Computer methods in dynamics of continua

(ME / AE 599)

Reza Abedi

www.rezaabedi.com/teaching/computer-methods-in-dynamics-of-continua/

University of Tennessee Space Institute

Department of Mechancial Aerospace and Biomedical Engineering

Contents

1 PDE type classification and analytical methods

1.1 Various aspects for classifications of PDEs

1.2 Linear, semilinear, quasilinear, nonlinear PDEs

1.3 Elliptic, parabolic, and hyperbolic PDEs

1.3.1 Hyperbolic PDEs and characteristics

1.3.2 Classification of a 2nd order PDE (scalar unknown) as a function of two or more independent parameters

1.3.3 D'Alembert solution of the wave equation

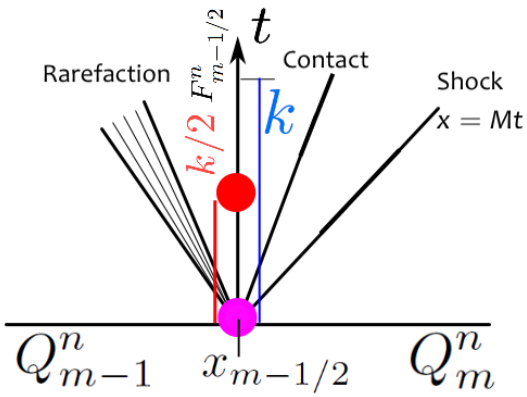
1.3.4 Systems of PDEs (multiple unknowns) as a function of two or more independent parameters

1.4 Method of characteristics: Shocks and rarefaction waves

1.5 Riemann solutions: linear and nonlinear hyperbolic PDES

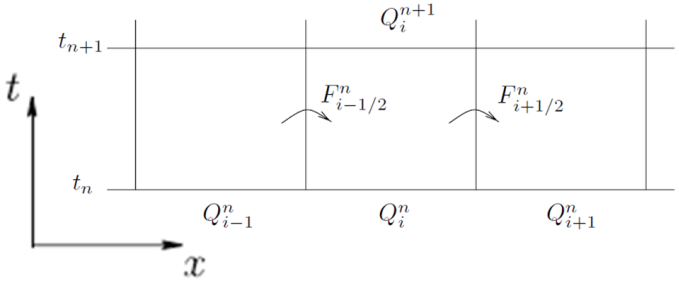
1.5.1 Introduction and Motivation

- In many instances we need to find the flux on an extruded facet in space time. For example, in Finite Volume (FV) method:
  - PDE from the balance law is $q_{,t}(x,t) + f_{,x}(q(x,t)) = 0$.
  - We need to calculate spatial flux $f(q(x,t))$ average
  - which for the boundary between cells $m-1$ and m is represented / approximated by numerical flux $F^n_{m-1/2}$.
- We solve a local problem with initial conditions Q^n_{m-1}, Q^n_m to find the value for solution at position in the figure. This is called a Riemann solution. Sometimes, we opt to choose substitutes or approximations for Riemann solution as it solution is to expensive or not available.

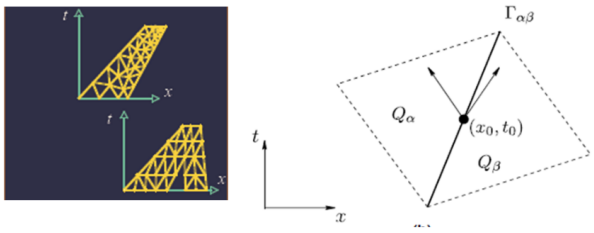


- The Riemann solution set-up in the figure is for a the Euler's equation, where for constant states at the left and right we obtain different regions in spacetime with distinct solutions.
- The red dot is the position where we position we may seek the solution for a FV scheme, Discontinuous Galerkin (DG), *etc.* (to represent the average of flux on the cell boundary).
- For quasilinear systems we may have complex solutions with rarefaction waves and shocks.
- This explains why we may use approximate Riemann solutions.
- Riemann solution may be required for nonvertical directions as well.

Vertical position: FV method & majority of DG methods



Nonvertical position: Unstructured spacetime grids



1.5.2 Approach 1: Using characteristic values (linear PDEs)

- Characteristic values $\omega = \mathbf{Lq}$ are constant (or solved as ODEs) along characteristic directions

- Primary field $\mathbf{q}$ has n components

$$\mathbf{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix}$$

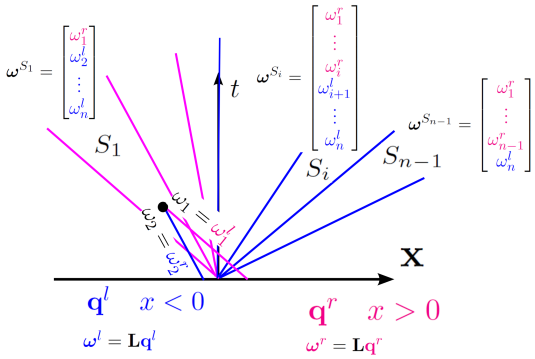
- Transfer to characteristic variables and directions For the system on $\mathbf{q}$ we transfer to characteristics by,

$$\text{For } \dot{\mathbf{q}} + \mathbf{A}\mathbf{q}_{,x} = 0 \quad \mathbf{L}\mathbf{A} = \mathbf{\Lambda}\mathbf{L}, \quad \text{for } \Lambda = \text{diag}(c_1, \cdots, c_n)$$

$\boldsymbol{\omega}$

- Eigenvalues (wave speeds) are $c_1 \le c_2 \le \cdots \le c_n$
- Initial conditions are

$$\mathbf{q}(x,0) = \mathbf{q}_0 = \begin{cases} \mathbf{q}^l & x < 0 \\ \mathbf{q}^r & x > 0 \end{cases} \Rightarrow \boldsymbol{\omega}(x,0) = \boldsymbol{\omega}_0 = \begin{cases} \boldsymbol{\omega}^l = \mathbf{L}\mathbf{q}^l & x < 0 \\ \boldsymbol{\omega}^r = \mathbf{L}\mathbf{q}^r & x < 0 \end{cases}$$



- Characteristic values $\boldsymbol{\omega} = \mathbf{L}\mathbf{q}$ are constant along characteristic directions; *i.e.*, ω_i is constant along the wave moving with speed c_i.
- Thus, for sample segments S^1, S^i, S^{n-1} we have,

$$\boldsymbol{\omega}^{S_1} = \begin{bmatrix} \omega_1^r \\ \omega_2^l \\ \vdots \\ \omega_n^l \end{bmatrix} \quad \boldsymbol{\omega}^{S_i} = \begin{bmatrix} \omega_1^r \\ \vdots \\ \omega_i^r \\ \omega_{i+1}^l \\ \vdots \\ \omega_n^l \end{bmatrix} \quad \boldsymbol{\omega}^{S_{n-1}} = \begin{bmatrix} \omega_1^r \\ \vdots \\ \omega_{n-1}^r \\ \omega_n^l \end{bmatrix}$$

- Transfer back to primary variables and fluxes is by using $\mathbf{L}$: $\mathbf{q}^{S_i} = \mathbf{L}^{-1}\boldsymbol{\omega}^{S_i}$

1.5.3 Approach 2: Using Jump shapes determined by right eigenvectors

- Again, consider the system,

$$\dot{\mathbf{q}} + \mathbf{A}\mathbf{q}_x = 0$$

- Recalling the general jump condition,

$$c = \frac{[\![\mathbf{f}_x]\!].\mathbf{n}_x}{[\![\mathbf{f}_t]\!]}$$

for spatial flux $\mathbf{f}_x$, and temporal flux $\mathbf{f}_t$. In 1D semi-linear problem, $\dot{\mathbf{q}} + \mathbf{A}\mathbf{q}_x = s$, $\mathbf{n}_x = 1$ (1D) and we have $\dot{\mathbf{q}} + \mathbf{A}\mathbf{q}_x = s + \mathbf{A}_x q$, that is spacetime fluxes are $\mathbf{f}_t = \mathbf{q}$, $\mathbf{f}_x = \mathbf{A}\mathbf{q}$. This yields the jump condition for this system as,

$$c = \frac{[\![\mathbf{A}\mathbf{q}]\!]}{[\![\mathbf{q}]\!]} \quad \Rightarrow [\![\mathbf{A}\mathbf{q}]\!] = c[\![\mathbf{q}]\!] \tag{1}$$

Now there are two cases,

- If $c = 0$ in fact $\mathbf{A}$ may jump because the flux matrix generally depend on material properties which may jump from the left to the right side of the "material interface" with has the speed $c = 0$. From (1) we get $[\![\mathbf{A}\mathbf{q}]\!] = c[\![\mathbf{q}]\!] = (\mathbf{A}\mathbf{q})^r - (\mathbf{A}\mathbf{q})^l = \mathbf{A}^r\mathbf{q}^r - \mathbf{A}^l\mathbf{q}^l = 0[\![\mathbf{q}]\!] = 0$. This condition is ONLY correct is the system is written in conservation law form. We will comment on this later. Otherwise, we need to directly check the "compatibility condition" between the two materials on the material interface. This condition needs to be checked only if
  - The materials on the two sides are different, OR
  - $\mathbf{A}$ has zero eigenvalue(s).

  Otherwise the jump condition on $c = 0$ is automatically satisfied.
- When $c \neq 0$ the jump manifold (line in 1D) lies only in one material (either left or right) and does not jump across the jump manifold. Thus in $[\![\mathbf{A}\mathbf{q}]\!]$, $\mathbf{A}$ can be taken out from the jump and from (1) we get $[\![\mathbf{A}\mathbf{q}]\!] = \mathbf{A}[\![\mathbf{q}]\!] = c[\![\mathbf{q}]\!]$, which basically implies $[\![\mathbf{q}]\!]$ is an eigenvector of $\mathbf{A}$ and c (speed of jump line) is an eigenvalue of $\mathbf{A}$. If $c < 0$, $\mathbf{A}^l$ is used, otherwise $\mathbf{A}^r$ is used.

Summary of jump conditions for a first order system of linear PDEs,

- Solve for right eigenvectors and eigenvalues of $\mathbf{A}$ (for a generic material) and find

$$\mathbf{A}\mathbf{r}^i = c_i \mathbf{r}^i \quad \text{no summation on } i$$

- Based on the sign of c_i group eigenvalues and vectors as follows,

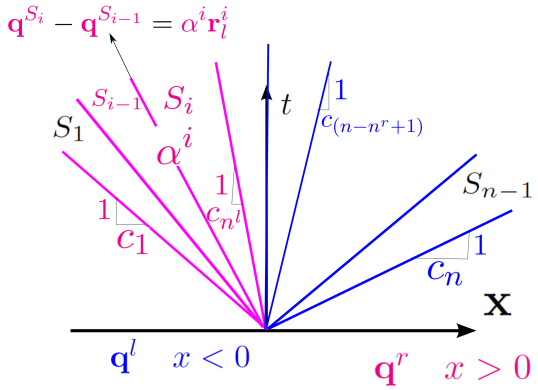
$$c_1 \le c_2 \le \cdots c_{n^l} < 0 \ (n - n^l - n^r) \text{ zero eigenvalues } < c_{(n-n^r+1)} \le \cdots \le c_n \tag{2}$$

where n^l is the number of negative eigenvalues, n^r the positive ones, and $(n - n^l - n^r)$ zero ones. Clearly, the number of any of the groups can be zero. We denote the eigenvalues from the left side $c < 0$ by $\mathbf{r}_l^i$ and the right ones $(c > 0)$ by $\mathbf{r}_r^i$ and the jump shapes for $c = 0$ by $\mathbf{r}_0^i$.

- Write the appropriate jump conditions starting from segment S^0 (adjacent of the left side IC) to segment S^{n+1} (adjacent of the right side IC),

$$\mathbf{q}^{S_i} - \mathbf{q}^{S_{i-1}} = \alpha^i \mathbf{r}_l^i \qquad \text{where } \mathbf{A}^l \mathbf{r}_l^i = c_i \mathbf{r}_l^i$$
$$\text{Left segments } S^i \qquad 1 \le i \le n^l \tag{3a}$$
$$\mathbf{q}^{S_{n-n^r}} - \mathbf{q}^{S_{n^l}} = \sum_{i=n^l+1}^{n-n-n^r} \alpha^i \mathbf{r}_0^i$$
$$\mathbf{r}_0^i \text{ jumps at material interface } S^i \qquad n^l + 1 \le i \le n - n^r \tag{3b}$$
$$\mathbf{q}^{S_i} - \mathbf{q}^{S_{i-1}} = \alpha^i \mathbf{r}_r^i \qquad \text{where } \mathbf{A}^r \mathbf{r}_r^i = c_i \mathbf{r}_r^i$$
$$\text{Right segments } S^i \qquad n - n^r + 1 \le i \le n \tag{3c}$$

For the jump conditions at $c = 0$ refer to comment on $c = 0$ below.



- By adding the jump conditions we have,

$$\mathbf{q}^r - \mathbf{q}^l = \mathbf{q}^{S_{n+1}} - \mathbf{q}^{S_0} = \sum_{i=1}^{n^l} \{\mathbf{q}^{S_i} - \mathbf{q}^{S_{i-1}}\} + \mathbf{q}^{S_{n-n^r}} - \mathbf{q}^{S_{n^l}} + \sum_{i=n-n^r+1}^{n} \{\mathbf{q}^{S_i} - \mathbf{q}^{S_{i-1}}\}$$
$$= \sum_{i=1}^{n^l} \alpha^i \mathbf{r}_l^i + \sum_{i=n^l+1}^{n-n^r} \alpha^i \mathbf{r}_0^i + \sum_{i=n-n^r+1}^{n} \alpha^i \mathbf{r}_r^i \tag{4}$$

That is,

$$\boldsymbol{\alpha} = \mathbf{r}^{-1} [\![\mathbf{q}]\!] \quad \text{where} \tag{5a}$$
$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha^1 & \cdots & \alpha^{n^l} & \alpha^{n^l+1} & \cdots & \alpha^{n-n^r} & \alpha^{n-n^r+1} & \cdots & \alpha^n \end{bmatrix}^{\mathrm{T}} \tag{5b}$$
$$\mathbf{r} = \begin{bmatrix} \mathbf{r}_l^1 & \cdots & \mathbf{r}_l^n & \mathbf{r}_0^{n^l+1} & \cdots & \mathbf{r}_0^{n-n^r} & \mathbf{r}_r^{n-n^r+1} & \cdots & \mathbf{r}_r^n \end{bmatrix} \tag{5c}$$
$$[\![\mathbf{q}]\!] = \mathbf{q}^r - \mathbf{q}^l \tag{5d}$$

- Once we have $\boldsymbol{\alpha}$ we can find the solution in segment S^i by,

$$\mathbf{q}^{S_i} = \mathbf{q}^l + \sum_{j=1}^{i} \alpha^j \mathbf{r}^j, \quad 0 \le i \le n+1 \tag{6}$$

where $\mathbf{r}^j$ is j^{th} eigenvector, *i.e.,* j^{th} column of $\mathbf{r}$ from (5c), α^j the j^{th} unknown value from (5b). Clearly, we recover trivial causal solutions $\mathbf{q}^{S_0} = \mathbf{q}^l$, $\mathbf{q}^{S_{n+1}} = \mathbf{q}^r$.

Comment on $c = 0$: Note that we have $n - n^l - n^r$ zero speeds (which can be a zero number) at the material interface that correspond to $n - n^l - n^r$ jump conditions from the middle equation. Even if there are no zero eigenvalues but there is a material interface (meaning that properties on the left and right side are distinct) we still need to verify the correct jump (matching) conditions are enforced across the material interface. For example, from balance of linear momentum, spatial flux s must be continuous across material interface $c = 0$. If we choose $\epsilon = u_{,x}$ (solid mechanics) as primary field, we need to enforce $s^l = E^l\epsilon^l = E^r\epsilon^r = s^r$. That is, in 1D elastodynamics where speeds are $c = -\sqrt{\frac{E^l}{\rho^l}}, \sqrt{\frac{E^r}{\rho^r}}$, *i.e.,* no zero eigenvalue, we need a nontrivial matching condition ($E^l\epsilon^l = E^r\epsilon^r$) at the material interface ($c = 0$). However, this jump condition does not introduce new unknown values α. That is, it is a matching condition with no added unknowns α to the system.

In any case, if there are nonzero number of zero eigenvalues the jump vector shapes $\mathbf{r}^i$ at $c = 0$ are obtained by correct matching conditions at the interface from the system of equations. As an example, in 2D (and 3D) from balance of linear momentum traction is continuous so off plane normal stresses such as s_{22} (and s_{33}) can have jumps. Direction 1 is normal to the interface.

1.5.4 Sample Riemann solution: Acoustic equation

1.5.4.1 Acoustic equation

- 1D acoustic equation for fluid / solid looks very much like the elastodynamic problem, especially, when the background velocity is zero. The primary field is given by,

$$\mathbf{q} = \begin{bmatrix} p \\ v \end{bmatrix} \tag{7}$$

where

  - p: pressure ($= -s$)
  - v: velocity.

- Closing the system of conservation laws:
  - Compatibility condition requires $\dot{p} = -\dot{s} = -Kv_{,x} \Rightarrow \dot{q}_1 + Kq_{2,x} = 0$ where K is the bulk modulus (similar to elastic modulus in solid mechanics).
  - Balance of linear momentum requires $(\rho v)_{,t} - s_{,x} = \rho\dot{v} + p_{,x} = 0 \Rightarrow \dot{q}_2 + \frac{1}{\rho}q_{1,x} = 0$.

  Thus the system of first order PDEs is,

$$\dot{\mathbf{q}} + \mathbf{A}\mathbf{q}_{,x} = 0, \text{ where } \mathbf{A} = \begin{bmatrix} 0 & K \\ \frac{1}{\rho} & 0 \end{bmatrix} \text{ (spatial flux matrix) and, } \mathbf{q} = \begin{bmatrix} p \\ v \end{bmatrix} \tag{8}$$

1.5.4.2 Riemann solutions approach 1: Acoustic equation

- We obtain the characteristic values for the acoustic problem (8) by forming the left eigenvalue eigenvector pairs,

$$\mathbf{A} = \begin{bmatrix} 0 & K \\ \frac{1}{\rho} & 0 \end{bmatrix}, \mathbf{L}\mathbf{A} = \boldsymbol{\Lambda}\mathbf{L}, \quad \text{where} \quad \mathbf{L} = \begin{bmatrix} -1 & Z \\ 1 & Z \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix} \tag{9}$$

where

$$Z = \sqrt{K\rho} = c\rho \qquad \text{Impedance} \tag{10a}$$

$$c = \sqrt{\frac{K}{\rho}} \qquad \text{Wave speed} \tag{10b}$$

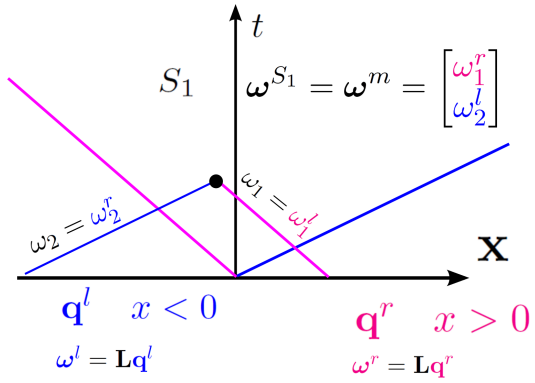
- Characteristic values $\boldsymbol{\omega} = \mathbf{L}\mathbf{q}$ are defined as,

$$\boldsymbol{\omega} = \mathbf{L}\mathbf{q} \quad \Rightarrow \quad \begin{cases} \omega_1 = -p + Zv \\ \omega_2 = p + Zv \end{cases} \tag{11}$$

- Direction of characteristics

$$\begin{cases} \text{Along } -c & \omega_1 = -p + Zv \quad \text{is constant (or varied if having source term)} \\ \text{Along } c & \omega_2 = p + Zv \quad \text{is constant (or varied if having source term)} \end{cases}$$

- Solution in terms of characteristics: According to the direction of characteristics the solution for the three regions shown is given by,



$$\boldsymbol{\omega}^L := \begin{bmatrix} \omega_1^l \\ \omega_2^l \end{bmatrix} \quad \Rightarrow \quad \mathbf{q}^L = \mathbf{L}^{-1}\boldsymbol{\omega}^L = \mathbf{q}^l \tag{12}$$

$$\boldsymbol{\omega}^R := \begin{bmatrix} \omega_1^r \\ \omega_2^r \end{bmatrix} \quad \Rightarrow \quad \mathbf{q}^R = \mathbf{L}^{-1}\boldsymbol{\omega}^R = \mathbf{q}^r \tag{13}$$

$$\boldsymbol{\omega}^m := \begin{bmatrix} \omega_1^r \\ \omega_2^l \end{bmatrix} = \begin{bmatrix} -p^r + Zv^r \\ p^l + Zv^l \end{bmatrix} \quad \Rightarrow \quad \mathbf{q}^m = \mathbf{L}^{-1}\boldsymbol{\omega}^m \quad \text{that is}$$

$$\begin{bmatrix} p^m \\ v^m \end{bmatrix} = \begin{bmatrix} \frac{p^r + p^l}{2} - \frac{Z}{2}(v^r - v^l) \\ -\frac{1}{2Z}(p^r - p^l) + \frac{v^r + v^l}{2} \end{bmatrix} \tag{14}$$

1.5.4.3 Riemann solutions approach 2: Acoustic equation

- Again we consider acoustic equation (8) whose flux matrix $\mathbf{A}$ and right eigenvectors are given by,

$$\mathbf{A} = \begin{bmatrix} 0 & K \\ \frac{1}{\rho} & 0 \end{bmatrix} \quad \mathbf{A}\mathbf{R} = \mathbf{R}\boldsymbol{\Lambda}, \quad \text{where} \quad \mathbf{R} = \begin{bmatrix} \mathbf{r}^1 & \mathbf{r}^2 \end{bmatrix} = \begin{bmatrix} -Z & Z \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix} \tag{15}$$

  Clearly, from (9)) $\mathbf{R} = \mathbf{L}^{-1}$, except possibly a constant factor. From (10), $Z = \sqrt{K\rho}$, $c = \sqrt{\frac{K}{\rho}}$.

- Since $c_1 = -c < 0$ it falls into the left side and $c_2 = c > 0$ on the right side. So,

$$\mathbf{r}_l^1 = \mathbf{r}^1 \text{ for } Z^l, c^l \quad \Rightarrow \quad \mathbf{r}_l^1 = \begin{bmatrix} -Z^l \\ 1 \end{bmatrix}, \quad c_1 = -c^l, \quad n^l = 1 \tag{16a}$$

$$\mathbf{r}_r^2 = \mathbf{r}^2 \text{ for } Z^r, c^r \quad \Rightarrow \quad \mathbf{r}_r^2 = \begin{bmatrix} Z^r \\ 1 \end{bmatrix}, \quad c_2 = c^r, \quad n^r = 1 \tag{16b}$$

- Handling Material interface $c = 0$: First, we note $n - n^l - n^r = 0$ so there are no jumps in the form (3b). Still, we need to make sure the $c = 0$ is handled correctly when the materials from the two sides are different.

  Recalling the comments below (1) for $c = 0$ we needed to satisfy $\mathbf{A}^l(\mathbf{q}^{S_1})^l = \mathbf{A}^r(\mathbf{q}^{S_1})^r$ where $(\mathbf{q}^{S_1})^l$ and $(\mathbf{q}^{S_1})^r$ are the values to the left and right of $c = 0$. This equation holds if the equation is written in the conservation law form. However, we realize that for example the equation for linear momentum is $\dot{v} + \frac{1}{\rho}p_{,x} = 0$ which is not in the conservation law form. The original

form of this equation is $\rho\dot{v} + p_{,x} = 0$ which would imply $[\![p]\!] = 0$ on the vertical interface. This is the correct condition by saying that traction on the material interface is the same from the two sides, which is implied by the Newton's third law. If we had taken the nonconservative form $\dot{v} + \frac{1}{\rho}p_{,x} = 0$ we would have required $[\![\frac{p}{\rho}]\!] = 0$ on the vertical line which clearly is not correct if $\rho^l \neq \rho^r$.

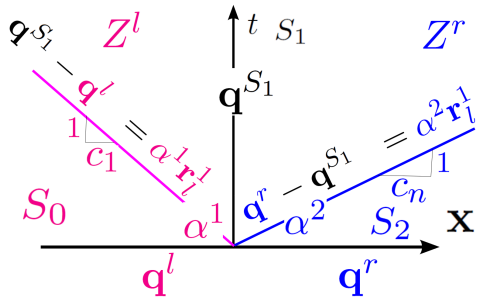
For this simple problem, from the balance of linear momentum we know that p must not suffer a jump on material interface. Same holds true for velocity v. So, if we express the equations in terms of p and v, by construction there will be no jump on $c = 0$ and we do not need to worry about it!

- Jump conditions are written as (*cf.* (1), (3a), (3c)),

$$\left.\begin{aligned} \mathbf{q}^{S_1} - \mathbf{q}^{S_0} = \mathbf{q}^{S_1} - \mathbf{q}^l &= \alpha^1 \mathbf{r}_l^1 \\ \mathbf{q}^{S_2} - \mathbf{q}^{S_1} = \mathbf{q}^r - \mathbf{q}^{S_1} &= \alpha^2 \mathbf{r}_l^1 \end{aligned}\right\} \quad \Rightarrow$$

$$\{\mathbf{q}^r - \mathbf{q}^{S_1}\} + \{\mathbf{q}^r - \mathbf{q}^l\} = \mathbf{q}^{S_2} - \mathbf{q}^l = [\![\mathbf{q}]\!] = \mathbf{r}\boldsymbol{\alpha}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha^1 \\ \alpha^2 \end{bmatrix} = \begin{bmatrix} \mathbf{r}_l^1 & \mathbf{r}_l^2 \end{bmatrix}^{-1} [\![\mathbf{q}]\!] = \begin{bmatrix} -Z^l & Z^r \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} [\![p]\!] \\ [\![v]\!] \end{bmatrix}$$



from here we obtain,

$$\alpha^1 = -\frac{1}{Z^l + Z^r}[\![p]\!] + \frac{Z^r}{Z^l + Z^r}[\![v]\!] \quad \Rightarrow$$

$$\mathbf{q}^m = \mathbf{q}^{S_1} = \begin{bmatrix} p^m \\ v^m \end{bmatrix} = \begin{bmatrix} \frac{Z^l p^r + Z^r p^l}{Z^l + Z^r} - \frac{Z^l Z^r}{Z^l + Z^r}(v^r - v^l) \\ -\frac{1}{Z^l + Z^r}(p^r - p^l) + \frac{Z^l v^l + Z^r v^r}{Z^l + Z^r} \end{bmatrix} \tag{17}$$

which matches the simpler case of $Z^l = Z^r$ from (14).

1.5.4.4 Importance of impedance Z: transmission and Reflection Coefficients

- We observed that if $Z^l = Z^r$ the Riemann solution (17) reduces to that of the same material on both sides (14).
- This implies Riemann solution only depends on Z not individual K and ρ values.
- In fact, by solving a Riemann problem with a right-going wave, we can find out how much of the wave is reflected from the interface and how much is transmitted. These values only depend on $Z^l = Z^r$.
- For example for a right-going wave with pressure value of p_0, $C_T p_0$ is transmitted and $C_R p_0$ is reflected. That is, after reflection the new pressure in the left region (region that the wave originates) is $C_R p_0$ and in the right region the pressure would be $C_T p_0$.
- The transmission and reflection coefficients for pressure field in acoustic equation are,

$$C_T = \frac{2Z^r}{Z^l + Z^r} \qquad C_R = \frac{Z^r - Z^l}{Z^l + Z^r} \tag{18}$$

- We observe when $Z^r = Z^l$ (impedance matching) $C_R = 0$ and no wave is reflected. That is, the interface behaves as if the same material is on both sides even if K, ρ are distinct.
- Conversely, when $Z^r = Z^l$ (impedance mismatch) $C_R \neq 0$ waves get reflected from the interface.
- That explains why we call Z impedance which is the impedance to wave motion.

- For more information refer to [LeVeque, 2002] "§9.8 Variable Impedance" and "§9.10 Transmission and Reflection Coefficients".
- Figures below ([LeVeque, 2002] Fig. 9.4) show a case where $K^r \neq K^l, \rho^r \neq \rho^l$ yet since $K^r = K^l$ no wave gets reflected.

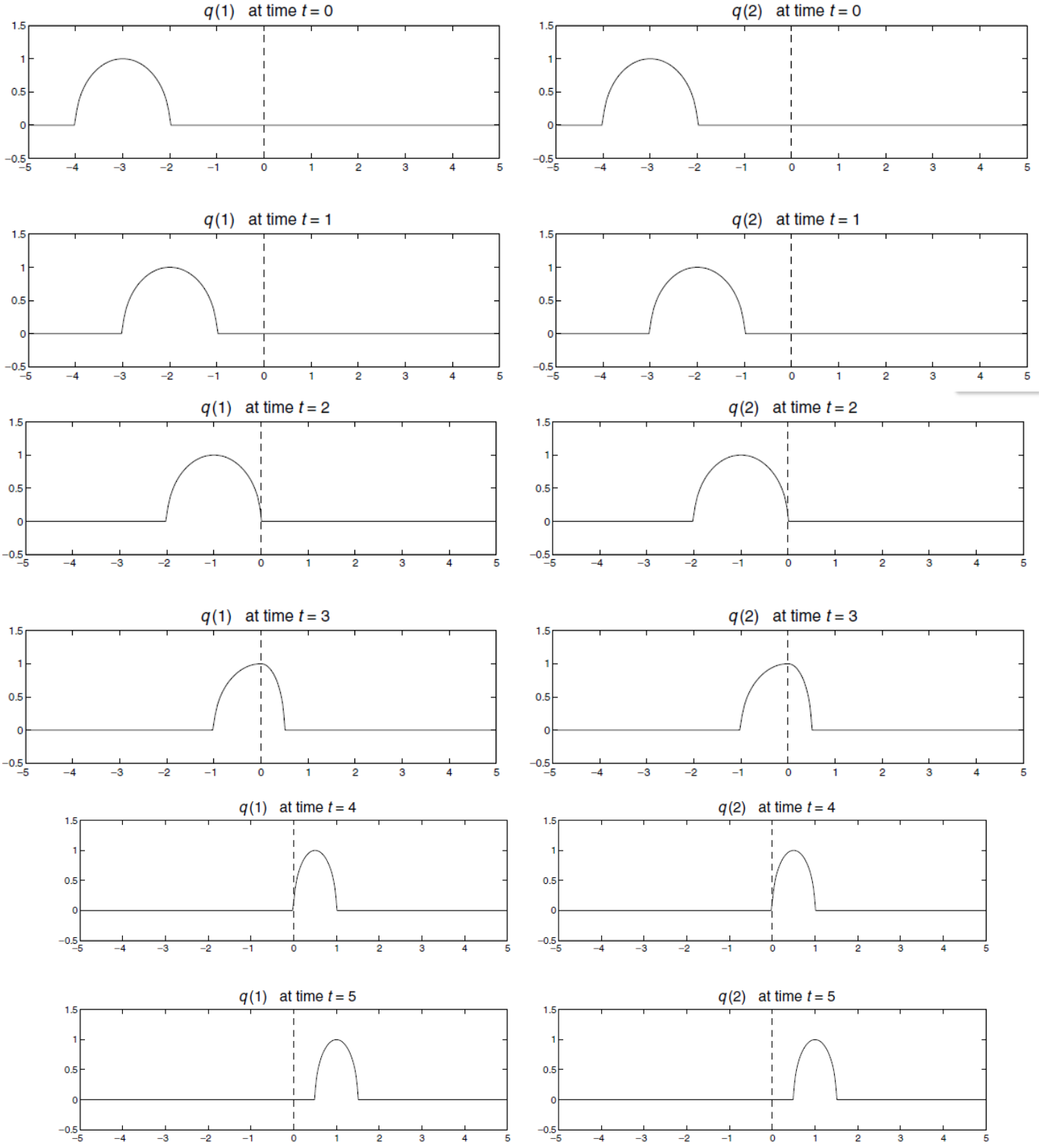


Fig. 9.4. Right-going acoustic pulse hitting a material interface (dashed line) where the sound speed changes from 1 to 0.5 but the impedance is the same. Left column: pressure. Right column: velocity. `[claw/book/chap9/acoustics/interface]`

- Figures below show another case where $K^r \neq K^l$ and we will have some wave reflection (source [LeVeque, 2002] Fig. 9.5).

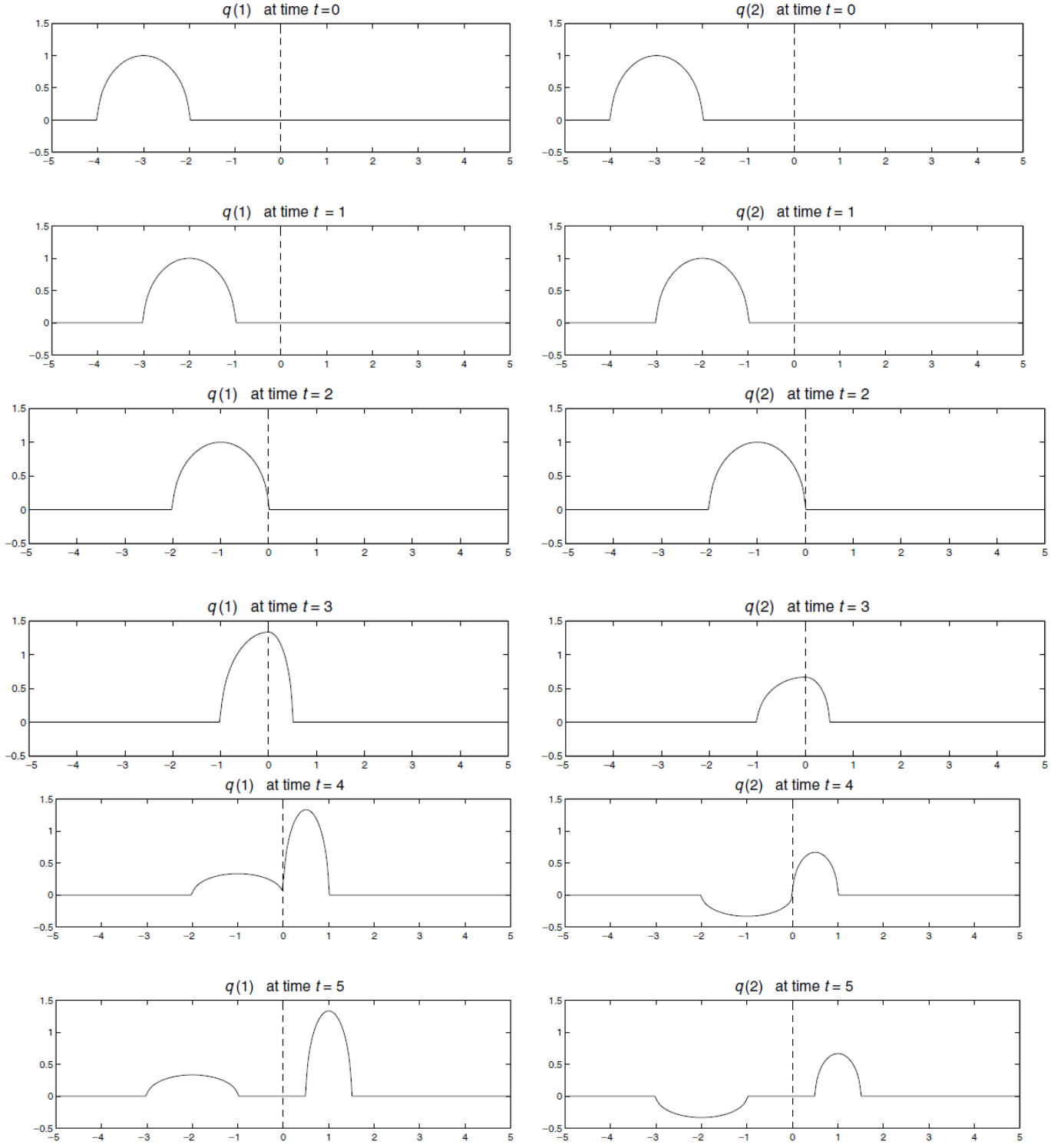


Fig. 9.5. Right-going acoustic pulse hitting a material interface (dashed line) where the sound speed changes from 1 to 0.5 and the impedance changes from 1 to 2. Part of the wave is reflected at the interface. Left column: pressure. Right column: velocity. `[claw/book/chap9/acoustics/interface]`

2 General solution schemes in space (or spacetime)

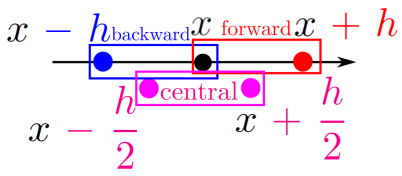
2.1 Finite Difference (FD)

2.1.1 Finite Difference (FD) operators

- $f(x)$ scalar function of x (either space or time).
- We define the following difference operators (with spacing h)

$$\Delta_h[f](x) = f(x+h) - f(x) \qquad \text{forward difference} \tag{19a}$$
$$\nabla_h[f](x) = f(x) - f(x-h) \qquad \text{backward difference} \tag{19b}$$
$$\delta_h[f](x) = f(x+\frac{h}{2}) - f(x-\frac{h}{2}) \qquad \text{central difference} \tag{19c}$$

- Note: Do not mix up ∇_h with gradient operator!



- Relation between difference operators and derivatives,

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{\Delta_h[f](x)}{h} \quad \Rightarrow$$
$$f'(x) \approx \frac{\Delta_h[f](x)}{h} \quad \text{or more precisely}$$
$$f'(x) = \frac{\Delta_h[f](x)}{h} + \mathcal{O}(h) = \frac{f(x+h)-f(x)}{h} + \mathcal{O}(h) \tag{20}$$

- The $\mathcal{O}(h)$ is obtained by the Taylor's expansion,

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \cdots + \frac{h^n}{n!}f^{(n)}(x) + \cdots \tag{21a}$$
$$= f(x) + hf'(x) + \mathcal{O}(h^2) \tag{21b}$$

- Similarly we have,

$$f'(x) = \frac{\Delta_h[f](x)}{h} + \mathcal{O}(h) = \frac{f(x+h)-f(x)}{h} + \mathcal{O}(h) \tag{22a}$$
$$f'(x) = \frac{\nabla_h[f](x)}{h} + \mathcal{O}(h) = \frac{f(x)-f(x-h)}{h} + \mathcal{O}(h) \tag{22b}$$
$$f'(x) = \frac{\delta_h[f](x)}{h} + \mathcal{O}(h^2) = \frac{f(x+\frac{h}{2})-f(x-\frac{h}{2})}{h} + \mathcal{O}(h^2) \tag{22c}$$

- Second order difference operators and their relation to second order derivative follows similarly. For example,

$$f''(x) \approx \frac{\Delta_h^2[f](x)}{h^2} = \frac{f(x+2h)-2f(x+h)+f(x)}{h^2} \qquad \text{Forward second difference} \tag{23a}$$
$$f''(x) \approx \frac{\nabla_h^2[f](x)}{h^2} = \frac{f(x)-2f(x-h)+f(x-2h)}{h^2} \qquad \text{backward second difference} \tag{23b}$$
$$f''(x) \approx \frac{\delta_h^2[f](x)}{h^2} = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} \qquad \text{Central second difference} \tag{23c}$$

- In general n-th order forward, backward, and central differences are written as,

$$\frac{\partial^n f}{\partial x^n}(x) = \frac{\Delta_h^n[f](x)}{h^n} + \mathcal{O}(h) \quad \text{where}$$
$$\Delta_h^n[f](x) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} f(x+(n-i)h) \qquad n\text{-th order forward difference} \tag{24a}$$

$$\frac{\partial^n f}{\partial x^n}(x) = \frac{\nabla_h^n[f](x)}{h^n} + \mathcal{O}(h) \quad \text{where}$$

$$\nabla_h^n[f](x) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} f(x - ih) \qquad n\text{-th order backward difference} \tag{24b}$$

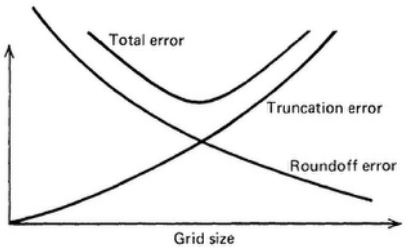
$$\frac{\partial^n f}{\partial x^n}(x) = \frac{\delta_h^n[f](x)}{h^n} + \mathcal{O}(h^2) \quad \text{where}$$

$$\delta_h^n[f](x) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} f(x + (\frac{n}{2} - i)h) \qquad n\text{-th order central difference} \tag{24c}$$

- Higher-order differences can also be used to construct better approximations.
- If necessary, the finite difference can be centered about any point by mixing forward, backward, and central differences.

2.1.2 Sources of error

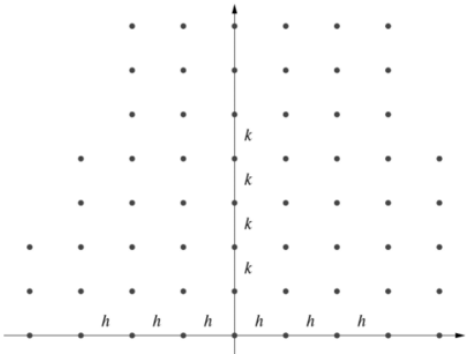
- Truncation (discretization) error: This is the error of the form $\mathcal{O}(h^p)$ which is due to discretization of differential operators with order of accuracy p. The error decreases as $h \searrow$ or $p \nearrow$.
- Roundoff error: This is due to finite precision calculations in computers, where $1 + \epsilon \to 1$ for ϵ being the machine epsilon. For FD $f(x+h) - f(x)$ takes the form $(1+\epsilon) - 1 = 1 - 1 = 0$ for small enough h.



2.1.3 Finite Difference grids

- FD grids can be nonuniform and be constructed for 2D, 3D problems. We only discuss the uniform grids, although extension to nonuniform grids is relatively trivial.
- For 1D × time problems the following notation is often used:
  - Spatial size: h (or Δx)
  - Temporal size: k (or Δt)
  - Temporal to spatial ratio $\lambda := \frac{k}{h}$.
- In 2D and 3D h_x, h_y, h_z (or $\Delta x, \Delta y, \Delta z$) are used.
- The FD difference notation is,

$$u_n^m := u(mh, nk) \tag{25}$$



2.1.4 Solution of 1D (semi-)linear advection equation

- Consider that we want to solve the advection equation,

$$u_{,t} + a(x,t)u_{,x} = 0 \qquad \text{PDE} \tag{26a}$$

$$u(x, t=0) = u_0(x) \qquad \text{IC} \tag{26b}$$

- We can discretize (26) with the following **explicit** schemes (v is used for discretized solution),

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_m^n}{h} = 0 \quad \text{forward-time, forward-space} \tag{27a}$$

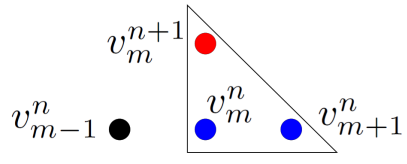


TABLE 3.1 *Finite difference approximations for various differentiations*

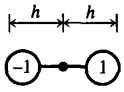
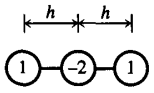
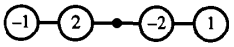
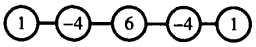
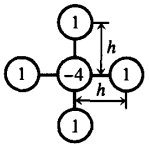
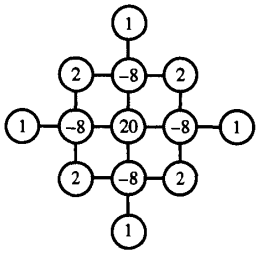
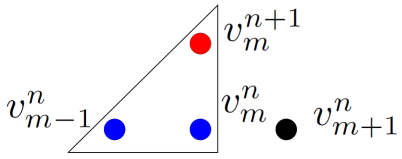
Differentiation	Finite difference approximation	Molecules
$\left.\frac{dw}{dx}\right\|_i$	$\frac{w_{i+1} - w_{i-1}}{2h}$	
$\left.\frac{d^2w}{dx^2}\right\|_i$	$\frac{w_{i+1} - 2w_i + w_{i-1}}{h^2}$	
$\left.\frac{d^3w}{dx^3}\right\|_i$	$\frac{w_{i+2} - 2w_{i+1} + 2w_{i-1} - w_{i-2}}{2h^3}$	
$\left.\frac{d^4w}{dx^4}\right\|_i$	$\frac{w_{i+2} - 4w_{i+1} + 6w_i - 4w_{i-1} + w_{i-2}}{h^4}$	

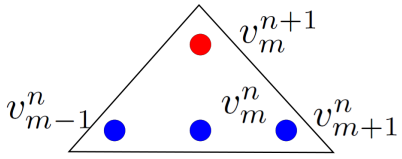
TABLE 3.1 *Finite difference approximations for various differentiations*

Differentiation	Finite difference approximation	Molecules
$\nabla^2 w\|_{i,j}$	$\frac{-4w_{i,j} + w_{i+1,j} + w_{i,j+1} + w_{i-1,j} + w_{i,j-1}}{h^2}$	
$\nabla^4 w\|_{i,j}$	$[20w_{i,j} - 8(w_{i+1,j} + w_{i-1,j} + w_{i,j+1} + w_{i,j-1}) + 2(w_{i+1,j+1} + w_{i-1,j+1} + w_{i-1,j-1} + w_{i+1,j-1}) + w_{i+2,j} + w_{i-2,j} + w_{i,j+2} w_{i,j-2}]/h^4$	

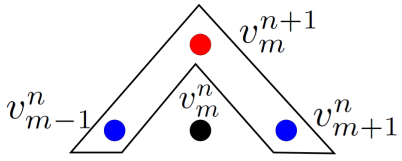
$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = 0 \quad \text{forward-time, backward-space} \tag{27b}$$



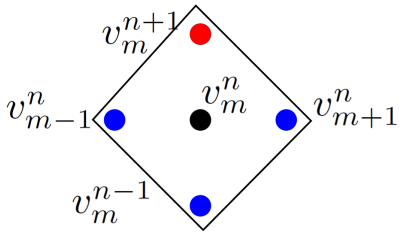
$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0 \quad \text{forward-time, central-space} \tag{27c}$$



$$\frac{v_m^{n+1} - \frac{1}{2}\left(v_{m-1}^n + v_{m+1}^n\right)}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0 \quad \text{Lax-Friedrichs} \tag{27d}$$



$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0 \quad \text{leapfrog} \tag{27e}$$



- Multistep schemes: If for the value of v_m^{n+1} more than time step n are required we call the scheme multi-step. For example leapfrog scheme is multi-step.
- We cannot directly apply multi-step schemes to compute first few time step values (*e.g.*, v_m^1 for leapfrog scheme.
- Solution for the first few steps:
  - First few steps are solved with with a single step approach, *e.g.*, (27d).
  - It is assumed first few step values are given (which in turn should be initialized with another method).

2.1.5 Explicit vs. Implicit schemes

- For all the schemes in (27) (to solve (26), $u_{,t} + a(x,t)u_{,x} = 0$), the only unknown in an equation is v_m^{n+1}. Solution stages are,
  - v_m^0 are initialized from IC (26b) $u(x, t = 0) = u_0(x)$.
  - Time-step by Time-step solution: Having the values v_m^n for all m:
    - Values at one end point of the domain $v_{m_{\min}}^{n+1}$ or $v_{m_{\max}}^{n+1}$ ($a > 0/a < 0$) is assigned by BC. In higher spatial orders stencils for $v_{,x}$, *etc.* may as well be used at the end points.
    - Inside the domain, stencils from (27) are used to find v_m^{n+1} (for all m) from previously solved/known values.
- The schemes are explicit as the solution can be solved by solving local equations one at a time.
- For explicit schemes the equations are written for time step n rather than $n + 1$.
- For the advection equation, we will see that explicit schemes must satisfy the CFL condition,

$$a\lambda_{\max} \le \alpha, \text{ for some } \alpha \le 1, \quad \text{where} \quad \lambda := \frac{k}{h} \qquad \text{that is} \quad \lambda_{\max} := \frac{k_{\max}}{h} \tag{28}$$

That is the maximum time step is limited for explicit schemes. Basically the solution CANNOT move SLOWER than the wave speed for explicit methods.

2.1.6 Violation of CFL condition for explicit methods (hyperbolic PDEs)

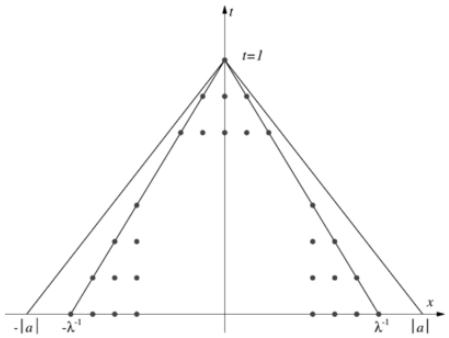
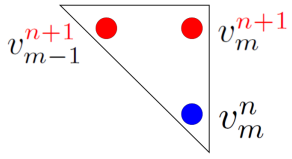


Figure 1.12. *The grid for an unstable scheme.*

- The FD grid for explicitly method has a slower speed than the wave speed a $\Rightarrow$ Scheme is nonconvergent and unstable

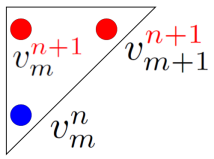
2.1.7 Explicit vs. Implicit schemes

- Below are some sample implicit schemes for the solution to (26) ($u_{,t} + a(x,t)u_{,x} = 0$),

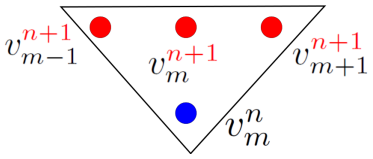


$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^{n+1} - v_{m-1}^{n+1}}{h} = 0 \qquad \text{backward-time, backward-space} \tag{29a}$$

$$\frac{v_m^{n+1}-v_m^n}{k}+a\frac{v_{m+1}^{n+1}-v_m^{n+1}}{h}=0 \qquad \text{backward-time, forward-space} \tag{29b}$$



$$\frac{v_m^{n+1}-v_m^n}{k}+a\frac{v_{m+1}^{n+1}-v_{m-1}^{n+1}}{2h}=0 \qquad \text{backward-time, central-space} \tag{29c}$$



- Unlike explicit schemes we cannot solve unknowns at a given point from its difference equation
- There are more unknowns that equations!
- Once we form the coupled system of unknowns in time step $n+1$, we can solve the entire values of this step by solving a coupled system
- Implicit methods are often stable meaning that we can choose any value for $\lambda := \frac{k}{h}$ for time stepping. The few implicit schemes that are not unconditionally stable are not used.

2.1.8 Examples for explicit methods

1. Consider the advection equation (26) $u_{,t}+a(x,t)u_{,x}=0$.
2. A Right-going wave with constant speed a is chosen. That is, $a(x,t)=a>0$.
3. We use a step function initial condition; *cf.* (26b),

$$u_0(x)=\begin{cases}1 & x<0\\ 0 & x\geq 0\end{cases} \tag{30}$$

4. Domain of computation & grid size:
   (a) Computation domain: $x\in[-3,\ 3]$.
   (b) FD grid size $h=1$

   So, there will be 7 grid points from point 0, 1, $\cdots$, 7, corresponding to $x=-3,-2,\cdots,3$, respectively.
5. Boundary Condition(s) (BC):
   (a) First order PDE in space $\Rightarrow$ only one spatial boundary condition.
   (b) Right going wave, BC is specified on the upstream of waves @ $x=-3$. BC is $u(x=-3,t)=1$.
   (c) No BCs for $x=3$ for this 1$^{\text{st}}$PDE. (Discussed further below).

2.1.8.1 FD1: Unconditionally unstable explicit method

- We use the forward-time, forward-space (FTFS) from (27a) to advance the solution in time,

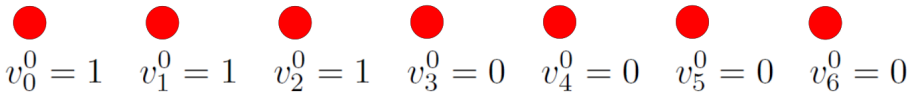
$$\frac{v_m^{n+1}-v_m^n}{k}+a\frac{v_{m+1}^n-v_m^n}{h}=0 \quad \Rightarrow \tag{31a}$$

$$v_m^{n+1}=(1+\bar{k})v_m^n-\bar{k}v_{m+1}^n \quad \text{where} \tag{31b}$$

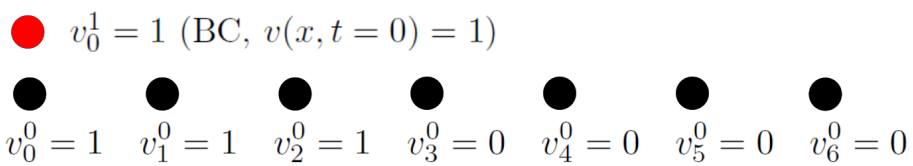
$$\bar{k}=a\lambda, \quad \lambda=\frac{k}{h}, \quad \text{recall (28)} \tag{31c}$$

- $\bar{k}$ is the normalized time step (it is nondimensional).
- Steps of the solution:

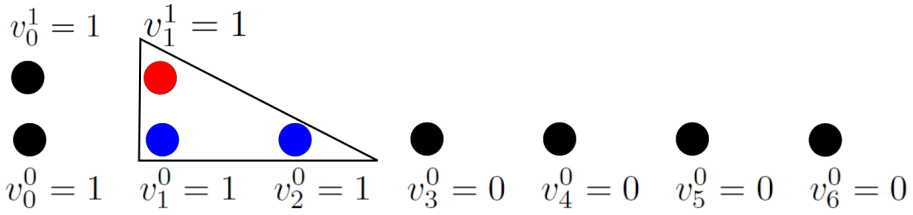
1. IC $\Rightarrow$ Solution for time step 0
   (a) We use IC (30) to set up values v_m^0, $-3\leq m\leq 3$ based on (30).
   (b) Note that $v_3^0=0$ rather than $v_3^0=1$ based on (30).



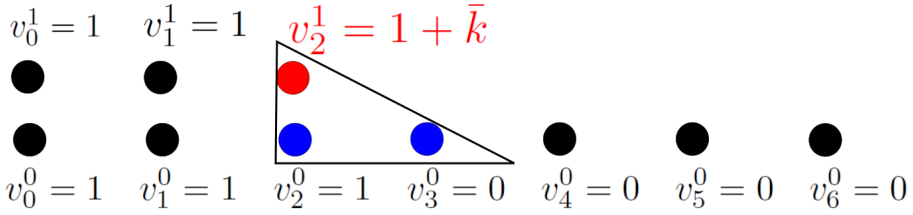
2. Time step 1: First we enforce the BC on the left boundary.



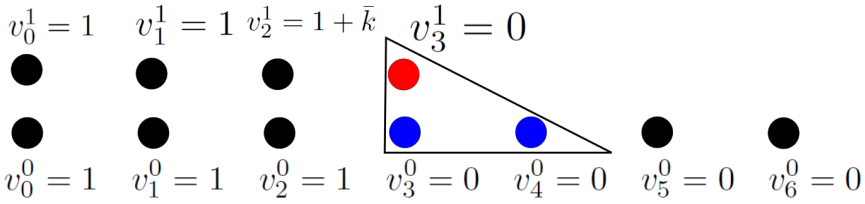
3. Using the equation for FTFS stencil (31b) ($v_m^{n+1} = (1 + \bar{k})v_m^n - \bar{k}v_{m+1}^n$), we obtain $v_1^1 = (1 + \bar{k})v_1^0 - \bar{k}v_2^0$ ($n = 1, m = 1$).



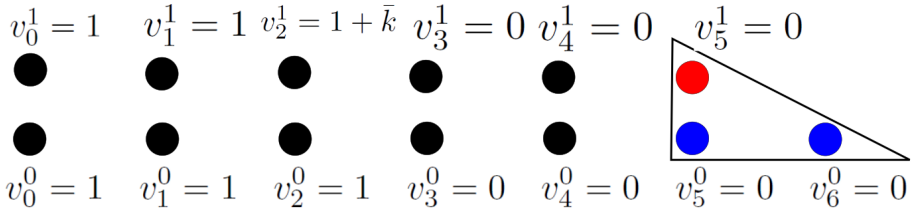
4. By a similar process we obtain $v_2^1 = (1 + \bar{k})v_2^0 - \bar{k}v_3^0 = (1 + \bar{k})$ ($n = 1, m = 2$).



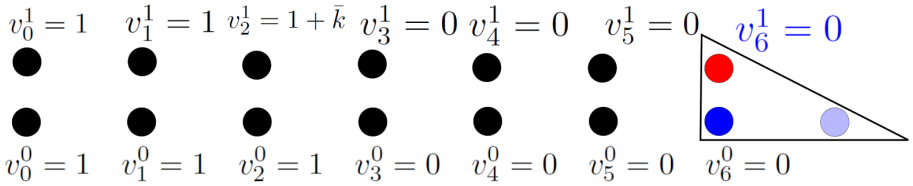
5. For $m = 3$ both previous values are zero: $v_3^1 = (1 + \bar{k})v_3^0 - \bar{k}v_4^0 = 0$.



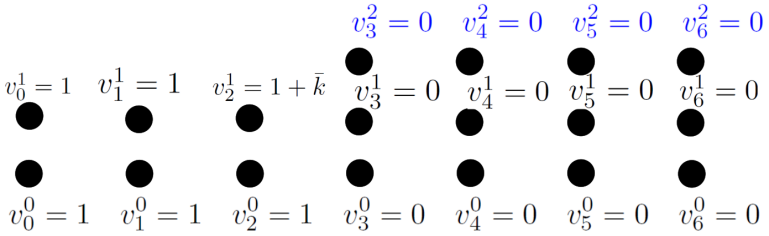
6. Process continues until point v_5^1



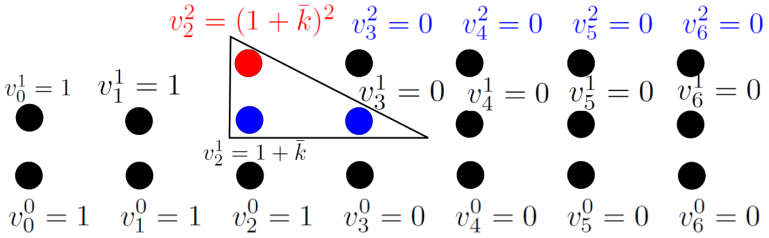
7. For v_6^1 v_7^0 does not exist! However, we will borrow one additional point for IC, knowing that the IC is zero for $x \geq 0$.



8. In a similar fashion we obtain $v_3^2 = v_4^2 = v_5^2 = v_6^2 = 0$

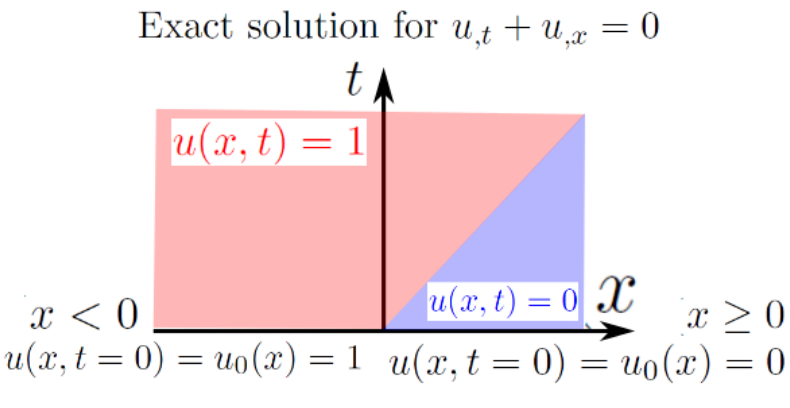


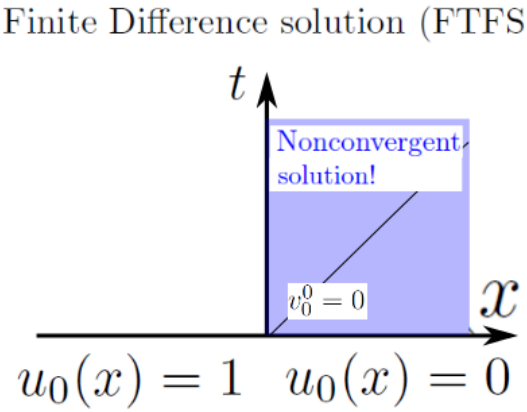
9. The value for v_2^2 $(n = 2, m = 2)$ is obtained as, $v_2^2 = (1 + \bar{k})v_2^1 - \bar{k}v_3^1 = (1 + \bar{k}).(1 + \bar{k}) = (1 + \bar{k})^2$.



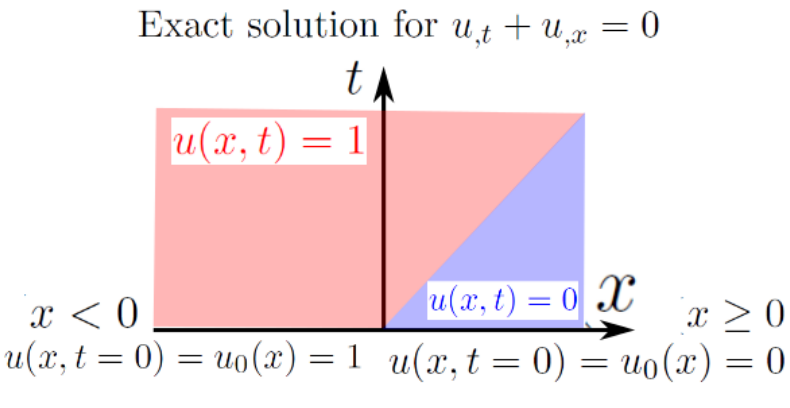
2.1.8.2 FD1: FTFS, Nonconvergence & unconditional instability!

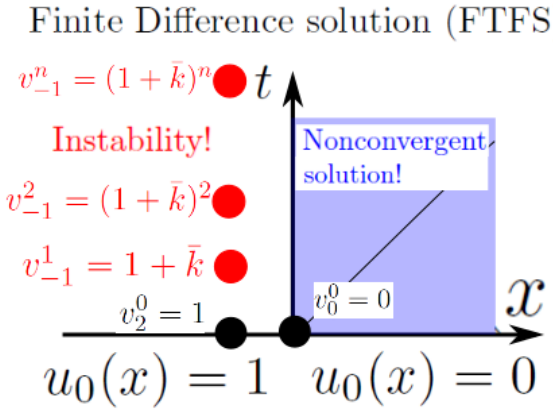
- We observe that for $x < at$ ($x < t$, for $a = 1$) exact solution is $u = 1$ but FTFS FD scheme gives us the value 0 for $x > 0$!





- Are there any other problems with the FD solution? For example the value of the point right to the left of $x = 0$ ($m = -1$)?
- We observe that for the point right next to $x = 0$ ($m = -1$) the value each time grows by the factor of $(1 + \bar{k})$!





2.1.8.3 Unconditional instability

- For a given time $T = nk$ we have,

$$v(x = -h, t = T) = v_h^k = (1 + \bar{k})^n = \left(1 + \frac{ak}{h}\right)^{\frac{T}{k}} \to \infty, \text{ as } T \to \infty \tag{32}$$

That is solution for this point approaches infinity in time. However, we know the correct solution remains bounded ($u = 1$) at this point!

- It is obvious that this problem hold for ANY time step k.
- We call this scheme UNCONDITIONALY UNSTABLE.
- On may be tempted to use very small time steps $\bar{k} \to 0$ to control the error, hoping that one may circumvent the problem. By recalling $\lim_{\bar{k}\to 0} \left(1 + \bar{k}\right)^{\frac{z}{k}} = e^z$ we have,

$$v(x = -h, t = T) \approx e^{\frac{Ta}{h}}, \text{ as } k \to 0 \tag{33a}$$

That is, even in the limit of very short time steps solution blows up. In fact, the smaller the spatial grid size h, the faster the solution blows up $v(-h, T) \approx e^{\frac{Ta}{h}}$!

- This occurs for FTFS FD because the waves move to the right but FD scheme goes from the right to the left.
- We will observe that even FTCS scheme (27c),

$$\frac{v_m^{n+1} - v_m^{n-1}}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0 \tag{34}$$

is unconditionally unstable. In this case, the difference takes contributions from the left and right, still it is unconditionally unstable. We will later have the proof on why FTFS scheme is unstable for the advection equation.

2.1.8.4 FD2: Conditionally stable explicit method

- We use the forward-time, backward-space (FTBS) from (27b) to advance the solution in time,

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = 0 \quad \Rightarrow \tag{35a}$$

$$v_m^{n+1} = (1 - \bar{k})v_m^n + \bar{k}v_{m-1}^n \quad \text{where as before (recall (31b), (28))} \tag{35b}$$

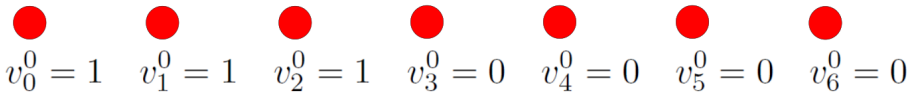
$$\bar{k} = a\lambda, \quad \lambda = \frac{k}{h}$$

Recall that $\bar{k}$ is the normalized time step.

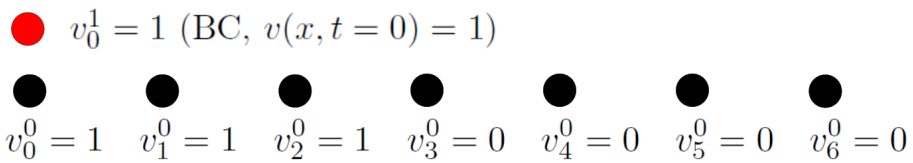
2.1.8.5 FD2: FTBS steps of solution

- Steps of the solution:

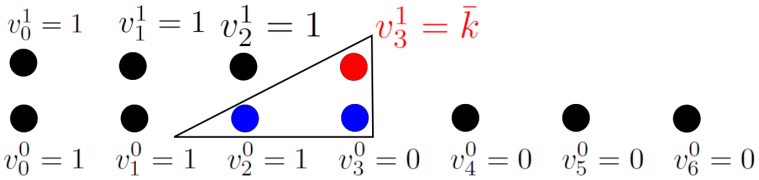
1. As before IC sets up values for time step 0



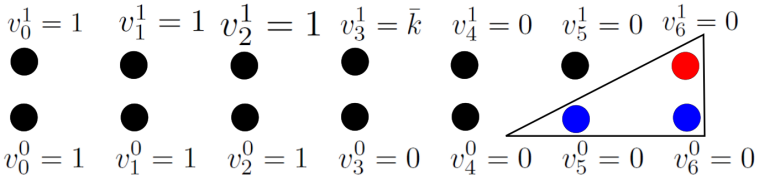
2. Again, similar to FTFS scheme, for time step 1, we enforce the BC on the left side,



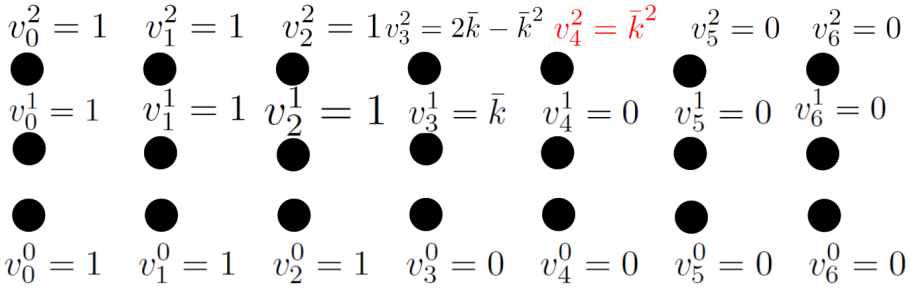
3. The formula for FTBS stencil (35b) yields $v_1^1 = v_2^1 = 1$. The value v_3^1 is obtained as, $v_3^1 = (1-\bar{k})v_3^0 + \bar{k}v_2^0 = \bar{k}$ for (m= 3, n = 1).



4. There should be no BC for the right end. The FTBS stencil basically matches the BC and the right end values does not need a value from v_7^0.



5. The solution values for time step 2 are computed similarly.



2.1.8.6 Discussion on the stability of FTFS & FTBS for $u_{,t} + au_{,x} = 0 (a > 0)$

- We observe that the value v_{n+2}^n grows with the factor $\bar{k}$. That is, $v_{n+2}^n = \bar{k}^n$.
- If $|\bar{k}| > 1$ we observe that this value blows up and the method *becomes unstable*.
- We call this scheme conditionally stable: It is stable for $|\bar{k}| \leq 1$.
- $\bar{k} = 1$ matches the maximum possible limit for explicit methods for hyperbolic problems. This corresponds to CFL number = 1 (discussed later).
- We will observe that for FD formulas of the type $v_m^{n+1} = \alpha v_{m-1}^n + \beta v_m^n$ stability is assured if $|\alpha| + |\beta| \leq 1$.
- For FTBS scheme we had $v_m^{n+1} = (1-\bar{k})v_{m-1}^n + \bar{k}v_m^n \Rightarrow$ stability requires $|1-\bar{k}| + |\bar{k}| \leq 1 \Rightarrow \bar{k} \leq 1$.
- Similarly for FD formulas of the type $v_m^{n+1} = \alpha v_m^n + \beta v_{m+1}^n$ stability again requires $|\alpha| + |\beta| \leq 1$.
- For FTFS scheme we had $v_m^{n+1} = (1+\bar{k})v_m^n + (-\bar{k})v_{m+1}^n \Rightarrow$ stability requires $|1+\bar{k}| + |\bar{k}| \leq 1$. This condition does not hold for any $\bar{k}$ meaning that FTFS is unconditionally stable (for $a > 0$).
- Clearly FTBS is unconditionally unstable for $a < 0$.
- We will observe FTCS (27c) $\frac{v_m^{n+1} - v_m^{n-1}}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$ is unconditionally unstable!

2.1.8.7 Discussion on Lax-Friedrichs and Leapfrog schemes

- Both Lax-Friedrichs (LF) and leapfrog schemes are conditionally stable.
- Figures below show solutions from [Strikwerda, 2004] for both methods for the hat shape IC,

$$u_0(x) = \begin{cases} 1 - |x| & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

  for $u_{,t} + u_{,x} = 0$ ($a = 1$ in $u_{,t} + au_{,x} = 0$. The normalized time step $\bar{k} = a\lambda = \lambda$ determines stability.
- Comparison of Lax-Friedrichs and Leapfrog when stable $\bar{k}$ is chosen ($\bar{k} = 0.8$) at $t = 0.8$:
- We observe leapfrog provides a better solution.

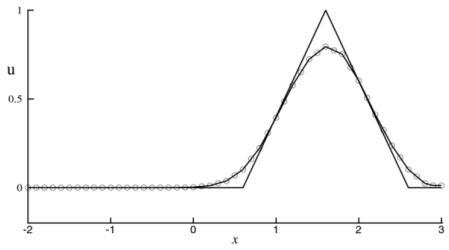


Figure 1.6. *A solution of the Lax–Friedrichs scheme,* $\lambda = 0.8$.

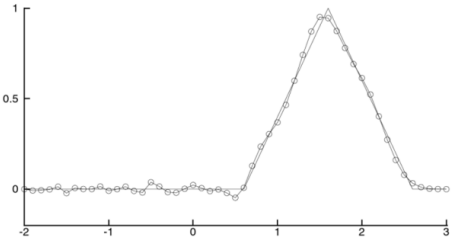


Figure 1.8. *A solution computed with leapfrog scheme,* $\lambda = 0.8$.

2.1.8.8 Development of instabilities from nonsmooth features

- If an unstable time stable is used $\bar{k} = \lambda = 1.6$ the solution will be unstable.
- If nonsmooth features exist in the solution (IC, BC, source term) instabilities often initiate from those locations (if the method is unstable):

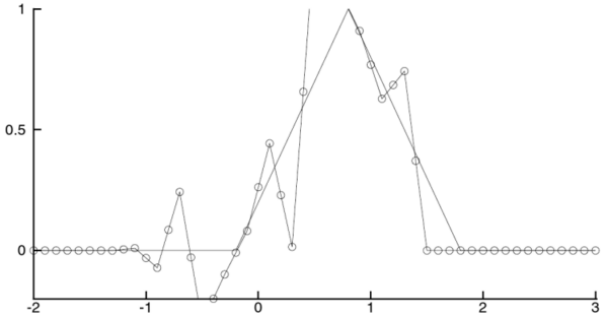
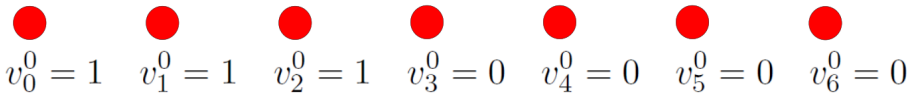


Figure 1.7. *A solution of the Lax–Friedrichs scheme,* $\lambda = 1.6$.

2.1.9 Examples for implicit methods

- We consider the same advection problem (26) $u_{,t} + a(x,t)u_{,x} = 0$, $a(x,t) = a > 0$, with IC (30) $u_0(x) = 1 - H(x)$, and BC $u(-3,t) = 1$ and the 7 point grid with $h = 1$ for the domain $x \in [-3,\ 3]$.



- The stencil for backward-time backward space (BTBS) scheme is (29a),

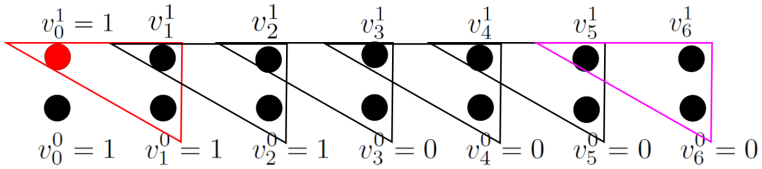
$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^{n+1} - v_{m-1}^{n+1}}{h} = 0 \quad \Rightarrow$$

$$(1 + \bar{k})v_m^{n+1} + (-\bar{k})v_{m-1}^{n+1} = v_m^n \quad \text{where} \tag{36a}$$

$$\bar{k} = \lambda a = \frac{ka}{h} \tag{36b}$$

2.1.9.1 Backward-time backward-space (BTBS) implicit method

- Stages of solutions:
  1. IC is set as before for FTFS & FTBS.
  2. Boundary condition on the left boundary is set as $v_0^1 = 1$.
  3. The equations for points 1 to 6 based on (36a) are,

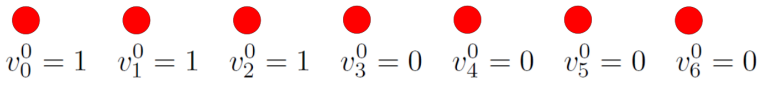


$$\left.\begin{array}{ll}(1+\bar{k})v_1^1-\bar{k}v_0^1 &= v_1^0\\ (1+\bar{k})v_2^1-\bar{k}v_1^1 &= v_2^0\\ (1+\bar{k})v_3^1-\bar{k}v_2^1 &= v_3^0\\ (1+\bar{k})v_4^1-\bar{k}v_3^1 &= v_4^0\\ (1+\bar{k})v_5^1-\bar{k}v_4^1 &= v_5^0\\ (1+\bar{k})v_6^1-\bar{k}v_5^1 &= v_6^0\end{array}\right\} \quad \Rightarrow \mathbf{A}\mathbf{v}^1 := \mathbf{v}^0 + \mathbf{b}^1, \quad \text{where} \tag{37a}$$

$$\mathbf{v}^n := \begin{bmatrix} v_1^n \\ v_2^n \\ v_3^n \\ v_4^n \\ v_5^n \\ v_6^n \end{bmatrix}, \quad \mathbf{b}^n = \bar{k}\begin{bmatrix} u(0,t_n) \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} (1+\bar{k}) & 0 & 0 & 0 & 0 & 0 \\ \bar{k} & (1+\bar{k}) & 0 & 0 & 0 & 0 \\ 0 & \bar{k} & (1+\bar{k}) & 0 & 0 & 0 \\ 0 & 0 & \bar{k} & (1+\bar{k}) & 0 & 0 \\ 0 & 0 & 0 & \bar{k} & (1+\bar{k}) & 0 \\ 0 & 0 & 0 & 0 & \bar{k} & (1+\bar{k}) \end{bmatrix} \tag{37b}$$

- Note that v_0^1, and all v_0^n, are known by the left BC. They values shown in red is moved to the RHS.
- The backward grid again nicely matches the grid for this problem, and no grid is needed for the right boundary (last stencil is shown in magenta).
- The solution of implicit system often requires a global system solution of the form $\mathbf{A}\mathbf{v}^{n+1} := \mathbf{v}^n + \mathbf{b}^{n+1}$, *cf.* (37a), where $\mathbf{A}$ is the size of unknowns in the (entire) spatial grid. The solution of this system can be expensive in 2D and 3D problems.
- For this 1D problem, the solution the solution can be obtain by judicious start of the system solve from the BC on the left. From (37a) we have,

$$\begin{aligned}
v_0^1 = 1 \qquad & \text{BC} \\
(1+\bar{k})v_1^1-\bar{k}v_0^1 = v_1^0 \qquad & \Rightarrow \qquad v_1^1 = \frac{\bar{k}}{1+\bar{k}}v_0^1 + \frac{1}{1+\bar{k}}v_1^0 = \frac{\bar{k}}{1+\bar{k}}1 + \frac{1}{1+\bar{k}}1 = 1 \\
(1+\bar{k})v_2^1-\bar{k}v_1^1 = v_2^0 \qquad & \Rightarrow \qquad v_2^1 = \frac{\bar{k}}{1+\bar{k}}v_1^1 + \frac{1}{1+\bar{k}}v_2^0 = \frac{\bar{k}}{1+\bar{k}}1 + \frac{1}{1+\bar{k}}1 = 1 \\
(1+\bar{k})v_3^1-\bar{k}v_2^1 = v_3^0 \qquad & \Rightarrow \qquad v_3^1 = \frac{\bar{k}}{1+\bar{k}}v_2^1 + \frac{1}{1+\bar{k}}v_3^0 = \frac{\bar{k}}{1+\bar{k}}1 + \frac{1}{1+\bar{k}}0 = \frac{\bar{k}}{1+\bar{k}} \\
(1+\bar{k})v_4^1-\bar{k}v_3^1 = v_4^0 \qquad & \Rightarrow \qquad v_4^1 = \frac{\bar{k}}{1+\bar{k}}v_3^1 + \frac{1}{1+\bar{k}}v_4^0 = \frac{\bar{k}}{1+\bar{k}}\frac{\bar{k}}{1+\bar{k}} + \frac{1}{1+\bar{k}}0 = \left(\frac{\bar{k}}{1+\bar{k}}\right)^2 \\
(1+\bar{k})v_5^1-\bar{k}v_4^1 = v_5^0 \qquad & \Rightarrow \qquad v_5^1 = \frac{\bar{k}}{1+\bar{k}}v_4^1 + \frac{1}{1+\bar{k}}v_5^0 = \frac{\bar{k}}{1+\bar{k}}\left(\frac{\bar{k}}{1+\bar{k}}\right)^2 + \frac{1}{1+\bar{k}}0 = \left(\frac{\bar{k}}{1+\bar{k}}\right)^3 \\
(1+\bar{k})v_6^1-\bar{k}v_5^1 = v_6^0 \qquad & \Rightarrow \qquad v_6^1 = \frac{\bar{k}}{1+\bar{k}}v_5^1 + \frac{1}{1+\bar{k}}v_6^0 = \frac{\bar{k}}{1+\bar{k}}\left(\frac{\bar{k}}{1+\bar{k}}\right)^3 + \frac{1}{1+\bar{k}}0 = \left(\frac{\bar{k}}{1+\bar{k}}\right)^4
\end{aligned}$$



2.1.9.2 Backward-time backward-space (BTBS): Unconditional stability

- As seen, for this particular 1D BTBS implicit method, we can solve the system locally starting from the left BC.
- Note that in explicit methods, we can do many updates from step n (and previous ones) to $n+1$ independent from each other. The updates can be done in parallel. Even in this fortunate case, that implicit method can be solve locally, we have do follow a certain order of solving the system. This can be observed by the particular form of $\mathbf{A}$ where there is only one nonzero value in the first row; *cf.* (37b).
- To discuss the stability of this scheme, we express the solution from step n to $n+1$ in the form of a linear transformation. For this IBVP solved by BTBS we need to solve an equation in the form (*cf.* (37a)),

$$\mathbf{A}\mathbf{v}^{n+1} := \mathbf{v}^n + \mathbf{b}^{n+1}, \quad \text{where} \tag{38}$$

where $\mathbf{b}^n = \bar{k}[u(0,t_n)\ 0\ 0\ 0\ 0\ 0]^{\mathrm{T}}$ is the contribution from the boundary condition.

- The solution to this system is obtained as follows,

$$\mathbf{v}^1 = \mathbf{A}^{-1}\mathbf{v}^0 + \mathbf{A}^{-1}\mathbf{b}^1$$
$$\mathbf{v}^2 = \mathbf{A}^{-1}\mathbf{v}^1 + \mathbf{A}^{-1}\mathbf{b}^2 = \mathbf{A}^{-1}\left\{\mathbf{A}^{-1}\mathbf{v}^0 + \mathbf{A}^{-1}\mathbf{b}^1\right\} + \mathbf{A}^{-1}\mathbf{b}^2 \qquad \mathbf{v}^2 = \mathbf{A}^{-2}\mathbf{v}^0 + \mathbf{A}^{-1}\mathbf{b}^2 + \mathbf{A}^{-2}\mathbf{b}^1$$

- By induction, we observe,

$$\mathbf{v}^n = \mathbf{A}^{-n}\mathbf{v}^0 + \mathbf{A}^{-1}\mathbf{b}^n + \cdots + \mathbf{A}^{-n}\mathbf{b}^1 \quad \Rightarrow \tag{39a}$$
$$\mathbf{v}^n = \mathbf{A}^{-n}\mathbf{v}^0 + \left\{\mathbf{A}^{-1} + \cdots \mathbf{A}^{-n}\right\}\mathbf{b}^1 = \mathbf{A}^{-n}\mathbf{v}^0 + \mathbf{A}^{-1}\left(\mathbf{I} - \mathbf{A}^{-n}\right)\left(\mathbf{I} - \mathbf{A}^{-1}\right)^{-1} \quad \text{for constant BC } u(-3,t) = 1 \text{ at } x = -3 \tag{39b}$$

- Let's for a moment, ignore the BC contributions (*e.g.,* for example when $u(x = -3, t) = 0$). Then we get the simplified equation,

$$\mathbf{v}^n = \mathbf{A}^{-n}\mathbf{v}^0 \quad \text{for zero BC on the left} \tag{40}$$

- When does $\mathbf{v}^n$ blow-up, *i.e.,* tend to infinity?
- Assume for the moment that $\mathbf{A}^{-1} = \mathbf{D}$ is diagonal (most general case by using Jordan decomposition is discussed in §5.3).

$$\mathbf{v}^n = \mathbf{D}^n\mathbf{v}^0 \quad \text{where} \quad \mathbf{D} := \begin{bmatrix} d_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & d_6 \end{bmatrix} \quad \text{where } d_1, \cdots, d_6 \text{ are diagonal values of } \mathbf{D} \tag{41}$$

- What can we say about the growth of $\mathbf{v}^n$?

2.1.9.3 Spectral radius of a matrix

- Based on these equations we get,

$$\mathbf{v}^n = \mathbf{D}^n\mathbf{v}^0 = \begin{bmatrix} d_1^n & 0 & 0 & 0 & 0 & 0 \\ 0 & d_2^n & 0 & 0 & 0 & 0 \\ 0 & 0 & d_3^n & 0 & 0 & 0 \\ 0 & 0 & 0 & d_4^n & 0 & 0 \\ 0 & 0 & 0 & 0 & d_5^n & 0 \\ 0 & 0 & 0 & 0 & 0 & d_6^n \end{bmatrix} \mathbf{v}^0 \tag{42}$$

- Let $n \to \infty$ (*i.e.,* $t_n = nk \to \infty$). When the solution goes to infinity at t_n (*i.e.,* components of $\mathbf{v}^n$ go to infinity)?
- Answer:

If $|d_i| > 1$ for ANY i the solution blows up!

- Notice that the physical system only propagates the solution with the speed a, and the solution physically does not blow up.
- But if this diagonal matrix $\mathbf{D}$ had any diagonal value larger than 1, the numerical solution would blow up.
- We define the **spectral radius** of $\mathbf{D}$ as the maximum of the absolute values of its diagonal values (remember $\mathbf{D}$ is diagonal). This is denoted by $\rho(\mathbf{D})$.
- In reality we know that $\mathbf{A}^{-1}$ is NOT diagonal for this problem. Then, how can we an argument similar to this?

2.1.9.4 Spectral (eigen-) decomposition of a matrix

- If we have the relation $\mathbf{v}^n = \mathbf{B}^n\mathbf{v}^0$, where $\mathbf{B} = \mathbf{A}^{-1}$ we can use right (or left) eigen decomposition of $\mathbf{B}$,

$$\mathbf{BR} = \mathbf{R}\boldsymbol{\Lambda}$$

where $\mathbf{R} = [\mathbf{r}^1 \ \cdots \mathbf{r}^N]$ is the right eigenvector matrix (N is the size of $\mathbf{B}$) and $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_N)$ is the eigenvalue matrix. We are assuming that $\mathbf{B}$ is diagonalizable.

In this case we have,

$$\mathbf{BR} = \mathbf{R}\boldsymbol{\Lambda} \quad \Rightarrow \qquad \mathbf{B} = \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^{-1} \quad \Rightarrow \tag{43a}$$
$$\mathbf{B}^n = \mathbf{R}\boldsymbol{\Lambda}^n\mathbf{R}^{-1} \tag{43b}$$

- Thus, from $\mathbf{v}^n = \mathbf{B}^n\mathbf{v}^0$ and (43b) we obtain,

$$\mathbf{v}^n = \mathbf{R}\mathbf{\Lambda}^n\mathbf{R}^{-1}\mathbf{v}^0 = \mathbf{R}\begin{bmatrix}\lambda_1^n & 0 & \cdots & 0\\ 0 & \lambda_2^n & \cdots & 0\\ \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & \lambda_N^n\end{bmatrix}\mathbf{R}^{-1}\mathbf{v}^0 \tag{44}$$

- Clearly, $\mathbf{v}^n$ blows up as $n \to \infty$ ($t_n = nk \to \infty$) iff $\max(|\lambda_i|) > 1$.
- For a general $N \times N$ matrix $\mathbf{B}$ we define the spectral radius as,

$$\rho(\mathbf{B}) = \max\{|\lambda_i| \; i \in \{1, \cdots, N\} | \lambda_i \text{ are eigenvalues of } \mathbf{B}\} \tag{45}$$

- Then the stability condition for the system of equation is,

$$\mathbf{v}^n = \mathbf{B}^n\mathbf{v}^0 \qquad \text{is stable iff } \rho(\mathbf{B}) \le 1 \quad \text{that is for (40) we have}$$

$$\mathbf{v}^n = \mathbf{A}^{-n}\mathbf{v}^0 \qquad \text{is stable iff } \rho(\mathbf{A}^{-1}) \le 1 \tag{46}$$

- It is easy to check that left BC terms also do not blow up in (39b) if $\rho(\mathbf{A}^{-1}) \le 1$.
- Remember that $\mathbf{A}$ was:

$$\mathbf{A} = \begin{bmatrix}(1+\bar{k}) & 0 & 0 & 0 & 0 & 0\\ \bar{k} & (1+\bar{k}) & 0 & 0 & 0 & 0\\ 0 & \bar{k} & (1+\bar{k}) & 0 & 0 & 0\\ 0 & 0 & \bar{k} & (1+\bar{k}) & 0 & 0\\ 0 & 0 & 0 & \bar{k} & (1+\bar{k}) & 0\\ 0 & 0 & 0 & 0 & \bar{k} & (1+\bar{k})\end{bmatrix}$$

2.1.9.5 Backward-time backward-space (BTBS): Unconditional stability

- What can we can say about $\rho(\mathbf{A})$?

$$\mathbf{A} = \begin{bmatrix}(1+\bar{k}) & 0 & 0 & 0 & 0 & 0\\ \bar{k} & (1+\bar{k}) & 0 & 0 & 0 & 0\\ 0 & \bar{k} & (1+\bar{k}) & 0 & 0 & 0\\ 0 & 0 & \bar{k} & (1+\bar{k}) & 0 & 0\\ 0 & 0 & 0 & \bar{k} & (1+\bar{k}) & 0\\ 0 & 0 & 0 & 0 & \bar{k} & (1+\bar{k})\end{bmatrix}$$

- Since $\mathbf{A}$ is lower triangular the diagonal values are the eigenvalues. In this case all eigenvalues of $\mathbf{A}$ are $(1+\bar{k})$ $\Rightarrow$
- Eigenvalues of $\mathbf{A}^{-1}$ are inverse of eigenvalues of $\mathbf{A}$:
- All eigenvalues of $\mathbf{A}$ are $(1+\bar{k})^{-1}$ $\Rightarrow$

$$\rho(\mathbf{A}^{-1}) = \frac{1}{1+\bar{k}}, \quad \text{stability requires} \quad \rho(\mathbf{A}^{-1}) = \frac{1}{1+\bar{k}} \quad \text{which holds for ALL } \bar{k} > 0. \tag{47}$$

but $\bar{k} = \frac{ak}{h} > 0$ because $a > 0$ (right-going wave).

- Thus, BTBS FD scheme for advection equation with $a > 0$ is **unconditionally stable**.

2.1.9.6 Backward-time forward-space: A conditional stable implicit method

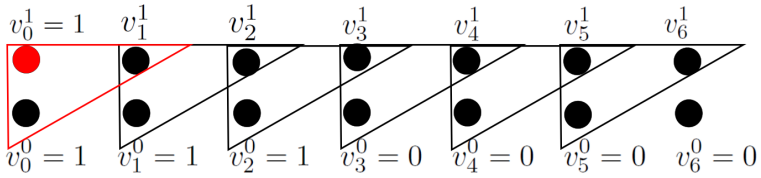
- For Backward-time forward-space (BTFS) scheme we have, *cf.* (29b)

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^{n+1} - v_m^{n+1}}{h} = 0 \quad \Rightarrow$$

$$(1-\bar{k})v_m^{n+1} + \bar{k}v_{m+1}^{n+1} = v_m^n \quad \text{where} \tag{48a}$$

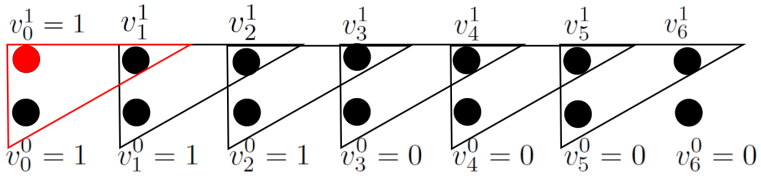
$$\bar{k} = \lambda a = \frac{ka}{h} \tag{48b}$$

- By writing the equations for points $0, 1, 2, 3, 4, 5$ (compare to $1, 2, 3, 4, 5, 6$ for BTBS),



- in this case the transformation matrix $\mathbf{A}$ ($\mathbf{v}^n = \mathbf{A}^{-n}\mathbf{v}^0 + \text{BC contribution}$)

$$\mathbf{A} = \begin{bmatrix} \bar{k} & 0 & 0 & 0 & 0 & 0 \\ (1-\bar{k}) & \bar{k} & 0 & 0 & 0 & 0 \\ 0 & (1-\bar{k}) & \bar{k} & 0 & 0 & 0 \\ 0 & 0 & (1-\bar{k}) & \bar{k} & 0 & 0 \\ 0 & 0 & 0 & (1-\bar{k}) & \bar{k} & 0 \\ 0 & 0 & 0 & 0 & (1-\bar{k}) & \bar{k} \end{bmatrix}$$



- Thus, all eigenvalues of $\mathbf{A}$ are $\bar{k}$ meaning that all eigenvalues of $\mathbf{A}^{-1}$ are $1/\bar{k}$ and

$$\rho(\mathbf{A}^{-1}) = \frac{1}{\bar{k}} \quad \Rightarrow$$

$$\text{BTFS scheme is stable if } \rho(\mathbf{A}^{-1}) = \frac{1}{\bar{k}} \leq \quad \Leftrightarrow \quad \bar{k} \geq 1 \tag{49}$$

- That is,

The IMPLICIT method of BTFS is CONDITIONALLY STABLE and large enough steps ($\bar{k} \geq 1$) must be taken for stability.

- This is a good example of an implicit method that IS NOT unconditionally stable. That is, it does not have the main advantage of most implicit methods (unconditional stability) yet is more expensive than explicit ones (in 2D and 3D) for this problem.
- The cause of this problem is again the wave (right-going with $a > 0$) not being consistent with FD grid. Although we cannot always make such arguments and stability of a method should be directly evaluated.
- Likewise BTBS method will only be conditionally stable for left-going wave.
- We will see that BTCS will be unconditionally stable regardless of the sign of a. This is in complete contrast to FTCS which was unconditionally unstable!
- In next sections we learn how to use von Neumann method to directly analyze the stability of linear FD methods.

2.1.10 Higher order PDEs: 2^{nd}order parabolic & hyperbolic PDEs

- Consider the solution of the parabolic PDE,

$$u_{,t} - Du_{,xx} = r \tag{50}$$

where $D = [L^2/T]$ (length squared / time) is the diffusion coefficient. The second term often is in the form $(Du_{,x})_{,x}$ which here for simplicity we have assumed D is constant.

- We can choose different options for FD difference to discretize this system.
- For example, Forward-time central-space (FTCS) scheme for the discrete solution v gives,

$$\frac{v_m^{n+1} - v_m^n}{k} - D\frac{v_{m+1}^n + v_{m-1}^n - 2v_m^n}{h^2} = r_m^n \quad \Rightarrow \tag{51a}$$

$$v_m^{n+1} = (1 - 2\bar{k})v_m^n + \bar{k}(v_{m-1}^n + v_{m+1}^n) + r_m^n \tag{51b}$$

$$\bar{k} = \frac{kD}{h^2} \qquad \text{Normalized time step for parabolic PDE} \tag{51c}$$

- Interestingly, as opposed to FTCS scheme for advection equation $\frac{v_m^{n+1}-v_m^n}{k}+a\frac{v_{m+1}^n-v_{m-1}^n}{2h}$ this scheme is conditionally stable, as one would expect from an explicit scheme.
- Equation 57c suggests $\bar{k}<\mathcal{O}(1)$ should provide the stable time step for this parabolic PDE, where $\mathcal{O}(1)$ is a constant number that we derive later. This number depends on the particular stencil used for the parabolic PDEs.
- Accordingly, we observe,

$$\bar{k}<\mathcal{O}(1) \qquad \Rightarrow k_{\max}\propto h^2 \qquad k_{\max}\text{ is the maximum stable time step} \tag{52}$$

  That is, we observe that for parabolic PDE $k_{\max}$ is proportional to k^2 rather than k for hyperbolic PDEs.
- This implies that for small grid sizes, the explicit parabolic FD schemes (and in fact FV, FEM, *etc.*) have a much more stringent time step requirement compared to explicit hyperbolic schemes.
- FD scheme can easily be applied to 2D and 3D diffusion equations as well. The 2D, 3D diffusion equation reads as,

$$u_{,t}-\nabla.(D\nabla u)=r \quad \text{for constant } u_{,t}-D\Delta u=u_{,t}-D(u_{,11}+u_{,22}+u_{,33})=r \tag{53}$$

- The forward time, forward space (FTFS) scheme for this equation is (2D version shown),

$$\frac{v_{m_xm_y}^{n+1}-v_{m_xm_y}^n}{k}-D\left\{\frac{v_{(m_x+1)m_y}^n+v_{(m_x-1)m_y}^n-2v_{m_xm_y}^n}{h_x^2}+\frac{v_{m_x(m_y+1)}^n+v_{m_x(m_y-1)}^n-2v_{m_xm_y}^n}{h_y^2}\right\}=r_{m_xm_y}^n \quad \Rightarrow \tag{54a}$$

$$v_m^{n+1}=(1-2\bar{k}_x-2\bar{k}_y)v_m^n+\bar{k}_x(v_{(m_x-1)m_y}^n+v_{(m_x+1)m_y}^n)++\bar{k}_y(v_{m_x(m_y-1)}^n+v_{m_x(m_y+1)}^n)+r_m^n \tag{54b}$$

$$\bar{k}_x=\frac{kD}{h_x^2},\quad \bar{k}_y=\frac{kD}{h_y^2} \qquad \text{Normalized time step for parabolic PDE} \tag{54c}$$

- Finally, to obtain an implicit scheme, we write FD equations at time step $n+1$ rather than n.
- For example in 1D, backward-time central-space (BTCS) scheme for the discrete solution v gives,

$$\frac{v_m^{n+1}-v_m^n}{k}-D\frac{v_{m+1}^{n+1}+v_{m-1}^{n+1}-2v_m^{n+1}}{h^2}=r_m^n \quad \Rightarrow \tag{55a}$$

$$(1+2\bar{k})v_m^{n+1}-\bar{k}(v_{m-1}^{n+1}+v_{m+1}^{n+1})=v_m^n+r_m^n \tag{55b}$$

$$\bar{k}=\frac{kD}{h^2} \qquad \text{Normalized time step for parabolic PDE (as in (57c))}$$

- This scheme will be stable for all $\bar{k}$.

2.1.10.1 Higher order PDEs: Hyperbolic wave equation

- Consider the wave equation,

$$u_{,tt}-c^2u_{,xx}=r \tag{56a}$$

$$\text{IC 1: } 0^{\text{th}}\text{ temporal derivative} \quad u(x,0)=u_0(x) \tag{56b}$$

$$\text{IC 2: } 1^{\text{st}}\text{temporal derivative} \quad \dot{u}(x,0)=\dot{u}_0(x) \tag{56c}$$

  where c is the wave speed.
- Similar to parabolic case we can discretize this system by FD. By using central-time central-space scheme we obtain,

$$\frac{v_m^{n+1}+v_m^{n-1}-2v_m^n}{k^2}-c^2\frac{v_{m+1}^n+v_{m-1}^n-2v_m^n}{h^2}=r_m^n \quad \Rightarrow \tag{57a}$$

$$v_m^{n+1}=-v_m^{n-1}+2(1-\bar{k}^2)v_m^n+\bar{k}^2(v_{m-1}^n+v_{m+1}^n)+r_m^n \tag{57b}$$

$$\bar{k}=\frac{kc}{h} \qquad \text{Normalized time step for hyperbolic wave equation} \tag{57c}$$

- Notice that this is a multi-step scheme, requiring the value of v_m^{n-1}.
- For $n=0$ (solution of first time step after IC) we need v_m^{-1} which does not exist!

- The trick is using initial 1st value at time step 0 by backward time difference

$$\dot{u}(mh,0) = \dot{u}_0(mh) = (\dot{u}_0)_m = \dot{v}(x = mh, 0) \approx \nabla_k[v_m^0] = \frac{v_m^0 - v_m^{-1}}{k} \quad \Rightarrow \tag{58a}$$

$$v_m^{-1} = v_m^0 - k(\dot{u}_0)_m = u_{0m} - k(\dot{u}_0)_m \tag{58b}$$

- Same process is applied to PDEs with higher temporal derivatives: by using initial temporal derivatives v_m^{-n} are formed.
- Similar to the parabolic case, and in contrast to the 1st order advection equation, this FTCS scheme is conditionally stable.
- The construction of implicit schemes is also straight forward. For example, by writing equations at time step $n+1$ rather than n and using backward time central space we obtain,

$$\frac{v_m^{n+1} + v_m^{n-1} - 2v_m^n}{k^2} - c^2 \frac{v_{m+1}^{n+1} + v_{m-1}^{n+1} - 2v_m^{n+1}}{h^2} = r_m^n \quad \Rightarrow \tag{59a}$$

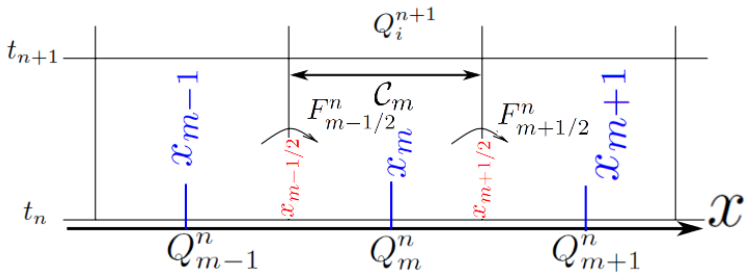
$$-\bar{k}^2 v_{m-1}^{n+1} (1 + 2\bar{k}^2) v_m^{n+1} - \bar{k}^2 v_{m+1}^{n+1} = -v_m^{n-1} + 2v_m^n + r_m^n \tag{59b}$$

$$\bar{k} = \frac{kc}{h} \qquad \text{As in (57c): normalized time step} \tag{59c}$$

2.2 Finite Volume (FV)

2.2.1 Finite Volume (FV) method description

- Consider the balance law with temporal flux q and spatial flux $f(q)$.



- The m^{th} grid cell is defined as,

$$\mathcal{C}_m = (x_{m-1/2}, x_{m+1/2}) \tag{60}$$

- Q_m^n: grid m average of q at time step n:

$$Q_m^n := \frac{1}{h_m} \int_{x_{m-1/2}}^{x_{m+1/2}} q(x, t_n)\, \mathrm{d}x = \frac{1}{h_m} \int_{\mathcal{C}_m} q(x, t_n)\, \mathrm{d}x \tag{61}$$

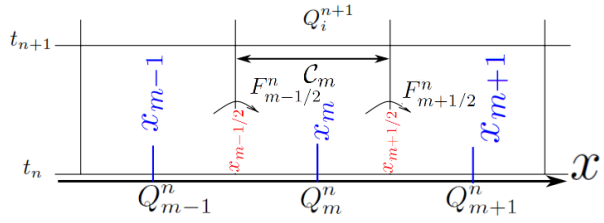
- Note that,
  - h_m is the average of the distances of the grid points to the left and right of x_m:

$$h_m = \frac{x_m - x_{m-1}}{2} + \frac{x_{m+1} - x_m}{2}$$

  - The only caveat is when FD time approximation of quantities such as $q_{,x}$ are required which stencils for nonuniform grids are used. Refer to [LeVeque, 2002] §6.17.1 for more information.
  - In general, however, definition of cell average values simplifies having varying grid sizes compared to FD scheme.
  - For uniform grid,

$$Q_m^n = q(x_m, t_n) + \mathcal{O}(h^2), \quad \text{when the exact solution } q \text{ is smooth enough } (q'' \text{ exists})$$

    that is the cell average is in fact a good approximation of the function value at the midpoint of the cell.



- The balance law for an equation in the form,

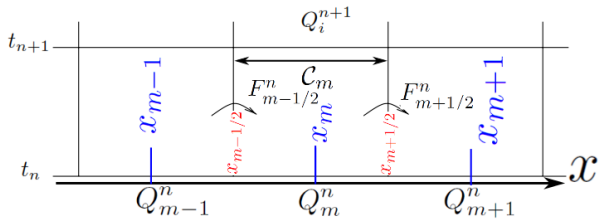
$$q_{,t} + \{f(q)\}_{,x} = 0 \tag{62}$$

is

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{C}_m} q(x,t)\,\mathrm{d}x = f(q(x_{m-1/2},t)) - f(q(x_{m+1/2},t)) \tag{63}$$

- By integrating over time step n: $[t_n\ t_{n+1}]$ we get,

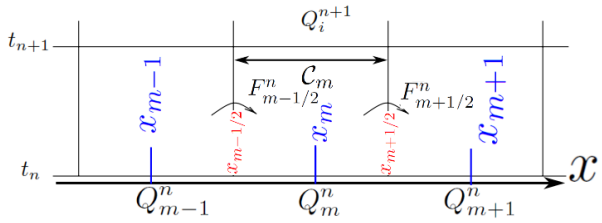
$$\int_{\mathcal{C}_m} q(x,t_{n+1})\,\mathrm{d}x - \int_{\mathcal{C}_m} q(x,t_n)\,\mathrm{d}x = \int_{t_n}^{t_{n+1}} f(q(x_{m-1/2},t))\,\mathrm{d}t - \int_{t_n}^{t_{n+1}} f(q(x_{m+1/2},t))\,\mathrm{d}t \tag{64}$$



- By rearranging and diving by h_m we obtain,

$$\frac{1}{h_m}\int_{\mathcal{C}_m} q(x,t_{n+1})\,\mathrm{d}x = \frac{1}{h_m}\int_{\mathcal{C}_m} q(x,t_n)\,\mathrm{d}x - \frac{1}{h_m}\left[\int_{t_n}^{t_{n+1}} f(q(x_{m+1/2},t))\,\mathrm{d}t - \int_{t_n}^{t_{n+1}} f(q(x_{m-1/2},t))\,\mathrm{d}t\right]$$

- We note that the LHS is the cell i average at time t_{n+1} which is expressed in terms of cell average at time t_n plus some spatial flux updates from the vertical boundaries of the cell at $x_{m-1/2}$ and $x_{m+1/2}$. However, the flux integrals from t_n to t_{n+1} cannot be evaluated exactly!

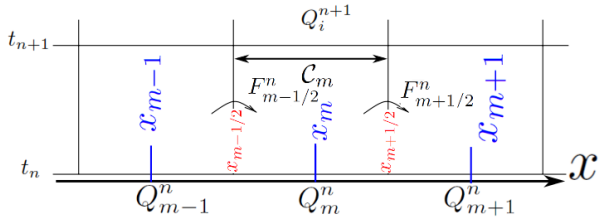


- This, however, suggests to study numerical methods of the form,

$$Q_m^{n+1} = Q_m^n - \frac{k}{h_m}(F^n_{m+1/2} - F^n_{m-1/2}), \text{where} \tag{65a}$$

$$k = t_{n+1} - t_n \qquad \text{time step size (which can easily be nonuniform)} \tag{65b}$$

$$F^n_{m\pm1/2} \approx \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m\pm1/2},t))\,\mathrm{d}t \qquad \text{some approximation of the average flux along } x_{m\pm1/2} \tag{65c}$$

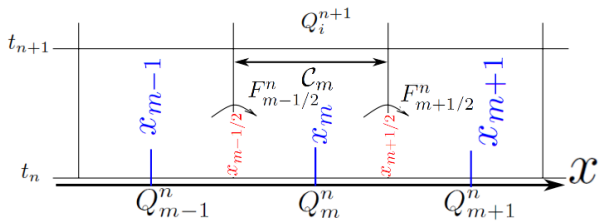


- For hyperbolic problems information propagates with finite speed, so it is reasonable to assume that we can obtain $F^n_{m-1/2}$ based on the values Q^n_{m-1} and Q^n_m: the cell averages on the two sides at the beginning of the time step,

$$F^n_{m-1/2} = \mathcal{F}(Q^n_{m-1}, Q^n_m), \qquad F^n_{m+1/2} = \mathcal{F}(Q^n_m, Q^n_{m+1}) \tag{66}$$

where $\mathcal{F}$ is some numerical flux function.

- The approximation of computing spatial flux integrals by the the cell averages at the beginning of the step results in explicit methods. This assumption can also be applied to parabolic equations (once stable time stables are used).



- Then the method (65a) becomes,

$$Q_m^{n+1} = Q_m^n - \frac{k}{h_m}\left[F^n_{m+1/2} - F^n_{m-1/2}\right] = Q_m^n - \frac{k}{h_m}\left[\mathcal{F}(Q_m^n, Q_{m+1}^n) - \mathcal{F}(Q_{m-1}^n, Q_m^n)\right] \tag{67}$$

- This explicit method, in general has a three-point stencil meaning that the value Q_m^{n+1} only depends on Q^n_{m-1}, Q^n_m, and Q^n_{m+1} from previous time step.

- Note that the method is in conservative form in discrete setting. By summing the equations (65a) (each multiplied by by h_m) for cells I to J the spatial flux integrals $F^n_{j+1/2}$ cancel out for the interior vertical boundaries and we obtain,

$$\sum_{m=I}^{J} h_m Q_m^{n+1} = \sum_{m=I}^{J} h_m Q_m^n - k\left(F^n_{J+1/2} - F^n_{I-1/2}\right) \tag{68}$$

- This is the counter part to the continuum version,

$$\int_{x_{I-1/2}}^{x_{J+1/2}} q(x, t_{n+1})\,\mathrm{d}x = \int_{x_{I-1/2}}^{x_{J+1/2}} q(x, t_n)\,\mathrm{d}x - \int_{t_n}^{t_{n+1}} f(q(x_{J+1/2}, t)\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} f(q(x_{I-1/2}, t)\,\mathrm{d}t \tag{69}$$

which is the summation of contiuum level (before discretization) of integrals (64) for cells I to J.

2.2.2 FV examples from 1st order hyperbolic PDEs

- To illustrate the importance of numerical flux function (66) we consider three difference options.

- We consider the hyperbolic system (62),

$$q_{,t} + \{f(q)\}_{,x} = 0 \qquad \text{PDE} \tag{70a}$$
$$q(x, t=0) = q_0(x) \qquad \text{IC} \tag{70b}$$

- Specifically, we consider the linear case of (70a) which is the advection equation,

$$q_{,t} + aq_{,x} = 0 \qquad f(q) = aq \tag{71}$$

where for simplicity it is assumed the wave speed $a(x,t) = a > 0$ is constant and positive. That is we consider a right-going wave. If $a(x,t)$ we get a source term of the form $-a_{,x}(x,t)q$ which is ignored here as it does not change the nature of the influence of different flux options (lower order derivatives).

2.2.2.1 1. Average fluxes

- The average flux option means that we use the average of the fluxes from the two sides,

$$\mathcal{F}(Q^n_{m-1}, Q^n_m) = \frac{1}{2}\left(f(Q^n_{m-1}) + f(Q^n_m)\right) \qquad \Rightarrow \mathcal{F}(Q^n_m, Q^n_{m+1}) = \frac{1}{2}\left(f(Q^n_m) + f(Q^n_{m+1})\right) \tag{72}$$

- Then, by plugging this into (67) we obtain,

$$\begin{aligned} Q_m^{n+1} &= Q_m^n - \frac{k}{h_m}\left[\mathcal{F}(Q_m^n, Q_{m+1}^n) - \mathcal{F}(Q_{m-1}^n, Q_m^n)\right] \\ &= Q_m^n - \frac{k}{h_m}\left[\frac{1}{2}\left(f(Q_m^n) + f(Q_{m+1}^n)\right) - \frac{1}{2}\left(f(Q_{m-1}^n) + f(Q_m^n)\right)\right] \quad \Rightarrow \\ Q_m^{n+1} &= Q_m^n - \frac{k}{2h_m}\left[f(Q_{m+1}^n) - f(Q_{m-1}^n)\right] \end{aligned} \tag{73}$$

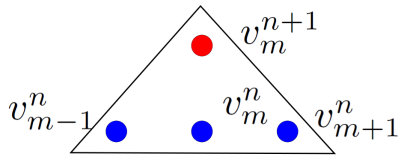
- This can be expressed in the FD form,

$$\frac{Q_m^{n+1}-Q_m^n}{k}+\frac{f(Q_{m+1}^n)-f(Q_{m-1}^n)}{2h_m}=0 \tag{74}$$

- Specifically, if we consider the simple linear advection PDE (71), equation (74) becomes,

$$\frac{Q_m^{n+1}-Q_m^n}{k}+a\frac{Q_{m+1}^n-Q_{m-1}^n}{2h_m}=0 \tag{75}$$

which is forward-time, central-space (FTCS) scheme discussed in (27c). As discussed under (34) this scheme is unconditionally unstable!



- So simply using the average fluxes not only may affect the accuracy (compared to correct fluxes) may also render the method unstable!

2.2.2.2 2. Lax-Friedrichs fluxes

- To simplify the discussion, we assume that the spatial grid is uniform (extensions to nonuniform can be done easily).
- To address the problem with average fluxes, Lax-Friedrichs fluxes modify them by adding a jump part of q values.

$$\mathcal{F}(Q_{m-1}^n,Q_m^n)=\frac{1}{2}\left(f(Q_{m-1}^n)+f(Q_m^n)\right)-\frac{h}{2k}\left(Q_m^n-Q_{m-1}^n\right) \quad \Rightarrow \tag{76a}$$

$$\mathcal{F}(Q_m^n,Q_{m+1}^n)=\frac{1}{2}\left(f(Q_m^n)+f(Q_{m+1}^n)\right)-\frac{h}{2k}\left(Q_{m+1}^n-Q_m^n\right) \tag{76b}$$

- Then, by plugging this into (67) we obtain,

$$\begin{aligned}
Q_m^{n+1}&=Q_m^n-\frac{k}{h}\left[\mathcal{F}(Q_m^n,Q_{m+1}^n)-\mathcal{F}(Q_{m-1}^n,Q_m^n)\right]\\
&=Q_m^n-\frac{k}{h}\left[\left\{\frac{1}{2}\left(f(Q_m^n)+f(Q_{m+1}^n)\right)-\frac{h}{2k}\left(Q_{m+1}^n-Q_m^n\right)\right\}-\left\{\frac{1}{2}\left(f(Q_{m-1}^n)+f(Q_m^n)\right)-\frac{h}{2k}\left(Q_m^n-Q_{m-1}^n\right)\right\}\right] \quad \Rightarrow\\
Q_m^{n+1}&=Q_m^n-\frac{k}{2h}\left[f(Q_{m+1}^n)-f(Q_{m-1}^n)\right]+\frac{1}{2}\left[Q_{m+1}^n+Q_{m-1}^n-2Q_m^n\right]
\end{aligned} \tag{77}$$

- This can be expressed in the FD form,

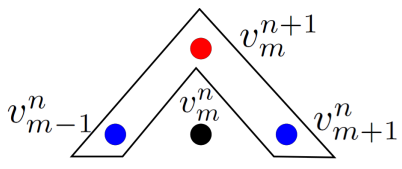
$$\frac{Q_m^{n+1}-\frac{1}{2}\left(Q_{m+1}^n+Q_{m-1}^n\right)}{k}+\frac{f(Q_{m+1}^n)-f(Q_{m-1}^n)}{2h}=0 \tag{78}$$

- This is the Lax-Friedrichs scheme we discussed in FD scheme. To see the connection, we consider the case that $f(q)=aq$, *i.e.,* the linear advection equation in (71). In this case, (78) becomes,

$$\frac{Q_m^{n+1}-\frac{1}{2}\left(Q_{m+1}^n+Q_{m-1}^n\right)}{k}+a\frac{Q_{m+1}^n-Q_{m-1}^n}{2h}=0 \tag{79}$$

- This is exactly, the Lax-Friedrichs scheme we had in (27d),

$$\frac{v_m^{n+1}-\frac{1}{2}\left(v_{m-1}^n+v_{m_1}^n\right)}{k}+a\frac{v_{m+1}^n-v_{m-1}^n}{2h}=0 \qquad \text{Lax-Friedrichs}$$



2.2.2.3 2. Lax-Friedrichs fluxes: Why does it work?

- We know from the discussions FD that Lax-Friedrichs scheme is conditionally stable. That is, the added jump terms in (76) stabilize the unconditionally unstable FTCS scheme.
- Why adding the jump conditions make the scheme (conditionally) stable?
- To understand the underlying mechanism, we rearrange (77) in the following form,

$$Q_m^{n+1} = Q_m^n - \frac{k}{2h}\left[f(Q_{m+1}^n) - f(Q_{m-1}^n)\right] + \frac{1}{2}\left[Q_{m+1}^n + Q_{m-1}^n - 2Q_m^n\right] \Rightarrow$$
$$\frac{Q_m^{n+1} - Q_m^n}{k} + \frac{f(Q_{m+1}^n) - f(Q_{m-1}^n)}{2h} - \frac{h^2}{2k}\frac{Q_{m+1}^n + Q_{m-1}^n - 2Q_m^n}{h^2} = 0 \tag{80}$$

- In the present form, it is obvious that the FV (FD) equation approximates the following equation,

$$q_{,t} + \{f(q)\}_{,x} - \frac{h^2}{2k} q_{,xx} = 0 \tag{81}$$

- Given that the term $\{f(q)\}_{,x}$ is first term derivative in x the original equation is hyperbolic, and
- We add the parabolic operator $-\frac{h^2}{2k} q_{,xx}$ to the original hyperbolic equation. Note that $q_{,t} - \frac{h^2}{2k} q_{,xx}$ is a parabolic equation.
- The addition of parabolic (diffusion) operator tends to damp instabilities that arise from FTCS scheme!
- This added diffusion operator results in the subtle change of $\frac{Q_m^{n+1} - Q_m^n}{k}$ in (74) (from the use of average fluxes) to $\frac{Q_m^{n+1} - \frac{1}{2}\left(Q_{m-1}^n + Q_{m-1}^n\right)}{k}$ in (79) from Lax-Friedrichs fluxes. This subtle change of the temporal difference, in fact as we saw correspond to an added diffusion operator.
- Question on consistency: We are actually solving the equation (81)

$$q_{,t} + \{f(q)\}_{,x} - \frac{h^2}{2k} q_{,xx} = 0$$

by using Lax-Friedrichs fluxes. This is clearly difference from the underlying PDE (62)

$$q_{,t} + \{f(q)\}_{,x} = 0$$

Here are some questions and observations,

  - The solved system (with added diffusion term) is different from the actual PDE. Is the system consistent, meaning that the solution of the new system converge to the correct solution?
  - We recall that stable time stable for the hyperbolic equation scales as $k_{\max} \propto h$ (*e.g.*, for $f(q) = aq$ for advection equation we have $\bar{k} = \frac{ak}{h} \leq 1$.
  - Thus in the term $\frac{h^2}{2k}$ in (81) if the ratio $\frac{k}{h}$ is kept fixed for stability concerns and we let $h \to 0$ *i.e.*, by mesh refinement the numerical advection coefficient $\frac{h^2}{2k}$ approaches zero as finer grids are used.
  - So, Lax-Friedrichs scheme is consistent with the underlying hyperbolic equation.
  - However, the added diffusion implies that the system can by overly diffusive!
  - The over damping of the Lax-Friedrichs can be seen in the comparisons shown with leapfrog scheme from FD section.

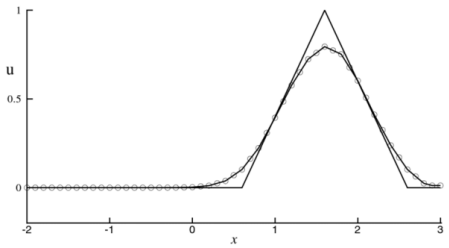


Figure 1.6. *A solution of the Lax–Friedrichs scheme,* $\lambda = 0.8$.

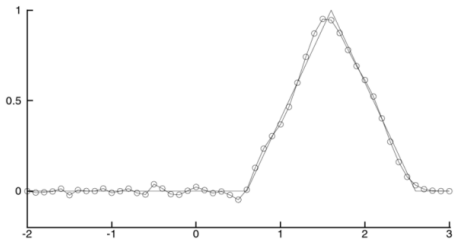


Figure 1.8. *A solution computed with leapfrog scheme,* $\lambda = 0.8$.

2.2.2.4 3. Upstream fluxes (Riemann solution)

- As a reminder from (65c) for a cell on its two cell boundaries we need to define flux average values (only shown for $x_{m-1/2}$),

$$F^n_{m-1/2} = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m-1/2},t))\,\mathrm{d}t$$

- In (66) we mentioned that for hyperbolic equations we can express this values as a function of the initial values (initial conditions) from time step n on the two sides of $x_{m-1/2}$,

$$F^n_{m-1/2} = \mathcal{F}(Q^n_{m-1}, Q^n_m), \qquad F^n_{m+1/2} = \mathcal{F}(Q^n_m, Q^n_{m+1})$$

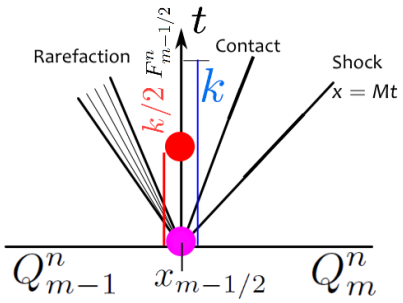
- Different choices for numerical flux function F^n can be used as we observed average and Lax-Friedrichs fluxes options before.
- $F^n_{m-1/2}$ can also be exactly solved from the initial conditions Q^n_{m-1} and Q^n_m.

2.2.2.5 Riemann problem set-up

This forms a Riemann problem set-up where,

- We solve the solution for all x, t with two distinct ICs Q^n_{m-1} and Q^n_m
- Having the solution along the vertical line $x_{m-1/2}$ we can calculate $q(x_{m-1/2}, t)$ for $t_n \le t \le t_{n+1}$
- Having q along this line, we can calculate $f(q(x_{m-1/2}, t))$ for $t_n \le t \le t_{n+1}$.
- We can then calculate $F^n_{m-1/2}$ from $f(q(x_{m-1/2}, t))$,

$$F^n_{m-1/2} = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m-1/2},t))\,\mathrm{d}t$$



Given that we can conceptually calculate $f(q(x_{m-1/2}, t))$ for $t_n \le t \le t_{n+1}$, there are (at least) three different levels to compute/approximate $F^n_{m-1/2}$

1. Exact integration of f(q)

$$F^n_{m-1/2} = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m-1/2},t))\,\mathrm{d}t$$

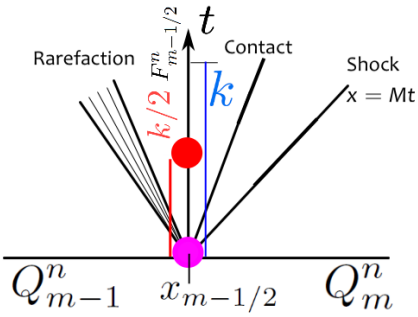
2. Using the mid-point value for f(q)

$$F^n_{m-1/2} = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m-1/2},t))\,\mathrm{d}t \approx f(q(x_{m-1/2}, \frac{t_n + t_{n+1}}{2}))$$

3. Using the start value f(q)

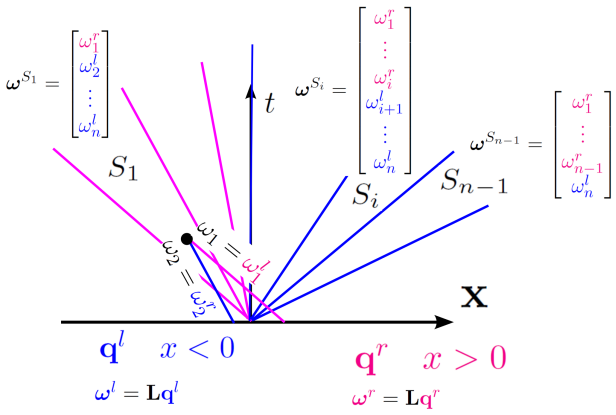
$$F^n_{m-1/2} = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m-1/2},t))\,\mathrm{d}t \approx f(q(x_{m-1/2}, t_n))$$

the advantage of the this approach is that we do not need to integrate source terms along the characteristics. For hyperbolic systems the difference between this and average option arises when source term is nonzero. This option, clearly is the least accurate one and special care must be taken in its use.



2.2.2.6 Solution of Riemann solution (Upstream solution)

- As discussed before (and shown in the previous figure) Riemann solution may involve regions with shocks and rarefaction waves for quasi-linear systems of hyperbolic PDEs.
- For semi-linear PDEs, we discussed two different approaches to derive Riemann solutions for different sectors in space time.
- The Riemann solution in each sectors are obtained by using the upstream fluxes from the characteristics coming to that sector



2.2.2.7 3. Upstream fluxes (Riemann solutions for first order hyperbolic PDEs)

- Again, we consider the hyperbolic system (62)

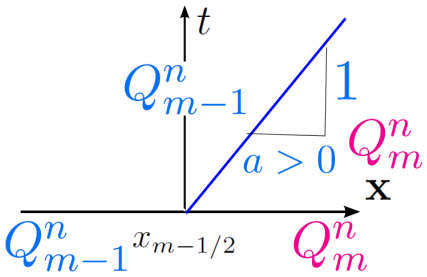
$$\begin{aligned} q_{,t} + \{f(q)\}_{,x} &= 0 && \text{PDE} \\ q(x, t=0) &= q_0(x) && \text{IC} \end{aligned}$$

- The Riemann solution for a general nonlinear flux $f(q)$ may not be trivial.
- For nonlinear flux, we either directly solve the nonlinear Riemann solution, or use approximate Riemann solutions similar to (76).
- We consider that we are solving the linear advection equation where $f(q) = aq$ (71),

$$q_{,t} + aq_{,x} = 0$$

  - We consider a case $a > 0$. The Riemann solution is shown,
  - On $x_{m-1/2}$ line we have $q(x_{m-1/2}, t) = Q^n_{m-1}$, $t_n \le t \le t_{n+1} \Rightarrow$
  - $f(q(x_{m-1/2}, t)) = f(Q^n_{m-1}) = aQ^n_{m-1}$.
  - From (65c) the flux average value is,

$$F^n_{m-1/2} = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m-1/2}, t))\,\mathrm{d}t = \frac{1}{k}\int_{t_n}^{t_{n+1}} aQ^n_{m-1}\,\mathrm{d}t = aQ^n_{m-1}$$



  - Similarly

$$F^n_{m+1/2} = aQ^n_m$$

- Then from the update equation (67) we have (consider uniform grid),

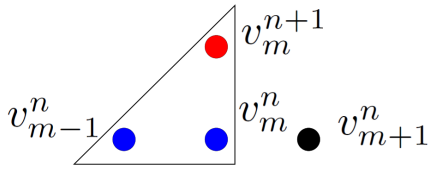
$$Q^{n+1}_m = Q^n_m - \frac{k}{h}\left[F^n_{m+1/2} - F^n_{m-1/2}\right] = Q^n_m - \frac{k}{h}\left[aQ^n_m - aQ^n_{m-1}\right] = Q^n_m - \bar{k}\left[Q^n_m - Q^n_{m-1}\right], \quad \Rightarrow \tag{82a}$$

$$Q^{n+1}_m = (1-\bar{k})Q^n_m + \bar{k}Q^n_{m-1}, \qquad \text{where} \bar{k} = \frac{ka}{h} \text{ is the normalized time step} \tag{82b}$$

- This is the FD FTBS update from (35b).
- To better see the connection, we can write (82a) in the FD form,

$$\frac{Q^{n+1}_m - Q^n_m}{k} + a\frac{Q^n_m - Q^n_{m-1}}{h} = 0 \tag{83}$$

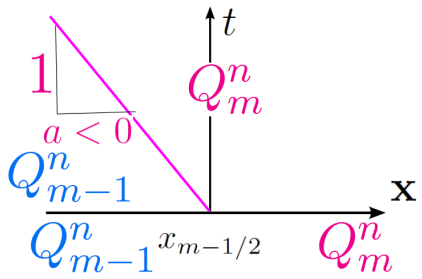
- which is the FD FTBS scheme (27b),



- As we know from FD section, FTBS scheme is conditionally stable for $a > 0$.
- So, by choosing the Riemann solution, *i.e.,* physically and mathematically correct flux, the also ensure the explicit method's conditional stability (compared to FTCS for example).
  - For $a < 0$ we have $q(x_{m-1/2}, t) = Q_m^n$, $t_n \le t \le t_{n+1} \Rightarrow$
  - $f(q(x_{m-1/2}, t)) = f(Q_m^n) = aQ_m^n \Rightarrow$

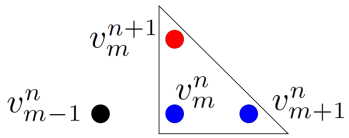
$$F_{m-1/2}^n = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m-1/2}, t))\,\mathrm{d}t = \frac{1}{k}\int_{t_n}^{t_{n+1}} aQ_m^n\,\mathrm{d}t = aQ_m^n$$

  - Similarly $F_{m+1/2}^n = aQ_{m+1}^n$



- The update equation (67) becomes,

$$Q_m^{n+1} = Q_m^n - \frac{k}{h}\left[F_{m+1/2}^n - F_{m-1/2}^n\right] = Q_m^n - \frac{k}{h}\left[aQ_{m+1}^n - aQ_m^n\right] = (1+\bar{k})Q_{m+1}^n - \bar{k}Q_m^n \tag{84}$$



- which is the FD FTFS scheme (27a). Equation (84) is written in FD form,

$$\frac{Q_m^{n+1} - Q_m^n}{k} + a\frac{Q_{m+1}^n - Q_m^n}{h} = 0 \tag{85}$$

2.2.3 Properties of the numerical flux function

- Through the advection equation $q_{,t} + aq_{,x} = 0$ we observed that upstream fluxes (through the use of Riemann solutions) pick the stable stencil (FTBS for $a > 0$, FTFS for $a < 0$).
- In general, to establish consistency of a FV scheme, The following conditions are required from the numerical flux:
  1. Consistent: $\mathcal{F}(Q, Q) = f(Q)$: If the cell averages are equal from both sides, the numerical flux recovers the exact flux function $f(Q)$.
  2. Continuous: If the two side cell averages are not equal, we still need a continuity condition that numerical flux $\mathcal{F}(Q^-, Q^+)$ converges to $f(Q)$ as $Q^- \to Q^+$. This continuity condition often is expressed in the form of Lipschitz continuity:

$$\exists L > 0 \text{ such that } |\mathcal{F}(Q^-, Q^+) - f(Q)| < L\max(|Q^- - Q|, |Q^+ - Q|) \tag{86}$$

     The Lipschitz continuity condition is also represented in an alternative form,

$$\exists L > 0 \text{ such that } |\mathcal{F}(Q^-, Q^+) - \mathcal{F}(\bar{Q}^-, \bar{Q}^+)| < L\max(|Q^- - \bar{Q}^-|, |Q^+ - \bar{Q}^+|) \tag{87}$$

  3. Conservative: In the 1D examples shown we did not explicitly included the normal vector of the interface $\mathbf{n}$ on the boundary of cells. In 2D and 3D it is more natural to include $\mathbf{n}$ in the formulation of numerical flux,

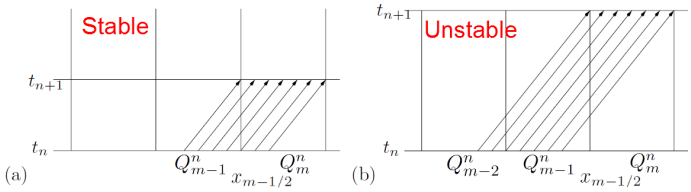
$$\mathcal{F}(Q^-, Q^+) \to \mathcal{F}(Q^-, Q^+, \mathbf{n})$$

     where Q^-, Q^+ are cell averages from the two sides of an interface and $\mathbf{n}$ is the (spatial) normal vector on the interface from $-$ to $+$ side. By incorporating the normal vector in the definition of the flux we require,

$$\mathcal{F}(Q^-, Q^+, \mathbf{n}) = -\mathcal{F}(Q^+, Q^-, -\mathbf{n}) \tag{88}$$

     As an example of the physical interpretation of this condition, consider that the traction vector from side $+$ to $-$ is the opposite from the side $-$ to $+$. This example, refers to the Newton's principle of action-reaction.

- It is easy to check consistency, continuity for all the fluxes considered if $f(q)$ is Lipschitz continuous: $\exists C \mid |f(q^+) - f(q^-)| \le C|q^+ - q^-|$. These fluxes are also conservative (although we did not carry normal vector $\mathbf{n}$ in our derivations for the simple 1D case).
- Consistency, continuity, and conservativity of numerical flux are often required for consistency, convergence, and stability of numerical methods using numerical fluxes (FV and discontinuous Galerkin).
- Still, as we saw with average flux option, these conditions are not sufficient for stability.
- These conditions are often assumed for numerical flux functions (or they are even essential) to facilitate the mathematical analysis of the method (consistency, convergence, and stability).



2.2.4 Stability limit of explicit finite volume methods (hyperbolic PDEs)

- For FD scheme we showed that for a first order hyperbolic PDE the stable time step must have been $\bar{k} \leq 1$, *i.e.,* the CFL condition was satisfied. FV update equations looked similar to FD onces, *e.g.,* FTBS, FTCS, FTFS, Lax-Friedrichs.
- This ensures that the numerical speed of information propagation is greater than or equal to the physical wave speed.
- Figure above shows that if CFL condition is violated on vertical boundary of cell $x_{m-1/2}$ information no longer is only influenced by the two side averages Q^n_{m-1} and Q^n_m but also influenced by Q^n_{m-2}.
- This manifests itself in instability as the influence of Q^n_{m-2} is not included in the numerical flux $F^n_{m-1/2} = \mathcal{F}(Q^n_{m-1}, Q^n_m)$.

2.2.5 FV example for 2nd order PDEs: Parabolic equation

- Consider the parabolic equation (50) (written for q),

$$q_{,t} - Dq_{,xx} = r \tag{89}$$

- This can be written as,

$$q_{,t} - (Dq_{,x})_{,x} = r \tag{90}$$

  In fact, (90) is the correct form of diffusion equation (where D appears inside the $(.)_{,x}$. For constant D this reduces to (89).
- In any case, (90) can be written as,

$$q_{,t} + \{f(q_{,x})\}_{,x} = r, \qquad \text{where} \qquad f(q_{,x}) = -Dq_{,x} \tag{91}$$

- This is the same as the PDE form of balance law (63), *i.e.,* (62) $q_{,t} + \{f(q)\}_{,x} = 0$,
- with the difference that the flux function f depends on $q_{,x}$ rather than q.
- In general for transient problems (hyperbolic, parabolic) we can express PDE and the flux function f in the form,

$$q_{,t} + \{f(q, q_{,x})\}_{,x} = r \tag{92}$$

- The formulation of FV is similar to the hyperbolic case. That is, we obtain,

$$Q^{n+1}_m = Q^n_m - \frac{k}{h}(F^n_{m+1/2} - F^n_{m-1/2}), \text{where} \tag{93a}$$

$$F^n_{m\pm1/2} \approx \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m\pm1/2}, t), q_{,x}(x_{m\pm1/2}, t))\,\mathrm{d}t \qquad \text{some approximation of the average flux along } x_{m\pm1/2} \tag{93b}$$

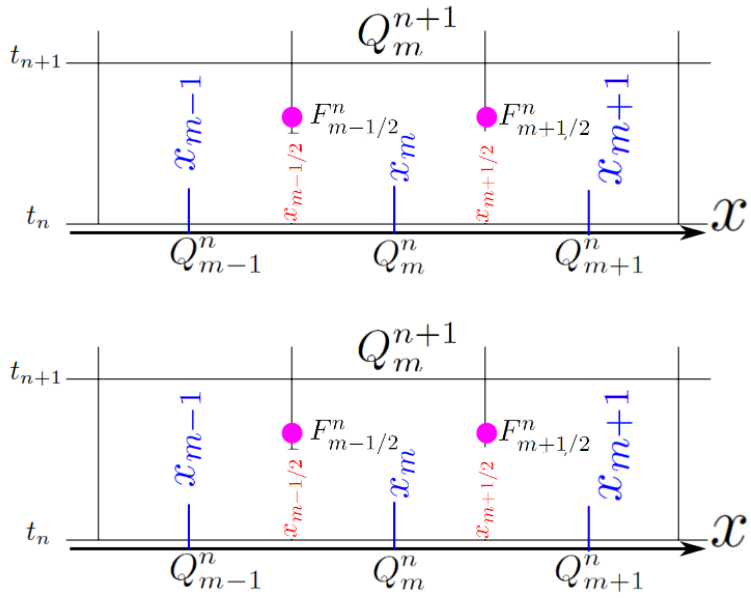
- Similar to (66), we can approximate average flux F, with numerical flux function $\mathcal{F}$, which depends only on inflow (previous step t_n) values:

$$F^n_{m-1/2} = \mathcal{F}(Q^n_{m-1}, Q^n_m), \qquad F^n_{m+1/2} = \mathcal{F}(Q^n_m, Q^n_{m+1}) \tag{94}$$

- We will demonstrate how we can apply FV method for the solution of diffusion problem (91). We assume uniform grid is used $h_m = h$ and D is constant.

2.2.5.1 FV example for Diffusion equation $q_{,t} - (Dq_{,x})_{,x} = 0$

- For the equation (91) the spatial flux is $f(q_{,x}) = -Dq_{,x}$.
- As shown in the figure , in the absence of hyperbolic equation characteristics, there are no upstream values. So,
- We cannot solve the Riemann solution for an underlying hyperbolic equation.
- We have two options:
  1. (Approximate flux): We use a reasonable numerical flux for the parabolic equation based on t_n values.
  2. (Exact flux): We actually solve a parabolic problem with Q^n_{m-1}, Q^n_m on the two sides of $x_{m-1/2}$ and evaluate an approximate value for (93b) (of exactly integrate it).



2.2.5.2 FV example for Diffusion equation: 1. Approximate flux

- In the absence of exact solution for $f(q, q_{,x}) = -Dq_{,x}$ at $x_{m-1/2}$ we use a FD approximation for the flux (uniform grid is assumed),

$$F^n_{m-1/2} = \mathcal{F}(Q^n_{m-1}, Q^n_m) = -D(q_{,x})^n_{m-1/2} \approx -D\frac{Q^n_m - Q^n_{m-1}}{h} \tag{95}$$

- Similarly $F^n_{m+1/2} = -D\frac{Q^n_{m+1} - Q^n_m}{h}$.
- By plugging the F's into (93a): $Q^{n+1}_m = Q^n_m - \frac{k}{h}(F^n_{m+1/2} - F^n_{m-1/2})$ we obtain,

$$Q^{n+1}_m = Q^n_m + \frac{kD}{h^2}\left(Q^n_{m+1} + Q^n_{m-1} - 2Q^n_m\right) = (1 - 2\bar{k})Q^n_m + \bar{k}(Q^n_{m-1} + Q^n_{m+1}) \tag{96}$$

- For normalized time step for the parabolic PDE

$$\bar{k} = \frac{kD}{h^2}$$

- This is the same as the FD update equation (51b): $v^{n+1}_m = (1 - 2\bar{k})v^n_m + \bar{k}(v^n_{m-1} + v^n_{m+1}) + r^n_m$ (here r is set to zero).
- In fact, we can be written in FD form,

$$\frac{Q^{n+1}_m - Q^n_m}{k} - D\frac{Q^n_{m+1} + Q^n_{m-1} - 2Q^n_m}{h^2} = 0 \tag{97}$$

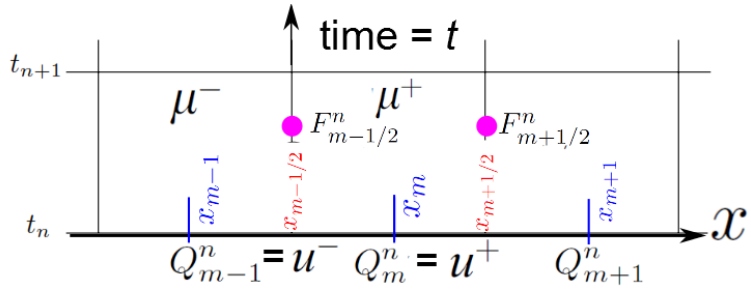
- which clearly is a FD approximation for (89): $q_{,t} - Dq_{,xx} = 0$ ($r = 0$).
- We observe, that for this parabolic equation, FV stencil update reduced to what we had obtained from FD FTCS equation, *cf.* §2.1.10.
- As discussed in FD schemes, the stability limit of this explicit scheme requires $\bar{k} < \alpha$ for a constant α which results in stable time stable scaling of $k_{\max} \propto \frac{h^2}{D}$.

2.2.5.3 FV example for Diffusion equation: 2. Exact flux

- The figure shows Riemann-like problem set up where the two values Q^n_{m-1} and Q^n_m from time step t_n are used as initial conditions for the diffusion equation:

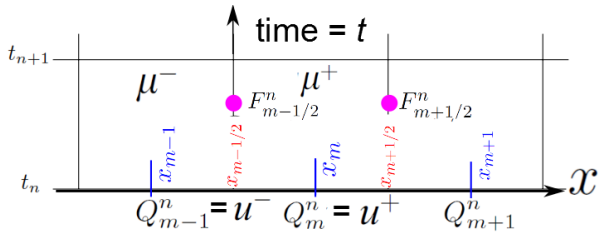
$$u_t - \mu u_{,xx} = 0$$

  where μ is the diffusion coefficient.
- To model a material interface with two distinct μ, μ^- and μ^+ left and right sides of an interface, respectively.



- In [Lorcher et al., 2008] a solution for this initial condition is obtained,

$$f(0,t) = \mu^+ \frac{\partial w^+}{\partial \xi_1}(0,t) = \frac{[\![u]\!]\sqrt{\mu^+\mu^-}}{\sqrt{\pi t}\left(\sqrt{\mu^+}+\sqrt{\mu^-}\right)} + \frac{\sqrt{\mu^+}f^-_{\vec{n}} + \sqrt{\mu^-}f^+_{\vec{n}}}{\sqrt{\mu^+}+\sqrt{\mu^-}}$$

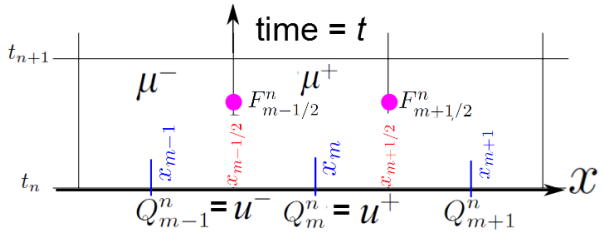


where $f(0,t)$ is the spatial flux function evaluated at the material interface, *e.g.*, $x_{m-1/2}$ in the figure and $f^\pm_{\vec{n}}$ are spatial fluxes from the two sides $f^\pm_{\vec{n}} = -\mu^\pm u^\pm_{,x}$.

- After integration of $f(0,t)$ in time (that is in the form of equation $F^n_{m\pm1/2} = \frac{1}{k}\int_{t_n}^{t_{n+1}} f(q(x_{m\pm1/2},t)$ [Lorcher et al., 2008] obtains ($g_{\vec{n}} = F^n_{m\pm1/2}$),

$$g_{\vec{n}} := \frac{1}{\Delta t}\int_0^{\Delta t} f(0,t)\mathrm{d}t = \frac{2[\![u]\!]\sqrt{\mu^+\mu^-}}{\sqrt{\Delta t}\sqrt{\pi}\left(\sqrt{\mu^+}+\sqrt{\mu^-}\right)} + \frac{\sqrt{\mu^+}f^-_{\vec{n}} + \sqrt{\mu^-}f^+_{\vec{n}}}{\sqrt{\mu^+}+\sqrt{\mu^-}}$$

- This flux can be used in the context of a FV or discontinuous Galerkin method.



- There are few points to observe,

1. Unlike hyperbolic problems where there is a finite wave speed and for small time steps $k = t_{n+1} - t_n$ solution and flux on $x_{m-1/2}$ only depends on the two side solutions Q^n_{m-1} and Q^n_m ($u^\pm$ in the figure) in parabolic equations solution depends on the values of ALL cells (*e.g.*, $\mathcal{C}_{m+1}$ in the figure) and the assumption of having constant IC on either side of the interface no longer makes sense for the duration of the solution k.
2. However, when the stable time step is used ($k_{\max} \propto \frac{h^2}{D}$) the error in ignoring nonadjacent cells does not cause numerical instability.
3. In fact the fluxes based on the exact solution of the diffusion equation is a much better option than the previous FD type flux approximation (95) $F^n_{m-1/2} = \mathcal{F}(Q^n_{m-1}, Q^n_m) = -D\frac{Q^n_m - Q^n_{m-1}}{h}$.

2.2.6 FV example for 2nd order PDEs: Hyperbolic equation

- Consider the balance of linear momentum in 1D,
$$p_{,t} + (-s)_{,x} = \rho b \tag{98}$$
- $p = \rho v$ is linear momentum, $s = E\epsilon$ is stress, and b is the body force.
- $v = \dot{u}$ is displacement, and $\epsilon = u_{,x}$ is strain, ρ is mass density, and E is elastic modulus.
- As discussed before, we can form a system of two first order equations,
$$p_{,t} + (-s)_{,x} = \rho b \tag{99a}$$
$$s_{,t} + -\frac{E}{\rho} p_{,x} = 0 \tag{99b}$$
- (100b) is in conservation law form, where as the compatibility equation (100a) should actually be written as $\epsilon_{,t} - v_{,x} = 0$. If ρ or E are variables their contributions from $\epsilon_{,t} - v_{,x} = 0$ should be added to (100a) (although $E_{,t}$ is often zero).
- We can use one from the pair s or ϵ and one from the pair p, v to have two first order PDEs.
- This needs careful formulation of the continuum problem (as in (100b)) solution to Riemann solutions (in the context of FV and DG methods), and discrete formulation.
- If material properties such as ρ and E are constant, the discussion simplifies.
- We choose s and v fields to write the system of equations.
- By using v instead of p in (100b) the equation no longer is in conservation law form.
- But this choice simplifies the Riemann solutions when E and ρ jump across an interface (because s and v do not jump across the material interface).
- Again, this practical consideration is not-important as we simply assume E and ρ to be constant which simplifies the formulation of the FV method.
- We also assume $b = 0$. The source term contributions can be easily added to FV formulation.
- We later comment on variable coefficient FV formulations.
- Using s and v for constant E and ρ and zero b (99) is written as,
$$s_{,t} + (-E)v_{,x} = 0 \tag{100a}$$
$$v,t + (-\frac{1}{\rho})s_{,x} = 0 \tag{100b}$$
- This, along with initial conditions, can be written in the form,
$$\mathbf{q}_{,t} + (\mathbf{A}\mathbf{q})_{,x} = 0, \qquad \text{where} \qquad \mathbf{q} = \begin{bmatrix} s \\ v \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & -E \\ -\frac{1}{\rho} & 0 \end{bmatrix} \qquad \text{PDE} \tag{101a}$$
$$\mathbf{q}(x, t = 0) = \mathbf{q}_0(x) \qquad \text{that is} \qquad \begin{bmatrix} s(x, t = 0) \\ v(x, t = 0) \end{bmatrix} = \begin{bmatrix} s_0(x) \\ v_0(x) \end{bmatrix} \qquad \text{IC} \tag{101b}$$
- $\mathbf{q}$ can be integrated in space, as in (61), and derive cell averages,
$$\mathbf{Q}_m^n = \int_{x_{m-1/2}}^{x_{m+1/2}} \mathbf{q}(x, t_n)\, \mathrm{d}x \qquad \Rightarrow \begin{bmatrix} S_m^n \\ V_m^n \end{bmatrix} = \int_{x_{m-1/2}}^{x_{m+1/2}} \begin{bmatrix} s(x, t_n) \\ v(x, t_n) \end{bmatrix} \mathrm{d}x \tag{102}$$
- Integration of (101a) in space and time similar to the process from (61) to (67), and using the average fluxes (102), we obtain,
$$S_m^{n+1} = S_m^n - \frac{k}{h}\left(F^n_{s_{m+1/2}} - F^n_{s_{m-1/2}}\right) \tag{103a}$$
$$V_m^{n+1} = V_m^n - \frac{k}{h}\left(F^n_{v_{m+1/2}} - F^n_{v_{m-1/2}}\right) \tag{103b}$$

  This is similar to (65a) and as in (65c) F_s and F_v stand for temporal averages of the spatial fluxes of s and v along the cell vertical lines $x_{m\pm1/2}$.

- Given, that the spatial flux of s is $-Ev$ and the spatial flux of v is $-\frac{1}{\rho}s$, *cf.* (101a) and (100), the spatial flux temporal averages are given as (similar to (65c)),

$$F^n_{s_{m-1/2}} = \frac{1}{k}\int_{t_n}^{t_{n+1}} -Ev(x_{m-1/2},t)\,\mathrm{d}t = -E\hat{V}_{m-1/2}, \qquad \text{where} \qquad \hat{V}_{m-1/2} := \frac{1}{k}\int_{t_n}^{t_{n+1}} v(x_{m-1/2},t)\,\mathrm{d}t \tag{104a}$$

$$F^n_{v_{m-1/2}} = \frac{1}{k}\int_{t_n}^{t_{n+1}} -\frac{1}{\rho}s(x_{m-1/2},t)\,\mathrm{d}t = -\frac{1}{\rho}\hat{S}_{m-1/2}, \qquad \text{where} \qquad \hat{S}_{m-1/2} := \frac{1}{k}\int_{t_n}^{t_{n+1}} s(x_{m-1/2},t)\,\mathrm{d}t \tag{104b}$$

- As in (66) ($F^n_{m-1/2} = \mathcal{F}(Q^n_{m-1}, Q^n_m)$) the temporal average of spatial flux, can be represented by numerical flux functions that depend only on time step n averages, motivated by hyperbolicity of the problem. That is,

$$\hat{V}_{m-1/2} = \mathcal{F}_v(\mathbf{Q}^n_{m-1}, \mathbf{Q}^n_m) = \mathcal{F}_v(S^n_{m-1}, S^n_m, V^n_{m-1}, V^n_m) \tag{105a}$$

$$\hat{S}_{m-1/2} = \mathcal{F}_s(\mathbf{Q}^n_{m-1}, \mathbf{Q}^n_m) = \mathcal{F}_s(S^n_{m-1}, S^n_m, V^n_{m-1}, V^n_m) \tag{105b}$$

- We will discuss difference choices for flux functions later.

- For now, we plug (104) into (103),

$$S^{n+1}_m = S^n_m + \frac{Ek}{h}\left(\hat{V}_{m+1/2} - \hat{V}_{m-1/2}\right) \tag{106a}$$

$$V^{n+1}_m = V^n_m + \frac{k}{h\rho}\left(\hat{S}_{m+1/2} - \hat{S}_{m-1/2}\right) \tag{106b}$$

- Next we discuss different options for (105).

2.2.6.1 1. Average fluxes for $\hat{V}$, $\hat{S}$

- The fluxes (105) can be obtained by actually solving the problem with left (S^n_{m-1}, V^n_{m-1}) and right (S^n_m, V^n_m), which will be covered later, or simply taking the average values which seems to be a reasonable choice.

- We consider average fluxes for S, V in (105). That is,

$$\hat{V}_{m-1/2} = \mathcal{F}^A_v(S^n_{m-1}, S^n_m, V^n_{m-1}, V^n_m) = \frac{V^n_{m-1} + V^n_m}{2} \qquad \text{Average velocity flux} \tag{107a}$$

$$\hat{S}_{m-1/2} = \mathcal{F}^A_s(S^n_{m-1}, S^n_m, V^n_{m-1}, V^n_m) = \frac{S^n_{m-1} + S^n_m}{2} \qquad \text{Average stress flux} \tag{107b}$$

- By using average fluxes (107) in (106) we obtain,

$$S^{n+1}_m = S^n_m + \frac{Ek}{2h}\left(V^n_{m+1} - V^n_{m-1}\right) \qquad \text{Stress update using average fluxes} \tag{108a}$$

$$V^{n+1}_m = V^n_m + \frac{k}{2h\rho}\left(S^n_{m+1} - S^n_{m-1}\right) \qquad \text{Velocity update using average fluxes} \tag{108b}$$

- To better observe, what equation (108) is solving, we express it in FD form (by multiplying it by $\frac{1}{k}$,

$$\frac{S^{n+1}_m - S^n_m}{k} - E\frac{V^n_{m+1} - V^n_{m-1}}{2h} = 0 \tag{109a}$$

$$\frac{V^{n+1}_m - V^n_m}{k} - \frac{1}{\rho}\frac{S^n_{m+1} - S^n_{m-1}}{2h} = 0 \tag{109b}$$

- Clearly, (109) FD equations approximate the equations,

$$s_{,t} + (-E)v_{,x} = 0$$

$$v,t + (-\frac{1}{\rho})s_{,x} = 0$$

  which is the equation we started with in (100).

- So, FV with average fluxes has the same update equations as a FD scheme with two fields of s, v would have had with FTCS stencils.

2.2.6.2 2. Exact fluxes for $\hat{V}$, $\hat{S}$ from Riemann problem solution

- The Riemann solution for elastodynamic problem can be solved with either of the solution processes discussed for linear system of first order PDEs in §1.5.
- In this case from (101a) the flux matrix is equal to

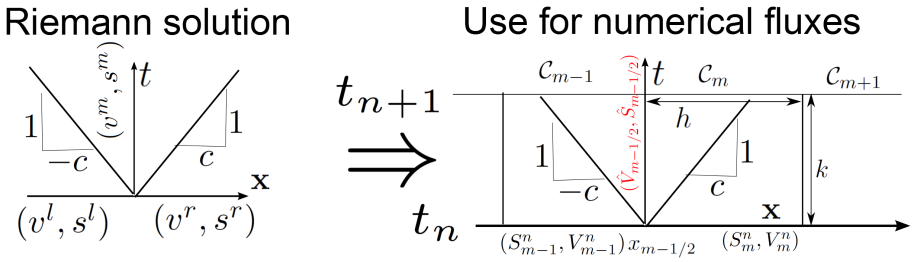
$$\mathbf{A} = \begin{bmatrix} 0 & -E \\ -\frac{1}{\rho} & 0 \end{bmatrix}$$

which similar to acoustic equation can be used in the derivation of characteristic values and Riemann problem solution yielding,

$$s^m = \frac{s^r + s^l}{2} + \frac{Z}{2}(v^r - v^l) \tag{110a}$$

$$v^m = \frac{1}{2Z}(s^r - s^l) + \frac{v^r + v^l}{2} \tag{110b}$$

- $(.)^m$ in (v^m, s^m) refers to the solution in the region between wave speeds $-c$ and $c = \sqrt{E/\rho}$ and (v^l, s^l) and (v^r, s^r) are Riemann problem initial conditions from the two sides.
- $Z = \sqrt{E\rho}$ is the impedance. This solution is given for same E, ρ in the two sides; similar to acoustic equation solutions for $Z^l \neq Z^r$ can be obtained; *cf.* (17).
- Note that Riemann solutions are the same as acoustic equations with just a change of sign for $p \to s$ in the two equations (compare (110) and (14)).



- As shown in the figure the Riemann solutions represent the exact solution for the vertical line $(x_{m-1/2}, t)$ for $t_n \leq t \leq t_{n+1}$.
- Not that the solution is constant along this line (when body force $b = 0$). That is, in (104) $\forall t \in [t_n \; t_{n+1}]$ $v(x_{m-1/2}, t)$ is constant and is equal to $\hat{V}_{m-1/2}$ and same holds for stress average.
- Thus, noting the Riemann solutions (110), the Riemann (exact) fluxes for S, V in (105) are,

$$\hat{V}_{m-1/2} = \mathcal{F}^R_v(S^n_{m-1}, S^n_m, V^n_{m-1}, V^n_m) = \frac{V^n_{m-1} + V^n_m}{2} + \frac{1}{2Z}\left(S^n_m - S^n_{m-1}\right) \qquad \text{Riemann velocity flux} \tag{111a}$$

$$\hat{S}_{m-1/2} = \mathcal{F}^R_s(S^n_{m-1}, S^n_m, V^n_{m-1}, V^n_m) = \frac{S^n_{m-1} + S^n_m}{2} + \frac{Z}{2}\left(V^n_m - V^n_{m-1}\right) \qquad \text{Riemann stress flux} \tag{111b}$$

- We observe that compared to (107), the terms in red are added to flux values.
- Plugging (111) in (106) we obtain,

$$S^{n+1}_m = S^n_m + \frac{k}{h}\left\{\frac{E}{2}\left(V^n_{m+1} - V^n_{m-1}\right) + \frac{E}{2Z}\left(S^n_{m+1} + S^n_{m-1} - 2S^n_m\right)\right\}$$

$$V^{n+1}_m = V^n_m + \frac{k}{h}\left\{\frac{1}{2\rho}\left(S^n_{m+1} - S^n_{m-1}\right) + \frac{Z}{2\rho}\left(V^n_{m+1} + V^n_{m-1} - 2V^n_m\right)\right\}$$

which noting $E/Z = Z/\rho = \sqrt{E/\rho} = c$ we have,

$$S^{n+1}_m = S^n_m + \frac{\bar{k}}{2}\left\{Z\left(V^n_{m+1} - V^n_{m-1}\right) + \left(S^n_{m+1} + S^n_{m-1} - 2S^n_m\right)\right\} \qquad \text{Stress update using Riemann fluxes} \tag{113a}$$

$$V^{n+1}_m = V^n_m + \frac{\bar{k}}{2}\left\{\frac{1}{Z}\left(S^n_{m+1} - S^n_{m-1}\right) + \left(V^n_{m+1} + V^n_{m-1} - 2V^n_m\right)\right\} \qquad \text{Velocity update using Riemann fluxes} \tag{113b}$$

- We observe that compared to update equations with average flux option (108), (113) has the additional terms in red.
- To better understand what equation (113) represents, we write in FD form (by multiplying (112) by $\frac{1}{k}$,

$$\frac{S_m^{n+1}-S_m^n}{k} - E\frac{V_{m+1}^n - V_{m-1}^n}{2h} - \frac{hc}{2}\frac{S_{m+1}^n + S_{m-1}^n - 2S_m^n}{h^2} = 0 \tag{114a}$$

$$\frac{V_m^{n+1}-V_m^n}{k} - \frac{1}{\rho}\frac{S_{m+1}^n - S_{m-1}^n}{2h} - \frac{hc}{2}\frac{V_{m+1}^n + V_{m-1}^n - 2V_m^n}{h^2} = 0 \tag{114b}$$

- The parameter $D_h := \frac{hc}{2}$ is a numerical dissipation factor. It has the dimension $[L^2/T]$.
- (114) FD equations approximate the equations,

$$s_{,t} + (-E)v_{,x} - D_h s_{,xx} = 0 \tag{115a}$$

$$v,t + (-\frac{1}{\rho})s_{,x} - D_h v_{,xx} = 0 \tag{115b}$$

- We observe that compared to (100) the diffusion terms with diffusion coefficient,

$$D_h = \frac{hc}{2} \qquad \text{Numerical diffusion coefficient} \tag{116}$$

  are added to both equations: ($s_{,t} - D_h s_{,xx}$ and $v,t - D_h v_{,xx}$).
- Here are some points about the added diffusion terms:
  - First, compared to average fluxes these fluxed are obtained by solving the exact fluxes on the cell interfaces.
  - The diffusion term tends to zero as grid is refined: $D_h = \frac{hc}{2} \to 0$ as $h \to 0$.
  - When h is large the diffusion terms further stabilize the solution by damping solution oscillations.

2.2.6.3 Discussion on the Riemann solutions

- Riemann solutions for material interfaces or slowly varying material property cases requires Riemann solutions for possibly distinct impedances $Z^l = \sqrt{E^l\rho^l}$ and $Z^r = \sqrt{E^r\rho^r}$.
- Riemann solutions, which can be derived similar to the acoustic equation (17), are

$$s^m = \frac{Z^l s^r + Z^r s^l}{Z^l + Z^r} + \frac{Z^l Z^r}{Z^l + Z^r}(v^r - v^l) \tag{117a}$$

$$v^m = \frac{1}{Z^l + Z^r}(s^r - s^l) + \frac{Z^l v^l + Z^r v^r}{Z^l + Z^r} \tag{117b}$$

- For the discussion on how variable coefficient linear equations are modeled with FV refer to [LeVeque, 2002], chapter 9.
- In fact, the advantages of using Riemann solutions are more pronounced specifically when
  1. Material properties change (*e.g.*, $Z^l \neq Z^r$).
  2. When solution has strong discontinuities.
- In these cases, Riemann solutions tend to much better control solution oscillations and capture the correct reflections and transitions of waves through an interface.

2.2.7 Use of one unknown field rather than two for 2nd order PDEs

- The FV (update) equations of the form (113) require multiple fields.
- This update corresponds to two first order PDEs and ICs in the system below; *cf.* (101),

$$\mathbf{q}_{,t} + (\mathbf{A}\mathbf{q})_{,x} = 0, \qquad \text{where} \qquad \mathbf{q} = \begin{bmatrix} s \\ v \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & -E \\ -\frac{1}{\rho} & 0 \end{bmatrix} \qquad \text{PDE}$$

$$\mathbf{q}(x, t=0) = \mathbf{q}_0(x) \qquad \text{that is} \qquad \begin{bmatrix} s(x,t=0) \\ v(x,t=0) \end{bmatrix} = \begin{bmatrix} s_0(x) \\ v_0(x) \end{bmatrix} \qquad \text{IC}$$

- This requires two unknowns V_m^n and S_m^n at each grid point.
- It may be tempting to reduce to the number of unknowns per point, under certain circumstances.
- Switching between mixed formulations (which has several fields discretized / interpolates such as v and s) versus one primary discretized / interpolated field is a subtle subject in scientific computing. At times special care should be taken in switching from one form to the other.
- For the elastodynamics problem, we can instead consider the one second order PDE in displacement u:

$$u_{,tt} - c^2 u_{,xx} = 0, \qquad \text{where } c = \sqrt{\frac{E}{\rho}} \text{ is the wave speed} \qquad \text{PDE} \tag{118a}$$

$$\begin{cases} u(x, t=0) = u_0(x) \\ \dot{u}(x, t=0) = \dot{u}_0(x) \end{cases} \qquad \text{IC} \tag{118b}$$

- We discuss how to use only one cell average U_m^n rather than two V_m^n and S_m^n in FV update equations (108) for average flux option and (113) for Riemann flux option.
- Since (113a) ((108a) for average flux option) is simply the compatibility equations and the the actual balance of linear momentum is (113b) ((108b) for average flux option), we only need to recast the latter update equation in terms of U_m^n.
- Instead of average flux option, we consider the Riemann flux option (113b) for this demonstration,

$$V_m^{n+1} = V_m^n + \frac{\bar{k}}{2}\left\{\frac{1}{Z}\left(S_{m+1}^n - S_{m-1}^n\right) + \left(V_{m+1}^n + V_{m-1}^n - 2V_m^n\right)\right\} \tag{119}$$

- Now we note that,

$$v = \dot{u}$$
$$s = E\epsilon = E u_{,x}$$

- So, we can replace V_m^n and S_m^n with appropriate FD approximations.
- For example, by using backward difference for V and central difference for S we have,

$$V_m^n \approx \frac{U_m^n - U_m^{n-1}}{k} \tag{120a}$$

$$S_m^n \approx E\frac{U_{m+1}^n - U_{m-1}^n}{2h} \tag{120b}$$

- By changing m and n in (120) and plugging in (119) we obtain,

$$\frac{U_m^{n+1} - U_m^n}{k} = \frac{U_m^n - U_m^{n-1}}{k} + \frac{\bar{k}}{2}\left\{\frac{E}{Z}\left(\frac{U_{m+2}^n - U_m^n}{2h} - \frac{U_m^n - U_{m-2}^n}{2h}\right) + \left(\frac{U_{m+1}^n - U_{m+1}^{n-1}}{k} + \frac{U_{m-1}^n - U_{m-1}^{n-1}}{k} - 2\frac{U_{m+1}^n - U_{m+1}^{n-1}}{k}\right)\right\} \quad \Rightarrow$$

$$\frac{U_m^{n+1} - 2U_m^n + U_m^{n-1}}{k^2} - c^2\frac{U_{m+2}^n - 2U_m^n + U_{m-2}^n}{4h^2} - D_h\frac{1}{k}\left\{\left[\frac{U_{m+1}^n - 2U_m^n + U_{m-1}^n}{h^2}\right] - \left[\frac{U_{m+1}^{n-1} - 2U_m^{n-1} + U_{m-1}^{n-1}}{h^2}\right]\right\} = 0 \tag{121}$$

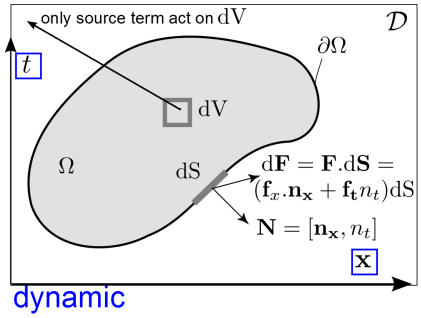
- Remember, *cf.* (116), $D_h = \frac{hc}{2}$ is the numerical coefficient.
- (121) is an explicit FD stencil that provides the value of U_m^{n+1} from 8 previous values: 5 U_{m-2}^n to U_{m+2}^n from time step n and 3 U_{m-1}^{n-1} to U_{m+1}^{n-1} from time step $n-1$.
- It is clear that (121) approximates the solution to,

$$u_{,tt} - c^2 u_{,xx} \; \underbrace{- D_h u_{,txx}}_{} = 0 \tag{122}$$

- Equation (122) is consistent with underlying elastodynamics second order PDE (118a) ($u_{,tt} - c^2 u_{,xx} = 0$) since the added term $-D_h u_{,txx}$ approaches zero as $h \to 0$ ($D_h = \frac{hc}{2}$).
- The added term stems from the use of Riemann solutions. Remember that FV update for V based on the use of Riemann solution approximated (115b): $v,t + (-\frac{1}{\rho})s_{,x} - D_h v_{,xx} = u_{,tt} - \frac{E}{\rho}u_{,xx} - D_h(u_{,t})_{,xx} = 0$ which is the same as (122).
- We can update the FV update equations for using only one field U by using average flux (108) option. In that case, the equation approximates $u_{,tt} - c^2 u_{,xx}$.
- The example of elastodynamics equation demonstrate that FV updates can naturally work on first order equation fields (*e.g.,* s and v) or by some manipulation solve the underlying field from a higher order PDE (*e.g.,* u), although in practice the direct solution of system of equation is the more common approach with FV methods.
- We also observed the flexibility of FV methods in choosing different numerical flux functions. When the exact flux solutions are not available or expensive, approximate numerical flux functions can be used.
- Taking the idea of flux functions and use of finite element methods with discontinuous basis functions, results in discontinuous Galerkin (DG) methods.
- Basically, DG methods can be considered as a higher order (finite element type) extension of FV methods.

2.3 Finite Element Method (FEM)

To derive the FEM formulation we follow the following steps:
Balance law ⇒ Strong form (+ BCs) ⇒ continuum weighted residual method ⇒ Continuum weak form ⇒ Discrete weak form ⇒ FEM method (by using shape functions)



2.3.1 Balance law (for FEM formulations)

- For solid mechanics the temporal flux, spatial flux, and source term are,

$$\mathbf{f}_t = \mathbf{p} = \rho\mathbf{v} \qquad p_i = \rho v_i \qquad \text{linear momentum density} \tag{123a}$$
$$\mathbf{f}_x = -\sigma = -\mathcal{C}\epsilon \qquad \sigma_{ij} = \mathcal{C}_{ijkl}\epsilon_{kl} \qquad \text{stress} \tag{123b}$$
$$\mathbf{r} = \rho\mathbf{b} \qquad \mathbf{b} = b_i\mathbf{e}_i \qquad \text{body force} \tag{123c}$$

  where $\mathcal{C}$ is the fourth order elasticity tensor and $\epsilon = \frac{1}{2}(\nabla\mathbf{u} + \nabla^{\mathrm{T}}\mathbf{u})$ is the strain tensor. Displacement is $\mathbf{u}$.
- Given that spacetime linear momentum density is,

$$\mathbf{F} = [\mathbf{f}_x|\mathbf{f}_t] = [-\sigma\ |\mathbf{p}]$$

- the balance of linear momentum for arbitrary domain Ω in spacetime is,

$$\forall\Omega \subset \mathcal{D} : \int_{\partial\Omega} \mathbf{F}.\mathrm{dS} - \int_{\Omega} \mathbf{r}\,\mathrm{dV} = \int_{\partial\Omega} (\mathbf{f}_x.\mathbf{n_x} + \mathbf{f_t}n_t)\mathrm{dS} - \int_{\Omega} \mathbf{r}\,\mathrm{dV} = \mathbf{0} \tag{124a}$$
$$\forall\Omega \subset \mathcal{D} = \int_{\partial\Omega} (-\sigma.\mathbf{n_x} + \mathbf{p}n_t)\mathrm{dS} - \int_{\Omega} \rho\mathbf{b}\,\mathrm{dV} = \mathbf{0} \tag{124b}$$

- Alternatively, in conventional expression of balance laws which are expressed for arbitrary domains ω in space rather than Ω in spacetime we have,

$$\forall\omega \subset \mathcal{D} \wedge \forall t : \int_{\omega} \mathbf{r}\,\mathrm{dv} - \int_{\partial\omega} \mathbf{f}_x.\mathrm{ds} = \int_{\omega} \mathbf{r}\,\mathrm{dv} - \int_{\partial\omega} (\mathbf{f}_x.\mathbf{n})\,\mathrm{ds} = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\omega} \mathbf{f}_t\,\mathrm{dv} \quad \Rightarrow \tag{125a}$$
$$\forall\omega \subset \mathcal{D} \wedge \forall t : \int_{\omega} \rho\mathbf{b}\,\mathrm{dv} - \int_{\partial\omega} (-\sigma.\mathbf{n})\,\mathrm{ds} = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\omega} \mathbf{p}\,\mathrm{dv} \quad \Rightarrow \tag{125b}$$

2.3.2 Derivation of Strong form from the balance law

- Strong form: In either case, by the application of divergence and localization theorems we obtain the strong form of the balance law,

$$\forall(\mathbf{x},t)\in\mathcal{D}:\nabla_{\text{st}}\mathbf{F}-\mathbf{r}=\left(\dot{\mathbf{f}}_t+\nabla.\mathbf{f}_x\right)-\mathbf{r}=\mathbf{0}\quad\text{Strong Form}\tag{126}$$

which for solid dynamics it reads as ($\mathbf{f}_t=\mathbf{p}$, $\mathbf{f}_x=-\sigma$),

$$\forall(\mathbf{x},t)\in\mathcal{D}:\dot{\mathbf{p}}-\nabla.\sigma-\rho\mathbf{b}=\mathbf{0}\quad\text{Solid mechanics Strong Form}\tag{127}$$

- Damping effects: In solid mechanics similar to fluids energy is dissipated by internal dissipative mechanisms. To motivate this, assume that the solid volume dv the opposing force to its motion is ($-\alpha\mathbf{v}$ dv) this force (in per volume form) is added to the body force

$$\rho\mathbf{b}\quad\rightarrow\quad\rho\mathbf{b}-\alpha\mathbf{v}\tag{128}$$

where α is the damping coefficient.

thus the final form of strong form for solid mechanics becomes,

$$\forall(\mathbf{x},t)\in\mathcal{D}:\begin{cases}\dot{\mathbf{p}}-\nabla.\sigma+\alpha\mathbf{v}-\rho\mathbf{b}=\mathbf{0} & \Rightarrow\\ \rho\ddot{\mathbf{u}}+\alpha\dot{\mathbf{u}}-\nabla.\sigma=\rho\mathbf{b} & \text{that is}\\ \rho\ddot{u}_i+\alpha\dot{u}_i-\sigma_{ij,j}=\rho\ddot{u}_i+\alpha\dot{u}_i-(\mathcal{C}_{ijkl}u_{k,l})_j=\rho b_i\end{cases}\tag{129}$$

Solid mechanics Strong Form with damping

2.3.3 Continuum weighted residual statement (WRS) from strong form

- Continuum weighted residual statement (WRS)
- We define the following,
  - space domain $\mathcal{D}$.
  - time interval $\mathcal{I}^t:=[t_{\min},\ t_{\max}]$ of the solution; IC is enforced at $t=t_{\min}$ and $t_{\max}$ is the terminal time.
  - spacetime domain $\mathcal{D}^t=\mathcal{D}\times\mathcal{I}^t$. Similarly we define,
    - spacetime domain boundary $\partial\mathcal{D}^t:=\partial\mathcal{D}\times\mathcal{I}^t$.
    - spacetime essential BC domain $\partial\mathcal{D}^t_u=\partial\mathcal{D}_u\times\mathcal{I}^t$ ($\partial\mathcal{D}_u$ is spatial essential BC domain).
    - spacetime natural BC domain $\partial\mathcal{D}^t_f=\partial\mathcal{D}_f\times\mathcal{I}^t$ ($\partial\mathcal{D}_f$ is natural essential BC domain).

The set of strong form equation for elastodynamics and boundary conditions are defined as,

$$\text{PDE (balance law: strong form)}\qquad\dot{\mathbf{p}}+\alpha\mathbf{v}-\nabla.\sigma=\rho\mathbf{b}\quad\forall\mathbf{x}\in\mathcal{D}^t\tag{130a}$$

$$\text{Boundary conditions (BCs))}\qquad\begin{cases}\mathbf{u}=\bar{\mathbf{u}} & \forall\mathbf{x}\in\partial\mathcal{D}^t_u\quad\text{Essential BC}\\ \mathbf{t}=\sigma.\mathbf{n}=\bar{\mathbf{t}} & \forall\mathbf{x}\in\partial\mathcal{D}^t_f\quad\text{Natural BC}\end{cases}\tag{130b}$$

- Residuals. Accordingly, we define the residuals,

$$\mathcal{R}_I(\mathbf{x})=-\dot{\mathbf{p}}-\alpha\mathbf{v}+\nabla.\sigma+\rho\mathbf{b}\ @(\mathbf{x})\qquad\forall\mathbf{x}\in\mathcal{D}^t\ \text{(interior residual)}\tag{131a}$$

$$\mathcal{R}_f(\mathbf{x})=\bar{\mathbf{t}}-\mathbf{t}=\bar{\mathbf{t}}-\sigma.\mathbf{n}\ @(\mathbf{x})\qquad\forall\mathbf{x}\in\partial\mathcal{D}^t_f\ \text{(Natural BC residual)}\tag{131b}$$

$$\mathcal{R}_u(\mathbf{x})=\bar{\mathbf{u}}-\mathbf{u}\ @(\mathbf{x})\qquad\forall\mathbf{x}\in\partial\mathcal{D}^t_u\ \text{(Essential BC residual)}\tag{131c}$$

- Weighted residual statement (WRS): For the exact solution all the residuals are zero and wise verse. In the weighted residual method (WRM) we have the option to satisfy (131a) and zero to two of BC residuals weakly by multiplying them with a weight function $\mathbf{w}$ and integrating it over their corresponding domains.

  As it is often done in the WRM, we will strongly (*i.e.*, "essentially") satisfy essential BC $\bar{\mathbf{u}}-\mathbf{u}$ on $\partial\mathcal{D}_u$. But $\mathcal{R}_I$ and $\mathcal{R}_f$ are satisfied weakly. The WRS is,

2.3.3.1 Weighted residual statement

$$\text{Find } \mathbf{u} \in \mathcal{V}^{\text{WRS}} = \{\mathbf{v} \in C^2(\mathcal{D}^t) \mid \forall \mathbf{x} \in \partial\mathcal{D}_u^t \ \mathbf{v}(\mathbf{x}) = \bar{\mathbf{u}}\}, \text{ such that,} \tag{132a}$$

$$\forall w \in \mathcal{W}^{\text{WRS}} = C^0(\mathcal{D}^t), \ \forall t \in \mathcal{I}^t \tag{132b}$$

no need to enforce the homogeneous essential BCs for WRS

$$\begin{aligned} 0 &= \int_{\mathcal{D}} \mathbf{w}.\mathcal{R}_I \, \mathrm{dv} + \int_{\partial\mathcal{D}_f} \mathbf{w}.\mathcal{R}_f \, \mathrm{ds} \\ &= \int_{\mathcal{D}} \mathbf{w}.(-\rho\ddot{\mathbf{u}} - \alpha\dot{\mathbf{u}} + \underbrace{\nabla.\sigma}_{C_{ijkl}\, u_{k,lj}\, \mathbf{e}_i} + \rho\mathbf{b}) \, \mathrm{dv} + \int_{\partial\mathcal{D}_f} \mathbf{w}.(\bar{\mathbf{t}} - \mathbf{t}) \, \mathrm{ds} \end{aligned} \tag{132c}$$

- equivalence of WRS (132) and strong form BVP (130): Clearly, if the strong form is satisfied, so is the WRS. The converse is also true. By the arbitrariness of $\mathbf{w}$ we can show that $\mathcal{R}_I$ and $\mathcal{R}_f$ are satisfies strongly (at every point), and $\mathcal{R}_u$ is already satisfied.

2.3.3.2 Implications on a discrete method

In a discrete method $\mathbf{u}$ is approximated in terms of n_{f} unknowns:

$$\mathbf{u}^h = \sum_{i=1}^{n_{\text{f}}} a_i(t)\phi_i(\mathbf{x}) + \mathbf{u}^{ph}$$

where for spatial dimension d (*e.g.*, $d = 3$ in 3D),

- $\phi_i(\mathbf{x}) = [\phi_i^1(\mathbf{x}) \ \phi_i^2(\mathbf{x}) \ \phi_i^d(\mathbf{x})]$ is the i^{th} spatial test function vector. $\phi_i(\mathbf{x})$ satisfy homogeneous essential BC: $\forall \mathbf{x} \in \partial\mathcal{D}_u^t \phi_i(\mathbf{x}) = 0$.
- $\mathbf{u}^{ph}$ is a particular solution satisfying essential BC: $\forall \mathbf{x} \in \partial\mathcal{D}_u^t \mathbf{u}^{ph}(\mathbf{x}, t) = \bar{\mathbf{u}}(\mathbf{x}, t)$.
- $\mathbf{a}(t) = [a_1(t) \ a_2(t) \ \cdots \ a_{n_{\text{f}}}(t)]$ are system unknown coefficients (they are scalars depending on time t).

We further discuss this topic in §2.3.5.

Since, the number of unknowns are limited (n_{f}) so would be the number of weight functions that we can choose (which is equal to n_{f}. Thus PDE and natural BC cannot (may not) be satisfied strongly (at every point) while Essential BCs are satisfied strongly even in a discrete setting.

- In the WRM there are many choices for the weight functions. The most relevant one to us is the Galerkin method where the weight functions are equal to test functions ($\mathbf{w}_i = \phi_i, i \leq n_{\text{f}}$).
- (Continuous / conventional) Finite Element Method (FEM) is a spatial type of Galerkin method where the test functions are equal to shape functions $\phi_i = N_i$ that satisfy a delta property. That is, $N_i(x_j) = \delta_{ij}$ where x_j are the nodes of the mesh (discrete grid).
- Side note: In more general form, *e.g.*, 4^{th} order beam problem, the shape functions have the delta property on the global degrees of freedom (dofs) rather than nodes.

2.3.4 Continuum weak statement (WK)

- Derivation of Continuum weak statement (WK) from weighted residual statement (WRS)

  The weight function vector $\mathbf{w}$ has zero derivatives and trial (solution) function $\mathbf{u}$ has two derivatives. As before, we want to transfer the derivative from $\mathbf{u}$ to $\mathbf{w}$ by using the divergence theorem. We want to show,

$$\int_{\mathcal{D}} \mathbf{w}.(\nabla.\sigma(\mathbf{u})) \, \mathrm{dv} = -\int_{\mathcal{D}} \epsilon(w) : \sigma(\mathbf{u}) \, \mathrm{dv} + \int_{\partial\mathcal{D}} \mathbf{w}.\mathbf{t}(\mathbf{u}) \, \mathrm{ds}$$

  where $\mathbf{t}$ is the traction vector and $\epsilon(w) = \frac{1}{2}\left(\nabla\mathbf{w} + (\nabla\mathbf{w})^{\text{T}}\right)$ is the weight strain field.

  We observe,

$$\mathbf{w}.(\nabla.\sigma) = w_i\sigma_{ij,j} = \frac{\partial w_i\sigma_{ij}}{\partial x_j} - w_{i,j}\sigma_{ij} = \frac{\partial w_i\sigma_{ij}}{\partial x_j} - \frac{w_{i,j} + w_{j,i}}{2}\sigma_{ij} - \underbrace{\frac{w_{i,j} - w_{j,i}}{2}}_{=z_{ij}}\sigma_{ij} \tag{133}$$

We note that $z_{ji} = \frac{w_{j,i}-w_{i,j}}{2} = -\frac{w_{i,j}-w_{j,i}}{2} = -z_{ij}$ and $\sigma_{ji} = \sigma_{ij}$ (symmetry of the stress tensor). Then we employ the asymmetry and symmetry of $\mathbf{z}$ and σ to show that:

$$\underbrace{z_{ji}w_{ji} = z_{ji}w_{ji}}_{\text{interchange of dummy indices}} = -(z_{ij})w_{ij} \;\Rightarrow\; 2z_{ij}w_{ij} = 0 \;\Rightarrow\; z_{ij}w_{ij} = 0 \tag{134}$$

also we note that,

$$\left(\frac{w_{i,j} + w_{j,i}}{2}\right)\sigma ij = \epsilon(\mathbf{w}) : \sigma(\mathbf{u}) \tag{135}$$

According to (133), (134), and (135) we have:

$$\begin{aligned}
\mathbf{w}.(\nabla.\sigma(\mathbf{u})) &= \frac{\partial w_i \sigma_{ij}}{\partial x_j} - \epsilon(\mathbf{w}) : \sigma \quad \Rightarrow \\
\int_{\mathcal{D}} \mathbf{w}.(\nabla.\sigma(\mathbf{u})) &= \int_{\mathcal{D}} \left(\frac{\partial w_i \sigma_{ij}}{\partial x_j}\right) \,\mathrm{dv} - \int_{\mathcal{D}} \epsilon(\mathbf{w}) : \sigma(\mathbf{u}) \,\mathrm{dv} \quad \Rightarrow \text{(Divergence theorem)} \\
\int_{\mathcal{D}} \mathbf{w}.(\nabla.\sigma(\mathbf{u})) &= \int_{\partial\mathcal{D}} w_i \underbrace{(\sigma_{ij} n_j)}_{t_i(\mathbf{u})} \,\mathrm{ds} - \int_{\mathcal{D}} \epsilon(\mathbf{w}) : \sigma \,\mathrm{dv} \\
&= \int_{\partial\mathcal{D}} \mathbf{w}.\mathbf{t}(\mathbf{u}) \,\mathrm{ds} - \int_{\mathcal{D}} \epsilon(\mathbf{w}) : \sigma(\mathbf{u}) \,\mathrm{dv}
\end{aligned}$$

which completes the proof

$$\int_{\mathcal{D}} \mathbf{w}.(\nabla.\sigma(\mathbf{u})) \,\mathrm{dv} = -\int_{\mathcal{D}} \epsilon(\mathbf{w}) : \sigma(\mathbf{u}) \,\mathrm{dv} + \int_{\partial\mathcal{D}} \mathbf{w}.\mathbf{t}(\mathbf{u}) \,\mathrm{ds} \tag{136}$$

We plug (136) in (132c) to obtain:

$$\begin{aligned}
0 &= \int_{\mathcal{D}} \mathbf{w}.\left(-\rho\ddot{\mathbf{u}} - \alpha\dot{\mathbf{u}} + \nabla.\sigma(\mathbf{u}) + \rho\mathbf{b}\right) \,\mathrm{dv} + \int_{\partial\mathcal{D}_f} \mathbf{w}.(\bar{\mathbf{t}} - \mathbf{t}) \,\mathrm{ds} \quad \Rightarrow \\
0 &= \left\{-\int_{\mathcal{D}} \epsilon(\mathbf{w}) : \sigma(\mathbf{u}) \,\mathrm{dv} + \int_{\partial\mathcal{D}} \mathbf{w}.\mathbf{t}(\mathbf{u}) \,\mathrm{ds}\right\} \\
&\quad - \int_{\mathcal{D}} \mathbf{w}.\left(\rho\ddot{\mathbf{u}} + \alpha\dot{\mathbf{u}} - \rho\mathbf{b}\right) \,\mathrm{dv} + \int_{\partial\mathcal{D}_f} \mathbf{w}.(\bar{\mathbf{t}} - \mathbf{t}) \,\mathrm{ds} \;\Rightarrow \\
0 &= -\int_{\mathcal{D}} \epsilon(\mathbf{w}) : \sigma(\mathbf{u}) \,\mathrm{dv} - \int_{\mathcal{D}} \mathbf{w}.\left(\rho\ddot{\mathbf{u}} + \alpha\dot{\mathbf{u}}\right) \,\mathrm{dv} + \int_{\mathcal{D}} \mathbf{w}.\rho\mathbf{b} \,\mathrm{dv} \\
&\quad + \int_{\partial\mathcal{D}_f} \mathbf{w}.\mathbf{t}(\mathbf{u}) \,\mathrm{ds} + \int_{\partial\mathcal{D}_f} \mathbf{w}.(\bar{\mathbf{t}} - \mathbf{t}(\mathbf{u})) \,\mathrm{ds} \\
&\quad + \int_{\partial\mathcal{D}_u} \mathbf{w}.\mathbf{t}(\mathbf{u}) \,\mathrm{ds}
\end{aligned}$$

We used the decomposition $\partial\mathcal{D} = \partial\mathcal{D}_f \cup \partial\mathcal{D}_u,\ \partial\mathcal{D}_f \cap \partial\mathcal{D}_u = \emptyset$. Next We get,

$$\begin{aligned}
\int_{\mathcal{D}} \left[\mathbf{w}.\left(\rho\ddot{\mathbf{u}} + \alpha\dot{\mathbf{u}}\right) + \epsilon(\mathbf{w}) : \sigma(\mathbf{u})\right] \,\mathrm{dv} &= \int_{\mathcal{D}} \mathbf{w}.\rho\mathbf{b} \,\mathrm{dv} + \int_{\partial\mathcal{D}_f} \mathbf{w}.\bar{\mathbf{t}} \,\mathrm{ds} \\
&\quad + \int_{\partial\mathcal{D}_u} \mathbf{w}.\mathbf{t}(\mathbf{u}) \,\mathrm{ds}
\end{aligned} \tag{137}$$

As customary in FE formulations, we eliminate the last integration term on $\partial\mathcal{D}_u$ by enforcing homogeneous boundary conditions for $\mathbf{w}$ on $\partial\mathcal{D}_u$.

2.3.4.1 Continuum weak statement (WK)

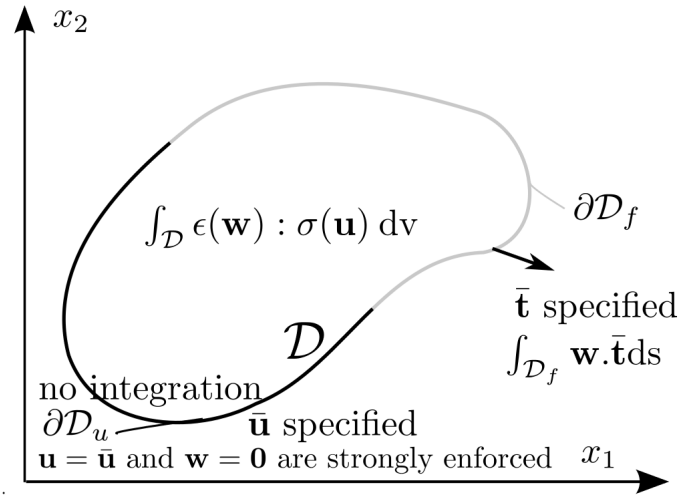
The weak statement for elastodynamics and the boundary conditions are:

$$\text{Find } \mathbf{u} \in \mathcal{V} = \{v \in C^1(\mathcal{D}^t) \mid \forall \mathbf{x} \in \partial\mathcal{D}_u^t \ \mathbf{v}(\mathbf{x}) = \bar{\mathbf{u}}(\mathbf{x})\}, \text{ such that,} \tag{138a}$$

$$\forall \mathbf{w} \in \mathcal{W} = \{v \in C^1(\mathcal{D}^t) \mid \forall \mathbf{x} \in \partial\mathcal{D}_u^t \ \mathbf{v}(\mathbf{x}) = 0\}, \ \forall t \in \mathcal{I}^t \tag{138b}$$

$$\int_{\mathcal{D}} \left[\rho\mathbf{w}.\ddot{\mathbf{u}} + \alpha\mathbf{w}.\dot{\mathbf{u}} + \epsilon(\mathbf{w}) : \sigma(\mathbf{u})\right] \,\mathrm{dv} = \int_{\mathcal{D}} \mathbf{w}.\rho\mathbf{b} \,\mathrm{dv} + \int_{\partial\mathcal{D}_f} \mathbf{w}.\bar{\mathbf{t}} \,\mathrm{ds} \tag{138c}$$

- Both $\mathcal{V}$ and $\mathcal{W}$ have the same regularity ($C^m(\mathcal{D})$): $m = M/2$, $M = 2$ is the order of the differential equation.
- The less demanding regularity conditions for the solution compared to the weighted residual statement ($C^M(\mathcal{D}) \rightarrow C^m(\mathcal{D})$) takes us to the same function space needed for the balance law (highest derivative is for $\sigma(\mathbf{u}) = C_{ijkl}\, u_{k,l}$ is 1).
- Both $\mathcal{V}$ and $\mathcal{W}$ exactly enforce the essential boundary conditions, with the difference that $\mathcal{W}$ satisfies the homogeneous version.



2.3.5 Discrete solution & weight function space

- Discretization of weak form: We approximate continuum solution $\mathbf{u}$ by discrete solution $\mathbf{u}^h$ in terms of n_{f} unknowns,

$$\boxed{\mathbf{u}^h(\mathbf{x},t) = \sum_{i=1}^{n_{\mathrm{f}}} a_i^f(t)\phi_i(\mathbf{x}) + \mathbf{u}^{ph}(\mathbf{x},t)} \tag{139}$$

  This is a semi-discrete expansion of $\mathbf{u}^h$ which is appropriate for time marching schemes and exact integration in time approaches.

  The terms in (139) are,

  - n_{f}: number of free degrees of freedom (dof).
  - $a_i^f(t)$: Unknown coefficients that are scalar and function of time t.
  - $\phi_i(\mathbf{x})$: trial functions which are vectors $\phi_i(\mathbf{x}) = [\phi_i^1(\mathbf{x}),\ \phi_i^2(\mathbf{x}), \phi_i^3(\mathbf{x})]^{\mathrm{T}}$ (for $d = 3$) which are functions of space $\mathbf{x}$.
    - $\phi_i(\mathbf{x})$ satisfy homogeneous essential BC: $\forall \mathbf{x} \in \partial\mathcal{D}_u^t :\ \phi_i(\mathbf{x}) = 0$. $\triangle$
    - ϕ_i is a vector because the unknown of the problem, displacement $\mathbf{u}$, is a vector. They will be a scalar for heat equation for temperature T, or any other order tensor depending on the tensor order of unknown of the problem.
  - $\mathbf{u}^{ph}(\mathbf{x},t)$ is a particular solution satisfying essential BC: $\forall \mathbf{x} \in \partial\mathcal{D}_u^t :\ \mathbf{u}^{ph}(\mathbf{x},t) = \bar{\mathbf{u}}(\mathbf{x},t)$. $\diamondsuit$
  - From $\triangle$ and $\diamondsuit$ $\forall \mathbf{x} \in \partial\mathcal{D}_u^t :\ \mathbf{u}^h(\mathbf{x},t) = \bar{\mathbf{u}}(\mathbf{x},t)$ that is $\mathbf{u}^h$ strongly satisfies the essential BC as we wanted.

- In a fully discrete spacetime scheme, we would have interpolated $\mathbf{u}^h$ as, $\mathbf{u}^h = \sum_{i=1}^{n_{\mathrm{f}}} a_i^f \phi_i(\mathbf{x},t) + \mathbf{u}^{ph}(\mathbf{x},t)$ that is the trial functions will be functions of space ($\mathbf{x}$) as well as time t. In that case the weak statement (138) should be expressed for $\mathcal{D}^t$ rather than $\mathcal{D}$ and being enforced for $t \in \mathcal{I}^t$.

  For the moment, we focus on discretization (139).

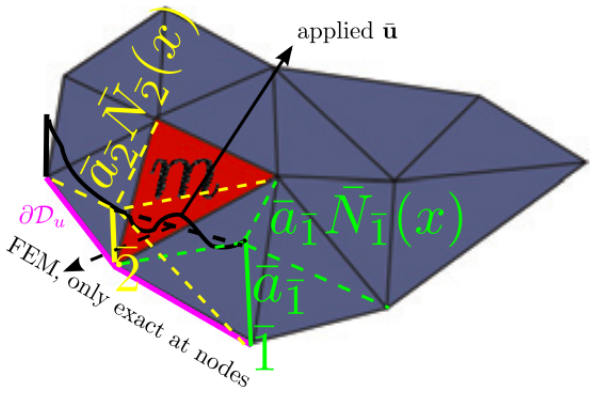
FEM expression of trial function and shape functions.

- Shape functions: In FEM trial functions are called shape functions,

$$\begin{aligned} &\mathbf{N}_i^f(\mathbf{x}) = \phi_i(\mathbf{x}) \quad \text{that is} \\ &\mathbf{N}_i^f(\mathbf{x}) = [N_i^{f^1}(\mathbf{x}),\ N_i^{f^2}(\mathbf{x}),\ N_i^{f^3}(\mathbf{x})]^{\mathrm{T}} = [\phi_i^1(\mathbf{x}),\ \phi_i^2(\mathbf{x}),\ \phi_i^3(\mathbf{x})]. \end{aligned} \tag{140}$$

- Note that shape functions are vectors and $N_i^{f^1}(\mathbf{x})$ is component (direction) j of shape function $\mathbf{N}_i^f(\mathbf{x})$.
- dof: Dofs in FEM correspond to displacement components of nodes of the grid (and their derivatives for shells, plates and other higher order elements).

- FEM delta property: shape function $\mathbf{N}_i(\mathbf{x})$ takes the value of 1 at dof i and zero elsewhere.



- Particular solution $\mathbf{u}^{ph}(\mathbf{x},t)$ is used to construct $\bar{\mathbf{u}}$ on $\partial\mathcal{D}_u^t$ for $\mathbf{u}^h$,

$$n_{\rm p} = \text{ number of prescribed dofs} \tag{141a}$$

$$\mathbf{a}^p(t) = \left[a_1^p(t),\ldots,a_{n_{\rm p}}^p(t)\right]^{\rm T} = \text{vector of prescribed values for prescribed dofs} \tag{141b}$$

$$\mathbf{N}^p(\mathbf{x}) = \left[\mathbf{N}_1^p(\mathbf{x}),\ldots,\mathbf{N}_{n_{\rm p}}^p(\mathbf{x})\right] = \text{(row) vector of shape functions for prescribed dofs} \tag{141c}$$

$$\mathbf{u}^{ph}(\mathbf{x},t) = \mathbf{N}^p(\mathbf{x})\mathbf{a}^p(t) = \sum_{i=1}^{n_{\rm p}} a_i^p(t)\mathbf{N}_i^p(\mathbf{x}) \tag{141d}$$

Size of these arrays are,

$\mathbf{a}^p(t)$	$1\times n_{\rm p}$
$\mathbf{N}_i^p(\mathbf{x}) = [\bar{N}_i^{p^1}(\mathbf{x}),\ \bar{N}_i^{p^2}(\mathbf{x}),\bar{N}_i^{p^3}(\mathbf{x})]^{\rm T}$	3×1 vector prescribed shape function i
$\mathbf{N}^p(\mathbf{x})$	$3\times n_{\rm p}$
$\mathbf{u}^{ph}(\mathbf{x},t)$	3×1 particular displacement vector

The construction of particular solution for a scalar problem, *e.g.*, T in thermal equation, is shown in the figure.

Combining free and prescribed dofs,

- Discrete solution function can be written as (*cf.* (139), (140), (141)),

$$\mathbf{u}^h(\mathbf{x},t) = \mathbf{u}^{fh} + \mathbf{u}^{ph}(\mathbf{x},t) \tag{142a}$$

$$= \sum_{i=1}^{n_{\rm f}} a_i^f(t)\mathbf{N}_i^f(\mathbf{x}) + \sum_{i=1}^{n_{\rm p}} a_i^p(t)\mathbf{N}_i^p(\mathbf{x}) \tag{142b}$$

$$= \mathbf{N}.\mathbf{a} \tag{142c}$$

where

$$\mathbf{N} = \begin{bmatrix}\mathbf{N}^f & \mathbf{N}^p\end{bmatrix} = \begin{bmatrix}\mathbf{N}_1^f & \cdots & \mathbf{N}_{n_{\rm f}}^f & |\mathbf{N}_1^p & \cdots & \mathbf{N}_{n_{\rm p}}^p\end{bmatrix} \qquad 3\times n \text{ array} \tag{143a}$$

$$\mathbf{a} = \begin{bmatrix}\mathbf{a}^f\\ \mathbf{a}^p\end{bmatrix} = \begin{bmatrix} a_1^f\\ \vdots\\ a_1^f\\ --\\ a_1^p\\ \vdots\\ a_{n_{\rm p}}^p\end{bmatrix} \qquad n\times 1 \text{ array} \tag{143b}$$

$$\mathbf{u}^h(\mathbf{x},t) \qquad 3\times 1 \text{ array} \tag{143c}$$

$$n = n_{\mathrm{f}} + n_{\mathrm{p}} \quad \text{total number of dofs} \tag{143d}$$

Notes:

- Unknown quantities, $\mathbf{a}^f$ are shown in this color. There are n_{f} unknowns in $\mathbf{a}$.
- The coefficient vector $\mathbf{a}^p$ is known because the prescribed coefficients are given ("prescribed").
- $\mathbf{u}^{fh}$ satisfies the homogeneous essential BC.
- $\mathbf{u}^{ph}$ satisfies the essential BC.
- So $\mathbf{u}^h(\mathbf{x},t)$ satisfies essential BC. strongly as we required.

Weight functions:

- We need two condition on weighted function set:
  - The number of weight functions should be equal to number of unknowns n_{f}.
  - Weight functions must satisfy essential boundary conditions.
  - If Galerkin method is used weight functions are equal to solution interpolant functions $\mathbf{w}_i = \phi_i$.
- The choice,

$$\mathbf{w}_i = a_i^f \quad i \le n_{\mathrm{f}}$$

satisfies the first two conditions and based on item 3 correspond to a Galerkin method.

2.3.6 Voigt stress and strain notation / displacement to strain operator (tensor)

- Second order strain tensor is,

$$\epsilon(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla^{\mathrm{T}}\mathbf{u}\right) \quad \text{that is}$$

$$\epsilon(\mathbf{u}) = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} = \begin{bmatrix} u_{1,1} & \frac{1}{2}(u_{1,2}+u_{2,1}) & \frac{1}{2}(u_{1,3}+u_{3,1}) \\ \frac{1}{2}(u_{1,2}+u_{2,1}) & u_{2,2} & \frac{1}{2}(u_{2,3}+u_{3,2}) \\ \frac{1}{2}(u_{1,3}+u_{3,1}) & \frac{1}{2}(u_{2,3}+u_{3,2}) & \frac{1}{2}u_{3,3} \end{bmatrix}$$

- Voigt-notation strain vector is,

$$\gamma(\mathbf{u}) = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{33} \\ 2\epsilon_{12} \\ 2\epsilon_{23} \\ 2\epsilon_{31} \end{bmatrix} = \begin{bmatrix} u_{1,1} \\ u_{2,2} \\ u_{3,3} \\ u_{1,2}+u_{2,1} \\ u_{2,3}+u_{3,2} \\ u_{3,1}+u_{1,3} \end{bmatrix} = L_m\mathbf{u} \quad \text{where} \tag{144a}$$

$$L_m = \begin{bmatrix} \frac{\partial}{\partial x_1} & 0 & 0 \\ 0 & \frac{\partial}{\partial x_2} & 0 \\ 0 & 0 & \frac{\partial}{\partial x_3} \\ \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_1} & 0 \\ 0 & \frac{\partial}{\partial x_3} & \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} & 0 & \frac{\partial}{\partial x_1} \end{bmatrix} \quad \textbf{displacement to strain differential operator} \tag{144b}$$

- Note the dimensions are,
  - Displacement vector $\mathbf{u}$: 3×1.
  - strain tensor ϵ: 3×3.
  - Voigt strain vector γ: 6×1.
  - Displacement to strain differential operator L_m: 6×3.

- Computation of strain in FEM,

$$\left.\begin{array}{rcl}(\mathbf{u}(\mathbf{x},t))_{3\times 1} &=& (\mathbf{N}(\mathbf{x}))_{3\times n}.(\mathbf{a}(t))_{n\times 1}\\ (\gamma(\mathbf{x},t))_{6\times 1} &=& (L_m)_{6\times 3}(\mathbf{u}(\mathbf{x},t))_{3\times 1} = L_m\mathbf{N}(\mathbf{x}).(\mathbf{a}(t))\end{array}\right\}\Rightarrow$$

$$\gamma(\mathbf{x},t) = \mathbf{B}(\mathbf{x})\mathbf{a}(t), \quad \text{where} \tag{145a}$$

$$(\mathbf{B}(\mathbf{x}))_{6\times n} = (L_m)_{6\times 3}(\mathbf{N}(\mathbf{x}))_{3\times n} \quad \text{displacement to strain array} \tag{145b}$$

- Second order stress tensor is,

$$\sigma = \begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{bmatrix}$$

- Voigt-notation stress vector is,

$$\mathbf{s} = \begin{bmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{12}\\ \sigma_{23}\\ \sigma_{31}\end{bmatrix} \tag{146a}$$

Constitutive equation:

- Constitutive equation for stress tensor (Voigt vector) can be obtained for complex models such as (nonlinear) hyperelasticity, hypoelasticity, plasticity, viscoplasticity, *etc.*.
- We focus on linear solid material where we have,

$$\sigma_{3\times 3} = \mathcal{C}_{3\times 3\times 3\times 3}\epsilon_{3\times 3} \quad \Rightarrow \quad \mathbf{s}_{6\times 1} = \bar{\bar{\mathbf{C}}}_{6\times 6}\gamma_{6\times 1} \tag{147}$$

where

$$\mathcal{C}_{3\times 3\times 3\times 3} = \mathcal{C}_{ijkl}\mathbf{e}_i\times\mathbf{e}_j\times\mathbf{e}_k\times\mathbf{e}_l \quad \text{fourth order elasticity tensor} \tag{148a}$$

$$\mathcal{C}_{ijkl} = \mathcal{C}_{klij} = \mathcal{C}_{jikl} = \mathcal{C}_{ijlk} \tag{148b}$$

$$\bar{\bar{\mathbf{C}}}_{6\times 6} = \bar{\bar{C}}_{ij}\mathbf{e}_i\times\mathbf{e}_j(1\le i,j\le 6) \quad \text{second order Voigt elasticity array} \tag{148c}$$

$$\bar{\bar{C}}_{ij} = \bar{\bar{C}}_{ji} \tag{148d}$$

Stress stress term in weak form

- Continuum: The continuum weak statement (138c) has a term, $\epsilon(\mathbf{w}):\sigma(\mathbf{u})$ which is virtual internal work from virtual displacement (weight function) $\mathbf{w}$ on solution stress $\sigma(\mathbf{u})$. For linear solid this term can be written as,

$$\epsilon(\mathbf{w}):\sigma(\mathbf{u}) = \epsilon(\mathbf{w}):\mathcal{C}\epsilon(\mathbf{u}) = \epsilon_{ij}(\mathbf{w})\mathcal{C}_{ijkl}\epsilon_{kl}(\mathbf{u}) \quad \text{or alternatively} \tag{149a}$$

$$\begin{aligned}\epsilon(\mathbf{w}):\sigma(\mathbf{u}) &= \gamma(\mathbf{w}).\mathbf{s}(\mathbf{u}) = \gamma(\mathbf{w})^{\mathrm{T}}\bar{\bar{\mathbf{C}}}\gamma(\mathbf{u})\\ &= \gamma_i(\mathbf{w})C_{ij}\gamma_j(\mathbf{u}) \quad (1\le i,j\le 6) \quad \text{Voigt notation}\end{aligned} \tag{149b}$$

The factor of two for shear strains in (144a) are necessary for the following two properties we used,

  - Symmetry of 6×6 Voigt elasticity $\bar{\bar{\mathbf{C}}}$ ($\bar{\bar{C}}_{ij} = \bar{\bar{C}}_{ji}$) in (147), (148d).
  - Ability to express $\epsilon(\mathbf{w}):\sigma(\mathbf{u})$ as $\epsilon(\mathbf{w}):\mathcal{C}\epsilon(\mathbf{u})$ in (149).

- Expression of strain from displacement in Discrete setting:
  - In discrete version as we will see $\mathbf{w}_{3\times 1} = \mathbf{N}^f_{3\times n_f}\hat{\mathbf{a}}$ where $\hat{\mathbf{a}}_{n_f\times 1}$ is an arbitrary vector that enables spanning the n_f dimensional space of the weight functions. In addition, $\mathbf{u}^h = \mathbf{N}_{3\times n}.\mathbf{a}_{n\times 1}$.
  - From (149b), the above line, noting that $\mathbf{a}_{n\times 1}, \mathbf{N}_{3\times n}$ go with solution $\mathbf{u}^h$ and $\hat{\mathbf{a}}_{n_f\times 1}, \mathbf{N}^f_{3\times n_f}$ go with the weight $\mathbf{w}$, and $\gamma = \mathbf{B}\mathbf{a}$ we have,

$$\epsilon(\mathbf{w}):\sigma(\mathbf{u}^h) = \gamma(\mathbf{w})^{\mathrm{T}}\bar{\bar{\mathbf{C}}}\gamma(\mathbf{u}^h) \qquad = (\mathbf{B}^p\hat{\mathbf{a}})^{\mathrm{T}}\bar{\bar{\mathbf{C}}}\mathbf{B}\mathbf{a} = \hat{\mathbf{a}}^{\mathrm{T}}\left(\mathbf{B}^{f\,\mathrm{T}}\bar{\bar{\mathbf{C}}}\mathbf{B}\right)\mathbf{a} \tag{150}$$

2.3.7 Discretization of weak form

- Trial solution and weight functions: Since $\mathbf{u}^{fh} = \sum_{i=1}^{n_\mathrm{f}} a_i^f(t)\mathbf{N}_i(\mathbf{x})$ satisfies the homogeneous essential BC and we use a Galerkin method $\mathbf{w}_i(\mathbf{x}) = \mathbf{N}_i(\mathbf{x})$. Clearly weight functions as required satisfy homogeneous essential BC. In addition, since $\mathbf{u}^{ph}(\mathbf{x},t)$ satisfies essential BC so does $\mathbf{u}^h(\mathbf{x},t)$ which again is required for the trial solution field. From this and (142), (143) we can write,

$$\begin{aligned} &\mathbf{u}^h(\mathbf{x},t) = \mathbf{N}(\mathbf{x}).\mathbf{a}(t) && \Rightarrow && \\ &\gamma(\mathbf{u}(\mathbf{x},t)) = \mathbf{B}(\mathbf{x}).\mathbf{a}(t) && \dot{\mathbf{u}}^h(\mathbf{x},t) = \mathbf{N}(\mathbf{x}).\dot{\mathbf{a}}(t) && \ddot{\mathbf{u}}^h(\mathbf{x},t) = \mathbf{N}(\mathbf{x}).\ddot{\mathbf{a}}(t) \qquad (151a) \\ &\mathbf{w}(\mathbf{x},t) = \mathbf{N}^f(\mathbf{x}).\hat{\mathbf{a}}(t) && \Rightarrow && \\ &\gamma(\mathbf{w}(\mathbf{x},t)) = \mathbf{B}^f(\mathbf{x}).\hat{\mathbf{a}}(t) && && \qquad (151b) \end{aligned}$$

- Discrete weak form: By plugging (151) into (138c), that is,

$$\int_{\mathcal{D}} \left[\rho \mathbf{w}.\ddot{\mathbf{u}} + \alpha \mathbf{w}.\dot{\mathbf{u}} + \epsilon(\mathbf{w}) : \sigma(\mathbf{u})\right] \,\mathrm{dv} = \int_{\mathcal{D}} \mathbf{w}.\rho\mathbf{b} \,\mathrm{dv} + \int_{\partial\mathcal{D}_f} \mathbf{w}.\bar{\mathbf{t}} \,\mathrm{ds}$$

and noting the expression for $\epsilon(\mathbf{w}) : \sigma(\mathbf{u})$ from (150) we obtain,

$$\forall \hat{\mathbf{a}}: \ \hat{\mathbf{a}}^{\mathrm{T}} \left\{ \underbrace{\left[\int_{\mathcal{D}} \rho \mathbf{N}^{f\,\mathrm{T}}\mathbf{N} \,\mathrm{dv}\right]}_{\mathbf{M}} \ddot{\mathbf{a}} + \underbrace{\left[\int_{\mathcal{D}} \alpha \mathbf{N}^{f\,\mathrm{T}}\mathbf{N} \,\mathrm{dv}\right]}_{\mathbf{C}} \dot{\mathbf{a}} + \underbrace{\left[\int_{\mathcal{D}} \mathbf{B}^{f\,\mathrm{T}}\bar{\bar{\mathbf{C}}}\mathbf{B} \,\mathrm{dv}\right]}_{\mathbf{K}} \mathbf{a} - \underbrace{\left[\int_{\mathcal{D}} \mathbf{N}^{f\,\mathrm{T}}\rho\mathbf{b} \,\mathrm{dv}\right]}_{\mathbf{F}_r} - \underbrace{\left[\int_{\partial\mathcal{D}_f} \mathbf{N}^{f\,\mathrm{T}}\bar{\mathbf{t}} \,\mathrm{ds}\right]}_{\mathbf{F}_N} \right\} = 0$$

Since $\hat{\mathbf{a}}$ is arbitrary (to span the n_f dimensional space of the weight functions), we conclude that,

$$\mathbf{M}\ddot{\mathbf{a}} + \mathbf{C}\dot{\mathbf{a}} + \mathbf{K}\mathbf{a} = \mathbf{F}_r + \mathbf{F}_N \qquad (152)$$

where,

$$\begin{aligned} \mathbf{M} &= \int_{\mathcal{D}} \rho \mathbf{N}^{f\,\mathrm{T}}\mathbf{N} \,\mathrm{dv} && \text{Mass matrix} \\ \mathbf{C} &= \int_{\mathcal{D}} \alpha \mathbf{N}^{f\,\mathrm{T}}\mathbf{N} \,\mathrm{dv} && \text{Damping matrix} \\ \mathbf{K} &= \int_{\mathcal{D}} \mathbf{B}^{f\,\mathrm{T}}\bar{\bar{\mathbf{C}}}\mathbf{B} \,\mathrm{dv} && \text{Stiffness matrix} \\ \mathbf{F}_r &= \int_{\mathcal{D}} \mathbf{N}^{f\,\mathrm{T}}\rho\mathbf{b} \,\mathrm{dv} && \text{Source term (body force) force vector} \\ \mathbf{F}_N &= \int_{\partial\mathcal{D}_f} \mathbf{N}^{f\,\mathrm{T}}\bar{\mathbf{t}} \,\mathrm{ds} && \text{Natural BC force vector} \end{aligned}$$

Since all matrices $\mathbf{M}, \mathbf{C}, \mathbf{K}$ have dimensions $n_\mathrm{f} \times n$ ($n = n_\mathrm{f} + n_\mathrm{p}$) and $\mathbf{a}$ has dimension of n rather than the dimension of unknowns n_f we separate the matrices and $\mathbf{a}$ to the part that corresponding to $\mathbf{a}^f$ and $\mathbf{a}^p$. For example,

$$\begin{aligned} \mathbf{M}\ddot{\mathbf{a}} &= \left\{\int_{\mathcal{D}} \rho \mathbf{N}^{f\,\mathrm{T}}\mathbf{N} \,\mathrm{dv}\right\} \ddot{\mathbf{a}} = \left\{\int_{\mathcal{D}} \rho \mathbf{N}^{f\,\mathrm{T}} \begin{bmatrix}\mathbf{N}^f & \mathbf{N}^p\end{bmatrix} \,\mathrm{dv}\right\} \begin{bmatrix}\ddot{\mathbf{a}}^f \\ \ddot{\mathbf{a}}^p\end{bmatrix} \\ &= \underbrace{\left\{\int_{\mathcal{D}} \rho \mathbf{N}^{f\,\mathrm{T}}\mathbf{N}^f \,\mathrm{dv}\right\}}_{\mathbf{M}^{ff}} \ddot{\mathbf{a}}^f + \underbrace{\left\{\int_{\mathcal{D}} \rho \mathbf{N}^{f\,\mathrm{T}}\mathbf{N}^p \,\mathrm{dv}\right\}}_{\mathbf{M}^{fp}} \ddot{\mathbf{a}}^p \end{aligned}$$

Similar process can be applied to $\mathbf{K}$ and $\mathbf{C}$ terms. Finally, the system can be summarized as follows,

$$\mathbf{M}^{ff}\ddot{\mathbf{a}}^f + \mathbf{C}^{ff}\dot{\mathbf{a}}^f + \mathbf{K}^{ff}\mathbf{a}^f = \mathbf{F}_r + \mathbf{F}_N - \mathbf{F}_D \quad \text{ODE where} \tag{153a}$$

$$\mathbf{a}^f(t=0) = \mathbf{a}_0^f, \dot{\mathbf{a}}^f(t=0) = \dot{\mathbf{a}}_0^f \quad \text{Initial condition (IC)} \tag{153b}$$

$$\mathbf{F}_r = \int_{\mathcal{D}} \mathbf{N}^{f\,\mathrm{T}} \rho \mathbf{b}\, \mathrm{dv} \quad \text{Source term (body force) force vector} \tag{153c}$$

$$\mathbf{F}_N = \int_{\partial\mathcal{D}_f} \mathbf{N}^{f\,\mathrm{T}} \bar{\mathbf{t}}\, \mathrm{ds} \quad \text{Natural (Neumann) BC force vector} \tag{153d}$$

$$\mathbf{F}_D = \mathbf{M}^{fp}\ddot{\mathbf{a}}^p + \mathbf{C}^{fp}\dot{\mathbf{a}}^p + \mathbf{K}^{fp}\mathbf{a}^p \quad \text{Essential (Dirichlet) BC force vector} \tag{153e}$$

$$\begin{aligned}
\mathbf{M}^{ff} &= \int_{\mathcal{D}} \rho \mathbf{N}^{f\,\mathrm{T}} \mathbf{N}^f\, \mathrm{dv}, \\
\mathbf{M}^{fp} &= \int_{\mathcal{D}} \rho \mathbf{N}^{f\,\mathrm{T}} \mathbf{N}^p\, \mathrm{dv} \quad \text{Mass matrices}
\end{aligned} \tag{153f}$$

$$\begin{aligned}
\mathbf{C}^{ff} &= \int_{\mathcal{D}} \alpha \mathbf{N}^{f\,\mathrm{T}} \mathbf{N}^f\, \mathrm{dv}, \\
\mathbf{C}^{fp} &= \int_{\mathcal{D}} \alpha \mathbf{N}^{f\,\mathrm{T}} \mathbf{N}^p\, \mathrm{dv} \quad \text{Damping matrices}
\end{aligned} \tag{153g}$$

$$\begin{aligned}
\mathbf{K}^{ff} &= \int_{\mathcal{D}} \mathbf{B}^{f\,\mathrm{T}} \bar{\bar{\mathbf{C}}} \mathbf{B}^f\, \mathrm{dv}, \\
\mathbf{K}^{fp} &= \int_{\mathcal{D}} \mathbf{B}^{f\,\mathrm{T}} \bar{\bar{\mathbf{C}}} \mathbf{B}^p\, \mathrm{dv} \quad \text{Stiffness matrices}
\end{aligned} \tag{153h}$$

Often , (153b) are written in the short form,

$$\mathbf{M}\ddot{\mathbf{a}} + \mathbf{C}\dot{\mathbf{a}} + \mathbf{K}\mathbf{a} = \mathbf{F}_r + \mathbf{F}_N - \mathbf{F}_D \quad \text{ODE where} \tag{154a}$$

$$\mathbf{a}(t=0) = \mathbf{a}_0, \quad \dot{\mathbf{a}}(t=0) = \dot{\mathbf{a}}_0 \quad \text{Initial condition (IC)} \tag{154b}$$

where for short the superscripts for free dofs is dropped knowing that free and prescribed dofs are handles according to (153).

- As we will see through the following example, we actually DO NOT form $\mathbf{M}^{fp}$, $\mathbf{C}^{fp}$, $\mathbf{K}^{fp}$ directly, rather computing their corresponding values from elements and assemble their effects to global force $\mathbf{F}_D$. Local versions of $\mathbf{F}_D$ is

$$\mathbf{f}_D^e = \mathbf{M}^e\ddot{\mathbf{a}}^e + \mathbf{C}^e\dot{\mathbf{a}}^e + \mathbf{k}^e\mathbf{a}^e \tag{155}$$

where $\mathbf{a}^e$ is the local displacement vector of element formed by

1. Having zero values for free dofs.
2. Having prescribed values for prescribed dofs.

2.3.8 Types of damping matrix

- Rayleigh damping matrix, generalizes the formula for $\mathbf{C}$ from (153g) by basically adding a coefficient of stiffness matrix. That is,

$$\mathbf{C} = \alpha\mathbf{M} + \beta\mathbf{K} \tag{156}$$

  - Justification for α is as before by modeling the equation of motion as in (128), that is $\rho\mathbf{b} \rightarrow \rho\mathbf{b} - \alpha\mathbf{v}$ and getting (129) which is,

$$\dot{\mathbf{p}} - \nabla.\sigma + \alpha\mathbf{v} - \rho\mathbf{b} = \mathbf{0}$$

  - Justification for β is modifying equation of motion in the form,

$$\sigma = \mathcal{C}(\epsilon + \beta\dot{\epsilon}) \tag{157}$$

2.3.9 Stiffness and mass matrices for 1D elastostatics

- Shape functions are given by,

$$\begin{aligned}
N^e &= [N_1^e \; N_2^e] \quad \text{where} \\
N_1^e &= \frac{x_{i+1} - x}{L}, \quad N_2^e = \frac{x - x_i}{L}, \quad L = x_{i+1} - x_i
\end{aligned}$$

- In 1D displacement to strain relation is $\epsilon = \frac{\partial}{\partial x}u$ so $L_m = \frac{\partial}{\partial x}$, *cf.* (144b) for 3D version of L_m, and B^e is given by,

$$B^e = L_m[N_1^e \; N_2^e] = \frac{\partial}{\partial x}[\frac{x_{i+1} - x}{L} \;\; \frac{x - x_i}{L}] = \frac{1}{L}[-1 \; 1]$$

- Stiffness matrix From (153h), but for an element level where all dofs would be free, we have,

$$\mathbf{k}^e = \int_{\mathcal{D}} {B^e}^{\mathrm{T}} \bar{\bar{\mathbf{C}}} B^e \,\mathrm{dv} = \int_{x_i}^{x_{i+1}} \frac{1}{L}\begin{bmatrix}-1\\1\end{bmatrix} E \begin{bmatrix}-1 & 1\end{bmatrix} (A\,\mathrm{d}x) = \frac{AE}{L}\begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix} \tag{158}$$

  where A is area section, E elastic modulus, and L length of the bar. A, E are assumed to be constant.

- (Consistent) mass matrix is obtained from (153f), again element level; all dofs being free,

$$\begin{aligned}\mathbf{M}^e &= \int_{\mathcal{D}} {N^e}^{\mathrm{T}} \rho N^e \,\mathrm{dv} = \int_{x_i}^{x_{i+1}} \begin{bmatrix}\frac{x_{i+1}-x}{L}\\ \frac{x-x_i}{L}\end{bmatrix} \rho \begin{bmatrix}\frac{x_{i+1}-x}{L} & \frac{x-x_i}{L}\end{bmatrix} (A\,\mathrm{d}x)\\ &= \frac{AL\rho}{6}\begin{bmatrix}2 & 1\\1 & 2\end{bmatrix} = \frac{m^e}{6}\begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}\end{aligned} \tag{159}$$

  where ρ is mass density and $m^e = AL\rho$ is the mass of element. Note that,

  - The sum of components of $\mathbf{M}^e$ is equal to m^e.
  - The mass matrix is NOT diagonal.
  - As we will see this results in NON-DIAGONAL global level mass matrix.

2.3.10 Lumped mass matrix

To obtain a diagonal mass matrix, that would result in system level mass matrix:

- Use quadrature rules whose points match finite element nodes.
- We have,

$$M^e_{ij} = \int_{\mathcal{D}} N^e_i(x)\rho N^e_j(x)\,\mathrm{dv} \tag{160}$$

- The numerical quadrature of M^e_{ij} is then,

$$M^e_{ij} \approx M^{e_q}_{ij}, \qquad \text{where numerical quadrature of mass matrix } M^{e_q}_{ij} \text{ is } \quad M^{e_q}_{ij} = \sum_{k=1}^{n_q} w_k N^e_i(x_k) N^e_j(x_k)\rho(x_k)J(x_k) \tag{161}$$

  where w_k are the weight values of the quadrature, n_q is the number of quadrature points, and x_k are the quadrature points. J is the Jacobian function for the transformation from the element coordinate to the quadrature parent element.

- Now if we use a quadrature scheme whose quadrature points match element nodal points by delta property of FEM shape functions (that is shape function i takes that value of 1 at node i and zero at other nodes) and the coincidence of nodal positions and quadrature points we have,

$$N_i(x_j) = \delta_{ij} \tag{162}$$

- Using in (161) it is clear that the mass matrix becomes diagonal and diagonal values are

$$M^{e_q}_{ii} = \sum_{k=1}^{n_q} w_k N^e_i(x_k) N^e_i(x_k)\rho(x_k)J(x_k) \qquad \Rightarrow \boxed{M^{e_q}_{ii} = w_i\rho(x_i)J(x_i) \qquad (\text{no summation on } i)} \tag{163}$$

- For a constant density and cross section (1D) or thickness (2D) we simply have,

$$\boxed{M^{e_q}_{ii} = w_i m^e} \quad m^e = \text{ element mass}, \qquad (\text{Uniform mass density and section (1D)/ thickness (2D)}) \tag{164}$$

  again there is no summation on i.

- As an example consider the lumped mass matrix for the first order 1D bar element.
- The quadrature scheme for a line of length L is:

$$\text{Quadrature}(\int_0^L f(x)\,\mathrm{d}x) = \frac{L}{2}f(0) + \frac{L}{2}f(L) \tag{165}$$

- That is this quadrature scheme uses the two end points of a line segment, which will be the end points of an element in computing the lumped mass of the element.

$$\mathbf{M}^e \approx \mathbf{M}^{e_q} = \text{Quadrature}\left(\int_{\mathcal{D}} {N^e}^{\mathrm{T}} \rho(x) N^e \,\mathrm{dv}\right) = \frac{L}{2}\begin{bmatrix} N_1(0) \\ N_2(0) \end{bmatrix} \rho(0)A(0) \begin{bmatrix} N_1(0) & N_2(0) \end{bmatrix} + \frac{L}{2}\begin{bmatrix} N_1(L) \\ N_2(L) \end{bmatrix} \rho(L)A(L) \begin{bmatrix} N_1(L) & N_2(L) \end{bmatrix} \Rightarrow$$

$$\mathbf{M}^{e_q} = \frac{L}{2}\begin{bmatrix} 1 \\ 0 \end{bmatrix} \rho(0)A(0) \begin{bmatrix} 1 & 0 \end{bmatrix} + \frac{L}{2}\begin{bmatrix} 0 \\ 1 \end{bmatrix} \rho(L)A(L) \begin{bmatrix} 0 & 1 \end{bmatrix} = \begin{bmatrix} \rho(0)A(0)\frac{L}{2} & 0 \\ 0 & \rho(L)A(L)\frac{L}{2} \end{bmatrix} \tag{166}$$

where $\mathrm{dv} = A\,\mathrm{d}x$.

- If further we assume that ρ and A are uniform we have,

$$\mathbf{M}^e \approx \mathbf{M}^{e_q} = \frac{m^e}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{diagonal mass matrix for constant } \rho \text{ and } A \tag{167}$$

This time we note,

- The sum of components of $\mathbf{M}^e$ is again equal to m^e.
- The mass matrix IS diagonal.
- As we will see this will results in DIAGONAL global level mass matrix.
- For the second order bar element we use the Simpson rule,

$$\text{Quadrature}(\int_0^L f(x)\,\mathrm{d}x) = \frac{L}{6}f(0) + \frac{4L}{6}f(L/2) + \frac{L}{6}f(L) \tag{168}$$

which results in

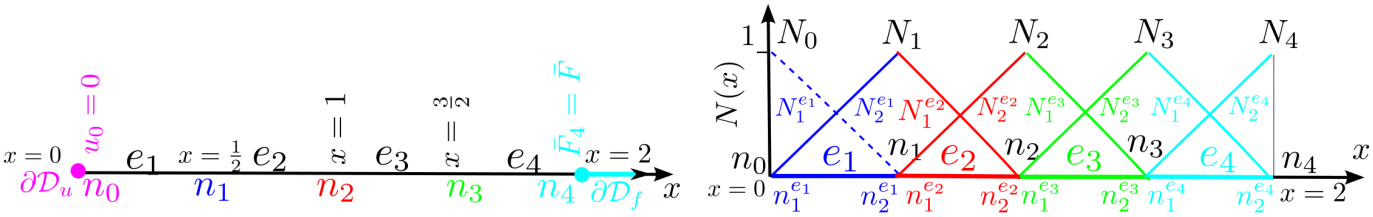
$$\mathbf{M}^e = \frac{m^e}{6}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{for second order 1D element} \tag{169}$$

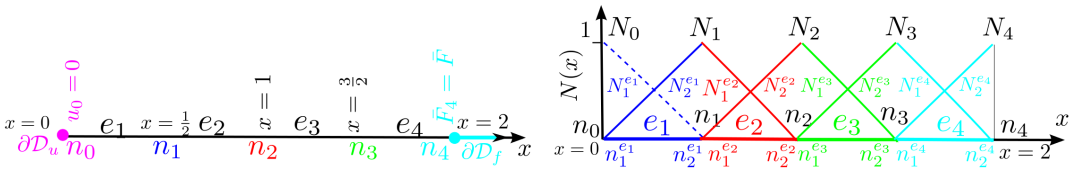
note that,

  - The lumped mass matrix is again diagonal.
  - The diagonal values may NOT necessarily be equal.

- For even higher order (p) bar elements, unfortunately using uniform distant internal element nodes at L/p positions and quadrature points corresponding to those points limits the order of accuracy in which the mass matrix is integrated.
- This in turn affects the FEM solutions convergence rates for the nodal solution **U** and other solution features such as modal quantities; *cf.* §3.1.7.
- To not sacrifice the order in which the element mass matrix is integrated we do two things:
  1. Choose the two end point of the element as two of the quadrature points because we have no freedom in changing the position of the element end nodes.
  2. Similar to Gauss quadrature formulation we optimize the position of high order element nodal (*i.e.*, quadrature) positions.
- As a result the corresponding quadrature scheme with these optimized points will have sufficient order of accuracy and the FEM solution convergence rates are not affected.
- The optimized scheme of quadrature points that include the end points is called Lobatto quadrature
- For more information refer to Section 7.3.2 [Hughes, 2012].

2.3.11 Example for the assembly of global matrix systems





- We recall that the element matrices for first order 1D bar elements were

$$\mathbf{k}^e = \frac{AE}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad \text{Stiffness matrix} \tag{170a}$$

$$\mathbf{M}^e = \frac{m^e}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \text{consistent mass matrix} \tag{170b}$$

$$\mathbf{M}^e = \frac{m^e}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{lumped mass matrix} \tag{170c}$$

- Local stiffness matrix: Since, $E=1, A=1, L=\frac{1}{2}$ for all elements, $\mathbf{k}^e$ is given by (170a)

$$\mathbf{k}^e = \frac{(1).(1)}{1/2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}$$

- Local mass matrix: Since, $m^e = 1 (A=1, L=\frac{1}{2}, \rho=2)$ for all elements, $\mathbf{M}^e$ is given by (170b) and (170c)

$$\mathbf{M}^e = \frac{1}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \text{consistent mass matrix} \tag{171a}$$

$$\mathbf{M}^e = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{lumped mass matrix} \tag{171b}$$

e	e_1	e_2	e_3	e_4
$\mathbf{k}^e$	$\bar{1}\ 1$ $\bar{1}$ 2 -2 1 -2 2	1 1 1 2 -2 1 -2 2	2 3 2 2 -2 3 -2 2	3 4 3 2 -2 4 -2 2
$\mathbf{M}^e$	$\bar{1}\ 1$ $\bar{1}$ $\frac{1}{3}$ $\frac{1}{6}$ 1 $\frac{1}{6}$ $\frac{1}{3}$	1 1 1 $\frac{1}{3}$ $\frac{1}{6}$ 1 $\frac{1}{6}$ $\frac{1}{3}$	2 3 2 $\frac{1}{3}$ $\frac{1}{6}$ 3 $\frac{1}{6}$ $\frac{1}{3}$	3 4 3 $\frac{1}{3}$ $\frac{1}{6}$ 4 $\frac{1}{6}$ $\frac{1}{3}$

e	e_1	e_2	e_3	e_4
$\mathbf{f}_D^e$	$\mathbf{k}^{e_1}\mathbf{a}_1^e + \mathbf{M}f^{e_1}\ddot{\mathbf{a}}_1^e =$ $\begin{matrix}\bar{1}\\1\end{matrix}\begin{bmatrix}0\\0\end{bmatrix}$	$\mathbf{k}^{e_2}\mathbf{a}_2^e + \mathbf{M}f^{e_2}\ddot{\mathbf{a}}_2^e =$ $\begin{matrix}1\\2\end{matrix}\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{matrix}2\\3\end{matrix}\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{matrix}3\\4\end{matrix}\begin{bmatrix}0\\0\end{bmatrix}$
$\mathbf{f}_e^e$	$\mathbf{f}_e^{e_1} = \mathbf{f}_r^{e_1} + \mathbf{f}_N^{e_1} - \mathbf{f}_D^{e_1} =$ $\begin{matrix}\bar{1}\\1\end{matrix}\begin{bmatrix}0\\0\end{bmatrix}$	$\mathbf{f}_e^{e_2} = \mathbf{f}_r^{e_2} + \mathbf{f}_N^{e_2} - \mathbf{f}_D^{e_2} =$ $\begin{matrix}1\\2\end{matrix}\begin{bmatrix}0\\0\end{bmatrix}$	$\mathbf{f}_e^{e_3} =$ $\begin{matrix}2\\3\end{matrix}\begin{bmatrix}0\\0\end{bmatrix}$	$\mathbf{f}_e^{e_4} =$ $\begin{matrix}3\\4\end{matrix}\begin{bmatrix}0\\0\end{bmatrix}$

$$\mathbf{K} = \begin{bmatrix} 2+2 & -2 & 0 & 0 \\ -2 & 2+2 & -2 & 0 \\ 0 & -2 & 2+2 & -2 \\ 0 & 0 & -2 & 2 \end{bmatrix} = \begin{bmatrix} 4 & -2 & 0 & 0 \\ -2 & 4 & -2 & 0 \\ 0 & -2 & 4 & -2 \\ 0 & 0 & -2 & 2 \end{bmatrix}$$

$$\mathbf{M} = \begin{bmatrix} \frac{1}{3}+\frac{1}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & \frac{1}{3}+\frac{1}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{1}{3}+\frac{1}{3} & \frac{1}{6} \\ 0 & 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix}, \quad \mathbf{F} = \mathbf{F}_N + \mathbf{F}_e = \begin{bmatrix} 0 \\ 0 \\ 0 \\ F(\bar{t}) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ F(\bar{t}) \end{bmatrix} \Rightarrow$$

$$\mathbf{M\ddot{a}+Ka=F} \quad \text{that is} \quad \begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} \ddot{a}_1 \\ \ddot{a}_2 \\ \ddot{a}_3 \\ \ddot{a}_4 \end{bmatrix} + \begin{bmatrix} 4 & -2 & 0 & 0 \\ -2 & 4 & -2 & 0 \\ 0 & -2 & 4 & -2 \\ 0 & 0 & -2 & 2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \bar{F}(t) \end{bmatrix} \quad \begin{matrix} \underline{\text{consistent}} \\ \text{mass} \\ \text{matrix} \end{matrix} \tag{172}$$

with lumped mass matrix (171b) we would have got

$$\mathbf{M\ddot{a}+Ka=F} \quad \text{that is} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \ddot{a}_1 \\ \ddot{a}_2 \\ \ddot{a}_3 \\ \ddot{a}_4 \end{bmatrix} + \begin{bmatrix} 4 & -2 & 0 & 0 \\ -2 & 4 & -2 & 0 \\ 0 & -2 & 4 & -2 \\ 0 & 0 & -2 & 2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \bar{F}(t) \end{bmatrix} \quad \begin{matrix} \underline{\text{lumped}} \\ \text{mass} \\ \text{matrix} \end{matrix} \tag{173}$$

3 General solution schemes in time (or spacetime)

3.1 Modal superposition

3.1.1 Modal analysis: Motivation

- We are interested in solving the system,
$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R} \tag{174}$$
where $\mathbf{M}, \mathbf{C}, \mathbf{K}$ are mass, damping, and stiffness matrices, $\mathbf{U}$ is the vector of nodal unknowns and $\mathbf{R}$ is the force vector.
- We are looking for a mapping of the form,
$$\mathbf{U}(t) = \mathbf{P}\mathbf{X}(t) \tag{175}$$
where
  - $\mathbf{P}$ is an $n \times n$ square matrix.
  - $\mathbf{X}(t)$ is the size n vector of generalized displacements.
- By plugging (175) into (174) and pre-multiplying by $\mathbf{P}^{\mathrm{T}}$ we obtain,
$$\tilde{\mathbf{M}}\ddot{\mathbf{X}} + \tilde{\mathbf{C}}\dot{\mathbf{X}} + \tilde{\mathbf{K}}\mathbf{X} = \tilde{\mathbf{R}}, \quad \text{where} \tag{176a}$$
$$\tilde{\mathbf{M}} = \mathbf{P}^{\mathrm{T}}\mathbf{M}\mathbf{P}, \qquad \tilde{\mathbf{C}} = \mathbf{P}^{\mathrm{T}}\mathbf{C}\mathbf{P}, \qquad \tilde{\mathbf{K}} = \mathbf{P}^{\mathrm{T}}\mathbf{K}\mathbf{P}, \qquad \tilde{\mathbf{R}} = \mathbf{P}^{\mathrm{T}}\mathbf{R} \tag{176b}$$
- The idea is to find the mapping $\mathbf{P}$ such that the modified mass, damping, and stiffness matrices $\tilde{\mathbf{M}}$, $\tilde{\mathbf{C}}$, $\tilde{\mathbf{K}}$ have a shorter bandwidth and it is computationally less expensive to solve (176a) than (174).
- Once $\mathbf{X}$ is solved we obtain original FEM nodal unknowns $\mathbf{U}(t)$ by (175) ($\mathbf{U}(t) = \mathbf{P}\mathbf{X}(t)$).

3.1.2 Modal analysis: Natural modes and frequencies

- Modal analysis facilitate an finding appropriate choice for $\mathbf{P}$.
- Consider (174) but without the damping term,
$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R} \qquad \text{Matrix equations without damping term} \tag{177}$$
- By plugging a solution in the form,
$$\mathbf{U} = \boldsymbol{\Phi} \sin \omega(t - t_0) \tag{178}$$
- $\boldsymbol{\Phi}$ is an $n \times 1$ vector called mode shape or natural mode which captures the shape of vibration.
- ω is the natural frequency.
- Substituting (178) in (177) we obtain a generalized eigenproblem,
$$\mathbf{K}\boldsymbol{\Phi} = \omega^2 \mathbf{M}\boldsymbol{\Phi} \tag{179}$$
- The eigenproblem (179) yields n eigen solutions,
$$(\omega_1^2, \boldsymbol{\Phi}_1), (\omega_2^2, \boldsymbol{\Phi}_2), \cdots, (\omega_n^2, \boldsymbol{\Phi}_n) \tag{180}$$
- The eigenvectors (mode shapes / natural modes) are M-orthonormalized,
$$\boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{M}\boldsymbol{\Phi}_j = \delta_{ij}, \qquad \boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{M}\boldsymbol{\Phi}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{181a}$$
$$0 \leq \omega_1^2 \leq \omega_2^2 \leq \cdots \leq \omega_n^2 \tag{181b}$$
- Natural modes are $\mathbf{M}$-orthogonal (so that they can be orthonormalized) follows from the symmetry of $\mathbf{M}$ and $\mathbf{K}$.
- Eigenvalues $\lambda_i = \omega_i^2$ are positive (or nonnegative) because $\mathbf{M}$ is positive definite and $\mathbf{K}$ is positive definite (or positive only when rigid body motion modes are not fully restricted).

- The natural mode matrix is formed by $\boldsymbol{\Phi}_i$ (its columns are natural modes)
- and the natural frequency matrix is the diagonal matrix of natural frequencies,

$$\boldsymbol{\Phi}^{\mathrm{T}} = [\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \cdots, \boldsymbol{\Phi}_n] \qquad \boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \omega_2, \cdots, \omega_n) = \begin{bmatrix} \omega_1 & & & & \\ & \omega_2 & & & \\ & & . & & \\ & & & . & \\ & & & & . \\ & & & & & \omega_n \end{bmatrix} \tag{182}$$

- Equations (179) to (182) yield,

$$\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{K}\boldsymbol{\Phi} = \boldsymbol{\Omega}^2 \tag{183a}$$

$$\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\boldsymbol{\Phi} = \mathbf{I} \qquad \boldsymbol{\Phi}^{-1} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M} \tag{183b}$$

- That is, $\boldsymbol{\Omega}$ is a very suitable transformation matrix $\mathbf{P}$ which in fact decouples the equations in (176) (when $\mathbf{C} = 0$).
- and (175) ($\mathbf{U}(t) = \mathbf{P}\mathbf{X}(t)$) becomes,

$$\mathbf{U}(t) = \boldsymbol{\Phi}\mathbf{X}(t) \Rightarrow \tag{184a}$$

$$\mathbf{X}(t) = \boldsymbol{\Phi}^{-1}\mathbf{U}(t) \qquad \text{that is} \qquad \mathbf{X}(t) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\mathbf{U}(t) \quad \textit{cf.}\ (183b)) \tag{184b}$$

- In general for $\mathbf{C} \neq 0$ (176a) with map (184a) and (183) gives the simplified system of ODEs in $\mathbf{X}(t)$ ((185a)).
- The initial conditions (ICs) are obtained by transforming ICs of the original system in terms of $\mathbf{U}$: $\mathbf{U}^0 = \mathbf{U}(t=0)$ for initial displacement and $\dot{\mathbf{U}}^0 = \dot{\mathbf{U}}(t=0)$ for initial velocity. The transformed values are shown in (185b).

$$\ddot{\mathbf{X}}(t) + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{C}\boldsymbol{\Phi}\dot{\mathbf{X}}(t) + \boldsymbol{\Omega}^2\mathbf{X}(t) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}(t) \qquad \text{System of ODEs} \tag{185a}$$

$$\mathbf{X}^0 = \mathbf{X}(t=0) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\mathbf{U}^0 \qquad \dot{\mathbf{X}}^0 = \dot{\mathbf{X}}(t=0) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\dot{\mathbf{U}}^0 \qquad \text{Initial conditions (ICs)} \tag{185b}$$

- Depending on the form of $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{C}\boldsymbol{\Phi}$ in (185) we have different scenarios,
  1. If $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{C}\boldsymbol{\Phi}$ is a full matrix x_i, *i.e.,* components of $\mathbf{X}$, are not decoupled.
  2. If $\mathbf{C} = 0$ or more generally $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{C}\boldsymbol{\Phi}$ is diagonal, (185a) can be solved as n decoupled equations.
- We first, discuss the case with no damping $\mathbf{C} = 0$, followed by diagonalizable damping matrix, and finally a damping matrix that cannot be diagonalized with the modal analysis.

3.1.3 Analysis with damping neglected or zero damping

- When $\mathbf{C} = 0$ or can be neglected, (185) becomes

$$\ddot{\mathbf{X}}(t) + \boldsymbol{\Omega}^2\mathbf{X}(t) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}(t) \qquad \text{System of ODEs}(\mathbf{C} = 0) \tag{186a}$$

$$\mathbf{X}^0 = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\mathbf{U}^0 \qquad \dot{\mathbf{X}}^0 = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\dot{\mathbf{U}}^0 \qquad \text{Initial conditions (ICs)} \tag{186b}$$

- This can be written in the form of n decoupled equations for components of $\mathbf{X}(t)$,

$$\left.\begin{aligned} \ddot{x}_i(t) + \omega_i^2 x_i(t) &= r_i(t) \\ r_i(t) &= \boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{R}(t) \end{aligned}\right\} \qquad i = 1, \cdots, n \qquad \text{Uncoupled ODEs} \tag{187a}$$

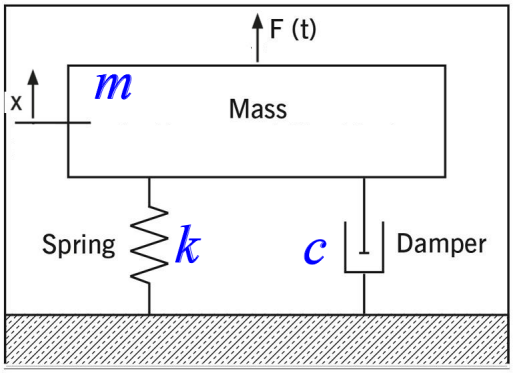
$$\left.\begin{aligned} x_i{}^0 = x_i|_{t=0} &= \boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{M}\mathbf{U}^0 \\ \dot{x}_i{}^0 = \dot{x}_i|_{t=0} &= \boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{M}\dot{\mathbf{U}}^0 \end{aligned}\right\} \qquad i = 1, \cdots, n \qquad \text{ICs} \tag{187b}$$

- This corresponds to a *single degree of freedom* (SDOF) mass-spring-damper system with mass m, spring stiffness k, and damping coefficient c:

$$F(t) - kx - c\dot{x} = m\ddot{x} \quad \Rightarrow \quad \ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = f(t) \tag{188}$$

where

1. ξ: damping coefficient. Transition from $\xi < \xi_{cr}$ to $\xi > \xi_{cr}$ for $\xi_{cr} = 1$ implies transition from a oscillatory-evanescent mode to purely evanescent mode.
2. critical damping $\xi_{cr} = 1$.
3. Natural frequency $\omega = \sqrt{\frac{k}{m}}$.
4. $f(t) = F(t)/m$.



- The solution to the SDOF undamped oscillator (187) can be obtained by analytical methods or by a variety of 2nd order initial value ODE solvers such as Newmark method.
- For example, we can use the Duhamel's integral to exactly express the solution,

$$x_i(t) = \frac{1}{\omega_i}\int_0^t r_i(\tau)\sin\omega_i(t-\tau)\mathrm{d}\tau + \alpha_i \sin\omega_i t + \beta_i \cos\omega_i t \tag{189}$$

where α_i and β_i are obtained by the ICs (187b),

$$\alpha_i = \frac{\dot{x}_i{}^0}{\omega_i}, \qquad \beta_i = x_i{}^0. \tag{190}$$

- Regardless on how we obtain $x_i(t)$ we can form the FE nodal solutions by,

$$\mathbf{U} = \sum_{i=1}^{n} \boldsymbol{\Phi}_i x_i(t) \tag{191}$$

- Thus, solution of a dynamic system with modal analysis requires,
  1. Natural modal analysis: Determination of $\omega_i, \boldsymbol{\Phi}_i$.
  2. Solution of decoupled SDOF systems (187) (if we can obtain decoupled SDOF systems with damping as in (188) those can also be solved independent of each other).
  3. Transferring the solutions back to $\mathbf{U}$ by (191) ($\mathbf{U} = \sum_{i=1}^{n} \boldsymbol{\Phi}_i x_i(t)$).
- For many applications, we are only interested in natural modes and frequencies and do not follow up with steps 2 and 3 above.
- Example below from [Bathe, 2006] Examples 9.6 and 9.7 demonstrate this process:

3.1.3.1 Solution of an undamped system using modal analysis ([Bathe, 2006])

EXAMPLE 9.6: Calculate the transformation matrix Φ for the problem considered in Examples 9.1 to 9.4 and thus establish the decoupled equations of equilibrium in the basis of mode shape vectors.

For the system under consideration we have

$$\mathbf{K} = \begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix}; \qquad \mathbf{M} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}; \qquad \mathbf{R} = \begin{bmatrix} 0 \\ 10 \end{bmatrix}$$

The generalized eigenproblem to be solved is therefore

$$\begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix}\boldsymbol{\phi} = \omega^2 \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\boldsymbol{\phi}$$

The solution is obtained by one of the methods given in Chapters 10 and 11. Here we simply give

the two solutions without derivations:

$$\omega_1^2 = 2; \qquad \boldsymbol{\phi}_1 = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}$$

$$\omega_2^2 = 5; \qquad \boldsymbol{\phi}_2 = \begin{bmatrix} \frac{1}{2}\sqrt{\frac{2}{3}} \\ -\sqrt{\frac{2}{3}} \end{bmatrix}$$

Therefore, considering the free-vibration equilibrium equations of the system

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \ddot{\mathbf{U}}(t) + \begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix} \mathbf{U}(t) = \mathbf{0} \qquad \text{(a)}$$

the following two solutions are possible:

$$\mathbf{U}_1(t) = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} \sin \sqrt{2}\,(t - t_0^1) \quad \text{and} \quad \mathbf{U}_2(t) = \begin{bmatrix} \frac{1}{2}\sqrt{\frac{2}{3}} \\ -\sqrt{\frac{2}{3}} \end{bmatrix} \sin \sqrt{5}\,(t - t_0^2)$$

That the vectors $\mathbf{U}_1(t)$ and $\mathbf{U}_2(t)$ indeed satisfy the relation in (a) can be verified simply by substituting $\mathbf{U}_1$ and $\mathbf{U}_2$ into the equilibrium equations. The actual solution to the equations in (a) is of the form

$$\mathbf{U}(t) = \alpha \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} \sin \sqrt{2}\,(t - t_0^1) + \beta \begin{bmatrix} \frac{1}{2}\sqrt{\frac{2}{3}} \\ -\sqrt{\frac{2}{3}} \end{bmatrix} \sin \sqrt{5}\,(t - t_0^2)$$

where α, β, t_0^1, and t_0^2 are determined by the initial conditions on $\mathbf{U}$ and $\dot{\mathbf{U}}$. In particular, if we impose initial conditions corresponding to α (or β) only, we find that the system vibrates in the corresponding eigenvector with frequency $\sqrt{2}$ rad/sec (or $\sqrt{5}$ rad/sec). The general procedure of solution for α, β, t_1^0, and t_2^0 is discussed in Section 9.3.2.

Having evaluated $(\omega_1^2, \boldsymbol{\phi}_1)$ and $(\omega_2^2, \boldsymbol{\phi}_2)$ for the problem in Examples 9.1 to 9.4, we arrive at the following equilibrium equations in the basis of eigenvectors:

$$\ddot{\mathbf{X}}(t) + \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix} \mathbf{X}(t) = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{2}\sqrt{\frac{2}{3}} & -\sqrt{\frac{2}{3}} \end{bmatrix} \begin{bmatrix} 0 \\ 10 \end{bmatrix}$$

or

$$\ddot{\mathbf{X}}(t) + \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix} \mathbf{X}(t) = \begin{bmatrix} \frac{10}{\sqrt{3}} \\ -10\sqrt{\frac{2}{3}} \end{bmatrix}$$

EXAMPLE 9.7: Use mode superposition to calculate the displacement response of the system considered in Examples 9.1 to 9.4 and 9.6.

(1) Calculate the exact response by integrating each of the two decoupled equilibrium equations exactly.

We established the decoupled equilibrium equations of the system under consideration in Example 9.6; i.e., the two equilibrium equations to be solved are

$$\ddot{x}_1 + 2x_1 = \frac{10}{\sqrt{3}}; \ddot{x}_2 + 5x_2 = -10\sqrt{\frac{2}{3}} \tag{a}$$

The initial conditions on the system are $\mathbf{U}|_{t=0} = \mathbf{0}$, $\dot{\mathbf{U}}|_{t=0} = \mathbf{0}$, and hence, using (9.46), we have

$$\begin{aligned} x_1|_{t=0} &= 0 \qquad \dot{x}_1|_{t=0} = 0 \\ x_2|_{t=0} &= 0 \qquad \dot{x}_2|_{t=0} = 0 \end{aligned} \tag{b}$$

Also, to obtain $\mathbf{U}$ we need to use the relation in (9.42), which, using the eigenvectors calculated in Example 9.6, gives

$$\mathbf{U}(t) = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{2}\sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & -\sqrt{\frac{2}{3}} \end{bmatrix} \mathbf{X}(t) \tag{c}$$

The exact solutions to the equations in (a) and (b) are

$$x_1 = \frac{5}{\sqrt{3}}(1 - \cos\sqrt{2}t); x_2 = 2\sqrt{\frac{2}{3}}(-1 + \cos\sqrt{5}t) \tag{d}$$

Hence, using (c), we have

$$\mathbf{U}(t) = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{2}\sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & -\sqrt{\frac{2}{3}} \end{bmatrix} \begin{bmatrix} \frac{5}{\sqrt{3}}(1 - \cos\sqrt{2}t) \\ 2\sqrt{\frac{2}{3}}(-1 + \cos\sqrt{5}t) \end{bmatrix} \tag{e}$$

Evaluating the displacements from (e) for times $\Delta t, 2\Delta t, \ldots, 12\Delta t$, where $\Delta t = 0.28$, we obtain

Time	Δt	$2\Delta t$	$3\Delta t$	$4\Delta t$	$5\Delta t$	$6\Delta t$	$7\Delta t$	$8\Delta t$	$9\Delta t$	$10\Delta t$	$11\Delta t$	$12\Delta t$
$^t\mathbf{U}$	0.003	0.038	0.176	0.486	0.996	1.66	2.338	2.861	3.052	2.806	2.131	1.157
	0.382	1.41	2.78	4.09	5.00	5.29	4.986	4.277	3.457	2.806	2.484	2.489

3.1.4 Use of the first few modes in the analysis

- To obtain the nodal solution $\mathbf{U}$ we need to solve for all $x_i(t), i \leq n$ to form (*cf.* (191)),

$$\mathbf{U} = \sum_{i=1}^{n} \boldsymbol{\Phi}_i x_i(t)$$

- where each $x_i(t)$ is obtained by solving an ODE (187a)

$$\ddot{x}_(t) + \omega_i^2 x_i(t) = r_i(t) = \boldsymbol{\Phi}_i^{\mathrm{T}} \mathbf{R}(t)$$

  or even one with the damping term as in (188) ($\ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = f(t)$) in general.

- In practice not the coefficients of all modes x_i are significant and in some application only a first few modes are considered in the solution,

$$\mathbf{U} \approx \sum_{i=1}^{p} \boldsymbol{\Phi}_i x_i(t)$$

- Then instead of solving n SDOF ODEs for $x_i(t)$, only p ODEs are solved. In some applications $p \ll n$.

- If we choose $p = n$ there is no difference in directly solving $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ or using modal dofs x_i and then forming $\mathbf{U}$ by $\mathbf{U} = \sum_{i=1}^{n} \boldsymbol{\Phi}_i x_i(t)$. That is the solution of these two systems analytically is the same!

- Obviously, once each scheme is numerically integrated, that is n SDOF ODEs with modal analysis and 1 size n $\mathbf{U}$ the source of error is the numerical integration of each system.

- Once we replace $\mathbf{U} = \sum_{i=1}^{n} \boldsymbol{\Phi}_i x_i(t)$ by $\mathbf{U} = \sum_{i=1}^{p} \boldsymbol{\Phi}_i x_i(t)$ in the modal analysis there is another source of error: Truncation of modes.

- The important question is when and how we can reduce the number of modes considered in the analysis. That is, how to decide p.

- To answer this question, we analyze the SDOF ODE (188), with and without c.

3.1.4.1 Analysis of a SDOF

- Consider (188) with zero damping $c = 0$ a harmonic force $f(t) = R\sin\hat{\omega}t$ and the following ICs,

$$\ddot{x} + \omega^2 x = R\sin\hat{\omega}t \qquad \text{ODE} \tag{192a}$$

$$x_i{}^0 = x_i|_{t=0} = 0, \qquad \dot{x}_i{}^0 = \dot{x}_{|_{t=0}} = 1, \qquad \text{IC} \tag{192b}$$

- Using (189) the solution to this system is,

$$x(t) = \frac{R}{\omega}\int_0^t \sin\hat{\omega}\tau \sin\omega(t-\tau)\mathrm{d}\tau + \alpha\sin\omega t + \beta\cos\omega t \tag{193}$$

  which after integration yields,

$$x(t) = \frac{R/\omega^2}{1-\hat{\omega}^2/\omega^2}\sin\hat{\omega}t + \alpha\sin\omega t + \beta\cos\omega t$$

- and by using ICs,

$$x|_{t=0} = \beta, \dot{x}|_{t=0} = \frac{R\hat{\omega}/\omega^2}{1-\hat{\omega}^2/\omega^2} + \alpha\sin\omega t$$

- which gives,

$$\beta = 0, \qquad \alpha = \frac{1}{\omega} - \frac{R\hat{\omega}/\omega^3}{1-\hat{\omega}^2/\omega^2}$$

- Substituting α and β in (193) we obtain,

$$x(t) = \frac{R/\omega^2}{1-\hat{\omega}^2/\omega^2}\sin\hat{\omega}t + \frac{1}{\omega}\left(1 - \frac{R\hat{\omega}/\omega^2}{1-\hat{\omega}^2/\omega^2}\right)\sin\omega t \tag{194}$$

- This can be written as,

$$x(t) = D x_{\text{stat}} + x_{\text{trans}} \tag{195a}$$

$$x_{\text{stat}} = \frac{R}{\omega^2}\sin\hat{\omega}t \qquad \text{static response} \tag{195b}$$

$$x_{\text{trans}} = \left(\frac{1}{\omega} - \frac{R\hat{\omega}/\omega^3}{1-\hat{\omega}^2/\omega^2}\right)\sin\omega t \qquad \text{transient response} \tag{195c}$$

$$D = \frac{1}{1-\hat{\omega}^2/\omega^2} \qquad \text{dynamic load factor} \tag{195d}$$

- The static response is obtained by ignoring inertia terms $\ddot{x}$ in (192a),

$$\omega^2 x_{\text{stat}} = R\sin\hat{\omega}t \qquad \Rightarrow \qquad x_{\text{stat}} = \frac{R\sin\hat{\omega}t}{\omega^2}$$

- x_{trans} specifically depends on the choice of IC ($x_i|_{t=0} = 0$ $\dot{x}_i|_{t=0} = 1$ herein or for example $x_i|_{t=0} = 1$ $\dot{x}_i|_{t=0} = 0$).
- In addition if the damping is nonzero $c \neq 0$ ($\xi \neq 0$) the transient response damps out.
- Given the dependence of x_{trans} on the form of IC and the fact that it damps out when damping is nonzero (as opposed to x_{stat}), makes x_{stat} the main term to consider.
- Dynamic load factor D shows how larger the dynamic response is relative to a static response (when inertia effects are ignored).
- The ratio of dynamic solution (total solution minus transient solution) to static solution (what we would have obtained by inertia effects) is dynamic amplification factor which took the value:

$$D = \frac{1}{1-\hat{\omega}^2/\omega^2} \qquad \text{undamped oscillator}$$

- There are three important ranges of $\hat{\omega}$ that we observe from this equation
  1. $\hat{\omega} \ll \omega$ very slow varying load: $D \approx 1$: That is, we are in quasi-static regime and ignoring inertia effects $\ddot{x}$ (and damping as we discuss later $\dot{x}$) is reasonable. Basically, loading rate is so slow that with any increment of loading the system has enough time to reach to a static equilibrium which is why we can ignore $\ddot{x}$ (and $\dot{x}$). In fact, for quasi-static loading regime, we can solve the solution by ignoring $\mathbf{M}$ (and $\mathbf{C}$) in (174) and have $\mathbf{K}\Delta\mathbf{U} = \Delta\mathbf{R}$ between time steps.
  2. $\hat{\omega} \approx \omega$ which is at or near resonance: We have the largest D. For an undamped oscillator $D \to \infty$ as $\hat{\omega} \to \omega$, *i.e.,* when the loading resonance occurs. Later, we show that D remains bounded when damping is added. Still D can get larger than unity for $\hat{\omega}$ near the undamped resonance frequency.
  3. $\hat{\omega} \gg \omega$ very fast varying/oscillating load: In this case the load oscillates so fast that the SDOF system does not have time to respond and basically dynamic respond would be close to zero. That is $D \to 0$ when $\hat{\omega}/\omega \to \infty$.
- To simplify the discussion, we define,

$$\Omega := \frac{\hat{\omega}}{\omega} \qquad \text{Frequency ratio} \tag{196}$$

- Summary:
  1. $\Omega \to 0$ very slow varying load: $D \approx 1$: A quasistatic solution (using $\mathbf{K}$ suffices): $\mathbf{K}\Delta\mathbf{U} = \Delta\mathbf{R}$.
  2. $\omega = \mathcal{O}(1)$ (or $0 \ll \Omega \ll \infty$) near resonance (or nontrivial solution): Need to find the dynamic solution and D can get very large.
  3. $\Omega \to \infty$ very fast varying load: $D \approx 0$: System response is almost zero and do not need to be considered.
- So far, we considered **one** SDOF oscillator but **varied loading frequency** $\hat{\omega}$.
- Conversely that we have **many** SDOF oscillators but **one or a short band of loading frequencies** $\hat{\omega}$.
- The latter is of more importance for the solution of (174) $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ where when the system could be diagonalized by modal decomposition we could get an equation of the form (187a) (in this equation $\mathbf{C}$ is assumed to be zero):

$$\ddot{x}_{(}t) + \omega_i^2 x_i(t) = r_i(t) = \boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{R}(t) \tag{197}$$

- For the moment assume $\mathbf{R}(t) = \mathbf{R}_0 \sin(\hat{\omega} t)$ (we will generalize the discussion to when there is a frequency band in the Fourier transform of the loading).
- In this case, for each SDOF i in (197) we have the loading to natural frequency ratio $\Omega_i = \frac{\hat{\omega}}{\omega_i}$. Since, $\hat{\omega}$ is fixed (or would represent a frequency band) and ω_i is varying, it would be more natural to consider the inverse ratio of Ω_i. There are three cases based on the previous discussion:

1. **High natural frequency** ω_i that is $\Omega_i = \frac{\hat{\omega}}{\omega_i} \to 0$: The contribution to the mode i can be obtained by a quasistatic analysis.
2. **Medium natural frequency** ω_i that is $0 \ll \Omega_i = \frac{\hat{\omega}}{\omega_i} \ll \infty$: The contribution to the mode i must be considered by solving the SDOF $\ddot{x}_i(t) + \omega_i^2 x_i(t) = r_i(t)$ (damping term $2\xi_i \omega_i^2 \dot{x}_i(t)$ can be added).
3. **Very low natural frequency** ω_i that is $\Omega_i = \frac{\hat{\omega}}{\omega_i} \to \infty$: The contribution to the mode i is almost zero and the SDOF for $x_i(t)$ does not need to be solved.

3.1.4.2 Use of first several natural modes

Now consider that loadings to the problem (ICs, BCs, and the source term) have a considerable frequency content between $[\hat{\omega}_m \ \hat{\omega}_M]$. That is their Fourier transform is almost zero or negligible outside this band. From the discussion above we conclude:

- Very high natural frequencies: $\omega_i \gg \hat{\omega}_M$: For these modes the loading frequency relative to them is very small. So, a quasi-static analysis would suffice. For example assume $\Delta\mathbf{R}^n$ is the part of force vector in time step t_n that we do incorporate by only including the modal solutions of mode one to mode p. The response of the missing modes (high frequency modes) can be added by solving $\mathbf{K}\Delta\mathbf{U}^n = \Delta\mathbf{R}^n$.
- Medium natural frequencies: $\hat{\omega}_m \lesssim \omega_i \lesssim \hat{\omega}_M$: The SDOF solutions for these modes have nontrivial dynamic solution and must be solved. **These modes must be considered in the p reduced set of modes**.
- Very low natural frequencies: $\omega_i \ll \hat{\omega}_m$: The loading frequency relative to these mode frequencies is very high and their effect is basically zero. Thus, these modes should not be considered in the p reduced set of modes and their effect is basically zero.

In practice loadings often include lowest frequency modes and the structure can easily be vibrated in its lowest modes. That is $\hat{\omega}_m$ should often be considered zero and the case 3 above would be irrelevant in most practices. As discussed later, in certain vibration problems even some lowest modes can be omitted.

3.1.4.3 Modal analysis: How to choose the reduced number of modes p

- Assume the highest relevant frequency content of the loading is $\hat{\omega}_M$. Choose p natural modes / frequencies such that $\omega_p \gg \hat{\omega}_M$.
- In fact, the solution of natural modes and frequencies can be done one at a time, until the last frequency become irrelevant instead of solving for all natural frequencies / modes which can be very expensive for a large system.
- If the purpose is only obtaining the relevant natural modes / frequencies to the problem we are done.
- If we need to solve a transient problem of the form (174):

$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$$

we solve the SDOF ODEs similar to (187) (here damping term is added to SDOFs assuming that the damping term can also be diagonalized. This is discussed later):

$$\left.\begin{aligned} \ddot{x}_i(t) + 2\xi_i\omega_i\dot{x}_i(t) + \omega_i^2 x_i(t) &= r_i(t) \\ r_i(t) &= \boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{R}(t) \end{aligned}\right\} \quad i = 1, \cdots, n \quad \text{Uncoupled ODEs} \tag{198a}$$

$$\left.\begin{aligned} x_i{}^0 = x_i|_{t=0} &= \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\mathbf{U}^0 \\ \dot{x}_i{}^0 = s\dot{x}_i|_{t=0} &= \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\dot{\mathbf{U}}^0 \end{aligned}\right\} \quad i = 1, \cdots, n \quad \text{ICs} \tag{198b}$$

ONLY for the first p modes ($1 \le p$)

- Compute the nodal solution vector $\mathbf{U}$ using the modal solutions,

$$\mathbf{U} \approx \sum_{i=1}^{p} \boldsymbol{\Phi}_i x_i(t) \tag{199}$$

- To add the quasistatic contribution of loading through the higher modes ($1 > p$) that we did need include in the modal analysis (199) we compute the error in the load vector. Since $\mathbf{R} = \sum_{i=1}^{n} r_i \mathbf{M}\boldsymbol{\Phi}_i$ the error in $\mathbf{R}$ is,

$$\Delta\mathbf{R} = \mathbf{R} - \sum_{i=1}^{p} r_i \mathbf{M}\boldsymbol{\Phi}_i \tag{200}$$

  The solution to the quasi-static problem, *i.e.,* it is time dependent but inertia and damping terms are ignored, corresponding to the modes that were not included in the modal analysis can be solved from,

$$\mathbf{K}\Delta\mathbf{U} = \Delta\mathbf{R} \tag{201}$$

- The total solution is the sum of the dynamic solution from (199) and the quasi-static solution from (201): $\mathbf{U}_{\text{tot}} = \sum_{i=1}^{p} \boldsymbol{\Phi}_i x_i(t) + \Delta\mathbf{U}$.

3.1.4.4 Modal analysis vs. Direct numerical integration of $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$

The choice between computing natural frequencies / modes vs. direct temporal integration of $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ depends on various aspects.

- **Need for natural frequencies / modes**: In Many applications, regardless of the need to solve $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$, we need to obtain natural frequencies and modes which warrants a modal analysis.
- **Load frequency band** The frequency band of the loadings (BCs, ICs, body force) to a large extend determine how many modes (p) should be included in a modal analysis. We can define two classes of problems:
  1. Structural dynamic problems: Only the first few terms are sufficient for an accurate solution with modal analysis. For example, for earthquake loading in some cases only the 10 lowest modes need to be considered[Bathe, 2006]. If instead of using modal analysis, we directly want to integrate $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ in time, an implicit scheme is preferred because from accuracy perspective large time steps can be taken without affecting the solution much. Thus, the very small time step restriction of explicit methods can render them inefficient.
  2. Wave propagation problems: The loading frequency is very broadband. For example, in blast of shock loading p can be as high as $2/3n$ [Bathe, 2006]. Often, for wave propagation problems explicit numerical integration schemes are used because they are inexpensive and their restrictive time step is not of major concern because from accuracy perspective small time steps should be taken.

  Note: For certain vibration problems where loading has a narrow frequency band but the content is high frequency, *i.e.,* that is both $\hat{\omega}_m$ $\hat{\omega}_M$ are high but close to each other, we can omit the lowest natural modes whose frequencies are much smaller than $\hat{\omega}_m$ in the analysis. This reduced the number of modes that need to be considered.
- **Linearity of the problem**: Modal analysis is restricted to linear problems. Although, there may be cases that the nonlinear response can be linearized about the current state or approaches that can expand the applicability of such eigen mode analyses.
- **Influence of damping term**: If the damping term is nonzero AND nondiagonalizable with modal analysis we cannot directly use modal analysis for the solution of (174). Although, under structural dynamic loading we still can consider a much fewer modes $p \ll n$ but in this case p x_i terms will be coupled through the damping terms in their corresponding temporal ODEs. For further discussion refer to [Bathe, 2006] Example 9.11.

3.1.5 Effect of damping matrix

3.1.5.1 Damping in a SDOF problem

- To better understand the behavior of $\mathbf{C}$ in the modal analysis and solution of (174) ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ we first discuss the response of a damped SDOF problem.
- Consider the damped SDOF problem (188),

$$\ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = f(t) \tag{202}$$

- It is easier to discuss the response of the system and the dynamic amplification factor in frequency domain.
- The Fourier transform and the inverse Fourier transform in 1D are defined as,

$$\check{f}(\hat{\omega}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{-\imath\hat{\omega}t}\,\mathrm{d}t \qquad \Leftrightarrow \tag{203a}$$

$$f(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \check{f}(\hat{\omega})e^{\imath\hat{\omega}t}\mathrm{d}\hat{\omega} \tag{203b}$$

  provided that the integrals are defined in their domains.
- There are many important identities in Fourier transform analysis. One of relevance here is,

$$\left(\frac{\mathrm{d}^{\mathrm{n}}\check{\mathrm{f}}}{\mathrm{dt}^{\mathrm{n}}}\right)(\hat{\omega}) = (\imath\hat{\omega})^n\check{f}(\hat{\omega}) \tag{204}$$

- Accordingly, by taking the Fourier transform of (202) and application of (204) we have,

$$(\imath\hat{\omega})^2\check{x}(\hat{\omega}) + 2\xi\omega(\imath\hat{\omega})\check{x}(\hat{\omega}) + \omega^2\check{x}(\hat{\omega}) = \check{f}(\hat{\omega}) \quad \Rightarrow \tag{205a}$$

$$\check{x}_{\mathrm{dyn}}(\hat{\omega}) = \frac{\check{x}(\hat{\omega})}{(\omega^2-\hat{\omega}^2)+2\imath\xi\omega\hat{\omega}} \tag{205b}$$

- The subscript dyn is added to the solution to emphasize that this is the full dynamic solution.
- Now, we consider a quasi-static solution that ignores the inertia term $\ddot{x}$ and the damping term $\dot{x}$. Clearly, the solution to this system is,

$$\omega^2\check{x}_{\mathrm{stat}}(\hat{\omega}) = \check{f}(\hat{\omega}) \qquad \Rightarrow \qquad \check{x}_{\mathrm{stat}}(\hat{\omega}) = \frac{\check{f}(\hat{\omega})}{\omega^2} \tag{206}$$

- Recalling that $\Omega = \frac{\hat{\omega}}{\omega}$ we define ratio of dynamic to static solution,

$$H(\Omega,\xi) := \frac{\check{x}_{\mathrm{dyn}}(\hat{\omega})}{\check{x}_{\mathrm{stat}}(\hat{\omega})} = \frac{1}{(1-\Omega^2)+2\imath\xi\Omega}, \qquad \Omega = \frac{\hat{\omega}}{\omega} \tag{207}$$

- The fact that the ratio of the solution is a complex number means that their solution has a phase difference when $\xi \neq 0$.
- The amplification factor then will be the magnitude of $H(\Omega,\xi)$:

$$\boxed{D(\Omega,\xi) = |H(\Omega,\xi)| = \frac{1}{\sqrt{(1-\Omega^2)^2+(2\xi\Omega)^2}}} \tag{208}$$

- We observe that when $\xi > 0$ (is nonzero) unlike the undamped oscillator D never approaches infinity at a resonance frequency.
- In fact, the maximum amplification factor is,

$$D_M|_\Omega(\Omega,\xi) = D(\Omega_M,\xi) = \begin{cases} \frac{1}{2\xi\sqrt{1-\xi^2}} & \xi \le \frac{\sqrt{2}}{2} \qquad (\Omega_M = \sqrt{1-2\xi^2}) \\ 1 & \text{otherwise} \quad (\Omega_M = 0;\ \textit{i.e.},\ \text{static loading}) \end{cases} \tag{209}$$

- Furthermore the dynamic solution to (202) ($\ddot{x}+2\xi\omega\dot{x}+\omega^2x = f(t)$) is obtained by the Duhamel integral:

$$x(t) = \frac{1}{\tilde{\omega}}\int_0^t f(\tau)e^{-\xi\omega(t-\tau)}\sin\tilde{\omega}(t-\tau)\mathrm{d}\tau + e^{-\xi\omega t}(\alpha\sin\tilde{\omega}t + \beta\cos\tilde{\omega}t), \qquad \text{where} \quad \tilde{\omega} := \omega\sqrt{1-\xi^2} \tag{210}$$

  where α and β are obtained from the ICs.
- Thus, we can analyze a SDOF with damping term.
- We will discuss when these SDOFs become relevant to solve (174) ($\mathbf{M\ddot{U}+C\dot{U}+KU=R}$).

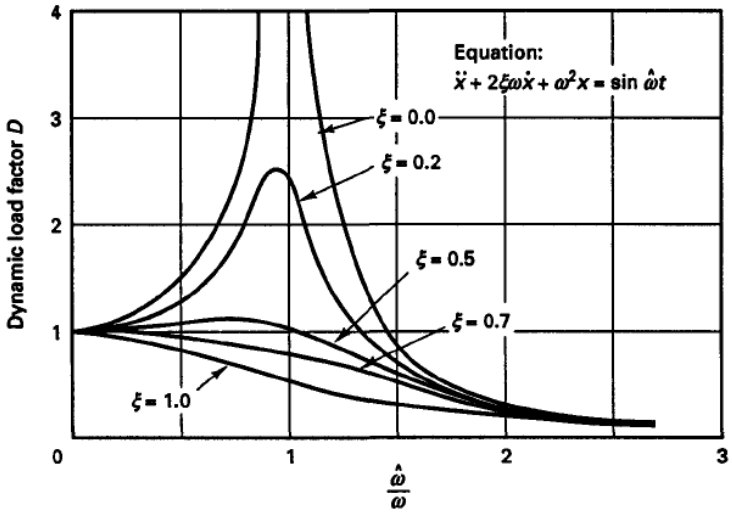


Figure 9.3 The dynamic load factor

3.1.5.2 Modal analysis of $\mathbf{M\ddot{U}} + \mathbf{C\dot{U}} + \mathbf{KU} = 0$ with $\mathbf{C} \neq \mathbf{0}$

- In practice **C** is often **not** compiled from local element matrices.
- This is unlike **M** and **C** matrices.
- In many applications, it is reasonable to actually start from

$$\ddot{x}_i(t) + 2\xi_i\omega_i\dot{x}_i(t) + \omega_i^2 x_i(t) = r_i(t), \qquad \text{where} \quad r_i(t) = \boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{R}(t) \tag{211}$$

- This means that we are assuming $\boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{C}\boldsymbol{\Phi}_j = 0$ for $i \neq j$ and $\boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{C}\boldsymbol{\Phi}_i = 2\xi_i\omega_i$ (no summation on i). That is, mode shapes are C-orthogonal.
- ω_i and $\boldsymbol{\Phi}_i$ are obtained with modal analysis of the undamped $\mathbf{M\ddot{U}} + \mathbf{KU} = 0$. That is, they satisfy $\mathbf{K}\boldsymbol{\Phi}_i = \omega_i^2\mathbf{M}\boldsymbol{\Phi}_i$ (no summation on i).
- FEM solution is again expressed as (191) ($\mathbf{U} = \sum_{i=1}^{n} \boldsymbol{\Phi}_i x_i(t)$) which as discussed before only the first p can be considered in the analysis.
- The main question is how ξ_i are chosen.
- The idea is by activating one damping at a time:
- By imposing ICs in the form $\mathbf{U}^0 = \boldsymbol{\Phi}_i$ and measuring the amplitude decay during the free damped vibration we obtain ξ_i.
- Again it is emphasized, with **C**-orthogonal $\boldsymbol{\Phi}_i$ assumption **C** is not even assembled when modal analysis is used and only (211) equations for x_i are solved.
- If for some reason, the explicit form of **C** is required, *e.g.*, when (174) ($\mathbf{M\ddot{U}} + \mathbf{C\dot{U}} + \mathbf{KU} = \mathbf{R}$) is numerically integrated in time by explicit or implicit methods, we can form **C** by Cauchy series,

$$\mathbf{C} = \mathbf{M}\sum_{k=0}^{r-1} a_k[\mathbf{M}^{-1}\mathbf{K}]^k, \qquad \text{where } a_k \text{ are solved from } r \text{ simultaneous equations :} \tag{212a}$$

$$\xi_i = \frac{1}{2}\left(\frac{a_0}{\omega_i} + a_1\omega_i + a_2\omega_i^3 + \cdots + a_{r-1}\omega_i^{2r-3}\right), \qquad i = 0, \cdots, (r-1) \tag{212b}$$

  and r is the number of damping coefficients given to define **C**.
- For $r = 2$ we recover,

$$\mathbf{C} = a_0\mathbf{M} + a_1\mathbf{K} \tag{213a}$$

$$\left.\begin{aligned} \xi_0 &= \tfrac{1}{2}\left(\tfrac{a_0}{\omega_0} + a_1\omega_0\right) \\ \xi_1 &= \tfrac{1}{2}\left(\tfrac{a_0}{\omega_1} + a_1\omega_1\right) \end{aligned}\right\} \quad \Rightarrow \left\{\begin{aligned} a_0 &= 2\omega_0\omega_0\tfrac{\xi_0\omega_1 - \xi_1\omega_0}{\omega_1^2 - \omega_0^2} \\ a_1 &= 2\tfrac{\xi_1\omega_1 - \xi_0\omega_0}{\omega_1^2 - \omega_0^2} \end{aligned}\right. \tag{213b}$$

- This is the Rayleigh damping matrix (156) $\mathbf{C} = \alpha\mathbf{M} + \beta\mathbf{K}$ that we previously derived assuming simple velocity and strain damping related terms where $a_0 = \alpha$ and $a_1 = \beta$.
- In practice if (174) $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ is directly integrated (with explicit or implicit methods) we rarely go beyond two damping terms used in forming $\mathbf{C}$ in (212) because it would make $\mathbf{C}$ a full matrix and inappropriate for solution update.
- Once $\mathbf{C}$ is formed from r chosen ξ_i we can compute other ξ_j by

$$2\xi_i\omega_i = \boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{C}\boldsymbol{\Phi}_i \qquad \Rightarrow \xi_i = \frac{1}{2\omega_i}\boldsymbol{\Phi}_i^{\mathrm{T}}\mathbf{C}\boldsymbol{\Phi}_i \qquad i > r \tag{214}$$

- For the simple Rayleigh damping matrix (213a) (that is for$r = 2$) from (214) we have,

$$\mathbf{C} = a_0\mathbf{M} + a_1\mathbf{K} \quad \Rightarrow \quad \xi_i = \frac{1}{2\omega_i}\boldsymbol{\Phi}_i^{\mathrm{T}}\left(a_0\mathbf{M} + a_1\mathbf{K}\right)\boldsymbol{\Phi}_i \quad \Rightarrow \quad \boxed{\xi_i = \frac{a_0}{2\omega_i} + \frac{a_1\omega_i}{2} \quad i > 2} \tag{215}$$

  For the small ω_i we have the mass proportional branch $a_0 = \alpha$ and for large ω_i we have the stiffness proportional branch.
- To conclude, there are cases that the assumption that $\mathbf{C}$ is proportional to factors of $\mathbf{M}$ and $\mathbf{K}$ is not reasonable (212).
- For example, the foundation of a building has significantly higher damping coefficients that the the surface structure.
- For each part of the structure with a given material its own specific coefficients may be used (*e.g.,* α, β for Rayleigh model) resulting in a modal-nondiagonalizable $\mathbf{C}$.
- One specific approach for decoupling $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ for modal-nondiagonalizable $\mathbf{C}$ is by seeking eigenmodes in the form $\mathbf{U} = \tilde{\mathbf{U}}e^{\imath\omega t}$.
- For further discussion on this topic refer to [Bathe, 2006] section 9.3.3.
- Examples 9.9 and 9.10 from [Bathe, 2006] provide an example of this process.

3.1.5.3 Example on the calibration and use of Rayleigh damping matrix

EXAMPLE 9.9: Assume that for a multiple degree of freedom system $\omega_1 = 2$ and $\omega_2 = 3$, and that in those two modes we require 2 percent and 10 percent critical damping, respectively; i.e., we require $\xi_1 = 0.02$ and $\xi_2 = 0.10$. Establish the constants α and β for Rayleigh damping in order that a direct step-by-step integration can be carried out.

In Rayleigh damping we have

$$\mathbf{C} = \alpha\mathbf{M} + \beta\mathbf{K} \tag{a}$$

But using the relation in (9.53) we obtain, using (a),

$$\boldsymbol{\phi}_i^T(\alpha\mathbf{M} + \beta\mathbf{K})\boldsymbol{\phi}_i = 2\omega_i\xi_i$$

or

$$\alpha + \beta\omega_i^2 = 2\omega_i\xi_i \tag{b}$$

Using this relation for ω_1, ξ_1 and ω_2, ξ_2, we obtain two equations for α and β,

$$\begin{aligned} \alpha + 4\beta &= 0.08 \\ \alpha + 9\beta &= 0.60 \end{aligned} \tag{c}$$

The solution of (c) is $\alpha = -0.336$ and $\beta = 0.104$. Thus, the damping matrix to be used is

$$\mathbf{C} = -0.336\mathbf{M} + 0.104\mathbf{K} \tag{d}$$

With the damping matrix given, we can now establish the damping ratio that is specified at any value of ω_i, when the Rayleigh damping matrix in (d) is used. Namely, the relation in (b) gives

$$\xi_i = \frac{-0.336 + 0.104\omega_i^2}{2\omega_i}$$

for all values of ω_i.

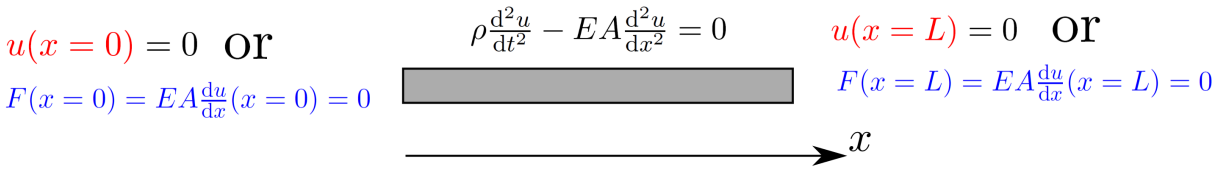
3.1.6 Continuum (exact) natural frequencies and modes

- We discussed how to obtain natural frequencies ω_i and modes $\boldsymbol{\Phi}_i$ for the discretized system $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = 0$ (if $\mathbf{C} = 0$ or natural modes are $\mathbf{C}$-orthogonal).
- That is, if we discretize the system with an n dof system there will be n natural frequency, natural mode pairs $(\omega_i, \boldsymbol{\Phi}_i)$. $\Rightarrow$
- As we have a system with more dofs we obtain more natural frequency / mode pairs.
- These in fact approximate the continuum level natural frequencies / modes
- Below contrasts continuum and discrete modal analysis:

System Type	Basis for natural mode analysis	Number of natural modes/frequencies
Continuum	PDE, *e.g.*, $\rho A\frac{\mathrm{d}^2u}{\mathrm{d}t^2} - EA\frac{\mathrm{d}^2u}{\mathrm{d}x^2} = 0$ for 1D elastic bar	∞
Discrete	ODE, $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = 0$	finite n

- From here for clarity continuum natural modes and frequencies will be undecorated (ω_i, Φ_i) and those obtained from the discrete system $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = 0$ are decorated by $(^h.)$ denoting them being discrete $(\omega_i^h, \boldsymbol{\Phi}_i^h)$.

3.1.6.1 Example: Continuum (exact) natural frequencies and modes of a 1D bar



- Consider the 1D bar example,

$$\rho A\frac{\mathrm{d}^2u}{\mathrm{d}t^2} - EA\frac{\mathrm{d}^2u}{\mathrm{d}x^2} = 0 \tag{216}$$

- On each of the end points 0 or L either **homogeneous** essential or natural BC is applied for modal analysis.
- **In modal analysis BCs and source terms are zero**.
- This is because the modal solutions are **increments** to an existing setup with possibly nonzero BCs and source terms.
- This way, we are looking for the **free vibration** of the structure possibly on top an already nonzero displacement field caused by the source terms and nonzero BCs.
- We solve the problem for fixed (u(0) = 0) - fixed (u(L) = 0) BCs:
- We can decompose the solution in the form,

$$u(x,t) = \Phi(x)T(t) \tag{217}$$

where ω is a natural mode.

- By plugging (217) in (216) we obtain,

$$\rho A\Phi(x)T''(t) - EA\Phi''(x)T(t) = 0 \qquad \Rightarrow \qquad c^2\frac{\Phi''(x)}{\Phi(x)} = \frac{T''(t)}{T(t)} = \alpha, \qquad \text{for} \quad c = \sqrt{\frac{E}{\rho}} \tag{218}$$

- The constant α is obtained given that it should be function of t only and x only at the same time.
- If $\alpha > 0$ that is $\alpha = \omega^2$ for some frequency (based on dimensions of the quantities) then $\Phi(x)$ takes the form $\Phi(x) = A\sinh(\omega x) + B\cosh(\omega x)$ which results in trivial $A = 0$ and $B = 0$ by requiring $u(x = 0, t) = 0$ and $u(x = L, t) = 0$.
- So we choose $\alpha \leq 0$ that is $\alpha = -\omega^2$.
- Clearly, from (218) the temporal solution becomes,

$$T(t) = \sin(\omega t + \theta_0) \qquad \text{where } \theta \text{ is a phase angle.} \tag{219}$$

- Then it is clear from (219) that ω is in fact the natural frequency of the vibration.
- To obtain the set of natural frequencies we refer to (218) and express spatial solution as,

$$\Phi(x) = A\sin\left(\frac{\omega x}{c}\right) + B\cos\left(\frac{\omega x}{c}\right) \tag{220}$$

- Depending on the BCs we find natural frequencies such that the solution is nonzero yet the homogeneous BCs are satisfied.
- For example, for the fixed-fixed condition assumed we have,

$$\left.\begin{aligned} u(x=0,t)=0 \\ u(x=L,t)=0 \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \Phi(x=0)=0 \\ \Phi(x=L)=0 \end{aligned}\right\} \Rightarrow \left.\begin{aligned} B=0 \\ A\sin\left(\tfrac{\omega L}{c}\right)+B\cos\left(\tfrac{\omega L}{c}\right)=0 \end{aligned}\right\}$$

which has the nontrivial solution ($A \neq 0$) if,

$$\sin\left(\frac{\omega L}{c}\right) = 0 \qquad \Rightarrow \frac{\omega L}{c} = n\pi, n > 0 \tag{221}$$

- That is, $\omega = n\pi\frac{c}{L}$ is a natural frequency for any natural number.
- We index natural frequencies and modes by n. The full solution is given by,

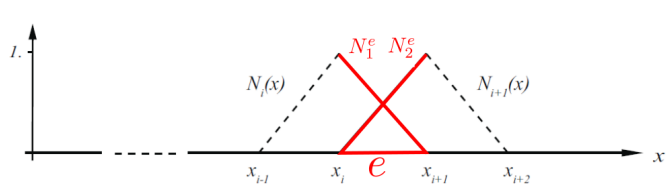
Natural frequency ω_n	Natural mode (Φ_n)	Temporal function $T_n(t)$	$u_n(x,t)$
$\omega_n = n\pi\frac{c}{L}$	$\sin\left(\frac{\omega_n L}{c}\right)$	$\sin(\omega_n t + \phi_{0_n})$	$\sin\left(\frac{\omega_n L}{c}\right)\sin(\omega_n t + \phi_{0_n})$

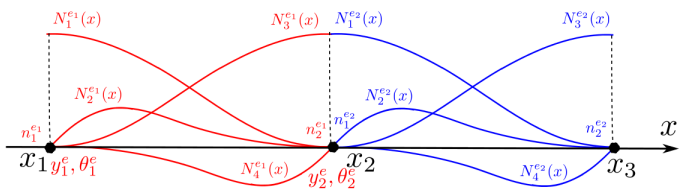
- The same analysis can be applied to other types of BC to obtain natural modes.
- For other problems a similar analysis can be followed. For example, For the beam problem $\rho A\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} - EI\frac{\mathrm{d}^4 y}{\mathrm{d}x^4} = 0$ we have a 4^{th} order in space and eigen modes will be build from $\sin\bar{x}, \cos\bar{x}, \sinh\bar{x}, \cosh\bar{x}$ for $\bar{x} = \frac{x}{L}$ and $\bar{L} = \sqrt[4]{\frac{EI}{\rho A\omega^2}}$. That is, $\Phi(x) = A_1\sin\bar{x} + A_2\cos\bar{x} + A_3\sinh\bar{x} + A_4\cosh\bar{x}$. Depending on the problem setup, we have essential or natural BC for either displacement y or rotation $\theta = \frac{\mathrm{d}y}{\mathrm{d}x}$ at either of the 4 end points. The natural modes will be obtained such that the homogeneous 4×4 system (4 BCs for 4 unknowns A_1 to A_4) has nonzero solutions for $\mathbf{A} = [A_1\ A_2\ A_3\ A_4]^{\mathrm{T}}$; *i.e.*, $|\mathbf{A}| \neq 0$ and $\Phi(x)$ is not trivially zero.

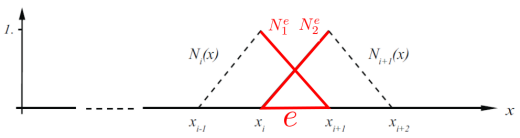
3.1.7 Error analysis for natural frequencies and natural modes

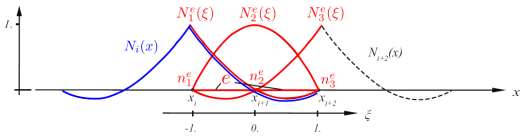
- If the differential equation has $2m$ highest spatial derivative, shape functions must be globally $\mathcal{C}^{m-1}$ continuous.
- Below, two cases for bar and beam examples are shown:

-	Bar	Beam
PDE	$\rho A\frac{\mathrm{d}^2 u}{\mathrm{d}t^2} - EA\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = 0$	$\rho A\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} - EI\frac{\mathrm{d}^4 y}{\mathrm{d}x^4} = 0$
$2m$	2	4
Global continuity $\mathcal{C}^{m-1}$	0	1









3.1.7.1 Preliminaries: FEM polynomial order p

-	Bar (1^{st}order)	Bar (2^{nd})
Sample shape function	$N_1 = 1 - \frac{x^1}{L}$	$N_1 = 1 - \frac{3}{2}x + \frac{1}{2}x^2$
Maximum element order p	1	2

- Note that the element maximum polynomial order p is not the same as minimum global required continuity $m-1$.
- For example, in the figure both elements are for the bar element with $m-1=0$ ($\mathcal{C}^0$ global continuity).
- Yet, the element on the left is 0^{th} order $(p=0)$ and on the right 1^{st}order $(p=1)$.

3.1.7.2 A priori error estimates for natural frequencies and modes

- *A priori* error estimates for natural frequencies and natural modes are in the form,

$$0 \leq \omega_i^h - \omega_i \leq Ch^{2(p+1-m)}\omega_i^{\frac{2p+2-m}{m}} \tag{222a}$$

$$||\Phi_i^h - \Phi_i||_m \leq Ch^{(p+1-m)}\omega_i^{\frac{p+1}{m}} \tag{222b}$$

grid resolution h = the largest element size (size of an element is the radius of its circumscribing circle (2D) / sphere (3D))

1. $0 \leq \omega_i^h - \omega_i$, *i.e.,* having $\omega_i \leq \omega_i^h$ is not preserved once the Galerkin rules are violated [Hughes, 2012] (e.g., when reduced integration or incompatible modes are employed or when lumped mass matrix is used).
2. The rate of convergence (*i.e.,* power of h) of eigenvalues is twice that of eigenfunctions in the H^m (Hilbert m norm) [compare (222a) and (222b)]. That is,

   Natural frequencies converge twice faster than natural modes

3. The appearance of powers of the natural frequencies on the right-hand sides of (222a) $\omega_i^{\frac{2p+2-m}{m}}$ and (222b) $\omega_i^{\frac{p+1}{m}}$ suggests that the **quality of approximation deteriorates for higher modes**. Recall that $\omega_0 < \omega_1 < \cdots < \omega_n$. This can be explained that higher modes have higher spatial variability (wave number) and for the same resolution of FEM mesh h it is more difficult to capture the exact solution.
4. **K**, **M** (and **C**) are often integrated numerically, *i.e.,* by quadrature.
   (a) For the convergence rates in h in (222) to hold:

   The quadrature rule must be accurate enough to exactly integrate all monomials through order $\bar{p}+p-2m$ where

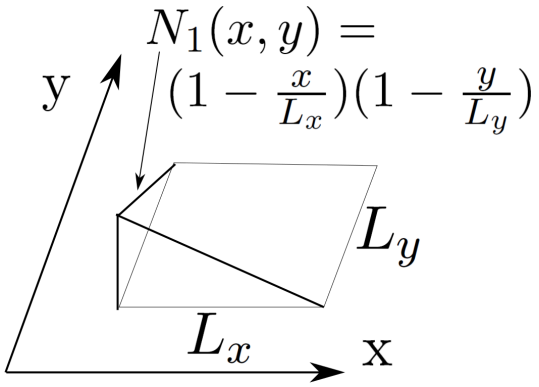
   $\bar{p}$ = Order of the highest-order monomial appearing in the element shape functions,
   p = Order of the element
   $m-1$ = Level of global continuity of FEM shape functions

   (b) A sufficient condition for the convergence of modal quantities (as $h \to 0$) is
   The quadrature rule must be accurate enough to exactly integrate all monomials through order $\bar{p}-m$ (a weaker condition that having the full convergence rates)

   If reduced order quadrature is used, we need to enforce condition (a) above to ensure preserving the maximum convergence rates for modal quantities and enforce condition (b) to ensure convergence of modal quantities to exact values as $h \to 0$ (the minimum requirement)
   This becomes critically important in the formation of lumped mass matrices where at times special quadrature rules, *e.g.,* Lobatto quadrature (it is similar to Gauss quadrature but maintains the end point values of the interval as quadrature points) need to be used to preserve maximum modal quantity convergence rates.

Example for evaluating p and $\bar{p}$ from 2D bilinear finite element:

- For this bilinear element we observe the highest monomial order in shape functions is two. For example, N_1 has a term $\frac{xy}{L_x L_y}$ $\bar{p} = 2$.
- At the same time, this element is considered linear $p = 1$ because the highest complete polynomial space covered (for this given Cartesian geometry) is 1st. For example for it to be second order it should have terms like x^2y^2 which the shape functions of this element do not have such monomials.



- Estimate (222) is presented in terms of natural frequencies ω_i^h which are the squares of the generalized eigenvalue problem $\mathbf{K}\boldsymbol{\Phi}_i^h = \lambda_i^h \mathbf{M}\boldsymbol{\Phi}_i^h$ (no summation on i) where $\lambda_i^h = (\omega_i^h)^2$.
- The general form of errors for eigenvalue eigenmode solutions in FEM is of the following form,

$$0 \leq \lambda_i^h - \lambda_i \leq Ch^{2(p+1-m)} \lambda_i^{\frac{p+1}{m}} \tag{223a}$$

$$||\Phi_i^h - \Phi_i||_m \leq Ch^{(p+1-m)} \lambda_i^{\frac{p+1}{2m}} \tag{223b}$$

- These eigenvalue / eigenmode error estimates are not only useful for natural mode error analysis, as cast in (222) form, but also find application in buckling analysis and other FEM problems that require eigen analysis.
- There is also an L2 norm (rather than H^m norm estimates for eigenmodes:

$$||\Phi_i^h - \Phi_i||_0 \leq Ch^{\sigma} \lambda_i^{\frac{p+1}{2m}}, \qquad \sigma = \min(p+1, 2(p+1-m)) \tag{224}$$

3.1.7.3 Examples for modal analysis error estimates

Example 1 Consider an elastic boundary value problem (m = 1) and assume linear elements are employed (k = 1). The errors estimates take the form:

$$\frac{\omega_l^h}{\omega_l} - 1 = O(h^2)$$

$$\|\boldsymbol{u}_{(l)}^h - \boldsymbol{u}_{(l)}\|_1 = O(h)$$

For quadratic-level elements ($k = 2$) we have

$$\frac{\omega_l^h}{\omega_l} - 1 = O(h^4)$$

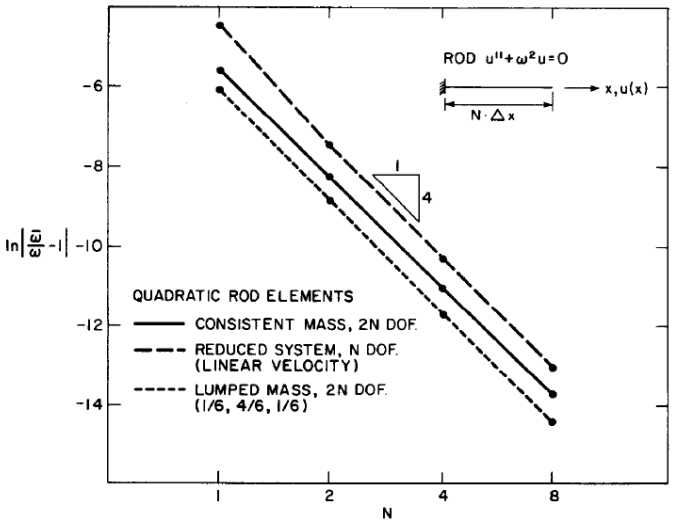
$$\|\boldsymbol{u}_{(l)}^h - \boldsymbol{u}_{(l)}\|_1 = O(h^2)$$

Example 2

Consider the Hermite cubic beam element ($m = 2$, $k = 3$):

$$\frac{\omega_l^h}{\omega_l} - 1 = O(h^4)$$

$$\|\boldsymbol{u}_{(l)}^h - \boldsymbol{u}_{(l)}\|_1 = O(h^2)$$

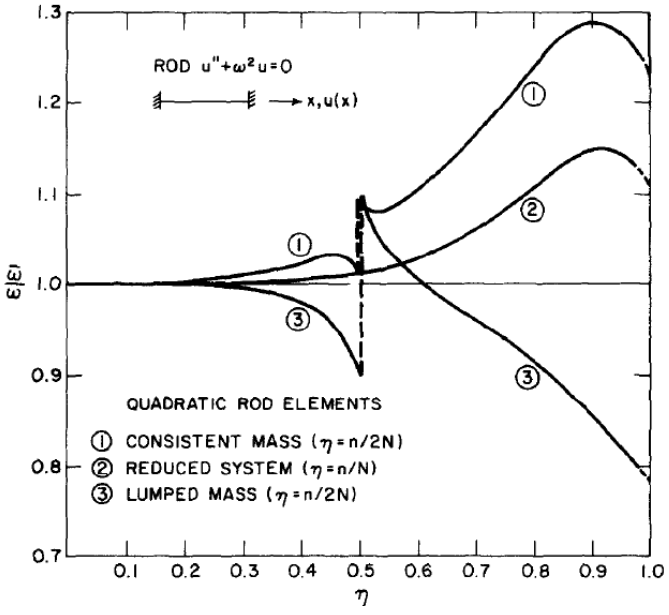


Convergence of fundamental (first) natural frequency for a quadratic 1D bar problem

- The plot demonstrate the convergence rate for the fundamental natural frequency of a bar problem [Hughes et al., 1976a].
- From Example 1, the convergence rate must be 4 if all the requirements are satisfied.
- The reduced integration order scheme still exactly integrates monomials of degree $\bar{p}+p-2m = 2+2-2\times 1 = 2$. Note that, in the full integration scheme, mass matrix $\mathbf{M}$ is integrated with fourth order quadrature because shape function are second order and the product of them is integrated in the mass matrix.
- We observe that by using lumped mass matrix we violate the Galerkin consistency required for the condition $\omega_i \leq \omega_i^h$ from (222b) ($i = 1$ for the fundamental mode).
- The lump mass matrix option also preserves the optimal convergence rate. It uses the element level mass matrix (169),

$$\mathbf{M}^e = \frac{m^e}{6}\text{diag}(1,4,1)$$

which again is based on third order accurate Simpson rule which is more accurate than 2^{nd}order required from $\bar{p}+p-2m = 2+2-2\times 1 = 2$.



- The plot demonstrate how the accuracy is modal frequencies decreases as higher mode frequencies are computed and compared with analytical ones.
- n is the mode number in the figure.

- N is the number of elements in the figure.
- As $\eta \propto \frac{n}{N}$ increases, *i.e.*, higher modes RELATIVE to the element size $h \propto 1/N$ are considered the error starts to increase.
- That is, the error is not merely a function of which natural mode is considered, but more on how accurately an element can approximate spatial variability (wave number) of a given natural mode.
- Again the violation of $\omega_i \leq \omega_i^h$ from (222b) for lumped mass matrix option is due to its violation of Galerkin methods required for of $\omega_i \leq \omega_i^h$ in (222b).
- Interestingly, we observe that the reduced integration order (of the mass matrix) performs better than full integration order with consistent mass option.
- The sharp jump of the error at $\eta = 0.5$ signals quick deterioration of calculating high natural frequencies (relative to the number of elements). This is one point of concern, and one of the advantages of isogeometric FEMs (a relatively new FE method) is having a much better performance in solving high natural frequencies (relative to the number of elements).
- To conclude figure below compares axial and flexural frequencies using different mass matrix options:
  - While relative error in the 1D bar problem is directly related to how many wave numbers an element can model, the absolute error quickly increases as mode number increases.
  - Consistent and lumped mass matrices option again provides higher and lower frequencies than exact ones while an average between two can provide both higher and lower values.
  - The errors in flexural natural frequencies are higher than axial ones for the same number of elements.
  - The unavailable natural frequency for beam problem with $n = 10$ and lumped mass matrix is that for this particular computational set-up natural modes beyond 4 are infinite, implying care that must be taken in using lumped mass matrix in practice.

[Cook, 2002]

TABLE 11.3-1. PERCENTAGE ERRORS OF COMPUTED NATURAL FREQUENCIES, USING DIFFERENT MASS MATRICES [11.7]. FOR BEAM ELEMENTS WITH PARTICLE-MASS LUMPING, $\alpha = 0$ IN EQ. 11.3-3. STRUCTURES ARE UNIFORM AND MODELED BY ELEMENTS OF EQUAL LENGTH.

Mode number	Type of mass matrix used		
	Particle-mass lumps (%)	Average [**m**] (%)	Consistent [**m**] (%)
	Axial vibration of an eight-element bar, one end fixed, the other free		
1	−0.16	0.00	+0.16
2	−1.44	−0.03	+1.45
3	−3.97	−0.20	+4.05
4	−7.69	−0.79	+7.92
8	−32.42	−17.43	+15.94
	Flexural vibration of a five-element cantilever beam		
1	−1.80	−0.91	0.00
2	−5.90	−3.07	+0.05
3	−9.31	−5.03	+0.36
4	−13.62	−7.69	+1.17
10	Unavailable	+91.77	+67.83

4 Overview of time matching schemes

4.1 Introduction to time marching schemes

- Our interest in this section is solving **ODEs in time**.
- Consider the following system of equations obtained by FEM discretization ((174))

$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R} \tag{225}$$

- This equation is an ODE in time.
- Spatial derivative terms are already eliminated by using FE discretization in space.
- In this section we discuss methods by which we can solve first and second order temporal ODEs, *i.e.,* ODEs with initial conditions.
- There are different aspects for which we want to classify the solution of (225):

1. Hyperbolic vs. Parabolic: FE discretization of a hyperbolic and parabolic problems are often reduced to the following ODEs,

$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R} \quad \text{Hyperbolic} \quad \text{Example: Elastodynamics } \mathbf{M} = \text{ mass, } \mathbf{C} = \text{ damping, } \mathbf{K} = \text{ stiffness matrices} \tag{226a}$$

$$\mathbf{M}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R} \quad \text{Parabolic} \quad \text{Example: Heat equation } \mathbf{M} = \text{ capacity, } \mathbf{K} = \text{ conductivity matrices} \tag{226b}$$

Note that any second order (or n^{th} order) ODE can be written in the form (226b) and the form itself does not directly imply if the underlying PDE is hyperbolic or parabolic. For example, we can express (226a) as,

$$\dot{\mathbf{U}} - \mathbf{V} = 0 \tag{227a}$$

$$\mathbf{M}\dot{\mathbf{V}} + (\mathbf{K}\mathbf{U} + \mathbf{C}\mathbf{V}) = \mathbf{R} \tag{227b}$$

where $\mathbf{V}$ represents the temporal derivative of $\mathbf{U}$, *i.e.,* velocity when $\mathbf{U}$ is displacement. We can express this in the form (226b):

$$\tilde{\mathbf{M}}\dot{\tilde{\mathbf{U}}} + \tilde{\mathbf{K}}\tilde{\mathbf{U}} = \tilde{\mathbf{R}}, \text{where} \tag{228a}$$

$$\tilde{\mathbf{U}} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}, \quad \tilde{\mathbf{M}} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ 0 & \mathbf{M} \end{bmatrix} \quad \tilde{\mathbf{K}} = \begin{bmatrix} \mathbf{0} & \mathbf{-1} \\ \mathbf{K} & \mathbf{C} \end{bmatrix} \quad \tilde{\mathbf{R}} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix} \tag{228b}$$

thus

- Having only one or two temporal derivatives on itself does not imply whether the underlying PDE is hyperbolic or PDE.
- Whether the underlying equation is hyperbolic or parabolic results in how the (smallest) element frequency scales versus its size.
  - Hyperbolic PDEs: $\omega^h \propto h \quad \Rightarrow \quad \Delta t \propto \frac{1}{h_{him}}$
  - Parabolic PDEs: $\omega^h \propto h^2 \quad \Rightarrow \quad \Delta t \propto \frac{1}{h^2_{him}}$

2. Single-degree-of-freedom vs. Multi-degree-of-freedom (SDOF vs. MDOF)

- For the temporal solution of PDEs we use a variety of different time marching schemes.
- As we observed from (226) they are often expressed in the form $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ or $\mathbf{M}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$.
- That is they are ODEs after FEM discretization.
- Similar to modal superposition from §3.1 for hyperbolic case ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$), we can reduce the general equation to n SDOF equations of the form:

$$\ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = f(t) \quad \text{corresponding to} \quad \mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R} \tag{229a}$$

$$\dot{x} + \lambda x = f(t) \quad \text{corresponding to} \quad \mathbf{M}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R} \tag{229b}$$

- A variety of time marching schemes can be directly applied to MDOF (226) or to their individual SDOF (188).

- In the next section, §5, we will use the SDOF form (188) for stability analysis of MDOF (226).

3. Single-step vs. Multi-step:
   - Single-step: only values from t_n are needed to obtain the solution for t_{n+1}.
   - Multi-step: values of t_n, t_{n-1}, and possibly more, are required to obtain the solution for t_{n+1}.
4. Single-stage vs. Multi-stage
   - Single-stage: Single We directly compute t_{n+1} from t_n (and possibly prior values) in one stage.
   - Multi-stage: To obtain t_{n+1} from previous solutions several intermediate values are computed from t_n to t_{n+1}. Runge-Kutta (RK) methods are Multi-stage methods.
5. Global versus local time step size: In explicit methods (below) smallest elements pose serious challenges in the time step of the entire domain. It is desirable to use smaller time step for smaller elements than larger ones. The same concept, but only from accuracy perspective and not for stability reasons, becomes relevant in implicit methods. The flexibility of a time marching scheme in this respect increases in the following order:
   - Global time step: All elements, small or large, share the same time step $\Delta t = t_{n+1} - t_n$.
   - Subcycling: Smaller elements take smaller time step, often in factors of 2^s ($\Delta t/2^s$) time smaller than the global time step $\Delta t = t_{n+1} - t_n$. Although elements have different time steps, at the end of the time step all have the same time value t_{n+1}. That, is from global time step to time step the method is synchronous.
   - Local Time Stepping (LTS) (asynchronous subcycling): Although LTS is also used for subcycling approaches, in general it is used for asynchronous time stepping schemes that each element takes a local time step (based on its stability limits in explicit methods) and element final times do not need to be synchronized.

   LTS/ subcycling approaches, along with IMEX schemes, will later be discussed as approaches to solve highly graded FE meshes and/or still problems where some operators of the PDE requires very small time steps with explicit methods.
6. Stability:
   - Unconditional unstable: The time marching scheme is unstable for any Δt. Clearly, such time integrators will not be used at all!
   - Conditional stable: Δt must satisfy $\Delta t_{\max} \leq \Delta t_{\max}$ (or more generally satisfy special conditions) for the solution to be stable.
   - Unconditional stable: For an underlying physical PDE, the numerical scheme is stable for any Δt.
7. Order of convergence:
   - For a method that is convergent we need to know how fast the numerical solution converges to the exact solution of the underlying problem. For example, the error between the two at a given time can temporally converge as Δt^q. We also have a spatial order of convergence h^p based on the order of of elements used and the type of error considered. The concept of order of convergence is closely related to "consistency", "local truncation error", and "convergence" discussed below.
   - In general, achieving high temporal convergence rates is much more difficult than spatial ones, as FE method can easily accommodate any spatial order of accuracy, but often FD type temporal update makes it challenging to achieve high temporal orders of accuracy.
8. Explicit vs. Implicit
   - Explicit: Refers to being able to solve solution for t_{n+1} "explicitly". Some aspects are:
     - No global matrix equation or nonlinear solution needs to be solved. That is,
     - Even if the underlying PDE is nonlinear, update equations will typically be linear.
     - Are often only conditionally stable.
     - In time stepping methods, explicit methods are often expressed for t_n, have only $\mathbf{M}$ (and $\mathbf{C}$) on the LHS of the update equation. If $\mathbf{C} = 0$ and mass matrix $\mathbf{M}$ is diagonal (or capacity diagonal in heat equation) is used the update equation is trivial and no global equations should be solved. In fact, if the solution cannot be reduced to a local and small matrix update equation, many still do not call the scheme explicit!. In these course notes, we label a larger group of problems explicit, basically by referring to schemes that write the equations at t_n rather than at a later time.
     - To enable local and no global matrix solution strategy and other reasons discussed later explicit methods are often matched with diagonal "mass" matrices.

- Implicit: Refers to being to solve solution for t_{n+1} "implicitly". Some aspects are:
  - Involve global matrix equations. Also,
  - If the underlying PDE is nonlinear, update equations will be nonlinear.
  - Are often unconditionally stable. The rare ones that are conditionally stable are not used. There are different levels and types of stability which we do not cover all in this course. But in general, the implicit methods in practice are stable for much wider ranges and stages of time step and PDE coefficient.
  - In time stepping methods, explicit methods are often expressed for t_{n+1}, have all $\mathbf{M}$, $\mathbf{K}$ (and $\mathbf{C}$) on the LHS of the update equation. Even for linear PDEs they require the solution of a full matrix equation for the solution of $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$.
  - For considerations that will be discussed later implicit methods are often matched with consistent "mass" matrices.
- Implicit-Explicit (IMEX): These are the schemes that use both explicit and implicit integration schemes, in one of both of the following modes:
  - Domain IMEX: Some parts of the domain are solved with explicit solver some parts with implicit solver. For example regions with small elements use implicit solver and large elements explicit solver. Another example, is using implicit solver for solid part and explicit method for a fluid part.
  - Operator IMEX: Example is using implicit integrators for more stringent operators such as parabolic ones and explicit integrators otherwise, *e.g.,* hyperbolic operator. Same concept can be applied to stiff PDEs where the stiff operators are integrated implicitly.

Two concepts that are closely related to the order of convergence above are loosely defined below:

- Consistency
  - Consistency refers to the concept that the update from step t_n (and previous ones in multi-step methods) to t_{n+1} is "consistent" with an underlying update of the exact solution from t_n to t_{n+1} and the local truncation error, *i.e.,* error of computational versus exact values for t_n, is $\mathcal{O}\left(\Delta t^p\right)$, $p > 1$.
- Convergence and rate of convergence
  - Convergence: The numerical method convergence to the exact solution as $\Delta t \to 0$, *i.e.,* $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ is exactly integrated. This does not mean that we converge to the underlying PDE solution. For that element size h must also approach zero.
  - Convergence rate: The rate in which the numerical solution convergences to the exact solution in terms of Δt^p.

In this section, we review various time marching schemes and in the following section we present their stability and convergence analysis.

4.2 A One-step single-field time stepping method: Generalized trapezoidal rule

- We consider the solution of a first order ODE of the form:

$$\mathbf{M}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F} \qquad \text{Temporal ODE (after FEM spatial discretization)} \tag{230a}$$

$$\mathbf{d}(t=0) = \mathbf{d}_0 \qquad \text{Initial Condition (IC)} \tag{230b}$$

  this can have been derived from the discretization of a parabolic equation (226b) or two-field representation of a second order hyperbolic equation (228).

- The update equation for the time $t = t_n + \alpha\Delta t$ is written as,

$$\dot{\mathbf{d}}^{t_n+\alpha\Delta t} = \frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \tag{231a}$$

$$\mathbf{d}^{t_n+\alpha\Delta t} = (1-\alpha)\mathbf{d}^n + \alpha\mathbf{d}^{n+1} \tag{231b}$$

- That is $\mathbf{d}^{t_n+\alpha\Delta t}$ is linearly interpolated from end point values at t_n and t_{n+1}.
- α-method includes three different FD update equations below as special cases:

α	Method	Explicit/Implicit	Temporal Order
0	Forward differences; forward Euler	Explicit	1
$\frac{1}{2}$	Trapezoidal rule; midpoint rule; Crank-Nicolson	Implicit	2
1	Backward differences; backward Euler	Implicit	1

- We satisfy (230a) at $t = t_n + \alpha\Delta t$:

$$\mathbf{M}\dot{\mathbf{d}}^{t_n+\alpha\Delta t} + \mathbf{K}\mathbf{d}^{t_n+\alpha\Delta t} = \mathbf{F}^{t_n+\alpha\Delta t} \tag{232}$$

where

$$\mathbf{F}^{t_n+\alpha\Delta t} = (1-\alpha)\mathbf{F}^n + \alpha\mathbf{F}^{n+1} \tag{233}$$

- Plugging (231) and (233) in (232) we obtain,

$$\tilde{\mathbf{M}}\mathbf{d}^{n+1} = \tilde{\mathbf{F}} \qquad \text{where} \tag{234a}$$

$$\tilde{\mathbf{M}} = \mathbf{M} + \alpha\Delta t\mathbf{K} \tag{234b}$$

$$\tilde{\mathbf{F}} = (\mathbf{M} - (1-\alpha)\Delta t\mathbf{K})\,\mathbf{d}^n + \Delta t\left((1-\alpha)\mathbf{F}^n + \alpha\mathbf{F}^{n+1}\right) \tag{234c}$$

- For a SDOF version of (230a), that is (229b) we have

$$\dot{d} + \lambda d = f(t) \tag{235}$$

- that is $\mathbf{M} = 1, \mathbf{K} = \lambda$, so the update can be written as,

$$d^{n+1} = Ad^n + L^n \tag{236a}$$

$$A = \frac{1-(1-\alpha)\Delta t\lambda}{1+\alpha\Delta t\lambda} \qquad \text{Amplification factor} \tag{236b}$$

$$L^n = \Delta t\frac{(1-\alpha)f^n + \alpha f^{n+1}}{1+\alpha\Delta t\lambda} \tag{236c}$$

- To start stability analysis of α-method we first obtain the exact solution to (235) is,

$$d(t_n) = d_0 e^{-\lambda t_n} \qquad \text{Exact solution}$$

which physically is stable for $\lambda \geq 0$ as,

$$\begin{cases} |d(t_{(n)}| \to 0, & \text{since} \quad d(t_n) = d_0 e^{-\lambda t_n} \quad \lambda > 0 \\ d(t_{(n)} = d_0 & \lambda = 0 \end{cases}$$

- For stability analysis of α-method in (236a) ($d^{n+1} = Ad^n + L^n$), consider the case that $f(t) = 0 \Rightarrow L^n = 0$. From (236a) we obtain

$$d^n = A^n d_0 \qquad \text{Solution from } \alpha\text{-method} \tag{237}$$

- Since the physical problem is only stable for $\lambda \geq 0$, we only consider stability limit of α-method for $\lambda \geq 0$.
- It is clear for (306) not to blow up we need to have,

$$|A| \leq 1 \qquad \Rightarrow \qquad -1 \leq \frac{1-(1-\alpha)\Delta t\lambda}{1+\alpha\Delta t\lambda} \leq 1 \tag{238}$$

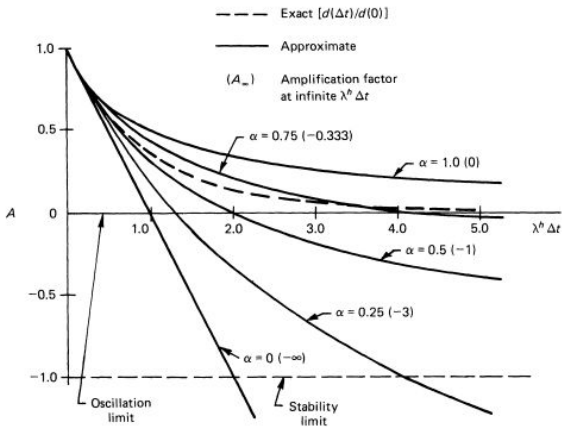
which results in the following conditions:

$$\begin{cases} \alpha < \frac{1}{2} & \text{Conditionally stable} \qquad \Delta t\lambda < \frac{2}{1-2\alpha} \\ \alpha \geq \frac{1}{2} & \text{Unconditionally stable} \end{cases} \tag{239}$$

- Some sample amplification factors for n steps are shown,

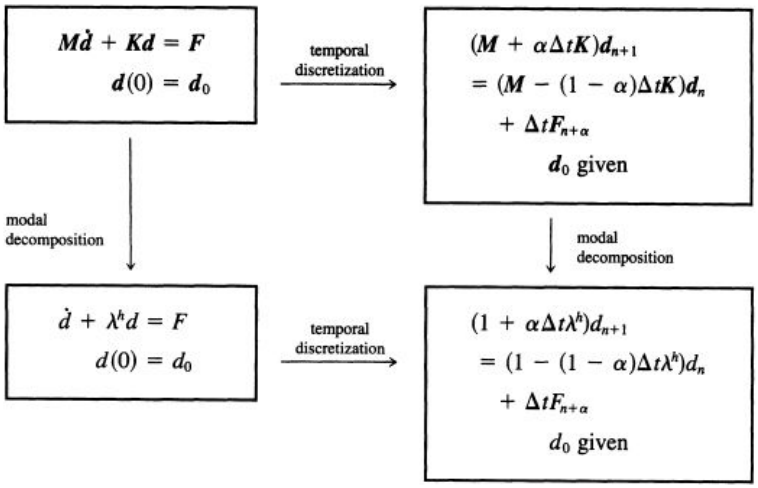
A \ n	100	1000
.99	.37	4.32×10^{-5}
1.01	2.70	2.09×10^{4}
.9	2.66×10^{-5}	1.75×10^{-46}
1.1	1.39×10^{4}	2.47×10^{41}

- Also figure below demonstrate how for unstable methods amplification factor A becomes less than one, resulting in oscillating and gowning solution.



Amplification factor for typical one-step methods.

- In terms of solving n SDOFs with α-method there are two routes to solve them with SDOF (shown in the figure):
  - We first derive n SDOF equations then solve them with α-method.
  - We first apply α method to (230a) ($\mathbf{M}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}$) then do modal decomposition to n SDOFs.



- For stability of a MDOF the maximum frequency is used in stability analysis as it provides the most stringent time step requirement.
- The stability limits of a MDOF is thus summarized as follows,

Summary: Stability for the generalized trapezoidal methods

Amplification factor: $A = \dfrac{1-(1-\alpha)\Delta t\lambda^h}{1+\alpha\Delta t\lambda^h}$

Stability requirement: $|A| < 1$ for $\lambda^h = \lambda^h_{n_{eq}}$ (= maximum eigenvalue)

Unconditional stability: $\alpha \geq \frac{1}{2}$

Conditional stability: $\alpha < \frac{1}{2}$, $\Delta t < \dfrac{2}{(1-2\alpha)\lambda^h_{n_{eq}}}$

- Other methods discussed in section will be analyzed in a similar manner in §5 to investigate whether they are unconditionally stable or if conditionally stable what their acceptable time step values must be.

4.3 Linear multi-step (LMS) methods

- Consider the first order ODE,

$$\dot{\mathbf{y}} = f(\mathbf{y},t) \qquad \text{General nonlinear first order ODE} \tag{240a}$$
$$= \mathbf{G}\mathbf{y} + \mathbf{H}(t) \qquad \text{Linear first order ODE} \tag{240b}$$

  where

  - $\mathbf{y}$: size n vector of unknowns.
  - $\mathbf{G}$: $n \times n$ constant matrix.
  - $\mathbf{H}$: size n vector of transient load.

- A k-step linear multi-step method is defined as,

$$\sum_{i=0}^{k} \{\alpha_i \mathbf{y}^{n+1-i} + \Delta t \beta_i f(\mathbf{y}^{n+1-i}, t_{n+1-i})\} = 0 \tag{241}$$

  - k: Number of steps that the method goes back from the time step t_{n+1} $\mathbf{y}$ value we want to solve for.
  - α_i and β_i are parameters that define the method. The method
  - Linearity in LMS does not refer to linearity of f, rather to the linear interpolation form in (241).
  - is called explicit if $\beta_0 = 0$. It is otherwise implicit.
  - is called backward-difference if $\beta_i = 0$ for $i \geq 1$.
  - An example of LMS scheme is the 1 step generalized trapezoidal (α) method §4.2:

$$\frac{\mathbf{y}^{n+1}-\mathbf{y}^n}{\Delta t} = f((1-\alpha)\mathbf{y}^n + \alpha\mathbf{y}^{n+1}) \approx (1-\alpha)f(\mathbf{y}^n) + \alpha f(\mathbf{y}^{n+1}) \quad \Rightarrow$$
$$(-1)\mathbf{y}^{n+1} + \alpha\Delta t f(\mathbf{y}^{n+1}) + (1)\mathbf{y}^n + (1-\alpha)\Delta t f(\mathbf{y}^n) = \alpha_0\mathbf{y}^{n+1} + \beta_0\Delta t f(\mathbf{y}^{n+1}) + \alpha_1\mathbf{y}^n + \beta_1\Delta t f(\mathbf{y}^n) = 0$$

  - That is $\alpha_0 = -1, \alpha_1 = 1, \beta_0 = \alpha, \beta_1 = 1-\alpha$ for generalized trapezoidal rule.
  - For $k = 1$ (*e.g.,* generalized trapezoidal rule) the scheme, rightfully, is often called linear single-step method.
  - Reminder: Generalized trapezoidal rule encompasses forward and backward Euler method and trapezoidal method.

- As for 2nd order ODEs, *i.e.,* those arising from structural dynamics, a linear second order ODE takes the form,

$$\ddot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}, \dot{\boldsymbol{y}}, t) = \boldsymbol{G}_0\boldsymbol{y} + \boldsymbol{G}_1\dot{\boldsymbol{y}} + \boldsymbol{H}(t) \tag{242}$$

- A k-step LMS method for linear second order ODE takes the form,

$$\sum_{i=0}^{k} \{\alpha_i \boldsymbol{y}_{n+1-i} + \Delta t\,\beta_i \boldsymbol{G}_1 \boldsymbol{y}_{n+1-i} + \Delta t^2 \gamma_i [\boldsymbol{G}_0 \boldsymbol{y}_{n+1-i} + \boldsymbol{H}(t_{n+1-i})]\} = \boldsymbol{0} \tag{243}$$

- where now $\alpha_i, \beta_i, \gamma_i$, $i \leq k$ are the model parameters. Similar to the first order ODE we have,
  - The method is explicit if $\beta_0 = \gamma_0 = 0$.
  - It is backward difference if $\beta_i = \gamma_i = 0$, $i \geq 1$.
- In the following, we describe a few LMS schemes for the solution of elastodynamics problem.
- We already used a 1-step LMS scheme of generalized trapezoidal rule for the solution of first order ODEs. That for example had direct application in the solution of parabolic PDEs.

4.3.1 Central Difference method for elastodynamics (an explicit LMS method)

- For the equation (226a) ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$) we use central difference approximations for both $\ddot{\mathbf{U}}$ and $\dot{\mathbf{U}}$:

$$
\begin{aligned}
{}^{t}\ddot{\mathbf{U}} &= \frac{1}{\Delta t^2}({}^{t-\Delta t}\mathbf{U} - 2\,{}^{t}\mathbf{U} + {}^{t+\Delta t}\mathbf{U}) \\
{}^{t}\dot{\mathbf{U}} &= \frac{1}{2\Delta t}(-{}^{t-\Delta t}\mathbf{U} + {}^{t+\Delta t}\mathbf{U})
\end{aligned}
\tag{244}
$$

- After plugging (244) in (226a) ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$) for time step t_n we obtain,

$$
\boxed{\left(\frac{1}{\Delta t^2}\mathbf{M} + \frac{1}{2\Delta t}\mathbf{C}\right)} {}^{t+\Delta t}\mathbf{U} = {}^{t}\mathbf{R} - \left(\mathbf{K} - \frac{2}{\Delta t^2}\mathbf{M}\right){}^{t}\mathbf{U} - \left(\frac{1}{\Delta t^2}\mathbf{M} - \frac{1}{2\Delta t}\mathbf{C}\right){}^{t-\Delta t}\mathbf{U}
\tag{245}
$$

- We observe that this is a 2-step LMS scheme by requiring values from t_n, t_{n-1} to obtain $\mathbf{U}$ for t_{n+1}.
- Starting point: Since for the first time step $n = 0$ update, $\mathbf{U}$ for t_{-1} does not exist we need to initialize this value accordingly:

$$
{}^{-\Delta t}U_i = {}^{0}U_i - \Delta t\ {}^{0}\dot{U}_i + \frac{\Delta t^2}{2}{}^{0}\ddot{U}_i
\tag{246}
$$

- Solution strategy from t_{n-1}, t_n to t_{n+1}:
  - From (245) the solution for $\mathbf{U}$ at t_{n+1} requires a linear system solution with matrix coefficient:

$$
\hat{\mathbf{M}} = \frac{1}{\Delta t^2}\mathbf{M} + \frac{1}{2\Delta t}\mathbf{C}, \quad \text{where} \quad \hat{\mathbf{M}}\mathbf{U}^{n+1} = \mathbf{R}^n
\tag{247}
$$

  - If the system is undamped we make the option to have the LHS matrix $\hat{\mathbf{M}} = \frac{1}{\Delta t^2}\mathbf{M}$. If $\mathbf{C} \propto \mathbf{M}$ we still have a similar problem.
  - If besides $\mathbf{C} = 0$ (or it being proportional to $\mathbf{M}$) we have a lumped mass matrix, we **do not need a matrix equation** and update is followed as,

$$
{}^{t+\Delta t}U_i = {}^{t}\hat{R}_i\left(\frac{\Delta t^2}{m_{ii}}\right) \quad \text{for} \quad {}^{t}\hat{\mathbf{R}} = {}^{t}\mathbf{R} - \left(\mathbf{K} - \frac{2}{\Delta t^2}\mathbf{M}\right){}^{t}\mathbf{U} - \left(\frac{1}{\Delta t^2}\mathbf{M}\right){}^{t-\Delta t}\mathbf{U}
\tag{248}
$$

  - Another advantage of lumped mass matrix is that:
    - Lumped mass matrix elongates period of moving waves.
    - Explicit method typically shorten the period of moving waves

    so **matching** explicit integrators and lumped mass matrices to some extend cancels the period error of the numerical method and is preferred from this perspective. On the other hand, if we had used consistent mass matrix that would as well would have shortened the period of moving waves and exaggerate the problem of explicit time integrators.
  - Another important implication of not having $\mathbf{K}$ appearing in $\hat{\mathbf{M}}$ is that we **do not need to** actually assemble $\mathbf{K}$.
  - We can directly add contributions from stiffness to the global force vector $\mathbf{R}$ at the element level:

$$
\mathbf{K}\,{}^{t}\mathbf{U} = \sum_i \mathbf{K}^{(i)}\,{}^{t}\mathbf{U} = \sum_i {}^{t}\mathbf{F}^{(i)}
\tag{249}
$$

  - The elimination of assembly of $\mathbf{K}$ (as its contributions can be directly added to global load vector at the element level) and assembly of a nontrivial $\mathbf{M}$ (since it's diagonal only the diagonal values are assembled) substantially reduces computational cost as well as memory as none of these matrices are stored in memory ($\mathbf{M}$ is assembled to a vector).

– Finally, we note that in fact any form of nonlinearity in material model would have easily been easily taken case of with an explicit method. That is, in the general nonlinear form of

$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{f}_k(\mathbf{U}) = \mathbf{R} \qquad \text{Nonlinear elastodynamics} \tag{250}$$

a general nonlinear stiffness part force $\mathbf{f}_k(\mathbf{U})$ would again have directly been assembled at the element level and would not make the update to t_{n+1} nonlinear!

- Summary of features of solution of central difference method (and other explicit methods are):
  1. These methods are stable and we must have $\Delta t \leq \Delta t_{\max}(h)$ where $\Delta t_{\max}(h)$ is the maximum allowable time step depending on the underlying PDE, explicit method, (min) element size h, *etc.*.
  2. $\mathbf{K}$ is not assembled, stiffness related forces are directly added to global force vector.
  3. Even for nonlinear elastodynamics the update equations are linear.
  4. If $\mathbf{C} = 0$ and lumped mass matrix is used:
     – Solution is trivial as no global matrix equation is solved.
     – The period shrinkage causes by using explicit integration scheme matches well with period elongation by using a lumped mass matrix.
  5. Since only diagonal members of $\mathbf{M}$ and no $\mathbf{K}$ are assembled there is not much memory overhead.
- The solution process can be summarized in the table below:

TABLE 9.1 *Step-by-step solution using central difference method (general mass and damping matrices)*

A. *Initial calculations:*

1. Form stiffness matrix $\mathbf{K}$, mass matrix $\mathbf{M}$, and damping matrix $\mathbf{C}$.
2. Initialize ${}^0\mathbf{U}$, ${}^0\dot{\mathbf{U}}$, and ${}^0\ddot{\mathbf{U}}$.
3. Select time step Δt, $\Delta t \leq \Delta t_{cr}$, and calculate integration constants:

$$a_0 = \frac{1}{\Delta t^2}; \qquad a_1 = \frac{1}{2\,\Delta t}; \qquad a_2 = 2a_0; \qquad a_3 = \frac{1}{a_2}$$

4. Calculate ${}^{-\Delta t}\mathbf{U} = {}^0\mathbf{U} - \Delta t\,{}^0\dot{\mathbf{U}} + a_3\,{}^0\ddot{\mathbf{U}}$.
5. Form effective mass matrix $\hat{\mathbf{M}} = a_0\mathbf{M} + a_1\mathbf{C}$.
6. Triangularize $\hat{\mathbf{M}}$: $\hat{\mathbf{M}} = \mathbf{L}\mathbf{D}\mathbf{L}^T$.

B. *For each time step:*

1. Calculate effective loads at time t:

$${}^t\hat{\mathbf{R}} = {}^t\mathbf{R} - (\mathbf{K} - a_2\mathbf{M})\,{}^t\mathbf{U} - (a_0\mathbf{M} - a_1\mathbf{C})\,{}^{t-\Delta t}\mathbf{U}$$

2. Solve for displacements at time $t + \Delta t$:

$$\mathbf{L}\mathbf{D}\mathbf{L}^T\,{}^{t+\Delta t}\mathbf{U} = {}^t\hat{\mathbf{R}}$$

3. If required, evaluate accelerations and velocities at time t:

$${}^t\ddot{\mathbf{U}} = a_0({}^{t-\Delta t}\mathbf{U} - 2\,{}^t\mathbf{U} + {}^{t+\Delta t}\mathbf{U})$$
$${}^t\dot{\mathbf{U}} = a_1(-{}^{t-\Delta t}\mathbf{U} + {}^{t+\Delta t}\mathbf{U})$$

4.3.2 Houbolt method (an implicit LMS method for elastodynamics)

- Houbolt method is a LMS ($k = 3$) implicit method where the FD stencils for $\ddot{\mathbf{U}}$ and $\dot{\mathbf{U}}$ are

$${}^{t+\Delta t}\dot{\mathbf{U}} = \frac{1}{6\Delta t}(11\,{}^{t+\Delta t}\mathbf{U} - 18\,{}^t\mathbf{U} + 9\,{}^{t-\Delta t}\mathbf{U} - 2\,{}^{t-2\Delta t}\mathbf{U})$$
$${}^{t+\Delta t}\ddot{\mathbf{U}} = \frac{1}{\Delta t^2}(2\,{}^{t+\Delta t}\mathbf{U} - 5\,{}^t\mathbf{U} + 4\,{}^{t-\Delta t}\mathbf{U} - {}^{t-2\Delta t}\mathbf{U}) \tag{251}$$

- After plugging these values in (226a) ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$) for t_{n+1} we obtain,

$$\left(\frac{2}{\Delta t^2}\mathbf{M} + \frac{11}{6\Delta t}\mathbf{C} + \mathbf{K}\right) {}^{t+\Delta t}\mathbf{U} = {}^{t+\Delta t}\mathbf{R} + \left(\frac{5}{\Delta t^2}\mathbf{M} + \frac{3}{\Delta t}\mathbf{C}\right) {}^{t}\mathbf{U} - \left(\frac{4}{\Delta t^2}\mathbf{M} + \frac{3}{2\Delta t}\mathbf{C}\right) {}^{t-\Delta t}\mathbf{U} + \left(\frac{1}{\Delta t^2}\mathbf{M} + \frac{1}{3\Delta t}\mathbf{C}\right) {}^{t-2\Delta t}\mathbf{U} \tag{252}$$

- We observe that that **K** appears at the LHS and must be assembled.
- In addition if the problem were nonlinear, this methods update equation would have been nonlinear.
- This implicit method is unconditionally stable. Table from [Bathe, 2006] summarized the system update for time step t_{n+1}:

TABLE 9.2 *Step-by-step solution using Houbolt integration method*

A. *Initial calculations:*

1. Form stiffness matrix $\mathbf{K}$, mass matrix $\mathbf{M}$, and damping matrix $\mathbf{C}$.
2. Initialize ${}^{0}\mathbf{U}$, ${}^{0}\dot{\mathbf{U}}$, and ${}^{0}\ddot{\mathbf{U}}$.
3. Select time step Δt and calculate integration constants:

$$a_0 = \frac{2}{\Delta t^2}; \quad a_1 = \frac{11}{6\Delta t}; \quad a_2 = \frac{5}{\Delta t^2}; \quad a_3 = \frac{3}{\Delta t}; \quad a_4 = -2a_0;$$

$$a_5 = \frac{-a_3}{2}; \quad a_6 = \frac{a_0}{2}; \quad a_7 = \frac{a_3}{9}$$

4. Use special starting procedure to calculate ${}^{\Delta t}\mathbf{U}$ and ${}^{2\Delta t}\mathbf{U}$.
5. Calculate effective stiffness matrix $\hat{\mathbf{K}}$: $\hat{\mathbf{K}} = \mathbf{K} + a_0\mathbf{M} + a_1\mathbf{C}$.
6. Triangularize $\hat{\mathbf{K}}$: $\hat{\mathbf{K}} = \mathbf{L}\mathbf{D}\mathbf{L}^T$.

B. *For each time step:*

1. Calculate effective load at time $t + \Delta t$:

$${}^{t+\Delta t}\hat{\mathbf{R}} = {}^{t+\Delta t}\mathbf{R} + \mathbf{M}(a_2\,{}^{t}\mathbf{U} + a_4\,{}^{t-\Delta t}\mathbf{U} + a_6\,{}^{t-2\Delta t}\mathbf{U}) + \mathbf{C}(a_3\,{}^{t}\mathbf{U} + a_5\,{}^{t-\Delta t}\mathbf{U} + a_7\,{}^{t-2\Delta t}\mathbf{U})$$

2. Solve for displacements at time $t + \Delta t$:

$$\mathbf{L}\mathbf{D}\mathbf{L}^T\,{}^{t+\Delta t}\mathbf{U} = {}^{t+\Delta t}\hat{\mathbf{R}}$$

3. If required, evaluate accelerations and velocities at time $t + \Delta t$:

$${}^{t+\Delta t}\ddot{\mathbf{U}} = a_0\,{}^{t+\Delta t}\mathbf{U} - a_2\,{}^{t}\mathbf{U} - a_4\,{}^{t-\Delta t}\mathbf{U} - a_6\,{}^{t-2\Delta t}\mathbf{U}$$

$${}^{t+\Delta t}\dot{\mathbf{U}} = a_1\,{}^{t+\Delta t}\mathbf{U} - a_3\,{}^{t}\mathbf{U} - a_5\,{}^{t-\Delta t}\mathbf{U} - a_7\,{}^{t-2\Delta t}\mathbf{U}$$

4.4 Multivariate single-step methods

- In contrast to the explicit central different and implicit Houbolt methods that require values for t_{n-1} and earlier for the solution of t_n, we are looking for solution schemes that only use values for t_n.
- To make this approach to work, we need to add $\dot{\mathbf{U}}$ and $\ddot{\mathbf{U}}$ (velocity and acceleration) to $\mathbf{U}$ as other variables of the problem that should be updated from step t_n to t_{n+1}.
- The values for $\dot{\mathbf{U}}$ and $\ddot{\mathbf{U}}$ may be kept in the formulation (as they are anyhow generally needed) or eliminated in the final form of update from t_n to t_{n+1}.
- We discuss two very important examples from these approaches: θ-Wilson and Newmark methods.

4.4.1 The θ-Wilson method

- In θ-Wilson method acceleration is linearly interpolated between time step t_n (t) to $\theta\Delta t$ after that

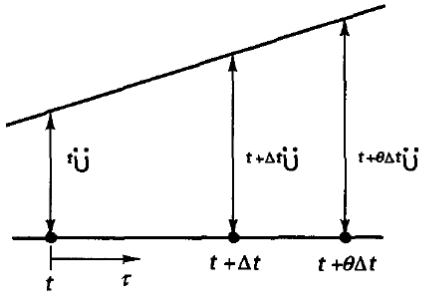


Figure 9.1 Linear acceleration assumption of Wilson θ method

- This linear acceleration interpolation is expressed as follows:

$$^{t+\tau}\ddot{\mathbf{U}} = {}^t\ddot{\mathbf{U}} + \frac{\tau}{\theta\,\Delta t}\left({}^{t+\theta\Delta t}\ddot{\mathbf{U}} - {}^t\ddot{\mathbf{U}}\right) \tag{253}$$

  Note that $^{t+\theta\Delta t}\ddot{\mathbf{U}}$ (encircled) is still an unknown that will be derived below.

- By twice integration of acceleration equation (253) we obtain equations for $\mathbf{U}$ and $\dot{\mathbf{U}}$:

$$\begin{aligned} {}^{t+\tau}\dot{\mathbf{U}} &= {}^t\dot{\mathbf{U}} + {}^t\ddot{\mathbf{U}}\tau + \frac{\tau^2}{2\theta\,\Delta t}\left({}^{t+\theta\Delta t}\ddot{\mathbf{U}} - {}^t\ddot{\mathbf{U}}\right) \\ {}^{t+\tau}\mathbf{U} &= {}^t\mathbf{U} + {}^t\dot{\mathbf{U}}\tau + \frac{1}{2}{}^t\ddot{\mathbf{U}}\tau^2 + \frac{1}{6\theta\,\Delta t}\tau^3\left({}^{t+\theta\Delta t}\ddot{\mathbf{U}} - {}^t\ddot{\mathbf{U}}\right) \end{aligned} \tag{254}$$

- To obtain $^{t+\theta\Delta t}\ddot{\mathbf{U}}$ and also have values for the next time step t_{n+1}, we plug in $t = t + \Delta t$ (t refers to t_n) in (254) to obtain,

$$\begin{aligned} {}^{t+\theta\Delta t}\dot{\mathbf{U}} &= {}^t\dot{\mathbf{U}} + \frac{\theta\,\Delta t}{2}\left({}^{t+\theta\Delta t}\ddot{\mathbf{U}} + {}^t\ddot{\mathbf{U}}\right) && \text{(a)} \\ {}^{t+\theta\Delta t}\mathbf{U} &= {}^t\mathbf{U} + \theta\,\Delta t\,{}^t\dot{\mathbf{U}} + \frac{\theta^2\,\Delta t^2}{6}\left({}^{t+\theta\Delta t}\ddot{\mathbf{U}} + 2\,{}^t\ddot{\mathbf{U}}\right) && \text{(b)} \end{aligned} \tag{255}$$

- To obtain $^{t+\theta\Delta t}\ddot{\mathbf{U}}$ and $^{t+\theta\Delta t}\dot{\mathbf{U}}$ we do:
  - First find $^{t+\theta\Delta t}\ddot{\mathbf{U}}$ from (255)(b).
  - Plug $^{t+\theta\Delta t}\ddot{\mathbf{U}}$ in (255)(a) to obtain $^{t+\theta\Delta t}\dot{\mathbf{U}}$.

- This provides values for the unknowns:

$$\begin{aligned} {}^{t+\theta\Delta t}\ddot{\mathbf{U}} &= \frac{6}{\theta^2\,\Delta t^2}\left({}^{t+\theta\Delta t}\mathbf{U} - {}^t\mathbf{U}\right) - \frac{6}{\theta\,\Delta t}{}^t\dot{\mathbf{U}} - 2\,{}^t\ddot{\mathbf{U}} \\ {}^{t+\theta\Delta t}\dot{\mathbf{U}} &= \frac{3}{\theta\,\Delta t}\left({}^{t+\theta\Delta t}\mathbf{U} - {}^t\mathbf{U}\right) - 2\,{}^t\dot{\mathbf{U}} - \frac{\theta\,\Delta t}{2}{}^t\ddot{\mathbf{U}} \end{aligned} \tag{256}$$

- Thus from (256) the unknowns $^{t+\theta\Delta t}\ddot{\mathbf{U}}$ and $^{t+\theta\Delta t}\dot{\mathbf{U}}$ are written in terms of one unknown vector $^{t+\theta\Delta t}\mathbf{U}$. Subsequently, we plug these values in (226a) ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$) for $t + \theta\Delta t$,

$$\begin{aligned} \mathbf{M}\,{}^{t+\theta\Delta t}\ddot{\mathbf{U}} + \mathbf{C}\,{}^{t+\theta\Delta t}\dot{\mathbf{U}} + \mathbf{K}\,{}^{t+\theta\Delta t}\mathbf{U} &= {}^{t+\theta\Delta t}\overline{\mathbf{R}} \\ {}^{t+\theta\Delta t}\overline{\mathbf{R}} &= {}^t\mathbf{R} + \theta\left({}^{t+\Delta t}\mathbf{R} - {}^t\mathbf{R}\right) \end{aligned} \tag{257}$$

- to obtain $^{t+\theta\Delta t}\mathbf{U}$ from the system below,

$$\hat{\mathbf{K}}\,{}^{t+\theta\Delta t}\mathbf{U} = {}^{t+\theta\Delta t}\,\hat{\mathbf{R}} \tag{258a}$$

$$\hat{\mathbf{K}} = \mathbf{K} + \frac{6}{(\theta\Delta t)^2}\mathbf{M} + \frac{3}{\theta\Delta t}\mathbf{C} \tag{258b}$$

$${}^{t+\theta\Delta t}\hat{\mathbf{R}} = {}^t\mathbf{R} + \theta\left({}^{t+\Delta t}\mathbf{R} - {}^t\mathbf{R}\right) + \mathbf{M}\left(a_0\,{}^t\mathbf{U} + a_2\,{}^t\dot{\mathbf{U}} + 2\,{}^t\ddot{\mathbf{U}}\right) + \mathbf{C}\left(a_1\,{}^t\mathbf{U} + 2\,{}^t\dot{\mathbf{U}} + a_3\,{}^t\ddot{\mathbf{U}}\right) \tag{258c}$$

- Once we obtain ${}^{t+\theta\Delta t}\mathbf{U}$ we obtain ${}^{t+\theta\Delta t}\dot{\mathbf{U}}$ and ${}^{t+\theta\Delta t}\ddot{\mathbf{U}}$ from (256).
- Finally, we plug $\tau = \Delta t$ in (254) to obtain ${}^{t+\Delta t}\mathbf{U}$ and ${}^{t+\Delta t}\dot{\mathbf{U}}$ and and in (253) to obtain ${}^{t+\Delta t}\ddot{\mathbf{U}}$ and be ready for the next time step.
- The θ-Wilson method is unconditionally stable for $\theta \geq 1.37$ and usually we use $\theta = 1.40$.
- The summary of the above algorithm is shown below,

TABLE 9.3 *Step-by-step solution using Wilson θ integration method*

A. *Initial calculations:*

1. Form stiffness matrix $\mathbf{K}$, mass matrix $\mathbf{M}$, and damping matrix $\mathbf{C}$.
2. Initialize ${}^{0}\mathbf{U}$, ${}^{0}\dot{\mathbf{U}}$, and ${}^{0}\ddot{\mathbf{U}}$.
3. Select time step Δt and calculate integration constants, $\theta = 1.4$ (usually):

$$a_0 = \frac{6}{(\theta\,\Delta t)^2}; \quad a_1 = \frac{3}{\theta\,\Delta t}; \quad a_2 = 2a_1; \quad a_3 = \frac{\theta\,\Delta t}{2}; \quad a_4 = \frac{a_0}{\theta};$$

$$a_5 = \frac{-a_2}{\theta}; \quad a_6 = 1 - \frac{3}{\theta}; \quad a_7 = \frac{\Delta t}{2}; \quad a_8 = \frac{\Delta t^2}{6}$$

4. Form effective stiffness matrix $\hat{\mathbf{K}}$: $\hat{\mathbf{K}} = \mathbf{K} + a_0\mathbf{M} + a_1\mathbf{C}$.
5. Triangularize $\hat{\mathbf{K}}$: $\hat{\mathbf{K}} = \mathbf{L}\mathbf{D}\mathbf{L}^T$.

B. *For each time step:*

1. Calculate effective loads at time $t + \theta\,\Delta t$:

$${}^{t+\theta\Delta t}\hat{\mathbf{R}} = {}^{t}\mathbf{R} + \theta({}^{t+\Delta t}\mathbf{R} - {}^{t}\mathbf{R}) + \mathbf{M}(a_0\,{}^{t}\mathbf{U} + a_2\,{}^{t}\dot{\mathbf{U}} + 2\,{}^{t}\ddot{\mathbf{U}}) + \mathbf{C}(a_1\,{}^{t}\mathbf{U} + 2\,{}^{t}\dot{\mathbf{U}} + a_3\,{}^{t}\ddot{\mathbf{U}})$$

2. Solve for displacements at time $t + \theta\,\Delta t$:

$$\mathbf{L}\mathbf{D}\mathbf{L}^T\,{}^{t+\theta\Delta t}\mathbf{U} = {}^{t+\theta\Delta t}\hat{\mathbf{R}}$$

3. Calculate displacements, velocities, and accelerations at time $t + \Delta t$:

$${}^{t+\Delta t}\ddot{\mathbf{U}} = a_4({}^{t+\theta\Delta t}\mathbf{U} - {}^{t}\mathbf{U}) + a_5\,{}^{t}\dot{\mathbf{U}} + a_6\,{}^{t}\ddot{\mathbf{U}}$$

$${}^{t+\Delta t}\dot{\mathbf{U}} = {}^{t}\dot{\mathbf{U}} + a_7({}^{t+\Delta t}\ddot{\mathbf{U}} + {}^{t}\ddot{\mathbf{U}})$$

$${}^{t+\Delta t}\mathbf{U} = {}^{t}\mathbf{U} + \Delta t\,{}^{t}\dot{\mathbf{U}} + a_8({}^{t+\Delta t}\ddot{\mathbf{U}} + 2\,{}^{t}\ddot{\mathbf{U}})$$

EXAMPLE 9.3: Calculate the displacement response of the system considered in Examples 9.1 and 9.2 using the Wilson θ method. Use $\theta = 1.4$.

First we consider the case $\Delta t = 0.28$. Following the steps of calculations in Table 9.3, we have

$${}^{0}\mathbf{U} = \begin{bmatrix}0\\0\end{bmatrix}; \quad {}^{0}\dot{\mathbf{U}} = \begin{bmatrix}0\\0\end{bmatrix}; \quad {}^{0}\ddot{\mathbf{U}} = \begin{bmatrix}0\\10\end{bmatrix}$$

where ${}^{0}\ddot{\mathbf{U}}$ was evaluated in Example 9.1. Then (listed to three digits)

$$a_0 = 39.0; \quad a_1 = 7.65; \quad a_2 = 15.3; \quad a_3 = 0.196; \quad a_4 = 27.9;$$

$$a_5 = -10.9; \quad a_6 = -1.14; \quad a_7 = 0.14; \quad a_8 = 0.0131$$

and

$$\hat{\mathbf{K}} = \begin{bmatrix}6 & -2\\-2 & 4\end{bmatrix} + 39.0\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix} = \begin{bmatrix}84.1 & -2\\-2 & 43.0\end{bmatrix}$$

For each time step we need to evaluate

$$
{}^{t+\theta\Delta t}\hat{\mathbf{R}} = \begin{bmatrix} 0 \\ 10 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}(39.0\ {}^{t}\mathbf{U} + 15.3\ {}^{t}\dot{\mathbf{U}} + 2\ {}^{t}\ddot{\mathbf{U}})
$$

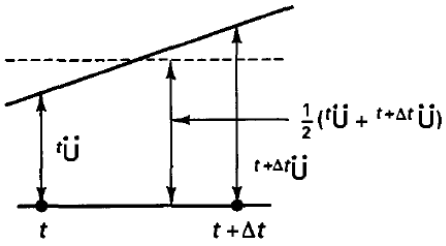
$$
\hat{\mathbf{K}}\ {}^{t+\theta\Delta t}\mathbf{U} = {}^{t+\theta\Delta t}\hat{\mathbf{R}}
$$

and then calculate

$$
\begin{aligned}
{}^{t+\Delta t}\ddot{\mathbf{U}} &= 27.9({}^{t+\theta\Delta t}\mathbf{U} - {}^{t}\mathbf{U}) - 10.9\ {}^{t}\dot{\mathbf{U}} - 1.14\ {}^{t}\ddot{\mathbf{U}} \\
{}^{t+\Delta t}\dot{\mathbf{U}} &= {}^{t}\dot{\mathbf{U}} + 0.14({}^{t+\Delta t}\ddot{\mathbf{U}} + {}^{t}\ddot{\mathbf{U}}) \\
{}^{t+\Delta t}\mathbf{U} &= {}^{t}\mathbf{U} + 0.28\ {}^{t}\dot{\mathbf{U}} + 0.0131({}^{t+\Delta t}\ddot{\mathbf{U}} + 2\ {}^{t}\ddot{\mathbf{U}})
\end{aligned}
$$

Time	Δt	$2\Delta t$	$3\Delta t$	$4\Delta t$	$5\Delta t$	$6\Delta t$	$7\Delta t$	$8\Delta t$	$9\Delta t$	$10\Delta t$	$11\Delta t$	$12\Delta t$
${}^{t}\mathbf{U}$	0.00605	0.0525	0.196	0.490	0.952	1.54	2.16	2.67	2.92	2.82	2.33	1.54
	0.366	1.34	2.64	3.92	4.88	5.31	5.18	4.61	3.82	3.06	2.52	2.29

4.4.2 The Newmark method



- In Newmark method, $\mathbf{U}$, $\dot{\mathbf{U}}$ are expressed in terms of $\mathbf{U}, \dot{\mathbf{U}}, \ddot{\mathbf{U}}$ at t_n and ${}^{t+\Delta t}\ddot{\mathbf{U}}$:

$$
\begin{aligned}
{}^{t+\Delta t}\dot{\mathbf{U}} &= {}^{t}\dot{\mathbf{U}} + [(1-\delta)\ {}^{t}\ddot{\mathbf{U}} + \delta\ {}^{t+\Delta t}\ddot{\mathbf{U}}]\ \Delta t \\
{}^{t+\Delta t}\mathbf{U} &= {}^{t}\mathbf{U} + {}^{t}\dot{\mathbf{U}}\ \Delta t + [(\tfrac{1}{2}-\alpha)\ {}^{t}\ddot{\mathbf{U}} + \alpha\ {}^{t+\Delta t}\ddot{\mathbf{U}}]\ \Delta t^2
\end{aligned}
\tag{259}
$$

  where α and δ are the two parameters of the Newmark method.

- Similar to θ-Wilson method we need to express ${}^{t+\Delta t}\dot{\mathbf{U}}$ and ${}^{t+\Delta t}\ddot{\mathbf{U}}$ in terms of ${}^{t+\Delta t}\mathbf{U}$.
- This is done by obtaining ${}^{t+\Delta t}\ddot{\mathbf{U}}$ from (259)(b), plugging back in (259)(a) to (259)(a) we obtain ${}^{t+\Delta t}\dot{\mathbf{U}}$.
- Thus in equation,

$$
\mathbf{M}\ {}^{t+\Delta t}\ddot{\mathbf{U}} + \mathbf{C}\ {}^{t+\Delta t}\dot{\mathbf{U}} + \mathbf{K}\ {}^{t+\Delta t}\mathbf{U} = {}^{t+\Delta t}\mathbf{R}
\tag{260}
$$

  we express ${}^{t+\Delta t}\dot{\mathbf{U}}$ and ${}^{t+\Delta t}\ddot{\mathbf{U}}$ in terms of ${}^{t+\Delta t}\mathbf{U}$ to obtain one matrix equation only in terms of ${}^{t+\Delta t}\mathbf{U}$.
- Once we have ${}^{t+\Delta t}\mathbf{U}$ we can go back to (259)(b) to obtain ${}^{t+\Delta t}\ddot{\mathbf{U}}$ and plug this back in (259)(a) to obtain ${}^{t+\Delta t}\dot{\mathbf{U}}$.
- That is, all values $\mathbf{U}, \dot{\mathbf{U}}, \ddot{\mathbf{U}}$ are obtained for the next time step t_{n+1} and we repeat this process until the final time.
- We further discuss the stability condition of Newmark method based on values of α and δ in §5.3.3; *cf.* (386).
- Family of Newmark method includes a variety of popular solution schemes for elastodynamic problem as shown below,

Method	Type	α	δ	Stability condition[2]	Order of accuracy[3]
Average acceleration (trapezoidal rule)	Implicit	$\frac{1}{4}$	$\frac{1}{2}$	Unconditional	2
Linear acceleration	Implicit	$\frac{1}{6}$	$\frac{1}{2}$	$\Omega_{\text{crit}} = 2\sqrt{3} \cong 3.464$	2
Fox-Goodwin (royal road)	Implicit	$\frac{1}{12}$	$\frac{1}{2}$	$\Omega_{\text{crit}} = \sqrt{6} \cong 2.449$	2
Central difference	Explicit[1]	0	$\frac{1}{2}$	$\Omega_{\text{crit}} = 2$	2

where Ω_{crit} is compared with $\overline{\Delta t} = \omega_{\max}\Delta t$ where $\omega_{\max}$ is the maximum frequency from modal analysis which can conservatively replaced by the highest element of the smallest element size $\omega_{h_{\min}}$ (if different elements are used the maximum natural frequency of the individual elements).

- Table below summarizes the steps of Newmark method.

TABLE 9.4 *Step-by-step solution using Newmark integration method*

A. *Initial calculations:*

1. Form stiffness matrix $\mathbf{K}$, mass matrix $\mathbf{M}$, and damping matrix $\mathbf{C}$.
2. Initialize ${}^0\mathbf{U}$, ${}^0\dot{\mathbf{U}}$, and ${}^0\ddot{\mathbf{U}}$.
3. Select time step Δt and parameters α and δ and calculate integration constants:

$$\delta \geq 0.50; \qquad \alpha \geq 0.25(0.5 + \delta)^2$$

$$a_0 = \frac{1}{\alpha\,\Delta t^2}; \qquad a_1 = \frac{\delta}{\alpha\,\Delta t}; \qquad a_2 = \frac{1}{\alpha\,\Delta t}; \qquad a_3 = \frac{1}{2\alpha} - 1;$$

$$a_4 = \frac{\delta}{\alpha} - 1; \qquad a_5 = \frac{\Delta t}{2}\left(\frac{\delta}{\alpha} - 2\right); \qquad a_6 = \Delta t(1 - \delta); \qquad a_7 = \delta\,\Delta t$$

4. Form effective stiffness matrix $\hat{\mathbf{K}}$: $\hat{\mathbf{K}} = \mathbf{K} + a_0\mathbf{M} + a_1\mathbf{C}$.
5. Triangularize $\hat{\mathbf{K}}$: $\hat{\mathbf{K}} = \mathbf{L}\mathbf{D}\mathbf{L}^T$.

B. *For each time step:*

1. Calculate effective loads at time $t + \Delta t$:

$${}^{t+\Delta t}\hat{\mathbf{R}} = {}^{t+\Delta t}\mathbf{R} + \mathbf{M}(a_0\,{}^t\mathbf{U} + a_2\,{}^t\dot{\mathbf{U}} + a_3\,{}^t\ddot{\mathbf{U}}) + \mathbf{C}(a_1\,{}^t\mathbf{U} + a_4\,{}^t\dot{\mathbf{U}} + a_5\,{}^t\ddot{\mathbf{U}})$$

2. Solve for displacements at time $t + \Delta t$:

$$\mathbf{L}\mathbf{D}\mathbf{L}^T\,{}^{t+\Delta t}\mathbf{U} = {}^{t+\Delta t}\hat{\mathbf{R}}$$

3. Calculate accelerations and velocities at time $t + \Delta t$:

$${}^{t+\Delta t}\ddot{\mathbf{U}} = a_0({}^{t+\Delta t}\mathbf{U} - {}^t\mathbf{U}) - a_2\,{}^t\dot{\mathbf{U}} - a_3\,{}^t\ddot{\mathbf{U}}$$

$${}^{t+\Delta t}\dot{\mathbf{U}} = {}^t\dot{\mathbf{U}} + a_6\,{}^t\ddot{\mathbf{U}} + a_7\,{}^{t+\Delta t}\ddot{\mathbf{U}}$$

EXAMPLE 9.4: Calculate the displacement response of the system considered in Examples 9.1 to 9.3 using the Newmark method. Use $\alpha = 0.25$, $\delta = 0.5$.

Consider first the case $\Delta t = 0.28$. Following the steps of calculations given in Table 9.4, we have

$${}^0\mathbf{U} = \begin{bmatrix}0\\0\end{bmatrix}; \qquad {}^0\dot{\mathbf{U}} = \begin{bmatrix}0\\0\end{bmatrix}; \qquad {}^0\ddot{\mathbf{U}} = \begin{bmatrix}0\\10\end{bmatrix}$$

The integration constants are (showing three digits)

$$a_0 = 51.0; \qquad a_1 = 7.14; \qquad a_2 = 14.3; \qquad a_3 = 1.00;$$

$$a_4 = 1.00; \qquad a_5 = 0.00; \qquad a_6 = 0.14; \qquad a_7 = 0.14$$

Thus the effective stiffness matrix is

$$\hat{\mathbf{K}} = \begin{bmatrix}6 & -2\\-2 & 4\end{bmatrix} + 51.0\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix} = \begin{bmatrix}108 & -2\\-2 & 55\end{bmatrix}$$

For each time step we need to evaluate

$$^{t+\Delta t}\hat{\mathbf{R}} = \begin{bmatrix} 0 \\ 10 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} (51\ ^{t}\mathbf{U} + 14.3\ ^{t}\dot{\mathbf{U}} + 1.0\ ^{t}\ddot{\mathbf{U}})$$

Then

$$\hat{\mathbf{K}}\ ^{t+\Delta t}\mathbf{U} = {}^{t+\Delta t}\hat{\mathbf{R}}$$

and

$$^{t+\Delta t}\ddot{\mathbf{U}} = 51.0(^{t+\Delta t}\mathbf{U} - {}^{t}\mathbf{U}) - 14.3\ ^{t}\dot{\mathbf{U}} - 1.0\ ^{t}\ddot{\mathbf{U}}$$

$$^{t+\Delta t}\dot{\mathbf{U}} = {}^{t}\dot{\mathbf{U}} + 0.14\ ^{t}\ddot{\mathbf{U}} + 0.14\ ^{t+\Delta t}\ddot{\mathbf{U}}$$

Performing these calculations, we obtain

Time	Δt	$2\Delta t$	$3\Delta t$	$4\Delta t$	$5\Delta t$	$6\Delta t$	$7\Delta t$	$8\Delta t$	$9\Delta t$	$10\Delta t$	$11\Delta t$	$12\Delta t$
$^{t}\mathbf{U}$	0.00673	0.0505	0.189	0.485	0.961	1.58	2.23	2.76	3.00	2.85	2.28	1.40
	0.364	1.35	2.68	4.00	4.95	5.34	5.13	4.48	3.64	2.90	2.44	2.31

4.5 Runge-Kutta (RK) methods

4.5.1 Runge-Kutta (RK) methods: Introduction

- Consider a general first order ODE expressed as follows,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y) \qquad \text{First order ODE} \tag{261a}$$

$$y(t = 0) = y_0 \qquad \text{Initial condition (IC)} \tag{261b}$$

  any nonlinear first order ODE can be expressed in the form (261).

- Explicit Runge-Kutta (RK) update the solution from time step t_n to t_{n+1} through $s \geq 1$ stages:

$$y_{n+1} = y_h + \Delta t \sum_{i=1}^{s} b_i k_i \qquad \text{where} \tag{262a}$$

$$k_i = f(t_n + \Delta t c_i, y_n + \Delta t \sum_{j=1}^{i-1} a_{ij} k_j), \qquad 1 \leq i \leq s \tag{262b}$$

- The intermediate values k_i represent intermediate slopes ($\frac{\mathrm{d}y}{\mathrm{d}t}$) at intermediate independent coordinate $t_n + \Delta t c_i$ (which fall between t_n and t_{n+1}) and dependent variable $\Delta t \sum_{j=1}^{i-1} a_{ij} k_j$.
- The fact that the upper limit of summation is $i - 1$ is that each k_i **only** depends on prior k_j ($j < i$) values.
- This enables a step-by-step solution strategy where k_i are solved for $i = 1$ to $i = s$ an finally the new update is computed from (262a) ($y_{n+1} = y_h + \Delta t \sum_{i=1}^{s} b_i k_i \qquad$ where).
- RK parameters are,
  1. Size $s \times s$ matrix a_{ij}
  2. Size s vector b_i
  3. Size s vector c_i
- For an explicit RK method $a_{ij} = 0$ for diagonal and upper diagonal members ($i \leq j$).
- This is what enables the method to become explicit and require a simple and linear update equation for each k_i **(even if f is nonlinear in y)**.

- **Butcher tableau**: The parameters of a RK method are shown in a butcher tableau:

$$\begin{array}{c|cccc} c_1 & 0 & \cdots & 0 & 0 \\ c_2 & a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & 0 \\ c_s & a_{s1} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & \cdots & b_{s-1} & b_s \end{array} \tag{263}$$

  which for an explicit RK method has upper and diagonal values of a zero.

- Comparison of RK methods and LMS:

  1. The purpose of having s stages in RK methods is to increase temporal order of accuracy of time integration.
  2. In LMS methods this is achieved by increasing the historical data pool, *i.e.,* number of prior steps required for updating the solution from t_n to t_{n+1}.
  3. However, RK methods, being simple step methods, are more flexible: They allow adjustment of order of accuracy in time and time step.
  4. This flexibility in changing or having high temporal order of accuracy and/or changing the time step size, provides very novel approaches to estimate local truncation errors and adjust time step or temporal order in adaptive RK schemes; *cf.* [Chapra and Canale, 2010] Chapter 25 (section 25.5) for more examples.

- The idea behind formulating values of 1) a, c, b, 2) Relation between s and temporal order of accuracy is based on Taylor series expansion of the exact solution and numerical update from t_n to t_{n+1} and matching factors of Δt^q.

4.5.2 Second order RK (RK2) methods

- We formulate EXRK2 (*i.e.,* explicit RK with $s = 2$).
- This scheme is written as,

$$y_{n+1} = y_h + \Delta t(b_1 k_1 + b_2 k_2) \qquad \text{where} \tag{264a}$$
$$k_1 = f(t_n, y_n) \tag{264b}$$
$$k_2 = f(t_n + c_2 \Delta t, y_n + \Delta t a_{21} k_1) \tag{264c}$$

- This corresponds to the first two rows and columns of Butcher tableau with $c_1 = 0$ so that k_1 is computed for t_n, y_n.
- There are 4 unknown values: b_1, b_2, c_2, a_{21}.
- We investigate what order of accuracy can be achieved with these four unknowns by matching Taylor series expansion of exact and numerical solutions.
- As usual we adopt the following notation,

$$y(t_n) \qquad \text{Exact solution at } t_n \text{ from the ODE (261):} \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y) \tag{265a}$$
$$y_n \qquad \text{RK (numerical) solution at } t_n \text{ from the RK update (262):} \qquad y_{n+1} = y_h + \Delta t \sum_{i=1}^{s} b_i k_i \tag{265b}$$

  note that both solutions start from the same initial condition (261b) ($y(t_0) = y_0 = y_0$).

- The purpose of the analysis in the following is,

  Let $y_n = y(t_n)$ (266a)
  Update exact solution to t_{n+1} ($y(t_{n+1})$) using (261a). (266b)
  Update numerical solution to t_{n+1} (y_{n+1}) using (262). (266c)
  Evaluate to what order Δt^q exact and numerical solutions can match by adjusting RK model parameters. (266d)

- First, we evaluate the Taylor expansion of the exact solution from t_n to t_{n+1},

$$y(t_{n+1}) = y(t_n) + \Delta t \frac{\mathrm{d}y}{\mathrm{d}t}(t_n) + \frac{1}{2}\Delta t^2 \frac{\mathrm{d}^2 y}{\mathrm{d}t^2}(t_n) + \cdots + \frac{1}{q!}\Delta t^q \frac{\mathrm{d}^{(q)} y}{\mathrm{d}t^{(q)}}(t_n) + \mathcal{O}\left(\Delta t^{q+1}\right) \tag{267}$$

- We use (261a) ($\frac{\mathrm{d}y}{\mathrm{d}t} = f(t,y)$) and the chain rule to compute $\frac{\mathrm{d}^{(q)} y}{\mathrm{d}t^{(q)}}(t_n)$ in (267).

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t,y) \quad \Rightarrow \tag{268a}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t}(t_n) := f \quad f \text{ is a shorthand for } f \text{ at } (t_n, y(t_n)) \text{ that is } f = f(t_n, y(t_n)) \text{ (the dependence on } t_n \text{ is not displayed)} \tag{268b}$$

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2}(t_n) = \frac{\mathrm{d}f}{\mathrm{d}t}(t_n) = \left(\frac{\partial f}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}t} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}\right)(t_n) \quad (\text{ from (268a)}) \quad \Rightarrow$$

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2}(t_n) := f_t + f_y f \qquad \text{note } \frac{\mathrm{d}y}{\mathrm{d}t}(t_n) = f \text{ from (268a) and shorthand notations } f_t := \frac{\partial f}{\partial t}(t_n, y(t_n)), f_y := \frac{\partial f}{\partial y}(t_n, y(t_n)) \tag{268c}$$

$$\frac{\mathrm{d}^3 y}{\mathrm{d}t^3}(t_n) := f_{tt} + f_t f_y + 2 f f_{ty} + f f_y^2 + f^2 f_{yy}, \qquad \text{(obtained in a similar fashion by the use of chain rule)} \tag{268d}$$

- By plugging (268b), (268c), and (268d) in (267) we obtain,

$$y(t_{n+1}) = y(t_n) + \Delta t f + \frac{1}{2}\Delta t^2 \left(f_t + f_y f\right) + \frac{1}{6}\Delta t^3 \left(f_{tt} + f_t f_y + 2 f f_{ty} + f f_y^2 + f^2 f_{yy}\right) + \mathcal{O}\left(\Delta t^4\right) \tag{269}$$

- Next, we form the Taylor series expansion of y_{n+1} starting from y_n.
- Note that from (266a) we let $y_n = y(t_n)$ to only characterize the local truncation error $= (y(t_{n+1}) - y_{n+1})/\Delta t$ (*i.e.*, the error in one local update step).

$$\begin{aligned}
y_{n+1} &= y_h + \Delta t(b_1 k_1 + b_2 k_2) \\
&= y_h + \Delta t(b_1 k_1 + b_2 \{f(t_n + c_2 \Delta t, y_n + a_{21}\Delta t k_1)\}) \\
&= y_h + \Delta t \left(b_1 k_1 + b_2 \left\{f + [(\Delta t c_2) f_t + (\Delta t a_{21} k_1) f_y] + \left[\frac{1}{2}(\Delta t c_2)^2 f_{tt} + \frac{1}{2}(\Delta t a_{21} k_1)^2 f_{yy} + (\Delta t c_2)(\Delta t a_{21} k_1) f_{ty}\right]\right\}\right) \\
&= y_h + \Delta t \{b_1 k_1 + b_2 f\} + \Delta t^2 b_2 \{c_2 f_t + a_{21} k_1 f_y\} + \Delta t^3 b_2 \left\{\frac{1}{2} c_2^2 f_{tt} + \frac{1}{2}(a_{21} k_1)^2 f_{yy} + c_2 a_{21} k_1 f_{ty}\right\} + \mathcal{O}\left(\Delta t^4\right)
\end{aligned}$$

- Noting that $k_1 = f$ by (264b), we have the final expression for y_{n+1},

$$y_{n+1} = y_h + \Delta t \{b_1 + b_2\} f + \Delta t^2 b_2 \{c_2 f_t + a_{21} f f_y\} + \Delta t^3 b_2 \left\{\frac{1}{2} c_2^2 f_{tt} + \frac{1}{2}(a_{21} f)^2 f_{yy} + c_2 a_{21} f f_{ty}\right\} + \mathcal{O}\left(\Delta t^4\right) \tag{270}$$

- The parameters of the RK2 (2-stage RK) method are derived by matching as many terms of exact (269) and numerical (270) factors of Δt^i:

$$\left.\begin{aligned}
&y(t_{n+1}) = y(t_n) + \Delta t f + \tfrac{1}{2}\Delta t^2 \left(f_t + f_y f\right) + \tfrac{1}{6}\Delta t^3 \left(f_{tt} + f_t f_y + 2 f f_{ty} + f f_y^2 + f^2 f_{yy}\right) + \mathcal{O}\left(\Delta t^4\right) && \underline{\text{exact}} \\
&y_{n+1} = y_h + \Delta t \{b_1 + b_2\} f + \Delta t^2 b_2 \{c_2 f_t + a_{21} f f_y\} + \Delta t^3 b_2 \left\{\tfrac{1}{2} c_2^2 f_{tt} + \tfrac{1}{2}(a_{21} f)^2 f_{yy} + c_2 a_{21} f f_{ty}\right\} + \mathcal{O}\left(\Delta t^4\right) && \underline{\text{RK}}
\end{aligned}\right\} \Rightarrow \tag{271a}$$

$$\left.\begin{aligned}
&\{b_1 + b_2\} f = f && \Delta t \text{ term} \\
&b_2 \{c_2 f_t + a_{21} f f_y\} = \tfrac{1}{2}\left(f_t + f f_y\right) && \Delta t^2 \text{ term}
\end{aligned}\right\} \quad \Rightarrow \tag{271b}$$

$$\begin{cases}
b_1 + b_2 &= 1 \\
b_2 c_2 &= \frac{1}{2} \\
b_2 a_{21} &= \frac{1}{2}
\end{cases} \tag{271c}$$

- Note that in (271b) we have only matched powers $\Delta t, \Delta t^2$ and not Δt^3 as there are not sufficient number of unknowns to make that possible.
- Still we have only **three** equations in (271c) and **four** unknowns.
- We let c_2 to be a free parameter and obtain families of EXRK2 methods:

$$c_2 \neq 0 \tag{272a}$$
$$b_1 = 1 - \frac{1}{2c_2} \tag{272b}$$
$$b_2 = \frac{1}{2c_2} \tag{272c}$$
$$a_{21} = c_2 \tag{272d}$$

In equation (273) we present some of the well-known members of RK2 methods by assigning different values of c_2.

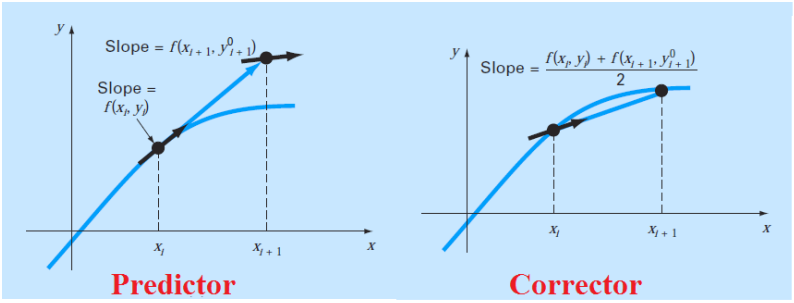
Name	c_2	RK parameters	RK update
Heun (Improved Euler)	1	$\begin{bmatrix} b_1 \\ b_2 \\ c_2 \\ a_{21} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 1 \end{bmatrix}$	$\begin{cases} y_{n+1} &= y_h + \frac{1}{2}\Delta t(k_1 + k_2) \\ k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \Delta t, y_n + \Delta t k_1) \end{cases}$ (273a)
Midpoint (Modified Euler)	$\frac{1}{2}$	$\begin{bmatrix} b_1 \\ b_2 \\ c_2 \\ a_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$	$\begin{cases} y_{n+1} &= y_h + \Delta t k_2 \\ k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \frac{1}{2}\Delta t, y_n + \frac{1}{2}\Delta t k_1) \end{cases}$ (273b)
Ralston	$\frac{3}{4}$	$\begin{bmatrix} b_1 \\ b_2 \\ c_2 \\ a_{21} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{3}{4} \\ \frac{3}{4} \end{bmatrix}$	$\begin{cases} y_{n+1} &= y_h + \Delta t \left(\frac{1}{3}k_1 + \frac{2}{3}k_2\right) \\ k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \frac{3}{4}\Delta t, y_n + \frac{3}{4}\Delta t k_1) \end{cases}$ (273c)

- Ralston [Ralston, 1962, Ralston and Rabinowitz, 1978] determined that choosing $c_2 = \frac{2}{3}$ ($c_2 = \frac{3}{4}$) provides a minimum bound on the truncation error for the second-order RK algorithms.
- Midpoint (Modified Euler) Uses the y_n to project the solution to the midpoint of the interval, and from there compute the slope k_2 that would project y_n to y_{n+1}. Note that this method is different from trapezoidal rule that is an implicit method and for which the update equation is written for the mid-point of the interval. Mid-point method, is often shown in the shorthand form below,

$$y_{n+1} = y_h + \Delta t\ f\left(t_n + \frac{1}{2}\Delta t, y_n + \frac{1}{2}\Delta t f(t_n, y_n)\right) \qquad \text{Midpoint (Modified Euler)} \tag{274}$$

- Improved Euler's method is a Heun's method without iteration (next figure). The update can be expressed as,

$$y_{n+1} = y_h + \frac{1}{2}\Delta t\left(f(t_n, y_n) + f\left(t_n + \Delta t, y_n + \Delta t f(t_n, y_n)\right)\right) \qquad \text{Heun (Improved Euler)} \tag{275}$$



Step 1: slope at the end point is estimated by forward Euler method

$y'_i = f(x_i, y_i)$

End value is estimated

$y^0_{i+1} = y_i + f(x_i, y_i)h$

Step 2: Slope is updated using predictor equation $y'_{i+1} = f(x_{i+1}, y^0_{i+1})$

Take the average slope from 2 y values

$\bar{y}' = \frac{y'_i + y'_{i+1}}{2} = \frac{f(x_i, y_i) + f(x_{i+1}, y^0_{i+1})}{2}$

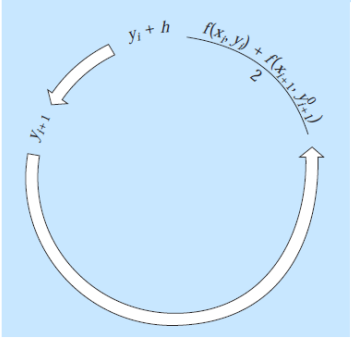
Update y using the corrector equation

$$y_{i+1} = y_i + \frac{f(x_i, y_i) + f(x_{i+1}, y^0_{i+1})}{2}h$$

Heun's method can become iterative:

Predictor (Fig. 25.9a): $y^0_{i+1} = y_i + f(x_i, y_i)h$

Corrector (Fig. 25.9b): $y_{i+1} = y_i + \frac{f(x_i, y_i) + f(x_{i+1}, y^0_{i+1})}{2}h$



[Chapra and Canale, 2010]

- To determine the order of accuracy and better understand the behavior of RK2 methods we define the local truncation error $\tau(t_n)$,

$$\begin{aligned}\tau(t_n) &:= \frac{y(t_{n+1}) - y_{n+1}}{\Delta t} \\ &= \frac{\Delta t^3 \left\{\left[\frac{1}{6}\left(f_{tt} + f_t f_y + 2ff_{ty} + ff_y^2 + f^2 f_{yy}\right)\right] - b_2\left[\frac{1}{2}c_2^2 f_{tt} + \frac{1}{2}(a_{21}f)^2 f_{yy} + c_2 a_{21} f f_{ty}\right]\right\} + \mathcal{O}\left(\Delta t^4\right)}{\Delta t} \\ &= \Delta t^2\left\{\left[\frac{1}{6}\left(f_{tt} + f_t f_y + 2ff_{ty} + ff_y^2 + f^2 f_{yy}\right)\right] - \frac{1}{2c_2}\left[\frac{1}{2}c_2^2 f_{tt} + \frac{1}{2}(c_2 f)^2 f_{yy} + c_2 c_2 f f_{ty}\right]\right\} + \mathcal{O}\left(\Delta t^3\right)\end{aligned} \tag{276}$$

- which takes the final form,

$$\tau(t_n) = \frac{y(t_n) - y_n}{\Delta t} = \Delta t^2\left\{\left(\frac{1}{6} - \frac{c_2}{4}\right)\left(f_{tt} + f_{yy}f^2\right) + \left(\frac{1}{3} - \frac{c_2}{2}\right) f f_{ty} + \frac{1}{6}\left(f_t f_y + f f_y^2\right)\right\} + \mathcal{O}\left(\Delta t^3\right) \tag{277}$$

- As usual the truncation error $\tau(t_n) = \frac{y(t_{n+1}) - y_{n+1}}{\Delta t}$ involves a division by Δt so that the order of accuracy of local truncation error would match the global order of convergence.
- We observe that the RK2 scheme is second order accurate in time.
- One thing that is clear from (277) is that we could not annihilate the $\mathcal{O}\left(\Delta t^2\right)$ term in $\tau(t_n)$ due to the lack of number of parameters for RK2 scheme, even though there was one free unknown value.
- This is often the case with RK schemes, that not parameters of an s-stage RK scheme are used in annihilating factors of Δt^i and for the ones that we can annihilate we often end up with more unknowns that equations. That, is why there may be variants of RK methods for a given stage number s.

4.5.2.1 Comparison of Various Second-Order RK Schemes[Chapra and Canale, 2010]

Problem Statement. Use the midpoint method [Eq. (25.37)] and Ralston's method [Eq. (25.38)] to numerically integrate Eq. (PT7.13)

$$f(x, y) = -2x^3 + 12x^2 - 20x + 8.5$$

from $x = 0$ to $x = 4$ using a step size of 0.5. The initial condition at $x = 0$ is $y = 1$. Compare the results with the values obtained using another second-order RK algorithm, that is, the Heun method without corrector iteration (Table 25.3).

Solution. The first step in the midpoint method is to use Eq. (25.37a) to compute

$$k_1 = -2(0)^3 + 12(0)^2 - 20(0) + 8.5 = 8.5$$

However, because the ODE is a function of x only, this result has no bearing on the second step—the use of Eq. (25.37b) to compute

$$k_2 = -2(0.25)^3 + 12(0.25)^2 - 20(0.25) + 8.5 = 4.21875$$

Notice that this estimate of the slope is much closer to the average value for the interval (4.4375) than the slope at the beginning of the interval (8.5) that would have been used for Euler's approach. The slope at the midpoint can then be substituted into Eq. (25.37) to predict

$$y(0.5) = 1 + 4.21875(0.5) = 3.109375 \qquad \varepsilon_t = 3.4\%$$

For Ralston's method, k_1 for the first interval also equals 8.5 and [Eq. (25.38*b*)]

$$k_2 = -2(0.375)^3 + 12(0.375)^2 - 20(0.375) + 8.5 = 2.58203125$$

The average slope is computed by

$$\phi = \frac{1}{3}(8.5) + \frac{2}{3}(2.58203125) = 4.5546875$$

which can be used to predict

$$y(0.5) = 1 + 4.5546875(0.5) = 3.27734375 \qquad \varepsilon_t = -1.82\%$$

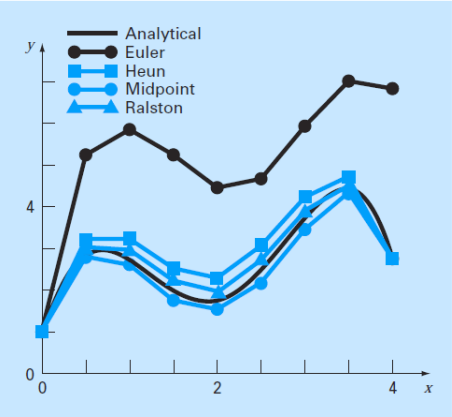


TABLE 25.3 Comparison of true and approximate values of the integral of $y' = -2x^3 + 12x^2 - 20x + 8.5$, with the initial condition that $y = 1$ at $x = 0$. The approximate values were computed using three versions of second-order RK methods with a step size of 0.5.

		Heun		Midpoint		Second-Order Ralston RK	
x	y_{true}	y	$\lvert\varepsilon_t\rvert$ (%)	y	$\lvert\varepsilon_t\rvert$ (%)	y	$\lvert\varepsilon_t\rvert$ (%)
0.0	1.00000	1.00000	0	1.00000	0	1.00000	0
0.5	3.21875	3.43750	6.8	3.109375	3.4	3.277344	1.8
1.0	3.00000	3.37500	12.5	2.81250	6.3	3.101563	3.4
1.5	2.21875	2.68750	21.1	1.984375	10.6	2.347656	5.8
2.0	2.00000	2.50000	25.0	1.75	12.5	2.140625	7.0
2.5	2.71875	3.18750	17.2	2.484375	8.6	2.855469	5.0
3.0	4.00000	4.37500	9.4	3.81250	4.7	4.117188	2.9
3.5	4.71875	4.93750	4.6	4.609375	2.3	4.800781	1.7
4.0	3.00000	3.00000	0	3	0	3.031250	1.0

4.5.3 Fourth order RK (RK4) method

- Perhaps the most popular RK method, is the 4-**stage** ($s = 4$) **fourth order accurate RK4** method below

$$y_{n+1} = y_n + \frac{1}{6}\Delta t(k_1 + 2k_2 + 2k_3 + k_4) \qquad \text{where} \tag{278a}$$

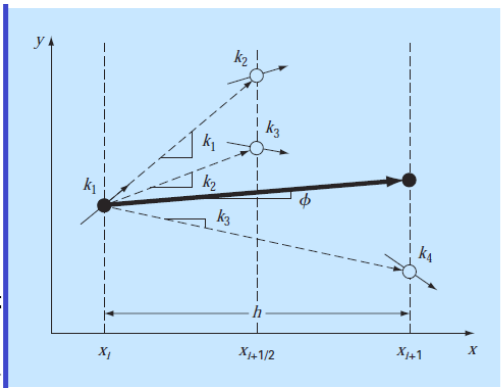
$$\left.\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + \tfrac{1}{2}\Delta t, y_n + \tfrac{1}{2}\Delta t k_1) \\ k_3 &= f(t_n + \tfrac{1}{2}\Delta t, y_n + \tfrac{1}{2}\Delta t k_2) \\ k_4 &= f(t_n + \Delta t, y_n + \Delta t k_3) \end{aligned}\right\} \tag{278b}$$

- When derivative is not a function of y, *i.e.*, when $f(t, y) = f(t)$ the solution to the ODE, is simply the integration of a scalar function.
- In such case, RK4 reduces to the Simpson rule for integration of an interval; *cf.* (168):

$$\text{Quadrature}(\int_0^L f(x)\,\mathrm{d}x) = \frac{L}{6}f(0) + \frac{4L}{6}f(L/2) + \frac{L}{6}f(L)$$

whose Butcher tableau and geometric schematic is shown below,

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$



$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h$$

where

$$k_1 = f(x_i, y_i)$$

$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right)$$

$$k_3 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_2 h\right)$$

$$k_4 = f(x_i + h, y_i + k_3 h)$$

[Chapra and Canale, 2010]

4.5.3.1 An example for RK4 method [Chapra and Canale, 2010]

(a) Use the classical fourth-order RK method [Eq. (25.40)] to integrate

$f(x, y) = -2x^3 + 12x^2 - 20x + 8.5$

using a step size of $h = 0.5$ and an initial condition of $y = 1$ at $x = 0$.

(b) Similarly, integrate

$f(x, y) = 4e^{0.8x} - 0.5y$

using $h = 0.5$ with $y(0) = 2$ from $x = 0$ to 0.5.

Solution.

(a) Equations (25.40*a*) through (25.40*d*) are used to compute $k_1 = 8.5$, $k_2 = 4.21875$, $k_3 = 4.21875$ and $k_4 = 1.25$, which are substituted into Eq. (25.40) to yield

$$y(0.5) = 1 + \left\{\frac{1}{6}[8.5 + 2(4.21875) + 2(4.21875) + 1.25]\right\} 0.5$$
$$= 3.21875$$

which is exact. Thus, because the true solution is a quartic [Eq. (PT7.16)], the fourth-order method gives an exact result.

(b) For this case, the slope at the beginning of the interval is computed as

$$k_1 = f(0, 2) = 4e^{0.8(0)} - 0.5(2) = 3$$

This value is used to compute a value of y and a slope at the midpoint,

$$y(0.25) = 2 + 3(0.25) = 2.75$$
$$k_2 = f(0.25, 2.75) = 4e^{0.8(0.25)} - 0.5(2.75) = 3.510611$$

This slope in turn is used to compute another value of y and another slope at the midpoint,

$$y(0.25) = 2 + 3.510611(0.25) = 2.877653$$
$$k_3 = f(0.25, 2.877653) = 4e^{0.8(0.25)} - 0.5(2.877653) = 3.446785$$

Next, this slope is used to compute a value of y and a slope at the end of the interval,

$$y(0.5) = 2 + 3.071785(0.5) = 3.723392$$
$$k_4 = f(0.5, 3.723392) = 4e^{0.8(0.5)} - 0.5(3.723392) = 4.105603$$

Finally, the four slope estimates are combined to yield an average slope. This average slope is then used to make the final prediction at the end of the interval.

$$\phi = \frac{1}{6}[3 + 2(3.510611) + 2(3.446785) + 4.105603] = 3.503399$$
$$y(0.5) = 2 + 3.503399(0.5) = 3.751699$$

which compares favorably with the true solution of 3.751521.

4.5.4 Butcher effect and higher order RK methods

- From these two results (RK2, RK4) one may be tempted to conclude that the order of accuracy is the same as number of stages s, which is not correct in general.
- The number of unknowns for an s-stage explicit RK method is $s-1$ (b's) + s (c's) + $(s-1)s/2$ (a's) = $(s^2+3s-2)/2$.
- The number of equations grow based on what f terms (and in what manner) appear as factors of Δt^i terms. For example, remember that the third order RK expansion was (269).

$$y(t_{n+1}) = y(t_n) + \Delta t f + \frac{1}{2}\Delta t^2 (f_t + f_y f) + \frac{1}{6}\Delta t^3 \left(f_{tt} + f_t f_y + 2 f f_{ty} + f f_y^2 + f^2 f_{yy}\right) + \mathcal{O}\left(\Delta t^4\right)$$

- Unfortunately, **there is no guarantee that an s-stage RK method will have s order of accuracy** given the different trends the number of equations and unknowns grow and due to the form of the equations.
- For example, if $S(o)$ is the number of RK stages needed for order o we have [Butcher, 1964],

$$N(o) = o \qquad o \leq 4 \tag{279a}$$
$$N(5) = 6 \tag{279b}$$
$$N(6) = 7 \tag{279c}$$

etc..

- That is, the **number of stages $s = N(o)$ increases more rapidly than o!**
- This phenomena is known as the Butcher's effect.
- Given the additional complexity of higher order RK methods and the Butcher's effect (the need of having higher number of stages than order of accuracy) limits the practical uses of higher order RK methods.

- For the fifth order of accuracy, from (279b) we observe $s = N(o) = N(5) = 6$ stages are required.
- Butcher's fifth order, six-stage RK update equation is given in (280).

$$y_{i+1} = y_i + \frac{1}{90}(7k_1 + 32k_3 + 12k_4 + 32k_5 + 7k_6)h$$

where

$$\begin{aligned}
k_1 &= f(x_i, y_i)\\
k_2 &= f\left(x_i + \frac{1}{4}h, y_i + \frac{1}{4}k_1 h\right)\\
k_3 &= f\left(x_i + \frac{1}{4}h, y_i + \frac{1}{8}k_1 h + \frac{1}{8}k_2 h\right)\\
k_4 &= f\left(x_i + \frac{1}{2}h, y_i - \frac{1}{2}k_2 h + k_3 h\right)\\
k_5 &= f\left(x_i + \frac{3}{4}h, y_i + \frac{3}{16}k_1 h + \frac{9}{16}k_4 h\right)\\
k_6 &= f\left(x_i + h, y_i - \frac{3}{7}k_1 h + \frac{2}{7}k_2 h + \frac{12}{7}k_3 h - \frac{12}{7}k_4 h + \frac{8}{7}k_5 h\right)
\end{aligned} \tag{280}$$

Problem Statement. Use first- through fifth-order RK methods to solve

$$f(x, y) = 4e^{0.8x} - 0.5y$$

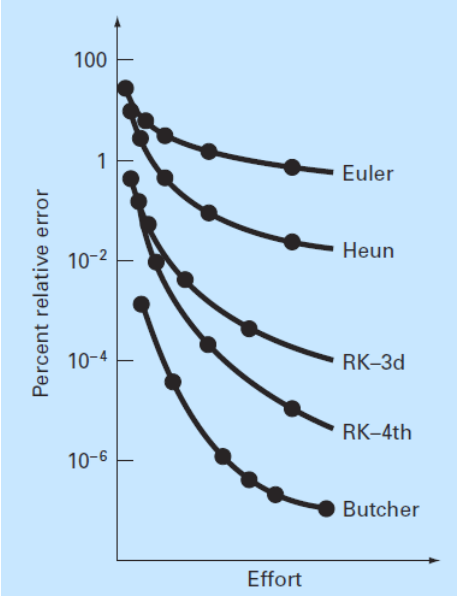
with $y(0) = 2$ from $x = 0$ to $x = 4$ with various step sizes. Compare the accuracy of the various methods for the result at $x = 4$ based on the exact answer of $y(4) = 75.33896$.

Solution. The computation is performed using Euler's, the noniterative Heun, the third-order RK [Eq. (25.39)], the classical fourth-order RK, and Butcher's fifth-order RK methods. The results are presented in Fig. 25.16, where we have plotted the absolute value of the percent relative error versus the computational effort. This latter quantity is equivalent to the number of function evaluations required to attain the result, as in

$$\text{Effort} = n_f \frac{b-a}{h}$$

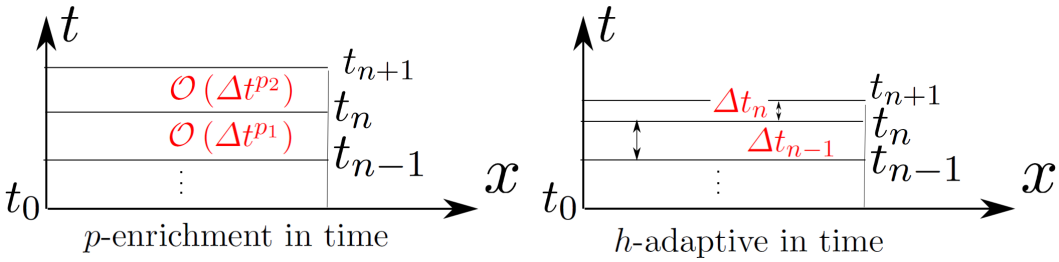
where n_f = the number of function evaluations involved in the particular RK computation. For orders ≤ 4, n_f is equal to the order of the method. However, note that Butcher's fifth-order technique requires six function evaluations [Eq. (25.41*a*) through (25.41*f*)]. The quantity $(b-a)/h$ is the total integration interval divided by the step size—that is, it is the number of applications of the RK technique required to obtain the result. Thus, because the function evaluations are usually the primary time-consuming steps, Eq. (E25.8.1) provides a rough measure of the run time required to attain the answer.

Inspection of Fig. 25.16 leads to a number of conclusions: first, that the higher-order methods attain better accuracy for the same computational effort and, second, that the gain in accuracy for the additional effort tends to diminish after a point. (Notice that the curves drop rapidly at first and then tend to level off.)



4.5.5 Adaptive RK methods

- Given that RK method is a **single-step** method, it is very suitable for adaptive operations in time.
- Since there is no historical data pool, as in LMS methods, we can easily accommodate one of the following two approaches to control the error in each time step:



1. *p*-enrichment (*p*-adaptivity) in time: The order of accuracy is changed from time step to time step. Due to the butcher effect and the need to have more complex RK formulation, adjusting the temporal order of accuracy by RK methods can be impractical for o beyond 5 to 6.
2. *h*-adaptivity in time: The time step can be changed based on local errors from time step t_n to t_{n+1}. This approach is more practical given the wide range of (stable) time steps that can be taking with stepping methods such as RK methods.

- In either case, we need an ***a posteriori* error indicator** to know
  1. when *p*-enrichment or *h*-refinement (when the error is too large) or *p*-reduction or *h*-coarsening (*h* stands for Δt for the time axis) is needed.
  2. which option is more favorable when both p and h options are available. The answer to this question, however is more difficult and in general depends on the regularity of the underlying problem we are solving. Besides for time stepping methods, similar to RK method discussed above, the *p*-enrichment option is often impractical and we are left with only *h*-refinement option. Thus, often we do not need to choose between *h*- or *p*-adaptivity in time.
- ***a posteriori* error indicators**: are obtained by the solution of the same time step (or in general local element, update, *etc.*) by comparing the base solution and a more accurate solution. The larger the difference between the two two solutions, the larger the local error.
- Examples for generating more accurate solutions in time, when time stepping methods are used:
  1. Step-halving methods or more generally schemes that cover the same time interval by two different resolutions of time steps. The one with finer step size, clearly represents the more accurate solution scheme.
  2. Different (successive) orders of accuracy: The same time step is solved with two schemes with successive orders of accuracy. The higher order scheme, clearly models the more accurate solution.
- Another use of *a posteriori* error indicators is the ability to improve the accuracy of the solution / or even local order of accuracy by updating the solution with a factor of the a posteriori error. The ability to use the error to improve the accuracy of the solution, requires some mathematical analysis of the time stepping method.
- Below, we present some excepts from [Chapra and Canale, 2010] section 25.5 that discussed both step-halving and different orders of accuracy approaches for formulating an *a posteriori* error indicator.

> *Step halving* (also called *adaptive RK*) involves taking each step twice, once as a full step and independently as two half steps. The difference in the two results represents an estimate of the local truncation error. If y_1 designates the single-step prediction and y_2 designates the prediction using the two half steps, the error Δ can be represented as
>
> $$\Delta = y_2 - y_1 \tag{25.43}$$
>
> In addition to providing a criterion for step-size control, Eq. (25.43) can also be used to correct the y_2 prediction. For the fourth-order RK version, the correction is
>
> $$y_2 \leftarrow y_2 + \frac{\Delta}{15} \tag{25.44}$$
>
> This estimate is fifth-order accurate.

Problem Statement. Use the adaptive fourth-order RK method to integrate $y' = 4e^{0.8x} - 0.5y$ from $x = 0$ to 2 using $h = 2$ and an initial condition of $y(0) = 2$. This is the same differential equation that was solved previously in Example 25.5. Recall that the true solutions is $y(2) = 14.84392$.

Solution. The single prediction with a step of h is computed as

$$y(2) = 2 + \frac{1}{6}[3 + 2(6.40216 + 4.70108) + 14.11105]2 = 15.10584$$

The two half-step predictions are

$$y(1) = 2 + \frac{1}{6}[3 + 2(4.21730 + 3.91297) + 5.945681]1 = 6.20104$$

and

$$y(2) = 6.20104 + \frac{1}{6}[5.80164 + 2(8.72954 + 7.99756) + 12.71283]1 = 14.86249$$

Therefore, the approximate error is

$$E_a = \frac{14.86249 - 15.10584}{15} = -0.01622$$

which compares favorably with the true error of

$$E_t = 14.84392 - 14.86249 = -0.01857$$

The error estimate can also be used to correct the prediction

$$y(2) = 14.86249 - 0.01622 = 14.84627$$

which has an $E_t = -0.00235$.

Aside from step halving as a strategy to adjust step size, an alternative approach for obtaining an error estimate involves computing two RK predictions of different order. The results can then be subtracted to obtain an estimate of the local truncation error. One shortcoming of this approach is that it greatly increases the computational overhead. For example, a fourth- and fifth-order prediction amount to a total of 10 function evaluations per step. The *Runge-Kutta Fehlberg* or *embedded RK* method cleverly circumvents this problem by using a fifth-order RK method that employs the function evaluations from the accompanying fourth-order RK method. Thus, the approach yields the error estimate on the basis of only six function evaluations!

For the present case, we use the following fourth-order estimate

$$y_{i+1} = y_i + \left(\frac{37}{378}k_1 + \frac{250}{621}k_3 + \frac{125}{594}k_4 + \frac{512}{1771}k_6\right)h \tag{281}$$

along with the fifth-order formula:

$$y_{i+1} = y_i + \left(\frac{2825}{27{,}648}k_1 + \frac{18{,}575}{48{,}384}k_3 + \frac{13{,}525}{55{,}296}k_4 + \frac{277}{14{,}336}k_5 + \frac{1}{4}k_6\right)h \qquad \text{where} \tag{282}$$

$$
\begin{aligned}
k_1 &= f(x_i, y_i) \\
k_2 &= f\left(x_i + \frac{1}{5}h, y_i + \frac{1}{5}k_1 h\right) \\
k_3 &= f\left(x_i + \frac{3}{10}h, y_i + \frac{3}{40}k_1 h + \frac{9}{40}k_2 h\right) \\
k_4 &= f\left(x_i + \frac{3}{5}h, y_i + \frac{3}{10}k_1 h - \frac{9}{10}k_2 h + \frac{6}{5}k_3 h\right) \\
k_5 &= f\left(x_i + h, y_i - \frac{11}{54}k_1 h + \frac{5}{2}k_2 h - \frac{70}{27}k_3 h + \frac{35}{27}k_4 h\right) \\
k_6 &= f\left(x_i + \frac{7}{8}h, y_i + \frac{1631}{55,296}k_1 h + \frac{175}{512}k_2 h + \frac{575}{13,824}k_3 h + \frac{44,275}{110,592}k_4 h + \frac{253}{4096}k_5 h\right)
\end{aligned} \tag{283}
$$

- The ODE is solved by using the fifth order scheme (282).
- *a posteriori* error estimate is obtained by computing the difference between 4^{th} and 5^{th} order solutions at each time step.

Problem Statement. Use the Cash-Karp version of the Runge-Kutta Fehlberg approach to perform the same calculation as in Example 25.12 from $x = 0$ to 2 using $h = 2$.

Solution. The calculation of the k's can be summarized in the following table:

	x	y	f(x, y)
k_1	0	2	3
k_2	0.4	3.2	3.908511
k_3	0.6	4.20883	4.359883
k_4	1.2	7.228398	6.832587
k_5	2	15.42765	12.09831
k_6	1.75	12.17686	10.13237

These can then be used to compute the fourth-order prediction

$$y_1 = 2 + \left(\frac{37}{378}3 + \frac{250}{621}4.359883 + \frac{125}{594}6.832587 + \frac{512}{1771}10.13237\right)2 = 14.83192$$

along with a fifth-order formula:

$$
\begin{aligned}
y_1 = 2 + \Bigg(&\frac{2825}{27,648}3 + \frac{18,575}{48,384}4.359883 + \frac{13,525}{55,296}6.832587 \\
&+ \frac{277}{14,336}12.09831 + \frac{1}{4}10.13237\Bigg)2 = 14.83677
\end{aligned}
$$

The error estimate is obtained by subtracting these two equations to give

$$E_a = 14.83677 - 14.83192 = 0.004842$$

4.5.6 Implicit RK methods

- The stability of explicit RK methods can be studies very similar to LMS methods.
- Similar to that case, explicit RK methods are only conditionally stable.
- Explicit Runge-Kutta methods are unsuitable for stiff systems or problems were mainly the first few modes are excited (*e.g.*, structural dynamic applications) because of their small region of absolute stability. That is, stability stipulates time steps that are much smaller than what is needed from accuracy perspectives for these problems.

- Implicit RK methods with very large regions of absolute stability, on the other hand, can be formulated by having a full matrix a matrix as shown in the following butcher tableau:

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \dots & \dots & \dots & \dots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array} \tag{284}$$

- The update can be written as,

$$y_{n+1} = y_h + \Delta t \sum_{i=1}^{s} b_i k_i \qquad \text{where} \tag{285a}$$

$$k_i = f(t_n + \Delta t c_i, y_n + \Delta t \sum_{j=1}^{s} a_{ij} k_j), \qquad 1 \le i \le s \tag{285b}$$

- To construct an s-stage implicit method ,which are A-stable (A-stability will be discussed in §5)

  1. c_i and b_i are set to the quadrature points and weights, respectively, in the Gauss quadrature formula in the evaluation of polynomials on [0 1],

$$\int_0^1 P(x)\, \mathrm{d}x = \sum_{i=1}^{s} b_i P(c_i) \tag{286}$$

     for polynomials up to order $2s-1$.

  2. The numbers a_{ij} can then be chosen so that the method has order 2s, and is A-stable.

- For example, the butcher tableau,

$$\begin{array}{c|cc} \frac{1}{6}(3-\sqrt{3}) & \frac{1}{4} & \frac{1}{12}(3-2\sqrt{3}) \\ \frac{1}{6}(3+\sqrt{3}) & \frac{1}{12}(3+2\sqrt{3}) & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{287}$$

  defines a 2-stage ($s=2$) A-stable method of order 4.

- Implicit RK methods are rarely used due to the following reasons,
  - Unlike explicit RK methods were k_i could be solved in succession ($k_i = 1, \dots, s$), for implicit RK methods k_i must be solved simultaneously.
  - That is, if we solved an m dof MDOF system with an s-stage implicit RK scheme, we need to solved a coupled system of size $m \times s$ for each time step!
  - This can be a huge drawback both from computational costs and memory perspectives.
  - If $f(t, y)$ is nonlinear in y the solution can become prohibitive as we need to solve now a $m \times s$ coupled system of nonlinear equations for each step update!

- For these reasons implicit Runge-Kutta methods cannot compete in efficiency with the Backward Differentiation methods (which are a group of LMS methods with very large absolute stability region), and their use is almost exclusively limited to stiff systems of ODEs.

5 Mathematical analysis of time marching schemes

5.1 Introduction

An informal overview of these three important topics (before discussing them for different methods and time integration schemes):

- Convergence: The numerical method convergence to the exact solution of the underlying problem as the relevant grid sizes decrease and/or interpolation degree increases. This is an analysis limit type argument, meaning that we can make the numerical solution as close as we want to the exact solution of the underlying equation by choosing small enough grid size(s) and/or interpolation order.

  The concept of the underlying equation is very important. For example for a time integration scheme that solves an FEM discretized equation $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ consistency refers to capturing the analytical solution of the underlying ODE $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ **not** the PDE that the FEM derived ODE $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ is based on. To converge to the exact solution of the underlying PDE: *e.g.*, $\rho A\frac{\mathrm{d}^2u}{\mathrm{d}t^2} - EA\frac{\mathrm{d}^2u}{\mathrm{d}x^2} = q$ for 1D elastodynamics (E is constant), we need to let spatial grid size $h \to 0$ so that the exact solution to the ODE $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ is close enough to $\rho A\frac{\mathrm{d}^2u}{\mathrm{d}t^2} - EA\frac{\mathrm{d}^2u}{\mathrm{d}x^2} = q$ then use small enough time step Δt so that the time-marching based numerical solution of $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ is closed to its exact value. Finally, by using triangular inequality, we can argue that for small enough $h, \Delta t$ the numerical solution to the ODE $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ is close enough to the exact solution of the PDE $\rho A\frac{\mathrm{d}^2u}{\mathrm{d}t^2} - EA\frac{\mathrm{d}^2u}{\mathrm{d}x^2} = q$.

  In short, convergence should be relative to an underlying exact solution, *e.g.,* only the exact solution of the ODE $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ (where only $\Delta t \to 0$ is needed) or the underlying PDE $\rho A\frac{\mathrm{d}^2u}{\mathrm{d}t^2} - EA\frac{\mathrm{d}^2u}{\mathrm{d}x^2} = q$ (where both $\Delta t, h \to 0$ are needed).

  Convergence rate: Is the rate in which the error between numerical and analytical solution goes to zero. For example for a method that the error is $\mathcal{O}(h^p) + \mathcal{O}(\Delta t^s)$ we call th convergence rate in space (h) is p and in time (Δt) is s.

- Consistency: Consistency is a concept that is relevant to step-by-step advancing schemes. This is particularly to any time marching method that advances the solution one time step at a time. Consistency is an easier condition that convergence and only requires that ONE time advance step be "consistent" with the underlying exact solution. It basically requires that for a sufficiently smooth exact solution from time step t_n to t_{n+1} if both exact and numerical solutions start from the same initial condition at t_n the truncation error which is the error at the end of step t_{n+1} between the exact an numerical time integration scheme goes "sufficiently fast" to zero. This "sufficiently fast" will be quantified in the context of different method.

  Consistency condition is a much easier condition to verify than convergence as it includes only algebraic operations. It also deals with once (time) advance step / local truncation error vs. total solution (*e.g.,* final solution time) / and global error which is used in convergence analysis. We will see that consistency is one of the two conditions used to prove convergence.

- Stability For a (time) advancing scheme stability requires that the solution at a time T is bounded by the solution at the initial time with a factor C_T which only depends on the given time T **not** the time step Δt. For a stable underlying PDE/ODE (where the physical solution does not blow up in time), stability requires the numerical solution too does not blow up in time.

Some notes on how these concepts are related:

- Lax-Richtmyer equivalence theorem in FD states that for a consistent FD scheme convergence and stability are equivalent:

$$\text{Consistency} \quad \Rightarrow \quad (\ \text{Stability} \quad \Leftrightarrow \quad \text{Convergence}\ ) \tag{288}$$

- The way this theorem is used in practice is as follows:

$$\text{Consistency and Stability} \quad \Rightarrow \quad \text{Convergence} \tag{289}$$

  - because eventually we want to have convergent numerical methods.
  - However, the proof of convergence is very difficult as we need to consider arbitrary initial and boundary condition and using analysis tools show that the limit of numerical solution as the grid resolution goes to zero (and/or interpolation order increases) the numerical solution tends to the exact solution.
  - The Lax-Richtmyer scheme shows that if we can prove the easier conditions of consistency and stability, which are generally more straightforward and require simple algebraic/arithmetic operations, we prove convergence.
  - Various form of similar theorems exist with other numerical methods, *e.g.,* FEM, FV, DG, *etc.*, where solution is discretized differently in space, yet the same conclusion is made for the dynamic solutions in time: consistency + stability $\Rightarrow$ convergence.
  - The proof and discussion of consistency and stability will be the focus of this section.
  - We will observe that the behavior of local truncation error in consistency verification also determines convergence rate!

5.2 Analysis of direct time integration methods (for FEMs): A sample analysis

Consider the hyperbolic $\mathbf{M}\ddot{\mathbf{U}}+\mathbf{C}\dot{\mathbf{U}}+\mathbf{K}\mathbf{U}=\mathbf{R}$ and parabolic (or two/multi-field first temporal order representation of a hyperbolic) $\mathbf{M}\dot{\mathbf{U}}+\mathbf{K}\mathbf{U}=\mathbf{R}$ n dof ODEs from (226). The analysis of time integration of FEM-based discrete ODEs requires the following steps:

1. Modal reduction to SDOF: We first reduce $\mathbf{M}\ddot{\mathbf{U}}+\mathbf{C}\dot{\mathbf{U}}+\mathbf{K}\mathbf{U}=\mathbf{R}$ or $\mathbf{M}\dot{\mathbf{U}}+\mathbf{K}\mathbf{U}=\mathbf{R}$ to n SDOFs in the form $\ddot{x}+2\xi\omega\dot{x}+\omega^2 x=f(t)$ or $\dot{x}+\lambda x=f(t)$, respectively; *cf.* (229). We show that the analysis of the underling matrix form ODE reduces to the analysis of n SDOFs.

2. Stability of SDOF: For the SDOFs we analyze their stability based on the time step Δt and SDOF parameters ξ,ω (2^{nd}order ODE), and λ for **all** modes 1 to n. If conditionally stable, the maximum time step $\Delta t_{\max}$ is chosen as the **minimum** of all SDOF time steps.

3. Consistency of SDOF: We show that local truncation error $\tau(t_n)$ is $\mathcal{O}\left(\Delta t^s\right)$, for $s>0$. This is only based on analyzing the numerical error for **one** time step.

4. Convergence of SDOF: Using consistency **and** stability results, we prove the convergence of the time integration scheme and show that the temporal convergence rate is k.

Some important considerations are:

- Worst SDOF system (*i.e.,* highest n natural frequency, *etc.*): Finding the worst SDOF that gives the lowest time step (for conditionally stable methods) and in general for error analysis itself is computationally prohibitive; it requires a complete modal analysis which is expensive!

  Fortunately, a simple analysis shows that for example for a second order temporal PDE, the highest natural frequency is **smaller** that the worst case element which is generally the smallest element in the domain. So, in fact, we do not need to solve for an n dof FEM model's modal parameters! We can use the worst case element parameters as conservative estimates. This is the practice for first or second order ODEs discussed above.

- Dissipation, dispersion, and other errors: One important consideration is how much the amplitude of moving waves decreases or basically energy is dissipated with a stable time integration Scheme. Equally important is how the period (or frequency) of a periodic moving wave is modified by the numerical time integration scheme. The latter error is called dispersion or period error. Both errors play important roles in the overall accuracy of the solution. We also comment on some other aspects of numerical error, *i.e.,* features such as overshoot and undershoot.

- We briefly repeat some material from §4.2 and complete the analysis of generalized trapezoidal rule.

- From (230) we consider the solution of an n dof first order ODE obtained by FEM spatial discretization: $\mathbf{M}\dot{\mathbf{d}}+\mathbf{K}\mathbf{d}=\mathbf{F}$ with IC $\mathbf{d}(t=0)=\mathbf{d}_0$.

- The update equation for the time $t=t_n+\alpha\Delta t$ was given by (231). That is, $\dot{\mathbf{d}}^{t_n+\alpha\Delta t}=\frac{\mathbf{d}^{n+1}-\mathbf{d}^n}{\Delta t}$ and $\mathbf{d}^{t_n+\alpha\Delta t}=(1-\alpha)\mathbf{d}^n+\alpha\mathbf{d}^{n+1}$.

- Below we describe how we can analyze the method by reducing it to n SDOF problems.

5.2.1 Generalized trapezoidal rule: Modal reduction to SDOF

- We perform modal analysis for the first order ODE below,

$$
\begin{aligned}
&\boldsymbol{M\dot{d}+Kd=F} \implies \text{Modal eigenproblem}\\
&(\boldsymbol{K}-\lambda_l^h\boldsymbol{M})\boldsymbol{\psi}_l=\boldsymbol{0}, \qquad l\in\{1,2,\ldots,n_{eq}\}\\
&\text{where}\\
&\begin{cases}
0\le\lambda_1^h\le\lambda_2^h\le\cdots\le\lambda_{n_{eq}}^h\\
\boldsymbol{\psi}_l^T\boldsymbol{M}\,\boldsymbol{\psi}_m=\delta_{lm} \qquad \text{(orthonormality)} \implies\\
\qquad\boldsymbol{\psi}_l^T\boldsymbol{K}\boldsymbol{\psi}_m=\lambda_l^h\delta_{lm} \qquad \text{(no sum)}
\end{cases}
\end{aligned}
\tag{290}
$$

  similar to the second order ODE $\mathbf{M}\ddot{\mathbf{U}}+\mathbf{C}\dot{\mathbf{U}}+\mathbf{K}\mathbf{U}=\mathbf{R}$ we observe modes ψ_i are **M**-orthonormal and **K** orthogonal with diagonal values λ_i^h which are natural eigenvalues.

- The total solution $d(t)$ (n_{eq} vector) is expressed in terms of mode m SDOF values $d_{(m)}(t)$,

$$\boldsymbol{d}(t) = \sum_{m=1}^{n_{eq}} d_{(m)}(t)\boldsymbol{\psi}_m \quad \Longrightarrow \quad d_{(l)}(t) = \boldsymbol{\psi}_l^T \boldsymbol{M} \boldsymbol{d}(t) \tag{291}$$

  which also provides how each individual mode $d_{(m)}(t)$ is expressed in terms of $d(t)$.

- This provides a means to find ICs for mode l:

$$\begin{aligned} d_{(l)}(0) &= \boldsymbol{\psi}_l^T \boldsymbol{M}\, \boldsymbol{d}(0) \\ &= \boldsymbol{\psi}_l^T \boldsymbol{M}\, \boldsymbol{d}_0 \\ &\overset{\text{def}}{=} d_{0(l)} \end{aligned} \tag{292}$$

- Finally we get ODE and IC for each SDOF mode which in fact can be solved with **any** appropriate time integration scheme

$$\left.\begin{aligned} \dot{d} + \lambda^h d &= F, \qquad t \in [0, T] \\ d(0) &= d_0 \end{aligned}\right\} \quad \textbf{(SDOF model problem)} \tag{293}$$

- Time integration scheme directly applied to $\mathbf{M}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}$ is **equivalent** to integrating SDOFs with the same integration scheme.

- This can be demonstrated for generalized trapezoidal rule:

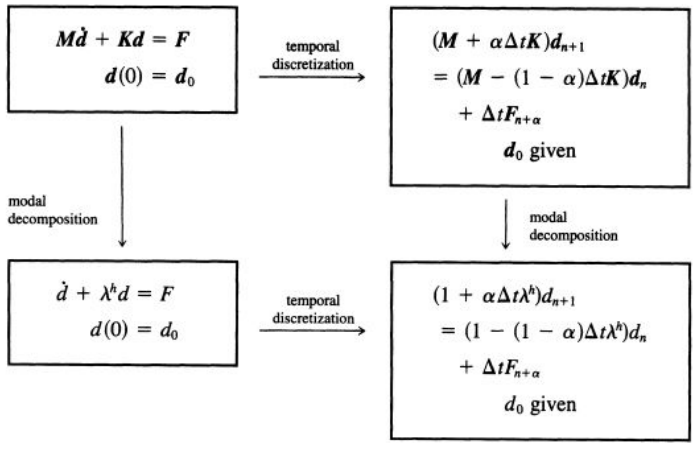
  - Equation (294)(e) is generalized trapezoidal rule applied to $\mathbf{M}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}$ premultiplied ψ_l^{T} where
  - $\mathbf{d}_n$ and $\mathbf{d}_{n+1}$ are expressed in terms of modal components.
  - Now, using $\mathbf{M}$-orthonormal and $\mathbf{K}$ orthogonal (with diagonal values λ_i^h) properties from (290) MDOF generalized trapezoidal method for MDOF system in (294)(e) (premultiplied by ψ_l) results SDOF generalized trapezoidal method for SDOFs in (294)(g).

$$\begin{aligned} \boldsymbol{d}_n &= \sum_{m=1}^{n_{eq}} d_{n(m)}\boldsymbol{\psi}_m && \text{(a)} \\ \boldsymbol{d}_{n+1} &= \sum_{m=1}^{n_{eq}} d_{n+1(m)}\boldsymbol{\psi}_m && \text{(b)} \\ d_{n(l)} &= \boldsymbol{\psi}_l^T \boldsymbol{M} \boldsymbol{d}_n && \text{(c)} \\ d_{n+1(l)} &= \boldsymbol{\psi}_l^T \boldsymbol{M} \boldsymbol{d}_{n+1} && \text{(d)} \\ \sum_{m=1}^{n_{eq}} &[d_{n+1(m)}\boldsymbol{\psi}_l^T(\boldsymbol{M} + \alpha\Delta t\, \boldsymbol{K})\boldsymbol{\psi}_m \\ &- d_{n(m)}\boldsymbol{\psi}_l^T(\boldsymbol{M} - (1-\alpha)\Delta t\boldsymbol{K})\boldsymbol{\psi}_m] = \Delta t\, \boldsymbol{\psi}_l^T \boldsymbol{F}_{n+\alpha} && \text{(e)} \\ \boldsymbol{F}_{n+\alpha} &= (1-\alpha)\boldsymbol{F}_n + \alpha \boldsymbol{F}_{n+1} && \text{(f)} \\ (1 + \alpha\Delta t\lambda_l^h)d_{n+1(l)} &= (1 - (1-\alpha)\Delta t\lambda_l^h)d_{n(l)} + \Delta t F_{n+\alpha(l)} && \text{(g)} \end{aligned} \tag{294}$$

- Thus, solution of MDOF $\mathbf{M}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}$ with generalized trapezoidal rule reduces to solving the following SDOF equations again with generalized trapezoidal rule

$$\left.\begin{aligned} (1 + \alpha\Delta t\lambda^h)d_{n+1} &= (1 - (1-\alpha)\Delta t\lambda^h)d_n + \Delta t F_{n+\alpha} \\ d_0 &\text{ given} \end{aligned}\right\} \quad \textbf{(temporally discretized SDOF model problem)} \tag{295}$$

  - The same can be shown for basically any ODE time integration scheme.
  - Basically, it does not matter if we first do modal decomposition, then apply generalized trapezoidal integration to SDOFs OR first apply generalized trapezoidal integration then modal decomposition, as shown in the figure:



- Error analysis, stability analysis, *etc.* of MDOF also reduces to the analysis of SDOF.
- For this reason we define the following for MDOF solution

$$\mathbf{d}_n = \text{Numerical vector solution for MDOF at time step } t_n \tag{296a}$$

$\mathbf{d}(t_n)$ = Exact vector solution for MDOF at time $t = t_n$ (evaluated at same dofs) (296b)

$e(t_n) = \mathbf{d}_n - \mathbf{d}(t_n)$ = Error vector for MDOF numerical ODE solution for relative to exact solution (296c)

- Similarly, we define numerical, exact, and error values for SDOF number l

$$e_{(l)}(t_n) = d_{n(l)} - d_{(l)}(t_n) \tag{297}$$

- We observe that MDOF error norm squared with kernel $\mathbf{M}$ is equal to the sum of squares of individual SDOF errors:

$$\begin{aligned} \boldsymbol{e}(t_n)^T \boldsymbol{M} \boldsymbol{e}(t_n) &= \sum_{l,m=1}^{n_{eq}} (e_{(l)}(t_n)\boldsymbol{\psi}_l)^T \boldsymbol{M} (e_{(m)}(t_n)\boldsymbol{\psi}_m) \\ &= \sum_{l,m=2}^{n_{eq}} e_{(l)}(t_n) e_{(m)}(t_n) \boldsymbol{\psi}_l^T \boldsymbol{M} \boldsymbol{\psi}_m \\ &= \sum_{l,m=1}^{n_{eq}} e_{(l)}(t_n) e_{(m)}(t_n) \delta_{lm} \qquad \text{(orthonormality)} \\ &= \sum_{l=1}^{n_{eq}} (e_{(l)}(t_n))^2 \end{aligned} \tag{298}$$

- Thus, the convergence of MDOF system (with $\mathbf{M}$ kernel) which requires $\mathbf{e}(t_n)^{\mathrm{T}}\mathbf{M}\mathbf{e}(t_n)$ is equivalent to individual convergence of SDOFs:

$$\boldsymbol{e}(t_n)^T \boldsymbol{M} \boldsymbol{e}(t_n) \longrightarrow 0 \text{ if and only if } e_{(l)}(t_n) \longrightarrow 0 \text{ for each } l \in \{1, 2, \ldots n_{eq}\} \tag{299}$$

- Finally, given that $\mathbf{M}$ is positive definite convergence of norm square with kernel $\mathbf{M}$ is equivalent to L2 convergence of $\mathbf{e}(t_n)$:

$$\boldsymbol{e}(t_n)^T \boldsymbol{M} \boldsymbol{e}(t_n) \longrightarrow 0 \text{ if and only if } \boldsymbol{e}(t_n) \longrightarrow \mathbf{0} \tag{300}$$

- **L2 Convergence of MDOF error is equivalent to convergence of scalar SDOF** $\Rightarrow$
- For convergence of MDOF we only need to investigate convergence of all SDOFs.

5.2.2 Generalized trapezoidal rule: Stability of SDOF

- As we observe, solution and even convergence analysis of MDOF $\mathbf{M}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}$ reduces to the solution and convergence analysis of SDOFs.
- To analyze the stability of the method, we first study how the exact solution behaves for a given modal value λ^h.

$$\dot{d} + \lambda^h d = 0 \tag{301}$$

- which has the solution:

$$d(t_{n+1}) = \exp\left(-\lambda^h (t_{n+1} - t_n)\right) d(t_n) \tag{302}$$

- The exact numerical solution is stable basically when $\lambda^h \geq 0$:

$$\left.\begin{aligned} |d(t_{n+1})| < |d(t_n)|, \qquad & \lambda^h > 0 \\ d(t_{n+1}) = d(t_n), \qquad & \lambda^h = 0 \end{aligned}\right\} \tag{303}$$

- To study the stability of the numerical method, we find the update of (301) based on a SDOF generalized trapezoidal rule; *cf.* (294)(g),

$$(1 + \alpha \Delta t\, \lambda^h) d_{n+1} = (1 - (1 - \alpha)\Delta t\, \lambda^h) d_n \tag{304}$$

- Which is,

$$d_{n+1} = A d_n, \qquad \text{where} \quad A = \frac{1-(1-\alpha)\Delta t\lambda}{1+\alpha\Delta t\lambda} \quad \text{Amplification factor} \tag{305}$$

- which from (305) we observe,

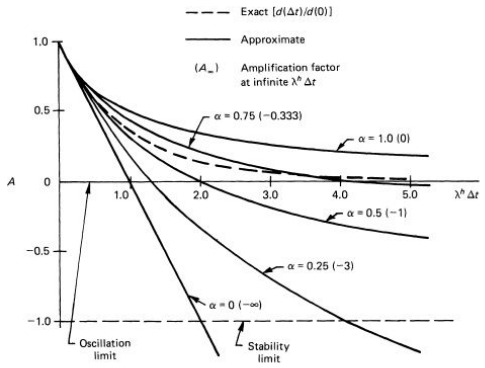
$$d^n = A^n d^n \tag{306}$$

- Clearly, numerical method is stable iff $A \leq 1$.
- now given that the exact solution is only stable for $\lambda^h \geq 0$, we are looking for conditions in which the numerical method is convergence for the same λ^h for which exact solution is stable.
- That is, we consider the condition $\lambda^h \geq 0$ and look for Δt for which $A \leq 1$:

$$|A| \leq 1 \qquad \Rightarrow \qquad -1 \leq \frac{1-(1-\alpha)\Delta t\lambda}{1+\alpha\Delta t\lambda} \leq 1 \tag{307}$$

which results in the following conditions:

$$\begin{cases} \alpha < \frac{1}{2} & \text{Conditionally stable} \qquad \Delta t\lambda < \frac{2}{1-2\alpha} \\ \alpha \geq \frac{1}{2} & \text{Unconditionally stable} \end{cases} \tag{308}$$



Amplification factor for typical one-step methods.

Summary: Stability for the generalized trapezoidal methods

Amplification factor: $A = \frac{1-(1-\alpha)\Delta t\lambda^h}{1+\alpha\Delta t\lambda^h}$

Stability requirement: $|A| < 1$ for $\lambda^h = \lambda^h_{n_{eq}}$ (= maximum eigenvalue)

Unconditional stability: $\alpha \geq \frac{1}{2}$

Conditional stability: $\alpha < \frac{1}{2}, \qquad \Delta t < \frac{2}{(1-2\alpha)\lambda^h_{n_{eq}}}$

- The maximum stable time stable can be chosen as follows

$$\alpha \geq \frac{1}{2} \quad \Rightarrow \quad \text{any } \Delta t \qquad \text{Unconditional stability} \tag{309a}$$

$$\alpha > \frac{1}{2} \quad \Rightarrow \quad \Delta t \leq \Delta t_{\max} = \frac{1}{\max_l(\lambda_l^h)} \frac{2}{1-2\alpha} \qquad \text{Conditional stability} \tag{309b}$$

- where $\max_l(\lambda_l^h)$ is the maximum modal eigen value obtained from modal analysis. The maximum value is chosen because smaller model eigenvalues result in more loose time constraint.
- In practice it is difficult/computationally expensive to actually compute $\max_l(\lambda_l^h)$ by a modal analysis.
- Instead, we can use

$$\lambda_e^m := \max_{e,i}(\lambda_e^i h) = \text{maximum of all element's } (e) \text{ maximum modal eigenvalue (index (i) in element)} \tag{310}$$

  - which is the maximum modal eigenvalue that any element can produce.
  - This value is very easy to be computed and values are already obtain for various element types in the literature.
  - λ_e^m It only depends on element size (geometry) and material properties.

- One can prove [Hughes, 2012, Bathe, 2006],

$$\boxed{\lambda_e^m \geq \max_l(\lambda_l^h)} \tag{311}$$

- that is, maximum eigenmode of all individual elements is larger than the maximum eigenmode of the MDOF system $\max_l(\lambda_l^h)$.

- That results in,

$$\Delta t \le \frac{1}{\lambda_e^m}\frac{2}{1-2\alpha} \quad \Rightarrow \quad \Delta t \le \Delta t_{\max} = \frac{1}{\max_l(\lambda_l^h)}\frac{2}{1-2\alpha} \quad \text{because} \quad \lambda_e^m \ge \max_l(\lambda_l^h) \tag{312}$$

- So, the more stringent condition $\Delta t \le \frac{1}{\lambda_e^m}\frac{2}{1-2\alpha}$ is sufficient for stable time step when $\alpha < \frac{1}{2}$.
- In practice, we use this more conservative value for choosing Δt of FEM time integration schemes because λ_e^m can easily be computed.
- So, we often conservatively reformulate (309) as,

$$\alpha \ge \frac{1}{2} \quad \Rightarrow \quad \text{any } \Delta t \qquad \text{Unconditional stability} \tag{313a}$$

$$\alpha < \frac{1}{2} \quad \Rightarrow \quad \Delta t \le \frac{1}{\lambda_e^m}\frac{2}{1-2\alpha} \qquad \text{Conditional stability} \tag{313b}$$

- The same process can be applied to other time integration schemes.
- Later, we provide analytical formulas for λ_e^m for some types of elements.
- As a note, we observe that,

$$\lambda_e^m \propto \frac{c}{h_{\min}} \qquad \text{Simple hyperbolic PDE, } e.g., \ u_{,tt} - c^2\nabla.\nabla u = 0,\ c = \text{wave speed} \tag{314a}$$

$$\lambda_e^m \propto \frac{D}{h_{\min}^2} \qquad \text{Simple parabolic problem, } e.g., \ u_{,t} - D\nabla.\nabla u = 0,\ D = \text{damping coefficient} \tag{314b}$$

where $h_{\min}$ is the minimum element size. For simplicity here it is assumed the domain is covered with the same element type. For more general cases, we need to consider the maximum eigen-frequency of all elements which may not necessarily correspond to that of the smallest element.

- From (313) and (314) we reach to the same conclusion we had reach for simple hyperbolic and parabolic PDEs with FD methods:

$$\Delta t \le C \begin{cases} h_{\min} & \text{Simple hyperbolic PDE, } e.g., \ u_{,tt} - c^2\nabla.\nabla u = 0 \\ h_{\min}^2 & \text{Simple parabolic PDE, } e.g., \ u_{,t} - D\nabla.\nabla u = 0 \end{cases} \tag{315a}$$

where C depends on material properties, $e.g.$, c, D, and the particular form of time integration scheme.

- The stable time step for more complex dynamic systems, $e.g.$, $\tau u_{,tt} + u_{,t} - D\nabla.\nabla u = 0$ will be discussed later.

5.2.3 Generalized trapezoidal rule: Consistency of SDOF

- In stability analysis, for brevity we assumed there was no source term $f = 0$. However, amplification factor A also applies to that term and if the physical system with f remains bounded so does the numerical one.
- Now for consistency analysis we provide the analysis for the more general case where $f \neq 0$:

$$d_{n+1} = Ad_n + L_n \quad \Rightarrow \quad \boxed{d_{n+1} - Ad_n - L_n = 0} \tag{316a}$$

$$L_n = \Delta t \frac{(1-\alpha)f_n + \alpha f_{n+1}}{1 + \alpha\Delta t\lambda} \tag{316b}$$

- As before A is the amplification factor. L_n is the load contribution from f at time step n.
- We insert the exact solution on the LHS of (316a) and obtain:

$$d(t_{n+1}) - Ad(t_n) - L_n = \Delta t\ \tau(t_n) \tag{317}$$

- Compared with (316a), where we have plugged in numerical solutions d_n, d_{n+1}, in (317) we have plugged in the exact solutions $d(t_{n+1})$ and $d(t_n)$.
- The term on the RHS, is a consequence of the exact solution not satisfying the numerical update equation.

- The error term is expected as we observed from (301) and (305) have different solutions and updates from t_n to t_{n+1}.
- We factor out one Δt from the time step update error on the RHS of (317) as that is needed for convergence analysis.
- $\tau(t_n)$ is called **truncation error**.
- consistency, and order of accuracy are defined as,

$$\text{Consistency}: \quad \text{If } \tau(t_n) = \mathcal{O}\left(\Delta t^k\right) \text{ for all } t \in [0,T],\ k > 0 \tag{318a}$$

$$\text{Order of accuracy / rate of convergence}: \quad k \tag{318b}$$

  where T is the final time considered.

- We will shortly see why k is called the order of accuracy of the time integration scheme.
- As we will see through the analysis of generalized trapezoidal method determination of k is just a (tedious) arithmetic calculation that often requires Taylor expansion of the equation for the exact solution, *e.g.,* (317).
- Generalized trapezoidal method is **consistent** and $k = 1$ for $\alpha \in [0,1]$ except $alpha = \frac{1}{2}$ (trapezoidal method) for which $k = 2$.
- For the proof, we use Taylor series expansion of $d(t_{n+1})$ and $d(t_n)$ terms in (317) around $t_n + \alpha\Delta t$:

$$\begin{aligned}
d(t_{n+1}) &= d(t_{n+\alpha}) + (1-\alpha)\Delta t\dot{d}(t_{n+\alpha}) \\
&\quad + \frac{((1-\alpha)\Delta t)^2}{2}\ddot{d}(t_{n+\alpha}) + \frac{((1-\alpha)\Delta t)^3}{3!}\dddot{d}(t_{n+\alpha}) + O(\Delta t^4) \\
d(t_n) &= d(t_{n+\alpha}) + (-\alpha\Delta t)\,\dot{d}(t_{n+\alpha}) + \frac{(-\alpha\Delta t)^2}{2}\ddot{d}(t_{n+\alpha}) \\
&\quad + \frac{(-\alpha\Delta t)^3}{3!}\dddot{d}(t_{n+\alpha}) + O(\Delta t^4)
\end{aligned} \tag{319}$$

- Plugging A from (305) in (317) and multiplying both sides by $(1+\alpha\Delta t\lambda^h)$:

$$\begin{aligned}
&\Delta t(1+\alpha\Delta t\lambda^h)\,\tau(t_n) \\
&\quad = (1+\alpha\Delta t\lambda^h)\,d(t_{n+1}) - (1-(1-\alpha)\Delta t\lambda^h)\,d(t_n) - \Delta t F_{n+\alpha} \\
&\quad = \{(1+\alpha\Delta t\lambda^h) - (1-(1-\alpha)\Delta t\lambda^h)\}d(t_{n+\alpha}) \\
&\qquad + \{(1+\alpha\Delta t\lambda^h)(1-\alpha)\Delta t - (1-(1-\alpha)\Delta t\lambda^h)(-\alpha\Delta t)\}\dot{d}(t_{n+\alpha}) \\
&\qquad + \left\{(1+\alpha\Delta t\lambda^h)\frac{((1-\alpha)\Delta t)^2}{2} - (1-(1-\alpha)\Delta t\lambda^h)\frac{(-\alpha\Delta t)^2}{2}\right\}\ddot{d}(t_{n+\alpha}) \\
&\qquad + \left\{(1+\alpha\Delta t\lambda^h)\frac{((1-\alpha)\Delta t)^3}{3!} - (1-(1-\alpha)\Delta t\lambda^h)\frac{(-\alpha\Delta t)^3}{3!}\right\}\dddot{d}(t_{n+\alpha}) \\
&\qquad - \Delta t\Bigg([\alpha+(1-\alpha)]F(t_{n+\alpha}) + [\alpha(1-\alpha)\Delta t + (1-\alpha)(-\alpha\Delta t)]\dot{F}(t_{n+\alpha}) \\
&\qquad + \left\{\alpha\frac{((1-\alpha)\Delta t)^2}{2} + (1-\alpha)\frac{(-\alpha\Delta t)^2}{2}\right\}\ddot{F}(t_{n+\alpha}) \\
&\qquad + \left\{\alpha\frac{((1-\alpha)\Delta t)}{3!} + (1-\alpha)\frac{(-\alpha\Delta t)^3}{3!}\right\}\dddot{F}(t_{n+\alpha})\Bigg) + O(\Delta t^4)
\end{aligned} \tag{320}$$

- Using $\dot{d} + \lambda^h d = F$ it can be shown,

$$\tau = (1-2\alpha)\mathcal{O}\left(\Delta t^1\right) + \mathcal{O}\left(\Delta t^2\right) \quad \Rightarrow \quad \begin{cases} k = 2 \quad \alpha = \frac{1}{2} \quad \text{(trapezoidal rule)} \\ k = 1 \quad \text{otherwise} \end{cases} \tag{321}$$

5.2.4 Generalized trapezoidal rule: Convergence of SDOF

- Remembering Lax-Richtmyer equivalence theorem for FD methods, we asserted in (288) that for a consistent method stability and convergence are equivalent.
- As shown in (289), in practice we prove the convergence of a method by establishing that it is both consistent and stable:

$$\text{Consistency and Stability} \quad \Rightarrow \quad \text{Convergence}$$

- Below we prove this for a SDOF problem with $\lambda^h \geq 0$ for a general 1-step time integration scheme in the form of (316a).
- Let $t_n = n\Delta t$ be fixed but Δt be allowed to vary. Assume the time integration is,
  1. stable, *i.e.*, $|A| \leq 1$.
  2. consistent, *i.e.*, there exists a $k > 0, c \geq 0$ such that $|\tau(t)| \leq c\Delta t^k$ for all $t \in [0, T]$; *cf.* (318a).

  Then the method is **convergent** ($e(t_n) \to 0$ as $\Delta t \to 0$) with the rate of convergence k.

Proof:

- First we want to form an update equation for the error from time step t_n to t_{n+1}:

$$\left.\begin{aligned} d_{n+1} - Ad_n - L_n &= 0 && \textit{cf.}\ (316a)\\ d(t_{n+1}) - Ad(t_n) - L_n &= \Delta t\tau(t_n) && \textit{cf.}\ (317)\\ e(t_{n+1}) = d_{n+1} - d(t_{n+1}),\quad e(t_n) &= d_n - d(t_n) && \text{Definition of error; } \textit{cf.}\ (296c) \end{aligned}\right\} \quad \Rightarrow \quad \boxed{e(t_{n+1}) = Ae(t_n) - \Delta t.\tau(t_n)} \tag{322}$$

- By using $n-1$ instead of n in equation (322) (*i.e.*, previous time step) we obtain,

$$\begin{aligned} &e(t_n) = Ae(t_{n-1}) - \Delta t.\tau(t_{n-1}) \quad \text{and knowing} \quad e(t_{n+1}) = Ae(t_n) - \Delta t.\tau(t_n) \quad \text{from (322)} \quad \Rightarrow \\ &e(t_{n+1}) = A^2 e(t_{n-1}) - \Delta t A\tau(t_{n-1}) - \Delta t\tau(t_n) \end{aligned}$$

- By repeating this equation to eliminate $e(t_{n-1})$ from the RHS (by writing (322) for $n \to n-2$) we obtain,

$$e(t_{n+1}) = A^3 e(t_{n-2}) - \Delta t A^2\tau(t_{n-2}) - \Delta t A\tau(t_{n-1}) - \Delta t\tau(t_n)$$

  and so on,

- So we would have,

$$\boxed{e(t_{n+1}) = A^{n+1}e(t_0) - \Delta t\sum_{i=0}^{n} A^i\tau(t_{n-i})} \tag{323}$$

- But $e(t_0) = 0$ because we initialize the time marching scheme at the first step with the exact solution, *i.e.*, IC.
- Expressing (323) for time step t_n instead of t_{n+1} and taking its absolute value we obtain,

$$\begin{aligned} |e(t_n)| &= \Delta t\left|\sum_{i=0}^{n-1} A^i\tau(t_{n-1-i})\right| && \text{(a)}\\ &\leq \Delta t\sum_{i=0}^{n-1} |A|^i\,|\tau(t_{n-1-i})| && \text{(b)}\\ &\leq \Delta t\sum_{i=0}^{n-1} |\tau(t_{n-1-i})| \quad \text{(stability)} && \text{(c)}\\ &\leq t_n \max|\tau(t)| \quad t \in [0, T] && \text{(d)}\\ &\leq t_n c\Delta t^k \quad \text{(consistency)} && \text{(e)} \end{aligned} \tag{324}$$

- We observe,

1. A stable SDOF one step time integration scheme is convergent iff it is stable and consistent (the converse, *i.e.,* convergence ⇒ stability and consistency is trivial; we only showed that stability and consistency proved convergence). Compare this with slightly different versions (288) and (289).
2. We observe, rate of convergence is the same as k in the definition of consistency in (317) ($d(t_{n+1})-Ad(t_n)-L_n = \Delta t\ \tau(t_n)$).
3. The extra Δt that we introduced in the definition of consistency condition on the RHS in (317) ($d(t_{n+1}) - Ad(t_n) - L_n = \Delta t\tau(t_n)$) is needed. Otherwise in (324)(d) we would have got $e(t_n) \leq n\max|\tau(t)|$, which clearly makes the RHS unbounded as we can have a very small time step Δt so that in $t_n = n\Delta t$, $n \to \infty$.
4. The bound on the error term in (324)(e) can be written as,

$$|e(t_n)| \leq C_{t_n}\Delta t^k, \qquad \text{for a fixed } t_n = n\Delta t, \text{where} \quad C_{t_n} = ct_n \tag{325}$$

we observe,

(a) We observe C_{t_n} in general **depends** on the time value t_n and can grow with the the observation time t_n.
(b) But for a **fixed** time t_n the error is bounded no matter what time step value (assuming stability is satisfied) is used!
(c) So, the error constant in general depends on time in convergence analysis, **but must NOT depend on the time step size** Δt.

SDOF to MDOF convergence rate:

– From (298) we have,

$$\mathbf{e}(t_n)^{\mathrm{T}}\mathbf{M}\mathbf{e}(t_n) = \sum_{i=1}^{n}\left(e_{\{i\}}(t_n)\right)^2 \tag{326}$$

where from (325) we know that all SDOF problems $i = 1$ to n have convergence rate of k for their error $e_{\{i\}}(t_n)$ if their local truncation convergence order is k and stable time step is used for all of them.

– As mentioned before, if time integration is conditionally stable, by using the most stringent time-step (from the highest λ_i^h we ensure that all SDOFs are stable.

– In addition, if we directly integrate the underlying MDOF with time step Δt it is equivalent to integrating all SDOFs with time step Δt.

– Given that all SDOFs have the same convergence rates but potentially different constants $(C_{t_n})_i$ we bound the RHS of (326) from (325) in the form,

$$\mathbf{e}(t_n)^{\mathrm{T}}\mathbf{M}\mathbf{e}(t_n) = \sum_{i=1}^{n}\left(e_{\{i\}}(t_n)\right)^2 \leq \sum_{i=1}^{n}(C_{t_n})_i^2\Delta t^{2k} = C_{t_n}^2\Delta t^{2k}, \qquad \text{where} \qquad C_{t_n}^2 = \sum_{i=1}^{n}(C_{t_n})_i^2 \qquad \Rightarrow$$
$$\sqrt{\mathbf{e}(t_n)^{\mathrm{T}}\mathbf{M}\mathbf{e}(t_n)} \leq C_{t_n}\Delta t^k \tag{327}$$

– but we know,

$$L_2(\mathbf{e}(t_n)) = \sqrt{\mathbf{e}(t_n).\mathbf{e}(t_n)} \leq \frac{1}{m_{\min}}\sqrt{\mathbf{e}(t_n)^{\mathrm{T}}\mathbf{M}\mathbf{e}(t_n)} \leq \frac{C_{t_n}}{m_{\min}}\Delta t^k \tag{328}$$

where $m_{\min} > 0$ is the minimum eigenvalue of $\mathbf{M}$. How do we get the first inequality in (328) and why $m_{\min} > 0$?

– Basically L_2 norm and norm with $\mathbf{M}$ are equivalent ⇒

– from (328) we observe that MDOF $\mathbf{M}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}$ error vector $\mathbf{e}(t_n)$ converges with order k in both L_2 norm and norm with $\mathbf{M}$ kernel provided that the integration scheme is stable and consistent with the rate k.

5.3 Stability analysis of SDOF problems involving matrix update equation

- In general we can have update equations from time step t_n to time step t_{n+1} which has an update equation of the form,
- As mentioned before, stability analysis of a MDOF problem, reduces to the stability analysis of its modal SDOFs:

MDOF	⇒	SDOF	
$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$		$\ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = f(t)$	(329b)
$\mathbf{M}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$		$\dot{x} + \lambda x = f(t)$	(329c)

- The stability, consistency, and convergence of (329c) was discussed in §5.2 in the context of generalized trapezoidal method.
- Herein, we analyze the stability of time marching methods from §4.1. by analyzing their corresponding SDOF from (329b).

- In the analysis of higher order ODEs, multi-step method methods with more than 2 steps, and many other instances we deal with update equations of the form,

$$^{t+\Delta t}\hat{\mathbf{X}} = \mathbf{A}\ ^{t}\hat{\mathbf{X}} + \mathbf{L}(^{t+\nu}r) \tag{330}$$

  where $^{t+\Delta t}\hat{X}$ and $^{t}\hat{X}$ correspond to generalized vector update values for time steps t_n and t_{n+1} and

- **A** is the **matrix** amplification factor.

- Examples of $^{t}\hat{X}$ are

  1. Value and subsequent time derivatives of x in (329b): $^{t}\hat{X} = [^{t}x\ ^{t}\dot{x}\ ^{t}\ddot{x}]$. Examples be from Newmark and θ-Wilson methods.
  2. Value and previous step values of x in (329b): $^{t}\hat{X} = [^{t+\Delta t}x\ ^{t}x\ ^{t-\Delta t}x\ \cdots]$. This will be the form of $^{t}\hat{X}$ for LMS methods.

- In either case, since $^{t}\hat{X}$ is a vector, unlike (316a) ($d_{n+1} = Ad_n + L_n$) where the update equation was for a scalar variable d_{n+1} and involved a scalar amplification factor A, in (330) **A** is a matrix.

- Applying (330) multiple times we obtain,

$$\begin{aligned} ^{t+n\Delta t}\hat{\mathbf{X}} = \mathbf{A}^n\ ^{t}\hat{\mathbf{X}} + \mathbf{A}^{n-1}\mathbf{L}(^{t+\nu}r) + \mathbf{A}^{n-2}\mathbf{L}(^{t+\Delta t+\nu}r) + \cdots \\ + \mathbf{A}\,\mathbf{L}(^{t+(n-2)\Delta t+\nu}r) + \mathbf{L}(^{t+(n-1)\Delta t+\nu}r) \end{aligned} \tag{331}$$

- For the moment, by assuming the force operator $\mathbf{L} = 0$ we obtain,

$$^{t+n\Delta t}\hat{\mathbf{X}} = \mathbf{A}^n\ ^{t}\hat{\mathbf{X}} \tag{332}$$

- The stability of the time marching scheme requires $\mathbf{A}^n$ does not blow up.

- For the moment assume **A** is diagonalizable: $\mathbf{A} = \mathbf{PJP}^{-1}$ ($\mathbf{P} = [\mathbf{p}_1\ \mathbf{p}_2 \cdots \mathbf{p}_m]$ is the matrix of right eigenvectors $\mathbf{p}_i$ and $\mathbf{J} = \text{diag}(a_1, \cdots, a_m)$ and a_i are the corresponding eigenvalues. The matrix **A** is $m \times m$.

- In this case, we have

$$\mathbf{A} = \mathbf{PJP}^{-1} \qquad \Rightarrow \mathbf{A}^n = \mathbf{PJ}^n\mathbf{P}^{-1} = \mathbf{P}\begin{bmatrix} a_1^n & & & \\ & a_2^n & & \\ & & \ddots & \\ & & & a_m^n \end{bmatrix}\mathbf{P}^{-1} \qquad \text{for diagonal } \mathbf{J} \tag{333}$$

- Recalling the definition of spectral radius (45)

$$\rho(\mathbf{A}) = \max\{|a_i|\ i \in \{1, \cdots, m\} | a_i \text{ are eigenvalues of } A\} \tag{334}$$

- Clearly, from the definition (334) the stability condition for a **diagonalizable** update matrix A is,

$$\text{Update by } \textbf{diagonalizable} \text{ matrix amplification factor } \mathbf{A} \text{ is stable iff} \qquad \rho(\mathbf{A}) \leq 1 \tag{335}$$

- Now, what happens if **A** is **not** diagonalizable and **when A** is **not** diagonalizable: **A** is **not** diagonalizable iff

  1. Eigenvalues a_k with eigenvalues with $n_k^A > 1$ **algebraic** multiplicity.
  2. Smaller **geometric** multiplicity $n_k^G < n_k^A$

- Basically, the matrix **A** is diagonalizable if it has repeated eigenvalues whose geometric multiplicity is smaller than algebraic one. Clearly, if **A** has distinct eigenvalues all algebraic and geometric multiplicities are are one and it's diagonalizable. Also, if it has repeated eigenvalues but with the same geometric and algebraic multiplicity that is not a problem either. A good example is α multiple of identity matrix $\alpha\mathbf{I}$ which clearly is already diagonal.

- So, what can we do if **A** is not diagonalizable?

- Can we still write $\mathbf{A} = \mathbf{PJP}^{-1} \Rightarrow \mathbf{A} = \mathbf{PJ}^n\mathbf{P}^{-1}$ for **J** being a more manageable matrix than **A**?

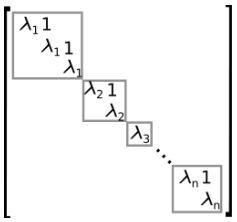
- The answer is yes. The transformation is done by Jordan normal form.
- Any complex matrix **A** can be decomposed in the form,

$$\text{for } \textbf{arbitrary } \mathbf{A}: \quad \mathbf{A} = \mathbf{PJP}^{-1}, \quad \text{where } \mathbf{J} \text{ is a } \textbf{Jordan normal form} \text{ that is similar to } \mathbf{A} \text{ by } \mathbf{P} \tag{336}$$

- where **Jordan normal form** J is a block diagonal matrix of the form,

$$J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{bmatrix} \quad \text{where each block } J_i \text{ is a square matrix of the form} \quad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix} \tag{337}$$

- an example can be seen below,



- Some comments about Jordan normal form:
  - For a given eigenvalue a_i with geometric multiplicity n_i^G (dimension of $\text{Ker}(\mathbf{A} - a_i\mathbf{I})$ there are **exactly** n_i^G Jordan blocks.
  - If the algebraic multiplicity of a_i n_i^A is equal to n_i^G then all these blocks are simply a 1×1 matrix with a_i.
  - If on the other hand $n_i^G < n_i^A$ then we can arrange $n_i^G - 1$ of blocks to be simply 1×1 Jordan blocks of a_i and the last block be a $(n_i^A - n_i^G + 1) \times (n_i^A - n_i^G + 1)$ Jordan block with $n_i^A - n_i^G$ super-diagonal values of 1.
- As an example consider,

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & 2 & 1 \\ 0 & 1 & -1 & -1 \\ -1 & -1 & 3 & 0 \\ 1 & 1 & -1 & 2 \end{bmatrix} \quad \Rightarrow$$

$a_1 = 1, a_2 = 2, a_3 = 4$(with **algebraic** multiplicity 2 $[n_3^A = 2]$ but **geometric** multiplicity 1 $[n_3^G = 1]$)

there is a $\mathbf{P}$ matrix: $\mathbf{A} = \mathbf{PJP}^{-1}$ such that $\mathbf{J} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}$

- So, the question of stability of update in (332) ${}^{t+n\Delta t}\hat{\mathbf{X}} = \mathbf{A}^{n\,t}\hat{\mathbf{X}}$.
- We have

$$\mathbf{A} = \mathbf{PJP}^{-1} \qquad \Rightarrow \qquad \mathbf{A}^n = \mathbf{PJ}^n\mathbf{P}^{-1}$$

- So the stability of update (332) reduces to the behavior (boundedness) of powers of Jordan block diagonal matrix $\mathbf{J}6n$.
- We can have the following statement for stability of the update:

Spectral stability:

$$\boxed{{}^{t+n\Delta t}\hat{\mathbf{X}} = \mathbf{A}^{n\,t}\hat{\mathbf{X}} \quad \text{is stable iff } \rho(\mathbf{A}) \leq 1 \text{ and if } \mathbf{A} \text{ is } \mathbf{not} \text{ diagonalizable eigenvalues } a_i \text{ with } n_i^A > n_i^G \text{ satisfy } |a_i| < 1} \tag{338}$$

- The condition that for eigenvalues a_i with $n_i^A > n_i^G$ we require $|a_i| < 1$ becomes apparent from the example below,

$$\mathbf{J} = \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix} \qquad \Rightarrow \mathbf{J}^n = \begin{bmatrix} a^n & na^{n-1} \\ 0 & a^n \end{bmatrix} \tag{339}$$

  which correspond to a 2×2 $\mathbf{A}$ with $n_1^A = 2, n_1^G = 1$. Same argument can be applied to $n_i^A > 2$ and $m > 2$ as $\mathbf{J}^n$ only involves the powers of the Jordan block diagonal matrices similar to the one in (339).

- There are three cases:

  1. If $a > 1$ both diagonal and off-diagonal (J_{12}) values blow up.
     - This instability is called exponential of "explosive" and very fast shows up in the numerical results.
     - The growth of this instability is of the form $\mathcal{O}(a^n)$.
  2. If $a < 1$ both diagonal and off-diagonal (J_{12}) are bounded (and in fact approach zero as $n \to \infty$).
  3. If $a = 1$, diagonal values are 1 but off-diagonal value $J_{12} = n$ **weakly** blows up:
     - This instability is called weak / algebraic and unlike exponential instability may not be easily detected in numerical results.
     - The growth of this instability is of the form $\mathcal{O}(n^s)$, $s = \max_i(n_i^A - n_i^G)$. Compare this with much more severe exponential instability in case 1: $\mathcal{O}(a^n)$.

- In the discussion of the stability of various methods that have a matrix amplification factor $\mathbf{A}$ we refer to (338).

5.3.1 Stability analysis of LMS methods

- General first and second order linear ODEs applied to a vector $\mathbf{y}$ can be represented as follows, (*cf.* (240) for general nonlinear first order expression of an ODE),

$$\dot{\mathbf{y}} = f(\mathbf{y}, t) = \mathbf{G}_0\mathbf{y} + \mathbf{H}(t) \qquad \text{Linear first order ODE} \tag{340a}$$

$$\ddot{\mathbf{y}} = f(\mathbf{y}, \dot{\mathbf{y}}, t) = \mathbf{G}_1\dot{\mathbf{y}} + \mathbf{G}_0\mathbf{y} + \mathbf{H}(t) \qquad \text{Linear second order ODE} \tag{340b}$$

- When a k-step LMS method is applied to an ODE $\mathbf{y}_{n+1}$ in terms of $\mathbf{y}_n, \mathbf{y}_{n-1}, \cdots, \mathbf{y}_{n-k+1}$.

- Formally, the expressions of an k-step LMS method applied to linear first and second order ODEs in (340) are,

$$\sum_{i=0}^{k} \{\alpha_i \mathbf{y}_{n+1-i} + \Delta t \beta_i [\mathbf{G}_0 \mathbf{y}_{n+1-i} + \mathbf{H}(t_{n+1-i})]\} = 0 \qquad \text{LMS applied Linear first order ODE} \tag{341a}$$

$$\sum_{i=0}^{k} \{\alpha_i \mathbf{y}_{n+1-i} + \Delta t \beta_i \mathbf{G}_1 \mathbf{y}_{n+1-i} + \Delta t^2 \gamma_i [\mathbf{G}_0 \mathbf{y}_{n+1-i} + \mathbf{H}(t_{n+1-i})]\} = 0 \qquad \text{LMS applied Linear second order ODE} \tag{341b}$$

- For the use of LMS methods, we focus on either second order $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ or first order $\mathbf{M}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ MDOF ODEs.

- As mentioned several times, for the analysis of these MDOF problems, we analyze their stability and convergence properties by reducing them to SDOFs,

MDOF	$\Rightarrow$	SDOF	Parameters in (340)	
$\mathbf{M}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$	$\Rightarrow$	$\dot{x} + \lambda x = f(t)$	$\mathbf{G}_0 = -\lambda, \mathbf{H}(t) = f(t)$	(342a)
$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$	$\Rightarrow$	$\ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = f(t)$	$\mathbf{G}_1 = -2\xi\omega, \mathbf{G}_0 = -\omega^2, \mathbf{H}(t) = f(t)$	(342b)

- Thus, plugging $\mathbf{H}(t), \mathbf{G}_0$ (and $\mathbf{G}_0$) from (342) in (341) we obtain the following k-step LMS method applied to SDOFs in (342),

$$\sum_{i=0}^{k} \{\alpha_i x_{n+1-i} + \Delta t \beta_i [-\lambda x_{n+1-i} + f(t_{n+1-i})]\} = 0 \qquad \text{LMS for } \dot{x} + \lambda x = f(t) \tag{343a}$$

$$\sum_{i=0}^{k} \{\alpha_i x_{n+1-i} - 2\Delta t \beta_i \xi\omega x_{n+1-i} + \Delta t^2 \gamma_i [-\omega^2 x_{n+1-i} + f(t_{n+1-i})]\} = 0 \qquad \text{LMS for } \ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = f(t) \tag{343b}$$

- For stability analysis, for simplicity of analysis herein, we assume $f(t) = 0$.
- Equation (343) for $f(t) = 0$ can be written in the short form,

$$c_0 x_{n+1} + c_1 x_n + c_1 x_{n-1} + \cdots + c_k x_{n-k+1} = 0 \qquad \text{where } c_i = \begin{cases} \alpha_i - \Delta t \beta_i \lambda & \text{LMS for } \dot{x} + \lambda x = 0 \\ \alpha_i - 2\Delta t \beta_i \xi \omega - \Delta t^2 \gamma_i \omega^2 & \text{LMS for } \ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = 0 \end{cases} \tag{344}$$

- We can express (353) as the following update equation,

$$x_{n+1} = \bar{c}_1 x_n + \bar{c}_1 x_{n-1} + \cdots + \bar{c}_k x_{n-k+1}, \quad \bar{c}_i = -\frac{c_i}{c_0} \tag{345}$$

  where the dependence of c_i on Δt, SDOF parameters such as λ or ω, ξ, and LMS parameters $\alpha_i, \beta_i, \gamma_i$ is shown in (353)
- There are different approaches to analyze the stability of the LMS update equation for a linear ODE based on (353).
- Herein, we present an approach that collects scalars x_i into a generalized $\hat{\mathbf{X}}$ update vector that is updated by a matrix amplification factor $\mathbf{A}$ so that we can apply the analysis tool from §5.3.
- For a k-step LMS method we define the vector of variables $\hat{\mathbf{X}}_{n+1}$ as,

$$\hat{\mathbf{X}}_n = \begin{bmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_{n-k+1} \end{bmatrix} \tag{346}$$

- In this case from (353) we can write the update for $\hat{\mathbf{X}}_{n+1}$:

$$\hat{\mathbf{X}}_{n+1} = \mathbf{A}\hat{\mathbf{X}}_n, \qquad \text{where} \qquad \mathbf{A} = \begin{bmatrix} \bar{c}_1 & \bar{c}_2 & \cdots & \bar{c}_{k-1} & \bar{c}_k \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & 1 & 0 \end{bmatrix} \tag{347}$$

  again c_i depend on Δt and parameters of SDOF parameters and time integration model and it parameters.
- It can be shown (will be a HW problem) that the eigenvalues of (347) satisfy,

$$a_i \text{ is an eigenvalue of } \mathbf{A} \quad \Leftrightarrow \tag{348a}$$
$$\exists \mathbf{v}_i \ \mathbf{A}\mathbf{v}_i = a_i \mathbf{v}_i \text{ (no summation on } i) \quad \Leftrightarrow \tag{348b}$$
$$-a_i^k + \bar{c}_1 a_i^{k-1} + \bar{c}_2 a_i^{k-2} + \cdots + \bar{c}_k = 0 \quad \Leftrightarrow \quad (\bar{c}_i = -\frac{c_i}{c_0}, \textit{cf. } (353)) \tag{348c}$$
$$\boxed{c_0 a_i^k + c_1 a_i^{k-1} + c_2 a_i^{k-2} + \cdots + c_k = 0} \tag{348d}$$

- It is easy to verify that all eigenvectors of $\mathbf{A}$ have geometric multiplicity of one ($n_i^G = 1$). Why?
- That is, if any eigenvalue a_i is repeated ($n_i^A > 1$) it corresponds to the case $n_i^G < n_i^A$ and it must be smaller than one for stability of the LMS scheme a_i according to stability statement (338).
- Accordingly, the stability analysis of LMS scheme is as follows,

$$|a_i| \leq 1, \text{ if } a_i \text{ is not repeated } (n_i^A = 1) \text{ otherwise } |a_i| < 1, \text{where} \tag{349a}$$
$$a_i \text{ are eigenvalues of } \mathbf{A}, \textit{ i.e., } \text{roots of } c_0 a_i^k + c_1 a_i^{k-1} + \cdots + c_k = 0 \tag{349b}$$

5.3.1.1 Stability analysis of LMS methods: Central Difference method

- Consider central difference method update equations,

$$\ddot{x_n} + 2\xi\omega\dot{x_n} + \omega^2 x_n = f(t_n) \tag{350a}$$

$$\dot{x_n} = \frac{1}{2\Delta t}\left(x_{n+1} - x_{n-1}\right) \tag{350b}$$

$$\ddot{x_n} = \frac{1}{\Delta t^2}\left(x_{n+1} - 2x_n + x_{n-1}\right) \tag{350c}$$

- By direct plugging (350b) and (350c) into (350a) we obtain the following update equation,

$$(1 + \xi\Delta t\omega)x_{n+1} + (-2 + (\Delta t\omega)^2)x_n + (1 - \xi\Delta t\omega)x_{n-1} = 0 \tag{351}$$

- We obtain the same equation (351) with a bit longer approach by formally obtaining the values $\alpha_i, \beta_i, \gamma_i$ for LMS method applied to the second order ODE (350a).
- Formally, the expressions of a 2-step LMS method applied to linear second order ODE in (343b) ($\sum_{i=0}^{k}\{\alpha_i x_{n+1-i} - 2\Delta t\beta_i\xi\omega x_{n+1-i} - \Delta t^2\gamma_i\omega^2 x_{n+1-i}\} = 0$ are ($f(t) = 0$),

$$\alpha_0 = 1 \qquad \alpha_1 = -2 \qquad \alpha_2 = 1 \qquad \ddot{x} \text{ (in (350c))} \tag{352a}$$

$$\beta_0 = -\frac{1}{2} \qquad \beta_1 = 0 \qquad \beta_2 = \frac{1}{2} \qquad \dot{x} \text{ (in (350b))} \tag{352b}$$

$$\gamma_0 = 0 \qquad \gamma_1 = -1 \qquad \gamma_2 = 0 \qquad x \text{ (inserted for } t(t_n) \text{ in (350a))} \tag{352c}$$

- If both β_0 and γ_0 were zero, this method formally would have been explicit *cf.* [Hughes, 2012] §9.3.2.
- The method, formally is not fully explicit because $\beta_0 \neq 0$ involves values for t_{n+1}.
- As we will see the method still is only conditionally stable. The main cause is $\gamma_0 = 0$.
- Based on the values $\alpha_i, \beta_i, \gamma_i, i = 0, 1, 2$ in (352) the update equation (353) is written as,

$$c_0 x_{n+1} + c_1 x_n + c_2 x_{n-1} = 0 \qquad \text{where } c_i = \alpha_i - 2\Delta t\beta_i\xi\omega - \Delta t^2\gamma_i\omega^2 \qquad \rightarrow \qquad \begin{cases} c_0 = 1 + \xi\Delta t\omega \\ c_1 = -2 + (\Delta t\omega)^2 \\ c_2 = 1 - \xi\Delta t\omega \end{cases} \tag{353}$$

  which is the same as (351).
- In either case, for the stability analysis, we can do the following,
  1. Directly evaluate roots a_i corresponding to the 2-step update equation in (353) ($c_0 x_{n+1} + c_1 x_n + c_2 x_{n-1} = 0$) based on (348d)

$$c_0 a^2 + c_1 a + c_2 = 0, \qquad \text{for} \qquad c_0 = 1 + \xi\Delta t\omega,\ c_1 = -2 + (\Delta t\omega)^2,\ c_2 = 1 - \xi\Delta t\omega \tag{354}$$

  which has the roots,

$$a^2 - 2A_1 a + A_2 = 0, \qquad A_1 = \frac{1 - \frac{1}{2}\overline{\Delta t}^2}{1 + \xi\overline{\Delta t}}, A_2 = \frac{1 - \xi\overline{\Delta t}}{1 + \xi\overline{\Delta t}}, \text{where} \quad \overline{\Delta t} := \Delta t\omega \text{ normalized time step} \quad \Rightarrow \tag{355a}$$

$$\boxed{a_{1,2} = A_1 \pm \sqrt{A_1^2 - A_2} = \frac{2 - \overline{\Delta t}^2 \pm \overline{\Delta t}\sqrt{\overline{\Delta t}^2 - 4(1 - \xi^2)}}{2(1 + \xi\overline{\Delta t})}} \tag{355b}$$

  2. By defining $\hat{\mathbf{X}}_n = [x_n\ x_{n-1}]$ and $c_0 x_{n+1} + c_1 x_n + c_2 x_{n-1} = 0$ from (353) we have (*cf.* (347)),

$$\hat{\mathbf{X}}_{n+1} = \mathbf{A}\hat{\mathbf{X}}_n, \qquad \text{where} \qquad \mathbf{A} = \begin{bmatrix} \bar{c}_1 & \bar{c}_2 \\ 1 & 0 \end{bmatrix}, \hat{\mathbf{X}}_n = \begin{bmatrix} x_n \\ x_{n-1} \end{bmatrix} \bar{c}_1 = -\frac{c_1}{c_0} = \frac{2 - \overline{\Delta t}^2}{1 + \xi\overline{\Delta t}}, \quad \bar{c}_2 = -\frac{c_1}{c_0} = -\frac{1 - \xi\overline{\Delta t}}{1 + \xi\overline{\Delta t}} \tag{356}$$

  Given that $\bar{c}_1 = -2A_1$, $\bar{c}_2 = -A_2$ the eigenvalue problem for $\mathbf{A}$ also results in the the polynomial (355a) and the solution (355b).

- For stability of central difference method, we require roots from (355b) satisfy stability condition (338).
- That is, $|a_{1,2}| \leq 1$ and if they are equal (*i.e.*, $A_1^2 = A_2$) $|a_1| = |a_2| < 1$.
- Clearly, it is difficult to obtain stability condition by directly checking the absolute values $|A_{1,2}|$.
- A theory, discussed in the next section, shows that for stability we must have,

$$
\begin{aligned}
&|A_2| \leq 1, |A_1| \leq \left|\frac{1+A_2}{2}\right| \quad \Rightarrow \\
&-1 \leq \frac{1-\xi\overline{\Delta t}}{1-\xi\overline{\Delta t}} \leq 1 \quad (|A_2| \leq 1 \text{ automatically satisfied} \\
&-\frac{1}{1+\xi\overline{\Delta t}} \leq \frac{1-\frac{1}{2}(\overline{\Delta t})^2}{1+\xi\overline{\Delta t}} \leq \frac{1}{1+\xi\overline{\Delta t}} \qquad \Rightarrow \qquad \boxed{\overline{\Delta t} \leq 2}
\end{aligned} \tag{357}
$$

- That is,

$$
\text{Central difference time integration for SDOF } \ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = 0 \quad \text{is stable if } \Delta t \leq \frac{2}{\omega} \tag{358a}
$$

$$
\text{Central difference time integration for MDOF } \mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = 0 \quad \text{is stable if} \qquad \Delta t \leq \frac{2}{\max_l(\omega_l^h)}
$$

$$
\text{or more conservatively \& conveniently} \qquad \Delta t \leq \frac{2}{\omega_e^m} \tag{358b}
$$

- Recall that ω_e^m is the maximum element level natural frequency which can be easily computed.
- $\max_l(\omega_l^h)$ is the maximum frequency of the MDOF problem which often is not computed.
- Since $\omega_e^m > \max_l(\omega_l^h)$ we can conservatively and conveniently use it in estimating stable time step of conditionally stable methods; *cf.* (311) and §5.2.2.
- One very interesting aspect is that the stable time step is not increased by increasing ξ which typically is the case.

5.3.1.2 Stability region for a 2×2 update equation (with real coefficients)

- Consider the update equation for a size two $\hat{\mathbf{X}}$ with real values (*cf.* (347) for a general size m $\hat{\mathbf{X}}$),

$$
\hat{\mathbf{X}}_{n+1} = \mathbf{A}\hat{\mathbf{X}}_n, \qquad \text{where} \qquad \mathbf{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{359}
$$

- The eigenvalues of $\mathbf{A}$ satisfy,

$$
a^2 - 2A_1 a + A_2 = 0 \quad \text{where} \tag{360a}
$$

$$
A_1 = \frac{1}{2}\text{trace}(\mathbf{A}), \qquad \text{trace}(\mathbf{A}) = A_{11} + A_{22} \qquad \text{first invariant of } \mathbf{A} \tag{360b}
$$

$$
A_2 = \det(\mathbf{A}), \qquad \det(\mathbf{A}) = A_{11}A_{22} - A_{12}A_{21} \qquad \text{second invariant of } \mathbf{A} \tag{360c}
$$

- On the other hand, in many instances we directly reach to a second order polynomial of the form (360a). See for example (355a).
- In this section we provide conditions in which the roots of the second order polynomial in (361) satisfy $|a_{1,2}| \leq 1$ or $|a_{1,2}| < 1$ and provide the full analysis (including when the coefficients A_1 and A_2 are complex in §6.4.6.
- In either case, whether the polynomial is directly derived or is from a size two $\hat{\mathbf{X}}_n$ update vector roots of (360a) must satisfy a stability condition which is of the form (338).
- For the resulting second order polynomial stability condition reduces to,

$a^2 - 2A_1 a + A_2 = 0$ correspond to a stable scheme iff

$$\begin{cases} |a_{1,2}| = \left|A_1 \pm \sqrt{A_1^2 - A_2}\right| \le 1 & \text{if } a_1 \neq a_2 \quad \text{that is } A_1^2 \neq A_2 \\ |a_{1,2}| = |A_1| < 1 & \text{if } a_1 = a_2 \quad \text{that is } A_1^2 = A_2 \end{cases} \tag{361a}$$

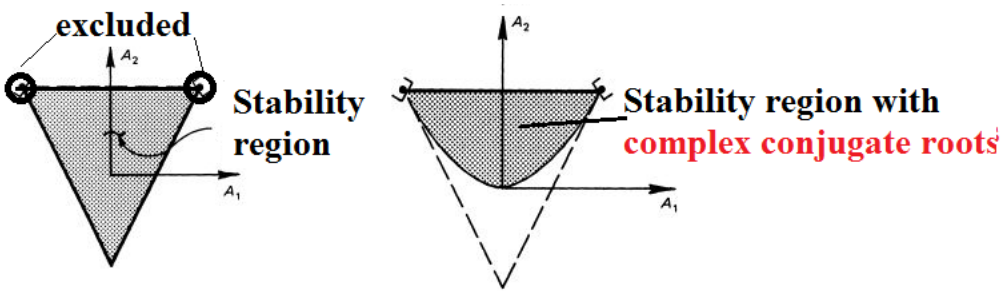
- One can show that for real A_1 and A_2 in the polynomial (361) ($a^2 - 2A_1 a + A_2 = 0$ where) (361) reduces to,

$$\boxed{-1 \le A_2 \le 1, \qquad -\frac{A_2+1}{2} \le A_1 \le \frac{A_2+1}{2}, \qquad \textbf{except the points } |A_1| = A_2 = 1} \tag{362}$$

- (362) is often used in various contexts for the stability of PDEs / ODEs with two temporal derivatives as in many cases the stability analysis reduces to a problem of the form (361)
- Another important condition is whether the roots are real or complex (in complex conjugate pairs):

$$a_{1,2} \text{ are complex iff } A_1^2 < A_2 \tag{363}$$

- Having complex conjugate roots can damp out high frequency content in various time marching schemes of the solution which is often desirable. This will be discussed further for Newmark methods.
- Overall stability region, and stability region with complex conjugate roots are shown below.



5.3.1.3 Stability analysis of LMS methods: Houbolt method

- We previously discussed the 3-step LMS Houbolt method in §4.3.2 with the FD operators:

$$\begin{aligned} {}^{t+\Delta t}\ddot{x} + 2\xi\omega\, {}^{t+\Delta t}\dot{x} + \omega^2\, {}^{t+\Delta t}x &= {}^{t+\Delta t}r \\ {}^{t+\Delta t}\ddot{x} &= \frac{1}{\Delta t^2}\left(2\, {}^{t+\Delta t}x - 5\, {}^{t}x + 4\, {}^{t-\Delta t}x - {}^{t-2\Delta t}x\right) \\ {}^{t+\Delta t}\dot{x} &= \frac{1}{6\Delta t}\left(11\, {}^{t+\Delta t}x - 18\, {}^{t}x + 9\, {}^{t-\Delta t}x - 2\, {}^{t-2\Delta t}x\right) \end{aligned} \tag{364}$$

- which results in the update equation,

$$\begin{bmatrix} {}^{t+\Delta t}x \\ {}^{t}x \\ {}^{t-\Delta t}x \end{bmatrix} = \mathbf{A}\begin{bmatrix} {}^{t}x \\ {}^{t-\Delta t}x \\ {}^{t-2\Delta t}x \end{bmatrix} + \mathbf{L}\, {}^{t+\Delta t}r \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} \frac{5\beta}{\omega^2\Delta t^2} + 6\kappa & -\left(\frac{4\beta}{\omega^2\Delta t^2} + 3\kappa\right) & \frac{\beta}{\omega^2\Delta t^2} + \frac{2\kappa}{3} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\beta = \left(\frac{2}{\omega^2\Delta t^2} + \frac{11\xi}{3\omega\,\Delta t} + 1\right)^{-1}; \qquad \kappa = \frac{\xi\beta}{\omega\,\Delta t} \qquad \mathbf{L} = \begin{bmatrix} \frac{\beta}{\omega^2} \\ 0 \\ 0 \end{bmatrix} \tag{365}$$

- Noticeably $\rho(\mathbf{A}) < 1$ for all Δt meaning that Houbolt method is unconditionally stable, as mentioned before.

5.3.2 Absolute stability, A-stable methods

5.3.2.1 Introduction: Properties of analytical solution

- Consider the p^{th} order ODE,
$$a_p \frac{\mathrm{d}^p u}{\mathrm{d}t^p} + a_{p-1} \frac{\mathrm{d}^{p-1} u}{\mathrm{d}t^{p-1}} + \cdots + a_1 \frac{\mathrm{d}u}{\mathrm{d}t} + a_0 u(t) = f(t) \tag{366}$$
with m initial conditions of $u(t=0) = u_0, \cdots, \frac{\mathrm{d}^{p-1}u}{\mathrm{d}t^{p-1}}(t=0) = u_0^{(p-1)}$.
- We assume a_i are constant values.
- We first discuss how the solution is obtained for homogeneous case ($f(t) = 0$) and later comment on the solution of (366).
- Considering $f(t) = 0$ and letting $u(t) = e^{\lambda t}$ we obtain,
$$u(t) = e^{\lambda t}, f(t) = 0, \text{ so (366)} \quad \Rightarrow \boxed{a_p \lambda^p + a_{p-1}\lambda^{p-1} + \cdots + a_1 \lambda + a_0 = 0} \tag{367}$$
- The solutions of the m^{th} order polynomial in (367) are $\bar{p}$ roots some of which may be repeated; *i.e.*, having algebraic multiplicity higher than 1.
- If the root i is repeated m_i times, the general solution corresponding to this root is $P_i(t)e^{\lambda_i t}$ where P_i is an $m_i - 1$ order polynomial.
- So, the solution is,
$$a_p \frac{\mathrm{d}^p u}{\mathrm{d}t^p} + a_{p-1} \frac{\mathrm{d}^{p-1} u}{\mathrm{d}t^{p-1}} + \cdots + a_1 \frac{\mathrm{d}u}{\mathrm{d}t} + a_0 u(t) = 0 \qquad \Rightarrow \qquad u(t) = \sum_{i=1}^{\bar{p}} P_i(t) e^{\lambda_i t} \tag{368}$$
where λ_i are distinct roots of (367) with multiplicities m_i.
- Given that $m_i - 1$ order polynomial P_i has m_i coefficients and $\sum_{i=1}^{\bar{p}} = p$, there are a total of p polynomial coefficients which is exactly the number of unknowns to match p ICs from (366).
- For distinct roots, $m_i - 1$ and the corresponding solution for it takes the form $p_i e^{\lambda_i t}$ where p_i is a constant.
- We are interested in stability (boundedness) of the physical solution of one of the modes in the form,
$$u_i(t) = P_i(t)e^{\lambda_i t}, \qquad P_i(t) \text{ is an order } m_i \text{ polynomial} \tag{369}$$
u_i's are components of $u(t)$. That is, $u(t) = \sum_{i=1}^{\bar{p}} u_i(t)$.
- To study the boundedness of (369) let,
$$\lambda_i = \lambda_i^R + \imath \lambda_i^I \qquad \Rightarrow \qquad u_i(t) = P_i(t) e^{(\lambda_i^R + \imath \lambda_i^I)t} = P_i(t) e^{\imath \lambda_i^I t} e^{\lambda_i^R t} \tag{370}$$
where λ_i^R, λ_i are real and imaginary parts of λ_i and $\imath$ is the imaginary number ($\imath^2 = -1$).
- Now, the boundedness of $u_i(t)$ depends on the sign of λ_i^R given that $|e^{\imath \lambda_i^I t}| = 1$:

$$\begin{cases} \lambda_i^R < 0 & \text{Bounded and diminishing, } (\lim_{t\to\infty} u_i(t) = 0) \\ \lambda_i^R = 0 \ \& \ m_i = 1 (\lambda_i \text{ simple root of (367)}) & \text{Bounded and oscillatory (the solution oscillated but not tending to zero)} \\ \lambda_i^R = 0 \ \& \ m_i > 1 (\lambda_i \text{ repeated root of (367)}) & \text{Weakly (algebraically) unbounded: } u_i(t) \text{ algebraically tends to } \infty \text{ as } t \to \infty \\ \lambda_i^R > 0 & \text{Strongly (exponentially) unbounded, } (\lim_{t\to\infty} u_i(t) = \infty) \end{cases} \tag{371}$$

- So for the (stability) boundedness of the solution to (368) with $f(t) = 0$ we have,

$$\begin{cases} \text{Exponentially unbounded} & \text{If } \underline{\text{any}} \text{ root has positive real part} \\ \text{Algebraically unbounded} & \text{If all } \lambda_i \le 0, \text{ but roots with } \lambda_i^R = 0 \text{ are repeated } m_i > 1 \\ \text{Bounded} & \text{If all } \lambda_i \le 0, \text{ and roots (if any) with } \lambda_i^R = 0 \text{ are simple } m_i = 1 \end{cases} \tag{372}$$

- Obviously, unstable modes may not get activated for particular ICs if their $P_i(t)$ is identically zero.
- An important question is if the IC is perturbed a bit, whether the perturbation result in unbounded changes in the solution as $t \to \infty$. This property is called dynamic-stability and for an p^{th} order linear ODE, it requires the satisfaction of (372).
- The solution of (366) for $f(t) \neq 0$ can be obtained by using Laplace transform. The solution will include convolutions of the kernels of the form (369) and $f(t)$. The details of the process can be found in any introductory ODE book.
- We are more interested in knowing when the exact solution (*i.e.,* physical system) is dynamically stable and afterward knowing when a numerical method can maintain "stability" in a numerical setting, meaning that the solution does not blowup.

5.3.2.2 Absolute stability

- Consider the first order ODE,

$$\dot{x} - \lambda x = 0 \tag{373}$$

- The solution of this ODE matches $u_i(t)$, a component of the p order ODE for u in (366) if λ is one of the roots of (367) (and λ is a simple root of it). Recall $u_i(t) = P_i(t)e^{\lambda_i t}$ and $u(t) = \sum_{i=1}^{\bar{p}} u_i(t)$.
- To be able to capture all the solutions of (367) with simple roots, λ in (373) can take any complex value $\lambda = \lambda^R + \imath\lambda^I$.
- Thus, understanding the behavior of a numerical ODE solver for (373) provides information how it would perform for a general p order ODE.
- Recall, even if the ODE solver only applies to first order ODEs, such as (373), we can write the p order ODE as a system of p first order ODEs.
- Given that (373) is physically (dynamically) **stable** for $\lambda < 0$ we are mainly interested in whether the numerical method is **stable** for $\lambda < 0$. We adopt the following definitions,
- An ODE solver whose update equation can be cast in the form $\hat{\mathbf{X}}_n = \mathbf{A}^n \hat{\mathbf{X}}_0$ ($\mathbf{A}$ is the amplification factor), *e.g.,* RK or any LMS scheme, is said to be **absolutely stable** at a fixed $\lambda\Delta t$ if the **spectral radius of A is strictly less than 1:** $\rho(\mathbf{A}) < 1$ for the solution of $\dot{x} - \lambda x = 0$ is stable with the time step Δt.
- Recall that **spectral stability** (338) in fact allows $\rho(\mathbf{A}) = 1$, whereas for **absolute stability** $\rho(\mathbf{A}) < 1$.
- To reiterate, spectral stability (338) is repeated here,
  - $\rho(\mathbf{A}) \leq 1$ and if $\mathbf{A}$ is **not** diagonalizable ,
  - eigenvalues a_i with $n_i^A > n_i^G$ satisfy $|a_i| < 1$.

  which clearly allows $\rho(\mathbf{A}) = 1$.
- This means that for an absolutely stable point $\lambda\Delta t$: $\rho(\mathbf{A}) < 1$, the numerical solution definitely dissipates and approaches zero as $n \to 0$ (n is the time step).
- The **region of absolute stability** refers to the collection of all $\lambda\Delta t$ in the complex plane for which the method is absolutely stable.
- For (373) we know that the exact solution is bounded if $\lambda \leq 0$ and is dissipative if $\lambda^R < 0$. Given that $\rho(\mathbf{A}) < 1$ is numerical counterpart of dissipative property, we seek numerical methods for which $\rho(\mathbf{A}) < 1$ for all Δt when $\lambda^R < 0$.
- That is, in cases that the physical system is well-posed and dissipative ($\lambda^R < 0$), an ideal numerical method would be stable for **ANY** Δt; *i.e.,* it is unconditionally stable (free Δt) for **all** λ with $\lambda^R < 0$. This is the motivation for the following definition,
- A method is called **A-stable** if its region of absolute stability contains the negative (left) complex half plane (all $\lambda\Delta t, \lambda^R < 0$).
- Unfortunately, the condition of A-stability is extremely demanding. [Dahlquist, 1963] has shown the following results (known as Dahlquist Second Barrier Theorem):
  1. No explicit linear multi-step method (LMS) is A-stable.
  2. No A-stable LMS can have order greater than 2.
  3. The second-order A-stable LMS with the smallest error constant is the trapezoid rule method.

- The result for explicit methods is expected because they are not even unconditionally stable for an equation in the form $\dot{x} - \lambda x = 0$ for real $\lambda < 0$. That is, they do not even cover the negative real axis for $\lambda \Delta t$ let alone the negative complex plane ($\lambda^R < 0$ arbitrary λ^I).

- The result for implicit LMS methods, however, is disappointing implying that if we want a LMS method that 100% preserves the well-posedness region of complex plane ($\lambda^R < 0$) by allowing arbitrary Δt we are at most offered a second order of accuracy, for which trapezoidal rul has the smallest error constant.

5.3.2.3 A-stability vs. unconditional stable / relaxing A-stability

- It seems A-stability is a very close concept to unconditional stability as for both the method is stable for any Δt.
- There are, however, two points to clarify,

1. When talking about unconditional stability we often deal with **a fixed ODE** for which any Δt can be used; *i.e.,* for any Δt the method is stable.

   - To elaborate on this, assume **fixed and negative real valued** λ_0: $\lambda_0^I = 0, \lambda_0^R < 0$ is given and we still want to solve (373) for this **fixed** λ_0,
   $$\dot{x} - \lambda_0 x = 0, \quad (\lambda_0^I = 0, \lambda_0^R < 0) \tag{374}$$
   - Clearly, (374) has a bounded and dissipative physical solution; it is a problem that comes up in many application and for which we want to have numerical methods with good stability properties.
   - For example, if we have a numerical method for the solution of (374) which is unconditionally stable it means that **any** Δt can be chosen, and yet the solution for would be stable.
   - Now, if we look at the set $\lambda \Delta t$ (with stability property) that is spanned and is of interest for this particular problem, we have $\lambda_0 \Delta t$ where here λ_0 is a **fixed** negative real number and $\Delta t > 0$ is an **arbitrary** (because the method is unconditionally stable).
   - That is, given the unconditional stability of the method and negative real number λ_0 the negative real axis in $\lambda \Delta t$ plane belongs to region of absolute stability (assuming that solution being spectrally stable ($\rho(\mathbf{A}) \leq 1$, *etc.* is absolutely stable $\rho(\mathbf{A}) < 1$; more about this in the second point).
   - From this example it is clear that if we only want to solve (374) fixed λ_0 with $\lambda_0^R < 0$ $\lambda^I = 0$ the A-stability is not even needed!
   - We only need the **negative real axis** to be in the **region of absolute stability**.
   - This provides the opportunity to look for numerical ODE solvers in a larger class of problems.
   - As will be seen shortly, we can find numerical methods with higher orders of accuracy, which are not entirely A-stable, yet negative real axis is covered in their region of absolute stability. So, they would result in an unconditionally solution method for the particular problem considered in (374).
   - Given that such schemes are not fully A-stable, for some fixed complex $\lambda_0 = r_0 e^{\imath\theta_0}$ the locus $\lambda \Delta t = (r_0 \Delta t) e^{\imath\theta_0}$, which is a ray of angle θ_0, is not entirely in the region of absolute stability. That is, not for all Δt the method is stable, so we cannot call it unconditionally for that particular choice of λ_0. Here r_0, θ_0 are the amplitude and phase of the complex value λ_0. $\pi/2 < \theta_0 < 3\pi/2$ to ensure $\lambda_0^R < 0$.
   - However, **not having unconditional stability for any** λ_0 **(with** $\lambda_0^R < 0$**)–which is what A-stability means** may pose no problem if our goal for example is only solving (374). In that case a higher order unconditionally stable scheme would still have no restriction on Δt.

2. In spectral stability, we let $\rho \mathbf{A} \leq 1$, which would allow $a_i = 1$, except cases where a_i is a repeated root with $n_i^G < n_i^A$. On the other hand, in absolute stability $\rho \mathbf{A} < 1$ which ensures the numerical solution is dissipative for $\lambda^R < 0$.

- As mentioned, there is a great interest in solving problems of the form (374) $\dot{x} - \lambda_0 x = 0$ ($\lambda_0^I = 0, \lambda_0^R < 0$) which correspond to the solution of an advection equation, or can even arise from the discretization of parabolic equations.

- Next, we provide some excepts from [Süli and Mayers, 2003], §12.11 "Stiff systems" that are high order and almost A-stable.

- These Backward Differentiation Formulae (BDF), which are LMS schemes, can have **orders of accuracy as high as 6**, yet covering **all negative real axis** in their region of absolute stability.

The coefficients are obtained by requiring that the order of accuracy of the method is as high as possible, *i.e.*, by making the coefficients C_j zero in (12.47) for $j = 0, 1, \ldots, k$. For $k = 1$ this yields the implicit Euler method (BDF1), whose order of accuracy is, of course, 1; the method is A-stable. The choice of $k = 6$ results in the sixth-order, six-step BDF method (BDF6):

$$147y_{n+6} - 360y_{n+5} + 450y_{n+4} - 400y_{n+3} + 225y_{n+2} - 72y_{n+1} + 10y_n = 60hf_{n+6}\,. \qquad (12.50)$$

Although the method (12.50) is not A-stable, its region of absolute stability includes the whole of the negative real axis (see Figure 12.5). For

To construct useful methods of higher order we need to relax the condition of A-stability by requiring that the region of absolute stability should include a large part of the negative half-plane, and certainly that it contains the whole of the negative real axis.

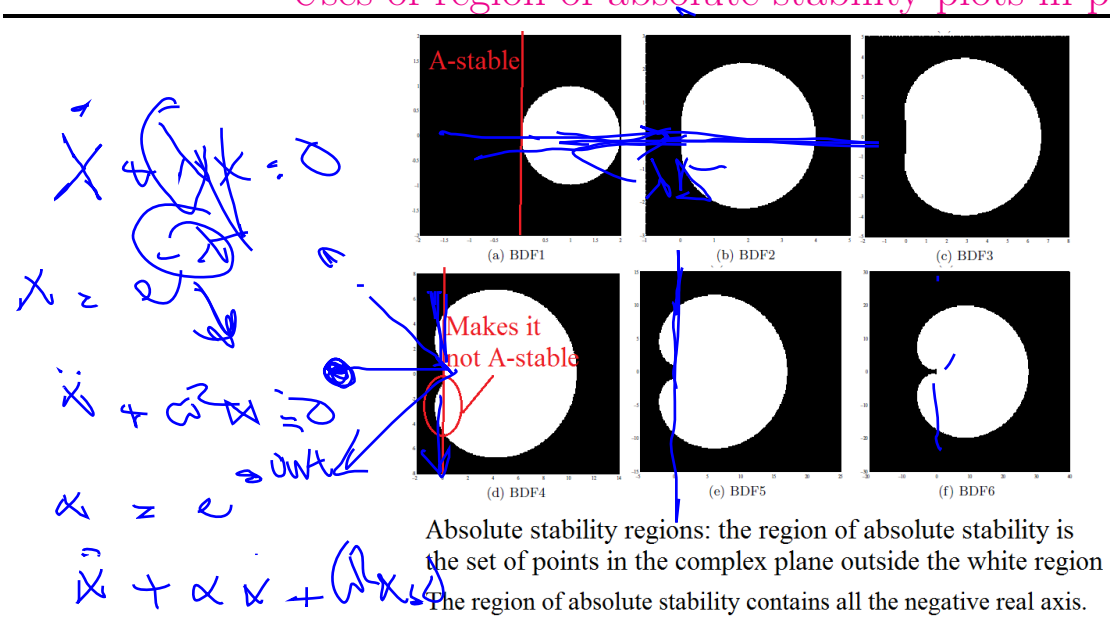
The most efficient methods of this kind in current use are the **Backward Differentiation Formulae**, or BDF methods. These are the linear multistep methods (12.35) in which $\beta_j = 0$, $0 \le j \le k-1$, $k \ge 1$, and $\beta_k \ne 0$. Thus,

$$\alpha_k y_{n+k} + \cdots + \alpha_0 y_n = h\beta_k f_{n+k}\,.$$

the intermediate values, $k = 2, 3, 4, 5$, we have the following kth-order, k-step BDF methods, respectively:

$$\begin{aligned}
3y_{n+2} - 4y_{n+1} + y_n &= 2hf_{n+2}\,,\\
11y_{n+3} - 18y_{n+2} + 9y_{n+1} - 2y_n &= 6hf_{n+3}\,,\\
25y_{n+4} - 48y_{n+3} + 36y_{n+2} - 16y_{n+1} + 3y_n &= 12hf_{n+4}\,,\\
137y_{n+5} - 300y_{n+4} + 300y_{n+3} - 200y_{n+2} + 75y_{n+1} - 12y_n &= 60hf_{n+5}\,,
\end{aligned}$$

referred to as BDF2, BDF3, BDF4 and BDF5. Their regions of absolute stability are also shown in Figure 12.5. In each case the region of absolute stability includes the negative real axis. Higher-order methods of this type cannot be used, as all BDF methods, with $k > 6$, are zero-unstable.



Absolute stability regions: the region of absolute stability is the set of points in the complex plane outside the white region The region of absolute stability contains all the negative real axis.

5.3.2.4 Uses of region of absolute stability plots in practice

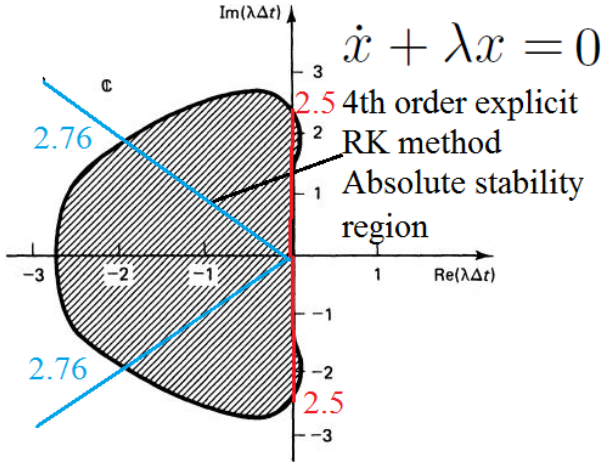
- The question that may arise from the discussion in (5.3.2.3) is that why worry absolute stability of a method in the whole left complex half plane ($\lambda\Delta t$, $\lambda^R < 0$) rather than only the negative real axis ($\lambda^R < 0, \lambda^T = 0$).
- For example, we solve equations of the form,
$$\dot{x} + 2x = 0, \qquad (\lambda = -2)$$
and not equations of the form
$$\dot{x} + \imath x = 0, \qquad (\lambda = -\imath), \qquad \text{or} \qquad \dot{x} + (1+\imath)x = 0, \qquad (\lambda = -1-\imath)$$
- However, let us consider the following second order ODEs,
$$\ddot{x} + x = 0 \tag{375a}$$
$$\ddot{x} + 2\dot{x} + 2x = 0 \tag{375b}$$
- These can be for example SDOFs (229a) ($\ddot{x} + 2\xi\omega\dot{x} + \omega^2 x = f(t)$) that are obtained from model decomposition of (225) $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ with $f(t) = 0$).
- (375a) is an undamped oscillator with $\omega = 1$, while (375b) is a damped oscillator with $\omega = \sqrt{2}$ and $\xi = 1/\sqrt{2}$.
- As mentioned, stability analysis of MDOF systems such as (225) ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ with $f(t) = 0$) reduces to the analysis of the corresponding SDOFs, even if we directly integrate the MDOF in time.
- So, the analysis of (375) is important both from the perspective of scalar ODEs and stability analysis of MDOF systems.
- Recalling from (367), the solution of constant coefficient ODEs $a_p \frac{\mathrm{d}^p u}{\mathrm{d}t^p} + a_{p-1}\frac{\mathrm{d}^{p-1}u}{\mathrm{d}t^{p-1}} + \cdots + a_1\frac{\mathrm{d}u}{\mathrm{d}t} + a_0 u(t) = 0$ can be written as the summation of $u(t) = e^{\lambda t}$ where λ are roots of $a_p\lambda^p + a_{p-1}\lambda^{p-1} + \cdots + a_1\lambda + a_0 = 0$. If a root i is repeated m_i times its corresponding solution is $P_i(t)e^{\lambda_i t}$.
- We follow with these steps and observe,
$$\ddot{x} + x = 0 \quad \Rightarrow \lambda^2 + 1 = 0 \quad \Rightarrow \lambda = \pm\imath \quad \Rightarrow x = a_1 e^{\imath t} + a_1 e^{-\imath t} \tag{376a}$$
$$\ddot{x} + 2\dot{x} + 2x = 0 \quad \Rightarrow \lambda^2 + 2\lambda + 2 = 0 \quad \Rightarrow \lambda = -1 \pm \imath \quad \Rightarrow x = a_1 e^{(\imath-1)t} + a_1 e^{(-\imath-1)t} \tag{376b}$$
- Interestingly, we observe an undamped hyperbolic PDE $\rho u_{,tt} - E u_{,xx} = 0$ (resulting in MDOF ODE $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ with $f(t) = 0$) corresponds to λ along $\pm\imath$ and the damped version of it $\rho u_{,tt} + c u_{,t} + c^2 u_{,xx} = 0$ (resulting in $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ with $f(t) = 0$) can generate λ with nonzero λ^R and λ^I parts.
- Back to stability analysis of (5.3.2.4), we observe that in (376a) $\lambda = \pm i$ and the maximum time step is obtained by finding the location that a ray with angles $\pi/2$ and $3\pi/2$ intersects region of absolute stability first.

- Similarly for (376b), the roots are $-1\pm\imath 1$ which make lines of angle $3\pi/4$ and $5\pi/4$. In this case, we need to find the intersections of the region of absolute stability with rays of these angles.
- Let us consider that we are using a RK4 method whose region of absolute stability versus $\lambda\Delta t$ is shown in the next figure.
- We first obtain RK4 stable time step for (376a) $\ddot{x}+x=0$, decoded with color red. The intersections of the roots $\pm\imath 1$ with region of absolute stability are both 2.5 (shown in the figure). So,

$$|\lambda\Delta t|\leq 2.5,\ \lambda=\pm\imath 1 \quad \Rightarrow |1|\Delta t\leq 2.5 \quad \rightarrow \boxed{\Delta t\leq 2.5} \tag{377}$$

- and for the damped equation (376b) $\ddot{x}+2\dot{x}+2x=0$, the intersection point of rays with angles $3\pi/4, 5\pi/4$ shown in blue in the figure are both 2.76. So,

$$|\lambda\Delta t|\leq 2.76,\ \lambda=-1\pm\imath 1 \quad \Rightarrow |\sqrt{2}|\Delta t\leq 2.76 \quad \rightarrow \boxed{\Delta t\leq 1.94} \tag{378}$$



Note: Having a more stringent time step for the damped system is not due to having damping, rather mainly due to having larger frequency ($\omega=\sqrt{2}$ compared to the damped case).

- Plots of regions of absolute stability are commonly used to determine stability limits for problems of the type discussed above, *e.g.*, damped SDOF oscillator that shows up in the modal decomposition of many MDOFs.

Reading source for A-stability

This section briefly discussed the following concept:

- Dahlquist's theorems that discuss the existence and properties of explicit and implicit LMS methods.
- Concept of region of absolute stability.
- Concepts of zero-stable, A-stable, and stiffly-stable

The following is a list of resources that provide more details on these topics:

- [Süli and Mayers, 2003] pages 329-341: Sections 12.6 Linear multi-step methods; §12.7 Zero-stability; §12.8 Consistency; §12.9 Dahlquist's theorems; §12.10 Systems of equations.
- [Hughes, 2012] section 9.3 (only §9.3.1 and 9.3.2)

5.3.3 Stability analysis of one-step multivariate methods

- As mentioned in (5.3) two of the cases that the amplification factor takes a matrix form $\mathbf{A}$ are,
  1. Value and previous step values of x in (329b): ${}^t\hat{X}=[{}^{t+\Delta t}x\ {}^t x\ {}^{t-\Delta t}x\ \cdots]$. This will be the form of ${}^t\hat{X}$ for LMS methods as we observed in §5.3.1.
  2. Value and subsequent time derivatives of x in (329b): ${}^t\hat{X}=[{}^t x\ {}^t\dot{x}\ {}^t\ddot{x}]$. Examples be from Newmark and θ-Wilson methods which will be discussed subsequently.

5.3.3.1 Stability analysis of one-step multivariate methods: Wilson-θ method

- The assumption of Wilson-θ method is that acceleration varies linearly over the interval t to $t+\theta\Delta t$, where $\theta \geq 1$ whose range of having a stable method will be obtained by the stability analysis below.
- Linear acceleration and its first and second integration yields,

$$
\begin{aligned}
{}^{t+\tau}\ddot{x} &= {}^{t}\ddot{x} + ({}^{t+\Delta t}\ddot{x} - {}^{t}\ddot{x})\frac{\tau}{\Delta t} \\
{}^{t+\tau}\dot{x} &= {}^{t}\dot{x} + {}^{t}\ddot{x}\,\tau + ({}^{t+\Delta t}\ddot{x} - {}^{t}\ddot{x})\frac{\tau^2}{2\Delta t} \\
{}^{t+\tau}x &= {}^{t}x + {}^{t}\dot{x}\,\tau + \tfrac{1}{2}\,{}^{t}\ddot{x}\,\tau^2 + ({}^{t+\Delta t}\ddot{x} - {}^{t}\ddot{x})\frac{\tau^3}{6\Delta t}
\end{aligned}
\tag{379}
$$

- which results the values below for the end of time step $t+\Delta t$,

$$
\begin{aligned}
{}^{t+\Delta t}\dot{x} &= {}^{t}\dot{x} + ({}^{t+\Delta t}\ddot{x} + {}^{t}\ddot{x})\frac{\Delta t}{2} \\
{}^{t+\Delta t}x &= {}^{t}x + {}^{t}\dot{x}\,\Delta t + (2\,{}^{t}\ddot{x} + {}^{t+\Delta t}\ddot{x})\frac{\Delta t^2}{6}
\end{aligned}
\tag{380}
$$

- As a reminder, in Wilson-θ method the equilibrium equation is written for $t+\theta\Delta t$ rather than $t+\Delta t$,

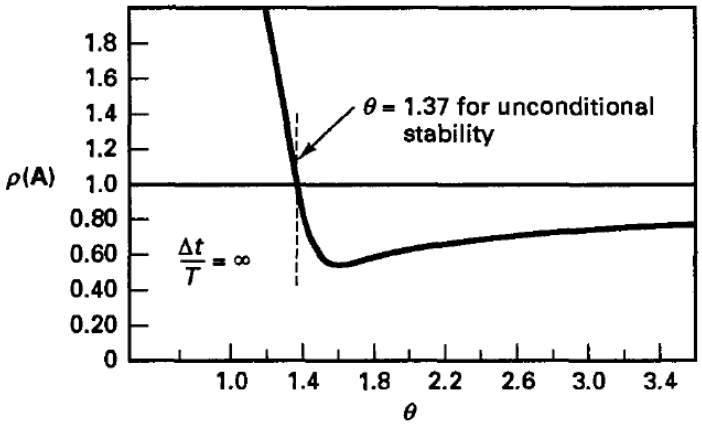
$$
{}^{t+\theta\Delta t}\ddot{x} + 2\xi\omega\,{}^{t+\theta\Delta t}\dot{x} + \omega^2\,{}^{t+\theta\Delta t}x = {}^{t+\theta\Delta t}r \tag{381}
$$

- By evaluating (380) at time $t+\theta\Delta t$ ($\tau = \theta\Delta t$) and substituting ${}^{t+\theta\Delta t}\ddot{x}$ in (381) we obtain an equation only for ${}^{t+\theta\Delta t}\ddot{x}$.
- Solving for ${}^{t+\theta\Delta t}\ddot{x}$ and substituting in (380) we obtain the following update equation (note: this process was in detail discussed in §4.4.1.

$$
\begin{bmatrix} {}^{t+\Delta t}\ddot{x} \\ {}^{t+\Delta t}\dot{x} \\ {}^{t+\Delta t}x \end{bmatrix} = \mathbf{A}\begin{bmatrix} {}^{t}\ddot{x} \\ {}^{t}\dot{x} \\ {}^{t}x \end{bmatrix} + \mathbf{L}\,{}^{t+\theta\Delta t}r \qquad
\mathbf{A} = \begin{bmatrix} 1-\frac{\beta\theta^2}{3}-\frac{1}{\theta}-\kappa\theta & \frac{1}{\Delta t}(-\beta\theta-2\kappa) & \frac{1}{\Delta t^2}(-\beta) \\ \Delta t\left(1-\frac{1}{2\theta}-\frac{\beta\theta^2}{6}-\frac{\kappa\theta}{2}\right) & 1-\frac{\beta\theta}{2}-\kappa & \frac{1}{\Delta t}\left(-\frac{\beta}{2}\right) \\ \Delta t^2\left(\frac{1}{2}-\frac{1}{6\theta}-\frac{\beta\theta^2}{18}-\frac{\kappa\theta}{6}\right) & \Delta t\left(1-\frac{\beta\theta}{6}-\frac{\kappa}{3}\right) & 1-\frac{\beta}{6} \end{bmatrix}
$$

$$
\beta = \left(\frac{\theta}{\omega^2\,\Delta t^2}+\frac{\xi\,\theta^2}{\omega\,\Delta t}+\frac{\theta^3}{6}\right)^{-1}; \qquad \kappa = \frac{\xi\beta}{\omega\,\Delta t} \qquad \mathbf{L} = \begin{bmatrix} \frac{\beta}{\omega^2\,\Delta t^2} \\ \frac{\beta}{2\omega^2\,\Delta t} \\ \frac{\beta}{6\omega^2} \end{bmatrix}
\tag{382}
$$

- Stability requires eigenvalues of $\mathbf{A}$ to satisfy $|a_i| \leq 1$ and if they have lower geometric multiplicity than algebraic multiplicity ($n_i^G < n_i^A$) satisfying $|a_i| < 1$ as discussed in (338).
- Unconditional stability for Wilson-θ method is obtained when $\theta \geq 1.37$.
- For example in the figure below it is shown that in the limit $\Delta t/T \to \infty$ (period $T = 2\pi/\omega$) amplification factor is larger than one for $\theta > 1.37$ necessitating $\theta \geq 1.37$ for unconditional stability.



5.3.3.2 Stability analysis of one-step multivariate methods: Newmark method

- In §4.4.2 we discuss the Newmark integration scheme.
- The Newmark method, we express the equilibrium equation at the end of time step $t + \Delta t$.

$$ {}^{t+\Delta t}\ddot{x} + 2\xi\omega\, {}^{t+\Delta t}\dot{x} + \omega^2\, {}^{t+\Delta t}x = {}^{t+\Delta t}r \tag{383} $$

- where ${}^{t+\Delta t}\dot{x}$ and ${}^{t+\Delta t}x$ are given by,

$$ \begin{aligned} {}^{t+\Delta t}\dot{x} &= {}^{t}\dot{x} + [(1-\delta)\, {}^{t}\ddot{x} + \delta\, {}^{t+\Delta t}\ddot{x}]\,\Delta t \qquad \text{(a)} \\ {}^{t+\Delta t}x &= {}^{t}x + {}^{t}\dot{x}\,\Delta t + [(\tfrac{1}{2}-\alpha)\, {}^{t}\ddot{x} + \alpha\, {}^{t+\Delta t}\ddot{x}]\,\Delta t^2 \quad \text{(b)} \end{aligned} \tag{384} $$

- From (384)(b) we obtain ${}^{t+\Delta t}\ddot{x}$ in terms of ${}^{t+\Delta t}x$ which by plugging in (384)(a) provides ${}^{t+\Delta t}\dot{x}$ in terms of ${}^{t+\Delta t}x$.
- Plugging all these values in (383) provides an equation for ${}^{t+\Delta t}x$ which can be solved in terms of ${}^{t}\hat{\mathbf{X}} = [{}^{t}x,\ {}^{t}\dot{x}, {}^{t}\ddot{x}]^{\mathrm{T}}$.
- Plugging ${}^{t+\Delta t}x$ in (384), we further obtain ${}^{t+\Delta t}\dot{x}$ and ${}^{t+\Delta t}\ddot{x}$ in terms of ${}^{t}\hat{\mathbf{X}}$ and will have the following update equation:

$$ \begin{bmatrix} {}^{t+\Delta t}\ddot{x} \\ {}^{t+\Delta t}\dot{x} \\ {}^{t+\Delta t}x \end{bmatrix} = \mathbf{A} \begin{bmatrix} {}^{t}\ddot{x} \\ {}^{t}\dot{x} \\ {}^{t}x \end{bmatrix} + \mathbf{L}\, {}^{t+\Delta t}r \qquad \mathbf{A} = \begin{bmatrix} -(\frac{1}{2}-\alpha)\beta - 2(1-\delta)\kappa & \frac{1}{\Delta t}(-\beta - 2\kappa) & \frac{1}{\Delta t^2}(-\beta) \\ \Delta t[1 - \delta - (\frac{1}{2}-\alpha)\delta\beta - 2(1-\delta)\delta\kappa] & 1 - \beta\delta - 2\delta\kappa & \frac{1}{\Delta t}(-\beta\delta) \\ \Delta t^2[\frac{1}{2} - \alpha - (\frac{1}{2}-\alpha)\alpha\beta - 2(1-\delta)\alpha\kappa] & \Delta t(1 - \alpha\beta - 2\alpha\kappa) & (1-\alpha\beta) \end{bmatrix} $$

$$ \beta = \left(\frac{1}{\omega^2\,\Delta t^2} + \frac{2\xi\delta}{\omega\,\Delta t} + \alpha\right)^{-1}; \qquad \kappa = \frac{\xi\beta}{\omega\,\Delta t} \qquad \mathbf{L} = \begin{bmatrix} \frac{\beta}{\omega^2\,\Delta t^2} \\ \frac{\beta\delta}{\omega^2\,\Delta t} \\ \frac{\alpha\beta}{\omega^2} \end{bmatrix} \tag{385} $$

- Again stability requires eigenvalues of $\mathbf{A}$ to satisfy $|a_i| \leq 1$ and if they have lower geometric multiplicity than algebraic multiplicity ($n_i^G < n_i^A$) satisfying $|a_i| < 1$ as discussed in (338).
- The analysis of eigenvalues of $\mathbf{A}$ results in the following conditions on the stability of Newmark method based on its parameters (α, δ) and Δt,

$$ \begin{cases} \text{Unconditional stable:} & 2\alpha > \delta \geq \frac{1}{2} \\ \text{Conditional stable:} & \begin{bmatrix} \delta \geq \frac{1}{2} \\ \alpha < \frac{\delta}{2} \\ \overline{\Delta t} = \omega \Delta t \leq \Omega_{\text{crit}} \end{bmatrix} \end{cases} \qquad \text{where} \tag{386a} $$

$$ \Omega_{\text{crit}} = \frac{\xi\left(\delta - \frac{1}{2}\right) + \left[\frac{\delta}{2} - \alpha + \xi^2\left(\delta - \frac{1}{2}\right)^2\right]^{\frac{1}{2}}}{\frac{\delta}{2} - \alpha}, \quad \text{critical normalized sampling frequency} \tag{386b} $$

- $\overline{\Delta t}$ is normalized time step (also called normalized frequency).
- As usual when the method is conditionally stable ω in (386a) will be the worst (*i.e.,* maximum) frequency that the MDOF discrete problem ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = 0$) can model $\max_l(\omega_l^h)$.
- In practice we replace this with more convenient and conservative value ω_e^m, *i.e.,* the highest frequency of individual elements.

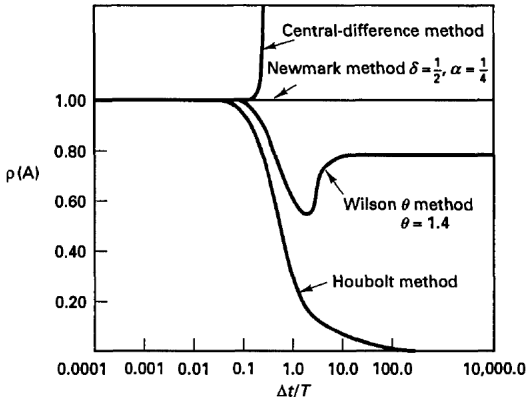
5.4 Practical considerations in using time marching methods

5.4.1 Control of high frequency numerical noise

- In the figure observe spectral radius of different time marching methods versus normalized element size.
- $T = \frac{\omega}{2\pi}$ is the period of a given SDOF.
- Clearly, as expected central-difference method becomes unstable for $\Delta t/T > \frac{1}{\pi}$: As we observed in (357) (also (358)) central difference method is stable if $\Delta t\omega \leq 2$, $T = \frac{\omega}{2\pi} \Rightarrow \Delta t/T \leq \frac{1}{\pi}$
- Other methods in the figure are unconditionally stable.
- One very important aspect of a time marching method in these plots is,

$$\rho_\infty = \lim_{\Delta t/T \to \infty} \rho(\mathbf{A}(\frac{\Delta t}{T})) \tag{387}$$

for example for Wilson-θ method $\rho_\infty \approx 0.8$



- $\Delta t/T \to \infty$ for individual SDOFs of a MDOF system (ω is in fact ω_i^h) can happen under two conditions which have important implications:
  1. $\Delta t \to \infty$ (**Too large of a time step**) which means very large time step is taken with respect to T. Often this can be a source of large numerical dissipation if $\Delta t \ll \max_i T_i$ (*i.e.,* time step is much larger than the period of the lowest natural mode) and $\rho_\infty < 1$. Having such high time steps can be afforded in unconditionally stable methods. If this condition occurs, this a sign that too large of a time step from numerical error perspective is taken.
  2. $T \to 0$ (*i.e.,* $\omega \to \infty$ **High frequency modes**): In this case, we are dealing with high frequency modes of the problem. Below, we discuss how by optimizing (having smallest) ρ_∞ we can effectively eliminate high frequency numerical noise.
- Assuming that case one is not of concern (*i.e.,* not too large of a time step is taken to quickly dissipate the solution by the numerical time integration when $\rho_\infty < 1$) a main concern of a numerical integration if the control of high frequency numerical noise.
- High-frequency behavior: "Because the higher modes of semi-discrete structural equations are artifacts of the discretization process and not representative of the behavior of the governing partial differential equations, it is generally viewed as desirable and often is considered absolutely necessary to have some form of algorithmic damping present to remove the participation of the high-frequency modal components." [Hughes, 2012].
- Figure below shows how low frequency part of the solution does not damp out much ($\Delta t/T_1 = 0.01, 0.1$) as for these low values of $\Delta t/T$ $\rho(\mathbf{A}) \lessapprox 1$. On the other hand, for high(er) frequency content (low(er) T) $\Delta t/T_i = 1, 10, 1001000$ $\rho(\mathbf{A}) \to \rho_\infty$ and these waves are almost entirely dissipated. This is the desired response as we want to maintain the physical part of the solution and dissipate / filter numerical noise.

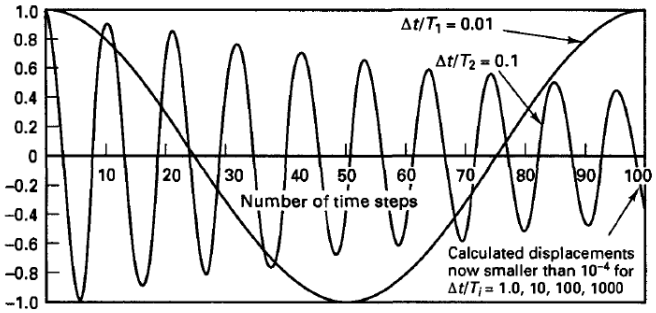
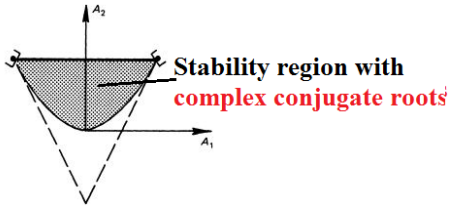


Figure 9.7 Displacement response predicted with increasing $\Delta t/T$ ratio; Wilson θ method, $\theta = 1.4$ [Bathe, 2006]

- The control of high frequency noise can be achieved by minimizing ρ_∞ to quickly dissipate any high frequency numerical noise that can be introduced in the solution.
- As an example, we optimize parameters for the Newmark method:
  - As was shown in the previous figure $\delta = \frac{1}{2}$ results in $\rho_\infty = 1$, *i.e.*, no dissipation of high frequency noise.
  - We need to choose $d > 0$ to have $\rho_\infty < 1$.
  - For a given δ we can optimize α such that high frequency dissipation is maximized (*i.e.*, ρ_∞ minimized).
  - This condition is created by requiring the eigenvalues of the amplification factor to be complex conjugate values.
  - Remembering for a 2×2 amplification matrix $\mathbf{A}$ such condition is $A_1^2 > A_2$; *cf.* (363).

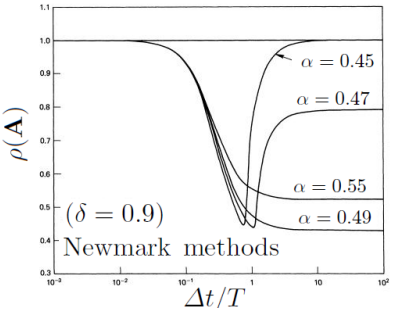


  - To ensure that the amplification factors are complex conjugate parameters α, δ from (386) are restricted as:

$$\begin{cases} \text{Unconditional stable:} & \begin{bmatrix} 0 \le \xi < 1 \\ \delta \ge \frac{1}{2} \\ \alpha \ge \frac{\left(\delta+\frac{1}{2}\right)^2}{4} \ge \frac{\delta}{2} \end{bmatrix} \\ \text{Conditional stable:} & \begin{bmatrix} 0 \le \xi < 1 \\ \delta \ge \frac{1}{2} \\ \alpha < \frac{\delta}{2} \\ \overline{\Delta t} = \omega \Delta t \le \ \Omega_{\text{bif}} \end{bmatrix} \end{cases} \quad \text{where} \tag{388a}$$

$$\Omega_{\text{bif}} = \frac{\frac{\xi}{2}\left(\delta - \frac{1}{2}\right) + \left[\left(\delta + \frac{1}{2}\right)^2/4 - \alpha + \xi^2\left(\alpha - \frac{\delta}{2}\right)\right]^{\frac{1}{2}}}{\left(\delta + \frac{1}{2}\right)^2 - \alpha} \tag{388b}$$

  - We observe that (388) is more restrictive than (386).
  - Ω_{bif} corresponds to normalized frequency (also called normalized time step) where bifurcation of real to complex complex eigenvalues occurs in the 2×2 matrix amplification factor for the Newmark method.
  - Equation (388) is for an under-damped problem ($\xi < 1$). For $\xi = 0$ we have $\Omega_{\text{bif}} = \left[\left(\delta + \frac{1}{2}\right)^2 - \alpha\right]^{-\frac{1}{2}}$.
  - Below we discuss how the value of α can be optimized based on δ (unconditionally stable case). There are three cases:
    - $\alpha < \frac{\left(\delta+\frac{1}{2}\right)^2}{4}$: For $\delta = 0.9 \Rightarrow \alpha < 0.49$. In this case eigenvalues of $\mathbf{A}$ bifurcate to complex conjugate values, but then past some $\Delta t/T$ value they bifurcate back to higher values. For the minimum $\alpha = \delta/2$ for unconditional stability we even have $\rho_\infty \to 1$ resulting in no dissipation for high frequency oscillations (noises). This can be see in cases $\alpha = 0.45$ and $\alpha = 0.47$.
    - $\alpha > \frac{\left(\delta+\frac{1}{2}\right)^2}{4}$: For $\delta = 0.9 \Rightarrow \alpha > 0.49$. In this case eigenvalues of $\mathbf{A}$ do not bifurcate but have $\rho_\infty < 1$. The case $\alpha = 0.55$ is shown.
    - $\alpha = \frac{\left(\delta+\frac{1}{2}\right)^2}{4}$: For $\delta = 0.9 \Rightarrow \alpha = 0.49$. This the optimum value: In this case eigenvalues bifurcate and result in an optimum (minimum) $\rho_\infty < 1$ for a given δ.



Accordingly, in practice to have the best dissipation of higher frequency noise we often set $\alpha = \frac{\left(\delta+\frac{1}{2}\right)^2}{4}$

- There are some other approaches to dissipatte high frequency noise:

1. Artificial damping: Similar to D_h in (114) in the context of FV methods various types of numerical damping operators can be added to a numerical method to control high frequency oscillations and other numerical artifacts. However, depending on the type of numerical method special case should be taken as at times they only damp an intermediate band of frequencies without significant effect in the all-important high modes.
2. α-method: Hilber-Hughes-Taylor (HHT) method:
   - The approach discussed above with choosing $\alpha = \frac{\left(\delta+\frac{1}{2}\right)^2}{4}$ requires $\delta > \frac{1}{2}$ which results in one order loss of accuracy compared to $\delta = \frac{1}{2}$ (first order compared to second order in Δt). For $\delta = \frac{1}{2}$ and $\alpha = \frac{1}{4}$ from $\frac{\left(\delta+\frac{1}{2}\right)^2}{4}$ yet the scheme is nondissipative and ρ_∞ rendering it ineffective in dissipative high frequency noise.
   - HHT suggested the following modification $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ by splitting the values to t_n and t_{n+1}:

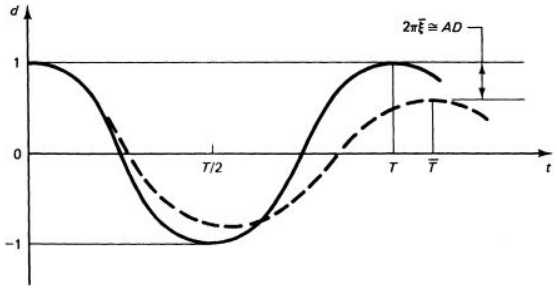
$$\mathbf{M}\ddot{\mathbf{U}}_{n+1} + (1+\alpha)\mathbf{C}\dot{\mathbf{U}}_{n+1} - \alpha\mathbf{C}\dot{\mathbf{U}}_n + (1+\alpha)\mathbf{K}\mathbf{U}_{n+1} - \alpha\mathbf{K}\mathbf{U}_n = \mathbf{R}(t_{n+1+\alpha}), \quad t_{n+1+\alpha} = (1+\alpha)t_{n+1} - \alpha t_n \tag{389}$$

   - While the MDOF is modified by α coefficient the Newmark update is done as usual by employing the following approximation: $\dot{\mathbf{U}}_{n+1} = \dot{\mathbf{U}}_n + \Delta t[(1-\delta)\ddot{\mathbf{U}}_n + \delta\ddot{\mathbf{U}}_{n+1}]$ and $\mathbf{U}_{n+1} = \mathbf{U}_n + \Delta t\dot{\mathbf{U}}_n + \frac{\Delta t^2}{2}[(1-2\bar{\alpha})\ddot{\mathbf{U}}_n + 2\bar{\alpha}\ddot{\mathbf{U}}_{n+1}]$ where $\bar{\alpha}$ is the usual α parameter that is marked by $\bar{(.)}$ to distinguish it from HHT α parameter.
   - If $\alpha \in [-\frac{1}{3}\ 0]$, $\delta = \frac{1-2\alpha}{2}$, and $\bar{\alpha} = \frac{(1-\alpha)^2}{4}$ an unconditionally stable, **second order** accurate scheme results.

5.4.2 Measures of accuracy: L2 error, numerical dissipation and dispersion

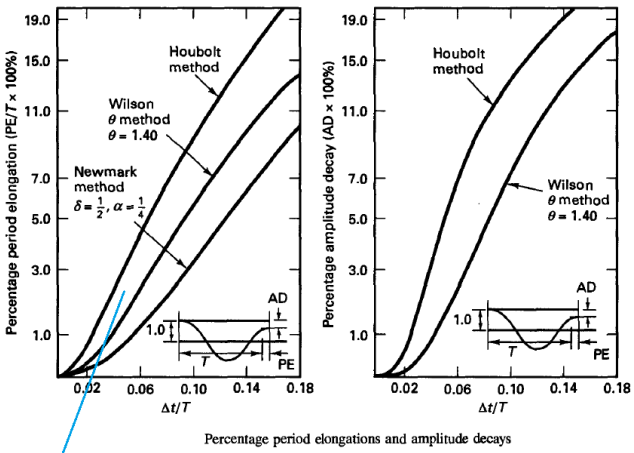
- Consider the solution for the undamped second order ODE shown below,

$$\left.\begin{aligned} \ddot{x} + \omega^2 x = 0 \\ {}^0x = 1.0; \qquad {}^0\dot{x} = 0.0; \qquad {}^0\ddot{x} = -\omega^2 \end{aligned}\right\} \tag{390}$$



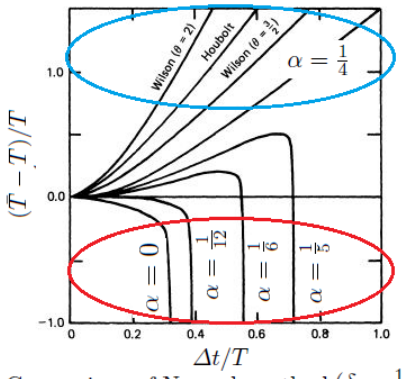
- The exact solution for this ODE is $x(t) = \cos(\omega t)$.
- The numerical solution may not be able to model the exact wave amplitude or period as shown in the figure.
- We have the following definitions,

1. Amplitude Decay (AD): The amount the amplitude of the wave decreases relative to the exact solution in one period. Note, the exact solution may actually be dissipative for example when damping is nonzero $\ddot{x}+2\xi\omega\dot{x}+\omega^2 x = 0$, yet we can formalize and separate physical dissipation from numerical one.
2. Period elongation (PE): The difference between numerical period $\bar{T}$ and exact period $T = \frac{2\pi}{\omega}$. that is $\bar{T} - T$. Interestingly, we often have the following trend:
   (a) Implicit integration $\Rightarrow$ PE ≥ 0: With implicit integration methods numerical period is often **longer** $\bar{T}$ (shorter frequency $\bar{\omega}$) than the exact period T. Examples are Wilson-θ, Houbolt, trapezoidal and unconditional stable Newmark methods.
   (b) Explicit integration $\Rightarrow$ PE ≤ 0: With explicit integration methods numerical period is often shorter $\bar{T}$ (longer frequency $\bar{\omega}$) than the exact period T.



Implicit methods increase wave period.

- More examples are shown in the figure:
  1. For Newmark method $\delta = \frac{1}{2}$, $\alpha < \frac{\delta}{2} = \frac{1}{4}$ correspond to conditional stable regime which as can be seen in the figure shorten the period / elongate frequency.
  2. On the other hand, implicit (unconditionally stable) Newmark method for $\alpha = \frac{1}{4}$ and Wilson and Houbolt methods elongate the period / shorten frequency.

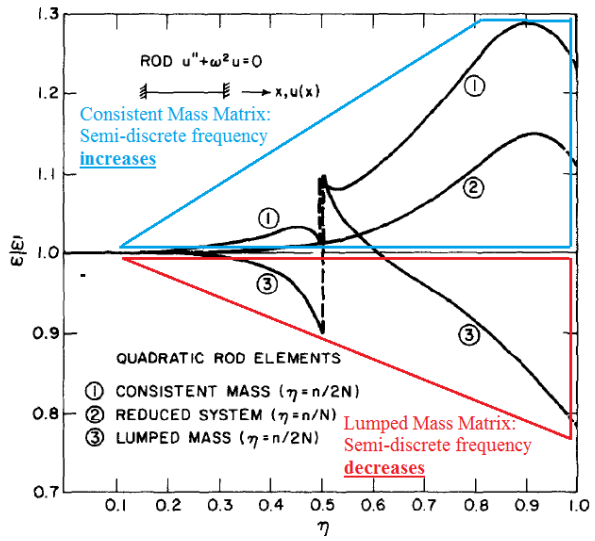


Implicit integrators: Period gets **longer**

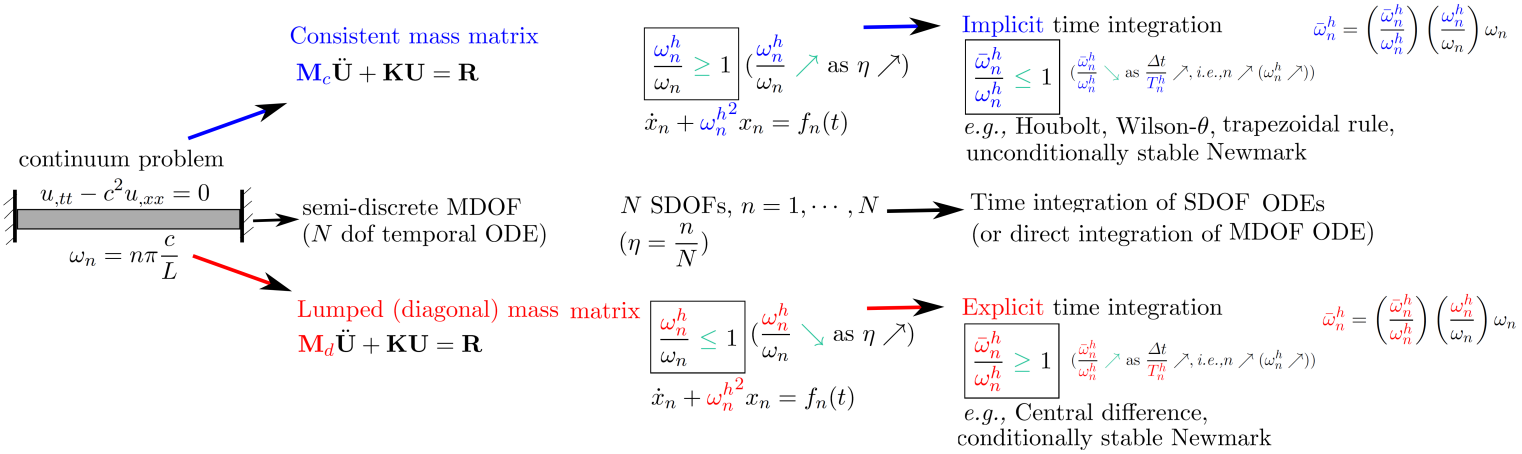
Explicit integrations: Period gets **shorter**

Comparison of Nemark method ($\delta = \frac{1}{2}$) and Wilson and Houbolt methods.

- Consider that we are solving 1D elastodynamic problem $u_{,tt} - c^2 u_{,xx} = 0$ for a double end fixed bar of length L.
- $\Rightarrow$ the exact natural frequencies are $\omega_n = n\pi \frac{c}{L}$.
- Numerical modes ω_n^h for an N dof MDOF spatial discretization $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{K}\mathbf{U} = 0$ take the following form:
  1. Consistent mass matrix: $\omega_n^h > \omega_n$.
  2. Lumped mass matrix: $\omega_n^h < \omega_n$.
- For this 1D problem, we observe that the relative error $\frac{\omega_n^h}{\omega_n}$ only depends on how far we are from the number of modes the MDOF model can capture through $\eta = \frac{n}{N}$.



- All the way from the continuum level frequencies to semi-discrete MDOF discretization $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$ to its numerical integration we deal with three groups of frequencies:
  1. ω: exact frequencies of the continuum problem.
  2. ω^h: semi-discrete frequencies are natural frequencies of MDOF FEM discretization $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$.
  3. $\bar{\omega}^h$: frequencies produced (realized) by time integration of semi-discrete MDOF $\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$.
- Eventually accuracy of the numerically integrated ODE in representing periods of waves depends on:
  1. How accurately semi-discrete frequencies are modeled: $\frac{\omega^h}{\omega}$.
  2. How accurately time-integration models frequencies: $\frac{\bar{\omega}^h}{\omega^h}$.
- We want to **match** time integration methods with appropriate mass matrix option so that the errors from the following two steps are in opposite directions and to some extend cancel each other out.
- This concept is shown in the next figure.



- Based on the results from previous slide, we make the following conclusions for the choice of mass matrix based on time integration model:

  1. Implicit Integration → Consistent mass matrix: Implicit integration methods often elongate T^h (shorten ω^h: $\frac{\bar{\omega}^h}{\omega^h} \leq 1$) which best is matched with consistent mass matrix as it shortens periods T (elongate ω: $\frac{\omega^h}{\omega} \geq 1$).
  2. Explicit Integration → Lumped mass matrix:
     (a) Implicit integration methods often shorten T^h (elongate ω^h: $\frac{\bar{\omega}^h}{\omega^h} \geq 1$) which best is matched with lumped mass matrix as it elongates periods T (shorten ω: $\frac{\omega^h}{\omega} \leq 1$).
     (b) In addition use of lump mass matrix + damping $\mathbf{C} = \mathbf{0}$ + explicit method enables a **local and trivial** linear system solve. For example, in (247) for central difference method we had $\hat{\mathbf{M}} = \frac{1}{\Delta t^2}\mathbf{M} + \frac{1}{2\Delta t}\mathbf{C}$ which for $\mathbf{C} = \mathbf{0}$ yielded $\hat{\mathbf{M}} = \frac{1}{\Delta t^2}\mathbf{M}$. If further a lumped mass matrix is used the update equation simply becomes $U_i^{n+1} = \frac{\Delta t^2}{m_{ii}} R_i^n$; *cf.* (248).

- That is, **use of lumped mass matrix for explicit integration methods not only can result in a local solution process but also is preferred from numerical error perspective (period elongation error)**.

- The idea that semi-discretization by FEM and time integration scheme errors in period (frequency) can cancel each other out has provided means to optimize the mass matrix such that $\bar{\omega}^h$ exactly matches ω.

- The details of this approach are described in [Hughes, 2012] where it is assumed Newmark method with $\delta = \frac{1}{2}$ is employed for 1D elastodynamic problem with no damping; *i.e.*, $u_{,tt} - c^2 u_{,xx} = 0$ for $c = 1$.

- The mass matrix is parameterized with r:

$$K^e = \frac{AE}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad M^e = \frac{A\rho}{L}\begin{bmatrix} \frac{1}{2} - r & r \\ r & \frac{1}{2} - r \end{bmatrix} \tag{391}$$

- For $r = \frac{1}{6}, 0, \frac{1}{12}$ we recover consistent mass, lumped mass, and a higher order mass matrix (resulting in higher convergence rates for natural modes / frequencies), respectively,

$$\begin{array}{lll} \text{Consistent mass}(r = \frac{1}{6}) & \text{Lumped mass}(r = 0) & \text{High order mass}(r = \frac{1}{12}) \\ M^e = \frac{A\rho}{6L}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} & M^e = \frac{A\rho}{2L}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & M^e = \frac{A\rho}{12L}\begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix} \end{array} \tag{392a}$$

- The goal is to optimize r based on $\frac{\Delta t}{h}$ and α(Newmark integration parameter).

$$\sin^2\frac{\bar{\omega}^h \Delta t}{2} = \frac{\sin^2(\omega h/2)}{4[\alpha - r(h/\Delta t)^2]\sin^2(\omega h/2) + (h/\Delta t)^2} \tag{393}$$

- We observe that if

  1. $\Delta t = h$ (for $c \neq 1 \Rightarrow \Delta t = \frac{h}{c}$).
  2. $\alpha = r$

  then (393) implies that $\bar{\omega}^h = \omega$. That is, we exactly preserve the continuum natural frequency all the way to the integration of the problem in time.

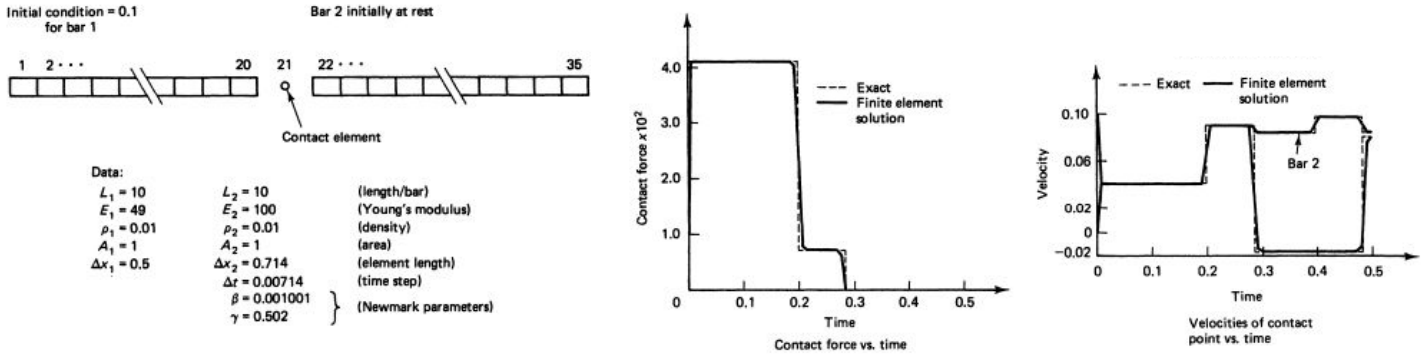
- In this case the errors introduced by finite element spatial discretization, the particular mass matrix and temporal algorithm all cancel to yield exact results.

- Time step $\Delta t = \Delta t/c$ is called the characteristic time step.

- It is interesting to note that reducing the size of the time step Δt while holding the mesh length h fixed can only worsen the results.

- In this case we converge to the exact solution of the spatially discrete, temporally continuous system (i.e., "mass points and springs") rather than the exact solution of $u_{,tt} - c^2 u_{,xx} = 0$.

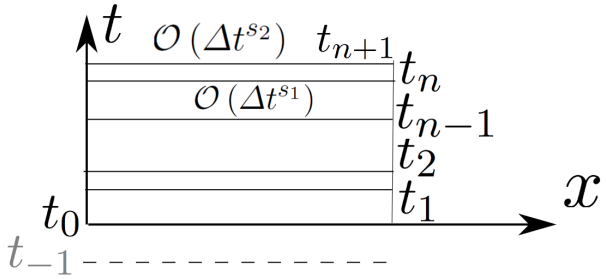
- In more general settings (*e.g.*,, unequal element lengths, variable material properties, multidimensional problems, etc.), results obtained by matched methods, such as central differences and lumped mass, will not be exact. However, it is felt that results obtained by matched methods will generally be superior to inappropriate combinations, such as consistent mass and central differences [Hughes, 2012].
- For example, results below from [Hughes et al., 1976b] show a perfect example of how optimizing the numerical model parameters and matching (optimizing) time integration method and mass matrix can result in excellent results.
- This problem demonstrates a contact problem between two dissimilar bars.
- Upon contact there is a sharp transition from traction free state to compressive stress state in the bars.
- Also, when the compressive waves reflect from the free end of the bar(s) they result in transition of the bars from contact to separation mode which reverses the stress state.
- This is a benchmark problem for checking contact algorithms and solving this problem is not trivial.
- The transitions often result in widespread numerical artifacts and noise in both bars, but we observe very good solutions herein with just a few elements.



5.4.3 Practical considerations in using time integration methods

- Implicit vs. explicit time integration: As discussed before the choice of implicit versus explicit integration depends on many factors:
  1. Structural dynamics vs. wave propagation problems: If the frequency content of the loads is not rich and for example limited to the first few natural modes, it is generally better and much more efficient to use an implicit time integration method either directly or applied to the first few SDOFs of modal decomposition.
  2. Linearity: One consideration is that even for nonlinear problems explicit integration results in a linear system update which can have a trivial solution if a lumped mass matrix is used and damping is zero. In any case, whether an implicit or explicit solver is better for a nonlinear problem may require a case-by-case decision.
  3. Memory constraints and parallel computing: The solution of explicit methods is local and even does not require the assembly of the stiffness matrix (and mass matrix), all being advantages in better use of memory and for parallel computing.

  Just as a reminder, if an explicit method is used, a lump mass matrix must be employed both for rendering the matrix equation trivial and having smaller period (frequency) error.



- Adaptivity in time, multi-step vs. single-step methods

- h-adaptivity and p-adaptivity in time: Changing the time step size Δt (h-adaptivity in time) or the order of accuracy (p-adaptivity/enrichment in time) are very challenging and often impractical with linear multi step methods (number of steps $k \geq 2$) these formulas are written for constant Δt and require previous time step values.
- Self-starting is another challenge of LMS methods as they need values for t_{-1} (and possibly beyond) at the first time steps. This is a burden in coding these methods and also can make their analysis more difficult.

Basically, one step methods such as Newmark and Wilson-θ that require only values of one step in the form of $\mathbf{U}, \dot{\mathbf{U}}$, (and $\ddot{\mathbf{U}}$) are more flexible in this respect.

- Order of accuracy in time $\mathcal{O}\left(\Delta t^s\right)$: Experience has indicated that in structural dynamics, second-order accurate methods are vastly superior to first-order accurate methods. In some applications, even higher orders of accuracy are required either for efficiency or accuracy considerations. Below, are a few comments on achieving high orders of accuracy in time:
  - LMS methods such as Houbolt, central difference, *etc.*. A consequence of Dahlquist's theorem is that there is no third-order accurate unconditionally stable linear multi-step method (LMS). Thus we must be content with second-order accuracy is an implicit LMS step is desired to be used. Furthermore, the second-order method with the smallest error constant is the trapezoidal rule. However, trapezoidal rule does not offer any dissipation of high frequency numerical artifacts. Other time integration methods that are high order, yet dissipate high order dissipation (such as Hilber-Hughes-Taylor (HHT) method) can for this reason be preferred to LMS methods if an implicit solver is desired. For explicit LMS methods achieving higher orders of accuracy becomes challenging as they require a much larger "historical data pool"; *cf.* [Süli and Mayers, 2003] §12.6 Linear multi-step methods and §12.9 Dahlquist's theorems for more information.
  - Runge-Kutta methods: Runge-Kutta (RK) methods, with RK4 perhaps the most popular member of them are designed to achieve high temporal orders of accuracy by updating the solution to the next step by computing solutions / values at intermediate stages. Some points to consider for RK methods are:
    - Butcher barrier: Given that for achieving higher orders of accuracy the number of matching equations between numerical integration scheme and Taylor expansion of exaction solution grows at a faster pace than the order of the method, the number of stages grow faster than the order of the RK methods, a feature that is formally expressed by the Butcher barrier [Butcher, 2005]. Accordingly, the efficiency of Runge-Kutta schemes drastically decreases for orders greater than four.
    - Application to second order temporal ODEs: RK methods are naturally formulated for first order ODEs. Although one can reformulate the update equations for a second order ODE (*e.g.,* either by directly using Taylor series expansion of the exact solution or employing the first order ODE update equations) the direct use of RK methods to second order ODEs, *e.g.,* elastodynamic problem, is very limited. However, one can express second order (or higher order) PDEs in time as a system of first order PDEs. In which case, RK method can be directly used. Unlike continuous FEMs where RK methods are rarely used, their use is common with discontinuous Galerkin methods; *cf.* Local DG (LDG) and RKDG methods [Cockburn and Shu, 1998a, Cockburn and Shu, 1998b].
    - Implicit RK methods: If an implicit numerical integration method is desired, implicit RK methods can be used. One advantage is that that relatively high temporal orders of accuracy can be modeled by these methods, whereas for example there is no unconditionally stable LMS method beyond second order accuracy. However, one disadvantage is the need to solve a sN coupled system with implicit RK methods where s is the number of stages and N the number of spatial dofs. This can be a major concern both from memory storage and computational cost perspectives, especially for very large N and nonlinear problems.
  - Cauchy-Kovalewski (CK) / Lax-Wendroff (LW): Achieving high temporal orders of accuracy in time is much more challenging than in space for time marching methods since the solution is only given at discrete time values as opposed to spatial representation of solution by basis functions. This is the main source of difficulty in achieving arbitrary high temporal orders of accuracy with the aforementioned methods.

    A successful approach to circumvent discrete representation of solution in time is Cauchy-Kovalewski (CK) or Lax-Wendroff approach which involves the following steps:
    - Expanding the solution in time using the Taylor series:

$$u(x,t+\Delta t) \approx u(x,t) + \Delta t u_{,t}(x,t) + \frac{\Delta t^2}{2} u_{,tt}(x,t) + \cdots + \frac{\Delta t^s}{s!} u^{(s)}(x,t) + \mathcal{O}\left(\Delta t^{s+1}\right) \tag{394}$$

    - Replacing temporal derivatives with spatial derivatives using the underlying PDE; *e.g.,*, for the advection equation

$$u_{,t} - a u_{,x} = 0 \quad \text{(advection equation)} \qquad \Rightarrow \qquad \frac{\mathrm{d}^s u}{\mathrm{d}t^s} = a^s \frac{\mathrm{d}^s u}{\mathrm{d}x^s} \tag{395}$$

    - For a given position x (which can correspond to a nodal position in an FEM mesh) obtain $\frac{\mathrm{d}^s u}{\mathrm{d}x^s}$ from FE spatial discretization. Since space is discretize with the more flexible FE method and elements with any desired shape function orders can be formulated, the ability of having high order $\frac{\mathrm{d}^s u}{\mathrm{d}x^s}$ is relatively trivial compared to having high temporal orders of accuracy with time marching schemes.

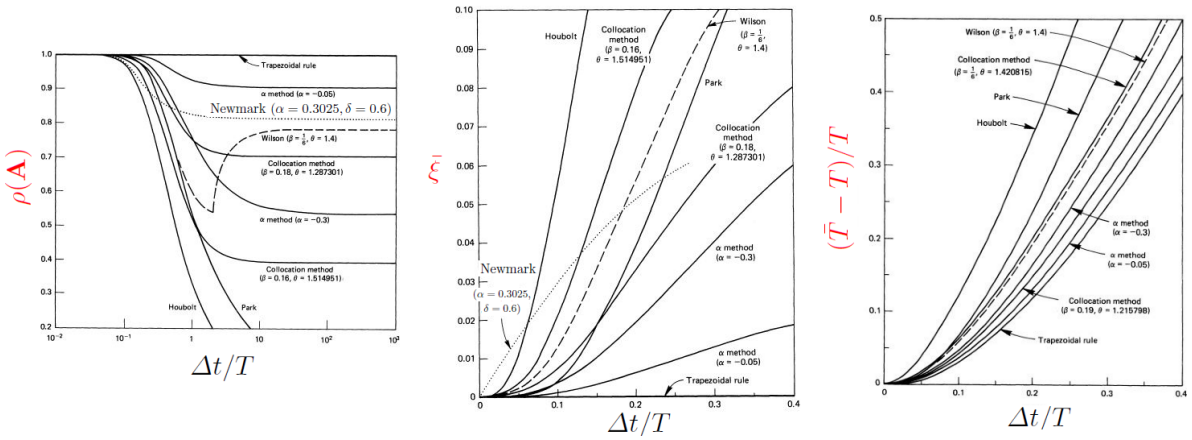
* Finally, by plugging spatial derivatives $\frac{\mathrm{d}^s u}{\mathrm{d}x^s}$ in (395) and corresponding $\frac{\mathrm{d}^s u}{\mathrm{d}t^s}$ (which are PDE dependent) in (394) we formulate a method with arbitrary high order of accuracy in time.

Refer to [De Basabe and Sen, 2010] for the application of CK method to second order elastodynamic problem, and [Dumbser and Munz, 2005, Dumbser and Munz, 2006] for the application of CK method in the context of discontinuous Galerkin methods.

– Other high order temporal integration methods. Some notable methods are:

* Spacetime FEMs: While spacetime finite element methods are not using a time marching scheme to advance the solution in time (they directly solve space and time with FEM), the expression of the solution in time with FE shape functions means that arbitrary high temporal orders of accuracy can be achieved with these methods.
* Methods based on analytical expansion of solution: Consider the problem $\dot{\mathbf{u}} + \mathbf{A}\mathbf{u} = \mathbf{0}$ where $\mathbf{u}$ is a vector and $\mathbf{A}$ a matrix. The solution of this ODE is $\mathbf{u}(t) = \mathbf{u}(t=0)e^{-\mathbf{A}t}$. Herein, $\mathbf{u}$ can represent the vector of unknowns that can be obtain by FEM discretization; *cf.* (226b) ($\mathbf{M}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$) or (227) for temporal first order representation of (226a) ($\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{R}$). Such approaches and exponential of a matrix, are the basis of achieving arbitrary high temporal order of accuracy in some methods, *e.g.,* [Fahs, 2012].

- Control of high frequency numerical artifacts, amplitude damping (AD), and period elongation (PE)



a) Spectral radius b) Algorithmic damping ratio c) Relative period error

In the figure spectral radius, AD, and PE of various time marching methods is displayed versus the normalized time step $\Delta t/T$ ($T = \frac{2\pi}{\omega}$ is the period for a SDOF). We make the following observation (for more details, especially on some other methods displayed in the figure, refer to [Hughes, 2012]): Control of high frequency artifacts and damping of the solution: The spectral radii of the Houbolt and Park methods approach zero as $\Delta t/T \to \infty$ as is typical of backward-difference schemes. While $\rho_\infty = 0$ implies that high frequency artifacts are dissipated, the quick transition of ρ to 0 has adverse effects. These two methods are seen to affect the low modes (*i.e.,*, $\Delta t/T = 0.1$) too strongly, which means even at moderate time step sizes all components of the solution including low frequency content can be severely dissipated. The quick approach of $\rho(\mathbf{A})$ to zero also manifests itself in high amplitude decay for these methods.

On the other hand, for trapezoidal rule we have $\rho(\mathbf{A}) = 1$ which is due to the fact that the method is nondissipative. While the method is second order accurate, having no ability to dissipate high frequency content (numerical artifacts) is an undesirable property. As mentioned before, methods such as Hilber-Hughes-Taylor (HHT) (α-method) not only are temporally second order accurate but provide some dissipation for high frequency content.

Finally, we observe that the Newmark method ($\alpha = 0.3025, \delta = 0.6$) and α method (HHT) have a good high frequency content dissipation behavior while not being too much dissipative for the lower frequency content; we observe $\rho(\mathbf{A}) \approx 1$ for $\Delta t/T$ is small; *i.e.,* for low frequency ω (high period T modes) and having $\rho_\infty < 1$ ensures high frequency numerical noise gets dissipated.

In short, the following is a list of properties are deemed desirable for structural dynamic problems (*i.e.,* when first few modes are excited) [Hughes, 2012]:

1. Unconditional stability when applied to linear problems.
2. No more than one set of implicit equations to be solved at each step.
3. Second-order accuracy.
4. Controllable algorithmic dissipation in the higher modes.
5. Self-starting.

5.5 Element natural frequencies vs MDOF maximum frequency

- The stability of a linear MDOF system can be reduced to the stability analysis of its modal SDOFs; *cf.* §5.2.2.
- For establishing the maximum time step for a MDOF system (generated by FEM discretization of space) we mentioned that the maximum $\max_l(\lambda_l^h)$ natural frequency (eigenvalue) must be considered.
- However, computation of $\max_l(\lambda_l^h)$ requires a full modal analysis which can be extremely expensive.
- Fortunately, equation (311) provides a means to find a conservative (*i.e.*, higher) estimate for $\max_l(\lambda_l^h)$ by observing $\lambda_e^m \geq \max_l(\lambda_l^h)$ where λ_e^m is the highest frequency (eigenvalue) of the elements in the computational domain.
- λ_e^m can be easily computed for different element types.
- In this section we do the following,
  1. Prove that $\lambda_e^m \geq \max_l(\lambda_l^h)$.
  2. Obtain and present λ_e^m for different element types, including the effect of using consistent or lumped mass matrices.
  3. Discuss how an element's highest frequency is changed based on its spatial order of accuracy and provide recommendations on the spatial resolution of FEM meshes.

5.5.1 Maximum bound of MDOF eigenvalue by its element eigenvalues

The complete background for this proof (including Rayleigh's quotient) can be found in [Bathe, 2006, Hughes, 2012].

Using the Rayleigh quotient (see Section 2.6), we have with $\mathbf{K}^{(m)}$ and $\mathbf{M}^{(m)}$ defined in (4.19) and (4.25),

$$(\omega_n)^2 = \frac{\boldsymbol{\phi}_n^T \left(\sum_m \mathbf{K}^{(m)} \right) \boldsymbol{\phi}_n}{\boldsymbol{\phi}_n^T \left(\sum_m \mathbf{M}^{(m)} \right) \boldsymbol{\phi}_n} \tag{b}$$

Let

$$\mathcal{U}^{(m)} = \boldsymbol{\phi}_n^T \mathbf{K}^{(m)} \boldsymbol{\phi}_n \qquad \text{and} \qquad \mathcal{T}^{(m)} = \boldsymbol{\phi}_n^T \mathbf{M}^{(m)} \boldsymbol{\phi}_n$$

then

$$(\omega_n)^2 = \frac{\sum_m \mathcal{U}^{(m)}}{\sum_m \mathcal{T}^{(m)}} \tag{c}$$

Now consider the Rayleigh quotient for a single element,

$$\rho^{(m)} = \frac{\boldsymbol{\phi}_n^T \mathbf{K}^{(m)} \boldsymbol{\phi}_n}{\boldsymbol{\phi}_n^T \mathbf{M}^{(m)} \boldsymbol{\phi}_n} = \frac{\mathcal{U}^{(m)}}{\mathcal{T}^{(m)}} \tag{d}$$

Since $\mathbf{M}^{(m)}$ and $\mathbf{K}^{(m)}$ are of the same size as $\mathbf{K}$, we could theoretically imagine $\mathcal{U}^{(m)}$ and $\mathcal{T}^{(m)}$ to be zero (but not for all m). However, in any case we have for each element (see Section 2.6)

$$\mathcal{U}^{(m)} \leq (\omega_n^{(m)})^2 \mathcal{T}^{(m)}$$

and therefore from (c),

$$(\omega_n)^2 \leq \frac{\sum_m (\omega_n^{(m)})^2 \mathcal{T}^{(m)}}{\sum_m \mathcal{T}^{(m)}} \leq \left[\max_m \left(\omega_n^{(m)} \right)^2 \right] \frac{\sum_m \mathcal{T}^{(m)}}{\sum_m \mathcal{T}^{(m)}}$$

which proves (a). Note that in (b) we used the $\mathbf{K}^{(m)}$ and $\mathbf{M}^{(m)}$ matrices of element m defined in (4.19) and (4.25), that is with all boundary conditions (and the actions of the other elements) removed. Of course, the same proof is applicable if some elements are constrained at certain degrees of freedom (applied to the assemblage of elements).

5.5.2 Different element frequencies (eigenvalues)

- We reiterate the equations for a 1D elastodynamic element stiffness matrix K^e and M^e (391), (repeated here),

$$K^e = \frac{AE}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad M^e = \frac{A\rho}{L}\begin{bmatrix} \frac{1}{2}-r & r \\ r & \frac{1}{2}-r \end{bmatrix}$$

- The parameter r enables generating different mass matrices (392) (repeated here),

Consistent mass$(r=\frac{1}{6})$	Lumped mass$(r=0)$	High order mass$(r=\frac{1}{12})$
$M^e = \frac{A\rho}{6L}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$	$M^e = \frac{A\rho}{2L}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$M^e = \frac{A\rho}{12L}\begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}$

  where the higher order mass option corresponds to a case that natural frequencies and modes have higher convergence rate compared to consistent and lumped mass options.

- Natural frequency of the element are obtained from the following generalized eigenvalue problem,

$$K^e\boldsymbol{\Phi} = \omega^2 M^e\boldsymbol{\Phi} \quad ((179)) \qquad \frac{AE}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\boldsymbol{\Phi} = \omega^2\frac{A\rho}{L}\begin{bmatrix} \frac{1}{2}-r & r \\ r & \frac{1}{2}-r \end{bmatrix} \Rightarrow \left\{\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} - \omega'^2\begin{bmatrix} \frac{1}{2}-r & r \\ r & \frac{1}{2}-r \end{bmatrix}\right\}\boldsymbol{\Phi} = \mathbf{0},$$

$$\text{where} \quad \omega' := \frac{L\omega}{c}, c = \sqrt{E/\rho}, \Rightarrow \qquad 1-\omega'^2\left(\frac{1}{2}-r\right) = \pm\left(-1-r\omega'^2\right) \qquad \Rightarrow \qquad \begin{cases} \omega'_1 = 0 \\ \omega'_2 = \frac{2}{\sqrt{1-4r}} \end{cases} \tag{396}$$

- Thus the maximum natural frequency of the element of length L is

$$\boxed{\omega_e = \frac{c}{L}\frac{2}{\sqrt{1-4r}}} \tag{397}$$

- This corresponds to the first natural frequency of the a 2 end free bar of length L.
- The exact natural frequencies for this problem are $\omega_n = n\pi\frac{c}{L}$.
- The following table summarizes what the maximum frequency of an element is for different values of r.

Case	r	$\frac{\omega_e}{(c/L)}$	$\frac{\omega_e}{\omega_1}$	
Lumped mass	0	2	$\frac{2}{\pi} = 0.637$	(398a)
Higher order mass	$\frac{1}{12} = 0.833$	$\sqrt{6} = 2.45$	$\frac{2}{\pi} = 0.780$	(398b)
Matching exact frequency	$\frac{1}{4}-\frac{1}{\pi^2} = 0.1487$	$\pi = 3.14$	1	(398c)
Consistent	$\frac{1}{6} = 0.1667$	$2\sqrt{3} = 3.464$	1.103	(398d)
Max r	$\left(\frac{1}{4}\right)^- = 0.25$	∞	∞	(398e)

- As expected the lumped mass matrix option provides a smaller ω_e than the exact one: $\frac{\omega_e}{\omega} = 0.637$ and
- the consistent mass matrix provides a larger ω_e than the exact one: $\frac{\omega_e}{\omega} = 1.103$.

- ω_e^m, *i.e.,* the maximum value of all elements' maximum frequency is given as (for a 1D problem where all the elements are 1D first order bar elements all with the same c),

$$\omega_e^m = \max_e(\omega_e) = \frac{c}{h_{\min}} \begin{cases} 2 & \text{Lumped mass matrix} \\ \sqrt{6} & \text{High order mass matrix} \\ 2\sqrt{3} & \text{Consistent mass matrix} \end{cases}, \quad \text{where} \quad h_{\min} = \min_e L_e \tag{399}$$

- Clearly, in a general case we simply use $\omega_e^m = \max_e(\omega_e)$ which may not correspond to the element of the smallest size.
- Now, for any given conditionally stable we can conservatively substitute the maximum frequency (eigenvalue) of the MDOF ODE $\max_l(\omega_l^h)$ by ω_e^m.
- This is exactly what we did for central difference method. Recalling from
- For example, if an explicit central difference method is used for time integration, recalling from (358b) we had,

$$\Delta t \leq \frac{2}{\omega_e^m}$$

- which based on the values in (403) we obtain (again to have an explicit value for ω_e^m we assume all elements are first order 1D bars with the same A, E),

$$\Delta t \leq \frac{2}{\omega_e^m} = \frac{h_{\min}}{c} \begin{cases} 1 & \text{Lumped mass matrix} \\ \sqrt{\frac{2}{3}} \approx 0.667 & \text{High order mass matrix} \\ \frac{1}{\sqrt{3}} \approx 0.577 & \text{Consistent mass matrix} \end{cases}, \quad \text{for } \textbf{central difference method} \tag{400}$$

- Interestingly we observe that consistent mass matrix option provides smaller critical time step than lumped mass matrix (0.577 times smaller in this particular case).
- This is because consistent mass matrix overestimates natural frequencies and lumped mass matrix underestimate them. $\Rightarrow$
- Consistent-mass matrices tend to yield smaller critical time steps than lumped-mass matrices [Hughes, 2012].
- Other important, and obvious, aspects are that the maximum time step also depends on the conditional stable method used
- and that for unconditional stable methods, there is no critical time step.
- Table below lists some critical time steps for central difference method.

E = Young's modulus, ν = Poisson's ratio, L = length (side length) of element, A = cross-sectional area of element, ρ = mass density, I = flexural moment of inertia, t = thickness of plane stress element, c = one-dimensional wave speed = $\sqrt{E/\rho}$

TABLE 9.5 *Central difference method critical time steps for some elements:* $\Delta t_{cr}^{(m)} = T_n^{(m)}/\pi = 2/\omega_n^{(m)}$

Two-node truss element:

$$\mathbf{K}^{(m)} = \frac{AE}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}; \qquad \mathbf{M}^{(m)} = \frac{\rho L}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Delta t_{cr}^{(m)} = \frac{L}{c};$$

Two-node beam element (see Example 4.1):

$$\mathbf{K}^{(m)} = \frac{EI}{L}\begin{bmatrix} \frac{12}{L^2} & -\frac{6}{L} & -\frac{12}{L^2} & -\frac{6}{L} \\ & 4 & \frac{6}{L} & 2 \\ & \text{Sym.} & \frac{12}{L^2} & \frac{6}{L} \\ & & & 4 \end{bmatrix}$$

$$\mathbf{M}^{(m)} = \frac{\rho AL}{24}\begin{bmatrix} 12 & 0 & 0 & 0 \\ & L^2 & 0 & 0 \\ & \text{Sym.} & 12 & 0 \\ & & & L^2 \end{bmatrix}$$

$$\Delta t_{cr}^{(m)} = \sqrt{\frac{A}{48I}}\frac{L^2}{c}$$

Four-node square plane stress element (see Example 4.6):

$$\mathbf{K}^{(m)} = \frac{Et}{1-\nu^2}\begin{bmatrix} \frac{3-\nu}{6} & & & \text{Elements are function of } \nu \\ & \ddots & & \\ \text{Sym.} & & & \\ & & & \frac{3-\nu}{6} \end{bmatrix}$$

$$\mathbf{M}^{(m)} = \frac{\rho L^2 t}{4}\begin{bmatrix} 1 & & & \\ & 1 & \text{Zeros} & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

$$\Delta t_{cr}^{(m)} = \frac{L}{c}\sqrt{1-\nu}$$

5.5.3 Effect of element order on maximum time step and other considerations

- For a lumped mass matrix and second order ($p = 2$) 1D bar element we obtain,

$$\omega_e = 2\sqrt{6}\frac{c}{L}, \qquad p = 2, \text{ lumped mass matrix} \tag{401}$$

- Recalling the maximum frequency from (398a) for $p = 1$ and lumped mass matrix we have the following,

$$\omega_e = \frac{c}{L}\begin{cases} 2 & p = 1 \\ 2\sqrt{6} & p = 2 \end{cases}, \quad \text{lumped mass matrix} \tag{402a}$$

- These values can be used to obtain the stable time step of any conditionally stable time step. For example assuming a central difference method is used for time integration, recalling from (358b) we have $\Delta t \leq \frac{2}{\omega_e^m}$. Thus, given the maximum time step (for lumped mass matrix option) becomes,

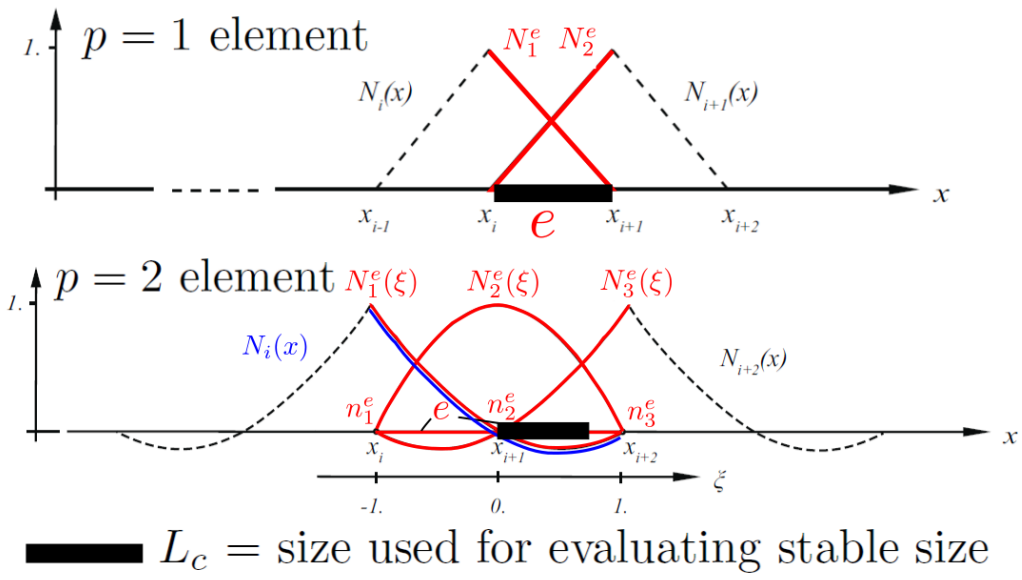
$$\Delta t \leq \frac{2}{\omega_e^m} = \frac{h_{\min}}{c}\begin{cases} 1 & p = 1 \\ \frac{1}{\sqrt{6}} \approx 0.408 & p = 2 \end{cases}, \quad \text{for central difference method and lumped mass matrix} \tag{403}$$

- So, The time step for $p = 2$ is only 0.408 times of $p = 1$.

- By considering the fact that for $p=2$ element, there is an internal node, one argues that a $p=2$ mesh should be compared with a $p=1$ mesh with half the element size.
- In fact, we can express the same equation for stable time step in the form,

$$\Delta t \le \frac{h_{c\min}}{c}, \quad \text{where} \quad h_{c\min} = \begin{cases} h_{\min} & p=1 \\ \frac{1}{\sqrt{6}} h_{\min} \approx 0.408\ h_{\min} & p=2 \end{cases}, \quad \text{for central difference method and lumped mass matrix} \tag{404}$$

- However, even in that case there will be a factor of $2 \times 0.408 = 0.816$ smaller time step for $p=2$ mesh compared to $p=1$ mesh with half the element size.
- That is, even if we match the nodal resolution between lower and higher order elements ($p=1$ and $p=2$) higher order elements still may require a smaller time step for stability.
- This concept is shown in the figure below,



- Another common way to express stability limit for different element orders is as,

$$\Delta t_{\max} = \begin{cases} C_H(p) \frac{h_{p\min}}{c} & \text{Hyperbolic PDE, } c = \text{ wave speed} \\ C_P(p) \frac{h_{p\min}^2}{D} & \text{Parabolic PDE, } D = \text{ diffusion coefficient} \end{cases} \quad \text{where} \quad h_{p\min} = \frac{h_{\min}}{p+1} \tag{405}$$

  - $h_{p\min}$ = is an effective element size based on the polynomial order that represent. This size in 1D is the distance between element nodes (if uniformly distributed) and in general represents the length-scale of a "wave", *i.e.,* region with a changed deflection, that an element can model.
  - $C_H(p)$ and $C_P(p)$ are correction factors that depend on,
    - Mass matrix option: *i.e.,* lumped mass, consistent mass, *etc.*.
    - Temporal integration scheme.
    - Underlying numerical method: For example, the same time of estimate can be applied to discontinuous Galerkin methods, *etc.* where for example a "mass matrix" (from item 1 above) may or may not exist, and same with the time integration order (*e.g.,* when spacetime FE methods are used). This can also depend on how many independent fields are interpolated (one-field versus multi-field) and possibly other details of a numerical method.
    - Spatial dimension d
    - Polynomial order p
  - For example, for 1D elastic bar, consistent mass matrix and central difference time integration scheme we have $C_H(1) = 1, C_H(2) = \sqrt{\frac{2}{3}} \approx 0.816$.

– Finally we note that the two estimates in (405) are for a purely hyperbolic equation, *e.g.*, $u_{,tt} - c^2 \nabla.\nabla u = 0$ and parabolic equation $u_{,tt} - D\nabla.\nabla u = 0$. For more general problems, the stability equation takes a more complex form.

Summary

- The stable time step of conditionally stable methods depend on mass matrix option, details of the spatial discretization method, time integration method, and **spatial polynomial order** p.
- Instead of the maximum frequency (eigenvalue) of a MDOF system $\max_l(\omega_l^h)$, conservatively the maximum frequency (eigenvalue) of the individual elements ω_e^m is chosen in evaluating stable time step.
- The definition, $h_{p_{\min}} = \frac{h_{\min}}{p+1}$ and many stability analysis for $p > 1$ are based on having p half a sine wave $0 - \pi$ for an order p element. This is for stability considerations. For accuracy reasons, it is suggested to have **at least 10 elements** for resolving a wave segment, *e.g.*, half a sine wave; [Shakib and Hughes, 1991].

6 Mathematical analysis of finite difference methods

6.1 Introduction: Analysis of FD methods

- The analysis of FD difference methods is similar to that of FE methods in the sense that again we are concerned with concepts of stability, consistency, and convergence.
- Similar to that case we use the easier conditions of stability and consistency to prove convergence.
- This is done by using Lax-Richtmyer theorem.
- The main difference in FD is that we directly solve space and time with FD methods where in FE methods we often discretize the space domain with FE and solve the resulting MDOF ODE, *e.g.,* (225) $\mathbf{M\ddot{U}} + \mathbf{C\dot{U}} + \mathbf{KU} = \mathbf{R}$, using a time marching scheme. In spacetime FE methods, FE discretizes space and time simultaneously eliminating a separate time marching scheme.
- The direct discretization of space and time with FD scheme simplifies the scheme to some extend, yet it does not have the flexibility of FE methods.
- In addition, analysis of FD schemes is more complex as both space and time are involved in the analysis opposed to FE methods where only an MDOF ODE is solved in time. On the other hand, since there are no two steps of FE spatial discretization and a separate time marching scheme one may argue that their analysis is simpler. This argument can be well made specially given the simple structure of FD schemes compared to complex shape functions in FE methods.

6.2 Convergence, consistency, and stability for FE methods

- The idea of convergence is having the FD solution for a given initial boundary value problem tend to the analytical one for any given time, provided that we let the mesh spatial and temporal resolution to zero.
- Again, the proof of convergence for given initial and boundary conditions and PDE is a challenging task as it involves ϵ, δ type limit analysis.
- Instead, as it's common in numerical solution of dynamic problems, we prove consistency and stability and indirectly prove convergence based on these two conditions.
- The following definitions are for one-step FD schemes applied to temporally first order PDEs taken from [Strikwerda, 2004] §1.4 and §1.5
- Formal definition of convergence, for one-step FD scheme applied to first order PDE, is

Definition 1 *A one-step FD scheme approximating a PDE is convergence if for any solution to the PDE $u(x,t)$ and solution to FD scheme v_m^n such that v_m^0 converges to $u_0(x)$ as mh converges to x, then v_m^n converges to $u(x,t)$ as (mh, nk) converges to (x,t) as h, k converge to zero.*

- Basically, definition 1 asserts that a FD scheme is convergent if for any IC, BC, source term, the numerical solution converges to the exact solution at any point if mesh grid sizes h, k approach zero. This idea is shown in the following figure from [Strikwerda, 2004].

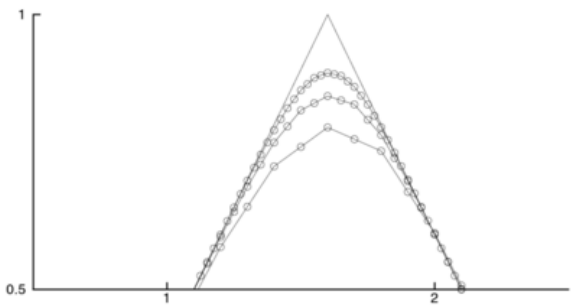


Figure 1.9. *Lax–Friedrichs scheme convergence.*

- As mentioned before, it is easier to prove convergence trough consistency and stability conditions.
- Consistency is a local condition which assert the finite difference operation is consistent with the underlying differential operator.

- Two difference between convergence and consistency are
  1. Convergence refers to the closeness of solutions while consistency refers to the closeness of differential operator occurring in the PDE.
  2. Convergence is a global condition by stipulating that the numerical and exact solutions are close at any point while consistency only requires the differential operator at a point to be close to the PDE differential operator.
- Accordingly, consistency is an algebraic / arithmetic condition to check and while often being tedious is easier to verify than convergence.
- The definition of consistency is as follows [Strikwerda, 2004],

Definition 2 *Given a partial differential equation $Pu = f$ and a FD scheme $P_{h,k} = f$, the FD scheme is consistent with the PDE if for any smooth function $\phi(x,t)$*

$$P\phi - P_{h,k}\phi \to 0 \quad as \quad h,k \to 0$$

the convergence is a point-wise condition at any given point (x,t).

- Note that the positions (x,t) do not need to match with grid points (mh,nk) and can be any arbitrary positions.

Example 1 *Proof of consistency for the Forward-Time Forward-Space (FTFS) scheme (source [Strikwerda, 2004] Example 1.4.1),*

For the one-wave wave equation (26a) $(u_{,t} + a(x,t)u_{,x} = 0)$, the differential operator P is $\frac{\partial}{\partial t} + a\frac{\partial}{\partial x}$ so that,

$$P\phi = \phi_t + a\phi_x$$

for the FTFS scheme (27a) the difference operator $P_{h,k}$ is given by,

$$P_{h,k} = \frac{\phi_m^{n+1} - \phi_m^n}{k} + a\frac{\phi_{m+1}^n - \phi_m^n}{h} \quad \text{where} \quad \phi_m^n = \phi(mh,nk)$$

We begin with the Taylor series of the function ϕ in x and t about $(x_m,t_n) = (mh,nk)$,

$$\phi_m^{n+1} = \phi_m^n + k\phi_{,t} + \frac{1}{2}k^2\phi_{,tt} + \mathcal{O}\left(k^3\right)$$
$$\phi_{m+1}^n = \phi_m^n + h\phi_{,x} + \frac{1}{2}h^2\phi_{,xx} + \mathcal{O}\left(h^3\right)$$

where the derivatives on the RHS are all evaluated at (x_m,t_n), and so,

$$P_{h,k} = \phi_{,t} + a\phi_{,x} + \frac{1}{2}k\phi_{,tt} + \frac{1}{2}ah\phi_{,xx} + \mathcal{O}\left(k^2\right) + \mathcal{O}\left(h^2\right)$$

$$P\phi - P_{h,k} = -\frac{1}{2}k\phi_{,tt} - \frac{1}{2}ah\phi_{,xx} + \mathcal{O}\left(k^2\right) + \mathcal{O}\left(h^2\right) \to \quad \text{as} \quad (h,k) \to 0$$

Therefore, the scheme is consistent.

Example 2 *Conditional consistency of the Lax-Friedrichs scheme (source [Strikwerda, 2004] Example 1.4.2),*

For the Lax-Friedrichs scheme (27d) the FD differential operator is,

$$P_{h,k} = \frac{\phi_m^{n+1} - \frac{1}{2}\left(\phi_{m-1}^n + \phi_{m+1}^n\right)}{k} + a\frac{\phi_{m+1}^n - \phi_{m-1}^n}{2h}$$

We use the Taylor series,

$$\phi_{m\pm1}^n = \phi_m^n \pm h\phi_{,x} + \frac{1}{2}h^2\phi_{,xx} \pm \frac{1}{6}h^3\phi_{,xxx} + \mathcal{O}\left(h^4\right)$$

where, as before, the derivatives are evaluated at (x_m,t_n) and we have,

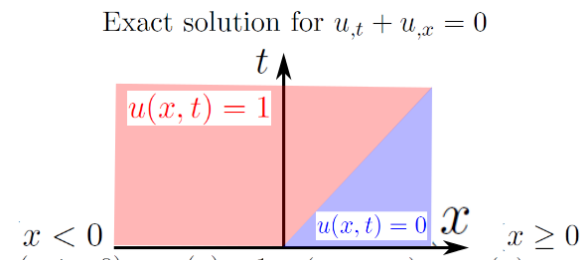
$$\frac{1}{2}\left(\phi_{m-1}^n + \phi_{m+1}^n\right) = \phi_m^n + \frac{1}{2}h^2\phi_{,xx} + \mathcal{O}\left(h^4\right) \quad \text{and}$$
$$\frac{\phi_{m+1}^n - \phi_{m-1}^n}{2h} = \phi_{,x} + \frac{1}{6}h^2\phi_{,xxx} + \mathcal{O}\left(h^4\right)$$

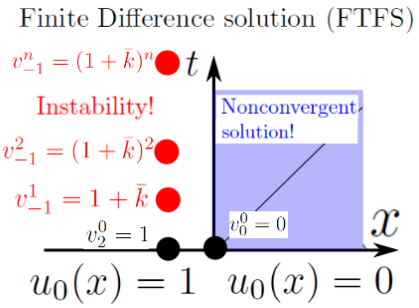
Substituting these expressions in the scheme, we obtain

$$P_{h,k} = \phi_{,t} + a\phi_{,x} + \frac{1}{2}k\phi_{,tt} - \frac{1}{2}k^{-1}h^2\phi_{,xx} + \frac{1}{6}ah^2\phi_{,xxx} + \mathcal{O}\left(h^4 + k^{-1}h^4 + k^2\right)$$

So $P_{h,k} - P\phi \to 0$ as $h,k \to 0$; *i.e.,* it is consistent as long as $k^{-1}h^2$ also tends to zero.

- Note that some schemes such as Lax-Friedrichs schemes are conditionally stable meaning that h, k must satisfy certain condition for the consistency pf the method.
- For the Lax-Friedrichs scheme as it is applied to hyperbolic equations we require $h \propto k$ (for stability) so the consistency condition $k^{-1}h^2 \to 0$ requires $h \to 0$ which is $h \to 0$ which is satisfied. Basically, as long as k does not tend to zero faster than h^2 as $h \to 0$ the Lax-Friedrichs scheme is consistent.
- Many schemes are unconditionally consistent that is as long as $P_{h,k} - P\phi \to$ for $h, k \to 0$ with any independent rate.
- For explicit schemes (and some implicit) schemes the more challenging part can be the stability condition, which for conditionally stable methods we obtain k for which the scheme is stable.
- As a reminder we studied the FD solution of advection equation (26a) ($u_{,t}+a(x,t)u_{,x} = 0$) in §2.1.8 for $a > 0$ and IC $u_0(x) = 1$ if $x \leq 0$ and 0 otherwise.
- The solutions for explicit methods FTBS, FTCS, FTFS were discussed therein.
- In particular, we observe that FTFS scheme was not convergent: For the whole region $x > 0$ FD solution was zero while the exact solution had the solution $u(x,t) = 1$ for $x < t$. This is shown in the figure below,

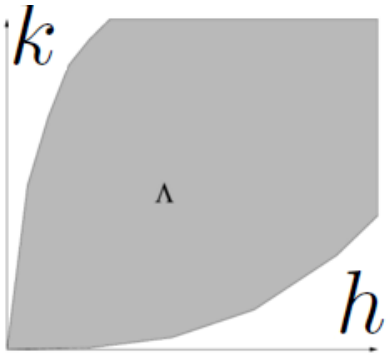




- From example 1 we observe that FTFS scheme is consistent.
- So how come FTFS scheme is consistent (that is the FD differential differential operator $P_{h,k}$ is a good approximation of $P(\phi)$ yet not convergent (FD solution v^n_m does not converge to exact solution $u(x,t) = u(mh, nk)$ for $x > t$ no matter how small h, k are)?
- The answer is that FTFS scheme is not stable for $a > 0$.
- This is shown on the grid points just to the left of $x = 0$ where the solution exponentially blows up with the rate $(1+\bar{k})^n$ (recall $\bar{k} = a\frac{k}{h}$ is the normalized time step from (28)).
- If the scheme was consistent and at the same time stable then from Lax-Richtmyer theorem it would have been convergent which is was a numerical analysis seeks, that is convergence of solution to the exact one as $h, k \to 0$.
- In the following we provide the definition of stability which will be required from FD schemes.

Definition 3 *Region of stability is any bounded nonempty region of the first quadrant of $\mathbb{R}^3$ that has origin as an accumulation point. That is, stability region must contain a sequence (h_ν, k_ν) that converges to the origin as $\nu \to \infty$.*

A common example is a region of the form $\{(h,k)|0 < k \leq ch \leq C\}$ for some positive constants c and C (c can represent a multiple of wave speed for hyperbolic problems for example). An example of stability region is shown below.



Definition 4 *Stability of temporally first order PDEs: A finite difference scheme $P_{h,k}v^n_m = 0$ for a temporally first-order PDE is stable in the stability region Λ if there an integer J such that for any positive time T, there is a constant C_T such that,*

$$h||v^n||^2_h \leq C_T h \sum_{j=0}^{J} ||v^j||^2_h \qquad \text{for } 0 \leq nk \leq T \quad \text{with} \quad (h,k) \in \Lambda. \tag{406}$$

There are a few points to clarity,

- The notation ||.|| refers to L2 norm of the FD solution,

$$||v^n||_h = \sqrt{h \sum_{m=-\infty}^{\infty} |v_m^n|^2} \tag{407}$$

  which corresponds to the discrete counterpart of L2 norm definition for functions on $(-\infty, \infty)$,

$$||u^t|| = \sqrt{\int_{\infty}^{\infty} |u(x,t)|^2 \, \mathrm{d}x} \tag{408}$$

- (419) is equivalent to

$$||v^n||_h \leq C_T^* \sum_{j=0}^{J} ||v^j||_h \tag{409}$$

  that is working with L2 norm quantities rather than their square roots. However, it is often easier to use the form with the squares of square roots as it will be shown in (6.3).

- The number J refers to the number of steps required in a multi-step method. For example for a 1-step method that only requires t_n for updating t_{n+1} J will be 0, that is only initial data will be used in eq:FD:Stability:FirstOrder.

$$||v^n||_h^2 \leq C_T ||v^0||_h^2, \qquad (J = 0) \text{ in (419) for single-step methods} \tag{410}$$

- Comments on the value C_T
  - The most important aspect is that C_T **only depends on T not k nor h**: This means that no matter what grid size is used the solution at time T does not blow-up by for example letting $k \to 0$.
  - For unstable FD methods by letting $k \to 0$ the FD grid can represent higher frequency content (as will be discussed in §6.3) and the limit C_T will grow as $k \to 0$. That is, there is no constant C_T only dependent on T for unstable methods.
  - Note that C_T can be larger than one and in fact norm $||v^n||_h \to \infty$ as $n \to \infty$. That is, the solution can tend to infinity. This type of stability limit ($C_T > 1$) can arise if the spatial norm of the underlying exact physical solution also tend to infinity.
  - If the solution of the underlying solution is in fact bounded or decaying (as in many physical problems called dynamically stable; *cf.* §6.5) the spatial norm of physical solution does not grow and the FD scheme may have a $C_T \leq 1$.
  - The stability condition of a numerical method is closely related to the concept of well-posedness or dynamic stability of a physical system which will be discussed in §6.5.
  - Stability is rarely directly checked. As will be discussed in §6.3 stability of a FD scheme is often investigated in the frequency domain. The example shows how stability can be checked directly.

Example 3 *Direct proof of stability of*

$$v_m^{n+1} = \alpha v_m^n + \beta v_{m+1}^n \tag{411}$$

is stable if $|\alpha| + |\beta| \leq 1$. *(source [Strikwerda, 2004] Example 1.5.1)*

This type of update for example was observed in FTBS scheme applied to advection equation (26a) $u_{,t} + a(x,t)u_{,x} = 0$ for constant $a(x,t) = a$ in (27b): $\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = 0 \Rightarrow v_m^{n+1} = (1 - \bar{k})v_m^n + \bar{k}v_{m-1}^n$ (*cf.* (35b)) with $\bar{k} = a\frac{k}{h}$ being the normalized time step. Thus, for FTBS scheme $\alpha = 1 - \bar{k}$ and $\beta = \bar{k}$. The analysis is as follows,

$$\begin{aligned}
\sum_{m=-\infty}^{\infty} |v_m^{n+1}|^2 &= \sum_{m=-\infty}^{\infty} |\alpha v_m^n + \beta v_{m+1}^n|^2 \\
&\leq \sum_{m=-\infty}^{\infty} |\alpha|^2 |v_m^n|^2 + 2|\alpha||\beta||v_m^n||v_{m+1}^n| + |\beta|^2 |v_{m+1}^n|^2 \\
&\leq \sum_{m=-\infty}^{\infty} |\alpha|^2 |v_m^n|^2 + |\alpha||\beta| \left(|v_m^n|^2 + |v_{m+1}^n|^2\right) + |\beta|^2 |v_{m+1}^n|^2
\end{aligned}$$

where we have used the inequality $2xy \leq x^2 + y^2$. The sum can be split over the terms with index m and those with index $m+1$. Subsequently, the index of the terms with index $m+1$ will be shifted one to the left without changing the summation, as the summation is from $m = -\infty$ to ∞:

$$\begin{aligned}
&\sum_{m=-\infty}^{\infty} |\alpha|^2|v_m^n|^2 + |\alpha||\beta|\left(|v_m^n|^2 + |v_{m+1}^n|^2\right) + |\beta|^2|v_{m+1}^n|^2 \\
&= \sum_{m=-\infty}^{\infty} |\alpha|^2|v_m^n|^2 + |\alpha||\beta||v_m^n|^2 + \sum_{m=-\infty}^{\infty} |\beta|^2|v_{m+1}^n|^2 + |\alpha||\beta||v_{m+1}^n|^2 \\
&= \sum_{m=-\infty}^{\infty} |\alpha|^2|v_m^n|^2 + |\alpha||\beta||v_m^n|^2 + \sum_{m=-\infty}^{\infty} |\beta|^2|v_m^n|^2 + |\alpha||\beta||v_m^n|^2 \\
&= \sum_{m=-\infty}^{\infty} \left(|\alpha|^2 + 2|\alpha||\beta| + |\beta|^2\right)|v_m^n|^2 \\
&= (|\alpha| + |\beta|)^2 \sum_{m=-\infty}^{\infty} |v_m^n|^2
\end{aligned}$$

The last two equations show that,

$$\sum_{m=-\infty}^{\infty} |v_m^{n+1}|^2 \leq (|\alpha| + |\beta|)^2 \sum_{m=-\infty}^{\infty} |v_m^n|^2$$

and since this applies to all n we have,

$$h \sum_{m=-\infty}^{\infty} |v_m^{n+1}|^2 \leq (|\alpha| + |\beta|)^{2n}\, h \sum_{m=-\infty}^{\infty} |v_m^0|^2 \qquad \Rightarrow \qquad ||v^n||_h^2 \leq (|\alpha| + |\beta|)^{2n}\, ||v^0||_h^2$$

based on the definition of discrete norm in (407).

Now if $|\alpha| + |\beta| \leq 1 \Rightarrow (|\alpha| + |\beta|)^{2n} \leq 1$ so $C_T = 1$ would work in the stability condition in (410). That is,

$$||v^n||_h^2 \leq ||v^0||_h^2, \qquad \text{if } |\alpha| + |\beta| \leq 1$$

- In fact, we observe that the norm of subsequent time steps is smaller than equal than that of the initial condition. That is, in this case the FD solution does not even grow.
- The converse that if $|\alpha| + |\beta| > 1$ we cannot find any C_T for which (410) ($||v^n||_h^2 \leq C_T||v^0||_h^2$) is a bit more difficult, but the Fourier analysis from §6.3 demonstrate that the converse is also true.
- Now, going back to FTBS scheme we had, $\alpha = 1 - \bar{k}$ and $\beta = \bar{k}$. Clearly, $|\alpha| + |\beta| = |1 - \bar{k}| + |\bar{k}| \leq 1$ can only hold true for $\bar{k} \leq 1$. That is, FTBS scheme is only stable when $\bar{k} = a\frac{k}{h} \leq 1$ or $k \leq \frac{h}{a}$.
- Also, note that if $a < 0$ then $|\alpha| + |\beta| = |1 - \bar{k}| + |\bar{k}| > 1$ for any k and FTBS scheme will be unconditionally unstable!
- The opposite also holds true for FTFS scheme where it is stable when $k \leq \frac{h}{|a|}$ when wave is left going, *i.e.,* $a < 0$ and is unconditionally unstable for $a > 0$.
- We demonstrated these concepts through a particular example with IC $u_0(x) = 1$ for $x < 0$ and zero otherwise in (2.1.8).
- While that example was demonstrative, it was not a proof of stability conditions of any of the methods discussed. Herein, we proved for a particular update equation of the form (411) ($v_m^{n+1} = \alpha v_m^n + \beta v_{m+1}^n$) the FD scheme is stable if $|\alpha| + |\beta| \leq 1$ (we defer the converse proof to §6.3).
- We recall the way consistency, stability, and convergence are generally related to each other from §5.1. In particular (288) and (289) stated that in general,

  Consistency $\Rightarrow$ (Stability $\Leftrightarrow$ Convergence) (Equation (288))

  and

  Consistency and Stability $\Rightarrow$ Convergence (Equation (289))
- As mentioned before, the final goal is to have a numerical method that is convergent to the exact solution.
- Convergence can be more easily proved by proving both consistency and stability (which are easier properties to prove in general as discussed in §5.1) and using a variant of Lax-Ritchmyer for the underlying numerical method conclude that it is convergent.

- An example of this process for first order temporal ODE was demonstrated in §5.2 for generalized trapezoidal rule.
- The Lax-Richtmyer reads as follows,

Theorem 1 *The Lax-Richtmyer equivalence theorem: A consistent FD scheme for a PDE for which the initial value problem is well-posed is convergent if any only if it is stable.*

- This is basically the type of relation between consistency, stability, and convergent stated in (288). In practice, the form (289) is used to prove convergence.
- Note that the condition of well-posedness for the underlying PDE is further discussed in (6.5). It basically limits how much the differences in the initial condition of a PDE can grow in time. Most problems with physical grounds are well-posed.
- For the proof of Lax-Richtmyer can be found in [Strikwerda, 2004] Chapter 10.

6.3 Analysis in frequency domain

6.3.1 Fourier transformation and Fourier series

- We recall the Fourier transform from (203),

$$\check{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-\imath\omega t} \, \mathrm{d}t \qquad \Leftrightarrow \tag{412a}$$

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \check{f}(\omega) e^{\imath\omega t} \mathrm{d}\omega \tag{412b}$$

- Equation (412b) has a profound meaning in that we can write a function $f(t)$ as a "summation" of its frequency modes with frequency ω and amplitude $\check{f}(\omega)$.
- That is a function $f(t)$ is expressed as a superposition of harmonic waves with different frequencies and different amplitudes.
- A very important identity in Fourier analysis is the Parseval's relation,

$$\int_{-\infty}^{\infty} |f(t)|^2 \, \mathrm{d}t = \int_{-\infty}^{\infty} |\check{f}(\omega)|^2 \mathrm{d}\omega \qquad \Rightarrow \boxed{||f||_2 = ||\check{f}||_2} \tag{413}$$

  the subscript 2 refers to L2 norm (408) $||u||_2 = \sqrt{\int_{\infty}^{\infty} |u(t)|^2 \, \mathrm{d}t}$ which for convenience we drop the subscript 2 when there is no confusion in the type of norm employed.
- The Parseval's relation is of utmost importance as it says the norm of a function is equal to the norm of its Fourier transform.
- In stability analysis it is often easier to establish the stability of simple harmonic solutions. Subsequently, we use the Parseval's relation to establish stability for any form of solution by basically decomposing it into its harmonic parts.
- In the stability analysis of FD methods we are interested in how the spatial (discrete) norm of the solution grows. See for example, (407) where $||v^n||_h = \sqrt{h \sum_{m=-\infty}^{\infty} |v_m^n|^2}$ and the stability condition (419) $||v^n||_h^2 \leq C_T \sum_{j=0}^{J} ||v^j||_h^2$.
- Accordingly, it is reasonable to apply the Fourier transform to x rather than t.
- By just writing the Fourier transform in x variable rather than t for a function u we express (412) as,

$$\hat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x) e^{-\imath\xi x} \, \mathrm{d}x \qquad \Leftrightarrow \tag{414a}$$

$$u(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{u}(\xi) e^{\imath\xi x} \mathrm{d}\xi \tag{414b}$$

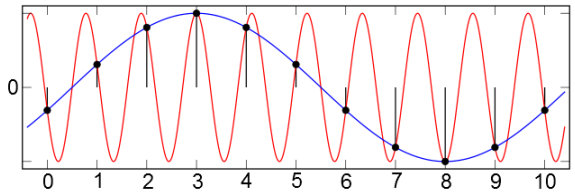
  the parameter ξ is spatial frequency which is also called wavenumber. The symbols k and ξ are often used for it. However, given that k is used for the time step herein, we use the latter notation for the wavenumber.
- Still, we cannot use the Fourier series analysis and Parseval's relation for the stability analysis of FD method as the solution in FD schemes is only provided at discrete points h apart.

- Instead, we need to use the discrete version of the Fourier transform, which is known as Fourier series,

$$\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} h \sum_{m=-\infty}^{\infty} e^{-\imath mh\xi} v_m, \quad \text{for} \quad \xi \in [-\frac{\pi}{h}, \frac{\pi}{h}] \quad \Leftrightarrow \tag{415a}$$

$$v_m = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{\imath mh\xi} \hat{v}(\xi) \mathrm{d}\xi \tag{415b}$$

- Several interesting points can be observed about the Fourier series,
  - When covering the concept of Fourier transform, we often deal with a function v_m that spans from $-\pi/h$ to π/h (*i.e.*, with period h centered at 0). The goal is to express the function $\hat{v}(\xi)$ as a sum of Fourier series harmonic terms in (415a): $\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} h \sum_{m=-\infty}^{\infty} e^{-\imath mh\xi} v_m$ for which the coefficients v_m are given from (415b) $v_m = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{-\imath mh\xi} \hat{v}(\xi) \mathrm{d}\xi$.
  - Herein, for the stability of FD methods, it is more natural to take a different perspective:
    * For a given FD solution at a time step the values at spatial grid points are given by v_m^n where for brevity here we drop n.
    * We note that the FD solution having these values at grid points corresponds to the summation of infinite simple harmonic waves with amplitudes $\hat{v}(\xi)$ and frequencies ξ (the factors $1/\sqrt{2\pi}$ should also be discussed in the amplitudes of the waves and v_m^n but that does not provide much insight to the physical breakdown of FD solution into infinite simple harmonic waves).
    * In this viewpoint, the frequency amplitudes are obtained from (415a) $\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} h \sum_{m=-\infty}^{\infty} e^{-\imath mh\xi} v_m$.
    * That is, v_m given, we can find frequency amplitudes $\hat{v}(\xi)$ from (415a) and re-express the values of v at grid points m by its superposition in terms of simple harmonic waves: $v_m = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{-\imath mh\xi} \hat{v}(\xi) \mathrm{d}\xi$.
    * It is this latter harmonic solution we will use in the stability analysis of linear FD schemes in what is known von Neumann analysis in §6.3.5. von Neumann analysis depends on the linearity of the PDE and the use of a Parseval's relation for the Fourier series problem in (415).
    * Obviously, one very important aspect in (415a) is the restriction of spatial frequencies (wavenumbers) ξ to $[-\frac{\pi}{h}, \frac{\pi}{h}]$. This restriction is clear if we take a Fourier series expansion of a function $\hat{v}(\xi)$ in (415b) which by assumption spans $[-\frac{\pi}{h}, \frac{\pi}{h}]$ in (415b). In introductory Fourier analysis books, as mentioned, we start from (415b) for a given h and prove v_m, *i.e.*, Fourier series coefficients, are given by (415a).
    * However, what is the physical reasoning of breaking down a discrete solution of spacing h into only wavenumbers in the range $[-\frac{\pi}{h}, \frac{\pi}{h}]$ not $[-\infty, \infty]$ as done in Fourier transformation case in (414b) ($u(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{u}(\xi) e^{\imath\xi x} \mathrm{d}\xi$) ?
    * There is a very simple explanation of this fact:
    * A grid of spacing h cannot distinguish any wavenumbers higher than π/h due to **aliasing effect**. That is for any higher order wavenumber there is a wavenumber in the interval $[-\frac{\pi}{h}, \frac{\pi}{h}]$ that can exactly capture the same values at grid points v_m. This is shown in the figure below,



Source: Wikipedia

It is easy to see why aliasing occurs,

$$e^{\imath mh(\xi + q(2\pi/h))} = e^{\imath mh\xi} e^{\imath 2(qm)\pi(h/h)} = e^{\imath mh\xi} e^{\imath 2(qm)} = e^{\imath mh\xi}$$

$e^{\imath mh(\xi+q(2\pi/h))}$ would represent the red line in the figure for $\xi_2 = \xi + q(2\pi/h)$ and the blue line the base frequency ξ. That is, a base frequency $\xi \in [-\frac{\pi}{h}, \frac{\pi}{h}]$ has the same solution value at all grid points with spacing h for any other larger frequency outside of $[-\frac{\pi}{h}, \frac{\pi}{h}]$ in the form $\xi_2 = \xi + q(2\pi/h)$ ($q \in \mathbb{Z}$) and this is the physical reasoning of sufficiency of only having frequencies in the range $[-\frac{\pi}{h}, \frac{\pi}{h}]$ in harmonic decomposition of the form (415b) of solution with grid values v_m.

In (6.3.1) we used $e^{\imath 2(qm)} = \cos(2(qm)) + \imath \sin(2(qm)) = 1$.

- The material in this section is restricted to 1D analysis, but without much difficulty all the analysis can be extended to 2D and 3D spatial domains, a topic not considered herein.

- The Fourier transform in higher dimensions ($d = 2$ in 2D and 3 in 3D) is,

$$\hat{u}(\vec{\xi}) = \frac{1}{2\pi^{d/2}} \int_{\mathbb{R}^d} u(\vec{\mathbf{x}}) e^{-\imath \vec{\xi}.\vec{\mathbf{x}}} \mathrm{d}\vec{\mathbf{x}} \qquad \Leftrightarrow \tag{416a}$$

$$u(\vec{\mathbf{x}}) = \frac{1}{2\pi^{d/2}} \int_{\mathbb{R}^d} \hat{u}(\vec{\xi}) e^{\imath \vec{\xi}.\vec{\mathbf{x}}} \mathrm{d}\vec{\xi} \tag{416b}$$

  where $\vec{\mathbf{x}}$ and $\vec{\xi}$ are vectors in $\mathbb{R}^d$ and . is the inner-product operator.

- Fourier series (415) can also be easily extended to 2D and 3D for grids with even different spacings h_1, h_2, h_3 which again we do not pursue them given the similarity of the analysis of 2D and 3D FD problems to 1D ones.

6.3.2 Parseval's relation for Fourier series

- Recall the Parseval's relation for functions (413) which stated that $||f||_2 = ||\hat{f}||_2$ where $||f||_2 = \sqrt{\int_{-\infty}^{\infty} |f(t)|^2 \, \mathrm{d}t}$ and $||\hat{f}||_2 = \sqrt{\int_{-\infty}^{\infty} |\hat{f}(\omega)|^2 \mathrm{d}\omega}$ are L2 norms of a function $f(t)$ and its Fourier transformation $\hat{f}(\omega)$.
- We seek a Parseval type inequality that can relate the discrete L2 norm of FD solution with spacing h spacing at time step t_n, *i.e.,* (407) where $||v^n||_h = \sqrt{h \sum_{m=-\infty}^{\infty} |v_m^n|^2}$ with the L2 norm of its frequency function $\hat{v}^n$.
- This would greatly simplify the stability analysis of FD schemes, *e.g.,* equations of the form (419) as we will see shortly.
- The Parseval's relation for Fourier series reads as,

$$h \sum_{m=-\infty}^{\infty} |v_m|^2 = \int_{-\pi/h}^{\pi/h} |\hat{v}(\xi)|^2 \mathrm{d}\xi \Rightarrow \boxed{||v||_h = ||\hat{v}||_h} \tag{417}$$

  where $||v||_h$ is the discrete L2 norm of v solution given at grid spacing h while $||\hat{v}||_h$ is the L2 norm of the function $\hat{v}(\xi)$ over the interval $[-\pi/h, \pi/h]$.

- Before we proceed with the uses of Parseval's relation above, we provide an informal proof of it,

$$\begin{aligned}
\|\hat{v}\|_h^2 &= \int_{-\pi/h}^{\pi/h} |\hat{v}(\xi)|^2 \, d\xi = \int_{-\pi/h}^{\pi/h} \overline{\hat{v}(\xi)} \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} e^{-imh\xi} v_m \, h \, d\xi \\
&= \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} \int_{-\pi/h}^{\pi/h} e^{-imh\xi} \overline{\hat{v}(\xi)} \, d\xi \; v_m \, h \\
&= \sum_{m=-\infty}^{\infty} \overline{\frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \hat{v}(\xi) \, d\xi} \; v_m \, h \\
&= \sum_{m=-\infty}^{\infty} \overline{v_m} v_m \, h = \|v\|_h^2.
\end{aligned} \tag{418}$$

  where $\bar{\cdot}$ refers to complex conjugate of a number ($\overline{a + b\imath} = a - b\imath$ for $a, b \in \mathbb{R}$).

- Now assume we want to prove the stability of a FD scheme from (419). That is finding C_T and J such that $||v^n||_h^2 \leq C_T \sum_{j=0}^{J} ||v^j||_h^2$ for $0 \leq nk \leq T$ with $(h, k) \in \Lambda$. However, given that $||v||_h = ||\hat{v}||_h$ from (417) we can alternatively show that,

$$||\hat{v}^n||^2 \leq C_T \sum_{j=0}^{J} ||\hat{v}^j||^2 \qquad \text{for } 0 \leq nk \leq T \;\; \text{with} \;\; (h, k) \in \Lambda. \tag{419}$$

- That is stability is analyzed in the frequency domain using $\hat{v}$ rather than v.
- The use of Parseval's relation and the stability in the form (419) becomes apparent in the next section.

6.3.3 Analysis in frequency domain: Amplification factor

- Consider FTBS scheme (27b),

$$\frac{v_m^{n+1}-v_m^n}{k}+a\frac{v_m^n-v_{m-1}^n}{h}=0$$

- which can be rewritten as (35b),

$$v_m^{n+1}=(1-\bar{k})v_m^n+\bar{k}v_{m-1}^n \quad \text{for normalized time step} \quad \bar{k}=a\frac{k}{h} \tag{420}$$

- By taking the Fourier series transform on both sides of (420) and recalling (415b) $v_m=\frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h}e^{\imath mh\xi}\hat{v}(\xi)\mathrm{d}\xi$ we obtain,

$$\begin{aligned}
v_m^{n+1}&=(1-\bar{k})v_m^n+\bar{k}v_{m-1}^n\\
&=(1-\bar{k})\left\{\frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h}e^{\imath mh\xi}\hat{v}^n(\xi)\mathrm{d}\xi\right\}+\bar{k}\left\{\frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h}e^{\imath(m-1)h\xi}\hat{v}^n(\xi)\mathrm{d}\xi\right\} \qquad \Rightarrow\\
v_m^{n+1}&=\frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h}e^{\imath mh\xi}\left[(1-\bar{k})+\bar{k}e^{-\imath h\xi}\right]\hat{v}^n(\xi)\mathrm{d}\xi
\end{aligned} \tag{421}$$

  note that the dependence to time step t_n for values v_m^n is shown as superscript for the grid point values and Fourier functions $\hat{v}^n$.

- On the other hand, again from the definition of Fourier transform we have,

$$v_m^{n+1}=\frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h}e^{\imath mh\xi}\hat{v}^{n+1}(\xi)\mathrm{d}\xi \tag{422}$$

- By comparing (421) and (422) we obtain,

$$\hat{v}^{n+1}(\xi)=g(h\xi)\hat{v}^n(\xi) \qquad \text{where} \tag{423a}$$

$$g(h\xi):=\left[(1-\bar{k})+\bar{k}e^{-\imath h\xi}\right] \qquad \text{amplification factor (for FTBS method)} \tag{423b}$$

- (423) shows that for this one-step scheme the update from time step t_n to t_{n+1} is equivalent to multiplying the Fourier transform of the solution by the **amplification factor** $g(h\xi)$.
- The terminology amplification factor is because Fourier space solution $\hat{v}^n(\xi)$ is multiply by it to reach the solution at the next time step $\hat{v}^{n+1}(\xi)$.
- Now by induction and use of (423) we obtain,

$$\hat{v}^n(\xi)=g^n(h\xi)\hat{v}^0(\xi) \tag{424}$$

- For the moment assume, we have some conditions on the time step k such that $|g(h\xi)|\le 1$.
- We will discuss what condition must hold such that $|g(h\xi)|\le 1$.
- Assuming that this condition is satisfied, we will have,

$$\begin{aligned}
||\hat{v}^n||_h^2&=\int_{-\pi/h}^{\pi/h}|\hat{v}^n(\xi)|^2\mathrm{d}\xi=\int_{-\pi/h}^{\pi/h}|g^{2n}(h\xi)\hat{v}^0(\xi)|^2\mathrm{d}\xi=\int_{-\pi/h}^{\pi/h}|g(h\xi)|^{2n}|\hat{v}^0(\xi)|^2\mathrm{d}\xi\\
&\le\int_{-\pi/h}^{\pi/h}|\hat{v}^0(\xi)|^2\mathrm{d}\xi=||\hat{v}^0||_h^2 \qquad (\text{if} \quad |g(h\xi)|\le 1)
\end{aligned}$$

- Thus, for conditions (will be derived below) for which $|g(h\xi)|\le 1$ we have,

$$||\hat{v}^n||_h^2\le||\hat{v}^0||_h^2 \qquad (\text{if} \quad |g(h\xi)|\le 1) \tag{425}$$

- Now here is where the Parseval's relation for Fourier series (417) $||v||_h = ||\hat{v}||_h$ is used:

$$\left.\begin{aligned} ||\hat{v}^n||_h^2 &= ||v^n||_h^2 \quad \text{Parseval's relation (417)} \\ ||\hat{v}^0||_h^2 &= ||v^0||_h^2 \quad \text{Parseval's relation (417)} \\ ||\hat{v}^n||_h^2 &\le ||\hat{v}^0||_h^2 \quad \text{from (425) (if } \ |g(h\xi)| \le 1) \end{aligned}\right\} \Rightarrow \quad \boxed{||v^n||_h^2 \le ||v^0||_h^2} \tag{426}$$

  for h, k for which $|g(h\xi)| \le 1$ for all $\xi \in [-\pi/h, \pi/h]$.

- Thus, the stability analysis of this FD scheme reduces in finding conditions where the absolute value of amplification factor $|g(h\xi)|$ is bounded by 1.

- Now, we focus on finding the conditions on h, k where $|g(h\xi)| \le 1$ for all $\xi \in [-\pi/h, \pi/h]$.

- We introduce the following notation,

$$\boxed{\theta := h\xi} \tag{427}$$

  later we further comment on the meaning of the parameter θ. Basically, its range is $\theta \in [-\pi, \pi]$ for which $\theta \to 0$ corresponds to a very refined mesh relative to wavenumber ξ and $\theta = \pm\pi$ corresponds to the coarsest grid, *i.e.,* the highest number of waves that the discrete FD grid can represent.

- With this notation (427) $g(h\xi) = g(\theta)$ in (423b) becomes,

$$\begin{aligned} &g(\theta) := (1 - \bar{k}) + \bar{k}e^{-\imath\theta} \qquad \Rightarrow \\ &g(\theta) = g_R + \imath g_I, \qquad \text{where} \qquad g_R = (1 - \bar{k}) + \bar{k}\cos\theta, \qquad g_I = -\bar{k}\sin\theta \end{aligned} \tag{428a}$$

- In (428a) g is written in terms of its real g_R and imaginary g_I components.

- Now, since we are seeking conditions where $|g(\theta)| \le 1$ we use its square value,

$$|g(\theta)| = \sqrt{g_R^2 + g_I^2} \le 1 \qquad \Leftrightarrow \qquad |g(\theta)|^2 = g_R^2 + g_I^2 \le 1 \tag{429a}$$

- Now using the identities,

$$1 - \cos\phi = 2\sin^2\frac{1}{2}\phi, \qquad \sin\phi = 2\sin\frac{1}{2}\phi\cos\frac{1}{2}\phi$$

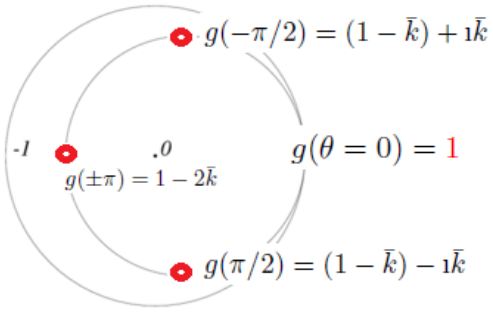
  and letting $\phi = \theta$ we obtain,

$$\begin{aligned} |g(\theta)|^2 = g_R^2 + g_I^2 &= (1 - \bar{k} + \bar{k}\cos\theta)^2 + \bar{k}^2\sin^2\theta \\ &= (1 - 2\bar{k}\sin^2\frac{1}{2}\theta)^2 + 4\bar{k}^2\sin^2\frac{1}{2}\theta\cos^2\frac{1}{2}\theta \\ &= 1 - 4\bar{k}\sin^2\frac{1}{2}\theta + 4\bar{k}^2\sin^4\frac{1}{2}\theta + 4\bar{k}^2\sin^2\frac{1}{2}\theta\cos^2\frac{1}{2}\theta \qquad \Rightarrow \\ &\boxed{|g(\theta)|^2 = 1 - 4\bar{k}(1 - \bar{k})\sin^2\frac{1}{2}\theta} \end{aligned} \tag{430}$$

- From (430) we observe that $|g(\theta)|^2 \le 1$ for all θ iff $0 \le \bar{k} \le 1$.

- The figure on the right shows the loci of $|g(\theta)|^2$ in the complex plane as θ changes from 0 to $\pm\pi$. The points corresponding to $\theta = 0, \pm\pi/2, \pm\pi$ are given in the equation below and marked in the figure,

$$g(\theta = 0) = 1 \tag{431a}$$
$$g(\pi/2) = (1 - \bar{k}) - \imath\bar{k} \tag{431b}$$
$$g(-\pi/2) = (1 - \bar{k}) + \imath\bar{k} \tag{431c}$$
$$g(\pm\pi) = 1 - 2\bar{k} \tag{431d}$$



The image of $g(\theta)$ for the forward-time backward-space scheme.

- When $\bar{k} > 1$ the loci of $g(\theta)$ of goes beyond the unit circle in the complex plane around $\theta = \pm\pi$ and around those θ. This is for this problem the instabilities initiate from the highest frequency modes $\bar{k}$. This corresponds to instable values for k.

- The fact that $g(\theta) = 1$ for $\theta = 0$ is always given if the FD scheme is consistent, as otherwise in the limit of $h, k \to 0$ (relative to wavenumber ξ) the FD scheme is not consistent with exact PDE.

- Recalling $\bar{k} = a\frac{k}{h}$ from (28) FTBS scheme is conditionally stable for the range $k \leq \frac{h}{a}$. This clearly matches our numerical example from §2.1.8 and example 3 earlier in this section.
- Clearly, in neither of these approaches / examples we have proven that the scheme is unstable for $k > \frac{h}{a}$.
- The next theorem demonstrates that $|g(\theta)|^2 > 1$ for some θ for this case where g is independent of k corresponds to instability.

6.3.4 Theorem on stability of FD methods

Theorem 2 *Stability analysis in frequency domain for one-step FD schemes: A one-step FD scheme (with constant coefficients) is stable in a stability region Λ if and only if there is a constant K (independent of θ, k and h) such that,*

$$|g(\theta, k, h)| \leq 1 + Kk \tag{432}$$

with $(h, k) \in \Lambda$. If $g(\theta, k, h)$ is independent of h and k, the stability condition (432) can be replaced with the restricted stability condition,

$$|g(\theta, k, h)| \leq 1 \tag{433}$$

Proof: We have the Parseval's relation (417) and the definition of g, that

$$||v^n||_h^2 = ||\hat{v}^n||_h^2 = \int_{-\pi/h}^{\pi/h} |\hat{v}^n(\xi)|^2 \mathrm{d}\xi \tag{434a}$$

$$= \int_{-\pi/h}^{\pi/h} |g^n(h\xi, k, h)\hat{v}^0(\xi)|^2 \mathrm{d}\xi = \int_{-\pi/h}^{\pi/h} |g^{2n}(h\xi, k, h)||\hat{v}^0(\xi)|^2 \mathrm{d}\xi \qquad \text{from (424)} \tag{434b}$$

Now from (432) ($|g(\theta, k, h)| \leq 1 + Kk$) we have,

$$||\hat{v}^n||_h^2 \leq \int_{-\pi/h}^{\pi/h} (1 + Kk)^{2n} |\hat{v}^0(\xi)|^2 \mathrm{d}\xi = (1 + Kk)^{2n} ||\hat{v}^0||_h^2 \tag{435}$$

now for $t = nk \leq T$ we have $n \leq T/k$. Thus from $(1 + \alpha)^\beta \leq e^{\alpha\beta}$ for $\alpha, \beta \geq 0$ we have,

$$(1 + Kk)^{2n} \leq (1 + Kk)^{2T/k} \leq e^{2KT} \tag{436}$$

and from (435) and (436) we obtain,

$$||\hat{v}^n||_h^2 \leq e^{2KT} ||\hat{v}^0||_h^2 \quad \Rightarrow \quad \boxed{||v^n||_h^2 \leq e^{2KT} ||v^0||_h^2} \quad \text{(using Parseval's relation (417))} \tag{437}$$

that is the stability condition (410) ($||v^n||_h^2 \leq C_T ||v^0||_h^2$) is satisfied for $C_T = e^{KT}$ and $J = 0$. That is the FD scheme is stable when (433) is satisfied.

The converse is more difficult to prove. For the converse, we prove that if (432) is not satisfied for $(h, k) \in \Lambda$ for any value of K, then the scheme is not stable in Λ. To do this we show that we can achieve any amount of growth in the solution, *i.e.*,, we show that the stability inequality for one step methods (410) ($||v^n||_h^2 \leq C||v^0||_h^2$) cannot be satisfied for any C.

If for some positive value C there is an interval of θ's ($\theta = h\xi$) $\theta \in [\theta_1 \ \theta_2]$ and $(h, k) \in \Lambda$ with $g(\theta, k, h) > 1 + Ck$, then we construct a function v_m^0 as,

$$\hat{v}^0(\xi) = \begin{cases} 0 & \text{if } \theta = h\xi \notin [\theta_1, \theta_2] \\ \sqrt{h(\theta_2 - \theta_1)^{-1}} & \text{if } \theta = h\xi \in [\theta_1, \theta_2] \end{cases} \tag{438}$$

Notice that $||\hat{v}^0||_h$ is equal to 1. Then,

$$\begin{aligned} ||v^n||_h^2 &= \int_{-\pi/h}^{\pi/h} |g(h\xi, k, h)|^{2n} |\hat{v}^0(\xi)|^2 \mathrm{d}\xi \\ &= \int_{\theta_1/h}^{\theta_2/h} |g^{2n}(h\xi, k, h)| \frac{h}{\theta_2 - \theta_1} \mathrm{d}\xi \\ &\geq (1 + Ck)^{2n} \\ &\geq e^{2TC} ||v^0||_h^2 \end{aligned}$$

for n near T/k. This shows the scheme to be unstable if C can be arbitrarily large. This, the scheme is unstable if there is no region in which $g(\theta, k, h)$ can be bounded in (432) ($|g(\theta, k, h)| \leq 1 + Kk$).

The proof for the case that g is explicitly independent of k and depends on a function of h, k (*i.e.*, $\bar{k} = ak/h$ for a simple hyperbolic problem $u_{,t} + au_{,x} = 0$ or $\bar{k} = Dk/h^2$ for $u_{,t} - Du_{,xx} = 0$) the proof follows a similar line. First, if (433) is satisfied, the proof of stability is the same as (433) but with $K = 0$. In this case, we will have the condition $||\hat{v}^n||_h^2 \leq ||\hat{v}^0||_h^2$ which beside the stability of the method it means that the norm of the numerical solution is bounded (by $||\hat{v}^0||_h^2$).

The converse, in this case again is the same as the process taken on and after (438) for explicitly k-dependent g. If (433) is violated, for some constant $\alpha > 0$, there is an interval$\theta \in [\theta_1, \theta_2]$ and $\bar{k}$ ranges such that $|g(\theta, \bar{k})| > 1 + \alpha$, we construct IC in the form (438) and following the same process we observe that $||v^n||_h^2 \geq (1+\alpha)^{2n} \geq \frac{1}{2}(1+\alpha)^{2T/k}||v^0||_h^2$ (again $||v^0||_h^2 = 1$) for $t_n = nk$ sufficiently close to T. Clearly, by letting $k \to 0$ ($n \to \infty$) for $\bar{k}$ values for which $|g(\theta, \bar{k})| > 1 + \alpha$ $||v^n||_h^2$ cannot be bounded by $||v^0||_h^2$ for $t_n = nk$ close to a chosen T. So the FD scheme is not stable.

6.3.5 Simplified use of von-Neumann analysis

We discuss how we can simply plug in simple harmonic solutions with wavenumbers $\xi \in [-\pi/h, \pi/h]$ in a given FD stencil to directly update amplification factor g.
The steps of this argument are as follows,

1. **Harmonic decomposition of the initial condition(s)**: First, the IC of the PDE can be written as **superposition** of waves with wavenumbers $\xi \in [-\pi/h, \pi/h]$ following the Fourier series (415b),

$$v_m^0 = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{\imath mh\xi} \hat{v}^0(\xi) \mathrm{d}\xi \tag{439}$$

   where from (415a) $\hat{v}^0(\xi) = h \sum_{m=-\infty}^{\infty} e^{-\imath mh\xi} v_m^0$.

   - That is the solutions at grid points at initial time v_m^0 is the superposition of harmonic waves with wavenumber ξ and amplitude $\hat{v}^0(\xi)$.
   - A single family of these waves can be written as $\frac{1}{\sqrt{2\pi}} e^{\imath x\xi} \hat{v}^0(\xi)$ which when evaluated for $x = x_m = mh$ generates the value $\frac{1}{\sqrt{2\pi}} e^{\imath mh\xi} \hat{v}^0(\xi)$. The sum (integral) of all frequency contributions for $\xi \in [-\pi/h, \pi/h]$ generates the initial value of v_m^0 in (441).
   - We can represent an IC with the single frequency ξ as,

$$\hat{v}_\xi^0(x) = \frac{1}{\sqrt{2\pi}} \hat{v}^0(\xi) e^{\imath \xi x} \tag{440}$$

   if the PDE is higher order in time, IC with higher order temporal derivatives will also be added to (440).

2. **Solution of individual problems with harmonic wave IC**:

   - From (440) for **linear PDEs**: If the underlying PDE is **linear** the solution to the problem with the total IC (v_m^0 is the **linear superposition** of the individual solutions of simple harmonic waves) is equal superposition (integral) of individual solutions with ICs from (440) .
   - The linearity of the problem implies that the **mode number content ξ of the solution at time step** t_n is the solution of the initial value problem with the initial condition (440).
   - That is in,

$$v_m^n = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{\imath mh\xi} \hat{v}^n(\xi) \mathrm{d}\xi \tag{441}$$

     the mode number component $\frac{1}{\sqrt{2\pi}} e^{\imath x\xi} \hat{v}^n(\xi)$ for all t_n is the solution of the numerical (FD) scheme with the IC (440).

   - This can be written as,

$$\text{For IC } v_\xi^0(x) = \frac{1}{\sqrt{2\pi}} \hat{v}^0(\xi) e^{\imath \xi x} \tag{442a}$$

$$\text{solution at } t_n \text{ is } \quad v_\xi^n(x) = \frac{1}{\sqrt{2\pi}} e^{\imath x\xi} \hat{v}^n(\xi) \quad (\text{only } x = x_m = mh \text{ are sampled in an FD scheme}) \tag{442b}$$

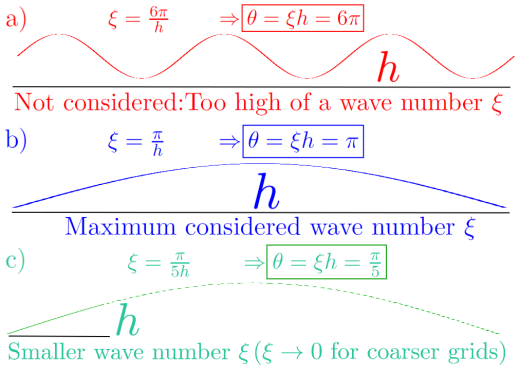
$$\text{where } \hat{v}^n(\xi) = \frac{1}{\sqrt{2\pi}} h \sum_{m=-\infty}^{\infty} e^{-\imath mh\xi} v_m \quad \text{from (415a)} \tag{442c}$$

   - Once question that was raised before was why the range of ξ is $[-\frac{\pi}{h}, \frac{\pi}{h}]$ which was discussed in detail that in the proof of Fourier series (415) we in fact start with (415b) and conclude the values of $\hat{v}^n$ from (415a). That is no proof was required. The physical justification why higher (absolute) values of wavenumbers ξ outside $[-\frac{\pi}{h}, \frac{\pi}{h}]$ was not required was also given by referring to the concept of aliasing in §415b.

- Recalling (427) ($\theta := h\xi$) and the range $\xi \in [-\frac{\pi}{h}, \frac{\pi}{h}]$ we realize,

$$\boxed{|\theta| \leq \pi} \tag{443}$$

- **θ represents how many half a sine wave a grid spacing h can represent.**
  - As shown in the figure in red, $|\theta| > \pi$ corresponds to too high of a wavenumber for a grid and the grid cannot model it. These values are not included in the stability analysis for the simple fact that the limits in (415b) are $-h/\pi$ and h/pi on ξ (thus $[-\pi, \pi]$ for θ). The aliasing phenomena just provides a physical justification why only $|\theta| \leq \pi$ is relevant.
  - The figure in blue ($\theta = \pi$) corresponds to exactly one half a sine wave for a grid spacing h. This is the highest frequency mode a FD grid can capture/model.
  - Finally, for the wave shown in green $\theta = \frac{1}{5}$ and only one fifth of the wave is modeled/captured per grid spacing h. while for stability analysis θ goes all the way to π **for the accuracy of numerical solutions at least 10 grid spacings/elements are recommended per wavelength [Shakib and Hughes, 1991]. This corresponds to** $\theta = \pi/5 \approx 0.6$ as shown for case c) in the figure



3. **Linking amplification factor of a single frequency solution to overall stability**:
   - From the sample solutions in §6.3.7 it will become apparent that the **amplification factor** g (423a), which updates solution from t_n to t_{n+1} in frequency domain ($\hat{v}^{n+1}(\xi) = g(h\xi)$), **can be directly obtained by plugging the solution of the form** (442b) **in the FD stencil**.
   - Now, assuming that g is well-behaved for all $\xi \in [-\pi/h, \pi/h]$, in a sense that the same K for all ξ and $(h, k) \in \Lambda$ can be found for stability condition (432) ($|g(\theta, k, h)| \leq 1 + Kk|$), the question becomes why the stability for the solution of individual problems $\hat{v}^n_\xi(x)$ in (442b) with harmonic IC (442a) $\hat{v}^0_\xi(x) = \frac{1}{\sqrt{2\pi}}\hat{v}^0(\xi)e^{\imath\xi x}$ (at $x_m = mh$) result the stability for any IC?
   - The answer lies in **use of L2 norm for the stability analysis and Parseval relation** in (434a) which states that $||v^n||_h^2 = ||\hat{v}^n||_h^2$ with the RHS being equal to $\int_{-\pi/h}^{\pi/h} |g^{2n}(h\xi, k, h)||\hat{v}^0(\xi)|^2 \mathrm{d}\xi$ from (434b) and subsequently $||\hat{v}^n||_h^2 \leq (1 + Kk)^{2n}||\hat{v}^0||_h^2$ in (435).
   - That is given that solutions to individual solutions with harmonic ICs are well behaved, owing to the Parseval's equality and the linearity of an underlying PDE (that enables superposition) so would be the solution of the problem to any IC.

6.3.6 Summary of the simpler use of von Neumann method

Below is a short summary on the steps that will be taken in the stability analysis (and later dissipation an dispersion error analysis) of FD methods and many other numerical methods **applied to linear PDEs**.

1. **Consider a simple harmonic solution with wavenumber** ξ, *i.e.,* (442b) but without the factor of $1/\sqrt{2\pi}$ (as a constant factor will not affect the stability and error analysis of the method, in fact one can interpret $\hat{v}^n(\xi)$ as $1/\sqrt{2\pi}$ times of the Fourier coefficient with wavenumber ξ at time step t_n, that is the factor is incorporated in the value of $\hat{v}^n(\xi)$).

$$v^n_\xi(x) = e^{\imath x\xi}\hat{v}^n(\xi) \qquad \text{for } \xi \in [-\pi/h, \pi/h] \tag{444}$$

   Note that again to have relevant frequencies that appear in Fourier series expressions (which are the ones that grid can represent without aliasing effect) only $\xi \in [-\pi/h, \pi/h]$ is considered.

2. **Insert the solution** (444) **in a given FD stencil equation** $P_{h,k}$. For example v^n_m is

$$v^n_m = v(x_m, t_n) = v(mh, t_n) = e^{\imath(mh)\xi}\hat{v}^n(\xi) = e^{\imath m(\overbrace{h\xi}^{\theta = h\xi})}\hat{v}^n(\xi) \qquad \Rightarrow$$
$$\boxed{v^n_m = e^{\imath m\theta}\hat{v}^n} \tag{445}$$

   note that in (445) the dependent of $\hat{v}^n$ on ξ is dropped for brevity as the dependence of the Fourier function at t_n on ξ is implicitly known.

Accordingly, some other examples of FD grid values for different m, n are,

$$v_{m+1}^n = e^{\imath(m+1)\theta}\hat{v}^n = e^{\imath\theta}\left(e^{\imath m\theta}\hat{v}^n\right) = e^{\imath\theta}v_m^n \Rightarrow \boxed{v_{m+1}^n = e^{\imath(m+1)\theta}\hat{v}^n = e^{\imath\theta}v_m^n} \tag{446a}$$

$$v_m^{n+1} = e^{\imath m\theta}\hat{v}^{n+1} \tag{446b}$$

In general we can write,

$$\boxed{v_{m+a}^{n+b} = e^{\imath(m+a)\theta}\hat{v}^{n+b}} \tag{447}$$

Now it may be tempting to use (423a) where $\hat{v}^{n+1}(\xi) = g(h\xi)\hat{v}^n(\xi) \Rightarrow \hat{v}^{n+b} = g^b\hat{v}^n$ in (447). Again, the dependent of g on its arguments θ, k, h are dropped for brevity.

However, this simple multiplicative relation between time step values t_n and t_{n+1} is only always true for **temporally single-step FD stencils**!

In that case, however, we can further simplify (447),

$$\boxed{v_{m+a}^{n+b} = e^{\imath a\theta}g^b v_m^n \qquad \text{Temporally one-step FD stencils}} \tag{448}$$

3. **Deriving and solving a recursive relation for the frequency solution $\hat{v}^{n+1}$:**

$$\alpha_q\hat{v}^{n+1} + \alpha_{q-1}\hat{v}^n + \cdots + \alpha_0\hat{v}^{n-q+1} = 0 \tag{449}$$

- This recursive equation is obtained from applying a q-step fD stencil in time.
- The coefficients $\alpha_q, \alpha_{q-1}, \ldots, \alpha_0$ depend on θ, h, k which for brevity are not shown here.
- Now for a one-step FD scheme (that only involves t_{n+1} and t_n values) (449) simplifies to

$$\hat{v}^{n+1} = g\hat{v}^n \qquad \text{as in (423a)} \quad \text{where} \quad g = -\frac{\alpha_n}{\alpha_{n+1}} \tag{450}$$

- For one-step methods that were obtained from the solution of temporally first order ODEs, we had the stability conditions on g from one of the forms (432) ($|g| \leq 1 + Kk$) or (433) ($|g| \leq 1$). That is once we obtain g from von Neumann analysis we can obtain the region of stability Λ from one the aforementioned conditions based on whether g explicitly depends on k or not.
- The analysis of stability of multi-step cases will be discussed in §6.4.

6.3.7 Sample stability analyses with von Neumann method

- In §6.3.3 we obtained the amplification factor of FTBS scheme $g(h\xi) := \left[(1-\bar{k}) + \bar{k}e^{-\imath h\xi}\right]$. As discussed for $a > 0$ the stability was achieved if $\bar{k} = a\frac{k}{h} \leq 1$.
- This resulted in a conditional stability condition in the form $k \leq h/a$.
- We provides a more straightforward computation of g by directly plugging a solution of the form (444) in the FD stencil.
- For FTBS scheme the update equation from (3) is,

$$v_m^{n+1} = (1-\bar{k})v_m^n + \bar{k}v_{m-1}^n$$

- For the FTBS scheme (an example of a one-step scheme) we can directly use (448) $v_{m+a}^{n+b} = e^{\imath a\theta}g^b v_m^n$ (rather than (447))

$$gv_m^n = (1-\bar{k})v_m^n + \bar{k}\left(e^{\imath(-1)\theta}v_m^n\right) \qquad \Rightarrow g = \left[(1-\bar{k}) + \bar{k}e^{-\imath\theta}\right]$$

- This is identical to (423b) but without the need to explicitly go through Fourier transform and use the Parseval's relation.
- The rest of the stability proof is similar to §6.3.3 where we showed for $\bar{k} \leq 1$ ($k \leq h/a$) $g \leq 1$ and FTBS is stable.

Example 4 *Stability of the Lax-Friedrichs scheme (source [Strikwerda, 2004] Example 2.2.4),*

- Consider the Lax-Friedrichs FD equation for the advection equation $u_{,t} + au_{,x} = 0$ from (27d),

$$\frac{v_m^{n+1} - \frac{1}{2}\left(v_{m-1}^n + v_{m+1}^n\right)}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$$

- The update equation for v_m^{n+1} is as follows,

$$v_m^{n+1} = -\frac{\bar{k}}{2}\left(v_{m+1}^n - v_{m-1}^n\right) + \frac{1}{2}\left(v_{m+1}^n + v_{m-1}^n\right) \qquad \text{for} \quad \bar{k} = a\frac{k}{h}$$

- Again, for this one-step scheme we can directly use (448) $v_{m+a}^{n+b} = e^{\imath a\theta} g^b v_m^n$ (rather than (447)),

$$g = -\frac{\bar{k}}{2}\left(e^{+\imath\theta} - e^{-\imath\theta}\right) + \frac{1}{2}\left(e^{+\imath\theta} + e^{-\imath\theta}\right) \tag{451}$$

- Now using the identities,

$$\sin\phi = \frac{e^{\imath\phi} - e^{-\imath\phi}}{2\imath} \tag{452a}$$

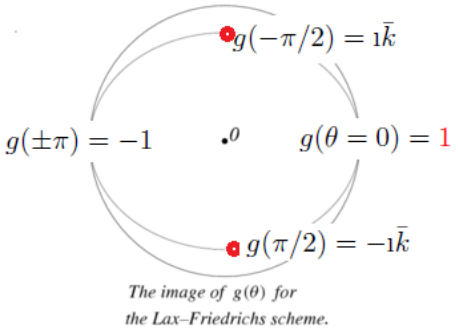
$$\cos\phi = \frac{e^{\imath\phi} + e^{-\imath\phi}}{2} \tag{452b}$$

- In (451) for $\phi = \theta$ we obtain,

$$g(\theta) = \cos\theta - \imath\bar{k}\sin\theta \qquad \text{which gives} \tag{453}$$

$$|g(\theta)|^2 = \cos^2\theta + \bar{k}^2\sin^2\theta \tag{454}$$

- Since $g(\theta)$ is explicitly independent from k we need to use the stability condition (433) ($|g(\theta, k, h)| \leq 1$) rather than (432).
- From (454) $|g(\theta)|^2 \leq 1$ for all $\bar{k}$ if any only if $|\bar{k}| = |a\frac{k}{h}| \leq 1$.
- Irrespective of sign of a the Lax-Friedrichs method is conditionally stable for $k \leq k/|a|$
- The figure shows the image of g in the complex plane as θ spans $[-\pi, \pi]$ when the image remains in the unit circle, that is for the case $k \leq k/|a|$.
- The points corresponding to $\theta = 0, \pm\pi/2, \pm\pi$ are given in the equation below and marked in the figure,



The image of $g(\theta)$ for the Lax–Friedrichs scheme.

$$g(\theta = 0) = 1 \tag{455a}$$

$$g(\pi/2) = -\imath\bar{k} \tag{455b}$$

$$g(-\pi/2) = \imath\bar{k} \tag{455c}$$

$$g(\pm\pi) = -1 \tag{455d}$$

- As discussed before, $g(0) = 1$ from consistency perspective.
- Interestingly, in this case $g(-\pi) = g(\pi) = -1$ and always remains on the unit circle, and instabilities arise from $\theta = \pm\pi/2$ as $g(\theta = \pm\pi) = \mp\imath\bar{k}$ can go outside the unit circle when $\bar{k} > 1$.
- Recall from example 2 that the Lax-Friedrichs method was **conditionally stable** with consistency satisfied for $k^{-1}h^2 \to 0$ when $(h, k) \to 0$.
- Now if, $\lambda = \frac{k}{h}$ remains constant as $h \to 0$ and $\lambda \leq 1/|a|$ (to maintain stability), the consistency condition becomes $k^{-1}h^2 = (h\lambda)^{-1}h^2 = h/\lambda \to$ as $h \to 0$ and λ remains constant and below $1/|a|$.
- Thus, as long as k does not go much faster than h (*e.g.*, $\lambda = k/h$ remaining constant) when $h \to 0$ and $\lambda = h/k < 1/|a|$ the Lax-Friedrich scheme is both consistent and stable thus from the Lax-Richtmyer theorem (theorem 1) it will be convergent.

Example 5 ***Numerical Stability*** *of the Lax-Friedrichs scheme applied to a dynamically unstable problem (source [Strikwerda, 2004] Example 2.2.3),*

- Consider the following problem,

$$u_{,t} + au_{,x} - u = 0 \tag{456}$$

which is an advection reaction equation $u_{,t} + au_{,x} + \beta u = 0$ with $\beta = -1$ where β is the reaction coefficient.

- Physically, the reaction coefficient $\beta \geq 0$ is observed in physical problems which results in decrease of the value of u in time.
- However, if $\beta < 0$ (as in this problem) the problem is what is called **dynamically unstable** meaning that the solution in time tends to infinity and does not remain bounded.

- To demonstrate this, consider that we look for harmonic solutions of the form

$$u(x,t) = e^{\imath\xi x+\omega t} \tag{457}$$

which by plugging into (456) yields,

$$\omega = -\imath\xi a + 1 \tag{458}$$

and by plugging into (457) yields,

$$u(x,t) = e^{\imath\xi(x-at)}e^{t} \tag{459}$$

which would correspond to an initial condition of the form $u(x,0) = u_0(x) = e^{\imath\xi x}$.

- Equation (459) corresponds to a wave of wavenumber ξ moving with speed a and amplified with the factor e^t.
- Due to the factor e^t the solution tends to infinity for any x. This type of the problem where the physical solution can tend to infinity is called *dynamically unstable* and is further discussed in §6.5.
- The concept of stability of a numerical scheme (which can be referred to as numerical stability) is a related by separate concept from dynamic stability of the physical problem:
  1. Numerical stability corresponds to finite times T where the solution for any finite T is bounded by the initial step discrete norms from (409) $||v^n||_h \leq C_T^* \sum_{j=0}^{J} ||v^j||_h$ whereas in dynamic stability we deal with behavior (boundedness) of the **physical (exact)** solution **as time** $t \to \infty$.
  2. The concept of (numerical) stability of a numerical method refers to the existence of the constant C_T **independent of temporal and spatial grid sizes** k, h **but possibly dependent on a given time** T. Numerical stability, ensures that numerical solution does not blow up for a fixed time T due to numerical aspects for example **letting** $k \to 0$**, *i.e.*, temporal step refinement**.
- As will be shown by the solution of this dynamically unstable problem, the numerical solution can be (numerically stable).
- Since, the physical solution tends to infinity, the numerical solution also is expected to tend to infinity, but for a numerically stable method, the possible tendency of the solution is only due to dynamic instability of the underlying exact solution **NOT numerical instability that can cause the solution to blow up even for a fixed time** T **as say** $k \to 0$.
- So, we can have a numerically stable method for a dynamically unstable problem!.
- Ideally, we want to have numerical methods that are numerically stable for both dynamically stable and unstable problems (albeit the physical problem still needs to be well-posed as further discussed in §6.5).
- Now, going back to the FD discretization of (457), the FD update equation is,

$$\frac{v_m^{n+1} - \frac{1}{2}\left(v_{m-1}^n + v_{m_1}^n\right)}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} - v_m^n = 0$$

- The rest of the analysis is very similar to that of example 4 on the stability of Lax-Friedrichs method without reaction term.
- The update equation (451) for v_m^{n+1} is slightly modified to,

$$v_m^{n+1} = -\frac{\bar{k}}{2}\left(v_{m+1}^n - v_{m-1}^n\right) + \frac{1}{2}\left(v_{m+1}^n + kv_{m-1}^n\right) + kv_m^n$$

for $\bar{k} = a\frac{k}{h}$

- Again, for this one-step scheme we can directly use (448) $v_{m+a}^{n+b} = e^{\imath a\theta}g^b v_m^n$ to obtain,

$$g = -\frac{\bar{k}}{2}\left(e^{+\imath\theta} - e^{-\imath\theta}\right) + \frac{1}{2}\left(e^{+\imath\theta} + e^{-\imath\theta}\right) + k \tag{460}$$

- By using (452) yields,

$$g(\theta) = \cos\theta - \imath\bar{k}\sin\theta + k \tag{461}$$

which gives,

$$|g(\theta,\bar{k},k)|^2 = (\cos\theta + k)^2 + \bar{k}^2\sin^2\theta \tag{462}$$

- **Interestingly in this case, in addition to its dependence on** $\theta, \bar{k}$**,** g **explicitly depends on** k.
- Accordingly, in this case we need to consider the more relaxed stability condition (432) $|g(\theta,k,h)| \leq 1 + Kk$ rather than (433) ($|g(\theta,k,h)| \leq 1$) which is for the case where g does not explicitly depend on k.

- Thus, we look for a K such that $|g(\theta,k,h)| \leq 1 + Kk$ for all $\theta \in [-\pi,\pi]$.
- However, this is trivial as we observe from (462),

$$|g(\theta,\bar{k},k)|^2 = (\cos\theta + k)^2 + \bar{k}^2\sin^2\theta \leq (1+k)^2 \tag{463}$$

  for $\bar{k} \leq 1$, thus the FD method is numerically stable for this dynamically unstable problem.
- To emphasize the meaning of numerical stability we observe that in the proof of theorem 2, we obtained the bound (437), that is $||v^n||_h^2 \leq e^{2KT}||v^0||_h^2$ **INDEPENDENT of k as $k \to 0$** when stability is satisfied.

6.4 von Neumann analysis for multi-step FD schemes

- Multi-step methods refer to those requiring beyond (earlier) time step values than t_n to obtain values for t_{n+1} solutions.
- Multi-step methods can be encountered,
  1. Higher than 1$^{\text{st}}$temporal oder for the PDE.
  2. Higher order stencils in time that require beyond (earlier) than time step t_n for updating values for t_{n+1}.
- Below, we provide examples from each category and discuss von Neumann stability analysis using these examples.

6.4.1 von Neumann analysis for leapfrog scheme

- Consider the leapfrog scheme (27e) for the advection equation $u_{,t} + au_{,x} = 0$,

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0 \qquad \Rightarrow \tag{464a}$$

$$v_m^{n+1} = \bar{k}(v_{m-1}^n - v_{m+1}^n) + v_m^{n-1} \tag{464b}$$

- Recall $\bar{k} = a\frac{k}{h}$. Herein, without loss of generality we assume $a > 0$. Given that the stencil is symmetric in x this does not change the following discussion on stability of the leapfrog method and only
- As discussed in §6.3.6, for von Neumann analysis we can simply plug a solution of the form (444) $\hat{v}_\xi^n(x) = e^{\imath x\xi}\hat{v}^n(\xi)$ in the FD stencil.
- This as shown results in (447) $v_{m+a}^{n+b} = e^{\imath(m+a)\theta}\hat{v}^{n+b}$.
- By using this equation in (464b) we obtain,

$$e^{\imath m\theta}\hat{v}^{n+1} = \bar{k}\left(e^{\imath(m-1)\theta}\hat{v}^n - e^{\imath(m+1)\theta}\hat{v}^n\right) + e^{\imath m\theta}\hat{v}^{n-1} \qquad \Rightarrow \qquad e^{\imath m\theta}\hat{v}^{n+1} = e^{\imath m\theta}\left\{-2\bar{k}\hat{v}^n\frac{e^{\imath\theta} - e^{-\imath\theta}}{2} + \hat{v}^{n-1}\right\} \tag{465}$$

- Using (452a) and canceling $e^{\imath m\theta}$ from both sides we obtain,

$$\boxed{\hat{v}^{n+1} = -2\imath\bar{k}\sin\theta\hat{v}^n + \hat{v}^{n-1}} \tag{466}$$

- This can also be recast in the form (449) $\alpha_q\hat{v}^{n+1} + \alpha_{q-1}\hat{v}^n + \cdots + \alpha_0\hat{v}^{n-q+1} = 0$ for this 2-step ($q = 2$) FD scheme,

$$\alpha_2\hat{v}^{n+1} + \alpha_1\hat{v}^n + \alpha_0\hat{v}^{n-1} = 0, \qquad \alpha_2 = 1, \alpha_1 = 2\imath\bar{k}\sin\theta, \alpha_0 = -1 \tag{467}$$

- Unlike one step methods we cannot always express the update from t_n to t_{n+1} by the amplification factor $\hat{v}^{n+1} = g\hat{v}^n$ as the update can take more complex form.
- However, to better understand what the update is in general setting, we let $\hat{v}^{n+1} = g\hat{v}^n \Rightarrow \hat{v}^n = g^n\hat{v}^0$ thus (475) becomes,

$$g^{n+1} + (2\imath\bar{k}\sin\theta)g^n - g^{n-1} = 0 \qquad \Rightarrow \qquad \boxed{g^2 + (2\imath\bar{k}\sin\theta)g - 1 = 0} \tag{468}$$

- Interestingly, for this 2-step FD scheme (468) represents a 2$^{\text{nd}}$order polynomial for the amplification factor g.
- The solutions to (468) are,

$$g_+ = -\imath\bar{k}\sin\theta + \sqrt{1 - \bar{k}^2\sin^2\theta} \tag{469a}$$

$$g_- = -\imath\bar{k}\sin\theta - \sqrt{1 - \bar{k}^2\sin^2\theta} \tag{469b}$$

- Later we comment on what happens when $g_+ = g_-$. For the moment when they are not equal. Recall for one g we have $\hat{v}^n = g^n \hat{v}^0$ as the assumption we started. A linear superposition of the solutions of the form $\hat{v}^n = g^n \hat{v}^0$ from the two g also satisfies (468) (and consequently (475)). Thus, for $g_+ \neq g_-$ we have,

$$\hat{v}^n = A_+(\xi) g_+^n(\theta) + A_-(\xi) g_-^n(\theta) \tag{470}$$

  where the superscript on g's are power as opposed to time step for $\hat{v}^n$.

- Now, it is easy to make the argument that any of the terms from g_+ and g_- must satisfy stability conditions similar to a 1-step FD schemes. This is because by choosing appropriate ICs only one can be left active.

- On the other hand, it is very easy to follow the same proof for the theorem 2 and show that if both g_+ and g_- satisfy 1-step stability conditions (that is (432) and (433)) then so would be $\hat{v}^n$ from (470). This, requires the use of triangle inequality.

- Now, to investigate stability conditions based on g_+ and g_- we observe both of them do not explicitly depend on k and so by fixing $\lambda = k/h$ fixed we can use (433) ($|g(\theta, k, h)| \leq 1$) is used rather than (432) ($|g(\theta, k, h)| \leq 1 + Kk$).

- Now there are two cases for the value of $\bar{k}$:

  1. $\bar{k} \leq 1$: In this case $1 - \bar{k}^2 \sin^2\theta \geq 0$ so $\sqrt{1 - \bar{k}^2 \sin^2\theta}$ is real and,

$$|g_+| = |g_-| = \left(\sqrt{1 - \bar{k}^2 \sin^2\theta}\right)^2 + \left(\bar{k} \sin\theta\right)^2 = 1 - \bar{k}^2 \sin^2\theta + \bar{k}^2 \sin^2\theta = 1. \tag{471}$$

     which satisfies $|g_\pm| \leq 1$ and corresponds to a stable time step value.

  2. $\bar{k} > 1$: In this case $1 - \bar{k}^2 \sin^2\theta < 0$ for $\pi/2 \geq \theta > \sin^{-1}(1/\bar{k})$ (*e.g.*, $\theta = \pi/2$). So $\sqrt{1 - \bar{k}^2 \sin^2\theta} = \imath\sqrt{\bar{k}^2 \sin^2\theta - 1}$ for such θ. In this case, $|g_-| > 1$ some θ so the scheme is not stable. For example, for $\theta = \pi$ we have,

$$g_-(\pi) = -\imath\bar{k}\sin(\pi/2) - \sqrt{1 - \bar{k}^2 \sin^2(\pi/2)} = -\imath\left(\bar{k} + \sqrt{\bar{k}^2 - 1}\right) > 1 \qquad (\text{since } \bar{k} > 1)$$

- Given that for $\bar{k} > 1$ the scheme is unstable and for $\bar{k} \leq 1$ it is stable with the analysis above, we may be tempted to call the leapfrog method to be stable for $\bar{k} \leq 1$.

- However, this is not entirely correct as we have not taken care of the case that $g_+ = g_-$.

- Obviously, given that $g_\pm = -\imath\bar{k}\sin\theta \pm \sqrt{1 - \bar{k}^2 \sin^2\theta}$ for $\bar{k} < 1$ $1 - \bar{k}^2 \sin^2\theta > 0$ and $g_+ \neq g_-$. That is the assumption of distinct roots for g_- and g_+ always holds for $\bar{k} < 1$ and the preceding stability condition for $\bar{k} < 1$ remains valid. That is, for $\bar{k} < 1$ the leapfrog method is stable.

- Now, we focus on the case where g_+ can become equal to g_- that is for $\bar{k} = 1$,

$$g_\pm = -\imath\sin\theta \pm \sqrt{1 - \sin^2\theta} \qquad \text{for } \bar{k} = 1 \tag{472}$$

- Clearly for $\sin^2\theta = 1$, *i.e.*, for $\theta = \pm\pi/2$, we have,

$$g_\pm := g = -\imath \qquad \text{for } \bar{k} = 1 \text{ and } \theta = \pm\pi/2 \tag{473}$$

- It is clear that the simple amplification factor $\hat{v}^{n+1} = g\hat{v}^n$ in this case results in final satisfaction of the recursive relation (475) for $\bar{k} = 1, \theta = \pm\pi/2$.

- However, there is another solution for the recursive relation (475) ($\alpha_2 \hat{v}^{n+1} + \alpha_1 \hat{v}^n + \alpha_0 \hat{v}^{n-1} = 0$) for $\alpha_2 = 1, \alpha_1 = 2\imath\bar{k}\sin\theta, \alpha_0 = -1$ in this particular case that there is a repeated root for (468).

- The form of general recursive relations will be further discussed in §6.4.3.

- For the moment, we observe beside $\hat{v}^n = A(\xi) g^n = A(\xi)(-\imath)^n$ there is another solution of the form,

$$\hat{v}^n = B(\xi) n g^n, \qquad \text{for } g = -\imath \quad \text{(obtained from (473) for } \bar{k} = 1 \text{ and } \theta = \pm\pi/2) \tag{474}$$

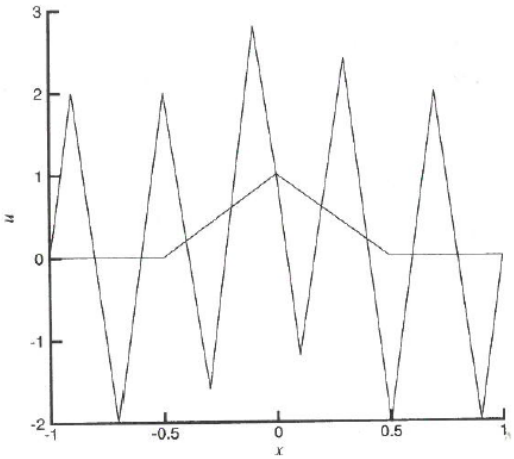
- This can be easily verified by plugging this solution in (475) for $\bar{k} = 1, \theta = \pm\pi/2$. That is, for $\alpha_2 = 1$, $\alpha_1 = 2\imath\bar{k}\sin\theta = 2\imath$, $\alpha_0 = -1$:

$$\begin{aligned}\alpha_2\hat{v}^{n+1} + \alpha_1\hat{v}^n + \alpha_0\hat{v}^{n-1} &= B(\xi)(n+1)(-\imath)^{n+1} + 2\imath B(\xi)n(-\imath)^n + (-1)B(\xi)(n-1)(-\imath)^{n-1} \\ &= B(\xi)(-\imath)^{n-1}\left\{(n+1)(-\imath)^2 + 2\imath n(-\imath)^1 - (n-1)\right\} = B(\xi)\imath^{n-1}\left\{(-n-1)+(2n)+(-n+1)\right\} = 0\end{aligned} \tag{475}$$

- So a general solution for $\bar{k} = 1$ and $\theta = \pm\pi/2$ takes the form,

$$\hat{v}^n = A(\xi)g^n + B(\xi)ng^n = A(\xi)(-\imath)^n + B(\xi)n(-\imath)^n \tag{476}$$

- The appearance of the the factor n in the solution and noting that $|g| = 1$ (so $|g|^n = 1$) mean that leapfrog scheme for $\bar{k} = 1$ is not stable. For $\theta = \pm\pi$ the solution **linearly (not exponentially)** is unstable. This type of instability called weak instability as opposed to strong (exponential) instability that would arise when $|g| > 1$. An example of weak instability is shown in the figure.



Leapfrog weak (algebraic) instability for $\bar{k} = 1$.

- Note that for a fixed time T by letting $k \to 0$ and choosing $t_n = nk \approx T$ we observe $n \approx T/k \to \infty$ for a fixed T and we have n grows in (476) and multiplies $B(\xi)$. The proof of instability for $\bar{k} = 1$ can be formally done through Exercise 4.1.5 in [Strikwerda, 2004].
- In this case, as opposed to one-step schemes considered the limiting value of instability for $\bar{k}$ itself is not included in stability zone.
- Rather, stability of leapfrog method requires $\bar{k} = ak/h < 1$ and the method is **NOT stable for** $\bar{k} = 1$

$$\bar{k} = \frac{ka}{h} < 1 \qquad \text{for the stability of leapfrog method} \tag{477}$$

6.4.2 von Neumann analysis for a temporally 2nd order PDE

- Before starting an example, we present the stability condition for temporally 2nd order PDEs.
- First recall that in the definition of stability of first order PDEs, *i.e.,* definition 4, in equation (419) we required $||v^n||_h^2 \leq C_T \sum_{j=0}^{J} ||v^j||_h^2$ for $0 \leq nk \leq T$ with $(h,k) \in \Lambda$. where the need to go beyond the initial values ($J > 0$) on the RHS can arise when multi-step methods were used for the solution of temporally first order PDEs.
- For temporally second order PDEs the stability condition is slightly modified as shown below,

Definition 5 *Stability of temporally first order PDEs: A finite difference scheme $P_{h,k}v_m^n = 0$ for a temporally second-order PDE is stable in the stability region Λ if there an integer J such that for any positive time T, there is a constant C_T such that,*

$$h||v^n||_h^2 \leq (1+n^2)C_T h\sum_{j=0}^{J}||v^j||_h^2 \qquad \textit{for } 0 \leq t_n = nk \leq T \quad \textit{with} \quad (h,k) \in \Lambda. \tag{478}$$

- The extra factor $(1+n^2)$ in (478) required for the temporally second order PDEs compared to first order ones in definition 4.
- This factor is introduced given that second order PDEs without $u_{,t}$ admit a linear growth of solution.
- For example, consider the wave equation (56a) ($u_{,tt} - c^2u_{,xx} = r$) but without the source term,

$$u_{,tt} - a^2u_{,xx} = 0 \tag{479}$$

where similar to the advection equation $u_{,t} + au_{,x} = 0$, a is the wave speed.

- This equation admits solutions of the form,

$$u(x,t) = At, \qquad \text{for a constant } A. \tag{480}$$

- That is the solution can grow linearly in time and the factor $(1+n^2)$ reflects the possibility of linear growth in t by the solution of temporally second order PDEs.
- Now consider a central-space central-time FD scheme being applied to the solution of (479). The FD equation will be,

$$\frac{v_m^{n+1}-2v_m^n+v_m^{n-1}}{k^2}-a^2\frac{v_{m+1}^n-2v_m^n+v_{m-1}^n}{h^2}=0 \tag{481}$$

- By using the notation $\bar{k}=a\frac{k}{h}$ (482) can be expressed as,

$$v_m^{n+1}-2v_m^n++v_m^{n-1}-\bar{k}^2\left(v_{m+1}^n-2v_m^n+v_{m-1}^n\right)=0 \tag{482}$$

- Again, similar to the analysis of leapfrog method in §6.4.1, for the von Neumann analysis we can simply plug a solution of the form (444) $\hat{v}_\xi^n(x)=e^{\imath x\xi}\hat{v}^n(\xi)$ in the FD stencil resulting in (447) ($v_{m+a}^{n+b}=e^{\imath(m+a)\theta}\hat{v}^{n+b}$).
- Thus, (482) becomes

$$e^{\imath m\theta}\hat{v}^{n+1}-2e^{\imath m\theta}\hat{v}^n+e^{\imath m\theta}\hat{v}^{n-1}-\bar{k}^2\left(e^{\imath(m+1)\theta}\hat{v}^n-2e^{\imath m\theta}\hat{v}^n+e^{\imath(m-1)\theta}\hat{v}^n\right)=0 \qquad \Rightarrow$$
$$\hat{v}^{n+1}-2\left\{1+2\bar{k}^2\left(\frac{e^{\imath\theta}+e^{-\imath\theta}-2}{2}\right)\right\}\hat{v}^n+\hat{v}^{n-1}=0$$

- Now from (452b) $(\cos\phi=\left(e^{\imath\phi}+e^{-\imath\phi}\right)/2)$ and $1-\cos\phi=2\sin(\phi/2)$ for $\phi=\theta$ we obtain,

$$\hat{v}^{n+1}-2\left(1-2\bar{k}^2\sin^2\frac{\theta}{2}\right)\hat{v}^n+\hat{v}^{n-1}=0$$

- As mentioned before, in §6.4.3 we further discuss how we can solve recursive relations of the form (483) in general.
- In §6.4.1 we observed that the solution for a recursive relation of the form (483) (with three consecutive steps involved) could be deduced from the roots of a polynomial that was derived from it.
- Assuming that amplification factor relation $\hat{v}^{n+1}=g\hat{v}^n$ holds (that is $\hat{v}^n=g^n\hat{v}^0$ by a similar process to (468) we reach at,

$$g^2-2A_1g+A_2=0, \qquad \text{for } A_1=1-2\bar{k}^2\sin^2\frac{\theta}{2}, \quad A_2=1 \tag{483}$$

- The roots of (483) are,

$$g_\pm=\left[1-2\bar{k}^2\sin^2\frac{\theta}{2}\right]\pm\left[2\bar{k}\sin\frac{\theta}{2}\sqrt{\bar{k}^2\sin^2\frac{\theta}{2}-1}\right] \tag{484}$$

- Clearly directly checking whether $|g_\pm|\le 1$ (needed for the stability analysis of this FD scheme where g would not depend on k and $\bar{k}$ is kept fixed as $k\to 0$) from (484) is difficult.
- Instead, without actually using the root solutions (484) we can verify whether $|g|\le 1$ (or $g<1$ if needed) is satisfied.
- Recall that the necessary and sufficient conditions for the roots of (483) to satisfy $|g|\le 1$ is (362) was $-1\le A_2\le 1$, $-\frac{A_2+1}{2}\le A_1\le\frac{A_2+1}{2}$ except the point with repeated roots of -1 or $+1$. That is,

$$-1\le|A_2|=|1|\le 1 \qquad \text{satisfied} \tag{485a}$$
$$-1\le 1-2\bar{k}^2\sin^2\frac{\theta}{2}\le 1 \qquad \bar{k}^2\sin^2\frac{\theta}{2}\le 1 \tag{485b}$$

- The condition (485b) is satisfied if $-1\le\bar{k}\sin\frac{\theta}{2}\le 1$ for $\theta\in[-\pi,\pi]$ which holds only if $\bar{k}\le 1$.
- Since we consider the case $\bar{k}=1$ below (as it will result in repeated roots for (483)) we first focus on the case $\bar{k}<1$.
- In this case, it is easy to verify the two roots for g are distinct and both satisfy $g\le 1$ so the FD scheme is stable for $|\bar{k}|\le 1$.
- Now, we consider the case $\bar{k}=1$ to evaluate whether the FD scheme can be stable for this normalized time step or not.
- Plugging $\bar{k}=1$ in (483) we obtain,

$$g^2-2(1-2\sin^2\frac{\theta}{2})g+1=0 \qquad \Rightarrow \qquad g^2-2(\cos\theta)g+1=0, \qquad \text{for } \bar{k}=1 \tag{486}$$

- Now for $\theta = 0, \pm\pi$ $\cos\theta = \pm 1$ and has the repeated roots of either +1 or -1.
- Without repeating the same process done in (475) it can easily be verified that for the two repeated root case g the solutions to (483) take the form $\hat{v}^n = g^n \hat{v}^0$ or $\hat{v}^n = n g^n \hat{v}^0$.
- Again, following the same process as in §6.4.1 (*i.e.,* (476)) the final solution for $\hat{v}^n$ for $\bar{k} = 1$, $\theta = 0, \pi, -\pi$ is,

$$\hat{v}^n = A(\xi) g^n + B(\xi) n g^n = A(\xi)(\pm 1)^n + B(\xi) n (\pm 1)^n \tag{487}$$

  where $+1$ is for $\theta = 0$ and -1 for highest wavenumbers $\theta = \pm\pi$.
- Now, it is evident as in (476) the solution grows linearly with n for $\bar{k} = 1$ and $\theta = 0, \pm\pi$ through the second term $B(\xi) n (\pm 1)^n$. Given that for t_n time is $t_n = nk$ a solution of the form $\hat{v}^n \propto n$ corresponds to a linear At grows of the solution in time for the second order PDE (479).
- However, as discussed in the stability condition for temporally second order ODEs (definition 5) and explicitly shown by (480) ($u(x,t) = At$ being a solution to (479) $u_{,tt} - a^2 u_{,xx} = 0$ such linear growth of the solution if physical in this case and $\bar{k} = 1$ actually results in a numerically stable scheme for the wave equation.
- In contrast, remember that for the leapfrog method stability condition was $|\bar{k} < 1$ and linear growth was not allowed. For the wave equation, however, the stability condition is $|\bar{k} \leq 1$ and allows linear growth of solution in time.
- Interestingly, for both explicit schemes the maximum allowable time step corresponding to CFL number equal to 1 ($\bar{k} = ak/h$) can be taken (still with leapfrog method we can take $\bar{k} < 1$ but not exactly equal to one).
- We know that this is not always true and for many explicit methods applied to hyperbolic problems smaller CFL numbers must be taken for stability reasons.

6.4.3 General solution of recursive relations related to stability analysis

- We observe that the stability analysis of a multi-step method reduces to solving a recursive equation of the form (449):

$$\alpha_q \hat{v}^{n+1} + \alpha_{q-1} \hat{v}^n + \cdots + \alpha_0 \hat{v}^{n-q+1} = 0 \tag{488}$$

- Example of this kind were observed for the leapfrog method applied to $u_{,t} + a u_{,x} = 0$ and central time central space difference applied to $u_{,tt} - a^2 u_{,xx} = 0$; *cf.* (466) and (483), respectively.
- We observed that in both cased assuming an amplification factor relation in the form (423a) $\hat{v}^{n+1}(\xi) = g(h\xi)\hat{v}^n(\xi)$ resulted in a second order polynomial whose solution for $\hat{v}^n(\xi) = g^n(h\xi)\hat{v}^0(\xi)$ were valid. Recall again that the superscript n on g is the power operation as opposed to that for $\hat{v}^n(\xi)$ which denotes time step number.
- When the roots of the second order polynomials (466) and (483) were repeated a term in the form $\hat{v}^n(\xi) = n g^n(h\xi)\hat{v}^0(\xi)$ was also a valid solution.
- Below, we derive general solutions to recursive relations of the form (488).
- Assume that recursive relation (488) admits a solution in the form with an amplification factor (423a) $\hat{v}^{n+1}(\xi) = g(h\xi)\hat{v}^n(\xi)$. That is, $\hat{v}^n(\xi) = g^n(h\xi)\hat{v}^0(\xi)$. Plugging this in (488) yields,

$$\alpha_q g^{n+1} + \alpha_{q-1} g^n + \cdots + \alpha_0 g^{n-q+1} = 0 \qquad \Rightarrow \qquad \boxed{\alpha_q g^q + \alpha_{q-1} g^{q-1} + \cdots + \alpha_1 g + \alpha_0 = 0} \tag{489}$$

- For a moment, assume that all q roots of q order polynomial in g are distinct. Since for any root g_j $\hat{v}^n = g_j^n \hat{v}^0$ is a solution any linear combination of these solutions with factors independent on n (*e.g.,* they can depend on ξ for example) can be a solution. So, a solution to (489) can be written as,

$$\hat{v}^n = \sum_{j=1^q} A_j g_j^n \tag{490}$$

  where again A_j can depend on any parameters that appear in α coefficients in (488) such as $\bar{k}$, θ that were present in (466) and (483).
- Basically, the q recursive relation (488) will have q unknowns A_j that will be obtained from the first q steps of the solution $\hat{v}^0, \cdots, \hat{v}^{q-1}$.
- However, when some roots of (489) are repeated we do not have all dofs A_j $1 \leq j \leq q$ present in (490).

- In such cases, there are some other nontrivial solutions to (488).
- Looking back at the two-step problems in §6.4.1 and §6.4.2 when (489) had repeated roots for $\bar{k} = 1$ (*e.g.*, $g_{1,2} = -1$ when $\theta = \pm\pi/2$ for the leapfrog method (473) and $g_{1,2} = 1$ or $g_{1,2} = -1$ $\theta = 0, \pm\pi$ for the central time central space scheme in (486)) we also had a secondary solution of the form $\hat{v}^n = An g^n$; *cf.* (476) ($\hat{v}^n = A(\xi)g^n + A_*(\xi)ng^n = A(\xi)(-1)^n + A_*(\xi)n(-1)^n$) and (487) ($\hat{v}^n = A(\xi)g^n + A_*(\xi)ng^n = A(\xi)(\pm 1)^n + A_*(\xi)n(\pm 1)^n$), respectively.
- This suggests, that if a root g to (489) is repeated m times then $\hat{v}^n p(n) g^n$ will be a solution to (488) for an arbitrary polynomial $p(n)$ of order $m-1$.
- This in fact is through and is formalized in the following theorem.

Theorem 3 *Solution to a recursive equation: Consider the q-order homogeneous linear recursive (recurrence) relation,*

$$\alpha_q \hat{v}^{n+1} + \alpha_{q-1}\hat{v}^n + \cdots + \alpha_0 \hat{v}^{n-q+1} = 0, \qquad n = 0, 1, \ldots \tag{491}$$

with $\alpha_q \neq 0$, $\alpha_0 \neq 0$ (if the end point coefficients are zero, the relation can be case in a recursive relation with smaller number of steps) and $\alpha_j \in \mathbb{R}$ and ***not dependent on*** *n (they can depend on any parameter other than n).*

We define the q-order ***characteristic polynomial****,*

$$\rho(g) = \alpha_q g^q + \alpha_{q-1} g^{q-1} + \cdots + \alpha_1 g + \alpha_0 \tag{492}$$

Let, g_r, $1 \leq r \leq l$, $l \leq q$, be the distinct roots of the polynomial $\rho(g)$, and let $m_r \geq 1$ denote the multiplicity of g_r, with $m_1 + \cdots + m_l = q$. If a sequence of complex numbers $\hat{v}^n$ satisfies (491), then

$$\boxed{\hat{v}^n = \sum_{r=1}^{l} p_r(n) g_r^n, \qquad \text{for all } n \leq 0} \tag{493}$$

where $p_r(.)$ is a polynomial in n of degree $m_r - 1$, $1 \leq r \leq l$. In particular, if a root r g_r is simple, that is $m_r = 1$, then p_r is constant and not dependent on n.

- The solution of the form (493) seems trivial based on previous examples.
- However, the interested reader can refer to Lemma 12.1, pages 333–5 of [Süli and Mayers, 2003] for the formal proof of this theorem.
- The interpretation of the impact of the values of roots g_r and their multiplicity m_r, when (491) is obtained by von Neumann analysis, on the stability of a linear PDE solved by FD methods is discussed in the next section.

6.4.4 von Neumann stability analysis based on the characteristic polynomial

- Assuming that (491) is obtained by von Neumann analysis by plugging the harmonic solution (444) ($\hat{v}^n_\xi(x) = e^{ix\xi}\hat{v}^n(\xi)$) of wavenumber $\xi \in [-\pi/h, \pi/h]$ in the FD stencil, the stability of the FD depends on the growth of $\hat{v}^n(\xi)$ which satisfy a q-order recursive relation of the form (491),

$$\alpha_q \hat{v}^{n+1} + \alpha_{q-1}\hat{v}^n + \cdots + \alpha_0 \hat{v}^{n-q+1} = 0, \qquad n = 0, 1, \ldots \tag{494}$$

  for a q-step FD scheme.
- As discussed in theorem (3), the coefficients $\hat{v}^n$ follow an evolution law of the form (493).

$$\hat{v}^n = \sum_{r=1}^{l} p_r(n) g_r^n, \qquad \text{for all } n \leq 0 \tag{495}$$

  where g_r are the distinct roots of characteristic polynomial (492),

$$\rho(g) = \alpha_q g^q + \alpha_{q-1} g^{q-1} + \cdots + \alpha_1 g + \alpha_0 \tag{496}$$

  with m_r being the multiplicity of root g_r.
- So, in principal we can obtain roots g_r their multiplicity and the form of (496) (*i.e.,* orders of p_r polynomials).
- Having this information we can analyze the growth and boundedness of FD solutions and discuss its stability.
- This is discussed in the following theorems:

Theorem 4 *Stability analysis in frequency domain for multi-step FD schemes: When the coefficients α in (494) are explicitly independent of h and k, then the necessary and sufficient condition for the FD scheme to be stable is that the roots g_r satisfy the following conditions,*

1. $|g_r(\theta)| \leq 1$ *for all* $1 \leq r \leq l$ *and* $\theta \in [-\pi, \pi]$.
2. *If* $|g_r(\theta)| = 1$ *then* g_r *is a simple root* ($m_r = 1$) *for a temporally first order ODE and at most of second multiplicity for temporally second order PDEs* ($m_r \leq 2$).

Some clarifications of the theorem 4 are,

- Allowing a multiplicity 2 for temporally second order PDEs is justified by the fact that for example $u_{,tt} - a^2 u_{,xx} = 0$ permits a temporally linearly growing solution of t. This topic was discussed in detail in §6.4.2.
- The statement that "α in (494) are explicitly independent of h and k" means that all the coefficients in (494) are made independent of h and k. For example, for the solution of $u_{,t} + a u_{,x} = 0$, and $u_{,tt} - a^2 u_{,xx}$ we can work with the nondimensional time step $\bar{k} = ak/h$. Similarly, for $u_{,t} - D u_{,xx} = 0$ we can define $\bar{k} = Dk/h^2$. If,
  1. α_j do not explicitly depend on h and k and only indirectly depend on $\bar{k}$,
  2. $\bar{k}$ is kept fixed as resolution is increased ($h \to 0$)

  then we can use theorem 4.
- As we saw before in some problems, for example when the underlying PDE was not dynamically stable in §5, the formula for g cannot be made independent of k. See for example, (461) ($g(\theta) = \cos\theta - \imath \bar{k} \sin\theta + k$).
- In such cases, the stability condition on the roots g must be an extension to the simple formula for one-step FD schemes (432) ($|g(\theta, k, h)| \leq 1 + Kk$) rather than the current theorem that is the generalization of the case where g was explicitly independent of h and k,*i.e.,* (433) ($|g(\theta, k, h)| \leq 1$) in theorem 2 for one step FD schemes.
- The extension of theorem 4 to the case where α's explicitly depend on h, k to multi-step schemes is a bit more complex and is not repeated here.
- The interested reader can refer to Theorem 4.2.2, p. 105 in [Strikwerda, 2004] for the form of the theorem for multi-step FD schemes applied to temporally first order PDEs when g depends on k, h.
- Perhaps, the most important aspect in the use of (4) is the evaluation of $|g_r|$ and m_r for the roots of (496) which can be quite complicated in multi-step methods even for a 2$^{\text{nd}}$order case with general complex coefficients for α_1 and α_0.
- In practice we do not need to solve for the roots g_r as the following discussion clarifies that all we need to do is evaluating conditions 1 and 2 above **rather than actually solving for the roots** g_r (condition 1 refers to $|g_r(\theta)| \leq 1$). This is further discussed below.

6.4.5 Theory of Schur and von Neumann polynomials

- Consider the characteristic polynomial (496),

$$\rho_q(z) = \alpha_q z^q + \alpha_{q-1} z^{q-1} + \cdots + \alpha_1 z + \alpha_0$$

  that arises from von Neumann analysis of a simply harmonic solution as discussed in §6.4.4. Again, we assume $\alpha_q \neq 0$ and $\alpha_0 \neq 0$ as otherwise the recursive relation in theorem 4 can result in a lower order polynomial.
- The subscript q for ρ_q is to denote its order being q.
- Based on the theorem 4 the stability analysis of a FD scheme reduces to the following two conditions,or the FD scheme to be stable is that the roots z_r satisfy the following conditions for the roots of (6.4.5),
  1. $|z_r(\theta)| \leq 1$ for all $1 \leq r \leq l$ and $\theta \in [-\pi, \pi]$.
  2. If $|z_r(\theta)| = 1$ then z_r is a simple root ($m_r = 1$) for a temporally first order ODE and at most of second
- A direct determination of the roots of (6.4.5) and evaluation of whether $|z_r(\theta)| \leq 1$ and if $|z_r(\theta)| = 1$ what their multiplicities are is a formidable task even for a second order (6.4.5) when α are general complex numbers.
- Fortunately, there is a well-developed theory and algorithm for checking whether these polynomials satisfy the conditions mentioned above.
- We begin with some definitions.

Definition 6 *Schur polynomial: The polynomial $\rho(z)$ is a Schur polynomial if all its roots z_r satisfy,*

$$|z_r| < 1 \tag{497}$$

Definition 7 *von Neumann polynomial: The polynomial $\rho(z)$ is a von Neumann polynomial if all its roots z_r satisfy,*

$$|z_r| \leq 1 \tag{498}$$

Definition 8 *Simple von Neumann polynomial: The polynomial $\rho(z)$ is a simple von Neumann polynomial if $\rho(z)$ is a von Neumann polynomial and its roots on the unit circle are simple roots. That is,*

$$|z_r| \leq 1 \quad \text{and when} \quad |z_r| = 1 \quad \Rightarrow \quad m_r = 1 \tag{499}$$

Definition 9 *Conservative polynomial: The polynomial $\rho(z)$ is a conservative polynomial if all its roots lie on the unit circle. That is,*

$$|z_r| = 1 \quad \text{for all roots} \tag{500}$$

- Now, for a polynomial $\rho_q(z)$ of exact order q, we define the polynomial ρ^* by,

$$\boxed{\rho^*(z) = \sum_{j=0}^{q} \bar{\alpha}_{q-j} z^j} \tag{501}$$

  where $\bar{(.)}$ is the complex conjugate operator. That is, $\overline{a + b\imath} = a - b\imath$ for $a, b \in \mathbb{R}$.

- Finally, for a polynomial $\rho_q(z)$ of degree q we define recursively the polynomial,

$$\rho_{q-1}(z) = \frac{\rho_q^*(0)\rho_q(z) - \rho_q(0)\rho_q^*(z)}{z} \tag{502}$$

- It is easy to see that the degree of ρ_{q-1} is less than that of ρ_q.
- The next two theorems give recursive tests for Schur polynomials and simple von-Neumann polynomials.

Theorem 5 *Recursive evaluation for Schur polynomials: ρ_q is a Schur polynomial of exact degree q if and only if ρ_{q-1} is a Schur polynomial of exact order $q-1$ and $|\rho_q(0)| < |\rho_q^*(0)|$.*

Theorem 6 *Recursive evaluation for simple von Neumann polynomials: ρ_q is a simple von Neumman polynomial if and only if either,*

1. *$|\rho_q(0)| < |\rho_q^*(0)|$ and ρ_{q-1} is a simple von Neumann polynomial or*
2. *ρ_{q-1} is identically zero and ρ_q' is a Schur polynomial.*

Theorem 7 *Recursive evaluation for von Neumann polynomials: ρ_q is a von Neumman polynomial of exact degree q if and only if either,*

1. *$|\rho_q(0)| < |\rho_q^*(0)|$ and ρ_{q-1} is a von Neumann polynomial or*
2. *ρ_{q-1} is identically zero and ρ_q' is a von Neumann polynomial.*

The proof of these theorems and a few more theorems on conservative polynomials can be found in [Strikwerda, 2004] section 4.3.

- To understand how these theorems will be utilized in evaluating the stability of a multi-step method, we refer to the two conditions in the theorem 4 which where,
  1. $|z_r(\theta)| \leq 1$ for all $1 \leq r \leq l$ and $\theta \in [-\pi, \pi]$.
  2. If $|z_r(\theta)| = 1$ then z_r is a simple root ($m_r = 1$) for a temporally first order ODE and at most of second multiplicity for temporally second order PDEs ($m_r \leq 2$).
- Basically, for a temporally first order PDE these two conditions correspond to ρ_q being a **simple von Neumann polynomial**. This condition can be **recursively verified with theorem 6**.

- For cases that $|z_r|=1$ can have $m_r>2$ (*e.g.,* temporally second order ODEs) and in other applications that this becomes relevant, we can first take case of $|z_r(\theta)|\leq 1$ by checking whether ρ_q is a Schur polynomial by theorem 5 and then separately take care of roots $|z_r|=1$ with multiplicity $m_r>1$.
- In either case, theorems (5) and (6) are extremely useful for general stability analysis of many methods as they provide the information on whether the roots are on or inside unit circle without actually solving them by recursively reducing the polynomial order to make the verification much simpler.
- In the following we discuss the conditions where a general second order characteristic is a Schur polynomial, simple von Neumann polynomial, or one with $|z_r|=1$ and $m_r=2$.

6.4.6 Analysis of $z^2+\alpha_1 z+\alpha_0=0$

- Consider the second order characteristic polynomial ($q=2$),

$$\rho_2(z)=z^2+\alpha_1 z+\alpha_0 \tag{503}$$

that is $\alpha_2=1$. where the values α_1,α_0 can be general complex numbers.

- We want to analyze the roots of this polynomial.
- A characteristic polynomial of this form occurs in the stability analysis of two-step FD schemes (*e.g.,* occurring for temporally second order PDEs and 2 step schemes for first order PDEs as those in §6.4.2 and §6.4.1, respectively).
- It also occurs in the stability analysis of second order ODEs, FEM implementation of second order PDEs and many other schemes.
- So, the analysis of the root of a polynomial of this form is of great practical importance.
- It also provides examples how theorems 5 and 6 are used.
- As the first step we compute ρ_2^* from (501):

$$\rho_2^*(z)=\overline{\alpha}_0 z^2+\overline{\alpha}_1 z+\overline{\alpha}_2=\overline{\alpha}_0 z^2+\overline{\alpha}_1 z+1 \qquad (\overline{\alpha}_2=1) \tag{504}$$

- Now having $\rho_2(z)$ and $\rho_2^*(z)$ we use (502) to compute ρ_1,

$$\rho_1(z)=\frac{\rho_2^*(0)\rho_2(z)-\rho_2(0)\rho_2^*(z)}{z}=\frac{\overline{\alpha}_2\left\{z^2+\alpha_1 z+\alpha_0\right\}-\alpha_0\left\{\overline{\alpha}_0 z^2+\overline{\alpha}_1 z+\overline{\alpha}_2\right\}}{z}\Rightarrow$$

$$\boxed{\rho_1(z)=\left(1-|\alpha_0|^2\right)z+\left(\alpha_1-\alpha_0\overline{\alpha}_1\right)} \tag{505a}$$

Note that we used $\alpha_2=1$.

- The root of $\rho_1(z)$, denoted by g^1 is,

$$g^1=\frac{\alpha_0\overline{\alpha}_1-\alpha_1}{1-|\alpha_0|^2} \tag{506}$$

- **Schur polynomial**: Now, we consider conditions that (503) ($\rho_2(z)=z^2+\alpha_1 z+\alpha_0$) is a Schur polynomial, *i.e.,* both roots z_r satisfy $|z_r|<1$.
  - From theorem 5 this occurs if,
    1. ρ_1 is a Schur polynomial of order 1.
    2. $|\rho_2(0)|<|\rho_2^*(0)|$.
  - From (503), (504), (505a), (506) these two conditions, respectively, are,
    1. $|g^1|<1\Rightarrow|\alpha_0\overline{\alpha}_1-\alpha_1|<|1-|\alpha_0|^2|$.
    2. $|\alpha_0|<1$.
  - These conditions can be summarized as,

$$\rho_2(z)=z^2+\alpha_1 z+\alpha_0 \text{ is a Schur polynomial (roots satisfy } |z_{1,2}|<1) \text{ iff}$$

$$\begin{cases}|\alpha_0|<1\\ |\alpha_1-\alpha_0\overline{\alpha}_1|<1-|\alpha_0|^2\end{cases} \tag{507}$$

- **simple von Neumann**: Now, we consider conditions that (503) ($\rho_2(z)=z^2+\alpha_1 z+\alpha_0$) is a Simple von Neumann polynomial, *i.e.,* both roots z_r satisfy $|z_r|<=1$ and if for any r $|z_r|=1$ then z_r is a simple root ($m_r=1$) .
  - From theorem 6 this occurs if one of the two conditions below hold
    - **Condition 1**:
      1. ρ_1 is a simple von Neumann polynomial and
      2. $|\rho_2(0)|<|\rho_2^*(0)|$.

      or
    - **Condition 2**:
      1. ρ_1 is identically zero.
      2. ρ_2' is a Schur polynomial.
  - **Condition 1**: The condition 1 is very similar to conditions for $\rho_2(z)$ being a Schur polynomial which was discussed just before this case with the difference that ρ_1 can be a von Neumann polynomial rather than only a Schur polynomial. That is, ρ_1 can admit root of magnitude 1.
  - Following condition 1 we reach at,

$$\begin{aligned}&|\alpha_0|<1\\&|\alpha_1-\alpha_0\overline{\alpha}_1|\le 1-|\alpha_0|^2\end{aligned}\qquad\text{Condition 1 for } \rho_2(z) \text{ being a simple von Neumann polynomial} \tag{508}$$

    note the slight difference of the inequality condition in the second equation relative to (507).
  - **Condition 2**: Before enforcing the conditions we write α_0 and α_1 in polar coordinate systems: $\alpha_0=R_0e^{\imath\phi_0}$ and $\alpha_1=R_1e^{\imath\phi_1}$. The satisfaction of condition 2 requires the following.
    1. ρ_1 is identically zero. From (505a) ($\rho_1(z)=\left(1-|\alpha_0|^2\right)z+(\alpha_1-\alpha_0\overline{\alpha}_1)$) this yields,
       (a) $1-|\alpha_0|^2=0 \;\Rightarrow\; |\alpha_0|=1 \;\Rightarrow\; R_0=1$, that is $\alpha_0=e^{\imath\phi_0}$.
       (b) $\alpha_1-\alpha_0\overline{\alpha}_1=0 \;\Rightarrow\; R_1e^{\imath\phi_1}-e^{\imath\phi_0}R_1e^{-\imath\phi_1}=0 \;\Rightarrow\; 2\phi_1=\phi_0+2\pi j$ (integer j) $\;\Rightarrow\; \alpha_1=\pm R_1e^{\imath\phi_0/2}$
    2. ρ_2' is a Schur polynomial: Given that $\rho_2(z)=z^2+\alpha_1 z+\alpha_0$ from (503) we have $\rho_2'(z)=2z+\alpha_1$ having a root $-\alpha 1/2$. For ρ_2' to be a Schur polynomial its root $-\alpha 1/2=\mp R_1/2e^{\imath\phi_0/2}$ must have magnitude less than 1. This gives $|\alpha 1/2=R_1/2e^{\imath\phi_0/2}|<1$ that is $R_1<2$.

    Collecting the three boxes in red above we have conditions on α_0 and α_1 need to satisfy so that "condition 2" of theorem 6 holds.

- Collecting conditions 1 and 2 above, we summarize the states of α_0,α_1 such that ρ_2 is a simple von Neumann polynomial,

$\rho_2(z)=z^2+\alpha_1 z+\alpha_0$ is a simple von Neumann polynomial (roots satisfy $(j=1,2)|z_j|\le 1$ and if $|z_j|=1$ z_j is simple) iff

$$\begin{cases}|\alpha_0|<1\\|\alpha_1-\alpha_0\overline{\alpha}_1|\le 1-|\alpha_0|^2\end{cases}\quad\textbf{OR}\quad\begin{cases}|\alpha_0|=1\\\alpha_1=\pm R_1\sqrt{\alpha_0}\quad\text{for real number } 0\le R_1<2\end{cases} \tag{509}$$

- **von Neumann**: Now, we consider conditions that (503) ($\rho_2(z)=z^2+\alpha_1 z+\alpha_0$) is a von Neumann polynomial.
  - That is its roots satisfy $|z|\le 1$ and they can even have multiplicity greater than 1, that is in this case repeated root of 2.
  - The use of von Neumann polynomial condition is for example when linear growth in the form t is allowed as the solution to the central time central space FD scheme applied to wave equation $u_{,tt}-a^2u_{,xx}=0$; *cf.*. §6.4.2 and (480).
  - The only difference that is realized by the application of theorems 7 and 6 applied to von Neumann polynomial and simple von Neumann polynomial applied to second order ρ_2 would be the change of inequality from $R_1<2$ to $R_1\le 2$ in the second item of condition 2 above, given that ρ_2' can be von Neumann polynomial rather than more restrictive Schur polynomial.
  - Thus, the conditions for van Neumann polynomial are,

$\rho_2(z)=z^2+\alpha_1 z+\alpha_0$ is a von Neumann polynomial (roots satisfy $(j=1,2)|z_j|\le 1$ allowing $z_1=z_2, |z_1|=1$) iff

$$\begin{cases}|\alpha_0|<1\\|\alpha_1-\alpha_0\overline{\alpha}_1|\le 1-|\alpha_0|^2\end{cases}\quad\textbf{OR}\quad\begin{cases}|\alpha_0|=1\\\alpha_1=\pm R_1\sqrt{\alpha_0}\quad\text{for real number } 0\le R_1\le 2\end{cases} \tag{510}$$

– The following equation ((511)) summarizes the results for all cases considered,

$\rho_2(z) = z^2 + \alpha_1 z + \alpha_0$ is a Schur polynomial (roots satisfy $|z_{1,2}| < 1$) iff

$$\begin{cases} |\alpha_0| < 1 \\ |\alpha_1 - \alpha_0\overline{\alpha}_1| < 1 - |\alpha_0|^2 \end{cases} \tag{511a}$$

$\rho_2(z) = z^2 + \alpha_1 z + \alpha_0$ is a simple von Neumann polynomial (roots satisfy $(j = 1, 2)|z_j| \leq 1$ and if $|z_j| = 1$ z_j is simple) iff

$$\begin{cases} |\alpha_0| < 1 \\ |\alpha_1 - \alpha_0\overline{\alpha}_1| \leq 1 - |\alpha_0|^2 \end{cases} \textbf{ OR } \begin{cases} |\alpha_0| = 1 \\ \alpha_1 = \pm R_1\sqrt{\alpha_0} \quad \text{for real number } 0 \leq R_1 < 2 \end{cases} \tag{511b}$$

$\rho_2(z) = z^2 + \alpha_1 z + \alpha_0$ is a von Neumann polynomial (roots satisfy $(j = 1, 2)|z_j| \leq 1$ allowing $z_1 = z_2, |z_1| = 1$) iff

$$\begin{cases} |\alpha_0| < 1 \\ |\alpha_1 - \alpha_0\overline{\alpha}_1| \leq 1 - |\alpha_0|^2 \end{cases} \textbf{ OR } \begin{cases} |\alpha_0| = 1 \\ \alpha_1 = \pm R_1\sqrt{\alpha_0} \quad \text{for real number } 0 \leq R_1 \leq 2 \end{cases} \tag{511c}$$

- There are many cases that actually the coefficients in (503) ($\rho_2(z) = z^2 + \alpha_1 z + \alpha_0$), *i.e.*, α_0, α_1 are in fact real numbers.
- In this case it will be easier to obtain reduced form of (511) by restricting α_0, α_1 to be real numbers. That is, $\overline{\alpha}_0 = \alpha_0$ and $\overline{\alpha}_1 = \alpha_1$.
- Simplification of (511) by using real α_0, α_1 results in the following equations:

$\rho_2(z) = z^2 + \alpha_1 z + \alpha_0$ is a Schur polynomial (roots satisfy $|z_{1,2}| < 1$) iff

$$\begin{cases} |\alpha_0| < 1 \\ |\alpha_1| < 1 + \alpha_0 \end{cases} \tag{512a}$$

$\rho_2(z) = z^2 + \alpha_1 z + \alpha_0$ is a simple von Neumann polynomial (roots satisfy $(j = 1, 2)|z_j| \leq 1$ and if $|z_j| = 1$ z_j is simple) iff

$$\begin{cases} |\alpha_0| < 1 \\ |\alpha_1| \leq 1 + \alpha_0 \end{cases} \textbf{ OR } \begin{cases} \alpha_0 = 1 \\ |\alpha_1| < 2 \end{cases} \tag{512b}$$

$\rho_2(z) = z^2 + \alpha_1 z + \alpha_0$ is a von Neumann polynomial (roots satisfy $(j = 1, 2)|z_j| \leq 1$ allowing $z_1 = z_2, |z_1| = 1$) iff

$$\begin{cases} |\alpha_0| < 1 \\ |\alpha_1| \leq 1 + \alpha_0 \end{cases} \textbf{ OR } \begin{cases} \alpha_0 = 1 \\ |\alpha_1| \leq 2 \end{cases} \tag{512c}$$

- In particular we compare (512b) with (362)

$$-1 \leq A_2 \leq 1, \qquad -\frac{A_2 + 1}{2} \leq A_1 \leq \frac{A_2 + 1}{2}, \qquad \textbf{except the point } A_1 = A_2 = 1$$

  which basically stated simple von Neumann polynomial conditions for (361) ($a^2 - 2A_1 a + A_2 = 0$ where).
- The matching of the coefficients between the two different representation of 2^{nd}order polynomials shows $\alpha_1 = -2A_1$ and $\alpha_0 = A_2$.
- It is clear that both equations (512b) with (362) specify the same stability regions (*i.e.*, $|z| < 1$ of $|z| = 1$ for simple roots).
- This region of stability is shown in schematics in §5.3.1.2.

6.5 Well-posedness, robustness, and dynamic stability of physical systems

6.5.1 Introduction to well-posedness, robustness, and dynamic stability

- Consider the following equation,
$$u_{,tt} = u_{,x}$$
- The question may arise why we deal will diffusion equation $u_{,t} = Du_{,xx}$ but not the equation above.
- Or as for another question why in diffusion equation ($u_{,t} = Du_{,xx}$) we have $D \geq 0$.
- The answer to these question is that (6.5.1) and $u_{,t} = Du_{,xx}$ for D are not well-posed.
- Well-posedness is the dependence of the solution to initial data (ICs) in a continuous way.
- Particularly, small errors such as those due to experimental error and interpolation or data should lead to small changes in the solution. This enables use to ignore the small uncertainties in the inputs to the model by knowing that due to well-posedness of the problem such initial noise does not uncontrollably blow up.
- Other concepts that will be discussed are dynamic stability and robustness.
- Dynamic stability requires the solution remain bounded in time and not tend to infinity.
- Dynamic stability is a stronger condition than well-posedness. For example for $u_{,t}+au_{,x} = bu$, $u(x,t) = e^{bt}u_0(x-at)$. Clearly, if $b > 0$ the solution for any t will grow to infinity, but as we will see this equation is still well-posed.
- The last concept is robustness which again is stated for well-posed problems.
- Robustness requires a well-posed PDE to remain well-posed if lower order differential terms are added to it.
- For example, $u_{,tt} + b^2u_{,xxxx} = 0$ (Euler-Bernoulli equation for $b^2 = EI/\rho$) is well-posed but not robust as adding the lower order term $-cu_{,xxx}$ (*i.e.*, $u_{,tt} + b^2u_{,xxxx} - cu_{,xxx} = 0$) makes it ill-posed.
- Robustness is important as we can with confidence assert that some lower order terms in the PDE can be ignored without much asserting the solution. However, for non-robust PDEs (such as Euler-Bernoulli) this becomes difficult.

6.5.2 Well-posedness of dynamic PDEs

- Consider a **scalar one dimensional, linear, temporally first order PDE**,
$$u_{,t} + \mathcal{L}u = 0 \tag{513}$$
where $\mathcal{L}$ is a spatially first order PDE.
- Some examples are,
$$u_{,t} - Du_{,xx} = 0 \qquad \mathcal{L}u = -Du_{,xx} \qquad \text{diffusion equation} \tag{514a}$$
$$u_{,t} + au_{,x} = bu \qquad \mathcal{L}u = au_{,x} - bu \qquad \text{advection reaction equation} \tag{514b}$$

Definition 10 ***Well-posedness / temporally first order PDE****: The initial value problem for a temporally first order equation is well-posed if for each time $t > 0$ there is a* ***constant*** *C_t such that,*
$$||u(.,t)|| \leq C_t||u(.,0)|| \tag{515}$$
holds for ***all*** *initial data $u(.,0)$.*

- The norm $||u(.,t)||$ is the spatial norm at time t. The norm can be an L_1, L_2, L_∞ norms but for convenience and due to the very useful Parseval's relation (413) ($\int_{-\infty}^{\infty} |f(x,t)|^2\,\mathrm{d}x = \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2\mathrm{d}\xi$, that is relating the function and its Fourier transform norms by $||f||_2 = ||\hat{f}||_2$) we greatly simplify the well-posed analysis by using the spatial Fourier transform of the function.
- The spatial L_2 norm is defined as,
$$||u(.,t)||_2 = \sqrt{\int_{-\infty}^{\infty} |u(x,t)|^2\,\mathrm{d}x}$$
- Note that in the definition 10 C_t is **independent** of initial condition $u(x,0) = u_0(x)$. Otherwise, for any given solution for a particular $u_0(x)$ we can find C_t for any given time in (517).

- If the problem is ill-posed (not well-posed) we can find t such that we can arbitrary make the solution at that time large (relative to initial condition) by choosing appropriate IC.
- Before relating well-posedness to propagation of error, we mention that (10) applies to first order PDEs that can even be nonlinear in space. The spatial domain can also be 2D, 3D and the unknown u can be a tensor. All extensions except the linearity are simple but for brevity not pursued here. The linearity basically required for certain discussion below (*e.g.,* propagation of error, Fourier analysis, *etc.*) but not required in the definition.
- Now we relate the definition 10 to propagation of error in an initial value problem.
- Basically, by well-posedness we require the very small errors in the IC, *e.g.,* noise, inaccuracy in experimental measurements, *etc.* **do not** affect the solution at later times in an unbounded manner.
- Assume if we deal with a problem that is not physically well-posed. This means that even the tiniest errors in measuring initial conditions can drastically change the solution at later times. That is, we can have no confidence in the system response and basically the problem is physically ill-posed / unstable.
- For a temporally first order and linear PDE the previous concept of propagation of error can be represented in the following form,

$$||u^1(.,t) - u^2(.,t)|| \leq C_t ||u^1(.,0) - u^2(.,0)|| \tag{516}$$

- Note that (516) says that the jump of the solutions u^1 and u^2 is bounded by C_t times of the jump of their corresponding initial conditions.
- That is, if we have the error bound of ϵ in measuring ICs we have the confidence that the error at time t is bounded by $C_t\epsilon$.
- Equation (516) can easily be derived from (517) by knowing that if u_1 and u_2 are both solution to a linear system so would be $u_1 - u_2$ (*i.e.,* $u_{1_{,t}} + \mathcal{L}u_1 = 0$, $u_{2_{,t}} + \mathcal{L}u_2 = 0$, *cf.* (531), $\Rightarrow \Delta u_{,t} + \mathcal{L}\Delta u = 0$ for $\Delta u = u_1 - u_2$ due to linearity of $\mathcal{L}$ and Δu can be plugged in (517).
- That is, if the PDE is linear we can express (516) in the following form,

$$||u(.,t)|| \leq C_t ||u(.,0)||$$

  which is the form that we adopted for for well-posedness of a temporally first order PDE in (6.5.2).
- One the first look, given that C_t can grow in time, it seems that it is trivial to find C_t based on the exact solution such that (6.5.2) is satisfied for a given IC.
- However, for ill-posed problems we can find high enough frequency initial conditions that for any proposed C_t for a given time, condition (6.5.2) does not hold.
- Well-posedness from definition 10 can be extended to second order or higher temporally order PDEs.

Definition 11 ***Well-posedness / temporally second order PDE****: The initial value problem for a temporally second order equation is well-posed if for each time $t > 0$ there is a* ***constant*** *C_t such that,*

$$||u(.,t)||_{H^p} + ||u_{,t}(.,t)||_{H^0} \leq C_t \left(||u(.,0)||_{H^p} + ||u_{,t}(.,0)||_{H^0} \right) \tag{517}$$

holds for ***all*** *initial data $u(.,0), u_{,t}(.,0)$ and $2p$ is the maximum differential order in space.*

where $||u_{,t}(.,0)||_{H^p}$ are Hilbert norms defined as

$$||f(x)||_{H^p} = \sqrt{\int_{-\infty}^{\infty} \left\{ |f(x)|^2 + |f_{,x}(x)|^2 + \cdots + \left| \frac{\partial^p f}{\partial x^p} \right|^2 \right\}} \tag{518a}$$

for example $||f(x)||_{H^0}$ is simply the L2 norm $||f(x)||_{L_2}$.

- The mathematical form of the norms, *etc.* in 11 are not as much as to emphasize and different forms can be found in the literature (*i.e.,* different norms being used).
- However, the main points is that for a temporally second order ODE since the IC is comprised of both values of $u(x,0)$ and $u_{,t}(x,0)$ both appear on the RHS.
- A general form linear second order scalar spatially 1D PDE can be written as,

$$u_{,tt} + \omega_0(x,t)u_{,t} + \mathcal{L}u = 0 \tag{519}$$

- For this linear PDE we can express the definition 11 as an equation in the form (516) ($||u^1(.,t) - u^2(.,t)|| \leq C_t||u^1(.,0) - u^2(.,0)||$) but this time with terms of the form $||u^1_{,t}(.,0) - u^2_{,t}(.,0)||$ and $||u^1_{,t}(.,t) - u^2_{,t}(.,t)||$ appearing on the RHS and LHS respectively (clearly difference terms will have the same norms applied to them as those appearing in the underlying well-posedness definition; *cf.* definition 11.
- Definition of well-posedness for higher temporal orders of PDE and in 2D, 3D can be followed similarly.
- For higher temporal orders higher temporal derivative terms will appear on both sides of the inequality.
- Direct proof of well-posedness of a PDE may be difficult as in the definition the IC is arbitrary.
- The next theorems make it much easier to investigate the well-posedness of a PDE.

6.5.3 Theorems for verifying the well-posedness of PDEs

- The well-posedness analysis of a linear PDE uses the Fourier transform of the PDE in a similar fashion that we analyzed the stability of numerical methods by using the Fourier transform of the solution.
- The idea is very simple: **If we understand the response of a system to simple harmonic ICs we can deduce its behavior for general IC given the linearity of an underlying PDE**.
- The use of L2 norm and Parseval's relation ensures that we can go back and forth between the norms of the Fourier transform of the solution and its inverse (*i.e.,* the PDE solution).
- For example consider the temporally first order and linear PDE (531),

$$u_{,t} + \mathcal{L}u = 0$$

- After the application of Fourier transformation for space variable, we cast the PDE in the form of an ODE for any given wavenumber ξ.
- Recalling the Fourier transform of $u(x,t)$ and its inverse from (414),

$$\begin{aligned} \hat{u}(\xi,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u(x,t)e^{-\imath\xi x}\,\mathrm{d}x \qquad \Leftrightarrow \\ u(x,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{u}(\xi,t)e^{\imath\xi x}\mathrm{d}\xi \end{aligned}$$

- and that (spatial) Fourier transform changes (spatial) differentiation to algebraic multiplication, that is,

$$\widehat{\left(\frac{\mathrm{d}^n u}{\mathrm{d}x^n}\right)}(\xi,t) = (\imath\xi)^n\hat{u}(\xi,t) \tag{520}$$

- we cast PDE (531) $u_{,t} + \mathcal{L}u = 0$ in the form of an ODE in time for a given wavenumber ξ,

$$\hat{u}_{,t}(\xi,t) = \omega(\xi)\hat{u}(\xi,t) \tag{521}$$

where $\omega(\omega)$ is the **algebraic terms** generated by the application of **Fourier transform on the spatial derivative operator** $\mathcal{L}$ in (531).

- The solution of the ODE (522) is very simple and is given by,

$$\boxed{\hat{u}(\xi,t) = e^{\omega(\xi)t}\hat{u}(\xi,0) = e^{\omega(\xi)t}\hat{u}_0(\xi)} \tag{522}$$

where $\hat{u}_0(\xi) = \hat{u}(\xi,0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u_0(x)e^{-\imath\xi x}\,\mathrm{d}x$ is the Fourier transform of the IC $u_0(x) = u(x,0)$.

- The equation (522) is very informative and basically asserts which what rate each harmonic wave with wavenumber ξ exponentially grows in the form $e^{\omega(\xi)t}$.
- Clearly, it is the real part of $e^{\omega(\xi)t}$ that determines the growth of the solution for a wavenumber ξ as,

$$|e^{\omega(\xi)t}| = |e^{(\mathrm{Re}\,\omega(\xi)t + \imath\mathrm{Im}\,\omega(\xi)t)}| = |e^{\mathrm{Re}\,\omega(\xi)t}e^{\imath\mathrm{Im}\,\omega(\xi)t}| = |e^{\mathrm{Re}\,\omega(\xi)t}||e^{\imath\mathrm{Im}\,\omega(\xi)t}| = e^{\mathrm{Re}\,\omega(\xi)t} \tag{523}$$

where $\mathrm{Re}\,\omega(\xi)$ and $\mathrm{Im}\,\omega(\xi)$ are real and imaginary parts of $\omega(\xi)$. That is, $\omega(\xi) = \mathrm{Re}\,\omega(\xi) + \imath\mathrm{Im}\,\omega(\xi)$.

- If $Re\ \omega(\xi) \leq 0$ for **all** ξ then we expect the solution for all simple harmonic waves with wavenumber (spatial frequency) ξ not to grow.
- Having assumed the linearity of $\mathcal{L}$ we expect for the solution $u(x,t)$ for the underlying PDE be non-growing for any arbitrary IC as the IC can be written as a linear combination of simple harmonic waves (by Fourier transform) and the solution in time is the summation of the solution of these simple harmonic waves (due to the linearity of the PDE).
- In fact, well-posedness requires a weaker condition than $Re\ \omega(\xi) \leq 0$ and only requires there to be bound $\bar{\omega}$ **INDEPDENDENT of** ξ.
- $\bar{\omega}$ can be a positive number in which the solution in fact can grow in time but there will be a C_t (possibly growing) in (517) ($||u(.,t)|| \leq C_t||u(.,0)||$) that is **independent** of IC.
- The formal form of this assessment on the relation between $\mathrm{Re}\ \omega(\xi)$ and well-posedness is as follows,

Theorem 8 *Well-posedness for linear and temporally first order PDEs The necessary and sufficient condition for well-posedness of a linear and temporally first order PDE from* (517) *is that there exists a real* ***constant*** $\bar{\omega}$ *such that*

$$\mathrm{Re}\ \omega(\xi) \leq \bar{\omega} \qquad \textit{for all} \qquad \xi \tag{524}$$

where $\mathrm{Re}\ \omega(\xi)$ *is the real part of the exponent in the solution of the Fourier transform* $\hat{u}(\xi,t)$ *of the solution* $u(x,t0$ *in* (522).

Proof:

- If the function $\omega(\xi)$ satisfies (524) for some $\bar{\omega}$ then from, (522) and (523) we have,

$$|\hat{u}(\xi,t)| = |e^{\omega(\xi)t}||\hat{u}_0(\xi)| = e^{\mathrm{Re}\ \omega(\xi)t}|\hat{u}_0(\xi)| \leq e^{\bar{\omega}t}|\hat{u}_0(\xi)| \qquad \Rightarrow \tag{525}$$

$$||\hat{u}(.,t)|| \leq e^{\bar{\omega}t}||\hat{u}_0(\xi)|| \tag{526}$$

  where $||\hat{u}(.,t)||$ and $||\hat{u}_0(\xi)||$ are the L2 norms of the Fourier transform at time t and at initial time, that is for the IC.
- Now, using the Parseval's relation ($||\hat{u}(.,t)|| = ||u(.,t)||$ for all t) we conclude,

$$||u(.,t)|| \leq e^{\bar{\omega}t}||u_0(\xi)||$$

  which is basically the proof of well-posedness in (6.5.2) $||u(.,t)|| \leq C_t||u(.,0)||$ for $C_t = e^{\bar{\omega}t}$ which clearly is independent of IC.
- The proof of the inverse is a bit more involved and can be skipped. However, the form of the solution that corresponds to ill-posedness is informative on what types of initial conditions result in ill-posedness.
- To prove the inverse, we want to show that for any arbitrary C there are some IC's $u_0(x) = u(x,0)$ such that $||u(.,t)|| > C||u(.,0)||$. That is, there exists no IC independent bound C in (6.5.2) $||u(.,t)|| \leq C||u(.,0)||$.
- To prove this for a given time $t > 0$ we choose $\bar{\omega} = 2(\log C)/t$. Since (524) is not satisfied, we can find a wavenumber ξ_0 such that $\mathrm{Re}\ \omega(\xi) > \bar{\omega} = 2(\log C)/t$.
- Then by continuity of $\mathrm{Re}\ \omega(\xi)$ (recall $\omega(\xi)$ is a polynomial) we can find a neighborhood $\xi \in [\xi_0 - h/2, \xi_0 + h/2]$ such that,

$$\mathrm{Re}\ \omega(\xi) > (\log C)/t \qquad \text{for} \qquad \xi \in [\xi_0 - h/2, \xi_0 + h/2] \tag{527}$$

- Now we choose the Fourier transform of the initial condition as,

$$\hat{u}_0(\xi) = \begin{cases} \frac{1}{\sqrt{h}} & \xi_0 - h/2 \leq \xi \leq \xi_0 + h/2 \\ 0 & \text{otherwise} \end{cases} \tag{528}$$

- Now using Parseval's relation (for $t = 0$) we have,

$$||u_0(.)|| = ||u(.,0)|| = ||\hat{u}(.,0)|| = ||\hat{u}_0(.)|| = \sqrt{\int_{-\infty}^{\infty} |\hat{u}_0(\xi)|^2 \mathrm{d}\xi} = \sqrt{\int_{\xi_0 - h/2}^{\xi_0 + h/2} \left(\frac{1}{\sqrt{h}}\right)^2 \mathrm{d}\xi} = 1 \tag{529}$$

- On the other hand for time t by using (522), (527) the range of $\hat{u}_0(\xi)$ in (528) and again using Parseval's relation we have,

$$||u(.,t)|| = ||\hat{u}(.,t)|| = \sqrt{\int_{-\infty}^{\infty} |\hat{u}(\xi,t)|^2 \mathrm{d}\xi} = \sqrt{\int_{-\infty}^{\infty} |e^{\omega(\xi)t}\hat{u}_0(\xi)|^2 \mathrm{d}\xi} = \sqrt{\int_{-\xi_0-h/2}^{\xi_0+h/2} |e^{\omega(\xi)t}\hat{u}_0(\xi)|^2 \mathrm{d}\xi}$$
$$\geq \sqrt{\int_{-\xi_0-h/2}^{\xi_0+h/2} e^{\{2(\log C/t)t\}} |\hat{u}_0(\xi)|^2 \mathrm{d}\xi} = C\sqrt{\int_{-\xi_0-h/2}^{\xi_0+h/2} |\hat{u}_0(\xi)|^2 \mathrm{d}\xi} = C\sqrt{\int_{-\infty}^{\infty} |\hat{u}_0(\xi)|^2 \mathrm{d}\xi} = C||\hat{u}_0(.)|| = C \quad \text{from (529)} \quad (530)$$

- From (529) we have $||u_0(.)|| = 1$ and from (530) $||u(.,t)|| = C$. Given, that C was an arbitrarily large number we have,

$$\forall t > 0 : \forall C > 0 \exists\ u_0(x) \text{ (IC) such that } ||u(.,t)|| > C||u_0(.)||$$

  so from (6.5.2) the PDE is not well-posed.

- Generally, the high wavenumber modes grow the fastest in ill-conditioned PDEs.

- Now from (524) we observe that a temporally first order PDE is stable if the real part of $\omega(\xi)$ is bounded from above.
- How about higher temporal PDEs?
- The analysis for temporally higher order PDEs is very similar.
- In such cases we again obtain ODEs in time similar to (522) ($\hat{u}_{,t}(\xi,t) = \omega(\xi)\hat{u}(\xi,t)$) for temporally first order PDE (531) $u_{,t} + \mathcal{L}u = 0$.
- For example, consider,

$$u_{,tt} + \mathcal{L}u = 0 \quad (531)$$

- The (spatial) Fourier transformation of this PDE yields,

$$\hat{u}_{,tt}(\xi,t) = \omega(\xi)\hat{u}(\xi,t) \quad (532)$$

  whose solution is of the form

$$e^{\pm\sqrt{\omega(\xi)}t}\psi_{\pm}(\xi) \quad (533)$$

  where $\psi_{\pm}(\xi)$ are obtained by the initial conditions, *i.e.,* $\hat{u}(\xi, t=0) = \hat{u}_0(\xi)$ and $\dot{\hat{u}}(\xi, t=0) = \dot{\hat{u}}_0(\xi)$.

- In general, for a temporally higher order PDE, the solution to a harmonic waves with wavenumber ξ is the summation of the solutions of the form

$$P_l(t)e^{\omega_l(\xi)t}\psi_l(\xi)e^{\imath\xi x} \quad (534)$$

  where $e^{\imath\xi x}$ corresponds to the spatial harmonic wave, $\psi_l(\xi)$ are obtained from ICs, and $P_l(t)e^{\omega_l(\xi)t}$ corresponds to the ODEs in time for the given ξ (*e.g., cf.* (532) (533)).

- The polynomial in time $P_l(t)$ is a constant when the root $\omega_l(\xi)$ is simple and for roots of multiplicity r it is a polynomial of order $r-1$.
- Now, the question is how we can decide whether the PDE is well-posed by investigate its solution for simple harmonic waves from (534).
- The answer is similar to the theorem 8 for temporally first order PDEs where the real part of $\omega(\xi)$ was bounded from above by a constant independent of the wave number.
- This can be express as,

Theorem 9 *Well-posedness for linear PDEs (first and higher temporal orders) The necessary and sufficient condition for well-posedness of a linear PDE with harmonic solutions of the form (534) ($P_l(t)e^{\omega_l(\xi)t}\psi_l(\xi)e^{\imath\xi x}$) is that there exists a real* ***constant*** *$\bar{\omega}$ such that*

$$\mathrm{Re}\ \omega_l(\xi) \leq \bar{\omega} \qquad \textit{for all } l \textit{ and } \xi \quad (535)$$

That is, again $\mathrm{Re}\ \omega_l(\xi)$ *are bounded from above* ***independent from*** ξ.

- Interestingly the polynomials, $P_l(t)$ if not constant correspond to algebraic (weak) growth of the solution (when $\mathrm{Re}\,\omega_l(\xi) \geq 0$), do not affect the well-posedness of the PDE as their growth is independent of ξ.
- In the next section we provide a few examples on well-posed and ill-posed problems.

6.5.4 Examples of well-posedness and ill-posed PDEs

Example 6 *Well-posedness of diffusion equation: The influence of the sign of diffusion coefficient on well-posedness of diffusion equation is investigated.*

- Consider the diffusion equation,
$$u_{,t} - Du_{,xx} = 0 \tag{536}$$
for constant D.
- D is the diffusion coefficient, *e.g.,* conductivity divided by heat capacity in Fourier heat equation.
- The term $-Du_{,xx} = (-Du_{,x})_{,x}$ (D is assumed to be constant) is the gradient of the spatial flux.
- That is the spatial flux is $-Du_{,x}$. If $D > 0$ the flux is from higher u value to lower u value. This is physically expected; for example heat flux from higher temperature to lower temperature, diffusion of contaminant from high concentration to low concentration, *etc.*.
- Now what happens if $D < 0$? The PDE becomes **ill-posed**.
- To demonstrate this, plug a harmonic solution in the form $e^{\omega(\xi)t}e^{\imath\xi x}$ in the PDE (which is the same as solving the ODEs resulting from Fourier transformation of the PDE for a given wavenumber ξ),
$$\left[\omega(\xi) - (\imath\xi)^2 D\right] e^{\omega(\xi)t}e^{\imath\xi x} = 0 \quad \Rightarrow \quad \omega(\xi) = -D\xi^2, \quad \text{that is} \quad \boxed{Re\,\omega(\xi) = -D\xi^2, \quad Im\,\omega(\xi) = 0} \tag{537}$$
- Now based on the sign of D we have two cases,
  1. $D \geq 0$: Then $Re\,\omega(\xi)$ is bounded from above by 0 which clearly is independent of ξ so the equation is well posed.
     In fact, we observe the higher wavenumber the faster it decays in time $e^{Re\,\omega(\xi)t} = e^{-D\xi^2 t}$ which is a testament of the very strong smoothening of diffusion equation and the annihilation of high frequency content.
  2. $D < 0$: Clearly in this case $Re\,\omega(\xi) = -D\xi^2$ is not bounded from above independent from ξ and it can be made arbitrary large by choosing high spatial frequency (wavenumber) ξ modes.
     Diffusion equation with a negative diffusion coefficient is ill-posed. This is physically expected. For example if the heat flux is from low to high temperature energy of all low temperature regions flow toward high temperature zones! In this problem, very small initial noises can grow with no bounds for any time.

Example 7 *Well-posedness of $u_{,tt} - u_{,x} = 0$: This equation is classified as parabolic if we compute the discriminant of the equation. Recall that for $Au_{,xx} + Bu_{,xt} + Cu_{,tt}$ + lower order terms is parabolic if $B^2 - AC = 0$ which is zero in this case. However, more appropriate definition of hyperbolicity and parabolicity in the context of dynamical system (cf. [Strikwerda, 2004] section 9.2 for these definitions of hyperbolicity and parabolicity) will not classify this PDE as either one as this it is in fact ill-posed.*

- The equation considered in this example is,
$$u_{,tt} - u_{,x} = 0 \tag{538}$$
- To demonstrate ill-posedness of this PDE, we plug harmonic solution $e^{\omega(\xi)t}e^{\imath\xi x}$ in the PDE.
- This is equivalent to Fourier analysis from theorem 9, if $\omega(\xi)$ is a repeated root, the polynomials $P_l(t)$ with order greater than 0 will appear. We also are not interested in the functions $e^{\imath\xi x}$ in (534) ($P_l(t)e^{\omega_l(\xi)t}\psi_l(\xi)e^{\imath\xi x}$) which depend on ICs and do not affect the well-posedness analysis of the PDE. So, by plugging $e^{\omega(\xi)t}e^{\imath\xi x}$ in (538) we obtain,
$$\left[\omega^2(\xi) - \imath\xi\right] e^{\omega(\xi)t}e^{\imath\xi x} = 0, \quad \Rightarrow \quad \omega^2(\xi) = \imath\xi \quad \Rightarrow \quad \omega(\xi) = \pm\left(\frac{\sqrt{2}}{2} + \imath\frac{\sqrt{2}}{2}\right)\xi \quad \Rightarrow \quad Re\,\omega(\xi) = \pm\frac{\sqrt{2}}{2}\xi \tag{539}$$
- So we observe that one of the roots has unbounded (in wavenumber ξ) real part $Re\,\omega(\xi) = \frac{\sqrt{2}}{2}\xi \Rightarrow$
- (538) $u_{,tt} - u_{,x} = 0$ is ill-posed.
- It is interesting to note that just by looking at $u_{,tt} - u_{,x} = 0$ from a parabolic type point of view and thinking that it should be well-posed with vanishing solution as diffusion equation from example (6) is wrong. Even though the signs of $u_{,tt}$ and $u_{,x}$ in the equation are opposite the interchange of x and t derivative makes the initial value problem ill-posed.

Example 8 *Well-posedness of $\frac{\partial^n u}{\partial t^n} - \mathcal{L}u = 0$, $n > 2$: This equation is ill-posed. It will have roots of the form,*

- Motivated from previous example, we consider the equation,
$$\frac{\partial^n u}{\partial t^n} - \mathcal{L}u = 0 \tag{540}$$

- Similar to (522) the linear spatial differential operator gives a term in the form $\omega(\xi)$ and the solution to a harmonic wave $e^{\omega_l(\xi)t}e^{\imath\xi x}$ is,

$$\omega_l = \sqrt[n]{\omega(\xi)}, \qquad \text{root number } l \text{ out of } n \text{ roots} \tag{541}$$

- Given that in complex plane the n^{th} root of a number are separated by polar angle $2\pi/n$ one of the roots will fall in the positive real side of the complex when $n > 2$.
- Since the roots are dependent and unbounded on ξ ($\omega(\xi)$ is a polynomial in ξ) those roots in the positive real side of the complex plan will not be bounded independent of ξ and the PDE (540) $\frac{\partial^n u}{\partial t^n} - \mathcal{L}u = 0$ is ill-posed.

Example 9 *The advection reaction equation $u_{,t} + au_{,x} = -bu$ is well-posed independent of the sign of b (and a).*

- Consider the advection-reaction equation,

$$u_{,t} + au_{,x} = -bu \tag{542}$$

- b is the reaction coefficient and physically it is generally positive signifying a diminishing solution.
- To study the well-posedness of the PDE we plug a harmonic solution in the form $e^{\omega(\xi)t}e^{\imath\xi x}$ in the PDE,

$$\left[\omega(\xi) + a\imath\xi + b\right] e^{\omega(\xi)t}e^{\imath\xi x} = 0 \quad \Rightarrow \quad \omega(\xi) = b - a\imath\xi \quad \Rightarrow \quad \boxed{Re\ \omega(\xi) = b, \quad Im\ \omega(\xi) = -a\imath\xi} \tag{543}$$

- Interestingly $Re\ \omega(\xi)$ is bounded by b which is independent of ξ meaning that the **advection-reaction equation is well-posed irrespective of the sign of b**.
- Now, whether the solution will grow or not depends on the sign of b given that the solution to the PDE is $u(x,t) = e^{-bt}u_0(x-at)$,
  1. $b \geq 0$: In this case the solution does not tend to infinity in time. In fact, for $b > 0$ it vanishes.
  2. $b < 0$: Solution grows to infinity but the main point is for any given time t **irrespective of the IC** the solution is bounded by e^{-bt} times the IC. This is exactly the stability condition from (517) $||u(.,t)|| \leq C_t||u(.,0)||$ for $C_t = e^{-bt}$ confirm our analysis that irrespective of the sign of b the PDE is well-posed.
- Later in §6.5.7 we further discuss the concept of **dynamical stability**. The reaction equation is dynamically stable (not tending to infinity) when $b \geq 0$ and dynamically unstable for $b < 0$.

Example 10 *Well-posedness of Wave $u_{,tt} - a^2u_{,xx} = 0$ and Euler-Bernoulli $u_{,tt} + b^2u_{,xxxx} = 0$ equations: Both of these equations are well-posed.*

- Consider the following two equations,

$$u_{tt} - a^2u_{,xx} = 0 \qquad \text{Wave equation} \tag{544a}$$
$$u_{tt} + b^2u_{,xxxx} = 0 \qquad \text{Euler-Bernoulli equation} \tag{544b}$$

- In the context of solid mechanics the equation of a 1D elasticity is modeled with (544a) where $a = \sqrt{E/\rho}$ and that of a beam with (545) where $b = \sqrt{EI/A\rho}$.
- To study the well-posedness of the PDE we plug a harmonic solution in the form $e^{\omega(\xi)t}e^{\imath\xi x}$ in the PDEs,

$$\left[\omega^2(\xi) - a^2(\imath\xi)^2\right] e^{\omega(\xi)t}e^{\imath\xi x} = 0 \quad \Rightarrow \quad \omega^2(\xi) + a^2\xi^2 = 0 \quad \omega(\xi) = \pm\imath a\xi, \quad \text{that is}$$
$$Re\ \omega(\xi) = 0, Im\ \omega(\xi) = \pm a\xi \quad \text{wave equation} \tag{545a}$$
$$\left[\omega^2(\xi) + b^2(\imath\xi)^4\right] e^{\omega(\xi)t}e^{\imath\xi x} = 0 \quad \Rightarrow \quad \omega^2(\xi) + b^2\xi^4 = 0 \quad \omega(\xi) = \pm\imath b\xi^2, \quad \text{that is}$$
$$Re\ \omega(\xi) = 0, Im\ \omega(\xi) = \pm b\xi^2 \quad \text{Euler-Bernoulli equation} \tag{545b}$$

- We observe that for both equations $Re\ \omega(\xi) = 0$, *i.e.,* being bounded by ξ independent value 0.
- In fact, since $Re\omega(\xi) = 0$ ant the roots are not repeated the solution is not growing nor decaying. These PDEs are conservative.
- In §6.5.5 we show that the Euler-Bernoulli equation is not robust.

6.5.5 Robustness of PDEs

- An important property of a PDE is that the qualitative response of the solution is unaffected by the **addition of or changes in lower order terms** and by sufficiently small changes in the coefficients [Strikwerda, 2004].
- This property is called robustness.
- Almost all derivation of equations to model physical processes make some assumptions that certain effects are not important to understand the physical process being studied [Strikwerda, 2004].
- Statements such as,
  - "assume the temperature of the body is constant"
  - "we may ignore gravitational forces"
  - "consider a homogeneous body"

  can be made because it is assumed that small variations in some quantities may be ignored without affecting conclusions of the analysis. That is the system is robust.
- Robustness is important when we consider numerical results:
  - FD method and other numerical methods may be regarded as perturbations, or approximation, of the equations **similar to modification of the equations by adding lower order terms**.
  - If the numerical method is **not robust** then the **construction of numerical methods will be more difficult**.
- Now as one example we should that the well-posed Euler-Bernoulli equation is not robust.

Example 11 *The Euler-Bernoulli equation $u_{,tt} + b^2 u_{,xxxx} = 0$ is not robust.*

- In Example 10 we observed that both wave equation and Euler-Bernoulli equations were well-posed and in fact conservative.
- Now, for the Euler-Bernoulli equation (545) ($u_{tt} + b^2 u_{,xxxx} = 0$) we consider the addition of a lower order term $-cu_{,xxx}$,

$$u_{,tt} + b^2 \ -cu_{,xxx} = 0 \tag{546}$$

- To study the well-posedness of the altered PDE (*i.e.*, robustness relative to the addition $-cu_{,xxx}$) we we plug a harmonic solution in the form $e^{\omega(\xi)t}e^{\imath\xi x}$ in (546),

$$\left[\omega^2(\xi) + b^2(\imath\xi)^4 - c(\imath\xi)^3\right] e^{\omega(\xi)t}e^{\imath\xi x} = 0 \qquad \Rightarrow \qquad \omega^2(\xi) + b^2\xi^4 + \ \imath c\xi^3 = 0 \tag{547}$$

so

$$\omega(\xi) = \pm\sqrt{-b^2\xi^4 - \imath c\xi^3} = \pm\imath b\xi^2\sqrt{1 + \frac{\imath c}{b^2\xi}} = \pm\imath b\xi^2\left[1 + \frac{1}{2}\frac{\imath c}{b^2\xi} + \mathcal{O}\left(\xi^{-2}\right)\right] \qquad \Rightarrow$$

$$\omega(\xi) = \pm\left[\imath b\xi^2 - \frac{c\xi}{2b} + \mathcal{O}(1)\right] \tag{548}$$

- So, the real part of $\omega(\xi)$ grows as $\mp\frac{c\xi}{2b}$ and for positive of negative c there will be no bound on $\mathrm{Re}\,\omega(\xi)$.
- So the Euler-Bernoulli equation is not robust as the addition of lower order differential terms make it ill-posed.
- On the other hand, the wave equation is robust; *cf.* [Strikwerda, 2004] Exercise 9.1.3.
- Now, this lack of robustness requires special attention in numerical solution of Euler-Bernoulli equation as numerical discretization may unintentionally "effectively" add such lower order differentiation terms that make the equation ill-posed.

6.5.6 Dynamic stability

- Let us recall the well-posedness analysis for the advection reaction equation (542),

$$u_{,t} + a u_{,x} = -bu$$

where we observed the solution for temporal frequency $\omega(\xi)$ for harmonic solutions $e^{\omega(\xi)t}e^{\imath\xi x}$ was,

$$Re\ \omega(\xi) = b, \quad Im\ \omega(\xi) = -a\imath\xi$$

- Given that $Re\ \omega(\xi) = b$ is bounded from above (by b) independent of ξ this equation is well-posed.
- However, depending on the sign of b is can be dynamically stable or not.
- Clearly, for $b < 0$ in $e^{\omega(\xi)t}e^{\imath\xi x} = e^{Re\ \omega(\xi)t}e^{Im\ \omega(\xi)t}$ is growing in time for all ξ as,

$$e^{Re\ \omega(\xi)t} = e^{-bt} \to \infty \quad \text{as} \quad t \to \infty$$

- This behavior can also be observed from the analytical solution of advection-reaction equation (6.5.6),

$$\left.\begin{array}{ll} u_{,t} + a u_{,x} = -bu & \text{PDE} \\ u(x, t=0) = u_0(x) & \text{IC} \end{array}\right\} \quad \Rightarrow \quad u(x,t) = e^{-bt}u_0(x) \quad \Rightarrow \quad \boxed{||u(.,t)|| = e^{-bt}||u_0||} \tag{549}$$

- Observations from the analytical solution agree with the well-posedness of the method verified by the analysis of harmonic waves,
  - For well-posedness from we required (517) $||u(.,t)|| \leq C_t||u(.,0)||$ we require a time-dependent, **initial condition (u_0) INDEPENDENT** constant C_t to exist bounding the norm at time t with the norm at time 0.
  - For the advection reaction equation this constant is $C_t = e^{-bt}$ which in fact grows in time for $b < 0$.
- Basically, regardless of the sign of b advection-reaction equation is well-posed because for any given time their a fixed time-dependent value ($C_t = e^{-bt}$) that limits the growth of the solution.
- This is unlike ill-posed problems $u_{,tt} - u_x = 0$ and $u_{,t} - Du_{,xx} = 0, D < 0$ that for any given time we could find ICs (which had high wavenumber content) that for any C $||u(.,t)|| > C||u_0||$.
- Still, within the class of well-posed problems we define the concept of **dynamic stability**,

$$\text{PDE is } \textbf{dynamically stable iff} \qquad \exists C \ \text{ such that } \ \forall t \ ||u(.,t)|| < C||u_0|| \tag{550a}$$

$$\text{PDE is } \textbf{dynamically unstable iff} \qquad ||u(.,t)|| \to \infty \quad \text{as} \quad t \to \infty \tag{550b}$$

basically,

  - Dynamic stability refer to the property that the solution does not blow up **in the limit of infinite time**.
  - that is, small variations from a reference state will decay, or at least not grow with time (using the linearity of a linear PDE and (550a)).
  - Note that dynamic stability deals with the limit of the solution norm as $t \to \infty$ while well-posedness for any finite time t (that the solution norm for any given time can be bounded by possibly time dependent but IC independent C_t).
  - Clearly, a **dynamically stable method is well-posed but not vice versa**,

$$\text{Dynamically stable} \quad \Rightarrow \quad \exists C \text{ such that } \forall t \ \ ||u(.,t)|| < C||u_0|| \quad \Rightarrow \tag{551a}$$

$$\forall t \ \ \exists C_t \text{ such that } ||u(.,t)|| < C_t||u_0|| \quad \text{where } C_t \text{ can be } \textbf{uniformly (independently from } t\textbf{)} \text{ be chosen as } C \tag{551b}$$

the inverse (a dynamically unstable but well-posed) example is the advection-reaction equation $u_{,t} + a u_{,x} = -bu$ with $b < 0$.

6.5.7 Numerical stability versus dynamic stability and well-posedness

- Now that we have discussed dynamic stability the question arises what is its relation with numerical stability.

- As an example recall the numerical stability for the FD implementation of a one-step ($J = 0$) temporally first order PDE (409) was,

$$||v^n||_h \leq C_T^* ||v^0||_h \tag{552}$$

where the discrete norm (from (408)) is defined as $||u^t|| = \sqrt{\int_\infty^\infty |u(x,t)|^2 \, \mathrm{d}x}$.

- We observe that the **numerical stability condition is basically a numerical counterpart of well-posedness definition** (compare C_T* in (552) in the definition with C_t in well-posedness definition (517) ($||u(.,t)|| \leq C_t ||u(.,0)||$).

- Following this we make the following remarks about numerical and dynamic stability and well-posedness:

- Numerical stability is only relevant to well-posed problems because if the underlying PDE is ill-posed there should not be any hope / expectation that the numerical method will make it numerically stable (*i.e.,* well-posed by having C_T^* in (552)).
- **Numerical methods should be able to solve well-posed PDE which may not be dynamically stable**. In this case, the numerical method's stability, (552), ensures that the **well-posedness of the PDE is preserved numerically** (*i.e.,* C_T^* in (552) exists as a numerical counterpart to C_t in the definition of well-posedness (517) ($||u(.,t)|| \leq C_t ||u(.,0)||$).
- For example, a good numerical scheme should be numerical stable to solve even a dynamically unstable PDE such as the advection-reaction equation (542) ($u_{,t} + a u_{,x} = -bu$) with $b < 0$. This was in fact demonstrated to be the case in the example 5 where the Lax-Friedrichs method with normalized time step $\bar{k} = ak/h \leq 1$ was successfully able to solve (456) $u_{,t} + a u_{,x} - u = 0$ (*i.e.,* advection-reaction equation with $b = -1$).
- In fact, the more relaxed stability constraint for the amplification factor (432) $|g(\theta, k, h)| \leq 1 + Kk$ (which permits a growth of the form Kk compared to the bounded stability condition (433) $|g(\theta, k, h)| \leq 1$ is directly relevant to cases when the underlying physical solution is dynamically unstable (and off course still well-posed). As an example, it was the stability condition $|g(\theta, k, h)| \leq 1 + Kk$ that was used for the analysis of the stability of Lax-Friedrichs method for the solution of $u_{,t} + a u_{,x} - u = 0$ in the example 5.
- **A numerical method is unstable if it does not honor the well-posedness of a PDE.** That is, when applied to a well-posed problem it cannot ensure that there exists a C_T^* in (552) that limits numerical solution at time t by IC at time 0; an **unstable numerical method makes the solution of a physically well-posed problem "numerically ill-posed"** by letting often high frequency content of the IC and prior solution (which are inevitable due to round-off and discretization errors) to grow without bound in time. .
- As an example, for the same $u_{,t} + a u_{,x} - u = 0$ discussed above if $\bar{k} > 1$ is used in the Lax-Friedrichs method the numerical method is unstable and for later times of the solution as $||u(.,t)||$ grows with no bounds with the existence of no C_T^* (552) that works for all IC and **most importantly all** h . Interestingly, if CFL condition is violated ($\bar{k} = ak/h > 1$) yet we let $h \to 0$ we often have worse blow up of the solution at later! This is due to the fact that for small h (and $\bar{k} > 1$ fixed) there are more time steps to get to a given time t and more instances that the solution numerically (and non-physically) grows.
- As two final remarks regarding the relation between numerical and dynamic stabilities we note,
  - Numerical stability (similar to well-posedness of the underlying PDE) deals with the response of the system at finite times t rather than the limiting response of the solution at $t \to \infty$ in the the definition of dynamic stability.
  - If a numerically unstable method is applied to a dynamically stable PDE the solutions will not be convergent (*i.e.,* for an unstable $\bar{k}$ letting $h \to 0$) and bounded in time (as $h \to 0$). Clearly, the same can be stated when a unstable numerical scheme is applied to a well-posed by dynamically unstable PDE.
- In conclusion, A stable numerical method is able to preserve well-posedness of an underlying PDE whether or not the PDE is dynamically stable or not.

7 Physical and numerical dispersion and dissipation

7.1 Analysis of general planar waves

- Consider the following equations,

$$u_{,t} + a u_{,x} = 0 \qquad \text{1D advection equation} \tag{553a}$$
$$u_{,tt} - a^2 u_{,xx} = 0 \qquad \text{1D wave equation} \tag{553b}$$
$$u_{,t} - D u_{,xx} = 0 \qquad \text{1D diffusion equation} \tag{553c}$$
$$\tau u_{,tt} + u_{,t} - D u_{,xx} = 0 \qquad \text{1D relaxed diffusion equation} \tag{553d}$$
$$\rho \mathbf{u}_{,tt} - \nabla.(\mathcal{C}\nabla\mathbf{u}) = \mathbf{0} \qquad \text{elastodynamic problem} \tag{553e}$$

- We are seeking planar waves of the form,

$$u = f(x - ct) \tag{554}$$

for the first four equations and a planar wave for the last equation. The value c represent the speed of the wave.

- Plugging (555) in (553b) we obtain $(c-a)f'(x-ct) = 0$. If $c = a$ the PDE is satisfied. So, the advection equation admits the propagation of an arbitrary planar wave with speed a. In fact, $u(x,t) = u_0(x-at)$ is the solution to the PDE for u_0 being the initial condition for $-\infty < x < \infty$.
- Similarly, by plugging (555) in (553b) we obtain $(c^2 - a^2)f''(x-ct) = 0$ which is satisfied if $c = \pm a$. So, 1D wave equation admits planar wave propagation with speeds $\pm a$.
- Consider the equation (553e) $(\rho\mathbf{u}_{,tt} - \nabla.(\mathcal{C}\nabla\mathbf{u}) = 0)$.
- This is the elastodynamic equation $\rho\mathbf{u}_{,tt} - \nabla.\sigma = \mathbf{0}$ where ρ is mass density, $\sigma = \mathcal{C}\epsilon(\mathbf{u})$ is stress, $(\mathcal{C})$ elasticity tensor, $\epsilon(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla^{\mathrm{T}}\mathbf{u}\right)$ is strain and $\mathbf{u}$ is displacement vector. The minor symmetry of elasticity tensor $\mathcal{C}_{jikl} = \mathcal{C}_{ijlk}$ (*cf.* (148b)) is used in (553e).
- This equation demonstrated how planar waves are modeled in 2D and 3D plus how the analysis is done for a vector PDE.
- Now, for the vector elastodynamic equation we look for a planar wave of the form,

$$\mathbf{u} = f(\mathbf{x}.\mathbf{n} - ct)\boldsymbol{\Phi} \qquad \text{that is } u_i = \Phi_i f(x_i n_i - ct) \tag{555a}$$
$$\boldsymbol{\Phi} = \text{mode vector} \tag{555b}$$
$$f = \text{planar wave function} \tag{555c}$$
$$c = \text{wave speed} \tag{555d}$$
$$\mathbf{n} = \text{direction of wave propagation (a unit vector)} \tag{555e}$$

- In this case, the problem is in 3D and the direction of the wave is specified by $\mathbf{n}$.
- In addition, $\mathbf{u}$ is a vector field rather than a scalar function, so the function f is multiplied by the mode (shape) vector $\boldsymbol{\Phi}$.
- The values c and $\boldsymbol{\Phi}$ will be determined from an eigenvalue problem as will become apparent below.
- Equation (553e) can be expanded in tensorial form,

$$\rho u_{i,tt} - \left(\mathcal{C}_{ijkl}u_{k,l}\right)_{,j} = 0$$

- Assuming that $\mathcal{C}$ is constant and plugging (555a) in the above equation we obtain,

$$\begin{aligned} \rho c^2 \Phi_i f''(x_i n_i - ct) - \mathcal{C}_{ijkl} n_l n_j \Phi_k f''(x_i n_i - ct) = 0 \quad &\Rightarrow \\ f''(x_i n_i - ct)\left[\rho c^2 \delta_{ik} - n_j \mathcal{C}_{jikl} n_l\right]\Phi_k = 0 \end{aligned} \tag{556}$$

where we have used the symmetry $\mathcal{C}_{jikl} = \mathcal{C}_{ijkl}$; *cf.* (148b).

- By defining second order Acoustic matrix for a given direction $\mathbf{n}$,

$$\mathbf{A}(\mathbf{n}) = \mathbf{n}\mathcal{C}\mathbf{n} \qquad \text{that is} \qquad A_{ik} = n_j \mathcal{C}_{jikl} n_l \tag{557}$$

- Equation (556) has a nontrivial solution if,

$$\boxed{\mathbf{A}(\mathbf{n})\boldsymbol{\Phi} = \rho c^2 \boldsymbol{\Phi}} \tag{558}$$

- That is for a wave propagating in direction $\mathbf{n}$, ρc^2 is the eigenvalue of the acoustic vector and $\boldsymbol{\Phi}$ is its eigenvector.
- For a given $\mathbf{n}$ we can have three eigenvalues c^2 for $\mathbf{A}(\mathbf{n})$ with their corresponding eigenvectors $\boldsymbol{\Phi}$.
- Since $\mathbf{A}(\mathbf{n})$ is a positive definite second order tensor (why?) all its eigenvalues are positive (that is why $\rho c^2 > 0$ makes sense) and three orthonormal eigenvectors $\boldsymbol{\Phi}$ can be formed. If two eigenvalues c^2 are repeated, clearly their corresponding eigenvector space form a plane.
- For isotropic solid we have,

$$\mathcal{C}_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}), \quad \text{isotropic solid,} \quad (\mu, \lambda \text{ Lamé parameters}) \tag{559}$$

- For isotropic solid we have,

$$\mathbf{A}(\mathbf{n}) = (\lambda + \mu)\mathbf{n} \otimes \mathbf{n} + \mu\mathbf{I} \qquad \text{for isotropic solid} \tag{560}$$

- It is easy to verify from (558) and (560) ,

$$c_1 = \sqrt{\frac{\lambda + 2\mu}{\rho}} \qquad \text{Longitudinal wave speed} \qquad \boldsymbol{\Phi} \parallel \mathbf{n} \tag{561a}$$

$$c_2 = \sqrt{\frac{\mu}{\rho}} \qquad \text{Shear wave speed} \qquad \boldsymbol{\Phi} \perp \mathbf{n} \tag{561b}$$

  correspond to c for normal ($\boldsymbol{\Phi} \parallel \mathbf{n}$) and shear ($\boldsymbol{\Phi} \parallel \mathbf{n}$) wave propagation speeds with c_2 is a repeated root of (558) of multiplicity 2 and all vectors $\boldsymbol{\Phi}$ normal to $\mathbf{v}$ correspond to shape modes that correspond to propagation speed c_2.
- For anisotropic solid, $\boldsymbol{\Phi}$ may not be normal or tangent to $\mathbf{n}$ for a given $\mathbf{n}$. In addition, wave speeds can be different from one direction to the other.
- In this case, the minimum and maximum wave speeds are

$$c_{\min} = \min_{|\mathbf{m}|=|\mathbf{n}|=1}[m_j A_{jk}(\mathbf{n}) m_k] = \min_{|\mathbf{m}|=|\mathbf{n}|=1}[m_j n_i \mathcal{C}_{ijkl} n_l m_k] \tag{562a}$$

$$c_{\max} = \max_{|\mathbf{m}|=|\mathbf{n}|=1}[m_j A_{jk}(\mathbf{n}) m_k] = \max_{|\mathbf{m}|=|\mathbf{n}|=1}[m_j n_i \mathcal{C}_{ijkl} n_l m_k] \tag{562b}$$

  for unit vectors $\mathbf{n}, \mathbf{m}$.
- Note that if nonlinear formulation is used (both geometric and material) $c_{\min}$ can become zero a condition that corresponds to **loss of hyperbolicity**. This condition can be used as a criterion for crack formation; *cf. e.g.,* [Belytschko et al., 2003].

7.2 Harmonic analysis of PDEs

- In §sec:genPlanarWave we had examples from 1D (1D advection equation, 1D wave equation) and 3D (elastodynamic) problems that permit the propagation of arbitrary functions f with specific wave speeds.
- Now, we want to evaluate if the diffusion equation permits the propagation of an arbitrary function f with speed c.
- We plug (555) ($u = f(x - ct)$) in the 1D diffusion equation (553c) $u_{,t} - Du_{,xx} = 0$ to obtain,

$$-cf'(x - ct) - Df''(x - ct) = 0 \tag{563}$$

- Clearly, unlike previous examples **an arbitrary function** f cannot be propagated by the advection equation.
- However, the form of the equation in (563) suggests exponential functions may satisfy (563).
- A common analysis for a general equation would be how would a PDE (physical problem) propagate harmonic waves of a given wave length ξ.
- The reason for the analysis of harmonic waves becomes apparent shortly.
- In addition to being able to satisfy (563), studying harmonic solutions enable us to determine what is the speed of wave propagation for different wavenumbers and how the wave is dissipated in time.
- For the analysis we consider functions of the form,

$$u(x,t) = e^{\imath(\xi x - \omega t)} = e^{\imath(\xi x - \omega_R t)} e^{\omega_I t} \qquad \text{which is the 1D version of general expression:} \tag{564a}$$

$$u(\mathbf{x},t) = e^{\imath(\boldsymbol{\xi}.\mathbf{x} - \omega t)} = e^{\imath(\boldsymbol{\xi}.\mathbf{x} - \omega_R t)} e^{\omega_I t} \qquad \text{for a scalar problem in 2D and 3D or} \tag{564b}$$

$$\mathbf{u}(\mathbf{x},t) = \boldsymbol{\Phi} e^{\imath(\boldsymbol{\xi}.\mathbf{x} - \omega t)} \qquad \text{for a vector (tensor) problems in 2D and 3D} \tag{564c}$$

- Herein,
  1. ξ is the wavenumber $\Rightarrow$ spatial wavelength $=\frac{2\pi}{\xi}$. In general the wavenumber vector is $\boldsymbol{\xi}=\xi_i\mathbf{e}_i=|\boldsymbol{\xi}|\mathbf{n}$ corresponding to a wave of spatial wavelength $\frac{2\pi}{|\boldsymbol{\xi}|}$ in direction $\mathbf{n}$.
  2. $\omega=\omega_R+\imath\omega_I$ is the temporal frequency.
     - The term $e^{\imath(\xi x-\omega_R t)}$ ($e^{\imath(\boldsymbol{\xi}.\mathbf{x}-\omega_R t)}$) corresponds to a bounded and nondiminishing harmonic oscillation.
     - and $e^{\omega_I t}$ an evanescent wave if $\omega_I<0$, exponentially growing if $\omega_I>0$ and one if $\omega_I=0$.
  3. $\boldsymbol{\Phi}$ is the mode shape:
     - Equation (564c) is multiplied by the tensor $\boldsymbol{\Phi}$ to make the RHS consistent with $\mathbf{u}$.
     - For a multi-field PDEs (564a) and (564b) they too may be needed by scalar coefficient so that solution of nonzero solutions (through an eigenvalue problem) becomes possible. In general solution of (564a) and (564b) accepts a constant factor Φ for linear PDEs which for brevity we have dropped in those equations as it does not affect dissipation and dispersion analysis.

- For the 1D advection equation (553a) ($u_{,t}+au_{,x}=0$) we plug (564a) in it to obtain,

$$-\imath\omega e^{\imath(\xi x-\omega t)}+a\imath\xi e^{\imath(\xi x-\omega t)}=0 \quad\Rightarrow\quad (-\imath\omega+\imath a\xi)e^{\imath(\xi x-\omega t)}=0 \quad\Rightarrow\quad \omega=a\xi \quad \text{that is} \quad\Rightarrow \tag{565a}$$

$$\omega_I=0,\qquad \omega_R=a\xi,\ \text{and} \tag{565b}$$

$$u(x,t)=e^{\imath(\xi x-\omega_R t)}e^{\omega_I t}=e^{\imath(\xi x-\xi a t)} \tag{565c}$$

- Next, for the scalar 1D diffusion equation (553c) $u_{,t}-Du_{,xx}=0$ we plug (564a) in it to obtain,

$$-\imath\omega e^{\imath(\xi x-\omega t)}-D(\imath\xi)^2e^{\imath(\xi x-\omega t)}=0 \quad\Rightarrow\quad (-\imath\omega+D\xi^2)e^{\imath(\xi x-\omega t)}=0 \quad\Rightarrow\quad -\imath(\omega_R+\imath\omega_I)+\xi^2D=0 \quad\Rightarrow$$

$$(\omega_I+\xi^2D)+\imath(-\omega_R)=0 \quad\Rightarrow\quad \boxed{\omega_I=-\xi^2D,\qquad \omega_R=0},\ \text{that is} \tag{566a}$$

$$\boxed{u(x,t)=e^{\imath(\xi x-\omega_R t)}e^{\omega_I t}=e^{\imath\xi x}e^{-\xi^2Dt}} \tag{566b}$$

- From (566a) and (566b) we observe that 1D diffusion equation in fact admits a harmonic solution while a general planar wave solution of the form (555) is not possible.

- Interesting observations on $\omega=\omega_R+\imath\omega_I$ are,
  - $\omega_R=0$ so harmonic waves of the form $e^{\imath\xi x}$ are stationary in time.
  - $\omega_I=-\xi^2D$ that is $\omega_I<0$ corresponding to an evanescent wave as physically expected from the diffusion equation and more importantly the coefficient of decay $\omega_I=-\xi^2D$ rapidly grows in magnitude for **higher wave numbers** ξ. That is the diffusion equation **damps out high wavenumber (spatial frequency) content of data much quicker than lower wavenumber ones**.

- Now, let us investigate the analysis of the harmonic solution for the relaxed diffusion equation (553d) ($\tau u_{,tt}+u_{,t}-Du_{,xx}=\mathbf{0}$).

- To better connect that relaxed parabolic equation of a wave equation we rewrite the underlying equation (553d) ($\tau u_{,tt}+u_{,t}-Du_{,xx}=0$) in the following form (by diving it by τ)

$$u_{,tt}+\omega_0u_{,t}-a^2u_{,xx}=0,\qquad a=\sqrt{\frac{D}{\tau}},\omega_0=\frac{1}{\tau} \tag{567}$$

- The rationale is that for $\tau=0$ (553d) ($\tau u_{,tt}+u_{,t}-Du_{,xx}=0$) reduces to diffusion equation (553c) $u_{,t}-Du_{,xx}=0$ which was analyzed in the previous example; *cf.* (566b).

- We plug (564a) ($u(x,t)=e^{\imath(\xi x-\omega t)}$) into (567) to get,

$$(-\imath\omega)^2e^{\imath(\xi x-\omega t)}-\imath\omega_0\omega e^{\imath(\xi x-\omega t)}-a^2(\imath\xi)^2e^{\imath(\xi x-\omega t)}=0 \ \Rightarrow\ (-\omega^2-\imath\omega_0\omega+a^2\xi^2)e^{\imath(\xi x-\omega t)}=0 \ \Rightarrow\ \omega^2+\imath\omega_0\omega-a^2\xi^2=0 \ \Rightarrow \tag{568a}$$

$$\omega=\frac{1}{2}\left(-\imath\omega_0\pm\sqrt{-\omega_0^2+(2a\xi)^2}\right)\qquad\Rightarrow \tag{568b}$$

$$\omega_{1,2}=\begin{cases}-\imath\frac{1}{2}\left(\omega_0\mp\sqrt{\omega_0^2-(2a\xi)^2}\right) & \xi<\frac{\omega_0}{2a}\\ \frac{1}{2}\left(-\imath\omega_0\pm\sqrt{(2a\xi)^2-\omega_0^2}\right) & \xi>\frac{\omega_0}{2a}\end{cases}\quad\text{that is}\quad\begin{cases}\omega_R=0 & \omega_I=-\frac{1}{2}\left(\omega_0\mp\sqrt{\omega_0^2-(2a\xi)^2}\right) & \xi<\frac{\omega_0}{2a}\\ \omega_R=\pm\frac{1}{2}\sqrt{(2a\xi)^2-\omega_0^2} & \omega_I=-\frac{1}{2}\omega_0 & \xi>\frac{\omega_0}{2a}\end{cases} \tag{568c}$$

- The case $\xi < \frac{\omega_0}{2a}$ corresponds to an over-damped oscillator where the wavenumber ξ is not large enough to induce any oscillatory mode ($\omega_R = 0$).
- The case $\xi > \frac{\omega_0}{2a}$ corresponds to an under-damped oscillator where the wavenumber ξ is large enough permit oscillatory modes ($\omega_R \neq 0$).
- Finally the case $\xi = \frac{\omega_0}{2a}$ corresponds to a critically-damped oscillator, where due to multiplicity 2 of the roots ω a solution of the form $te^{\imath(\xi x - \omega t}$ is also permitted.
- In all three cases $\omega_I < 0$ for both roots. That is the waves are evanescent as expected from a relaxed diffusion equation.
- With the new notation, the critical wavenumber will be $\frac{1}{2\sqrt{\tau D}} = \frac{\omega_0}{2a}$.
- The following equation summarizes the harmonic solutions (568) that the relaxed diffusion equation admits.

$$u(x,t) = A_1 e^{\imath\xi x} e^{-\frac{1}{2}t\left(\omega_0 + \sqrt{\omega_0^2 - (2a\xi)^2}\right)} + A_2 e^{\imath\xi x} e^{-\frac{1}{2}t\left(\omega_0 - \sqrt{\omega_0^2 - (2a\xi)^2}\right)} \quad \text{Over-damped wavenumber} \quad \xi < \frac{\omega_0}{2a} \tag{569a}$$

$$u(x,t) = A_1 e^{\imath\xi x} e^{-\frac{t\omega_0}{2}} + A_2 t e^{\imath\xi x} e^{-\frac{t\omega_0}{2}} \quad \text{Critically damped wavenumber} \quad \xi = \frac{\omega_0}{2a} \tag{569b}$$

$$u(x,t) = A_1 e^{\imath\left(\xi x + t\frac{1}{2}\sqrt{(2a\xi)^2 - \omega_0^2}\right)} e^{-\frac{t\omega_0}{2}} + A_2 e^{\imath\left(\xi x - t\frac{1}{2}\sqrt{(2a\xi)^2 - \omega_0^2}\right)} e^{-\frac{t\omega_0}{2}} \quad \text{Under-damped wavenumber} \quad \xi > \frac{\omega_0}{2a} \tag{569c}$$

7.3 Transition between different PDE modes at different spacetime scales

- It is of importance to compute the asymptotic limits when $\xi \ll \frac{\omega_0}{2a}$ and $\xi \gg \frac{\omega_0}{2a}$,

$$u(x,t) \approx A_1 e^{\imath\xi x} e^{-\omega_0 t} + A_2 e^{\imath\xi x} e^{-\frac{(a\xi)^2}{\omega_0}t} = A_1 e^{\imath\xi x} e^{-\omega_0 t} + A_2 e^{\imath\xi x} e^{-\xi^2 Dt} \quad \xi \ll \frac{\omega_0}{2a} \tag{570a}$$

$$u(x,t) \approx A_1 e^{\imath(\xi x + a\xi t)} + A_2 e^{\imath(\xi x - a\xi t)} \quad \xi \gg \frac{\omega_0}{2a} \tag{570b}$$

- The solutions for $\xi \ll \frac{\omega_0}{2a}$ and $\xi \gg \frac{\omega_0}{2a}$ resemble two problems we studies before.
  - Diffusion equation:
    - Recall that for the diffusion equation (553c) $u_{,t} - Du_{,xx} = 0$ the harmonic solution was of the form (566b) ($u(x,t) = e^{\imath\xi x} e^{-\xi^2 Dt}$) which is exactly what we observe for the relaxed hyperbolic equation (553d) ($\tau u_{,tt} + u_{,t} - Du_{,xx} = 0$) in (570a) through the term multiplying A_2.
    - Although in (570a) the other term of $A_1 e^{\imath\xi x} e^{-\omega_0 t}$ is also present to a diffusion equation term (that is $A_2 e^{-\frac{(a\xi)^2}{\omega_0}t} = A_2 e^{\imath\xi x} e^{-\xi^2 Dt}$). However, since $\xi \ll \frac{\omega_0}{2a}$ we have $\omega_0 \gg \omega(a\xi)^2\omega_0$ the term $A_1 e^{\imath\xi x} e^{-\omega_0 t}$ tends to zero much quicker than $A_2 e^{\imath\xi x} e^{-\xi^2 Dt}$ and the latter term (which matches the parabolic solution) dominate the former term.
    - That is, **in small wavenumber ξ mode (large length scale $1/\xi$) the solution to a relaxed diffusion equation** (553d) $\tau u_{,tt} + u_{,t} - Du_{,xx} = 0$ **is well approximated with the solution of the (unrelaxed) diffusion equation** (553c) $u_{,t} - Du_{,xx} = 0$.
  - Undamped wave equation:
    - Consider the (undamped) wave equation (553b) ($u_{,tt} - a^2 u_{,xx} = 0$).
    - The harmonic solution for this equation can directly be obtained by letting $\omega_0 = 0$ in (567) ($u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0$) and taking the under-damped solution (569c) $u(x,t) = A_1 e^{\imath\left(\xi x + t\frac{1}{2}\sqrt{(2a\xi)^2 - \omega_0^2}\right)} e^{-\frac{t\omega_0}{2}} + A_2 e^{\imath\left(\xi x - t\frac{1}{2}\sqrt{(2a\xi)^2 - \omega_0^2}\right)} e^{-\frac{t\omega_0}{2}}$ which in fact holds for all ξ as any ξ satisfies $\xi > \omega_0 2a = 0$. By letting $\omega_0 = 0$ in this equation we recover,

$$u(x,t) = A_1 e^{\imath(\xi x + a\xi t)} + A_2 e^{\imath(\xi x - a\xi t)} \quad \text{Harmonic solutions for 1D wave equation} \tag{571}$$

    - This is the limiting case for $\xi \gg \frac{\omega_0}{2a}$ for a general damped wave equation (567) $u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0$ as demonstrated in (570b).
    - That is, **in large wavenumber ξ mode (small length scale $1/\xi$) the solution to a damped wave equation** (567) $u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0$ **(*i.e.*, relaxed diffusion equation** (553d) $\tau u_{,tt} + u_{,t} - Du_{,xx} = 0$**) is well approximated with the solution of the (undamped) wave equation** (553b) $u_{,tt} - a^2 u_{,xx} = 0$**.**
- Given that a length scale is provided by the wavenumber by its wavelength (spatial period) L through the relation $\xi L = 2\pi$ and the speed $a = \sqrt{D/\tau}$ we can define the following length, time, and frequency scales for transition of the PDE mode,

$$u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0((567)) \qquad \tau u_{,tt} + u_{,t} - Du_{,xx} = 0((553d))$$

Frequency scale $\tilde{\xi}$	$\frac{\omega_0}{2a}$	$\frac{1}{2\sqrt{\tau D}}$	(572a)
Length scale $\tilde{L}$	$\frac{4\pi a}{\omega_0}$	$4\pi\sqrt{\tau D}$	(572b)
Time scale $\tilde{T}$	$\frac{4\pi}{\omega_0}$	$4\pi\tau$	(572c)

where the notation $\tilde{P}$ corresponds to a relevant scale of that physical quantity.

- Note that even from dimension analysis we could have deduced the length, time, and frequency scales for possible mode transitions for the PDE:
  - The time scale would be $\tilde{T} = 1/\omega_0$ $(= \tau)$.
  - Wave speed is given by a $(= \sqrt{D/\tau})$.
  - Length scale is computed from $\tilde{L} = a\tilde{T} = a/\omega_0$ $(= \sqrt{\tau D})$.
  - Wavenumber (spatial frequency) scale $\tilde{\xi} = \frac{2\pi}{\tilde{L}} = 2\pi\omega_0/a$ $(= \frac{2\pi}{\sqrt{\tau D}})$.
- Note that there is a factor of 4π difference between the values deduced by simple dimensional analysis and those based on harmonic mode transition (in this case from over-damped to under-damped mode).
- In fact, many other types of analysis can be performed to deduce the scales for time, length, *etc.* for this problem.
- All these scales would agree in terms of their parameters and may only differ by their constant factors.
- The important part is the two limiting cases of this damped hyperbolic equation (567) ($u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0$) alternatively expressed in the form of a relaxed hyperbolic equation (553d) ($\tau u_{,tt} + u_{,t} - D u_{,xx} = 0$):
  1. Very large length scale ($L \gg \tilde{L}$), time scale ($T \gg \tilde{T}$), low wavenumber $\xi \ll \tilde{\xi}$: Solution can be approximated with the PDE maintaining the **LOWEST ORDER** terms of the PDE for each argument $(x, t,\ etc.)$. In this case between $u_{,t}$ and $u_{,tt}$ we can discarded $u_{,tt}$ at high spacetime scales with reasonable accuracy. So, the equation can be approximated by,

$$\tau u_{,tt} + u_{,t} - D u_{,xx} = 0 \quad \text{Very large spacetime scales } \Rightarrow \text{ can be approximated by} \quad u_{,t} - D u_{,xx} = 0 \tag{573}$$

     note that for this problem the large spacetime scale limiting equation is parabolic.
  2. Very small length scale ($L \gg \tilde{L}$), time scale ($T \gg \tilde{T}$), high wavenumber $\xi \ll \tilde{\xi}$: Solution can be approximated with the PDE maintaining the **HIGHEST ORDER** terms of the PDE for each argument $(x, t,\ etc.)$. In this case between $u_{,t}$ and $u_{,tt}$ we can discarded $u_{,t}$ at high spacetime scales with reasonable accuracy. So, the equation can be approximated by,

$$u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0 \quad \text{Very small spacetime scales } \Rightarrow \text{ can be approximated by} \quad u_{,tt} - a^2 u_{,xx} = 0 \tag{574}$$

     note that for this problem the large spacetime scale limiting equation is hyperbolic.
- This type of argument on which terms can be discarded at very small or large time scales can also be justified from dimensional analysis perspective by expressing a PDE at particular spacetime scales and observe that if large spacetime scales are used the highest derivative terms (*e.g.,* $u_{,tt}$) will have smaller coefficient than their corresponding lower order derivative terms ((*e.g.,* $u_{,t}$). In contrast, if PDE is scaled by very small spacetime scales, the coefficient of the lowest order terms will be larger and these terms will dominate.
- To demonstrate PDE mode transition of application of different PDE types to different length and time scales consider the balance of energy for thermal response,

$$C\dot{T} + \nabla \cdot \mathbf{q} = Q \tag{575}$$

  where C is the volumetric heat capacity, T is temperature, $\mathbf{q}$ is heat flux vector, and Q is volumetric heat supply.
- If we employ Fourier's constitutive model,

$$\mathbf{q} = -\kappa \nabla T \qquad \text{Fourier heat conduction constitutive model} \tag{576}$$

  where κ is the thermal conductivity tensor, we obtain the familiar parabolic heat equation,

$$C T_{,t} - \nabla \cdot (\kappa \nabla T) = Q \qquad \text{Fourier thermal conduction} \tag{577}$$

- One of the main shortcomings of the Fourier's heat model is its poor performance at very small length and time scales and also the fact that parabolic equations imply a nonphysical wave speed of infinity.
- Several hyperbolic heat conduction models are proposed to remedy the infinite wave speed implied by the Fickian heat equation.
- A popular thermal wave model was independently proposed by Maxwell [Maxwell, 1867], Cattaneo [Cattaneo, 1948], and Vernotte [Vernotte, 1958].
- For a homogeneous material this so-called *MCV constitutive model* is expressed as,

$$\tau\dot{\mathbf{q}} + \mathbf{q} = -\kappa\nabla T \qquad \text{MCV heat conduction constitutive model} \tag{578}$$

  which basically "relaxes" the Fourier thermal constitutive model (576) by the term $\tau\dot{\mathbf{q}}$. The values τ is called the relaxation time.
- Using (577) and (576) we can show,

$$\tau CT_{,tt} + CT_{,t} - \nabla\cdot(\kappa\nabla T) = Q + \tau Q_t \qquad \text{MCV thermal conduction} \tag{579}$$

- Unlike the Fourier heat equation which is parabolic, the MCV equation (579) corresponds to a hyperbolic equation.
- In fact, being a hyperbolic equation, the thermal wave speed $c = \sqrt{\frac{\Lambda}{\tau C}}$ (Λ is the maximum eigenvalue of κ) models the physical nature of thermal wave propagation mode at very small space and time scales. In thermal physics c is referred to as *second-sound speed*; *cf.* [Dedeurwaerdere et al., 1996, Compte and Jou, 1996] for more information.
- Having the higher temporal derivatives makes the MCV model particularly a better model than Fourier model at very small spacetime scales.
- There are a variety of other thermal models that are also formulated to better model heat condition at small spacetime scales. They too often involve higher space and/or time derivative terms.
- The discussion in the first part of §7.3 clarifies why the higher derivative terms are important at small spacetime scales.
- Now, if we are interested in finding a reasonable PDE that corresponds to the MCV equation at large space/time scales we can discard the highest temporal derivative as we just did for a general relaxed hyperbolic equation.
- In such case, it is apparent that the MCV equation (579) can be accurately approximated by the Fourier heat conduction model (577) at large space/time scales.

7.4 Relation between Fourier transform and harmonic solutions

- To motivate the relation between harmonic analysis and Fourier analysis consider the advection equation (553a) ($u_{,t}+au_{,x}=0$).
- We already know that the advection equation has a solution of the form,

$$u(x,t) = u_0(x-at) \tag{580}$$

  where u_0 is the initial condition (IC) function ($u(x,t=0)=u_0(x)$).
- The advection equation **does not disperse** the solution as it simply advects any initial data by the speed of a in space.
- This can also be observed with the solution of harmonic solutions and connecting it to the Fourier transform of the solution.
- The spatial Fourier transform of $u(x,t)$ and its inverse are (*cf.* (412)),

$$\hat{u}(\xi,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u(x,t)e^{-\imath\xi x}\,\mathrm{d}x \tag{581a}$$

$$u(x,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{u}(\xi,t)e^{\imath\xi x}\mathrm{d}\xi \tag{581b}$$

- Now, the use of analysis of harmonic solution of PDEs from §7.2 becomes apparent:
  - Initial Conditions (ICs) can be written as a linear combination of waves.
  - For linear PDEs the solution of the PDE to any arbitrary IC is simply the linear combination of the PDE to these simple harmonic ICs.

– The solution to these ICs is obtained by the harmonic analysis discussed before.

- The formalized form of this process is illustrated by the solution to the advection equation (553a) ($u_{,t}+au_{,x}=0$).
- The Fourier transform (581a) applied to the advection equation yields,

$$\widehat{u_{,t}+au_{,x}}=0 \qquad \Rightarrow \qquad \hat{u}_{,t}+a\widehat{u_{,x}}=\hat{u}_{,t}+a(\imath\xi)\hat{u}=0 \tag{582}$$

in which we have used $\widehat{u_{,x}}=(\imath\xi)\hat{u}$ which turns derivatives to algebraic multiplication by Fourier transformation; *cf.* (204).

- Equation (582) is now an ODE in time,

$$\hat{u}_{,t}(\xi,t)+\imath a\xi\hat{u}(\xi,t)=0 \qquad \text{ODE in } t \textbf{ for each } \xi \tag{583a}$$

$$\hat{u}(\xi,t=0)=\hat{u}_0(\xi) \qquad \text{IC} \tag{583b}$$

- The **enormous advantage of Fourier transform is the ability to change this two argument PDE to an ODE**: For a **fixed wavenumber** ξ we are dealing with harmonic solutions whose evolution in time was discussed in §7.2.
- The solution to the initial value ODE (583) is,

$$\hat{u}(\xi,t)=e^{-\imath a\xi t}\hat{u}_0(\xi) \tag{584}$$

- Now the contribution from one frequency in (581b) ($u(x,t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\hat{u}(\xi,t)e^{\imath\xi x}\mathrm{d}\xi$) to the solution $u(x,t)$ is

$$u(x,t)\sim\hat{u}(\xi,t)e^{\imath\xi x}=e^{\imath\xi x}e^{-\imath a\xi t}\hat{u}_0(\xi)=e^{\imath(x\xi-(a\xi)t)}\hat{u}_0(\xi)$$

- No surprise that the term $e^{\imath(x\xi-(a\xi)t)}$ is exactly what we derived by the harmonic analysis of the advection equation in (565c) ($u(x,t)=e^{\imath(\xi x-\xi at)}$).
- In the comparison above, we have not incorporated the factors $1/\sqrt{2\pi}$ in Fourier transform equations.
- That is the Fourier analysis:
  1. **Decomposes** arbitrary initial condition(s) to harmonics with wavenumber ξ and amplitude $\hat{u}_0(\xi)$ (This is done by the **Fourier transform** of the initial data).
  2. Solving an ODE on t obtained by Fourier transform of the PDE (which basically solved harmonic solutions in the form discussed in §7.2).
  3. **Combine (add)** contributions from the solution of problems with simple harmonic ICs. This is equivalent to **inverse Fourier transform** of the solution from wavenumber (spatial frequency) to space coordinate.
- It should be emphasized that if the problem is nonlinear different frequencies interact and the solution using Fourier transformation would not be straightforward.
- Now that we have made the connection between Fourier transform and harmonic analysis in §7.2 we continue with the solution (583) by having the solution in Fourier space from (584) ($\hat{u}(\xi,t)=e^{-\imath a\xi t}\hat{u}_0(\xi)$).
- The Fourier inverse of $\hat{u}(\xi,t)$ is computed from (581b)

$$u(x,t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\hat{u}(\xi,t)e^{\imath\xi x}\mathrm{d}\xi=\int_{-\infty}^{\infty}e^{-\imath a\xi t}\hat{u}_0(\xi)e^{\imath\xi x}\mathrm{d}\xi=\int_{-\infty}^{\infty}\hat{u}_0(\xi)e^{\imath\xi(x-at)}\mathrm{d}\xi=u_0(x-at) \tag{585}$$

given that $u_0(y)=\int_{-\infty}^{\infty}e^{\imath\xi y}\hat{u}_0(\xi)\mathrm{d}\xi$ and choosing $y=x-at$.

- Clearly, for this simple problem using Fourier transformation makes the solution of the advection problem more difficult.
- In fact, from §7.1 we recall that advection equation not only move any harmonic wave with speed a, but also with any arbitrary profile. That is what $u(x,t)=u_0(x-at)$.
- As the second equation we consider the 1D wave equation (553b) with initial condition,

$$u_{,tt}-a^2u_{,xx}=0 \qquad \text{PDE of 1D wave equation} \tag{586a}$$

$$\begin{cases} u(x,t=0)=u_0(x) \\ u_{,t}(x,t=0)=\dot{u}_0(x) \end{cases} \qquad \text{ICs} \tag{586b}$$

- The Fourier transform of (586) given,

$$\hat{u}_{,tt} - a^2(\imath\xi)^2\hat{u} = 0 \qquad \text{PDE of 1D wave equation in terms of wavenumber } \xi \text{ and time } t \tag{587a}$$

$$\begin{cases} \hat{u}(\xi, t=0) = \hat{u}_0(\xi) \\ \hat{u}_{,t}(\xi, t=0) = \hat{\dot{u}}_0(\xi) \end{cases} \qquad \text{ICs of the ODE in time } t \tag{587b}$$

- The solution to this ODE (587a) is,

$$\hat{u} = A_1(\xi)e^{\imath\xi at} + A_2(\xi)e^{-\imath\xi at} \tag{588}$$

- Note that again similar to the previous case, effectively for fixed wavenumber the solutions $u(x,t) \sim \left(A_1(\xi)e^{-\imath\xi t} + A_2(\xi)e^{\imath\xi t}\right)e^{\imath\xi x} = A_1(\xi)e^{\imath(\xi t + a\xi t)} + A_2(\xi)e^{\imath(\xi t + a\xi t)}$ which is basically the undamped wave equation harmonic solutions we obtained before in (571) again ignoring the $1/\sqrt{2\pi}$ factors in Fourier transforms.
- So, this is another example that the solutions of the PDE for a fixed frequency corresponds to simple harmonic solutions we discussed in §7.2.
- The values $A_1(\xi)$ and $A_2(\xi)$ are obtained by the ICs (587b).
- After the solution of $A_1(\xi)$ and $A_2(\xi)$ the Fourier transform of the solution takes the form,

$$\hat{u}(\xi, t) = \frac{\hat{u}_0(\xi)}{2}\left(e^{\imath a\xi t} + e^{-\imath a\xi t}\right) + \frac{\hat{\dot{u}}_0(\xi)}{2\imath\xi}\left(e^{\imath a\xi t} - e^{-\imath a\xi t}\right) \tag{589}$$

- The solution $u(x,t)$ is obtained by the inverse Fourier transform of $\hat{u}(\xi,t)$ using (581b)

$$\begin{aligned} u(x,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\hat{u}(\xi,t)e^{\imath\xi x}\mathrm{d}\xi \\ &= \int_{-\infty}^{\infty}\left(\frac{\hat{u}_0(\xi)}{2}e^{\imath a\xi t}\right)e^{\imath\xi x}\mathrm{d}\xi + \int_{-\infty}^{\infty}\left(\frac{\hat{u}_0(\xi)}{2}e^{-\imath a\xi t}\right)e^{\imath\xi x}\mathrm{d}\xi + \int_{-\infty}^{\infty}\left(\frac{\hat{\dot{u}}_0(\xi)}{2\imath\xi}e^{\imath a\xi t}\right)e^{\imath\xi x}\mathrm{d}\xi - \int_{-\infty}^{\infty}\left(\frac{\hat{\dot{u}}_0(\xi)}{2\imath\xi}e^{-\imath a\xi t}\right)e^{\imath\xi x}\mathrm{d}\xi \qquad (590a) \\ &= \frac{1}{2}\int_{-\infty}^{\infty}\hat{u}_0(\xi)e^{\imath\xi(x+at)}\mathrm{d}\xi + \frac{1}{2}\int_{-\infty}^{\infty}\hat{u}_0(\xi)e^{\imath\xi(x-at)}\mathrm{d}\xi + \frac{1}{2}\int_{-\infty}^{\infty}\frac{\hat{\dot{u}}_0}{\imath\xi}(\xi)e^{\imath\xi(x+at)}\mathrm{d}\xi - \frac{1}{2}\int_{-\infty}^{\infty}\frac{\hat{\dot{u}}_0}{\imath\xi}(\xi)e^{\imath\xi(x-at)}\mathrm{d}\xi \\ &= u_0(x+at) + u_0(x-at) + \dot{U}_0(x+at) - \dot{U}_0(x-at) \qquad (590b) \end{aligned}$$

where $\dot{U}_0$ is the anti-derivative of $\dot{u}_0$ (the initial velocity). That is, $\dot{U}_0(x) = \int_{-\infty}^{x}\dot{u}_0(y)\mathrm{d}y \Leftrightarrow \frac{\mathrm{d}\dot{U}_0(x)}{\mathrm{d}x} = \dot{u}(x)$. The reason for this is the division by $\imath\xi$ in the last two terms of the integrals in (590a) (recall $\frac{\widehat{\mathrm{d}f}}{\mathrm{d}x} = \imath\xi\hat{f}$; (204)).

- Given that $\dot{U}_0(x+at) - \dot{U}_0(x-at) = \int_{-\infty}^{x+at}\dot{u}_0(y)\mathrm{d}y - \int_{-\infty}^{x-at}\dot{u}_0(y)\mathrm{d}y = \int_{x-at}^{x+at}\dot{u}_0(y)\mathrm{d}y - \int_{-\infty}^{x-at}\dot{u}_0(y)\mathrm{d}y$ the solution to the wave equation from (590b) is,

$$u(x,t) = \frac{1}{2}\left(u_0(x-at) + u_0(x+at)\right) + \frac{1}{2}\int_{x-at}^{x+at}\dot{u}_0(y)\mathrm{d}y \tag{591}$$

  for initial value $u_0(x)$ and speed (time rate) $\dot{u}_0(x)$.
- This matches the D'Alembert solution which we derived before using the method of characteristics.

7.5 Dispersion and dissipation for a harmonic wave

- In §7.4 through the solution of two sample problems of advection equation (553a) ($u_{,t} + au_{,x} = 0$) and wave equation (553b) ($u_{,tt} - a^2u_{,xx} = 0$) we demonstrated that Fourier transform basically "adds" the solution of the PDE to simple harmonic ICs together for a linear PDE.
- The same process can be used for the solution of diffusion equation (553c) $u_{,t} - Du_{,xx} = 0$ and relaxed diffusion equation (damped wave equation) (553d) $\tau u_{,tt} + u_{,t} - Du_{,xx} = \mathbf{0}$.
- In these cases the harmonic solutions (566b) ($u(x,t) = e^{\imath\xi x}e^{-\xi^2 Dt}$) for diffusion equation and (570) for relaxed diffusion equation will be effectively added by the Fourier analysis by decomposing initial condition(s) to simple harmonic waves.

- Now, to better understand the concepts of dispersion and diffusion, consider the 1D harmonic solution (564a) rewritten as follows,

$$u(x,t) = e^{\imath(\xi x - \omega t)} = e^{\imath\xi(x-\omega/\xi t)} \tag{592}$$

- This implies a harmonic wave of wavelength (spatial frequency) ξ propagates with speed ω/ξ as along the direction ω/ξ the solution is unaltered.
- We define **phase velocity** as follows,

$$\boxed{c_p(\xi) := \frac{\omega}{\xi} \qquad \text{phase velocity}} \tag{593}$$

$c_p(\xi)$ is wavenumber dependence because for (592) to be a solution to a PDE, ω becomes a function of ξ.

- By the definition of phase velocity in (592) we observe,

$$u(x,t) = e^{\imath\xi(x-c_p(\xi)t)}, \qquad c_p(\xi) = \frac{\omega}{\xi} \tag{594}$$

- To demonstrate the importance of phase velocity we plug (592) into advection and wave equations to obtain,

$$u(x,t) = e^{\imath(\xi x-\omega t)} \qquad \text{in (553a)}: \; u_{,t} + a u_{,x} = 0 \qquad \Rightarrow \qquad \omega(\xi) = a\xi \qquad \Rightarrow \qquad c_p(\xi) = \frac{\omega(\xi)}{\xi} = a \tag{595a}$$

$$u(x,t) = e^{\imath(\xi x-\omega t)} \qquad \text{in (553b)}: \; u_{,tt} - a^2 u_{,xx} = 0 \qquad \Rightarrow \qquad \omega^2(\xi) = a^2\xi^2 \qquad \Rightarrow \qquad c_p(\xi) = \frac{\omega(\xi)}{\xi} = \pm a \tag{595b}$$

- An equation relating ω and ξ, *e.g.*, $\omega(\xi) = a\xi$ or $\omega^2(\xi) = a^2\xi^2$ in (595), is called a **dispersion relation**.
- Dispersion relation specifies how temporal and spacial frequencies are related for a given PDE.
- For the two problem in (595) we observe two things:
  1. The phase velocity is independent from wavenumber.
  2. ω (and $c_p(\xi)$) are both entirely real.
- The first property means that wave speed is independent of wavenumber.
- This means any solution feature propagates unaltered: The arbitrary shape feature is broken into a summation of harmonic waves (using Fourier series) and all those harmonic waves move with the same speed c independent of ξ.
- So, at a later time the same shape of the solution feature is reconstructed by summation of individual harmonic waves having moved the same distance (*i.e.*, using inverse Fourier series for a linear PDE).
- The advection equation $u_{,t} + a u_{,x} = 0$ and wave equation $u_{,tt} - a^2 u_{,xx} = 0$ correspond to a **nondispersive** response as solution features propagate unaltered and do not disperse.
- The second property ($\omega_I = 0$) imply that the the solution is non-dissipative. This is better understood by writing the solution as follows; *cf.* (564a),

$$u(x,t) = e^{\imath(\xi x-\omega t)} = e^{\imath(\xi x-\omega_R t)} e^{\omega_I t}$$

- Beside observing that only ω_R contributes to oscillatory wave propagation, we notice that the imaginary part of ω contributes to an evanescent mode (if $\omega_I < 0$) and an exponentially growing mode if $\omega_I > 0$.
- For the advection equation and wave equation $\omega_I = 0$ and the solution is non-dissipative.
- On the other hand, as we observed in (566a), for diffusion equation $u_{,t} - D u_{,xx} = 0$ $\omega_I = -\xi^2 D$ signifying a dissipative response whose strength in fact growth for higher frequency modes, explaining the smoothening effect of diffusion equation.
- Not all equations and problems are nondispersive and nondissipative.
- For example, let us consider the damped wave equation (567) $u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0$ which if $\omega_0 = 0$ will be match the solutions we obtained for the undamped wave equation.
- The dispersion relation for this equation was given in (568a) and (568b),

$$\omega^2 + \imath\omega_0\omega - a^2\xi^2 = 0 \qquad \Rightarrow \qquad \omega = \frac{1}{2}\left(-\imath\omega_0 \pm \sqrt{-\omega_0^2 + (2a\xi)^2}\right)$$

- The solutions for over-damped ($\xi < \omega_0/2a$) and under-damped ($\xi > \omega_0/2a$) cases are given in (568c).
- For the discussion here we only consider the under-damped solutions, which for a slightly damped wave equation the range ($\xi > \omega_0/2a$) may hold in a wide range of cases where the system response is better represented by the wave equation rather than the diffusion equation. The frequency for this case is (*cf.* (568c)),

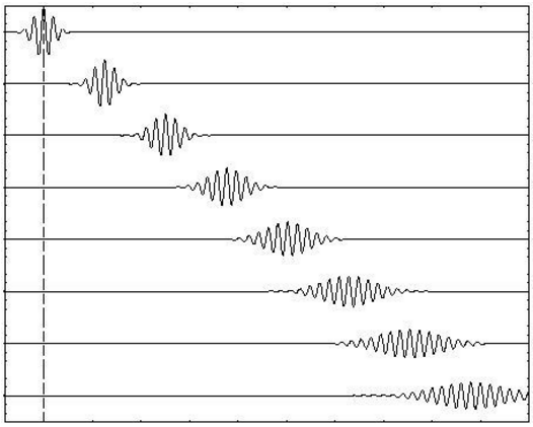
$$\omega = \omega_R + \imath\omega_I \qquad \text{where} \qquad \omega_R = \pm\frac{1}{2}\sqrt{(2a\xi)^2 - \omega_0^2} \text{ and } \omega_I = -\frac{1}{2}\omega_0, \qquad (\xi > \omega_0/2a) \tag{596}$$

- Accordingly, phase velocity is,

$$c_p(\xi) = \frac{\omega}{\xi} = a\left\{\pm\sqrt{1 - \left(\frac{\omega_0}{2a\xi}\right)^2} - \imath\left(\frac{\omega_0}{2a\xi}\right)\right\} \tag{597}$$

- We observe that in this case phase velocity $c_p(\xi)$ is wavenumber ξ-dependent.
- That is, the PDE response for the damped wave equation is dispersive:
  1. As $\xi \to \left(\frac{\omega_0}{2a}\right)^+$: $c_p(\xi) \to -\imath a$. That is the (oscillatory) wave speed approaches zero and the imaginary part of $c_p(\xi)$ correspond to the evanescent response.
  2. As $\xi \to \infty$: $c_p(\xi) \to \pm a$ *i.e.,* the response approaches that of and undamped wave equation; *cf.* (595b). The response in this limit is also nondissipative.

- The fact that different wave numbers propagate with different speeds (and also in this case dissipate at different rates) mean that a given solution feature (wave profile) **disperses**.
- The figure below show how a unit square is propagated with a nondispersive wave equation and a dispersive one.
- Since different parts (frequency components) of a wave propagate at different speeds, a wave pocket in fact **spreads and disperses** for a nondispersive problem / media.
- Interestingly, we observe the wave pocket in the figure has an overall speed that moves it to the right.



- This speed is called **group velocity** and in contrast to phase velocity, which is defined as $c_p(\xi) = \omega/\xi$ and corresponds to a simple harmonic wave, is defined as

$$c_g(\xi) = \frac{\mathrm{d}\omega}{\mathrm{d}\xi} \tag{598}$$

  The subscript g and p are to differentiate between group and phase velocity.
- The wavenumber used in evaluating group speed of a wave pocket is the dominant wavelength of the wave pocket (which can be determined from Fourier transform of the wave pocket).
- Finally, as comparison to phase velocity, group velocity of the damped wave equation is; *cf.* (596) and (597),

$$c_g(\xi) = \frac{\mathrm{d}\omega}{\mathrm{d}\xi} = \frac{\mathrm{d}\left(\pm\frac{1}{2}\sqrt{(2a\xi)^2 - \omega_0^2} - \imath\frac{1}{2}\omega_0\right)}{\mathrm{d}\xi} = \pm a\frac{1}{\sqrt{1 - \left(\frac{\omega_0}{2a\xi}\right)^2}} \tag{599}$$

  which is interesting to compare with phase velocity from (597) $c_p(\xi) = \frac{\omega}{\xi} = a\left\{\pm\sqrt{1 - \left(\frac{\omega_0}{2a\xi}\right)^2} - \imath\left(\frac{\omega_0}{2a\xi}\right)\right\}$.
- For more information about dispersion, phase velocity, group velocity, and dissipation refer to [Morin, 2010] which is also shared with you on dropbox.
- Finally consider $\omega = \omega_R + \imath\omega_I$. The solution is,

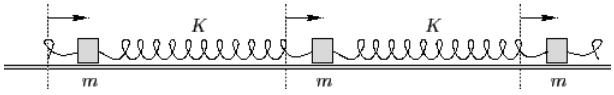
$$u(x,t) = e^{\imath(\xi x + \omega_R t)}e^{\omega_I t} \tag{600}$$

- If $\omega_R \neq 0$ one period of oscillation corresponds to $\omega_R T = 2\pi \Rightarrow T = \frac{2\pi}{\omega_R}$.
- In that one period the solution amplitude decay is $e^{\omega_I T}$. That is,

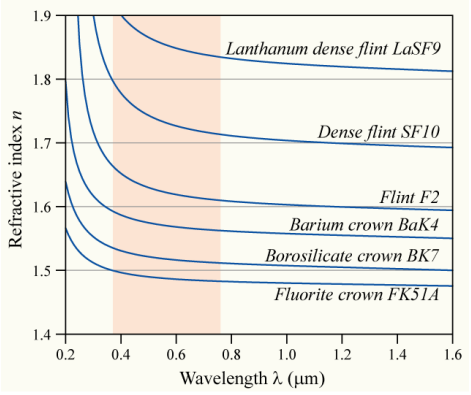
$$\boxed{\text{Amplitude decay } = e^{\left(2\pi\frac{\omega_I}{\omega_R}\right)} \qquad \text{for one period} \quad T = \frac{2\pi}{\omega_R}} \tag{601}$$

7.6 Dispersive media

- We observed that if the wave equation is damped (*i.e.*, $u_{,tt} - a^2 u_{,xx} = 0$ being modified to $u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0$) the phase velocity $c_p(\xi)$ is wavenumber dependence unlike the original wave equation that does not disperse wave pockets.
- Physical dispersion correspond to a media propagating waves with **different frequencies** with **different speeds**.
- Frequency here can refer to both spatial frequency (wavenumber) ξ or temporal frequency ω_R.
- This concept can also be viewed from their corresponding period perspective, *i.e.*, propagating waves with different wavelengths / period with different speeds.
- Many physical phenomena are modeled by wave propagation.
- Some examples are electromagnetics, acoustics, and elastodynamics, all clearly being models in dynamic mode (time-dependent).
- Physically, dispersive media correspond to those whose constitutive equation is **frequency-dependent**.
- Dispersion can physically occur if the wavelength is comparable to material microstructure length scales.
- If the wavelength is much larger than material microstructure length scales it can macroscopically viewed as homogeneous and the large wavelengths basically "do not feel" microstructure.
- Examples (they are not inclusive and certain examples belong to multiple groups):
  - In the damped hyperbolic problem considered $c_p(\xi) = a\left\{\pm\sqrt{1-\left(\frac{\omega_0}{2a\xi}\right)^2} - \imath\left(\frac{\omega_0}{2a\xi}\right)\right\}$ for $\xi > \omega_0/2a$. For wavenumber $\xi \approx \omega_0/2a$ and $\xi < \omega_0/2a$ (wavelength $\lambda \gtrsim 4\pi a/\omega_0$) the behavior is highly dispersive ($\lambda = 2\pi/\xi$). Note that in this case, it is the high wavelength range that corresponds to dispersive response.
  - Beaded string / mass spring: as a classical example in [Morin, 2010] beads with distance l are connected by a spring. If the wavelength $\lambda = 2\pi/\xi$ is much larger than the distance between the beads l the sting can effectively be modeled as a continuum string, whereas when $\lambda \lesssim l$ the material response is dispersive. A similar behavior is observed in mass spring systems.

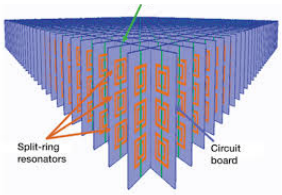


  - Complex media: Again, the wave propagation in any media in which the wavelength is of the same order or smaller than the microstructure can exhibit dispersive response. Some examples are electromagnetic, acoustic, and elastodynamic wave propagation in complex media, *e.g.*, composites, with fine enough microstructure for the exposed wavelength. Acoustic wave propagation through schools of fish in the ocean is another interesting example.
  - Some natural materials in optics, general electromagnetics, elastodynamics, and acoustics can exhibit dispersive response. Figure below shows how optics refractive index varies based on the wavelength.

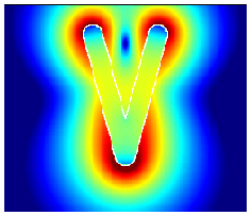


The variation of refractive index vs. vacuum wavelength for various glasses. The wavelengths of visible light are shaded in red. (source: wikipedia)

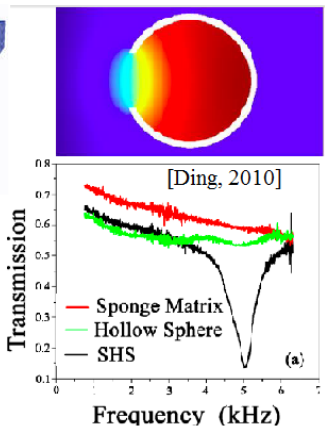
  - Metamaterials are man-made materials whose microstructure is engineered to exhibit desired behavior. They find applications in electromagnetic (optic) and acoustic cloaks, perfect lense, specific waveguides, *etc.*.



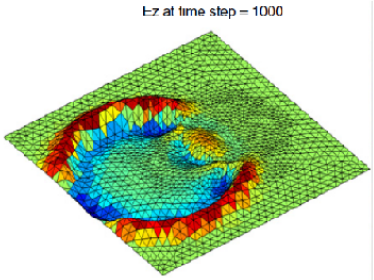
Split-ring resonators (SRRs)



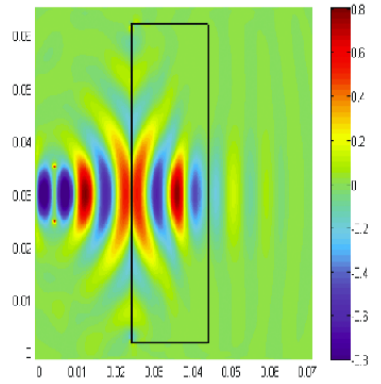
Silver V resonator
[Stannigel, 2009]



Transmission coefficient for
Split Hollow Sphere (SHS)



Electromagnetic cloaking (wave reflected
from the circular zone on the right)
[Li, 2011]



Wave propagation in a metamaterial
(boxed region) [Huang, 2013]

* The figures are taken from Wikipedia, and [Stannigel et al., 2009, Ding et al., 2010, Li, 2011, Huang et al., 2013].
* Those on the left show some common repeating features in metamaterials (*e.g.,* Split-ring resonator (SRR) for electromagnetics, Split-hollow sphere (SHS) for acoustic, *etc.*).
* Figures on the left show some interesting wave phenomena in metamaterials such as cloaking and backward wave propagation.
* Constitutive and wave propagation parameters of metamaterials, *e.g.,* electric permittivity and magnetic permeability (electromagnetics), and bulk & elastic modulus, mass density for acoustics / electromagnetics, can be both frequency dependent and have imaginary components.
* An example of frequency-dependent transmission coefficient is shown in the figure.

- For modeling dispersive media the main point is that material properties are frequency dependent.
- This frequency can be either spatial frequency (wavenumber ξ) or temporal (ω).
- It is more convenient to express material properties in temporal frequency domain.
- For example, in many electromagnetic applications a fixed frequency is encountered and for the analysis a static analysis of the temporal Fourier analysis of the problem is sufficient.
- Basically, Fourier transform is used to cast an equation in the **frequency domain (FD)** from a **time domain (TD)**.
- While in harmonic wave analysis we were dealing with harmonic waves with fixed wavenumber (spatial frequency) ξ and the Fourier transform in (581) was considered in our analysis (*e.g.,* $\hat{u}(\xi,t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}u(x,t)e^{-\imath\xi x}\,\mathrm{d}x$ in (581a)) it is often more convenient to consider the temporal Fourier transform of the equations.
- This provides more flexibility in dealing with practical problems where domains are actually finite and have complex geometries in space.
- Temporal Fourier transform transforms a dynamic PDE into a sequence of static PDEs (*i.e.,* time-independent) each one solving a problem with a fixed frequency.
- If the problem is already for a fixed frequency, *e.g., Alternative Current (AC)* in many electromagnetic applications, frequency domain analysis is even more natural and convenient.
- To continue, using the general definition of Fourier transformation, we define temporal Fourier transformation of a function,

$$\breve{u}(\mathbf{x},\omega)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}u(\mathbf{x},t)e^{-\imath\omega t}\,\mathrm{d}t \tag{602a}$$

$$u(\mathbf{x},t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\breve{u}(\omega,t)e^{\imath\omega t}\mathrm{d}\omega \tag{602b}$$

- We have used the notations of $\hat{u}$ and $\breve{u}$ to better distinguish between spatial and temporal Fourier transforms although once their arguments are provided there would be no ambiguity: $\hat{u}(\boldsymbol{\xi},t)$ vs. $\breve{u}(\mathbf{x},\omega)$.
- For a 1D problem $\mathbf{x}$ is simple x in (602).
- In general, temporal Fourier transform (602) turns a dynamic PDE to a static PDE only in spatial domain.

- The resulting PDE and its boundary conditions on the boundaries of the spatial domain, then can be solved with numerical methods such as FD, FEM, *etc...*
- To demonstrate how dispersive media can be modeled consider the following wave equation for a media in which relevant material properties to wave speed are isotropic and homogeneous,

$$u_{,tt} - a^2 \nabla.\nabla u = 0, \qquad \text{In 1D the equation is} \qquad u_{,tt} - a^2 u_{,xx} = 0 \tag{603}$$

- The temporal Fourier transform of (604) is,

$$(\imath\omega)^2 \check{u} - a^2 \nabla.\nabla \check{u} = 0, \qquad \text{that is in 1D} \qquad (\imath\omega)^2 \check{u} - a^2 \check{u}_{,xx} = 0 \tag{604}$$

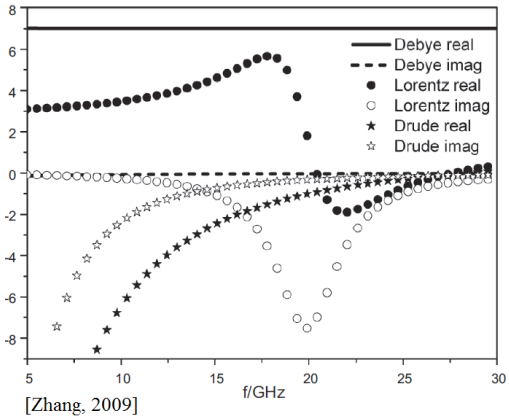
  noting that $\check{\ddot{u}} = (\imath\omega)^2 \check{u}$.
- Note that is basically a static problem with boundary conditions on the boundary o the spatial domain.
- To solve a dynamic problem, a **frequency domain (FD)** approach turns the problem into a sequence of static problems whose solutions can be turned back to **time domain (TD)** using the inverse Fourier transform (602b).
- Now limiting ourselves to the 1D version wave equation we have,

$$(\imath\omega)^2 \check{u} - a^2 \check{u}_{,xx} = 0 \qquad \text{1D wave equation in frequency domain for a nondispersive media}$$

- We observe that the wave speed a is **frequency (ω) independent**.
- What would change if the material is dispersive?
- In this case the wave speed a will depend on frequency ω,

$$(\imath\omega)^2 \check{u} - a^2(\omega) \check{u}_{,xx} = 0 \qquad \text{1D wave equation in frequency domain for a dispersive media} \tag{605}$$

- That is the wave speed $a(\omega)$ is frequency-dependent.
- For example, in electromagnetics wave speed is given by $a = 1/\sqrt{\epsilon\mu}$, where ϵ and μ are electric permittivity and magnetic permeability, respectively.
- In many realistic materials these material properties (ϵ, μ) are dispersive(*i.e.,* frequency-dependent $\epsilon(\omega), \mu(\omega)$) resulting in many interesting phenomena including dispersive wave speed $a(\omega)$ for example.
- Two question arise in modeling dispersive media:
  - How dispersive constitutive parameters such as ϵ and μ in electromagnetic are expressed? The figure (from [Zhang and Ge, 2009]) shows the three commonly used models (Debye, Lorentz, Drude models) in electromagnetics and in general for dispersive media. A general complex dispersive relations can be approximated with a summation of Debye, Lorentz, and two other types of dispersive relations; *cf.* [Baumann et al., 2009, Viquerat et al., 2013] for example.
  - How are dispersive media modeled / solved in time domain (TD)? This is further discussed below.



[Zhang, 2009]
Real and imaginary part of the relative complex permittivity versus frequency for three typical kinds of dispersive medium.

- Let us consider equation (605), the 1D wave equation in frequency domain for a dispersive media,

$$(\imath\omega)^2 \check{u} - a^2(\omega) \check{u}_{,xx} = 0$$

- As mentioned, dispersive wave speed results from dispersive material properties: $a(\omega) = 1/\sqrt{\epsilon(\omega)\mu(\omega)}$ in electromagnetics, $a(\omega) = \sqrt{K(\omega)}\rho(\omega)$ in acoustics and elastodynamics (K is the bulk modulus and ρ the density; in 2D, 3D elastodynamics we also deal with shear modulus and shear wave speed).
- The solution to (7.6) is straightforward in frequency domain (FD) as for each frequency simply a different wave of wave speed is used.

- However, modeling dispersive models in time domain (TD) becomes challenging.
- There are a few approaches for the modeling / solution of dispersive media in TD from which only 2 are mentioned here:
  - Convolution integrals: To recast (7.6) in TD we need to apply the inverse Fourier transform(that is (602b) on it). This will result on a **convolution integral** on the term involving $a(\omega)$ (and dispersive parts in general). Given that convolution integrals involve integration from time $t = -\infty$ to ∞ the modeling and numerical solution of dispersive media in TD when convolution terms are explicitly models is a challenging task. For more information see [Luebbers et al., 1990, Luebbers and Hunsberger, 1992, Bui et al., 1991] for Recursive Convolution (RC) method and [Kelley and Luebbers, 1996] for Piecewise Linear Recursive Convolution (PLRC) method wherein the convolution term is explicitly modeled in the numerical setting.
  - Auxiliary differential equations (ADEs): In this approach the convolution terms are eliminated by the introduction of additional differential equations to the system. This approach was first introduced by [Kashiwa and Fukai, 1990, Kashiwa et al., 1990, Joseph et al., 1991]. The ADE approach is more flexible and general than those that explicitly maintain the convolution term. It can have higher orders of accuracy, applied for nonlinear problems, and their computational modeling can be much less troublesome.
- Clearly, the overview of these approaches is beyond the scope of this course.
- However, we provide an example how a dispersive response changes the form of PDEs in TD.
- Consider the following wave equation with dispersive wave speed,

$$(\imath\omega)^2\check{u} - a^2(\omega)\check{u}_{,xx} = 0, \qquad \text{for} \qquad a(\omega) = \frac{a_0}{\sqrt{1+\frac{\omega_0}{\imath\omega}}} \tag{606}$$

- We observe that the wave speed $a(\omega)$ is dispersive, *i.e.,* frequency-dependent, and
- as $\omega \to \infty \Rightarrow a(\omega) \to a_0$.
- That is at high frequencies ω the response tends to a nondispersive media with wave speed a_0.
- Now, we are interested in casting (606) in TD.
- By squaring a and multiplying the equation by $(1+\omega_0/(\imath\omega)$ we obtain,

$$\begin{aligned}
(\imath\omega)^2\check{u} - \left(\frac{a_0}{\sqrt{1+\frac{\omega_0}{\imath\omega}}}\right)^2 \check{u}_{,xx} = 0, \qquad & \Rightarrow \\
(\imath\omega)^2\check{u} - a_0^2\frac{1}{1+\frac{\omega_0}{\imath\omega}}\check{u}_{,xx} = 0, \qquad & \Rightarrow \qquad \text{(Multiplying by } 1+\frac{\omega_0}{\imath\omega}\text{)} \\
(\imath\omega)^2\check{u} + \omega_0(\imath\omega)\check{u} - a_0^2\check{u}_{,xx} = 0 \qquad &
\end{aligned} \tag{607}$$

- Equation (607) is the Fourier transform of,

$$u_{,tt} + \omega_0 u_{,t} - a_0^2 u_{,xx} = 0 \tag{608}$$

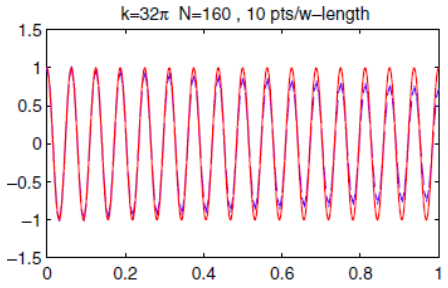
  given that $\check{\dot{u}} = (\imath\omega)\check{u}$ and $\check{\ddot{u}} = (\imath\omega)^2\check{u}$.
- This result is very interesting, showing that for this problem the dispersive velocity corresponds to the addition of a damping term to the wave equation; *cf.* (567) ($u_{,tt} + \omega_0 u_{,t} - a^2 u_{,xx} = 0, \qquad a = \sqrt{\frac{D}{\tau}}, \omega_0 = \frac{1}{\tau}$) and its subsequent harmonic analysis in previous sections.
- Unfortunately, this is an exception where dispersive material properties result to simple addition of differential terms in the nondispersive equation in TD.
- In general, as mentioned before, the pull-back of the equations of dispersive media from FD to TD introduces convolution terms.
- If Auxiliary differential equation (ADEs) are used the pull-back of these equations in time domain will involve additional temporal ODEs that complement an nondispersive equation in TD.
- In this simple case of a damped wave equation, obviously, we do not need to introduce ADEs or convolution terms as the starting equation (damped wave equation) already has a simple form in TD.
- However, it is interesting to observe that the addition of a damping term ($u_{,tt} + \omega_0 u_{,t} - a_0^2 u_{,xx} = 0$ instead of $u_{,tt} - a_0^2 u_{,xx} = 0$) makes the wave equation dispersive by having a dispersive wave speed from (606) ($a(\omega) = \frac{a_0}{\sqrt{1+\frac{\omega_0}{\imath\omega}}}$).

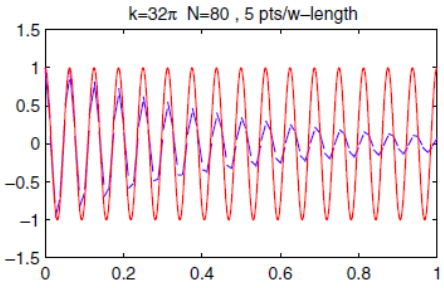
- The same concept for this equation was also discussed in the context of phase velocity where the damped wave equation has a wavenumber (spatial frequency ξ) dependent value. See (597) where $c_p(\xi) = \frac{\omega}{\xi} = a\left\{\pm\sqrt{1-\left(\frac{\omega_0}{2a\xi}\right)^2} - \imath\left(\frac{\omega_0}{2a\xi}\right)\right\}$.
- It is just when we were computing phase velocity we expressed quantities as functions of wavenumber (spatial frequency) ξ rather than temporal frequency in (606) ($a(\omega) = \frac{a_0}{\sqrt{1+\frac{\omega_0}{\imath\omega}}}$).
- In any case, apart from the interesting physics of dispersive media and whether such the response is characterized from wavenumber (ξ) or temporal frequency (ω) perspective, it should be emphasized that the **dispersion and dissipation of waves with given wavenumber ξ (or temporal frequency ω) can depend on ξ (ω)** .
- We will observe that numerical solution of PDEs introduces **numerical dispersion and dissipation** which should be differentiated from the **physical dispersion and dissipation**. *e.g.*,,
- The advection ($u_{,t} + au_{,x} = 0$) and wave equation ($u_{,tt} - a^2u_{,xx} = 0$) are both physically nondissipative and nondispersive.
- However, as we will observe numerical solution of these equations introduces numerical dissipation and numerical dispersion.
- For a numerical methods to be effective, it is important for the errors in dissipation and dispersion not to overshadow physical dissipation and dispersion (if any).

7.7 Numerical dispersion and dissipation

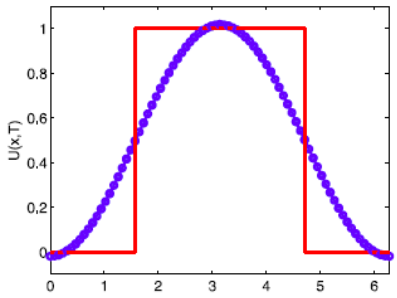
7.7.1 Introduction and motivation

- At the end of §7.6 we discussed the importance of distinguishing between **physical** dispersion and dissipation and **numerical** dispersion and dissipation.
- Depending on the application and accuracy demanded, we require numerical dissipation and dispersion to not overshadow their corresponding physical values.
- For example in many electromagnetic applications dispersion and dissipation of the waves should be very accurately modeled.
- Also, dissipation corresponds to energy loss and in many applications we seek conservative numerical methods (if the underlying physical problem is conservative) or those who have very small numerical dissipation.
- As a side note recall that in §5.4.1 we discussed often we want the numerical method to be dissipative in such manner to dissipate high frequency numerical artifacts.
- To demonstrate the concept of numerical dissipation and dispersion we refer to sample numerical results below.
- In both cases a simple wave equation $u_{,tt} - a^2u_{,xx} = 0$ is corresponds to a nondispersive and non-dissipative (*i.e.*, conservative) response.





[Ainsworth, 2006]
Display of **numerical dissipation**: The numerical solution (blue line) dissipates in time unlike the exact solution (red line)



[Yang, 2013]
Numerical dispersion: The exact square is dispersed numerically.

1. In the two figures on the left from [Ainsworth et al., 2006]. We observe, albeit the numerical having no dispersion error (as wave lengths are not altered), is dissipative.
2. The example from [Yang et al., 2013] displayed the numerical solution of a square wave that is dispersed due to numerical dispersion not physical dispersion (as $u_{,tt} - a^2u_{,xx} = 0$ is nondispersive).
   Interestingly, a square pulse has a very rich Fourier transform. Each component of the Fourier transforms with a different wave speed because of numerical dispersion. Accordingly, the shape of the wave is altered just because of numerical dispersion in this example.

7.7.2 Definition of numerical dispersion and dissipation

- The basis of numerical dispersion and dissipation error is on harmonic wave analysis.
- Consider the simple 1D harmonic solution for a scalar field $u(x,t)$ from (564a),

$$u(x,t) = e^{\imath(\xi x-\omega t)} = e^{\imath(\xi x-\omega_R t)}e^{\omega_I t} \qquad \text{Analytical solution} \tag{609}$$

- **The wavenumber ξ is given** and temporal frequency $\omega = \omega_R + \omega_I$ is **obtained by seeking harmonic solutions for underlying PDE** (plugging (609) in the PDE).
- This process wave done for various scalar and vector equations in §7.2.
- Herein, we focus on a scalar equation, although the treatment in 2D and 3D is exactly the same. That is, in equations (564b) ($u(\mathbf{x},t) = e^{\imath(\boldsymbol{\xi}.\mathbf{x}-\omega t)} = e^{\imath(\boldsymbol{\xi}.\mathbf{x}-\omega_R t)}e^{\omega_I t}$) and (564b) ($\mathbf{u}(\mathbf{x},t) = \boldsymbol{\Phi}e^{\imath(\boldsymbol{\xi}.\mathbf{x}-\omega t)}$) again $\boldsymbol{\xi}$ (now a vector) is fixed for a given direction $\mathbf{n}$ and the dispersion and dissipation of numerical scheme are analyzed for the given direction.
- In any of these cases for a given wavenumber ξ ($\boldsymbol{\xi}$ in 2D/3D) ω_R corresponds to wave propagation mode (dispersion related) and ω_I to evanescent mode (if $\omega_I < 0$) (dissipation related).
- The analytical solution is basically written as $u(x,t) = e^{\imath(\xi x-\omega_R t)}e^{\omega_I t}$; *cf.* (609).
- Now, for numerical solution we solve what **numerical temporal frequency will be for a given wave number** ξ,

$$u^h(x,t)v(x,t) = e^{\imath(\xi x-\omega^h t)} = e^{\imath(\xi x-\omega_R^h t)}e^{\omega_I^h t} \qquad \text{Numerical solution} \tag{610}$$

  where v the notation for approximate solution u^h.
- The determination of numerical dissipation and dispersion will be discussed in §7.7.4.
- For the moment, we focus on the abstract notation of numerical dissipation and dispersion errors as $\Delta\omega := \omega^h - \omega$. That is,

$$\left.\begin{array}{ll} u(x,t) = e^{\imath(\xi x-\omega_R t)}e^{\omega_I t} & \text{Analytical solution} \\ u^h(x,t) = e^{\imath(\xi x-\omega_R^h t)}e^{\omega_I^h t} & \text{Numerical solution} \end{array}\right\} \Rightarrow \left\{\begin{array}{lll} \Delta\omega_R := \omega_R^h - \omega_R & \textbf{Dispersion error} \\ \Delta\omega_I := \omega_I^h - \omega_I & \textbf{Dissipation error} \end{array}\right. \tag{611}$$

- It is clear that dispersion error is related to error in frequency and wave speed (implied by $e^{\imath(\xi x-\omega_R t)}$ is $\frac{\omega_R}{\xi}$).
- Dispersion error is related to how much numerical wave speed differs from analytical wave speed for a given wavenumber.
- On the other hand, dispersion error determines how much more a numerical method is dissipative than the underlying analytical solution. Clearly, $\Delta\omega_I \leq 0$ (which is often is the case) implies higher dissipation in numerical method than analytical solution. For example, if analytical solution is conservative, numerical solution would be either conservative or dissipative.

7.7.3 Numerical dispersion / dissipation to period elongation (PE) / amplitude decay (AD)

- We are interested in turning dispersion and dissipation errors into period elongation and amplitude decay measures which may be considered more physically understandable.
- From (611) we observe that analytical and numerical solutions for a harmonic wave with wavenumber ξ are,

$$\begin{aligned} u(x,t) &= e^{\imath(\xi x-\omega_R t)}e^{\omega_I t} && \text{Analytical solution} \\ u^h(x,t) &= e^{\imath(\xi x-\omega_R^h t)}e^{\omega_I^h t} && \text{Numerical solution} \end{aligned}$$

- The terms $e^{\imath(\xi x-\omega_R t)}$ and $e^{\imath(\xi x-\omega_R^h t)}$ correspond to oscillatory wave propagation with frequencies ω_R^h and ω_R.
- The corresponding exact and numerical periods are defined as,

$$T(\xi) = \frac{2\pi}{\omega_R(\xi)} \qquad \text{Analytical period for wavenumber } \xi \tag{612a}$$

$$T^h(\xi) = \frac{2\pi}{\omega_R^h(\xi)} \qquad \text{Numerical period for wavenumber } \xi \tag{612b}$$

  The same wave we defined dispersion and dissipation errors we define **period elongation (pe)**,

$$\boxed{\Delta T(\xi) := T^h(\xi) - T(\xi) \qquad \text{period elongation (PE)}} \tag{613}$$

- Now we want to relate frequency error to period error by taking the increments of period and frequency of the exact solution to numerical solution,

$$T\omega_R = 2\pi \quad \Rightarrow \quad \Delta(T\omega_R) = \Delta T\omega_R + T\Delta\omega_R + \mathcal{O}\left((\Delta\omega_R)^2\right) = 0 \quad \Rightarrow \quad \text{(diving by } T\omega_R\text{)}$$
$$\frac{\Delta T}{T} = -\frac{\Delta\omega_R}{\omega_R} + \mathcal{O}\left((\Delta\omega_R)^2\right)$$

  Recall $\Delta\omega_R = \omega_R - \omega_R^h$ is the dispersion (frequency) error.

- From the preceding equation an ignoring lower order term $\mathcal{O}\left((\Delta\omega_R)^2\right)$ we obtain,

$$\frac{\Delta\omega_R}{\omega_R} = \frac{\omega_R^h - \omega_R}{\omega_R} \qquad \text{Relative frequency error} \tag{614a}$$
$$\frac{\Delta T}{T} = \frac{T^h - T}{T} = -\frac{\Delta\omega_R}{\omega_R} \qquad \text{Relative period elongation (error)} \tag{614b}$$

- Next we consider the amplitude decay error.
- From (611) we relate analytical ($u(x,t)$) and numerical solutions for a harmonic wave with wavenumber ξ,

$$\left.\begin{aligned} u^h(x,t) &= e^{\imath(\xi x - \omega_R^h t)}e^{\omega_I^h t} = e^{\imath(\xi x - (\omega_R + \Delta\omega_R)t)}e^{(\omega_I + \Delta\omega_I)t} = e^{\imath(\xi x - \omega_R t)}e^{\omega_I t}e^{-\imath\Delta\omega_R t}e^{\Delta\omega_I t} \\ u(x,t) &= e^{\imath(\xi x - \omega_R t)}e^{\omega_I t} \end{aligned}\right\} \quad \Rightarrow$$

$$\boxed{u^h(x,t) = u(x,t)e^{-\imath\Delta\omega_R t}e^{\Delta\omega_I t}} \tag{615}$$

- Now, we are interesting in finding out how much the numerical solution will differ in **one period** of wave propagation T.
- This is achieved by evaluating $u^h(x,t)$ and $u(x,t)$ for a given x and two times t and $t+T$:

$$\left.\begin{aligned} u^h(x,t+T) &= u(x,t+T)e^{-\imath\Delta\omega_R(t+T)}e^{\Delta\omega_I(t+T)} \\ u^h(x,t) &= u(x,t)e^{-\imath\Delta\omega_R t}e^{\Delta\omega_I t} \end{aligned}\right\} \quad \Rightarrow$$

$$\boxed{\overbrace{\frac{u^h(x,t+T)}{u^h(x,t)}}^{\text{numerical sln. ratio}} = \overbrace{\frac{u(x,t+T)}{u(x,t)}}^{\text{exact sln. ratio}} \overbrace{e^{\imath(-\Delta\omega_R T)}}^{\text{phase error}} \overbrace{e^{\Delta\omega_I T}}^{\text{amplitude ratio}}} \tag{616}$$

- We observe that the numerical solution ratio for the same x **after a full cycle** of a wave passes through it (t to $t+T$) is broken to three parts,

  1. **exact solution ratio**: How much the **exact** numerical solution changes in one period. Note that if the underlying exact solution is dissipative the magnitude of the exact solution can decrease for the complex harmonic wave $e^{\imath(\xi x - \omega_R t)}e^{\omega_I t}$ because of the ω_I term if $\omega_I < 0$.
  2. **Phase error**: As shown this is do to the error in capturing correct frequency (period) in the numerical method. This is a complex number of magnitude 1. This type of error is characterized by dispersion (frequency) error / amplitude decay from (614).
  3. **Amplitude ratio**: This corresponds how much the numerical solution is decayed (dissipated) **in addition to physical dissipation** . This is the **net dissipation implied by a numerical method**. Amplitude decay corresponds to the case when $e^{\Delta\omega_I(T)} \leq 1$ that is $\Delta\omega_I < 0$.

- To formalize the definition of amplitude decay, we take the magnitude of the both sides of (616),

$$\left|\frac{u^h(x,t+T)}{u^h(x,t)}\right| = \left|\frac{u(x,t+T)}{u(x,t)}\right| \left|e^{\imath - \Delta\omega_R T}\right| \left|e^{\Delta\omega_I T}\right|, \qquad (|e^{\imath a}| = 1, a \in \mathbb{R})$$
$$\left|\frac{u^h(x,t+T)}{u^h(x,t)}\right| = \left|\frac{u(x,t+T)}{u(x,t)}\right| (1 - A_d) \qquad \text{where} \tag{617a}$$
$$A_d = 1 - \left|e^{\Delta\omega_I T}\right| \tag{617b}$$

- **Amplitude decay** A_d is defined such that when $\left|e^{\Delta\omega_I T}\right| = 1$, that is when numerical solution does not change the magnitude of the exact solution it takes the value of one. In this case $A_d = 0$ as there is no decay.
- Basically $A_d > 0$ specifies what ratio of the wave amplitude decays for one full period of wave propagation.
- Given that $T = 2\pi/\omega_R$ and $e^x = 1 + x + \mathcal{O}\left(x^2\right)$ (*cf.*eq:TaylorExp) from (617b) we have,

$$\boxed{A_d = 1 - e^{2\pi\frac{\Delta\omega_I}{\omega_R}} \approx -2\pi\frac{\Delta\omega_I}{\omega_R}, \qquad \frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\pi}\log(1 - A_d) \approx -\frac{1}{2\pi}A_d} \tag{618}$$

- If $\Delta\omega_I > 0$ then amplitude decay is negative meaning that the numerical solution amplifies the solution relative to what physical dissipation is.
- That is, if the physical problem is conservative (*e.g.,* advection equation $u_{,t} + au_{,x} = 0$ and wave equations $u_{,tt} - a^2u_{,xx} = 0$) for $\Delta\omega_I > 0$ numerical solution will be amplified by roughly the ratio $2\pi\frac{\Delta\omega_I}{\omega_R}$ for each period of the harmonic wave.
- Basically, for a numerical method to be dissipative / conservative relative to the physical problem we need $\Delta\omega_I \leq 0 (A_d \geq 0)$.
- Finally we note that all values discussed ($\Delta\omega_R$ and ΔT for dispersion (frequency) / period error; $\Delta\omega_I$ and A_d for dissipation / amplitude decay errors) **they all depend on wavenumber** ξ which for brevity is omitted from preceding equations.
- In (7.7.4) we quantify these values from amplification factor g which often is obtained from stability analysis.

7.7.4 Determination of numerical dispersion and dissipation

- The analytical dispersion relation (*i.e.,* expression of the form $\omega = \omega(\xi)$) is obtained by seeking analytic harmonic solutions of the form $u(x,t) = e^{\imath(\xi x - \omega t)}$ (*i.e.,* by plugging the harmonic solution in the underlying PDE).
- **Numerical dispersion relation** is also determined in a similar fashion.
- Basically, we seek harmonic solutions of the form (610) ($u(x,t) = e^{\imath(\xi x - \omega^h t)}$) in the formulated numerical method corresponding to the underlying PDE.
- As an illustration assume that in FD /FEM / *etc.* scheme, the numerical solution matches harmonic solution (610) ($u(x,t) = e^{\imath(\xi x - \omega^h t)}$) at grid points (x_m, t_n),

$$u^h(x_m, t_n) = e^{\imath(\xi x_m - \omega^h t_n)} \tag{619}$$

- Now, by considering the solution at two subsequent times, *i.e.,* grid points t_n and t_{n+1} we obtain,

$$\left.\begin{aligned} u^h(x_m, t_n) &= e^{\imath(\xi x_m - \omega^h t_n)} = e^{\imath\xi x_m}e^{-\imath\omega^h t_n} \\ u^h(x_m, t_{n+1}) &= e^{\imath(\xi x_m - \omega^h t_{n+1})} = e^{\imath\xi x_m}e^{-\imath\omega^h t_{n+1}} = e^{\imath\xi x_m}e^{-\imath\omega^h (t_n + k)} = e^{\imath\xi x_m}e^{-\imath\omega^h t_n}e^{-\imath\omega^h k} \end{aligned}\right\} \quad \Rightarrow$$

$$\boxed{u^h(x_m, t_{n+1}) = e^{-\imath\omega^h k}u^h(x_m, t_n)} \tag{620}$$

- Equation (620) basically implies the existence of an amplification factor between two successive time steps.
- This is, the factor $e^{-\imath\omega^h k}$ is the amplification factor that for example we observed in the stability analysis of FD methods.
- That is,

$$\boxed{g = e^{-\imath\omega^h k}} \tag{621}$$

  where g is the FD amplification factor.
- This relation is intuitive, but if the reader is interested in relating it to previous development the explanation is as follows.
- In fact, from the discussion in §6.3.5 this connection is clear. We basically showed that to obtain $g(\xi)$ we simply plug solutions of the form (610) in the PDE and the corresponding amplification factor is basically $g(\xi)$ which is what equation (621) is.
- However, below will be a short overview of the related material from that section.
- Equation (423a) expressed the amplification factor of single-step FD methods,

$$\hat{v}^{n+1}(\xi) = g\hat{v}^n(\xi) \tag{622}$$

  where v is the notation we used for numerical solution of u. That is, $v = u^h$.

- Here $\hat{v}^n$ corresponds to Fourier transform of FD solution for time step t_n.
- The solution for time step t_n is obtained by the inverse Fourier transform of the FD solution from (415b),

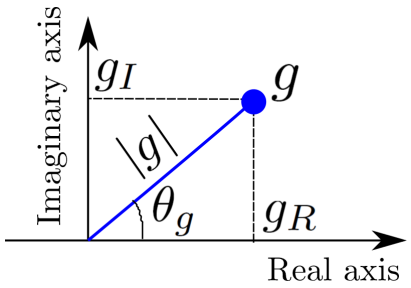
$$v_m^n = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{\imath m h \tilde{\xi}} \hat{v}^n(\tilde{\xi}) \mathrm{d}\tilde{\xi} \tag{623}$$

  where $\tilde{\xi}$ is the wavenumber dummy parameter in the integral.
- Now given that the FD solution including its IC only involves one wavenumber ξ due to the harmonic form in (619) $u^h(x_m, t_n) = e^{\imath(\xi x_m - \omega^h t_n)}$ the integrand in (623) is basically only nonzero (a delta function form) for $\tilde{\xi} = \xi \Rightarrow$
- Basically when only one wavenumber ξ is active in FD solution (*i.e.,* a harmonic wave is solved), (622) takes the form,

$$v_m^{n+1} = g v_m^n \tag{624}$$

- Noting that v_m^n is a shorthand for $u^h(x_m, t_n)$, from (620) and (624) we obtain the obvious relation (621) ($g = e^{-\imath\omega^h k}$).
- Now returning to the main part of the analysis which yields ω_R^h and ω_I^h in terms of g we express g and $e^{-\imath\omega^h}$ in polar coordinate,

$$\left.\begin{aligned} & g = e^{-\imath\omega^h k} && g \text{ in terms of } \omega, \textit{ cf.} \text{ (621)} \\ & g = |g| e^{\imath\theta_g} = |g|\cos\theta_g + \imath|g|\sin\theta_g && \text{Polar coordinate expression of } g \\ & e^{-\imath\omega^h k} = e^{-\imath k\left(\omega_R^h + \imath\omega_I^h\right)} = e^{k\omega_I^h} e^{-\imath k\omega_R^h} \end{aligned}\right\} \Rightarrow$$



$$\boxed{|g| e^{\imath\theta_g} = e^{k\omega_I^h} e^{-\imath k\omega_R^h}} \tag{625a}$$

- By matching complex number magnitudes and phase angles from both sides of (625a), we obtain,

$$|g| = e^{k\omega_I^h} \quad \Rightarrow \quad k\omega_I^h = \log(|g|) \tag{626a}$$

$$e^{\imath\theta_g} = e^{-\imath k\omega_R^h} \quad \Rightarrow \quad k\omega_R^h = -\theta_g \tag{626b}$$

- We relate g polar and Cartesian coordinates as generally g is expressed in terms of its real and imaginary parts g_R and g_I,

$$g = |g| e^{\theta_g} = |g|\cos\theta_g + \imath|g|\sin\theta_g = g_R + \imath g_I \quad \Rightarrow \quad |g| = \sqrt{g_R^2 + g_I^2}, \quad \theta_g = \tan^{-1}\left(\frac{g_I}{g_R}\right) \tag{627}$$

- This from (626) and (627) we conclude,

$$\omega_I^h = \frac{1}{k}\log(|g|) = \frac{1}{2k}\log(|g|^2) = \frac{1}{2k}\log(g_R^2 + g_I^2) \tag{628a}$$

$$\omega_R^h = -\frac{1}{k}\theta_g = \frac{1}{k}\tan^{-1}\left(\frac{-g_I}{g_R}\right) \tag{628b}$$

- For a given FD method we compute g_R and g_I from von Neumann analysis.
- Once ω_I^h, ω_R^h are computed from (628) we compute dispersion and dissipation errors from (611) ($\Delta\omega_R = \omega_R^h - \omega_R$ and $\Delta\omega_I = \omega_I^h - \omega_I$).
- There are a few points to clarify about the value and uniqueness of g,
  - Multi-step methods: As will be discussed in §7.7.7 a multi-step method, even when applied to a temporally first order PDE, may have more than g for the numerical method. The question arises which g to choose in the dispersion error analysis.

    We choose the one g that at the limit of $(h, k) \to 0$ approaches 1 (due to consistency one of the g's satisfy this condition). The other g's are parasitic and either decay much faster than the physical g or from the ICs their contribution is very small from the beginning of the simulation. In short, we take the one g value that corresponds to the exact solution and tends to one as grid size $(h, k) \to 0$; *cf.* the leapfrog example in §7.7.7.

- Higher order temporal PDEs: For the wave equation $u_{,tt} - a^2 u_{,xx} = 0$ there are two ω corresponding to one ξ, namely $\omega = \pm a\xi$; *cf.* (571).

  Corresponding to multiple physical values of ω for the physical solution, there will be multiple non-parasitic g's from the FD solution. For example in the analysis of Central Space Central Time method in §6.4.2 we had two roots for g from (483) ($g^2 - 2(1 - 2\bar{k}^2 \sin^2 \frac{\theta}{2})g + 1 = 0$). One of these roots corresponds to $\omega = a\xi$ and one to $\omega = -a\xi$. In this case both g's are non-parasitic meaning that $g \to 1$ as $(h, k) \to 0$. For dispersion analysis, we correctly match ω and g pairs. In some cases (*e.g.,* this second order PDE), the analysis of one of the pairs would be sufficient due to certain symmetries in the PDE and the underlying numerical method.

- Repeated roots: As we have seen many times if the root ω is a root with multiplicity greater than one, then in addition to simple harmonic waves there are solutions of the same harmonic wave multiplied by polynomials in t. An example was given in the analysis of critically damped, damped wave equation in equation (569b) ($u(x,t) = A_1 e^{\imath\xi x} e^{-\frac{t\omega_0}{2}} + A_2 t e^{\imath\xi x} e^{-\frac{t\omega_0}{2}}$).

  The same form of solution is also observed in FD (or other numerical method) update from time step t_n to t_{n+1} where instead of a simple multiplicative factor relation g^n the repeated roots have an additional polynomial in n in the update equation as shown (495) ($\hat{v}^n = \sum_{r=1}^{l} p_r(n) g_r^n$). Obviously, from the discussion on multi-step methods, and temporally higher order PDEs, each time only one non-parasitic term from the summation above will be matched with its corresponding ω in dispersion analysis.

  The main point related to repeated roots is that both exact and numerical results have simple harmonic waves (those without multiplying polynomials of order 1 and higher in t and n) which fall into the dispersion analysis discussed above. That is, we can proceed with the same solution proceed at least for the simple harmonic wave types when repeated roots are encountered for a given ξ.

- Finally, we note that often we are interested in **dispersion and dissipation convergence rates** which correspond to the limit $(h, k) \to 0$. The analysis from §7.7.5 shows how convergence rates are obtained.

7.7.5 A sample dispersion / dissipation analysis for FTBS FD method

- To compute dispersion and dissipation errors from (611),

$$\Delta\omega_I = \omega_I^h - \omega_I \quad \text{(Dissipation error).} \qquad \Delta\omega_R = \omega_R^h - \omega_R, \quad \text{(Dispersion error)}$$

  we do the following,

  1. Compute ω_I and ω_R: Using harmonic analysis from §7.2 to compute ω_I and ω_R for a given PDE.
  2. Compute ω_I^h and ω_R^h: Recall equation (628) where ω_I^h and ω_R^h were given by the amplification factor g,

$$\omega_I^h = \frac{1}{2k} \log(g_R^2 + g_I^2), \qquad \omega_R^h = \frac{1}{k} \tan^{-1}\left(\frac{-g_I}{g_R}\right)$$

  this in turn requires the computation of g which can be done by von Neumann analysis for FD methods.

- We demonstrate this process for the advection equation (553a) where by seeking harmonic solutions of the form (564a) ($u(x,t) = e^{\imath(\xi x - \omega t)}$) we obtained ω_R, ω_I in (565b),

$$u_{,t} + a u_{,x} = 0, \qquad \text{and} \qquad u(x,t) = e^{\imath(\xi x - \omega t)} \qquad \Rightarrow \omega_I = 0, \quad \omega_R = a\xi \tag{629}$$

- To obtain ω_R^h, ω_I^h from g we refer to the values of amplification factor g computed for the advection equation in §6.3.7. Below, is the summary of such results,

FD scheme	g_R	g_I	$\lvert g\rvert^2 = g_R^2 + g_I^2$	Equation source	
FTBS	$(1 - \bar{k}) + \bar{k}\cos\theta$	$-\bar{k}\sin\theta$	$1 - 4\bar{k}(1 - \bar{k})\sin^2\frac{\theta}{2}$	(428a), (430)	(630a)
Lax-Friedrichs	$\cos\theta$	$-\bar{k}\sin\theta$	$\cos^2\theta + \bar{k}^2\sin^2\theta$	(453)(454)	(630b)
Leapfrog	$\pm\sqrt{1 - \bar{k}^2\sin^2\theta}$	$-\bar{k}\sin\theta$	1	(469), (471)	(630c)

- Recall that formula for normalized time step for advection equation from (28) and θ from (427)

$$\bar{k} = a\frac{k}{h} \qquad \text{Nondimensional time step} \tag{631a}$$

$$\theta = h\xi \qquad \text{Nondimensional spatial resolution} \tag{631b}$$

- For explicit methods $\bar{k} \leq 1$. At times stability analysis requires even more stringent values. The value of 1 corresponds to the maximum possible value $\bar{k}$ can take.
- From the stability analysis in §6.4.1 and §6.3.7 we recall that $\bar{k} \leq 1$ for FTBS ($a > 0$) and Lax-Friedrichs methods and $\bar{k} < 1$ for the leapfrog method .
- The parameter θ also takes values in the interval $\theta \in [-\pi, \pi]$ where the value $\theta \to 0$ corresponds to $h \to 0$ (the fine resolution) and $|\theta| = \pi$ to the coarsest grid, *i.e.,* the highest number ξ of waves that the discrete FD grid can represent; *cf.* §6.3.1, (415b), (443), and §6.3.5.
- The analysis for the FTBS scheme is given in this section and that for leapfrog method in §7.7.7.
- For the FTBS scheme applied to the advection equation we use the values $g_R = (1-\bar{k}) + \bar{k}\cos\theta$, $g_I = -\bar{k}\sin\theta$, $|g|^2 = g_R^2 + g_I^2 = 1 - 4\bar{k}(1-\bar{k})\sin^2\frac{\theta}{2}$ (from (630a)) in (628) to obtain,

$$\omega_I^h = \frac{1}{2k}\log(g_R^2 + g_I^2) = \frac{1}{2k}\log\left(1 - 4\bar{k}(1-\bar{k})\sin^2\frac{\theta}{2}\right) \tag{632a}$$

$$\omega_R^h = \frac{1}{k}\tan^{-1}\left(\frac{-g_I}{g_R}\right) = \frac{1}{k}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{(1-\bar{k}) + \bar{k}\cos\theta}\right) \tag{632b}$$

- Finally, from the analytical solutions for advection equation $\omega_I = 0$ and $\omega_R = a\xi$ from (629) and ω_I^h, ω_R^h from (632) we compute dispersion and dissipation errors from (611),

$$\Delta\omega_I = \omega_I^h - \omega_I = \frac{1}{2k}\log\left(1 - 4\bar{k}(1-\bar{k})\sin^2\frac{\theta}{2}\right) \tag{633a}$$

$$\Delta\omega_R = \omega_R^h - \omega_R = \frac{1}{k}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{(1-\bar{k}) + \bar{k}\cos\theta}\right) - a\xi \tag{633b}$$

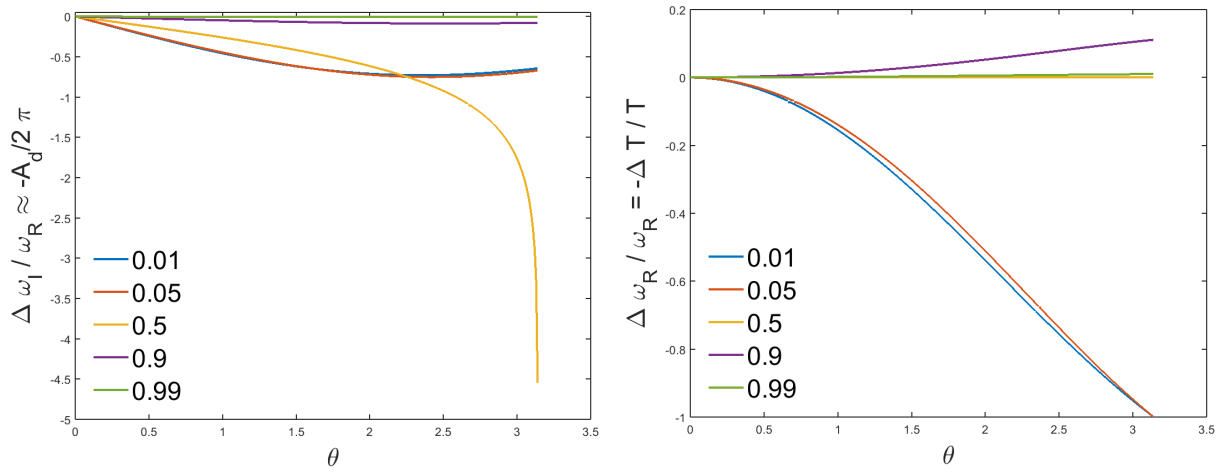
- It may be more useful to express these errors in nondimensional and normalized form.
- We express these errors in the form $\Delta\omega_I k$ and $\Delta\omega_R/\omega_R$ and noting that $1/(k\omega_R) = 1/(ka\xi) = \frac{h}{ak}\frac{1}{h\xi} = \frac{1}{\bar{k}\theta}$ (*cf.* (631)) we observe,

$$\omega_I^h k = \frac{1}{2}\log\left(1 - 4\bar{k}(1-\bar{k})\sin^2\frac{\theta}{2}\right) \quad\Rightarrow\quad \frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\pi}\log(1-A_d) \approx -\frac{1}{2\pi}A_d = \frac{1}{2\bar{k}\theta}\log\left(1 - 4\bar{k}(1-\bar{k})\sin^2\frac{\theta}{2}\right) \tag{634a}$$

$$\frac{\omega_R^h}{\omega_R} = \frac{1}{\bar{k}\theta}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{(1-\bar{k}) + \bar{k}\cos\theta}\right) \quad\Rightarrow\quad \frac{\Delta\omega_R}{\omega_R} = -\frac{\Delta T}{T} = \frac{1}{\bar{k}\theta}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{(1-\bar{k}) + \bar{k}\cos\theta}\right) - 1 \tag{634b}$$

note that we have used (614b) ($\frac{\Delta T}{T} = \frac{T^h - T}{T} = -\frac{\Delta\omega_R}{\omega_R}$) for relative period elongation (error) and (618) ($A_d = 1 - e^{2\pi\frac{\Delta\omega_I}{\omega_R}} \approx -2\pi\frac{\Delta\omega_I}{\omega_R}$ *i.e.,* $\frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\pi}\log(1-A_d) \approx -\frac{1}{2\pi}A_d$) for amplitude decay.

- The normalization of $\Delta\omega_I$ by k instead of ω_I is due to the fact that $\omega_I = 0$. We could have normalized $\Delta\omega_I$ by ω_R which will be done in §7.7.6.
- The plots below show $\frac{\Delta\omega_I}{\omega_R} = -\frac{1}{2\pi}A_d =$ and $\frac{\Delta\omega_R}{\omega_R} = -\frac{\Delta T}{T}$ for (634) from different normalized time steps $\bar{k} = 0.01, 0.05, 0.5, 0.9, 0.99$ versus normalized spatial resolution $\theta \in [0, \pi]$.



Some comments on the results are,

- Range of θ: Instead of plotting the results for the range of $\theta \in [-\pi, \pi]$ we only consider the range of $[0, \pi]$. This is due to the fact that both $\Delta\omega_I k$ and $\frac{\omega_R^h}{\omega_R}$ are even in θ.
- $\Delta\omega_I \leq 0$:
  - We observe that $\Delta\omega_I/\omega_R \leq 0$ which means numerical results are dissipative (given that the advection equation itself is conservative).
  - This is no surprise because in analyzing the stability of the advection equation in *cf.* discussion below (430) and §6.3.7 we in fact required $|g| \leq 1$.
  - However, from (633a) $|g|^2 = g_R^2 + g_I^2 \leq 1$ corresponds to $\omega_I^h k = \Delta\omega_I k \leq 0$.
- Behavior for $\bar{k} = 0.5$: The results for this case are shown in orange. From the figure on the left we observe the method is highly dissipative ($\Delta\omega_I k \searrow$ quickly as $\theta \to \pi$) and from the one on the right $\frac{\omega_R^h}{\omega_R} = 1$ that is for for $\bar{k} = 0.5$ FTBS scheme has zero dispersion error. These are demonstrated analytically below using (634),

$$\bar{k} = 0.5 \quad \text{and} \quad \Delta\omega_I k = \frac{1}{2}\log\left(1 - 4\bar{k}(1-\bar{k})\sin^2\frac{\theta}{2}\right) \quad \Rightarrow$$

$$k\Delta\omega_I(\bar{k}=0.5,\theta) = \frac{1}{2}\log\left(1-\sin^2\frac{\theta}{2}\right) = \frac{1}{2}\log\left(\cos^2\frac{\theta}{2}\right) = \log\left(\cos\frac{\theta}{2}\right) \quad \Rightarrow \quad \lim_{\theta\to\pi} k\Delta\omega_I(\bar{k}=0.5,\theta) = \infty \tag{635a}$$

$$\bar{k} = 0.5 \quad \text{and} \quad \frac{\omega_R^h}{\omega_R} = \frac{1}{\bar{k}\theta}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{(1-\bar{k})+\bar{k}\cos\theta}\right)$$

$$\frac{\omega_R^h(\bar{k}=0.5,\theta)}{\omega_R} = \frac{1}{(1/2)\theta}\tan^{-1}\left(\frac{\sin\theta}{(1+\cos\theta)}\right) = \frac{2}{\theta}\tan^{-1}\left(\frac{2\sin(\theta/2)\cos(\theta/2)}{2\cos^2(\theta/2)}\right) = \frac{2}{\theta}\tan^{-1}\left(\tan(\frac{\theta}{2}\right) = \frac{2}{\theta}\frac{\theta}{2} \quad \Rightarrow$$

$$\omega_R^h(\bar{k}=0.5,\theta) = \omega_R \tag{635b}$$

  - These results match what can be observed in the plots.
  - That is for half-stability limit time step $\bar{k} = 0.5$ FTBS FD scheme for advection equation has **zero dispersion error**.
  - The method also is highly dissipative for high frequency content of the solution relative to grid size (*i.e.,* for $\theta = \xi h$ $\theta \to \pm\pi$).
  - Accordingly, if a coarse grid h is used relative to the frequency of the data (IC), FTBS may be overly dissipative for $\bar{k} = 0.5$.
- **Non-dispersive and non-dissipative (*i.e.*, conservative) response for** $\bar{k} = 1$:
  - By a simple calculation in (634) we observe that for $\bar{k} = 1$ we have $\Delta\omega_I = \omega_I^h = 0$ and $\Delta\omega_R = 0$ ($\omega_R^h = \omega_R$).
  - That is the FTBS FD scheme is non-dispersive and non-dissipative for $\bar{k} = 1$ at the limit of CFL condition for this explicit method.
  - This behavior can be observed in many other cases where the numerical method is non-dispersive and conservative right at the CFL limit.
  - We are fortunate in this case that stability condition actually permits going all the way to $\bar{k} = 1$ as in some cases explicit method have a more stringent stability limit ($\bar{k} < 1$).
  - Interestingly, for this case the FD scheme **exactly advects any IC** (with the resolution h sampled at IC) **without dispersing or dissipating the wave**.
  - The message from this example is, in many cases we want to stay close to the CFL limit $\bar{k} = 1$ (if stability permits) as often it corresponds to small or zero dispersive and dissipative errors.
  - In 2D and 3D we generally do not preserve such an interesting feature, however, still staying close to CFL limit $\bar{k} = 1$ often is a good idea.
  - That is in mesh refinement and convergence to the exact solution it is recommended to **fix** $\bar{k} \lesssim 1$ **and let** $h \to 0$.
  - Having better dispersion and dissipation errors for $\bar{k} = 0.9, 0.99$ compared to $\bar{k} = 0.01$ and $\bar{k} = 0.05$ in the plots also support this assertion (although this feature many not hold for all numerical methods).
  - Also by having $\bar{k} \lesssim 1$ rather than $\bar{k} \to 0$ we are taking much larger time steps which correspond to much lower computational costs. This is another advantage of taking time steps close to CFL condition of 1 for explicit methods.
- Converge rates for dissipation and dispersion results / behavior as $h \to 0$ ($\theta \to 0$)
  - While for stability analysis we consider $\theta \in [-\pi, \pi]$ from accuracy perspective **we want to limit** $|\theta|/5$; *cf.* §6.3.5 .

- Again as mentioned in §6.3.5 (and depicted in θ 3 schematics), $|\theta|/5$ corresponds to at least 10 grid spacings/elements are recommended per wavelength which as suggested in [Shakib and Hughes, 1991] is the minimum reasonable resolution of the spatial grid.
- Recall $\theta = h\xi$, and ξ is dictated by the wave length of the solution features needed to be solved. This $\theta \leq 1/5$ results in refining h based on the highest relevant / important wavelengths in the problem.
- Thus, **in the plots depicted in practice we are dealing with ranges** $\theta \lesssim \pi/5 \approx 0.6$ where dispersion and dissipation errors are actually small.
- In the small range of $\bar{k}$ ($\bar{k} \to 0$) we can in fact obtain **convergence rates** of **dispersion and dissipation errors** which is discussed in §7.7.6 for FTBS FD scheme applied to advection equation.
- Having convergence rates of dissipation and dispersion errors enable us to determine how fast these errors tend to zero as the mesh is refined which is of invaluable practical importance in choosing the right grid size for achieving desired level of accuracy and designing effective adaptive operations.

7.7.6 Orders of convergence for dispersion / dissipation errors

- As mentioned at the end of §7.7.5 in practice from accuracy perspectives θ is a small number and as mentioned it is recommended to take values $\theta \lesssim \pi/5 \approx 0.6$.
- We want to investigate dispersion and dissipation errors in the limit $h, k \to 0$ ($\theta \to 0$ for a given $\bar{k}$).
- We recall the values of $\Delta\omega_I, \Delta\omega_R$ from (634),

$$\frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\bar{k}\theta}\log\left(1 - 2\bar{k}(1-\bar{k})(1-\cos\theta)\right) \tag{636a}$$

$$\frac{\Delta\omega_R}{\omega_R} = \frac{1}{\bar{k}\theta}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{(1-\bar{k}) + \bar{k}\cos\theta}\right) - 1 \tag{636b}$$

  where we have used $\sin^2\frac{\theta}{2} = (1-\cos\theta)/2$ in (632a).
- Now we need to expand these errors in powers of θ using Taylor series expansion of certain functions appearing in (636).
- In the following expansion of (636) we refer to Taylor expansion of some of the following functions,

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + x^4 + \cdots \qquad x \in (-1, 1) \tag{637a}$$

$$\log(1+x) = \sum_{n=0}^{\infty}(-1)^{n+1}\frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \qquad x \in (-1, 1] \tag{637b}$$

$$e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \qquad x \in \mathbb{R} \tag{637c}$$

$$\cos x = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} + \cdots \qquad x \in \mathbb{R} \tag{637d}$$

$$\sin x = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} + \cdots \qquad x \in \mathbb{R} \tag{637e}$$

$$\tan x = x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + +\frac{17}{315}x^7 + \frac{62}{2835}x^9 + \cdots \qquad x \in \mathbb{R}\ \&\ x \neq \pi\frac{2m-1}{2} \tag{637f}$$

$$\tan^{-1} x = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} + \cdots \qquad x \in [-1, 1] \tag{637g}$$

- Given that we are considering the case $\theta \to 0$ it is easy to see all the Taylor expansions will be convergent in the equations in the remainder of the section.
- We first start with the expansion of $\Delta\omega_I k$,

$$\frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\bar{k}\theta}\log\left(1 - 2\bar{k}(1-\bar{k})(1-\cos\theta)\right) \qquad \text{(from (637b))} \tag{638}$$

$$= \frac{1}{2\bar{k}\theta}\left\{\left(-2\bar{k}(1-\bar{k})(1-\cos\theta)\right) + \left(-2\bar{k}(1-\bar{k})(1-\cos\theta)\right)^2\right\} \qquad \text{(from (637d))} + \mathcal{O}\left(\theta^6\right) \tag{639}$$

$$= \frac{1}{2\bar{k}\theta}\left(-2\bar{k}(1-\bar{k})(1-\{1-\theta^2/2+\mathcal{O}(\theta^4)\})+\mathcal{O}(\theta^3)\right) \tag{640}$$

$$= -\frac{1}{2\bar{k}\theta}\bar{k}(1-\bar{k})\theta^2+\mathcal{O}(\theta^3) \qquad \Rightarrow \quad \text{(from (634a))} \tag{641}$$

$$\frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\pi}\log(1-A_d) \approx -\frac{1}{2\pi}A_d = -\frac{1-\bar{k}}{2}\theta+\mathcal{O}(\theta^3) \quad \text{and} \quad \omega_I^h k = -\bar{k}\frac{1-\bar{k}}{2}\theta^2+\mathcal{O}(\theta^4) \tag{642}$$

where we have used $\Delta\omega_I = \omega_I^h - \omega_I = \omega_I^h$ and $\omega_I^h k = \frac{\Delta\omega_I}{\omega_R}\omega_R k = \frac{\Delta\omega_I}{\omega_R}(\bar{k}\theta)$ since $\omega_R k = a\xi k = \frac{ka}{h}(h\xi) = \bar{k}\theta$; *cf.* (565a), (427), (28).

- Next compute the relative dispersion error from (634b),

$$\begin{aligned}
\frac{\omega_R^h}{\omega_R} &= \frac{1}{\bar{k}\theta}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{(1-\bar{k})+\bar{k}\cos\theta}\right) = \frac{1}{\bar{k}\theta}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{1-\bar{k}(1-\cos\theta)}\right) && \text{(using (637e)(637a))}\\
&= \frac{1}{\bar{k}\theta}\tan^{-1}\left(\bar{k}\left\{\theta-\frac{\theta^3}{6}+\mathcal{O}(\theta^5)\right\}\left\{1+\bar{k}(1-\cos\theta)+\left(\bar{k}(1-\cos\theta)\right)^2+\mathcal{O}\left(\left(\bar{k}(1-\cos\theta)\right)^3\right)\right\}\right) && \text{(using (637d))}\\
&= \frac{1}{\bar{k}\theta}\tan^{-1}\left(\bar{k}\left\{\theta-\frac{\theta^3}{6}+\mathcal{O}(\theta^5)\right\}\left\{1+\frac{\bar{k}\theta^2}{2}+\mathcal{O}(\theta^4)\right\}\right)\\
&= \frac{1}{\bar{k}\theta}\tan^{-1}\left(\bar{k}\theta+\bar{k}\theta^3\left(\frac{\bar{k}}{2}-\frac{1}{6}\right)+\mathcal{O}(\theta^5)\right) && \text{(using (637g))}\\
&= \frac{1}{\bar{k}\theta}\left(\left\{\bar{k}\theta+\bar{k}\theta^3\left(\frac{\bar{k}}{2}-\frac{1}{6}\right)+\mathcal{O}(\theta^5)\right\}-\frac{1}{3}\left\{\bar{k}\theta+\bar{k}\theta^3\left(\frac{\bar{k}}{2}-\frac{1}{6}\right)+\mathcal{O}(\theta^5)\right\}^3+\mathcal{O}(\theta^5)\right)\\
&= \frac{1}{\bar{k}\theta}\left(\bar{k}\theta+\bar{k}\theta^3\left(-\frac{1}{3}\bar{k}^2+\frac{\bar{k}}{2}-\frac{1}{6}\right)+\mathcal{O}(\theta^5)\right) \qquad \Rightarrow\\
\frac{\Delta\omega_R}{\omega_R} &= \frac{\omega_R^h}{\omega_R}-1 = \left(1-\frac{\theta^2}{6}\left(2\bar{k}^2-3\bar{k}+1\right)+\mathcal{O}(\theta^4)\right)-1
\end{aligned}$$

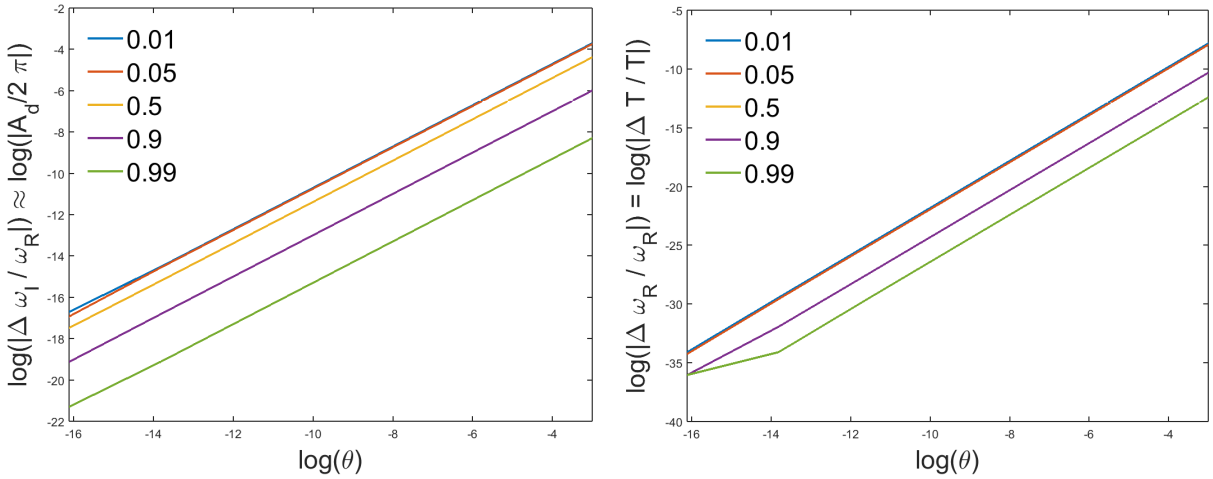
by combining the preceding result and repeating (642) we summarize dissipation / amplitude decay and dispersion / frequency / period errors

$$\frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\pi}\log(1-A_d) \approx -\frac{1}{2\pi}A_d = -\frac{1-\bar{k}}{2}\theta+\mathcal{O}(\theta^3) \quad \text{and} \quad \omega_I^h k = -\bar{k}\frac{1-\bar{k}}{2}\theta^2+\mathcal{O}(\theta^4) \tag{643a}$$

$$\frac{\Delta\omega_R}{\omega_R} = -\frac{\Delta T}{T} = -\frac{\theta^2}{6}(1-\bar{k})(1-2\bar{k})+\mathcal{O}(\theta^4) \tag{643b}$$

- Some interesting observation about the plots are,
  - Convergence rates: We observe the convergence rates are as follows,
    - Dissipation error in the form $\omega_I^h k$ is one in θ (and h since $\theta = \xi h$).
    - Dissipation error and amplitude decay A_d Form $\frac{\Delta\omega_I}{\omega_R} = -\frac{1}{2\pi}A_d$ both have second convergence rate.
    - Dispersion (frequency) error and period elongation in normalized form have second convergence rate in θ (and h).
  - Factors of the leading terms of errors:
    - The factor of the leading term in $\frac{\Delta\omega_I}{\omega_R}$ is $-\frac{1-\bar{k}}{2}\theta$ which is zero for $\bar{k}=1$. This is consistent with $\Delta\omega_I = 0$ for $\bar{k}=1$. This was discussed under "Non-dispersive and non-dissipative (*i.e.*,, conservative) response for $\bar{k}=1$" at the end of §7.7.5. In fact, the higher order terms (in θ) Taylor expansion of $\Delta\omega_I$ also will have factors of $1-\bar{k}$.
    - $-\frac{\theta^2}{6}(1-\bar{k})(1-2\bar{k})$ the leading term in the expansion of $\frac{\Delta\omega_R}{\omega_R}$ is zero for $\bar{k}=\frac{1}{2}$ and $\bar{k}=1$. Again, in this case we observe that in fact $\Delta\omega_R$ is identically zero for $\bar{k}=0.5$ (*cf.* (635b)) and $\bar{k}-1$ (again the discussion "Non-dispersive and non-dissipative (*i.e.*,, conservative) response for $\bar{k}=1$" at the end of §7.7.5).

    In general the leading terms of dissipation / dispersion errors may be zero for specific values of $\bar{k}$. Obviously, if such errors are identically zero for the given $\bar{k}$ values those are preferred for the analysis (*e.g.*, $\bar{k}=1$ for FTBS FD scheme). Otherwise, still by improving the order of convergence of dispersion or dissipation errors for such special values of $\bar{k}$ we can greatly improve the performance of the numerical method.
  - Dissipative response for $\bar{k}<1$: By investigating (643a) we observe that dissipation error $\Delta\omega_I \leq 0$ and amplitude decay A_d which based on their definition correspond to a dissipative numerical solution. If $\bar{k}=1$ dissipation error is zero and for $\bar{k}<1$ is a positive number.
- The following plots depict convergence plots for dissipation and dispersion errors.



- For the range of θ depicted the straight lines in the log-log plot imply reaching the asymptotic convergence rates (for $\theta \to 0$). We observe that convergence rates in (643) hold in the plots.

7.7.7 Dispersion / dissipation errors for multi-step methods / parasitic modes

7.7.7.1 Limiting value of amplification factor as mesh grid size tends to zero

Let us recall that for the leapfrog method the two roots for the amplification factor from (469) were $g_+ = -\imath\bar{k}\sin\theta + \sqrt{1-\bar{k}^2\sin^2\theta}$ and $g_- = -\imath\bar{k}\sin\theta - \sqrt{1-\bar{k}^2\sin^2\theta}$. g_+ corresponds to physical mode while g_- a parasitic mode as their limiting values when $\theta \to 0$ are 1 and -1 respectively. Below, we describe why for the physical mode $\lim_{\theta\to 0} = 1$.
Consider that we have a one step method. From von Neumann analysis we always have a relation of the form

$$\hat{v}^{n+1}(\xi) = g(\theta)\hat{v}^n(\xi) \tag{644}$$

for $\theta = h\xi$. Below we describe why,

$$\lim_{\theta\to 0} g(\theta) = 1, \qquad \text{for a a single-step method} \tag{645}$$

- From consistency we know that as $h, k \to 0$ $v_m^{n+1} \to v_m^n$.
- So for harmonic waves used in von Neumann stability analysis (*cf.* (445)) we have $v_m^{n+1} = e^{\imath m\theta}\hat{v}^{n+1} \to v_m^n = e^{\imath m\theta}\hat{v}^n$ as $h, k \to 0$.
- That is, $\hat{v}^{n+1} \to \hat{v}^n$ as $h, k \to 0$.
- Given the definition $\hat{v}^{n+1} = g\hat{v}^n$ above we have,
- $g \to 1$ as $h, k \to 0$.
- But when $\theta = \xi h \to 0$ (for a given ξ considered in stability analysis) we have $h \to 0$.
- k also tends to zero as $\theta \to 0$ because $\bar{k} = ka/h$ is bounded (assumed bounded independent of h for the analysis of an implicit method and for explicit method it must be at most one from the CFL condition) so $\theta \to 0 \Rightarrow h \to 0 \Rightarrow k0$.
- In summary,

$$\left.\begin{array}{ll} \theta \to 0 \Rightarrow h, k \to 0 & (\theta = \xi h, k = \bar{k}h/a, \bar{k} \text{ bounded, } e.g., \leq 1 \text{ for explicit methods}) \\ \lim_{h,k\to 0} g = 1 & \text{consistency} \end{array}\right\} \Rightarrow \boxed{\lim_{\theta\to 0} g(\theta) = 1} \tag{646}$$

7.7.7.2 Parasitic modes: Introduction

- Let us consider limiting values of g for a variety of methods considered so far,

$$g(\theta) = \left[(1-\bar{k}) + \bar{k}\cos\theta\right] - \imath\left[\bar{k}\sin\theta\right] \qquad \text{FTBS} \tag{647a}$$

$$g(\theta) = \left[\cos\theta\right] - \imath\left[\bar{k}\sin\theta\right] \qquad \text{Lax-Friedrichs} \tag{647b}$$

$$g_\pm(\theta) = \left[\pm\sqrt{1-\bar{k}^2\sin^2\theta}\right] - \imath\left[\bar{k}\sin\theta\right] \qquad \text{Leapfrog} \tag{647c}$$

in which the advection equation $u_{,t} + au_{,x} = 0$ is considered.

- It is clear that for the single-step methods of FTBS and Lax-Friedrichs $\lim_{\theta\to 0} g(\theta) = 1$.
- For the two step leapfrog method also has one root g_+ satisfying $\lim_{\theta\to 0} g_+(\theta) = 1$.
- However, for the other root g_- we have $\lim_{\theta\to 0} g_-(\theta) = -1$.
- Unfortunately, the appearance of the parasitic mode can have adverse effects in numerical solutions.
- The source of these modes is having a larger "historical data pool" with multi-step methods.
- Multi-step methods provide many benefits, including the potential to increase the order of accuracy in time but one unwanted outcome is the appearance of the parasitic modes.
- As a minimum we want the parasitic modes to disappear in the limit of mesh refinement.
- In the remainder of this section we demonstrate that this is in fact the case through the analysis of leapfrog method.
- Now we emphasize that it is not always the case that only one g corresponds to a physical amplification factor.
- For example consider the FTCS scheme applied to the wave equation $u_{,tt} - a^2 u_{,xx} = 0$ from §6.4.2.
- For this temporally second order PDE, the von Neumann analysis resulted in the second order polynomial on g (483) $g^2 - 2A_1 g + A_2 = 0$ for $A_1 = 1 - 2\bar{k}^2 \sin^2\frac{\theta}{2}$ and $A_2 = 1$.
- The solution for g was given in (484) $g_\pm = \left[1 - 2\bar{k}^2 \sin^2\frac{\theta}{2}\right] \pm \left[2\bar{k}\sin\frac{\theta}{2}\sqrt{\bar{k}^2\sin^2\frac{\theta}{2} - 1}\right]$.
- In the limit of $\theta \to 0$ (in fact for $\theta \leq 1$) the values are

$$g_\pm = \left[1 - 2\bar{k}^2 \sin^2\frac{\theta}{2}\right] \pm \imath \left[2\bar{k}\sin\frac{\theta}{2}\sqrt{1 - \bar{k}^2\sin^2\frac{\theta}{2}}\right], \qquad \text{for} \quad \theta \leq 1 \tag{648}$$

- It is clear that unlike leapfrog method for **both roots we have** $\lim_{\theta\to 0} g_\pm(\theta) = 1$.
- This in fact makes perfect sense because for the wave equation physically for a given ξ there are are two harmonic waves one moving to the left and one to the right with speeds $\pm a$ and **each g corresponds to one of these harmonic waves**.
- That there where two waves physically moving with speeds $\pm a$ was demonstrated in (595b) where for the wave equation $u_{,tt} - a^2 u_{,xx} = 0$ a harmonic solution $u(x,t) = e^{\imath(\xi x - \omega t)}$ resulted in wave speeds (phase velocities)$c_p(\xi)\frac{\omega(\xi)}{\xi} = \pm a$.
- In short, if a multi-step method has the number of steps more than temporal order of the PDE, we can have parasitic modes.
- For example, for leapfrog method temporal order is one and number of steps two resulting in one parasitic mode.
- However, for FTCS applied to the wave equation both number of steps and temporal orders were two resulting in no parasitic modes.
- Our goal is to show that the solution corresponding to the parasitic modes tend to disappear in the limit of mesh refinement.
- Accordingly, dispersion and dissipation of the method is governed by its physical amplification factors g, at least in the limit of mesh refinement where we are interested in the convergence rate of dispersion and dissipation errors.

7.7.7.3 When parasitic modes are negligible?

- In §7.7.7.1 and §7.7.7.2 we discussed that in FD discretized form, some of the amplification factors and their corresponding terms in the expansion of $\hat{v}^n$ are parasitic.
- Ideally, we prefer these parametric modes do not contribute much to the solution and at least in the limit of mesh refinement disappear.
- Fortunately, the contribution of parasitic modes diminish in the limit of the mesh refinement. This is demonstrated by the analysis of the leapfrog method.

- As a short overview, we refer to the solution of advection equation $u_{,t} + au_{,x} = 0$ with leapfrog method (464)

$$\frac{v_{n+1}^m - v_{n-1}^m}{2k} + a\frac{v_n^{m+1} - v_n^{m-1}}{2h} = 0 \qquad \Rightarrow$$
$$v_{n+1}^m = \bar{k}(v_n^{m-1} - v_n^{m+1}) + v_{n-1}^m k$$

  where $\bar{k} = a\frac{k}{h}$.

- The von Neumann analysis from §6.4.1 yielded, (466)

$$\hat{v}^{n+1} = -2\imath\bar{k}\sin\theta\hat{v}^n + \hat{v}^{n-1} \tag{649}$$

  note that to obtain the above equation we plug harmonic solution (445) $v_m^n = e^{\imath m\theta}\hat{v}^n$ in the FD equations *cf.* §6.3.6.

- Following the discussion in §6.4.1, assuming that time step update taking the form $\hat{v}^{n+1} = g\hat{v}^n$ we obtained the second order equation (468),

$$g^2 + (2\imath\bar{k}\sin\theta)g - 1 = 0$$

- The roots of this equation are (469),

$$g_+ = -\imath\bar{k}\sin\theta + \sqrt{1 - \bar{k}^2\sin^2\theta}$$
$$g_- = -\imath\bar{k}\sin\theta - \sqrt{1 - \bar{k}^2\sin^2\theta}$$

- As explained in §7.7.7.1 and §7.7.7.2 the root g_- generates a **parasitic mode** implied by the fact that $\lim_{\theta\to 0} g_-(\theta) = -1 \neq 1$.

- Through equations (470) and (476) we observed that the solution to (649) can be written in the following form,

$$\hat{v}^n = A_+(\xi)g_+^n(\theta) + A_-(\xi)g_-^n(\theta) \qquad g_- \neq g_+ \tag{650a}$$
$$\hat{v}^n = A(\xi)g^n + A_*(\xi)ng^n(\theta) \qquad g_- = g_+ = g \tag{650b}$$

  the solution (650b) for $g_- = g_+$ means that a simple multiplicative relation $\hat{v}^{n+1} = g\hat{v}^n$ is no longer correct in this case.

- However, (649) admits solution of the form (650) in general; *cf.* §6.4.3 for general solution of recursive relations.

- In any case, we can write the solution of (650) in the form,

$$\hat{v}^n = A(\xi)g_+^n(\theta) + B(\xi)\left[\frac{g_-^n(\theta) - g_+^n(\theta)}{g_-(\theta) - g_+(\theta)}\right] \qquad g_- \neq g_+ \tag{651a}$$
$$\hat{v}^n = A(\xi)g_+^n(\theta) + +B(\xi)ng^{n-1}(\theta) \qquad g_- = g_+ = g \tag{651b}$$

  by relating the factors of g between (651) and (650).

- Whether $g_- \neq g_+$ or not, by plugging $n = 0$ (IC) and 1 (first step) we obtain,

$$A(\xi) = \hat{v}^0 \qquad B(\xi) = \hat{v}^1 - \hat{v}^0 g_+^n(\theta) + \tag{652a}$$

- As with other multi-step methods the value of the first steps requires special initialization.

- In this case for the temporally first order PDE with two steps the first step should be advanced with a single-step method.

- For example, assume that we use forward time central space (FTCS) scheme (27c),

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0 \tag{653}$$

- From §2.1.8 we know that FTCS is unconditionally stable, but for one step there will be no problem in using FTCS to initialize the first step of the leapfrog method. The following argument would work with any other initialization as well.

- In any case, to relate $\hat{v}^0$ and $\hat{v}^1$ by using the fact that in von Neumann analysis we plug solutions in the form (445) $v_m^n = e^{\imath m\theta}\hat{v}^n$ in the FD stencils (*cf.* §6.3.6 for discussion). By plugging this in the stencil (653) for $n = 0$ we obtain,

$$e^{\imath m\theta}\hat{v}^{n+1} - e^{\imath m\theta}\hat{v}^n + \frac{\bar{k}}{2}\left\{e^{\imath(m+1)\theta}\hat{v}^n - e^{\imath(m-1)\theta}\hat{v}^n\right\} = 0 \qquad \Rightarrow$$
$$\hat{v}^{n+1} = (1 - \bar{k}\sin\theta)\hat{v}^n \quad \text{which for } n = 0 \text{ gives} \qquad \hat{v}^1 = (1 - \bar{k}\sin\theta)\hat{v}^0 \tag{654}$$

- Plugging $\hat{v}^1$ from (654) to (655) yields,

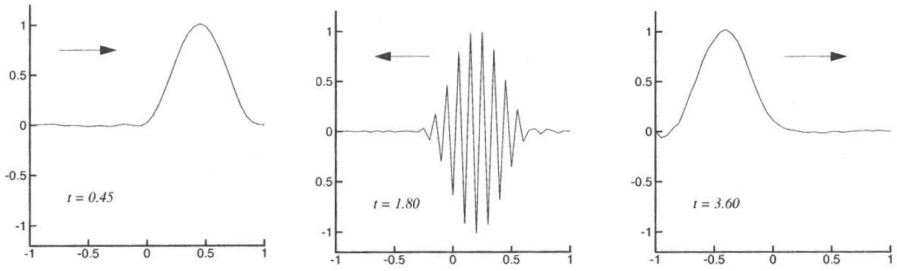
$$A(\xi) = 1\hat{v}^0 \tag{655a}$$

$$B(\xi) = \left[\frac{1}{2}\bar{k}^2 \sin^2\theta + \mathcal{O}\left(\theta^4\right)\right]\hat{v}^0 = \mathcal{O}\left(\theta^2\right)\hat{v}^0 \tag{655b}$$

- So, this equation says $B(\xi)/A(\xi) = \mathcal{O}\left(\theta^2\right)$ that is as for $\theta \ll 1$ $B(\xi)$ is much smaller (by θ^2 ratio) smaller than $A(\xi)$.
- In §7.7.7.1 we discussed that as $\theta h\xi \to 0$ for a given ξ so does h. In addition if $\bar{k} = ka/h$ is bounded independent of h (which is automatically done for explicit methods due to CFL limit $\bar{k} \leq 1$) k also tends to zero as $\theta \to 0$; *cf.* §7.7.7.1 and (644).
- Thus, **as** $h, k \to 0$ $\theta \to 0$ (for all active wavenumbers ξ in the IC) and **the parasitic term corresponding to** $B(\xi)$ **in** (651) **becomes negligible compared to physical term corresponding to** $A(\xi)$**.**
- Now, the question is what would happens if the grid size is coarse relative to relevant wavenumbers in the solution.
- Unfortunately in this case, the parasitic modes can have a very adverse effect. The also are activated due to inconsistent boundary condition as demonstrated in the next example.

Example 12 *Adverse effect of parasitic modes (copied from [Strikwerda, 2004] example 4.1.1). An interesting way to see the parasitic mode and also to illustrate the effect of inconsistent BCs is shows below. The figures show the solution computed by the leapfrog scheme with initial data as a pulse given by,*

$$v_m^n - \begin{cases} \cos^2 \pi x_m & \text{if } |x_m| \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$



Leapfrog parasitic mode

The wave speed a for the advection equation $u_{,t} + au_{,x} = 0$ is $a = 0.9$ and $x \in [-1, 1]$. At both boundaries v^n is set to zero. At the right boundary this is inconsistent with the PDE as the waves move from left to right and BCs should be enforced only on the left boundary. This inconsistency will generate substantial parasitic mode in the solution.

- The top left plot shows the solution at $t = 0.45$ with the pulse moving to the right.
- The top right plot shows a wave at $t = 1.80$ moving to the left.
- The inconsistent boundary condition has generated a solution having a significant parasitic mode, as indicated by the **oscillatory nature of the pulse** and **its "wrong" direction of travel**.
- The bottom plot shows the solution at $t = 3.6$ with the original pulse shape nearly restored.
- The parasitic is converted to the non-parasitic mode by the boundary condition at the left endpoint of the interval.

In Summary we mention,

- With multi-step schemes if the number of steps is greater than the temporal order of PDE there will be parasitic modes.
- The effect of parasitic modes typically negligible.
- However, when coarse grids are used, inconsistent boundary conditions are used, or in some other scenarios they can substantially pollute the numerical solution.
- The solution features that correspond to parasitic modes are nonphysical. For example parasitic modes travel in the wrong direction as demonstrated by the preceding example.

- Given that multi-step methods can enhance the temporal order of accuracy they can be favored over lower order methods. However, different approaches may be taken to reduce the effect of parasitic modes.
- For example, numerical dissipation may be introduced to reduce their effect; *cf.* [Strikwerda, 2004] chapter 5 for more details.
- Finally, to the point relevant to dispersion and dissipation analysis in this section, the analysis is only carries for physical amplification factor given that as $h, k \to 0$ parasitic modes vanish.

7.7.7.4 Dispersion and dissipation analysis of leapfrog method

- Recall the values for amplification factors of the leapfrog method are given as, (*cf.* (469)),

$$\begin{aligned} g_+ &= -\imath\bar{k}\sin\theta + \sqrt{1-\bar{k}^2\sin^2\theta} \\ g_- &= -\imath\bar{k}\sin\theta - \sqrt{1-\bar{k}^2\sin^2\theta} \end{aligned}$$

- As mentioned before the fact that $\lim_{\theta\to 0} g_-(\theta) = -1 \neq 1$ corresponds to the fact that that amplification factor corresponds to a parasitic mode.
- For clarity points corresponding to $\theta = 0, \pm\pi/2, \pm\pi$ for both g_+, g_- are given in the equation below and marked in the figure,

$$g_+(\theta = 0) = 1 \tag{656a}$$
$$g_+(\pi/2) = \sqrt{1-\bar{k}^2} - \imath\bar{k} \tag{656b}$$
$$g_+(-\pi/2) = \sqrt{1-\bar{k}^2} + \imath\bar{k} \tag{656c}$$
$$g_+(\pm) = 1 \tag{656d}$$

- Referring to (655) and the discussion at the end of §7.7.7.3 we observe that when $h, k \to 0$ the parasitic mode (corresponding to g_-) is negligible.
- This implies that to determine the convergence rates of dissipation and dispersion rates, which are for the limiting case $h, k \to 0$, the contribution of parasitic modes to the solution are negligible to that of physical mode (corresponding to g_+).
- In any case, it is the g_+ amplification factor that represents dispersion and dissipation of the advection equation, *i.e.*, approximating physical frequencies. $\omega_I = 0$, $\omega_R = a\xi$ for the advection $u_{,t} + au_{,x} = 0$ equation from (565b).
- We follow the same solution process from §7.7.5 (for FTBS scheme therein) applied to physical amplification factor g_+,
- Given that $g_+ = -\imath\bar{k}\sin\theta + \sqrt{1-\bar{k}^2\sin^2\theta}$ from (628) we obtain,

$$\omega_I^h = \frac{1}{2k}\log(g_R^2 + g_I^2) = \frac{1}{2k}\log\left[\bar{k}^2\sin^2\theta + \left(1-\bar{k}^2\sin^2\theta\right)\right] = \frac{1}{2k}\log 1 = 0 \tag{657a}$$
$$\omega_R^h = \frac{1}{k}\tan^{-1}\left(\frac{-g_I}{g_R}\right) = \frac{1}{k}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{\sqrt{1-\bar{k}^2\sin^2\theta}}\right) \tag{657b}$$

- From the analytical solutions for advection equation $\omega_I = 0$ and $\omega_R = a\xi$ from (629) and ω_I^h, ω_R^h from (657) we compute dispersion and dissipation errors from (611),

$$\Delta\omega_I = \omega_I^h - \omega_I = 0 \tag{658a}$$
$$\Delta\omega_R = \omega_R^h - \omega_R = \frac{1}{k}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{\sqrt{1-\bar{k}^2\sin^2\theta}}\right) - a\xi \tag{658b}$$

- Similar to (634) we express these errors in nondimensional and normalized form $\Delta\omega_I k$ and $\Delta\omega_R/\omega_R$. Recall $1/(k\omega_R) = 1/(ka\xi) = \frac{h}{ak}\frac{1}{h\xi} = \frac{1}{k\theta}$ (*cf.* (631)) to obtain,

$$\omega_I^h k = 0 \quad \Rightarrow \qquad \frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\pi}\log(1-A_d) = 0 \qquad \Rightarrow A_d = 0 \tag{659a}$$

$$\frac{\omega_R^h}{\omega_R} = \frac{1}{\bar{k}\theta}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{(1-\bar{k})+\bar{k}\cos\theta}\right) \quad \Rightarrow \quad \frac{\Delta\omega_R}{\omega_R} = -\frac{\Delta T}{T} = \frac{1}{\bar{k}\theta}\tan^{-1}\left(\frac{\bar{k}\sin\theta}{\sqrt{1-\bar{k}^2\sin^2\theta}}\right) - 1 \tag{659b}$$

Similar to the solution process in §7.7.5 we have used (614b) ($\frac{\Delta T}{T} = \frac{T^h - T}{T} = -\frac{\Delta\omega_R}{\omega_R}$) for relative period elongation (error) and (618) ($A_d = 1 - e^{2\pi\frac{\Delta\omega_I}{\omega_R}} \approx -2\pi\frac{\Delta\omega_I}{\omega_R}$ *i.e.*, $\frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\pi}\log(1-A_d) \approx -\frac{1}{2\pi}A_d$) for amplitude decay.

- **Convergence rates**: We can follow the solution process of §7.7.6 to expand dispersion and dissipation errors in (659) in θ by employing (637) series.
- Given that dissipation error is identically zero, we only expand the dispersion (frequency / period) error from (659b) to obtain (details of derivation are not provided for brevity),

$$\frac{\Delta\omega_R}{\omega_R} = -\frac{\Delta T}{T} = -\frac{\theta^2}{6}(1-\bar{k}^2) + \mathcal{O}\left(\theta^4\right) \tag{660}$$

Some observations on dispersion and dissipation error analysis of Leapfrog method are:

- **Zero dissipation error**: As evident from (657a) (and (658a), (659a)) dissipation and amplitude decay errors are both zero. This is due to the fact that $|g_+(\theta)| = 1$.
- **No dissipation or high frequency content**: As explained in detail in §5.4.1, in the context of time marching methods applied to ODEs, it is actually favorable to have some dissipation active for high frequency content of the solution that often is generated due to numerical noise. Basically, adding numerical dissipation for higher frequency content will dissipate high frequency noise which is a favorable property. Since leapfrog method is conservative, it does not provide any dissipative mechanisms for high frequency content. A solution to this is adding numerical dissipation which can also reduce the effect of parasitic modes for the leapfrog method. For more details refer to [Strikwerda, 2004] chapter 5.
- **Dispersion error and its order of convergence**: Referring to (660) and in comparison with the asymptotic dispersion error of FTBS scheme from (643b) ($\frac{\Delta\omega_R}{\omega_R} = -\frac{\theta^2}{6}(1-\bar{k})(1-2\bar{k}) + \mathcal{O}\left(\theta^4\right)$) we observe that both methods have second order of accuracy for dispersion error.

  In this case we also observe that $\Delta\omega_R = 0$ for CFL number 1 ($\bar{k} = 1$) from (659b). That is again CFL number 1 is the best possible case in control dispersion error. That is, again it is beneficial to take a time step exactly equal to CFL = 1. We also observe that asymptotic expansion of the dispersion error (660) is also consistent in having the expended term(s) be zero for $\bar{k} = 1$ (which must be the case).

  However, for the leapfrog method due to multiplicity of roots $g_- = g_+ = -1$ for $\bar{k} = 1$ and $\theta = \pm\pi/2$ from (476) we observed the solution took the form $\hat{v}^n = A(\xi)g^n + B(\xi)ng^n = A(\xi)(-1)^n + B(\xi)n(-1)^n$ in which the algebraic growth n corresponded to a weak instability. Accordingly, from (477) we require $\bar{k} < 1$ for the stability of the leapfrog method and we cannot choose $\bar{k} = 1$ for the leapfrog method to completely annihilate dispersion error. Still, we can take $\bar{k}$ as close as possible to 1 to decrease dispersion error.

References

[Ainsworth et al., 2006] Ainsworth, M., Monk, P., and Muniz, W. (2006). Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second-order wave equation. *Journal of Scientific Computing*, 27(1-3):5–40.

[Bathe, 2006] Bathe, K.-J. (2006). *Finite element procedures*. Klaus-Jurgen Bathe.

[Baumann et al., 2009] Baumann, D., Fumeaux, C., Hafner, C., and Li, E. P. (2009). A modular implementation of dispersive materials for time-domain simulations with application to gold nanospheres at optical frequencies. *Optics Express*, 17(17):15186–200.

[Belytschko et al., 2003] Belytschko, T., Chen, H., Xu, J., and Zi, G. (2003). Dynamic crack propagation based on loss of hyperbolicity and a new discontinuous enrichment. *International Journal for Numerical Methods in Engineering*, 58(12):1873 – 1905.

[Bui et al., 1991] Bui, M. D., Stuchly, S. S., and Costache, G. I. (1991). Propagation of transients in dispersive dielectric media. *IEEE Transactions on Microwave Theory and Techniques*, 39(7):1165 – 1172.

[Butcher, 1964] Butcher, J. C. (1964). On Runge-Kutta processes of high order. *Journal of the Australian Mathematical Society*, 4(02):179–94.

[Butcher, 2005] Butcher, J. C. (2005). *The Numerical Analysis of Ordinary Differential Equations, Runge-Kutta and General Linear Methods*. Whiley, Chichester.

[Cattaneo, 1948] Cattaneo, C. (1948). On the conduction of heat. *Atti del Seminario Matematico e Fisico dell'Università di Modena*, 3:3–21.

[Chapra and Canale, 2010] Chapra, S. C. and Canale, R. P. (2010). *Numerical methods for engineers*, volume 2. McGraw-Hill. 6th edition.

[Cockburn and Shu, 1998a] Cockburn, B. and Shu, C.-W. (1998a). Local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM Journal on Numerical Analysis*, 35(6):2440–63.

[Cockburn and Shu, 1998b] Cockburn, B. and Shu, C.-W. (1998b). The Runge-Kutta discontinuous Galerkin method for conservation laws. V multidimensional systems. *Journal of Computational Physics*, 141(2):199–224.

[Compte and Jou, 1996] Compte, A. and Jou, D. (1996). Non-equilibrium thermodynamics and anomalous diffusion. *Journal of Physics A: Mathematical and General*, 29(15):4321–9.

[Dahlquist, 1963] Dahlquist, G. G. (1963). A special stability problem for linear multistep methods. *BIT Numerical Mathematics*, 3(1):27–43.

[De Basabe and Sen, 2010] De Basabe, J. D. and Sen, M. K. (2010). Stability of the high-order finite elements for acoustic or elastic wave propagation with high-order time stepping. *Geophysical Journal International*, 181(1):577–590.

[Dedeurwaerdere et al., 1996] Dedeurwaerdere, T., Casas-Vazquez, J., Jou, D., and Lebon, G. (1996). Foundations and applications of a mesoscopic thermodynamic theory of fast phenomena. *Physical Review E*, 53(1):498–506.

[Ding et al., 2010] Ding, C., Hao, L., and Zhao, X. (2010). Two-dimensional acoustic metamaterial with negative modulus. *Journal of Applied Physics*, 108(7).

[Dumbser and Munz, 2005] Dumbser, M. and Munz, C.-D. (2005). ADER discontinuous Galerkin schemes for aeroacoustics. *Comptes Rendus de l'Academie des Sciences Serie II b/Mecanique*, 333(9):683–7.

[Dumbser and Munz, 2006] Dumbser, M. and Munz, C.-D. (2006). Building blocks for arbitrary high order discontinuous Galerkin schemes. *Journal of Scientific Computing*, 27(1-3):215–30.

[Fahs, 2012] Fahs, H. (2012). Investigation on polynomial integrators for time-domain electromagnetics using a high-order discontinuous Galerkin method. *Applied Mathematical Modelling*, 36(11):5466–81.

[Huang et al., 2013] Huang, Y., Li, J., and Yang, W. (2013). Modeling backward wave propagation in metamaterials by the finite element time-domain method. *SIAM Journal On Scientific Computing*, 35(1):B248–B274.

[Hughes et al., 1976a] Hughes, T., Hilber, H., and Taylor, R. (1976a). A reduction scheme for problems of structural dynamics. *International Journal of Solids and Structures*, 12(11):749–67.

[Hughes et al., 1976b] Hughes, T., Taylor, R., Sackman, J., Curnier, A., and Kanoknukulchai, W. (1976b). A finite element method for a class of contact-impact problems. *Computer Methods in Applied Mechanics and Engineering*, 8(3):249 – 276.

[Hughes, 2012] Hughes, T. J. (2012). *The finite element method: linear static and dynamic finite element analysis*. Courier Corporation.

[Joseph et al., 1991] Joseph, R. M., Hagness, S. C., and Taflove, A. (1991). Direct time integration of Maxwell's equations in linear dispersive media with absorption for scattering and propagation of femtosecond electromagnetic pulses. *Optics Letters*, 16(18):1412–4.

[Kashiwa and Fukai, 1990] Kashiwa, T. and Fukai, I. (1990). A treatment by the FD-TD method of the dispersive characteristics associated with electronic polarization. *Microwave and Optical Technology Letters*, 3(6):203–205.

[Kashiwa et al., 1990] Kashiwa, T., Ohtomo, Y., and Fukai, I. (1990). A finite-difference time-domain formulation for transient propagation in dispersive media associated with Cole-Cole's circular arc law. *Microwave and Optical Technology Letters*, 3(12):416–9.

[Kelley and Luebbers, 1996] Kelley, D. and Luebbers, R. (1996). Piecewise linear recursive convolution for dispersive media using FDTD. *IEEE Transactions on Antennas and Propagation*, 44(6):792–7.

[LeVeque, 2002] LeVeque, R. L. (2002). *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press.

[Li, 2011] Li, J. (2011). Development of discontinuous Galerkin methods for Maxwell's equations in metamaterials and perfectly matched layers. volume 236, pages 950–61.

[Lorcher et al., 2008] Lorcher, F., Gassner, G., and Munz, C.-D. (2008). An explicit discontinuous Galerkin scheme with local time-stepping for general unsteady diffusion equations. *Journal of Computational Physics*, 227(11):5649–70.

[Luebbers and Hunsberger, 1992] Luebbers, R. and Hunsberger, F. (1992). FDTD for nth-order dispersive media. *IEEE Transactions on Antennas and Propagation*, 40(11):1297–301.

[Luebbers et al., 1990] Luebbers, R., Hunsberger, F. P., Kunz, K. S., Standler, R. B., and Schneider, M. P. (1990). A frequency-dependent finite-difference time-domain formulation for dispersive materials. *IEEE Transactions on Electromagnetic Compatibility*, 32(3):222–7.

[Maxwell, 1867] Maxwell, J. (1867). On the dynamical theory of gases. *Philosophical Transactions of the Royal Society of London*, 157:49–88.

[Morin, 2010] Morin, D. (2010). Dispersion (chapter 6 of the coursenotes). morin@physics.harvard.edu.

[Ralston, 1962] Ralston, A. (1962). Runge-Kutta methods with minimum error bounds. *Mathematics of computation*, 16(80):431–437.

[Ralston and Rabinowitz, 1978] Ralston, A. and Rabinowitz, P. (1978). *A First Course in Numerical Analysis*. McGraw-Hill, New York. 2d ed.

[Shakib and Hughes, 1991] Shakib, F. and Hughes, T. (1991). A new finite element formulation for computational fluid dynamics. IX. fourier analysis of space-time Galerkin/least-squares algorithms. *Computer Methods in Applied Mechanics and Engineering*, 87(1):35–58.

[Stannigel et al., 2009] Stannigel, K., Konig, M., Niegemann, J., and Busch, K. (2009). Discontinuous Galerkin time-domain computations of metallic nanostructures. *Optics Express*, 17(17):14934–47.

[Strikwerda, 2004] Strikwerda, J. C. (2004). *Finite difference schemes and partial differential equations*. SIAM.

[Süli and Mayers, 2003] Süli, E. and Mayers, D. F. (2003). *An introduction to numerical analysis*. Cambridge university press.

[Vernotte, 1958] Vernotte, P. (1958). Les paradoxes de la theorie continue de lequation de la chaleur. *Comptes Rendus Nebdomadaires des Seances de l Academie des Sciences*, 246(22):3154–5.

[Viquerat et al., 2013] Viquerat, J., Klemm, M., Lanteri, S., and Scheid, C. (2013). Theoretical and numerical analysis of local dispersion models coupled to a discontinuous Galerkin time-domain method for Maxwell's equations. Technical Report INRIA, 2013 Tech. rep.

[Yang et al., 2013] Yang, H., Li, F., and Qiu, J. (2013). Dispersion and dissipation errors of two fully discrete discontinuous Galerkin methods. *Journal of Scientific Computing*, 55(3):552–74.

[Zhang and Ge, 2009] Zhang, Y.-Q. and Ge, D.-B. (2009). A unified FDTD approach for electromagnetic analysis of dispersive objects. *Progress In Electromagnetics Research*, 96:155–72.

Homework assignments

1. (**20 Points**) Classify the following PDEs as linear (L), semi-linear (SL), quasi-linear (QL), and fully-nonlinear (FN),
   (a) $u_t + x^2 u_x = 0$
   (b) $u_t + u_{xxx} + uu_x = 0$
   (c) $u_t + a(u)u_x = 0$
   (d) $u_x^2 + u_y^2 = 1$
   (e) $\text{div}\left(\frac{\nabla \mathbf{u}}{\sqrt{1+|\nabla \mathbf{u}|^2}}\right) = 0$

2. (**60 Points**) (a) Verify that the equation,
$$3u_{xx} + 7u_{xy} + 2u_{yy} = 0 \tag{1}$$
is hyperbolic for all x and y, (b) find the new characteristic coordinates η, ξ, and (c) express the PDE in the canonical form (find Φ),
$$U_{\xi\eta} = \Phi(\xi, \eta, u, u_\xi, u_\eta) \tag{2}$$

3. (**10 Points**) Determine the type of Tricomi PDE based on the coordinate values (x, y) in terms of being hyperbolic, parabolic, and elliptic,
$$\frac{\partial^2 u}{\partial x^2} + y\frac{\partial^2 u}{\partial y^2} = 0 \tag{3}$$

4. (**20 Points**) Determine the type (hyperbolic, parabolic, or elliptic) of the following constant-coefficient PDE,
$$u_{xx} + 2u_{xy} - 3u_{yz} + 5u_{zz} = 0 \tag{4}$$

5. (**60 Points**) Consider the initial value problem for the equation,
$$u_t + au_x = f(x, t) \tag{5}$$
with $u(0, x) = 0$ and
$$f(x,t) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$
Assume that a is positive. Show that the solution is given by,
$$u(x,t) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{a} & x \geq 0 \text{ and } x - at \leq 0 \\ t & x \geq 0 \text{ and } x - at \geq 0 \end{cases} \tag{7}$$

6. (**80 Points**) Find a solution to the initial-value problem.
$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}_t + \begin{bmatrix} 1 & 4 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}_x = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{8a}$$
$$\begin{bmatrix} u_1(x,0) \\ u_2(x,0) \end{bmatrix} = \begin{bmatrix} \sin x \\ \cos x \end{bmatrix} \tag{8b}$$
Note: You need to express $u_1(x,t)$, and $u_2(x,t)$. For example, $u_1(x,t) = \frac{1}{2}\sin(x - 5t) + \cdots$. Do not turn the system to a second order PDE. Solve it as a system of first order PDEs by deriving characteristics ω_1, ω_2, and finally expression the solution in terms of (x, t).

We want to solve the advection equation,

$$u_{,t} + a(x,t)u_{,x} = 0 \qquad \text{PDE} \tag{1a}$$
$$u(x, t=0) = u_0(x) \qquad \text{IC} \tag{1b}$$

The specific of the problem are,

1. $a = 0.1$.
2. Computational domain is $[-10,\ 10]$.
3. Initial condition (IC): Two different cases are considered,

$$u_0(x) = \begin{cases} 1 & x < 0 \\ 0 & 0 \le x \end{cases} \qquad \textbf{nonsmooth IC} \text{ option (not even a } \mathcal{C}^0 \text{ function)} \tag{2a}$$

$$u_0(x) = \begin{cases} 1 & x < -1 \\ -x^3(6x^2 + 15x + 10) & -1 \le x < 0 \\ 0 & 0 \le x \end{cases} \qquad \textbf{smooth IC} \text{ option } (\mathcal{C}^2 \text{ function}) \tag{2b}$$

1. ($(90 + 20 + 10 =)$ **120 Points**) Numerical solution of (26) with given computational domain size and $a = 0.1$:
   - Solve **ONLY the nonsmooth IC**.
   - Consider FD spatial grid size $h = 0.1$.
   - Consider **two** different normalized time steps,

$$\bar{k} = 0.8, 1.1, \qquad \bar{k} = \frac{ka}{h}$$

   - Consider 3 target times of $T_t = 3, 6, 20$. Depending on the value of $\bar{k}$ you need to choose a time step value just above target values. For example, $\bar{k} = 0.8 \Rightarrow 0.8 = \bar{k} = \frac{ka}{h} = \frac{0.1k}{0.1} \Rightarrow k = 0.8 \Rightarrow$ for $T_t = 3$ number of time steps $=$ ceiling(T_t/k) = ceiling$(3/0.8)$ = ceiling$(3.75) = 4 \Rightarrow 4$ time steps are needed and the *computed final time* instead of 3 would be $4k = 4 \times 0.8 = 3.2$. Different $\bar{k}$ given slightly different final times than $3, 6, 20$. Below, is the list of these values for your reference (T refers to the time that samples will be taken for visualization):

$$T = 3.2, 6.4, 20 \quad \text{for} \quad T_t = 3, 6, 20, \text{ and } \bar{k} = 0.8$$
$$T = 3.3, 6.6, 20.9 \quad \text{for} \quad T_t = 3, 6, 20, \text{ and } \bar{k} = 1.1$$

     You can interpolate values between different time steps to exactly match T_t but the values T above suffice for this study.
   - Solve for the schemes **BTFS, BTBS, FTFS, FTCS, FTBS**.
   - submit all the plots in one or several file(s) AND the code that you have used to generate the results by email or by using dropbox.
   - Each plot shows numerical results u^h versus x along with the exact solution for the given time.
   - Since some solutions blow up, use the following limits for your x and u (y) axes:

x axis	$[-1.5\ 4]$
y axis	$[-4\ 4]$

   - For boundary conditions (BC) use $u = 1$ at $x = -10$ and $u = 0$ for $x = 10$ (or beyond as may be needed for FTFS scheme.
   - The three explicit schemes (FTFS, FTCS, FTBS) can be solved for one point at a time u^n_m. The two implicit schemes can also be solved starting from the left side going to the right side as demonstrated in the class notes (part 2) for this particular problem.

Material that should be submitted are:

(a) Present your results in 12 plots $\{2\ \bar{k}\ (0.8,\ 1.1)\} \times \{3\ T_t\ (3,\ 6,\ 20)\} \times \{2$ (Implicit schemes BTFS , BTBS versus explicit schemes FTFS, FTBS, FTCS)$\}$ names as kn($\bar{k}$ value),T =(T), (Explicit or Implicit) where terms in () are replaced by the values displayed in the plots. If it is difficult to group explicit and implicit results together, you can submit the plot for each scheme separately.

(b) For each of the 5 schemes (BTFS, BTBS, FTFS, FTCS, FTBS) write the stencil equation you use for the solution of a new value in terms of $\bar{k}$ parameter. For example, from the course notes for this problem the equation for BTFS is $(1-\bar{k})u_m^{n+1} + \bar{k}u_{m+1}^{n+1} = u_m^n \Rightarrow u_{m+1}^{n+1} = \frac{1}{\bar{k}}u_m^n - \frac{1-\bar{k}}{\bar{k}}u_m^{n+1}$. Right the update equations for all the schemes in your solution sheets.

(c) In the solution sheet, write the exact solution to this problem for the nonsmooth initial condition.

2. ($(5 \times 6 \times 3 =$) **90 Points**) Continuing with the same problem and using your solutions, class notes, and convergence plots provided to you in the folder "HW02/P1,2/ConvergencePlots", answer the following questions for EACH of the schemes BTFS, BTCS, BTBS, FTFS, FTCS, FTBS:

(a) Classify the method as i) Unconditionally unstable (UU), ii) Conditionally stable (CS), iii) Unconditionally stable (US).

(b) List the set $S_{\bar{k}}$ of stability in terms of $\bar{k}$. For example, for an unconditionally unstable scheme $S_{\bar{k}} = \emptyset$, conditional stable with CFL number 0.5 it is $S_{\bar{k}} = [0,\ 0.5]$.

(c) For the schemes that are conditionally stable or unconditionally unstable comment on how from your solutions instabilities develop and grow in time T_t from 3 to 6 to 20 for either $\bar{k} = 0.8$ or $\bar{k} = 1.1$ (which one(s) that apply). Also, comment whether there are regions where the solution does not blow up, but does not converge to the correct solution.

(d) Referring to convergence plots (in HW2_FD_convergencePlots folder) comment on the rate of convergence of the error versus h for the ranges of $\bar{k}$ (kn in the plots) where the method is stable. Specifically i) by visual inspection of convergence plots provide approximate values for both smooth and nonsmooth IC cases; ii) comment whether nonsmooth and smooth IC problems have different convergence rates or not; iii) specify if the method has the highest spatial convergence rate among those considered; iv) if the method is unconditionally stable comment whether for any k the convergence plots may erroneously suggest a convergence of error to zero as $h \to 0$.

Background: The L2 error norm is defined as,

$$L_2(u^h - u) = \sqrt{\int_{-10}^{10} (u^h(x) - u(x))^2\ \mathrm{d}x} \tag{3}$$

In many numerical methods, *e.g.,* FV, FD, FEM, the convergence rate of an error (*e.g.,* the L2 error norm above) is expressed as,

$$e \propto C_h h^p + C_k k^s \tag{4}$$

where p and s are spatial and temporal convergence rates and C's are constants depending on the details of the method, grid, *etc.*. If the exact solution is NOT smooth enough the convergence rates decrease depending of the level of regularity of the exact solution (*e.g.,* if it is a $\mathcal{C}^0, \mathcal{C}^1, \cdots, \mathcal{C}^r$ function). For a static problem the temporal part drops. If the error is entirely from the spatial part (or is dominated by it), we can write

$$e \approx C_h h^p \qquad \Rightarrow \log(e) \approx \log(C_h) + p \log(h) \tag{5}$$

So, the convergence rate is the slope at which $\log(e)$ decreases as h decreases in log-log plots (or obtained from mathematical analysis). Note that to compute convergence rate, the ranges of small h should be considered where the asymptotic convergence rates are achieved. At large h the convergence rates do not hold.

(e) Comment on the role of normalized time step $\bar{k}$ on the convergence plots. Is there a value that beyond of $\bar{k}$ beyond which the temporal discretization error is ignorable compared to spatial discretization error? Does taking smaller $\bar{k}$ improve the convergence rate. Specifically for BTCS method discuss why the convergence rate changes as $\bar{k} \to 0$.

- List your solutions as BTFS(a), BTFS(b), $\cdots$, BTFS(e), BTCS(a), $\cdots$, BTCS(e), $\cdots$.
- Be brief in your answers.

3. ($(6 \times 15 =$) **90 Points**) Consider the 1D elastodynamic problem,

$$u_{,tt} - c^2 u_{,xx} = 0 \tag{6}$$

for zero body $b = 0$.

- The problem can be solved as a **1-field** problem where u (*i.e.,* U_m^n being the unknowns) is directly discretized or as a **3-field** problem where u, v, s (*i.e.,* U_m^n, V_m^n, S_m^n are unknowns) are discretized. We use F1 and F3 to refer to these two approaches.

- For each F1 or F3 we can use average fluxes which recovers a normal FD scheme or use Riemann solutions which in the following discussion we label it as FV. So, FD means we have used average fluxes, FV Riemann fluxes.
- In the class notes (part 2) we demonstrated that the FV update for the stress, velocity (and added displacement here) in a 3F scheme is,

$$\frac{S_m^{n+1}-S_m^n}{k} - E\frac{V_{m+1}^n - V_{m-1}^n}{2h} - \frac{hc}{2}\frac{S_{m+1}^n + S_{m-1}^n - 2S_m^n}{h^2} = 0 \tag{7a}$$

$$\frac{V_m^{n+1}-V_m^n}{k} - \frac{1}{\rho}\frac{S_{m+1}^n - S_{m-1}^n}{2h} - \frac{hc}{2}\frac{V_{m+1}^n + V_{m-1}^n - 2V_m^n}{h^2} = 0 \tag{7b}$$

$$\frac{U_m^{n+1}-U_m^n}{k} = V_m^n \tag{7c}$$

  the terms in the first two equations after $\frac{hc}{2}$ (terms in red) are added terms by using Riemann fluxes (labeled as FV scheme herein) compared to average fluxes (labeled as FD scheme).
- By eliminating S and V from (**??**) we obtain the 1F update:

$$\frac{U_m^{n+1} - 2U_m^n + U_m^{n-1}}{k^2} - c^2\frac{U_{m+2}^n - 2U_m^n + U_{m-2}^n}{4h^2} - D_h\frac{1}{k}\left\{\left[\frac{U_{m+1}^n - 2U_m^n + U_{m-1}^n}{h^2}\right] - \left[\frac{U_{m+1}^{n-1} - 2U_m^{n-1} + U_{m-1}^{n-1}}{h^2}\right]\right\} = 0 \tag{8}$$

  the terms in red (after D) are added when Riemann fluxes are used. That is, with them we solve FV1F and without them FD1F.
- The solution to all four combinations FD3F, FV3F, FD1F, FV1F options are provided in the folder P3/F3 and P3/F1 folders with subfolders FD and FV.
- Solutions are given for three different grid spatial resolutions $h = 0.001, 0.01, 0.1$ to represent refined, medium, and coarse resolutions.
- To numerically observe the stability limits of these methods, solutions are given for $\bar{k} = 0.1, 0.49, 0.5, 0.51, 0.99, 1.01, 1.99, 2.01$.
- The initial condition consists of a hat function (with peak 1) spanning $[0.4,\ 0.6]$ in the computational domain $[0,\ 1]$ with period boundary conditions.
- Material properties are $E = 0.04, \rho = 1 \Rightarrow c = Z = 0.2$.
- Results are presented for the target time $T_t = 1.75$ (slightly different values are realized based on the closest time step).
- Within 1F and 3F folders sample Matlab files are provided that would generate the results for problem 3. They can help you in writing compute code for the first problem, although in problem 1 many aspects (*e.g.,* boundary conditions) are different.

Answer the following questions using the provided numerical results:

(a) Classify FD3F, FV3F, FD1F, FV1F schemes as as i) Unconditionally unstable (UU), ii) Conditionally stable (CS), iii) Unconditionally stable (US).

(b) If one of the schemes is unconditionally unstable, compare the instability with a corresponding FD stencil instability when only one first order PDE is solved.

(c) For each scheme list the set $S_{\bar{k}}$ of stability in terms of $\bar{k}$.

(d) Given that results are compared for a fixed target time (slightly different times are realized for different $\bar{k}$), discuss which one of small or large grid sizes h expose instabilities better and justify your answer.

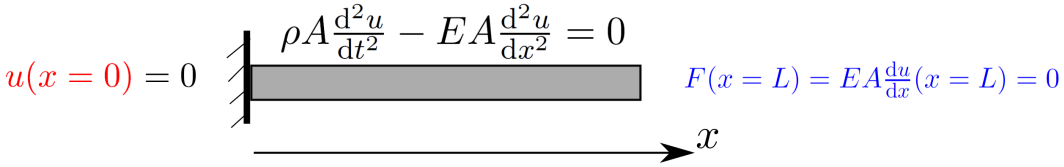
(e) Between FD and FV schemes describe which one is more dissipative, *i.e.,* dampens the sharp peaks. Explain the cause for this, if one scheme is more dissipative.

(f) Based on your observations briefly (less than 3 sentences) describe what scheme (FV vs. FD) you would choose for 1F and 3F choices.

1. (**30 Points**) Consider 1D bar problem for constant area A, Young's modulus E, and density ρ,

$$\rho A \frac{\mathrm{d}^2 u}{\mathrm{d}t^2} - EA \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = 0 \tag{1}$$

which as shown in the figure has prescribed displacement boundary on the left and free stress on the right. This boundary condition configuration is denoted by P1F1 (1 free and 1 prescribed BCs),



(a) Show that exact natural frequencies are,

$$\omega_n = (n - \frac{\pi}{2})\frac{c}{L}, \qquad \text{where } c = \sqrt{\frac{E}{\rho}} \tag{2}$$

(b) Obtain mode shapes $\Phi_i(x)$.

Hint: Use separation of variables $u(x,t) = \Phi(x)T(t)$ and follow the same process used in the course notes for a double fixed bar example.

2. ((20 + 10 + 20 =) **50 Points**) Recalling that element stiffness and mass matrices are (refer to the course notes),

$$\mathbf{k}^e = \frac{AE}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{3a}$$

$$\mathbf{m}_c^e = \frac{AL_e\rho}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \mathbf{m}_d^e = \frac{AL_e\rho}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3b}$$

where subscripts c and d refer to consistent and diagonal (lumped) mass matrices, and L_e is the element length.

(a) Find natural modes and frequencies for one element for **both** consistent and diagonal mass matrix options.

(b) Describe why one of the natural frequencies is zero. Use corresponding natural mode to explain what the natural mode corresponds to.

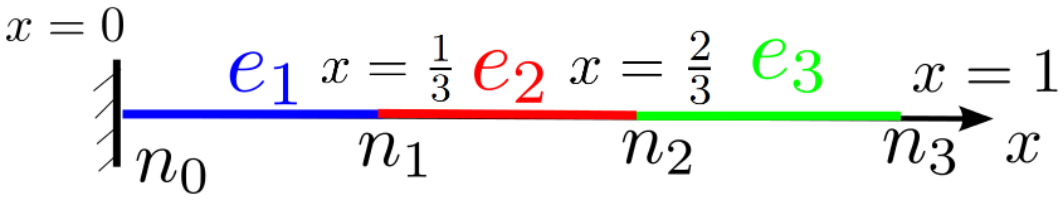
(c) Exact natural frequencies of the same element and corresponding modes will be $\omega_n = \frac{n\pi c}{L_e}$, $\Phi_n = \cos(n\pi\frac{x}{L_e})$ (you do not to prove this, the proof is similar to that of problem 1). Compare the first nonzero natural frequency that you obtain from either consistent or diagonal mass matrix options. We call them $\omega^h_{c\,1}$ for consistent mass matrix option and $\omega^h_{d\,1}$ for diagonal (lumped) mass matrix option. Compute $\omega^h_{c\,1}/\omega_1$ and $\omega^h_{d\,1}/\omega_1$ and comment on the values you obtain.. That is, discuss what the error of one element natural frequencies are with respect to the exact values.

Hint: Remember that natural frequencies for a FEM mesh with no damping ($\mathbf{C} = 0$) are obtained by solving a generalized eigenvalue problem for,

$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{K}\mathbf{U} = 0 \tag{4}$$

For one element $\mathbf{M} = \mathbf{m}_c^e$ (consistent mass matrix) or $\mathbf{M} = \mathbf{m}_d^e$ (diagonal mass matrix) and $\mathbf{K}$ is also taken from element value. Solve the corresponding natural mode, frequency problem for one element.

3. ((6 × 15 =) **90 Points**) First, the solution of a sample natural mode analysis is outlined for the one side prescribed one side free displacement boundary condition. Consider that this problem is discretized with three elements of equal size meaning that the element size is $L_e = L/3$. Also for simplicity assume $L = 1, A = 1, E = 1, \rho = 1$. Then local element matrices are (only consistent mass matrix option is shown):



$$\mathbf{m}_c^{e_1} = \frac{AL_e\rho}{6}\begin{matrix}0\\1\end{matrix}\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}\begin{matrix}0 & 1\end{matrix} = \begin{matrix}0\\1\end{matrix}\begin{bmatrix}\frac{1}{9} & \frac{1}{18}\\ \frac{1}{18} & \frac{1}{9}\end{bmatrix}\begin{matrix}0 & 1\end{matrix}, \text{ similarly } \mathbf{m}_c^{e_2} = \begin{matrix}1\\2\end{matrix}\begin{bmatrix}\frac{1}{9} & \frac{1}{18}\\ \frac{1}{18} & \frac{1}{9}\end{bmatrix}\begin{matrix}1 & 2\end{matrix}, \quad \mathbf{m}_c^{e_3} = \begin{matrix}2\\3\end{matrix}\begin{bmatrix}\frac{1}{9} & \frac{1}{18}\\ \frac{1}{18} & \frac{1}{9}\end{bmatrix}\begin{matrix}2 & 3\end{matrix} \tag{5}$$

after assembling free degrees of freedom into the global matrix we obtain,

$$\mathbf{M}_c = \begin{bmatrix}\frac{1}{9} & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} + \begin{bmatrix}\frac{1}{9} & \frac{1}{18} & 0\\ \frac{1}{18} & \frac{1}{9} & 0\\ 0 & 0 & 0\end{bmatrix} + \begin{bmatrix}0 & 0 & 0\\ 0 & \frac{1}{9} & \frac{1}{18}\\ 0 & \frac{1}{18} & \frac{1}{9}\end{bmatrix} = \frac{1}{18}\begin{bmatrix}4 & 1 & 0\\ 1 & 4 & 1\\ 0 & 1 & 2\end{bmatrix} \tag{6}$$

Similarly, given that for three elements,

$$\mathbf{k}^e = \frac{AE}{L_e}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix} = \begin{bmatrix}3 & -3\\ -3 & 3\end{bmatrix}$$

which after assembly to the global system (similar to the mass matrix) we get,

$$\mathbf{K} = 3\begin{bmatrix}2 & -1 & 0\\ -1 & 2 & -1\\ 0 & -1 & 1\end{bmatrix} \tag{7}$$

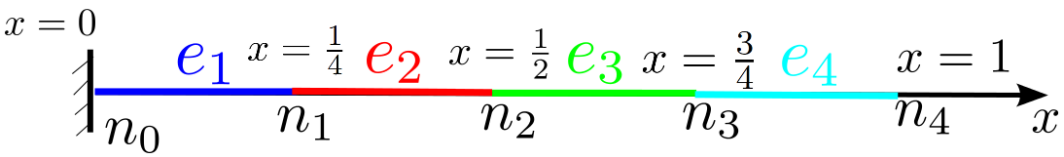
To obtain the 3 natural frequencies and modes for this 3 dof FEM model we solve the following generalized eigenvalue problem,

$$\mathbf{K}\boldsymbol{\Phi}_i^h = {\omega_i^h}^2\mathbf{M}\boldsymbol{\Phi}_i^h, i \leq 3, \quad \text{that is} \quad \begin{bmatrix}6 & -3 & 0\\ -3 & 6 & -3\\ 0 & -3 & 3\end{bmatrix} = {\omega_i^h}^2\begin{bmatrix}\frac{4}{18} & \frac{1}{18} & 0\\ \frac{1}{18} & \frac{4}{18} & \frac{1}{18}\\ 0 & \frac{1}{18} & \frac{2}{18}\end{bmatrix}\boldsymbol{\Phi}_i^h, i \leq 3 \tag{8}$$

just for reference we get $\omega_1^h = 1.5888$ and $\omega_3^h = 9.4266$. The exact natural frequencies 1 and 3 are given from (**??**) ($L = c = 1$), $\omega_1 = \pi/2 = 1.5708$ and $\omega_3 = 5\pi/2 = 7.8540$. We observe that ω_1^h has much smaller error than ω_3^h.

Another quantity of great practical important is finding how the highest frequency of this 3 dof system, *i.e.*, $\omega_3^h = 9.4266$, compares with the highest frequency of its SMALLEST element (here all elements are of the same size). We call the latter $\omega_{h_{\min}}$. From the solution of problem 2 and using the correct element size (1/3) you will realize that $\omega_{h_{\min}}^c = 10.3923$ (consistent mass option). We observe, $\omega_3^h/\omega_{h_{\min}}^c < 1$.

Based on the following background answer the following questions for the four dof system shown below,



(a) Obtain 4×4 mass matrices $\mathbf{M}_c$ (consistent) $\mathbf{M}_d$ (diagonal) and stiffness matrix $\mathbf{K}$.

(b) For EACH option (consistent and diagonal mass), find all 4 natural frequencies and natural modes using the generalized eigenvalue problem. Matlab and many other packages support the solution of a generalized eigenvalue problem.

(c) List the highest natural frequency of the 4 dof system for consistent and diagonal mass matrix options. That is, $\omega^h_{c\,4}$ and $\omega^h_{d\,4}$ (These are computed in previous item). Compare them with the fourth exact natural frequency ω_4 obtain from (**??**). That is, provide values for $\omega^h_{c\,4}/\omega_4$ and $\omega^h_{d\,4}/\omega_4$. This is to demonstrate how accurate the last modes are.

(d) Compare the minimum element size ($L_e = 0.25$) maximum frequency for both consistent mass and diagonal mass option: $\omega^c_{h_{\min}}$, $\omega^d_{h_{\min}}$.

(e) Compute the ratio of the system's maximum computed frequency to the smallest element's maximum frequency. That is, $\omega^h_{c\,4}/\omega^c_{h_{\min}}$ and $\omega^h_{d\,4}/\omega^d_{h_{\min}}$. Comment on their values.

(f) Refer to the two files "Ratio of maximum frequency to that of the smallest element_consistent.png" and "Ratio of maximum frequency to that of the smallest element_lumped" where $\omega^h_{c\,n}/\omega^c_{h_{\min}}$ and $\omega^h_{d\,n}/\omega^d_{h_{\min}}$ are computed for an N segment bar for three different boundary conditions:

- P0F2: Two end points are free displacement condition.
- P1F1: One prescribed and one free displacement conditions (problem considred in this HW).
- P2F0: Two end points are prescribed displacement condition.

The values to the left in the horizontal axis (closes to zero) the smaller the elements are. Based on these plots and the figures provided comment on $\omega^h_{c\,n}/\omega^c_{h_{\min}}$ and $\omega^h_{d\,n}/\omega^d_{h_{\min}}$ and how these values change in a uniform FE mesh this ratio changes as elements get smaller (use the figure).

4. ((10 + 20 + 10 =) **40 Points**) In the provided files starting with "Natural frequency convergence" convergence rates of natural frequencies of mode 1, 2, and 8 and three different boundary conditions (P0F2, P1F1, P2F0), and consistent or diagonal mass matrix options are provided. The plots are generated by solving the solutions with different element sizes h so that we can numerically investigate the convergence rate of natural frequencies. Answer the following questions,

(a) From these plots discuss what is the convergence rate of ω^h_i (based on results for $\omega^h_1, \omega^h_2, \omega^h_8$) for different mass matrix options and boundary conditions. Refer to the plots for answering this question. (note: some or all convergence rates may be identical).

(b) Refer to *a priori* error estimate provided in course notes

$$0 \le \omega^h_i - \omega_i \le C h^{2(p+1-m)} \omega_i^{\frac{2p+2-m}{m}} \tag{9}$$

Note that the proof of this condition is based on having Galerkin FEM formulation (*i.e.,* consistent mass) and full integration order.

Based on the discussion in the course notes, provide values for p and m.

(c) Having p and m, from (**??**) discuss what the convergence rate for ω^h_i for the bar problem with linear elements should be. How does this value compare with convergence rates you obtained for modes $i = 1, 2, 8$?

5. ((80 + 10 =) **90 Points**) Description of spectra plot for natural frequencies. **Part one requires computational code writing. Even without generating the results you can answer the second question of this problem with the provided plot**

This plot shows ω^h/ω (vertical axes) versus normalized wave number $\eta = n/N$.

- n is the mode number.
- N is the number of dof of the discrete FEM mesh.

For example for N dof discretization the mode one is shown in x coordinate $\eta = 1/N$ and its y value is ω^h_1/ω_1 that is the ratio of the numerical first natural frequency to the first exact natural frequency. Similarly, for the very last numerical natural frequency that this N dof FEM grid can capture, we have $n = N$. For $n = N$ x value is $\eta = N/N = 1$, and y value is ω^h_N/ω_N. For an N dof discretization of the problem we get N points $(\eta = n/N, \omega^h_n/\omega_n)$, $n = 1, \cdots, N$. A sample plot that superimposes results from different FEM dofs $N = 2, 8, 32, 128, 256$ and two different mass matrix options (consistent and diagonal) is "Spectra3_2 fixed ends.png".

(a) Use a computational tool, *e.g.,* Matlab, to generate spectral plots for P1F1 BC with results for $N = 2, 8, 32, 128$ and two mass matrix options (consistent and lumped) all super-imposed on in plot (similar to "Spectra3_2 fixed ends.png").

(b) Why $\omega^h/\omega \le 1$ when using diagonal mass matrix albeit $0 \le \omega^h_n - \omega_n$ in (**??**)? Explain the source of apparent discrepancy. You can use "Spectra3_2 fixed ends.png" is not solving the previous part.

6. **(EXTRA CREDIT)**((25 + 25 =) **50 Points**) Explanation of the meaning of spectral plot obtained in previous question.

(a) In your own word, explain why ω^h/ω becomes only a function of $\eta = n/N$ that is **the relative accuracy of natural frequency ω^h/ω is a direct function of the relative order of natural mode the FEM can capture $\eta = n/N$.** **Hint**: Consider that $\eta = 1/N$ is the lowest (and most accurate) mode an N dof system can capture, while $\eta = n/N$ for $n = N$ is the very last mode (and the least accurate) it would capture. Points that would facilitate the discussion is that for N dof system element size scales as $h = L/N$ (in 1D). At the same time, for a mode n the length scale in which the solution oscillates in space is L/n (can be easily seen in all mode n shape functions we obtained). The ratio of these two ratio is related to how accurately an N dof FEM can capture mode n?

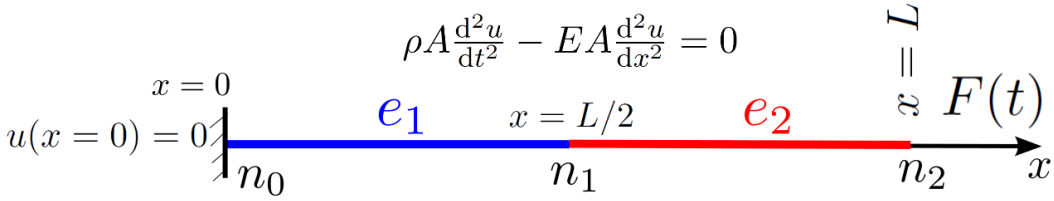
(b) Mathematically demonstrate this point. Use the convergence equation for natural frequency (**??**) (written for mode n)

$$0 \leq \omega_n^h - \omega_n \leq C h^{2(p+1-m)} \omega_n^{\frac{2p+2-m}{m}} \tag{10}$$

- For the bar problem we considered plug in the value of m. Use this equation to demonstrate that at least for small h where the asymptotic expression (**??**) holds we can demonstrate ω_n^h/ω_n is a function of η only.

Hint: First divide the equation by ω_n so to get relative value ω_n^h/ω_n on the LHS. For a 1D problem of length L and uniform element size h we have $N_e h = L$ where N_e is the number of elements. Given that the number of dof N is proportional to number of elements (why?) $h \propto \frac{1}{N}$. At the same time, $\omega_n \propto n$ (at least for larger n). By these two substitutions on the RHS ($h \propto \frac{1}{N}$ and $\omega_n \propto n$) we can express RHS only in terms of n and N. Verify the simplified expression results in ω_n^h/ω_n is a function of $\eta = n/N$ only for small h (large N).

1. ($((25(a)+4\times 15(b)+3\times 5(c)+10(d)=)$ **110 Points**) Consider 1D bar problem shown in the figure for constant area $A=1$, Young's modulus $E=1$, and density $\rho=1$, and length $L=1$. The spatial domain is discretized with two elements e_1 and e_2,



   (a) Show that stiffness matrix matrix $\mathbf{K}$, mass matrices (consistent $\mathbf{M}_c$ and diagonal (lumped) $\mathbf{M}_d$), and force vector $\mathbf{R}$ are,

$$\mathbf{K}=\begin{bmatrix}4 & -2\\ -2 & 2\end{bmatrix},\qquad \mathbf{M}_c=\begin{bmatrix}\frac{1}{3} & \frac{1}{12}\\ \frac{1}{12} & \frac{1}{6}\end{bmatrix},\qquad \mathbf{M}_d=\begin{bmatrix}\frac{1}{2} & 0\\ 0 & \frac{1}{4}\end{bmatrix},\qquad \mathbf{R}=\begin{bmatrix}0\\ F(t)\end{bmatrix} \tag{1}$$

   You can compute local element stiffness and mass matrices (consistent and lumped) from equations (158) (159) (167) in §2.3.9 . The 1D example from §2.3.11 is a good reference on how the local matrices are assembled to the global system.

   (b) Recall that $\max_l(\lambda_l^h)$ denotes the maximum frequency of the global system (a 2 dof system herein) and λ_e^m denotes the maximum element-wise frequency (*i.e.,* for each element compute the maximum frequency and compute the maximum of those). λ_e^m is much easier to compute in practice because as described in the course notes we know what these values will be for given element types. Both $\max_l(\lambda_l^h)$ and λ_e^m can be directly computed by modal analysis of the global system and one element (worst element, *i.e.,* element with highest frequency), respectively.

   The mass matrix (either in the global or element local level) can be consistent or diagonal (lumped) mass matrix and are decorated by c and d below.

   Compute the following,

   i. $\max_l(\lambda_{c_l}^h)$: Maximum global system frequency with consistent mass matrix.
   ii. $\lambda_{c_e}^m$: Maximum element frequency with consistent mass matrix.
   iii. $\max_l(\lambda_{d_l}^h)$: Maximum global system frequency with diagonal (lumped) mass matrix.
   iv. $\lambda_{d_e}^m$: Maximum element frequency with diagonal (lumped) mass matrix.

   for element level quantities do not use the formulas in the course note and solve the values by modal analysis of the smallest element with free end points.

   (c) Answer the following questions,

   i. Is the worst element maximum frequency larger or the system maximum frequency? Use both consistent and diagonal mass matrix options.
   ii. Which mass option (consistent versus diagonal) provides a higher frequency? Use global and element level values for your comparison.
   iii. For the previous question for the element level values compare $\lambda_{c_e}^m$ and $\lambda_{d_e}^m$. Then compare them with analytical value for the maximum frequency the element can model and state which one overestimates the frequency and which one underestimate. The same assertion can be made for the global system but is not asked to be checked in this problem.

   Hint: (for Q3): For a bar of free ends (which is similar to one element boundary conditions when its frequencies are computed) analytical natural frequencies are $\omega_n=n\pi c/L_e$ where L_e is the element length and $c=\sqrt{E/\rho}$ the wave speed.

   (d) Compute the system (global) Rayleigh damping matrix $\mathbf{C}$ (using consistent mass matrix $\mathbf{M}$) for $a_0=0.05$, $a_1=0.01$. You **will not add** this damping matrix to FEM equations in the next questions. For Rayleigh damping matrix refer to (213a) which is the same as (156).

2. ($([5(i)+5(ii)+30(iii)](a)+[6\times 5(i,ii,iii,iv,vi,vii)+10(v)](b)+[105(i,iii)+5(ii)](c)=$) **190 Points**) Time marching of the 2 dof problem in the figure.

   Solve the MDOF ODE corresponding to the problem from previous questions:

$$\mathbf{M}\ddot{\mathbf{U}}+\mathbf{K}\mathbf{U}=\mathbf{R}$$

   that is the system with no damping. The mass matrix $\mathbf{M}$ takes either the consistent form $\mathbf{M}_c$ or diagonal (lumped) form $\mathbf{M}_d$ depending on the method used. First part of this problem is about using the Newmark method as an example of a single-step

method and the second part the central difference method as an example of a linear multi step (LMS) method. Use the force function,

$$F(t) = 1$$

Note: I suggest to simulate the same problems below but with a load $F(t)$ that takes $t = 1000$ to attain its final value of 1 (final time should also be set to $T_f = 1000$) as an example on how the static solution would look like. Also, you can use $F(t) = \sin(6t)$ and an example of a harmonic load to observe the differences between the two methods. You **do not** need to return results for the two $F(t)$ mentioned here and only $F(t) = 1$ is requested to be analyzed.

Some other parameters that you need for both questions below are:

- Final time $T_f = 10$.
- Time step $\Delta t = 0.01$ (fine time step) and $\Delta t = 0.543$ (coarse time step).
- ICs: Both initial displacement and velocity are zero: $\mathbf{U}(t=0) = 0, \dot{\mathbf{U}}(t=0) = 0$.

(a) **Newmark method**: Use the parameters $\alpha = 1/4$ and $\delta = 1/2$. This Corresponds to "Average acceleration", *i.e.,* trapezoidal rule as one of the methods that is a special case of Newmark method.

i. Is the Newmark method with given α, δ conditionally stable or unconditionally stable?

ii. Which mass matrix (consistent or diagonal) will you use with this method and why (answer in one sentence)?

iii. Using equation (259) from the course notes (shown below)

$$\begin{aligned} {}^{t+\Delta t}\dot{\mathbf{U}} &= {}^{t}\dot{\mathbf{U}} + [(1-\delta)\,{}^{t}\ddot{\mathbf{U}} + \delta\,{}^{t+\Delta t}\ddot{\mathbf{U}}]\,\Delta t \\ {}^{t+\Delta t}\mathbf{U} &= {}^{t}\mathbf{U} + {}^{t}\dot{\mathbf{U}}\,\Delta t + [(\tfrac{1}{2}-\alpha)\,{}^{t}\ddot{\mathbf{U}} + \alpha\,{}^{t+\Delta t}\ddot{\mathbf{U}}]\,\Delta t^2 \end{aligned} \tag{2}$$

and knowing the values of α and δ obtain $\ddot{\mathbf{U}}^{n+1}$ in terms of $\mathbf{U}^{n+1}$ from the second equation, plug it in the first equation to obtain $\dot{\mathbf{U}}^{n+1}$ in terms of $\mathbf{U}^{n+1}$ and finally plug both values in the update equation for t_{n+1},

$$\mathbf{M}\ddot{\mathbf{U}}^{n+1} + \mathbf{C}\dot{\mathbf{U}}^{n+1} + \mathbf{K}\mathbf{U}^{n+1} = \mathbf{R}^{n+1}$$

to obtain,

$$\hat{\mathbf{K}}\mathbf{U}^{n+1} = \hat{\mathbf{R}}, \qquad \text{where} \tag{3a}$$

$$\hat{\mathbf{K}} = a_k\mathbf{K} + a_c\mathbf{C} + a_m\mathbf{M}, \qquad \text{and} \tag{3b}$$

$$\hat{\mathbf{R}} = \mathbf{R}^{n+1} + \mathbf{M}(m_0\mathbf{U}^n + m_1\dot{\mathbf{U}}^n + m_2\ddot{\mathbf{U}}^n) + \mathbf{C}(c_0\mathbf{U}^n + c_1\dot{\mathbf{U}}^n + c_2\ddot{\mathbf{U}}^n) \tag{3c}$$

provide numerical values for $a_k, a_c, a_m, m_0, m_1, m_2, c_0, c_1, c_2$ for the specific values of α, δ given herein. **Do not** directly used the formulas in table 9.4 (pages 302 and 303 of the course notes) to calculate these values. Rather, directly follow from the process described above to compute them.
Note: For this answer use symbolic $\mathbf{M}, \mathbf{K}, \mathbf{C}$. For numerical calculation we use $\mathbf{K}$ and $\mathbf{M}$ from (**??**) (with lumped or consistent mass matrix used based on your answer from question 2 and damping matrix $\mathbf{C} = \mathbf{0}$).

(b) **Central difference method**: Use the scheme §4.3.1 and equations (**??**) and (245) to solve the 2 dof problem with provided system matrices, time step, and final time.

i. Is this scheme explicit or implicit?

ii. If the central difference method is explicit name an implicit LMS method and if implicit name an explicit LMS method for structural dynamics.

iii. Is it conditionally stable or unconditionally stable?

iv. What mass matrix should be used with central difference method (consistent or diagonal)? Provide at least **two reasons** for your answer, including comments on the efficiency of the method **and** which one is more appropriate with this method in terms of handling **period elongation error (*i.e.,* frequency error)**.

v. If the method is conditionally stable what will be the absolute maximum time step we can take with this method if consistent or lumped mass matrices are used. Use frequencies from global system to answer this question (that is questions 1(b)i and 1(b)iii above). Label these two time steps as Δt_{c_M} and Δt_{d_M}.
Note: Note that the maximum time step of central difference method is not simply $1/\omega_m$ with ω_m being a representative maximum frequency. It may be a factor of it (obviously if it is conditionally stable). You need to refer to the course notes to find the factor in front of $1/\omega_m$ (if any).

vi. Based on the answer to the previous question, which mass matrix will require a more stringent time step (if any). That is which one of Δt_{c_M} and Δt_{d_M} is smaller?

vii. Now given that we often do not find the actual frequencies of a structure, use the maximum element frequencies (from item 1(b)ii and 1(b)iv) to compute maximum time steps $\Delta t^e_{c_M}$ and $\Delta t^e_{d_M}$ in which instead of the correct MDOF system maximum frequency we use element level maximum frequencies (from item 1(b)ii and 1(b)iv). Is this approach conservative (That is $\Delta t^e_{c_M} < \Delta t_{c_M}$ and $\Delta t^e_{d_M} < \Delta t_{d_M}$)? What are the values of $\Delta t^e_{c_M}$ and $\Delta t^2_{d_M}$?

(c) Numerically solve both **Newmark method** (with given parameters α, δ) and **central difference** method for the given $\mathbf{K}, \mathbf{M}$ and $\mathbf{R}$ (**??**), Δt, T_f and $F(t)$, and ICs. For each method choose its appropriate mass matrix (consistent or lumped) from (**??**). Deliverables are,

i. Plots of the free end point displacement ($U_2(t)$) (y axis) versus time (x axis) for **both Newmark and central difference methods** for $t = 0$ to T_f and **both time steps** Δt provided.

ii. In terms of the results you obtain with the two values of Δt explain the type of solutions you observe? What type of problems we can encounter by using large time steps with conditionally stable and unconditionally stable methods?

iii. Source code(s) (Matlab, Mathematica, C++, *etc.*) used for your computations.

3. ($(5(a) + 90(b, c) + 5(d) =$) **100 Points**) **Runge-Kutta method**: Solve the SDOF initial value problem,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y) = yt^3 - 1.5y \tag{4}$$

for $t = 0$ to $T_f = 2$ and IC $y(0) = 1$. Display all your results on the same plot.

- Analytically.
- Forward Euler method with $\Delta t = 0.5$ and $\Delta t = 0.125$.
- Midpoint method with $\Delta t = 0.5$.
- Fourth-order explicit RK method (RK4) with $\Delta t = 0.5$.

Deliverables are,

(a) Derivation of the analytical solution.

(b) A plot showing all the solutions mentioned above.

(c) Source code(s) used for your computations.

(d) Between Euler method with $\Delta t = 0.125$ and RK4 method with $\Delta t = 0.5$ both have the same "stop points" while advancing the solution. With the former they are time steps and in the latter one in every four is a time step and the other three are stages. Which one is more accurate? Explain why?

Hint: It will be much easier to computer code the equation (262) (shown below) for general a, b, c, s:

$$y_{n+1} = y_h + \Delta t \sum_{i=1}^{s} b_i k_i \qquad \text{where} \tag{5a}$$

$$k_i = f(t_n + \Delta t c_i, y_n + \Delta t \sum_{j=1}^{i-1} a_{ij} k_j), \qquad 1 \le i \le s \tag{5b}$$

All needed to be done is provide the values for vectors b, c and matrix a in the computer code and switch based on the value of s (and potentially options within a given s such as Heun and Midpoint methods for $s = 2$) in your code. All methods mentioned (forward Euler, Midpoint, and RK4) are RK methods with different s, a, b, c. You can very easily code a general purpose RK method that has s, Δt, T_f and the function f (or an option number for the coded functions f) as input arguments to your program.

1. ((10(*a*) + 50(*b*) =) **60 Points**) **Simple von Neumann analysis**: Stability analysis of forward-time central-space (FTCS) method.

   (a) Using the FD equation (27c)

   $$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$$

   for the advection equation,

   $$u_{,t} + au_{,x} = 0 \tag{1}$$

   show that the update equation is,

   $$v_m^{n+1} = v_m^n + \frac{\bar{k}}{2}v_{m-1}^n - \frac{\bar{k}}{2}v_{m+1}^n \quad \text{for} \tag{2a}$$

   $$\bar{k} = \frac{ka}{h} \qquad \text{Normalized time step for advection equation} \tag{2b}$$

   (b) Using von-Neumann analysis and by using (448),

   $$v_{m+a}^{n+b} = e^{\imath a\theta} g^b v_m^n \qquad \text{Temporally one-step FD stencils} \tag{3}$$

   for this one step method show that the amplification factor $g > 1$ for all $\bar{k}$ so the method is unconditionally unstable.

2. ((10(*a*) + 10(*b*) + 10(*c*) + 20(*d*) =) **50 Points**) **PDE mode transition and design of a FD stencil**: Consider the following PDE from (553d),

   $$\tau u_{,tt} + u_{,t} - Du_{,xx} = 0 \qquad \text{1D relaxed diffusion equation} \tag{4}$$

   where $\tau \geq 0, D > 0$. This PDE is hyperbolic for $\tau > 0$ and if $\tau = 0$ the equation is a simple diffusion equation whose response is characterized by damping and diffusion solution u. On the other hand if we do not have $u_{,t}$ term the PDE is $\tau u_{,tt} - Du_{,xx} = 0$ which is a wave equation which splits and propagates IC on value $(u_0(x))$ to the left and right with speeds $\pm a$, $a = \sqrt{\frac{D}{\tau}}$ (*cf.* D'Alembert solution (591)). That is the response of these systems are vastly different. If appropriate explicit FD schemes are used their corresponding maximum time steps are proportional to D/h^2 and c/h respectively.

   Limiting cases for the PDE (553d): Using dimensional analysis and discussion in §7.3 describe that for length scale $\tilde{L}$,

   $$\tilde{L} = \sqrt{\tau D} \tag{5}$$

   we have the following two limiting cases for equation (553d)

   $$\tilde{L}_0 \ll \tilde{L} \qquad u_{,tt} - a^2 u_{,xx} = 0 \qquad k_{\max} \propto \frac{h}{a} \qquad \text{Undamped hyperbolic limit} \tag{6a}$$

   $$\tilde{L}_0 \gg \tilde{L} \qquad u_{,t} - Du_{,xx} = 0 \qquad k_{\max} \propto \frac{h^2}{D} \qquad \text{Diffusion (parabolic) limit(??)} \tag{6b}$$

   where

   $$a = \sqrt{\frac{D}{\tau}} \quad \text{is the wave speed} \tag{7}$$

   , $\tilde{L}_0$ is a length scale of interest, *e.g.,* observation length scale (*i.e.,* element length h or length scale relevant to a particular problem considered) and $k_{\max}$ is the maximum time step of an explicit method.

   Note: For the dimensional analysis provide response for the following three items:

   (a) The length scale implied by the PDE (553d) is $\tilde{L} = \sqrt{\tau D}$.

   (b) The time step limit for (**??**) (if the scheme is conditionally stable with a maximum time step $k_{\max}$) is proportional to $\frac{h}{a}$.

   (c) The time step limit for $k_{\max}$ in (**??**) (again if a maximum time step $k_{\max}$ exists) is proportional to $\frac{h^2}{D}$.

   In the dimensional analysis you need to use the scales of parameters involved. For example, $[\tau] = T, [a] = L/T$ where [.] is the physical dimension of a quantity and L, T are length and time respectively.

   Fourth item (d): For the explanation of why for small length (and time) scales the undamped hyperbolic limit is apporached and why for large ones the diffusion limit is approached you can refer to §7.3 and equations (570b), (571), and (572) (particularly (572b)). Be **very brief (less than 4-5 sentences) in your explanation**.

3. (((25 + 25 + 3 × 10)(a) + (25 + 25 + 4 × 10)(b) =) **170 Points**) **FD formulation for a problem of the form** (553d): We want to formulate an appropriate explicitly FD formulation for (553d). The schemes considered are both consistent so the proof of stability will be sufficient for establishing their convergence. The stability analysis also provides the maximum time step $k_{\max}$ which is not clear what it would be for a problem of the type (553d).

   The stability analysis for both cases involves von Neumann analysis which plugs (445),

$$v_m^n = e^{\imath m\theta}\hat{v}^n \tag{8}$$

   in the FD stencil. For both methods considered we obtain an equation for amplification factor g ($\hat{v}^{n+1} = g\hat{v}^n$) in the form,

$$g^2 - 2A_1 g + A_2 = 0 \tag{9}$$

   where coefficients A_1 and A_2 will be obtained based on the von Neumman analysis (*i.e.,* insertion of (**??**) in the FD stencil).

   If the coefficients A_1 and A_2 are real (which will be the case for the examples in this HW), the condition $g \leq 1$ is equivalent to,

$$-1 \leq A_2 \leq 1, \qquad -\frac{A_2+1}{2} \leq A_1 \leq \frac{A_2+1}{2} \tag{10}$$

   Side Note (FYI): Equation (**??**) is basically the same as (362) but allowing the case $|A_1| = A_2 = 1$ given that we allow repeated root of $g_1 = g_2 = \pm 1$ since for a hyperbolic equation (553d) growth in the form $\hat{v}^{n+1} = (\pm 1)t\hat{v}^n$ is allowed; *cf.* §6.4.2 (480) and (487) for further discussion. Finally, since in all the problems considered in this assignment g does not explicitly depend on k the simpler stability condition $|g| \leq 1$ is used rather than (432) $|g| < 1 + Kk$ for one-step methods and its generalization for the two step methods herein.

   Below we consider two different stencils for the solution of (553d). By your analysis you will observe that one is more appropriate for the solution of this PDE.

   (a) **CCC scheme**: We use central difference for $u_{,tt}$, $u_{,t}$, and $u_{,xx}$ in (553d) to obtain,

$$\tau\frac{v_m^{n+1} - 2v_m^n + v_m^{n-1}}{k^2} + \frac{v_m^{n+1} - v_m^{n-1}}{2k} - D\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} = 0 \tag{11}$$

   Demonstrate the following:

   i. By von Neumann analysis (*i.e.,* plugging $v_m^n = e^{\imath m\theta}\hat{v}^n$ from (**??**)) demonstrate

$$g^2 - 2A_1 g + A_2 = 0 \qquad \text{where} \quad \begin{cases} A_1 = \frac{2\bar{\tau} - 4\bar{\bar{k}}\sin^2\left(\frac{\theta}{2}\right)}{2\bar{\tau}+1} \\ A_2 = \frac{2\bar{\tau}-1}{2\bar{\tau}+1} \end{cases} \tag{12a}$$

$$\bar{\bar{k}} = \frac{kD}{h^2} \qquad \text{normalized time step for parabolic PDE } u_{,t} - Du_{,xx} = 0 \tag{12b}$$

$$\bar{\tau} = \frac{\tau}{k} \qquad \text{normalized } \tau \text{ by } k \tag{12c}$$

   ii. Using (**??**) and the condition (**??**) show that the stable time step for the scheme (**??**) is,

$$k \leq k_{\max}, \quad \text{for} \quad k_{\max} = \frac{h}{a}, \quad \text{where from (??)} \quad a = \sqrt{\frac{D}{\tau}}, \quad \text{that is} \tag{13a}$$

$$\bar{k} \leq 1, \quad \text{where} \tag{13b}$$

$$\bar{k} = \frac{ka}{h}, \qquad \text{normalized time step for the wave equation } u_{,tt} - a^2 u_{,xx} = 0 \tag{13c}$$

   Note: Clearly, if you cannot derive (**??**) you can still use it and (**??**) to demonstrate (**??**).

   iii. Compare the stability condition (**??**) with that of the central time central space stencil for the undamped hyperbolic equation $u_{,tt} - a^2 u_{,xx} = 0$ ($a = \sqrt{D/\tau}$) which is discussed in §6.4.2.

   iv. Based on your answer from previous item, does the term $u_{,t}$ in (553d) (discretized by central time term $\frac{v_m^{n+1} - v_m^{n-1}}{2k}$ in (**??**)) affect the stability of $\tau u_{,tt} + u_{,t} - Du_{,xx} = 0$ in the FD scheme (**??**) ? Later, you will comment whether this behavior is favorable or not.

   v. Now, let us focus on another feature of (**??**) when $\tau = 0$, that is when we solve the diffusion equation $u_{,t} - Du_{,xx} = 0$. Your analysis from (**??**) and (**??**) still holds for this case, with the difference that $g_1 = g_2 = \pm 1$ is not acceptable anymore given that for this temporally first order PDE (similar to any other temporally first order PDE) FD growth in the form $\hat{v}^{n+1} = (\pm 1)t\hat{v}^n$ is not permitted; *cf.* the analysis of leapfrog method in §6.4.1.

In any case, apart from this minor point (that $g_1 = g_2 = \pm 1$ is not acceptable for $\tau = 0$) show that (**??**) for $\tau = 0$ is **unconditionally unstable**. That is, the discretization

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} - D\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} = 0 \tag{14}$$

for $u_{,t} - Du_{,xx} = 0$ is unconditionally unstable.
Hint: As mentioned above, the analysis above ((**??**) and (**??**)) for general τ would carry for $\tau = 0$. So, you do not need to do the von Neumann analysis for (**??**) from the beginning.

(b) **CFC scheme**: We use central difference for $u_{,tt}$ and $u_{,xx}$ and forward difference for $u_{,t}$ in (553d) to obtain,

$$\tau\frac{v_m^{n+1} - 2v_m^n + v_m^{n-1}}{k^2} + \frac{v_m^{n+1} - v_m^n}{k} - D\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} = 0 \tag{15}$$

Demonstrate the following:

i. By von Neumann analysis (*i.e.,* plugging $v_m^n = e^{im\theta}\hat{v}^n$ from (**??**)) demonstrate

$$g^2 - 2A_1 g + A_2 = 0 \qquad \text{where} \quad \begin{cases} A_1 = \frac{\bar{\tau} + \frac{1}{2} - 2\bar{\bar{k}}\sin^2\left(\frac{\theta}{2}\right)}{\bar{\tau}+1} \\ A_2 = \frac{\bar{\tau}}{\bar{\tau}+1} \end{cases} \tag{16}$$

where as in (**??**) $\bar{\bar{k}} = \frac{kD}{h^2}$ and $\bar{\tau} = \frac{\tau}{k}$.

ii. Using (**??**) and the condition (**??**) show that the stable time step for the scheme (**??**) is,

$$k \le k_{\max}, \quad \text{for} \quad k_{\max} = \frac{h^2}{2D}\left(\frac{1 + \sqrt{1 + \left(4\frac{\tilde{L}}{h}\right)^2}}{2}\right) \tag{17a}$$

recall $\tilde{L} = \sqrt{\tau D}$ from (**??**).
Hint: The application of (**??**) on (**??**) yields $\bar{\bar{k}} \le \bar{\tau} + \frac{1}{2}$ whose solution results in (**??**). Also again you can directly proceed from (**??**) if you fail to obtain the values A_1, A_2 from (**??**).

iii. **Very small grid size limit**: Consider the limiting case $\frac{h}{\tilde{L}} \ll 1$, *i.e.,* a very small grid size relative to the length scale $\tilde{L}$ for the PDE (553d). What time step we obtain from (**??**) and does it match what is expected from (**??**) for the undamped hyperbolic $\tau u_{,tt} - Du_{,xx} = 0$ limit of (553d) ($\tau u_{,tt} + u_{,t} - Du_{,xx} = 0$) for $h(= \tilde{L}_0) \ll \tilde{L}$.

iv. **Very large grid size limit**: Consider the limiting case $\frac{h}{\tilde{L}} \gg 1$, *i.e.,* a very large grid size relative to the length scale $\tilde{L}$ for the PDE (553d). What time step we obtain from (**??**) and does it match what is expected from (**??**) for the diffusion equation $u_{,t} - Du_{,xx} = 0$ limit of (553d) ($\tau u_{,tt} + u_{,t} - Du_{,xx} = 0$) for $h(= \tilde{L}_0) \gg \tilde{L}$.

v. In light of the previous two answers, compare the efficiency of CCC scheme (**??**) and CFC scheme (**??**) for the solution of (553d) ($\tau u_{,tt} + u_{,t} - Du_{,xx} = 0$) in the limit $h \gg \tilde{L}$. Which scheme gives smaller time step which happens to also be consistent with the physics of the problem from (**??**)?

vi. Based on all previous questions which scheme (CCC or CFC) would you use for the solution of (553d)?

Side Note (FYI): The solution of problems that have different limiting PDEs as the relevant length/time scales vary (*i.e.,* grid space and time steps) is a very active research topic. For example the relaxed advection-diffusion-reaction problem $\tau u_{,tt} + u_{,t} + vu_{,x} - Du_{,xx} = -ru$ (v is advection speed and r the reaction rate) shows several PDE mode transitions. The design of numerical methods that can consistently solve all limiting cases in a unified manner is a challenging task. For more information refer to the discussion on "asymptotic preserving" schemes in [Jin, 2010] (shared with you in dropbox folder Dynamics of continua/Books_courseNotes/StiffSystems/Relaxation).

4. ((25(a) + 25(b) + 40(c) + 15(d) + 15(e) + 20(f, **extra credit**) =) **120 + 20 e.c. Points**) **An unconditionally stable explicit method (with conditional consistency)!:** FD schemes that can be explicitly solved and have no time step constraint are not encountered except in special cases in 1D where the update of an implicit method can be done explicitly; *cf.* BTBS scheme applied to 1D advection equation $u_{,t} + au_{,x} = 0$ in §2.1.9 as one example. Genuinely explicit methods often have a time step limit imposed by stability constraint. Below, we show an explicit method that does not have any stability limits but its conditional consistency instead limits its time step.

Consider the diffusion equation,

$$u_{,t} - Du_{,xx} = 0 \tag{18}$$

solved with Dufort-Frankel FD method,

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} - D\frac{v_{m+1}^n - (v_m^{n-1} + v_m^{n+1}) + v_{m-1}^n}{h^2} = 0 \tag{19}$$

which is basically similar to (**??**) but $-2v_m^n$ in the stencil for $u_{,xx}$ being replaced by $(v_m^{n-1} + v_m^{n+1})$.

(a) **von Neuman stability analysis**: By the insertion of (**??**) in the FD stencil (**??**) show that the equation for g is,

$$g^2 - 2A_1 g + A_2 = 0 \qquad \text{where} \quad \begin{cases} A_1 = \frac{2\bar{\bar{k}}\cos\theta}{2\bar{\bar{k}}+1} \\ A_2 = \frac{2\bar{\bar{k}}-1}{2\bar{\bar{k}}+1} \end{cases} \tag{20}$$

(b) By applying (**??**) to (**??**) (the equation (**??**) cannot have repeated roots $g_1 = g_2 = \pm 1$ so we do not need to worry about such repeated roots for the temporally first order ODE (**??**)) demonstrate that the explicit method of Dufort-Frankel is **unconditionally stable!**

(c) **Conditional consistency**: By plugging Taylor series expansion of terms in (**??**) for an infinitely smooth function ϕ, *e.g.*, $\phi_m^{n+1} = \phi_m^n + k\dot{\phi} + \sum_{i=2}^{\infty} \frac{k^i}{i!}\frac{\partial^i \phi}{\partial t^i}$, $\phi_{m+1}^n = \phi_m^n + h\phi_{,x} + \sum_{i=2}^{\infty} \frac{h^i}{i!}\frac{\partial^i \phi}{\partial x^i}$, form numerical PDE operator $P_{h,k}\phi$. Then separate the exact PDE operator $P\phi = u_{,t} - Du_{,xx}$ and show,

$$P_{h,k}\phi - P\phi = \frac{1}{6}k^2\frac{\partial^3\phi}{\partial t^3} - h^2\frac{D}{12}\frac{\partial^4\phi}{\partial x^4} + \frac{k^2}{h^2}D\ddot{\phi} + \text{H.O.T.} \tag{21}$$

(d) From (**??**) show that (**??**) is conditionally consistent. That is for this scheme $P_{h,k}\phi - P\phi \to 0$ when $h, k \to 0$ only **when k tends to zero faster than h.**

(e) Inconsistency of a FD scheme basically means that we are solving another PDE. From (**??**) it is evident that if $D\frac{k^2}{h^2}$ is bounded and not tending to zero; *e.g.*, $k \propto h$ as for an explicit scheme for a hyperbolic PDE, we basically solve $D\frac{k^2}{h^2}u_{,tt} + u_{,t} - Du_{,xx}$. This can be demonstrated in another way too; we rewrite (**??**) as,

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} - D\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + \text{ other terms} = 0 \tag{22}$$

Show that the "other terms" actually correspond to $D\frac{k^2}{h^2}u_{,tt}$. Clearly the first two terms are the FD stencils for $u_{,t}$ and $Du_{,xx}$.

Side Note (FYI): In fact, many FD schemes basically add other differential terms to the PDE that for a consistent scheme they must tend to zero as $h, k \to 0$. For example, in the discussion of FV solution for elastodynamic problem in §2.2.6.2 we discussed that the use of Riemann fluxes effectively adds a numerical diffusion coefficient (116) $D_h = \frac{hc}{2}$ whose value (and its contribution to the PDE) vanish as grid size $h \to 0$. For this problem, the relaxation term $D\frac{k^2}{h^2}u_{,tt}$ only vanishes if the scheme is consistent; *i.e.*, in this case k tending to zero faster than h.

(f) **Relation to CCC scheme above (extra credit, do not need to return)**: As discussed in previous question Dufort-Frankel is basically solving $D\frac{k^2}{h^2}u_{,tt} + u_{,t} - Du_{,xx} = 0$. A closer examination of what you obtain from (**??**) (*i.e.*, when other terms are determined) shows that this equation is identical to CCC scheme (**??**) for $\tau = \frac{Dk^2}{h^2}$. Show that the time step constraint (**??**) is basically trivially satisfied.

Side Note (FYI): This is another confirmation that Dufort-Frankel method is unconditionally stable, but obviously when $\tau = \frac{Dk^2}{h^2}$ does not tend to zero when $h, k \to 0$ we have a stable method that converges to the solution to a relaxed diffusion equation rather than the underlying diffusion equation (**??**) ($u_{,t} - Du_{,xx} = 0$). Sometimes detection of inconsistencies of numerical methods can be difficult since the solution does not blow up as with strongly unstable method and we may get reasonably well-behaved solutions but with the solutions corresponding to another PDE!

Side Note (FYI): As another interesting feature of Dufort-Frankel we observe all needed for consistency is that k tend to zero faster than h. That is, $k/h \to 0$ as $h, k \to 0$. Clearly, typical time step of explicit methods for diffusion equation (**??**) ($u_{,t} - Du_{,xx} = 0$) requires $k < h^2/D$. So, a scaling of the form $k \propto h^2/D$ makes Dufort-Frankel scheme consistent.

What can be achieved beyond this is that any time step scaling with $k/h \to 0$ works for Dufort-Frankel scheme. For example, we can have $k \propto h^{1.5}$ or even $k \propto h^{1.001}$ and still have a consistent scheme with much less stringent time step than a scaling of the form $k \propto h^2/D$ which is typical for explicit solvers of diffusion equation. Clearly, the slower the relaxation time $\tau = \frac{Dk^2}{h^2}$ tends to zero the slower the convergence of FD solution will be to the exact solution but again the interesting feature is that such scalings would result in a convergent scheme while typcially for explicit schemes applied to diffusion equation $k \propto h^2$ is required for convergence.

1. ((20(*a*) + 30(*b*) + 30(*b*) =) **80 Points**) **Dispersion relation / phase and group velocity**: Consider the relaxed *advection-diffusion-reaction*(ADR) PDE,

$$\tau u_{,tt} - \nu u_{,xx} + u_{,t} + v u_{,x} = -ru \tag{1}$$

where $\tau, \nu, v, r \geq 0$ are the relaxation time, diffusion coefficient, advection velocity, and reaction coefficient ratio and have the units of $[T], [L]^2/[T], [L]/[T], 1/[T]$ with $[L]$ and $[T]$ being the length(space) and time scales, respectively.

   (a) Obtain the dispersion relation for (**??**); *cf.* §7.5.

   (b) Obtain the phase velocity from (593) and group velocity from (598) for the right-moving waves (*i.e.,* $\omega_R > 0$).

   (c) Write a finite difference stencil for (**??**) by using central difference stencils for $\ddot{u}^n_j, (u_{,xx})^n_j$, forward Euler for $\dot{u}^n_j$, and backward Euler for $v(u_{,x})^n_j$.
   Note the von Neumann analysis of this FD scheme provides the time step of this complex PDE and demonstrates that at different space/time scales time scale matches what the physical limits of the PDE dictated (*i.e.,* wave equation, diffusion equation, reaction equation, etc). The derivation of the time step and von Neumman analysis can be time consuming and only the expression of the FD scheme is requested.

2. ((10(*a*)+10(*b*)+30(*c*)+10(*d*) =) **60 Points**) **Dispersion and dissipation error analysis**: Consider the advection equation,

$$u_{,t} + a u_{,x} = 0$$

solved by the Lax-Friedrich's scheme. By referring to (630b) for the values of g_R, g_I and following the solution scheme discussed in sections §7.7.5 and §7.7.6 for dispersion and dissipation analysis answer the following.

   (a) Show that numerical real and imaginary components of frequency for Lax-Friedrich's method ω^h are,

$$\omega^h_I = \frac{1}{2k}\log(g_R^2 + g_I^2) = \frac{1}{2k}\log\left(\cos^2\theta + \bar{k}^2\sin^2\theta\right) \tag{2a}$$

$$\omega^h_R = \frac{1}{k}\tan^{-1}\left(\frac{-g_I}{g_R}\right) = \frac{1}{k}\tan^{-1}\left(\bar{k}\tan\theta\right) \tag{2b}$$

   (b) For value of $\bar{k}$, ω^h_R, ω^h_I match the exact values $\omega_R = a\xi, \omega_I = 0$?

   (c) Show the asymptotic expressions, *i.e.,* $\theta \to 0$, for dissipation (and amplitude decay) and dispersion (and period elongation) for Lax Friedrich's method are,

$$\frac{\Delta\omega_I}{\omega_R} = \frac{1}{2\pi}\log(1 - A_d) \approx -\frac{1}{2\pi}A_d = \frac{1}{8}\left(\frac{\bar{k}^2 - 1}{\bar{k}}\right)\theta + \mathcal{O}\left(\theta^3\right) \tag{3a}$$

$$\frac{\Delta\omega_R}{\omega_R} = -\frac{\Delta T}{T} = \frac{1}{3}(1 - \bar{k}^2)\theta^2 + \mathcal{O}\left(\theta^4\right) \tag{3b}$$

   (d) The leading terms are zero for $\bar{k} = 1$. What do you expect the values of the higher order terms be for $\bar{k} = 1$ and why?

3. ((20(*a*) + 20(*b*) + 20(*c*) =) **60 Points**) **Well-posedness, dynamic stability, and robustness**:

   (a) Show the following two equations are ill-posed.

$$u_{,tt} - u_{,x} = 0$$
$$u_{,ttt} - u_{,xx} = 0$$

   (b) Show that the following wave equation with negative reaction coefficient $r < 0$ is well-posed but not dynamically stable,

$$u_{,tt} - a^2 u_{,xx} = -ru \tag{4}$$

   (c) Show that Euler-Bernoulli equation (545),

$$u_{tt} + b^2 u_{,xxxx} = 0, \tag{5}$$

   while being well-posed, is not robust.
   Side Note (FYI): You can refer to §6.5 for the discussion of these concepts. Two of the examples above are directly discussed and solved in the course notes.