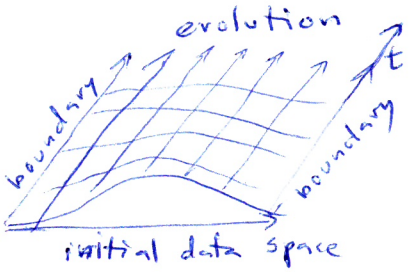
Numerical solution of partial differential equations

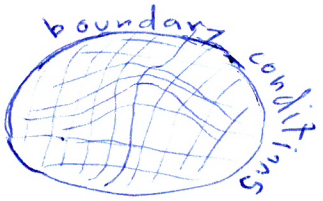
- instead of classification as elliptic, parabolic or hyperbolic equations or problems, in numerical analysis, it is better to distinguish

1) initial value problems



- examples: wave equations, heat equation
- usually parabolic or hyperbolic eq.
- initial data at certain time + boundary conditions in space
- importance of stability

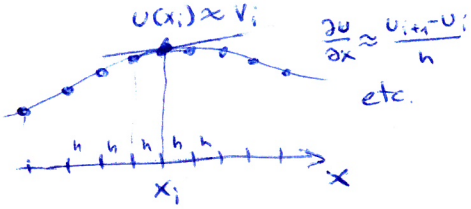
2) boundary value problems



- Laplace, Poisson equations
- usually elliptic problems
- only boundary conditions
- stationary solution

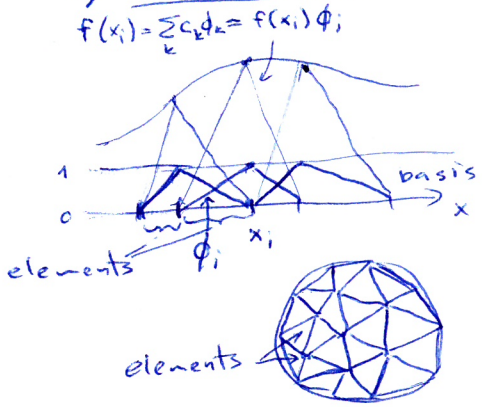
- basic numerical techniques

1) finite-difference method (FDM)



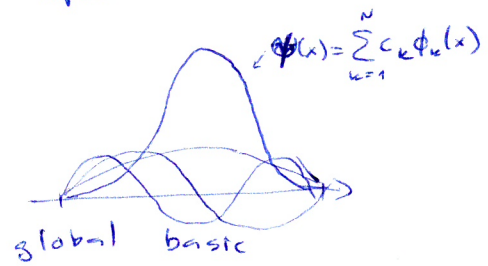
- discretization of continuous quantities in space and time
- approximation of derivatives using finite differences
- either explicit formulas for evolution or solution of (large) systems of linear (algebraic) equations
- difficult to deal with boundary conditions in space

2) finite-element method (FEM)

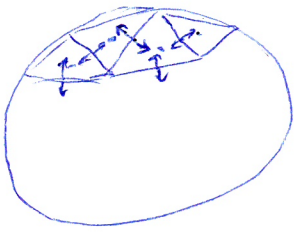


- expansion of solutions into a basis with a compact support based on separation of space into so-called finite elements
- weak formulation for easier use of non-smooth basis functions
  ⇒ large, but sparse systems of linear equations
- easier treatment of boundary conditions

3) spectral methods - expansion of solution into a global basis => full matrices
- useful especially for sufficiently smooth solutions when high accuracy is required



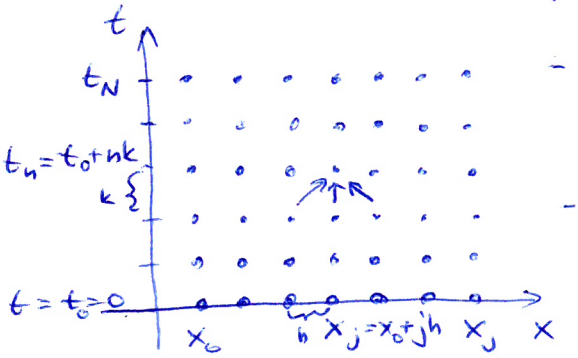
4) finite-volume method (FVM)



- division of space into finite cells around chosen points, averaging over cells
- integrals over cells replaced by surface integrals => balancing currents on surfaces
- direct use of conservation laws

Finite-difference method for linear PDEs

- discretization of space and time



- points (x_j, t_n) are usually distributed on equidistant grids
- at these points we approximate a solution of a certain PDE (or of a system of PDEs)

$$u(x_j, t_n) \simeq V_j^n$$

- for example we want to solve the heat equation in its simplest form

$$u_t = u_{xx} \quad \text{or} \quad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

with some initial condition

$$u(x, t=0) = u_0(x)$$

and boundary conditions at x_0 and x_J (this can be a Dirichlet or von Neumann condition)

- generalization to more space variables (dimensions) is straightforward, but in general more difficult if the region boundary is complicated, espacially if we want to use higher-order methods

- goal: to transform PDE into a finite system of linear equations and get the best approximation of $u(x,t)$ at points (x_j, t_n)

  - explicit methods - straightforward solution
    - each equation contains only one unknown
  - implicit methods - at each step in time it is necessary to solve a system of linear equations for all space points
    - more stable, but demanding

• formulas for finite differences

- in general, we consider s-step difference scheme, in which we use s known previous values at each point in space $v_j^{n-(s-1)}, v_j^{n-(s-2)}, \dots, v_j^n$ for $j=0,\dots,J$

  to determine values v_j^{n+1} at time t_{n+1}

- explicit formula if there is just one unknown v_j^{n+1}
- implicit formula if there are several v_j^{n+1} for different j in the formula
- the higher s and more points in space, the more accurate approximation in terms of a local discretization error, but the method will be
  1) more memory consuming (we need to save s values at each point
  2) less stable (usually)
  3) more difficult to initiate ($v_0(x)$ is not enough)

• sometimes the following operators are used to express various difference formulas:

- translations in time $Zv_j^n = v_j^{n+1}$, in space $Kv_j^n = v_{j+1}^n$

  ⇒ Operators of forward, backward and centered differences in space $\delta_+(h)v_j^n = \frac{1}{h}(K-I)v_j^n = \frac{1}{h}(v_{j+1}^n - v_j^n)$ with I being identity operator

  $\delta_-(h) = \frac{1}{h}(I-K^{-1})$, $\delta_0(h) = \frac{1}{2h}(K-K^{-1})$

  where $K^{-1}v_j^n = v_{j-1}^n$

  and averaging operators in space

  $\mu_+ = \frac{1}{2}(I+K)$, $\mu_- = \frac{1}{2}(K^{-1}+I)$, $\mu_0 = \frac{1}{2}(K^{-1}+K)$

  and similarly in time, but K is replaced by Z and an upper index is used, e.g. $\delta_+^t(k) = \frac{1}{k}(Z-I)$

- second order operators: $\delta_x(h) = \frac{1}{h^2}(K-2I+K^{-1}) = \delta_+\delta_- = \delta_-\delta_+$, $\delta^x(k) = \frac{1}{k^2}(Z-2I+Z^{-1})$

• examples of formulas for the simplest wave equation $u_t = u_x$

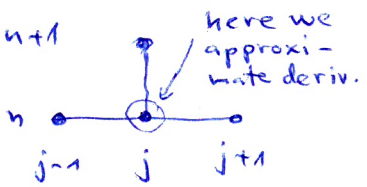
1) Euler explicit scheme (order 1, unstable)

$$\delta_+^t v_j^n = \frac{1}{k}\left(v_j^{n+1} - v_j^n\right) = \delta_0 v_j^n = \frac{1}{2h}\left(v_{j+1}^n - v_{j-1}^n\right)$$

forward difference in time at v_j^n $\approx \left.\frac{\partial u}{\partial t}\right|_{(x_j, t_n)}$

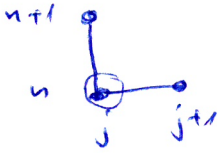
centered difference in space at v_j^n $\approx \left.\frac{\partial u}{\partial x}\right|_{(x_j, t_n)}$

stencil



2) upwind method (explicit) (order 1, stable for $\lambda = \frac{k}{h} \le 1$)

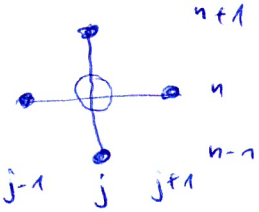
$$\delta_+^t v = \delta_+ v \quad \text{or} \quad \frac{1}{k}\left(v_j^{n+1} - v_j^n\right) = \frac{1}{h}\left(v_{j+1}^n - v_j^n\right)$$



3) leap-frog method (explicit, order 2, stable for $\lambda < 1$)

$$\delta_0^t v = \delta_0 v \quad \text{or} \quad \frac{1}{2k}\left(v_j^{n+1} - v_j^{n-1}\right) = \frac{1}{2h}\left(v_{j+1}^n - v_{j-1}^n\right)$$

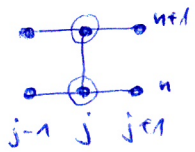
centered differences in space and time $\Rightarrow$ 2-step method



4) Crank-Nicolson method (implicit, order 2, stable)

$$\delta_+^t v = \mu^+ \delta_0 v \quad \text{or} \quad \frac{1}{k}\left(v_j^{n+1} - v_j^n\right) = \frac{1}{2}\left[\frac{1}{2h}\left(v_{j+1}^n - v_{j-1}^n\right) + \frac{1}{2h}\left(v_{j+1}^{n+1} - v_{j-1}^{n+1}\right)\right]$$

averaging in time of space centered differences

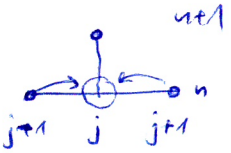


5) Lax-Friedrichs method (explicit, order 1, stable for $\lambda \le 1$)

- invented to stabilize Eurel method

$$\frac{1}{k}\left(zv - \mu_0 v\right) = \delta_0 v \quad \text{or} \quad \frac{1}{k}\left(v_j^{n+1} - \frac{1}{2}\left(v_{j+1}^n + v_{j-1}^n\right)\right) = \frac{1}{2h}\left(v_{j+1}^n - v_{j-1}^n\right)$$

averaging instead of v_j^n



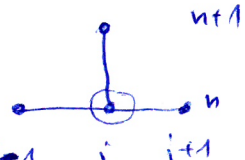
6) Lax-Wendroff method (explicit, order 2, stable for $\lambda \le 1$)

$$\delta_+^t v = \delta_0 v + \frac{k}{2}\delta_x v$$

or

$$\frac{1}{k}\left(v_j^{n+1} - v_j^n\right) = \frac{1}{2h}\left(v_{j+1}^n - v_{j-1}^n\right) + \frac{k}{2h^2}\left(v_{j+1}^n - 2v_j^n + v_{j-1}^n\right)$$

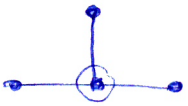
dissipative term for stability



• examples of formulas for the heat equation $u_t = u_{xx}$

1) **Euler method** (explicit, order 1, stable for $\frac{k}{h^2} \le \frac{1}{2}$)

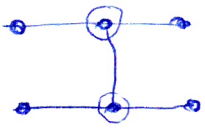
$$\delta^+ v = \delta_x v \quad \text{or} \quad \frac{1}{k}\left(v_j^{n+1} - v_j^n\right) = \frac{1}{h^2}\left(v_{j+1}^n - 2v_j^n + v_{j-1}^n\right)$$



2) **Crank-Nicolson method** (implicit, order 2, stable)

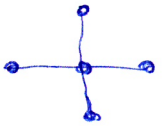
$$\delta^+ v = \mu^+ \delta_x v$$

average of the second differences



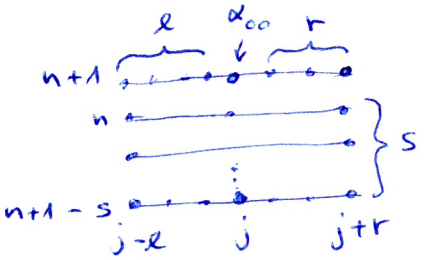
3) **Leap-frog method** (explicit, order 2, unstable)

$$\delta^\circ v = \delta_x v$$



• general s-step linear scheme of finite differences

$$\sum_{\nu=0}^{s} \sum_{\mu=-\ell}^{r} \alpha_{\mu\nu} v_{j+\mu}^{n-(\nu-1)} = 0$$



where $\alpha_{\mu\nu}$ are either constants

(for equations with constant coefficients)

or functions of x (and t)

- coefficient $\alpha_{00} \neq 0$ and $\alpha_{-\ell,\nu_1} \neq 0$ for at least one ν_1

  and $\alpha_{r,\nu_2} \neq 0$ for at least one ν_2

- if $\alpha_{\mu 0} = 0$ for $\forall\, \mu \neq 0$, the method is explicit

  otherwise it is implicit

- if $u(x,t)$ is a vector function and thus each v_j^n is also a vector

  then $\alpha_{\mu\nu}$ are matrices $M \times M$ (if $u(x,t)$ has M elements)

  and the condition $\det \alpha_{00} \neq 0$ etc.

- example: Crank-Nicolson method for $u_t = u_{xx}$

  - we have $s=1$, $\mu = -1, 0, 1$ ($\ell = 1$, $r = 1$) and $\nu = 0, 1$

$$\frac{1}{k}\left(v_j^{n+1} - v_j^n\right) = \frac{1}{2}\left[\frac{1}{h^2}\left(v_{j+1}^n - 2v_j^n + v_{j-1}^n\right) + \frac{1}{h^2}\left(v_{j+1}^{n+1} - 2v_j^{n+1} + v_{j-1}^{n+1}\right)\right] \qquad \alpha = \begin{pmatrix} -\frac{\sigma}{2} & -\frac{\sigma}{2} \\ 1+\sigma & -1+\sigma \\ -\frac{\sigma}{2} & -\frac{\sigma}{2} \end{pmatrix} \quad \text{with} \quad \sigma = \frac{k}{h^2}$$

• order of accuracy

- it is given by local discretization error of the finite-difference scheme, usually expressed in time step as $\mathcal{O}(k^{p+1})$ where p is then order of accuracy of the method
- space step h is usually considered as a function of k for example $h = \frac{k}{\lambda}$ with λ constant for $u_x = u_t$
- local discretization error can be determined in the same way as for numerical methods for ODEs, i.e. by Taylor series when we substitute into a scheme $u(x,t)$ for v_j^n, $u(x,t+k)$ for v_j^{n+1} etc. and then we find the first non-zero term of order k^{p+1}

examples:

1) Euler method for $u_t = u_x$

from scheme: $v_j^{n+1} = v_j^n + \frac{k}{2h}(v_{j+1}^n - v_{j-1}^n)$

we set

$$u(x,t+k) - u(x,t) - \frac{k}{2h}\left(u(x+h,t) - u(x-h,t)\right) =$$

$$= k u_t + \frac{k^2}{2} u_{tt} + \mathcal{O}(k^3) - \frac{k}{2h}\left(2h u_x + \frac{2h^3}{3!} u_{xxx} + \mathcal{O}(h^5)\right) =$$

$$= \underbrace{k(u_t - u_x)}_{0} + \frac{k^2}{2} u_{tt} + \mathcal{O}(k^3) \quad \text{including } \mathcal{O}(kh^2) \text{ if } h = \frac{k}{\lambda}$$

$$= \frac{k^2}{2} u_{tt} + \mathcal{O}(k^3) = \mathcal{O}(k^{1+1}) \Rightarrow p = 1$$

order of accuracy is 1, but it is unstable (see demonstration in Mathematica)

2) Leap-frog method for $u_t = u_x$

here $v_j^{n+1} = v_j^{n-1} + \frac{k}{h}(v_{j+1}^n - v_{j-1}^n)$

and thus $u(x,t+k) - u(x,t-k) - \frac{k}{h}\left(u(x+h,t) - u(x-h,t)\right) =$

$$= 2k u_t + 2\frac{k^3}{3!} u_{ttt} - \frac{k}{h}\left(2h u_x + 2\frac{h^3}{3!} u_{xxx}\right) + \mathcal{O}(k^5) \quad \text{if } h = \frac{k}{\lambda}$$

$= \mathcal{O}(k^3)$ for general λ

or $= 0$ for $\lambda = 1$ $(h = k)$ when it is basically exact method (in exact arithmetics)

3) Lax-Wendroff method for $u_t = u_x$

now

$$v_j^{n+1} = v_j^n + \frac{k}{2h}\left(v_{j+1}^n - v_{j-1}^n\right) + \frac{k^2}{2h^2}\left(v_{j+1}^n - 2v_j^n + v_{j-1}^n\right)$$

and thus (omitting terms $u(x,t)$ which cancel each other)

$$k u_t + \frac{k^2}{2} u_{tt} + \frac{k^3}{3!} u_{ttt} - \frac{k}{2h}\left(2h u_x + 2\frac{h^3}{3!} u_{xxx}\right)$$

$$- \frac{k^2}{2h^2}\left(2\frac{h^2}{2} u_{xx}\right) + O(k^4) = \underbrace{\left(\frac{k^3}{3!} - \frac{kh^2}{3!}\right)}_{0 \text{ for } k=h} u_{xxx} + O(k^4)$$

otherwise $p = 2$

Convergence and stability of finite-difference method

- similarly as for numerical methods for ODEs, also here

  convergence means that as $k \to 0$ and $h \to 0$ we get a correct solution of the given PDEs and

  stability means that the errors are not increased during numerical solution of the PDEs

- for ODEs, Dahlquist theorem about equivalence of convergence and of stability of consistent method can be proven for an arbitrary non-linear ODE which has a unique solution

- for PDEs, we have similar theorem, but only for linear eqs.

  here it is called Lax theorem and we do not have a general methods as for ODEs, but we work with specific schemes (methods) for each linear equation

- details can be found in literature, e.g. (if you are interested)

  Lax, Richtmyer: Comm. on Pure and Appl. Math. 9 (1956) 267

  Richtmyer, Morton: Difference methods for Initial-Value Problems, 2ed, Wiley 1967

  Trefethen: Finite Difference and Spetral Methods for ODE and PDE (incomplete), 1996 - online

  Thomée: SIAM Review 11 (1969) 152-195

- here we only provide summary of some results and terminology without proofs

• general formulation of the initial value problem

- let us consider a system of PDEs in the form

$$\frac{\partial u(t)}{\partial t} = A u(t), \quad 0 \le t \le T, \quad u(0) = u_0$$

where $A : \mathcal{B} \to \mathcal{B}$ is a linear operator on a Banach space $\mathcal{B}$ with a certain norm $\|\cdot\|$

- we assume that there exists one unique solution $u(t) \in \mathcal{B}$ for an arbitrary initial condition $u(0) = u_0$ and that the solution $u(t)$ depends continuously on u_0
- boundary conditions are included in the proper choice of $\mathcal{B}$
- the function $u(t)$ can depend on an arbitrary number of space variables (again hidden in the choice of $\mathcal{B}$) and it can be a vector function
- even systems with higher time derivatives are included by adding new components into $u(t)$

e.g. the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u = A u$$

can be rewritten as:

$$U(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \begin{pmatrix} u(t) \\ \frac{\partial u(t)}{\partial t} \end{pmatrix}$$

$$\frac{\partial U(t)}{\partial t} = \frac{\partial}{\partial t}\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \begin{pmatrix} u_2(t) \\ c^2 \Delta u_1(t) \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & 1 \\ c^2\Delta & 0 \end{pmatrix}\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \mathbb{A}\, U(t)$$

- for example $\mathcal{B}$ can be $L^2(\mathbb{R})$ and $A = \frac{\partial}{\partial x}$ for transfer or $A = \frac{\partial^2}{\partial x^2}$ for diffusion or heat transfer etc.

• general formulation of finite-difference methods

- to each method corresponds a class of bounded linear operators $S_k : \mathcal{B} \to \mathcal{B}$ which depend on time step k (we usually consider the space step h as a function of k, e.g. $h = \frac{k}{\lambda}$ for $u_t = u_x$ or $h = \sqrt{\frac{k}{\sigma}}$ for $u_t = u_{xx}$)

- for example for Euler method for $v_t = v_x$

we have $$v_j^{n+1} = v_j^n + \frac{k}{2h}\left(v_{j+1}^n - v_{j-1}^n\right)$$

to which the following operator S_k corresponds

$$u(x,t+k) \doteq S_k u(\cdot,t) = u(x,t) + \frac{k}{2h}\left(u(x+h,t) - u(x-h,t)\right)$$

- although S_k is an operator on functions from $\mathcal{B}$

we can use it also to write the formula as

$$v^{n+1} = S_k v^n \quad \text{or} \quad v^n = S_k^n v^0$$

(v^n: numerical solution after n time steps; v^0: initial vector representing initial data on the grid)

because the operator S_k evaluates the function $u(t)$ only at discrete points (e.g. x, $x+h$ and $x-h$)

- in general, the operators S_k could be dependent on time and we can rewrite in this way both explicit and implicit methods (using the inverse of a suitable operator)

and also multistep methods by introducing a new vector $w^n = (v^n, v^{n-1}, \dots, v^{n-(s-1)})$ and choosing a proper Banach space $\mathcal{B}$

• order and consistency of the method

- a certain method $\{S_k\}$ has the order of accuracy p if

$$\| u(t+k) - S_k u(t) \| = \sigma(k^{p+1}) \quad \text{for } k \to 0$$

and for all $t \in \langle 0, T \rangle$, where $u(t)$ is an arbitrary, sufficiently smooth solution of the initial value problem

$$(*) \qquad \frac{\partial u(t)}{\partial t} = A u(t), \quad u(0) = u_0 \quad \text{for } 0 \le t \le T$$

- $\{S_k\}$ is consistent if its order of accuracy $p \ge 1$, which means that the given scheme really approximates $(*)$

- convergence and stability

- the method described by $\{S_k\}$ is convergent

if $$\lim_{\substack{k\to 0 \\ nk=t}} \| S_k^n v(0) - v(t) \| = 0 \quad \text{for } t \in \langle 0, T \rangle$$

where $v(t)$ is the solution of $(*)$ for initial values $v(0)$ (arbitrary)

- in other words, if we decrease the time step k sufficiently and at the same time increase the number of steps to get to time t, we get better approximation and in the limit $k \to 0$ we get the solution of $(*)$

- the main difference between ODEs and PDEs is that for ODEs, the method was convergent if such a limit was true for an arbitrary ODEs

- the method is stable, if there exists $C > 0$ such that $\| S_k^n \| \leq C$ for all n and k satisfying $0 \leq nk \leq T$

or $$\| v^n \| = \| S_k^n v^0 \| \leq C \| v^0 \| \quad \text{for all initial vectors } v^0 \text{ and } 0 \leq nk \leq T$$

- even for the finite T we get $C \to \infty$ if $k \to 0$ and $n \to \infty$ and a method is then unstable!

- Lax equivalence theorem

- if $\{S_k\}$ is a consistent approximation of a well-conditioned linear initial value problem $(*)$ then $\{S_k\}$ is convergent if and only if $\{S_k\}$ is stable.

(for the proof see the literature above)

- notes: 1) definition of stability is independent of the particular system of PDEs and to determine if $\{S_k\}$ is stable is usually much simpler than to prove convergence, that is why Lax theorem is important

2) the result is general for parabolic, hyperbolic and other types of PDEs, only linearity is important and the problem must be well-conditioned

(for problems that are ill-conditioned, we can get very different solution by introducing even very small errors although the metod is stable)

3) as for ODEs, also here it is true that the order of accuracy p means that numerical solution $v(t)$ satisfies

$$\|v(t) - U(t)\| = \mathcal{O}(k^p) \quad \text{as } k \to 0$$

uniformly for all $t \in \langle 0, T \rangle$

(only locally the error is $\mathcal{O}(k^{p+1})$)

- intuitively we apply $N = \frac{T}{k}$ times a formula with error $\mathcal{O}(k^{p+1})$ and get the global error $\mathcal{O}(Nk^{p+1}) \approx \mathcal{O}(k^p)$

- in practice, stability is often shown by finding that $\|S_k\| \leq 1$ for a certain k and thus $\|S_k^n\| \leq 1 = C$

or very often it is used von Neumann analysis of stability (see later), even though it is not fully general,

that determines the stability using reciprocal space (Fourier image)

• Conditions of stability for finite differences

- if we have a certain method described by $\{S_k\}$,

i.e. $v^{n+1} = S_k v^n$ or $v^n = S_k^n v^0$ (***)

then it is stable if there exists $C > 0$ such that

$$\| S_k^n \| \leq C \quad \text{for all } n \text{ and } k,\ 0 \leq nk \leq T$$

or $$\| v^n \| \leq C \| v^0 \| \quad \text{for all } v^0 \text{ and } n$$

- we can estimate the error after n steps by subtracting

$$v^{n+1} = S_k v^n + k\tau^n \quad (\tau^n \sim O(k^{p+1}))$$

↑ local discretization error satisfying $\|\tau^n\| \to 0$ as $k \to 0$ if the method is consistent

from (***)

we get $$e^{n+1} = S_k e^n - k\tau^n$$

and after N steps

$$e^N = S_k^N e^0 - k \sum_{n=1}^{N} S_k^{N-n} \tau^{n-1}$$

thus we can estimate

$$\| e^N \| \leq \| S_k^N \| \| e^0 \| + k \sum_{n=1}^{N} \| S_k^{N-n} \| \| \tau^{n-1} \|$$

and for a stable method and $Nk \leq T$ finally

$$\| e^N \| \leq \underbrace{C \| e^0 \|}_{\text{initial error typically multiplied by } C=1} + \underbrace{TC \max \| \tau^{n-1} \|}_{\to 0 \text{ for } k \to 0}$$

but in general C can be arbitrary, $< \infty$

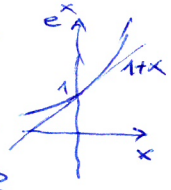
if $\| e^0 \| = 0$ we have convergence

- if $\| S_k \| \leq 1$ we talk about strong stability

but it is sufficient if $\boxed{\| S_k \| \leq 1 + \alpha k}$, where α is constant

because then $\| S_k^n \| \leq (1+\alpha k)^n \leq e^{\alpha T} = C$ where $nk = T$

↑ it follows from $1 + x \leq e^x$ for all $x \in \mathbb{R}$



or it is sufficient to be valid

$$\| v^{n+1} \| \leq (1 + \alpha k) \| v^n \|$$

• von Neumann analysis of stability

- instead of checking the condition (here we use the Euclid norm explicitly)

$$\|v^{n+1}\|_2 \le (1+\alpha k)\|v^n\|_2$$

we work with

$$\|\hat{v}^{n+1}\|_2 \le (1+\alpha k)\|\hat{v}^n\|_2$$

which is equivalent with the first condition thanks to Parseval relation $\|\hat{v}\|_2 = \|v\|_2\sqrt{2\pi}$

if v and $\hat{v}$ are Fourier images of each other defined as

$$v_j = \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \hat{v}(\xi)\, e^{ijh\xi}\, d\xi \quad \Longleftrightarrow \quad \hat{v}(\xi) = h\sum_{j=-\infty}^{\infty} v_j\, e^{-ijh\xi}$$

($x_j = jh$)

if v is ℓ_2 function i.e. $\|v\|_2 < \infty$

and norms are given by

$$\|v\|_2 = \sqrt{h\sum_{j=-\infty}^{\infty} |v_j|^2}\,, \qquad \|\hat{v}\|_2 = \sqrt{\int_{-\pi/h}^{\pi/h} |\hat{v}(\xi)|^2 d\xi}$$

- the analysis in the reciprocal space is simpler because for the Fourier image $\hat{v}(\xi)$ we get

$$\hat{v}^{n+1}(\xi) = \hat{a}(\xi)\,\hat{v}^n(\xi) \quad \text{for all } \xi$$

since (at least for explicit formulas, for implicit method situation will be more complicated by more or less similar)

$$v_j^{n+1} = (S_k v^n)_j = \sum_{\mu=-\infty}^{\infty} \alpha_\mu v_{j+\mu}^n = \underbrace{(a * v)_j}_{\text{convolution of } a \text{ and } v} = h\sum_{\mu=-\infty}^{\infty} a_{j-\mu} v_\mu^n$$

where $a_\mu = \frac{1}{h}\alpha_{-\mu}$

and thus

$$\widehat{v^{n+1}}(\xi) = \widehat{S_k v^n}(\xi) = \widehat{a * v^n}(\xi) = \hat{a}(\xi)\,\hat{v}^n(\xi)$$

as the Fourier image of a convolution is the product of the Fourier images of convolved functions

- if the amplification factor $\hat{a}(\xi)$ satisfies

$$\boxed{|\hat{a}(\xi)| \le 1+\alpha k}$$ for a certain α independent of ξ

then we have a stable method since then $|\hat{v}^{n+1}(\xi)| \le (1+\alpha k)|\hat{v}^n(\xi)|$ for all ξ

and by integration of squares $\|\hat{v}^{n+1}\|_2 \le (1+2\alpha k)\|\hat{v}^n\|_2$.

Example: Euler method for $v_t = v_{xx}$

we have $$v_j^{n+1} = v_j^n + \sigma(v_{j+1}^n - 2v_j^n + v_{j-1}^n) \quad \text{where } \sigma = \frac{k}{h^2}$$

and thus $$a = \frac{1}{h}(\ldots\ 0\ \ \sigma\ \ 1-2\sigma\ \ \sigma\ \ 0\ \ldots)$$
$$\qquad\qquad -2\ -1\ \ 0\ \ \ 1\ \ 2 \longrightarrow \mu$$

to have $$v_j^{n+1} = a_{-1} v_{j+1}^n + a_0 v_j^n + a_1 v_{j-1}^n = a * v^n$$

by Fourier transform we get

$$\hat{a}(\xi) = h \sum_{j=-\infty}^{\infty} e^{-ijh\xi} a_j = \underbrace{e^{ih\xi}}_{a_{-1}}\sigma + \underbrace{1(1-2\sigma)}_{a_0} + \underbrace{e^{-ih\xi}}_{a_1}\sigma =$$

$$= 1 - 2\sigma(1-\cos h\xi) = 1 - 4\sigma \sin^2\frac{h\xi}{2}, \quad \xi \in (-\tfrac{\pi}{h}, \tfrac{\pi}{h})$$

to get $|\hat{a}(\xi)| \le 1$ for all ξ

we must have $4\sigma \le 2$ and thus $\frac{k}{h^2} \le \frac{1}{2}$

-but we can get the same result in a simpler way

by substituting "plane wave" solution $e^{ijh\xi} = v_j^n$

directly into the finite-difference formula

to set $v_j^{n+1} = \hat{a}(\xi)\, e^{ijh\xi} = \hat{a}(\xi)\, v_j^n$

we get

$$v_j^{n+1} = \hat{a}(\xi) e^{ijh\xi} = e^{ijh\xi}\left[1 + \sigma\left(e^{ih\xi} - 2 + e^{-ih\xi}\right)\right]$$

or $\hat{a}(\xi) = 1 - 2\sigma(1-\cos h\xi)$ as above

- von Neumann analysis for implicit one-step method

- such a method can be, in general, written in the form

$$\sum_{\mu=-\ell}^{r} \beta_\mu v_{j+\mu}^{n+1} = \sum_{\mu=-\ell}^{r} \alpha_\mu v_{j+\mu}^{n} \qquad \text{with some } \beta_\mu \neq 0 \text{ for } \mu \neq 0 \text{ (to be implicit)}$$

- if we have an infinit grid we deal with an infinite set of linear equations $B v^{n+1} = A v^n$ where B is banded

it can be shown that if v^n and v^{n+1} are l_h^2-sequences $\left(h \sum_{j=-\infty}^{\infty} v_j^2 < \infty\right)$ than there exists a unique solution

if $\hat{b}(\xi) \neq 0$ for $\xi \in \langle -\frac{\pi}{h}, \frac{\pi}{h} \rangle$

where $\hat{b}(\xi)$ is the Fourier image of $b_\mu = \frac{1}{h} \beta_{-\mu}$

since then again the right-hand side is $\sum_{\mu=-\infty}^{\infty} \beta_\mu v_{j+\mu}^{n+1} = b * v^{n+1}$ - convolution

- under this assumption we can write

$$\widehat{b * v^{n+1}}(\xi) = \widehat{a * v^n}(\xi)$$

$$\hat{b}(\xi)\, \widehat{v^{n+1}}(\xi) = \hat{a}(\xi)\, \hat{v}^n$$

or $$\widehat{v^{n+1}}(\xi) = g(\xi)\, \widehat{v^n}(\xi) \qquad \text{with } g(\xi) = \frac{\hat{a}(\xi)}{\hat{b}(\xi)}$$

- $g(\xi)$ is again the amplification factor which is a continuous function on $\langle -\frac{\pi}{h}, \frac{\pi}{h} \rangle$ and thus it has a finite maximum $\|g\|_\infty = \max_\xi \left| \frac{\hat{a}(\xi)}{\hat{b}(\xi)} \right| < \infty$

- in space l_h^2, the vector is uniquely determined by its Fourier image $\Rightarrow$ uniqueness of solution

$v^{n+1} \in l_h^2$ and we can write $v^{n+1} = S_k v^n$

where S_k is a bounded linear operator

with $\|S_k^n\| = \|g^n\|_\infty = (\|g\|_\infty)^n$ for $n \geq 0$

or $\|v^n\| \leq (\|g\|_\infty)^n \|v^0\|$

- as one can use von Neumann analysis for explicit methods simply by assuming $v_j^n = g^n e^{i\xi jh}$

also here, for implicit methods, we can do the same

Example: 1) Crank-Nicolson method for $u_t = u_{xx}$

$$v_j^{n+1} - v_j^n = \frac{1}{2}\sigma\left(v_{j+1}^n - 2v_j^n + v_{j-1}^n\right) + \frac{1}{2}\sigma\left(v_{j+1}^{n+1} - 2v_j^{n+1} + v_{j-1}^{n+1}\right), \quad \sigma = \frac{k}{h^2}$$

using $v_j^n = g^n e^{i\xi jh}$ we get ($g^n e^{i\xi jh}$ already cancelled)

$$g - 1 = \frac{1}{2}\sigma\left(e^{i\xi h} - 2 + e^{-i\xi h}\right) + \frac{1}{2}\sigma\left(g e^{i\xi h} - 2g + g e^{-i\xi h}\right)$$

or $$g\left(1 + \sigma(1 - \cos\xi h)\right) = 1 - \sigma(1 - \cos\xi h)$$

and thus $$g(\xi) = \frac{1 - 2\sigma \sin^2\frac{\xi h}{2}}{1 + 2\sigma \sin^2\frac{\xi h}{2}}$$ ← always > 1 and thus $\neq 1$ as should be

we can see that $|g(\xi)| \leq 1$ for all k (h) and ξ

and thus the method is unconditionally stable

2) C-N method for $u_t = u_x$

in a similar way as above, we would get

$$g(\xi) = \frac{1 + \frac{ik}{2h}\sin\xi h}{1 - \frac{ik}{2h}\sin\xi h}$$

and $|g(\xi)| = 1$ for all $\frac{k}{h}$ and ξ

this follows from

$$\left|\frac{a+ib}{a-ib}\right| = \left|\frac{(a^2+b^2)^{1/2} e^{i\varphi}}{(a^2+b^2)^{1/2} e^{-i\varphi}}\right| = |e^{2i\varphi}| = 1$$

also stable

3) C-N method for $u_t = iu_{xx}$ (Schrödinger equation)

$$g(\xi) = \frac{1 - 2i\frac{k}{h^2}\sin^2\frac{\xi h}{2}}{1 + 2i\frac{k}{h^2}\sin^2\frac{\xi h}{2}} \Rightarrow |g(\xi)| = 1 \text{ for all } k, h, \xi$$

and again we have stability

- **von Neumann analysis for vector functions and multistep methods**

- because multistep methods can be rewritten as a one-step method for a vector function by introducing new unknown variables, we can analyze them in the same way as the latter methods

Example: leap-frog for $u_t = u_x$

$$v_j^{n+1} = v_j^{n-1} + \frac{k}{h}\left(v_{j+1}^n - v_{j-1}^n\right) \Leftrightarrow \begin{pmatrix} v_j^{n+1} \\ v_j^n \end{pmatrix} = \begin{pmatrix} -\lambda & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_{j-1}^n \\ v_{j-1}^{n-1} \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} v_j^n \\ v_j^{n-1} \end{pmatrix} + \begin{pmatrix} \lambda & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_{j+1}^n \\ v_{j+1}^{n-1} \end{pmatrix}$$

or $$w_j^{n+1} = \alpha_{-1} w_{j-1}^n + \alpha_0 w_j^n + \alpha_1 w_{j+1}^n \quad \text{where } w_j^n = \begin{pmatrix} v_j^n \\ v_j^{n-1} \end{pmatrix}$$

again we have convolution $w^{n+1} = a * w^n$ with $a_m = h^{-1}\alpha_{-m}$

by doing the Fourier transform, we get

$$\widehat{w^{n+1}}(\xi) = \hat{a}(\xi)\,\widehat{w^{n}}(\xi)$$

where $\hat{a}(\xi)$ is a 2×2 matrix with elements that are Fourier images of vectors consisting from all elements of a_{jn} at the same positions, e.g.

$$\hat{a}_{11}(\xi) = h\sum_{j=-\infty}^{\infty}(a_j)_{11}\,e^{-ijh\xi}$$

in the case of the leap-frog method for $v_t = v_x$ we get

$$\hat{a}(\xi) = \begin{pmatrix} 2i\lambda\sin\xi h & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} ix & 1 \\ 1 & 0 \end{pmatrix},\quad x = 2\lambda\sin\xi h$$

called ~~the~~ <u>amplification matrix</u>

- here the necessary condition for stability is

$$\varrho(\hat{a}(\xi)) \le 1 + O(k)$$

eigenvalues of the matrix $\hat{a}(\xi)$

where $\varrho(\hat{a}(\xi))$ is the spectral radius of $\hat{a}(\xi)$, i.e. $\max_j |\lambda_j|$

- eigenvalues are solutions of $\det(z\mathbb{1} - \hat{a}(\xi)) = z^2 - ixz - 1 = 0$

$$\Rightarrow\quad z_{1,2} = i\frac{x}{2} \pm \sqrt{1-\left(\frac{x}{2}\right)^2}$$

- both roots satisfy $|z| \le 1$ if and only if $\left|\frac{x}{2}\right| \le 1$

or $|\lambda\sin\xi h| \le 1 \Rightarrow \lambda = \frac{k}{h} \le 1$

- in general, even for implicit methods, we can write

$$\|v^n\| \le \left(\max_{\xi}\|G(\xi)\|_2\right)\|v^0\|\quad\text{, where}\quad G(\xi) = \hat{b}^{-1}(\xi)\,\hat{a}(\xi)$$

$\mathbb{1}$ for explicit methods

where $\hat{b}(\xi)$ is a Fourier image of matrices on the left-hand side of implicit methods

- for vector functions of M components (or M-step method) $G(\xi)$ is a $M\times M$ matrix and in order the method to be stable it must be true that

$$\|G(\xi)^n\| \le C \quad\text{for all } \xi\in\left\langle -\tfrac{\pi}{h}, \tfrac{\pi}{h}\right\rangle$$

and n,k satisfying $0 \le nk \le T$

where C is a constant $< \infty$ independent of ξ and $\|\cdot\|$ is a suitable matrix norm

- because in general

$$\rho(G(\xi))^n \le \|G(\xi)^n\| \le \|G(\xi)\|^n$$

where $\rho(A)$ is the spectral radius of A, i.e. $\max_j |\lambda_j(A)|$ (eigenvalues)

we can see that

(1) $\rho(G_k(\xi)) \le 1 + \mathcal{O}(k)$ as $k \to 0$ for all ξ uniformly — is the necessary condition of stability

(2) $\|G_k(\xi)\| \le 1 + \mathcal{O}(k)$ as $k \to 0$ for all ξ uniformly — is the sufficient condition of stability

- condition (1) is called von Neumann condition

- note that (2) is only sufficient, but not necessary

which means that a method can be stable even if this condition is not satisfied

(it can be shown that for the leap-frog method for $u_t = u_x$ we get $\|G(\xi)\| > 1$ for most ξ even though $\rho(G(\xi)) \le 1$ for $\lambda \le 1$)

- on the other hand, the condition (1) is only necessary

meaning that it must be valid for a method to be stable but it does not guarantee that a method is stable

- the necessary and sufficient condition is

$$\rho_\varepsilon(G_k(\xi)) \le 1 + \mathcal{O}(\varepsilon) + \mathcal{O}(k) \quad \text{for all } \xi \in \left\langle -\tfrac{\pi}{h}, \tfrac{\pi}{h} \right\rangle$$

as $k \to 0$ and $\varepsilon \to 0$

where $\rho_\varepsilon(A) = \sup_{\lambda_\varepsilon \in \Lambda_\varepsilon(A)} |\lambda_\varepsilon|$ is called ε-pseudospectral radius

and $\Lambda_\varepsilon(A)$ is ε-pseudospectrum of A,

i.e. eigenvalues of a matrix $A+E$ for a certain $E \in \mathbb{C}^{M \times M}$ for which $\|E\| \le \varepsilon$

(for details see Trefethen, chap. 4.5)

Example: von Neumann analysis of the leap-frog method of order 4 in space for $v_t = v_x$

- it is a 2-step method given by (explicit, 5-point approx. of $\frac{\partial}{\partial x}$)

$$v_j^{n+1} = v_j^{n-1} + \frac{4}{3}\lambda\left(v_{j+1}^n - v_{j-1}^n\right) - \frac{1}{6}\lambda\left(v_{j+2}^n - v_{j-2}^n\right), \lambda = \frac{k}{h}$$

- we introduce $w_j^n = \begin{pmatrix} v_j^n \\ v_j^{n-1} \end{pmatrix}$ and rewrite it as (with the second equation $v_j^n = v_j^n$)

$$w_j^{n+1} = \begin{pmatrix} v_j^{n+1} \\ v_j^n \end{pmatrix} = \begin{pmatrix} \frac{\lambda}{6} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_{j-2}^n \\ v_{j-2}^{n-1} \end{pmatrix} + \begin{pmatrix} -\frac{4}{3}\lambda & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_{j-1}^n \\ v_{j-1}^{n-1} \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} v_j^n \\ v_j^{n-1} \end{pmatrix} + \begin{pmatrix} \frac{4}{3}\lambda & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_{j+1}^n \\ v_{j+1}^{n-1} \end{pmatrix} + \begin{pmatrix} -\frac{\lambda}{6} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_{j+2}^n \\ v_{j+2}^{n-1} \end{pmatrix}$$

$$\alpha_{-2} w_{j-2}^n + \alpha_{-1} w_{j-1}^n + \alpha_0 w_j^n + \alpha_1 w_{j+1}^n + \alpha_2 w_{j+2}^n$$

- again to write it as a convolution $v_j^{n+1} = a * w^n$, we set $a_\mu = \frac{1}{h}\alpha_{-\mu}$

- by doing Fourier transform we get

$$\widehat{w^{n+1}}(\xi) = \hat{a}(\xi)\,\widehat{w^n}(\xi)$$

where $\hat{a}(\xi)$ is again a 2×2 matrix with elements being Fourier transforms of sequences of corresponding matrix elements of a, the result is the amplification matrix

$$\hat{a}(\xi) = \begin{pmatrix} \frac{\lambda}{6}\left(e^{2i\xi h} - e^{-2i\xi h}\right) - \frac{4}{3}\lambda\left(e^{i\xi h} - e^{-i\xi h}\right) & 1 \\ 1 & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} i\left(\frac{\lambda}{3}\sin 2\xi h - \frac{8\lambda}{3}\sin\xi h\right) & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} ix & 1 \\ 1 & 0 \end{pmatrix}$$

- eigenvalues are (as for the leap-frog of order 2)

$$z = i\frac{x}{2} \pm \sqrt{1 - \left(\frac{x}{2}\right)^2} \quad \text{with} \quad x = \frac{\lambda}{3}\sin 2\xi h - \frac{8\lambda}{3}\sin\xi h$$

- if $\left|\frac{x}{2}\right| \le 1$ we get also $|z| \le 1$ for all ξ, otherwise there will be a root $|z| > 1$

or we have the condition

$$\left|\lambda\left(\underbrace{\tfrac{1}{6}\sin 2\xi h - \tfrac{4}{3}\sin\xi h}_{f(\xi h)}\right)\right| \le 1$$

$f(\xi h)$ has a maximum when

$$f'(\xi h) = \frac{1}{3}\left(\cos 2\xi h - 4\cos\xi h\right) = 0$$

or $\quad 2\cos^2\xi h - 4\cos\xi h - 1 = 0$

$$\cos\xi h = 1 \pm \sqrt{\frac{3}{2}}$$

finally we get

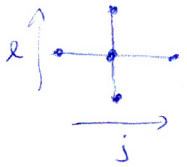
$$\lambda \le \left[\left(1 + \frac{\sqrt{6}}{6}\right)\sqrt{\sqrt{6} - \frac{3}{2}}\right]^{-1} \approx 0.72875\ldots$$

necessary condition, but also sufficient

• example of von Neumann analysis in more dimensions

- let us consider Euler method for the heat equation $u_t = u_{xx} + u_{yy}$

i.e. $$v_{je}^{n+1} = v_{je}^{n} + \sigma\left(v_{j+1,e}^{n} + v_{j,e+1}^{n} - 4v_{je}^{n} + v_{j-1,e}^{n} + v_{j,e-1}^{n}\right), \quad \sigma = \frac{k}{h^2}$$



(we use the same space step in x and y)

- conditions on stability can be obtained again using Fourier transforms of 2D sequences or directly by von Neumann analysis, in which we use $$v_{je}^{n} = \hat{a}(\xi_x, \xi_y)^n e^{i(\xi_x jh + \xi_y eh)}$$

after canceling v_{je}^{n} we set

$$\hat{a}(\xi_x, \xi_y) = 1 + \sigma\left(e^{i\xi_x h} + e^{i\xi_y h} - 4 + e^{-i\xi_x h} + e^{-i\xi_y h}\right) =$$

$$= 1 - 2\sigma\left(1 - \cos\xi_x h + 1 - \cos\xi_y h\right) =$$

$$= 1 - 4\sigma\underbrace{\left(\sin^2\frac{\xi_x h}{2} + \sin^2\frac{\xi_y h}{2}\right)}_{\in\langle 0,2\rangle}$$

and thus to be $|\hat{a}(\xi_x, \xi_y)| \le 1$ for all ξ_x and ξ_y

we must have $4\sigma \le 1$ (in 1D we had $4\sigma \le 2$)

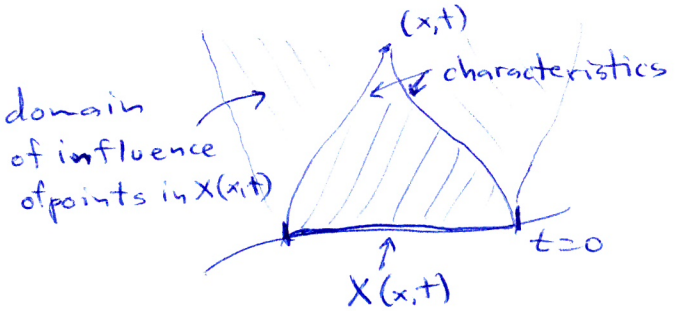
or $\sigma = \frac{k}{h^2} \le \frac{1}{4}$ (in 1D it was $\frac{1}{2}$)

- we see that more dimensions typically mean more strict conditions on step size

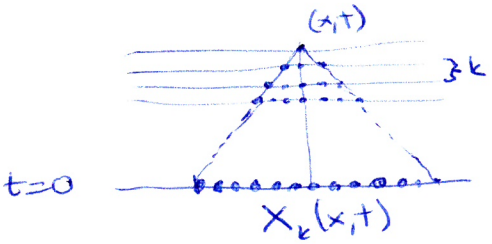
• Courant-Friedrichs-Lewy condition (CFL condition)

- played an important role historically as condition of stability for solution of (hyperbolic) PDEs
- it basically says that time step cannot be larger than time needed for the solution (wave etc) to propagate across the space step, but exact condition depends on PDE and chosen method

- general formulation is based on the domain of dependence
  - mathematical domain of dependence $X(x,t)$ of a solution $u(x,t)$ of a given PDE at a point (x,t) is a region which has an influence on the solution at (x,t) and it depends on the speed propagation of signal



  - numerical domain of dependence $X_k(x,t)$ for a certain time step k
    - initial points (at $t=0$) which have an influence on the numerical solution at (x,t)

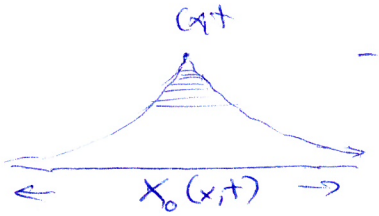


    - for explicit methods, it is a finite number of points for a finite k
    - for implicit methods, it comprises of all grid points
  - for stability, it is important the limit as $k \to 0$ or the limiting numerical domain of dependence $X_0(x,t)$
    - it can be finite (explicit methods for hyperbolic problems) or infinite for implicit methods or explicit methods for example for the heat equation

- let us consider Euler method for $u_t = u_{xx}$

- for each finite k, $X_k(x,t)$ is also finite

- but as $k \to 0$ we keep $r = \frac{k}{h^2}$ constant $\leq \frac{1}{2}$



and thus h can decrease more slowly, typically $h \sim \sqrt{k}$

and $X_0(x,t)$ is infinite

CFL condition says that it must be

$$X(x,t) \subseteq X_0(x,t) \quad \text{for all } (x,t)$$

for a method to be stable

it is necessary condition, not sufficient