3.2 Initial value problems: numerical solution – Finite differences - Truncation errors, consistency, stability and convergence – Criteria for computational stability – Explicit and implicit time schemes –Table of time schemes

Hyperbolic and parabolic PDEs are *initial value* (IV) or *marching* problems: The solution is obtained by using the known initial values and marching or advancing in time.

If boundary values are necessary, they are called “mixed initial-boundary value problems”.

Again, the simplest prototypes of these IV problems are:

$$\frac{\partial u}{\partial t} = -c\frac{\partial u}{\partial x}$$

wave or advection (hyperbolic) equation, with solution

$u(x,t) = u(x-ct,0)$, showing that the information moves along the characteristics x-ct=const,

and

$$\frac{\partial u}{\partial t} = \sigma\frac{\partial^2 u}{\partial x^2}$$

the diffusion equation, a parabolic equation.

Finite difference method:

We take discrete values for x and t: $x_j = j\Delta x$, $t_n = n\Delta t$.

The solution of the finite difference equation is also defined at the discrete points ($j\Delta x$, $n\Delta t$): $U_j^n = U(j\Delta x, n\Delta t)$.

We will use small u to denote the solution of the PDE (continuous) and capital U to denote the solution of the Finite Difference Equation (FDE), a discrete solution.

Consider again the advection equation $\frac{\partial u}{\partial t} = -c\frac{\partial u}{\partial x}$.

Suppose that we choose to approximate this PDE with the following FDE (called “upstream scheme”):

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + c\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0 \qquad (2.3)$$

Note that both differences are non-centered with respect to the point ($j\Delta x$, $n\Delta t$).

We should now ask two fundamental questions:

1) Is this FDE **consistent** with the PDE?

2) For any given time t>0, will the solution U of the FDE **converge** to u as $\Delta x \to 0$, $\Delta t \to 0$?

Let’s clarify the questions:

1) We say that the FDE is **consistent** with PDE if, in the limit when $\Delta x \to 0$, $\Delta t \to 0$ the FDE coincides with the PDE.

Obviously, this is a first requirement that the FDE should fulfill if its solutions are going to be good approximations of the solutions of the PDE.

The difference between the PDE and the FDE is the discretization error or **local** (in space and time) **truncation error**.

Consistency is rather simple to verify:

Substitute U by u in the FDE, and evaluate all terms using a Taylor series expansion centered on the point (j,n), and then subtract the PDE from the FDE.

If the difference (or local truncation error τ) goes to zero as $\Delta x \to 0$, $\Delta t \to 0$, then the FDE is consistent with the PDE.

Example: We verify the consistency of

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + c\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0 \text{ with}$$

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

by performing a Taylor series expansion around the point x_j, t_n:

$$u_j^{n+1} = (u + u_t \Delta t + u_{tt} \frac{\Delta t^2}{2} + \ldots)_j^n$$

$$u_{j-1}^n = (u - u_x \Delta x + u_{xx} \frac{\Delta x^2}{2} + \ldots)_j^n \qquad (2.1)$$

Replace in the FDE $\frac{U_j^{n+1} - U_j^n}{\Delta t} + c \frac{U_j^n - U_{j-1}^n}{\Delta x} = 0$

the Taylor series (2.1)

$$(u_t + u_{tt} \frac{\Delta t}{2} + \ldots + cu_x - cu_{xx} \frac{\Delta x}{2} + \ldots)_j^n = 0 \qquad (2.2)$$

When we subtract the PDE

$$(u_t + cu_x)_j^n = 0$$

we get the (local) **truncation error**

$$\tau = u_{tt} \frac{\Delta t}{2} - cu_{xx} \frac{\Delta x}{2} + \text{higher order terms} = 0(\Delta t) + 0(\Delta x)$$

(2.3)

So that $\lim\limits_{\Delta t \to 0, \Delta x \to 0} \tau \to 0$.

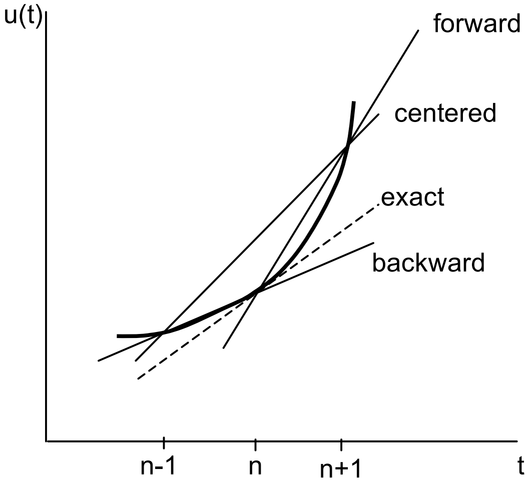
Therefore the FDE is consistent.

Note that both the *time and the space truncation errors are of first order*, because the finite differences are uncentered in both space and in time (not centered around around x_j, t_n).

Truncation errors for *centered differences are second order*, and therefore centered differences are more accurate than uncentered differences.

See schematic Fig. 3.2a and the leapfrog scheme, based on centered differences in space and in time, later in this section.

Fig. 3.2a: Schematic of centered $(\partial u / \partial t)_n \approx (u_{n+1} - u_{n-1}) / (2\Delta t)$, forward $(\partial u / \partial t)_n \approx (u_{n+1} - u_n) / \Delta t$ and backward $(\partial u / \partial t)_n \approx (u_n - u_{n-1}) / \Delta t$ finite differences estimating the time derivative $\partial u / \partial t$ at time $t_n = n\Delta t$. The three estimates are *consistent* with $\partial u / \partial t$ since they all converge to $\partial u / \partial t$ as $\Delta t \to 0$. However, the slope calculated from centered differences is much closer to the exact derivative because its truncation errors are second order.



The second question, **convergence**, i.e., whether

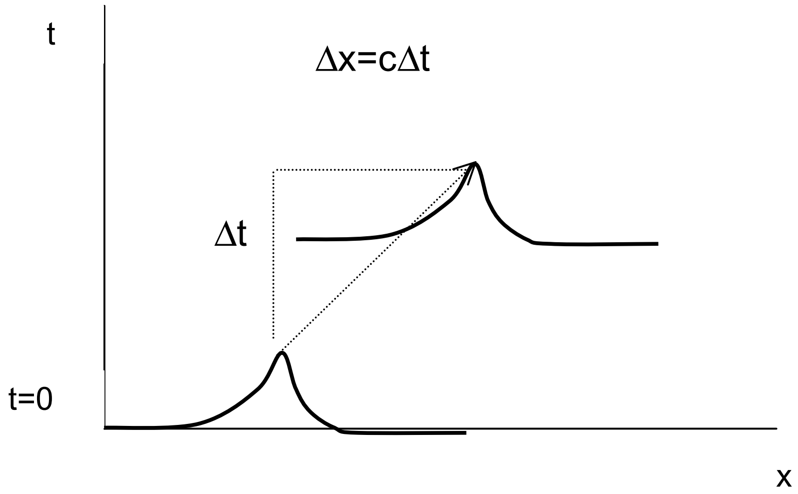
$$U(j\Delta x, n\Delta t) \to u(x,t)$$

when $j\Delta x \to x, n\Delta t \to t, \Delta x \to 0, \Delta t \to 0$

is of evident practical importance, but can only be answered after considering another problem, that of **computational stability.**

Consider again the advection equation $\frac{\partial u}{\partial t} = -c\frac{\partial u}{\partial x}$, which has the solution $u(x,t) = u(x-ct,0)$, shown schematically in Fig. 3.2b (the initial shape of u translates with velocity c).

Fig. 3.2b: Schematic of the solution of the wave equation moving along the characteristic



The upstream FDE

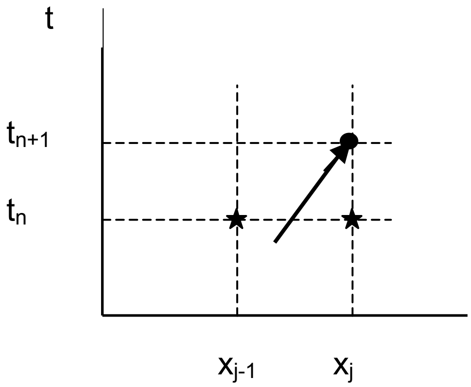
$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + c\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0$$

can be written as

$$U_j^{n+1} = (1-\mu)U_j^n + \mu U_{j-1}^n \qquad (2.4)$$

where $\mu = \dfrac{c\Delta t}{\Delta x}$ is the *Courant number*.

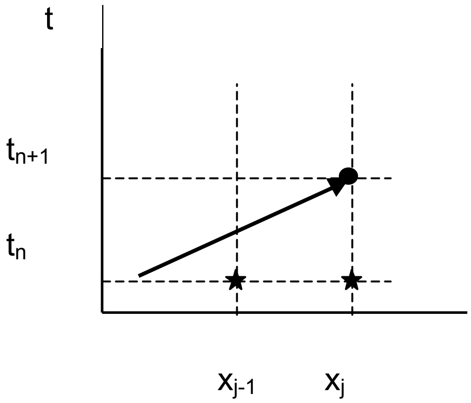
Assume that $0 < \mu = \frac{c\Delta t}{\Delta x} < 1$, as in Fig. 3.3a.



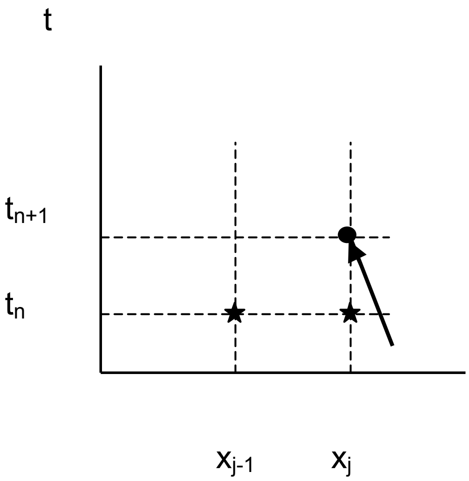
Then $U_j^{n+1} = (1-\mu)U_j^n + \mu U_{j-1}^n$ shows that the FDE solution at the new time level U_j^{n+1} is **interpolated** between the values U_j^n and U_{j-1}^n (where the stars are).

In this case the advection scheme works the way it should, because we know the true solution is in between those values (moving along the characteristic).

However, if this condition is not satisfied, and $\mu = \frac{c\Delta t}{\Delta x} > 1$ (as in Fig. 3.3b) or if



$\mu = \frac{c\Delta t}{\Delta x} < 0$ (as in Fig. 3.3c),



then the value of U_j^{n+1} is extrapolated from the values U_j^n and U_{j-1}^n.

The problem with extrapolation is that the maximum absolute value of the solution U_j^n increases with each time step.

Taking absolute values of

$$U_j^{n+1} = (1-\mu)U_j^n + \mu U_{j-1}^n$$

and letting $U^n = \max_j |U_j^n|$, we get

$$|U_j^{n+1}| \le |U_j^n||1-\mu| + |U_{j-1}^n||\mu|,$$ so that

$$U^{n+1} \le \{|1-\mu| + |\mu|\}U^n$$

Then $U^{n+1} \le U^n$ if and only if $0 \le \mu \le 1$.

If the condition $0 \le \mu \le 1$ is not satisfied, then the solution is not bounded and it grows with n.

If we let $\Delta t, \Delta x \to 0$ with μ=const, it only makes things worse, because then $n \to \infty$.

In practice, if the condition $0 \le \mu \le 1$ is not satisfied, the FDE "blows up" in a few time steps, faster for nonlinear problems.

We define now computational stability: we say that an FDE is computationally stable if the solution of the FDE at a fixed time t=nΔt remains bounded as Δt→0.

We can now state the fundamental **Lax-Richtmyer theorem**: "Given a **properly posed linear** initial value problem, and a finite difference scheme that satisfies the **consistency** condition, then the **stability** of the FDE is the necessary and sufficient condition for **convergence**".

The theorem is useful because it allows us to establish convergence by examining separately the easier questions of consistency and stability.

We are interested in convergence not because we want to let $\Delta t, \Delta x \to 0$, but because we want to make sure that if $\Delta t, \Delta x$ are small, then the errors $u(j\Delta x, n\Delta t) - U_j^n$ (accumulated or **global truncation errors** at a finite time) are acceptably small.

To determine the necessary condition for stability of the FDE (2) we used the "criterion of the maximum" method.

We can also use the “criterion of the maximum” to study the stability condition of the following FDE which approximates the parabolic diffusion equation

$$\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2}$$

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \sigma \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2} \qquad (2.5)$$

The verification of consistency is immediate. Note that, because the differences are centered in space but forward in time, the truncation error is first order in space and second order in time $O(\Delta t) + O(\Delta x)^2$.

We can write (2.5) as

$$U_j^{n+1} = \mu U_{j+1}^n + (1 - 2\mu) U_j^n + \mu U_{j-1}^n$$

where $\mu = \sigma \frac{\Delta t}{\Delta x^2}$.

If we take absolute values, and let $U^n = \max_j |U_j^n|$, we get

$$U^{n+1} \le \{|\mu| + |1 - 2\mu| + |\mu|\} U^n \qquad (2.6)$$

If all the terms inside the absolute values are positive, we can drop them, and get $U^{n+1} \le U^n$.

So we obtain a condition $0 \le \mu \le 1/2$ to insure that the solution remains bounded as n→∞, i.e., as the necessary condition for stability of the FDE.

Exercise 2: The condition on the wave equation $0 \le \mu \le 1$ for the upstream FDE is interpreted as "the time step should be chosen so that a signal cannot travel more than one grid size in one time step." Give a physical interpretation of the stability condition and the equivalent "Courant number" $\mu = \sigma \frac{\Delta t}{\Delta x^2} \le 1/2$ for the diffusion equation.

Unfortunately, the criterion of the maximum, which is intuitively very clear, can only be applied in very few cases.

In most FDE some coefficients of the equations analogous to (2.6) are negative, and the criterion cannot be applied.

Another stability criterion that has much wider application is the **von Neumann stability criterion**:

Assume that the boundary conditions allow expansion of the solution of the FDE in an appropriate set of eigenfunctions.

For simplicity we will assume an expansion into Fourier series (e.g., periodic BC):

$$U(x,t) = \sum_{\mathbf{k}} Z_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} \qquad (2.7)$$

The space variable, x, and the wave number k can be multidimensional, e.g., $\mathbf{x} = (x_1, x_2, x_3)$, $\mathbf{k} = (k_1, k_2, k_3)$. The dependent variable U can also be a vector for a system of equations.

Let $x_j = j\Delta x$ (or $\mathbf{x}_j = (j_1\Delta x_1, j_2\Delta x_2, j_3\Delta x_3)$).

We define p as the discrete wave number for the finite differences space: $p = k\Delta x = \frac{2\pi\Delta x}{L}$.
p is non-dimensional and varies between zero (for infinitely long waves) and π (for the shortest wave $L = 2\Delta x$)

For multidimensional problems, $\mathbf{p} = (k_1\Delta x_1, k_2\Delta x_2, k_3\Delta x_3)$.

Let $t_n = n\Delta t$.

Then the Fourier expansion is

$$U_j^n = \sum_p Z_p^n e^{ipj} \qquad (2.8)$$

(where for multiple dimensions $\mathbf{p}\bullet\mathbf{j} = p_1 j_1 + p_2 j_2 + p_3 j_3$).

When we substitute this Fourier expansion into a linear FDE, we obtain a system of equations for each wavenumber p

$$Z_p^{n+1} = G_p Z_p^n$$

G_p is an “amplification matrix” that, when applied to the *p* Fourier component of the solution at time nΔt “advances“ it to the time (n+1)Δt.

G_p depends on p, Δt and Δx.

If we know the initial conditions

$$U_j^0 = \sum_p Z_p^0 e^{ipj} \qquad (2.9),$$

then the solution of the FDE is

$$Z_p^n = (G_p)^n Z_p^0 \qquad (2.10).$$

since we can plug it back into

$U_j^n = \sum_p Z_p^n e^{ipj}$ and compute the solution.

Therefore, stability, i.e., boundedness of the solution for any permissible initial condition at any fixed time, is guaranteed if the matrix $\left(G_p\right)^n$ is bounded for all p when Δt→0 and n→∞.

So, for some norm, we must have $\left\|\left(G_p\right)^n\right\| < M$ for all p, as n→∞.

If $\sigma\left(G_p\right) = \max_i \left|\lambda_i\right|$ is the spectral radius of G, i.e., the maximum eigenvalue of G, then it can be shown that for any norm,

$$\left[\sigma\left(G_p\right)\right]^n \le \left\|\left[G_p\right]^n\right\| \le \left\|G_p\right\|^n \qquad (2.11)$$

Then a necessary condition for stability of an FDE, and therefore a necessary condition for convergence, is that

$$\lim_{\Delta t\to 0, n\Delta t\to t} [\sigma(G)]^n = \text{finite} = e^{\text{CONST}} \qquad (2.12)$$

Then

$$\sigma(G) \le \left[\sigma(G)^n\right]^{1/n} \le e^{\frac{\text{CONST}}{\text{n}}} = e^{\frac{\text{CONST}\,\Delta t}{\text{t}}} \approx 1 + \frac{\text{CONST}\,\Delta t}{t}$$

Or

$$\sigma(G) \le 1 + O(\Delta t) \qquad (2.13),$$

So, the von Neumann necessary condition for computational stability states that the maximum eigenvalue of the amplification matrix should be $\left|\lambda_{\max}\right| \le 1 + O(\Delta t)$.

Note: The term O(Δt) allows bounded growth with time if this growth is “legitimate”, i.e., if it arises from a physical instability present in the PDE. If the exact solution grows with time, then the FDE cannot both satisfy $\sigma(G) \le 1$ and be consistent with the PDE. However, in practice we just check that $\left|\lambda_{\max}\right| \le 1$.

Note: Sufficient conditions are very complicated, and are known only for special cases. In practice it is generally observed that eliminating the equal sign in (2.13) is enough to ensure computational stability.

Note: In principle this method can also be used to study the stability of the boundary conditions, if they are appropriately included in the amplification matrix. In practice this is complicated, and computational stability of the BC is usually obtained by ensuring well-posedness, and testing the stability experimentally.

For simple canonical equations the von Neumann criterion can be simplified by assuming solutions with an

amplification factor ρ_p rather than a matrix.

The solution for the amplification factor ρ_p then coincides with the eigenvalues of the amplification matrix, and the von Neumann stability criterion is $\rho \le 1$.

Example of von Neumann's stability criterion:

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

FDE: $\frac{U_j^{n+1} - U_j^n}{\Delta t} + c\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0$ (upstream scheme)

We studied already the consistency, and used the criterion of the maximum to get a sufficient condition for stability. Let's apply now the von Neumann criterion:

Assume $U_j^n = \sum_p Z_p^n e^{ipj} = \sum_p A\left(\rho_p\right)^n e^{ipj}$

We replace in the FDE, cancel out the common factor Ae^{ipj} and obtain

$$\frac{\left(\rho_p\right)^{n+1} - \left(\rho_p\right)^n}{\Delta t} + c\frac{\left(\rho_p\right)^n (1-e^{-ip})}{\Delta x} = 0, \text{ for all } p$$

so that at every step the solution gets amplified by a factor $\rho_p = \left[1 - \mu(1-e^{-ip})\right]$

The amplification factor ρ_p is the 1x1 amplification matrix G , so $\left|\rho_p\right| = \sigma(G)$, and the stability condition is $\left|\rho_p\right| \le 1$ for all wave numbers *p*.

We need to estimate the maximum value of the spectral radius (or amplification factor in this case):

$$\rho_p = 1 - \mu(1-e^{-ip}) = 1 - \mu(1 - \cos p + i \sin p)$$ or

$$\left|\rho_p\right|^2 = (1-\mu(1-\cos p))^2 + \mu^2 \sin^2 p$$

We make use of the trigonometrical relationships

$$\cos p = \cos^2 \frac{p}{2} - \sin^2 \frac{p}{2}, \quad \sin p = 2 \sin \frac{p}{2} \cos \frac{p}{2}$$

and obtain

$$\left|\rho_p\right|^2 = 1 - 4\mu(1-\mu)\sin^2 \frac{p}{2}$$

Recall that the non-dimensional wavenumber p varies between zero (for infinitely long waves) and π (for the shortest wave $L = 2\Delta x$). The maximum value of

$\sin^2 \frac{p}{2} = 1$ is for $p = \pi$ ($L = 2\Delta x$).

$\mu(1-\mu)$, is a parabola whose maximum value is 0.25 when μ=0.5. So the amplification factor squared will remain less or equal to 1 as long as 0≤μ≤1.

This coincides with the condition we obtained from the criterion of the maximum (and also with the notion that we should not extrapolate but interpolate the new values at time level t=(n+1)Δt, cf. schematic Fig. 3.3).

It is important to note that the amplification factor ρ_p indicates how much the amplitude of each wavenumber *p* will decrease or increase with each time step. The upstream scheme decreases the amplitude of all Fourier wave components of the solution, since, if 0<μ<1, ρ_p <1.

This is therefore a very dissipative FDE: it has strong "numerical diffusion." Fig. 3.4 shows the decrease in amplitude when using the upstream scheme after one time step and after 100 time steps for each wavenumber *p* , using a Courant number μ=0.5, a reasonable practical choice.

Since its truncation errors are large (of first order), the upstream is in general not recommended except for special situations (e.g., for outflow boundary conditions, or when modified in such a way that the dissipation rate becomes lower).

An alternative, less damping scheme known as Matsuno or Euler-backward, frequently used in combination with the Leap-Frog scheme is also shown. (Note that a "downstream" scheme (Fig. 3.3c) is unstable).

Matsuno scheme (a.k.a. Euler-backward), a predictor-corrector scheme:

$$\frac{U_j^* - U_j^n}{\Delta t} + c\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = 0$$ predictor step (forward-centered)

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + c\frac{U_{j+1}^* - U_{j-1}^*}{2\Delta x} = 0$$ corrector step

Exercise: calculate the amplification factor ρ_p

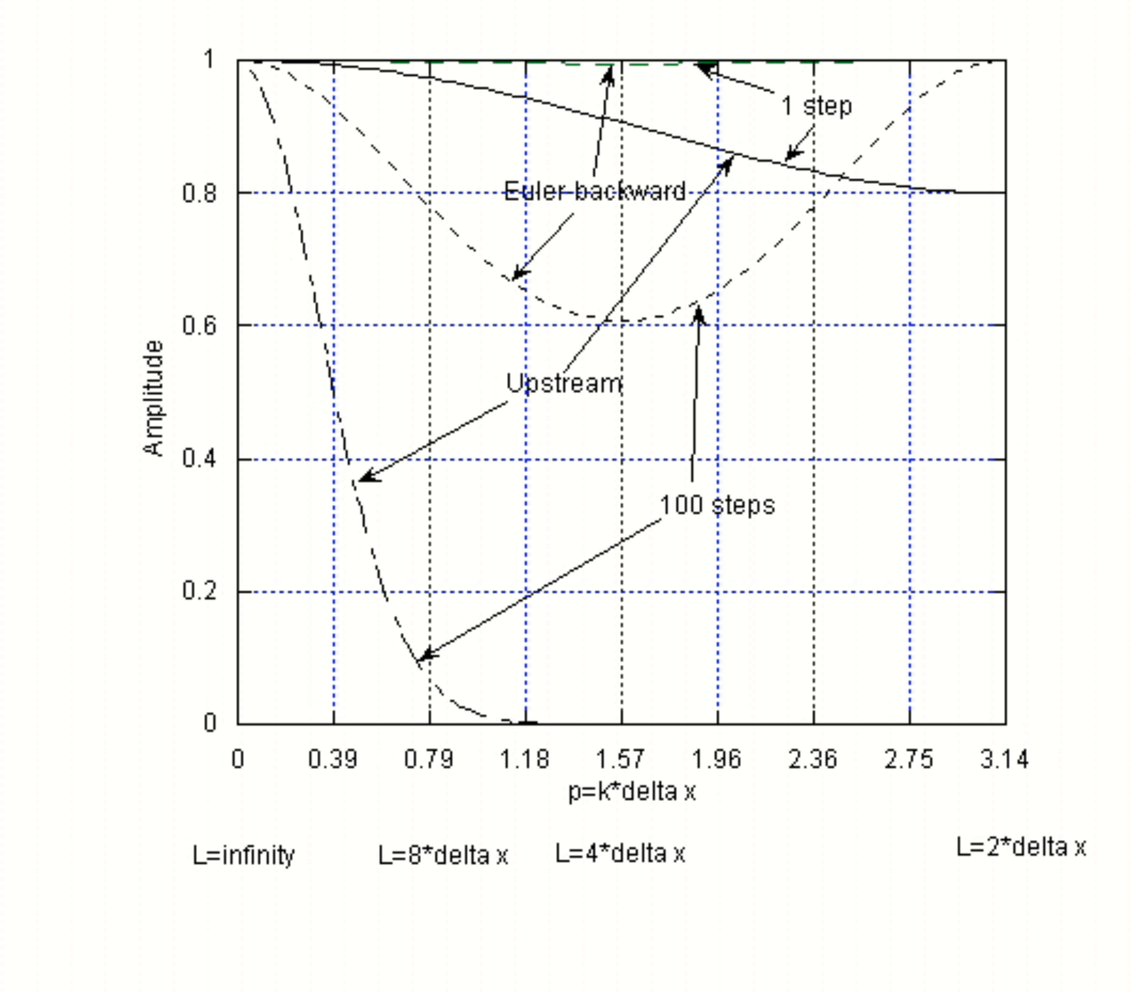


Fig. 3.4: Amplification factor of wave components of the wave equation using either the "upstream" FDE, and the Matsuno or Euler-backward schemes with μ=0.5. L is the wavelength in units of Δx .