor product methods, the Fourier series methods, and a few others. All these methods have one ore more of the following disadvantages: limited to rectangular domains like $L-$ or T-shapes; limited to boundary conditions; large storage requirements; not adaptable to non-Cartesian coordinates; field size limited (restricting the magnitude of I or J) due to accumulation of round-off error; field specification limited (e.g., I or J limited to the form 2^k, k integer); elaborate preliminary calculations specific to the mesh required; complex to program; difficult to understand. However, the direct methods, particularly the Fourier-series methods, are increasing in use especially for large problems (I and J large). There exist commonly available efficient FORTRAN program libraries for the direct (and also iterative) solution of elliptic equations e.g. FISHPACK and ELLPACK. By comparison, the various iterative methods are very easy to understand and to program, and are quite flexible. The iterative methods have historically been used most often in computational fluid dynamics and will undoubtedly continue to be important. The latest developments are the (iterative) multigrid methods.

5.3 Richardsons Method

The solution of elliptic equations by iteration is analogous to solving a time-dependent problem to an asymptotic steady state. Suppose we consider a time-dependent diffusion equation for ψ with a source term, ζ, and with a diffusion coefficient of unity

$$\frac{\partial \psi}{\partial t} = \nabla^2 \psi - \zeta. \tag{5.4}$$

We are not interested in the physical significance of the transient solution, but as a steady state of this solution is approached, $\frac{\partial \psi}{\partial t} \to 0$, the solution approaches the desired solution for the Poisson equation $\nabla^2 \psi = \zeta$. In some cases the analogy is exact. To illustrate the equivalence, we shall proceed to derive Richardsons iterative method for the elliptic Poisson equation from the Euler (forward) time-step with centred space differences time-dependent method for the parabolic diffusion equation. Applying forward in time, centred in space (FTCS) differencing to (5.4) we obtain

$$\frac{\psi_{i,j}^{k+1} - \psi_{i,j}^{k}}{\Delta t} = (\delta_{xx} + \delta_{yy})\psi_{i,j}^{k} - \zeta_{i,j}. \tag{5.5}$$

To avoid confusion, we momentarily restrict ourselves to $\Delta x = \Delta y = \Delta$. Then Eq. (5.5) gives

$$\psi_{i,j}^{k+1} = \psi_{i,j}^{k} + \frac{\Delta t}{\Delta^2}\left[\psi_{i+1,j}^{k} + \psi_{i-1,j}^{k} + \psi_{i,j+1}^{k} + \psi_{i,j-1}^{k} - 4\psi_{i,j}^{k} - \Delta^2 \zeta_{i,j}\right]. \tag{5.6}$$

We first demonstrate that $\zeta_{i,j}$ does not affect the stability properties of Eq. (5.6). We consider Dirichlet boundary conditions and denote the exact solution of the finite-difference equation of the Poisson Eq. (5.2) by ψ^{∞}. The errors $e_{i,j}^{k}$ in the iterative values are then

$$e_{i,j}^{k} = \psi_{i,j}^{\infty} - \psi_{i,j}^{k}. \tag{5.7}$$

Substituting this into (5.2) gives

$$\delta_{xx}\psi^{\infty} + \delta_{yy}\psi^{\infty} - \delta_{xx}e^{k} - \delta_{yy}e^{k} = \zeta. \tag{5.8}$$

Since ψ^{∞} exactly satisfies (5.2), then (5.8) reduces to the Laplace equation,

$$\delta_{xx}e^{k} + \delta_{yy}e^{k} = 0. \tag{5.9}$$

Since the boundary values are known, $\psi^{k} = \psi^{\infty}$ or $e = 0$ on all boundaries. Then the iteration (5.6) can be written for the error e as

$$e_{i,j}^{k+1} = e_{i,j}^{k} + \frac{\Delta t}{\Delta^2}\left[e_{i+1,j}^{k} + e_{i-1,j}^{k} + e_{i,j+1}^{k} + e_{i,j-1}^{k} - 4e_{i,j}^{k}\right]. \tag{5.10}$$

So the iteration (5.6) for ψ is equivalent to the iteration (5.10) for e, which clearly does not depend on ζ. The stability restriction on Eq. (5.6) is just $\Delta t \leq \Delta^2/4$ (Exercise 5.2). Since we wish to approach the asymptotic condition as rapidly as possible, we consider taking the largest possible $\Delta t = \Delta^2/4$. Substituting into (5.6) then gives

$$\psi_{i,j}^{k+1} = \psi_{i,j}^{k} + \frac{1}{4}\left[\psi_{i+1,j}^{k} + \psi_{i-1,j}^{k} + \psi_{i,j+1}^{k} + \psi_{i,j-1}^{k} - 4\psi_{i,j}^{k} - \Delta^2\zeta_{i,j}\right]. \tag{5.11}$$

The terms inside and outside the brackets cancel, giving

$$\psi_{i,j}^{k+1} = \frac{1}{4}\left[\psi_{i+1,j}^{k} + \psi_{i-1,j}^{k} + \psi_{i,j+1}^{k} + \psi_{i,j-1}^{k} - \Delta^2\zeta_{i,j}\right]. \tag{5.12}$$

This is *Richardson's method* for $\Delta x = \Delta y = \Delta$. The method is also known as the Jacobi method or as *iteration by total steps*, or *method of simultaneous displacements*, since each $\psi_{i,j}^{k+1}$ is calculated independent of the sequence in (i, j) and therefore, in a sense, simultaneously. This distinction is important to the consideration of techniques for parallel-processing computers. For the special case of Laplace's equation ($\zeta_{i,j} = 0$), the method simply involves setting the new iterative value equal to the arithmetic average of its four neighbours. Equation (5.12) is the same result that is obtained from simply solving the (steady) elliptic Eq. (5.3) for $\psi_{i,j}$ and evaluating this term on the left side of the equation at $(k+1)$ and all terms on the right-hand side at k. Defining the mesh aspect ratio $\beta = \Delta x/\Delta y$ this gives

$$\psi_{i,j}^{k+1} = \frac{1}{2(1+\beta^2)}\left[\psi_{i+1,j}^{k} + \psi_{i-1,j}^{k} + \beta^2\psi_{i,j+1}^{k} + \beta^2\psi_{i,j-1}^{k} - \Delta x^2\zeta_{i,j}\right]. \tag{5.13}$$

This is Richardson's method for $\Delta x \neq \Delta y$. The analysis of the convergence rate can proceed as in the analysis of stability for the vorticity equation, now writing

$$e^{k+1} = \lambda e^{k} \tag{5.14}$$

for the error equation (5.10). The highest and lowest wavelength error components damp most slowly ,i.e. have largest $|\lambda(\theta)|$.

5.4 Liebmanns method

Equation (5.13) is a two-level equation, requiring storage of ψ^{k+1} and ψ^k. If we sweep in $i \uparrow$, $j \uparrow$ and use new values wherever available in Eq. (5.13), we obtain

$$\psi_{i,j}^{k+1} = \frac{1}{2(1+\beta^2)} \left[\psi_{i+1,j}^{k} + \psi_{i-1,j}^{k+1} + \beta^2 \psi_{i,j+1}^{k} + \beta^2 \psi_{i,j-1}^{k+1} - \Delta x^2 \zeta_{i,j} \right], \quad (5.15)$$

which is *Liebmann's method*, also known as *Gauss-Seidel method*. This method may be programmed with only one storage level, interpreting (5.15) as a FORTRAN replacement statement

$$\ldots$$

$$do\ j = 2, jj - 1$$

$$do\ i = 2, ii - 1$$

$$\psi(i,j) = fac^*(\psi(i+1,j)+\psi(i-1,j)+bb^*\psi(i,j+1)+bb^*\psi(i,j-1)-dd^*\zeta(i,j))$$

$$\text{enddo}$$

$$\text{enddo}$$

$$\ldots$$

For the most resistant high and low wavenumber error components,

$$\lambda \text{ (Liebmann)} = [\lambda \text{ (Richardson)}]^2 . \quad (5.16)$$

Asymptotically, k Liebman iterations are worth $2k$ Richardson iterations, and only require half the core storage.

5.5 Southwell's Residual Relaxation Method

The original name for the method was just the relaxation method, but here we use the term *residual relaxation* to distinguish it from Liebman's method and other iterative procedures, which today are sometimes called *relaxation methods*. The simplest form of Southwell's residual relaxation method uses the same equation as Richardson's method (5.13) to evaluate new $\psi_{i,j}^{k+1}$. The difference is that Eq. (5.13) is not applied indiscriminately to all mesh points in a sweep of the matrix. Rather the residual $r_{i,j}$ is defined by

$$r_{i,j} = \frac{\psi_{i+1,j} - 2\psi_{i,j} + \psi_{i-1,j}}{\Delta x^2} + \frac{\psi_{i,j+1} - 2\psi_{i,j} + \psi_{i,j-1}}{\Delta y^2} - \zeta_{i,j}. \tag{5.17}$$

When $r_{ij} = 0$, the Poisson equation (5.3) is satisfied, but only at the point (i, j). Thus, $|r_{i,j}|$ is indicative of how much the present estimate for all $\psi_{i,j}$ is in error at (i, j). One then scans the field for the largest $|r_{i,j}|$, and sets this $r_{ij} = 0$ by calculating a new $\psi_{i,j}$ from (5.13). This in turn changes r at all neighbouring points and the scan is repeated. This method is not used in today's computers because the time required to scan for Max(r_{ij}) and recalculate neighbouring $r's$ is not sufficiently shorter than the time needed to apply Eq. (5.13). Southwell's method was historically important because a more refined version which evolved suggested the extrapolated Liebmann method, more commonly known as the successive over-relaxation method.

5.6 Successive Over-Relaxation (SOR) Method

Frankel (1950) and independently Young (1954) developed a method of applying overrelaxation to Liebmann's method in a methodical manner suited to electronic computers. Frankel called it the extrapolated Liebmann method, and Young called it successive over-relaxation. Adding to Eq. (5.15) and regrouping gives

$$\begin{aligned} &\psi_{i,j}^{k+1} = \psi_{i,j}^{k} + \tfrac{1}{2(1+\beta^2)} \\ &\left[\psi_{i+1,j}^{k} + \psi_{i-1,j}^{k+1} + \beta^2\psi_{i,j+1}^{k} + \beta^2\psi_{i,j-1}^{k+1} - \Delta x^2\zeta_{i,j} - 2(1+\beta^2)\psi_{i,j}^{k}\right]. \end{aligned} \tag{5.18}$$

Now as a solution is approached, $\psi_{i,j}^{k+1} \to \psi_{i,j}^{k}$ for all (i, j) the bracketed term becomes zero identically by Eq. (5.3) and (5.18) becomes a statement of convergence, $\psi_{i,j}^{k+1} = \psi_{i,j}^{k}$. If Liebmann's method is used and the bracketed term is set identically to zero at the point (i, j), $\psi_{i,j}^{k+1} = \psi_{i,j}^{k}$. That is, the residual $r_{ij} = 0$. In the SOR method, the bracketed term in (5.18) is multiplied by a relaxation factor ω, where $\omega \neq 0$ thus, $r_{ij} \neq 0$, but $r_{ij} \to 0$ as $\psi_{i,j}^{k+1} \to \psi_{i,j}^{k}$, as before.

$$\begin{aligned} &\psi_{i,j}^{k+1} = \psi_{i,j}^{k} + \tfrac{\omega}{2(1+\beta^2)} \\ &\left[\psi_{i+1,j}^{k} + \psi_{i-1,j}^{k+1} + \beta^2\psi_{i,j+1}^{k} + \beta^2\psi_{i,j-1}^{k+1} - \Delta x^2\zeta_{i,j} - 2(1+\beta^2)\psi_{i,j}^{k}\right]. \end{aligned} \tag{5.19}$$

For convergence, it is required that $1 \leq \omega < 2$. Frankel and Young both determined an optimum value, ω_0, basing their optimality criterion on the

asymptotic reduction of the most resistant error. The optimum value depends on the mesh, the shape of the domain, and the type of boundary conditions. For the Dirichlet (constant boundary values) problem in a rectangular domain of size $(I-1)\Delta x$ by $(J-1)\Delta y$ with constant Δx and Δy, it may be shown that

$$\omega_0 = 2\left(\frac{1-\sqrt{1-\xi}}{\xi}\right) \tag{5.20}$$

where

$$\xi = \left[\frac{1}{1+\beta^2}\left\{\cos\left(\frac{\pi}{I-1}\right)+\beta^2\cos\left(\frac{\pi}{J-1}\right)\right\}\right]^2. \tag{5.21}$$

With $\omega = \omega_0$, the k required to reduce the error to some specified level varies directly with the total number of equations $N = (I-2)\times(J-2)$ for the Liebmann method $k \propto N^2$. So the SOR method with the optimum ω_0, sometimes referred to as the *optimum over-relaxation method*, is much better for large problems. Analytic evaluation of ω_0 exists only for slightly more general problems (e.g. Warlick and Young, 1970). Miyakoda (1962) has shown that ω_0 increases for the case of the Neumann conditions at all boundaries. For Dirichlet conditions along some boundaries and the Neumann conditions along others, for varying Δ_x or Δ_y, and for L-shaped and most other non-rectangular regions, no analytic evaluation exists. In such cases, ω_0 may be determined experimentally, by solving successions of Laplace equations with zero boundary conditions using different values of ω over $1 < \omega < 2$, and monitoring convergence toward $\psi^k_{i,j} = 0$ for large k (the value of ω_0 does not change with the source term, ζ). It is important to assure that all error components are present in the initial condition. This is readily met by choosing $\psi^0_{i,j} = 1$ at all interior points. The process of experimentally finding ω_0 is tedious because the convergence rate is usually very sensitive to ω near ω_0. An example is given in Fig. 5.1.

The curvature near ω_0 shows that it is usually best to slightly overestimate ω_0 than to underestimate it. The use of some guessed value, say $\omega = 1.1$, is seen to have very little good effect. The experimental determination of an approximation to ω_0 is then always worthwhile, when the Poisson equation must be solved many times, e.g. at every time step of the (barotropic) vorticity equation. The SOR method described here is the original *point* SOR of Frankel and Young. It uses advance $(k+1)$ values at neighbouring points $(I-1, j)$ and $(i, j-1)$. It is possible to slightly improve the convergence rate further by line SOR, which uses advance $(k+1)$ values at 3 neighbouring points. Ames (1969, page 147) states that line SOR will converge in fewer

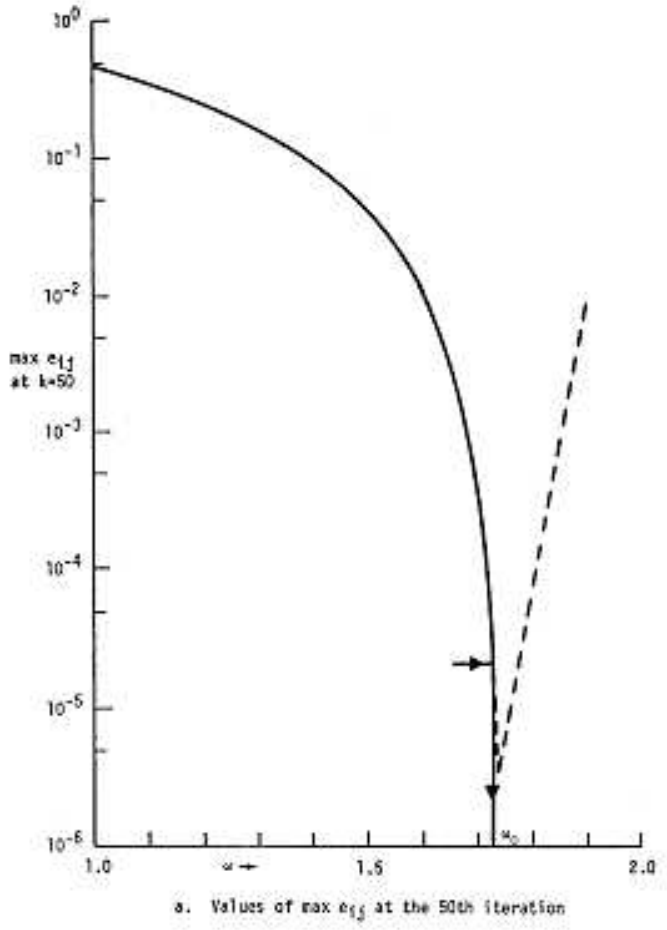

Figure 5.1: Behaviour of SOR iteration for values of the relaxation parameter ω. $I = J21$, $\Delta x = \Delta Dy$, which gives $\omega_0 = 1.7295$.

iterations than point SOR by a factor of $1/\sqrt{2}$ for $\Delta x = \Delta y \to 0$. However, each iteration takes longer, because of the implicit tridiagonal solution itself. Because of the simplicity and effectiveness, the point SOR method has been the most popular of the iterative methods for solving the Poisson equation in computational fluid dynamics problems.

A simple modification of the SOR that is very easy to program is to use the Liebmann method for only the first iteration by setting $\omega = 1$ thereafter, $\omega = \omega_0$ is used.

5.7 Tactics and Strategy for 'SOR'

The ω_0 given by Eqs. (5.20) - (5.21) is optimum in the sense of a strategic or long-range manoeuvre. That is the total error $\sum_{i,j} e^k_{i,j} \to 0$ asymptotically (as $k \to \infty$) the fastest for $\omega = \omega_0$. But for a finite k the optimum may be something less than ω_0, depending on the initial error distribution (initial guess). In fact, if we limit $k = 1$, then $\omega = 1$ (Liebmann's method) gives the greatest reduction in total error in a single sweep; thus Liebmann's method is the optimum tactical or short-range manoeuvre. Numerical experiments by Roache (1971) in a 21×21 grid indicated that the advantage of $\omega = 1$ over $\omega = \omega_0$, based on total error, could persist to $k = 6$ or 8. Further $\omega = \omega_0$, gave errors highly skewed along the diagonal, compared with the results of $\omega = 1$. If the direction were changed at each iteration (say, from $i \uparrow i$, $j \uparrow$, to $i \downarrow$, $j \downarrow$, or others), convergence in the average is not affected, but the skewness is decreased. The more efficient iterative methods do not produce a truly symmetric calculation for symmetric boundary conditions, a situation which may be viewed as another *behavioural error*. Note that the crude Richardson method has the advantage of giving truly symmetric calculations.

5.8 Other Iterative Methods

Alternating direction implicit (ADI) Methods

In analogy to Eq. (5.4) we write

$$\psi^{k+1/2} = \psi^k + \frac{\Delta t}{2}\left[\delta_{xx}\psi^{k+1/2} + \delta_{yy}\psi^k - \zeta\right] \tag{5.22}$$

$$\psi^{k+1} = \psi^{k+1/2} + \frac{\Delta t}{2}\left[\delta_{xx}\psi^{k+1/2} + \delta_{yy}\psi^{k+1} - \zeta\right]. \tag{5.23}$$

There are two fractional steps. The first equation is implicit in x with a tridiagonal form. The second is implicit in y with the same tridiagonal form. Convergence is assured, provided the same Dt is used in both half-steps. It might be thought that very large Δt would hasten the asymptotic "ime" convergence, but actually an optimum Δt exists. The optimum Δt gives convergence in slightly fewer iterations than the optimum SOR. However, since each iteration of ADI takes longer, the optimum SOR method actually takes less computer time than this *single parameter* ADI method. The real strength of the ADI method comes in choosing a sequence of iteration

parameters $\rho = 2\Delta x^2/\Delta t$. The corresponding Dt starts off large and drops off. The optimum sequence is generally not attainable. The determination of good parameters is beyond the scope of this text, the reader is referred to Wachspress (1966).

In SOR methods, the number of iterations required for convergence increases with N. For ADI methods applied to square regions, k_{max} is almost independent of N, so that for large enough N, ADI methods are preferable. Other iterative methods are the *conjugate gradient* method and the *multigrid methods*. Implementations of these methods are available as FORTRAN subprograms, e.g in the program-library ITPACK (package of iterative procedures). Because of its simplicity and acceptable convergence rate, the basic SOR method (with $\omega = 1$ for the first iteration) will probably continue to be the most popular iterative method for non-rectangular regions.

5.9 Fourier-Series Methods

The Fourier-series methods are based on the fact that an exact solution of the finite difference equation (5.3) can be expressed in terms of finite eigenfunction expansions. For example, on a rectangular region of dimensions $X \times Y$ with $M \times N$ interior points ($N = I - 2, M = J - 2$), with constant Δx and Δy, and $\psi 0$ an all boundaries, the exact solution of Eq. (5.3) can be expressed as

$$\psi_{i,j} = \sqrt{\frac{2}{N+1}} \sum_{p=1}^{N} H_{p,j} \sin \frac{p\pi x_i}{X}, \tag{5.24}$$

where $x_i = (i-1)\Delta x$. The H_{pj} are solutions, for $1 \leq p \leq N$, of the tridiagonal difference equations,

$$\frac{1}{\Delta y^2} \left(H_{p,j-1} - 2H_{p,j} + H_{p,j+1} \right) + \lambda_p H_{p,j} = V_{p,j} \tag{5.25}$$

with

$$H_{p,1} = H_{p,J} = 0 \tag{5.26}$$

and

$$V_{p,j} = \sqrt{\frac{2}{N+1}} \sum_{q=1}^{N} \zeta_{q+1,j+1} \sin \frac{q\pi p \Delta x}{X}, \tag{5.27}$$

$$\lambda_p = \frac{2}{\Delta x^2}\left(\cos\frac{p\pi\Delta x}{X} - 1\right). \tag{5.28}$$

The Fourier-series-method is reviewed by Dorr (1970). The method becomes very efficient through the use of *fast Fourier transforms* which can be applied for $N = 2^r 3^s 5^t$ with r, s, t integer. The method is simpler for periodic boundary condition.

We do note, however, that although these methods in their basic forms are quite restricted in the boundary formulation of the problem, they may be applied in a modified procedure to more general problems. Consider first the case of a rectangular region with Dirichlet condition $\psi = f(x, y)$ on the boundary, where $f \neq 0$ everywhere. An auxiliary function, ψ^1, is defined by obtaining the exact solution of $\nabla^2\psi^1 = \zeta$, with boundary conditions of $\psi^1 = 0$ everywhere. Then a second auxiliary function, ψ^{11}, is defined by obtaining the exact solution of the finite-difference Laplace equation, , with the correct boundary conditions $\psi^{11} = f(x, y)$. The exact solution is obtained by the separation of variables method of partial differential equations applied to the finite-difference equation. Then the final solution ψ is obtained, because of the linearity of the problem, by superposition. That is since $\nabla^2\psi^1 = \zeta$ and $\nabla^2\psi^{11} = 0$, then $\nabla^2(\psi^1 + \psi^{11} = \zeta$, and since $\psi^1 = 0$ and $\psi^{11} = f(x, y)$ on boundaries, then $\psi^1 + \psi^{11} = f(x, y)$. So $\psi\psi^1 + \psi^{11}$ satisfies $\nabla^2\psi = \zeta$ and $\psi = f(x, y)$ on boundaries.

If the zero-gradient Neumann boundary conditions are used, expansion in a cosine series is appropriate. The problem of non-zero boundary specification of the normal gradient $\partial\psi/\partial n = g(x, y)$ can be solved as follows. An auxiliary function ψ is written as $\psi^1 = 0$ at all internal points, $\psi^1 = g(x, y) \times \Delta n$ at the extremes $i = I$ and $j = J$, and $\psi^1 = -g(x, y) \times \Delta n$ at the extremes $i = 1$ and $j = 1$. This ψ^1 is a solution of the auxiliary discretized Poisson equation $\nabla^2\psi^1 = \zeta^1$, with on boundaries, and $\zeta^1 = 0$ everywhere except at points adjacent to boundaries, where $\zeta^1\nabla^2\psi^1 \neq 0$. At a node two positions in from a boundary, $\nabla^2\psi^1 = 0$, since $\psi^1 = 0$ at all the neighbouring points. Defining $\psi^{11} = \psi - \psi^1$ and $\zeta^{11} = \zeta - \zeta^1$, the original problem is then converted to finding the finite-difference equation solution of $\nabla^2\psi^{11} = \zeta^{11}$ with conditions $\partial\psi^{11}/\partial n = 0$ on all boundaries, which can be handled by a cosine expansion. The desired solution is then $\psi\psi^1 + \psi^{11}$.

In a similar manner, non-rectangular regions may be solved by using a rectangular mesh which overlaps the desired region. Consider the region shown in Fig. 5.2a, formed by taking a small corner off of a rectangular region. The boundary point (2,2) is not on the overlapping rectangle. Consider $\psi = 0$ on all boundaries. A first auxiliary function ψ^1, is defined by obtaining a solution to $\nabla^2\psi^1 = \zeta^1$ in the overlapping mesh with $\zeta^1_{2,2} = 0$.

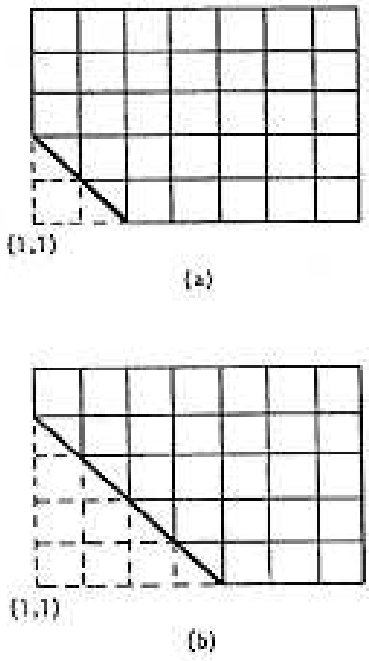

Figure 5.2: Non-rectangular regions by a Fourier series method.

A second auxiliary function, ψ^{11}, is defined by obtaining a solution to the equation $\nabla^2\psi^{11} = \zeta^{11}$, where ζ^{11} is defined as $\zeta^{11}_{2,2} = 1$, all other $\zeta^{11}_{i,j} = 0$. In electrical applications where ψ is the potential and ζ is the charge, this technique is known as the *unit charge method*. Then a linear combination is found which gives the desired $\psi_{2,2} = 0$ value without altering ζ at internal points. That is

$$\psi_{2,2} = 0 = 1 \cdot \psi^1_{2,2} + a^{11} \cdot \psi^{11}_{2,2} \tag{5.29}$$

or

$$a^{11} = -\psi^1_{2,2}/\psi^{11}_{2,2}. \tag{5.30}$$

The denominator is never equal to zero. The final solution is, by superposition,

$$\psi = \psi^1 + a^{11}\psi^{11}. \tag{5.31}$$

Although the ζ at $(2,2)$ of the composite solution is a^{11} + (true) $z_{2,2}$, the solution is not affected, since $(2,2)$ is a boundary point of the composite

problem and therefore $\zeta_{2,2}$ does not enter into that problem.

When more than one boundary point is not on the overlapping rectangle, each point requires an additional Poisson solution. In Fig. (5.2b), auxiliary solutions ψ^{11}, ψ^{111}, ψ^{1111}, are associated with $\zeta = 1$ solutions at (2,4), (3,3), (4,2), respectively. Corresponding to Eq. (5.29), the linear system defined by

$$\begin{aligned}
\psi_{2,4} &= 0 = \psi^1_{2,4} + a \cdot \psi^{11}_{2,4} + b \cdot \psi^{111}_{2,4} + c \cdot \psi^{1111}_{2,4} \\
\psi_{3,3} &= 0 = \psi^1_{3,3} + a \cdot \psi^{11}_{3,3} + b \cdot \psi^{111}_{3,3} + c \cdot \psi^{1111}_{3,3} \\
\psi_{4,2} &= 0 = \psi^1_{4,2} + a \cdot \psi^{11}_{4,2} + b \cdot \psi^{111}_{4,2} + c \cdot \psi^{1111}_{4,2}
\end{aligned} \tag{5.32}$$

must be solved for a, b, c, establishing an *influence coefficient matrix* to zero the boundary values for later solution with new ζ. For m boundary points, m auxiliary Poisson equations must be solved, and a m-order linear system like (5.32) must be solved by Gaussian elimination. However this need be done only for the first solution in a family of problems with different ζ, all in the same mesh.

5.10 Neumann Boundary conditions in SOR Method

Neumann conditions require two special formulations: first the incorporation of the gradient condition in the SOR method. An obvious method of solution is to sweep the mesh for the new $(k+1)$ iterative estimates at all interior points, and then set new iterative $(k+1)$ estimates for the boundary values from the known slope and the newly calculated adjacent interior points. Assume (i, jc) is on a boundary, then for a point using (5.19), we would have the following. During interior sweep:

$$\begin{aligned}
&\psi^{k+1}_{i,jc+1} = \psi^k_{i,jc+1} + \tfrac{\omega}{2(1+\beta^2)} \\
&\left[\psi^k_{i+1,jc+1} + \psi^{k+1}_{i-1,jc+1} + \beta^2 \psi^k_{i,jc+2} + \beta^2 \underbrace{\psi^k_{i,jc}}_{\text{boundary}} - \Delta x^2 \zeta_{i,jc+1} - 2(1+\beta^2)\psi^k_{i,jc+1}\right]
\end{aligned} \tag{5.33}$$

boundary-value determination:

$$\psi^{k+1}_{i,jc} = \psi^{k+1}_{i,jc+1} - \frac{\partial \psi}{\partial n} \cdot \Delta y. \tag{5.34}$$

This plausible method does not converge. The solution drifts, slowly but endlessly. The meteorologist Miyakoda (1962) recommends that the derivative

boundary condition is incorporated directly into the SOR difference scheme at interior points adjacent to the boundaries. Thus, an equation of the form (5.19) is used only at interior points more than one node removed from the boundary. At points adjacent to the boundary, Eq. (5.33) is replaced by

$$\psi^{k+1}_{i,jc+1} = \psi^{k}_{i,jc+1} + \tfrac{\omega}{2(1+\beta^2)} \left[\psi^{k}_{i+1,jc+1} + \psi^{k+1}_{i-1,jc+1} + \beta^2\psi^{k}_{i,jc+2} + \beta^2\left(\psi^{k+1}_{i,jc+1} - \left(\tfrac{\partial\psi}{\partial n}\right)_{i,jc}\cdot\Delta y\right) - \Delta x^2\zeta_{i,jc+1} - 2(1+\beta^2)\psi^{k}_{i,jc+1}\right] \tag{5.35}$$

which is solved algebraically for appearing on both sides. After convergence is completed, the final boundary values may be computed from Eq. (5.34). Equation (5.35) differs from the cyclic use of (5.33) and (5.34) only in the time level of the term

$$\beta^2\left(\psi^{k+1}_{i,jc+1} - \left(\frac{\partial\psi}{\partial n}\right)_{i,jc}\cdot\Delta y\right).$$

The second special formulation for Neumann conditions is that the boundary gradients should be compatible with the source term, this applies not only to SOR but also to direct methods. By Greens theorem, a continuum solution to $\nabla^2\psi - \zeta = 0$ over the area R exists only if

$$E = \int_R \varsigma\cdot dR - \oint_{\partial R} \tfrac{\partial\psi}{\partial n}\cdot dl = 0.$$

Because of truncation error, the boundary values usually will fail to meet this constraint, causing a slow divergence of SOR (and other iterative methods) iteration. Miyakoda (1962) recommends that $\partial\psi/\partial n$ be determined to meet this constraint. Usually, however, the discretized value of E is computed, and then the modified equation $\nabla^2\psi = \zeta - E/R$ is solved.

Provided that $\partial\psi/\partial n$ is determined to second-order accuracy, the overall method appears to be second-order accurate also. Experiments (Miyakoda, 1962) indicate that the optimum ω_0 is increased by the Neumann conditions. With Neumann conditions if $\psi_{i,j}$ is a solution, then $\psi_{i,j} + C$, where C is a constant, is also a solution. The particular solution is chosen by specifying ψ at one point. This one point may be omitted from the iteration scheme.

5.11 Example of Relaxation Methods

We demonstrate the convergence speed of relaxation methods using a simple test case. We prescribe the solution $\psi^{\infty}_{i,j}$ everywhere in a domain of 41 by 21

grid points. The initial distribution is displayed in Fig. 5.3. Then the right hand side of the Poisson equation $\zeta_{i,j}$ is calculated from (5.3). We compare the performance of Richardson's method (5.12), Liebmann's method (5.15) or (5.18) and the SOR method (5.19) in obtaining the diskrete solution of the Poisson-equation (5.3) for after k iterations with given $\zeta_{i,j}$. We can then compare the numerical solutions with , considering the maximum error $\varepsilon_{\max} = \underset{i,j}{Max}\left(\left|\psi^{\infty}_{i,j} - \psi^{k}_{i,j}\right|\right)$, the mean error

$$\varepsilon_{mean} = \frac{1}{I \times J} \sum_{j=1}^{J} \sum_{i=1}^{I} \left(\psi^{\infty}_{i,j} - \psi^{k}_{i,j}\right),$$

or the standard deviation

$$\sigma = \left(\frac{1}{I \times J - 1} \sum_{j=1}^{J} \sum_{i=1}^{I} \left(\psi^{\infty}_{i,j} - \psi^{k}_{i,j}\right)^2 \right)^{1/2} .$$

Figure 5.3 shows the analytic solution and the convergence of Richardson's method, Liebmann's method and the SOR method after $k = 25$ and $k = 100$ iterations. The accelerated SOR with $\omega = 1$ for $k = 1$ (which corresponds with Liebmann's method for the first iteration only) can be shown theoretically to converge faster than the original SOR method. In our example

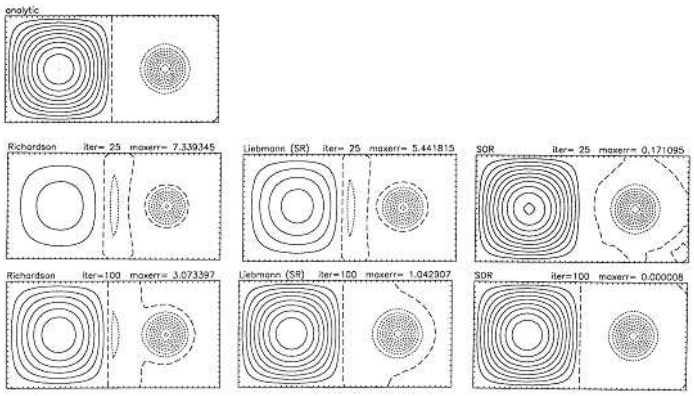

Figure 5.3: Convergence of relaxation methods. a) the true solution. Middle panels: 25 iterations. b) Richardson's method. c) Liebmann's method. d) SOR method. Bottom panels: after 100 iterations. The maximum error is displayed on top of the panels.

however, this acceleration has no (or negligible) beneficial effect. This is obvious in Fig. 5.4 where the maximum error is graphed versus the iteration

cycles. The accelerated SOR is indicated by a dashed line. The SOR converges much faster than the other two methods up to 70 iterations roughly. Then we have reached machine accuracy of our computer, smaller corrections are indistinguishable. On a computer with higher precision (increased word length) the error of the SOR would decrease further with the same slope. When the analytic distribution is changed to a random number distribution prescribed on every grid point, the convergence is similar except that the accelerated SOR starts off with slightly less maximum error that persists up to the 10th iteration.

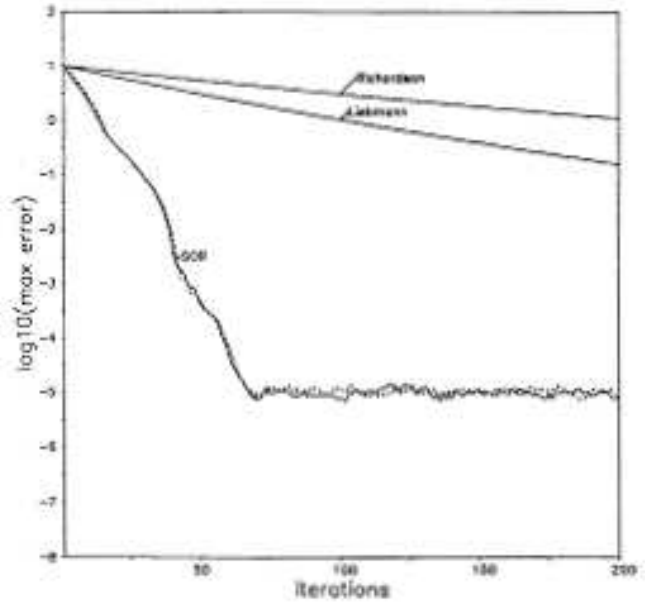

Figure 5.4: Maximum error of relaxation methods versus iteration cycles. The dashed line corresponds with the accelerated SOR method. It has no (negligible) beneficial effect. The error of SOR does not drop below 10-5 because of limited machine accuracy of our computer.

5.12 General elliptic equations

The most general two-dimensional elliptic equations that can be solved by an efficient direct method ("fast elliptic solver" sometimes simply referred to as

"fast Poisson solver") in the FISHPACK fortran subprogram (efficient FORTRAN subprograms for the solution of elliptic partial differential equations) package (subroutine BLKTRI) solves the difference equivalent corresponding with the elliptic differential equation

$$a(x)\frac{\partial^2\psi}{\partial x^2} + b(x)\frac{\partial\psi}{\partial x} + c(x)\psi + d(y)\frac{\partial^2\psi}{\partial y^2} + e(y)\frac{\partial\psi}{\partial y} + f(y)\psi = g(x,y)$$

More general forms are possible to be solved by recursive direct methods where coefficients may vary with x and y (however not in FISHPACK). Three-dimensional elliptic partial differential equations may be reduced in dimension to two-dimensional slices by application of the Fourier-series method. Fast elliptic solvers may be useful for the solution of any elliptic partial differential equation by use of an iteration. Let us assume that we want to solve the general elliptic equation $\Lambda[\psi] = RHS$, where $\Lambda[\psi]$ is a general not necessary linear elliptic operator (more precisely, its difference-equation equivalent) and RHS is the right hand side. Then we chose the closest available operator (fast elliptic solver) $L[\psi] = RHS$ (e.g the one from BLKTRI) and establish an iteration

$$\Lambda\left[\psi^{k+1}\right] = L\left[\psi^{k}\right] + \omega\left(\Lambda\left[\psi^{k}\right] - RHS\right),$$

where ω is an over-relaxation factor (typical $\omega \approx$ 1-1.3). Without proof it is evident that if the iteration has converged $\psi^{k+1} = \psi^k$, ψ^{k+1} is a solution of $\Lambda[\psi^{k+1}] = RHS$ as desired. This approach works for all kind of weakly nonlinear Operators $\Lambda[\psi]$. These kind of problems normally arises when coordinate transformations are introduced (e.g. bottom topography). Typically there are only a few iterations necessary to obtain convergence in case of mildly nonlinear problems.

Exercises

5.1 Write (5.3) as a matrix equation $\mathbf{A}\psi = \zeta$ for $I = 4$, $J = 3$. Explain the band-structure of matrix $\mathbf{A}$. Assume that $\psi_{ij} = 0$ at the boundaries of the domain.

5.2 Show that the stability restriction on Eq. (5.10) is $\Delta t \leq \Delta^2/4$.

5.3 Write Eqs.(5.22)- (5.23) with subscripts and use

$$\rho = \frac{2\Delta x^2}{\Delta t}.$$

Collect all new values on the left-hand-side, so that the tridiagonal structure is obvious.

5.4 Show that (5.24)-(5.28) solve the difference equation (5.3). Consider the limiting case $\Delta x \to 0$.

5.5 Find suitable expressions that replace (5.24)-(5.28) in case of periodic boundaries in the x-direction.

Chapter 6

Nonoscillatory Advection Schemes

One complaint against traditional numerical methods for advection transport is that they usually lead to negative values in the positive-definite scalar fields (e.g., water substance, chemical tracers) or, in more general terms, to spurious numerical over- and under- shoots in regions of steep gradients in these variables. Traditional techniques that are free of these problems are only first order accurate (e.g. the upwind method). In the last two decades, advanced finite-difference methods for solving the transport problem have been designed that are essentially free of spurious oscillations characteristic of higher-order methods yet do not suffer from the excessive implicit diffusion characteristic of low order methods. Although these essentially nonlinear algorithms are far more complex than the traditional linear schemes, they offer strong (linear and nonlinear) stability, maintain steep gradients of the transported fields, and preserve the monotonicity and/or sign of the advected variable. These properties make them potentially attractive tools for applications which combine transport with processes responsible for coupling the advected variables. In atmospheric applications, methods suppressing spurious oscillations are usually considered in the context of advection transport equations and their most advertised properties are sign-preservation and overall smooth appearance of of the solutions. Complexity of atmospheric equations associated with multidimensionality, earth rotation, sphericity, orography, phase-change processes, radiation, chemical reactions, and other natural effects makes stability, second order accuracy, generality, and computational efficiency the primary concerns of atmospheric models. Although there is a wide availability of nonoscillatory techniques suitable for constant coefficient advection in one spatial dimension, there are few techniques adequate for general flows.

6.1 Computational stability of nonoscillatory advection schemes

The basic equation to be solved is the continuity equation describing the transport of a nondiffusive scalar quantity in M-dimensional space,

$$\frac{\partial \psi}{\partial t} + \sum_{I=1}^{M} \frac{\partial}{\partial x^I} \left(\psi u^I \right) = 0, \tag{6.1}$$

where $\psi = \psi \left(t, x^1, \ldots, x^M \right)$ is the nondiffusive scalar quantity $u^I = u^I \left(t, x^1, \ldots, x^M \right)$ is the I^{th} velocity component, $I = 1, \ldots, M$ and $(t, x) = \left(t, x^1, \ldots, x^M \right)$ are the time- and space-independent variables. Equation (6.1) is usually referred to as *flux form of the advection equation*. In order to compress the notation in the following numerical equations, we choose traditional n and i indices to denote, respectively, the temporal and spatial position on a uniform computational grid $(t^n, \mathbf{x}_i) = \left(n\Delta t, \; i_1 \Delta X^1, \ldots \; i_M \Delta X^M \right)$ and adopt e_I for the unity vector in the I direction. A conservative advection algorithm for integration of (6.1) may then be compactly written as:

$$\psi_i^{n+1} = \psi_i^n - \sum_{I=1}^{M} \left(F^I_{i+1/2e_I} - F^I_{i-1/2e_I} \right), \tag{6.2}$$

where $F^I_{i+1/2e_I} \Delta X^I$ is an approximation to the I component of the integrated advective flux, evaluated at $\mathbf{i} + \frac{1}{2}\mathbf{e}_I$ position on the grid based on the information provided in a local neighbourhood of the $\mathbf{i} + \frac{1}{2}\mathbf{e}_I$ grid point. Since in (6.2) the time level of the fluxes may be taken at any position, this equation represents the general form of an arbitrary finite-difference flux-form scheme. For the sake of illustration, the fluxes from selected, elementary, one-dimensional advection schemes are explicitly written as follows:

$$F_{i+1/2} = \left[\alpha^n_{i+1/2} \right]^+ \psi_i^n + \left[\alpha^n_{i+1/2} \right]^- \psi_{i+1}^n, \tag{6.3}$$

$$F_{i+1/2} = \frac{1}{2} \alpha^n_{i+1/2} \left(\psi^n_{i+1} + \psi^n_i \right) - \frac{1}{2} \left(\alpha^n_{i+1/2} \right)^2 \left(\psi^n_{i+1} - \psi^n_i \right), \tag{6.4}$$

$$F_{i+1/2} = \frac{1}{2} \alpha^{n+1/2}_{i+1/2} \left(\psi^{n+1/2}_{i+1} + \psi^{n+1/2}_i \right), \tag{6.5}$$

where $\alpha = (u\Delta t)/\Delta X$ is a local Courant number, defined on a *staggered grid*, and $[\;\;]^+$, $[\;\;]^-$ denote the non-negative- and non-positive-part operators, respectively. Equations (6.3)-(6.5) provide fluxes from, correspondingly, the first-order-accurate donor-cell (alias upwind, upstream) scheme (6.3), the second-order-accurate (for uniform flow) Lax-Wendroff scheme (6.3b), and

the second-order-accurate, centred in time and space (leapfrog) scheme (6.5). A simple, compact form of the advective fluxes in (6.3)-(6.5) is typical of elementary advection schemes.

In order to assess the computational stability of nonoscillatory schemes, consider first a *constant-sign* field $\psi(t, \mathbf{x})$ in (6.1) and an arbitrary, sign-preserving advection algorithm in (6.2). For simplicity, assume that both analytic and numerical fluxes vanish at the boundaries of a computational domain. Then the conservation form (6.2) implies

$$\underset{n}{\forall} \sum_{\mathbf{i}} \psi_{\mathbf{i}}^n = \sum_i \psi_{\mathbf{i}}^0. \tag{6.6}$$

Since the scheme preserves the sign of the transported quantity by assumption, (6.6) is equivalent to

$$\forall n \sum_{\mathbf{i}} |\psi_{\mathbf{i}}^n| \sum_{\mathbf{i}} \left|\psi_{\mathbf{i}}^0\right|. \tag{6.7}$$

Recalling that $\sum |\quad| \geq \left(\sum (\quad)^2\right)^{1/2}$, (6.7) implies that

$$\sum_{\mathbf{i}} (\psi_{\mathbf{i}}^n)^2 \leq \left(\sum_{\mathbf{i}} \left|\psi_{\mathbf{i}}^0\right| \right)^2 \equiv B, \forall n. \tag{6.8}$$

In other words, total "energy" ("entropy") of the sign-preserving solution is uniformly bounded in time, which is to say that the sign-preserving solution is computationally stable. Since for a sufficiently small time step advection schemes can be designed which are sign-preserving for arbitrary flows (e.g., Smolarkiewicz, 1984; Smolarkiewicz and Clark, 1986), the inequality (6.8) is a statement of both linear and nonlinear stability for such schemes.

The simplicity of the result in (6.8) is a direct consequence of the assumption that ψ is of a constant sign. For variable-sign fields, a similar result may be obtained by noting that every ψ may be uniquely decomposed into the nonpositive and nonnegative part, $\psi = [\psi]^+ + [\psi]^-$. Since the two parts have disjoined supports, they are independent of each other, and they both satisfy (6.1). Applying a sign-preserving advection scheme to both parts ensures uniform boundedness of both $[\psi]^+$ and $[\psi]^-$, and consequently of their sum. Such arguments can be further elaborated (over a single time step) for the transport equation with forcings and/or sources leading to the conclusion that *sign-preserving advection schemes*[1] *offer the means of controlling nonlinear stability in numerical models.* This result is rarely appreciated in the

[1]This also concerns monotonicity-preserving schemes, as every monotonicity-preserving scheme is also sign-preserving (the oppsite is not true).

meteorological literature, where sign- and/or shape-preserving schemes are usually considered in the context of their elementary (defining) properties, and where Arakawa-type schemes (Arakawa, 1966) are thought to be the obvious choice insofar as the nonlinearly stable, finite-difference advection transport algorithms are concerned.

6.2 General flux-corrected-transport (FCT) procedure

Among a variety of existing monotonicity-preserving methods, the FCT schemes originated by Boris and Book (1973), and later generalized by to fully multi-dimensional algorithms by Zalesak (1979), become perhaps the only modern nonlinear approach that has been adopted on a permanent basis in several atmospheric and oceanic research models. It is important to realize that FCT is a general concept that allows several degrees of freedom and may lead to many different schemes. A number of known nonoscillatory techniques may be viewed as either a particular realization of the FCT approach or as implementing certain conceptual elements of FCT in their design. The basic idea of the FCT approach is simple: The generic reason for the appearance of the oscillations in the numerically generated higher-order-accurate solutions to (6.1) is that the magnitude of certain fluxes is overestimated with respect to their analytic value. In contrast, the magnitude of the fluxes given by first-order-accurate schemes is underestimated, which results in monotone but heavily damped solutions (Zalesak, 1979). The FCT procedure overcomes the problem of false oscillations by imposing appropriate limits on the transport fluxes from the higher-order-accurate algorithms. Consider an arbitrary, higher-order-accurate advection algorithm for the integration of (6.1):

$$\psi_{\mathbf{i}}^{n+1} = \psi_{\mathbf{i}}^{n} - \sum_{I=1}^{M} \left(FH^{I}_{\mathbf{i}+1/2e_I} - FH^{I}_{\mathbf{i}-1/2e_I} \right). \tag{6.9}$$

The high-order FH-flux may be arbitrarily cast into the sum of the flux from a certain low-order nonoscillatory scheme and the residual, i.e.,

$$FH^{I}_{i+1/2e_I} = FL^{I}_{i+1/2e_I} + A^{I}_{i+1/2e_I}, \tag{6.10}$$

where (6.10) defines the residual A-flux, which has a sense of correcting at least the first order truncation error terms in the transport fluxes of the low-order scheme, i.e.,

$$A^{I}_{\mathbf{i}+1/2e_I} \propto \Delta t \cdot O(\Delta X, \Delta t) + HOT, \tag{6.11}$$

where HOT has the usual meaning of "higher order terms". Because of this compensation of the leading truncation-error term in a low-order scheme, the A-flux is traditionally referred to as the "antidiffusive" flux. Using (6.10) in (6.9) results in

$$\psi_{\mathbf{i}}^{n+1} = \Psi_{\mathbf{i}}^{n+1} - \sum_{I=1}^{M} \left(A_{\mathbf{i}+1/2e_I}^{I} - A_{\mathbf{i}-1/2e_I}^{I} \right) \tag{6.12}$$

where 'ψ' denotes the solution given by the low-order scheme, which by assumption satisfies

$$\psi_{\mathbf{i}}^{MAX} \geq \Psi_{\mathbf{i}}^{n+1} \geq \psi_{\mathbf{i}}^{MIN}, \tag{6.13}$$

where $\psi_{\mathbf{i}}^{MAX}$ and $\psi_{\mathbf{i}}^{MIN}$ are yet unspecified maximal and minimal values of the scalar within the the $\mathbf{i}^{\text{th}}$ grid box that achieve the monotonicity of the scheme. Their explicit form will be discussed later in this section. Inasmuch as $\Psi_{\mathbf{i}}^{n+1}$ preserves the monotone character of the transported field [by means of (6.13)], the eventual oscillatory behaviour in $\psi_{\mathbf{i}}^{n+1}$ comes from overestimating the magnitude of certain A-fluxes in (6.12). Thus to ensure ripple-free solutions it is sufficient to appropriately limit A-fluxes such that

$$\overline{A}_{\mathbf{i}+1/2e_I}^{I} = C_{\mathbf{i}+1/2e_I}^{I} \cdot A_{\mathbf{i}+1/2e_I}^{I}, \tag{6.14}$$

where C-coefficients, that in general are functions of the low- and high-order solutions on the grid, are determined from the set of constraints

$$0 \leq C_{\mathbf{i}+1/2e_I}^{I} \leq 1 \tag{6.15}$$

and

$$\psi_{\mathbf{i}}^{MAX} \geq \overline{\psi}_{\mathbf{i}}^{n+1} = \Psi_{\mathbf{i}}^{n+1} - \sum_{I=1}^{M} \left(\overline{A}_{\mathbf{i}+1/2e_I}^{I} - \overline{A}_{\mathbf{i}-1/2e_I}^{I} \right) \geq \psi_{\mathbf{i}}^{MIN}. \tag{6.16}$$

When $C_{\mathrm{i}+\frac{1}{2}\mathbf{e}_{\mathrm{I}}}^{\mathrm{I}}$ is equal to zero or unity the resulting transport flux in (6.16) becomes or respectively. The assumed convergence of the low-order schemes involved in (6.10) together with (6.11), (6.14), and (6.15) ensure the convergence of the $\overline{\psi}$-solutions in (6.16) as ΔX, $\Delta t \to 0$.

The constraints in (6.15) and (6.16) allow one to derive formally the explicit form of the C-coefficients and, consequently, the explicit form of the limited anti-diffusive fluxes in (6.14). The derivation provides maximized $\overline{A}$-fluxes in (6.14) satisfying constraints (6.15) and (6.16):

$$\overline{A}^I_{i_{+1/2\mathbf{e}_I}} = \min\left(1, \beta^{\downarrow}_{\mathbf{i}}, \beta^{\uparrow}_{\mathbf{i}+1/2\mathbf{e}_I}\right)\left[A^I_{\mathbf{i}+1/2\mathbf{e}_I}\right]^+ + \min\left(1, \beta^{\uparrow}_{\mathbf{i}}, \beta^{\downarrow}_{\mathbf{i}+1/2\mathbf{e}_I}\right)\left[A^I_{\mathbf{i}+1/2\mathbf{e}_I}\right]^-, \tag{6.17}$$

where

$$\beta^{\uparrow}_{\mathbf{i}} = \frac{\psi^{MAX}_{\mathbf{i}} - \Psi^{n+1}_{\mathbf{i}}}{A^{IN}_{\mathbf{i}} + \varepsilon}, \tag{6.18}$$

$$\beta^{\downarrow}_{\mathbf{i}} = \frac{\Psi^{n+1}_{\mathbf{i}} - \psi^{MIN}_{\mathbf{i}}}{A^{OUT}_{\mathbf{i}} + \varepsilon}, \tag{6.19}$$

and $A^{IN}_{\mathbf{i}}$, $A^{\mathrm{OUT}}_{\mathbf{i}}$ are the absolute values of the total incoming and outgoing A-fluxes from the $\mathbf{i}^{\mathrm{th}}$ grid box,

$$A^{IN}_{\mathbf{i}} = \sum_{I=1}^{M}\left(\left[A^I_{\mathbf{i}-1/2\mathbf{e}_I}\right]^+ - \left[A^I_{\mathbf{i}+1/2\mathbf{e}_I}\right]^-\right),$$

and

$$A^{OUT}_{i} = \sum_{I=1}^{M}\left(\left[A^I_{\mathbf{i}+1/2\mathbf{e}_I}\right]^+ - \left[A^I_{\mathbf{i}-1/2\mathbf{e}_I}\right]^-\right),$$

respectively. ε is a small value, e.g. $\sim 10^{-15}$, which has been introduced herein to allow for efficient coding of β-ratios when $A^{IN}_{\mathbf{i}}$ or $A^{\mathrm{OUT}}_{\mathbf{i}}$ vanish. Equations (6.16), (6.17), (6.10), and Eqs. (6.18), (6.19) constitute a general, arbitrary dimensional form of the FCT-algorithm discussed by Zalesak (1979). The arbitrary dimensionality of this procedure contrasts with alternate-direction approach utilized by most other monotone schemes. In order to determine $\beta^{\uparrow}_{\mathbf{i}}$ and $\beta^{\downarrow}_{\mathbf{i}}$ uniquely, one must specify the limiter $\psi^{\mathrm{MAX}}_{\mathbf{i}}$, $\psi^{\mathrm{MIN}}_{\mathbf{i}}$ in (6.18), (6.19). The simple standard limiter (Zalesak, 1979) is

$$\psi^{MAX}_{\mathbf{i}} = \max_I\left(\psi^n_{\mathbf{i}-\mathbf{e}_I}, \psi^n_{\mathbf{i}}, \psi^n_{\mathbf{i}+\mathbf{e}_I}, \Psi^{n+1}_{\mathbf{i}-\mathbf{e}_I}, \Psi^{n+1}_{\mathbf{i}}, \Psi^{n+1}_{i+\mathbf{e}_I}\right) \tag{6.20}$$

$$\psi^{MAX}_{\mathbf{i}} = \max_I\left(\psi^n_{\mathbf{i}-\mathbf{e}_I}, \psi^n_{\mathbf{i}}, \psi^n_{\mathbf{i}+\mathbf{e}_I}, \Psi^{n+1}_{\mathbf{i}-\mathbf{e}_I}, \Psi^{n+1}_{\mathbf{i}}, \Psi^{n+1}_{\mathbf{i}+\mathbf{e}_I}\right). \tag{6.21}$$

The low-order, nonoscillatory Ψ-solutions appearing in Eqs. (6.20), (6.21) constitute the original limiter of Boris and Book (1973). The limiter effectively prevents development of spurious oscillations in an arbitrary flow field. Zalesak (1979) extended the original limiter onto the local extrema of the solution at the previous time step. The goal of this extension is to improve the predictions in incompressible flows where the only extrema allowed in an arbitrary grid point are those that were present in its immediate

environment (determined by the CFL stability criteria) at the previous time step; in principle, i.e., with accuracy to round-off errors, the original Boris and Book limiter is redundant in incompressible flows. Note that the limiters (6.20), (6.21) ensure uniform boundedness of the solution, providing uniformly bounded ψ. Since uniform boundedness of the low-order solutions is easily achievable, a nonlinear stability of the FCT approximations is assured essentially by design. Note also that the FCT limiters in (6.20), (6.21) impose tighter bounds on the finite-difference solutions than the "energy" ("entropy") constraint (6.8) of sign-preserving schemes considered in the preceding section.

There are several degrees of freedom in the approach outlined. First, the constraints (6.15) and (6.16) can be supplemented with some additional conditions which may need to be satisfied by a final approximation. Second, so long as (6.13) holds, the limiters themselves may be, in principle, arbitrary. This emphasizes that a nonoscillatory character of the solutions is understood in a relative sense with respect to ψ and the limiters, and that the limiters essentially define "monotonicity" of the resulting FCT scheme (for instance, selecting *weaker* limiters $\psi_{\mathbf{i}}^{MAX} = \infty$ and $\psi_{\mathbf{i}}^{MIN} = 0$ leads to positive, but apparently oscillatory, advection schemes). This degree of freedom is particularly useful where synchronized limiting of a system of equations is concerned. Finally, the third degree of freedom is in the choice of the high- and low-order schemes mixed by the FCT procedure. As there are no essential restrictions imposed on the mixed schemes, a variety of FCT algorithms may be designed. The formalism outlined does not even necessarily require mixing of low- and high-order schemes but provides a general framework for mixing *any* two algorithms attractive for a particular problem at hand. Insofar as atmospheric applications are concerned, y evaluated with the first-order-accurate upwind scheme is usually considered as a low-order solution , whereas the leapfrog advection schemes are usually selected (following Zalesak, 1979) for the high-order-accurate algorithms. However, as the resulting FCT algorithm mixes the two schemes of different distributions of the truncation errors, it is not necessarily optimal in terms of overall accuracy. The leapfrog schemes yield excellent amplitude behaviour but large phase errors, whereas the upwind scheme suffers from large amplitude errors but relatively small phase errors. Because the low- and high-order solutions are shifted in phase, the FCT mixing of the two solutions (which should eliminate dispersive ripples essentially without degrading the accuracy of the high-order solutions) is inefficient.

The latter point is illustrated in Fig. 6.1, which shows the result of one-dimensional uniform advection of the irregular signal (heavy solid line). The dashed line corresponds to the leapfrog solution, whereas the thin solid

line displays the solution obtained with the upwind/leapfrog FCT scheme. Although the FCT procedure efficiently removes the dispersive ripples, the amplitudes of the initial perturbations are severly damped at the cost of improving the phase error. The overall accuracy of the FCT solutions may be improved by employing for the high-order scheme an algorithm of phase-error characteristics similar to those of the low-order scheme. The nonlinear

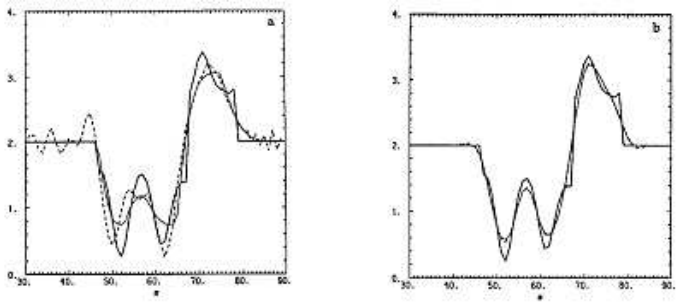

Figure 6.1: Uniform advection of the irregular signal (heavy solid line) with oscillatory scheme (dashed line) and with their non-oscillatory FCT versions (thin solid lines).

MPDATA methods (Smolarkiewicz and Clark, 1986) are an excellent choice for this pupose as they yield phase-errors similar to those in the upwind-scheme. The MPDATA schemes, whose design is conceptually unrelated to the FCT approach, are at least second-order-accurate for arbitrary flows and already preserve the sign of the transported field (therefore, for many practical applications they do not require additional FCT enhancement). Figure 6.1b shows the MPDATA solution (dashed line) and the fully nonoscillatory solution obtained with the FCT version of this scheme (thin solid line), as well as the analytic solution (heavy solid line). Comparison of the three curves shows that the primary effect of the FCT modification is to remove the overshoot and the undershoot present in the original MPDATA solution.

6.3 Semi-Lagrangian approximations for atmospheric flows using nonoscillatory advection schemes

Having established general FCT procedure and its degrees of freedom, we narrow the discussion here to a special class of applications which emphasizes the utility and importance of all those one-dimensional nonoscillatory

schemes which are suitable for integrating merely a constant-coefficient advection equation. Although such schemes are not particularly competitive in applications addressing complex natural problems in Eulerian formulation, they become accurate and powerful tools as nonoscillatory interpolation operators inherent in semi-Lagrangian representation of equations governing atmospheric phenomena.

The notion of a semi-Lagrangian approach is simple. It originates with a Lagrangian form of the evolution equation for a fluid variable ψ

$$\frac{d\psi}{dt} = R \tag{6.22}$$

and the particular interpretation of its integral

$$\psi\left(\mathbf{x}_i, t\right) = \psi\left(\mathbf{x}_0, t_0\right) + \int\limits_T R \cdot dt \tag{6.23}$$

where $(\mathbf{x_i}, t)$ and $(\mathbf{x_0}, t_0)$ are the two points connected by a parcel's trajectory T, and R combines all sources and forces. Assuming that $(\mathbf{x_i}, t)$ represents points of a regular mesh, one computes backward-in-time the trajectories arriving at $(\mathbf{x_i}, t)$ and finds their departure points $(\mathbf{x}_0, t_0)$. Since the departure points do not necessarily coincide with the points of the mesh, information therein is not immediately available, and one has to invoke an *interpolation procedure* to determine $\psi(\mathbf{x_0}, t_0)$. In essence, similar action must be taken with respect to the source term, with details depending upon a selected method of the integral's approximation and the particular problem at hand. The advertised advantages of semi-Lagrangian techniques are a circumvention of the CFL stability condition (a typical restriction of Eulerian methods) and the convenience of a regular-mesh discretization (contrasting with purely Langrangian techniques). A variety of semi-Lagrangian methods has been designed exploiting different interpolation techniques and different discretizations (see Staniforth and Cˆté, 1991 for a review).

Smolarkiewicz and Rasch (1990) suggested that semi-Lagrangian methods represent a special case of a more general approach. Referring to elementary properties of differential forms and viewing a ψ variable on the time-space continuum, one can relate two values of the variable at any two points of the continuum through Stokes' theorem, gaining thereby a large degree of freedom associated with the selection of points and connecting contours. Selecting contours whose elements coincide with parcels' trajectories leads to a variety of semi-Lagrangian formulations. Moreover, this approach documents that interpolation procedures - inherent to semi-Lagrangian approximations - may be replaced with Eulerian algorithms suitable for integrating constant-coefficient advection equations while retaining their formal accuracy and such

attractive properties as monotonicity and/or sign preservation. In order to show this latter point (which is of particular interest to applications), consider the following interpolation problem: let us assume that at an arbitrary instance t_0 (for conciseness, the dependence of ψ on t_0 will be dropped in the following discussion) a sufficiently continuous field ψ is known a priori in $\mathbf{x}_i$ points of a uniform mesh in $\mathbf{R}^M$. For the sake of clarity, we restrict ourselves to regular meshes with a constant grid interval $\Delta\mathbf{X} = (\Delta X^1, \ldots, \Delta X^M)$ such that $\mathbf{x}_i = \mathbf{i} \times \Delta X$. The problem addressed is to provide an estimate of ψ, with certain desired accuracy, in a point of interest $\mathbf{x}_0$ and provide adequate estimates of the derivatives using information available on the grid, or equivalently, to evaluate ψ at $\mathbf{x}_0$ based on the assumption that ψ fits a p-th order Lagrangian polynomial in between the grid points. We consider here, however, an alternate approach (Smolarkiewicz and Grell, 1991).

As a consequence of the Stokes' theorem,

$$\psi(\mathbf{x}_0) - \psi(\mathbf{x}_i) = \int_c d\mathbf{x} \cdot \nabla\psi(\mathbf{x}) \tag{6.24}$$

where C denotes an arbitrary contour connecting the point of interest $\mathbf{x}_0$ with an arbitrary $\mathbf{x}_i$ of the computational grid. Exploiting the arbitrariness of the contour selection in (6.24), we choose for C a line segment of the parametric equation

$$x(\mathbf{x_i}, \tau) = -(\mathbf{x_i} - \mathbf{x}_0)\tau + \mathbf{x_i} \tag{6.25}$$

where the parameter $\tau \in [0, 1]$. Implementing (6.25) in (6.24) gives

$$\phi(\mathbf{x}_i, 1) = \phi(\mathbf{x}_i, 0) - \int_0^1 (\mathbf{x}_i - \mathbf{x}_0) \cdot \nabla\phi(\mathbf{x}_i, \tau) d\tau, \tag{6.26}$$

where

$$\phi(\mathbf{x_i}, 1) = \phi(\mathbf{x}_i, 0) - \int_0^1 (\mathbf{x}_i - \mathbf{x}_0) \cdot \nabla\phi(\mathbf{x}_i, \tau) d\tau. \tag{6.27}$$

Since for fixed $\mathbf{x}_0$ and $\mathbf{x}_i$, the first element of the scalar product appearing under the integral in (6.26) is constant, (6.26) may be rewritten as

$$\phi(\mathbf{x}_i, 1) = \phi(\mathbf{x}_i, 0) - \int_0^1 \nabla \cdot (\mathbf{U}\phi(\mathbf{x}_i, \tau) d\tau, \tag{6.28}$$

where

$$\mathbf{U} = \mathbf{x}_i - \mathbf{x}_0. \tag{6.29}$$

Equation (6.28) represents the integral of the equation

$$\frac{\partial \phi}{\partial \tau} + \nabla \cdot (\mathbf{U}\phi) = 0, \tag{6.30}$$

over the τ interval $[0, 1]$ at $\mathbf{x}_i$ grid-point. In other words, (6.28) is a formal solution of the advection equation (6.30) in which the free parameter τ plays the role of a time independent variable, and the vector $\mathbf{U}$ defined in (6.29) plays the role of a velocity field. Therefore, the interpolation problem has been expressed as the equivalent advection problem, and (6.28) may be approximated using, in principle, any known advection algorithm. Among the variety of available advection algorithms, the foreward-in-time schemes are the most attractive for applications, as they require information only from $x(\mathbf{x}_i, \tau = 0)$ points in (6.25), where $\phi(\mathbf{x}_i, 0) = \psi(\mathbf{x}_i)$ are known. Such an approximation may be compactly written as

$$\phi^1_{\mathbf{i}} = \phi^0_{\mathbf{i}} - AS_{\mathbf{i}}(\phi^0, \beta), \tag{6.31}$$

where

$$\boldsymbol{\beta} = \frac{\mathbf{U}}{\Delta \mathbf{X}} = \left(\frac{U^1}{\Delta X^1}, ..., \frac{U^M}{\Delta X^M} \right)$$

is an effective Courant number vector (recall that $\Delta\tau = 1$), and the finite difference flux-divergence operator AS identifies an advection scheme defined by its particular dependence on the values of ϕ and β available on the mesh in a neighbourhood of the $\mathbf{x}_i$ grid point. The truncation error in the approximation in (6.31) represents the error of estimating ψ at the point of interest $\mathbf{x}_0$. Choosing an arbitrary grid point $\mathbf{x}_i$ that appears in the definition of the effective advecting velocity (6.29), as the closest grid point to $\mathbf{x}_0$,

$$\mathbf{x}_i = [\mathbf{x}_0] = (NINT(x^1_0/\Delta x^1) \cdot \Delta x^1, \ldots, NINT(x^M_0/\Delta x^M) \cdot \Delta x^M) \tag{6.32}$$

(where (NINT denotes the nearest integer value), and using definitions (6.29) and (6.27), the resulting interpolation algorithm may be compactly written as

$$\psi(\mathbf{x}_0) = \psi\left([\mathbf{x}_0]\right) - AS_{[\mathbf{x}_0]}\left(\psi, \frac{[\mathbf{x}_0] - \mathbf{x}_0}{\Delta \mathbf{x}}\right). \tag{6.33}$$

There is no gain in using (6.33) with most linear forward-in-time advection algorithms, however, where preservation of the monotonicity and sign of the interpolated variable is concerned, (6.33) becomes a useful tool as it allows to implement advanced monotone and sign-preserving advection

schemes. The simplicity and efficiency of (6.33) is a consequence of the constancy of the effective Courant number in (6.33), which allows for straightforward, alternate-direction (time-split) applications of one-dimensional advection schemes without degrading the formal accuracy of their constant coefficient limit. In contrast to fully multidimensional algorithms, the time-split approximations employing one-dimensional advection schemes are simple and versatile. Since

$$\underset{I}{\forall} -\tfrac{1}{2} \leq \frac{\left[\mathbf{x}_0^I\right] - \mathbf{x}_0^I}{\Delta \mathbf{x}^{\mathrm{I}}} \leq \tfrac{1}{2} \tag{6.34}$$

the computational stability is ensured for all one-dimensional forward-in-time advection schemes, securing thereby the stability of a time-split algorithm for arbitrary M.

A paradigm for the applicability: The integration of two-dimensional Boussinesq-Equations

The utility of simple nonoscillatory schemes in (6.33) for semi-Lagrangian integrations is illustrated in the example of a thermal convection. The governing system of equations consists of the momentum and temperature evolution equations for an ideal, nonrotating, two-dimensional Boussinesq fluid

$$\frac{d\mathbf{v}}{dt} = -\nabla \prod + g\frac{\theta\prime}{\theta_0}\nabla z \tag{6.35}$$

$$\frac{d\theta}{dt} = 0. \tag{6.36}$$

Here, $\mathbf{v} = (u, w)$ is the velocity vector in the (x, z) vertical plane,$\overline{\prod}(z)$ is the pressure (normalized by a reference density,$\prod = p/\rho_0$), and g is the gravity. θ denotes the potential temperature of a fluid parcel, whereas $\theta' = \theta - \overline{\theta}(z)$ represents its deviation from ambient, hydrostatic profile $\overline{\theta}(z)$. $\theta_0 = \overline{\theta}(z)$ is a reference potential temperature. The prognostic equations (6.35-6.36) are accompanied by the incompressibility constraint

$$\nabla \times v = 0 \tag{6.37}$$

characteristic of the Boussinesq approximation. Free-slip, rigid-lid upper and lower boundaries, and open lateral boundaries are assumed. The adapted semi-Lagrangian approximation to Eqs. (6.35)-(6.36) and (6.37) consists of three distinct steps. First, departure points of the trajectories are evaluated using the second-order-accurate implicit midpoint rule

$$\mathbf{x}_0 = \mathbf{x}_i - \Delta v(\mathbf{x}_{\mathbf{m}}, t_{\mathbf{m}})\Delta t, \tag{6.38}$$

where the velocities at the midpoints of the trajectories $(\mathbf{x}_m, t_m)$ are predicted with the first-order-accurate integral of the momentum equations

$$\mathbf{v}(\mathbf{x_m}, t_m) + 0(\Delta t^2) = \tilde{\mathbf{v}}(\mathbf{x_0}, t_0) \equiv \mathbf{v}(\mathbf{x_0}, t_0) + \tfrac{1}{2}\Delta t \frac{d\mathbf{v}}{dt}(\mathbf{x_0}, t_0). \tag{6.39}$$

The implicit nature of the trajectory algorithm consisting of (6.38) and (6.39) requires an iterative solution; the iteration converges provided

$$B = \left|\frac{\partial \tilde{\mathbf{v}}}{\partial \mathbf{x}}\right| \Delta t < 1 \tag{6.40}$$

and one iteration (assuming the Euler backward approximation for the first guess) suffices for the second-order-accurate approximation to the departure points of the trajectories. In the second step, (6.35-6.36) are approximated assuming the trapezoidal rule

$$\int_T R dt \approx \tfrac{1}{2}\Delta t \left(R\left(\mathbf{x_i}, t_0 + \Delta t\right) + R(\mathbf{x_0}, t_0)\right) \tag{6.41}$$

for evaluating the trajectory integral (6.23); updating the temperature prior to the momentum ensures availability at t0 + Dt in the vertical equation of motion. While integrating momentum equations (6.35), pressure forces at $t_0 + \Delta t$ are yet unknown and must be determined from the incompressibility constraint (6.37). In this third step, the approximate solution to (6.35) is written as

$$\mathbf{v}(\mathbf{x_i}, t_0 + \Delta t) = \mathbf{v}^*(\mathbf{x_i}) - \tfrac{1}{2}\Delta t \, \nabla \prod(\mathbf{x_i}, t_0 + \Delta t) \tag{6.42}$$

where $\mathbf{v}^*(\mathbf{x_i})$ combines all known terms appearing on the right hand side of (6.41) (i.e., velocity, pressure gradient, and buoyancy terms evaluated at (x0, t0), as well as the buoyancy term at $(\mathbf{x_i}, t_0 + \Delta t)$. Applying the divergence (6.37) to (6.42) leads to the Poisson equation for the pressure

$$\nabla^2 \prod(\mathbf{x_i}, t + \Delta t) = \frac{2}{\Delta t} \nabla \cdot \mathbf{v}^*(\mathbf{x_i}) \tag{6.43}$$

which is solved with standard methods, subject to (Neumann) boundary conditions. Having established pressure at $t_0 + \Delta t$, (6.42) completes the procedure. The entire semi-Lagrangian algorithm for Eqs. (6.35)-(6.36) and (6.37) is implicit with respect to all ordinary differential equation integrations, ensuring thereby computational stability regardless of the time step employed; Δt is solely controlled by the convergence condition (6.40).

Figure 6.2a shows a solution to Eqs. (6.35)-(6.36) and (6.37) obtained with the above-outlined solver which employs (6.33) in all interpolations inherent in semi-Lagrangian integrations. The AS operator is from the FCT

(flux-corrected-transport) advection scheme employed in the simple example Fig. 6.1c. The experimental set-up assumes a symmetric thermal of the ini-

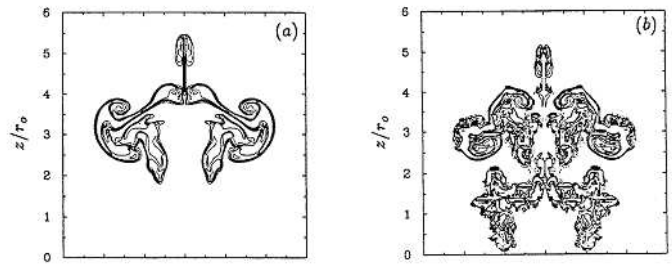

Figure 6.2: Isolines of the temperature perturbation θ of the inviscid buoyant thermal at dimensionless time $\tau \approx 7$ simulated with the semi-Lagrangian model. The contour interval is 0.2 of the initial θ amplitude: a) experiment with nonoscillatory interpolation implementing FCT-advection scheme in Fig. 6.1c; b) as in plate a), but with oscillatory interpolators.

tial radius $r_0 = 250$ m; the thermal is placed in the $8r_0 \times 8r_0$ domain resolved with 200×200 uniform grid intervals $\Delta X = \Delta Z = 0.04r_0$. The centre of the thermal is located at the vertical symmetry axis of the domain $r_0 + \Delta Z$ above the bottom boundary. The thermal has a uniform, initial temperature excess of 0.5K relative to the neutrally stratified environment $\overline{\theta}(z) = 300$ K. The field is shown after 7 characteristic time scales $T = r_0/U$, where U is the characteristic velocity scale $U = (2gBr_0)^{1/2}$ with B representing initial buoyancy of the thermal. The time step of the computations is $\Delta t \approx 0.06$T, which results in $B \approx 0.5$ in (6.40). The excellent performance of the algorithm is evident in its ability to resolve fine interfacial structures and maintain sharp across-the-interface gradients while introducing no spurious oscillations. The highly inviscid yet smooth nature of computations is clear in Fig. 6.2a; the initial thermal's core may still be identified within narrow ($\sim 2\Delta X$ wide) interfacial convolutions engulfing ambient fluid. This performance is primarily controlled by the monotonicity of the interpolation procedure in (6.33), which by design preserves local convexity/concavity and sign, and ensures uniform boundedness, of the interpolated variables. In the absence of forcing, these properties hold along parcel trajectories consistently with analytic equations. Non-monotonic schemes lead to poor results with excessive numerical noise and a spurious secondary thermal in the wake of the main structure, evident in Fig. 6.2b. Due to the Lagrangian format of the approximated equations, the monotonicity constraint is ultimately imposed along a parcel trajectory.

Exercise

6.1. Reformulate the integration of the Boussinesq fluid (6.35)-(6.36)-(6.43) for an anelastic approximation $\nabla \cdot (\rho_0(z)\,\mathbf{v}) = 0$. Energetic considerations require that the same pressure variable $\pi = p/\rho_0$ as in (6.35) must be used.

Chapter 7

Spectral Methods

7.1 Introduction

The numerical integration methods discussed thus far are based on the discrete representation of the data on a grid or mesh of points covering the space over which a prediction of the variables is desired. Then a local time derivatives of the quantities to be predicted are determined by expressing the horizontal and vertical advection terms, sources etc., in finite difference form. Finally, the time extrapolation is achieved by one of many possible algorithms, for example leapfrog. The finite-difference technique has a number of associated problems such as truncation error, linear and nonlinear instability. Despite these difficulties, the finite-difference method has been the most practical method of producing forecasts numerically from the dynamical equations.

There is another approach called the *spectral* method which avoids some of the difficulties cited previously, in particular, nonlinear instability; however the method is less versatile and the required computations are comparatively time consuming. In a general sense, the mode of representation of data depends on the nature of the data and the shape of the region over which the representation is desired. An alternative to depiction on a mesh or grid of discrete points is a representation in the form of a series of orthogonal functions. This requires the determination of the coefficients of these functions, and the representation is said to be *spectral* representation or a series expansion in *wavenumber space*. When such functions are used, the *space derivatives can be evaluated analytically, eliminating the need for approximating them with finite-differences.*

As indicated earlier the choice of orthogonal functions depends in part on the geometry of the region to be represented, and for meteorological data

a natural choice is a series of spherical harmonics. The first published work on the application of this technique to meteorological prediction is that of Silberman (1954). He considered the barotropic vorticity equation

$$\frac{\partial \zeta}{\partial t} = -\mathbf{v} \cdot \nabla(\zeta + f), \tag{7.1}$$

in spherical coordinates, where

$$\mathbf{e}_\theta \frac{1}{a}\frac{\partial}{\partial \theta} + e_\lambda \frac{1}{a \sin\theta}\frac{\partial}{\partial \lambda}$$

and $\mathbf{v} = v_\theta \mathbf{e}_\theta + v_\lambda \mathbf{e}_\lambda$

$$\frac{\partial \zeta}{\partial t} = -\frac{1}{a}\left(v_\theta \frac{\partial}{\partial \theta} + \frac{v_\lambda}{\sin\,\theta}\frac{\partial}{\partial \lambda}\right)(\zeta + 2\Omega\ \cos\,\theta), \tag{7.2}$$

$$\zeta = \frac{1}{a \sin\theta}\left[\frac{\partial}{\partial \theta}(v_\lambda \sin\theta) - \frac{\partial v_\theta}{\partial \lambda}\right], \tag{7.3}$$

where a is the earth's radius, θ is the co-latitude, λ is the longitude, and v_λ and v_θ are the velocity components in the directions of increasing λ and θ. In terms of a stream function ψ, the velocity components from $\mathbf{v} = \mathbf{k} \times \nabla\psi$ are

$$v_\lambda = \frac{1}{a}\frac{\partial \psi}{\partial \theta}, \quad v_\theta = -\frac{1}{a \sin\theta}\frac{\partial \psi}{\partial \lambda}$$

and the vorticity equation becomes with $\zeta = \nabla_s^2 \psi$

$$\nabla_s^2 \frac{\partial \psi}{\partial t} = \frac{1}{a^2 \sin\theta}\left(\frac{\partial \psi}{\partial \lambda}\frac{\partial}{\partial \theta} - \frac{\partial \psi}{\partial \theta}\frac{\partial}{\partial \lambda}\right)(\nabla_s^2 \psi + 2\Omega \cos\theta), \tag{7.4}$$

where

$$\nabla_s^2 = \frac{1}{a^2 \sin\theta}\left[\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial}{\partial \theta}\right) + \frac{1}{\sin\theta}\frac{\partial^2}{\partial \lambda^2}\right].$$

The stream function is represented in terms of spherical harmonics which are the solutions Y_n^m of the equation

$$a^2 \nabla_s^2 Y_n^m + n(n+1) Y_n^m = 0. \tag{7.5}$$

These functions are expressible in the form

$$Y_n^m = e^{im\lambda} p_n^m. \tag{7.6}$$

If depend only on q if (7.6) is substituted into (7.5), the result is the ordinary differential equation (Exercise 7.1)

$$\frac{d^2P_n^m}{d\theta^2} + \cot\theta \frac{dP_n^m}{d\theta} + \left[n(n+1) - \frac{m^2}{\sin^2\theta} \right] P_n^m = 0. \tag{7.7}$$

This is a form of the Legendre equation, and its solutions P_n^m are known as *Legendre* functions of order m and degree n. The characteristic solutions of (7.7) are expressible as a set orthonormal functions such that

$$\int_0^\pi p_n^m p_s^m \sin\theta d\theta = \delta_{ns}, \tag{7.8}$$

where δ_{ns} is the Kronecker delta,

$$\delta_{ns} = 1 \quad \text{if} \quad n = s$$

and

$$\delta_{ns} = 0 \quad \text{if} \quad n \neq s.$$

The order m may take on negative values for which the legendre function is

$$P_n^{-m} = (-1)^m P_n^m.$$

For integer values of n and m, the Legendre functions are simply polynomials, the orders of which increase with n.

At any particular time t' the stream function may be expressed as the finite sum

$$\psi_{t=t'} = a^2\Omega \sum_{n=|m|}^{n'} \sum_{m=-m'}^{m'} K_n^m Y_n^m. \tag{7.9}$$

Since the series is finite, disturbances of sufficiently small-scale are not represented, which constitutes a truncation error; however, this is not necessarily undesirable. The harmonic coefficients K_n^{-m} are complex, and the condition for a real ψ is that

$$K_n^{-m} = (-1)^m K_n^{-m*}$$

where the * denotes the complex conjugate. The stream function tendency is given by

$$\left(\frac{\partial\psi}{\partial t} \right)_{t=t'} = a^2\Omega \sum_{n|m|}^{n''} \sum_{m=-m''}^{m''} \left(\frac{dK_n^m}{dt} \right)_{t=t'} Y_n^m. \tag{7.10}$$

When the expressions for ∂
partialpsi and ψ/t are substituted into the vorticity equation (7.4) and then

this equation is multiplied by $Y_n^{-m}\sin\theta$ and integrated from 0 to 2π with respect to λ and from 0 to π with respect to θ, the result is

$$\left(\frac{dK_n^m}{dt}\right)_{t=t'} = \frac{2i\Omega m K_n(t\prime)}{n(n+1)} \tag{7.11}$$
$$+\frac{i\Omega}{2}\sum_{s=|r|}^{n'}\sum_{r=-m'}^{m'}\sum_{k=|j|}^{n'}\sum_{j=-m'}^{m'} K_k^j(t')K_s^r(t')H_{kns}^{jmr},$$

where H_{kns}^{jmr} is zero unless $j+r=m$, in which case,

$$H_{kns}^{jmr} = \frac{s(s+1)-k(k+1)}{n(n+1)}\int_0^\pi P_n^m\left(jP_k^j\frac{dP_s^r}{d\theta} - r\frac{dP_k^j}{d\theta}P_s^r\right)d\theta. \tag{7.12}$$

The quantities H_{kns}^{jmr} are called *interaction coefficients*, which are zero unless

$$k+n+s = \text{odd integer} \quad \text{and} \quad |k-s| < n < k+s.$$

Also $m''2m'$ and $n'' \le 2n'-1$. After the right side of the tendency equation (7.11) has been calculated, the future values of the expansion coefficients may be determined by extrapolating forward in time as in the finite-difference technique; for example, the leapfrog method:

$$K_n^m(t+\Delta t) = K_n^m(t-\Delta t) + 2\Delta t\frac{dK_n^m(t)}{dt}. \tag{7.13}$$

Robert (1966) pointed out that the components u and v of the wind field constitute pseudo-scalar fields on the globe, and as such are not well suited to representation in terms of scalar spectral expansions; he suggested that the variables

$$U = u\cos f \quad \text{and} \quad V = v\cos f, \tag{7.14}$$

where ϕ denotes latitude, would be more appropriate for global spectral representation.

The nondivergent flow field may be represented as usual in terms of a scalar stream function ψ as

$$\mathbf{V} = \mathbf{k}\times\nabla\psi.$$

Accordingly ζ is seen to be expressed as

$$\zeta = \mathbf{k} \cdot \nabla \times \mathbf{V} = \nabla^2 \psi$$

and the quantities U and V as

$$U = -\frac{\cos\phi}{a}\frac{\partial \psi}{\partial \phi} \quad \text{and} \quad V = \frac{1}{a}\frac{\partial \psi}{\partial \lambda} \tag{7.15}$$

The conservation of absolute vorticity may now be rewritten, with substitution of Eqs. (2) and (4), and an expansion into spherical polar coordinates as

$$\frac{\partial}{\partial t}\nabla^2\psi = -\frac{1}{a\cos^2\phi}\left[\frac{\partial}{\partial \lambda}(U\nabla^2\psi) + \cos\phi\frac{\partial}{\partial\phi}(V\nabla^2\psi)\right] - 2\Omega\frac{V}{a}, \tag{7.16}$$

where ϕ, λ, a, and Ω denote, respectively, latitude, longitude, and the earth's radius and rotation rate. The linear equations (7.15) provide specification of the diagnostic quantities U and V in terms of the prognostic ψ.

The more usual form of the barotropic vorticity equation is seen on substitution of Eqs. (7.15) into (7.16) to be

$$\frac{\partial}{\partial t}\left(\nabla^2\psi\right) = \frac{1}{a^2\cos\phi}\left[\frac{\partial\nabla^2\psi}{\partial\lambda}\frac{\partial\psi}{\partial\phi} - \frac{\partial\psi}{\partial\lambda}\frac{\partial\nabla^2\psi}{\partial\phi}\right] - \frac{2\Omega}{a^2}\frac{\partial\psi}{\partial\lambda}. \tag{7.17}$$

The calculation of the interaction coefficients is a lengthy task. An advantage of the method, however, is that nonlinear instability is avoided completely because all nonlinear interactions are computed analytically and all contributions to wave numbers outside the truncated series are automatically eliminated.

Robert (1966) proposed a modification to Silbermann's method for numerical integration of the primitive equations in which some simpler functions are substituted for the spherical harmonics. These functions are in fact the basic elements required to generate spherical harmonics, namely,

$$G_n^m(\alpha,\varphi) = e^{im\lambda}\cos^M\varphi\sin^n\varphi.$$

Here λ and φ are the longitude and latitude respectively, M is the absolute value of m and gives the number of waves along a latitude circle, and both M and n are either positive integers or zero.

7.2 An Example of the Spectral Method

As an illustration of the spectral method, we present a simple example due to Lorenz (1960). Consider the vorticity equation (7.1) applied to a plane region over which the stream function is doubly periodic, that is,

$$\psi\left(x+\frac{2\pi}{k},\, y+\frac{2\pi}{l},\, t\right)=\psi(x,y,t), \tag{7.18}$$

where k and l are constants. Thus the area is finite but unbounded, and in that respect it is analogous to the spherical earth. Note also that (7.1) applies with a constant coriolis parameter so that rotation is not excluded. Next assume that the stream function can be represented in terms of the characteristic solutions of the equation

$$\nabla^2\psi - c\psi = 0, \tag{7.19}$$

which is the analogue to (7.5). The solutions are trigonometric functions; thus ψ is expressible as a double Fourier series, which for convenience may be written as

$$\psi=\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}-\frac{1}{m^2k^2+n^2l^2}\left[A_{mn}\cos(mkx+nly)+B_{mn}\sin(mkx+nly)\right]. \tag{7.20}$$

The coefficients are as yet unknown functions of time, except for the initial values which are assumed known. It is apparent from (7.16) that the characteristic values of (7.19) are $c_{mn}=-(m^2k^2+n^2l^2)$.

For actual prediction purposes a finite series of the form (7.20) is used and the time derivatives of the coefficients must be determined from the vorticity equation. Consider a simple case where m and n take on only the values of $0, 1, -1$.

Then after combining like terms, (7.20) is expressible in the form

$$\begin{aligned}\psi \;=& -\tfrac{A_{10}}{k^2}\cos kx-\tfrac{A_{01}}{l^2}\cos ly-\tfrac{A_{11}}{k^2+l^2}\cos(kx+ly)\\ &-\tfrac{A_{1-1}}{k^2+l^2}\cos(kx-ly)-\tfrac{B_{10}}{k^2}\sin kx-\tfrac{B_{01}}{l^2}\sin ly\\ &-\tfrac{B_{11}}{k^2+l^2}\sin(kx+ly)-\tfrac{B_{1-1}}{k^2+l^2}\sin(kx-ly).\end{aligned}$$

It turns out that, if the B's are zero initially, they will remain so; also if $A_{1-1}-A_{11}$ initially, it will remain so. With these assumptions, Lorenz obtains the 'maximum specification' of the stream function for use with (7.1), namely,

$$\psi=-\frac{A}{l^2}\cos ly-\frac{F}{k^2}\cos kx-\frac{2G}{k^2+l^2}\sin ly\sin kx. \tag{7.21}$$

Substituting this streamfunction into the vorticity equation (7.1), utilizing trigonometric identities, and finally equating coefficients of like terms leads to the following differential equations for the coefficients:

$$\begin{aligned}
\frac{dA}{dt} &= -\left(\frac{1}{k^2} - \frac{1}{k^2+l^2}\right) klFG \equiv K_1 FG, \\
\frac{dF}{dt} &= \left(\frac{1}{l^2} - \frac{1}{k^2+l^2}\right) klAG \equiv K_2 AG, \\
\frac{dG}{dt} &= -\frac{1}{2}\left(\frac{1}{l^2} - \frac{1}{k^2}\right) klAF \equiv K_3 AF.
\end{aligned} \tag{7.22}$$

Note that the interaction coefficients K_1, K_2, and K_3 are constants and hence, remain the same throughout the period of integration. The set (7.22), which is analogous to (12.53), can be solved analytically; however when the spectral technique is applied, say to a hemisphere with real data, the resulting system of equations is much more complex and must be solved by numerical methods. If the leapfrog scheme is used here, the numerical integration of (7.22) would be analogous to (7.13), that is,

$$A^{n+1} = A^{n-1} + 2\Delta t K_1 F^n G^n, \tag{7.23}$$

etc., where the superscript denotes the time step. As usual, a forward step must be used for the first time step from the initial values of the coefficients A_0, F_0, and G_0. To avoid linear instability, Δt must be a fairly small fraction of the period of the most rapidly oscillating variable. Equation (7.23) and similar ones for F and G permit the calculation of future values of the coefficient of the series (7.21); thus the prediction of a stream function is achieved.

Chapter 8

Finite Element Methods

8.1 Introduction

In designing a numerical weather prediction model, one of the most fundamental aspects is the choice of discretization technique in each of the spatial dimensions. In the vertical, by far the most popular choice is the finite-difference method; while in the horizontal, both finite-difference and (especially for global models) spectral methods are widely employed. A third possibility is the finite element method.

8.2 What is the finite element method?

The essence of the finite element method can be seen by considering various ways of representing a function $f(x)$ on an interval $a \leq x \leq b$. In the *finite-difference* method the function is defined only on a set of grid points; i.e. $f(x_j)$ is defined for a set of $x_j e[a, b]$, but there is no explicit information about how the function behaves between the grid points. In the *spectral* method, on the other hand, the function is defined in terms of a finite set of basis functions:

$$f(x) = \sum_{k=0}^{N} a_k e_k(x) \tag{8.1}$$

where the basis functions $e_k(x)$ are *global* (e.g. Fourier series, or spherical harmonics for two-dimensional functions on the surface of a sphere), and the a_k are the spectral coefficients. Equation (8.1) defines $f(x)$ *everywhere* on the interval, and the representation is independent of any set of grid points.

In the *finite-element* method, the function is again in terms of a finite set of basis functions:

$$f(x) = \sum_{k=0}^{N} a_k e_k \tag{8.2}$$

but this time basis functions $e_k(x)$ are local, i.e. they are non-zero only on a small-sub-interval. As in the spectral method, the a_k are the coefficients of the basis functions, and $f(x)$ is defined everywhere; but as in the finite-difference method, there is an underlying mesh of gridpoints (nodes) involved in the representation.

To clarify this idea we consider the simple choice of linear finite elements. The interval $[a, b]$ is divided into subintervals by specifying a set of mesh points, say $x_0, x_1, \ldots, x_N$. The basis function $e_k(x)$ is defined to be 1 at x_k, decreasing linearly to zero at x_{k-1} and x_{k+1}, and to be zero outside the interval $[x_{k-1}, x_{k+1}]$. Thus the defining equations are:

$$\begin{aligned} e_k(x) &= \frac{x - x_{k-1}}{x_k - x_{k-1}}, \quad x \,\epsilon[x_{k-1},\ x_k] \\ &= \frac{x_{k+1} - x}{x_{k+1} - x_k}, \quad x \,\epsilon[\ x_k,\ x_{k+1}] \\ &= 0 \ \text{otherwise} \end{aligned}$$

The situation is illustrated in Fig. 8.1; note that the mesh may be non-uniform.

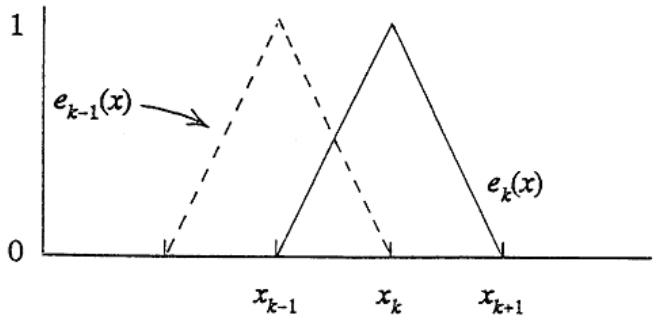

Figure 8.1: Linear finite element basis functions.

Suppose now that $f(x)$ is given at the set of gridpoints x_k, as in Fig. 8.2; how do we determine the coefficients a_k of the basis functions, in order to use

the representation given by Eq. (8.2)? The answer is almost obvious: since $e_k(x)$ is the only basis function which is non-zero at x_k, we must have

$$a_k = f(x_k), 0 \leq k \leq N.$$

Between the grid points, say between x_k and x_{k+1}, just two of the basis functions (e_k and e_{k+1}) are non-zero; since both are linear, $f(x)$ as defined by Eq. (8.2) is linear on the subinterval $[x_k, x_{k+1}]$. Thus the behaviour of $f(x)$ between grid points is determined simply by linear interpolation. At this point, it may seem that we have gained very little over the simple grid point representation of $f(x)$. The benefits of the representation in terms of linear basis functions will become clear in the next section, when we consider elementary operations with functions.

Finally in this section, we note that finite element basis functions can be even simpler (piecewise *constant* on subintervals) or more complicated (piecewise quadratic, cubic, and so on).

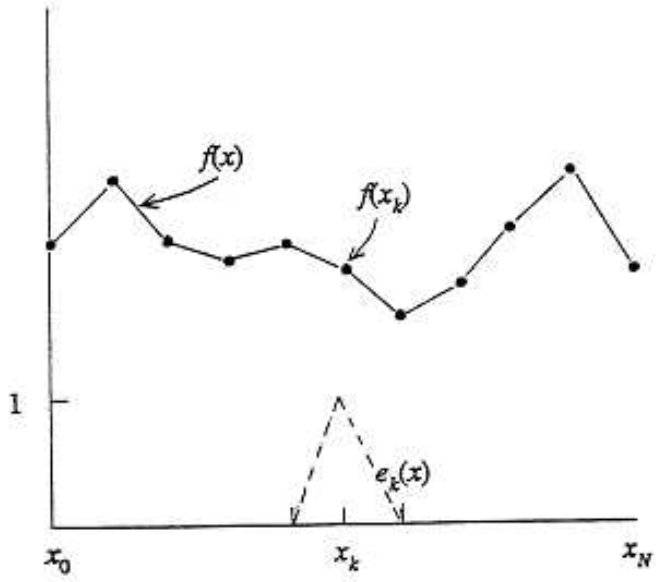

Figure 8.2: Representation of f(x).

8.3 Simple operations with linear finite elements

In this section we show how to perform three simple operations with linear elements in one dimension: differentiation, multiplication, and taking second derivatives.

8.3.1 Differentiation

Suppose we are given $u_i = u(x_i)$ at a set of nodes $x_i (0 \leq i \leq N)$, and we wish to differentiate u to find $v = u_x$. We start by expanding u and v in terms of the basis functions $e_i(x)$, as shown in section 2:

$$u(x) = \sum_{i=0}^{N} u_i e_i(x), \;\; v(x) = \sum_{i=0}^{N} v_i e_i(x),$$

where the coefficients $v_i (0 \leq i \leq N)$ are the unknowns of our problem. The series for u can be differentiated term by term, so that $v = u_x$ becomes

$$\sum_{i=0}^{N} v_i e_i(x) = \sum_{i=0}^{N} u_i e_i'(x). \tag{8.3}$$

In Eq. (8.3) we use the notation $e_i^{'}(x)$ to denote the x-derivative of the basis function, to avoid multiple subscripts. From the definition of the (piecewise linear) basis functions, $e_i^{'}(x)$ is piecewise *constant*.

The next step is to apply the Galerkin technique, namely to orthogonalize the error in (8.3) to the basis: i.e., set

$$\langle v - u_x, \;\; e_k \rangle = 0 \text{ for all } k (0 \leq k \leq N)$$

where the inner product, $\langle \;\; , \;\; \rangle$, is defined by

$$\langle f(x), g(x) \rangle = \int_{x_0}^{x_N} f(x) g(x) dx.$$

It is easily seen that this is equivalent simply to multiplying both sides of (8.3) by $e_k(x)$ and integrating from x_0 to x_N:

$$\int_{x_0}^{x_N} \sum_{i=0}^{N} v_i e_i(x) e_k(x) dx = \int_{x_0}^{x_N} \sum_{i=0}^{N} u_i e_i'(x) e_k(x) dx. \tag{8.4}$$

Since everything is well-behaved we can exchange the order of the integration and the summation.

Moreover, since v_i and u_i are coefficients of basis functions, they are not functions of x, so we can take them outside the integrals. Hence (8.4) becomes:

$$\sum_{i=0}^{N} v_i \int_{x_0}^{x_N} \mathrm{e_i(x)e_k(x)dx} = \sum_{i=0}^{N} u_i \int_{x_0}^{x_N} \mathrm{e'_i(x)e_k(x)dx}. \tag{8.5}$$

The integrands remaining in (8.5) depend only on the mesh, not on the functions u and v. It is easily seen that for example $e_i(x)e_k(x)$ is zero everywhere unless $i = k$ or $i = k \pm 1$; in fact it is a simple exercise to show that:

$$\begin{array}{ll} \int_{x_0}^{x_N} e_i(x)e_k(x)dx = \frac{h_{k-1}}{6} & \text{for } i = k-1 \\ = \frac{h_{k-1}+h_k}{3} & \text{for } i = k \\ = \frac{h_k}{6} & \text{for } i = k+1 \\ = 0 \quad \text{otherwise} & \end{array}$$

where $h_k = x_{k+1} - x_k$. Similarly,

$$\begin{array}{ll} \int_{x_0}^{x_N} e'_i(x)e_k(x)dx = -\frac{1}{2} & \text{for } i = k-1 \\ = \frac{1}{2} & \text{for } i = k+1 \\ = 0 \quad \text{otherwise.} & \end{array}$$

The outcome is most easily expressed in matrix/vector notation: let

$$\mathbf{u} = (u_0, u_1, \ldots, u_N)^T$$

$$\mathbf{v} = (v_0, v_1, \ldots, v_N)^T,$$

then using the above results Eq. (8.5) becomes

$$\mathbf{Pv} = \mathbf{P_x u} \tag{8.6}$$

where $\mathbf{P}$ is tridiagonal; $\mathbf{P_x}$ is also tridiagonal, but with zeros on the diagonal itself. The matrix $\mathbf{P}$ is diagonally dominant, and the usual algorithm for solving tridiagonal systems can safely be used.

It is very instructive to consider (8.6) in the case of a uniform mesh; at interior points, we have just

$$\frac{1}{6}(\mathrm{v_{k-1}+4v_k+v_{k+1}})\frac{1}{2h}(\mathrm{u_{k+1}-u_{k-1}}), \tag{8.7}$$

where h is the grid length.

If we perform a Taylor series analysis of (8.7), we find that on a uniform mesh we have 4th-order accuracy. General finite-element theory tells us that we only have a right to expect 2nd-order accuracy in the derivative if the basis functions are linear; here we have an example of 'superconvergence' in which the (second-order) errors happen to cancel by symmetry.

Notice also that the manipulations leading to (8.6) automatically yield the appropriate equations for determining $v = u_x$ at the boundary points x_0 and x_N in contrast to the finite-difference case (especially for fourth-order schemes) where in effect we have to 'invent' additional values beyond the boundary.

The right-hand side of (8.7) is of course just the usual second-order finite-difference approximation to $v = u_x$, which is known to be 'smooth' (Fourier analysis shows that derivative is underestimated, the degree of underestimation depending on the number of grid points per wavelength). The left-hand side of (8.7) - i.e., the matrix $\mathbf{P}$ - is itself a three-point smoothing filter. To solve (8.6) we multiply the right-hand side by the inverse of $\mathbf{P}$, which is thus a 'de-smoothing' or 'sharpening' operation. This simple argument provides some insight into the superiority of linear finite elements over second-order finite-differences. In finite element parlance, $\mathbf{P}$ is often called the 'mass matrix'.

8.3.2 Multiplication

Suppose now we are given $u_i = u(x_i)$ and $v_i = v(x_i)$ at a set of nodes $x_i (0 \leq i \leq N)$ and we wish to find the product $w = u_v$. Again we expand u and v in terms of the basis functions $e_i(x)$, and similarly let

$$w(x) = \sum_{i=0}^{N} w_i e_i(x).$$

Applying the Galerkin techniques as before,

$$\int_{x_0}^{x_N} \sum_{i=0}^{N} w_i e_i(x) e_k(x) dx = \int_{x_0}^{x_N} \left(\sum_{i=0}^{N} u_i e_i(x) \right) \left(\sum_{j=0}^{N} v_j e_j(x) \right) e_k(x) dx. \quad (8.8)$$

and we obtain a matrix problem

$$\mathbf{Pw} = \mathbf{r}$$

where **P** is the same tridiagonal matrix as in the previous sub-section. The right-hand side can also be expanded in a similar way; it is easily seen for example that the integral is zero unless both i and j are equal to $k-1$, k or $k+1$. In practice it is more efficient (*much* more efficient in 2 or 3 dimensions) to evaluate the integral by numerical quadrature. The right-hand side of (8.8) is a piecewise cubic, and can be evaluated exactly by using an appropriate integration formula. In fact we have a choice between Gaussian quadrature and using Simpson's formula; Staniforth and Beaudoin (1986) show that the second alternative is twice as efficient than a straightforward term-by-term evaluation).

Notice that in deriving (8.8) we orthogonalized the error to the basis, so that the result $w(x)$ is an *alias-free* representation of the product uv, just as in the case of the spectral method. The product is fourth-order accurate on a uniform mesh (Cullen and Morton, 1980).

8.3.3 Second derivatives

Finally, suppose we are given $u_i = u(x_i)$ at the nodes $x_i (0 \leq i \leq N)$, and we wish to obtain the second derivative $v = u_{xx}$. If we let

$$v(x) = \sum_{i=0}^{N} v_i e_i(x)$$

and proceed exactly as in Section 8.3.1, we obtain the folllowing analogue of Eq. (8.4):

$$\int_{x_0}^{x_N} \sum_{i=0}^{N} v_i e_i(x) e_k(x) dx = \int_{x_0}^{x_N} \sum_{i=0}^{N} u_i e_i^{''}(x) e_k(x)(dx). \tag{8.9}$$

Clearly we are in trouble here, since $e_i^{''}(x)$ is zero everywhere. The trick is to rewrite right-hand side of (8.9) as

$$\int_{x_0}^{x_N} u_{xx} e_k(x) dx$$

and then integrate by parts. We obtain

$$\int_{x_0}^{x_N} \sum_{i=0}^{N} v_i e_i(x) e_k(x)(dx) = [u_x e_k(x)_{x_0}^{x_N} - \int_{x_0}^{x_N} u_x e_k^{'}(x) dx. \tag{8.10}$$

Now we can use the expansion of $\mathbf{u}$ in terms of the basis functions to replace $\mathbf{u}_x$ in the integration:

$$\int_{x_0}^{x_N} \sum_{i=0}^{N} v_i e_i(x) e_k(x)(dx) = [u_x e_k(x)_{x_0}^{x_N} - \int_{x_0}^{x_N} \sum_{i=0}^{N} u_i e_i^{'}(x) e_k^{'}(x) dx,$$

and thus

$$\sum_{i=0}^{N} v_i \int_{x_0}^{x_N} e_i(x) e_k(x)(dx) = [u_x e_k(x)_{x_0}^{x_N} - \sum_{i=0}^{N} u_i \int_{x_0}^{x_N} e_i^{'}(x) e_k^{'}(x) dx. \qquad (8.11)$$

The left-hand side of (8.11), in matrix/vector notation, is just the familiar Pv again. The first term on the right-hand side is zero except at the boundary points. The second term on the right-hand side contains the easily-evaluated integrals.

$$\begin{array}{ll} \int_{x_0}^{x_N} e_i^{'}(x) e_k^{'}(x) dx = -\frac{1}{h_{k-1}} & \text{for } i = k-1 \\ = \left(\frac{1}{h_{k-1}} + \frac{1}{h_k}\right) & \text{for } i = k \\ = -\frac{1}{h_k} & \text{for } i = k+1 \end{array}$$

where again $h_k = x_{k+1} - x_k$.Thus (8.11) has the form

$$\mathbf{Pv} = \mathbf{P_{xx}u}$$

where $\mathbf{P}$ and $\mathbf{P_{xx}}$ are both tridiagonal.

On a uniform grid at interior points, (8.11) becomes

$$\frac{1}{6}(v_{k-1} + 4v_k + v_{k+1}) = \frac{1}{h^2}(u_{k-1} - 2u_k + u_{k+1}) \qquad (8.12)$$

and again the right-hand side is just the usual second-order finite difference approximation. Unlike the case of taking a first derivative, however, inverting the mass matrix does not provide fourth-order accuracy the accuracy remains second-order.

There is a way round this problem, provided we are willing to bend the rules of the Galerkin technique. If we replace the left-hand side of (8.12) by

$$\frac{1}{12}(v_{k-1} + 10v_k + v_{k+1}),$$

in effect using different 'mass matrix', then we recover fourth-order accuracy on a uniform grid, with no extra computational work.

8.4 Efficiency, accuracy and conservation

8.4.1 Efficiency

In the one-dimensional examples of the previous section, there was only one way to divide the domain into subintervals. As soon as we move to higher-dimensional problems, there is a choice. For example, in two dimensions we might choose to subdivide the domain into triangles (especially if the boundary of the domain is irregular) or into rectangles. In either case, linear finite-element basis functions can be defined on the subdivisions. Staniforth (1987) has powerfully argued the case for using a rectangular mesh if the geometry of the problem allows it.

The fundamental reason is the cost of inverting the mass matrix $\mathbf{P}$. On a *rectangular* mesh, the linear finite-element basis functions are *separable*:

$$e_{kl}(x, y) = e_k(x)e_l(y)$$

where e_{kl} is the basis function centred at the mesh point (k, l). as a result, the two-dimensional mass matrix can be inverted simply by solving a set of tridiagonal systems in the x-direction, followed by a set of tridiagonal systems in the y-direction (or vice-versa). Another way of looking at this is that the two-dimensional mass matrix is just a tensor product ($\mathbf{P^x} \otimes \mathbf{P^y}$), where $\mathbf{P^x}$ and $\mathbf{P^y}$ are the mass matrices associated with the one-dimensional sets of basis functions $e_k(x)$ and $e_l(y)$. This tensor product algorithm for inverting the two-dimensional mass matrix was first demonstrated by Staniforth and Mitchell (1977).

On a triangular mesh the separability of the basis functions is lost, and inverting the mass matrix is much more difficult. If the mesh is completely regular then a reasonably efficient FFT-based direct method could be used, but generally it is necessary to resort to iterative methods or approximate inverses (Cullen, 1974b). In engineering applications of the finite element method, the mass matrix is sometimes simply approximated by the identity matrix to circumvent this problem ('mass lumping'), but as we have seen in Section 8.3.1 this can seriously compromise the accuracy of the solution.

Staniforth (1987) also points out several other efficiency 'tricks' which exist for linear finite elements on rectangles, but which do not carry over to triangular meshes. In three dimensions, the arguments for trilinear basis functions on bricks' rather than on tetrahedra are even stronger.

8.4.2 Accuracy

So far we have considered only methods based on *linear* finite elements. Is it worth using higher-order elements? In several respects, the answer seems to be no. Quadratic elements can be less accurate than linear elements (no superconvergence properties), so the extra expense is not likely to be justified. Cubic elements do have superconvergence properties and can provide a high order of accuracy, but they are much more expensive to compute with than linear elements; also the additional degrees of freedom can result in computational modes (noise) being excited. Staniforth (1987) puts it succinctly: 'the law of diminishing returns seems to apply'.

8.4.3 Conservation

Finite-difference schemes for nonlinear problems are often designed to conserve invariants of the original partial differential equations. In general, such conservation properties are not maintained in simple linear finite element schemes. They can be recovered for example by spatial staggering of the elements, or by choosing different orders of element for different variables (Lee *et al.*, Cliffe, 1981; Girard, 1983; Steppeler, 1987b). The extent to which the extra computation is worthwhile seems to be rather debatable, and in any case is certain to be problem-dependent.

Bibliography

Ames, W. F. (1969). *Numerical methods for partial differential equations*. Barnes and Noble Inc., New York.

Arakawa, A. (1966). Computational design for long-term numerical integrations of the equations of atmospheric motion. *J. Comput. Phys.*, 1:119–143.

Arakawa, A. (1972). Design of the ucla general circulation model. Numerical simulation of weather and climate, Dept of Meteorology, Univ. of California, Los Angeles.

Arakawa, A. et al. (1974). The ucla atmospheric general circulation model. Technical report, University of California, Los Angeles.

Arakawa, A. and Lamb (1976). Computational design of the ucla general circulation model. In *Methods in Computational Physics*. Academic Press, New York.

Asselin (1972). Frequency filter for time integrations. *Mon. Wea. Rev.*, 100:487–490.

Bjerknes, V. (1904). Das problem von der wettervorhersage, betrachtet vom standpunkt der mechanik und der physik. *Meteor. Z.*, 21:1–7.

Boris, J. P. and Book, D. L. (1973). Flux-corrected transport I: SHASTA a fluid transport algorithm that works. *J. Comput. Phys.*, 11:38–69.

Charney, J., Fjortoft, R., and von Neumann, J. (1950). Numerical integration of the barotropic vorticity equations. *Tellus*, 2:237–254.

Charney, J. G. (1966). Some remaining problems in numerical weather prediction. In *Advances in numerical weather prediction*, pages 61–70, Hartford, Conn. Travelers Research Center, Inc.

Courant, R., Friedrichs, K., and Lewy, H. (1928). über die partiellen differenzengleichungen der mathematischen physik. *Mathematische Annalen*, 100:32–74.

Cullen, M. J. P. (1974). Integration of the primitive equations on a shere using the finite element method. *Quart. J. Roy. Meteor. Soc.*, 100:555–562.

Dorr, F. W. (1970). The direct solution of the discrete poisson equation on a rectangle. *SIAM Review*, 12:248–263.

ECMWF (1991). Numerical methods in atmospheric models. In *Seminar proceedings*, volume 1. European Centre for Medium-Range Weather Forecasts.

Eliassen, A. (1956). A procedure for numerical integration of the primitive equations of the two-parameter model of the atmosphere. Sci. Rep 4, Dept. of Meteorology, UCLA.

Fjortoft, R. (1953). On the changes in the spectral distribution of kinetic energy for two-dimensional, non-divergent flows. *Tellus*, 5:225–230.

Frankel (1950). Convergence rates of iterative treatments of partial differential equations. *Math tables and other aids to computation*, 4:65–75.

Gillespie, R. P. (1955). *Integration*. Oliver and Boyd, Edinburgh.

Kasahara, A. (1969). Simulation of the earths atmosphere. Ncar manuscript 69-27, National Center for Atmospheric Research, Boulder, Colo.

Kwizak, M. and Robert, A. (1971). A semi implicit scheme for grid point atmospheric models of the primitive equations. *Mon. Wea. Rev.*, 99:32–36.

Marchuk (1967). *Numerical methods in weather prediction*. Academic Press.

Mesinger (1973). A method for construction of second-order accuracy difference schemes permitting no false two-grid-interval wave in the height field. *Tellus*, 25:444–458.

Mesinger, F. and Arakawa, A. (1976). Numerical methods used in atmospheric models. In *GARP Publications Series*, number 14, page 64 pp. WMO/ICSU Joint Organizing Committee.

Miyakoda (1962). Contribution to the numerical weather prediction computation with finite difference. *Japanese J. of Geophysics*, 3:75–190.

Orszag (1971). On the elimination of aliasing in finite-difference schemes by filtering high wave number components. *J. Atmos. Sci.*, 28:1074 pp.

Phillips, N. A. (1956). The general circulation of the atmosphere: a numerical experiment. *Quart. J. Roy. Meteor. Soc.*, 82:123 pp.

Richardson, L. F. (1922). *Weather prediction by numerical process*. Cambridge Univ. Press, London. Reprinted by Dover.

Roache, P. J. (1972). *Computational fluid dynamics*. Hermosa publishers, Albuquerque, New Mexico.

Robert, A., J. (1966). The integration of a low order spectral form of the primitive meteorological equations. *J. Meteor. Soc. Japan*, 44:237–245.

Silberman, I. (1954). Planetary waves in the atmosphere. *J. Meteor.*, 11:27–34.

Smolarkiewicz, P. K. (1984). A fully multidimensional positive definite advection transport algorithm with small implicit diffusion. *J. Comp. Phys.*, 54:325–362.

Smolarkiewicz, P. K. and Clark (1986). The multidimensional positive definite advection transport algorithm: Further development and applications. *T. J. Comp. Phys.*, 86:355–375.

Smolarkiewicz, P. K. and Grell, G. A. (1992). A class of monotone interpolation schemes. *J. Comp. Phys.*, 101:431–440.

Smolarkiewicz, P. K. and Rasch, P. J. (1990). Monotone advection on the sphere: An eulerian versus semi-lagrangian approach. *J. Atmos. Sci.*, 48:793–810.

Staniforth and Mitchell (1977). A semi-implicit finite-element barotropic model. *Mon. Wea. Rev.*, 105:154–169.

Staniforth, A. (1987). Review: Formulating efficient finite-element codes for flows in regular domains. *Int. J. Num. Meth. Fluids*, 7:1–16.

Staniforth, A. and Cote, J. (1991). Semi-lagrangian integration schemes for atmospheric models a review. *Mon. Wea. Rev.*, 119:2206–2223.

Steppeler, J. (1987a). Energy conserving galerkin finite element schemes for the primitive equations of numerical weather prediction. *J. Comp. Phys.*, 69:258–264.

Steppeler, J. (1987b). Quadratic galerkin finite element schemes for the vertical discretization of numerical forecast models. *Mon. Wea. Rev.*, 115:1575–1588.

Wachspress, E. L. (1966). *Iterative solution of elliptic systems*. Prentice-Hall, Englewood Cliffs, New Jersey.

Wurtele (1961). On the problem of truncation error. *Tellus*, 13:379–391.

Young, D. (1954). Iterative methods for solving partial differential equations of elliptic type. *Transactions of the American Mathematical Society*, 76:92–111.

Zalesak, S. T. (1979). Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comp. Phys.*, 31:335–362.