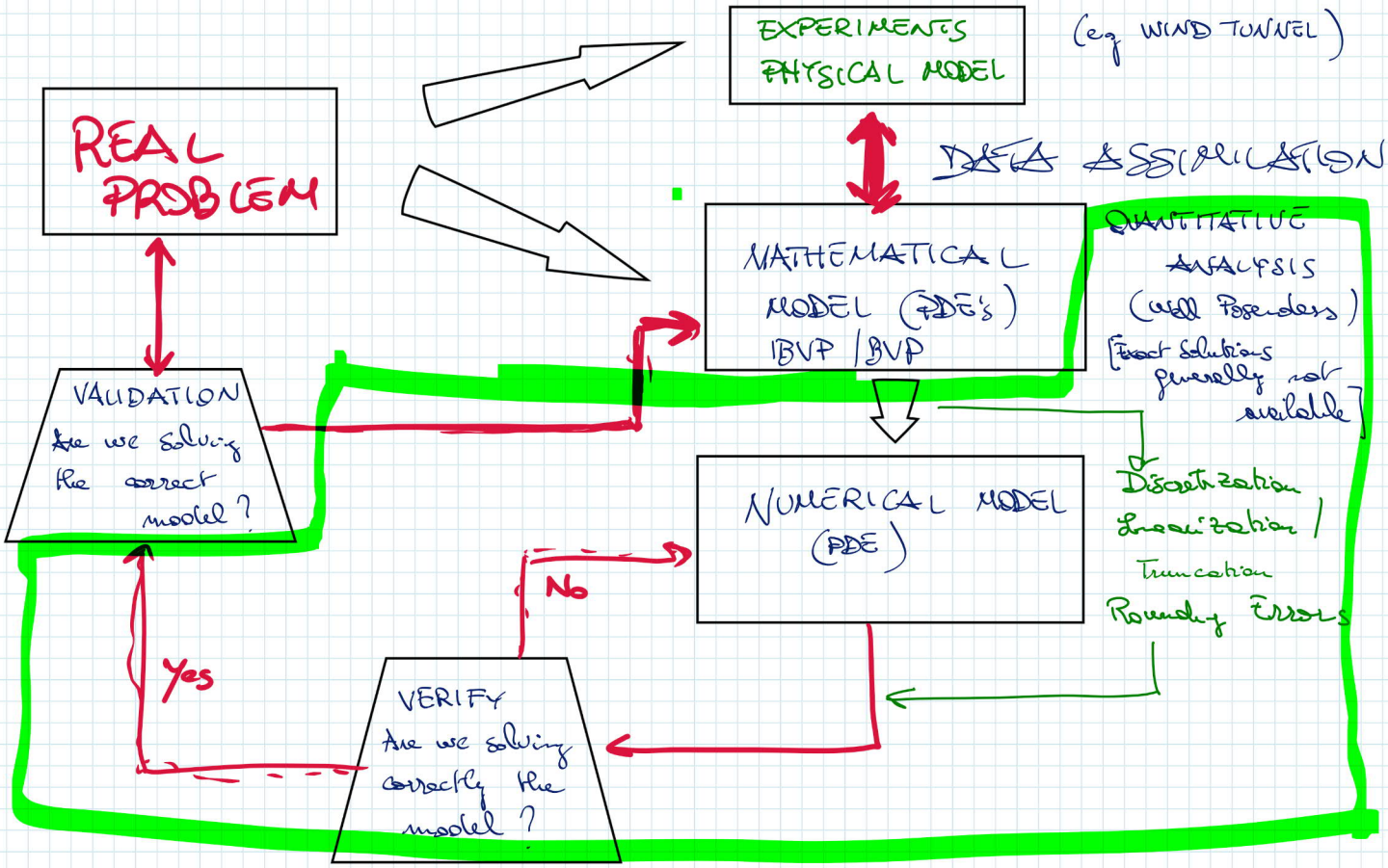
Finite Difference discretization of a simple Cauchy Problem (ODE): analysis, implementation, verification

REFERENCE : Langtangen - Linge, Chapter 1 + Appendix B

INTRODUCTION

In this first week, we consider some differential problem numerically approximated by the finite difference method. However, before we consider an example, we introduce some basic concept useful for the entire course. This will help to understand the features we want to investigate in a numerical method for solving a partial differential equation (and beyond).
Let's start with a practical approach leading to the numerical approximation of a Boundary Value Problem (BVP) or an Initial Boundary Value Problem (IBVP).

Modeling a Real problem with Numerical PDE's



In this course we are concerned with the region highlighted in green.

Source of Errors

Discretization

With "discretization" we mean the approximation of the differential components (derivatives) of the problem with algebraic operations. This step is necessary to manage the problem with a finite machine like a computer, yet it introduces some errors.
To give a simple example:

$$u'(\bar{x}) = \underbrace{\left[\frac{du}{dx}(\bar{x})\right]}_{\text{Problem}} \approx \underbrace{\frac{u(\bar{x}+h) - u(\bar{x}-h)}{2h}}_{\text{Numerical Problem}} = \left(u'(\bar{x})\right)_h$$

The derivative $\left(\lim_{\Delta x \to 0} \frac{u(\bar{x}+\Delta x) - u(\bar{x})}{\Delta x}\right)$ is approximated by an algebraic operation $(+, -, :, \times)$

Now, reasonable questions are (h = discretization parameter)

⇒ what is the error $\left| u'(\bar{x}) - (u'(\bar{x}))_h \right|$?

More precisely:

- $\lim_{h \to 0} (u'(\bar{x}))_h = u'(\bar{x})$?

  CONVERGENCE

  This is the fundamental feature: WE NEED CONVERGENCE to trust a method

- if yes: $u'(\bar{x}) - (u'(\bar{x}))_h = O(h^p)$

  what is p

  CONVERGENCE ORDER

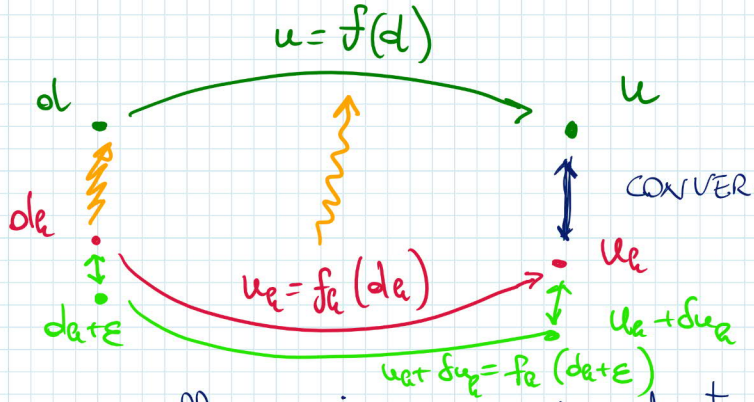
  This helps ranking different convergent methods: higher p and faster is the method to converge

The analysis of the discretization errors $\left| u'(\bar{x}) - (u'(\bar{x}))_h \right|$ is one of the specific topics of this course for the different methods proposed.

However, let's generalize the ideas:

Mathematical Problem: $P(u; d) = 0$ [u = unknown, d = data, parameters]

Numerical (Discretized) Problem: $P_h(u_h; d_h) = 0$ [u_h = numerical solution, d_h = approximation of data]



CONVERGENCE: $u_h \xrightarrow[h \to 0]{(?)} u$

The convergence generally requires some ingredients

CONSISTENCY: for $h \to 0$ $(d_h \to d)$

then $f_h \to f$

STABILITY: The numerical problem is "stable" for perturbations:

there exists $C > 0$ s.t.

$$\|\delta u_h\| \leq C \|\varepsilon\|$$

with $\|\cdot\|$ proper norms. C in general may depend on h.

(but it is bad if $C \to +\infty$ when $h \to 0$)

For linear problems and linear methods (Lax-Richtmyer Theorem)

CONVERGENCE = STABILITY + CONSISTENCE

For nonlinear problems, consistency and stability are only necessary, not sufficient (COMPACTNESS needed).

In the example above we have that

$$\left(u'(\bar{x})\right)_h = \lim_{h\to 0} \frac{u(\bar{x}+h)-u(\bar{x}-h)}{2h} = u'(\bar{x}) \quad \text{by definition (convergence)}$$

and with a Taylor expansion we notice that

$$\left|u'(\bar{x}) - \left(u'(\bar{x})\right)_h\right| = O(h^2).$$

(2^{nd} order method).

CONSISTENCY ERROR

With "consistency error" we mean specifically the residual (= quantity $\neq 0$) we obtain when we enforce the exact solution of a problem to the numerical scheme.

For instance, for $u = e^{\lambda x}$, $u'(\bar{x}) = \lambda e^{\lambda \bar{x}}$ and:

$$(u)'_h(\bar{x}) = \frac{u(\bar{x}+h)-u(\bar{x}-h)}{2h} = \frac{e^{\lambda(\bar{x}+h)} - e^{\lambda(\bar{x}-h)}}{2h} = \lambda e^{\lambda\bar{x}} + \boxed{O(h^2)}$$

Consistency error

Notice that in this case consistency and convergence coincide, because the solution of the problem (the derivative) is the problem itself.

However, notice that this analysis does not include the rounding errors.

We will follow-up on this.

For the moment, let's assume that $u(x)$ can be differentiated and that we perturb $u(x)$ with a differentiable function δ s.t. both $\|\delta\|_\infty$ and $\|\delta'\|_\infty$ are bounded so that $\varepsilon = \max\left(\|\delta\|_\infty, \|\delta'\|_\infty\right)$.

Then $$(u+\delta)'_h(\bar{x}) = \underbrace{\frac{u(\bar{x}+h)-u(\bar{x}-h)}{2h}}_{(u)'_h(\bar{x})} + \frac{\delta(\bar{x}+h)-\delta(\bar{x}-h)}{2h}$$

So we have:

$$\left|(u+\delta)'_h(\bar{x}) - (u)'_h(\bar{x})\right| \le \frac{\varepsilon}{h} \quad \text{for } h>0 \qquad (1)$$

(this estimate doesn't require the differentiability of δ)

If we let $h \to 0$ we have

$$\left|(u+\delta)'_h(\bar{x}) - (u)'_h(\bar{x})\right| \le \left|\delta'(\bar{x})\right| \le \varepsilon$$

(this estimate requires the differentiability of δ)

So, if the perturbation is (reasonably) regular, we have stability (same regularity of the solution)

IN THIS COURSE WE MAINLY INVESTIGATE DISCRETIZATION ERRORS

Linearization / Truncation

As we will see, the final numerical solution may be affected by other errors.

For instance, in most of the cases, we need to solve – after the discretization, a linear system. The numerical solution of this system introduces some errors. In the vast majority of practical problems, the numerical methods for linear systems are ITERATIVE (GMRES, Conjugate Gradient, etc.), so this error is under control with the convergence test (see MATH 517).

Also, when we have a non-linear differential problem, we need to "linearize" it, e.g. with a Newton approach.

We have, generally, two options: we DISCRETIZE first and then LINEARIZE (D-L) or the other way around (L-D). The two approaches are not equivalent, D-L is generally preferred. The linearization process introduces other errors, because we need to solve non-linear algebraic equations with some method investigated in the courses of numerical analysis. These errors are not specifically analyzed here, yet they are important in some cases. Since numerical methods for nonlinear problems are iterative, also in this case we can control the error with the stopping criterion.

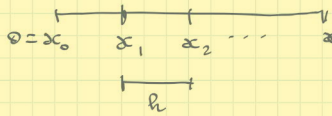
EXAMPLE:

$$-\frac{\partial^2 u}{\partial x^2} + u\frac{\partial u}{\partial x} + u^3 = f \qquad x \in (0,1)$$

$$u(0) = u(1) = 0$$

A possible DISCRETIZE-then-LINEARIZE approach:

GRID:



Definition: $u_h(x_j) = u_j$

$$\begin{cases} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + u_j \dfrac{u_{j+1} - u_{j-1}}{2h} + u_j^3 = f_j \\ u_0 = 0 \qquad u_N = 0 \end{cases}$$

This is a nonlinear algebraic system that we can write as:

$\underline{F}(\underline{u}) = \underline{0}$ with $\underline{u} = [u_0, u_1, \dots u_N]^T$

Newton method:

Denote by $\bar{J}$ the Jacobian matrix $\left[\bar{J}_{ij} = \dfrac{\partial F_i}{\partial u_j}\right]$

Then we start with a guess $\underline{u}^{(0)}$ and we start the loop:

```
k = 0
while (Not convergence)
    Compute J(u^(k))
    Solve the linear system:
        J(u^(k)) δ = - F(u^(k))
    Update: u^(k+1) = u^(k) + δ
    k++
    Convergence = Test_Convergence(u^(k+1), u^(k), F(u^(k+1)))
end_while
```

[There is another possible scheme:

$$u\frac{\partial u}{\partial x} = \frac{1}{2}\frac{\partial (u^2)}{\partial x} \to \frac{1}{2}\frac{u_{j+1}^2 - u_{j-1}^2}{2h}$$

This approach is called CONSERVATIVE]

In the Linearize-then-Discretize approach, we need to differentiate the equation with respect to a function (Gateaux Derivative). Although possible, we do not follow this approach here.

Rounding Errors

All the numbers represented on a computer are subject to numerical approximation. This approximation depends on the hardware (32 bit vs 64 bit for instance) and it is quantified by the MACHINE-ε, i.e. the smallest number ε such that $1 + \varepsilon > 1$ for the computer. Generally, $\varepsilon \sim 10^{-16}$.

We do not consider the presence of rounding errors when we do our convergence analysis, but we need to keep in mind that rounding errors are everywhere!

EXAMPLE:

$$e^{\bar{x}} = \left.\frac{de^x}{dx}\right|_{x=\bar{x}} \approx \frac{e^{\bar{x}+h} - e^{\bar{x}}}{h}$$

Theory: for $h \to 0$, $\lim_{h\to 0} \dfrac{e^{\bar{x}+h} - e^{\bar{x}}}{h} = e^{\bar{x}}$

Practice: $fl\left(e^{\bar{x}+h}\right)$ (= floating point approximation of $e^{\bar{x}+h}$) $= e^{\bar{x}+h}(1+\varepsilon_1)$

$$fl\left(e^{\bar{x}}\right) = e^{\bar{x}}(1+\varepsilon_2)$$

with $|\varepsilon_1|, |\varepsilon_2| \le \varepsilon$

So:

$$\frac{fl\left(e^{\bar{x}+h}\right) - fl\left(e^{\bar{x}}\right)}{h} = \frac{e^{\bar{x}+h} - e^{\bar{x}}}{h} + \frac{e^{\bar{x}}\left(e^{h}\varepsilon_1 - \varepsilon_2\right)}{h}$$

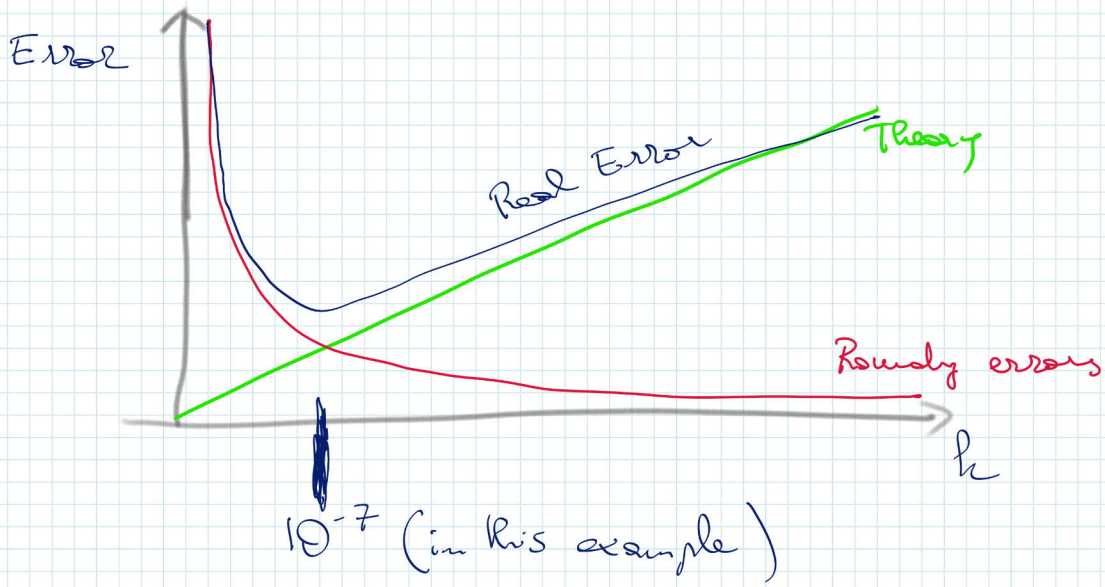
$\frac{e^{\bar{x}+h} - e^{\bar{x}}}{h}$ for $h \to 0$ this $\to e^{\bar{x}}$

$\left|\frac{e^{\bar{x}}\left(e^{h}\varepsilon_1 - \varepsilon_2\right)}{h}\right| < \frac{e^{\bar{x}} 2\varepsilon}{h} \xrightarrow[h\to 0]{} \infty$

Keep in mind this aspect in your numerical simulations.
TRY: implement this simple example

Python:
```
import numpy as np
x = 2
h = 0.1
np.abs( np.exp(x) - ((np.exp(x+h) - np.exp(x))/h) )
```

Test the error for decreasing h: for $h \geq 10^{-7}$ everything goes as expected (linear reduction of the error), for $h < 10^{-7}$ you see the error going up



Question: So, why we do not consider rounding errors in our analysis?

Answer: in general, practical problems are OK with errors greater than the range of h where rounding errors show up ($h >> 10^{-7}$).

FINITE DIFFERENCES

The first method we consider is the simplest one: simplicity means that we do not need much background, but also that there are inherent limitations (we will go over when studying other methods). The basic idea is to replace the derivatives with incremental quotients in specific points selected with some criterion. We start with a simple INITIAL VALUE problem, so to avoid some technicalities, yet to get familiar with some concepts and analysis tools.

LESSONS FROM A SECOND ORDER CAUCHY PROBLEM

We start considering the numerical solution of

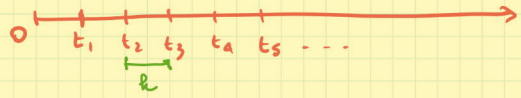
$$\frac{d^2u}{dt^2} + \omega^2 u = 0 \qquad t > 0$$
$$u(0) = I \qquad \frac{du}{dt}(0) = 0$$

Clearly, we have the analytical solution $u(t) = I\cos(\omega t)$. This will help us understand how accurate are the methods we consider.

To perform a numerical approximation, let's "discretize" the differential equation, replacing the derivative with algebraic operations:

$$P(u) = 0: \quad \frac{d^2u}{dx^2} + \omega^2 u = 0 \; ; \quad u(0) = I \; , \quad u'(0) = 0.$$

Numerical Approximation



$\Delta t = h$ = user defined

$u(t_j) = u_j$

Let's collocate the problem in the points $t_j = jh$ and replace the derivatives with the incremental quotients:

$$(2) \qquad \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \omega^2 u_j = 0$$

$u_0 = I$ For $u'(0)$ we need an approximation of the first derivative in 0

Let's start with the option in the book: $u'(0) \approx \frac{u_1 - u_{-1}}{2h}$ → This is clearly a numerical trick, because this point doesn't exist. However, it can work.

So, we can write sequentially:

I.C.
$$u_0 = I$$
$$\frac{u_1 - u_{-1}}{2h} = u'(0) = 0 \Rightarrow u_{-1} = u_1$$

First Step
$$\frac{u_1 - 2u_0 + u_{-1}}{h^2} + \omega^2 u_0 = 0 \Rightarrow \frac{2u_1 - 2I}{h^2} + \omega^2 I = 0$$
$$u_1 = \left(1 - \frac{\omega^2 h^2}{2}\right) I$$

Subsequent steps
$$u_{j+1} = \left(2 - \omega^2 h^2\right) u_j - u_{j-1} \qquad \boxed{j \geq 2}$$

The method is clear and pretty immediate: how does it perform?

SEE THE PYTHON CODE ex1.py

If we use the code and we measure the error as:

$$e_\infty \overset{def}{=} \max_j |u_e(t_j) - u_j|$$

we find that $e_\infty \sim O(h^2)$.

```
def solver(I, w, dt, T):
    """
    Solve u'' + w**2*u = 0 for t in (0,T], u(0)=I and u'(0)=0,
    by a central finite difference method with time step dt.
    """
    dt = float(dt)
    Nt = int(round(T/dt))
    u = np.zeros(Nt+1)
    t = np.linspace(0, Nt*dt, Nt+1)

    u[0] = I
    u[1] = u[0] - 0.5*dt**2*w**2*u[0]

    for n in range(1, Nt):
        u[n+1] = 2*u[n] - u[n-1] - dt**2*w**2*u[n]
    return u, t
```

i.c.

core of the method

CONVERGENCE ANALYSIS

Let's try to explain our numerical results.

If we use a Taylor expansion, we have

$$\begin{cases} u_{j+1} = u_j + u'_j h + u''_j \frac{h^2}{2} + u'''_j \frac{h^3}{3!} + u^{iv}_j \frac{h^4}{4!} + \text{high order terms} \\ u_{j-1} = u_j - u'_j h + u''_j \frac{h^2}{2} - u'''_j \frac{h^3}{3!} + u^{iv}_j \frac{h^4}{4!} + H.O.T. \end{cases}$$

If we sum these two, we have:

$$u''_j = \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + u^{iv} \frac{h^2}{12} + H.O.T.$$

So, we easily define the LOCAL TRUNCATION ERROR:

EXACT PROBLEM: $u''_e(x_j) + w^2 u_e(x_j) = \frac{u_{ex_{j+1}} - 2u_{ex_j} + u_{ex_{j-1}}}{h^2} + w^2 u_{ex_j} = -\frac{u^{iv}_e h^2}{12}$

NUMERICAL PROBLEM: $\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + w^2 u_j = 0$

The Local Truncation Error $\xrightarrow[h \to 0]{} 0$. This states that the method is CONSISTENT with the original problem.

Unfortunately this is necessary but not sufficient for the convergence.

In this simple problem, we can work out the convergence directly.

Set $e_j = u_{ex_j} - u_j$. Our goal is to prove that the $|e_j| \xrightarrow[h \to 0]{} 0$.

By subtracting the exact problem and the numerical one and setting the LTE: $\frac{u_e(t_j)}{12} h^2 = \gamma_j h^2$ we have:

$$\begin{cases} e_{j+1} - (2 - w^2 h^2) e_j + e_{j-1} = \gamma_j h^4 \\ e_0 = 0 \qquad e_1 = D h^2 \Rightarrow \boxed{\frac{u_1 - u_{-1}}{2h} + D h^2 = u'(0)} \end{cases}$$

By Taylor expansion we can move this

Let's solve this 2nd order difference equation. Before, a Remark

REMARK: On the Taylor Expansion (TE)

When we do TE, we postulate "regularity", i.e. that all the derivatives we need exist. In this case, this is true because $u_e = I \cos(wt)$ is infinitely regular. However, when we don't know the solution, LACK OF REGULARITY may undermine our analysis. This may be the root of discrepancies between the expectations (Theory) and the results (practice).

If we take $C_j \simeq C$, we have an error equation in the form: (C can be $C = \max_j C_j$)

$$e_{j+1} - (2 - \omega^2 h^2) e_j + e_{j-1} = C h^4$$

(*) Difference equation of order 2: please check how to solve it.

A particular solution to this difference equation is for e = constant:

$$e = \frac{C h^4}{1 - 2 + \omega^2 h^2 + 1} \simeq \frac{C}{\omega^2} h^2$$

Then, we add errors from the homogenous part: to do this, we need to solve:

$$\rho^2 - (2 - \omega^2 h^2)\rho + 1 = 0 \Rightarrow \rho_{1,2} = \frac{2 - \omega^2 h^2 \pm \sqrt{4 + \omega^4 h^4 - 4\omega^2 h^2 - 4}}{2} = \frac{2 - \omega^2 h^2 \pm \omega h \sqrt{\omega^2 h^2 - 4}}{2}$$

Then, the solution will be:

(a) for $\rho_1 \neq \rho_2$: $e_j = A\rho_1^j + B\rho_2^j + \frac{C}{\omega^2} h^2$ where A and B depend on the initial conditions.

check with a Taylor expansion

In our case, $e_0 = 0$ $e_1 = Dh^2$ (D is another constant). In fact $\begin{cases} u_0'(0) = 0 = \frac{u_1 - u_{-1}}{2h} + Dh^2 \\ u'(0) = \frac{u_1 - u_{-1}}{2h} \end{cases}$

(b) for $\rho_1 = \rho_2 = \rho$ $e_j = (A + tB)\rho^j + \frac{C}{\omega^2} h^2$

Let's start considering the case:

(1) $\rho_1 \neq \rho_2$ real: this occurs for $\omega^2 h^2 - 4 > 0 \Rightarrow h > \frac{2}{\omega}$ (a time step larger than the inverse of half of the frequency)

$$\rho_{1,2} = \frac{2 - \omega^2 h^2 \pm \omega h \sqrt{\omega^2 h^2 - 4}}{2}$$

$\Rightarrow$ Notice that $\rho_1 \rho_2 = 1$, so $\rho_1 = \frac{1}{\rho_2}$. If $|\rho_2| < 1$, then $|\rho_1| > 1$ (or the other way around)

So either ρ_1 or ρ_2 is $|\rho_j| > 1$

This means $|e_j| \xrightarrow[j \to +\infty]{} +\infty$

We do not have convergence because of the error accumulation

(2) $\rho_1 = \rho_2$: $h = \frac{2}{\omega} \Rightarrow \rho = \frac{2 - \omega^2 h^2}{2} = -1$

$$e_j = (A + jhB)(-1)^j + \frac{Ch^2}{\omega^2} \Rightarrow e_0 = A + C\frac{4}{\omega^4} = 0 \rightarrow A = -\frac{C4}{\omega^4}$$

$$\Rightarrow e_1 = h\frac{Ch^2}{\omega^2} - hB = Dh^2$$

if $B = 0$, we expect an oscillatory behaviour because of the $(-1)^j$ term.

if $B \neq 0$, we may expect an explosion

So, in general $h = \frac{2}{\omega}$ can be unstable

(3) $\rho_1 = \rho_2$ complex conjugate: $h < \frac{2}{\omega}$

$$\rho_{1,2} = \frac{2 - \omega^2 h^2 \pm i\omega h \sqrt{4 - h^2\omega^2}}{2}$$

Notice that: $|\rho_i| = \frac{1}{2}\sqrt{4 + \omega^4 h^4 - 4\omega^2 h^2 + 4\omega^2 h^2 - \omega^4 h^4} = 1$

So we can conclude that:

$$|e_j| \leq |A||\rho_1|^j + |B||\rho_2|^j + \left|\frac{Ch^2}{\omega^2}\right| \leq A + B + \frac{Ch^2}{\omega^2}$$

$$0 = e_0 = A + B + \frac{Ch^2}{\omega^2} \qquad A + B = -\frac{Ch^2}{\omega^2}$$

$$Dh^2 = e_1 = A\rho_1 + B\rho_2 + \frac{Ch^2}{\omega^2} \qquad A(\rho_1 - 1) + B(\rho_2 - 1) = Dh^2$$

$$A(\rho_1 - \rho_2) = \left(D + \frac{(\rho_2 - 1)C}{\omega^2}\right)h^2 \qquad B(\rho_2 - \rho_1) = \left(D - \frac{(\rho_1 - 1)C}{\omega^2}\right)h^2$$

$\Rightarrow |A|, |B| \sim O(h^2)$

We conclude that : for $h < \frac{2}{\omega}$ the method is consistent and stable, so it converges and, in particular, with order 2.

→ CONDITIONALLY STABLE METHOD

REMARK

Notice that the convergence rate is dictated by the consistency error for the equation, but also for the initial data.

EXERCISE 1:

Approximate the initial derivative with no ghost point : $u'(0) \approx \frac{u_1 - u_0}{h} = 0 \implies u_1 = u_0$

Test and prove that the convergence rate becomes 1 in this case.

EXERCISE 2 :

If we want a second order formula, with no ghost point, we can use the formula :

$$u'(0) \approx \frac{-3u_0 - 4u_1 + u_2}{2h}$$

(1) Prove that this is a 2nd order approximation of $u'(0)$

(2) Test the method with this formula

(3) Prove that the method converges with order 2.

ANOTHER (MORE SPECIFIC) CONVERGENCE ANALYSIS

The solution to this problem is PERIODIC. When we think to the Fourier decomposition, we know that we can do a spectral or frequency analysis.

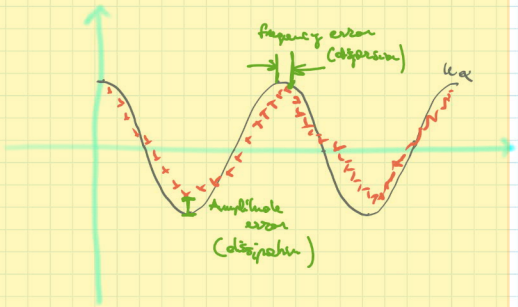
$$f \text{ periodic or on a bounded interval} = \sum_{j=0}^{\infty} \hat{f}_j e^{ijwt}$$

In our particular case, the solution is a co-sinusoidal function $u_e = I \cos(wt)$.

In this case, we may be interested specifically to the errors on the amplitude I and the frequency w.
In other terms assuming that the numerical solution reads like : $u_j = I_e \cos(w_e h j)$
where we have a numerical amplitude I_e and a numerical frequency w_e, what are the :

(1) AMPLITUDE ERROR $|I - I_e|$

(2) FREQUENCY ERROR $|w - w_e|$ or $\frac{w}{w_e}$?



Amplitude errors go under the name of DISSIPATION ERRORS

Frequency errors go under the name of DISPERSION ERRORS

We can categorize the error according to these two components (Von Neumann analysis).

REMARK : The following analysis is equivalent but different of the book. It is more general

Let's compute directly the numerical solution and see if we can identify AMPLITUDE I_e and FREQUENCY w_e.
Based on this, we can compute specifically the different components of the errors.

The numerical solution solves :

$$u_{j+1} - (2 - w^2h^2)u_j + u_{j-1} = 0$$

(I denote w_e as $\tilde{w}$ for easiness of notation)

Based on what we learned with the error equation, for $h < \frac{2}{\omega}$ we have :

$$u_j = A\rho_1^j + B\rho_2^j \qquad \rho_{1,2} = \frac{2 - w^2h^2}{2} \pm i\frac{wh\sqrt{4 - w^2h^2}}{2}$$

Notice that : $|\rho| = \frac{4 + w^2h^4 - 4w^2h^2 + 4w^2h^2 - w^4h^4}{4} = 1$

Using the polar form : $\rho_1 = |\rho| e^{i\tilde{w}h}$ $\rho_2 = |\rho| e^{-i\tilde{w}h}$ where $\tilde{w} = \frac{1}{h}\arctan\frac{wh\sqrt{4 - w^2h^2}}{2 - w^2h^2}$

For simplicity, set $\gamma = \frac{wh\sqrt{4 - w^2h^2}}{2 - w^2h^2}$

Now, let us compute A and B: in our scheme:

$$u_0 = I = A + B$$

$$u_1 = u_{-1} \Rightarrow u_1 - u_{-1} = 0 \Rightarrow Ae^{i\omega h} + Be^{-i\omega h} = Ae^{-i\omega h} + Be^{i\omega h} \Rightarrow A(e^{i\omega h} - e^{-i\omega h}) = B(e^{i\omega h} - e^{-i\omega h}) \Rightarrow A = B$$

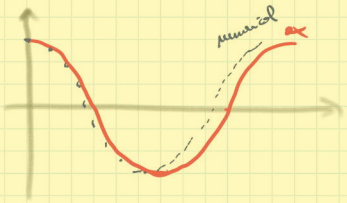
$$\Rightarrow A = B = \frac{I}{2}$$

$$u_j = \frac{I}{2}\left(e^{i\tilde{\omega}t_j} + e^{-i\tilde{\omega}t_j}\right) = I\cos(\tilde{\omega}hj)$$

The exact solution is $u_{ex}(t_j) = I\cos(\omega hj)$

From here it is apparent that (in exact arithmetic) $I_e = I$ and $\omega_e = \tilde{\omega} = \frac{1}{h}\arctan(y)$

We do not expect dissipation error, we do expect dispersion errors.
This is consistent with the numerical results.



Let's investigate $\frac{1}{h}\arctan(y)$

(1°) arctan is an odd function. For $h \to 0$, $y \to 0$ and

$$\arctan(y) \simeq y - \frac{1}{3}y^3 + \frac{1}{5}y^5$$

(2°) $y = \frac{\omega h\sqrt{4-\omega^2h^2}}{2-\omega^2h^2} = x\frac{\sqrt{4-x^2}}{2-x^2}$ (set $x = \omega h$) $\Rightarrow$ This is odd in x too, so for $x \to 0$ (even derivatives are 0)

$$y(x) \simeq x + \frac{3}{8}x^3 + cx^5$$

Putting together (1°) and (2°) we have: $\arctan(y(x)) \simeq y - \frac{1}{3}y^3 \simeq x + \frac{3}{8}x^3 - \frac{1}{3}x^3 = \omega h\left(1 + \frac{1}{24}\omega^2h^2\right)$

$$\Rightarrow \tilde{\omega} = \omega\left(1 + \frac{1}{24}\omega^2h^2\right) \quad (+ O(h^4))$$

This is the same conclusion of the book:

The dispersion error $\frac{\tilde{\omega}}{\omega} = \left(1 + \frac{\omega^2h^2}{24}\right)$ so (1) $\tilde{\omega} > \omega$ (2) $\tilde{\omega} \to \omega$ with h^2.

This analysis will subject of Homework!

POSTPROCESSING

Suppose that we are interested in $u'(x)$. The exact derivative is $-\omega I\sin(\omega t)$.

Once we have u_j we can use a formula like:

$$u'_j = \frac{u_{j+1} - u_{j-1}}{2h}$$

‖ We know this is a 2nd order formula for the exact values of u_j

$$u'_{ex}(t_j) = \frac{u_{ex_{j+1}} - u_{ex_{j-1}}}{2h} + O(h^2)$$

$$u'_j = \frac{u_{j+1} - u_{j-1}}{2h} = \frac{u_{ex_{j+1}} - e_{j+1} - u_{ex_{j-1}} + e_{j-1}}{2h} = \frac{u_{ex_{j+1}} - u_{ex_{j-1}}}{2h} - \frac{e_{j+1} - e_j}{2h} = u'_{ex}(t_j) + O(h^2) + \frac{e_{j+1} - e_j}{2h}$$

We need to understand how $\frac{e_{j+1} - e_j}{2h}$ scales with h.

$$e_{j+1} = A\delta_1^j + B\delta_2^j + Ch^2 \quad \text{with} \quad |\delta_{1,2}| = 1 \text{ and } A, B \sim O(h^2)$$

$$e_{j+1} - e_j = \underbrace{A(\delta_1^2 - 1)}_{O(h^2)}\delta_1^{j-1} + \underbrace{B(\delta_2^2 - 1)}_{O(h^2)}\delta_2^{j-1} = O(h^3)$$

By direct computation $O(h)$

$$\frac{e_{j+1} - e_j}{2h} \sim O(h^2)$$

We can verify this numerically (hands on session)

REMARK: Pay Attention to the different ways of implementing the computation of the derivative.

A Different approach for the Numerical Discretization

From the analytical point of view, we can reduce the problem to a first order set of equations with the introduction of the auxiliary variable v:

$$\begin{cases} u'' + \omega^2 u = 0 \\ u(0) = I \\ u'(0) = 0 \end{cases} \iff \begin{cases} u' = v \\ v' = -\omega^2 u \\ \begin{bmatrix} u(0) = I \\ v(0) = 0 \end{bmatrix} \end{cases}$$

We can design now different methods with different approximations for u and v with 1st order methods.

When we get to first order problems, we have a lot of options: see MATH 516

Quick Recap:

$$\frac{d\underline{y}}{dt} = \underline{f}(\underline{y}) \quad , \quad \underline{y}(0) = \underline{y}_0$$

→ One step linear methods:

$$\underline{y}^{n+1} = \underline{y}^n + h\left(\theta \underline{f}(t^{n+1}, \underline{y}^{n+1}) + (1-\theta)\underline{f}(t^n, \underline{y}^n)\right)$$

for $\theta = 0$: Explicit/Forward Euler, conditionally stable, First Order

for $\theta = 1$: Implicit/Backward Euler, unconditionally stable, First Order

For $\theta = \frac{1}{2}$: Crank-Nicolson, unconditionally stable, Second Order

For $\theta \neq 0$ we need to solve at each step a system (nonlinear if $\underline{f}$ is nonlinear, linear otherwise)

→ Linear Multi-Step Methods (Adams: Adams Bashforth / Adams Moulton, BDF)

→ Runge-Kutta : e.g., Heun:

$$\underline{y}^{n+1} = \underline{y}^n + \frac{h}{2}\left(f(t^{n+1}, \underline{u}^n + \Delta t f^n) + f^n\right)$$

In our case:

$$\underline{f} = A\underline{y} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix}\underline{y} \quad \text{with} \quad \underline{y} = \begin{bmatrix} u \\ v \end{bmatrix}$$

[Notice that A has two distinct eigenvalues: $\lambda = \pm i\omega$ expected for an oscillatory solution]

For explicit Euler, we have:

$$\underline{y}^{n+1} = \underline{y}^n + h A \underline{y}^n = (I + hA)\underline{y}^n = \begin{bmatrix} 1 & h \\ -\omega^2 h & 1 \end{bmatrix} \underline{y}^n$$

Eigenvalues: $(\lambda - 1)^2 + \omega^2 h^2 = 0 \qquad \lambda^2 - 2\lambda + 1 + \omega^2 h^2 = 0 \qquad \lambda_{1,2} = 1 \pm i\omega h \qquad |\lambda_{1,2}| = 1 + \omega^2 h^2$

Notice also that:

$$\underline{y}^{n+1} = \underline{y}^n + h A \underline{y}^n \Rightarrow \begin{cases} u^{n+1} = u^n + h v^n \\ v^{n+1} = -\omega^2 h\, u^n + v^n = -\omega^2 h\, u^n + (u^{n+1} - u^n)/h \end{cases}$$

$$u^{n+2} = u^{n+1} + h v^{n+1} = u^{n+1} - \omega^2 h^2 u^n + u^{n+1} - u^n = 2u^{n+1} - (1 + \omega^2 h^2) u^n$$

In summary, the method reads (adjusting the index)

$$u^{n+1} = 2u^n - (1 + \omega^2 h^2) u^{n-1}$$

(The method we studied was: $u^n = (2 - \omega^2 h^2) u^{n-1} - u^{n-1}$)

The associated difference equation reads

$$\rho^2 - 2\rho + (1 + \omega^2 h^2) = 0 \qquad \rho_{1,2} = 1 \pm i\omega^2 h^2 \qquad |\rho_{1,2}| = 1 + \omega^2 h^2$$

This method for this problem is unable of containing the error for the amplitude $|\rho_j| \longrightarrow +\infty \quad \forall h > 0$.

A similar analysis can be done for the three methods.

It is interesting to note that we can see:

Explicit Euler: $u^{n+1} = 2u^n - u^{n-1}(1 + \omega^2 h^2)$

Implicit Euler: $u^{n+1} = -\omega^2 h^2 u^{n+1} + 2u^n - u^{n-1}$

Crank Nicolson: $u^{n+1} = 2u^n - u^{n-1} - \frac{\omega^2 h^2}{4}\left(u^{n+1} + 2u^n + u^{n-1}\right)$

Our previous method: $u^{n+1} = 2u^n - u^{n-1} - \omega^2 h^2 u^n$

Notice that: $\dfrac{u^{n+1} + 2u^n + u^{n-1}}{4} \simeq u^n + O(h^2)$

Another Way for Measuring the Accuracy (Based on Physics)

The problem we are considering has a clear physical interpretation:

$F = ma$

$$m\frac{d^2u}{dt^2} = -Ku \Rightarrow \frac{d^2u}{dt^2} + \frac{K}{m}u = 0 \qquad \omega^2 \quad (K, m > 0)$$

No friction
No air resistance

The energy of the system is the kinetic energy of the mass $\left(\frac{1}{2}m(u')^2\right)$ + the elastic energy of the spring $\frac{1}{2}Ku^2$

For our "normalized" equation we can compute the energy with the virtual work principle.

$$E = \frac{1}{2}(u')^2 + \frac{1}{2}\omega^2 u^2 \qquad \frac{dE}{dt} = \frac{dE}{du}\frac{du}{dt} = u'u'' + \omega^2 u u' = u'\left(u'' + \omega^2 u\right) = 0$$

$\hookrightarrow u'' + \omega^2 u = 0$

Not surprisingly, the energy of the system is constant:

$$E(t) = E(0) = \frac{1}{2}0 + \frac{1}{2}\omega^2 I^2 = \frac{1}{2}\omega^2 I^2$$

A way to assess the performance/accuracy of a method is to compute the numerical energy E_j

$$E_j = \frac{1}{2}\left(\frac{u_{j+1} - u_{j-1}}{2h}\right)^2 + \frac{1}{2}\omega^2 u_j^2$$

If $\left| E_j - \frac{1}{2}\omega^2 I^2 \right|$ is small, we can conclude that the method is accurate.

EXERCISE: Test this on the method's considered here

Summary: What we learned from this example?

(1) There are many options in the discretization of an equation: use your fantasy

(2) Consistency is not enough, we need (at least) STABILITY.

(3) Direct error analysis are possible and necessary for a full educated interaction with your computer

(4) For propagative dynamics, the Fourier analysis helps understanding the error

(5) Analytical solutions are crucial for assessing the performances of a method

FINITE DIFFERENCES FOR A FIRST BOUNDARY VALUE PROBLEM (Elliptic problem)

Ref.: Quarteroni / Sacco / Saleri
NUMERICAL MATHEMATICS

We consider here how our numerical approximation may work when we consider a boundary value problem, a classical one:

$$-Ku'' = f(x) \qquad x \in (0,1)$$

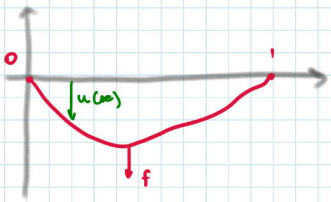
BOUNDARY CONDITIONS
$u(0) = 0$ $\qquad$ $u(1) = 0$

This is the problem of an elastic string displacement under the force f when small deformations are assumed.

The analytical Problem

Before we give details on the numerical approximation, let's see some insights on the nature of this problem.
Assume that we want to minimize the following functional

$$J(u) = \frac{1}{2}\int_0^1 K(u')^2 - \int_0^1 fu \qquad \text{with } u(0)=u(1)=0$$



This represents the total energy of a system where the first term is the potential elastic energy and the second term is the energy to win the external force f to generate the displacement u.
(Let's assume $K>0$ constant).

To compute the minimum of J we compute $\dfrac{J(u+\varepsilon v) - J(u)}{\varepsilon}$ where v is a function with the same regularity of u and $v(0)=v(1)=0$.

$$\frac{\frac{1}{2}K\left(\int_0^1 u'^2 + 2\varepsilon uv + \varepsilon^2 v'^2\right) - \int_0^1 fu - \varepsilon\int_0^1 fv - \frac{1}{2}K\int_0^1 u'^2 + \int_0^1 fu}{\varepsilon} = K\int_0^1 u'v' - \int_0^1 fv + \frac{\varepsilon}{2}K\int_0^1 (v')^2$$

Now, compute $\dfrac{\partial J}{\partial u} \overset{\text{def}}{=} \lim_{\varepsilon\to 0}\dfrac{J(u+\varepsilon v)-J(u)}{\varepsilon} = K\int_0^1 u'v' - \int_0^1 fv$

Virtual Work principle: The solution of energy minimization solves the problem (with $u(0)=u(1)=0$)

$$K\int_0^1 u'v' = \int_0^1 fv \qquad \forall\, v \text{ s.t. } v(0)=v(1)=0$$

Now, we have $K\int_0^1 u'v' \overset{\text{B.P.}}{=} K\underbrace{[u'v]_0^1}_{=0\ (v(0)=v(1)=0)} - K\int_0^1 u''v$

$\Rightarrow$ if u is regular enough (u'' exists) then the problem reads:

$$\int_0^1 (-Ku'' - f)v = 0 \quad \Rightarrow \quad -Ku'' = f \quad \text{for } x\in(0,1).$$

(v is arbitrary)

In other terms, the equation originates from a MINIMIZATION process or, more in general, from a VARIATIONAL procedure (i.e. finding the solution as $J(u) \le J(u+\text{variation})$).

This problem will be proved to be well posed later. For now, trust me.

The analytical solution can be found by double integration:

$$-Ku'' = f \Rightarrow u'(x) = -\frac{1}{K}\int_0^x f + C \Rightarrow u(x) = -\frac{1}{K}\int_0^x dz\int_0^z f(\xi)\,d\xi + Cx + D$$

ξ, z = dummy integration variables, C, D = integration constants

With the boundary conditions: $u(0)=0 \Rightarrow D=0$

$$u(1) = -\frac{1}{K}\int_0^1 dz\int_0^z f(\xi)\,d\xi + C = 0 \Rightarrow u(x) = \frac{1}{K}\left(x\int_0^1 F(z) - \int_0^x F(z)\right)$$

$F(z) = \int_0^z f(\xi)\,d\xi$ (anti-derivative of f)

Note that

$$\int_0^1 F(z)\,dz = \int_0^1 F(z)\cdot 1\,dz \overset{\text{B.P.}}{=} [zF(z)]_0^1 - \int_0^1 zf(z) = \int_0^1 f(\xi)\,d\xi - \int_0^1 \xi f(\xi)$$
(we can change name to dummy variables)

$$\int_0^x F(z)\,dz \overset{\text{B.P.}}{=} [zF(z)]_0^x - \int_0^x \xi f(\xi) = x\int_0^x f(\xi)\,d\xi - \int_0^x \xi f(\xi)$$

$$u(x) = \frac{x}{K}\int_0^1 (1-\xi)f(\xi) - \frac{1}{K}\int_0^x (x-\xi)f(\xi) = \int_x^1 \frac{x(1-\xi)}{K}f(\xi) + \int_0^x \frac{\xi(1-x)}{K}f(\xi) = \int_0^1 G(x,\xi)f(\xi)$$

where $G(x,\xi) = \begin{cases} \frac{x(1-\xi)}{K} & \text{for } x \le \xi \le 1 \\ \frac{\xi(1-x)}{K} & \text{for } 0 \le \xi \le x \end{cases}$ is called Green's function.

If $f \in C^0([0,1])$, then we have an explicit formula for the solution. Also, since $u'' \propto f \Rightarrow$ $u'' \in C^2([0,1])$

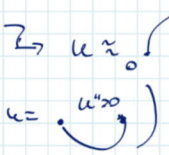
Remarks:

(1) By a direct computation, one can see that

$$\int_0^1 G(x,\xi)\,d\xi = \frac{1}{2}x(1-x) \Rightarrow \max_{0\le x\le 1} \int_0^1 G(x,\xi)\,d\xi = \frac{1}{8} \Rightarrow \boxed{\|u\|_\infty \le \frac{1}{8}\|f\|_\infty}$$

(Continuous dependence on the data).

(2) If f is non-negative, then u is non-negative too. This means that the minimum is on the boundary, as easily inferred by the fact that $-u'' = f \Rightarrow u'' = -f \le 0$

$\Rightarrow$ 

Similarly, if f is non positive, the maximum is on the boundary $\left(u = \begin{smallmatrix} u''>0 \\ u''<0 \end{smallmatrix}\right)$

(If $f=0$, both max and min are on the boundary).

The maximum principle holds also in multiple dimensions

‖ Maximum Principle

Finite Difference Numerical Approximation

As in the previous chapter/Week, we perform the following steps:

(1) Split $0 \to 1$ in sub-intervals with lenght h, so that $h = \frac{1}{N}$ N = # of intervals and the vertices are $x_j = jh$ $j = 0, 1, \dots N$

(2) Collocate the problem in x_j: $-u''(x_j) = f(x_j)$ (Set $K=1$ for simplicity)

(3) Discretize the derivative, e.g.: $-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f_j$ $u_j \simeq u(x_j)$, $f_j = f(x_j)$

with $u_0 = u_N = 0$

In this case, we cannot proceed in a specific direction finding u_{j+1} from u_j and u_{j-1}. But if we write all the equations together, we have a linear system with the matrix:

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots \\ 0 & 0 & -1 & 2 & -1 \\ \cdots & & & & \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_{N-1} \end{bmatrix}$$

i.e. $A\underline{u} = \underline{f}$ (with $u_0 = u_N = 0$)

With the notation introduced earlier:

$u = P(df)$ is the original problem $-u'' = f$, $u(0) = u(1) = 0$

$u_h = P_h(df_h)$ is the numerical problem $A\underline{u} = \underline{f}$, $u_0 = u_N = 0$

We found that the original problem is well-posed.

What about the numerical one?

The numerical problem is a linear system

$$A\underline{u} = \underline{f}$$

$$A = \frac{1}{h^2}\,\text{tridiag}(-1, 2, -1)$$

The question is: is this matrix invertible? We can answer in different ways. Let's use a straightforward one.

Proposition: A is symmetric positive definite, so it is invertible

Proof: The symmetry is trivial.

To prove that the matrix is p.d. we can use the following method: a p.d. matrix is such that $\underline{x}^T A \underline{x} > 0 \quad \forall \underline{x} \neq \underline{0}$

In our case:

$$\underline{x}^T A \underline{x} = \frac{1}{h^2}\left(2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + x_3^2 + \dots - 2x_{n-1}x_n + 2x_n^2\right) =$$

$$= \frac{1}{h^2}\left(x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + \dots + (x_{n-1} - x_n)^2 + x_n^2\right) > 0 \quad \forall \underline{x} \neq 0$$

(Notice that we can set $x_1 = x_2 = \dots x_n$ so that we get $\underline{x}^T A \underline{x} = \frac{1}{h^2}\left(x_1^2 + x_n^2\right) = \frac{2}{h^2}x_1^2$ if we take $x_1 = 0$ then $x_n = 0$ and $\underline{x} = \underline{0}$ so for $\underline{x} \neq \underline{0}$, $\underline{x}^T A \underline{x} > 0$)

A s.p.d. matrix is invertible (all the eigenvalues are > 0) so we have:

$$\underline{u}_h = A^{-1}\underline{f} \qquad \text{well posed.}$$

Convergence Analysis

Based on what we did in the previous chapter, we have:

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = -u''(x_j) + ch^2$$

so we can write an error system:

$$-u_{ex}'' = -\frac{u_{ex}(x_{j+1}) - 2u_{ex}(x_j) + u_{ex}(x_{j-1})}{h^2} + ch^2 = f(x_j)$$

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j)$$

→ This states the CONSISTENCY

$$\Rightarrow \underline{e} = \underline{u}_{ex} - \underline{u} \qquad : \qquad A\underline{e} = -\underline{c}h^2 \qquad \underline{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{bmatrix}$$

$\underline{u}_{ex}$: vector of the exact values in the nodes.

The question of the convergence refers to the stability. With linear systems, the stability is related to boundness

$$\left.\begin{aligned} A\underline{x}_1 &= \underline{b} \\ A(\underline{x} + \delta\underline{x}) &= \underline{b} + \delta\underline{b} \end{aligned}\right\} \quad A\,\delta\underline{x} = \delta\underline{b} \Rightarrow \delta\underline{x} = A^{-1}\delta\underline{b} \Rightarrow \|\delta\underline{x}\| \leq \|A^{-1}\|\,\|\delta\underline{b}\|$$

We have $\|A^{-1}\|$, so we have stability. However, the $\|A^{-1}\|$ may depend on h, so the convergence rate is not necessarily $O(h^2)$. Also, consider that if the matrix is close to be singular, the stability may be true but useless. In this case, we are in a good shape.

Proposition : The smallest eigenvalue of A is independent of h

Proof

It's technical. Let's consider the matrix $T = \text{tridiag}\left([-1, 2, -1]\right)$. The eigenvalues of A scales $\frac{1}{h^2}$ * eigenvalues of T

We can compute directly the eigenvalues noting that $T\underline{x} = \lambda \underline{x}$ reads

$$x_{j-1} - (2-\lambda)x_j + x_{j+1} = 0 \qquad x_0 = x_m = 0$$

Solving this, we have:

$$\xi^2 - (2-\lambda)\xi + 1 = 0 \qquad \xi_{1,2} = \frac{(2-\lambda) \pm \sqrt{4+\lambda^2-4\lambda-4}}{2} = \frac{(2-\lambda) \pm \sqrt{\lambda}\sqrt{\lambda-4}}{2}$$

If $\lambda > 4$:

$$x_j = A\xi_1^j + B\xi_2^j \qquad A+B=0$$
$$A\left(\xi_1^m - \xi_2^m\right) = 0 \Rightarrow A = 0 \quad \left(\text{for a generic } m \text{ with } \xi_1, \xi_2 \text{ real}\right)$$

So, NO EIGENVALUES > 4.

If $\lambda < 4$: $\xi_{1,2} = \frac{2-\lambda \pm i\sqrt{\lambda}\sqrt{4-\lambda}}{2}$ Notice that $|\xi| = 1$, so we can write $\xi_{1,2} = \cos(\theta) \pm i\sin(\theta)$

To have eigenvalues, we need that $\underline{x} \neq \underline{0}$. When we prescribe the boundary conditions, we have:

$$\left.\begin{aligned} A + B &= 0 \\ A\xi_1^m + B\xi_2^m &= 0 \end{aligned}\right\} \quad A = -B \qquad \xi_1^m - \xi_2^m = 0$$
$$\Downarrow$$
$$\cos(m\theta) + i\sin(m\theta) - \cos(m\theta) + i\sin(m\theta) = 0$$
$$\Downarrow$$
$$\sin(m\theta) = 0 \Rightarrow \theta = \frac{j\pi}{m} \qquad (j = 1 \dots m-1)$$

So the possible eigenvalues are

$$\lambda_j = -\frac{1}{\xi} - \xi + 2 = 2\left(1 - \cos\left(\frac{j\pi}{m}\right)\right) \qquad j = 1, \dots m-1$$

Recall: $\cos(2\alpha) = 1 - 2\sin^2(\alpha) \Rightarrow 2\sin^2(\alpha) = 1 - \cos(2\alpha)$

$$\lambda_j = 4\sin^2\left(\frac{j\pi}{2m}\right) \qquad j = 1, \dots m-1$$

For the matrix A: $\mu_j = \frac{1}{h^2}\lambda_j = \frac{4}{h^2}\sin^2\left(jh\frac{\pi}{2}\right)$ ($mh = 1$)

Notice that the smallest eigenvalue is for $j = 1$, and $\sin^2\left(h\frac{\pi}{2}\right) \sim h^2\frac{\pi^2}{4}$

$$\mu_{min} \sim \frac{4}{h^2}h^2\frac{\pi^2}{4} \sim \pi^2$$

This proves the Proposition.

Now, we have some points to make:

(i) $\underline{e} = A^{-1}\underline{\tau}(h) \Rightarrow \|\underline{e}\|_2 \leq \|A^{-1}\|_2 \|\underline{\tau}\|_2$

(ii) A is s.p.d. $\Rightarrow A^{-1}$ is s.p.d. and $\|A^{-1}\|_2$ is the spectral radius of A^{-1}

(iii) So spectral radius of $A^{-1} = \max\left(\left|\frac{1}{\text{eigen}(A)}\right|\right) = \frac{1}{\text{min eigen}(A)} = \frac{1}{\mu_{min}}$

Proposition 1 + (i) + (ii) + (iii) $\Rightarrow \|A^{-1}\|$ is independent of h (stability)

The question now is: how $\|\tau\|_2$ scales with h?

$$\|\tau\|_2^2 = \sum_{i,j}^{n} C_j^2 h^4 \leq \max_j \left(C_j^2 h^4\right) \sum_{i,j}^{n+1} 1 \leq \underbrace{\max_j C_j^2 h^4}_{\|\tau\|_\infty} \sum_{j=1}^{n} 1 = C_{max} h^4 n = C_{max} h^3$$

($nh = 1$)

$$\boxed{\|\tau\|_2 \sim O\left(h^{\frac{3}{2}}\right)}$$

With the stability result we have: $\|e\|_2 \sim O\left(h^{3/2}\right)$

There are other ways to prove the stability and then the convergence. One is based on the definition of the Green function for the discrete problem (see Quarteroni - Sacco - Saleri)

So, one can prove that

$$\|e\|_\infty \leq \frac{1}{8} \max_j |\tau_j| \sim O(h^2)$$

There is another technique (similar to the Finite Element Method)

Let's define the discrete scalar product $\left(u, v\right)_h = h \sum_0^n c_j u_j v_j$ $\quad c_{0,n} = \frac{1}{2}$, $c_j = 1$ if $0, n$

This is basically the trapezoidal integration formula of two functions with nodal values u_j, v_j

$$(T)\left(\int_0^r fg\right) = \frac{h}{2}\left(u_0 v_0 + u_1 v_1\right) + \frac{h}{2}\left(u_1 v_1 + u_2 v_2\right) + \dots + \frac{h}{2}\left(u_{n-1} v_{n-1} + u_n v_n\right).$$

Now, with some technicalities one can prove the following statements.

Denote

$$L_h u_h(x_j) = -\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}$$

Then:

(1) $\left(L_h u_h, v_h\right)_h = \left(u_h, L_h v_h\right)_h$ $\quad \forall\, u_h, v_h$ s.t. $u_0 = v_0 = u_n = v_n = 0$

(2) $\left(L_h v_h, v_h\right) \geq 0$ $\quad \forall\, v_h$ s.t. $v_0 = v_n = 0$

(3) $\left(L_h v_h, v_h\right) = \frac{1}{h} \sum_j \left(v_{j+1} - v_j\right)^2 = h^2 \sum_j \left(\frac{v_{j+1} - v_j}{h}\right)^2$

→ In general, this quantity is ≥ 0. However, since we have $v_0 = v_n = 0$, we can say in detail that

$$\left(L_h v_h, v_h\right) = 0 \iff v_h = 0$$

Notice that $h^2 \sum_j \left(\frac{v_{j+1} - v_j}{h}\right)^2$ is therefore a norm for the finite difference of order 1 of v_h ($\sim$ first derivative)

(4) $\left(v_h, v_h\right)_h \leq \frac{1}{2}\left(L_h v_h, v_h\right)$ or $\|v_h\|_h \leq \frac{1}{\sqrt{2}} |||v_h|||_h$

having set: $\|v_h\|_h = \left(v_h, v_h\right)_h^{1/2}$ $\quad |||v_h|||_h = \left(L_h v_h, v_h\right)_h^{1/2}$

Now, if we put together the consequences of (1-4) we obtain:

Exact problem: $L_h u_{ex,h} = f_j + c_j h^2$

Numerical problem: $L_h u_h = f_j$

Error problem: $L_h e_h = c_j \cdot h^2$

Cauchy Schwartz

$$(L_h e_h, e_h)_h = h^2 (\underline{c}, e_h)_h = h^2 \left(h \sum c_j e_j \right) \leq h^2 \, h^{1/2} \left(\sum c_j^2 \right)^{1/2} h^{1/2} \left(\sum e_j^2 \right)^{1/2}$$

$$\leq h^2 \left(\max_j c_j \right) \| e_h \|_h \, C$$

$$|||e_h|||_h^2 \leq h^2 C \| e_h \|_h$$

$$\sqrt{2} \| e_h \|_h^2 \leq h^2 C \| e_h \|$$

$$\hookrightarrow \| e_h \|_h \leq \frac{C}{\sqrt{2}} h^2$$

Again we found that

CONSISTENCY + STABILITY $\Rightarrow$ CONVERGENCE

In particular, in these linear problems, stability is equivalent to boundness. For the energy, we have

$$\sqrt{2} \| u_h \|_h^2 \leq |||u_h|||_h^2 = (L_h u_h, u_h)_h = (f, u_h)_h \leq \| f \|_h \| u_h \|_h$$

$$\| u_h \|_h \leq \frac{\| f \|_h}{\sqrt{2}}$$

This is the boundness that gives stability to the numerical problem.

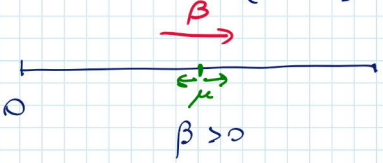
Advection-Diffusion Problems

When we extend our method to more general problems, we can find that, in practice, stability and convergence are not enough. We are happy to know that the error vanishes with h, but certainly we do not want to take h too small (poor computer!!!)

Let us consider this problem:

$$\begin{cases} -\mu u'' + \beta u' = f & x \in (0,1) \\ u(0) = 0 \quad u(1) = 1 \end{cases}$$

This is a model problem for drift $(\beta u')$ + diffusion $(\mu u'')$ problems.



μ = diffusion (Brownian motion)

β = drift / convection (or advection)

Model: pollutant in a river.

Before we study the numerical approximation, let's do some preliminary considerations.

Assume $\beta = 0$: $-\mu u'' = 0$, $u(0) = 0$, $u(1) = 1$ } Solution is rectilinear $u = x$

For $\mu = 0$, the solution is constant, but we have a problem with the boundary conditions.

For $\beta \neq 0$ we can compute the solution explicitly : Set $w = u'$ and assume $\beta \neq 0$, $\mu > 0$:

$$-\mu w' + \beta w = 0 \quad \Rightarrow \quad \frac{w'}{w} = \frac{\beta}{\mu} \quad \Rightarrow \quad \frac{d}{dx}\log(w) = \frac{\beta}{\mu}$$

$$\text{so} \quad \log(w) = \frac{\beta}{\mu}x + A \quad \Rightarrow \quad w = e^{\frac{\beta}{\mu}x} B$$

(A: constant; e^A: constant)

Now $u' = w \quad \Rightarrow \quad u = B\frac{\mu}{\beta}e^{\frac{\beta}{\mu}x} + C$ (C: constant)

$u(0) = 0 \Rightarrow B\frac{\mu}{\beta} + C = 0$

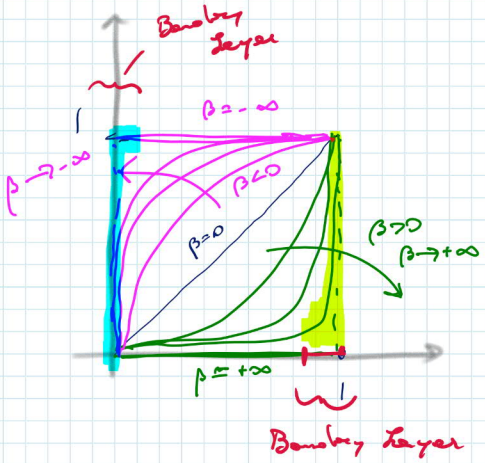
$u(1) = 1 \Rightarrow B\frac{\mu}{\beta}e^{\frac{\beta}{\mu}} + C = 1$

$B\frac{\mu}{\beta}\left(e^{\frac{\beta}{\mu}} - 1\right) = 1$

$B = \frac{\beta}{\mu}\left(e^{\frac{\beta}{\mu}} - 1\right)^{-1}$

(check: $u'' = B\frac{\beta}{\mu}e^{\frac{\beta}{\mu}x}$, $u' = Be^{\frac{\beta}{\mu}x}$ } $-\mu u'' + \beta u' = 0$)

$$u = \frac{e^{\frac{\beta}{\mu}x} - 1}{e^{\frac{\beta}{\mu}} - 1}$$



In fluid mechanics, the part of the solution connecting the "almost" 0 (for $\beta > 0$) to the 1 (or the "almost" 1 to 0 for $\beta < 0$) is called BOUNDARY LAYER.

The lenght of the boundary layer scales with $\frac{\mu}{\beta}$: larger is β vs μ, and smaller is the B.L.

We found the solution to the problem, and it is unique and bounded in $0 \div 1$.

Also, notice that $u' = \frac{\beta}{\mu}\frac{e^{\frac{\beta}{\mu}x}}{e^{\frac{\beta}{\mu}} - 1}$ so the derivative is bounded by $\frac{\beta}{\mu}\left(1 + \frac{1}{e^{\frac{\beta}{\mu}} - 1}\right)$

How can we approximate this problem numerically ?

Idea :

$$-\mu\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \beta\frac{u_{j+1} - u_{j-1}}{2h} = 0 \qquad j = 1, \dots n-1$$

$$u_0 = 0 \qquad u_n = 1$$

Let's solve this explicity, noting that each finite difference has a truncation error scaling with $O(h^2)$.

$$\left(\mu - \frac{\beta h}{2}\right)u_{j+1} - 2\mu u_j + \left(\mu + \frac{\beta h}{2}\right)u_{j-1} = 0 \qquad \left(\text{assume } h : \mu - \frac{\beta h}{2} \neq 0\right)$$

$$\frac{\mu \pm \sqrt{\mu^2 - \mu^2 + \beta^2 h^2/4}}{\mu - \frac{\beta h}{2}} = \begin{cases} 1 \\ \frac{\mu + \frac{\beta h}{2}}{\mu - \frac{\beta h}{2}} \end{cases} \qquad u_j = A\rho_1^j + B\rho_2^j \qquad \begin{matrix} u_0 = 0 & A = -B \\ u_n = 1 & A = \left(\rho_1^n - \rho_2^n\right)^{-1} \end{matrix}$$

Let's introduce the following adimensional number: $\text{Pe}clet \equiv \frac{|\beta| h}{2\mu}$

Then, we can write: $\rho_1 = 1$, $\rho_2 = \frac{1+\text{Pe}}{1-\text{Pe}}$, so that $u_j = \frac{\rho_1^j - \rho_2^j}{\rho_1^n - \rho_2^n} = \frac{1 - \frac{(1+\text{Pe})^j}{(1-\text{Pe})^j}}{1 - \frac{(1+\text{Pe})^n}{(1-\text{Pe})^n}}$

Notice that we have a numerical solution and it is possible to prove that the linear system

$$A\underline{u} = \underline{f}$$

with $A = \frac{\mu}{h^2}\,\text{tridiag}\left([-1, 2, -1]\right) + \frac{\beta}{2h}\,\text{tridiag}\left([1, 0, -1]\right)$ is INVERTIBLE.

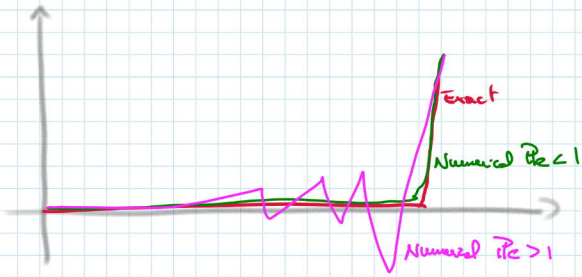
We will test that for $h \to 0$, the method has order 2 in $\|\cdot\|_\infty$ and $\|\cdot\|_{L^2}$

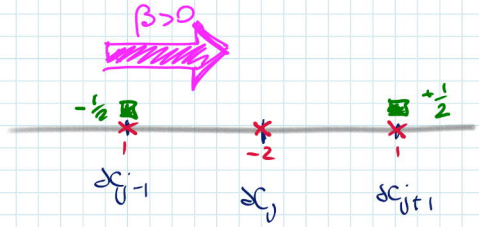
However, we notice that for large h the solution features unphysical oscillations. Why?

Note that for $\text{Pe} > 1$ $\left(h > \frac{2\mu}{\beta}\right)$, $\rho_2 < 0$, so ρ_2^j is $\begin{cases} > 0 & \text{for } j \text{ even} \\ < 0 & \text{for } j \text{ odd} \end{cases}$.

This introduces oscillations that make the solution unphysical.
Even if the solution is clear (take $h < \frac{2\mu}{\beta}$), this may be practically impossible: in real problems we may have $\frac{\mu}{\beta} \sim 10^{-6}$ or less!

This is not an instability of the numerical problem, yet for h large we need some practical solution. Even if not completely appropriate, these solutions go under the name of "stabilization".





x = 2nd derivative
■ = 1st derivative

The reason of the "instability" is the lack of "physical consistency" of the scheme with the problem.

In this problem we have a "symmetric" dynamics (the diffusion) + a "directional" one (the convection).

The numerical scheme is "symmetric" and it does not include the directionality

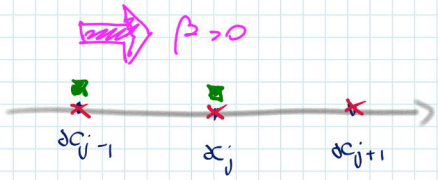
If you focus on x_j, the "wind" is blowing from the left to the right (for β positive), but the coefficients "upwind" (x_{j-1}) and "downwind" (x_{j+1}) are the same.

This is not "physically" consistent with the directional nature of the convective term.

The upwind Scheme

As a first idea, we break the "symmetry" by going upwind, i.e. in the direction where the wind is blowing from. For instance:

$(\beta > 0)$: $u'(x_j) \approx \frac{u_j - u_{j-1}}{h}$



|| We take data from "fresh" information

[When you go hunting, the prey must be "upwind" otherwise it smells you!]

The scheme reads:

$$-\frac{\mu}{h^2}u_{j+1} + \left(\frac{2\mu}{h^2} + \frac{\beta}{h}\right)u_j - \left(\frac{\mu}{h^2} + \frac{\beta}{h}\right)u_{j-1} = 0$$

$$-\frac{2\mu}{h^2}\frac{1}{2}u_{j+1} + \frac{2\mu}{h^2}\left(1 + \frac{\beta h}{2\mu}\right)u_j - \frac{2\mu}{h^2}\left(\frac{1}{2} + \frac{\beta h}{2\mu}\right)u_{j-1} = 0$$

$$\left(1 + \frac{\beta h}{2\mu}\right) = (1 + \mathrm{Pe}) \qquad \left(\frac{1}{2} + \frac{\beta h}{2\mu}\right) = \left(\frac{1}{2} + \mathrm{Pe}\right)$$

$$u_{j+1} - 2(1+\mathrm{Pe})u_j + (1+2\mathrm{Pe})u_{j-1} = 0$$

$$\beta_{1,2} = (1+\mathrm{Pe}) \pm \sqrt{1 + 2\mathrm{Pe} + \mathrm{Pe}^2 - 1 - 2\mathrm{Pe}} =$$

$$= \begin{cases} 1 + 2\mathrm{Pe} \\ 1 \end{cases} \Big\} \text{ both } > 0$$

$$\Downarrow$$

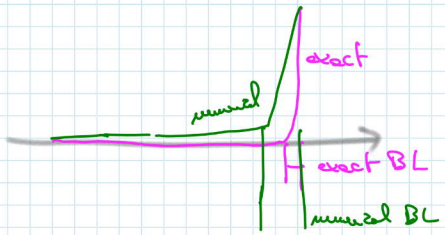
NO OSCILLATIONS

Test Upwind in the hands-on sessions.

The drawback of upwind is that the finite difference scheme we get is only 1st order:

We got stability by reducing accuracy (it happens sometimes)

In the numerical tests, we also notice that the numerical boundary layer is always overestimated vs. the exact one:



We want to explain why.

Notice that:

$$\beta\frac{u_j - u_{j-1}}{h} = \beta\frac{u_{j+1} - u_{j-1}}{2h} - \beta\frac{u_{j+1} - 2u_j + u_{j-1}}{2h} = \boxed{\beta\frac{u_{j+1} - u_{j-1}}{2h}} - \boxed{\frac{\beta h}{2}\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}}$$

Centred FD — extra numerical diffusion

UPWIND is like solving with centred finite difference with an extra numerical diffusion: it is like solving the problem:

$$-\mu_R u'' + \beta u' = 0 \qquad \text{with} \quad \mu_R = \mu + \frac{\beta h}{2} = \mu(1+\mathrm{Pe})$$

Numerical Viscosity (over diffusion)

The numerical BL is therefore $O\left(\frac{\mu_R}{\beta}\right) = O\left(\frac{\mu}{\beta} + \frac{h}{2}\right)$

From here, it is also evident that:

$$\mathrm{Pe}_{\text{upwind}} = \frac{\beta h}{2\mu_R} = \frac{\beta h}{2\mu(1+\mathrm{Pe})} = \frac{\mathrm{Pe}}{1+\mathrm{Pe}} < 1 \quad \forall h$$

Question: Can we buy stability without loosing accuracy?

Two ingredients:

(1) we need to go "upwind"

(2) we need a 2nd order approximation of the first derivative.

Remark: Reaction problems

A complete real problem may include also a so called reactive term:

$$-\mu\frac{\partial^2 u}{\partial x^2} + \beta\frac{\partial u}{\partial x} + \sigma u = 0 \qquad \left(\begin{array}{c} x \in (0,1) \\ + \text{ B.C.} \end{array}\right)$$

Diffusion — Convection (Advection) — Reaction

For instance, in the dynamics of a pollutant in a river, we have:

- diffusion is the dilution of the pollutant (random walk)
- convection is the movement of the water stream (β = water velocity)
- reaction is any chemical reaction with other agents (weighted by σ)

We found that if β is large vs μ we have numerical problems. What if $\sigma >> \mu$?

For instance:

$$-\mu\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \sigma u_j = 0 \quad \Rightarrow \quad \mu u_{j+1} - \left(2\mu + \sigma h^2\right)u_j + \mu u_{j-1} = 0$$

$$\rho^2 - 2\left(1 + \frac{\sigma h^2}{2\mu}\right)\rho + 1 = 0 \qquad \rho_{1,2} = 1 + \frac{\sigma h^2}{2\mu} \pm \sqrt{\left(1 + \frac{\sigma h^2}{2\mu}\right)^2 - 1}$$

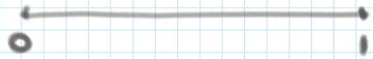
$\rho_{1,2} > 0 \quad \forall h$

No numerical oscillations!

(Reaction is less problematic than Convection)

Other Types of Boundary Conditions

So far, we considered only conditions like $u(0) = \text{data}$. In fact, we may have different conditions.



Let's keep the example of the river pollutant.

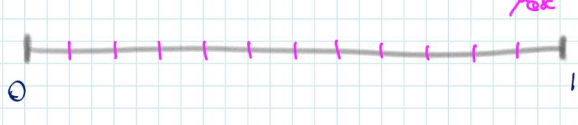
- Dirichlet (or Essential) Conditions: $u(\text{boundary}) = \text{data}$ — We prescribe a value for the concentration
- Neumann (or Natural) " : $\mu\frac{\partial u}{\partial x}(\text{boundary}) = \text{data}$ — We prescribe the flux of pollutant entering/leaving the domain
- Robin (or Mixed) " : $\mu\frac{\partial u}{\partial x} + \chi u\,(\text{boundary}) = \text{data}$ — Typically data $= \chi u_{\text{external}}$, so that: $\mu\frac{\partial u}{\partial x} = \chi(u_{\text{external}} - u)$. We prescribe a flux proportional to the difference between the external concentration and the concentration at the boundary.

Let's see an example:

$$-\mu\frac{\partial^2 u}{\partial x^2} + \beta\frac{\partial u}{\partial x} + \sigma u = 0$$

$u(0) = 1$

$$\left[\mu\frac{\partial u}{\partial x} + \chi u\right](1) = \chi u_{ex} \qquad (\text{for } \chi = 0 \text{ we have Neumann homogeneous})$$



A first order (stable) scheme:

Assuming $\beta > 0$:

$$\begin{cases} -\mu \dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \beta \dfrac{u_j - u_{j-1}}{h} + \sigma u_j = 0 & j = 1, \dots n-1 \\ u_0 = 1 \\ -\mu \dfrac{u_{n-1} - u_n}{h} + \chi u_n = \chi u_{ext} \end{cases}$$

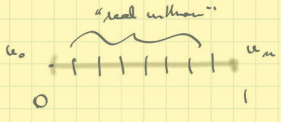
We can set a $(n+1)\times(n+1)$ system solving this problem.

What if we want a 2nd order stable scheme? (HW2)

REMARK

The Dirichlet conditions $u(\text{boundary}) = \text{data}$ are quite immediate to consider: we don't have an unknown at the boundary!!!

When we work on paper, we simply "eliminate" these unknowns and work on a smaller system:



We eliminate u_0 and u_n and solve an $(n-1)\times(n-1)$ system.

In practice, in real coding, we prefer another strategy: we retain $\begin{cases} u(0) = u_{left} \\ u(1) = u_{right} \end{cases}$ as equations and solve a $(n+1)\times(n+1)$ system

In a small example:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = f_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = f_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = f_3 \end{cases}$$

with $x_1 = \alpha$, $x_2 = \beta$

Theory → $a_{22}x_2 = f_2 - a_{21}\alpha - a_{23}\beta$ (for $a_{22} \neq 0$)

Practice →

$$\begin{cases} \gamma x_1 = \gamma\alpha \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = f_2 \\ \gamma x_3 = \gamma\beta \end{cases}$$

This coefficient intend to be of the same order of a_{21}, a_{22}, a_{23} to have good conditioning properties of the matrix.

This allows a better management of the memory and coding.
Solving a system $(n+1)\times(n+1)$ has basically the same order of $(n-1)\times(n-1)$.

In general, when coding a PDE problem, we have four steps:

Pre-processing: Definition of physical and numerical parameters

Assembly: (1) Construction of the matrix A and the right hand side W/OUT the boundary conditions
(2) Prescription of the boundary conditions

→ The two-step structure allows a more effective coding (no "if" statements in "for" loops).

Solving: the solution of the linear system(s)

Post-processing: Visualization / Verify

Week 3

Parabolic Problems: Heat (Diffusion) Equation and Similar

Langtangen - Chap. 3
Quarteroni - Chap. 5

Introduction

In this Chapter, we consider the FD approximation of the famous equation:

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} = f(x,t) \qquad x \in (0,1) \quad t > 0$$

($\mu > 0$ is a constant)

with Boundary Conditions:

$$u(0,t) = u(1,t) = 0 \qquad \forall t > 0$$

an Initial Condition:

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

It is, basically, the unsteady version of the problem considered in Week 2.
We do not consider the analytical solution, if can be obtained by separation of variables (see undergraduate courses on PDEs).

However, we give a little insight to the analysis of the solution, even without finding it explicitly. We will find a similar pattern in the analysis of the numerical solution.

Then, we will recall some basics for finding the analytical solution.

Physical Meaning of the equation

This is the classical "diffusion" equation: it may represent the evolution of a pollutant in time, driven only by the random molecular agitation. We can generalize it to the unsteady Advection-Diffusion-Reaction:

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} + \beta \frac{\partial u}{\partial x} + \sigma u = f(x,t), \qquad x \in (0,1), \quad t > 0$$

$$+ \text{B.C.} \qquad + \text{I.C.}$$
$$u(0,t) = u(1,t) = 0 \qquad u(x,0) = u_0(x)$$

Another interpretation is the "heat equation": in this case, $u(x,t)$ is the temperature of a rod. The temperature evolves for the thermal conductivity, represented by the coefficient μ. In this case, the boundary conditions have the following meaning:

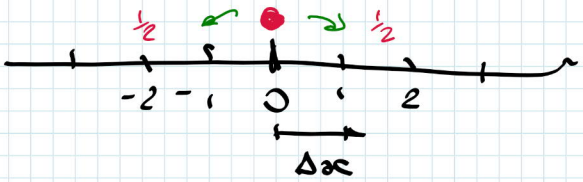
(1) $u(\text{boundary}) = \text{given}$ $\Rightarrow$ we measure the temperature on the boundary

(2) $\mu \frac{\partial u}{\partial x}(\text{boundary}) = \text{given}$ $\Rightarrow$ we prescribe the thermal flux ($= 0$ for insulated boundary)

(3) $\mu \frac{\partial u}{\partial x} + \chi u = \chi u_{ext}$ $\Rightarrow$ Radiation conditions.
(Thermal flux proportional to the difference between internal and external temperature).

It is worth recalling the following fact.

If you consider a 1D random walk, i.e. a game where a mass is located at the centre of an interval (say in 0) and at each time step the mass can move to the left with probability α and to the right with probability $1-\alpha$.



Denoting by Δx the step between two cells or positions and by Δt the time step, and by $p(x_j, t^n)$ the probability that the mass is in $x_j = j\Delta x$ at time $t^n = n\Delta t$, for $\Delta t \to 0$ and $\Delta x \to 0$ this probability obeys the diffusion equation with the initial condition

$$p(x,0) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}$$

This object is not a function, it is a distribution (called Dirac-δ).

This justifies the attention in the Langtangen book to 1D Random Walks in Chapter 3.
The fundamental solution of the heat equation on $x \in \mathbb{R}$ (no boundary conditions) with the Dirac-δ as initial condition is a Gaussian function - not surprisingly..

For a complete proof, see S. Salsa, PDEs in Action, Chap. 2
To prove this, you need to postulate that:

$$\frac{1}{2} \lim_{\substack{\Delta t \to 0 \\ \Delta x \to 0}} \frac{\Delta x^2}{\Delta t} = \text{constant}$$

and this constant turns out to be μ.

Analyzing the solution w/out computing it

Let's assume we know that a solution exists. What can we say about it, without computing it explicitly? The techniques we see here will be useful to understand also our numerical solution.

We know that u is such that

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} = f \quad , \quad u(0,t) = u(1,t) = 0 \qquad u(x,0) = u_0(x) \, .$$

For simplicity, let's start with the case $\boxed{f \equiv 0}$.

If we multiply the equation by u, we integrate over $x \in (0,1)$ and use integration by parts, we have:

$$\int_0^1 u \frac{\partial u}{\partial t} - \mu \int_0^1 u \frac{\partial^2 u}{\partial x^2} = 0 \quad \Rightarrow \quad \frac{1}{2} \int_0^1 \frac{\partial (u^2)}{\partial t} + \mu \int_0^1 \left(\frac{\partial u}{\partial x} \right)^2 = 0$$

In fact:

$$u \frac{\partial u}{\partial t} = \frac{1}{2} \frac{\partial (u^2)}{\partial t}$$

$$-\mu \int_0^1 u \frac{\partial^2 u}{\partial x^2} = -\mu \left[u \frac{\partial u}{\partial x} \right]_0^1 + \mu \int_0^1 \frac{\partial u}{\partial x} \frac{\partial u}{\partial x}$$

$u(0,t) = u(1,t) = 0$

$$\Rightarrow \quad \frac{1}{2} \frac{\partial}{\partial t} \int_0^1 u^2 + \mu \int_0^1 \left(\frac{\partial u}{\partial x} \right)^2 = 0 \qquad (*)$$

Using a notation that will be clear later with Finite Elements, for a generic function g (regular enough):

$$\| g \|^2 \stackrel{\text{def}}{=} \int_0^1 g^2 \, dx$$

Then (*) reads: $\frac{d}{dt}\|u\|^2 + \mu\left\|\frac{\partial u}{\partial x}\right\|^2 = 0$

Now, let's integrate in time between 0 and a final time T_f (generic)

$$\int_0^{T_f} \frac{d}{dt}\|u\|^2 dt + \int_0^{T_f} \mu\left\|\frac{\partial u}{\partial x}\right\|^2 dt = 0$$

(≥ 0)

$$\|u\|^2(T_f) + \int_0^{T_f} \mu\left\|\frac{\partial u}{\partial x}\right\|^2 dt = \|u\|^2(0) = \|u_0\|^2$$

$$\Rightarrow \quad \|u\|(T_f) \leq \|u_0\|$$

(given)

The solution is bounded by the initial condition!

If we have $f \neq 0$, we have the following inequality, following similar steps:

$$\|u\|^2(T_f) + \int_0^{T_f} \mu\left\|\frac{\partial u}{\partial x}\right\|^2 = \|u_0\|^2 + \int_0^{T_f}\int_0^1 f\,u\,dx\,dt$$

The Cauchy-Schwartz inequality states that: $\left|\int_0^1 f u\right| \leq \|f\|\,\|u\|$

So we have:

$$\|u\|^2(T_f) \leq \|u_0\|^2 + \int_0^{T_f} \|f\|\,\|u\|\,dt$$

At this point we call a result called: Gronwall Lemma:

GRONWALL LEMMA

If $\left.\begin{array}{l} y(t) \leq f(t) + \int_0^T g(t)y(t)\,dt \\ \text{with } f \geq 0,\ g \geq 0 \end{array}\right\}$ then $y(t) \leq f(t) + \int_0^T f g\, e^{\int_t^T g(s)ds}\,dt$

By a direct application of this Lemma, we have that the solution is bounded.

Notice that there exists a "discrete" version of this Lemma (we can use it for our numerical solution).

Calculating the exact solution (in 1D)

In 1D we can compute the exact solution using the method of SEPARATION OF VARIABLES. We recall here the basic principles.

(1) We use the superposition of effects (that holds because the problem is linear) to postulate a solution in the form:

$$u(x,t) = \sum_{j=1}^{\infty} X_j(x)\,T_j(t)$$

(2) We work out each component.

Let's look for the component $X_j(x)\,T_j(t)$ and assume that we can also write $f(x,t) = \sum F_j(t)\,X_j(x)$

We will see in a second who are the $X_j(x)$.

If we plug our solution in the equation, we have:

$$(A) \qquad \sum_{j=1}^{\infty} \frac{dT_j}{dt} X_j - \mu \sum_{j=1}^{\infty} \frac{d^2X_j}{dx^2} T_j = \sum_{j=1}^{\infty} F_j X_j$$

$$\text{with} \quad u_0(x) = \sum_{j=1}^{\infty} C_j X_j(x)$$

Now, let's assume we can solve the problem:

$$(\square) \qquad \frac{d^2X_j}{dx^2} = -\lambda_j X_j \qquad \text{with } X_j(0) = X_j(1) = 0$$

Then we have from (A)

$$\sum_{j=1}^{\infty} \left(\frac{dT_j}{dt} + \mu \lambda_j T_j - F_j \right) X_j = 0$$

So that we solve the problem by solving the ODEs:

$$\frac{dT_j}{dt} = -\mu \lambda_j T_j + F_j \qquad \text{with } T_j(0) = C_j$$

Can we solve $(\square)$?

The specific answer is easy: non trivial X_j are in the form:

$$X_j = \sin\left(\sqrt{\lambda_j}\, x\right) \qquad \text{with } \sqrt{\lambda_j} = j\pi$$

In general the problem $(\square)$ is well understood and it goes under the name of STURM-LIOUVILLE EIGENVALUE PROBLEM.

It is possible to prove that the eigenfunctions X_j form an orthogonal basis, i.e., for a generic function $g(x)$ with enough regularity, we can always write:

$$g(x) = \sum_{j=1}^{\infty} G_j X_j(x) \qquad \text{with} \quad \int_0^1 X_j X_k = 0 \quad \text{for } j \neq k$$

and $$G_j = \frac{\int_0^1 g X_j}{\int_0^1 X_j^2}$$ (generalized Fourier expansion).

So to summarize:

(1) one computes: X_j: $\dfrac{d^2X_j}{dx^2} = -\lambda_j X_j$ + B.C.

(2) Then: $$C_j = \frac{\int_0^1 u_0 X_j}{\int_0^1 X_j^2}, \quad F_j = \frac{\int_0^1 f X_j}{\int_0^1 X_j^2}$$

(3) Then:

$$\frac{dT_j}{dt} = -\mu \lambda_j T_j + F_j \qquad T_j(0) = C_j$$

(4) $$u(x,t) = \sum_{j=1}^{\infty} T_j X_j$$

Finite Differences for the Heat Equation

The heat equation is the first truly PDE problem we have to solve: we take advantage of the experience of the previous two chapters.

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} = f$$

We need to discretize both the space and the time derivative.

SPACE (SEMI) DISCRETIZATION

Let $u_j(t)$ be the approximation of $u(x_j, t)$ and let us COLLOCATE the problem in x_j, with the usual approximation for the 2nd derivative:

$$(\bullet) \quad \frac{du_j}{dt} - \frac{\mu}{h^2}\left(u_{j+1}(t) - 2u_j(t) + u_{j-1}(t)\right) = f(x_j, t)$$

with $u_j(t=0) = u_0(x_j)$ and $u_0(t) = u_m(t) = 0$.

[Notation $h = \Delta x$]

If we define the vector (function of time):

$$\underline{u}(t) = \begin{bmatrix} u_j(t) \end{bmatrix}$$

we can recognize in the semi-discrete problem $(\bullet)$ the system of ODEs:

$$\frac{d\underline{u}}{dt} = A\underline{u} + \underline{f} \qquad (\blacktriangle) \qquad (\text{with } u_j(0) = u_0(x_j))$$

where:

$$A = \frac{\mu}{h^2}\,\text{tridiag}(1, -2, 1) \qquad \underline{f} = \begin{bmatrix} f(x_j, t) \end{bmatrix}$$

We know that $\text{tridiag}(-1, 2, -1)$ is s.p.d.

Since $\frac{\mu}{h^2} > 0$, we conclude that A is s.n.d. (negative) and we already know that the eigenvalues are all real and negative

(if a generic matrix B has eigenvalues λ, $-B$ has eigenvalues $-\lambda$)

How do we discretize in time the system $(\blacktriangle)$?

The answer is an entire Chapter of MATH 516!

We faced a 2nd order Cauchy problem in Week 1. Here we have a first order problem and the solution is simpler.

It is particularly "educational" to look at a simple class of methods called θ-methods. In fact, Runge-Kutta methods are always an excellent option.

θ-Methods

The following presentation is different if compared with the books, but I think it is more rigorous.

The integral form of the Cauchy problem reads:

$$\int_{t_1}^{t_2} \frac{d\underline{u}}{dt} = \int_{t_1}^{t_2} A\underline{u} + \int_{t_1}^{t_2} \underline{f} \quad \Rightarrow \quad \underline{u}(t_2) = \underline{u}(t_1) + \int_{t_1}^{t_2} \left(A\underline{u} + \underline{f} \right) dt$$

Two words to recall basics of numerical quadratures (integrals).

We can approximate the integral of a generic function $y(t)$ as follows:

$$\int_{t_1}^{t_2} y(t)\,dt \approx w_1 y(a) + w_2 y(b)$$

with w_1, w_2 "weights" and $t_1 \le a \le b \le t_2$.

If w_1, w_2, a and b can be all be determined we can maximize the accuracy of the formula (Gauss).

Also, notice that in order to find the correct integral of a constant,

$$w_1 + w_2 = (t_2 - t_1)$$

In fact, if y is constant:

$$\int_{t_1}^{t_2} y = \text{constant} \cdot (t_2 - t_1) = (w_1 + w_2) \cdot \text{constant}$$

So, from now on we set:

$$w_1 = (1-\theta)(t_2 - t_1) \qquad w_2 = \theta (t_2 - t_1).$$

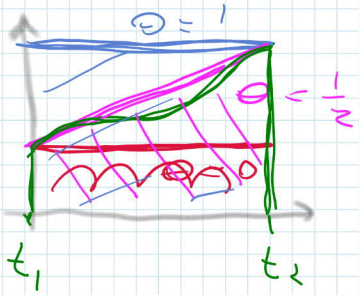
The nodes a and b can be selected "optimally" (for the accuracy) or with any practical convenience.

A common choice is: $a = t_1$ $\quad b = t_2$.

Then, we have:

$$\int_{t_1}^{t_2} y \approx \left(y(t_1)(1-\theta) + y(t_2)\theta \right)(t_2 - t_1)$$

For $\theta = 0, 1$ we approximate the integral with a rectangular area (in red like in the figure), for $\theta = \frac{1}{2}$ we have the trapezoidal rule (in magenta).



Notice that the only formula exact also for linear functions is for $\theta = \frac{1}{2}$

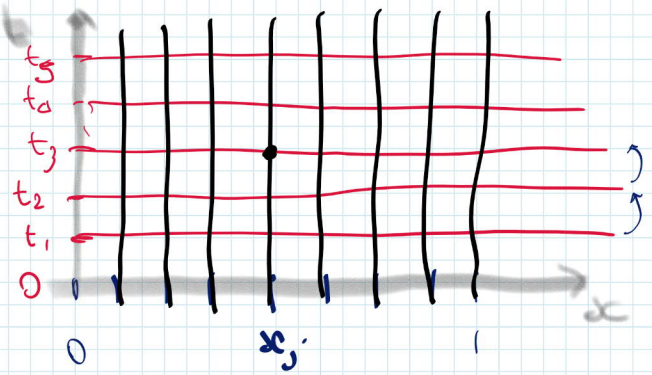
$$\frac{1}{2} = \int_0^1 x\,dx = \theta \cdot 0 + (1-\theta)\cdot 1 = 1-\theta$$

$$\boxed{\theta = \frac{1}{2}}$$

The best accuracy is attained with $\theta = \frac{1}{2}$.

Full Discretization

Now, let's discretize the time axis and apply the θ method to advance at each time step:



$t^n \to t^{n+1}$ $\quad \frac{d\underline{u}}{dt} = A\underline{u} + \underline{f}$

$$\underline{u}^{n+1} = \underline{u}^n + \int_{t^n}^{t^{n+1}} (A\underline{u} + \underline{f})\,dt \approx$$

$$\approx \underline{u}^n + \Delta t\, A\underline{u}^n(1-\theta) + \Delta t\,\underline{f}^n(1-\theta) + \Delta t\, A\underline{u}^{n+1}\theta + \Delta t\,\underline{f}^{n+1}\theta$$

where $\underline{f}^n = \left[f(x_j, t^n) \right]_{j=1,\dots,m-1}$

We can rewrite the step as:

$$(I - \Delta t\,\theta A)\,\underline{u}^{n+1} = (I + \Delta t(1-\theta)A)\,\underline{u}^n + \Delta t\left(\theta \underline{f}^{n+1} + (1-\theta)\underline{f}^n\right)$$

Notice that for $\theta = 0$ we do not have a linear system to solve.

For $\theta \neq 0$ we need to solve a linear system AT EACH TIME STEP: it is critical to have an efficient linear solver.

On the choice of the linear solver ($\theta \neq 0$)

The system has matrix:

$$I - \Delta t\,\theta A = \text{tridiag}\left[\, 1,\ 1 - 2\frac{\Delta t}{h^2}\theta,\ 1 \,\right]$$

It's a tridiagonal matrix: sparse with structured pattern
It's a time-independent matrix

If we perform an LU factorization, this holds for all the time steps!

So, a reasonable choice for the linear solver is:

(1) Factorize $(I - \Delta t \Theta A) = LU$

(2) Time loop:

for $t^n < T_f$:

$\underbrace{LU}\ u^{n+1} = (I + \Delta t(1-\theta)A)u^n + \theta f^{n+1}$

two easy systems (Order of n operations each)

}

The real questions now are:

(1) is the method stable?

(2) how the accuracy changes with θ, h and Δt?

Stability Analysis of the Implict Euler scheme

Let's preliminary consider the Implicit / Backward Euler scheme: we will see that it is unconditionally stable. The technique resembles the one used for the analytical solution.

$$u^n - \Delta t A u^n = u^{n-1} + f^n$$

(1) Multiply the two sides by u^n

$$\|u^n\|_2^2 - \Delta t\, u^n A u^n = u^n \cdot u^{n-1} + u^n \cdot f^n$$

(2) Recall that $-A$ is s.p.d.

$$\Rightarrow -\Delta t\, u^n A u^n > 0$$

(3) Also, recall that

$$0 \le \|u^n - u^{n-1}\|_2^2 = (u^n - u^{n-1})^T \cdot (u^n - u^{n-1}) = \|u^n\|_2^2 + \|u^{n-1}\|_2^2 - 2\, u^n \cdot u^{n-1}$$

$$u^n \cdot u^{n-1} \le \frac{1}{2}\|u^n\|_2^2 + \|u^{n-1}\|_2^2$$

$$\Rightarrow \frac{1}{2}\|u^n\|_2^2 - \Delta t\, u^n A u^n \le \frac{1}{2}\|u^{n-1}\|_2^2 + u^n \cdot f^n$$

Now, we sum for $n = 1, 2, \dots N_t$ where $N_t = \frac{T_f}{\Delta t}$

and we get

$$\|\underline{u}^{N_t}\|^2 + \|\underline{u}^{N_t-1}\|^2 + \dots \|\underline{u}^1\|_2^2 \underbrace{-2\Delta t \sum_{n=1}^{N_t} \underline{u}^{n\,T} A \underline{u}^n}_{>0} \leq \|\underline{u}^{N_t-1}\|^2 + \|\underline{u}^{N_t-2}\|^2 + \dots$$

$$+ \|\underline{u}^0\|^2 + 2\sum_n \|\underline{f}^n\| \|\underline{u}^n\|$$

(Discrete) Cauchy - Schwarz inequality

After cancelling the corresponding terms, and dropping the positive term:

$$\|\underline{u}^{N_t}\|^2 \leq \|\underline{u}^0\|^2 + 2\sum_n \|\underline{u}^n\| \|\underline{f}^n\|$$

$\Rightarrow$ For $\underline{f} = \underline{0}$ we have a bound on the solution at any time, based on initial data:

THE SOLUTION CANNOT EXPLODE!

$\Rightarrow$ For $\underline{f} \neq \underline{0}$ we get to the bound by using the "discrete" equivalent of the Gronwall Lemma

$\Rightarrow$ Backward Euler is unconditionally stable.

Stability Analysis of the θ-Methods

The technique for Backward Euler works because the term

$$-2\Delta t \sum_{n=1}^{N_t} \underline{u}^{n\,T} A \underline{u}^n \quad > 0$$

What can we do in general?

Some BASIC FACTS

If $A = \Delta t \frac{\mu}{h^2}$ tridiag $(1, -2, 1)$ is s.p. negative:

specifically we found that it has all negative and distinct eigenvalues: $\lambda_j = -\frac{\Delta t \mu}{h^2} 4 \sin\left(jh\frac{\pi}{2}\right)$ $j = 1, \dots N_x - 1$

where N_x is the number of space intervals $\left(N_x \cdot h = 1\right)$.

When a matrix has all distinct eigenvalues, then it is diagonalizable, i.e., there exists a matrix T s.t.:

$$AT = TD \qquad D = \text{diag}\left(\lambda_j\right)$$

T is invertible, and in the case of a symmetric matrix, it is also orthogonal, so we can write:

$$T^{-1}AT = D \Rightarrow T^T AT = D$$

because

$$TT^T = I \Rightarrow T^T = T^{-1}.$$

We are ready to analyze the general Θ method
(the analysis of Langtangen book is different: Von Neumann)

Generic step of the Θ method:

$$\left(I - \Delta t\,\Theta A\right)\underline{u}^n = \left(I + \Delta t\,(1-\Theta)A\right)\underline{u}^{n-1} + \Theta\underline{f}^n + (1-\Theta)\underline{f}^{n-1} \quad (\bullet)$$

Let's target the asymptotic behavior of the solution when $\underline{f} = 0$. The exact solution is a dissipative one, so we know that for $\underline{f} = 0$, $\underline{u}_{ex} \longrightarrow 0$.

In fact, for $\underline{f} = 0$:

$$\frac{d\underline{u}}{dt} = A\underline{u} \qquad \text{with eigenvalues of } A \text{ real and negative.}$$

Now, let's rewrite the step $(\bullet)$ as follows:

$$\left(TT^{-1} - \Delta t\,\Theta T\,\underbrace{T^{-1}AT}_{D}\,T^{-1}\right)\underline{u}^n = \left(TT^{-1} + \Delta t\,(1-\theta)T\,\underbrace{T^{-1}AT}_{D}\,T^{-1}\right)\underline{u}^{n-1}$$

$$\Downarrow$$

$$T\left(I - \Delta t\,\Theta D\right)T^{-1}\underline{u}^n = T\left(I + \Delta t\,(1-\theta)D\right)T^{-1}\underline{u}^{n-1}$$

T is invertible; also, denote:

$$T^{-1}\underline{u}^n = \underline{w}^n \quad \text{and} \quad T^{-1}\underline{u}^{n-1} = \underline{w}^{n-1}$$

Notice that if we know $\underline{w}^j \ (\forall j)$, then we know $\underline{v}^j$ $\left(\underline{v}^j = T\underline{w}^j\right)$ and if $\underline{w}^j$ blows up, so does $\underline{v}^j$ (and the other way around)

So we have:

$$\left(I - \Delta t\theta D\right)\underline{w}^n = \left(I + \Delta t(1-\theta) D\right)\underline{w}^{n-1}$$

But this is a diagonal system, so we can write the single equation:

$$\left(1 - \Delta t\theta \lambda_j\right) w_j^n = \left(1 + \Delta t(1-\theta)\lambda_j\right) w_j^{n-1}$$

$$w_j^n = \frac{1 + \Delta t(1-\theta)\lambda_j}{1 - \Delta t\theta\lambda_j} w_j^{n-1} \Rightarrow$$

$$\Rightarrow \quad w_j^n = \left(\frac{1 + \Delta t(1-\theta)\lambda_j}{1 - \Delta t\theta\lambda_j}\right)^n w_j^0$$

We have that the solution remains stable if and only if

$$\left|\frac{1 + \Delta t(1-\theta)\lambda_j}{1 - \Delta t\theta\lambda_j}\right| < 1$$

$$\Downarrow$$

$$-1 + \Delta t\theta\lambda_j < 1 + \Delta t(1-\theta)\lambda_j < 1 - \Delta t\theta\lambda_j$$

recall that $\lambda_j < 0$

$$1 - \Delta t(1-\theta)|\lambda_j| < 1 + \Delta t\theta|\lambda_j|$$

$$1 - \Delta t|\lambda_j| + \theta\Delta t|\lambda_j| < 1 + \Delta t\theta|\lambda_j|$$

ALWAYS TRUE

$$-\Delta t\theta|\lambda_j| + \Delta t(1-\theta)|\lambda_j| < 2$$

$$\Delta t|\lambda_j|(1-2\theta) < 2 \Rightarrow$$

if $\theta \geq \frac{1}{2}$ always true

if $\theta < \frac{1}{2}$ true for $\Delta t < \frac{2}{|\lambda_j|(1-2\theta)}$

Conclusion 1 (Theorem)

The θ method for the heat equation is

(1) unconditionally absolutely stable for $\theta \geq \frac{1}{2}$

(2) conditionally stable for $\theta < \frac{1}{2}$ with the condition:

$$\Delta t \leq \frac{1}{|\lambda_j|} \frac{2}{1-2\theta} \qquad j = 1, \dots N_x$$

This condition must be taken in the most restrictive sense:

$$\Delta t \leq \min_j \left| \frac{1}{\lambda_j} \right| \frac{2}{1-2\theta} = \frac{1}{\max |\lambda_j|} \frac{2}{1-2\theta}$$

Now, recalling that $|\lambda_j| \sim \frac{4}{h^2} \sin^2\left(\frac{j\pi}{2} h\right)$ we conclude that

$$\max_j \sin^2\left(\frac{j\pi}{2} h\right) \simeq \sin^2\left(\frac{\pi}{2}(1-h)\right) \approx 1 - O(h^2) \text{ so that}$$

$$j = N_x - 1$$

$$\max_j |\lambda_j| \sim \frac{1}{h^2}$$

Conclusion 2

For $\theta < \frac{1}{2}$, the stability condition relates the discretization parameters:

$$\Delta t \leq C h^2$$

Overall, this is a real constraint: if you want to refine your spatial mesh $\left(h \to \frac{h}{2}\right)$, the temporal mesh must be refined more $\left(\Delta t \to \frac{\Delta t}{4}\right)$.

For this reason, methods with $\theta \geq \frac{1}{2}$ are more popular.

We will check this conclusions with the hands-on session.

Accuracy Analysis.

We just give some results with a "sketch" of the proof.

The space discretization has an error $O(h^2)$

The time discretization has order (we didn't prove this here):

$O(\Delta t^p)$ where p depends on θ: $\begin{cases} p(\frac{1}{2}) = 2 \\ p(\theta) = 1 \quad \forall \theta \neq \frac{1}{2} \end{cases}$

We can prove the theorem:

θ methods with centred discretization of the 2nd derivative have an error:

$$O\left(\Delta t^{p(\theta)} + h^2\right) \quad \text{with} \begin{cases} p(\frac{1}{2}) = 2 \\ p(\theta) = 1 \quad \forall \theta \neq \frac{1}{2} \end{cases}$$

As usual, we get to the proof by combining stability and consistency.

Remarks

(1) We have two sources of errors, one controlled by Δt, the other by h.
In the numerical experiments it can be not easy to test the order.
You can:

(a) change Δt and h simultaneously
($\Delta t \to \frac{\Delta t}{2}$ and $h \to \frac{h}{2}$)

(b) take an extremely small parameter (e.g., Δt) and play with the other (h) so to "kill" one error and play with the other

(2) The errors of the time discretization is generally of "dissipative" nature (in particular for Implicit Euler): they tend to dissipate the energy of the problem.
Keep in mind this in Practice sessions!!

A Final Word on More Complex Problems.

We may now consider the problem:

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} + \beta \frac{\partial u}{\partial x} + \sigma u = f \qquad x \in (0,1), \ t > 0$$

$$+ \text{B.C.} + \text{I.C.}$$

In this problem, u may be the concentration of pollutant in a river:

μ is the diffusivity of the pollutant

β the velocity of the river

σ the reaction of the pollutant with the agents.

A possible set of boundary conditions is

$$u(0,t) = g(t) \qquad \frac{\partial u}{\partial x}(1,t) = 0$$

(the concentration is space invariant at the outlet)

Then, one can replace the condition in 1 with:

$$\frac{1}{h}\left(\frac{3}{2}u_{N_x} - 2u_{N_x-1} + \frac{1}{2}u_{N_x-2}\right) = 0$$

and write the system:

$$\begin{cases} u_0^n = g(t^n) \\ u_j^n = u_j^{n-1} - \Delta t\,\theta\left(\frac{\mu}{h^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) + \frac{\beta}{2h}\left(u_{j+1}^n - u_{j-1}^n\right) + \sigma u_j^n\right) \\ \quad - \Delta t(1-\theta)\left(\frac{\mu}{h^2}\left(u_{j+1}^{n-1} - 2u_j^{n-1} + u_{j-1}^{n-1}\right) + \frac{\beta}{2h}\left(u_{j+1}^{n-1} - u_{j-1}^{n-1}\right) + \sigma u_j^{n-1}\right) \\ \quad + \Delta t\,\theta f_j^n + \Delta t(1-\theta) f_j^{n-1} \qquad \text{for } j = 1, \dots N_x - 1 \\ \frac{3}{2}u_{N_x} - 2u_{N_x-1} + \frac{1}{2}u_{N_x-2} = 0 \end{cases}$$

- This is expected to be a scheme of order $O\left(h^2 + \Delta t^{p(\theta)}\right)$ unconditionally stable for $\theta \geq \frac{1}{2}$.

Remark

Of course, the centred scheme may suffer from oscillations when the $Pe > 1$. So, either we enforce $Pe < 1$ with h small enough, or we need some upwind scheme (possibly preserving the accuracy).

However, notice that: $\frac{\partial u}{\partial t}(t^n) \approx \frac{u^n - u^{n-1}}{\Delta t} - \underbrace{\frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}} + \text{H.O.T.}$

dropping this term amounts to adding a form of numerical dissipation that can delay or even suppress the Pe-related instabilities.

However, it is safe to enforce $Pe < 1$.

Week 4

Hyperbolic Problems

(1) Scalar Conservation Laws (Linear
for non-linear problems, see specific books)

(2) Systems of Conservation laws

(3) Wave Equation

Most of the material is in the slides.
Langtangen book: Chapt. 4 and 2
Quarteroni: Chapt. 13
Quarteroni-Valli: Chapt. 14

Here I address specific topics not covered in the slides

(1) Derivation of the Lax-Wendroff Scheme

The LW scheme is an explicit scheme very popular in this field.
The derivation of the scheme has a STRONG EDUCATIONAL CONTENT that goes beyond the scheme itself.

Equation:

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \qquad \begin{matrix} x \in \mathbb{R} \\ t > 0 \end{matrix}$$

LW is an example of "Taylor-based" method.

Taylor expansion of the solution in time:

$$u(x, t^{n+1}) = u(x, t^n) + \frac{\partial u}{\partial t}(x, t^n)\,\Delta t + \frac{\partial^2 u}{\partial t}(x, t^n)\frac{\Delta t^2}{2} + H.O.T.$$

This is true for any regular function.
But in the case of our problem:

$$\frac{\partial u}{\partial t} = -a \frac{\partial u}{\partial x} \qquad \text{(from the equation)}$$

and

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

(a is constant; the two derivatives commute)

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

So we have:

$$u(x,t^{n+1}) = u(x,t^n) - a\frac{\partial u}{\partial x}(x,t^n)\Delta t + a^2\frac{\Delta t^2}{2}\frac{\partial^2 u}{\partial x^2}(x,t^n) + H.O.T.$$

Now, we drop the H.O.T. (scaling with $O(\Delta t^3)$) and approximate

$$\frac{\partial u}{\partial x}(x_j,t^n) = \frac{1}{2h}\left(u_{j+1}^n - u_{j-1}^n\right) \quad ; \quad \frac{\partial^2 u}{\partial x^2}(x_j,t^n) = \frac{1}{h^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right)$$

(no surprise)

The final scheme reads:

$$u_j^{n+1} = u_j^n - \frac{a\lambda}{2}\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{a^2\lambda^2}{2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right)$$

Explicit Euler — Numerical Viscosity

(2) Wave Equation as 2x2 System

We noticed that the wave equation can be written as a first order system:

$$w_1 = \frac{\partial u}{\partial x} \qquad w_2 = \frac{\partial u}{\partial t}$$

The two mixed derivatives commute

$$\begin{cases} \frac{\partial w_1}{\partial t} = \frac{\partial^2 u}{\partial t \partial x} = \frac{\partial^2 u}{\partial x \partial t} = \frac{\partial w_2}{\partial x} \\ \frac{\partial w_2}{\partial t} = \frac{\partial^2 u}{\partial t^2} = \gamma^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial w_1}{\partial x} \end{cases}$$

$$\frac{\partial \underline{w}}{\partial t} + A\frac{\partial \underline{w}}{\partial x} = 0$$

$$A = \begin{bmatrix} 0 & -1 \\ -\gamma^2 & 0 \end{bmatrix} \qquad \det(\lambda I - A) = \lambda^2 - \gamma^2 = 0$$

$$\lambda_{1,2} = \pm\gamma$$

Eigenvectors:

$$A\underline{x} = \gamma\underline{x} \Rightarrow \begin{cases} -x_2 = \gamma x_1 \\ -\gamma^2 x_1 = \gamma x_2 \end{cases} \qquad \underline{x}_1 = \begin{bmatrix} 1 \\ -\gamma \end{bmatrix}$$

$$A\underline{x} = -\gamma\underline{x} \Rightarrow \begin{matrix} -x_2 = -\gamma x_1 \\ -\gamma^2 x_1 = -\gamma x_2 \end{matrix} \qquad \underline{x}_2 = \begin{bmatrix} 1 \\ \gamma \end{bmatrix}$$

$$T = \begin{bmatrix} 1 & 1 \\ -\gamma & \gamma \end{bmatrix} \qquad T^{-1} = \frac{1}{2\gamma}\begin{bmatrix} \gamma & -1 \\ \gamma & 1 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} \gamma & 0 \\ 0 & -\gamma \end{bmatrix}, \quad |\Lambda| = \begin{bmatrix} \gamma & 0 \\ 0 & \gamma \end{bmatrix}$$

$$T^{-1}AT = \Lambda \Rightarrow A = T\Lambda T^{-1} \qquad |A| = T|\Lambda|T^{-1} =$$

$$= \frac{1}{2\gamma}\begin{bmatrix} 1 & 1 \\ -\gamma & \gamma \end{bmatrix}\begin{bmatrix} \gamma & 0 \\ 0 & \gamma \end{bmatrix}\begin{bmatrix} \gamma & -1 \\ \gamma & 1 \end{bmatrix} = \frac{1}{2\gamma}\begin{bmatrix} 1 & 1 \\ -\gamma & \gamma \end{bmatrix}\begin{bmatrix} \gamma^2 & -\gamma \\ \gamma^2 & \gamma \end{bmatrix} = \begin{bmatrix} \gamma & 0 \\ 0 & \gamma \end{bmatrix}$$

Wave + LW:

$$\underline{w}_j^{n+1} = \underline{w}_j^n - \frac{\lambda}{2}A\left(\underline{w}_{j+1}^n - \underline{w}_{j-1}^n\right) + \frac{\lambda^2}{2}A^2\left(\underline{w}_{j+1}^n - 2\underline{w}_j^n + \underline{w}_{j-1}^n\right)$$

Then $\frac{\partial u}{\partial t}(x_j,t^n) = w_{2j}^n \Rightarrow u_j^n = u_j^{n-2} + 2\Delta t\, w_{2j}^{n-1}$ (for example, for $n \geq 2$)

(3) On the Boundary Conditions

We have noticed that some schemes involve points upwind and downwind.

Example

LW: $u_j^{n+1} = u_j^n - \frac{a\Delta t}{2h}\left(u_{j+1}^n - u_{j-1}^n\right) + \left(\frac{a\Delta t}{h}\right)^2 \frac{1}{2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right)$

for $a>0$: down, up, down, up

For UPW this is not happening:

$a>0$: $u_j^{n+1} = u_j^n - \frac{a\Delta t}{h}\left(u_j^n - u_{j-1}^n\right)$

$a<0$: $u_j^{n+1} = u_j^n - \frac{a\Delta t}{h}\left(u_{j+1}^n - u_j^n\right)$

This creates a problem with the B.C. because the data refer only to the INFLOW, we do not have data at the outflow.

For instance for $a>0$:

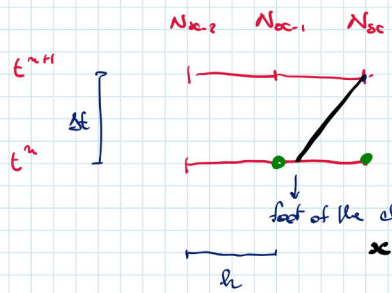
LW: $u_{N_{sc}-1}^{n+1} = u_{N_{sc}-1}^n - \frac{a\Delta t}{2h}\left(u_{N_{sc}}^n - u_{N_{sc}-2}^n\right) + \frac{a^2\Delta t^2}{2h^2}\left(u_{N_{sc}}^n - 2u_{N_{sc}}^n + u_{N_{sc}-2}^n\right)$

What data we have here?

There are several possible approximations.
We consider here an approach based on the analytical solution and the method of characteristics.

(1) For the equation $u_t + a u_x = 0$

$u(t^{n+1}, x_{N_{sc}}) = u(t^n, \text{foot of ch.})$



$x = \text{foot} + a(t - t^n)$

$x_{N_{sc}} = \text{foot} + a\Delta t$

$\text{foot} = x_{N_{sc}} - a\Delta t$

(2) For the general equation:

$u_t + a u_x + a_0 u = f$

we solve:

$$\begin{cases} \frac{du}{dt} = f - a_0 u & t \in (t^n, t^{n+1}) \\ x = a(t - t^n) + \text{foot} \\ u(\text{foot}, t^n) \text{ given} \end{cases}$$

How do we compute $u(\text{foot}, t^n)$?

By interpolation of $[u_j^n]$.
With a piecewise linear interpolation, for instance:

$$u(\text{foot}, t^n) = u_{N_{sc}-1}\frac{x_{N_{sc}} - \text{foot}}{h} + u_{N_{sc}}\frac{\text{foot} - x_{N_{sc}-1}}{h}$$

We need to perform an interpolation accurate enough to maintain the accuracy.

In short, we extrapolate the missing values along the characteristics.

This approach can be extended (with some approximation) to the non-linear case.

(4) On the Boundary Conditions (2):

Reflecting conditions / Non Reflecting Conditions

When we have a system of advection equations, we found that some components travel from the left to the right, others travel back.

For instance, for the wave equation we found that one component travels forward, the other travels backward.

The components traveling in one direction are linear combination of the physical variable $\underline{u}$

$$\frac{\partial \underline{u}}{\partial t} + A\underline{u} = 0 \quad \Rightarrow \quad T^{-1}\frac{\partial \underline{u}}{\partial t} + T^{-1}AT T^{-1}\underline{u} = 0$$

$$\Rightarrow \quad \frac{\partial \underline{w}}{\partial t} + D\underline{w} = 0$$

$$\underline{w} = T^{-1}\underline{u}$$

Each $\underline{w}$:

(1) is a linear combination of the u_i;

(2) has a well defined direction of propagation.

The physical variables $\underline{u}$ do not have a propagation direction

The mathematical variables $\underline{w}$ do have a propagation direction.

If we have a 2×2 system with 1 positive e 1 negative eigenvalue, like in the wave equation :

$\Rightarrow$ if we prescribe a condition on u_1, or u_2, we are prescribing a condition on both the mathematical variables w_1 and w_2, so we TRIGGER COMPONENTS TRAVELING IN BOTH THE DIRECTION

In general, conditions on the physical variables are "reflecting" as they turn on components traveling in any direction.

- what if we do not want reflections ?

EXAMPLE:

$$\begin{cases} \frac{\partial u_1}{\partial t} - \frac{\partial u_2}{\partial x} = 0 \\ \frac{\partial u_2}{\partial t} - \frac{\partial u_1}{\partial x} = 0 \end{cases} \qquad A = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \qquad \lambda^2 - 1 = 0 \quad \lambda_{1,2} = \pm 1$$

$$T = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \qquad T^{-1} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \qquad \underline{w} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} \frac{u_1 - u_2}{2} \\ \frac{u_1 + u_2}{2} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = T\underline{w} = \begin{bmatrix} w_1 + w_2 \\ -w_1 + w_2 \end{bmatrix}$$

If we prescribe a condition on either u_1 or u_2, we affect w_1 and w_2.

Imagine that a boundary is fictitious, so it is not a real boundary but it is just a limitation of a physical domain. For instance, assume we have a forward-propagating solution

If we prescribe conditions on $\underline{u}$, we introduce some components travelling in both the directions, so we have some SPURIOUS REFLECTIONS.

To avoid artifacts induced by the boundary conditions, we need to write conditions on the $\underline{w}$!

In the example, this means:

Inflow: $w_1 = \text{data}$ (travelling from left to right)

Outflow: $w_2 = 0$ (no backward propagating component)

Conditions on the $\underline{w}$ (inflow: $\frac{u_1 - u_2}{2} = \text{data}$; outflow: $\frac{u_1 + u_2}{2} = 0$) are practically more complex to prescribe, but doable.

These are called "Non-Reflecting Conditions".

Week 5 : Multi Dimensional Problems

When we go in more than 1D, we have conceptually the same problems but with many more technicalities to work out.

Before we face these topics, some notation you are probably already familiar with, but... just in case.

NOTATION:

Space domain: Ω

Boundary of Ω: $\partial\Omega$

OPERATORS (referring to problems in 3D)

If u is a scalar function of x_1, x_2, x_3 (and possibly t):

∇u is usually referred to space variables:

$$\nabla u = \begin{bmatrix} \partial_{x_1} u \\ \partial_{x_2} u \\ \partial_{x_3} u \end{bmatrix}$$ (gradient) in 2D is a 2D vector, in 3D has 3 components

If $\underline{v}$ is a vector, $\nabla \cdot \underline{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}$

Clearly: $\nabla \cdot (\nabla u) = \Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2}$

Einstein Notation: (1) an index refers to the component of a vector

EXAMPLE: u_i = i^{th} component of vector $\underline{u}$

(2) a comma + index i refers to the derivation w.r.t. x_i

EXAMPLE: $u_{,i}$ = i^{th} component of ∇u $\left(= \frac{\partial u}{\partial x_i}\right)$

(3) If two indexes are repeated, a sum (saturation) is understood

EXAMPLE

$$u_{i,i} = \sum \frac{\partial u_i}{\partial x_i} = \nabla \cdot \underline{w} \quad \left(= \frac{\partial w_1}{\partial x_1} + \frac{\partial w_2}{\partial x_2} + \frac{\partial w_3}{\partial x_3}\right)$$

So, we have:

$$\Delta u = u_{,ii}$$

($,$ ← derivation; ii ← index saturation)

Integration by parts in multiD (Green Formula):

$$\int_\Omega (\nabla \cdot \underline{v})\, u = \int_{\partial\Omega} \underline{v} \cdot \underline{n}\, u - \int_\Omega \underline{v} \cdot \nabla u$$

($\underline{n}$: outward normal unit vector)

[Einstein: $\int_\Omega v_{i,i}\, u = \int_{\partial\Omega} v_i n_i u - \int_\Omega v_i u_{,i}$]

Corollary: If $\underline{v} = \nabla w$ (w scalar):

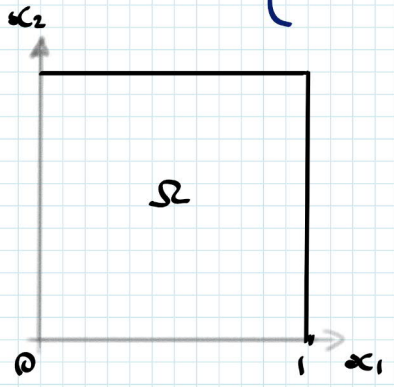
$$\int_\Omega \Delta w\, u = \int_{\partial\Omega} (\nabla w \cdot \underline{n})\, u - \int_\Omega \nabla w \cdot \nabla u$$

[Einstein: $\int_\Omega w_{,ii}\, u = \int_{\partial\Omega} w_{,i} n_i u - \int_\Omega w_{,i} u_{,i}$]

(1) Poisson (Laplace) Problem in 2D

Let Ω be a square domain $((0,1)\times(0,1))$: we want to solve

$$\begin{cases} -\Delta u = f(x_1, x_2) & \text{in } \Omega \\ u(\partial\Omega) = 0 \end{cases}$$



First, recall that the problem stems from the minimization of an energy:

$$J(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 - \int_\Omega f u$$

↑ elastic energy ; ↑ work to win f

Physically, this may be the static deformation of a square membrane clamped at the boundaries (DRUM?) under the action of the force f.

If we perturb J with $u+\varepsilon v$, we get: $(\varepsilon \in \mathbb{R})$

$$J(u+\varepsilon v) = \frac{1}{2}\int_\Omega |\nabla u|^2 + \int_\Omega \nabla u \cdot \nabla v + \frac{\varepsilon^2}{2}\int_\Omega |\nabla v|^2 - \int_\Omega f u - \varepsilon \int_\Omega f v$$

If we take the differential:

$$\frac{1}{\varepsilon}\left(J(u+\varepsilon v) - J(u)\right) = \int_\Omega \nabla u \nabla v + \frac{\varepsilon}{2}\int_\Omega |\nabla v|^2 - \int_\Omega f v$$

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(J(u+\varepsilon v) - J(u)\right) = \int_\Omega \nabla u \nabla v - \int_\Omega f v$$

At the equilibrium, the differential is 0 :

$$(1) \quad \boxed{\int_\Omega \nabla u \nabla v = \int_\Omega f v \qquad \forall v \text{ s.t. } v(\partial\Omega) = 0}$$

(the perturbation must be consistent with the b.c.)

With a standard application of the Green formula:

$$\int_\Omega \nabla u \nabla v = \cancel{\int_\Omega \nabla u \cdot \underline{n}\, v} - \int_\Omega \Delta u\, v \quad \Rightarrow \quad \int_\Omega (-\Delta u - f) v = 0$$

$=0$ because $v(\partial\Omega)=0$

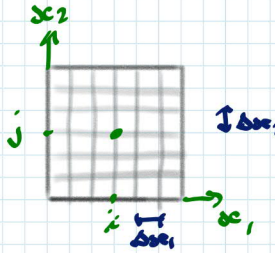
v is arbitrary

$$(2) \quad \boxed{\text{in } \Omega : \quad -\Delta u = f \qquad (u(\partial\Omega)=0)}$$

(1) and (2) are formally equivalent
(but in (2) u is differentiated twice, in (1) only once.

Numerical Approximation with FD

We can rely on cartesian directions:



$$u_{ij} \approx u(x_{1,i}, x_{2,j})$$

$$x_{1,i} = i\Delta x_1 \qquad x_{2,j} = j\Delta x_2$$

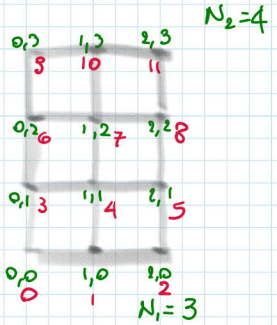
We "collocate" the problem in the vertex of the grid

$$-\Delta u(x_{1,i}, x_{2,j}) = f(x_{1,i}, x_{2,j}) = f_{ij}$$

$$-\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x_1^2} - \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta x_2^2} = f_{ij}$$

Clearly, we need to solve a linear system, but the real issue is the organization of the solution and the data.

We need to move from a 2D space representation (grid) to a vector (Real) (1D)



Total Number of points: $N_1 \times N_2$ (from 0 to $N_1 \times N_2 - 1$)

Number of horizontal points: N_1

$$0 \le i < N_1$$

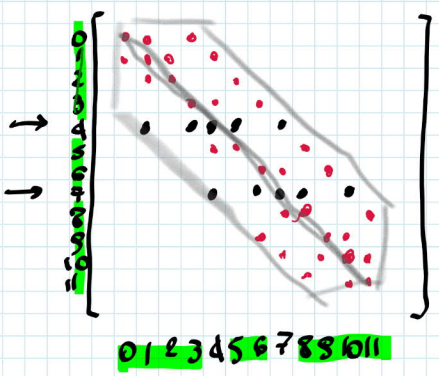
$$\text{Real} = N_1 j + i \qquad 0 \le j < N_2$$

We start from 0, as this is more consistent with computer architectures.
(MATLAB starts from 1
Python " " 0)

We use the Real index for our unknown:

$$u(\text{Real}(i,j)) = u_{ij}$$

Similarly for the matrix: in



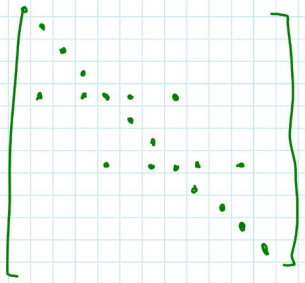
the only nodes not on the boundary are 4 and 7

However we can construct the matrix without including the B.C. The pattern of the matrix (red and black)

The matrix is banded

Main Diagonal ± 3 $(=N_1)$

Pattern after B.C.:



Diagonal coefficients $a_{ii} = -\frac{2}{\Delta x_1^2} - \frac{2}{\Delta x_2^2}$

± 1 coefficients: $a_{i,i\pm1} = \frac{1}{\Delta x_1^2}$

± 3 " $a_{i,i\pm3} = \frac{1}{\Delta x_2^2}$

After B.C. $a_{ii} = 1$, $a_{ij} = 0$ $j \neq i$, $i \in [0,1,2,3,5,6,8,9,10,11]$.

(1) We can construct the matrix by bands

$$\text{spdiags}\left(\left[-3,-1,0,1,3\right], \left[\frac{1}{\Delta x_2^2}, \frac{1}{\Delta x_1^2}, -2\left(\frac{1}{\Delta x_1^2}+\frac{1}{\Delta x_2^2}\right), \dots\right]\right)$$

(2) The nodes marked to be "boundary" are treated accordingly)

The right hand side:

Before the B.C.: $b(\text{Red}(i,j)) = f(x_1(i), x_2(j))$

After the B.C.: $b[\text{Boundary Nodes}] = \text{Boundary Values}$

Remark

inverting the map: $j = \text{integer division}\left(\frac{\text{Red}}{N_1}\right)$

$i = \text{Resto of integer division}\left(\frac{\text{Red}}{N_1}\right)$

On the solution of the system

When we include the boundary conditions and look at the matrix only for the internal nodes, we have (assuming for simplicity $\Delta x_1 = \Delta x_2 = h$)

$$A = \frac{1}{h^2}\begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix}$$ which is S.P.D. $(\lambda_1 = 5,\ \lambda_2 = 3)$

We will see that there are features of the matrix reflecting features of the physical problem. This is critical for the solution of the linear system and, in particular, for the choice of the method.

As you can see, it's easy in multiple dimensions to have large linear systems. Direct Methods are an option only if the matrix is not too large, otherwise the fill-in of sparse matrices can fill up the memory.

When the matrix is SPD, the Cholesky factorization is better,

because it takes half of a memory of LU

For large matrices, the only option is iterative methods.
The Book of Langtangen considers methods like Jacobi and S.O.R. These are pretty old methods.
The current choices are:

SPD Matrices: CONJUGATE GRADIENT

General Matrices: GMRES (powerful and well known)

BiCGStab (efficient but with a behavior of the error sometimes unexpected).

Remark On the Dirichlet B.C.

As you notice, the boundary nodes are now distributed everywhere in the solution, not just first and last as in the 1D case.
They need to be LABELED to be processed in the code and the sequence:

(1) Build the system w/out Dirichlet B.C.

(2) Process the Dirichlet B.C.

is much more efficient than the merging of the two steps.

General Boundary Conditions.

The Poisson problem may feature other types of B.C.:

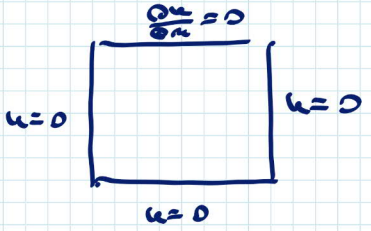
$$\mu \nabla u \cdot \underline{n} = \text{data} \qquad (\underline{n} = \text{normal unit vector})$$

($\mu \nabla u \cdot \underline{n}$: outward normal flux)

or

$$\mu \nabla u \cdot \underline{n} + \alpha u = \alpha u_\infty \qquad \text{where } u_\infty \text{ is given.}$$

For instance, we can prescribe on the unit square:



In this case:

$$\underline{n} = [0, 1]$$

and

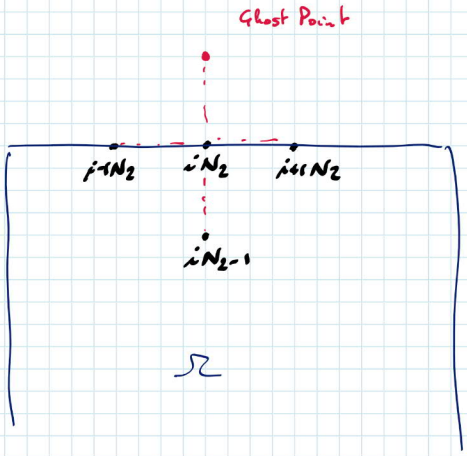
$$\frac{\partial u}{\partial \underline{n}} \equiv \nabla u \cdot \underline{n} = \frac{\partial u}{\partial x_2}$$

To preserve the second-order nature of our scheme, we could write for the nodes with $x_2 = 1$:

$$\frac{1}{h}\left(\frac{3}{2} u_{i N_2} - 2 u_{i N_2 - 1} + u_{i N_2 - 2}\right) = 0$$

(j when $x_2 = 1$)

or we can write the equation also on the boundary by using ghost points:



$$-\frac{1}{\Delta x_1^2}\left(u_{i-1N_2} - 2u_{iN_2} + u_{i+1N_2}\right)$$

$$-\frac{1}{\Delta x_2^2}\left(u_{iN_2+1} + 2u_{iN_2-1} - 2u_{iN_2} + u_{iN_2-1}\right) = f_{iN_2}$$

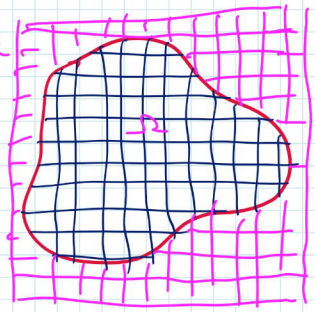
Ghost

B.C.: $\dfrac{u_{iN_2+1} - u_{iN_2-1}}{2\Delta x_2} = 0 \Rightarrow u_{iN_2+1} = u_{iN_2-1}$

Clearly, FD are strongly Cartesian-oriented.

Generic domains may create troubles:

Note: A possible approach is to embed the generic domain into a cartesian frame and then solve a fictitious problem that reduces to the physical one on the domain Ω



We need many local adjustments to manage this case.

↓

We do not cover them here, as with Finite Elements they are gone!

(2) Advection-Diffusion Problems in 2D (and in 3D also)

If we want to consider a problem in multiple dimensions with diffusion and advection:

$$-\mu \Delta u + \underline{\beta} \cdot \nabla u = f \quad \text{in } \Omega \qquad (\mu > 0)$$

$$+ \text{B.C.}$$

where $\underline{\beta}$ is a vector $[\beta_1, \beta_2]$ with the direction and strength of the wind.

In detail, the problem reads:

$$-\mu \frac{\partial^2 u}{\partial x_1^2} - \mu \frac{\partial^2 u}{\partial x_2^2} + \beta_1 \frac{\partial u}{\partial x_1} + \beta_2 \frac{\partial u}{\partial x_2} = f(x_1, x_2)$$

when $\|\underline{\beta}\| >> \mu$, should we expect numerical instabilities?

The answer is clearly positive, for the same reason of 1D: with $\|\beta\| > \mu$ the problem gets closer to change its nature, the first order term winning against the 2nd order one.

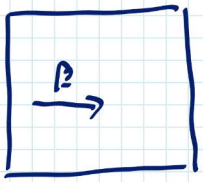
However, in 1D we noticed that upwind was equivalent to adding a numerical viscosity to the problem:

$$\beta \frac{u_i - u_{i-1}}{h} = \underbrace{\beta \frac{u_{i+1} - u_{i-1}}{2h}}_{\text{centred}} - \underbrace{\frac{\beta h}{2}\left(\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}\right)}_{\text{numerical viscosity (stabilizing)}}$$

$(\beta > 0)$

This is **not** true in multiple dimensions.

To see this, let's take the case $\underline{\beta} = [\beta_1, 0]$, so the problem reads:



$$-\mu \frac{\partial^2 u}{\partial x_1^2} - \mu \frac{\partial^2 u}{\partial x_2^2} + \beta_1 \frac{\partial u}{\partial x} = 0$$

Before we see numerical stabilizations, let's notice that this problem is:

- advection-diffusion along x_1 (for x_2 constant)
- purely diffusion along x_2 (= for x_1 constant)

This suggests that we don't need stabilization along x_2.
In fact, if we do UPWIND when $|\beta_1| h / 2\mu > 1$, we do:

$$-\mu \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{\Delta x_1^2} - \mu \frac{u_{i,j+1} - 2u_{ij} + u_{ij-1}}{\Delta x_2^2} + \beta_1 \frac{u_{ij} - u_{i-1,j}}{\Delta x_1} = f_{ij} \qquad \text{for } \beta_1 > 0$$

$$+ \beta_1 \frac{u_{i+1,j} - u_{ij}}{\Delta x_1} = f_{ij} \qquad \text{for } \beta_1 < 0$$

This corresponds to:

$$-\mu(1 + \mathrm{Pe}) \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{\Delta x_1^2} - \mu \frac{u_{i,j+1} - 2u_{ij} + u_{ij-1}}{\Delta x_2^2} + \beta_1 \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x_1} = f_{ij}$$

We are adding viscosity only along x_1 (where really needed)

This is untouched

The "artificial viscosity" here would lead to:

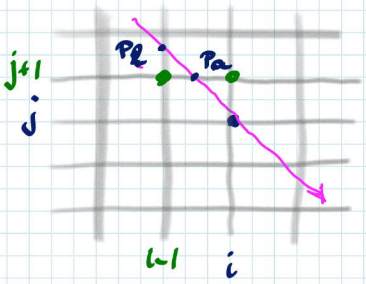
$$-\mu(1 + \mathrm{Pe}) \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{\Delta x_1^2} - \mu(1 + \mathrm{Pe}) \frac{u_{i,j+1} - 2u_{ij} + u_{ij-1}}{\Delta x_2^2} + \beta_1 \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x_1} = f_{ij}$$

This is USELESS

In other terms, UPWIND ADDS VISCOSITY ONLY ALONG THE WIND DIRECTION (a.k.a. streamlines) while adding viscosity $\mu^* = \mu(1+\alpha_E)$ just adds viscosity in all the directions, including the CROSSWIND one. However, along the crosswind direction we don't need it for stabilizing, so we just add a <u>useless numerical error</u>.

Upwind along a generic direction

How do we implement upwind along a generic direction?



There are different strategies. We see here one, while FE we will see another one.

Let's assume to have $\underline{\beta}$ constant for the moment.

(streamline)

If β is constant, the upwind direction is denoted by

$$x_1 = \beta_1 \tau + x_{1i} \qquad x_2 = \beta_2 \tau + x_{2j}$$

where τ is a parameter

For $\tau = 0$ we are in x_{1i}, x_{2j}

Assuming (just for the sake of the example) that $\beta_1 > 0$ $\beta_2 < 0$ we can identify the upwind point. We can find two points: one for $x_1 = x_{1i-1}$ ($\beta_1 > 0$), the other with $x_2 = x_{2j-1}$ ($\beta_2 < 0$)

$$P_a: \quad \bar{\tau}_a = \frac{x_{2j+1} - x_{2j}}{\beta_2} \; ; \; x_1 = \beta_1 \bar{\tau}_a + x_{1i} \qquad (\bar{\tau}_a < 0)$$

$$P_b: \quad \bar{\tau}_b = \frac{x_{1i-1} - x_{1i}}{\beta_1} \; ; \; x_2 = \beta_2 \bar{\tau}_b + x_{2j}$$

Now, to make the approximation of the first derivative, we choose the point closest to x_{1i}, x_{2j}, in this case P_a.

Then:

$$\underline{\beta} \cdot \nabla u \simeq \beta_1 \frac{u_{ij} - u(P_a)}{\Delta x_1} + \beta_2 \frac{u(P_a) - u_{ij}}{\Delta x_2}$$

We have the problem that $u(P_a)$ in general is not known. We can proceed by interpolation, for instance:

$$u(P_a) \simeq u_{i-1,j+1} \frac{x_{1i} - x_{1a}}{\Delta x_1} + u_{ij+1} \frac{x_{1a} - x_{1i-1}}{\Delta x_1}$$

using the green points for the interpolation.

As you may notice, the upwind of order 2 requires a lot of technicalities, all related to the "Cartesian" structure of the method.

With Finite Elements we will see a more general approach, still based on adding numerical viscosity, but not isotropically as done before.

Week 6 Module 2: The Finite Element Method

(1) Some Recall of Functional Analysis

Quarteroni: Chap. 2
Formaggia et al.: Chap. 1

In this week we will introduce some basic tools for the analysis of differential equations of interest.

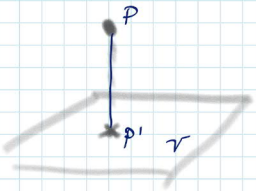
We need some concepts of functional analysis to have the right framework.

Functional Analysis, at the lowest level, is the branch of mathematics trying to extend to functions some Geometrical concepts like "distance" and "orthogonality".

These concepts are extremely helpful for us:

⇒ "distance" between two functions is related to measuring the error between the exact solution and the numerical one.

⇒ "orthogonality" between two functions is related to the optimality of approximation:



P' is the "best approximation of P on the plane V

⇕

PP' is orthogonal to V

Clearly, we do not cover many things here, Functional Analysis is a course per se, but we recall here some basic tool.

Functional: a functional is an application associating a real number to a function:

For instance, given a continuous function f on an interval (a,b),

$$\int_a^b f\,dx \quad \text{is a functional.}$$

If a functional $\mathcal{J}$ is s.t., given two functions f_1, f_2:

$$\mathcal{J}(\lambda f_1 + \mu f_2) = \lambda \mathcal{J}(f_1) + \mu \mathcal{J}(f_2)$$

we say that the functional is linear.

Form: suppose to have an application between a couple of functions and a real number:

$$a(u,v) \longrightarrow \mathbb{R} \qquad \text{for given } u \text{ and } v.$$

This is called a Form.

The form is bilinear if:

$$a(\lambda_1 u_1 + \mu_1 u_2, v) = \lambda_1 a(u_1,v) + \mu_1 a(u_2,v)$$
$$a(u, \lambda_2 v_1 + \mu_2 v_2) = \lambda_2 a(u,v_1) + \mu_2 a(u,v_2)$$

A bilinear form is symmetric when (for all the possible arguments):

$$a(u,v) = a(v,u)$$

Functional Spaces

A set is a collection of objects

A "space" is a set with "operations", i.e. a set of objects that can be combined:

X is a space: $u \in X$, $v \in V$

$\Downarrow$

$\exists u \in X \quad \forall \lambda \in \mathbb{R}$

$u + v \in X \quad \forall u, v \in X$

$(\exists \lambda u + \mu v \in X \quad \forall u, v \in X, \ \forall \lambda, \mu \in \mathbb{R})$

A space is "metric" if we can introduce a non-negative functional that can measure the distance between two objects:

$a \in X$, $b \in X$ $\quad d(a,b) \geq 0$ and $d(a,b) = 0 \iff a = b$

$d(a,b) = d(b,a)$

$d(a,c) \leq d(a,b) + d(b,c)$

With the distance, we can introduce a "topology"

A "closed" space is such that for a convergent sequence a_n, such that

$$\lim_{n \to +\infty} d(a, a_n) = 0 \quad \text{then} \quad a \in X.$$

A "Cauchy sequence" is a sequence such that

$$\lim_{m,n \to +\infty} d(a_m, a_n) = 0$$

In general, being complete is not sufficient (but necessary) for the convergence in the space. If it is also sufficient, the space is called "Complete".

A norm is a functional s.t.

$$\|f\|_X \geq 0 \quad \text{and} = 0 \text{ for } f = 0$$

$$\|\lambda f\|_X = |\lambda| \|f\|_X \quad \forall \lambda \in \mathbb{R} \text{ and } f \in X$$

$$\|f_1 + f_2\|_X \leq \|f_1\|_X + \|f_2\|_X$$

If a space has a norm, it is called "normed". In fact, a norm induces the following definition of distance:

$$d(a,b) = \|a - b\|_X$$

Definition: A normed complete space is called a Banach space

A **scalar product** $(\cdot,\cdot)_X$ is a symmetric bilinear form defined with both the arguments in a given space such that:

$$(f,f) \geq 0 \quad (=0 \text{ only for } f=0), \qquad |(f_1,f_2)| \leq (f_1,f_1)^{1/2}(f_2,f_2)^{1/2}$$

A scalar product induces a norm:

$$\|f\|_X^2 = (f,f)_X \qquad \left(\text{e.g. } (f_1+f_2,f_1+f_2)_X = (f_1,f_1)+2(f_1,f_2)+(f_2,f_2) \leq \|f_1\|^2 + 2\|f_1\|\|f_2\| + \|f_2\|^2 = \left(\|f_1\|+\|f_2\|\right)^2\right)$$

So, a space with a scalar product is automatically normed.

A **Hilbert space** is a Banach space equipped with a scalar product.

The importance of scalar products is in the definition of ORTHOGONALITY: from a geometrical to a functional definition:

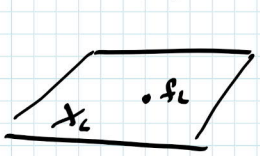
$$f \perp g \text{ in a Hilbert space} : \quad (f,g)=0$$

Thanks to the scalar product, we can define

- the "angle" between two functions: $\sphericalangle(f_1,f_2)$: $\cos(\varphi) = \dfrac{(f_1,f_2)}{\|f_1\|\|f_2\|} = \dfrac{(f_1,f_2)}{(f_1,f_1)^{1/2}(f_2,f_2)^{1/2}}$
- the "optimal approximation of a function in a subspace

Let $X_L \subset X$ where X_L is "small", both Hilbert spaces:

we want to find the best approximation f_L of f in X_L:



$$(f-f_L, v_L) = 0 \qquad \forall v_L \in X_L$$

This is a consequence of the Pythagoras Theorem:

If g_L is in X_L:

$$\|f-g_L\|^2 = (f-g_L, f-g_L) = (f-f_L+f_L-g_L, f-f_L+f_L-g_L) =$$

$$= \|f-f_L\|^2 + 2\underbrace{(f-f_L, f_L-g_L)}_{=0 \text{ (orthogonality)}} + \|f_L-g_L\|^2 =$$

$$= \|f-f_L\|^2 + \|f_L-g_L\|^2 \geq \|f-f_L\|^2.$$

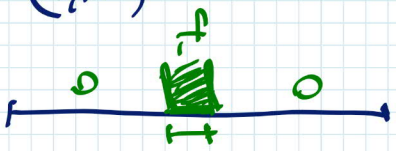
Lebesgue Integration and some functional space

Differential calculus as we learned in undergraduate courses has some limiting when approaching PDE's. It's strongly related to the concept of "continuity", but we know that reality and real problems are plenty of discontinuities.

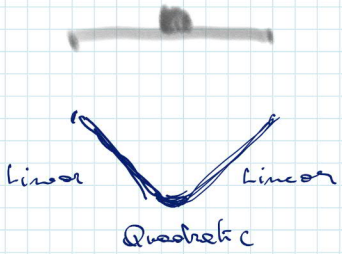
For instance, when we have a string (steady conditions):

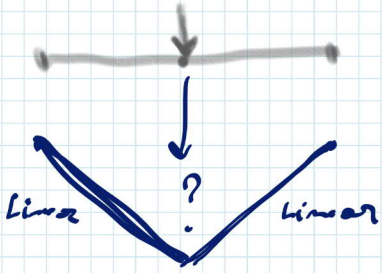
$$-u'' = f \quad \text{in } (a,b)$$

f may be discontinuous:

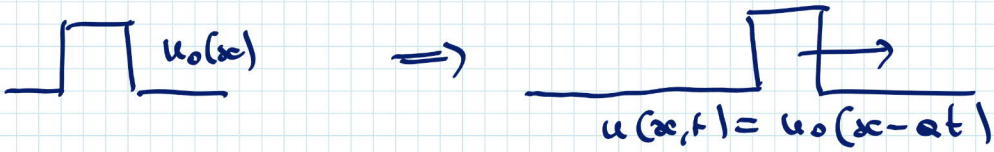


We can compute the solution, but we have some formal inconsistency, particularly if we let the support of the force f (= the interval where $f \neq 0$) to have an infinitesimal lenght





We may have only an intuitive understanding but with some formal problems, like for the solution $u_t + a\,u_x = 0$ with a discontinuous initial condition.



We know that this is the solution, even if formally this cannot be differentiated in space.

We need new conceptualizations for integrals and derivatives.

Lebesgue integrals are a new and somehow different way of thinking of integrals, based on a new theory of measure introduced by the French mathematician Lebesgue.

Instead of focusing on the function it'self, Lebesgue integration focuses also on the "dx"

$$(L)\int_a^b f(x)\,dx$$

If dx is null-measure, then locally the behaviour of the integral

may be "pathologic" but the function can be integrated.

The traditional and the Lebesgue integrals are largely the same (the same on regular functions) but with Lebesgue we can face functions with local lack of regularity, exactly what we need for PDE's.

Based on this, we can introduce the following spaces:

$$L^p(a,b): \left\{ f: (L)\int_a^b f^p\,dx < +\infty \right\} \quad \text{with } p = 1,2,\ldots \in \mathbb{N}$$

(in the sense of Lebesgue)

These are BANACH spaces with the norm $\|f\|_p = \left(\int_a^b |f|^p dx\right)^{1/p}$

(similar definition in $\Omega \subset \mathbb{R}^n$ $n>1$).

$$L^\infty(a,b): \left\{ f: \text{bounded (up to null-measure intervals) in } (a,b) \right\}$$

Banach space with $\|f\|_\infty = \max_{(a,b)} |f|$

These definitions can be extended to $(a,b)=\mathbb{R}$ or, in general, unbounded domains.

In particular: $L^2(a,b)$ is a Hilbert space with the scalar product

$$(f,g) = \int_a^b f g\,dx \qquad \left(\text{Norm: } \left(\int_a^b f^2 dx\right)^{1/2}\right).$$

$L^2(a,b)$ or $L^2(\Omega)$ will be our "home" for most of our pde's studies.

Notice that if we have a function $f \in L^p(a,b)$ and $g \in L^q(a,b)$ with $\frac{1}{p}+\frac{1}{q}=1$ then

$$(L)\int_a^b fg < +\infty \qquad \left(\text{or } fg \in L^1(a,b)\right)$$

L^p and L^q are conjugate spaces.

Clearly, the only space conjugate with itself is L^2, and we have

$$|(f,g)_{L^2}| \le \|f\|_2 \|g\|_2 .$$

Distributions and distributional derivatives

The set of linear bounded functionals defined on a space X is per se a functional space, that we call <u>dual</u> X':

$$X' = \{ \mathcal{F}(f) \quad \text{for } f \in X \text{, linear and s.t. } |\mathcal{F}(f)| \leq C \|f\|_X \}$$

Notice that for linear functionals <u>boundness and continuity</u> coincide.

Notice also that in L^2, the dual of L^2 is isometric to L^2.
(this is true for any Hilbert spaces).

this means that any linear and continuous functional in L^2 corresponds to an element within L^2 itself (Riesz representation Theorem).

The space $\mathcal{D}$

We define "support" of a function the interval or region of the domain of definition when the function is $\neq 0$.
In particular, if the support is a closed bounded region, we say that the function has a COMPACT support.

With the letter $\mathcal{D}$ we denote the space of functions in $C^\infty(\mathbb{R}^n)$ $(n \geq 1)$ with compact support

Let introduce the dual space of $\mathcal{D}$, $\mathcal{D}'$ with the following notation:

$$\mathcal{D}'\langle T, f \rangle_{\mathcal{D}} \qquad \begin{matrix} T \in \mathcal{D}' \\ f \in \mathcal{D} \end{matrix}$$

EXAMPLES:

① Let $H(x)$ be the step function $= \begin{cases} 0 & \text{for } x \leq 0 \\ 1 & \text{for } x > 0 \end{cases}$

$$\langle T, f \rangle = \int_{-\infty}^{+\infty} H f \, dx = \int_0^{+\infty} f \, dx$$

② The functional $\langle T_0, f \rangle = f(0)$ belongs to $\mathcal{D}'$

in fact $\left| \langle T_0, f \rangle \right| \leq \max_{x \in \mathbb{R}} |f(x)|$

The objects of $\mathcal{D}'$ are called <u>DISTRIBUTIONS</u> or Generalized Functions.

We can introduce a differential calculus for distributions.

In fact, we formally define:

$$\langle T', f \rangle = - \langle T, f' \rangle$$

This is inspired by integration by parts. In facts, if we take a L^2 function for instance, called g, we have:

$$\int_{\mathbb{R}} g' f = \Big[gf \Big]_{-\infty}^{+\infty} - \int_{\mathbb{R}} g f'$$

$\Rightarrow 0$ because $\lim_{x \to \pm\infty} f = 0$

$$\Big\| \Big\downarrow$$

$$\langle g', f \rangle = - \langle g, f' \rangle$$

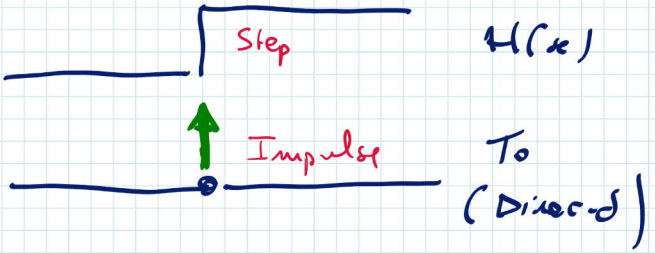
The definition obviously applies to higher derivatives.
Notice that being $f \in C^\infty$, we can differentiate as many times as we want.

EXAMPLE

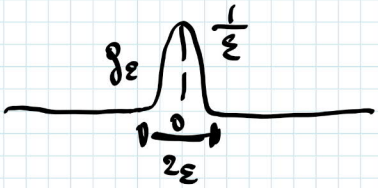
$$\int_{-\infty}^{+\infty} H' f = - \int_{-\infty}^{+\infty} H f' = - \int_0^{+\infty} f = -\Big[f \Big]_0^{+\infty} = f(0) = \langle T_0, f \rangle$$

T_0 is called Dirac-δ and it is the mathematical description of an impulse

It is not a function in the traditional sense



Dirac-δ is the limit of functions:



$$\lim_{\varepsilon \to 0} g_\varepsilon = \delta$$

when $\int_{\mathbb{R}} g_\varepsilon = 1 \quad (\forall \varepsilon)$

The definition of derivatives in the sense of distributions is very general and with a wide range of applications.

$$L^2 \subset \mathcal{D}'$$

FROM NOW ON, WHEN WE DIFFERENTIATE A FUNCTION IN L^2, WE INTEND A DISTRIBUTIONAL DERIVATIVE.

Sobolev Spaces

Once we have obtained the new concept of derivative that doesn't require a strong constraint like continuity, we can characterize functions that can be differentiated

Space $H^k(a,b)$ (or $H^k(\mathbb{R})$) is the space of functions f s.t. the belong to L^2 together with all the derivatives up to the order k.

$$H^k(a,b) \equiv \left\{ f \text{ s.t. } f, f' \dots f^{(k)} \in L^2 \equiv H^0 \right\}$$

In multiple dimensions, the derivatives are all the derivatives of order k.

$$\frac{\partial^{\alpha_1+\alpha_2+\dots+\alpha_n} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_n^{\alpha_n}} \in L^2 \qquad (*)$$

with $\sum \alpha_i = k$

These spaces are called SOBOLEV SPACES (Russian mathematician) and are a special case of the more general definition:

$$W^{k,p} \equiv \left\{ f \text{ s.t. } f, f' \dots f^{(k)} \in L^p \right\}$$

$(H^k = W^{k,2})$.

There are many properties of spaces H^k. Let's just recall some of them:

⇒ $H^k(a,b)$ is a Hilbert space with the scalar product

$$(f,g)_{H^k} = \sum_{i=0}^{k} \left(f^{(i)}, g^{(i)} \right)_{L^2}$$

(i: differentiation index)

(in 1D, in multi-dimension the sum is extended to all the derivatives in the form $(*)$).

Specifically, $H^1(a,b)$ has the norm:

$$\|f\|_{H^1}^2 = \|f\|_{L^2}^2 + \|f'\|_{L^2}^2$$

Clearly $H^1 \subset L^2$ so $\|f\|_{H^1}^2 \geq \|f\|_{L^2}^2$ (as obvious from the definition).

In general $H^{k+1} \subset H^k \subset \dots L^2$.

Notice that the definition of Sobolev spaces H^s can be extended to s real and negative.

The Space H^1_0

In general, the value of a L^2 function in a point or on a line (in 2D) or a surface (in 3D) doesn't matter because L^2 is defined up to null-measure sets and a point/line/surface has null measure in 1D/2D/3D.

However, in PDE's we prescribe boundary conditions, so we need to understand in what sense we can prescribe BC.

This has been clarified by the introduction of the concept of TRACE. TRACE is an operator that maps a function in H^α to its value of the boundary and, in general, reduces regularity:

$$\gamma : H^\alpha(\Omega) \longrightarrow H^{\alpha-\frac{1}{2}}(\partial\Omega)$$

is the (continuous) trace operator.

From now on, BC will be intended in the sense of traces.

The space $H^1_0(\Omega)$ is the space of H^1 functions with null trace on $\partial\Omega$.

This will be very important for our problems.

The scalar product and norm of H^1_0 are like H^1. However we have a fundamental property in H^1_0:

POINCARE' INEQUALITY: $\exists C>0$ s.t. $\forall f \in H^1_0(\Omega)$:

$$\|f\|_{L^2} \leq C \|\nabla f\|_{L^2}$$

This allows to write the following sequence of inequalities:

$$\|\nabla f\|^2_{L^2} \leq \|f\|^2_{L^2} + \|\nabla f\|^2_{L^2} \leq (1+C^2)\|\nabla f\|^2_{L^2}$$

$\Rightarrow$ In $H^1_0(\Omega)$, the norm of ∇f in L^2 and of u in H^1 are equivalent

This is true also if the function has only a portion of the boundary (with measure >0) where it vanishes.

The reason why the Poincaré inequality has to be expected in H^1_0 is that in general we may have a function with the norm $\geq$ of the gradient: the constants! However, in H^1_0 the only constant possible is 0.

The space of linear and continuous functionals on H^1_0 is denoted by H^{-1} and we have that

$$H^{-1} \text{ contains } L^2 \text{ contains } H^1_0$$

Space-Time Functional Spaces:

For time-dependent problems we have spaces that describe the two dependencies

For instance

$L^p(0,T; H^k(\Omega))$ is the space of functions f such that:

the norm $\|f\|_{H^k}(t) \in L^p(0,T)$ [in short: $L^p(H^k)$].

We will use $L^2(H^1_0)$ and $L^\infty(L^2)$

Embedding Theorems

There is a relation between belonging to H^k and L^p with $p \neq 2$.
Also, there is a relation between H^k and the traditional continuity.
These relations are summarized by a sequence of Theorems called "embedding theorems".

If $u \in H^s$ in $\Omega \subset \mathbb{R}^n$

$0 < 2s < n$: $H^s \subset L^q$ (continuous embedding) $\quad 1 \leq q \leq q^* = \frac{2n}{n-2s}$

$2s = n$: $H^s \subset L^q$ $\quad \forall q \quad 1 \leq q < +\infty$

$2s > n$: $H^s \subset C^0(\bar{\Omega})$

Also

$$H^s \subset C^m(\bar{\Omega}) \quad \text{if} \quad s > m + \frac{n}{2}$$

Notice the role of the number of space-dimensions!
This explains why some results of well-posedness (e.g. in fluid mechanics) are available only in 2D and not in 3D.

Question

Let u be a function in $H^1(\Omega)$ s.t. $u(\partial\Omega) = 1$

Can I define the space

$$H^1_1(\Omega) = \{ f \in H^1,\ f(\partial\Omega) = 1 \} ?$$

NO! It's not a space! In fact: $f_1 \in H^1_1$, $f_2 \in H^1_1 \Rightarrow f_1 + f_2 \notin H^1_1$.

(2) The Poisson Problem in 1D

We have already noted that the classical problem

$$-u'' = f \qquad x \in (0,1)$$
$$u(0) = u(1) = 0$$

is in fact the result of the minimization of

$$J = \frac{1}{2}\int_0^1 (u')^2 - \int_0^1 f u$$

By the way, we found that this corresponds to solving the problem:

$$\int_0^1 u'v' = \int_0^1 f v \qquad \forall v \text{ s.t. } v(0) = v(1) = 0.$$

Let's give a more rigorous formulation.
With the notation already introduced we have:

> find $u \in H^1_0(0,1)$: $\int_0^1 u'v'\,dx = \int_0^1 f v\,dx \qquad \forall v \in H^1_0(0,1)$
>
> where $f(x) \in L^2(0,1)$.

This is what we call the "weak" formulation of the problem.
Notice that:

→ the BC are encoded in the space, H^1_0

→ $f \in L^2$ is not the most general space, as a matter of fact we can take $f \in H^{-1}(0,1)$

⇒ each term is well defined:

$\int_0^1 u'v'\,dx$ — $u'v'$ is summable

$u' \in L^2$, $v' \in L^2$

$\int_0^1 f v$

$v \in H^1_0$

→ f can be a continuous-linear functional on H^1_0

⇒ We can prescribe $f = \delta(\frac{1}{2})$ for the problem depicted on the right

Is this problem well-posed?

Let's give a general formulation

$$\text{Find } u \in V : a(u,v) = f(v) \quad \forall v \in V \qquad (1)$$

where : (i) $a(\cdot,\cdot)$ is a bilinear form : $V \times V \to \mathbb{R}$

(ii) $f(\cdot)$ is a linear-continuous function on V : $V \to \mathbb{R}$.

THE LAX-MILGRAM LEMMA :

For the problem (1), let's assume :

(A) $a(\cdot,\cdot)$ continuous : $\exists M : |a(u,v)| \le M \|u\|_V \|v\|_V$

(B) $f(\cdot)$ " : $\exists C : |f(v)| \le C \|v\|_V$

and in particular :

(C) $\exists \gamma > 0 : a(u,u) \ge \gamma \|u\|_V^2 \quad \forall u \in V$

COERCIVITY OF THE BILINEAR FORM

(A), (B), (C) are sufficient to conclude that :

(I) u exists

(II) u is unique

(III) u depends continuously on the data

Conclusion (I) can be found in functional analysis books.

Conclusion (II) by contradiction :

if we have u_1, u_2 solving the same problem :

$$H_0' \ni u_1 - u_2 : \quad a(u_1 - u_2, v) = 0 \quad \forall v \in H_0'$$

Then : for $v = u_1 - u_2$:

$$\alpha \|u_1 - u_2\|_V^2 \le a(u_1 - u_2, u_1 - u_2) \le 0$$

$$\Downarrow$$

$$\|u_1 - u_2\|_V = 0 \implies u_1 - u_2 = 0 \quad \text{(contradiction)}$$

Also : We have for $v = u$:

$$\alpha \|u\|_V^2 \le a(u,u) \le C \|u\|_V \implies \|u\|_V \le \frac{C}{\alpha}$$

(stability bound for the solution)

(III) Let u_δ: $a(u_\delta, v) = f(v) + \delta f(v)$ → Continuous perturbation

$$\Downarrow$$

$$a(u_\delta - u, v) = \delta f(v) \qquad \forall v \in H^1_0(0,1)$$

For $v = u_\delta - u$:

$$\alpha \|u_\delta - u\|^2_V \le a(u_\delta - u, u_\delta - u) \le \delta C \|u_\delta - u\|_V$$

$$\Rightarrow \quad \|u_\delta - u\|_V \le \delta \frac{C}{\alpha}$$

so, for $\delta \to 0$, $\|u_\delta - u\|_V \to 0$.

For our problem, Lax-Milgram Lemma can be applied ⟹ EXERCISE

WARNING: The Lax-Milgram lemma is a SUFFICIENT (but not NECESSARY) condition. If it is not applicable, then the problem can be still well-posed.

LAX-MILGRAM Disease = to think that LM is also necessary ☺

The LM Lemma is a particular case of a more general result called BNB (Babuška, Nečas, Banach) Theorem (that, conversely, is necessary & sufficient.)

More General Boundary Conditions

(1) Non homogeneous Dirichlet conditions

$$u(0) = d_0 \,, \quad u(1) = d_1$$

As we noted, we cannot just introduce a set of functions fulfilling the B.C., because this set is not a space.

The simple solution is to introduce a LIFTING or EXTENSION.
Let $l(x)$ to be an arbitrary function of H^1 s.t.

$$l(0) = d_0 \,, \quad l(1) = d_1$$

Then, let $\mathring{u} = u - l$, so that:

$$\mathring{u}(0) = \mathring{u}(1) = 0$$

and

$$a(\mathring{u}, v) = \underbrace{f(v) - a(l, v)}_{\text{known}}$$

So, formally we can write the problem for $\mathring{u}$: find $\mathring{u} \in H^1_0$ s.t.

$$a(\mathring{u}, v) = f(v) - a(l, v) \qquad \forall v \in H^1_0$$

and then we reconstruct the solution $u = \mathring{u} + l$.
Since the solution is unique, we know that regardless of the specific lifting

the solution u will be the same.

How do we select the lifting function $\ell(x)$?

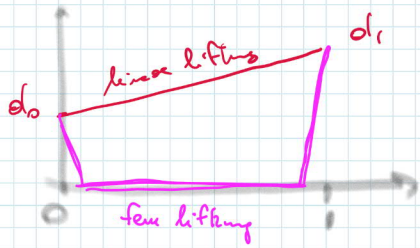
In 1D there is a natural choice, linear interpolation:

$$\ell(x) = d_0 + (d_1 - d_0)\,x \qquad x \in (0,1)$$

By the way, this is such that $\ell'' = 0$, so:

$$-u'' = -\mathring{u}'' - \ell'' = -\mathring{u}'' = f$$

For the numerical approximation, there are better choices.



We will see that using the piecewise polynomial lifting in the picture (magenta) the finite element method can manage the lifting automatically, with no additional effort.

(2) Neumann (or Natural) conditions

Let's assume that in $x = 1$ we have the boundary condition $u'(1) = r \in \mathbb{R}$

In the "strong" form we have:

$$-u'' = f \quad x \in (0,1), \quad u(0) = 0, \quad u'(1) = r$$

For the weak form, instead of using the minimization argument, we use a formal practical approach.

(a) multiply the equation by v: $-u''v = fv$

(b) integrate over the domain $(0,1)$: $-\int_0^1 u''v = \int_0^1 fv$

(c) integrate by parts:

$$-\left[u'v\right]_0^1 + \int_0^1 u'v' = \int_0^1 fv$$

$$-u'(1)v(1) + u'(0)v(0)$$

(d) manage the boundary conditions:

- take $v(0) = 0$ (like u vanishes in 0)
- notice that $u'(1) = r$

Weak form: let $H^1_D = \{ f \in H^1 \text{ s.t. } f(0) = 0 \}$:

find $u \in H^1_D$ s.t. $\int_0^1 u'v' = \int_0^1 fv + rv(1) \quad \forall v \in H^1_D$

Also in this case, the Lax-Milgram lemma can be applied and we have well posedness.

Notice that HOMOGENEOUS Neumann conditions ($r=0$) lead to:

$$\int_0^1 u'v' = \int_0^1 fv \qquad H^1_D$$

so these conditions do not require any modification of the functional and of the functional space either.

They are called "do-nothing" conditions.

REMARK

Let's consider this problem:

$$-u'' = f \qquad u'(0) = r_0 \qquad u'(1) = r_1$$

The weak formulation reads:

find $u \in H^1(0,1)$ s.t. $\int_0^1 u'v' = \int_0^1 fv + r_1 v(1) - r_0 v(0)$

Is this problem well posed?

- Direct answer: u occurs in the equation and the B.C. only under differentiation, so it can be replaced by u + constant and the equation is still the same $\Rightarrow$ NO UNIQUENESS.

- Functional analysis answer: if you test the coercivity of the bilinear form:

$$a(u,v) = \int_0^1 u'v'$$

notice that:

$$a(u,u) = \int_0^1 u'^2 = \|u'\|_{L^2}^2$$

The Poincaré inequality states that $\|u'\|$ is equivalent to $\|u\|_{H^1}$ in H^1_0 and H^1_D, not in H^1.

$\Rightarrow$ The Lax-Milgram lemma doesn't apply.

(3) Robin or Mixed Conditions:

Let's assume the condition:

$$u'(1) + \sigma u(1) = r \qquad (\text{and } u(0) = 0)$$

The weak formulation reads:

find $u \in H^1_D$ s.t.

$$\int_0^1 u'v' = \int_0^1 fv + \left[u'v\right]_0^1 =$$

$$= \int_0^1 fv + \left(2 - \sigma u(1)\right) v(1) \qquad \forall v \in H^1_D$$

($= 2$ because $v(0) = 0$)

or:

$$\int_0^1 u'v' + \sigma u(1) v(1) = \int_0^1 fv \qquad \forall v \in H^1_D.$$

(3) The Poisson Problem in 2D (or 2+D)

In this case, we noticed that the problem reads:

$$-\Delta u = f \text{ in } \Omega \qquad u(\partial\Omega) = 0$$

The weak formulation with our formal approach follows the same guidelines:

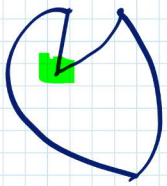
(1) $\quad -\Delta u\, v = f v \qquad$ with $v(\partial\Omega) = 0$

(2) $\quad -\int_\Omega \Delta u\, v = \int_\Omega fv$

(3) $\quad -\int_\Omega \Delta u\, v = -\int_{\partial\Omega} (\nabla u \cdot \vec{n}) v + \int_\Omega \nabla u \nabla v$

(4) $\Rightarrow$ Find $u \in H^1_0(\Omega)$ s.t. $\int_\Omega \nabla u \nabla v = \int_\Omega fv \quad \forall v \in H^1_0(\Omega)$

Now, the Lax-Milgram Lemma can be applied only with some constraints on Ω:

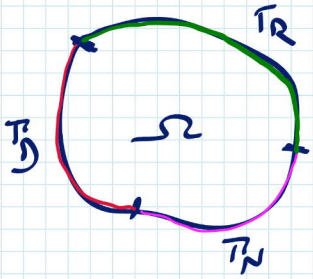


Concavities can undermine the continuity of the bilinear forms and the functionals.

In general, for polygonal domains, the well posedness is guaranteed.

We can consider more general sets of boundary conditions.

Let's consider a general problem with different boundary conditions.



$$-\Delta u = f \quad \text{in } \Omega$$
$$u(\partial\Omega) = g_D \quad \text{on } \Gamma_D$$
$$\nabla u \cdot \underline{n} = g_N \quad \text{on } \Gamma_N$$
$$\nabla u \cdot \underline{n} + \sigma u = g_R \quad \text{on } \Gamma_R$$

$$\partial\Omega = \overline{\Gamma_D \cup \Gamma_N \cup \Gamma_R}$$
$$\Gamma_N \cap \Gamma_D = \emptyset$$
$$\Gamma_N \cap \Gamma_R = \emptyset$$
$$\Gamma_R \cap \Gamma_D = \emptyset$$

(1^{st}) Let's introduce a lifting, i.e. a regular function l in Ω, s.t.

$$l(\partial\Omega) = g_D .$$

Then, define $\mathring{u} = u - l$:

$$-\Delta(\mathring{u} + l) = f \Rightarrow -\Delta\mathring{u} = f \quad \text{in } \Omega$$
$$\mathring{u}(\Gamma_D) = 0$$
$$\nabla\mathring{u} \cdot \underline{n} = g_N - \nabla l \cdot \underline{n} \quad \text{on } \Gamma_N$$
$$\nabla\mathring{u} \cdot \underline{n} + \sigma\mathring{u} = g_R - \nabla l \cdot \underline{n} - \sigma l$$

Then, with the usual procedure, notice that the Green formula here applies as :

$$-\int_\Omega \Delta u\, v = -\int_{\Gamma_D} \nabla u \cdot n\, v - \int_{\Gamma_N} \nabla u \cdot \underline{n}\, v - \int_{\Gamma_R} \nabla u \cdot \underline{n}\, v$$

So, we have : Find $\mathring{u} \in H^1_{\Gamma_D}(\Omega)$ $\left(= f \in H^1 \text{ s.t. } u(\Gamma_D) = 0 \right)$ s.t.

$$\int_\Omega \nabla\mathring{u} \cdot \nabla v + \int_{\Gamma_R} \sigma \mathring{u} v = \int_\Omega f v + \int_\Omega \Delta l\, v + \int_{\Gamma_N} (g_N - \nabla l \cdot \underline{n}) v + \int_{\Gamma_R} (g_R - \nabla l \cdot \underline{n} - \sigma l) v$$

Alternative RHS : $\int_\Omega f v - \int_\Omega \nabla l \cdot \nabla v + \int_{\Gamma_N} g_N v + \int_{\Gamma_R} (g_R - \sigma l) v \qquad \forall v \in H^1_{\Gamma_D}(\Omega)$

(if we decide to apply the Green formula to l too).

The analysis of this problem where we have several boundary conditions at the same time is not trivial $\Rightarrow$ An entire book by P. Grisvard is dedicated to this.

REMARK 1 : You may notice that the same problem has two different weak formulations depending on the treatment of the lifting.

Keep in mind that, in general, after numerical approximation they may lead to slightly different numerical results.

REMARK 2 More in general, it is important to stress that a weak formulation does not refer to an equation, but to the entire Boundary Value problem : DIFFERENT BOUNDARY CONDITIONS LEAD TO DIFFERENT WEAK

FORMULATIONS.

REMARK 3: In all the formulations above you can check that if $f \in L^2(\Omega)$ (or also to the dual of $H^1_{\Gamma_D}$), then the terms of the weak formulation are all well defined.

(4) More General Problems.

In general, let us consider the following problem:

$$-\nabla\cdot\left(\mu(x_1,x_2)\nabla u\right) + \underline{\beta}(x_1,x_2)\cdot\nabla u + \sigma(x_1,x_2)\,u = f \quad \text{in } \Omega$$

with
- $u(\Gamma_D) = 0$ (for non homogeneous data we need a lifting)
- $\mu(x_1,x_2)\nabla u = g_N$ on Γ_N
- $\mu(x_1,x_2)\nabla u + \gamma(x_1,x_2)\,u = g_R$ on Γ_R

Let's construct the weak formulation formally. Then, we will discuss the well posedness.

Green Formula
$$-\int_\Omega \nabla\cdot(\mu\nabla u)\,v = -\int_{\partial\Omega} \mu\nabla u\cdot\underline{n}\,v + \int_\Omega \mu\nabla u\nabla v$$

Find $u \in H^1_{\Gamma_D}(\Omega)$ s.t.

$$\int_\Omega \mu\nabla u\nabla v + \int_\Omega(\underline{\beta}\cdot\nabla u)v + \int_\Omega \sigma u v + \int_{\Gamma_R}\gamma u v = \int_\Omega fv + \int_{\Gamma_N} g_N v + \int_{\Gamma_R} g_R v$$

$\forall v \in H^1_{\Gamma_D}$

To analyze the well posedness, we need first to guarantee that each term makes sense ($=$ is $<+\infty$).

Let's assume $f \in L^2(\Omega)$

$$\int \mu\,\nabla u\nabla v$$

$\nabla u \in L^2$, $\nabla v \in L^2$, so $\nabla u\nabla v \in L^1$

Since $(\nabla u\nabla v) \in L^1$ ← Then we need $\mu \in L^\infty(\Omega)$

When we have
$$\int_\Omega \varphi\psi$$
with $\varphi \in L^p$, to have this term bounded we need $\psi \in L^q$ with $q = (1-\frac{1}{p})^{-1} = \frac{p}{p-1}$

If $\varphi \in L^1$, $\psi \in L^\infty$.

$\int_\Omega \underline{\beta} \cdot \nabla u \, v$ — $\nabla u \in L^2$, $v \in H^1$

↳ For the embedding theorem, $v \in L^p$ where p depends on the number of dimensions.

↓

In 2D and 3D $v \in L^6(\Omega)$, so we can assume $\underline{\beta} \in (L^3)^n$ (each component of the vector belongs to L^3).

$\int_\Omega \sigma \, u \, v$ — $u \in H^1$, $v \in H^1$, $uv \in L^3$

↳ We need $\sigma \in L^{\frac{3}{2}}(\Omega)$ at least.

$\int_{\Gamma_R} \gamma \, u \, v$ — $u \in H^{1/2}$ (traces), $v \in H^{1/2}$ (traces), $uv \in L^1$

→ we need $\gamma \in L^2(\Gamma_R)$

So, as you can see, the "collocation" in the most general spaces is not easy.

In general, in the applications, it's common to have:

$$\mu \in L^\infty(\Omega), \quad \underline{\beta} \in (L^\infty(\Omega))^n, \quad \sigma \in L^\infty(\Omega), \quad \gamma \in L^\infty(\Gamma_R)$$

So we are good to go, because $L^\infty(\Omega)$ is in all the spaces indicated above.

Lax-Milgram-based well posedness

Assuming that all the functions $\mu, \beta, \sigma, \gamma$ and f are regular enough, so the bilinear forms and the functionals are all continuous,

UNDER WHAT CONDITIONS, the bilinear form:

$$a(u,v) = \int_\Omega \mu \nabla u \nabla v + \int_\Omega (\underline{\beta} \cdot \nabla u) v + \int_\Omega \sigma u v + \int_{\Gamma_R} \gamma u v$$

is COERCIVE?

If we limit for simplicity to the case $\mu > 0$, $\sigma \geq 0$, $\gamma \geq 0$ it is immediately realized that the 1st, 3rd, 4th of the bilinear form

$a(u,u)$ are positive.

We need to focus on:

$$\int_\Omega (\underline{\beta}\cdot\nabla u)u \qquad \left(\text{Einstein notation: } \int_\Omega \beta_i u_{j,i} u_j\right)$$

Notice that we can write:

$$\frac{1}{2}\int_\Omega \underline{\beta}\cdot\nabla(u^2) \overset{\text{Green}}{=\!=\!=} \frac{1}{2}\int_{\partial\Omega} \underline{\beta}\cdot\underline{n}\,u^2 - \frac{1}{2}\int_\Omega (\nabla\cdot\underline{\beta})u^2$$

This term for generic boundary conditions is complicated.

For Dirichlet conditions ($u(\partial\Omega)=0$), the first term is 0 and we conclude that

$$\text{if } (\nabla\cdot\underline{\beta}) \le 0 \implies \int_\Omega (\underline{\beta}\cdot\nabla u)u \ge 0.$$

As you may notice, the Lax-Milgram analysis gets complicated!!!

REMARK

In the construction of the weak formulation, we can integrate by parts also the term:

$$\int_\Omega \underline{\beta}\cdot\nabla u\, v = \int_{\partial\Omega} \underline{\beta}\cdot\underline{n}\,uv - \int_\Omega \nabla\cdot(\underline{\beta}v)u =$$

$$= \int_{\partial\Omega} \underline{\beta}\cdot\underline{n}\,uv - \int_\Omega (\nabla\cdot\underline{\beta})vu - \int_\Omega \underline{\beta}\cdot\nabla v\, u$$

$v(\Gamma_0)=0$ ↑ so this reduces to $\Gamma_N\cup\Gamma_R$

Since the term on the boundary changes, this approach makes sense for a different boundary condition, for instance:

$$\mu\nabla u\cdot n + \underline{\beta}\cdot\underline{n}\,u = \tilde{d}_N \qquad \text{on } \Gamma_N$$

or

$$\mu\nabla u\cdot\underline{n} + \underline{\beta}\cdot\underline{n}\,u + \gamma u = \tilde{d}_R \qquad \text{on } \Gamma_R$$

} in some applications, these are, in fact, the physical conditions to prescribe.

Recalling the Remark a couple of pages back, FOR THE SAME EQUATION WE MAY HAVE DIFFERENT WEAK FORMULATIONS SUITABLE FOR DIFFERENT BOUNDARY CONDITIONS.

Week 7

Quarteroni: Chap. 3
QV - Chap. 5

The Galerkin Method

In this Chapter, we have a general view and introduce the Galerkin method for the generic problem:

$$\text{find } u \in V: \quad a(u,v) = f(v) \qquad \forall v \in V$$

where $a(\cdot,\cdot)$ is a bilinear form: $V \times V \longrightarrow \mathbb{R}$

$f(\cdot)$ a linear and continuous functional: $V \longrightarrow \mathbb{R}$

We assume in general that V is an infinite dimensional space.

In the FD method, our approximation was based on the idea of approximating the differential operators (with incremental quotients), after "collocating" the problem in special points.

With the Galerkin method we change completely our perspective.

The approximation is not in the differential operators, but in the search space.

Let V_h be a FINITE DIMENSIONAL SUBSPACE of V, where each function v_h can be represented as the linear combination of BASIS FUNCTIONS:

$$v_h = \sum_{j=1}^{N_h} c_j \varphi_j$$

where N_j is the DIMENSION of V_h.

Then, we construct the APPROXIMATE PROBLEM

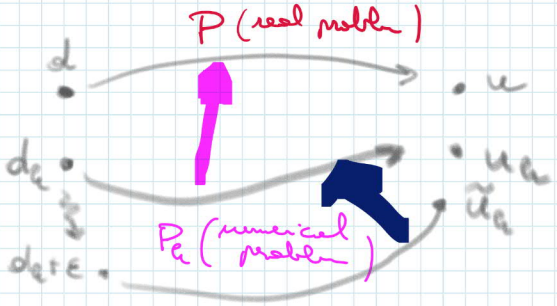
$$\text{find } u_h \in V_h \text{ s.t. } \quad a(u_h, v_h) = f(v_h) \qquad \forall v_h \in V_h$$

This is the Galerkin approximation of our problem.

Before we look into the specific construction of V_h, we can state some general properties of this approach.

You recall that for an approximation of a differential problem we introduced the concepts of

- CONSISTENCY $P_h \to P$ as $h \to 0$
- STABILITY $\tilde{u}_h \to u_h$ as $\epsilon \to 0$



For the consistency, if we construct $V_h \subset V$ in such a way that $V_h \xrightarrow[h\to 0]{} V$, we are ok.

But we can characterize the consistency error in a better way:

in general, if we pretend the exact solution to fulfill the numerical problem, we have a RESIDUAL:

$$\left.\begin{aligned} P(u;d) &= 0 \\ P_h(u_h;d_h) &= 0 \quad \text{(in exact arithmetic)} \end{aligned}\right] \; P_h(u,d_h) \neq 0$$

$\Downarrow$ Truncation error

You may recall that CONSISTENCY means that the truncation error vanishes with h.

EXAMPLE: $\dfrac{d(e^x)}{dx} = e^x$

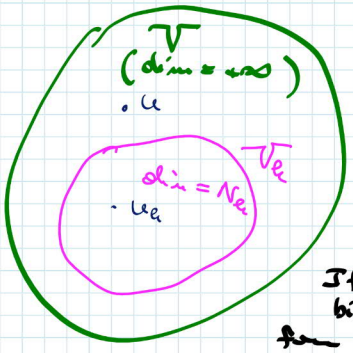
$$\frac{e^{x+h} - e^x}{h} \neq e^x \quad \text{but} \quad e^{x+h} = e^x + e^x h + \frac{e^x}{2} h^2 + \dots$$

error: $\dfrac{e^x}{2} h \xrightarrow[h\to 0]{} 0$

The Galerkin method is more than consistent, it is STRONGLY CONSISTENT: the truncation error is 0.

P: $u \in V$: $a(u,v) = f(v) \quad \forall v \in V$

P_h: $u_h \in V_h \subset V$: $a(u_h, v_h) = f(v_h) \quad \forall v_h \in V_h$



If we test P_h with u, we have (for $V_h \subset V$):

$$a(u, v_h) = f(v_h)$$

because $v_h \in V$

If the bilinear form is already in V, it is in V_h as well

So we have STRONG CONSISTENCY and:

$$\boxed{a(u - u_h, v_h) = 0} \qquad \boxed{\forall v_h \in V_h.}$$

This is a fundamental relation.

Let's recall a definition:

a bilinear form is SYMMETRIC if:

$$a(u,v) = a(v,u) \quad \forall u, v \in V$$

Now, let's assume that our problem P (find $u \in V$: $a(u,v) = f(v) \;\forall v \in V$) fulfills under the LAX-MILGRAM lemma.

Under this assumption, it is immediate to prove that also P_h is well posed! ☆

Now, we have

$$|a(u,v)| \leq M \|u\|_V \|v\|_V$$

$$a(u,u) \geq \alpha \|u\|_V^2 \Rightarrow \text{Notice that } a(u,u) \leq M\|u\|_V^2$$

In addition, let's assume $a(u,v)$ to be <u>Symmetric</u>.

Then, we can state that $a(\cdot,\cdot)$ defines a <u>scalar product</u> having all the features to be so.

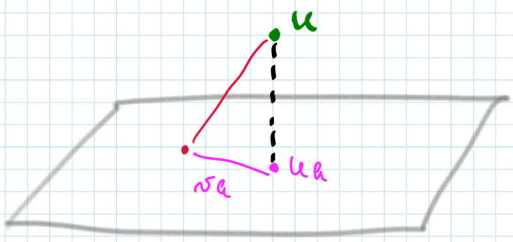
So, we can define a norm too,

$$\|u\|_a \equiv \left(a(u,u)\right)^{1/2}$$

and this norm is equivalent to the norm in V, because

$$\alpha \|u\|_V^2 \leq a(u,u) \leq M \|u\|_V^2$$

The relation $a(u-u_h, v_h) = 0$ is therefore an <u>orthogonality</u> relation this states that for the topology induced by $a(\cdot,\cdot)$, u_h is the "best" approximation of u in V_h:



$$a(u-u_h, v_h) = 0 \Downarrow u-u_h \perp V_h$$

- The strong consistency is independent of the symmetry.
- The geometrical interpretation (highly suggestive) holds only for symmetric forms

However, the strong consistency leads to an important result.

$$\alpha \|u-u_h\|_V^2 \leq a(u-u_h, u-u_h) = a(u-u_h, u-w_h+w_h-u_h) =$$

($u-u_h \in V$, $u-u_h \in V$; $\forall w_h \in V_h$)

$$= a(u-u_h, u-w_h) + a(u-u_h, w_h-u_h)$$

($w_h - u_h \in V_h$; $a(u-u_h, w_h-u_h) = 0$ for strong consistency)

$$\leq M \|u-u_h\| \|u-w_h\|$$

$$\Downarrow$$

$$\|u-u_h\|_V \leq \frac{M}{\alpha} \|u-w_h\|_V$$

Since u_h is arbitrary, we can take u_h to be the "best" approximation of u in V_h in the topology of V:

$$u_h: \quad \|u - u_h\|_V \leq \|u - v_h\|_V \qquad \forall v_h \in V_h$$

or

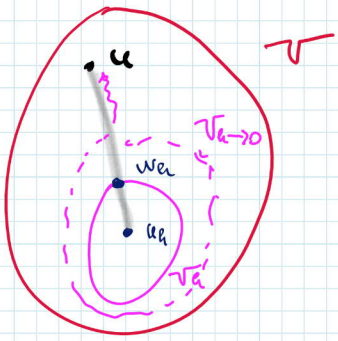
$$u_h: \quad \inf_{v_h \in V_h} \|u - v_h\|_V = \|u - u_h\|_V$$

So we have that:

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V$$

This is called **Cèa Lemma.**

The Cèa Lemma states that the Galerkin solution is not necessarily the best one, but converges to the exact one with the same rate of the best approximation.



If $a(\cdot,\cdot)$ is symmetric, u_h is the "best" in the norm $\|\cdot\|_a$ and it is possible to prove that this norm is equivalent to the norm in V

$$u_h: \quad \|u - u_h\|_a = \inf_{\substack{\forall v_h \\ \in V_h}} \|u - v_h\|_a$$

In general: $\|u - u_h\|_V \leq C \inf_{v_h} \|u - v_h\|_V$

(2) On Possible Choices for the space V_h.

We can now become more specific and specify some possible choices for V_h.

In general, we can construct the solution as a polynomial function.

If we do this, the theory of polynomial approximation (MATH516) can help understanding the term

$$\inf_{v_h \in V_h} \|u - u_h\|_V$$

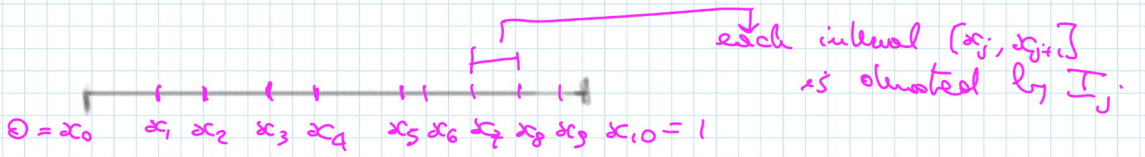
One possible option is using Gaussian Interpolation theory to construct the function u_h. As well known from Interpolation theory, the collocation of the interpolation nodes is critical for the convergence of the interpolation. Gaussian interpolation is an option leading to a class of methods called SPECTRAL METHODS.

The book of Langtangen-Mardal "Introduction to Numerical Methods for

variational problems treat this topic in Chapter 2 and 4.
Here, we follow a different approach (we will explain why later on).
We select V_h in this way. Let's start with 1D problems.

The Finite Element Method in 1D.

(1) We take the interval ($[0,1]$ or $[a,b]$ in general) and split it into subintervals



Notice that the distance $|x_{j+1} - x_j|$ does not need to be constant.
We set

$$h = \max_j |x_{j+1} - x_j|$$

T_h denotes the set of nodes $\{x_j\}$

For simplicity, we will first assume $|x_{j+1} - x_j| = h \quad \forall j$
(uniform reticulation)

(2) We construct V_h as the space of continuous piecewise polynomial functions with degree p on the reticulation:

$$V_h(T_h) \equiv \left\{ v_h \in C^0([0,1]) \text{ s.t. } v_h(I_j) \in \mathbb{P}^p \right\}$$

(3) We represent each function v_h with a Lagrange piecewise polynomial basis:

$$v_h = \sum_{j=1}^{N_h} c_j \varphi_j(x)$$

where

$$\varphi_j(x) \in V_h \quad \text{and} \quad \varphi_j(x_i) = \delta_{ij}$$

We will see examples very soon.

(4) How do we use all of this?
We want to solve

$$\text{find } u \in V_h : a(u_h, v_h) = \ell(v_h) \quad \forall v_h \in V_h$$

(4.1) Do we need to test against all the functions $v_h \in V_h$?

No! Clearly, if a generic function is:

$$v_h = \sum v_j \cdot \varphi_j(x) \qquad (*)$$

it is enough to use only the functions φ_j:

$$\text{Find } u_h: \ a(u_h, \varphi_i) = f(\varphi_i) \quad \forall i = 1, \dots N_h \qquad (\square)$$

In fact, all the other equations are a linear combination of these, thanks to the linearity of the arguments in v_h and the $(*)$.

(4.2) But what is $(\square)$ from a practical point of view?

Also the solution u_h reads:

$$u_h = \sum u_j \varphi_j(x)$$

so that we obtain:

$$\text{Find } \underline{u} = [u_j] \ \text{ s.t. } \ \sum_{j=1 \dots N_h} u_j a(\varphi_j, \varphi_i) = f(\varphi_i) \quad \forall i = 1 \dots N_h$$

Now, set

$$A = \left[a_{ij} = a(\varphi_j, \varphi_i) \right] \quad N_h \times N_h \text{ matrix (called STIFFNESS MATRIX)}$$

$$\underline{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_{N_h} \end{bmatrix} \qquad \underline{b} = \begin{bmatrix} f(\varphi_1) \\ \vdots \\ f(\varphi_{N_h}) \end{bmatrix} \text{ (called LOAD VECTOR)}$$

Then, it is easily verified that we have a linear system:

$$A\underline{u} = \underline{b}$$

Problem $(\square)$ is just a linear system: we know how to solve it!

What are the features of this linear system?

① First of all, we already proved that the problem P_E is well posed, so we argue that the matrix A is non-singular.

Also notice that the coercivity implies:

$$a(u_h, u_h) \geq \alpha \|u_h\|$$

Also, by contradiction, assume that $\exists\, \lambda = 0$ eigenvalue of A (so A is singular) then let $\underline{u}_0$ be the associated eigenvector and $u_{h0} = \sum u_{j0} \varphi_j$.

Then

$$\alpha \|u_{h0}\|^2 \leq a(u_{h0}, u_{h0}) = \underline{u}_0^T A \underline{u}_0 = \underline{u}_0^T 0 = 0$$

(coercivity) (100 equivalent)

CONTRADICTION: all $\lambda \neq 0 \Rightarrow A$ is not singular.

② Notice that if $a(\cdot,\cdot)$ is symmetric, it follows that

$$A \text{ is } s.p.d.$$

③ On the sparsity pattern of A we'll discuss in the description of specific finite elements.

(1) $\Rightarrow$ the linear system is non-singular

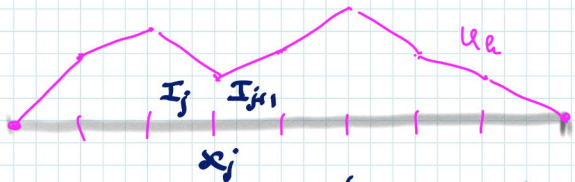
(2) $\Rightarrow$ CG or Cholesky can be used

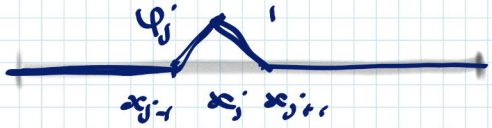
The question to answer is the CONVERGENCE RATE.

However, before this, let's see some particular examples.

Linear Finite Elements

In this case, we approximate the solution with a continuous piecewise linear function. We call this case as P1.





The Lagrange basis functions take the form:

$$\varphi_j(x) = \begin{cases} 0 & x \notin I_j \cup I_{j+1} = [x_{j-1}, x_{j+1}] \\ \dfrac{x - x_{j-1}}{x_j - x_{j-1}} & x \in I_j = [x_{j-1}, x_j] \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j} & x \in I_{j+1} = [x_{j+1}, x_j] \end{cases}$$

(if the mesh is uniform $x_j - x_{j-1} = x_{j+1} - x_j = h$)

Notice that:

(1) φ_j are non zero only on two elements (or intervals)
We say that φ_j has a small support

(2) the coefficients u_j in $u_h = \sum u_j \varphi_j(x)$ have the meaning of:

$$u_j = u_h(x_j)$$

This is the consequence of having chosen a Lagrange basis function.

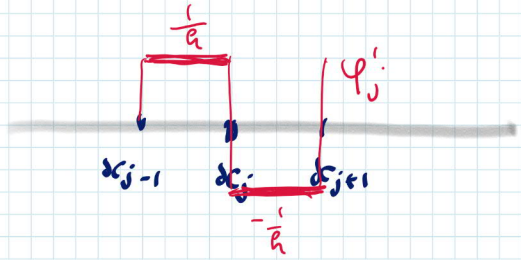
Using Lagrange polynomials ($\varphi_j(x_i) = \delta_{ij}$ (Kronecker delta)) is not the only possible choice. We will see this later.

Property (1) has an important consequence:

Let's look at the entries of A:

$a(\varphi_j, \varphi_i)$ in the Laplace operator is

$\int_0^1 \varphi_j' \varphi_i'$ $\Rightarrow$ if $j \neq i, i+1, i-1$ this integral is automatically 0, because the support of φ_j' and the one of φ_i' do not intersect.



$\Downarrow$

$a_{ij} = 0$ if $j \neq i, i+1, i-1$

$\Downarrow$

A is tridiagonal
(like for FD)

$$A = \begin{bmatrix} & & 0 \\ & & \\ 0 & & \end{bmatrix}$$

A note on the number of degrees of freedom (DOF):

In P1 we have a coefficient for each vertex:

If we have N intervals, we have $N+1$ nodes ($N_h = N+1$) but if we have 2 Dirichlet conditions, the real size of the system reduces to $N_h - 2$ ($= N-1$) equations (in theory ... practice is different).

If we adopt an elementwise perspective,

- On each element we want a linear function
- A linear function is specified by 2 coefficients

→ locally, we have 2 DOF needed to allow the computation of the unique function.

Choosing "Lagrange" interpolation means that we use the 2 DOF to prescribe the values of the function in two nodes (if we knew the function u, we would say that we "interpolate" u in the 2 nodes)

Choosing the vertexes guarantees the continuity of the functions

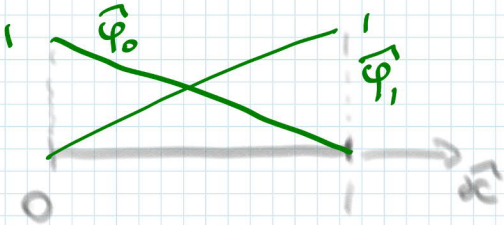
→ This is important because this guarantees that $V_h \equiv P1 \subset H^1$ ("conformal" finite element).

The Reference Element

With the element-wise perspective, we may introduce an auxiliary tool.

Let's assume for a while that we have only 1 element (and also not link to the B.C. for now).

We write this in the reference variable $\hat{x}$:



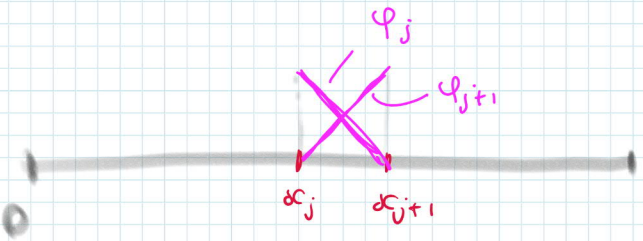
Here we have two basis functions

$$\hat{\varphi}_0 = 1 - \hat{x}$$

$$\hat{\varphi}_1 = \hat{x}$$

On this element, $A_{loc} = \begin{bmatrix} a(\hat{\varphi}_0, \hat{\varphi}_0) & a(\hat{\varphi}_1, \hat{\varphi}_0) \\ a(\hat{\varphi}_0, \hat{\varphi}_1) & a(\hat{\varphi}_1, \hat{\varphi}_1) \end{bmatrix}$

Notice that we have the same situation on each element of a generic reticulation:



$$\varphi_j(x) = \hat{\varphi}_0(\hat{x})$$

where $x = x_j + \hat{x}(x_{j+1} - x_j)$

$$\left(\hat{x} = \frac{x - x_j}{x_{j+1} - x_j}\right)$$

$$\varphi_{j+1}(x) = \hat{\varphi}_1(\hat{x})$$

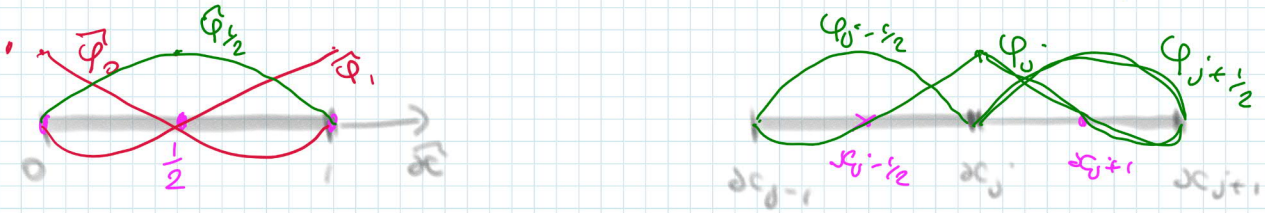
While in 1D this use of the Reference Finite Element is redundant, it is critical in multidimension.

Quadratic Finite Elements

Let's adopt the elementwise perspective.

On each interval, we want to approximate the solution with a quadratic function:

$$V_h = \mathbb{P}2 = \left\{ v_h \in C^0([0,1]) \text{ s.t. } v_h(I_j) \in \mathbb{P}^2 \right\}$$



We need 3 DOF on each interval to be unisolvent (= uniquely find the function).

We use three Lagrangian DOF, so that the coefficient of the expansion is the value of the solution in the point

We choose as interpolation points:

- the vertices (so we have continuity between two intervals)
- the mid-point (for symmetry and to have better numerical properties)

The corresponding Lagrange polynomials are depicted in the figures:

The total number of DOF (3 per interval but the ones in the vertices coincide on pairs of intervals)

$$N_h = 3 \times N - (N-1) = 2N+1$$

($N-1$: # vertices)

or: 1 dof per vertex, 1 dof per interval

$$N_h = \overbrace{N+1}^{\text{vertices}} + \overbrace{N}^{\text{intervals}} = 2N+1$$

If we include the Dirichlet B.C.: $N_h = 2N-1$

$$\hat{\varphi}_0 = \frac{\left(\hat{x}-\frac{1}{2}\right)(\hat{x}-1)}{\frac{1}{2}} = 2\left(\hat{x}-\frac{1}{2}\right)(\hat{x}-1) = 2\left(\frac{1}{2}-\hat{x}\right)(1-\hat{x})$$

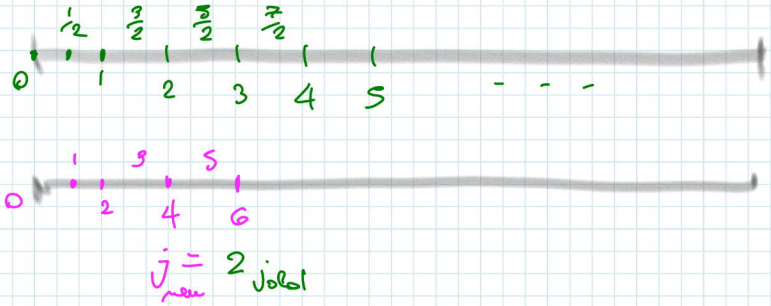
$$\hat{\varphi}_{1/2} = \frac{\hat{x}\,(\hat{x}-1)}{-\frac{1}{4}} = 4\,\hat{x}\,(1-\hat{x})$$

$$\hat{\varphi}_1 = \frac{\hat{x}\left(x-\frac{1}{2}\right)}{-\frac{1}{2}} = 2\,\hat{x}\left(\frac{1}{2}-\hat{x}\right)$$

Then we can φ_j as the map with $x = x_j + \hat{x}\,(x_{j+1} - x_j)$

The function $\hat{\varphi}_{1/2}$ (and $\varphi_{j\pm 1/2}$) are called "bubbles" because they vanish at the boundary.

Notice that in this case, if we re-number the degrees of freedom:



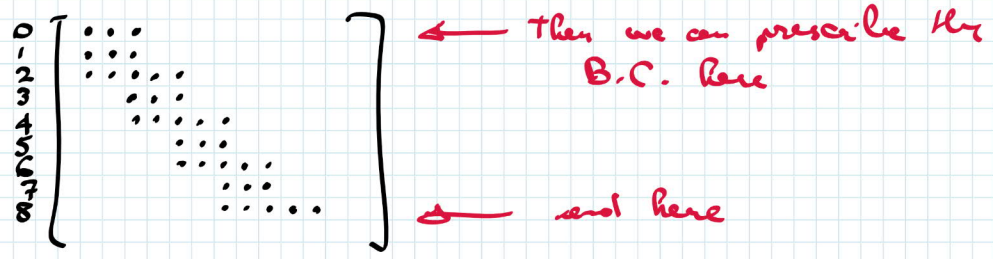
we have that:

- the support of φ_{2k} is $I_k \cup I_{k+1}$
- the support of φ_{2k+1} is I_{k+1}

Consequently:

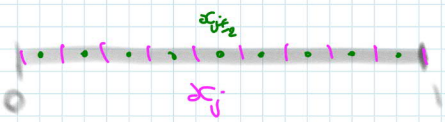
$$a(\varphi_j, \varphi_i) = \begin{cases} i \text{ even}: & 0 \text{ for } j \neq i, i\pm 1, i\pm 2 \\ i \text{ odd}: & 0 \text{ for } j \neq i, i\pm 1 \end{cases}$$

In this case, the matrix is penta-diagonal, alternating 5 and 3 non zero entries



Different Choices for the Basis

For what we have done so far, if we compute the solution with P1 and then we want to recompute the solution with P2, we have to recompute the matrix



Assuming to have two Dirichlet conditions, we have and N = # of intervals

P1 : $N_q = N-1$ dof

P2 : $N_q = 2N-1$ dof

Size of matrices :

A_{P_1} : $(N-1)\times(N-1)$ with $(N-1)+2(N-2) = 3N-5$ nonzero entries

A_{P_2} : $(2N-1)\times(2N-1)$ with $3N+5(N-1) = 8N-5$ nonzero entries

TASK: construct a basis for P2 "prone to recycling" so that A_{P_1} is a submatrix of A_{P_2}.

In the reference element, we can take :

$\hat{\varphi}_0 = 1-\hat{x}$, $\hat{\varphi}_1 = \hat{x}$ } same as P_1 $\qquad \hat{\varphi}_{\frac{1}{2}} = \hat{\varphi}_0\,\hat{\varphi}_1$

This basis is unisolvent, any quadratic function can be written as :

$$p_2(\hat{x}) = c_0\hat{\varphi}_0 + c_1\hat{\varphi}_1 + c_{\frac{1}{2}}\hat{\varphi}_{\frac{1}{2}}$$

where $c_0 = p_2(0)$, $c_1 = p_2(1)$

while $c_{\frac{1}{2}}$ has no precise physical meaning.

for instance : $p_2 = \hat{x}^2 + \hat{x} + 1 =$

$$= 1-\hat{x} + 3\hat{x} - (1-\hat{x})\hat{x}$$

or in general : $p_2 = \alpha\hat{x}^2 + \beta\hat{x} + \gamma =$

$$= \gamma(1-\hat{x}) + (\alpha+\beta+\gamma)\hat{x} - \alpha(1-\hat{x})\hat{x}$$

When we construct the matrix, we may write:

$$A = \begin{bmatrix} a(\hat{\varphi}_0, \hat{\varphi}_0) & a(\hat{\varphi}_1, \hat{\varphi}_0) & a(\hat{\varphi}_{1/2}, \hat{\varphi}_0) \\ a(\hat{\varphi}_0, \hat{\varphi}_1) & a(\hat{\varphi}_1, \hat{\varphi}_1) & a(\hat{\varphi}_{1/2}, \hat{\varphi}_1) \\ a(\hat{\varphi}_0, \hat{\varphi}_{1/2}) & a(\hat{\varphi}_1, \hat{\varphi}_{1/2}) & a(\hat{\varphi}_{1/2}, \hat{\varphi}_{1/2}) \end{bmatrix}$$

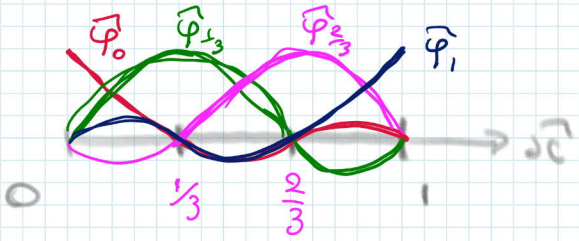
The submatrix highlighted in yellow is the P1 matrix !!!

We have a "nested" structure that may be helpful in case of an automatic selection of the degree (p-adaptivity).

This hierarchical basis is just an example of non-Lagrangian basis and it is available in FEM1D (Advanced Mode).

Cubic Finite Elements

We can proceed rapidly now



Lagrangian DOF:

value of the function in:

$0, \frac{1}{3}, \frac{2}{3}, 1$

$$\hat{\varphi}_0 = \frac{(\hat{x}-\frac{1}{3})(\hat{x}-\frac{2}{3})(\hat{x}-1)}{(-\frac{1}{3})(-\frac{2}{3})(-1)}$$

$$\hat{\varphi}_1 = \frac{\hat{x}(\hat{x}-\frac{1}{3})(\hat{x}-\frac{2}{3})}{1(1-\frac{1}{3})(1-\frac{2}{3})}$$

$$\hat{\varphi}_{1/3} = \frac{\hat{x}(\hat{x}-\frac{2}{3})(\hat{x}-1)}{\frac{1}{3}(\frac{1}{3}-\frac{2}{3})(\frac{1}{3}-1)}$$

$$\hat{\varphi}_{2/3} = \frac{\hat{x}(\hat{x}-\frac{1}{3})(\hat{x}-1)}{\frac{2}{3}(\frac{2}{3}-\frac{1}{3})(\frac{2}{3}-1)}$$

$$V_h = P_3 \equiv \{ v_h \in C^0([0,1]) : v_h(I_j) \in P^3 \ \forall I_j \}$$

Number of DOFs (N = # intervals)

$4 \times N$ − $(N-1)$ [−2 for b.c.] =

($4 \times N$: dof per interval; $(N-1)$: continuity at the vertices)

$= 3N+1$ [−2 for b.c.]

or

$N+1$ vertices + $2 \times N$ (2 internal dof per interval) [−2 for b.c.]

$= 3N+1$ [−2 for b.c.]

You can verify that with the proper numbering, the pattern is a sequence with 7, 4, 4 non zero entries. (7 for rows with an index multiple of 3 and 4 for the others).

Convergence Rate Theory

The starting point is the Céa lemma:

$$\|u - u_h\|_V \leq C \inf_{w_h \in V_h} \|u - w_h\|_V .$$

Let's use the interpolation theory for piecewise polynomial interpolation to quantify the right hand side.

Theorem

Let's assume $f \in H^{s+1}([a,b])$ $s \geq 1$ and let $\{x_i\}$ a reticulation of nodes such that $x_i \neq x_j$ for $i \neq j$ and set $h = \max |x_{i+1} - x_i|$.

Then, if f_h is the piecewise polynomial interpolation of order p (i.e., f_h:

$$f_h(x_h) = f(x_h)$$ in the d.o.f. (vertices for $p=1$, vertices and element dof for $p>1$)

$$f_h(I_j) \in \mathbb{P}^p(I_j)$$)

Then, $\exists\, C > 0$ s.t.

$$\|f - f_h\|_{H^1(a,b)} \leq C h^r \left\| \frac{\partial^{(s)} f}{\partial x^{(s)}} \right\|_{L^2(a,b)}$$

with $r = \min(p, s)$

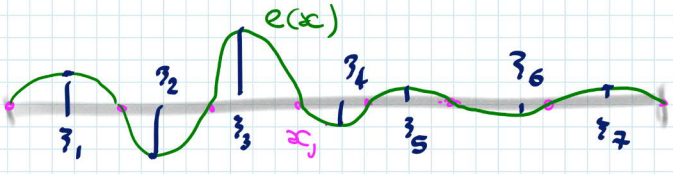
Also

$$\|f - f_h\|_{L^2(a,b)} \leq C h^{r+1} \left\| \frac{\partial^{(s)} f}{\partial x^{(s)}} \right\|_{L^2(a,b)} .$$

PROOF

We consider only the case $f \in H^2$, $p = 1$ (so $s = 1$).

$f_h(x_i) = f(x_i)$ means that $e(x) \equiv f(x) - f_h(x)$ has $N+1$ zeros (where N is the number of intervals).

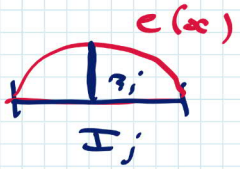


The Rolle's theorem implies that in each interval I_j there exists at least 1 point ξ_j s.t.

$$e'(\xi_j) = 0$$

Also, notice that in each interval $f_h'' = 0$ so:

$$e''(x) = f''(x) - 0 = f''(x)$$



Then:

$$e'(x) = \int_{z_j}^{x} e''(\sigma)\,d\sigma \quad \text{for } x \in I_j \text{ (in fact, } e'(z_j) = 0 \text{ with this formula)}$$

$$|e'(x)| \le \left(\int_{z_j}^{x} 1\right)^{1/2}\left(\int_{z_j}^{x} (e'')^2\right)^{1/2} \le h^{1/2}\,\|e''\|_{L^2(I_j)}$$

$z_j \mapsto x \le I_j$

$$|e'(x)|^2 \le h\,\|e''\|^2_{L^2(I_j)} \qquad x \in I_j$$

$$\|e'\|^2_{L^2(I_j)} = \int_{I_j} |e'(x)|^2\,dx \le h\,\|e''\|^2_{L^2} \int_{I_j} 1\,dx = h^2\,\|e''\|^2_{L^2(I_j)}$$

$$\Rightarrow \quad \|e'(x)\|^2_{L^2(a,b)} = \sum_j \|e'\|^2_{L^2(I_j)} \le h^2 \sum_j \|e''\|^2_{L^2(I_j)} = h^2\,\|e''\|^2_{L^2(a,b)}$$

i.e.,

$$\boxed{\|e'\|_{L^2(a,b)} \le h\,\|e''\|_{L^2(a,b)} = h\,\|f''\|_{L^2(a,b)}} \qquad (*)$$

Now, we have:

$$|e(x)| = \left|\int_{x_j}^{x} e'(x)\,dx\right| \le \left|\int_{x_j}^{x} 1\left(\int_{x_j}^{x} (e')^2\right)^{1/2}\right| \le h^{1/2}\,\|e'\|_{L^2(I_j)}$$

$e(x_j) = 0$

$$\|e\|^2_{L^2(I_j)} = \int_{I_j} e^2\,dx \le h\,\|e'\|^2_{L^2(I_j)} \int_{I_j} 1\,dx = h^2\,\|e'\|^2_{L^2(I_j)}$$

$$\Rightarrow \quad \|e\|^2_{L^2(a,b)} = \sum_j \|e\|^2_{L^2(I_j)} \le h^2 \sum_j \|e'\|^2_{L^2(I_j)} = h^2\,\|e'\|^2_{L^2(a,b)} \le C h^4\,\|f''\|^2_{L^2(a,b)}$$

i.e.

$$\boxed{\|e\|_{L^2(a,b)} \le C h^2\,\|f''\|_{L^2(a,b)}} \qquad (**)$$

$$\left(I + II\right) \Rightarrow \quad \|e\|_{H^1}^2 = \|e\|_{L^2}^2 + \|e'\|_{L^2}^2 \leq C\left(h^4 + h^2\right) \|f''\|_{L^2}^2$$

$$\Downarrow$$

$$\|e\|_{H^1} \leq \tilde{C} h \|f''\|_{L^2}$$

Notice that $\|f''\|_{L^2(a,b)}$ is also called the semi-norm H^2 (like a norm without the property $|u|_{H^2} = 0 \in \mathbb{R} \Rightarrow u = 0$)

$$\|e\|_{H^1} \leq \tilde{C} h |f|_{H^2}.$$

Solution regularity (H^{s+1})

Fe degree

p \ s	0	1	2	3	0	1	2	3
1	Conv	1	1	1	Conv	2	2	2
2	Conv	1	2	2	Conv	2	3	3
3	Conv	1	2	3	Conv	2	3	4

H^1 $\qquad$ L^2

This table provides the guide to the selection of the most appropriate FE degree as a function of the regularity.

In fact, from the Céa lemma:

$$\|u - u_h\|_{H^1} \leq C \inf_{w_h \in V_h} \|u - w_h\|_{H^1} \leq C \|u - u_h^*\|_{H^1}$$

where u_h^* is the interpolant of u on the dof.

From here, we notice that selecting high-order FE pays only for regular solutions, otherwise the convergence rate is the same of low order FE.

So, the boxes outlined in green are somehow "optimal" choices.

A last piece of information is given by the following Theorem.

Theorem

The condition number of the stiffness matrix scales with h^{-2}, i.e.:

$$\text{cond}(A) \propto h^{-2}$$

■

This means that when we refine the mesh $h \to \frac{h}{2}$ we expect $\text{cond}(A)$ to be multiplied by a factor of 4, regardless the degree of FE.

A FINAL REMARK

If we want to increase the quality of a FE solution, we have two options.

h- strategy : REFINE THE MESH

p - strategy : INCREASE THE DEGREE [Hierarchical Basis very useful here]

The second strategy works only for regular solutions

The first one increases the computational costs as stated by this last theorem.

The two strategies can be

- automatized (ADAPTIVITY : we need an a-posteriori error estimator)
- combined (hp - adaptivity)

If you are interested in this topic, see Quarteroni Chap. 4

Talking about ADAPTIVITY, M. Fortin (great mathematician) stated that a mesh should not be data but an "unknown" of a FE solution.

Week 8

Finite Elements in 2+D for Elliptic Problems

We want to extend our procedure to the 2+D case.
The backbone of our approach is the same.

Find u : $\quad a(u, v) = f(v) \qquad \forall v \in V$

We introduce $V_h \subset V$ with $\dim(V_h) = N_h < +\infty$
and assume that

$$u_h = \sum_{j=1}^{N_h} c_j \varphi_j(\underline{x}) \qquad \underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

Then, the Galerkin formulation reads:

$$\boxed{\text{Find } u_h \in V_h \qquad a(u_h, v_h) = f(v_h) \qquad \forall v_h \in V_h}$$

and this is equivalent to

$$\text{Find } \underline{c} = [c_j] : \quad \sum_{j=1}^{N_h} a(\varphi_j, \varphi_i) c_j = f(\varphi_i) \qquad i = 1 \dots N_h$$

In its algebraic form this problem reads:

$$A\underline{c} = \underline{b}$$

A : $[a(\varphi_j, \varphi_i)]$ (stiffness matrix)

$\underline{b}$: $[f(\varphi_i)]$ (load vector)

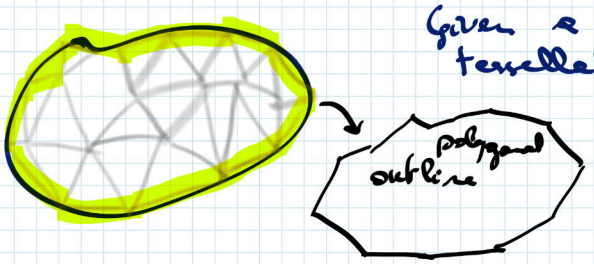
A inherits the features of $a(\cdot,\cdot)$ as in the 1D case.

Finite Elements

The basic idea of FE is to construct the space $V_h \subset V \; (= H^1_0)$ as a **piecewise polynomial** approximation.

One important difference with the 1D case is the construction of the "pieces" or of the MESH

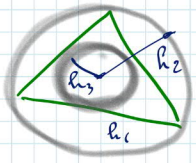
Triangular Meshes in 2D



Given a generic domain, we can think of a tessellation with triangles.

Triangles are really versatile and can cover regions with complex shapes. However, we have some WARNINGS.

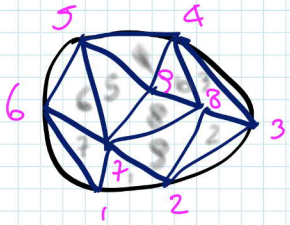
(1) Denote by T_h the tessellation with Triangles $\{T_K\}$ $K = 1 \dots N$
Here h denotes a characteristic lenght of the triangles.
A possible choice is the longest edge, or the radius of the inscribed circle or the circumscribed



(2) A mesh, in general, consists of

- VERTEXES
- EDGES (2 vertexes)
- TRIANGLES (3 vertexes, 3 edges)

Now, the way a mesh file is organized differs for different softwares. However, you find this information.



List of Vertexes: x_1, x_2 (z in the 3D case)

Pattern of Elements:

1	1	2	7
2	2	3	8
3	3	4	8
4	4	5	9
5	5	7	9
6	5	6	7
7	1	7	6
8	7	8	9
9	2	8	7

10 8 48

The nodes are always COUNTER CLOCKWISE

List of Boundary Edges

1	2
2	3
3	4
4	5
5	6
6	1

+ Labels

Additional information may be added to identify the neighbor elements:

Element 1) 9 7 (-1 = boundary : - boundary edge)

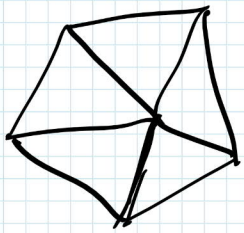
Element 2) 3 9 (-2)

(3) If we want continuous finctions, we have some restrictions.
Why we want continuous functions?

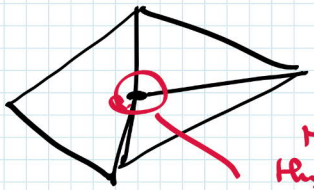
Theorem If $f \in C^0(\overline{\Omega})$ and it's piecewise H^1, then $\in H^1(\Omega)$
So, continuity guarantees conformity ($V_h \subset V$).

Each geometrical entity must be the same for all the elements sharing it:

a VERTEX is a vertex for all the elements sharing it:



OK



NO

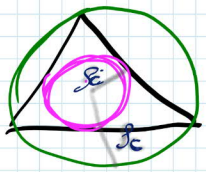
HANGING NODE
this point is a vertex only for some elements

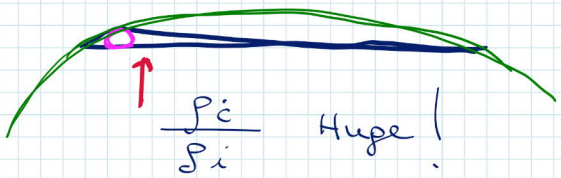
Triangles should not self-intersect or overlap

is not acceptable.

REMARK
In some cases we can live with HANGING NODES, but they need a particular care. We do not cover this here.

(4) A good mesh generally prefers "regular" meshes, so triangles with a good "ratio" or quotient between the inscribed and the circumscribed circles.



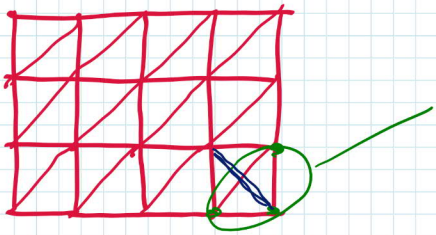


We consider here meshes where

$$\frac{\rho_c}{\rho_i} \text{ is bounded}$$

Nevertheless, meshes with extremely skewed elements are sometimes useful (specific theory required).

(5) It's good avoiding all vertices of a triangle on the boundary



having 3 nodes on the boundary, the solution here is completely determined by the B.C., no equation is solved inside $\Rightarrow$ SWAP THE DIAGONAL (Bhe)

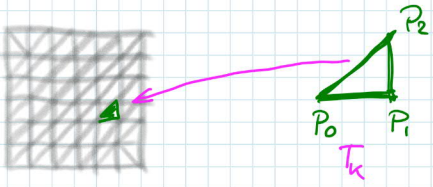
Linear FE in 2D

If we use a triangular mesh with the features mentioned above, we can postulate the construction of a solution in the space:

$$V_h = \{ v_h \in C^0(\bar{\Omega}) : v_h(T_k) \in \mathbb{P}^1 \}$$

where T_k is the generic triangle of the mesh

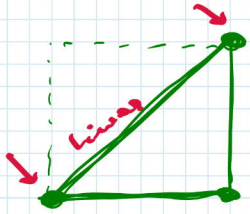
Here, we are assuming that the mesh is covering exactly the domain ($\Omega_h = \Omega$) like in a square or a rectangle:



In the generic triangle, we assume the solution to be linear

$$u_h(T_k) = ax + by + c$$

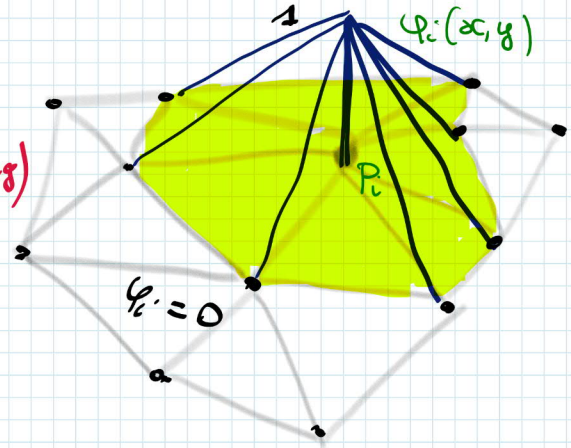
We have 3 parameters to identify, so a natural choice that guarantees the continuity of u_h is to use the value in the 3 vertexes.



3 values of u_h in the vertexes identify the 3 coefficients a, b, c

A function in the neighbor triangle with the same values at the vertexes is automatically continuous.

In a modernise perspective, we are constructing the Lagrange piecewise polynomials as illustrated below:



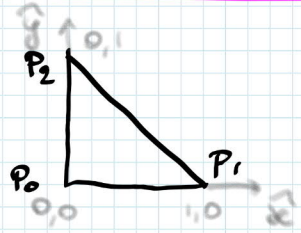
In yellow, the support of $\varphi_i(x,y)$

$$u_h = \sum c_j \varphi_j(x,y)$$

with this choice of basis functions: $c_j = u_h(P_j)$

Formally, the solution of the problem is like in the 1D case. However, we have some technical details to cover.

The Reference Finite Element



The factual representation of the basis functions is in two-steps

- representation on the reference element
- map current ← reference

Representation on the Reference FE

We have to construct 3 linear functions:

$$\left.\begin{aligned}\hat{\varphi}_0(\hat{P}_0) &= 1\\ \hat{\varphi}_0(\hat{P}_1) = \hat{\varphi}_0(\hat{P}_2) &= 0\end{aligned}\right\}$$

Vanishing on the nodes $\hat{P}_1$ and $\hat{P}_2$, and being linear, clearly $\hat{\varphi}_0$ vanishes on the entire segment $\hat{P}_1\hat{P}_2$.

This line has equation: $\hat{x}+\hat{y}=1$

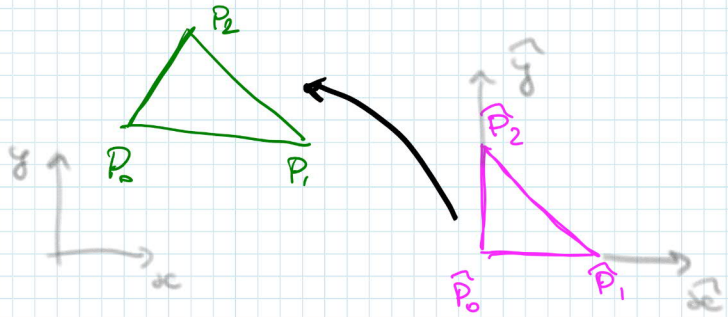
$$\Rightarrow \hat{\varphi}_0(\hat{x},\hat{y}) = 1-\hat{x}-\hat{y}$$

(vanishes on $\hat{P}_1\hat{P}_2$ automatically)

Similarly: $\hat{\varphi}_1(\hat{x},\hat{y})=\hat{x}$, $\hat{\varphi}_2(\hat{x},\hat{y})=\hat{y}$

(BTW, notice that $\hat{\varphi}_0+\hat{\varphi}_1+\hat{\varphi}_2=1$. The same holds in 1D and 3D $\Rightarrow$ partition of unity property)

Map Current ← Reference



The map consists of two functions

$$\left.\begin{aligned}x(\hat{x},\hat{y})\\ y(\hat{x},\hat{y})\end{aligned}\right\}$$

s.t.

$$\begin{cases} x(0,0) = x_0\\ y(0,0) = y_0\\ x(1,0) = x_1\\ y(1,0) = y_1\\ x(0,1) = x_2\\ y(0,1) = y_2\end{cases}$$

We have 3×2 conditions, if $x(\hat{x},\hat{y})$ and $y(\hat{x},\hat{y})$ are linear functions, we have 3×2 parameters, so we have a perfect match!!!

The explicit solution is easily constructed by using the same $\hat{\varphi}_0, \hat{\varphi}_1, \hat{\varphi}_2$ of the reference finite element:

$$\begin{aligned}x &= x_0\hat{\varphi}_0(\hat{x},\hat{y}) + x_1\hat{\varphi}_1(\hat{x},\hat{y}) + x_2\hat{\varphi}_2(\hat{x},\hat{y})\\ y &= y_0\hat{\varphi}_0(\hat{x},\hat{y}) + y_1\hat{\varphi}_1(\hat{x},\hat{y}) + y_2\hat{\varphi}_2(\hat{x},\hat{y})\end{aligned}$$

This is an AFFINE map, mapping triangles in Triangles.

Finite elements using this map are called affine.

Notice that the jacobian matrix of this transform is

$$J = \begin{bmatrix} \frac{\partial x}{\partial \hat{x}} & \frac{\partial y}{\partial \hat{x}} \\ \frac{\partial x}{\partial \hat{y}} & \frac{\partial y}{\partial \hat{y}} \end{bmatrix} \qquad |J| = \frac{\partial x}{\partial \hat{x}}\frac{\partial y}{\partial \hat{y}} - \frac{\partial x}{\partial \hat{y}}\frac{\partial y}{\partial \hat{x}}$$

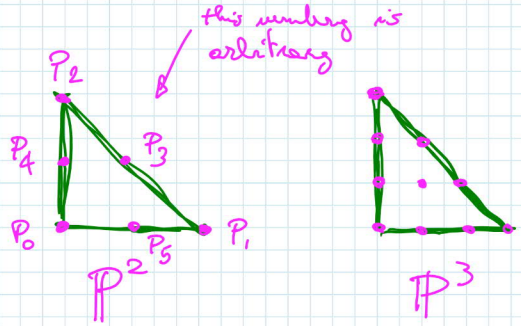
and in general we need to keep in mind that:

$$\nabla_{x,y} u = \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{\partial u}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial x} + \frac{\partial u}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial x} \\ \frac{\partial u}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial y} + \frac{\partial u}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{\partial \hat{x}}{\partial x} & \frac{\partial \hat{y}}{\partial x} \\ \frac{\partial \hat{x}}{\partial y} & \frac{\partial \hat{y}}{\partial y} \end{bmatrix} \nabla_{\hat{x},\hat{y}} u$$

Affine Finite Elements of Order > 1 (2D)

We can obviously introduce finite elements of degree 2 and 3 on elements obtained by an affine map.

On the reference element



• = Lagrangian degrees of freedom

$u = a\hat{x}^2 + b\hat{y}^2 + c\hat{x}\hat{y} + d\hat{x} + e\hat{y} + f$

6 parameters

$u = a\hat{x}^3 + b\hat{x}^2\hat{y} + c\hat{x}\hat{y}^2 + d\hat{y}^3 + e\hat{x}^2 + f\hat{y}^2 + g\hat{x}\hat{y} + h\hat{x} + k\hat{y} + j$

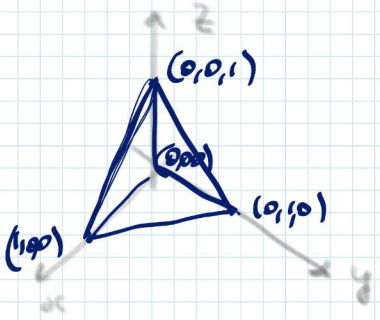
10 parameters

$$\hat{\varphi}_0 = 2(1-\hat{x}-\hat{y})\left(\tfrac{1}{2}-\hat{x}-\hat{y}\right) \qquad \hat{\varphi}_3 = 4\hat{x}\hat{y}$$
$$\hat{\varphi}_1 = \hat{x}\left(\hat{x}-\tfrac{1}{2}\right)2 \qquad \hat{\varphi}_4 = 4\hat{y}(1-\hat{x}-\hat{y})$$
$$\hat{\varphi}_2 = \hat{y}\left(\hat{y}-\tfrac{1}{2}\right)2 \qquad \hat{\varphi}_5 = 4\hat{x}(1-\hat{x}-\hat{y})$$

9 dof on the boundary + 1 bubble.

Affine Finite Elements in 3D

In 3D we work directly on the reference element, the unit tetrahedron

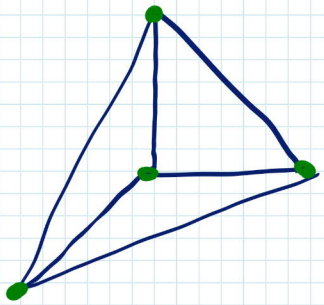


4 vertices

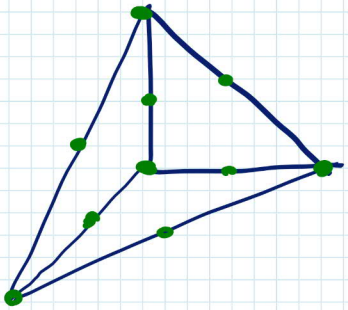
linear functions are in the form:

$a\hat{x} + b\hat{y} + c\hat{z} + d$ ⇒ 4 parameters

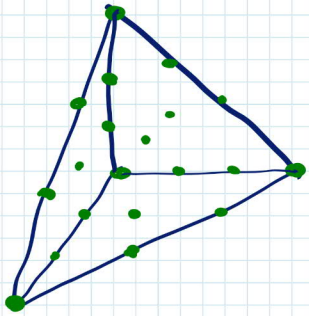
The 4 vertices are used for d.o.f. in the linear case



4 vertices

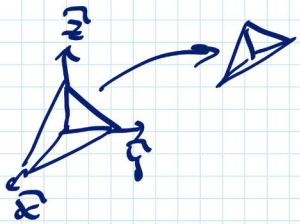


4 vertices +
6 edges = 10 dof



4 vertices + 12 edges +
4 faces = 2 dof

Also in this case the map current ← reference is done by using an affine map with the P1 basis functions:



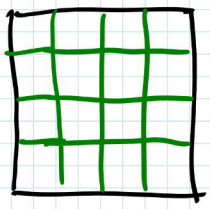
$$\begin{cases} x = x_0 \hat{\varphi}_0(\hat{x},\hat{y},\hat{z}) + x_1 \hat{\varphi}_1(\hat{x},\hat{y},\hat{z}) + x_2 \hat{\varphi}_2(\hat{x},\hat{y},\hat{z}) + x_3 \hat{\varphi}_3(\hat{x},\hat{y},\hat{z}) \\ y = y_0 \hat{\varphi}_0 + y_1 \hat{\varphi}_1 + y_2 \hat{\varphi}_2 + y_3 \hat{\varphi}_3 \\ z = z_0 \hat{\varphi}_0 + z_1 \hat{\varphi}_1 + z_2 \hat{\varphi}_2 + z_3 \hat{\varphi}_3 \end{cases}$$

Remarks

(1) Quadrilateral finite elements (and more)

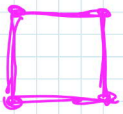
We can consider also different basic geometries, like QUADRILATERALS.

For instance, if we cover a domain with squares:



we need to consider different functions:

(1) Bilinear functions: $axy + bx + cy + d$ ⟹ (linear in each variable)



4 d.o.f.
=
4 equations = 4 Vertices

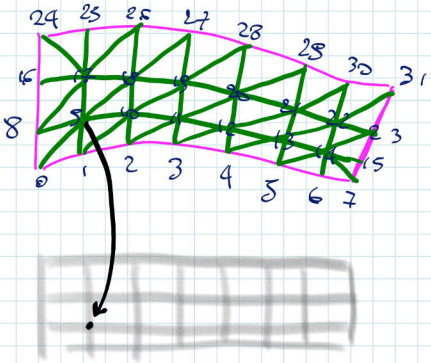
(2) Biquadratic or more functions.

We do not go in details here: bilinear f.e. may have good properties, but the meshing is less versatile.

Recently, Brezzi et al. introduced the so-called MIMETIC FINITE DIFFERENCES that are smart finite elements working on polygons (or, in general, polytopes).

(2) Structured and Unstructured Meshes

When a mesh has a STRONG CARTESIAN FOOTPRINT (like in Finite Differences) we can easily infer a relation between the position and the numbering of the nodes/elements.

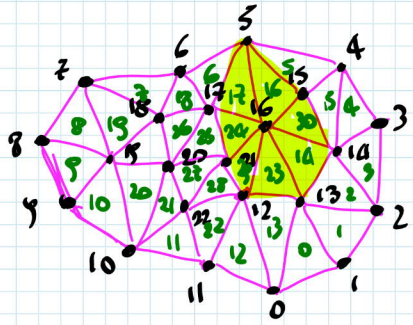


In this case (see the Figure) we say that the mesh is STRUCTURED.

Having a rule between position and numbering can make the coding more efficient.

However, in most of the cases this is not possible.

In general, the mesh is then called UNSTRUCTURED.
In this case, the connection between NUMBERING & COORDINATES is all based on the tables in the mesh file.



IMPORTANT

A typical feature of FEM is that the stiffness matrix is sparse.

With unstructured meshes the sparsity pattern is however disordered.

Nevertheless, it is completely determined by the mesh numbering. This means that we know the pattern of the matrix directly from the mesh.

For instance, in the mesh above we have 22 nodes, 12 of them on the boundary.

For instance, node 16 "sees" the nodes:

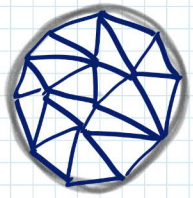
(16) 15 - 5 - 17 - 21 - 12 - 13 - 14 (row 16 has nonzero entries in these columns)

(18) 18 - 7 - 8 - 9 - 10 - 20

As you can see, the pattern (command SPY) will be with no structure.

Isoparametric Finite Elements

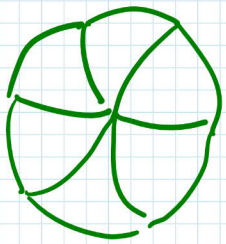
So far, we assumed that $\Omega_h = \Omega$. However, in general this is not true: a circle, for instance, cannot be covered exactly by linear triangles



This introduces an additional error that affects the actual convergence rates

=> if you are interested,
ERN, GUERMOND, Theory & Practice of F E, SPRINGER.

This can be avoided using "curvilinear finite elements":



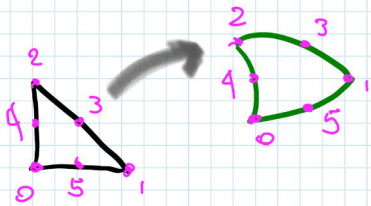
How can we generate those elements.

Well, we can still start from the usual Reference Element.

However, we postulate an either quadratic or cubic map that can be described by the same basis functions of the finite elements.

EXAMPLE

Quadratic Elements:



$$\begin{cases} x = x_0\hat{\varphi}_0 + x_1\hat{\varphi}_1 + x_2\hat{\varphi}_2 + x_3\hat{\varphi}_3 + x_4\hat{\varphi}_4 + x_5\hat{\varphi}_5 \\ y = y_0\hat{\varphi}_0 + \dots + y_5\hat{\varphi}_5 \end{cases}$$

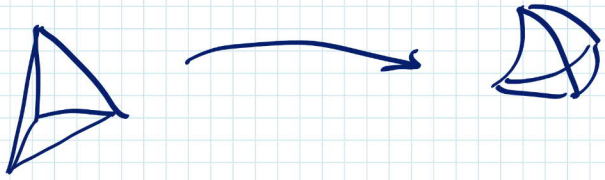
If we use $\mathbb{P}^2$ with quadratic elements, we have ISOPARAMETRIC ELEMENTS (iso = equal, same; parametric = # of parameters)

=> we use the same basis functions for the map and the solution u_h.

(Linear Affine Finite Elements are 'isoparametric too).

Using these elements, the geometrical error for Ω_h being $\neq \Omega$ is significantly reduced.

We can obviously consider isoparametric finite elements in 3D too



Convergence Rate and Other Theoretical Aspects in 2+D.

In spite of an increased complexity, many theoretical results considered in 1D hold in multiple dimensions.

CONVERGENCE RATE:

$$\|u - u_h\|_{H^1} \leq C h^{r} |u|_{H^{s+1}} \qquad r = \min(s, p)$$

for $u \in H^{s+1}(\Omega)$ and with FE of order p.

This can be proved by combining two results:

- Deny-Lions Lemma
- Bramble-Hilbert Theorem

] with regular mesh $\left(\frac{\delta_c}{\delta_i} \leq T < +\infty\right)$ and $\Omega_h = \Omega$.

CONDITION NUMBER

$$\text{cond}(A) \propto h^{-2}$$ also in 2+D.

Remark: ISOGEOMETRIC ANALYSIS (IGA)

Recently ($\approx$ 2005) T. Hughes introduced the so-called ISOGEOMETRIC ANALYSIS

The basic idea is to use NURBS instead of regular polynomials on the elements. NURBS stands for Non-Uniform Rational B-Splines.

These are (piecewise) regular functions much more regular and complicated that can attain high convergence rates.

The big advantage of IGA is that NURBS are the basis of CAD softwares. With IGA the combination of CAD (Computer Aided Design) and CAE (Computer Aided Engineering) is very immediate and having the same approach of isoparametric finite elements, the geometrical accuracy of the domain Ω is very high.

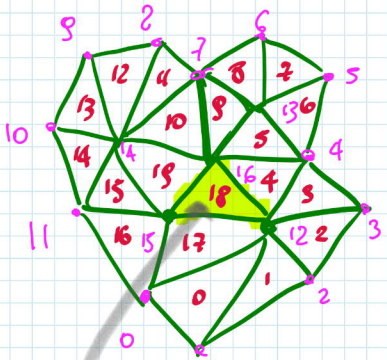
Some Details on the Implementation

As we noticed, the assembly of the stiffness matrix is done elementwise, with a "think globally, act locally" approach.

We first assemble the local matrix on the single element: the local matrix is small and full.

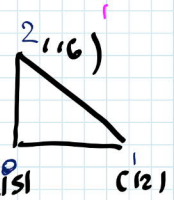
Then, we map the contribution of the local matrix to the global one (exploiting the additivity of the integral operator, such that for instance

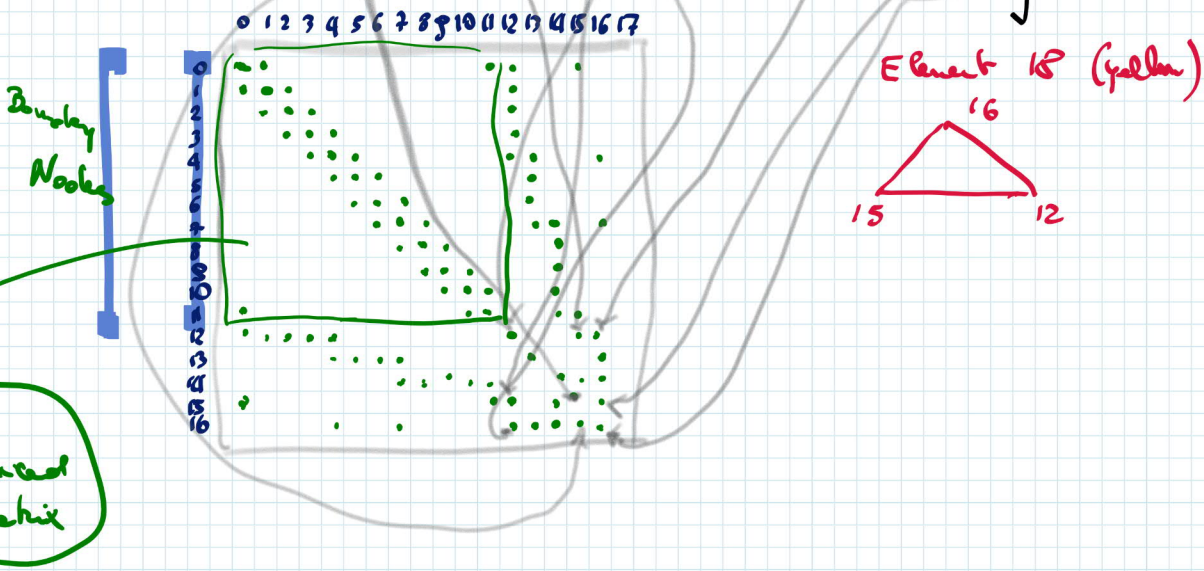
$$\int_{\Omega_h} u_h v_h = \sum_K \int_{T_K} u_h v_h$$



For instance, for linear FEM in 2D, the local matrix is 3×3:

$$A_{loc} = \begin{bmatrix} a(\hat{\varphi}_0,\hat{\varphi}_0) & a(\hat{\varphi}_1,\hat{\varphi}_0) & a(\hat{\varphi}_2,\hat{\varphi}_0) \\ a(\hat{\varphi}_0,\hat{\varphi}_1) & a(\hat{\varphi}_1,\hat{\varphi}_1) & a(\hat{\varphi}_2,\hat{\varphi}_1) \\ a(\hat{\varphi}_0,\hat{\varphi}_2) & a(\hat{\varphi}_1,\hat{\varphi}_2) & a(\hat{\varphi}_2,\hat{\varphi}_2) \end{bmatrix} \begin{matrix} 15 \\ 12 \\ 16 \end{matrix}$$





We address the entries by using the mesh DS as described earlier.

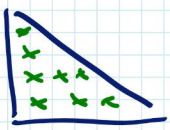
Element 18: 15 12 16

↑ ↑ ↑

rows and columns touched in A when processing element 18.

Computing the Integrals

We used Gaussian or Gauss-Lobatto formulas



$$\int_T f \approx \sum w_i f(x_i, y_i)$$

weights, nodes tabulated.

Since we do not want errors in the integrals, we try to use a formula with many nodes (= high degree of exactness).

However, be careful! The computational costs rapidly increases.

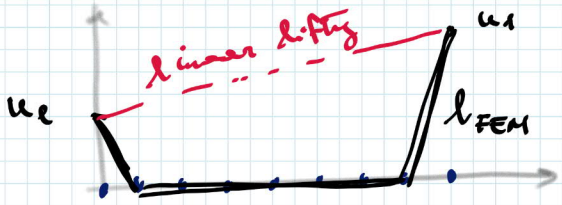
Boundary Conditions.

Let's visualize this in 1D, just for simplicity.

When we identify the Dirichlet boundary nodes and set:

$$A(\text{label\_D}, \text{label\_D}) = 1$$
$$A(\text{label\_D}, \text{off-diagonal}) = 0$$

is equivalent to introducing the following lifting of the data:



you can test yourself this idea with an exercise.

Homework: install Fenics

Week 9

Quarteroni Chap. 12

Advection - Diffusion - Reaction with FEM

We have noticed with FD that sometimes the real numbers require restrictions to the numerical discretization.
In particular, with the problem:

$$-\mu u'' + \beta u' = 0 \qquad x \in (0,1) \qquad (*)$$
$$u(0) = 0 \qquad u(1) = 1$$

we found that the classical centered 2nd order difference scheme:

$$-\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + \beta \frac{u_{i+1} - u_{i-1}}{2h} = 0$$
$$u_0 = 0 \qquad u_{N_h} = 1 \qquad (Nh = 1)$$

is oscillating for $\frac{|\beta| h}{2\mu} > 1$

while the solution is monotone.

A natural question is: what happens with FEM?
Do we have the same problem?
What are the solutions?
What to do in 2D?
In this week, we will answer to these questions.

Advection Dominated Problems in 1D

To start with, we can check what we obtain if we discretize the same problem (*) with linear finite elements.

The variational form reads:

$$\mu \int_0^1 u_h' \varphi_i' + \beta \int_0^1 u_h' \varphi_i = 0 \qquad \varphi_i(0) = \varphi_i(1) = 0$$

The entries for the first integral are:

$$\mu \int_0^1 \varphi_j' \varphi_i' = \begin{cases} \frac{2}{h}\mu & i = j \\ -\frac{1}{h}\mu & j = i \pm 1 \\ 0 & j \neq i, i \pm 1 \end{cases}$$

For the second one:

$$\beta \int_0^1 \varphi_i \varphi_j' = \begin{cases} 0 & j = i \text{ and for } j \neq i \pm 1 \\ \beta & j = i+1 \\ -\beta & j = i-1 \end{cases}$$

So, at the end we get the scheme:

$$-\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{h} + \beta \left(u_{i+1} - u_{i-1} \right) = 0$$

$$u_0 = 0 \qquad u_{N+1} = 1$$

But THIS IS EXACTLY THE FD SCHEME, just multiplied by h.

$\Rightarrow$ Also with FEM we should expect oscillations for $\frac{\beta h}{2\mu} > 1$.

We already know that UPWIND is a remedy, but the upwind physics (= go toward the upwind direction) doesn't apply to the variational formulation. In fact, that interpretation works only for FD.

To apply the concept to FEM, we resort to the ARTIFICIAL VISCOSITY.

We found that Upwind in 1D is equivalent to an increment of the coefficient μ, so that:

$$\mu^* = \mu (1 + \text{Pe})$$

With FE, we can therefore suppress the oscillation by solving the following problem:

$$\mu \int_0^1 u_h' v_h' + \frac{|\beta| h}{2} \int_0^1 u_h' v_h' + \beta \int_0^1 u_h' v_h = 0 \qquad \begin{matrix} u_h(0) = 0 \\ u_h(1) = 1 \end{matrix} \qquad \forall v_h \in H_0^1(0,1).$$

For the FD problem, we know that this increment of the viscosity was obtained by reducing the accuracy and that we need a second-order upwind scheme to preserve the overall 2nd-order of the FD scheme.

What can we say for the FE perspective (i.e., error analysis)?

A first remark: the red term is a "perturbation" of the real problem we want to solve. However, it's a legal perturbation as long as it vanishes with h.

Let's try to formalize the concept.

The original problem reads:

find $u \in V$ s.t. $a(u,v) = f(v) \quad \forall v \in V$

where $a(u,v) = \mu \int_0^1 u'v' + \beta \int_0^1 u'v$ and $f(v) = 0$.

The numerical problem with upwind reads:

find $u_h \in V_h$ s.t. $a_h(u_h, v_h) = f(v_h) \quad \forall v_h \in V_h$

where $a_h(u_h, v_h) = a(u_h, v_h) + \frac{|\beta| h}{2} \int_\Omega u_h' v_h'$

(so that $a_h(u_h, v_h) \to a(u_h, v_h)$ for $h \to 0$).

A first important consequence is that THE CÉA LEMMA DOES NOT HOLD ANY LONGER.

While $a(u, v_h) = a(u_h, v_h)$, we have:

$$a_h(u, v_h) \neq f(v_h), \quad \text{so} \quad a_h(u - u_h, v_h) \neq 0.$$

$\Rightarrow$ THE UPWIND FE METHOD IS NOT STRONGLY CONSISTENT.

However $\left| a_h(u_h, v_h) - a(u_h, v_h) \right| = \left| \frac{|\beta| h}{2} \int_\Omega u_h' v_h' \right| \xrightarrow{h \to 0} 0.$

We conclude that $a_h(u - u_h, v_h) = f(v_h) - f(v_h) - \frac{|\beta| h}{2} \int_\Omega u_h' v_h' = \tau(h) \xrightarrow{h \to 0} 0$

so upwind is CONSISTENT (not only consistent).

From the Lax-Richtmyer Theorem (convergence $\Leftrightarrow$ stability + consistency) we know that this is enough for the convergence.

As a matter of fact, it is easy to prove that our upwind scheme is stable:

$a_h(u_h, v_h) = a(u_h, v_h) + \frac{|\beta| h}{2} \int u_h' v_h'$ is

$$\begin{cases} \text{BILINEAR} \\ \text{CONTINUOUS} \\ \text{COERCIVE} \end{cases} \quad \left(a_h(u_h, u_h) = \underbrace{a(u_h, v_h)}_{\text{coercive}} + \underbrace{\frac{|\beta| h}{2} \int u_h'^2}_{>0} \geq \alpha \|u_h\|_V^2 \right)$$

$\Rightarrow \alpha \|u_h\|_V^2 \leq |a_h(u_h, u_h)| \leq |f(u_h)| \leq M \|u_h\|_V \Rightarrow \|u_h\|_V \leq M/\alpha.$

OK, we have the convergence, but what about the convergence rate?
We don't have the Céa Lemma.
We need a new theoretical tool.
This is given by the so-called STRANG-LEMMA.

Generalized Galerkin Schemes (GG)

Let's consider, in general, the following class of schemes for our original problem, called Generalized Galerkin:

Find $u_h \in V_h$, s.t. $a_h(u_h, v_h) = f_h(v_h) \quad \forall v_h \in V_h$ (**)

where

$$a_h(u_h, v_h) - a(u_h, v_h) = b_h(u_h, v_h)$$
$$f_h(v_h) - f(v_h) \approx g_h(u_h, v_h)$$

(for upwind this term is 0)

STRANG LEMMA

Let's assume that $a_h(\cdot,\cdot)$ is bilinear, continuous and coercive

$f_h(\cdot,\cdot)$ is linear, continuous

(with continuity and coercivity constants independent of h)

Then, the Generalized Galerkin problem $(**)$ is well-posed and $\exists$ Constants C_1, C_2, C_3 s.t.

$$\|u - u_h\|_V \le \inf_{w_h \in V_h} \left(C_1 \|u - w_h\|_V + C_2 \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{|b_h(w_h, v_h)|}{\|v_h\|_V} \right) + C_3 \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{|G_h(v_h)|}{\|v_h\|_V}$$

- $\inf_{w_h \in V_h} C_1 \|u - w_h\|_V$: discretization error (as with the Céa Lemma)
- $C_2 \sup \frac{|b_h(w_h, v_h)|}{\|v_h\|_V}$ and $C_3 \sup \frac{|G_h(v_h)|}{\|v_h\|_V}$: error for the changes in a_h and f_h

This Lemma generalizes to GG the Céa Lemma, showing that the total error is the combination of the discretization and the modifications of $a(\cdot,\cdot)$ and $f(\cdot)$.

In the case of UPWIND, we have

$$b_h(w_h, v_h) = \frac{|\beta| h}{2} \int_0^1 w_h' v_h \quad \Rightarrow \quad \frac{|b_h(w_h, v_h)|}{\|v_h\|_V} \le h \frac{|\beta|}{2} \|w_h\|_V$$

$$G_h(v_h) = 0$$

so we conclude that $\|u - u_h^{UPW}\|_V \le Ch$ regardless of the FE degree.

REMARK

The Strang Lemma is useful also for estimating the impact of numerical quadrature on the performance of the method.

In fact, we can write:

$$a(u,v) = \int_0^1 \text{"some terms"}$$

$$a_h(u,v) = \text{Numerical Quadrature}\left(\text{"same terms"}\right)$$

In this respect, $b_h(\cdot,\cdot)$ and $G_h(\cdot)$ are the Quadrature Errors.

If we use Gaussian formulas with an adequate number of nodes, the convergence rate of these terms will be always $> r$ $(= \min(p, s))$ of the discretization term. This requires an adequate number of nodes to use, which can be computationally demanding.

However, there are cases where high-order quadrature formulas are not necessary. Or not even convenient.

For instance, for constant coefficients problems with FEM of order q

$$a(\varphi_j, \varphi_i) = \mu \int_0^1 \varphi_j' \varphi_i' + \beta \int_0^1 \varphi_j' \varphi_i + \sigma \int_0^1 \varphi_j \varphi_i$$

$\varphi_j' \varphi_i'$: deg. $q-1 \times q-1$, $2q-2$ — $\varphi_j' \varphi_i$: $q-1 \times q$, $2q-1$ — $\varphi_j \varphi_i$: $q \times q$, $2q$

With a quadrature with DEGREE OF EXACTNESS $2q$ or more we do not have any error for the quadrature.

This shows how important is to know the concept of Degree of Exactness (the highest polynomial degree integrated exactly by a quadrature formula).

As we said, in general, we can write:

$$b_h(u_h, v_h) = a(u_h, v_h) - \text{Numerical Integration of } a(u_h, v_h)$$
$$g_h(v_h) = f(u_{h1}, v_h) - \text{" " " } f(u_h, v_h)$$

If the quadrature formula is accurate enough, these terms feature a dependence on h^{qf} with $qf >$ discretization order, so the quadrature error is present but it doesn't affect the convergence order due to the discretization.

This obviously requires a certain number of Gaussian quadrature nodes, with an impact on the computational costs.

High Order Upwind (in 1D) with FE

We noticed that with FD we can get higher order stable schemes by using not-centred high order formulas. What can we do with FE, where we do not have a "centred" or "uncentred" choice?

In 1D we can stick to the concept of artificial viscosity: we found that the basic upwind is an artificial viscosity s.t.

$$\mu^* = \mu(1 + Pe)$$

The idea is to consider an artificial viscosity:

$$\mu^{**} = \mu(1 + \Phi(Pe))$$

where $\Phi(Pe)$ is a function with the following features:

✓ $\Phi(Pe)$ "large enough" (stabilization: $\frac{|\beta| h}{2\mu^{eff}} < 1 \Rightarrow \frac{Pe}{1+\Phi(Pe)} < 1 \Rightarrow \Phi(Pe) - Pe + 1 > 0$)

✓ $\lim_{Pe \to 0} \Phi(Pe) = 0$ and possibly $= O(h^p)$ $p > 1$

(accuracy: $|u_e| \sim O(h^p)$)

A possible function with these features is the so-called SCHARFETTER-GUMMEL stabilization:

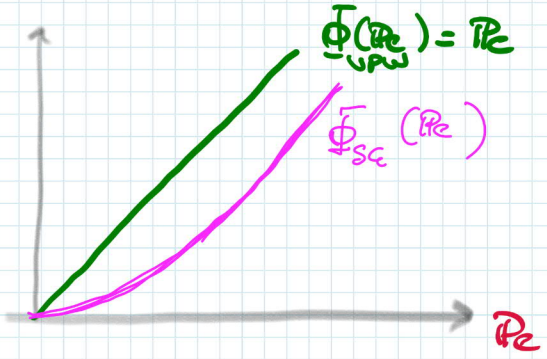
$$\Phi_{SG}(Pe) = Pe - 1 + \frac{Pe}{e^{Pe} - 1} \qquad \left(\text{with } \Phi_{SG}(0) = 0 \text{ to manage the singularity}\right)$$

This function is "smart": clearly

$$\Phi_{SG}(Pe) - Pe + 1 = \frac{Pe}{e^{Pe} - 1} > 0$$

and

$$\lim_{h \to 0} \Phi_{SG}(Pe) \sim O(h^2)$$



For high Pe, $\Phi_{SG}(Pe) \simeq \Phi_{UPW}(Pe)$, for h small, it converges to 0 faster than UPW.

SG is so good that it is possible to prove that in the problem:

$$-\mu u'' + \beta u' = f$$

for a piecewise constant f, it gives a "nodally exact" solution.

Upwind and SG as Petrov-Galerkin methods.

The basic Galerkin method reads:

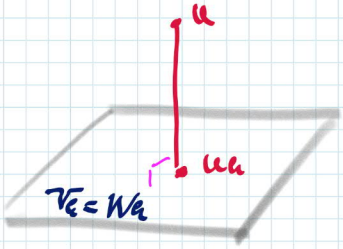
$$a(u_h, v_h) = \ell(v_h)$$

with u_h, v_h belonging to the same space.

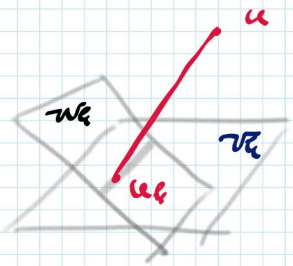
This leads to $a(u - u_h, v_h) = 0$ that is an orthogonal projection.

We can generalize to an oblique projection with:

$$a(u_h, w_h) = \ell(w_h)$$ where w_h has the same dimension (to have a square problem) but it is not $= v_h$.



Orthogonal Projection (Galerkin)



Oblique Projection (Petrov-Galerkin)

It is possible to see for LINEAR FINITE ELEMENTS, that UPW and SG can be regarded as Petrov-Galerkin methods, where

$$W_h = V_h + \{B^\alpha\}$$

where B are "bubble" functions and $\alpha = 1$ for UPW and another "strange" value for SG.

See Quarteroni 12.8.2

Reaction Dominated Problems

Let's consider this problem:

$$-\mu u'' + \sigma u = 0 \qquad x \in (0,1)$$
$$u(0) = 0 \quad u(1) = 1 \qquad (\mu > 0, \sigma \geq 0)$$

Should we expect numerical instabilities when $\sigma >> \mu$ as it happens for advection dominated problems?

The problem is certainly well posed:

$$a(u,v) = \mu \int_0^1 u'v' + \sigma \int uv$$

- Continuous with continuity constant: $M = \max(\sigma, \mu)$
- Coercive with coercivity constant: $\alpha = \min(\sigma, \mu)$

so the numerical approximation is such that

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \inf_{w_h \in V_h} \|u - w_h\|_V$$

If $\sigma >> \mu$ we have $\frac{M}{\alpha} = \frac{\sigma}{\mu} >> 1$ and we may expect some inaccuracies.

Now, let's consider in detail the discretization of the problem.

(1) Exact solution:

$$u'' = \frac{\sigma}{\mu} u \qquad u = c_1 e^{\sqrt{\frac{\sigma}{\mu}}x} + c_2 e^{-\sqrt{\frac{\sigma}{\mu}}x}$$

$$\left.\begin{aligned} u(0)=0 &: \quad c_1 = -c_2 \\ u(1)=1 &: \quad c_1 e^{\sqrt{\frac{\sigma}{\mu}}} + c_2 e^{-\sqrt{\frac{\sigma}{\mu}}} = 1 \end{aligned}\right\} \; c_1 = \frac{1}{e^{\sqrt{\frac{\sigma}{\mu}}} - e^{-\sqrt{\frac{\sigma}{\mu}}}}$$

$$u_a = \frac{e^{\sqrt{\frac{\sigma}{\mu}}x} - e^{-\sqrt{\frac{\sigma}{\mu}}x}}{e^{\sqrt{\frac{\sigma}{\mu}}} - e^{-\sqrt{\frac{\sigma}{\mu}}}}$$

2) Finite Difference solution:

$$\left.\begin{aligned} -\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + \sigma u_i = 0 \\ u_0 = 0 \qquad u_{n+1} = 1 \end{aligned}\right\} \quad u_{i+1} - 2\left(1 + \frac{\sigma h^2}{2\mu}\right) u_i + u_{i-1} = 0$$

$$\beta_{1,2} = \left(1 + \frac{\sigma h^2}{2\mu}\right) \pm \sqrt{\left(1 + \frac{\sigma h^2}{2\mu}\right)^2 - 1} =$$

$$= 1 + \frac{\sigma h^2}{2\mu} \pm \sqrt{\frac{\sigma h^2}{\mu} + \frac{\sigma h^4}{4\mu^2}} > 0 \quad \forall h$$

Since β_1, β_2 are always > 0, no oscillations!

(3) Linear Finite Elements discretization

We already know that the diffusive term leads to:

$$-\mu u'' \Rightarrow \mu \int_0^1 u' v' \Rightarrow -\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{h}$$

For the reactive term:

$$\sigma u \Rightarrow \sigma \int_0^1 uv \Rightarrow ?$$

$$\sigma \int_0^1 \varphi_i^2 = \sigma \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1})^2}{h^2} + \sigma \int_{x_i}^{x_{i+1}} \frac{(x_{i+1} - x)^2}{h^2} =$$

$$= \frac{\sigma}{h^2} \frac{1}{3} h^3 + \frac{\sigma}{h^2} \frac{1}{3} h^3 = \frac{2}{3} \sigma h$$

$$\sigma \int_0^1 \varphi_i \varphi_{i-1} = \sigma \int_{x_{i-1}}^{x_i} \frac{1}{h^2} (x_{i+1} - x)(x - x_i) = \ldots = \sigma h \frac{1}{6}$$

$$\sigma \int_0^1 \varphi_i \varphi_{i+1} = \qquad\qquad = \sigma h \frac{1}{6}$$

So, at the end we have:

$$-\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{h} + \frac{\sigma h}{6}\left(u_{i+1} + 4u_i + u_{i-1}\right) = 0$$

$$\frac{\mu}{h}\left(\frac{\sigma h^2}{6\mu} - 1\right) u_{i+1} + \frac{\sigma h}{6}\left(1 + \frac{\sigma h^2}{3\mu}\right) u_i + \frac{\mu}{h}\left(\frac{\sigma h^2}{6\mu} - 1\right) u_{i-1} = 0$$

$$u_{i+1} - 2\underbrace{\frac{\left(1+\frac{\sigma h^2}{3\mu}\right)}{1-\frac{\sigma h^2}{6\mu}}}_{\gamma} u_i + u_{i-1} = 0$$

$$g^2 - 2\gamma g + 1 = 0 \qquad g_{1,2} = \gamma \pm \sqrt{\gamma^2 - 1}$$

Notice that, if $1 - \frac{\sigma h^2}{6\mu} > 0$, then $\gamma > 0$ and precisely:

$$\gamma = \frac{1+\frac{\sigma h^2}{3\mu}}{1-\frac{\sigma h^2}{6\mu}} = \frac{1 - \frac{\sigma h^2}{6\mu} + \frac{\sigma h^2}{3\mu} + \frac{\sigma h^2}{6\mu}}{1-\frac{\sigma h^2}{6\mu}} > 1 \quad \forall h$$

then:

$$g_{1,2} = \gamma \pm \sqrt{\gamma^2 - 1} > 0 \quad \forall h$$

In the case $1 - \frac{\sigma h^2}{6\mu} < 0$, then:

$$|\gamma| = \frac{1+\frac{\sigma h^2}{3\mu}}{\frac{\sigma h^2}{6\mu} - 1} = \frac{\frac{\sigma h^2}{6\mu} - 1 + \frac{\sigma h^2}{6\mu} + 2}{\frac{\sigma h^2}{6\mu} - 1} = 1 + \frac{\frac{\sigma h^2}{6\mu} + 2}{\frac{\sigma h^2}{6\mu} - 1} > 1$$

Then

$$g_{1,2} = -|\gamma| \pm \sqrt{\gamma^2 - 1}$$ are real but one is negative

$\Rightarrow$ We have oscillations!

Linear Finite Elements present oscillations if $\frac{\sigma h^2}{6\mu} > 1$

We can introduce the Reactive Peclét number and state that to avoid oscillations we need:

$$Pe_R = \frac{\sigma h^2}{6\mu} < 1 \Rightarrow h < \sqrt{\frac{6\mu}{\sigma}}$$

Differently from FD, FE do suffer in the case $\frac{\mu}{\sigma} \ll 1$ (as expected by our analysis).

Questions: do we have a workaround?
What the FD can teach us?

The difference with the FD method is in the term σh^2 that

with FD leads to σu_i, with FE to $\frac{\sigma h}{6}(u_{i+1} + 4u_i + u_{i-1})$.

Notice that the reactive terms is given by the matrix

$$M_{ij} = \begin{cases} \frac{\sigma h}{6} & j = i \pm 1 \\ \frac{2}{3}h\sigma & j = i \\ 0 & j \neq i, i\pm 1 \end{cases}$$

$M_{ij} = \sigma \int_0^1 \varphi_i \varphi_j$ that is called MASS MATRIX (we will see why)

Also, we can notice that the integrand $f = \varphi_i \varphi_j$ is a quadratic term so a quadrature formula with Degree of Exactness ≥ 2 is enough to compute it exactly. Simpson for instance:

$$(\text{Simpson}) \int_a^b f = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$$

In fact:

$$\sigma \int_{x_i}^{x_{i+1}} \varphi_i \varphi_{i+1} \overset{\text{Simpson}}{\approx} = \frac{\sigma h}{6}\left(\cancel{f(x_i)}^{0} + 4f\left(\frac{x_i + x_{i+1}}{2}\right) + \cancel{f(x_{i+1})}^{0}\right) =$$

$$= \frac{\sigma h}{6}\frac{1}{h^2}\frac{h^2}{4}4 = \frac{\sigma h}{6} \quad \text{EXACT}$$

$$\sigma \int_{x_{i-1}}^{x_{i+1}} \varphi_i^2 = \sigma \int_{x_{i-1}}^{x_i} \varphi_i^2 + \sigma \int_{x_i}^{x_{i+1}} \varphi_i^2 \overset{(\text{Simpson})}{\approx} = \frac{\sigma h}{6}\left(\cancel{\varphi_i^2(x_{i-1})}^{0} + 4\varphi_i^2\left(\frac{x_i + x_{i-1}}{2}\right) + 2\varphi_i^2(x_i)\right)$$

$$+ 4\varphi_i^2\left(\frac{x_i + x_{i+1}}{2}\right) + \cancel{\varphi_i^2(x_{i+1})}^{0}\Big) =$$

$$= \frac{\sigma h}{6}\left(4\frac{1}{4} + 2 + 4\frac{1}{4}\right) = \frac{4\sigma h}{6} \quad \text{EXACT}$$

On the contrary, if we use a formula like the Trapezoidal one, we get a quadrature error:

$$(T) \int_x^y f = \frac{b-a}{2}\left(f(a) + f(b)\right) \qquad (DOE = 1)$$

$$\sigma \int_{x_{i-1}}^{x_i} \varphi_i \varphi_{i-1} \overset{(T)}{\approx} 0 \qquad \left(\varphi_i(x_{i-1}) = 0 \text{ and } \varphi_{i-1}(x_i) = 0\right)$$

$$\sigma \int_{x_{i-1}}^{x_{i+1}} \varphi_i^2 = \sigma \int_{x_{i-1}}^{x_i} \varphi_i^2 + \sigma \int_{x_i}^{x_{i+1}} \varphi_i^2 \overset{(T)}{\approx} \frac{\sigma h}{2}\left(\cancel{\varphi_i^2(x_{i-1})}^{0} + 2\varphi_i^2(x_i) + \cancel{\varphi_i^2(x_{i+1})}^{0}\right) =$$

$$= \sigma h$$

So, if we use the Trapezoidal Rule for FE (linear) we obtain:

$$-\frac{\mu}{h}\left(u_{i+1} - 2u_i + u_{i-1}\right) + \sigma h u_i = 0$$

that is

$$h\left(-\frac{\mu}{h^2}\left(u_{i+1} - 2u_i + u_{i-1}\right) + \sigma u_i\right) = 0$$

FD scheme!!!

For this problem : Trapezoidal + Linear FE $\equiv$ FD !!

How can we analyze this case?

We can look at the Trapezoidal integration as a perturbation of the bilinear form $a(\cdot,\cdot)$.

Keep in mind that, in general:

SIMPLE TRAPEZOIDAL : $\left| \int_a^b f - \frac{(b-a)}{2}\left(f(a)+f(b)\right) \right| =$

$$= \left| \int_a^b f(a) + \int_a^b f'(a)(x-a) + \int_a^b f''(\xi)\frac{(x-a)^2}{2} - \frac{(b-a)}{2}\left(f(a)+f(b)\right) \right| =$$

$$= \left| f(a)(b-a) + f'(a)\frac{(b-a)^2}{2} - \frac{(b-a)}{2}\left(f(a) + f(a) + f'(a)(b-a)\right) + \right.$$
$$\left. + f''(\xi)\frac{(b-a)^2}{2} + \int_a^b f''' \frac{(x-a)^2}{2} \right|$$

$$\sim O\left((b-a)^3\right)$$

COMPOSITE TRAPEZOIDAL : $\left| \int_a^b f - \sum(T) \int_{x_i}^{x_{i+1}} f \right| \sim O(h^2)$

In this way, we can state:

$$a(u,v) = \underbrace{\text{Trapezoidal}\left(a(u,v)\right)}_{a_h(u,v)} + O(h^2)$$

When we use the trapezoidal Rule, we fall into the class of Generalized Galerkin method with b_h and G_h scaling with $O(h^2)$.

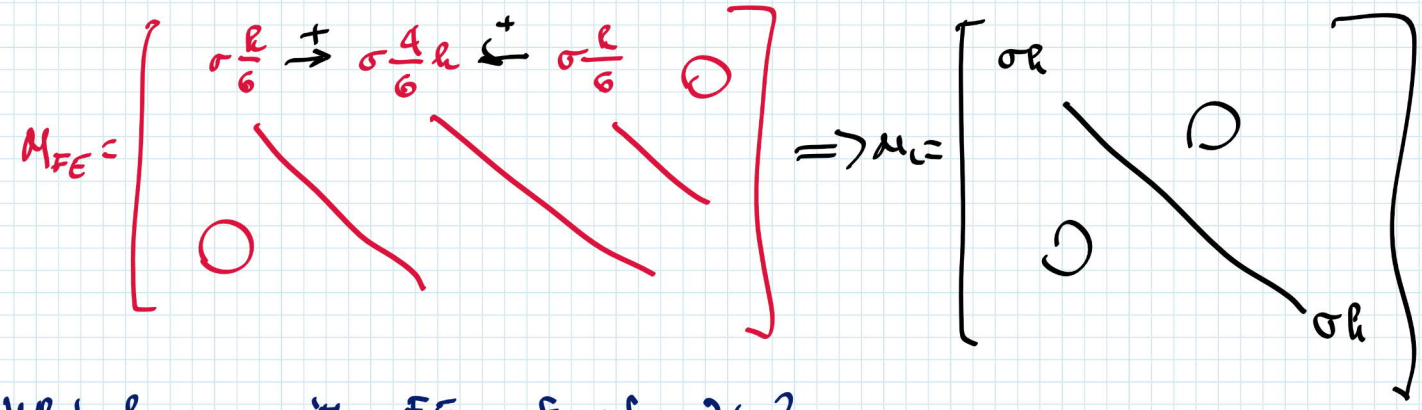
Then we have from the Strang Lemma:

$$\|u - u_h\|_V \leq \underbrace{C_1 h}_{\text{discretization}\ (p=1)} + \underbrace{C_2 h^2 + C_3 h^2}_{\text{quadrature}}$$

Our method will still be first order in the H^1-norm.
The behavior in the L^2-norm will be tested in our numerical session.

Linear FE + Trapezoidal Rule is called MASS LUMPING :
in fact, it is like summing up the off-diagonal elements to the diagonal

element of the mass-matrix:



What happens with FE of order 2+ ?

The mass-lumping doesn't work in the same way: trapezoidal rule would lead to a SINGULAR mass matrix.

However, mass-lumping strategies are available => see Quarteroni-Book.

WEEK 10

Advection - Diffusion - Reaction Problems in 2+ dimensions.

Quarteroni Chap. 12

Let's consider this problem:

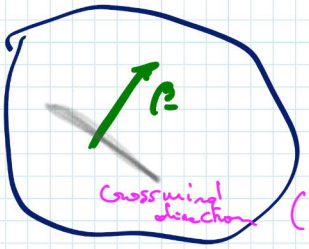
$$-\mu \Delta u + \underline{\beta} \cdot \nabla u \quad (+\sigma u) = f$$

$$u(\partial\Omega) = \text{data}.$$

We know already that:

(1) for $\|\underline{\beta}\| >> \mu$ we should expect instabilities (from the 1D case and the FD experience)

(2) the action of a possible numerical viscosity should not go in all the directions, as only the streamline direction (identified by $\underline{\beta}$) requires special treatments.



(orthogonal to $\underline{\beta}$)

Along the crosswind direction, we do not need additional viscosity.

How can we organize the "directional viscosity" in the variational formulation?

(To be honest, I think it's easier with the variational formulation than with the strong one).

Problem: $$-\nabla \cdot \left(\mu \nabla u + \underline{\beta} u\right) \quad (+\sigma u) = f$$

The artificial viscosity consists of adding the (isotropic) term:

$$-\nabla \cdot \left(\frac{\|\underline{\beta}\| h}{2} \nabla u \right)$$

but in fact we need a stabilization only along $\underline{\beta}$, so we should consider to add the term:

$$-\nabla \cdot \left(\delta h \, (\underline{\beta} \cdot \nabla u) \frac{\underline{\beta}}{\|\underline{\beta}\|} \right)$$

stabilization along $\underline{\beta}$

$\frac{\underline{\beta}}{\|\underline{\beta}\|}$ = unit vector in the direction of $\underline{\beta}$.

δ is an arbitrary parameter to be tuned manually.

In the variational form, this leads to:

$$\mu \int_\Omega \nabla u \nabla v + \int_\Omega (\underline{\beta} \cdot \nabla u) v + \boxed{\delta h \int_\Omega (\underline{\beta} \cdot \nabla u)\, \underline{\beta} \cdot \nabla v}$$

→ This is the term actually added (regardless of the B.C. when we use the

Streamline Diffusion (SD) method.

In this way, we resort to a Generalized Galerkin method

$$a_h(u_h, v_h) = \ell(v_h)$$

where $a_h(u_h, v_h) = a(u_h, v_h) + \frac{\delta h}{\|\beta\|}\int_\Omega (\beta\cdot\nabla u_h)(\beta\cdot\nabla v_h)$

To be explicit: $\frac{\delta h}{\|\beta\|}\int_\Omega (\beta\cdot\nabla u_h)(\beta\cdot\nabla v_h) =$

$$= \frac{\delta h}{\|\beta\|}\int_\Omega \left(\beta_1\frac{\partial u_h}{\partial x_1} + \beta_2\frac{\partial u_h}{\partial x_2}\right)\left(\beta_1\frac{\partial v_h}{\partial x_1} + \beta_2\frac{\partial v_h}{\partial x_2}\right) =$$

$$= \frac{\delta h}{\|\beta\|}\int_\Omega \beta_1^2\frac{\partial u_h}{\partial x_1}\frac{\partial v_h}{\partial x_1} + \beta_1\beta_2\left(\frac{\partial u_h}{\partial x_1}\frac{\partial v_h}{\partial x_2} + \frac{\partial u_h}{\partial x_2}\frac{\partial v_h}{\partial x_1}\right) + \beta_2^2\frac{\partial u_h}{\partial x_2}\frac{\partial v_h}{\partial x_2}$$

Notice that if $\beta_2 = 0$, we get exactly the stabilization we expect only along x_1 (and the other way around for $\beta_1 = 0$).

The Streamline Diffusion method is very popular, it looses the strong consistency of Galerkin and it can be analyzed by the Strang Lemma.

We find straightforwardly that the method is 1^{st} order regardless of the polynomial degree of finite elements.

On the other hand, if $a(u,v)$ is coercive,

$$a_h(u,v) = a(u,v) + \frac{\delta h}{\|\beta\|}\int_\Omega (\beta\cdot\nabla u)(\beta\cdot\nabla v)$$

(with $\delta > 0$) is trivially (more) coercive, with a coercivity constant

$$\alpha_{SD} = \alpha + C\delta h$$

so that by an appropriate tuning of δ we can avoid spurious oscillations.

The natural question however is: can we obtain a stable high order (>1) method?

Instead of tuning a different coefficient δ (as with Scharfetter–Gummel) we change paradigm:

CAN WE FIND A STRONGLY CONSISTEN STABLE METHOD?

To answer this question, we need some preliminary definitions.

Symmetric and Skew-symmetric parts of an operator.

We already know that if a bilinear form is such that

$$a(u,v) = a(v,u) \qquad \forall u,v \in V$$

we say that it is <u>symmetric</u>.

We also say that if a bilinear form is s.t.

$$a(u,v) = -a(v,u)$$

we say that it is <u>skew-symmetric</u>

In linear algebra, similarly, we have symmetric and skew-symmetric matrices:

$$A = A^T \quad \text{(symmetric)}$$

$$A = -A^T \quad \text{(skew-symmetric)}$$

Now, in linear algebra, for a generic matrix A, we have:

$$A = \underbrace{\frac{1}{2}(A + A^T)}_{\text{symmetric}} + \underbrace{\frac{1}{2}(A - A^T)}_{\text{skew-symmetric}}$$

Do we have the same for bilinear forms?

Let's work on the strong formulation:

$$\underbrace{-\mu \Delta u}_{\int_\Omega \mu \nabla u \cdot \nabla v \text{ (symmetric)}} + \underbrace{\underline{\beta} \cdot \nabla u}_{?} + \underbrace{\sigma u}_{\int_\Omega \sigma u v \text{ (symmetric)}} = f$$

$$\underbrace{\int_\Omega \underline{\beta} \cdot \nabla u \, v}_{a_C(u,v)} = \int_\Gamma \underline{\beta} \cdot \underline{n} \, u v - \int_\Omega u \nabla \cdot (\underline{\beta} v) = \text{Boundary term} - \int_\Omega (\nabla \cdot \underline{\beta}) u v - \underbrace{\int_\Omega (\underline{\beta} \cdot \nabla v) u}_{a_C(v,u)}$$

$\Downarrow$ Symmetric

So, from here and dropping the boundary term:

$$\underbrace{\int_\Omega (\underline{\beta} \cdot \nabla u) v + \frac{1}{2} \int_\Omega (\nabla \cdot \underline{\beta}) u v}_{\text{skew symmetric}} = -\int_\Omega (\underline{\beta} \cdot \nabla v) u - \frac{1}{2} \int_\Omega (\nabla \cdot \underline{\beta}) u v$$

$$\Rightarrow \quad \int_\Omega (\underline{\beta}\cdot\nabla u)v = \underbrace{\int_\Omega (\underline{\beta}\cdot\nabla u)v + \frac{1}{2}\int (\nabla\cdot\underline{\beta})uv}_{\text{Skew Symmetric}} \underbrace{- \frac{1}{2}\int (\nabla\cdot\underline{\beta})uv}_{\text{Symmetric}}$$

REMARK

In the book of Quarteroni, you find the convective term in the form:

$$\int_\Omega \nabla\cdot(\underline{\beta}u)v$$

In this case:

$$\int_\Omega \nabla\cdot(\underline{\beta}u)v = \int_{\partial\Omega} \underline{\beta}\cdot\underline{n}\,uv - \int (\underline{\beta}\cdot\nabla v)u = \text{Boundary Term} - \int_\Omega \nabla\cdot(\underline{\beta}u)u + \int_\Omega (\nabla\cdot\underline{\beta})uv$$

Again, skipping the boundary term:

$$\int_\Omega \nabla\cdot(\underline{\beta}u)v - \frac{1}{2}\int (\nabla\cdot\underline{\beta})uv = -\int_\Omega \nabla\cdot(\underline{\beta}v)u + \frac{1}{2}\int (\nabla\cdot\underline{\beta})uv$$

Skew symmetric:

$$\int_\Omega \nabla\cdot(\underline{\beta}u)v = \underbrace{\int_\Omega \nabla\cdot(\underline{\beta}u)v - \frac{1}{2}\int (\nabla\cdot\underline{\beta})uv}_{SS} + \underbrace{\frac{1}{2}\int (\nabla\cdot\underline{\beta})uv}_{S}$$

In strong form:

$$L = -\mu\Delta u + (\underline{\beta}\cdot\nabla u) + \sigma u = \overbrace{-\mu\Delta u + \sigma u - \frac{1}{2}(\nabla\cdot\underline{\beta})u}^{L_S} + \underbrace{(\underline{\beta}\cdot\nabla u) + \frac{1}{2}(\nabla\cdot\underline{\beta})u}_{L_{SS}}$$

or

$$L = -\mu\Delta u + \nabla\cdot(\underline{\beta}u) + \sigma u = \overbrace{-\mu\Delta u + \sigma u + \frac{1}{2}(\nabla\cdot\underline{\beta})u}^{L_S} + \underbrace{\nabla\cdot(\underline{\beta}u) - \frac{1}{2}(\nabla\cdot\underline{\beta})u}_{L_{SS}}$$

Notice that $\nabla\cdot(\underline{\beta}u) - \frac{1}{2}(\nabla\underline{\beta})u = \frac{1}{2}\nabla\cdot(\underline{\beta}u) + \frac{1}{2}\underline{\beta}\cdot\nabla u$

So also for differential operators (or bilinear forms), we can identify a symmetric and a skew-symmetric part.

Strongly Consistent Stabilization Methods

Let's try to "design" a stabilization method from scratch.

$$a_h(u_h, v_h) + b_h(u_h, v_h) = f(v_h) + g_h(v_h)$$

To maintain the strong consistency, we need that for u, exact solution:

$$(*) \quad b_h(u, v_h) - g_h(v_h) = 0 \qquad \forall v_h \in V_h$$

But we have also:

$$(**) \quad a(u, v_h) - f(v_h) = 0.$$

So $(*)$ should be related to $(**)$ but how?

Clearly, $(*)$ cannot be $(**)$ because this is strongly consistent but add no stabilization.

(1) → What if we consider the $(**)$ in its strong form?

$$\int_\Omega (-\mu \Delta u + \underline{\beta} \cdot \nabla u + \sigma u - f) v \qquad (***)$$

This is formally not correct, because the functions are not regular enough! (In general, u is not H^2).

However, recall that the SD was $\int_\Omega (\underline{\beta} \cdot \nabla u)\, \underline{\beta} \cdot \nabla v$. This means that if we choose as test function $(\underline{\beta} \cdot \nabla v)$, we have a potentially stabilizing term.

(2) OK, let's break $(***)$ by elements:

$$\sum_{T_K} \delta_{Ku} \int_{T_K} (-\mu \Delta u_h + \underline{\beta} \cdot \nabla u_h + \sigma u_h - f)\, w_h$$

Formally this is correct because on each element the functions are regular.

We are close to the final solution.

Let's define $b_h(u_h, v_h) - g_h(v_h) = \sum_K \delta_{Ku} \int_{T_K} \big((L_s + L_{ss}) u_h - f\big)\big(L_{ss} v_h + \rho L_s v_h\big)$

where δ_{Ku} is a parameter (obtained by the user), ρ is a parameter with values $\rho = 0, 1, -1$.

Notice that:

(1) The terms is strongly consistent, being proportional to the residual (elementwise, in its strong form).

(2) The method includes the SD, since $\sum_K \delta_K \int_{T_K} L_{SS} u_h \cdot L_{SS} v_h$ includes exactly the SD terms.

We are legitimated to hope that this is a strongly consistent stabilization.

For $\beta = 0$ this is called SUPG (Streamline Upwind Petrov-Galerkin)

For $\beta = 1$ " " " GaLS (Galerkin Least Squares)

For $\beta = -1$ " " " DW (Douglas-Wang).

Remark

For piecewise linear finite elements $-\Delta u_h = 0$ on each element, so the actual terms are simpler.

For instance, in the case of $\nabla \cdot \beta = 0$ with linear fem we have:

$$a(u_h, v_h) + \sum_{T_K} \delta_K \int_{T_K} \left(\underbrace{\beta \cdot \nabla u_h}_{\text{SD}} + \overbrace{\sigma u_h - f}^{\text{strongly consistent terms}}, \underbrace{\beta \cdot \nabla v_h}_{\text{SD}} + \beta \sigma v_h \right) = \int_\Omega f v_h$$

Analysis of Strongly Consistent Methods

The strong consistency should enable to get results similar to the ones for the regular Galerkin approach. This is true, but we need to introduce the right norms (the strong consistency is only elementwise).

Let's consider the case of GaLS ($\beta = 1$)

We may consider the bilinear form:

$$a_{GaLS}(u_h, v_h) = a(u_h, v_h) + \sum_{T_K} \delta_K \left(L u_h, L v_h \right)_{L^2(T_K)}$$

It is possible to prove the following results:

(1) a_{GaLS} is continuous in V_h

(2) $a_{GaLS}(u_h, u_h) \geq \mu_0 \| \nabla u_h \|_{L^2}^2 + \gamma_0 \| u_h \|_{L^2}^2 + \sum \delta_K \| L u_h \|_{L^2}^2$

$\mu_0 = \min_{x \in \Omega} \mu(x)$ $\quad \gamma_0 = \min \left((\nabla \cdot \beta) + \frac{1}{2} \sigma \right)$

Consequently, we can introduce the norm:

$$\|u_h\|^2_{GALS} = \mu_0 \|\nabla u_h\|^2_{L^2} + \gamma_0 \|u_h\|^2_{L^2} + \sum_K \delta_K \|L u_h\|^2_{L^2(T_K)}$$

The GALS problem is continuous and coercive w.r.t. this norm, so that it is well posed.

Moreover, when $Pe > 1$, for an appropriate choice of δ_K, we can prove:

$$\|u - u_h\|_{GALS} \le C h^{k+\frac{1}{2}} |u|_{H^{k+1}} \qquad k = \min(p, s)$$

In this norm, the accuracy depends again on the degree of the FE polynomials.

The parameters δ_K depend on the constant C_0 of the inverse inequality

$$\sum h_K^2 \|\Delta u_h\|^2_{L^2(T_K)} \le C_0 \|\nabla u_h\|^2_{L^2(\Omega)}$$

This is called "inverse inequality" because while in general in $H^s(\Omega)$ we have the second derivative bounding the first one with an h at the denominator, in discrete spaces V_h we can prove it.

For SUPG and DW similar results can be proven (but much more complicated). ⇒ See the book Quarteroni-Valli, Chap. 8

REMARK 1

Also in dimension 2+ we can interpret the stabilization methods as Petrov-Galerkin methods (oblique projections) where

$$W_h = V_h \oplus \text{Bubbles}.$$

This interpretation leads to different stabilization methods called "subgrid viscosity". Find a description in the excellent but advanced book:

Ern, Guermond, Theory and Practice of Finite Elements, Springer

REMARK 2

Reaction Dominated problems give similar oscillations as in 1D also in multiple dimensions.

The Quadrature Formula in 2D on a triangle:

$$\int_D f = \frac{1}{3}\left(f(P_0) + f(P_1) + f(P_2)\right)|A| \qquad \text{where } |A| = \text{area of the triangle}$$

or in 3D on a tetrahedron

$$\int_\Delta f = \frac{1}{4}\left(f(P_0) + f(P_1) + f(P_2) + f(P_3)\right)|V| \qquad \text{where } |V| = \text{volume of the } T$$

can be used for the mass matrix with linear finite elements.

(in general, on a simplex in n-dimensions:

$$\int_{S^n} f \approx \frac{1}{n+1}\left(\sum_{i=0}^{n} f(P_i)\right)|V| \qquad \left(\text{where } |V| = \int_{S^n} 1\right)$$

Clearly, (linear) finite element functions are <u>orthogonal</u> with respect to the scalar product induced by this quadrature formula, and the resulting matrix is <u>diagonal</u>

$$(T)\int_{\Omega} \varphi_i \varphi_j \begin{cases} = 0 & i \neq j \\ \neq 0 & i = j \end{cases}$$

This fixes the oscillations.

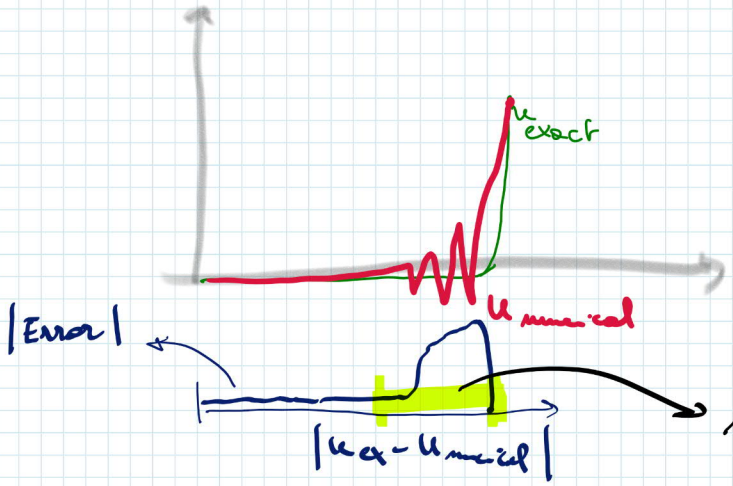
Other (more sophisticated) strategies are needed for higher order elements.

Stabilization and Grid-Adaptivity

An alternative or complementary approach to the Advection/Reaction dominated problem is to use Grid-Adaptivity.

Since FEM are strongly Grid-Dependent, we may decide to select the mesh size h in a non-uniform way, so that we can adapt it to the Péclet number.

If $\underline{\beta}$ is function of space and we estimate a large local error, we refine h so to make the local $Pe < 1$.



refine h here until we eliminate locally the oscillations.

to do this, we need: (i) a good error estimator

(ii) mesh refinement / coarsening methods.

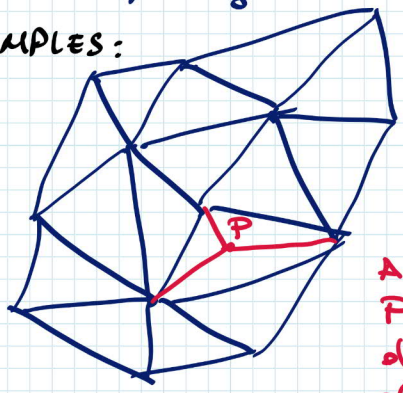
(1) The error estimation theory is huge.

In MATH 516 we consider sometimes some adaptivity and local error estimators are generally based on two solutions with different accuracy. Other approach, like the so-called "goal-oriented" ones are very popular (see e.g. Zienkiewicz-Zhu).

See specific literature (Strikwerda, Ramacher, Bräck, ecc.)

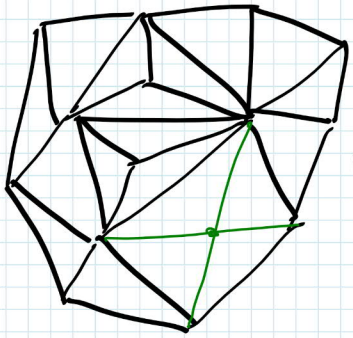
(2) Refinement and coarsening are trivial only in 1D.
In 2+ dimensions, adding or removing a node needs to be done respecting the conformity constraints

EXAMPLES:



(a) Refinement

P
Adding the node P requires the correct definition and numbering of the elements.



(b) Coarsening

What is the right node to eliminate for coarsening?
After removal, how we organize the elements?

Keep in mind that the great numerical analyst Michel Fortin stated that the mesh should not be considered as given, but as an unknown of the problem.

REMARK

It is well known that for some differential problems we have a principle of maximum or minimum. This means that we know that maxima or minima are attained only at the boundary.
For the Laplace problem $-\Delta u = 0$, max and min are on $\partial\Omega$.

When we get to the numerical solution, having a similar principle may be useful, because this guarantees that the approximate solution cannot be negative if it is supposed to be positive, if it vanishes on the boundary for instance.

This requires that in the linear system

$$A\underline{u} = \underline{f}$$

A fulfills some properties that guarantee

the monotonicity of the solution.

This leads to the concept of M-matrix (i.e. $a_{ij} \leq 0 \quad i \neq j$, $Re(\lambda_i) > 0$)

The construction of a non-oscillatory solution can be characterized by having at the end of the discretization a M-matrix.

Week 11

In this week we will learn how to solve the problem

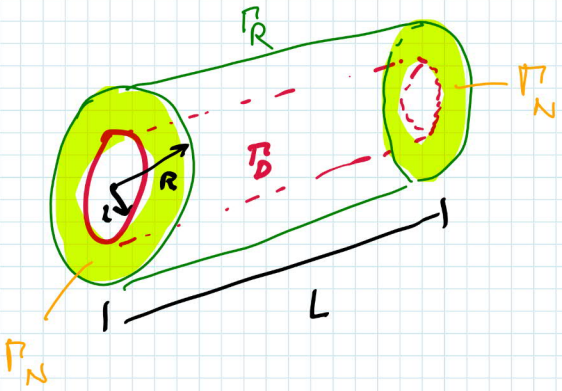
$$-\nabla\cdot(K\nabla u) = 0 \qquad \text{in } \Omega$$

$$u(\Gamma_D) = u_w$$

$$K\nabla u \cdot \underline{n}\,(\Gamma_N) = 0$$

$$K\nabla u \cdot \underline{n} + \alpha u = \alpha u_{ext} \quad (\Gamma_R)$$

Ω :



Data

$R = 0.025$ m

$r = 0.015$ m

$L = 0.1$ m

$u_w = 60$ °C

$\alpha = 10 \frac{kg}{s^3 °C}$

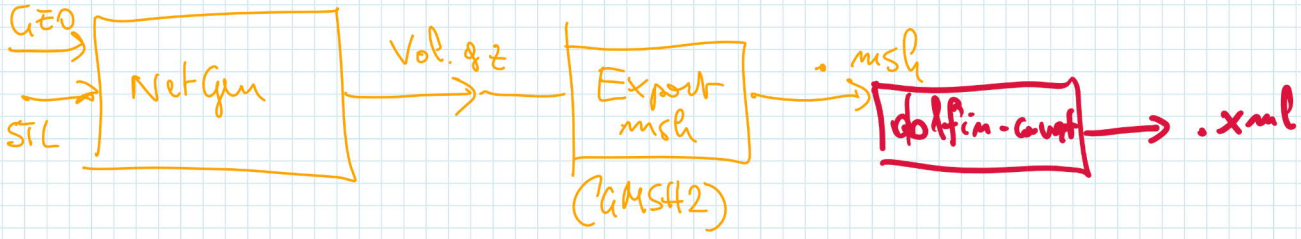
$K = 0.05 \frac{kg\,m}{s^3 °C}$

$u_{ext} = 15$ °C

Preprocessing

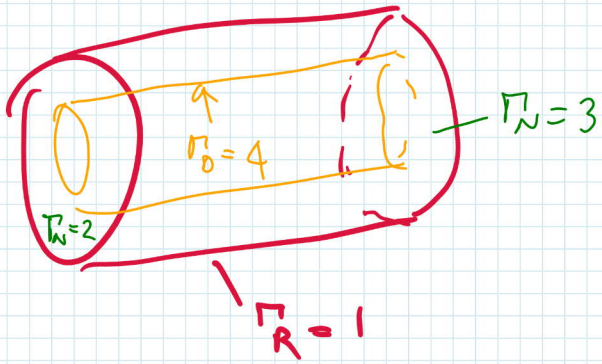
NetGen is a software for meshing with some elementary geometrical subroutines. Generally, an input file is in the STL format, but the constructive geometry format is given in files with extension .geo.

In our case, the geometry is given by two cylinders, internal and external and the internal is subtracted to the external one.



After the mesh is generated with NetGen you need to:

(a) Edit the Boundary Conditions (name "Mesh")



Hint

Use the inverted lexicographic order:

$$D < N < R$$

↑ Highest index

↑ Lowest Index

In fact, Dirichlet conditions are the LAST ONES to be processed in your solver (they are the only ones that affect the pattern of the matrix).

(b) Export to the FENICS format using the GMSH format .msh as an intermediary step.

. GEO FILE:

```
#
## a cylinder
#
algebraic3d

# cut cylinder by planes:

solid extcyl = cylinder ( 3, 0, 0; -1, 0, 0; 0.025 )
     and plane (0, 0, 0; -1, 0, 0)
     and plane (0.1, 0, 0; 1, 0, 0);

solid intcyl = cylinder ( 3, 0, 0; -1, 0, 0; 0.015 )
     and plane (-0.1, 0, 0; -1, 0, 0)
     and plane (0.2, 0, 0; 1, 0, 0);

solid insul = extcyl and not intcyl;

tlo insul;
```

Week 12

Quarteroni Chap. 5

Parabolic Problems with FEM

Variational Formulation and estimates

In this chapter, we consider problems like:

$$\frac{\partial u}{\partial t} - \mu \Delta u + \underline{\beta} \cdot \nabla u + \sigma u = f \qquad \text{in } \Omega \times [0,T]$$

+ B.C. like:

$$u(\Gamma_D, t) = u_D \,, \quad \mu \nabla u \cdot \underline{n}(\Gamma_N, t) = \phi_N$$

$$\mu \nabla u \cdot \underline{n} + \chi u(\Gamma_R, t) = \phi_R$$

+ I.C.

$$u(\underline{x}, 0) = u_0(\underline{x}) \qquad \text{where } \underline{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

To study this problem, already investigated with the Finite-Difference Method, with Finite Elements, we need to give a variational formulation.

Notice that in 1D we already gave a variational formulation, without mentioning it. Here, we generalize to multiple dimensions and with the right functional spaces. The Quarteroni's book is only for $\underline{\beta} = 0$, $\sigma = 0$, here we carry out a more general analysis.

Let's proceed formally. For the moment, we assume to have only Dirichlet conditions,

$$u(\partial\Omega) = 0$$

So, we postulate $u \in H_0^1(\Omega)$ and we specify the regularity in time later.

Let $v \in H_0^1(\Omega)$ function, time-independent. With the usual steps, we multiply the equation by v, integrate over Ω.

$$\boxed{\int_\Omega \frac{\partial u}{\partial t} v} + \int_\Omega \mu \nabla u \nabla v + \int_\Omega \underline{\beta} \cdot \nabla u \, v + \int_\Omega \sigma u v = \int_\Omega f v$$

$$u(\underline{x}, 0) = u_0(\underline{x})$$

↳ if Ω is time-independent:

$$\frac{d}{dt} \int_\Omega u v$$

(if Ω is not time-dependent, we need a thing called Reynolds Theorem:

$$\int_{\Omega_t} \frac{\partial u}{\partial t} v = \frac{d}{dt} \int_{\Omega_t} u v - \int_{\partial\Omega_t} \underline{w} \cdot \underline{n} \, u v \qquad \underline{w} = \frac{\partial(\partial\Omega_t)}{\partial t}$$

From now on Ω is time-invariant.

So we can write the problem as:

Find $u \in H_0^1(\Omega)$ (time dependent to obtain) s.t.

$$\frac{d}{dt}\int_\Omega uv + a(u,v) = f(v) \qquad \forall v \in H_0^1(\Omega) \quad (*)$$

$$f(v) \to \int_\Omega fv$$

$$+ \text{I.C.}$$

Lax-Milgram for Parabolic Problems

A sufficient condition for (*) to be well-posed is the continuity of the bilinear form and functionals **and**:

$$\exists \lambda > 0, \ \exists \alpha > 0 \quad \text{s.t.} \quad a(u,u) + \lambda \|u\|_{L^2}^2 \geq \alpha \|u\|_{H^1}^2$$

This condition is called WEAK COERCIVITY, for $\lambda = 0$ we get the old coercivity, this is less restrictive.

REMARK Notice that a problem like $\frac{\partial u}{\partial t} - \mu \Delta u - |\sigma| u = f$ in this case is immediately well posed (while with no time-derivative, the Lax-Milgram lemma requires some assumptions on the coefficients)

To complete the picture, we need to specify the functional spaces.

Let us integrate in time the (*):

$$\int_0^T \frac{d}{dt}\int_\Omega uv + \int_0^T a(u,v) = \int_0^T \int_\Omega fv \longrightarrow$$

we need $f \in L^2$ in both time and space

$$\int_\Omega u(x,T)v - \int_\Omega u_0(x)v$$

$$\left(u(T), v\right)_{L^2} - \left(u_0, v\right)_{L^2}$$

INTUITION: we need $\|u\|_{L^2}(t)$ to be bounded

→ INTUITION: we need the $\|u\|_{H^1}(t)$ to be $L^2(0,T)$

We will confirm the intuitions with the estimates on the solution.

For now, let's introduce the spaces:

$L^2(0,T; H_0^1(\Omega))$ = functions g with $\|g\|_{H^1}(t) \in L^2(0,T)$ (and trace null on $\partial\Omega$)

$L^\infty(0,T; L^2(\Omega))$ = " g with $\|g\|_{L^2(\Omega)}(t) \in L^\infty(0,T)$.

$L^2(0,T;L^2(\Omega))$ = functions g s.t. $\|g\|_{L^2(\Omega)} \in L^2(0,T)$.

VARIATIONAL FORMULATION with nonhomogeneous Dirichlet conditions.

Find $u \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T; \underbrace{\ell}_{(\text{lift of } g)} + H^1_{\Gamma_D}(\Omega))$ s.t.

given $f \in L^2(0,T;L^2(\Omega))$

$u_0(x) \in L^2(\Omega)$

$u_g(\Gamma_D) \in L^2(0,T; H^{1/2}(\partial\Omega))$

$$\frac{d}{dt}(u,v)_{L^2(\Omega)} + a(u,v) = f(v) \qquad \forall v \in H^1_{\Gamma_D}(\Omega)$$

$$u(0) = u_0$$

or

$$(u(T),v)_{L^2(\Omega)} + \int_0^T a(u,v) = \int_0^T f(v) + (u_0,v)_{L^2(\Omega)}$$

A-priori estimates

To get a-priori estimates, we need to go back to the form:

$$\int_\Omega \frac{\partial u}{\partial t} v + a(u,v) = \int_\Omega f v$$

Now, take $v = u$ (THIS IS TIME-DEPENDENT!):

$$\int_\Omega \frac{\partial u}{\partial t} u + a(u,u) = \int_\Omega f u$$

Now: $\frac{\partial u}{\partial t} u = \frac{1}{2}\frac{\partial (u^2)}{\partial t}$ so $\int_\Omega \frac{\partial u}{\partial t} u = \frac{1}{2}\frac{d}{dt}\|u\|^2_{L^2(\Omega)}$

We have therefore:

$$\frac{1}{2}\frac{d}{dt}\|u\|^2_{L^2} + a(u,u) = \int_\Omega f u$$

If we integrate in time:

$$\|u\|^2_{L^2}(T) + 2\int_0^T a(u,u) = 2\int_0^T \int_\Omega f u + \|u_0\|^2_{L^2(\Omega)}$$

Let's recall the Young Inequality: for a, b given, $\varepsilon > 0$ arbitrary

$$0 \le \left(\varepsilon a - \frac{b}{2\varepsilon}\right)^2 = \varepsilon^2 a^2 + \frac{b^2}{4\varepsilon^2} - ab$$

$$ab \le \varepsilon^2 a^2 + \frac{b^2}{4\varepsilon^2}$$

Then:

$$\int_\Omega f u \le \varepsilon^2 \|u\|^2_{L^2(\Omega)} + \frac{1}{4\varepsilon^2} \|f\|^2_{L^2(\Omega)}$$

So we can write (under weak coercivity)

$$\frac{1}{2}\frac{d}{dt}\|u\|^2_{L^2(\Omega)} + \alpha \|u\|^2_{H^1(\Omega)} - \lambda \|u\|^2_{L^2(\Omega)} \le \varepsilon^2 \|u\|^2_{L^2(\Omega)} + \frac{1}{4\varepsilon^2}\|f\|^2_{L^2(\Omega)}$$

If we can find ε small enough s.t. $\lambda + \varepsilon^2 < \alpha$, we have:

$$\|u^2\|_{L^2}(T) + 2(\alpha - (\lambda+\varepsilon^2)) \int_0^T \|u\|^2_{H^1} \le \frac{1}{4\varepsilon^2}\int_0^T \|f\|^2_{L^2(\Omega)} + \|u_0\|^2_{L^2(\Omega)}$$

For $\alpha - (\lambda+\varepsilon^2) \ge k > 0$ we immediately get:

$$\|u\|_{L^\infty(L^2)} \le \text{data} \qquad (T \text{ can be replaced with any instant})$$
$$\|u\|_{L^2(H^1)} \le \text{data}$$

For a bilinear form coercive ($\lambda = 0$) this is an option: we can choose $\varepsilon^2 = \frac{\alpha}{2}$).

In general, if we cannot say that $\alpha - (\lambda+\varepsilon^2) > 0$, then we can use the Gronwall Lemma:

$$\|u\|^2_{L^2(\Omega)}(t) + 2\alpha \int_0^t \|u\|^2_{H^1} \le \int_0^t \|f\|^2 + \int_0^t (\lambda+1)\|u\|^2_{L^2} d\tau + \|u_0\|^2_{L^2}$$

Gronwall Lemma:

$$g(t) \le \alpha + \int_0^t \rho\, g(\tau) d\tau \Rightarrow g(t) \le \alpha + \int_0^t \alpha\, \rho\, e^{\int_\tau^t \beta} d\tau$$

($\rho \ge 0$)

REMARK

Notice that for an i.c. in $L^2(\Omega)$, the solution for $t > 0 \in L^2(H^1)$, so it is more regular: the diffusion operator gives more regularity to the solution (discontinuities are smoothened).

Semi-Discretization in Space

Let's start using FEM for the derivative in space.
To do this, we postulate for the solution a form similar to the one we use with the separation of variables in 1D:

$$u(x,t) = \sum_{j=1}^{\infty} X_j(x)\, T_j(t) \qquad (*)$$

Here we assume:

$$u_h(\underline{x},t) = \sum_{j=1}^{N} u_j(t)\, \varphi_j(\underline{x}) \qquad (**)$$

where $\varphi_j(\underline{x})$ are the piecewise polynomial (Lagrange) functions
$u_j(t)$ are functions to find.

The major differences between $(*)$ and $(**)$ are that:

(i) $(*)$ is a series, $(**)$ is a FINITE sum

(ii) the X_j are eigenfunctions of the problem, here the $\varphi_j(\underline{x})$ are generic polynomial functions with no specific shape for the problem at hand.

At this point, we can use $(**)$ into our problem in the form:

$$\frac{d}{dt}(u_h, v_h)_{L^2} + a(u_h, v_h) = f(v_h) \qquad \text{where } v_h \text{ is a finite element function in } V_h$$

More precisely, we write:

$$\sum_j^N \frac{d}{dt}\left(u_j(t)\,(\varphi_j, \varphi_i)\right)_{L^2} + \sum_j^N u_j(t)\, a(\varphi_j, \varphi_i) = f(\varphi_i) \qquad i = 1 \dots N$$

that becomes:

$$\sum_j^N (\varphi_j, \varphi_i)_{L^2} \frac{du_j}{dt} + \sum_j^N a(\varphi_j, \varphi_i)\, u_j = f(\varphi_i) \qquad i = 1 \dots N$$

At the bottom line, this reads

$$M \frac{d\underline{u}}{dt} + A\,\underline{u} = \underline{b}$$

where $\underline{u} = [u_j]$, M is the mass matrix $[(\varphi_j, \varphi_i)]$,
A is the stiffness matrix $[a(\varphi_j, \varphi_i)]$
$\underline{b}$ is the vector $[f(\varphi_i)]$.

Not surprisingly, we obtain an ordinary differential equation system.

Remark The name "mass matrix" comes from here, because typically the time-derivative comes from "inertial terms" (e.g.: mass × acceleration when $\underline{u}$ is a velocity).

Full Discretization

At this point, we can use our Finite difference methods for the time discretization.

However, methods are typically written for $\frac{d\underline{y}}{dt} = B\underline{y} + \underline{c}$

Here we have the mass matrix:

$$M\frac{d\underline{u}}{dt} = -A\underline{u} + \underline{b} \quad \text{with } \underline{u}_0 = \text{approximation of } u_0 \text{ in } V_h$$

(for instance $\underline{u}_0 = [u_0(x_j)]$)

Two options:

(1) The mass matrix is s.p.d., so we can invert it (at least formally, not in practice)

$$\frac{d\underline{u}}{dt} = -M^{-1}A\underline{u} + M^{-1}\underline{b}$$
$$= B\underline{u} + \underline{c}$$

At this point, we have $\approx$ 200 of methods: Runge-Kutta, Adams, etc.

For instance, we can use Forward Euler:

→ time step Δt
→ collocation instants t^n

$$\left.\begin{array}{l}\\ \end{array}\right\} \frac{1}{\Delta t}(\underline{u}^{n+1} - \underline{u}^n) = B\underline{u}^n + c^n$$

in practice, this reads:

$$M\underline{u}^{n+1} = (M - \Delta t A)\underline{u}^n + \Delta t \underline{b}^n$$

Backward Euler:

$$(M + \Delta t A)\underline{u}^{n+1} = \Delta t \underline{b}^n$$

(here, we assume A to be time-independent, but this is not necessarily true)

Heun: recall that this is the most basic RK method:

$$\frac{1}{\Delta t}(\underline{u}^{n+1} - \underline{u}^n) = \frac{1}{2}\left(B\underline{u}^n + B\left(\underline{u}^n + \Delta t B\underline{u}^n + \Delta t c^n\right)\right) + \frac{1}{2}\left(\underline{c}^{n+1} + \underline{c}^n\right)$$

(Crank-Nicolson + Forward Euler)

This reads

$$M\underline{u}^{n+1} = \left(M + \Delta t A + \frac{1}{2}\Delta t^2 A M^{-1} A\right)\underline{u}^n + \frac{\Delta t^2}{2} A M^{-1}\underline{b}^n + \frac{1}{2}\left(\underline{b}^{n+1} + \underline{b}^n\right)$$

In practice, we set:

$$\begin{cases} M\tilde{\underline{u}} = (M - \Delta t A)\underline{u}^n + \Delta t \underline{b}^n \\ M\underline{u}^{n+1} = \frac{1}{2}\left(A\underline{u}^n + A\tilde{\underline{u}}\right) + \frac{1}{2}\left(\underline{b}^{n+1} + \underline{b}^n\right) \end{cases}$$

Notice that with the Finite Difference method, we have $M = I$, so an explicit method reads

$$\underline{u}^{n+1} = (I - \Delta t A)\underline{u}^n + \Delta t \underline{b}^n \qquad \text{(Forward Euler)}$$

$$\tilde{\underline{u}} = (I - \Delta t A)\underline{u}^n + \Delta t \underline{b}^n$$
$$\underline{u}^{n+1} = \frac{\Delta t}{2} A(\underline{u}^n + \tilde{\underline{u}}) + \frac{\Delta t}{2}(\underline{b}^{n+1} + \underline{b}^n) \qquad \text{(Heun)}$$

In this case, we do not have any system to solve. In the FEM case, we have N_t systems to solve with the matrix M. The mass matrix is s.p.d., so it can be solved easily with CG, however it is an additional cost. The FEM is computationally more expensive.

(ii) Second Option: if we use the MASS LUMPING (Linear FEM), the problem of explicit methods is solved, as M_L is diagonal, so the linear systems are trivially solved.

⇒ The mass lumping is useful not only for reaction dominated problems, but also for time-dependent problems.

The Θ-method

As done for FD, we will focus on the Θ-method.

We get to the method by using a quadrature formula:

$$M\int_{t^n}^{t^{n+1}} \frac{d\underline{u}}{dt} = M(\underline{u}^{n+1} - \underline{u}^n) = \int_{t^n}^{t^{n+1}} \underline{g} \approx \left(\Theta \underline{g}^{n+1} + (1-\Theta)\underline{g}^n\right)\Delta t$$

we get:

$$(M + \Delta t \Theta A)\underline{u}^{n+1} = (M - \Delta t(1-\Theta)A)\underline{u}^n + \Delta t(\Theta \underline{b}^{n+1} + (1-\Theta)\underline{b}^n)$$

Remark

What the weak coercivity is enough?

Let us consider the following problem - weakly coercive:

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) = f(v)$$

Let's change the unknown: $u \longrightarrow e^{-\lambda t}u \equiv w \qquad (\lambda > 0)$

We have then:

$$e^{-\lambda t}\left(\frac{\partial u}{\partial t}, v\right) + e^{-\lambda t}a(u,v) = e^{-\lambda t}f(v) \equiv g(v)$$

$$\left(\frac{\partial w}{\partial t}, v\right) + \underbrace{a(w,v) + \lambda(w,v)}_{\tilde{a}(w,v)} = g(v) \qquad \left(\frac{\partial w}{\partial t} = e^{-\lambda t}\frac{\partial u}{\partial t} - \lambda e^{-\lambda t} u\right)$$

$$\left.\begin{aligned} \tilde{a}(w,w) &= \underbrace{a(w,w)}_{\geq \alpha \|w\|_H^2 - \lambda \|w\|_{L^2}^2} + \lambda \|w\|_{L^2}^2 \end{aligned}\right\} \geq \alpha \|w\|_{L^2}^2$$

Formally, a weakly coercive problem can be reformulated into a coercive one (rounding errors here do not matter).

Analysis of the Semi-discrete problem

Based on the previous remark, we can focus on a coercive problem (in the traditional sense).

Now, we have the following problem:

Find $u \in L^2(H^1) \cap L^\infty(L^2)$:

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) = f(v) \qquad \forall v \in V$$

and

Find $u_h \in L^2(V_h) \cap L^\infty(L^2)$:

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) = f(v_h) \qquad \forall v_h \in V_h$$

After taking v_h in the first one and subtraction membre-a-membre, we get:

$$\left(\frac{\partial e_h}{\partial t}, v_h\right) + a(e_h, v_h) = 0 \qquad \forall v_h \in V_h.$$
$$e_h(0) = u_0 - u_{h0} = e_{h0}$$

where $e_h \equiv u - u_h$

Now we have:

$$\alpha \|e_h\|_{H^1}^2 \leq a(e_h, e_h) = a(e_h, u - w_h) + a(e_h, w_h - u_h) =$$

$$= a(e_h, u - w_h) - \left(\frac{\partial e_h}{\partial t}, w_h - u_h\right)$$

At this point, we need to find a bound for the two terms on the right hand side. The first is standard:

$$|a(e_h, u - w_h)| \leq M \|e_h\| \|u - w_h\| \leq \frac{\alpha}{2}\|e_h\|^2 + \gamma \|u - w_h\|^2$$

The second one requires much more work (see Quarteroni)

At the end, one can prove that

$$\|e_h\|_{L^2}^2(T) + \int_0^T \|e_h\|_{H^1}^2 \leq C\left(\|w\|_{H^s}, \|u\|_{H^1(H^{s+1})}\right) \quad \text{with } r = \min(p, s),\ u \in H^1(H^{s+1})$$

Stability Analysis of the Θ-method

In this case, we work as done for the FD method:

- discretize in time
- diagonalize the problem
- write the diagonal problem as a set of independent equations

At this point, we found again that the method

$$\left(M+\Theta\Delta t A\right)\underline{u}^{n+1} = \left(M-(1-\Theta)\Delta t A\right)\underline{u}^{n} + \Delta t\Theta\underline{b}^{n+1} + \Delta t(1-\Theta)\underline{b}^{n}$$

is

- unconditionally stable for $\Theta \geq \frac{1}{2}$
- conditionally stable for $\Theta < \frac{1}{2}$ with $\Delta t \leq C h^2$

Notice that, in this case, the eigenvalues have to be intended the generalized eigenvalues:

$$A\underline{x} = \lambda M \underline{x} \qquad \left(\text{eigenvalues of } M^{-1}A\right).$$

Analysis of the Fully Discrete Problem

In this case, we proceed by combining all the previous results for the error $\underline{e}^n = \left[\left(\underline{u}_{\text{Fully Discrete}} - \underline{u}_{\text{semi-discrete}}\right)(\underline{x}_j, t^n)\right]$

The proof is technical and boring. We need many tricks. For instance, for a, b two generic vectors:

$$\begin{aligned}(b-a,b) &= (b-a,b-a) + (b-a,a) = \\ &\frac{1}{2}\|b-a\|^2 + \frac{1}{2}(b-a,b-a) + (b,a) - (a,a) = \\ &= \frac{1}{2}\|b-a\|^2 + \frac{1}{2}\|b\|^2 - \frac{1}{2}\|a\|^2\end{aligned}$$

so that $$\left(e^{n+1}-e^n, e^{n+1}\right) \geq \frac{1}{2}\|e^{n+1}-e^n\|^2 + \frac{1}{2}\|e^{n+1}\|^2 - \frac{1}{2}\|e^n\|^2$$

and summing over n:

$$\frac{1}{2}\|e^1-e^0\|^2 + \frac{1}{2}\|e^1\|^2 - \frac{1}{2}\|e^0\|^2$$

$$\frac{1}{2}\|e^2-e^1\|^2 + \frac{1}{2}\|e^2\|^2 - \frac{1}{2}\|e^1\|^2$$

$$\ldots$$

$$\sum_{n=1}^{N}\left(e^{n+1}-e^n, e^{n+1}\right) = \frac{1}{2}\|e^N\|^2 + \frac{1}{2}\sum\|e^{n+1}-e^n\|^2 - \frac{1}{2}\|e^0\|^2$$

Assembling all the results, we obtain that:

$$\|\underline{e}^N\|_{L^2}^2 + \alpha \sum \Delta t \, \|\underline{e}^n\|_{H^1}^2 \leq C(\text{data}) \left(h^{2r} + \Delta t^{\sigma(\theta)} \right)$$

where $r = \min(p, s)$, $\sigma(\theta) = \begin{cases} 1 & \theta \neq \frac{1}{2} \\ 2 & \theta = \frac{1}{2} \end{cases}$.

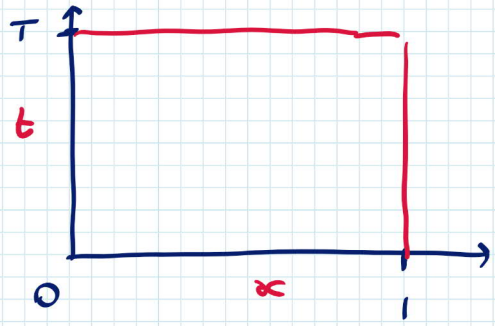
Space-Time Finite Elements

Let us consider the 1D problem:

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} + \beta \frac{\partial u}{\partial x} + \sigma u = f \qquad \begin{array}{l} x \in (0,1) \\ 0 < t \leq T \end{array}$$

for simplicity.

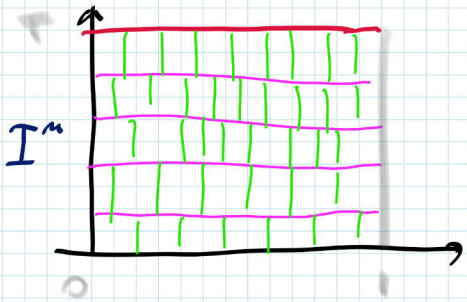
As we said, we could consider space and time at the same way:



However, this leads to a problem where all the time instants are solved simultaneously, which is problematic for simulations with a long time-interval.

We prefer to create a sequential solution in time.

For this reason, we do not mesh the domain with an unstructured mesh but with a sequence of time-slabs:



The nodes (green lines) do not need to be collocated in the same places in each slab.

For simplicity, however, we assume they are.

The finite dimensional space (in each slab) can be defined as given by functions such that:

$$v_h^n(x,t) \equiv \left\{ \underbrace{\sum_{k=0}^{d_t} c_k t^k}_{\text{time dependence}} \underbrace{\sum_{j=0}^{\tilde{N}} d_j \varphi_j(x)}_{\text{space dependence}} \right\} \quad \| \text{ Out of the slab } I^n \text{ these functions are } 0.$$

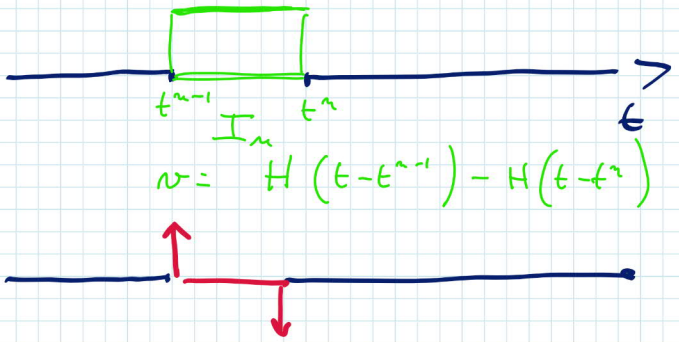
If we now consider the space-time problem:

$$\int_0^T \int_0 \frac{\partial u}{\partial t} v + \int_0^T \int_0^T (-\mu) \frac{\partial^2 u}{\partial x^2} + \ldots = \int_0^T \int_0 f \, dx \, dt$$

The function $u_h = \sum_{\text{slab}} \sum c_i t^k \sum d_j \varphi_j(x)$ is discontinuous in time, so we need to specify the derivative in the distributional sense.

This step requires some technicalities. We cannot write the problem as a sequence of slabs, the would out to be independent one to the other, which is wrong.

Notice that if we have a piecewise constant function in time:



the time derivative reads:

$$\frac{dv}{dt} = \delta_{t^{n-1}} - \delta_{t^n}$$

For a piecewise linear function: $v = (\alpha + \beta t)\chi$

where χ is the characteristic function of the slab (1 in the slab, 0 out), we have:

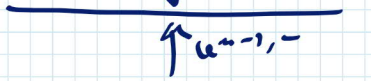
$$\frac{dv}{dt} = \beta\chi + (\alpha+\beta t)\frac{d\chi}{dt} = \beta\chi + \alpha - (\alpha+\beta\Delta t)$$

↑ the two Dirac deltas in zero

With these definition (see Formaggia, Saleri, Veneziani Chapter 5) one writes the slabwise version of the problem:

$$\int_{I^n}\int_0^{\Sigma}\left(\frac{\partial u}{\partial t}v + \mu\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \beta\frac{\partial u}{\partial x}v + \sigma uv\right) + \int_0^{\Sigma}[u^{n-1}]v = \int_{\Omega}\int_{I_n} f$$

where $[u^{n-1}] = u^{n-1,+} - u^{n-1,-}$



It is possible to prove that this formulation is unconditionally stable.

The case with piecewise-constant in time functions.

In this case, in each time slab we assume the solution to be constant in time, so that

$$u_h^n(x,t) = \sum u_j^n \varphi_j(x)$$

↳ classical Lagrange piecewise polynomials.

So, in this case $u^{n-1,+} = u^n$ and we have $\left.\frac{\partial u}{\partial t}\right|_{I_n} = 0$.

$$\int_{I_n}\int_\Omega' \frac{\partial u_\ell}{\partial t}\varphi_i + \int_{I_n}\int_\Omega' \mu \frac{\partial u_\ell}{\partial x}\frac{\partial \varphi_i}{\partial x} + \beta \frac{\partial u_\ell}{\partial x}\varphi_i + \sigma u_\ell \varphi_i = \int_{I_n}\int_\Omega' f\varphi_i$$

($\frac{\partial u_\ell}{\partial t} = 0$ (u_ℓ is constant))

$$+ \int_\Omega' [u_\ell^{n-1}]\varphi_i$$

$$\Downarrow$$

$$\left(u_\ell^n - u_\ell^{n-1}, \varphi_i\right) + \Delta t\, a\left(u_\ell^n, \varphi_i\right) = \int_{I_n}\int_\Omega' f\varphi_i$$

If f is 0 or constant in time we have the classical Backward Euler scheme.

The difference, in general, is in the right hand side: the classical BE introduces an additional approximation:

$$\int_{I_n}\int_\Omega' f\varphi_i \approx \left(f^n, \varphi_i\right)_{L^2(\Omega)}$$

Space-Time FEM — BE

BE can be regarded as an approximation of Space-Time FEM of order 0 in time.

The case of piecewise linear in time functions.

In this case, we have in each slab:

$$u_\ell(x, I^n) = \alpha_n(x)\psi_0(t) + \beta_n(x)\psi_1(t)$$

where

$$\psi_0 = \frac{t^n - t}{\Delta t} \qquad \psi_1 = \frac{t - t^{n-1}}{\Delta t} \qquad \left(\frac{d\psi_0}{dt} = -\frac{1}{\Delta t};\ \frac{d\psi_1}{dt} = \frac{1}{\Delta t}\right)$$

so that

$$u_\ell\left(x, t^{n-1}\right)_{|I_n} = u_\ell^{n-1,+} = \alpha_n(x)$$

$$u_\ell(x, t^n)_{|I_n} = u_\ell^{n,-} = \beta_n(x)$$

So, in the variational formulation we have for the time discretization

$$\begin{cases} \int_{I^n}\int_\Omega' \left(u_\ell^{n-1,+}\frac{d\psi_0}{dt} + u_\ell^{n,-}\frac{d\psi_1}{dt}\right)\psi_0 v_\ell + \int_{I^n} a\left(u_\ell^{n-1,+}\psi_0 + u_\ell^{n,-}\psi_1, \psi_0 v\right) \\ \qquad + \int_\Omega' \left(u_\ell^{n-1,+} - u_\ell^{n-1,-}\right)\psi_0(t^{n-1}) v = \int_{I^n}\int_\Omega' f\psi_0 v \\ \int_{I^n}\int_\Omega' \left(u_\ell^{n-1,+}\frac{d\psi_0}{dt} + u_\ell^{n,-}\frac{d\psi_1}{dt}\right)\psi_1 v_\ell + \int_{I^n} a\left(u_\ell^{n-1,+}\psi_0 + u_\ell^{n,-}\psi_1, \psi_1 v\right) = \int_{I^n}\int_\Omega' f\psi_1 v \end{cases}$$

Notice that $\int_{I_n} \frac{d\psi_0}{dt}\psi_0 = \int_{t^{n-1}}^{t^n} -\frac{1}{\Delta t^2}(t^n - t) = +\frac{1}{2\Delta t^2}(t^n - t)^2\Big]_{t^{n-1}}^{t^n} = -\frac{1}{2}$

$$\int_{I_n} \frac{d\psi_1}{dt}\psi_0 = \dots = -\frac{1}{2}$$

$$\int_{I_n} \frac{d\psi_0}{dt}\psi_1 = \dots = \frac{1}{2}$$

$$\int_{I_n} \frac{d\psi_1}{dt}\psi_1 = \dots = \frac{1}{2}$$

$$\int_{I_n} \psi_0^2 = \frac{1}{3}\Delta t \qquad \int_{I_n} \psi_0\psi_1 = \frac{1}{6}\Delta t \qquad \int_{I_n} \psi_1^2 = \frac{1}{3}\Delta t$$

Proceeding with the discretization, we find the system:

$$\begin{cases} \left(\frac{1}{2}M + \frac{\Delta t}{3}A\right)\underline{U}^{n,-} + \left(\frac{1}{2}M + \frac{\Delta t}{6}A\right)\underline{U}^{n,+} = M\,U^{n-1,-} + \int_{I_n}(f, \psi_0\varphi_i)_{L^2} \\ \left(-\frac{1}{2}M + \frac{\Delta t}{6}A\right)\underline{U}^{n,-} + \left(\frac{1}{2}M + \frac{\Delta t}{3}A\right)\underline{U}^{n-1,-} = \int_{I_n}(f, \psi_1\varphi_i) \end{cases}$$

At each time step, we need to solve this $(2\times N)\times(2\times N)$ linear system.

Accuracy: for linear finite elements:

$\mathbb{P}^0$: $\|e\|_{L^\infty(L^2)} \le C\left(\Delta t\,\|u\|_{L^\infty(L^\infty)} + h^2\,\|u\|_{L^\infty(H^2)}\right)$

$\mathbb{P}^1$: $\|e\|_{L^\infty(L^2)} \le C\left(\Delta t^2\,\|u\|_{L^\infty(L^\infty)} + h^2\,\|u\|_{L^\infty(H^2)}\right)$

sorry, typo in my book (it is written 1).

Week 13 : Hyperbolic Problems with FEM

(see the slides)

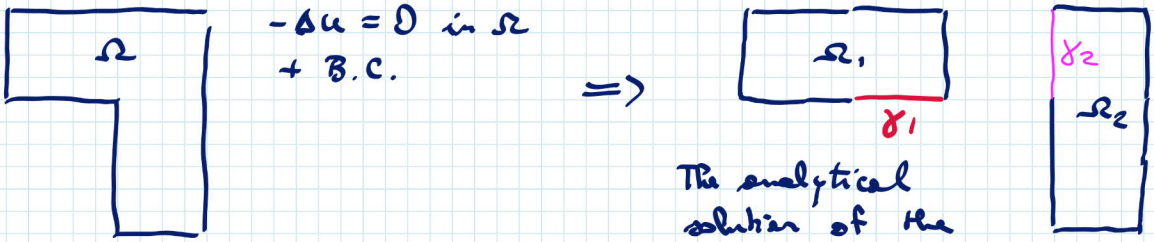
Chapt. 18 Quarteroni
Chapt. 6 Quarteroni-Valli
Chapt. 3 Formaggia-Saleri-V.

A short intro to Domain-Decomposition

Domain Decomposition (DD) is a technique for the efficient numerical solution of problems :

- in complex domains
- with a multiphysics nature
- with parallel processors.

(1) → surprisingly enough, the first idea of DD is *before* scientific computing, Schwartz in the XIX century :



The analytical solution of the Laplace problem is found by separation of variables :

IDEA

for loop : initial guess for $u(\gamma_1)$

solve : $-\Delta u_1^{(n+1)} = 0$ + B.C. + $u_1^{(n+1)}(\gamma_1) = \lambda^{(n)}$

solve : $-\Delta u_2^{(n+1)} = 0$ + B.C. + $u_2^{(n+1)}(\gamma_2) = u_1^{(n+1)}(\gamma_2)$

set $\lambda^{(n+1)} = u_2^{(n+1)}(\gamma_1)$

This is a first example of DD with overlapping.

Notice that if we have an initial guess for $u(\gamma_2)$ too, in we can write

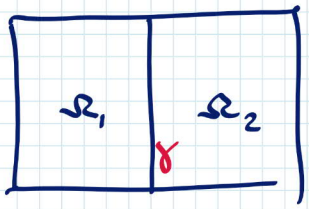
$$u_2^{(n+1)}(\gamma_2) = u_1^{(n)}(\gamma_2)$$

and the two solution processes are completely independent ⇒ PARALLEL COMPUTING!

One may argue that the conditions used on γ_i $(i=1,2)$ also not need to be necessarily Dirichlet.

Also, the critical question is : does the iterative process converge ?

(2) Another example is for multiphysics problems:



For instance:

$$-\mu_1 \Delta u_1 = f_1 \quad \text{in } \Omega_1$$
$$+ \text{B.C.}$$
$$-\mu_2 \Delta u_2 = f_2 \quad \text{in } \Omega_2$$
$$+ \text{B.C.}$$
$$+ \text{interface conditions on } \gamma$$

The interface conditions may reflect physical processes.

In this case we have NON-OVERLAPPING DD, and it can be used for single physics problem too.

For instance, the Poisson problem:

$$-\Delta u = f \quad \text{in } \Omega$$
$$+ \text{B.C.}$$

is equivalent to

$$-\Delta u_1 = f_1 \quad \text{in } \Omega_1 \qquad\qquad -\Delta u_2 = f_2 \quad \text{in } \Omega_2$$
$$+ \text{B.C.}_1 \qquad\qquad + \text{B.C.}_2$$

$$u_1(\gamma) = u_2(\gamma)$$
$$\nabla u_1 \cdot \underline{n}_1(\gamma) = -\nabla u_2 \cdot \underline{n}_2(\gamma)$$
$$\hookrightarrow \underline{n}_1 = -\underline{n}_2$$

} interface conditions.

From here, we have the following non-overlapping DD scheme:

for loop:

solve $-\Delta u_1^{(k+1)} = f_1$
$+ \text{B.C.}_1 +$ $\boxed{u_1^{(k+1)}(\gamma) = \lambda^{(k)}}$ Dirichlet

solve $-\Delta u_2^{(k+1)} = f_2$
$+ \text{B.C.} +$ $\boxed{\nabla u_2 \cdot \underline{n}(\gamma) = \nabla u_1 \cdot \underline{n}(\gamma)}$ Neumann

Set $\lambda^{(k+1)} = u_2^{(k+1)}(\gamma)$ (*)

Again, the critical question is about the convergence.

We will see the convergence analysis.

There are several strategies to improve the convergence.

(1) Replace (*) with $\lambda^{(n+1)} = \Theta\, u_2^{(n+1)}(\gamma) + (1-\Theta)\,\lambda^{(n)}$ (Relaxation)

(2) Replace the interface conditions with equivalent Robin-Robin conditions:

$$\nabla u_1^{n+1} \cdot \underline{n}_1 + \alpha u_1|_\gamma = -\nabla u_2 \cdot \underline{n}_2 + \alpha u_2|_\gamma$$

$$\nabla u_1 \cdot \underline{n}_1 + \beta u_1|_\gamma = -\nabla u_2 \cdot \underline{n} + \beta u_2|_\gamma$$

where α and β are parameters to tune

If they are "optimized", we have "nonoverlapping Schwarz".

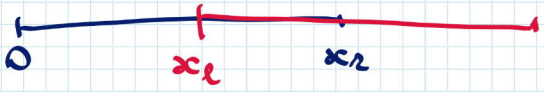
As you see, the topic is vast and full of potential:

- preconditioning full problems
- parallel computing and load balancing.

Here, we limit to just some simple analysis in 1D.

1D Analysis

Overlapping Method.



$$\begin{cases} -u_1''^{(n+1)} = 1 \\ u_1^{(n+1)}(0) = 0 \quad u_1^{(n+1)}(x_r) = \lambda^{(n)} \end{cases}$$

$$-u_2''^{(n+1)} = 1$$

$$u_2^{(n+1)}(1) = 0 \quad u_2^{(n+1)}(x_\ell) = u_1^{(n+1)}(x_\ell)$$

Problem:

$$-u'' = 1$$

$$u(0) = u(1) = 0$$

Test: $\int_{x_\ell}^{x_1} (u_1^{(n+1)} - u_2^{(n+1)})^2$ else then $\lambda^{(n+1)} = u_2^{(n+1)}(x_r)$

We know that the exact solution of the problem reads: $\frac{1}{2}x(1-x)$

Let's introduce the errors: $e_1 = u_1 - u$, $e_2 = u_2 - u$

$$-e_1''^{(n+1)} = 0$$

$$e_1^{(n+1)}(0) = 0 \quad e_1^{(n+1)}(x_r) = \varepsilon^{(n)} = \lambda^{(n)} - \frac{1}{2}x_r(1-x_r)$$

$$-e_2''^{(n+1)} = 0$$

$$e_2^{(n+1)}(1) = 0 \quad e_2^{(n+1)}(x_r) = e_1^{(n+1)}(x_r)$$

We have:

$$e_1^{(k+1)}(x_1) = \varepsilon^{(k)} \frac{x}{x_2}$$

↑ linear function

$$e_2^{(k+1)} = (1-x)\,\varepsilon^{(k)} \frac{x_2}{x_2} \frac{1}{1-x_1}$$

so that $$\varepsilon^{(k+1)} = e_2^{(k+1)}(x_2) = \frac{1-x_2}{1-x_1} \frac{x_1}{x_2} \varepsilon^{(k)}$$

To have convergence, i.e. $\varepsilon^{(k)} \xrightarrow[k\to\infty]{} 0$ we need:

$$\left| \frac{1-x_2}{1-x_1} \frac{x_1}{x_2} \right| < 1$$

$$\left| \frac{\frac{1}{x_1} - 1}{\frac{1}{x_2} - 1} \right| < 1 \Rightarrow 1 - \frac{1}{x_1} < \frac{1}{x_2} - 1 < \frac{1}{x_1} - 1$$

$$\downarrow$$

$$< 0$$

This is always true for $0 < x_1 < x_2 \leq 1$, NOT for $x_1 = x_2$

Also, for either $x_1 = 0$ or $x_2 = 1$, we have convergence in one step.

This is all expected:

(1) for $x_1 = x_2$, we are prescribing two problems with no overlapping, using the same condition twice

(2) for $x_1 = 0$ or $x_2 = 1$, we are covering the entire domain (fake DD).

In general, if $|x_2 - x_1|$ gets large (more overlapping), the convergence accelerates:

Study: $\frac{1-x_1-h}{1-x_1} \frac{x_1}{x_1+h}$ as function of h to see this.

REMARK

This is the so-called "multiplicative version". If we use the parallel or additive version with:

$$u_2^{(k+1)}(x_1) = u_1^{(k)}(x_1) \quad \text{not } (k+1)$$

we still have the same convergence results, but we need as much as twice iterations (try).

Dirichlet Neumann Method

Let's consider the scheme:

$$\begin{cases} -u_1''^{(k+1)} = 1 & x \in [0, \gamma) \quad \text{where } \gamma \in (0,1) \\ u_1^{(k+1)}(0) = 0, \; u_1^{(k+1)}(\gamma) = \lambda^{(k)} \end{cases}$$

$$\begin{cases} -u_2''^{(k+1)} = 1 \\ u_2^{(k+1)}(1) = 0, \; \dfrac{\partial u_2^{(k+1)}}{\partial x} = \dfrac{\partial u_1^{(k+1)}}{\partial x} \end{cases}$$

$$\left[\lambda^{(k+1)} = u_2^{(k+1)}(\gamma) \right.$$

Proceeding with the same notation as before:

$$\begin{cases} -e_1''^{(k+1)} = 0 \\ e_1^{(k+1)}(0) = 0 \qquad e_1^{(k+1)}(\gamma) = \varepsilon^{(k)} \end{cases} \Rightarrow \quad e_1^{(k+1)} = \frac{\varepsilon^{(k)}}{\gamma} x$$

$$\begin{cases} -e_2''^{(k+1)} = 0 \\ e_2^{(k+1)}(1) = 0 \qquad e_2'^{(k+1)} = \dfrac{\varepsilon^{(k)}}{\gamma} \end{cases} \Rightarrow \quad e_2^{(k+1)} = -\frac{\varepsilon^{(k)}}{\gamma}(1-x)$$

$$\varepsilon^{(k+1)} = \frac{\varepsilon^{(k)}}{\gamma}(\gamma - 1)$$

The convergence is guaranteed for

$$1 - \gamma < \gamma \quad \Rightarrow \quad \gamma > \frac{1}{2}$$

If we have the Dirichlet domain LARGER than the Neumann domain, we have convergence.

Otherwise, we can use the relaxation:

$$\varepsilon^{(k+1)} = \left(\theta \frac{(\gamma - 1)}{\gamma} + (1 - \theta) \right) \varepsilon^{(k)}$$

$$\downarrow$$

$$\frac{1}{\gamma}\left(\cancel{\theta\gamma} - \theta + \gamma - \cancel{\theta\gamma} \right) \varepsilon^{(k)}$$

$$\left| \frac{\gamma - \theta}{\gamma} \right| < 1 \quad \Rightarrow \quad -\gamma < \gamma - \theta < \gamma$$

(the left inequality is trivial)

$$\boxed{\theta < 2\gamma}$$

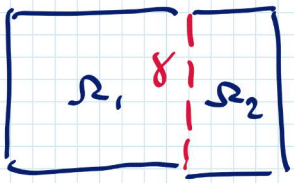
With this choice we converge and the optimal choice is $\theta = \gamma$.

Remark

Computing the flux $\nabla u \cdot \underline{n}$ on a Dirichlet boundary.

This topic is somehow overlooked in the literature.
However, this may lead to some useless and inaccurate procedures.
As a matter of fact, computing the flux on a Dirichlet boundary, as required by a DN domain-decomposition can be done with no need of numerical differentiation (that leads to an accuracy degradation).

Let's see:



Dirichlet: $-\Delta u_1^{(n+1)} = f_1$ + B.C. in $\Omega_1, \partial\Omega_1 \setminus \gamma$

$u_1^{(n+1)}(\gamma) = \lambda^{(n)}$

Neumann: $-\Delta u_2^{(n+1)}$ + B.C. in $\Omega_2, \partial\Omega_2 \setminus \gamma$

$$\frac{\partial u_2^{(n+1)}}{\partial \underline{n}} = \boxed{\frac{\partial u_1^{(n+1)}}{\partial \underline{n}}}$$ we need this

Let's focus on the Problem 1, "Dirichlet". For simplicity $u(\partial\Omega) = 0$

$$-\int_{\Omega_1} \Delta u_1 v = -\underbrace{\int_{\partial\Omega} \nabla u \cdot \underline{n}\, v}_{=0 \text{ because } v(\partial\Omega_1)=0} + \int_{\Omega} \nabla u \nabla v = \int_{\Omega} f_1 v$$

However, notice that before we apply the Dirichlet boundary conditions, this term is alive.
So, for $v \in H^1(\Omega_1)$ and not $v \in H^1_0(\Omega_1)$ we have this term.

In $H^1_0(\Omega_1)$: $\int_{\Omega_1} \nabla u \nabla v = \int_{\Omega_1} f_1 v$ $\quad v \in H^1_0(\Omega)$, $u \in \mathcal{L} + H^1_0(\Omega)$

Once we solve this problem, what is:

$$\int_{\Omega_1} \nabla u \nabla v - \int_{\Omega_1} f_1 v = ?$$

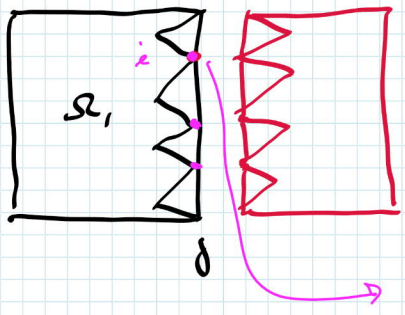
$\begin{cases} = 0 & \text{for } v \in H^1_0(\Omega) \\ \neq 0 & \text{for } v \in H^1(\Omega) \text{ but } \notin H^1_0(\Omega) \end{cases}$

The residual is exactly $\int_{\partial\Omega_1} \nabla u \cdot \underline{n}\, v$

Now, if we take v vanishing on $\partial\Omega_1 \setminus \gamma$, we have

$$\int_{\gamma} \nabla u \cdot \underline{n}\, v$$

This happens with the Lagrange polynomials with nodes on the interface γ.



$$\int_{\Omega_1} \nabla u \nabla \varphi_i - \int_{\Omega_1} f_1 \varphi_i = \int_{\gamma} \nabla u \cdot n \, \varphi_i$$

So if we compute

$$A_{\text{no bc}}\, u - b = \underline{r}$$

↓ No Boundary Conditions

$\underline{r}(i)$ is $\int_{\gamma} \nabla u \cdot \underline{n} \, \varphi_i$

If the mesh is conformal (same nodes from left and right) we can use this number as b.c. for Ω_2.

Week 14

What I didn't tell you!

Non Linear Problems

When we have a non-linear problem, clearly we need to adjust the procedure with an extra-layer of an iterative root finding step.

Let's see an example:

$$-\Delta u + u^3 = f \qquad \text{in } \Omega$$

$$+ \text{B.C.}$$

Clearly, if we apply the usual procedure, we have to deal with a non-linear algebraic system. We can simplify the problem, by writing

$$u^3 \simeq \sum u_i^3 \varphi_i$$

So that we get the non-linear system:

$$K\underline{u} + M\underline{u}^3 = \underline{b} \qquad (\underline{u}^3 \text{ in the sense of Matlab } \underline{u}.\hat{}\,3)$$

At this point, we can call any method for solving non-linear algebraic equation. For instance, the Newton method.

$$\underline{F}(\underline{u}) = 0 \quad \Rightarrow \quad \bar{J}\,\delta\underline{u}^{(\text{curr})} = -F(\underline{u}^{(k)}) \qquad \bar{J} = \text{Jacobian of } \underline{F}$$

$$\underline{u}^{(\text{curr})} = \underline{u}^{(k)} + \delta\underline{u}^{(\text{curr})} \qquad \left(\bar{J}_{ij} = \frac{\partial F_i}{\partial u_j}\right)$$

In the example above:

$$J(\underline{u}) = K + 3M\,\text{diag}(u.\hat{}\,2)$$

So we have:

$$\left(K + 3M\,\text{diag}(u^{(k)}.\hat{}\,2)\right)\left(u^{(\text{curr})} - \underline{u}^{(k)}\right) = -K\underline{u}^{(k)} - M u^{(k)}.\hat{}\,3 + \underline{b}$$

(Linear System)

This needs to be iterated until convergence.

In this approach, we first <u>discretize</u> then <u>linearize</u>. In particular, we approximate

$$\left(\sum_i u_i \varphi_i\right)^3 \simeq \sum u_i^3 \varphi_i$$

This last step is optional.

There is another approach: LINEARIZE then DISCRETIZE.

$-\Delta u + u^3 = f$ can be written as $-\Delta u + u^{*2} u = f$

where $u^* \approx u$. For instance:

- give $u^{(0)}$
- loop k: $-\Delta u^{(k+1)} + \left(u^{(k)}\right)^2 u^{(k+1)} = f$
  until convergence (if possible)
- we discretize this problem:

$$K \underline{u}^{(k+1)} + M\left(\underline{u}^{(k)2}\right) \underline{u}^{(k+1)} = \underline{f}$$

where $$M\left(u^{(k)2}\right)_{ij} = \int u^{(k)2} \varphi_j \varphi_i$$

($u^{(k)2}$: given)

As usual, the problem here is to guarantee the convergence. A relaxation step may help:

$$K \underline{\tilde{u}} + M\left(u^{(k)2}\right) \underline{\tilde{u}} = \underline{f}$$

$$u^{(k+1)} = \theta \tilde{u} + (1-\theta) u^{(k)} \qquad \theta \in (0,1]$$

This is called Picard iteration, and it is in general:

- easy
- robust (converges for a large neighborhood of the solution)
- slow

A Newton method can be used with the Lim-Then-Discr approach, but it requires the differentiation of the operator.

To find the so-called TANGENT PROBLEM we differentiate:

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(-\Delta(u + \varepsilon v) + (u+\varepsilon v)^3 + \Delta u - u^3 \right) =$$

$$= \lim \frac{1}{\varepsilon} \left(-\varepsilon \Delta v + u^3 + 3\varepsilon u^2 v + 3\varepsilon^2 u v^2 + \varepsilon^3 v - u^3 \right) =$$

$$= \lim_{\varepsilon \to 0} \left(-\Delta v + 3u^2 v + 3\varepsilon u v^2 + \varepsilon^2 v \right) = -\Delta v + 3u^2 v$$

So the tangent problem reads:

$$-\Delta \delta u + 3u^2 \delta u = f + \Delta u - u^3$$

Iteratively, one has to solve:

given $\underline{u}^{(0)}$;

loop: $-\Delta \underline{u}^{(u+1)} + 3\underline{u}^{(u)2}\,\underline{u}^{(u+1)} - 3\underline{u}^{(u)3} =$

$= f - u^{(u)3}$

$\Downarrow$

$-\Delta u^{(u+1)} + 3u^{(u)2} u^{(u+1)} = f + 2u^{(u)3}$

CHECK: if $u^{(u+1)} = u^{(u)} = \bar{u}$ then

$-\Delta \bar{u} + \bar{u} = f \Rightarrow$ Yes!!!

The Newton method is more elaborate, it requires a good initial guess to converge (better than Picard), but it is faster.

A reasonable approach is

$u^{(0)} \longrightarrow$ Picard for a few iterations $\longrightarrow u_{Picard} = u^{(0)}_{Newton} \longrightarrow$ Newton.

To summarize:

DISCRETIZE-THEN-LINEARIZE $\Rightarrow$ Nonlinear Algebraic Solver

LINEARIZE-THEN-DISCRETIZE $\Rightarrow$ Sequence of nonlinear problems.

The linearization can be done in other ways, for instance fixed point iterations.

$$-\Delta u + \sin(u)\,u = f$$

could lead to:

$-\Delta u^{(u+1)} + \sin(u^{(u)})\,u^{(u+1)} = f$ (Picard)

$-\Delta u^{(u+1)} = f - \sin(u^{(u)})\,u^{(u)}$ (Probably not convergent or slow)

$-\Delta u^{(u+1)} + \cos(u^{(u)})\,u^{(u)}\,(u^{(u+1)} - u^{(u)}) + \sin(u^{(u)})\,(u^{(u+1)} - u^{(u)})$

$= f - \sin(u^{(u)})\,u^{(u)}$ (Newton)

$\downarrow$

$-\Delta u^{(u+1)} + \cos(u^{(u)})\,u^{(u)}\,u^{(u+1)} + \sin(u^{(u)})\,u^{(u+1)} =$

$= f + \cos(u^{(u)})\,u^{(u)2}$

Personally, I prefer Linearize-then-Discretize.

EXAMPLE

Viscous steady Burgers Equation (similar to Navier-Stokes)

$$(\underline{u}\cdot\nabla)\underline{u} - \mu\Delta\underline{u} = \underline{f} \qquad \underline{u} = \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix}$$

rewrite:

$$\left(\sum_i\right) u_i \frac{\partial u_j}{\partial x_i} - \mu \sum_i \frac{\partial^2 u_j}{\partial x_i^2} = f_j$$

↑ understood in the Einstein notation

$$u_0\frac{\partial u_j}{\partial x_0} + u_1\frac{\partial u_j}{\partial x_1} + u_2\frac{\partial u_j}{\partial x_2} - \mu\frac{\partial^2 u_j}{\partial x_0^2} - \mu\frac{\partial^2 u_j}{\partial x_1^2} - \mu\frac{\partial^2 u_j}{\partial x_2^2} = f_j$$

Picard: $\quad (\underline{u}^{(k)}\cdot\nabla)\underline{u}^{(k+1)} - \mu\Delta\underline{u}^{(k+1)} = \underline{f}$

Newton: TANGENT PROBLEM:

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(((\underline{u}+\varepsilon\underline{v})\cdot\nabla)(\underline{u}+\varepsilon\underline{v}) - \mu\Delta(\underline{u}+\varepsilon\underline{v}) - \underline{u}\cdot\nabla\underline{u} + \mu\Delta\underline{u}\right) =$$

$$= (\underline{v}\cdot\nabla)\underline{u} + (\underline{u}\cdot\nabla)\underline{v} - \mu\Delta\underline{v}$$

In the Newton iteration, we have $\{\ \delta\underline{u} = \underline{u}^{(k+1)} - \underline{u}^{(k)} = \underline{v}\ \}$

$$(\delta\underline{u}\cdot\nabla)\underline{u}^{(k)} + (\underline{u}^{(k)}\cdot\nabla)\delta\underline{u} - \mu\Delta\delta\underline{u} = \underline{f} - (\underline{u}^{(k)}\cdot\nabla)\underline{u}^{(k)} + \mu\Delta\underline{u}^{(k)}$$

$$(\underline{u}^{(k+1)}\cdot\nabla)\underline{u}^{(k)} + (\underline{u}^{(k)}\cdot\nabla)\underline{u}^{(k+1)} - \mu\Delta\underline{u}^{(k+1)} = \underline{f} + (\underline{u}^{(k)}\cdot\nabla)\underline{u}^{(k)}$$

REMARK

The FENICS tutorial suggests that there is a DISCRETIZE-THEN-NEWTON approach encoded.
However, LINEARIZE-THEN-DISCRETIZE approaches are manually implemented.

The case of Time-Dependent Problems.

In the case of time-dependent problems, non-linearity may introduce some outstanding computational costs.

As a matter of fact, we may have 3 nested loops!!:

- for time
  - for nonlinear equations
    - for linear solver (if iterative)

However notice that this cost applies only for implicit time-advancing.

In fact, if we have:

$$\frac{\partial u}{\partial t} - \mu \Delta + u^3 = f$$

an explicit scheme reads:

$$u^{n+1} = u^n + \Delta t K \tilde{u}^{n+1} + \Delta t (u^3)^n + \Delta t f^n$$

The implicit one (linearize then discretize)

$$u^{n+1} - \Delta t \mu \Delta u^{n+1,k+1} + \Delta t 3 (u^{n+1,k})^2 u^{n+1,k+1} = \Delta t f^n + \Delta t 2 (u^{n+1,k})^3$$

$\Rightarrow$ Linear System:

$$\left[I + \Delta t K + \Delta t 3 M(\underline{u}^{n+1,k}) \right] \underline{u}^{n+1,k+1} = \Delta t \underline{b} + \Delta t 2 \underline{C}(\underline{u}^{n+1,k})$$

An explicit scheme, however, may have poor stability properties.

A possible TRADE-OFF is a SEMI-IMPLICIT SCHEME, where we use the time advancing to linearize the problem:

$$\frac{1}{\Delta t}(u^{n+1} - u^n) - \mu \Delta u^{n+1} + (u^n)^2 u^{n+1} = f^{n+1}$$

RATIONALE: $u^n = u^{n+1} + O(\Delta t)$ — $O(\Delta t)$ is also the discretization in time error

With a second order scheme we could write:

$$u^* u^{n+1} \quad \text{where } u^* = 2u^n - u^{n-1}$$

Notice that, in this way, the stability is CONDITIONAL, but as long as the linearization concerns terms of order 0 (reactive) or 1 (convective) the constraint $\Delta t < Ch^2$ is avoided.

EXAMPLE FOR BURGERS:

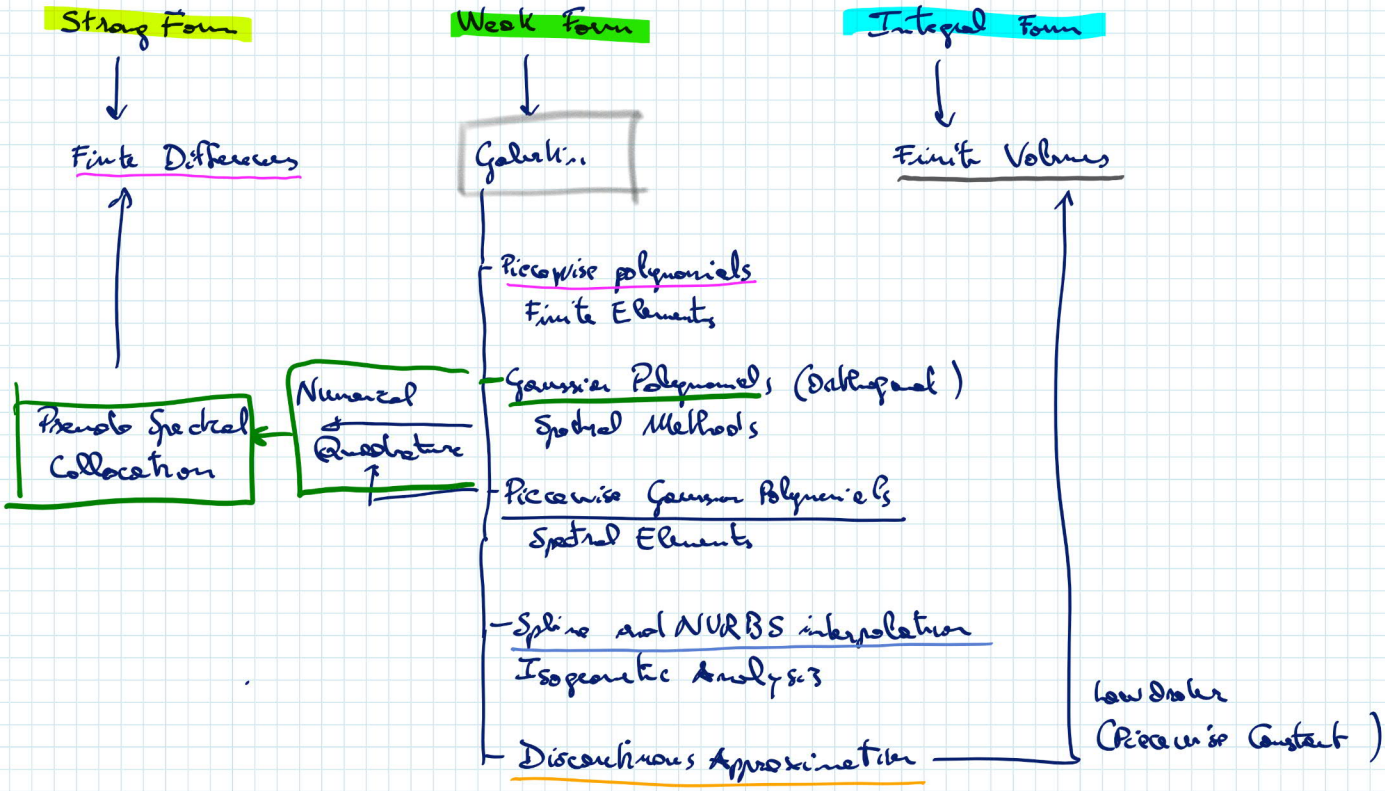
$$\frac{\partial \underline{u}}{\partial t} - \mu \Delta \underline{u} + (\underline{u} \cdot \nabla) \underline{u} = f$$

$$\frac{1}{\Delta t}(\underline{u}^{n+1} - \underline{u}^n) - \mu \Delta \underline{u}^{n+1} + (\underline{u}^* \cdot \nabla) \underline{u}^{n+1} = \underline{f}$$

$\underline{u}^*$ extrapolation in time: $\underline{u}^n$
(of the same order of the time discretization).

A quick look to the methods

Differential Problems:



Two words on Spectral Methods

When we use Gaussian interpolation as the backbone, we have some advantages and some drawbacks.

Spectral Methods are strongly accurate: the Gaussian interpolation has an exponential accuracy (N = degree of the Lagrange polynomials)

$$\|u - u_N\|_{H^1} \leq C \|u\|_{H^s} N^{-s} \quad \text{for } u \in H^s$$

$$\left(s = +\infty \quad \leq C \|u\|_{H^s} e^{-s}\right)$$

so the only limiting factor is the regularity of the solution

The identification of the nodes for the Gaussian interpolation is easy in 1D and simple geometries. In a complex geometry, it is not trivial (we should map a complex geometry into a simple one)

$\Rightarrow$ this led to the development of Spectral Elements.

Quadrilaterals in Spectral Methods are popular as they are obtained as the tensor-product of the 1D case

Spectral basis functions are L^2-orthogonal (Chebyshev, Legendre) $\Rightarrow$

the mass matrix is diagonal! $(\varphi_i, \varphi_j) = 0$ for $i \neq j$!

There is no way that for polynomials of order $N \longrightarrow +\infty$ we do exact integration $\Rightarrow$ All the integrals are approximated by suitable quadrature formulas.

This means that the method is actually a pseudo-spectral and the analysis tool is the STRANG-LEMMA (not Céa).

If we approximate:

$$(\nabla\varphi_i, \nabla\varphi_j) \approx \sum_{k=1}^{nqn} w_k \nabla\varphi_i \cdot \nabla\varphi_j (x_k, y_k, z_k)$$

(nqn = number of quadrature nodes)

it is possible to reinterpret the entire method as a collocation method (Finite Difference) applied to the strong form.

ASIDE NOTE

Finite Elements have been recently extended to polytopes by Brezzi and co-workers (Virtual Element Method or Mimetic Finite Differences).

Saddle Point Problems

When facing elliptic (and also parabolic) problems, we look for the solution of a free minimization:

$$\min_{u \in H^1_0} J = \frac{1}{2}\int_\Omega |\nabla u|^2 - \int_\Omega f u \quad \Rightarrow \quad \begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

This leads to a coercivity-based formulation:

$$a(u,v) = (f,v) \quad \forall v \in H^1_0$$

with

$$a(u,u) \geq \alpha \|u\|^2$$

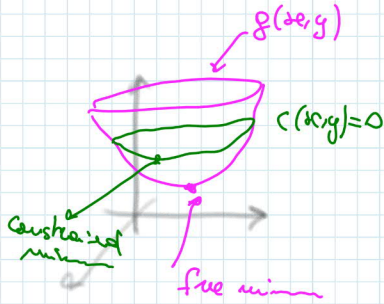
The coercivity plays the role of convexity in minimization:

$$\min_{x \in I} g \quad \Leftrightarrow \quad g'(x) = 0 \text{ with } g'' > 0$$

What happens if the minimization is not free?

In calculus we may have:

$$\min f(x,y) \quad \text{with the constraint} \quad c(x,y) = 0$$



In the specific case, one could find

$$c(x,y) = 0 \quad \Rightarrow \quad y = y(x) \qquad \left[\text{Dini theorem}\right]$$

then

$$f(x,y)\Big|_{c(x,y)=0} = f(x, y(x)) = \tilde{f}(x)$$

and minimize $\tilde{f}(x)$.

To pursue a general strategy, one can introduce the Lagrange multiplier approach:

$$\min g(x,y) \quad \text{with constraint} \quad c(x,y) = 0$$

$$\Downarrow$$

$$\max_{\lambda} \min_{x,y} \mathcal{L}(x,y;\lambda) = g(x,y) - \lambda\, c(x,y) \qquad \text{(free)}$$

In fact, this leads to:

$$\frac{\partial \mathcal{L}}{\partial \lambda} : \quad c(x,y) = 0 \qquad \text{(constraint)}$$

$$\frac{\partial \mathcal{L}}{\partial (x,y)} : \quad \frac{\partial g}{\partial (x,y)} - \lambda \frac{\partial c}{\partial (x,y)} = 0$$

What about problems involving functionals?

AN IMPORTANT EXAMPLE

$$J(\underline{u}) = \frac{1}{2}\int_\Omega |\nabla \underline{u}|^2 - \underline{f}\cdot\underline{u} \qquad \underline{u} \text{ vector}$$

($\nabla \underline{u}^T \cdot \nabla \underline{u}$)

with the constraint $\nabla \cdot \underline{u} = 0$

$$\Rightarrow \quad \mathcal{L} = \int_\Omega \nabla \underline{u} \cdot \nabla \underline{v} - \int_\Omega p\, \nabla \cdot \underline{u} = 0$$

For $\underline{u}(\partial\Omega) = \underline{0}$:

$$\begin{cases} -\Delta \underline{u} + \nabla p = 0 \\ \nabla \cdot \underline{u} = 0 \end{cases}$$

STOKES PROBLEM (incompressible fluids)

Now, the coercivity is not enough to be well posedness. We need something more.

FORMAL DISCUSSION

Find $(\underline{u}, p) \in \left(H_0^1(\Omega), L^2(\Omega)\right)$ s.t.

$$\begin{cases} \int_\Omega \nabla \underline{u} \cdot \nabla \underline{v} - \int_\Omega p\, \nabla \cdot \underline{v} = 0 \\ \int_\Omega q\, \nabla \cdot \underline{u} = 0 \end{cases} \qquad \forall\, \underline{v} \in H_0^1(\Omega),\ q \in L^2(\Omega)$$

Find $(\underline{u}, p) \in V \times Q$: $a(\underline{u}, \underline{v}) + b(\underline{v}, p) = 0 \quad \forall \underline{v} \in V$

$b(\underline{u}, q) = 0 \quad \forall q \in Q$

Sufficient conditions:

$\exists \alpha > 0$ s.t. $\forall \underline{v} \in V$ $\quad a(\underline{v}, \underline{v}) \geq \alpha \|\underline{v}\|_V^2$

inf-sup condition: $\forall p \in Q$, $\exists \underline{v} \in V$ and $\beta > 0$ s.t.

$$b(p, \underline{v}) \geq \beta \|p\|_Q \|\underline{v}\|_V$$

These conditions are true for Stokes.

But... but... what happens when we perform a Galerkin approach.

Simply coercive case:

coercivity in V $\Rightarrow$ coercivity in V_h

$a(u,u) \geq \alpha \|u\|_V^2 \Rightarrow a(u_h, u_h) \geq \alpha \|u_h\|_V^2$ i.e. $V_h \subset V$

In the saddle point case:

$$a(u,u) \geq \alpha \|u\|_V^2 \Rightarrow a(u_h, u_h) \geq \alpha \|u_h\|_V^2 \quad i.e.\ V_h \subset V$$

but:

$\forall p \in L^2$, $\exists \underline{v} \in V$, $\beta > 0$ s.t.

$$b(\underline{v}, p) \geq \beta \|p\|_Q \|\underline{v}\|_V$$

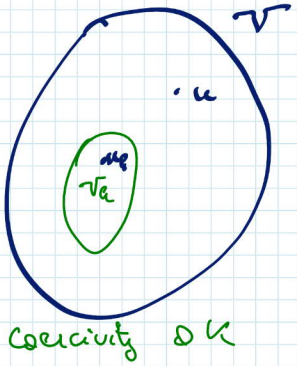
$\forall p_h \in Q_h$, $\exists \underline{v}_h \in V_h$, $\beta > 0$ s.t.

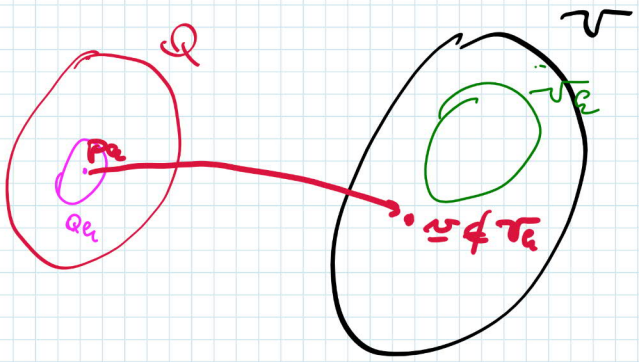
$$b(\underline{v}_h, p_h) \geq \beta \|p_h\|_0 \|\underline{v}\|_V$$

even when $Q_h \subset Q$ and $V_h \subset V$

This requires some extra effort.

In diagrams:





We need to enforce that V_h contains all the corresponding elements of Q_h for the inf-sup condition.

This excludes some pairs of finite elements.

For instance, for Stokes:

$\underline{u}_h \in \mathbb{P}^1$, $p_h \in \mathbb{P}^1$ is NOT inf-sup compatible

$\underline{u}_h \in \mathbb{P}^2$, $p_h \in \mathbb{P}^1$ is inf-sup compatible.