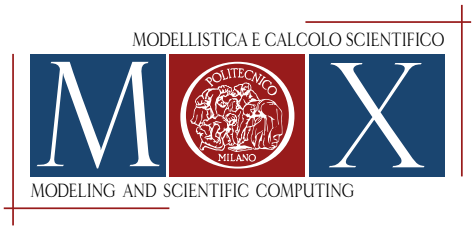
Numerical Analysis of Partial Differential Equations

Alfio Quarteroni

MOX, Dipartimento di Matematica
Politecnico di Milano



Lecture Notes
A.Y. 2022-2023

Parabolic equations

Cfr [Q], Chap. 5

We consider parabolic equations of the form

$$\frac{\partial u}{\partial t} + Lu = f, \qquad \mathbf{x} \in \Omega,\ t > 0, \tag{1}$$

where:
Ω is a domain of $\mathbb{R}^d$, $d = 1, 2, 3$,
$f = f(\mathbf{x}, t)$ is a given function,
$L = L(\mathbf{x})$ is a generic elliptic operator acting on the unknown $u = u(\mathbf{x}, t)$.

[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

When solved only for a bounded temporal interval, say for $0 < t < T$, the region $Q_T = \Omega \times (0, T)$ is called *cylinder* in the space $\mathbb{R}^d \times \mathbb{R}^+$.

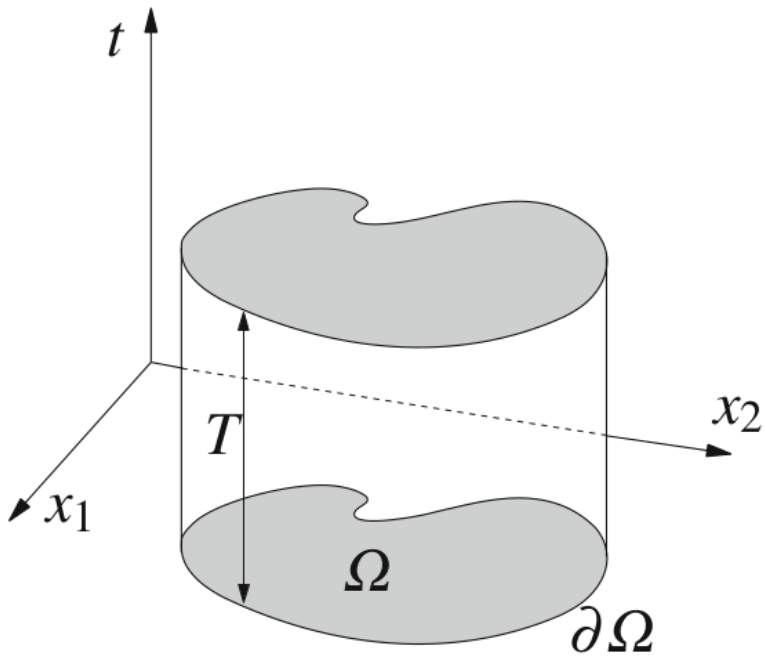


Figure: The cylinder $Q_T = \Omega \times (0, T)$, $\Omega \subset \mathbb{R}^2$

In the case where $T = +\infty$, $Q = \{(\mathbf{x}, t) : \mathbf{x} \in \Omega, t > 0\}$ will be an infinite cylinder.

Equation (1) must be completed by assigning an initial condition

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega, \tag{2}$$

together with boundary conditions, which can take the following form:

$$\begin{aligned} u(\mathbf{x}, t) &= \varphi(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma_D \text{ and } t > 0, \\ \frac{\partial u(\mathbf{x}, t)}{\partial n} &= \psi(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma_N \text{ and } t > 0, \end{aligned} \tag{3}$$

where u_0, φ and ψ are given functions and $\{\Gamma_D, \Gamma_N\}$ provides a boundary partition, that is $\Gamma_D \cup \Gamma_N = \partial\Omega$, $\mathring{\Gamma}_D \cap \mathring{\Gamma}_N = \emptyset$. For obvious reasons, Γ_D is called Dirichlet boundary and Γ_N Neumann boundary.

In the one-dimensional case, the problem:

$$\begin{aligned} &\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} = f, \quad 0 < x < d, \quad t > 0, \\ &u(x, 0) = u_0(x), \quad 0 < x < d, \\ &u(0, t) = u(d, t) = 0, \quad t > 0, \end{aligned} \tag{4}$$

describes the evolution of the temperature $u(x, t)$ at point x and time t of a metal bar of length d occupying the interval $[0, d]$, whose thermal conductivity is ν and whose endpoints are kept at a constant temperature of zero degrees.
The function u_0 describes the initial temperature, while f represents the heat generated (per unit length) by the bar.

For this reason, (4) is called heat equation.

Weak formulation and its approximation

We proceed formally, by multiplying for each $t > 0$ the differential equation by a test function $v = v(\mathbf{x})$ and integrating on Ω. We set $V = \mathrm{H}^1_{\Gamma_D}(\Omega)$ and for each $t > 0$ we seek $u(t) \in V$ such that

$$\int_\Omega \frac{\partial u(t)}{\partial t} v \, d\Omega + a(u(t), v) = \int_\Omega f(t) v \, d\Omega \qquad \forall v \in V, \tag{5}$$

where

- $u(0) = u_0$
- $a(\cdot, \cdot)$ is the bilinear form associated to the elliptic operator L
- we have supposed for simplicity $\varphi = 0$ and $\psi = 0$.

The modification of (5) in the case where $\varphi \neq 0$ and $\psi \neq 0$ is left for exercise

Definition

A bilinear form $a(\cdot,\cdot)$ is said *weakly coercive* if

$$\exists \lambda \geq 0,\ \exists \alpha > 0 : \quad a(v,v) + \lambda \|v\|^2_{\mathrm{L}^2(\Omega)} \geq \alpha \|v\|^2_V \quad \forall v \in V,$$

yielding for $\lambda = 0$ the standard definition of coercivity.

Rationale for weak coercivity:

$$\frac{\partial u}{\partial t} + \mathcal{L}u = f$$

(change fo variable: $u = e^{\lambda t} w$)

$$\frac{\partial w}{\partial t} + \mathcal{L}w + \lambda w = e^{-\lambda t} f$$

New bilinear form:

$$\tilde{a}(w,v) := a(w,v) + \lambda (w,v)_{L^2(\Omega)}$$
$$\Longrightarrow \tilde{a}(w,w) := a(w,w) + \lambda \|w\|^2_{L^2(\Omega)}$$

Theorem

Suppose that the bilinear form $a(\cdot,\cdot)$ is continuous and weakly coercive. Moreover, we require $u_0 \in \mathrm{L}^2(\Omega)$ and $f \in \mathrm{L}^2(Q)$. Then, problem (5) admits a unique solution $u \in C^0(\mathbb{R}^+; \mathrm{L}^2(\Omega))$; moreover $u \in \mathrm{L}^2(\mathbb{R}^+; V)$ and $\frac{\partial u}{\partial t} \in L^2(\mathbb{R}^+; V')$ (that is $u \in H^1(\mathbb{R}^+; V, V')$ – notation of Prof. Salsa).

Proof. see [QV, Sect. 11.1.1]

For the definition of these functional spaces, see [Q, Sect. 2.7] and [S].

Some a priori estimates of the solution u will be provided later.

[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

[QV] A. Quarteroni, A. Valli *Numerical Approximation of Partial Differential Equations*, Springer, 1994

[S] S. Salsa, *Partial Differential Equations in Action – from Modelling to Theory*, Springer, 2008

Galerkin approximation

For each $t > 0$, find $u_h(t) \in V_h$ such that

$$\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, d\Omega + a(u_h(t), v_h) = \int_\Omega f(t) v_h \, d\Omega \quad \forall v_h \in V_h \tag{6}$$

with $u_h(0) = u_{0h}$, where $V_h \subset V$ is a suitable space of finite dimension and u_{0h} is a convenient approximation of u_0 in the space V_h.

Such problem is called semi-discretization of (5), as the temporal variable has not yet been discretized.

Algebraic formulation

We introduce a basis $\{\varphi_j\}$ for V_h and we observe that it suffices that (6) is verified for the basis functions in order to be satisfied by all the functions of the subspace.

Moreover, since for each $t > 0$ the solution to the Galerkin problem belongs to the subspace as well, we will have

$$u_h(\mathbf{x}, t) = \sum_{j=1}^{N_h} u_j(t)\varphi_j(\mathbf{x}),$$

where the coefficients $\{u_j(t)\}$ represent the unknowns of problem (6).

Denoting by $\dot{u}_j(t)$ the derivatives of the function $u_j(t)$ with respect to time, (6) becomes

$$\int_\Omega \sum_{j=1}^{N_h} \dot{u}_j(t)\varphi_j\varphi_i \, d\Omega + a\left(\sum_{j=1}^{N_h} u_j(t)\varphi_j, \varphi_i\right) = \int_\Omega f(t)\phi_i \, d\Omega,$$
$$i = 1, 2, \ldots, N_h,$$

that is

$$\sum_{j=1}^{N_h} \dot{u}_j(t) \underbrace{\int_\Omega \varphi_j\varphi_i \, d\Omega}_{m_{ij}} + \sum_{j=1}^{N_h} u_j(t) \underbrace{a(\varphi_j, \varphi_i)}_{a_{ij}} = \underbrace{\int_\Omega f(t)\phi_i \, d\Omega}_{f_i(t)}, \tag{7}$$
$$i = 1, 2, \ldots, N_h.$$

If we define the vector of unknowns $\mathbf{u} = (u_1(t),\ u_2(t),\ \ldots, u_{N_h}(t))^T$, the *mass matrix* $\mathrm{M} = [m_{ij}]$, the stiffness matrix $\mathrm{A} = [a_{ij}]$ and the right-hand side vector $\mathbf{f} = (f_1(t),\ f_2(t), \ldots, f_{N_h}(t))^T$, the system (7) can be rewritten in matrix form as

$$\mathrm{M}\dot{\mathbf{u}}(t) + \mathrm{A}\mathbf{u}(t) = \mathbf{f}(t).$$

Time discretization

For the numerical solution of this ODE system, many finite difference methods are available. See, e.g., [QSS, Chap. 11]. Here we limit ourselves to considering the so-called θ-method.

The latter discretizes the temporal derivative by a simple difference quotient and replaces the other terms with a linear combination of the value at time t^k and of the value at time t^{k+1}, depending on the real parameter θ ($0 \leq \theta \leq 1$),

$$\mathrm{M}\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + \mathrm{A}[\theta \mathbf{u}^{k+1} + (1-\theta)\mathbf{u}^k] = \theta \mathbf{f}^{k+1} + (1-\theta)\mathbf{f}^k. \qquad (8)$$

The real positive parameter $\Delta t = t^{k+1} - t^k$, $k = 0, 1, \ldots$, denotes the discretization step (here assumed to be constant), while the superscript k indicates that the quantity under consideration refers to the time t^k.

[QSS] A. Quarteroni, R. Sacco, F. Saleri, *Numerical Mathematics*, Springer, 2007

Let us see some particular cases of (8):

- for $\theta = 0$ we obtain the forward Euler (or *explicit* Euler) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + \mathrm{A}\mathbf{u}^k = \mathbf{f}^k$$

  which is accurate to order one with respect to Δt;

- for $\theta = 1$ we have the backward Euler (or *implicit* Euler) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + \mathrm{A}\mathbf{u}^{k+1} = \mathbf{f}^{k+1},$$

  also of first order with respect to Δt;

- for $\theta = 1/2$ we have the Crank-Nicolson (or trapezoidal) method

$$\mathrm{M}\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + \frac{1}{2}\mathrm{A}\left(\mathbf{u}^{k+1} + \mathbf{u}^k\right) = \frac{1}{2}\left(\mathbf{f}^{k+1} + \mathbf{f}^k\right)$$

  which is of second order in Δt. (More precisely, $\theta = 1/2$ is the only value for which we obtain a second-order method.)

Let us consider the two extremal cases, $\theta = 0$ and $\theta = 1$. For both, we obtain a system of linear equations:

1. if $\theta = 0$, the system to solve has matrix $\frac{\mathrm{M}}{\Delta t}$
2. if $\theta = 1$, the system to solve has matrix $\frac{\mathrm{M}}{\Delta t} + A$

We observe that the M matrix is invertible, being positive definite.

In the $\theta = 0$ case, the scheme is not unconditionally stable (see [Q, Sect. 5.4]) and in the case where V_h is a subspace of finite elements we have the following stability condition (see [Q, Sect. 5.4])

$$\exists c > 0 \quad : \quad \Delta t \leq c h^2 \qquad \forall h > 0,$$

so Δt cannot be chosen irrespective of h.

In this case, if we make M diagonal, we actually decouple the system. This operation is performed by the so-called *lumping* of the mass matrix (see [Q, Sect. 13.5]).

[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

In case $\theta > 0$, the system will have the form $K\mathbf{u}^{k+1} = \mathbf{g}$, where

- $\mathbf{g}$ is the source term
- $\mathrm{K} = \frac{\mathrm{M}}{\Delta t} + \theta \mathrm{A}$

Such matrix is however invariant in time (the operator L, and therefore the matrix A, being independent of time); if the spacial mesh does not change, it can then be factorized once and for all at the beginning of the process.

Since M is symmetric, if A is symmetric too, the K matrix associated to the system will also be symmetric. Hence, we can use, for instance, the Cholesky factorization, K=H H^T, H being lower triangular. At each time step, we will therefore have to solve two triangular systems in N_h unknowns:

$$\begin{aligned} &\mathrm{H}\mathbf{y} = \mathbf{g}, \\ &\mathrm{H}^T\mathbf{u}^{k+1} = \mathbf{y} \end{aligned}$$

(see [QSS, Chap. 3]).

[QSS] A. Quarteroni, R. Sacco, F. Saleri, *Numerical Mathematics*, Springer, 2007

A priori estimates

Let us consider problem (5); since the corresponding equations must hold for each $v \in V$, it will be legitimate to set $v = u(t)$ (t being given), solution of the problem itself, yielding

$$\int_\Omega \frac{\partial u(t)}{\partial t} u(t)\, d\Omega + a(u(t), u(t)) = \int_\Omega f(t) u(t)\, d\Omega \qquad \forall t > 0. \tag{9}$$

Considering the individual terms, we have

$$\int_\Omega \frac{\partial u(t)}{\partial t} u(t)\, d\Omega = \frac{1}{2}\frac{\partial}{\partial t}\int_\Omega |u(t)|^2 d\Omega = \frac{1}{2}\frac{\partial}{\partial t}\|u(t)\|^2_{L^2(\Omega)}. \tag{10}$$

If we assume for simplicity that the bilinear form is coercive (with coercivity constant equal to α), we obtain

$$a(u(t), u(t)) \geq \alpha \|u(t)\|^2_V,$$

while thanks to the Cauchy-Schwarz inequality, we find

$$(f(t), u(t)) \leq \|f(t)\|_{\mathrm{L}^2(\Omega)} \|u(t)\|_{\mathrm{L}^2(\Omega)}. \tag{11}$$

In the remainder, we will often use *Young's inequality*

$$\forall a, b \in \mathbb{R}, \qquad ab \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2 \qquad \forall \varepsilon > 0, \tag{12}$$

which descends from the elementary inequality

$$\left(\sqrt{\varepsilon}\, a - \frac{1}{2\sqrt{\varepsilon}}\, b \right)^2 \geq 0.$$

Using first Poincaré' inequality and Young's inequality, we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha \|\nabla u(t)\|^2_{L^2(\Omega)} &\leq \|f(t)\|_{L^2(\Omega)} \|u(t)\|_{L^2(\Omega)} \\ &\leq \frac{C_\Omega^2}{2\alpha}\|f(t)\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\nabla u(t)\|^2_{L^2(\Omega)}. \end{aligned} \tag{13}$$

Poincaré inequalilty

If Γ_D is a set of positive measure, then:

$$\exists\, C_\Omega > 0 \colon \|v\|_{L^2(\Omega)} \leq C_\Omega \|\nabla v\|_{L^2(\Omega)} \qquad \forall\, v \in H^1_{\Gamma_D}(\Omega)$$

Then, by integrating in time we obtain, for all $t > 0$,

$$\|u(t)\|^2_{L^2(\Omega)} + \alpha \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)} ds \leq \|u_0\|^2_{L^2(\Omega)} + \frac{C_\Omega^2}{\alpha} \int_0^t \|f(s)\|^2_{L^2(\Omega)} ds \tag{14}$$

This is an a priori energy estimate. Different kinds of a priori estimates can be obtained as follows. Note that

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} = \|u(t)\|_{L^2(\Omega)} \frac{d}{dt}\|u(t)\|_{L^2(\Omega)}$$

Then from (9), using (10) and (11) we obtain (still using the Poincaré inequality)

$$\|u(t)\|_{L^2(\Omega)} \frac{d}{dt}\|u(t)\|_{L^2(\Omega)} + \frac{\alpha}{C_\Omega}\|u(t)\|_{L^2(\Omega)}\|\nabla u(t)\|_{L^2(\Omega)} \leq \|f(t)\|_{L^2(\Omega)}\|u(t)\|_{L^2(\Omega)}, \qquad t > 0.$$

If $\|u(t)\|_{L^2(\Omega)} \neq 0$ (otherwise we should proceed differently, even though the final result is still true) we can divide by $\|u(t)\|_{L^2(\Omega)}$ and integrate in time to obtain

$$\|u(t)\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \quad t > 0. \tag{15}$$

This is a further a priori estimate.

Let us now use the first inequality in (13) and integrate in time to yield

$$
\begin{aligned}
&\|u(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u(s)\|^2 ds \\
&\overset{(13)}{\leq} \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \|u(s)\|_{L^2(\Omega)} ds \\
&\overset{(15)}{\leq} \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \cdot (\|u_0\|_{L^2(\Omega)} + \int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau) ds \\
&= \|u_0\|^2_{L^2(\Omega)} + 2\int_0^t \|f(s)\|_{L^2(\Omega)} \|u_0\|_{L^2(\Omega)} \\
&+ 2\int_0^t \|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau \\
&= (\|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\| ds)^2. \qquad (16)
\end{aligned}
$$

The latter equality follows upon noticing that

$$
\|f(s)\|_{L^2(\Omega)} \int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau = \frac{d}{ds} (\int_0^s \|f(\tau)\|_{L^2(\Omega)} d\tau)^2.
$$

We therefore conclude with the additional a priori estimate

$$
\begin{aligned}
&\left(\|u(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)} ds\right)^{\frac{1}{2}} \\
&\qquad \leq \|u_0\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \qquad t > 0.
\end{aligned} \tag{17}
$$

A priori estimate for the Galerkin problem

We have seen that we can formulate the Galerkin problem (6) for problem (5) and that the latter, under suitable hypotheses, admits a unique solution. Similarly to what we did for problem (5) (see estimate (14)) we can prove the following a priori (stability) estimates for the solution to problem (6):

$$\|u_h(t)\|^2_{L^2(\Omega)} + \alpha \int_0^t \|\nabla u_h(s)\|^2_{L^2(\Omega)} ds \leq \|u_{0h}\|^2_{L^2(\Omega)} + \frac{C_\Omega^2}{\alpha} \int_0^t \|f(s)\|^2_{L^2(\Omega)} ds, \qquad t > 0. \tag{18}$$

For its proof we can take, for every $t > 0$, $v_h = u_h(t)$ and proceed as we did to obtain (14). Then, by recalling that the initial data is $u_h(0) = u_{0h}$, we can deduce the following discrete counterparts of (15) and (17):

$$\|u_h(t)\|_{L^2(\Omega)} \leq \|u_{0h}(t)\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \qquad t > 0 \tag{19}$$

and

$$\begin{aligned} &(\|u_h(t)\|^2_{L^2(\Omega)} + 2\alpha \int_0^t \|\nabla u_h(s)\|^2_{L^2(\Omega)} ds)^{\frac{1}{2}} \\ &\qquad \leq \|u_{0h}\|_{L^2(\Omega)} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds, \qquad t > 0. \end{aligned} \tag{20}$$

Convergence analysis of the semi-discrete problem

Theorem

There exists a constant $C > 0$ independent of both t and h such that

$$\{\|u(t) - u_h(t)\|^2_{L^2(\Omega)} + \alpha \int_0^t \|\nabla u(s) - \nabla u_h(s)\|^2_{L^2(\Omega)} ds\}^{1/2}$$
$$\leq Ch^r \{|u_0|^2_{H^r(\Omega)} + \int_0^t |u(s)|^2_{H^{r+1}(\Omega)} ds + \int_0^t |\frac{\partial u(s)}{\partial s}|^2_{H^{r+1}(\Omega)} ds\}^{1/2}.$$

Proof. [Q, Sect. 5.3]

Further error estimates are proven, e.g. in [QV, Chap. 11]

[QV] A. Quarteroni, A. Valli *Numerical Approximation of Partial Differential Equations*, Springer, 1994

[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

Stability analysis of the θ-method

We now analyze the stability of the fully discretized problem. Applying the θ-method to the Galerkin problem (6) we obtain

$$\begin{aligned}\left(\frac{u_h^{k+1}-u_h^k}{\Delta t}, v_h\right) + a\left(\theta u_h^{k+1} + (1-\theta)u_h^k, v_h\right) \\ = \theta F^{k+1}(v_h) + (1-\theta)F^k(v_h) \qquad \forall v_h \in V_h,\end{aligned} \tag{21}$$

for each $k \geq 0$, with $u_h^0 = u_{0h}$.

F^k indicates that the functional is evaluated at time t^k.

We will limit ourselves to the case where $F = 0$ and start to consider the case of the implicit Euler method ($\theta = 1$) that is

$$\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}, v_h \right) + a\left(u_h^{k+1}, v_h\right) = 0 \qquad \forall v_h \in V_h.$$

By choosing $v_h = u_h^{k+1}$, we obtain

$$(u_h^{k+1}, u_h^{k+1}) + \Delta t \; a\left(u_h^{k+1}, u_h^{k+1}\right) = (u_h^k, u_h^{k+1}).$$

By exploiting the following inequalities

$$a(u_h^{k+1}, u_h^{k+1}) \geq \alpha \|u_h^{k+1}\|_V^2, \qquad (u_h^k, u_h^{k+1}) \leq \frac{1}{2}\|u_h^k\|_{\mathrm{L}^2(\Omega)}^2 + \frac{1}{2}\|u_h^{k+1}\|_{\mathrm{L}^2(\Omega)}^2,$$

the former deriving from the coercivity of the bilinear form $a(\cdot, \cdot)$, and the latter from the Cauchy-Schwarz and Young inequalities, we obtain

$$\|u_h^{k+1}\|_{\mathrm{L}^2(\Omega)}^2 + 2\alpha\Delta t \|u_h^{k+1}\|_V^2 \leq \|u_h^k\|_{\mathrm{L}^2(\Omega)}^2. \tag{22}$$

Observing that $\|u_h^{k+1}\|_V \geq \|u_h^{k+1}\|_{\mathrm{L}^2(\Omega)}$, we deduce from (22) that:

$$(1 + 2\alpha\Delta t)\,\|u_h^{k+1}\|^2_{\mathrm{L}^2(\Omega)} \leq \|u_h^k\|^2_{\mathrm{L}^2(\Omega)}.$$

hence:

$$\|u_h^{k+1}\|_{\mathrm{L}^2(\Omega)} \leq \frac{1}{\sqrt{1 + 2\alpha\Delta t}}\|u_h^k\|_{\mathrm{L}^2(\Omega)}$$

which entails:

$$\|u_h^k\|_{\mathrm{L}^2(\Omega)} \leq \left(\frac{1}{\sqrt{1 + 2\alpha\Delta t}}\right)^k \|u_{0h}\|_{\mathrm{L}^2(\Omega)}$$

and therefore

$$\lim_{k\to\infty} \|u_h^k\|_{\mathrm{L}^2(\Omega)} = 0,$$

that is the backward Euler method is absolutely stable without any restriction on the time step Δt.

Assume now $f \neq 0$

$$\underbrace{\left(\frac{u_h^{k+1} - u_h^k}{\Delta t}, u_h^{k+1}\right)}_{(I)} + \underbrace{a(u_h^{k+1}, u_h^{k+1})}_{(II)} = \underbrace{\int_\Omega f^{k+1} u_h^{k+1}}_{(III)}$$

$$(I) \geq \frac{1}{2\,\Delta t}\left(\|u_h^{k+1}\|_{L^2}^2 - \|u_h^k\|_{L^2}^2\right)^{1}$$

$$(II) \geq \alpha \|u_h^{k+1}\|_V^2 \qquad \text{(coercivity of } a(\cdot,\cdot)\text{)}$$

$$(III) \overset{\text{(C.S.)}}{\leq} \|f^{k+1}\|_{L^2}\|u_h^{k+1}\|_V \overset{\text{(Young)}}{\leq} \frac{1}{2\alpha}\|f^{k+1}\|_{L^2}^2 + \frac{\alpha}{2}\|u_h^{k+1}\|_V^2$$

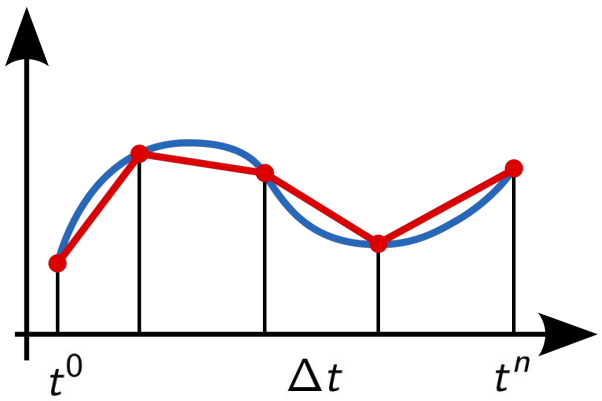
[1] indeed: $(a-b, a) \geq \frac{1}{2}(\|a\|^2 - \|b\|^2) \quad \forall\, a, b$

Then, after summation on k, for $k = 0, \dots, n-1$:

$$\|u_h^n\|_{L^2}^2 + \underbrace{\alpha \sum_{k=1}^{n} \Delta t \|u_h^k\|_V^2}_{\simeq \alpha \int_0^{t^n} \|u_h(t)\|_V^2 \, dt} \leq \|u_{0,h}\|_{L^2}^2 + \underbrace{\frac{1}{\alpha} \sum_{k=1}^{n} \Delta t \|f^k\|_{L^2}^2}_{\simeq \frac{1}{\alpha} \int_0^{t^n} \|f(t)\|_{L^2}^2 \, dt}$$

$\rightarrow$ Unconditional stability (no restriction on Δt)

Remainder: composite trapezoidal rule



$$I = \int_0^{t^n} \int_\Omega f^2(t) dx\, dt$$

$$I \simeq I_{\Delta t} := \frac{1}{2}\Delta t \|f^{(0)}\|^2_{L^2(\Omega)} + \sum_{k=1}^{n-1} \Delta t \|f^{(k)}\|^2_{L^2(\Omega)} + \frac{1}{2}\Delta t \|f^{(n)}\|^2_{L^2(\Omega)} + \mathcal{O}(\Delta t^4)$$

Therefore:

$$\sum_{k=1}^{n} \Delta t \|f^{(k)}\|^2_{L^2(\Omega)} \leq 2\, I_{\Delta t} \simeq 2\|f\|^2_{L^2(0,t^n;L^2(\Omega))} + \mathcal{O}(\Delta t^4)$$

Before analyzing the general case where θ is an arbitrary parameter ranging between 0 and 1, we introduce the following definition.

Definition

We say that the scalar λ is an eigenvalue of the bilinear form $a(\cdot,\cdot) : V \times V \mapsto \mathbb{R}$ and that $w \in V$ is its corresponding *eigenfunction* if it turns out that

$$a(w, v) = \lambda(w, v) \qquad \forall v \in V.$$

If the bilinear form $a(\cdot,\cdot)$ is symmetric and coercive, it has positive, real eigenvalues forming an infinite sequence; moreover, its eigenfunctions form a basis of the space V.

The eigenvalues and eigenfunctions of $a(\cdot,\cdot)$ can be approximated by finding the pairs $\lambda_h \in \mathbb{R}$ and $w_h \in V_h$ which satisfy

$$a(w_h, v_h) = \lambda_h (w_h, v_h) \quad \forall v_h \in V_h. \tag{23}$$

From an algebraic viewpoint, problem (23) can be formulated as follows

$$\mathrm{A}\mathbf{w} = \lambda_h \mathrm{M}\mathbf{w},$$

where A is the stiffness matrix and M the mass matrix. We are therefore dealing with a generalized eigenvalue problem.
Such eigenvalues are all positive and N_h in number (N_h being as usual the dimension of the subspace V_h); after ordering them in ascending order, $\lambda_h^1 \leq \lambda_h^2 \leq \ldots \leq \lambda_h^{N_h}$, we have

$$\lambda_h^{N_h} \to \infty \qquad \text{for } N_h \to \infty.$$

Moreover, the corresponding eigenfunctions form a basis for the subspace V_h and can be chosen to be *orthonormal* with respect to the scalar product of $\mathrm{L}^2(\Omega)$. This means that, denoting by w_h^i the eigenfunction corresponding to the eigenvalue λ_h^i, we have $(w_h^i, w_h^j) = \delta_{ij}$ $\forall i, j = 1, \ldots, N_h$.

Thus, each function $v_h \in V_h$ can be represented as follows

$$v_h(\mathbf{x}) = \sum_{j=1}^{N_h} v_j w_h^j(\mathbf{x})$$

and, thanks to the eigenfunction orthonormality,

$$\|v_h\|^2_{\mathrm{L}^2(\Omega)} = \sum_{j=1}^{N_h} v_j^2. \tag{24}$$

Let us consider an arbitrary $\theta \in [0, 1]$ and let us limit ourselves to the case where the bilinear form $a(\cdot, \cdot)$ is symmetric (otherwise, although the final stability result holds in general, the following proof would not work, as the eigenfunctions would not necessarily form a basis).

Since $u_h^k \in V_h$, we can write

$$u_h^k(\mathbf{x}) = \sum_{j=1}^{N_h} u_j^k w_h^j(\mathbf{x}).$$

We observe that in this modal expansion, the u_j^k no longer represent the nodal values of u_h^k.

If we now set $F = 0$ in (21) and take $v_h = w_h^i$, we find

$$\frac{1}{\Delta t}\sum_{j=1}^{N_h}[u_j^{k+1} - u_j^k]\left(w_h^j, w_h^i\right) + \sum_{j=1}^{N_h}[\theta u_j^{k+1} + (1-\theta)u_j^k]a(w_h^j, w_h^i) = 0,$$

for each $i = 1, \ldots, N_h$. For each pair $i, j = 1, \ldots, N_h$ we have

$$a(w_h^j, w_h^i) = \lambda_h^j(w_h^j, w_h^i) = \lambda_h^j \delta_{ij} = \lambda_h^i,$$

and thus, for each $i = 1, \ldots, N_h$,

$$\frac{u_i^{k+1} - u_i^k}{\Delta t} + [\theta u_i^{k+1} + (1-\theta)u_i^k]\lambda_h^i = 0.$$

Solving now for u_i^{k+1}, we find

$$u_i^{k+1} = u_i^k \frac{1 - (1-\theta)\lambda_h^i \Delta t}{1 + \theta \lambda_h^i \Delta t}.$$

Recalling (24), we can conclude that for the method to be absolutely stable, we must impose the inequality

$$\left| \frac{1 - (1-\theta)\lambda_h^i \Delta t}{1 + \theta \lambda_h^i \Delta t} \right| < 1,$$

that is

$$-1 - \theta\lambda_h^i \Delta t < 1 - (1-\theta)\lambda_h^i \Delta t < 1 + \theta\lambda_h^i \Delta t.$$

Hence,

$$-\frac{2}{\lambda_h^i \Delta t} - \theta < \theta - 1 < \theta.$$

The second inequality is always verified, while the first one can be rewritten as

$$2\theta - 1 > -\frac{2}{\lambda_h^i \Delta t}.$$

If $\theta \geq 1/2$, the left-hand side is non-negative, while the right-hand side is negative, so the inequality holds for each Δt. Instead, if $\theta < 1/2$, the inequality is satisfied (hence the method is stable) only if

$$\Delta t < \frac{2}{(1-2\theta)\lambda_h^i}. \tag{25}$$

As such relation must hold for all the eigenvalues λ_h^i of the bilinear form, it will suffice to require that it holds for the largest among them, which we have supposed to be $\lambda_h^{N_h}$.

To summarize, we have:

- if $\theta \geq 1/2$, the θ-method is unconditionally absolutely stable, i.e. it is absolutely stable for each Δt;
- if $\theta < 1/2$, the θ-method is absolutely stable only for $\Delta t \leq \frac{2}{(1-2\theta)\lambda_h^{N_h}}$.

Thanks to the definition of eigenvalue (23) and to the continuity property of $a(\cdot,\cdot)$, we deduce

$$\lambda_h^{N_h} = \frac{a(w_{N_h}, w_{N_h})}{\|w_{N_h}\|^2_{\mathrm{L}^2(\Omega)}} \leq \frac{M\|w_{N_h}\|^2_V}{\|w_{N_h}\|^2_{\mathrm{L}^2(\Omega)}} \leq M(1 + C^2 h^{-2}).$$

The constant $C > 0$ which appears in the latter step derives from the following inverse inequality [QV, Chap. 3]

$$\exists C > 0 \ : \ \|\nabla v_h\|_{\mathrm{L}^2(\Omega)} \leq C h^{-1} \|v_h\|_{\mathrm{L}^2(\Omega)} \qquad \forall v_h \in V_h,$$

[QV] A. Quarteroni, A. Valli *Numerical Approximation of Partial Differential Equations*, Springer, 1994

Hence, for h small enough, $\lambda_h^{N_h} \leq Ch^{-2}$. In fact, we can prove that $\lambda_h^{N_h}$ is indeed of the order of h^{-2}, that is

$$\lambda_h^{N_h} = max_i \lambda_h^i \simeq ch^{-2}.$$

Keeping this into account, we obtain that for $\theta < 1/2$ the method is absolutely stable only if

$$\Delta t \leq C(\theta)h^2, \tag{26}$$

where $C(\theta)$ denotes a positive constant depending on θ.

The latter relation implies that for $\theta < 1/2$, Δt cannot be chosen arbitrarily but is bound to the choice of h.

Convergence analysis of the θ-method

Theorem

Under the hypothesis that u_0, f and the exact solution are sufficiently regular, the following *a priori error estimate* holds: $\forall n \geq 1$,

$$\|u(t^n) - u_h^n\|^2_{\mathrm{L}^2(\Omega)} + 2\alpha\Delta t \sum_{k=1}^{n} \|u(t^k) - u_h^k\|^2_V \leq C(u_0, f, u)(\Delta t^{p(\theta)} + h^{2r}),$$

where $p(\theta) = 2$ if $\theta \neq 1/2$, $p(1/2) = 4$ and C depends on its arguments but not on h and Δt.

Proof. See [Q, Sect. 5.5]

[Q] A. Quarteroni, *Numerical Models for Differential Problems*, 3rd Ed. , Springer, 2018

Parbolic ADR equation

Consider the parabolic PDE, where $\Omega \subset \mathbb{R}^2$ is an open bounded domain:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mu \Delta u + \boldsymbol{\beta} \cdot \nabla u + \sigma u = f & \text{in } \Omega \times (0, \mathrm{T}), \\ u = 0 & \text{on } \partial\Omega \times (0, \mathrm{T}), \\ u(0) = u_0 & \text{in } \Omega, \end{cases} \tag{27}$$

and where μ, $\boldsymbol{\beta}$, σ and f are regular functions, satisfying:

$$\begin{aligned} 0 < \mu_0 \leq \mu \leq \mu_1 \quad & \text{a.e. in } \Omega \\ |\boldsymbol{\beta}| \leq b_1 \quad & \text{a.e. in } \Omega \\ 0 < \sigma_0 \leq \sigma \leq \sigma_1 \quad & \text{a.e. in } \Omega \end{aligned}$$

Introducing a finite dimensional space $V_h \subset H_0^1(\Omega)$, the semi-discrete Galerkin formulation reads: for all $t \in (0, T]$ find $u_h(t) \in V_h$ such that

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial u_h(t)}{\partial t} v_h \, dx + \int_\Omega \mu \nabla u_h(t) \cdot \nabla v_h + \int_\Omega \boldsymbol{\beta} \cdot \nabla u_h(t) v_h + \int_\Omega \sigma u_h(t) v_h \\ \qquad = \int_\Omega f v_h \quad \forall \, v_h \in V_h, \end{cases} \tag{28}$$

and such that $u_h(0) = u_{0,h}$, where $u_{0,h}$ is the projection of the initial condition into V_h.

A semimplicit scheme (Please cover this part even if not carried out during classes)

We consider a time-advancing scheme, where the diffusion and reaction terms are treated implicitly, while the advection term is treated explicitly. Let us denote $t_k = k\Delta t$, for $k = 0, \dots, N$, where $\Delta t = T/N$. Let $u_h^{(k)}$ be the approximation of $u(t_k)$. A fully discretized version of (27) reads:

$$\begin{cases} \left(\dfrac{u_h^{(k+1)} - u_h^{(k)}}{\Delta t}, v_h\right) + \left(\mu \nabla u_h^{(k+1)}, \nabla v_h\right) + \left(\boldsymbol{\beta} \cdot \nabla u_h^{(k)}, v_h\right) \\ \qquad + \left(\sigma u_h^{(k+1)}, v_h\right) = (f, v_h) \quad \forall\, v_h \in V_h, \quad k = 0, \dots, N-1 \\ u_h^{(0)} = u_{0,h} \end{cases} \tag{29}$$

where $(\cdot, \cdot)$ denotes the $L^2(\Omega)$ scalar product.

Stability analysis of the semimplicit scheme

Theorem

If the coefficients of the problem satisfy

$$b_1^2 < 4\mu_0\sigma_0, \tag{30}$$

then the semimplicit scheme (29) is absolutely stable for any choice of Δt. Consider now the case $\sigma = 0$. If the coefficients of the problem satisfy (C_p being the Poicaré constant)

$$b_1 < \mu_0 / C_p, \tag{31}$$

then the scheme is absolutely stable for any choice of Δt.

Proof.

Let us choose $v_h = u_h^{(k+1)}$. We have:

$$\left(\mu \nabla u_h^{(k+1)}, \nabla u_h^{(k+1)}\right) \geq \mu_0 \|\nabla u_h^{(k+1)}\|^2,$$
$$\left(\sigma u_h^{(k+1)}, u_h^{(k+1)}\right) \geq \sigma_0 \|u_h^{(k+1)}\|^2,$$

which entails, for every k

$$\|u_h^{(k+1)}\|^2 + \Delta t\, \mu_0 \|\nabla u_h^{(k+1)}\|^2 + \Delta t\, \sigma_0 \|u_h^{(k+1)}\|^2 \leq \left|\left(u_h^{(k)}, u_h^{(k+1)}\right)\right| + \Delta t \left|\left(\boldsymbol{\beta} \cdot \nabla u_h^{(k)}, u_h^{(k+1)}\right)\right|$$

The two right-hand side terms can be bounded by combining the Cauchy-Schwarz and the Young inequalities:

$$\left|\left(u_h^{(k)}, u_h^{(k+1)}\right)\right| \leq \frac{1}{2\eta_1}\|u_h^{(k)}\|^2 + \frac{\eta_1}{2}\|u_h^{(k+1)}\|^2,$$

$$\left|\left(\boldsymbol{\beta}\cdot\nabla u_h^{(k)}, u_h^{(k+1)}\right)\right| \leq \frac{b_1}{2\eta_2}\|\nabla u_h^{(k)}\|^2 + \frac{\eta_2 b_1}{2}\|u_h^{(k+1)}\|^2,$$

where the positive constants η_1 and η_2 will be later fixed according to our best convenience.
We end up with the following inequality:

$$\underbrace{\left[1 + \Delta t\,\sigma_0 - \frac{\eta_1}{2} - \frac{\Delta t\,\eta_2 b_1}{2}\right]}_{A}\|u_h^{(k+1)}\|^2 + \underbrace{\Delta t\,\mu_0}_{B}\|\nabla u_h^{(k+1)}\|^2$$

$$\leq \underbrace{\frac{1}{2\eta_1}}_{A'}\|u_h^{(k)}\|^2 + \underbrace{\frac{\Delta t\, b_1}{2\eta_2}}_{B'}\|\nabla u_h^{(k)}\|^2$$

In order to prove stability, we need $A > A'$ and $B > B'$. Indeed, if this were true, then we would have

$$A\|u_h^{(k+1)}\|^2 + B\|\nabla u_h^{(k+1)}\|^2 \leq \max\left(\frac{A'}{A}, \frac{B'}{B}\right)\left[A\|u_h^{(k)}\|^2 + B\|\nabla u_h^{(k)}\|^2\right],$$

which is the seeked stability result in the norm $\|\cdot\|_{A,B} := (A\|\cdot\|^2 + B\|\nabla\cdot\|^2)^{1/2}$, equivalent to the standard V_h norm. Therefore, we look for a suitable choice (if it exists) of η_1 and η_2 that ensures $A > A'$ and $B > B'$. The second inequality is satifsied if and only if

$$\eta_2 = \frac{b_1 + \epsilon}{2\mu_0}$$

for some $\epsilon > 0$.

Hence, the first inequality reads

$$1 + \Delta t\, \sigma_0 - \frac{\Delta t\, b_1(b_1 + \epsilon)}{4\mu_0} > \frac{1}{2\eta_1} + \frac{\eta_1}{2}$$

The right-hand side is minimized for $\eta_1 = 1$, thus leading to the condition

$$4\frac{\mu_0\sigma_0}{b_1(b_1 + \epsilon)} > 1. \tag{32}$$

Clearly, it is possible to find $\epsilon > 0$ such that (32) holds if and only if

$$b_1^2 < 4\mu_0\sigma_0. \tag{33}$$

In conclusion, whenever the coefficients of the problem satisfy the condition (33), the scheme (29) is absolutely stable, for any choice of Δt.

Let us consider now the case $\sigma = 0$. Proceeding as before, we have:

$$\left[1 - \frac{\eta_1}{2} - \frac{\Delta t\, \eta_2 b_1}{2}\right] \|u_h^{(k+1)}\|^2 + \Delta t\, \mu_0 \|\nabla u_h^{(k+1)}\|^2$$
$$\leq \frac{1}{2\eta_1} \|u_h^{(k)}\|^2 + \frac{\Delta t\, b_1}{2\eta_2} \|\nabla u_h^{(k)}\|^2$$

Let us introduce a constant $\omega \in (0, 1)$ (to be fixed later). By the Poincaré inequality, we have:

$$\|\nabla u_h^{(k+1)}\|^2 = (1-\omega)\|\nabla u_h^{(k+1)}\|^2 + \omega \|\nabla u_h^{(k+1)}\|^2$$
$$\geq \frac{1-\omega}{C_p^2} \|u_h^{(k+1)}\|^2 + \omega \|\nabla u_h^{(k+1)}\|^2.$$

Combining the latter inequalities, we have

$$\underbrace{\left[1 - \frac{\eta_1}{2} - \frac{\Delta t\, \eta_2 b_1}{2} + \frac{(1-\omega)\Delta t\, \mu_0}{C_p^2}\right]}_{A} \|u_h^{(k+1)}\|^2 + \underbrace{\omega \Delta t\, \mu_0}_{B} \|\nabla u_h^{(k+1)}\|^2$$

$$\leq \underbrace{\frac{1}{2\eta_1}}_{A'} \|u_h^{(k)}\|^2 + \underbrace{\frac{\Delta t\, b_1}{2\eta_2}}_{B'} \|\nabla u_h^{(k)}\|^2$$

As in the previous point, we look for conditions on the coefficients such that $A > A'$ and $B > B'$. The second inequality is satifsied if and only if

$$\eta_2 = \frac{b_1 + \epsilon}{2\omega\mu_0}$$

for some $\epsilon > 0$.

Then, the first inequality reads

$$1 - \frac{\Delta t\, b_1(b_1+\epsilon)}{4\omega\mu_0} + \frac{(1-\omega)\Delta t\,\mu_0}{C_p^2} > \frac{1}{2\eta_1} + \frac{\eta_1}{2}$$

The right-hand side is minimized for $\eta_1 = 1$. Rearranging the terms, we get

$$-\omega^2 + \omega - \frac{b_1(b_1+\epsilon)C_p^2}{4\mu_0^2} > 0 \tag{34}$$

Real solutions $\omega \in (0,1)$ exists whenever the discriminant is positive, that is:

$$b_1(b_1+\epsilon)C_p^2 < \mu_0^2.$$

The latter condition can be satisfied (by suitable choosing ϵ) if and only if

$$b_1 < \mu_0/C_p. \tag{35}$$

In conclusion, if (35) is satisfied, the scheme is absolutely stable for any choice of Δt. □