CHAPTER 4 A closer look at the advection equation

4.1 Introduction

Most of this chapter is devoted to a discussion of the one-dimensional advection equation,

$$\frac{\partial A}{\partial t} + c\frac{\partial A}{\partial x} = 0 . \tag{4.1}$$

Here A is the advected quantity, and c is the advecting current. This is a linear, first-order, partial differential equation with a constant coefficient, namely c. Both space and time differencing are discussed in this chapter, but more emphasis is placed on space differencing.

We have already presented the exact solution of (4.1). Before proceeding, however, it is useful to review the physical nature of advection, because the design or choice of a numerical method should always be motivated as far as possible by our understanding of the physical process at hand.

In Lagrangian form, the advection equation, in any number of dimensions, is simply

$$\frac{DA}{Dt} = 0 . \tag{4.2}$$

This means that the value of A does not change following a particle. We say that A is "conserved" following a particle. In fluid dynamics, we consider an infinite collection of fluid particles. According to (4.2), each particle maintains its value of A as it moves. If we do a survey of the values of A in our fluid system, let advection occur, and conduct a "follow-up" survey, we will find that exactly the same values of A are still in the system. The locations of the particles presumably will have changed, but the maximum value of A over the population of particles is unchanged by advection, the minimum value is unchanged, the average is unchanged, and in fact *all of the statistics of the distribution of A over the mass of the fluid are completely unchanged by the advective process*. This is an important characteristic of advection.

Here is another way of describing this characteristic: If we worked out the probability density function (pdf) for A, by defining narrow "bins" and counting the mass associated with

particles having values of A falling within each bin, we would find that the pdf was unchanged by advection. For instance, if the pdf of A at a certain time is Gaussian (or "bell shaped"), it will still be Gaussian at a later time (and with the same mean and standard deviation) if the only intervening process is advection and if no mass enters or leaves the system.

Consider a simple function of A, such as A^2. Since A is unchanged during advection, for each particle, A^2 will also be unchanged. Obviously, any other function of A will also be unchanged. It follows that the pdf for any function of A is unchanged by advection.

In many cases of interest, A is non-negative more or less by definition. For example, the mixing ratio of water vapor cannot be negative; a negative mixing ratio would have no physical meaning. Some other variables, such as the zonal component of the wind vector, can be either positive or negative; for the zonal wind, our convention is that positive values denote westerlies and negative values denote easterlies.

Suppose that A is conserved under advection, following each particle. It follows that if there are no negative values of A at some initial time, then, to the extent that advection is the only process at work, there will be no negative values of A at any later time either. This is true whether the variable in question is non-negative by definition (like the mixing ratio of water vapor) or not (like the zonal component of the wind vector).

Typically the variable A represents an "intensive" property, which is defined per unit mass. An example is the mixing ratio of some trace species, such as water vapor. A second example is temperature, which is proportional to the internal energy per unit mass. A third example, and a particularly troublesome one, is the case in which A is a component of the advecting velocity field itself; here A is a component of the momentum per unit mass.

Of course, in general these various quantities are not really conserved following particles; various sources and sinks cause the value of A to change as the particle moves. For instance, if A is temperature, one possible source is radiative heating. To describe more general processes which include not only advection but also sources and sinks, we replace (4.2) by

$$\frac{DA}{Dt} = S, \tag{4.3}$$

where S is the source of A per unit time. (A negative value of S represents a sink.) We still refer to (4.3) as a "conservation" equation; it says that A is conserved *except* to the extent that sources or sinks come into play.

In addition to conservation equations for quantities that are defined per unit mass, we need a conservation equation for mass itself. This can be written as

$$\frac{\partial \rho}{\partial t} = -\nabla\bullet(\rho \mathbf{V}), \tag{4.4}$$

where ρ is the density (mass per unit volume) and $\mathbf{V}$ is the velocity vector. Using the velocity vector, we can expand (4.3) into the Eulerian advective form of the conservation equation for A:

$$\frac{\partial A}{\partial t} = -(\mathbf{V} \bullet \nabla)A + S. \tag{4.5}$$

Multiply (4.4) by A, and (4.5) by ρ and add the results to obtain

$$\frac{\partial}{\partial t}(\rho A) = -\nabla\bullet(\rho \mathbf{V} A) + \rho S. \tag{4.6}$$

This is called the flux form of the conservation equation for A. Notice that if we put $A \equiv 1$ and $S \equiv 0$ then (4.6) reduces to (4.4). This is an important point that can and should be used in the design of advection schemes.

If we integrate (4.4) over a closed domain R ("closed" meaning that R experiences no sources or sinks of mass) then we find, using Gauss's Theorem, that

$$\frac{d}{dt}\int_R \rho \, dR = 0. \tag{4.7}$$

This simply states that mass is conserved within the domain. Similarly, we can integrate (4.6) over R to obtain

$$\frac{d}{dt}\int_R \rho A \, dR = \int_R \rho S \, dR. \tag{4.8}$$

This says that the mass-weighted average value of A is conserved within the domain, except for the effects of sources and sinks. We can say that (4.6) and (4.8) are integral forms of the conservation equations for mass and A, respectively.

It may seem that the ideal way to simulate advection in a model is to define a collection of particles, to associate various properties of interest with each particle, and to let the particles be advected about by the wind. In such a Lagrangian model, the properties associated with each particle would include its spatial coordinates, e.g. its longitude, latitude, and height. These would change in response to the predicted velocity field. Such a Lagrangian approach will be discussed later in this chapter.

At the present time, virtually all models in atmospheric science are based on Eulerian methods, although the Eulerian coordinates are sometimes permitted to "move" as the circulation evolves (e.g. Phillips, 1957; Hsu and Arakawa, 1990).

When we design finite-difference schemes to represent advection, we strive for

accuracy, stability, simplicity, and computational economy, as always. In addition, it is often required that a finite-difference scheme for advection be conservative in the sense that

$$\sum_j \rho_j^{n+1}\, dR_j = \sum_j \rho_j^{n}\, dR_j \tag{4.9}$$

and

$$\sum_j (\rho A)_j^{n+1}\, dR_j = \sum_j (\rho A)_j^{n}\, dR_j + \Delta t \sum_j (\rho S)_j^{n}\, dR_j . \tag{4.10}$$

These are finite-difference analogs to the integral forms (4.7) and (4.8), respectively. In (4.10) we have assumed for simplicity that the effects of the source, S, are evaluated using forward time differencing, although this need not be the case in general.

We may also wish to require conservation of some function of A, such as A^2. This might correspond, for example, to conservation of kinetic energy. Energy conservation can be arranged, as we will see.

Finally, we may wish to require that a non-negative variable, such as the water vapor mixing ratio, remain non-negative under advection. An advection scheme with this property is often called "positive-definite" or "sign-preserving" positive. Definite schemes are obviously desirable, since negative values that arise through truncation errors will have to be eliminated somehow before any moist physics can be considered, and the methods used to eliminate the negative values are inevitably somewhat artificial (e.g. Williamson and Rasch, 1994). As we will see, most of the older advection schemes do not come anywhere near satisfying this requirement. Many newer schemes do satisfy it, however.

There are various additional requirements that we might like to impose. Ideally, for example, the finite-difference advection operator would not alter the pdf of A over the mass. Unfortunately this cannot be guaranteed with Eulerian methods, although we can minimize the effects of advection on the pdf, especially if the shape of the pdf is known *a priori*. This will be discussed later. Note that in a model based on Lagrangian methods, advection does not alter the pdf of the advected quantity.

4.2 Conservative finite-difference methods

Let A be a "conservative" variable, satisfying the following one-dimensional conservation law:

$$\frac{\partial}{\partial t}(mA) + \frac{\partial}{\partial x}(muA) = 0 . \tag{4.11}$$

Here m is a mass variable, which might be the density of the air, or might be the depth of shallow water, and mu is a mass flux. Putting $A \equiv 1$ in (4.11) gives mass conservation:

$$\frac{\partial m}{\partial t} + \frac{\partial}{\partial x}(mu) = 0 . \tag{4.12}$$

Approximate (4.11) and (4.12) with:

$$\frac{d}{dt}(m_j A_j) + \frac{(mu)_{j+\frac{1}{2}} A_{j+\frac{1}{2}} - (mu)_{j-\frac{1}{2}} A_{j-\frac{1}{2}}}{\Delta x_j} = 0 , \tag{4.13}$$

$$\frac{dm_j}{dt} + \frac{(mu)_{j+\frac{1}{2}} - (mu)_{j-\frac{1}{2}}}{\Delta x_j} = 0 . \tag{4.14}$$

These are called differential-difference equations (or sometimes semi-discrete equations), because the time-change terms are in differential form, while the spatial derivatives have been approximated using a finite-difference quotient. The variables m and A are defined at integer points, while u and mu are defined at half-integer points. See Fig. 4.1. This is an example of

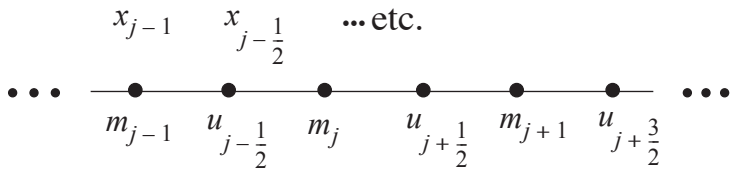


Figure 4.1: The staggered grid used in (4.13) and (4.14).

a "staggered" grid. $A_{j+\frac{1}{2}}$ and $A_{j-\frac{1}{2}}$ must be interpolated somehow from the predicted values of A. Note that if we put $A \equiv 1$, (4.13) reduces to (4.14). This is an important point which will be discussed further later.

Multiply (4.13) and (4.14) through by Δx_j, and sum over the domain:

$$\frac{d}{dt}\sum_{j=0}^{J}(m_j A_j \Delta x_j) + (mu)_{J+\frac{1}{2}} A_{J+\frac{1}{2}} - (mu)_{-\frac{1}{2}} A_{-\frac{1}{2}} = 0 , \tag{4.15}$$

$$\frac{d}{dt}\sum_{j=0}^{J}(m_j \Delta x_j) - (mu)_{J+\frac{1}{2}} + (mu)_{-\frac{1}{2}} = 0 . \tag{4.16}$$

If

$$(mu)_{J+\frac{1}{2}} A_{J+\frac{1}{2}} = (mu)_{-\frac{1}{2}} A_{-\frac{1}{2}} , \tag{4.17}$$

and

$$(mu)_{J+\frac{1}{2}} = (mu)_{-\frac{1}{2}}, \tag{4.18}$$

then we obtain

$$\frac{d}{dt}\sum_{j=0}^{J}(m_j A_j \Delta x_j) = 0, \tag{4.19}$$

$$\frac{d}{dt}\sum_{j=0}^{J}(m_j \Delta x_j) = 0, \tag{4.20}$$

which express conservation of mass [Eq. (4.20)] and of the mass-weighted value of A [Eq. (4.19)]. Compare with (4.9) and (4.10). Note that (4.19) holds regardless of the form of the interpolation used for $A_{j+\frac{1}{2}}$.

By combining (4.11) and (4.12), we obtain the advective form of our conservation law:

$$m\frac{\partial A}{\partial t} + mu\frac{\partial A}{\partial x} = 0. \tag{4.21}$$

From (4.13) and (4.14), we can derive a finite-difference "advective form," analogous to (4.21):

$$m_j\frac{dA_j}{dt} + \frac{(mu)_{j+\frac{1}{2}}\left(A_{j+\frac{1}{2}} - A_j\right) + (mu)_{j-\frac{1}{2}}\left(A_j - A_{j-\frac{1}{2}}\right)}{\Delta x_j} = 0. \tag{4.22}$$

Since (4.22) is consistent with (4.13) and (4.14), use of (4.22) and (4.14) will allow conservation of the mass-weighted value of A (and of mass itself). Also note that if A is uniform over the grid, then (4.22) gives $\frac{dA_j}{dt} = 0$, which is "the right answer." Note that this is ensured ***because*** (4.13) reduces to (4.14) when A is uniform over the grid. *If the flux-form advection equation does not reduce to the flux-form continuity equation when A is uniform over the grid, then a uniform tracer field will not remain uniform under advection.*

We have already discussed the fact that, for the continuous system, conservation of A itself implies conservation of *any function* of A, e.g., A^2, A^n, $\ln(A)$, etc. This is most easily seen from the Lagrangian form of (4.21):

$$\frac{DA}{Dt} = 0 . \tag{4.23}$$

According to (4.23), A is conserved "following a particle." As discussed earlier, this implies that

$$\frac{D}{Dt}[F(A)] = 0 , \tag{4.24}$$

where $F(A)$ is an arbitrary function of A only. We can derive (4.24) by multiplying (4.23) by dF/dA.

In a finite difference system, we can force conservation of at most one nontrivial function of A, in addition to A itself. Let F_j denote $F(A_j)$, where F is an arbitrary function, and let F'_j denote $\dfrac{d[F(A_j)]}{dA_j}$. Multiplying (4.22) by F'_j gives

$$m_j \frac{dF_j}{dt} + \frac{(mu)_{j+\frac{1}{2}} F_j'\left(A_{j+\frac{1}{2}} - A_j\right) + (mu)_{j-\frac{1}{2}} F_j'\left(A_j - A_{j-\frac{1}{2}}\right)}{\Delta x_j} = 0 . \tag{4.25}$$

Now use (4.14) to rewrite (4.25) in "flux form":

$$\frac{d}{dt}(m_j F_j) + \frac{1}{\Delta x_j}\left\{(mu)_{j+\frac{1}{2}}\left[F_j'\left(A_{j+\frac{1}{2}} - A_j\right) + F_j\right] - (mu)_{j-\frac{1}{2}}\left[-F_j'\left(A_j - A_{j-\frac{1}{2}}\right) + F_j\right]\right\} = 0 . \tag{4.26}$$

Inspection of (4.26) shows that, to ensure conservation of $F(A)$, we must choose

$$F_{j+\frac{1}{2}} = F_j'\left(A_{j+\frac{1}{2}} - A_j\right) + F_j , \tag{4.27}$$

$$F_{j-\frac{1}{2}} = -F_j'\left(A_j - A_{j-\frac{1}{2}}\right) + F_j . \tag{4.28}$$

Let $j \to j+1$ in (4.28), giving

$$F_{j+\frac{1}{2}} = -F'_{j+1}\left(A_{j+1} - A_{j+\frac{1}{2}}\right) + F_{j+1} . \tag{4.29}$$

Eliminating $F_{j+\frac{1}{2}}$ between (4.27) and (4.29), we obtain

$$A_{j+\frac{1}{2}} = \frac{(F'_{j+1}A_{j+1} - F_{j+1}) - (F'_j A_j - F_j)}{F'_{j+1} - F'_j} . \tag{4.30}$$

By choosing $A_{j+\frac{1}{2}}$ accordingly to (4.30), we can guarantee conservation of both A and $F(A)$ (apart from time-differencing errors).

As an example, suppose that $F(A) = A^2$. Then $F'(A) = 2A$, and we find that

$$A_{j+\frac{1}{2}} = \frac{(2A^2_{j+1} - A^2_{j+1}) - (2A^2_j - A^2_j)}{2(A_{j+1} - A_j)} = \frac{1}{2}(A_{j+1} + A_j) . \tag{4.31}$$

This arithmetic-mean interpolation allows conservation of the square of A. It may or may not be an *accurate* interpolation for $A_{j+\frac{1}{2}}$. Note that x_{j+1}, x_j, and $x_{j+\frac{1}{2}}$ do not appear in (4.31). This means that our spatial interpolation does not contain any information about the spatial locations of the various grid points involved -- a rather awkward and somewhat strange property of the scheme. If the grid spacing is uniform, (4.31) gives second-order accuracy in space. If the grid spacing is nonuniform, however, the accuracy drops to first order. The strength of the first-order error depends on how rapidly the grid spacing changes. Substituting (4.31) back into (4.22) gives

$$m_j \frac{dA_j}{dt} + \frac{1}{2\Delta x_j}\left[(mu)_{j+\frac{1}{2}}(A_{j+1} - A_j) + (mu)_{j-\frac{1}{2}}(A_j - A_{j-1})\right] = 0 . \tag{4.32}$$

This is the advective form that allows conservation of A^2 (and of A).

One point to be noticed here is that there are infinitely many ways to interpolate a variable. We can spatially interpolate A itself in a linear fashion, e.g.

$$A_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} A_j + \left(1 - \alpha_{j+\frac{1}{2}}\right) A_{j+1} , \tag{4.33}$$

where $\alpha_{j+\frac{1}{2}}$ is a weighting factor which might be a constant, as in (4.31), or might be a function of x_j, x_{j+1}, and $x_{j+\frac{1}{2}}$. We can interpolate so as to conserve an arbitrary function of A, as in (4.30). We can compute an arbitrary function of A, interpolate the function using a

form such as (4.33), and then extract an interpolated value of A by applying the inverse of the function to the result. A practical example of this would be interpolation of the water vapor mixing ratio by computing the relative humidity, interpolating the relative humidity, and then converting back to mixing ratio. We can also make use of "averages" which are different from the simple and familiar arithmetic mean given by (4.31). Examples are the "*geometric mean*,"

$$A_{j+\frac{1}{2}} = \sqrt{A_j A_{j+1}}, \tag{4.34}$$

and the "*harmonic mean*,"

$$A_{j+\frac{1}{2}} = \frac{2A_j A_{j+1}}{A_j + A_{j+1}}. \tag{4.35}$$

Note that both (4.34) and (4.35) give $A_{j+\frac{1}{2}} = C$ if both A_{j+1} and A_j are equal to C, which is what we expect from an "average." They are both nonlinear averages. For example, the geometric mean of A plus the geometric mean of B is not equal to the geometric mean of $A + B$, although it will usually be close. The geometric mean and the harmonic mean both have the potentially useful property that if either A_{j+1} or A_j is equal to zero, then $A_{j+\frac{1}{2}}$ will also be equal to zero. More generally, both (4.34) and (4.35) tend to make the interpolated value close to the smaller of the two input values.

Here is another interesting interpolation that has the opposite property, i.e., it makes the interpolated value close to the larger of the two input values:

$$A_{j+\frac{1}{2}} = \frac{A_j + A_{j+1} - 2A_j A_{j+1}}{2 - (A_j + A_{j+1})}. \tag{4.36}$$

In short, there are infinitely many ways to average and/or interpolate. This is good because it means that we have the opportunity to choose the *best* way for our particular application.

Fig. 4.2 shows four interpolations as functions of the two input values.

In this section, we have considered truncation errors only insofar as they affect conservation properties. We must also consider how they affect the various other aspects of the solution. This is taken up in the next section.

4.3 Examples of schemes with centered space differencing

A centered-difference quotient already discussed in Chapter 2 is

$$\frac{\partial u^A}{\partial x} \cong \frac{u^A{}_{j+1} - u^A{}_{j-1}}{2\Delta x} \text{ at } x_j, \tag{4.37}$$

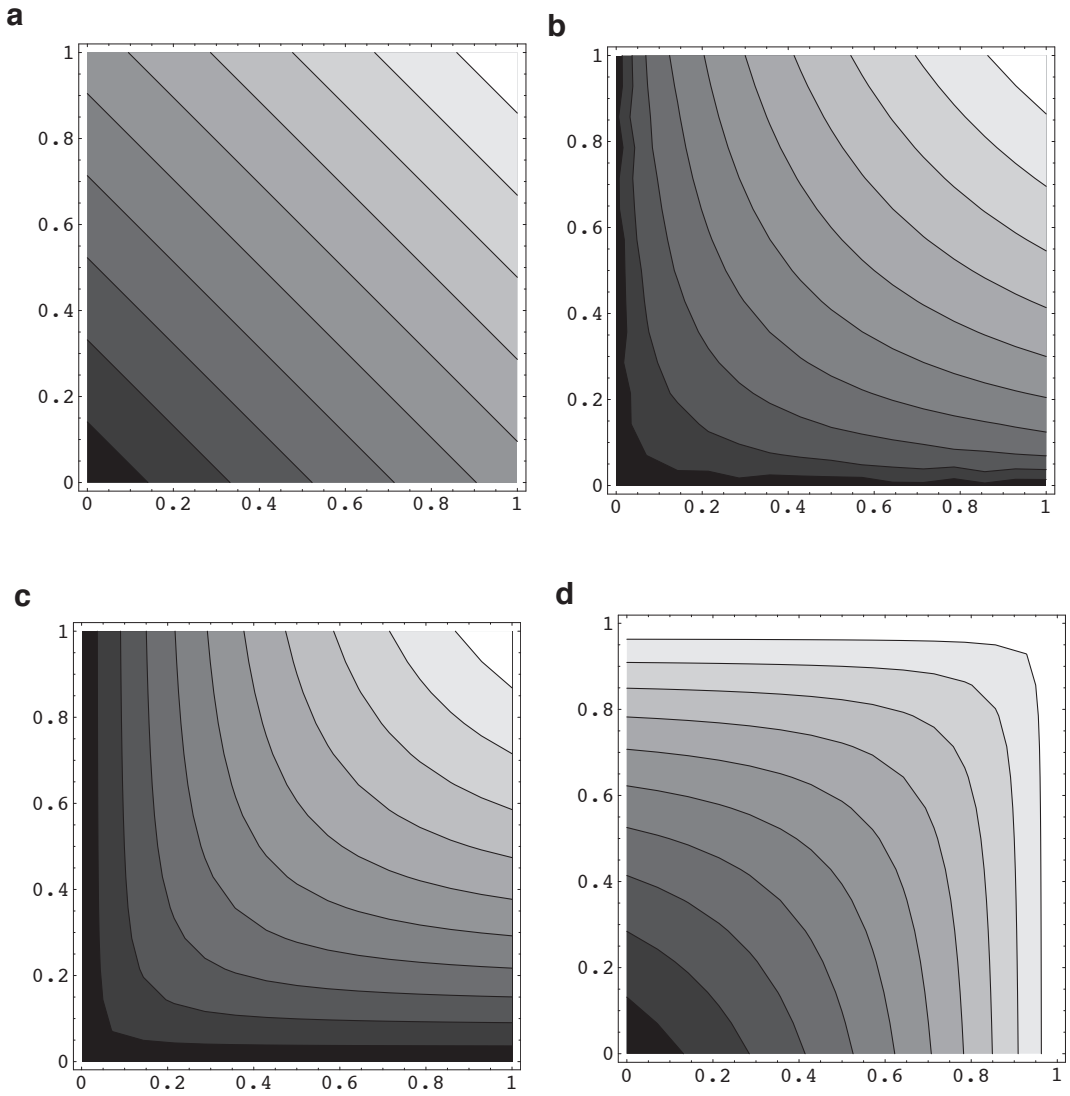


Figure 4.2: **Four interpolations as functions of the input values. a) arithmetic mean, b) geometric mean, c) harmonic mean, d) Eq. (4.36), which makes the interpolated value close to the larger of the two input values. In all plots, black is close to zero, and white is close to one.**

where j is the spatial index and n is the time index. If $u(x, t)$ has the wave form $u(x, t) = \hat{u}(t)e^{ikj\Delta x}$, where k is the wave number, then

$$\frac{u_{j+1} - u_{j-1}}{2\Delta x} = ik\frac{\sin k\Delta x}{k\Delta x}\hat{u}(t)e^{ikj\Delta x} . \tag{4.38}$$

Therefore, the advection equation becomes

$$\frac{d\hat{u}}{dt} + ikc\frac{\sin k\Delta x}{k\Delta x}\hat{u} = 0 . \tag{4.39}$$

If we define $\omega \equiv -kc\frac{\sin k\Delta x}{k\Delta x}$, then (4.39) reduces to the oscillation equation, which was discussed at length in Chapter 3. Note that $\frac{\sin k\Delta x}{k\Delta x} \to 1$ as $k\Delta x \to 0$.

We can now study the properties of the various time-differencing schemes, as we did in Chapter 3, but we are now able to obtain an explicit relationship between Δx and Δt as a condition for stability, based on the use of (4.39) The forward time scheme is unstable when combined with the centered space scheme. You should prove this fact and remember it. It was mentioned already in Chapter 3. In the case of the leapfrog scheme, the finite-difference analogue of (4.1) is

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + c\left(\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}\right) = 0 . \tag{4.40}$$

If we assume that u_j^n has the wave form for which (4.38) holds, then (4.40) can be written as

$$\hat{u}^{n+1} - \hat{u}^{n-1} = 2i\Omega\hat{u}^n , \tag{4.41}$$

where

$$\Omega \equiv -kc\frac{\sin k\Delta x}{k\Delta x}\Delta t . \tag{4.42}$$

We recognize Eq. (4.41) as the leapfrog scheme for the oscillation equation. Recall from Chapter 3 that $|\Omega| \le 1$ is necessary for (4.41) to be stable. Therefore,

$$\left| c\frac{\sin k\Delta x}{\Delta x}\Delta t \right| \le 1 \tag{4.43}$$

must hold for stability, for any and all k. The "worst case" is $|\sin k\Delta x| = 1$. This occurs for wavelength $L = 4\Delta x$, so

$$|c|\frac{\Delta t}{\Delta x} \le 1 \tag{4.44}$$

is the necessary condition for stability, i.e. stability for all modes. Note that the $2\Delta x$ wave is not the problem here. It is the $4\Delta x$ wave that is most likely to cause trouble. Eq. (4.44) is the famous "CFL" stability criterion associated with the names Courant, Friedrichs and Lewy.

Recall that the leapfrog scheme gives a numerical solution with two modes -- a

physical mode and a computational mode. We can write these two modes as in Chapter 3:

$$\hat{u}_1 = \lambda_1^n \hat{u}_1, \qquad \hat{u}_2 = \lambda_2^n \hat{u}_2 . \tag{4.45}$$

For $|\Omega| \le 1$, we find, as discussed in Chapter 3, that

$$\lambda_1 = e^{i\theta}, \qquad \theta = \tan^{-1} \frac{\Omega}{\sqrt{1-\Omega^2}}, \tag{4.46}$$

$$\lambda_2 = e^{i(\pi-\theta)} = -e^{-i\theta} . \tag{4.47}$$

Both modes are neutral. For the physical mode,

$$(\hat{u}_j^n)_1 = \lambda_1^{(n)} \hat{u}_1^0 e^{ikj\Delta x} = \hat{u}_1^0 exp\left[ik\left(j\Delta x + \frac{\theta}{k\Delta t} n\Delta t \right) \right] . \tag{4.48}$$

For the computational mode, similarly, we obtain

$$(u_j^n)_2 = \hat{u}_2^0 (-1)^n exp\left[ik\left(j\Delta x - \frac{\theta}{k\Delta t} n\Delta t \right) \right] . \tag{4.49}$$

Note the nasty factor of $(-1)^n$, which comes from the leading minus sign in (4.47). Comparing (4.48) and (4.49) with the expression $u(x, t) = \hat{u}(0) e^{ik(x-ct)}$, which is the true solution, we see that the speeds of the physical and computational modes are $-\frac{\theta}{k\Delta t}$ and $\frac{\theta}{k\Delta t}$, respectively, for even time steps. It is easy to see that as $(\Delta x, \Delta t) \to 0$, $\theta \to \Omega \to -kc\Delta t$, and so the speed of the physical mode approaches c, while that of the computational mode approaches $-c$. The computational mode goes backwards!

Note that the finite-difference approximation to the phase speed depends on k, while the true phase speed, c, is independent of k. The spurious dependence of phase speed on wave number with the finite-difference scheme is an example of *computational dispersion*, which will be discussed in detail later.

Further examples of schemes for the advection equation can be obtained by combining this centered space differencing with the two-level time-differencing schemes (see Chapter 3). In the case of the Matsuno scheme, the first approximation to u_j^{n+1} comes from

$$\frac{(u_j^{n+1})^* - u_j^{(n)}}{\Delta t} + c \frac{u_{j+1}^{(n)} - u_{j-1}^{(n)}}{2\Delta x} = 0 , \tag{4.50}$$

and the final value from

$$\frac{u_j^{(n+1)} - u_j^{(n)}}{\Delta t} + c\frac{(u_{j+1}^{(n+1)})^* - (u_{j-1}^{(n+1)})^*}{2\Delta x} = 0 . \tag{4.51}$$

Eliminating the terms with $(\ )^*$ from (4.51) by using (4.50) twice (first with j replaced by $j+1$, then with j replaced by $j-1$), we obtain

$$\frac{u_j^{(n+1)} - u_j^{(n)}}{\Delta t} + c\frac{u_{j+1}^{(n)} - u_{j-1}^{(n)}}{2\Delta x} = \frac{c^2\Delta t}{(2\Delta x)^2}(u_{j+2}^{(n)} - 2u_j^{(n)} + u_{j-2}^{(n)}) . \tag{4.52}$$

The term on the right side of (4.52) approaches zero as $\Delta t \to 0$, and thus (4.52) is consistent with (4.1). If we let $\Delta x \to 0$ (and $\Delta t \to 0$ to keep stability), this term approaches $\frac{c^2\Delta t}{4}\frac{\partial^2 u}{\partial x^2}$. In effect, *it acts as a diffusion term that damps disturbances*. The "diffusion coefficient" is $\frac{c^2\Delta t}{4}$, so it goes to zero as $\Delta t \to 0$.

We now examine a similarly diffusive scheme, called the Lax-Wendroff scheme, which has second-order accuracy. Consider an explicit two-level scheme of the form:

$$\frac{u^{n+1}{}_j - u_j^n}{\Delta t} + \frac{c}{\Delta x}(\alpha u_{j-1}^n + \beta u_j^n + \gamma u_{j+1}^n) = 0 . \tag{4.53}$$

Note that there are only two time levels, so this scheme does not have any computational modes. Replace the various u's by the corresponding values of the true solution and then expand them into the Taylor series around the point $(j\Delta x, n\Delta t)$. The result is

$$\begin{aligned}
&\left(u_t + \frac{\Delta t}{2!}u_{tt} + \frac{\Delta t^2}{3!}u_{ttt}\right)_j + \cdots \\
&\quad + \frac{c}{\Delta x}\left[\alpha\left(u - \Delta x u_x + \frac{\Delta x^2}{2!}u_{xx} - \frac{\Delta x^3}{3!}u_{xxx} + \text{-} \cdots\right)_j\right. \\
&\qquad + \beta u_j \\
&\qquad \left. + \gamma\left(u + \Delta x u_x + \frac{\Delta x^2}{2!}u_{xx} + \frac{\Delta x^3}{3!}u_{xxx} + \cdots\right)_j\right] \\
&= \varepsilon ,
\end{aligned} \tag{4.54}$$

where all quantities are evaluated at $(x, t) = (x_j, t^n)$, and ε is the truncation error. Make sure that you understand where (4.54) comes from. From the consistency condition, we must require

$$\alpha + \beta + \gamma = 0, \text{ and } -\alpha + \gamma = 1. \tag{4.55}$$

These conditions ensure first-order accuracy. If we further require second-order accuracy *in both time and space*, we must require that

$$\frac{\Delta t}{2} c^2 + \frac{c\Delta x}{2}(\alpha + \gamma) = 0. \tag{4.56}$$

Here we have used

$$u_{tt} = c^2 u_{xx}, \tag{4.57}$$

which follows from the exact advection equation provided that c is constant. Solving, we get

$$\alpha = \frac{-1-\mu}{2}, \ \beta = \mu \text{ and } \gamma = \frac{1-\mu}{2}, \tag{4.58}$$

where, as usual, $\mu \equiv \dfrac{c\Delta t}{\Delta x}$.

The scheme can be written as

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c}{2\Delta x}(u_{j+1}^n - u_{j-1}^n) = \frac{c^2 \Delta t}{2\Delta x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \tag{4.59}$$

Note that although (4.59) is second-order accurate in time, it involves only two time levels. On the other hand, the scheme achieves second-order accuracy in space through the use of three grid points. Compare (4.59) with (4.52).

In this example, we used three grid points $(j-1, j, j+1)$ to construct our approximation to $\dfrac{du}{dx}$, each with a coefficient (α, β, γ). In a similar way, we could have used any number of grid points, each with a suitably chosen coefficient, to construct a scheme of arbitrary accuracy in both space and time. The result would still be a two-time-level scheme! This illustrates that *a non-iterative two-level scheme is not necessarily a first-order scheme.*

Eq. (4.59) was proposed by Lax and Wendroff (1960), and recommended by Richtmeyer (1963). The right-hand-side of (4.59) looks like a diffusion term. It tends to smooth out small-scale noise. The Lax-Wendroff scheme is equivalent to and can be

interpreted in terms of the following procedure: First calculate $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ and $u_{j-\frac{1}{2}}^{n+\frac{1}{2}}$ from

$$\frac{u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}(u_{j+1}^{n} + u_{j}^{n})}{\frac{1}{2}\Delta t} = -c\left(\frac{u_{j+1}^{n} - u_{j}^{n}}{\Delta x}\right), \tag{4.60}$$

$$\frac{u_{j-\frac{1}{2}}^{n+\frac{1}{2}} - \frac{1}{2}(u_{j}^{n} + u_{j-1}^{n})}{\frac{1}{2}\Delta t} = -c\left(\frac{u_{j}^{n} - u_{j-1}^{n}}{\Delta x}\right), \tag{4.61}$$

and then use these to obtain u_j^{n+1} from

$$\frac{u_j^{n+1} - u_j^{n}}{\Delta t} = -c\left(\frac{u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - u_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x}\right). \tag{4.62}$$

Note that (4.63) is "centered in time." If (4.60) and (4.61) are substituted into (4.62), we recover (4.59). This derivation helps us to rationalize why it is possible to obtain second-order accuracy in time with this two-time-level scheme.

To test the stability of the Lax-Wendroff scheme, we use von Neumann's method. Assuming $u_j = \hat{u}e^{ikj\Delta x}$ in (4.59), we get

$$\hat{u}^{n+1} - \hat{u}^{n} = -\mu\left[i\sin(k\Delta x) + 2\mu\sin^2\left(\frac{k\Delta x}{2}\right)\right]\hat{u}^{n}. \tag{4.63}$$

Here we have used the trigonometric identity $2\sin^2\left(\frac{\theta}{2}\right) = 1 - \cos\theta$. The amplification factor is

$$\lambda = 1 - 2\mu^2\sin^2\left(\frac{k\Delta x}{2}\right) - i\mu\sin(k\Delta x), \tag{4.64}$$

so that

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

$$= \left[1 - 4\mu^2(1-\mu^2)\sin^4\left(\frac{k\Delta x}{2}\right)\right]^{\frac{1}{2}}. \tag{4.65}$$

If $|\mu| < 1$, $|\lambda| < 1$ and the scheme is dissipative. Fig. 4.3 shows how $|\lambda|^2$ depends on μ and L. The scheme strongly but selectively damps the short waves. Compare with the

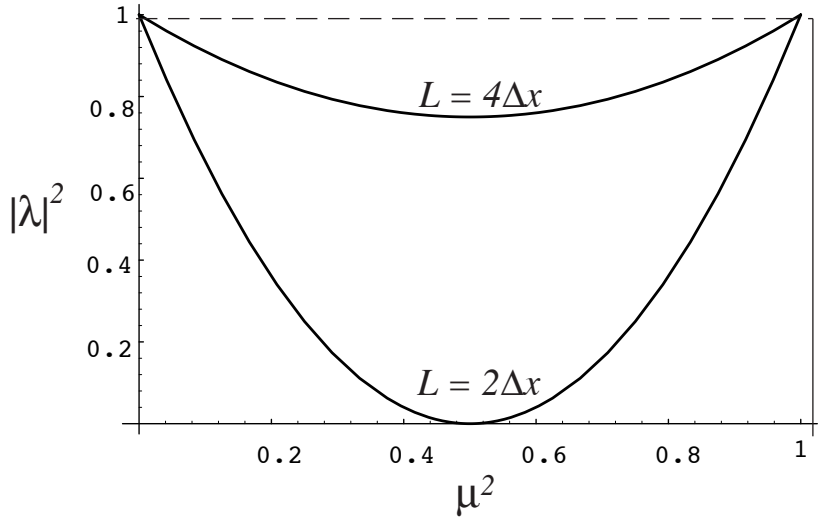


Figure 4.3: The amplification factor for the Lax–Wendroff scheme, for two different wavelengths, plotted as a function of μ^2. Compare with Fig. 2.4.

corresponding plot for the upstream scheme, given earlier. The Lax-Wendroff scheme is comparably dissipative for the shortest wave, but less dissipative for the longer waves.

There are also various implicit schemes, such as the trapezoidal implicit scheme, which are neutral and unconditionally stable, so that in principle any Δt can be used if the error in phase can be tolerated. Such implicit schemes have the drawback that an iterative procedure may be needed to solve the system of equations involved. In many cases, the iterative procedure may take as much computer time as a simpler non-iterative scheme with a smaller Δt.

4.4 Computational dispersion

Consider the differential-difference equation

$$\frac{du_j}{dt} + c\left(\frac{u_{j+1} - u_{j-1}}{2\Delta x}\right) = 0 \,. \tag{4.66}$$

Using $u_j = \hat{u} e^{\imath k j \Delta x}$, as before, we can write (4.66) as

$$\frac{d\hat{u}_j}{dt} + cik\frac{\sin(k\Delta x)}{k\Delta x}\hat{u}_j = 0 \,. \tag{4.67}$$

If we had retained the differential form (4.1), we would have obtained $\frac{\partial \hat{u}}{\partial t} + cik\hat{u} = 0$. Comparison with (4.67) shows that the phase speed is not simply c, as with (4.1), but c^*, given by

$$c^* \equiv c\frac{\sin(k\Delta x)}{k\Delta x} \,. \tag{4.68}$$

Because c^* depends on the wave number k, we have *computational dispersion* that arises from the space differencing. Note that the true phase speed, c, is independent of k. A plot of c^*/c versus $k\Delta x$ is given by the upper curve in Fig. 4.4. (The second (lower) curve in the

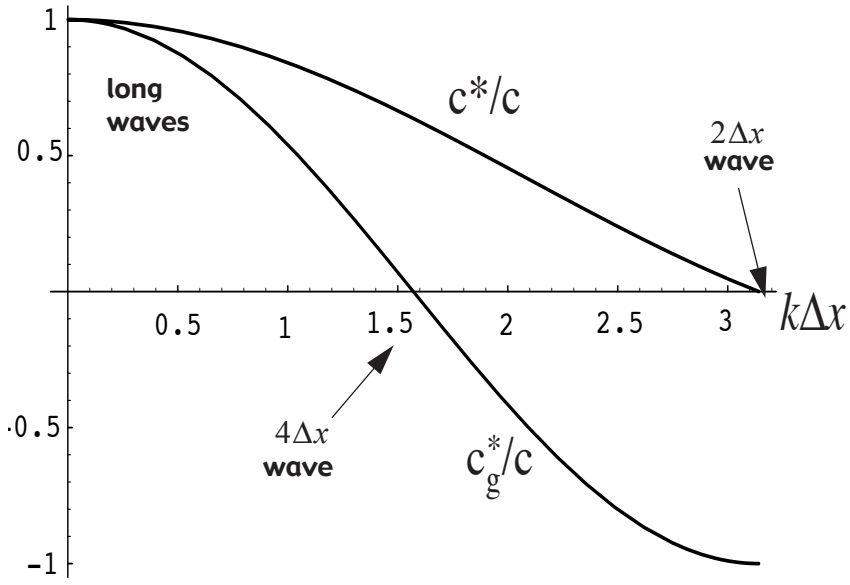


Figure 4.4: **The ratio of the computational phase speed to the true phase speed, and also the ratio of the computational group speed to the true group speed, both plotted as functions of wave number.**

figure, which illustrates the computational group velocity, is discussed later.)

If k_s is defined by $k_s\Delta x = \pi$, then $L_s \equiv \frac{2\pi}{k_s} = 2\Delta x$ is the smallest wave length that our grid can resolve. Therefore, we need only be concerned with $0 < k\Delta x < \pi$. Because $k_s\Delta x = \pi$, $c^* = 0$ for this wave, and so *the shortest possible wave is stationary!* This is actually obvious from the form of the space difference. Since $c^* < c$ for all k, all waves move slower than they should according to the exact equation. Moreover, if we have a number of wave components superimposed on one another, each component moves with a different phase speed, depending on its wave number. The total "pattern" formed by the superimposed waves will break apart, as the waves separate from each other.

Now we briefly digress to explain the concept of group velocity, in the context of the continuous equations. Suppose that we have a superposition of two waves, with slightly different wave numbers k_1 and k_2, respectively. Define

$$k \equiv \frac{k_1 + k_2}{2},\ c \equiv \frac{c_1 + c_2}{2},\ \Delta k \equiv \frac{k_1 - k_2}{2},\ \Delta(kc) \equiv \frac{k_1 c_1 - k_2 c_2}{2}. \tag{4.69}$$

See Fig. 4.5. Note that $k_1 = k + \Delta k$, and $k_2 = k - \Delta k$. Similarly, $c_1 = c + \Delta c$.

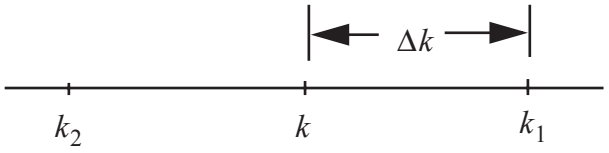


Figure 4.5: Sketch defining notation used in the discussion of the group velocity.

$c_2 = c - \Delta c$. You should be able to show that

$$k_1 c_1 \cong kc + \Delta(kc) \text{ and } k_2 c_2 \cong kc - \Delta(kc). \tag{4.70}$$

Here we neglect terms involving the product $\Delta k \Delta c$. This is acceptable when $k_1 \cong k_2$ and $c_1 \cong c_2$. Using (4.70), we can write the sum of two waves, each with the same amplitude, as

$$
\begin{aligned}
&\exp[ik_1(x-c_1t)] + \exp[ik_2(x-c_2t)] \\
&\cong \exp(i\{(k+\Delta k)x - [kc+\Delta(kc)]t\}) + \exp(i\{(k-\Delta k)x - [kc-\Delta(kc)]t\}) \\
&= \exp[ik(x-ct)](\exp\{i[\Delta kx - \Delta(kc)]t\} + \exp\{-i[\Delta kx - \Delta(kc)]t\}) \\
&= 2\cos[\Delta kx - \Delta(kc)t]\exp[ik(x-ct)] \\
&= 2\cos\left\{\Delta k\left[x - \frac{\Delta(kc)}{\Delta k}t\right]\right\}\exp[ik(x-ct)] \quad .
\end{aligned}
\tag{4.71}
$$

If Δk is small, the factor $\cos\left\{\Delta k\left[x - \frac{\Delta(kc)}{\Delta k}t\right]\right\}$ may appear schematically as the outer, slowly varying envelope in Fig. 4.6. The envelope "modulates" wave k, which is represented by the inner, rapidly varying curve in the figure. The short waves move with phase speed c,

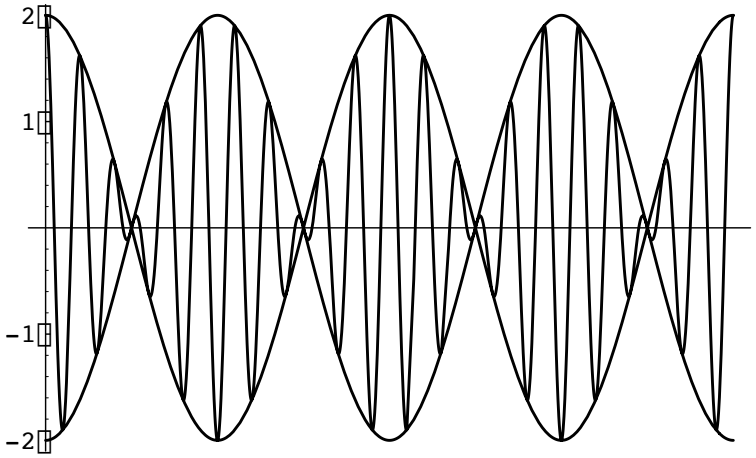


Figure 4.6: **Sketch used to illustrate the concept of group velocity. The short waves are modulated by longer waves.**

but the wave packets, i.e., the envelopes of the short waves, move with speed $\frac{\Delta(kc)}{\Delta k}$. The differential expression $\frac{d(kc)}{dk} \equiv c_g$ is called the "group velocity." Note that $c_g = c$ if c does not depend on k. For the problem at hand, i.e., advection, the "right answer" is $c_g = c$, i.e. the group velocity and phase velocity are the same. For this reason, we usually do not discuss the group velocity for advection.

With our finite-difference scheme, however, we have

$$
c_g^* = \frac{d(kc^*)}{dk} = c\frac{d}{dk}\left(\frac{\sin k\Delta x}{\Delta x}\right) = c\cos k\Delta x \ . \tag{4.72}
$$

A plot of c_g^* versus $k\Delta x$ is given in Fig. 4.4. Note that $c_g^* = 0$ for the $4\Delta x$ wave, and is negative for the $2\Delta x$ wave. This means that wave groups with wavelengths between $L = 4\Delta x$ and $L = 2\Delta x$ have negative group velocities. Very close to $L = 2\Delta x$, c_g^* actually approaches $-c$, when in reality it should be equal to c for all wavelengths. For all waves, $c_g^* < c^* < c = c_g$. This problem arises from the space differencing; it has nothing to do with time differencing.

Fig. 4.7, which is a modified version of Fig. 4.6, illustrates this phenomenon in a different way, for the particular case $L = 2\Delta x$. Consider the upper solid curve and the thick

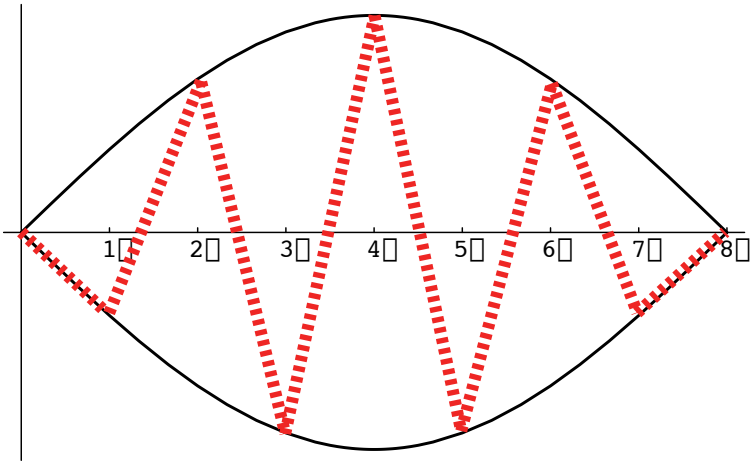


Figure 4.7: Yet another sketch used to illustrate the concept of group velocity. The short wave has wavelength 2Δx.

dashed curve. If we denote points on the thick dashed curve (corresponding to our solution with $L = 2\Delta x$) by u_j, and points on the upper solid curve (the envelope of the thick dashed curve, moving with speed c^*_g) by U_j, we see that

$$U_j = (-1)^j u_j . \tag{4.73}$$

(This is true only for the particular case $L = 2\Delta x$.) Using (4.73), Eq. (4.66) can be rewritten as

$$\frac{\partial U_j}{\partial t} + (-c)\left(\frac{U_{j+1} - U_{j-1}}{2\Delta x}\right) = 0 . \tag{4.74}$$

Eq.(4.74) shows that the upper solid curve will move with speed -*c*.

Recall that when we introduce time differencing, the computed phase change per

time step is generally not equal to $-kc\Delta t$. This leads to changes in c^* and c_g^*, although the formulas discussed above remain valid for $\Delta t \to 0$.

We now present an *analytical* solution of (4.66), which illustrates dispersion error in a very clear way, following an analysis by Matsuno (1966). If we write (4.66) in the form

$$2\frac{du_j}{d\left(\frac{tc}{\Delta x}\right)} = u_{j-1} - u_{j+1}\ , \tag{4.75}$$

and define a non-dimensional time $\tau \equiv \frac{tc}{\Delta x}$, we obtain

$$2\frac{du_j}{d\tau} = u_{j-1} - u_{j+1}\ . \tag{4.76}$$

This is a recursion formula satisfied by the Bessel functions of the first kind of order j, which are usually denoted by $J_j(\tau)$. (See any handbook of functions.) These functions have the property that $J_0(0) = 1$, and $J_j(0) = 0$ for $j \neq 0$. Because the $J_j(\tau)$ satisfy (4.76), each $J_j(\tau)$ represents the solution at a particular grid point, j, as a function of the nondimensional time, τ.

As an example, set $u_j = J_j(\tau)$, which is consistent with and in fact implies the initial conditions that $u_0(0) = 1$ and $u_j(0) = 0$ for all $j \neq 0$. This initial condition is an isolated "spike" at $j = 0$. The solution of (4.76) for the points $j = 0$, 1, and 2 is illustrated in Fig. 4.8. By using the identity

$$J_{(-j)} = (-1)^j J_j\ , \tag{4.77}$$

we can obtain the solution at the points $j = -1$, -2, -3, etc.

The solution of (4.76) for $\tau = 5$ and $\tau = 10$, for $-15 \le j \le 15$, with these "spike" initial conditions, is shown in Fig. 4.9, which is taken from a paper by Matsuno (1966). Computational dispersion, schematically illustrated earlier in Fig. 4.4 and Fig. 4.7, is seen directly here. The figure also shows that c_g is negative for the shortest wave.

A similar type of solution is shown in Fig. 4.10, which is taken from a paper by Wurtele (1961). Here the initial conditions are slightly different, namely,

$$u_{-1} = 1,\ \ u_0 = 1,\ \ u_1 = 1 \text{ and } u_j = 0 \text{ for } j \le -2,\ \ j \ge 2\ .$$

This is a "top hat" or "box" initial condition. We can construct it by combining

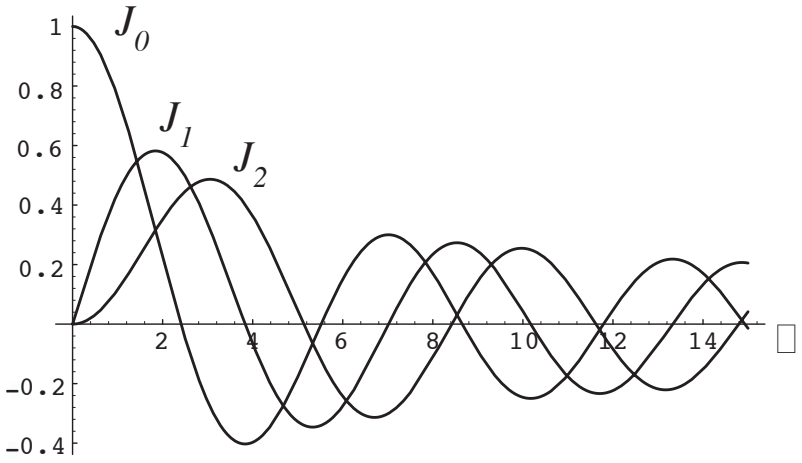


Figure 4.8: **The time evolution of the solution of (4.76) at grid points *j* = 0, 1, and 2.**

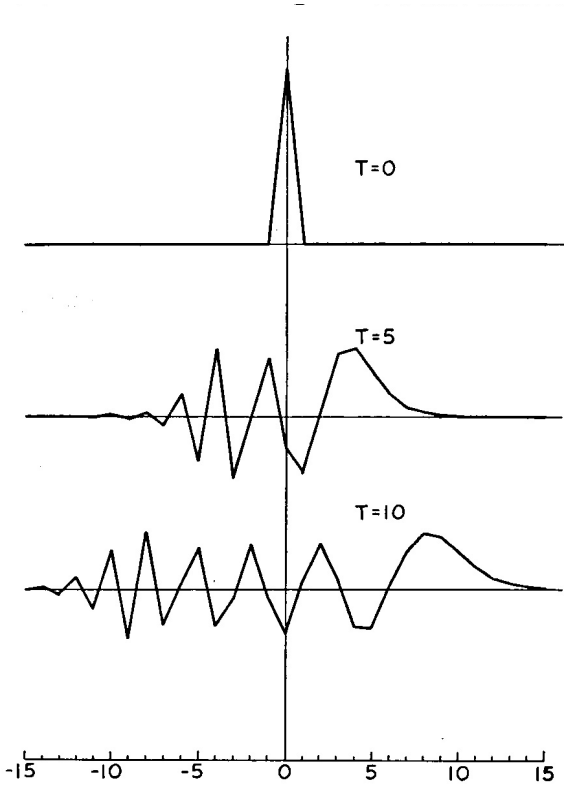


Figure 4.9: **The solution of (4.76) for *t* = 5 and t = 10 for *j* in the range −15 to 15, with "spike" initial conditions. From Matsuno (1966).**

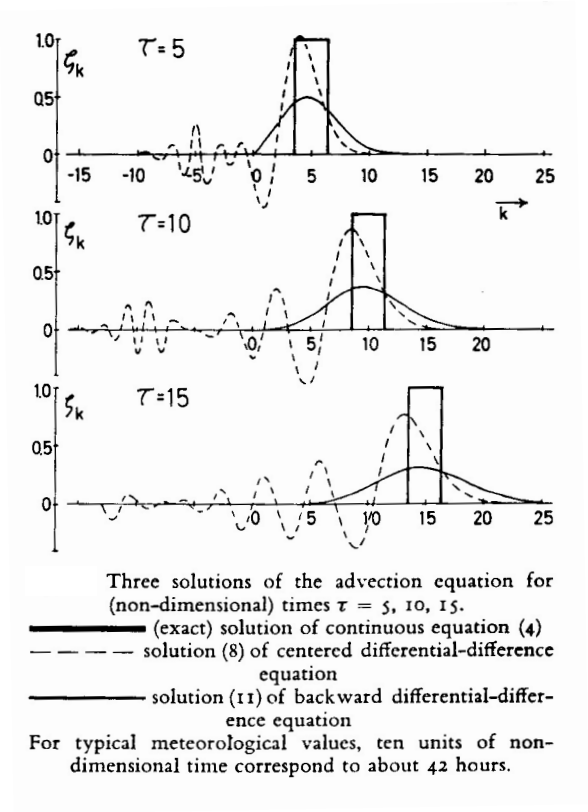
$$J_{j-1}(0) = 1 \text{ for } j = 1 \text{ and zero elsewhere,}$$

$J_j(0) = 1$ for $j = 0$ and zero elsewhere,

$J_{j+1}(0) = 1$ for $j = -1$ and zero elsewhere,

so that the full solution is given by

$$u_j(\tau) = J_{j-1}(\tau) + J_j(\tau) + J_{j+1}(\tau) . \tag{4.78}$$



Three solutions of the advection equation for (non-dimensional) times $\tau = 5, 10, 15$.
(exact) solution of continuous equation (4)
solution (8) of centered differential-difference equation
solution (11) of backward differential-difference equation
For typical meteorological values, ten units of non-dimensional time correspond to about 42 hours.

Figure 4.10: The solution of (4.73) with "box" initial conditions. From Wurtele (1961).

Dispersion is evident again in Fig. 4.10. The dashed curve is for centered space differencing, and the solid curve is for an uncentered scheme, which corresponds to the upstream scheme. (The solution for the uncentered case is given in terms of the Poisson distribution rather than Bessel functions; see Wurtele's paper for details.) The principal disturbance moves to the right, but the short-wave components move to the left.

Do not confuse computational dispersion with instability. A noisy solution is not necessarily unstable. At least initially, both dispersion and instability can lead to "noise." In the case of dispersion, the waves are not growing in amplitude, but are becoming separated from one another ("dispersing"), each at its own speed.

4.5 The effect s of fourth-order space differencing on the phase speed

As discussed in Chapter 2, the fourth-order difference quotient takes the form

$$\left(\frac{\partial u}{\partial x}\right)_j = \frac{4}{3}\left(\frac{u_{j+1}-u_{j-1}}{2\Delta x}\right) - \frac{1}{3}\left(\frac{u_{j+2}-u_{j-2}}{4\Delta x}\right) + O[(\Delta x)^4] \tag{4.79}$$

Recall that in our previous discussion concerning the second-order scheme, we derived an expression for the phase speed of the numerical solution given by

$$c^* = c\left(\frac{\sin k\Delta x}{k\Delta x}\right). \tag{4.80}$$

Now we can also derive a similar equation for this fourth-order scheme. It is

$$c^* = c\left(\frac{4}{3}\frac{\sin k\Delta x}{k\Delta x} - \frac{1}{3}\frac{\sin 2k\Delta x}{2k\Delta x}\right). \tag{4.81}$$

Fig. 4.11 shows a graph of c^*/c versus $k\Delta x$ for each scheme. We see that the fourth-order scheme gives a considerable improvement in the accuracy of the phase speed, for long waves. There is no improvement for wavelengths close to $L = 2\Delta x$, however, and the problems that we have discussed in connection with the second-order schemes become more complicated with the fourth-order scheme.

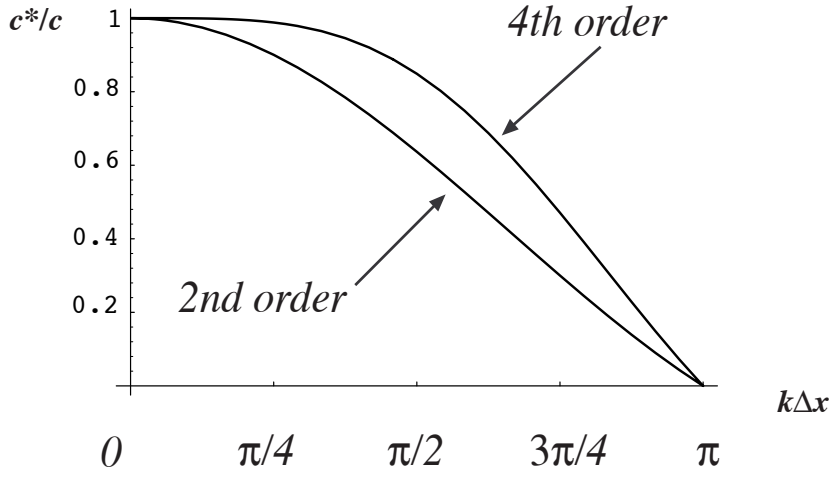


Figure 4.11: The ratio of the computational phase speed, c^*, to the true phase speed, c, plotted as a function of $k\Delta x$, for the second-order and fourth-order schemes.

4.6 Space-uncentered schemes

One way in which computational dispersion can be reduced in the numerical solution

of (4.1) is to use uncentered space differencing, as, for example, in the upstream scheme. Recall that in Chapter 2 we defined and illustrated the concept of the "domain of dependence." By reversing the idea, we can define a "domain of influence." For example, the domain of influence for explicit non-iterative space-centered schemes expands in time as is shown by the union of Regions I and II in Fig. 4.12.

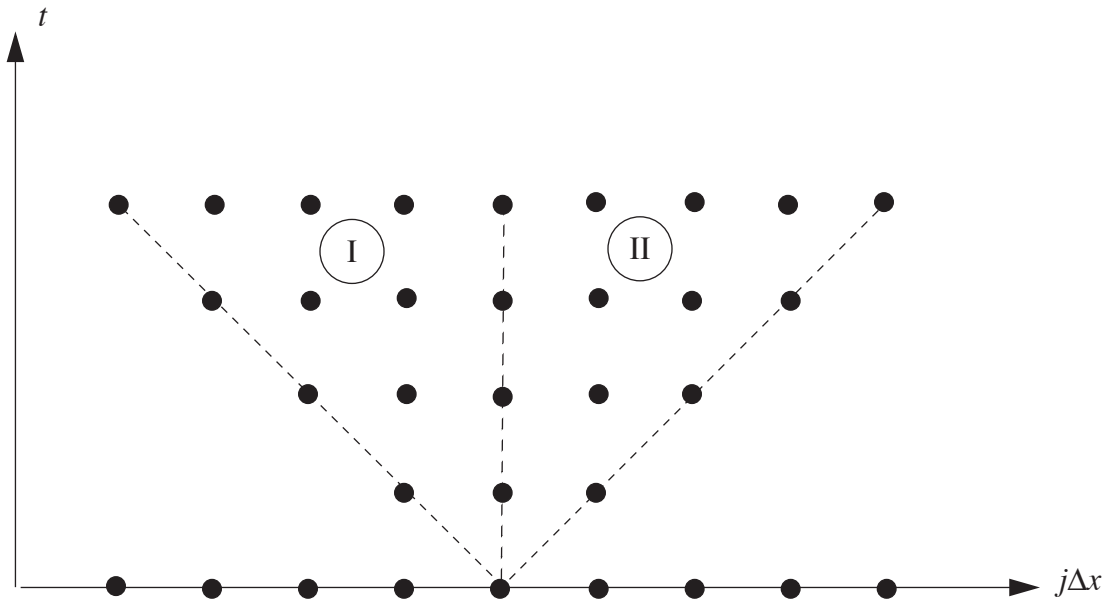


Figure 4.12: The domain of influence for explicit non-iterative space-centered schemes expands in time, as is shown by the union of Regions I and II.

The "upstream scheme," given by

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \text{ for } c > 0\,. \tag{4.82}$$

or

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_j^n}{\Delta x} = 0 \text{ for } c < 0\,. \tag{4.83}$$

is an example of a space-uncentered scheme, and has already been discussed. As shown earlier, we can write (4.83) as

$$u_j^{n+1} = (1-\mu)u_j^n + \mu u_{j-1}^n\,, \tag{4.84}$$

which has the form of an interpolation. Obviously, for the upstream scheme, Region II alone is the domain of influence when $c > 0$, and Region I alone is the domain of influence when $c < 0$. This is good. The scheme produces strong damping, however, as shown in Fig. 4.10. The damping results from the linear interpolation. Although we can reduce the undesirable

effects of computational dispersion by using the upstream scheme, usually the disadvantages of the damping outweigh the advantages of reduced dispersion.

As discussed earlier, the stability condition for the upstream scheme is $\mu \le 1$. When this stability condition is met, (4.84) guarantees that

$$Min\{u_j^n, u_{j-1}^n\} \le u_j^{n+1} \le Max\{u_j^n, u_{j-1}^n\} \ . \tag{4.85}$$

This means that u_j^{n+1} cannot be smaller than the smallest value of u_j^n, or larger than the largest value of u_j^n. The finite-difference advection process associated with the upstream scheme cannot produce any new maxima or minima. As discussed in the introduction to this chapter, real advection also has this property.

In particular, real advection cannot produce negative values of u if none are present initially, and neither can the upstream scheme, provided that $\mu \le 1$. This means that *the upstream scheme is a sign-preserving scheme* when the stability criterion is satisfied. This is very useful if the advected quantity is intrinsically non-negative, e.g. the mixing ratio of some trace species. Even better, *the upstream scheme is a monotone scheme* when the stability criterion is satisfied. This means that it cannot produce new maxima or minima, like those associated with the dispersive ripples seen in Fig. 4.10. This monotone property is expressed by (4.85).

It is easy to show that all monotone schemes are sign-preserving schemes. The converse is not true.

As discussed by Smolarkiewicz (1991), sign-preserving schemes tend to be stable. To see why, suppose that we have a linearly conservative scheme for a variable q, so that

$$\sum_i q_i^n \ = \ \sum_i q_i^0 , \tag{4.7}$$

where the sum represents a sum over the whole spatial domain, the superscripts $n > 0$ and 0 represent two time levels. For simplicity we indicate only a single spatial dimension here, but the following argument holds for any number of dimensions. Suppose that the scheme that takes us from q^0 to q^n through n time steps is sign-preserving and conserves q. If q_i^0 is of one sign everywhere, it follows from (4.7) that

$$\sum_i \left| q_i^n \right| \ = \ \sum_i \left| q_i^0 \right| . \tag{4.8}$$

Recall that for an arbitrary variable A, we have

$$\sum_i (A_i)^2 \ge \left(\sum_i |A_i| \right)^2 . \tag{4.9}$$

Then from (4.8) and (4.9) we see that

$$\sum_i (q_i^n)^2 \le \left(\sum_i \left| q_i^0 \right| \right)^2 . \tag{4.10}$$

Note that the right-hand side of (4.10) is a constant. Eq. (4.10) demonstrates that $(q_i^n)^2$ is bounded for all time, and so it proves stability by the energy method discussed in Chapter 2. The essence of (4.10) is that there is an upper bound on $\sum (q_i^n)^2$. This bound is rather weak, however; try some numbers to see for yourself. So, although (4.10) does demonstrate absolute stability, it does not ensure good behavior!

In the preceding discussion we assumed that q_i^0 is everywhere of one sign, but this assumption is not really necessary. For variable-sign fields, a similar result can be obtained by decomposing q into positive and negative parts, i.e.

$$q = q^+ + q^- . \tag{4.11}$$

We can consider that q^+ is positive where q is positive, and zero elsewhere; and similarly that q^- is negative where q is negative, and zero elsewhere. The total of q is then the sum of the two parts, as stated in (4.11). Advection of q is equivalent to advection of q^+ and q^- separately. If we apply a sign-preserving scheme to each part, then each of these two advections is stable by the argument given above, and so the advection of q itself is also stable.

Although the upstream scheme is sign-preserving, it is only first-order accurate and strongly damps, as we have seen. Can we find more accurate schemes which are sign-preserving or nearly so? A spurious negative value is customarily called a "hole." Second-order advection schemes that produce relatively few holes are given by (4.13) with either the geometric mean given by (4.34), or the harmonic mean given by (4.35). Both of these schemes have the property that as either A_j or A_{j+1} goes to zero, $A_{j+\frac{1}{2}}$ also goes to zero. If the time step were infinitesimal, this would be enough to prevent the property denoted by A from changing sign. Because time-steps are finite in real models, however, such schemes do not completely prevent hole production, although they do tend to minimize it.

Better results can be obtained as follows: Replace (4.13) by

$$\frac{d}{dt}(m_j A_j) + \frac{\left[(mu)^+_{j+\frac{1}{2}} A^+_{j+\frac{1}{2}} + (mu)^-_{j+\frac{1}{2}} A^-_{j+\frac{1}{2}} \right] - \left[(mu)^+_{j-\frac{1}{2}} A^+_{j-\frac{1}{2}} + (mu)^-_{j-\frac{1}{2}} A^-_{j-\frac{1}{2}} \right]}{\Delta x_j} = 0 \quad ,$$

(4.86)

where

$$(mu)^{+}_{j+\frac{1}{2}} \equiv \frac{(mu)_{j+\frac{1}{2}} + \left|(mu)_{j+\frac{1}{2}}\right|}{2} \geq 0 \ , \tag{4.87}$$

$$(mu)^{-}_{j+\frac{1}{2}} \equiv \frac{(mu)_{j+\frac{1}{2}} - \left|(mu)_{j+\frac{1}{2}}\right|}{2} \leq 0 \ , \tag{4.88}$$

$$A^{+}_{j+\frac{1}{2}} \equiv \frac{2A_j^{n+1}A_{j+1}^{n}}{(A_j^n + A_{j+1}^n)} \ , \tag{4.89}$$

$$A^{-}_{j+\frac{1}{2}} \equiv \frac{2A_j^{n}A_{j+1}^{n+1}}{(A_j^n + A_{j+1}^n)} \ . \tag{4.90}$$

Note that this scheme is implicit and must be solved as a coupled system over all grid points, but it remains linear. The "upstream" values of A are implicit in the numerators of (4.89) and (4.90), while the "downstream" values are explicit. Care must be taken to avoid division by zero when both A^n_{j+1} and A^n_j are zero; in such a case we simply set $A^{+}_{j+\frac{1}{2}} = A^{-}_{j+\frac{1}{2}} = 0$.

4.12 Hole filling

If a non-sign-preserving advection scheme is used, and holes are produced, then a procedure is needed to fill the holes. To make the discussion concrete, we consider here a scheme to fill "water holes", in a model which advects water vapor mixing ratio.

Simply replacing negative mixing ratios by zero is unacceptable because it leads to a systematic increase in the mass-weighted total water. Hole-filling schemes therefore "borrow" mass from elsewhere on the grid. They take from the rich, and give to the poor.

There are many possible borrowing schemes. Some borrow systematically from nearby points, but of course borrowing is only possible from neighbors with positive mixing ratios, and it can happen that the nearest neighbors of a "holy" grid cell have insufficient water to fill the hole. Logic can be invented to deal with such issues, but hole-fillers of this type tend to be complicated and computationally slow.

An alternative is to borrow from *all* points on the mesh that have positive mixing ratios. The "global multiplicative hole-filler" is a particularly simple and computationally fast algorithm. The first step is to add up all of the negative water on the mesh:

$$N \equiv \sum_{\text{where } q_j < 0} m_j q_j \le 0 \ . \tag{4.91}$$

Here q_j is the mixing ratio in grid cell j, and m_j is the mass of dry air in that grid cell (in kg, say), so that the product $m_j q_j$ is the mass of water in the cell. Note that m_j is *not* the density of dry air in the cell; rather it is the product of the density of dry air and the volume of the cell. The total amount of water on the mesh is given by

$$T \equiv \sum_{\text{all points}} m_j q_j \ . \tag{4.92}$$

Both T and N have the dimensions of mass. Define the nondimensional ratio

$$\Phi \equiv \frac{T+N}{T} \le 1 \ ; \tag{4.93}$$

normally Φ is just very slightly less than one, because there are only a few holes and they are not very "deep". We replace all negative values of q_j by zero, and then set

$$q_j^{\text{new}} = \Phi q_j \ . \tag{4.94}$$

In this way, we are ensured of the following:

- No negative values of q_j remain on the mesh.
- The total mass of water in the adjusted state is the same as that in the "holy" state.
- Water is borrowed most heavily from grid cells with large mixing ratios, and least from cells with small mixing ratios.

Hole-filling is ugly. Any hole-filling procedure is necessarily somewhat arbitrary, because we cannot mimic any natural process; nature has no holes to fill. In addition, hole-filling tends to be "quasi-diffusive" because it remove water from wet cells and adds it to dry cells, so that it reduces the total variance of the mixing ratio. The best approach is to choose an advection scheme that does not make holes in the first place. At the very least, we should insist that an advection scheme digs holes slowly, so that, like a Maytag repairman, the hole-filler will have very little work to do.

4.13 Flux-corrected transport

The upstream scheme is monotone and sign-preserving, but, unfortunately, as we have seen, it is strongly damping. Damping is in fact characteristic of all monotone and sign-preserving schemes. Much work has been devoted to designing monotone or sign-preserving schemes that produce *as little damping as possible*. The following discussion, abstracted from the paper of Zalesak (1979), indicates how this is done.

Monotone and sign-preserving schemes can be derived by using the approach of "flux-corrected transport," sometimes abbreviated as FCT, which was invented by Boris and Book (1973) and extended by Zalesak (1979) and many others. Suppose that we have a "high-order" advection scheme, represented schematically by

$$\psi_i^{n+1} = \psi_i^n - \left(FH_{i+\frac{1}{2}} - FH_{i-\frac{1}{2}} \right) . \tag{4.95}$$

Here *FH* represents the "high-order" fluxes associated with the scheme. Note that (4.95) is in "conservation" form, and that forward time-differencing has been used. Suppose that we have at our disposal a monotone or sign-preserving low-order scheme, whose fluxes are denoted by $FL_{i+\frac{1}{2}}$. This low-order scheme could be, for example, the upstream scheme. (From this point on we say "monotone" with the understanding that we mean "monotone or sign-preserving.") We can write

$$FH_{i+\frac{1}{2}} \equiv FL_{i+\frac{1}{2}} + A_{i+\frac{1}{2}} . \tag{4.96}$$

Here $A_{i+\frac{1}{2}}$ is a "residual" flux, sometimes called an "anti-diffusive" flux. Eq. (4.96) is essentially the definition of $A_{i+\frac{1}{2}}$. According to (4.96), the high-order flux is the low-order flux plus an anti-diffusive correction. We know that the low-order flux is diffusive in the sense that it damps the solution, but on the other hand by assumption it is monotone. The high-order flux is presumably less diffusive, and more accurate, but does not have the nice monotone property that we want.

Suppose that we take a time-step using the low-order scheme. Let the result be denoted by Ψ_i^{n+1}, i.e.

$$\Psi_i^{n+1} = \psi_i^n - \left(FL_{i+\frac{1}{2}} - FL_{i-\frac{1}{2}} \right) . \tag{4.97}$$

Since by assumption the low-order scheme is monotone, we know that

$$\psi_i^{MAX} \geq \Psi_i^{n+1} \geq \psi_i^{MIN} , \tag{4.98}$$

where ψ_i^{MAX} and ψ_i^{MIN} are suitably chosen upper and lower bounds, respectively, on the value of ψ within the grid-box in question. For instance, ψ_i^{MIN} might be zero, if ψ is a non-negative scalar like the water vapor mixing ratio. Other possibilities will be discussed below.

There are two important points in connection with the inequalities in (4.98). First, the inequalities must actually be true for the low-order scheme that is being used. Second, the

inequalities should be strong enough to ensure that the solution obtained is in fact monotone.

From (4.95), (4.96), and (4.97) it is easy to see that

$$\psi_i^{n+1} = \Psi_i^{n+1} - \left(A_{i+\frac{1}{2}} - A_{i-\frac{1}{2}} \right). \tag{4.99}$$

This simply says that we can obtain the high-order solution from the low-order solution by adding the anti-diffusive fluxes.

We now define some coefficients, denoted by $C_{i+\frac{1}{2}}$, and "scaled-back" anti-diffusive fluxes, denoted by $\hat{A}_{i+\frac{1}{2}}$, such that

$$\hat{A}_{i+\frac{1}{2}} \equiv C_{i+\frac{1}{2}} A_{i+\frac{1}{2}} \ . \tag{4.100}$$

In place of (4.99), we use

$$\psi_i^{n+1} = \Psi_i^{n+1} - \left(\hat{A}_{i+\frac{1}{2}} - \hat{A}_{i-\frac{1}{2}} \right). \tag{4.101}$$

To see the idea, consider two limiting cases. If $C_{i+\frac{1}{2}} = 1$, then $\hat{A}_{i+\frac{1}{2}} = A_{i+\frac{1}{2}}$, and so (4.101) will reduce to (4.99) and so will simply give the high-order solution. If $C_{i+\frac{1}{2}} = 0$, then $\hat{A}_{i+\frac{1}{2}} = 0$, and so (4.101) will simply give the low-order solution. We enforce

$$0 \le C_{i+\frac{1}{2}} \le 1 , \tag{4.102}$$

and try to make $C_{i+\frac{1}{2}}$ as close to one as possible, so that we get as much as possible of the high-order scheme and as little as possible of the low-order scheme, but we require that

$$\psi_i^{MAX} \ge \psi_i^{n+1} \ge \psi_i^{MIN} \tag{4.103}$$

be satisfied. Compare (4.103) with (4.98). We can always ensure that (4.103) will be satisfied by taking $C_{i+\frac{1}{2}} = 0$; this is the "worst case." Quite often it may happen that (4.103) is

satisfied for $C_{i+\frac{1}{2}} = 1$; that is the "best case."

It remains to specify the upper and lower bounds that appear in (4.103) and (4.98). Zalesak (1979) proposed limiting ψ_i^{n+1} so that it is bounded by the largest and smallest values of its neighbors at time level n, and also by the largest and smallest values of the low-order solution at time level n+1. In other words, he took

$$\psi_i^{MAX} = Max\{\psi_{i-1}^n, \psi_i^n, \psi_{i+1}^n, \Psi_{i-1}^{n+1}, \Psi_i^{n+1}, \Psi_{i+1}^{n+1}\}, \tag{4.104}$$

and

$$\psi_i^{MIN} = Min\{\psi_{i-1}^n, \psi_i^n, \psi_{i+1}^n, \Psi_{i-1}^{n+1}, \Psi_i^{n+1}, \Psi_{i+1}^{n+1}\}. \tag{4.105}$$

Smolarkiewicz (1991) shows how (4.104) and (4.105) can be combined with (4.102) and (4.103) to obtain the largest feasible anti-diffusive fluxes.

The "limiter" denoted by (4.104) and (4.105) is not unique. Other possibilities are discussed by Smolarkiewicz (1991).

Our analysis of FCT schemes has been given in terms of one spatial dimension, but all of the discussion given above can very easily be extended to two or three dimensions, without time splitting. The literature on FCT schemes is very large and rapidly growing, although their application to atmospheric science is still fairly new.

FCT schemes are, philosophically, not that different from hole-fillers. The high-order scheme makes a hole, and the low-order scheme is used to fill it, immediately, before the end of the time step. Hole? What hole?

4.14 Lagrangian schemes

Lagrangian schemes, in which particles are tracked through space without the use of an Eulerian grid, have been used in the atmospheric sciences, as well as other fields including astrophysics and weapons physics (e.g. Mesinger, 1971; Trease, 1988; Monaghan, 1992; Norris, 1996; Haertel and Randall, 2001). The Lagrangian approach has a number of attractive features:

- The pdf of the advected quantity (and all functions of the advected quantity) can be preserved "exactly" under advection. Here "exactly" is put in quotation marks because of course the result is actually approximate in the sense that, in practice, only a finite number of particles can be tracked.

- As a consequence of the first point mentioned above, Lagrangian schemes are monotone and positive definite.

- Time steps can be very long without triggering computational instability, although of course long time steps still lead to large truncation errors.

- Aliasing instability does not occur with Lagrangian schemes. Aliasing instability will be discussed later.

On the other hand, Lagrangian schemes encounter a number of practical difficulties. Some of these problems have to do with "voids" that develop, i.e. regions with few particles. Others arise from the need to compute spatial derivatives (e.g. the pressure gradient force, which is needed to compute the acceleration of each particle from the equation of motion) on the basis of a collection of particles that can travel literally anywhere within the domain, in an uncontrolled way.

One class of Lagrangian schemes, called "smoothed particle hydrodynamics" (SPH), has been widely used by the astrophysical research community and is reviewed by Monaghan (1992). The approach is to specify a way to compute a given field at any point in space, given the value of the field at a collection of particles which can be located anywhere in the domain. For an arbitrary field A, let

$$A(\boldsymbol{r}) = \int A(\boldsymbol{r}')W(\boldsymbol{r}-\boldsymbol{r}', h)\ d\boldsymbol{r}' , \tag{4.106}$$

where the integration is over the whole domain (e.g. the whole atmosphere), and W is an interpolating "kernel" such that

$$\int W(\boldsymbol{r}-\boldsymbol{r}', h)\ d\boldsymbol{r}' = 1 \tag{4.107}$$

and

$$\lim_{h \to 0} W(\boldsymbol{r}-\boldsymbol{r}', h) = \delta(\boldsymbol{r}-\boldsymbol{r}') , \tag{4.108}$$

where $\delta(\boldsymbol{r}-\boldsymbol{r}')$ is the Dirac delta function. In (4.106) - (4.108), h is a parameter, which is a measure of the "width" of W. We can interpret $W(\boldsymbol{r}-\boldsymbol{r}', h)$ as a "weighting function" that is strongly peaked at $\boldsymbol{r}-\boldsymbol{r}' = 0$. For example, we might use the Gaussian weighting function given by

$$W(\boldsymbol{r}-\boldsymbol{r}', h) = \frac{e^{-[x(\underset{\sim}{r}-\underset{\sim}{r}')^2/h^2]}}{h\sqrt{\pi}} , \tag{4.109}$$

which can be shown to satisfy (4.108).

In a discrete model, we replace (4.106) by

$$A(\boldsymbol{r}) = \sum_b m_b \frac{A_b}{\rho_b} W(\boldsymbol{r}-\boldsymbol{r}_b, h) . \tag{4.110}$$

Here the index b denotes a particular particle, m_b is the mass of the particle, and ρ_b is the density of the particle. To see what is going on in (4.110), consider the case $A \equiv \rho$. Then

(4.110) reduces to

$$\rho(\boldsymbol{r}) = \sum_b m_b W(\boldsymbol{r} - \boldsymbol{r}_b, h) , \tag{4.111}$$

which simply says that the density at a point $\boldsymbol{r}$ is a weighted sum of the masses of particles in the vicinity of $\boldsymbol{r}$. In case there are no particles near the point $\boldsymbol{r}$, the density there will be small.

We can now perform spatial differentiation simply by taking the appropriate derivatives of $W(\boldsymbol{r} - \boldsymbol{r}_b, h)$, e.g.

$$\nabla A(\boldsymbol{r}) = \sum_b m_b \frac{A_b}{\rho_b} \nabla W(\boldsymbol{r} - \boldsymbol{r}_b, h) . \tag{4.112}$$

This follows because m_b, A_b, and ρ_b are associated with particular particles and are, therefore, not functions of space.

Further discussion of SPH and related methods is given by Monaghan (1992) and the other references cited above.

4.15 Semi-Lagrangian schemes

Recently there has been considerable interest in a particular family of advection schemes called "semi-Lagrangian schemes" (e.g. Robert et al., 1985; Staniforth and Cote, 1991; Bates et al., 1993; Williamson and Olson, 1994). These schemes are of interest in part because they allow very long time steps, and in part because they can easily maintain such properties as monotonicity.

The basic idea is very simple. At time step $n + 1$, values of the advected field, at the various grid points, are considered to be characteristic of the particles which reside at those points. We ask where those particles were at time step n. This question can be answered by using the (known) velocity field, averaged over the time interval (n, n+1), to track the particles backward in time from their locations at the various specified grid points, at time level n+1, to their "departure points" at time level n. Naturally, the departure points are typically located in between grid cells. The values of the advected field at the departure points, at time level n, can be determined by spatial interpolation. If advection is the only process occurring, then the values of the advected field at the departure points at time level n will be identical to those at the grid points at time level n+1.

As a simple example, consider one-dimensional advection of a variable q by a constant current, c. A particle that resides at $x = x_j$ at time level $t = t^{n+1}$ has a departure point given by

$$(x_{departure})^n_j = x_j - c\Delta t . \tag{4.113}$$

Here the superscript n is used to indicate that the departure point is the location of the particle at time level n. Suppose that $c > 0$, and that

$$x_{j-1} < (x_{departure})^n{}_j < x_j \ . \tag{4.114}$$

Then the simplest linear interpolation for q at the departure point is

$$\begin{aligned}(q_{departure})^n{}_j &= q^n{}_{j-1} + \left[\frac{(x_{departure})^n{}_j - x_{j-1}}{\Delta x}\right](q^n{}_j - q^n{}_{j-1}) \\ &= q^n{}_{j-1} + \left(\frac{\Delta x - c\Delta t}{\Delta x}\right)(q^n{}_j - q^n{}_{j-1}) \\ &= q^n{}_{j-1} + (1-\mu)(q^n{}_j - q^n{}_{j-1}) \\ &= \mu q^n{}_{j-1} + (1-\mu)q^n{}_j\end{aligned} \ . \tag{4.115}$$

Here we assume for simplicity that the mesh is spatially uniform, and $\mu \equiv \frac{c\Delta t}{\Delta x}$, as usual. The semi-Lagrangian scheme uses

$$q^{n+1}{}_j = (q_{departure})^n{}_j \ , \tag{4.116}$$

so we find that

$$q^{n+1}{}_j = \mu q^n{}_{j-1} + (1-\mu)q^n{}_j \ . \tag{4.117}$$

This is simply the familiar upstream scheme. Note that (4.114), which was used in setting up the spatial interpolation, is equivalent to

$$0 < \mu < 1 \ . \tag{4.118}$$

As shown earlier, this is the condition for stability of the upstream scheme.

What if (4.114) is not satisfied? This will be the case if the particle is moving quickly and/or the time step is long or, in other words, if $\mu > 1$. Then we might have, for example,

$$x_{j-a} < (x_{departure})^n{}_j < x_{j-a+1} \ , \tag{4.119}$$

where a is an *integer* greater than 1. For this case, we find in place of (4.115) that

$$(q_{departure})^n{}_j = \hat{\mu} q^n{}_{j-a} + (1-\hat{\mu})q^n{}_{j-a+1} \ , \tag{4.120}$$

where

$$\hat{\mu} \equiv 1 - a + \mu \,. \tag{4.121}$$

Notice that we have assumed again here, for simplicity, that both the mesh and the advecting current are spatially uniform. It should be clear that

$$0 \le \hat{\mu} \le 1 \,. \tag{4.122}$$

For $a = 1$, $\mu = \hat{\mu}$. Eq. (4.116) gives

$$q^{n+1}{}_j = \hat{\mu} q^n{}_{j-a} + (1 - \hat{\mu}) q^n{}_{j-a+1} \,. \tag{4.123}$$

It is easy to prove that we still have computational stability. This means that *the semi-Lagrangian scheme is computationally stable regardless of the size of the time step*. You should also be able to see that the scheme is monotone.

It is clear that the semi-Lagrangian scheme outlined above is very diffusive, because it is more or less equivalent to a "generalized upstream scheme," and we know that the upstream scheme is very diffusive. By using higher-order interpolations, the strength of this computational diffusion can be reduced, although it cannot be eliminated completely.

Is the semi-Lagrangian scheme conservative? To prove that the scheme is conservative, it would suffice to show that it can be written in a "flux form." Note, however, that in deriving the scheme we have used the Lagrangian version of the advective form very directly, by considering the parcel trajectory between the mesh point at time level $n+1$ and the departure point at time level n. Because the advective form is used in their derivations, most semi-Lagrangian schemes are not conservative.

4.16 Two-dimensional advection

Variable currents more or less have to be multi-dimensional. Before we discuss variable currents, in a later chapter, it is useful to consider constant currents in two-dimensions.

Let q be an arbitrary quantity advected, in two dimensions, by a constant basic current. The advection equation is

$$\frac{\partial q}{\partial t} + U\frac{\partial q}{\partial x} + V\frac{\partial q}{\partial y} = 0 \,, \tag{4.124}$$

where U and V are the x and y components of the current, respectively. Let i and j be the indices of grid points in the x and y directions. Replacing $\frac{\partial q}{\partial x}$ and $\frac{\partial q}{\partial y}$ by the corresponding centered difference quotients, we obtain

$$\frac{\partial q_{i,j}}{\partial t} + U\frac{1}{2\Delta x}(q_{i+1,j} - q_{i-1,j}) + V\frac{1}{2\Delta y}(q_{i,j+1} - q_{i,j-1}) = 0 \,. \tag{4.125}$$

Assume that q has the form

$$q_{ij} = R_e\left[Q(t)e^{i(kI\Delta x + lJ\Delta y)} \right], \tag{4.126}$$

where $i \equiv \sqrt{-1}$, and k and l are wave numbers in the x and y directions. Substitution gives the oscillation equation again:

$$\frac{dQ}{dt} = i\omega Q, \qquad \omega \equiv -\left(U\frac{\sin k\Delta x}{\Delta x} + V\frac{\sin l\Delta y}{\Delta y} \right). \tag{4.127}$$

We have already analyzed the oscillation equation in detail, in Chapter 3. When we apply the leapfrog scheme, the stability criterion is $|\Omega| \le 1$, where $\Omega \equiv \omega\Delta t$. Therefore, we must require

$$\left| U\frac{\sin k\Delta x}{\Delta x} + V\frac{\sin l\Delta y}{\Delta y} \right| \Delta t \le 1 . \tag{4.128}$$

Since

$$\left| U\frac{\sin k\Delta x}{\Delta x} + V\frac{\sin l\Delta y}{\Delta y} \right| \Delta t \le \left(\left| U\frac{\sin k\Delta x}{\Delta x} \right| + \left| V\frac{\sin l\Delta y}{\Delta y} \right| \right)\Delta t \le \left(\frac{|U|}{\Delta x} + \frac{|V|}{\Delta y} \right)\Delta t , \tag{4.129}$$

a condition sufficient to satisfy (4.128) is

$$\left(\frac{|U|}{\Delta x} + \frac{|V|}{\Delta y} \right)\Delta t \le 1 . \tag{4.130}$$

If we require the scheme to be stable for all possible k and l, and for all combinations of U and V, then (4.130) is also a necessary condition.

Put

$$|U| = C\cos\alpha , \; |V| = C\sin\alpha , \tag{4.131}$$

where $0 \le \alpha \le \frac{\pi}{2}$. Note that with this definition C is the wind speed, and $C \ge 0$. For $\alpha = 0$, the flow is zonal, and for $\alpha = \pi/2$ it is meridional. Then (4.130) becomes

$$C\left(\frac{\cos\alpha}{\Delta x} + \frac{\sin\alpha}{\Delta y} \right)\Delta t \le 1 . \tag{4.132}$$

In order for the scheme to be stable for any orientation of the current, we must have

$$C\left(\frac{\cos\alpha_m}{\Delta x}+\frac{\sin\alpha_m}{\Delta y}\right)\Delta t \le 1\ , \tag{4.133}$$

where α_m is the "worst-case" α which makes the left hand side of (4.132) a maximum. We can show that α_m satisfies

$$\tan\alpha_m = \frac{\Delta x}{\Delta y}. \tag{4.134}$$

As shown in Fig. 4.13, α_m measures the angle of the "diagonal" across a grid box. For

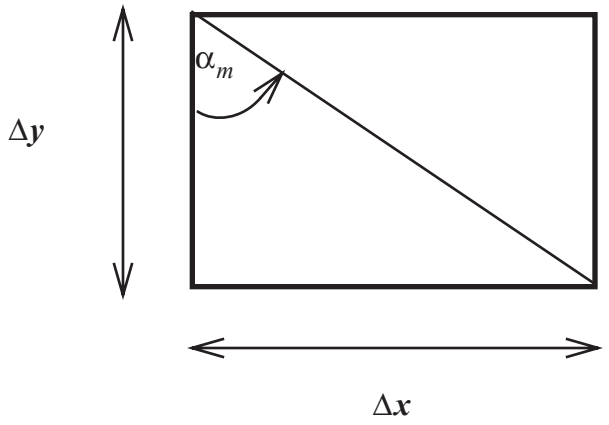


Figure 4.13: Sketch illustrating the angle α_m on a rectangular grid.

instance, when $\frac{\Delta y}{\Delta x} \ll 1$, α_m corresponds to a flow that is mainly meridional, because the worst case is the direction in which the grid cell is "narrowest." As a second example, for $\Delta x = \Delta y$, we get $\alpha_m = \frac{\pi}{4}$. From (4.133) and (4.134) we see that the stability criterion can be written as

$$\frac{C\Delta t}{\sqrt{\Delta x^2+\Delta y^2}}\left(\frac{\Delta y}{\Delta x}+\frac{\Delta x}{\Delta y}\right)\le 1\ . \tag{4.135}$$

In particular, for $\Delta x = \Delta y = d$,

$$\frac{C\Delta t}{d} \le \frac{1}{\sqrt{2}} < 1\ . \tag{4.136}$$

4.17 Summary

Finite-difference schemes for the advection equation can be designed to allow "exact" or "formal" conservation of mass, of the advected quantity itself (such as potential temperature), and of one arbitrary function of the advected quantity (such as the square of the potential temperature). Conservative schemes mimic the "form" of the exact equations. In addition, they are often well behaved computationally. Since bugs often lead to failure to conserve, conservative schemes can be easier to de-bug than non-conservative schemes.

When we solve the advection equation, space-differencing schemes can introduce diffusion-like damping. If this damping is sufficiently scale-selective, it can be beneficial.

Computational dispersion arises from space differencing. It causes waves of different wavelengths to move at different speeds. In some cases, the phase speed can be zero or even negative, when it should be positive. Short waves generally move slower than longer waves. The phase speeds of the long waves are well simulated by the commonly used space-time differencing schemes. The group speed, which is the rate at which a wave "envelope" moves, can also be adversely affected by space truncation errors. Space-uncentered schemes are well suited to advection, which is a spatially asymmetric process, and they can minimize the effects of computational dispersion.

Higher-order schemes simulate the well resolved modes more accurately, but do not improve the solution for the shortest modes (e.g. the $2\Delta x$ modes) and can actually make the problems with the short modes worse, in some ways. Of course, higher-order schemes involve more arithmetic and so are computationally more expensive than lower-order schemes. An alternative is to use a lower-order scheme with more grid points. This may be preferable in many cases.

Problems

1. Find a one-dimensional advection scheme that conserves both A and $\ln(A)$. Keep the time derivative continuous.

2. Adopt the continuity equation

$$\frac{\partial m_j}{\partial t} + \frac{(\hat{mu})_{j+\frac{1}{2}} - (\hat{mu})_{j-\frac{1}{2}}}{\Delta x} = 0 ,$$

and the advection equation

$$\frac{\partial A_j}{\partial t} + \frac{1}{2}\left(u_{j+\frac{1}{2}} + u_{j-\frac{1}{2}}\right)\left(\frac{A_{j+1} - A_{j-1}}{2\Delta x}\right) = 0 .$$

Determine whether or not this scheme conserves the mass-weighted average value of A.

3. Program the following one-dimensional model:

$$\frac{(hA)_j^{n+1} - (hA)_j^{n-1}}{2\Delta t} + \frac{(\hat{hu})_{j+\frac{1}{2}}^{n} \hat{A}_{j+\frac{1}{2}}^{n} - (\hat{hu})_{j-\frac{1}{2}}^{n} \hat{A}_{j-\frac{1}{2}}^{n}}{\Delta x} = 0 ,$$

$$\frac{h_j^{n+1} - h_j^{n-1}}{2\Delta t} + \frac{(\hat{hu})_{j+\frac{1}{2}}^{n} - (\hat{hu})_{j-\frac{1}{2}}^{n}}{\Delta x} = 0 ,$$

$$\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^{n-1}}{2\Delta t} + \left(\frac{k_{j+1}^{n} - k_j^{n}}{\Delta x} \right) + g \left(\frac{h_{j+1}^{n} - h_j^{n}}{\Delta x} \right) = 0 .$$

Use a forward time step for the first step only. Take

$$\Delta x = 10^5 \text{ m} ,$$
$$g = 0.1 \text{ m s}^{-2} ,$$

$$\hat{h}_{j+\frac{1}{2}} = \frac{1}{2}(h_j + h_{j+1}) ,$$

$$k_j = \frac{1}{4}\left(u_{j+\frac{1}{2}}^2 + u_{j-\frac{1}{2}}^2 \right) .$$

Use 100 grid points, with periodic boundary conditions. Let the initial condition be

$$u_{j+\frac{1}{2}} = 0 \quad \text{for all } j ,$$

$$h_j = 1000 + 50 \cdot \sin\left(\frac{2\pi j}{20} \right),$$

$$A_j = 100 + 10 \cdot \cos\left(\frac{2\pi j}{4} \right).$$

Use von Neuman's method to estimate the largest time step that is consistent with numerical stability. Experiment with time steps "close" (within a factor of 2) to

the predicted maximum stable Δt, in order to find a value that is stable in practice.

Run for the following two choices of $\hat{A}_{j+\frac{1}{2}}$:

$$\hat{A}_{j+\frac{1}{2}} = \frac{1}{2}(A_j + A_{j+1}) ,$$

$$\hat{A}_{j+\frac{1}{2}} = \sqrt{Max\{0, A_j A_{j+1}\}} .$$

Run out to $t = 1.5 \text{ x } 10^6$ seconds. If you encounter $A < 0$, invent or choose a method to enforce $A \geq 0$ without violating conservation of A. Explain your method. Check conservation of A and A^2 for both cases. Explain how you do this.

4. Consider a domain $0 \leq j \leq 100$, with initial conditions

$q_i = 100, 45 \leq i \leq 55$,

$q_i = 0$ otherwise.

Solve

$$\frac{\partial q}{\partial t} + c\frac{\partial q}{\partial x} = 0$$

using

1) Leapfrog in time, centered in space;

2) Lax Wendroff;

3) Upstream.

Choose $\mu = 0.7$ in each case. Compare the solutions.

5. The advective form of the finite-difference advection equation is:

$$m_j \frac{dA_j}{dt} + \frac{(mu)_{j+\frac{1}{2}}\left(A_{j+\frac{1}{2}} - A_j\right) + (mu)_{j-\frac{1}{2}}\left(A_j - A_{j-\frac{1}{2}}\right)}{\Delta x} = 0 \quad . \tag{4.137}$$

Here we have assumed that Δx is spatially constant. If the advecting mass flux is spatially constant, this reduces to

$$m_j \frac{dA_j}{dt} + mu\left(\frac{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}}{\Delta x}\right) = 0 \ , \tag{4.138}$$

from which we see that we are using the approximation

$$\left(\frac{\partial A}{\partial x}\right)_j \cong \frac{A_{j+\frac{1}{2}} - A_{j-\frac{1}{2}}}{\Delta x} \ . \tag{4.139}$$

Suppose that we adopt

$$A_{j+\frac{1}{2}} = \frac{2A_j A_{j+1}}{(A_j + A_{j+1})} \ . \tag{4.140}$$

Determine the order of accuracy of the approximation (4.139), in case (4.140) is used for interpolation to the cell walls.

6. Discuss Eq. (4.30) for the case $F(A) = A$.