Simulation Methods

Numerical Methods for parabolic and hyperbolic PDE

Joan Carles Tatjer

Departament de Matemàtiques i Informàtica

Universitat de Barcelona

Outline

Introduction

We consider two types of PDE: parabolic (diffusion equation) and hyperbolic (wave equation).

The heat equation

Let $\Omega \subset \mathbb{R}^n$ an open and bounded set with 'regular' boundary Γ, and $Q_T = \Omega \times (0, T)$, $\Sigma_T = \Gamma \times (0, T)$, for $T > 0$. Consider the following problem

$$\frac{\partial u}{\partial t} - \Delta u = f \text{ in } Q_T \text{ (heat equation)},$$
$$u = 0 \text{ in } \Sigma_T \text{ (boundary condition)},$$
$$u(\cdot, 0) = u_0 \text{ in } \Omega \text{ (initial condition)}.$$

If we multiply the equation by a test function $v \in \mathcal{D}(\Omega)$ (regular function $v : \Omega \to \mathbb{R}$ s.t. $\operatorname{supp} u \subset \Omega$) and integrates over Ω :

$$\int_\Omega \frac{\partial u}{\partial t}(x,t)v(x)\,dx - \int_\Omega \Delta u(x,t)v(x)\,dx = \int_\Omega f(x,t)v(x)\,dx.$$

By using the Green formula, we have

$$\frac{d}{dt}\int_\Omega u(x,t)v(x)\,dx + \sum_{i=1}^n \int_\Omega \frac{\partial u}{\partial x_i}(x,t)\frac{\partial v}{\partial x_i}(x)\,dx = \int_\Omega f(x,t)v(x)\,dx.$$

If we define for all $\varphi, \psi \in L^2(\Omega)$

$$(\varphi, \psi) = \int_\Omega \varphi(x)\psi(x)\, dx$$

and for $\varphi, \psi \in L^2(\Omega)$ such that their partial derivatives belong also to $L^2(\Omega)$ (that is $\varphi, \psi \in H^1(\Omega)$)

$$a(\varphi, \psi) = \sum_{i=1}^{n} \int_\Omega \frac{\partial \varphi}{\partial x_i} \frac{\partial \psi}{\partial x_i}\, dx.$$

The we have the variational formulation of the heat problem:
Find a function $u : t \in [0, T] \mapsto u(t) \in H_0^1(\Omega) = \overline{\mathcal{D}(\Omega)}$ such that

$$\forall\, v \in H_0^1(\Omega), \quad \frac{d}{dt}(u(t), v) + a(u(t), v) = (f(t), v),$$

$$u(0) = u_0,$$

where if $\varphi \in H_0^1(\Omega)$ then $\varphi \in H^1(\Omega)$ and it vanishes at Γ.

Abstract parabolic problems

One introduces:

- Two Hilbert spaces V and H (over $\mathbb{R}$) s.t. (i) $V \subset H$ with continuous injection; (ii) V is dense in H.
- A bilinear form $u, v \mapsto a(u, v)$ continuous on $V \times V$.

A general parabolic problem is: Given $u_0 \in H$ and $f \in L^2(0, T; H)$, find a function u such that

1. $u \in L^2(0, T; V) \cap C^0(0, T; H)$,
2. $\forall\, v \in V,\ \frac{d}{dt}(u(t), v) + a(u(t), v) = (f(t), v)$.
3. $u(0) = u_0$.

We add the coercivity condition: $\exists\, \alpha > 0$ and $\lambda \in \mathbb{R}$ such that

$$\forall\, v \in V, \quad a(v, v) + \lambda |v|^2 \geq \alpha \|v\|^2,$$

and that the injection from V to H is compact (the image of a bounded set is relatively compact) and the bilinear form $a(\cdot, \cdot)$ is symmetric. Here $|\cdot|$ is the norm in H, and $\|\cdot\|$ is the norm in V.

Theorem

Under the previous hypothesis there exists a unique solution of the abstract parabolic equation given by

$$u(t) = \sum_{i \geq 1} \{(u_0, w_i)e^{-\lambda_i t} + \int_0^t (f(s), w_i)e^{-\lambda_i (t-s)}\, ds\} w_i,$$

where (w_i) is an orthonormal hilbertian basis of eigenvectors of eigenvalues $-\lambda < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_i \leq \cdots$, such that $\forall\, v \in V$, $a(w_i, v) = \lambda_i (w_i, v)$.

Comment

$(w_i)_{i\geq}$ is an orthonormal Hilbertian basis of H if $(w_i, w_j) = \delta_{ij}$, for all $i, j \geq 1$ and the linear subspace generated by $(w_i)_{i \geq 1}$ is dense in H. One can prove that if $u \in H$ then

$$u = \sum_{i=1}^{\infty} (u, w_i) w_i, \qquad |u|^2 = \sum_{i=1}^{\infty} |(u, w_i)|^2.$$

Proof.

We only see how to obtain the formula. Let u be a solution. As (w_i) is an orthonormal hilbertian basis, we have

$$u(t) = \sum_{i \geq 1} (u(t), w_i) w_i.$$

As $u(t) \in V$ (a. e.), we have

$$a(u(t), w_i) = \lambda_i (u(t), w_i).$$

From the differential equation applied to $v = w_i$, and defining $\alpha_i(t) = (u(t), w_i)$, we have that $\alpha_i(t)$ is the solution of the linear ode

$$\left\{ \begin{array}{rcl} \frac{d}{dt}\alpha_i(t) + \lambda_i \alpha_i(t) & = & (f(t), w_i), \\ \alpha_i(0) & = & (u_0, w_i), \end{array} \right.$$

with solution $\alpha_i(t) = (u_0, w_i) e^{-\lambda_i t} + \int_0^t (f(s), w_i) e^{-\lambda_i (t-s)}\, ds$. Then, we obtain the formula for $u(t)$. $\square$

Semi-discretization method

Let V_h a subspace of V of finite dimension $I = I(h)$. We consider the following approximate problem: Given $u_{0,h} \in V_h$, find a function $u_h : t \in [0, T] \mapsto u_h(t) \in V_h$ solution of the system of ode's:

$$\begin{aligned} \forall\, v_h \in V_h, \quad \frac{d}{dt}(u_h(t), v_h) + a(u_h(t), v_h) &= (f(t), v_h), \\ u_h(0) &= u_{0,h} \end{aligned}$$

Theorem

There exists an increasing sequence of eigenvalues $-\lambda < \lambda_{1,h} \leq \lambda_{2,h} \leq \cdots \leq \lambda_{I,h}$ *and an orthonomal basis* $(w_{i,h})$ *of* V_h *on* H *such that*

$$\forall\, v_h \in V_h, \quad a(w_{i,h}, v_h) = \lambda_{i,h}(w_{i,h}, v_h).$$

Proof.

We consider only the case $\lambda = 0$.
Let $(\varphi_i)_{1\le i\le I}$ a basis of V_h. We look for $u_h = \sum_{j=1}^{I} \xi_j \varphi_j$ s.t.

$$\sum_{j=1}^{I} a(\varphi_j,\varphi_i)\xi_j = \mu \sum_{j=1}^{I} (\varphi_j,\varphi_i)\xi_j, \quad 1 \le i \le I.$$

Let $R = (a(\varphi_j,\varphi_i))_{1\le i,j\le I}$, $M = ((\varphi_j,\varphi_i))_{1\le i,j\le I}$ be (resp.) the rigidity matrix and the mass matrix. Then R and M are symmetric and positive definite. If $\xi = (\xi_1,\dots,\xi_I)^T$, we have $R\xi = \mu M\xi$. Let $M = LL^T$ the Cholesky factorization of M. Then:

$$L^{-1}R(L^T)^{-1}L^T\xi = \mu L^T\xi,$$

and if we define $\eta = L^T\xi$, we have $L^{-1}R(L^T)^{-1}\eta = \mu\eta$. As $L^{-1}R(L^T)^{-1}$ is symmetric definite positive, there exist eigenvalues $\lambda_i = \lambda_{i,h}$ such that $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_I$, and orthonormal eigenvectors η_i, $1 \le i \le I$ in $\mathbb{R}^I$.

$\square$

Proof.

(cont.) Let $\xi_i = (L^T)^{-1}\eta_i$, and $\xi_i = (\xi_{i1}, \dots, \xi_{iI})^T$. Then

$$R\xi_i = \lambda_i M\xi_i, \quad 1 \leq i \leq I,$$

$$\xi_j^T M\xi_i = \eta_j^T L^{-1} M (L^T)^{-1}\eta_i = \eta_j^T \eta_i = \delta_{ij}, \quad 1 \leq i,j \leq I.$$

Defining $w_{i,h} = \sum_{j=1}^{I} \xi_{ij}\varphi_j$, one gets for $1 \leq i,j \leq I$

$$a(w_{i,h}, \varphi_j) = \lambda_i (w_{i,h}, \varphi_j),$$

$$(w_{i,h}, w_{j,h}) = \delta_{ij}.$$

$\square$

Comment

If λ_i are the eigenvalues corresponding to the abstract parabolic problem, it is possible to prove that $\lambda_i \leq \lambda_{i,h}$.

Consider again the approximate problem: given $u_{0,h} \in V_h$, find a function $u_h : t \in [0, T] \mapsto u_h(t) \in V_h$ solution of the system of ode's:

$$\begin{array}{rcl} \forall\, v_h \in V_h, \quad \dfrac{d}{dt}(u_h(t), v_h) + a(u_h(t), v_h) & = & (f(t), v_h), \\ u_h(0) & = & u_{0,h} \end{array}$$

Theorem

The approximate problem has a unique solution u_h given by

$$u_h(t) = \sum_{i=1}^{I} \left\{ (u_{0,h}, w_{i,h}) e^{-\lambda_{i,h}} + \int_0^t (f(s), w_{i,h}) e^{-\lambda_{i,h}(t-s)}\, ds \right\} w_{i,h}.$$

Comment

In order to find the solution, we introduce as before a basis $(\varphi_i)_{1\leq i\leq I}$ *of* V_h *and we write*

$$u_h(t) = \sum_{j=1}^{I} \xi_j(t)\varphi_j, \quad u_{0,h} = \sum_{j=1}^{I} \xi_{0,j}\varphi_j.$$

Then, we have to solve:

$$\sum_{j=1}^{I}(\varphi_j,\varphi_i)\frac{d\xi_j}{dt}(t) + \sum_{j=1}^{I} a(\varphi_j,\varphi_i)\xi_j(t) = (f(t),\varphi_i), \quad 1 \leq i \leq I,$$

$$\xi_i(0) = \xi_{0,i}, \qquad 1 \leq i \leq I.$$

Comment

Using the rigidity matrix $R = (a(\varphi_j, \varphi_i))_{1 \le i,j \le I}$ *and the mass matrix* $M = ((\varphi_j, \varphi_i))_{1 \le i,j \le I}$, *we write*

$$M \frac{d\xi}{dt}(t) + R\xi(t) = \beta(t),$$

$$\xi(0) = \xi_0,$$

where $\xi(t) = (\xi_1(t), \ldots, \xi_I(t))^T$, $\beta(t) = (\beta_1(t), \ldots, \beta_I(t))^T$, $\beta_i(t) = (f(t), \varphi)$. *Our goal will be to solve numerically this system of ordinary differential equations.*

Comment

Another way to solve numerically this problem is to obtain the eigenvalues and eigenvectors as we have seen before, and compute the previous formula.

Comment

Under the previous conditions with $\lambda = 0$ and a symmetric if

1. *The solution of the abstract parabolic problem $u \in C^1(0, T; V)$*
2. $\lim_{h\to 0} |u_{h,0} - u_0| = 0$,
3. $\forall\, v \in V$, $\lim_{h\to 0} \inf_{v_h \in V_h} \|v - v_h\| = 0$,

then

$$\forall\, t \in [0, T], \qquad \lim_{h\to 0} |u_h(t) - u(t)| = 0.$$

Total discretization of parabolic problems

First we recall some facts about the approximate solution of the Cauchy problem:

$$y'(t) = \varphi(t, y(t)), \quad 0 \le t \le T,$$
$$y(0) = y_0,$$

where $\varphi : [0, T] \times \mathbb{R} \to \mathbb{R}$ is a continuous map. We define $\Delta t = T/N$, and $t_n = n\Delta t$, $0 \le n \le N$. We compute $\forall\ n = 1, \ldots, N$ an approximation y_n of $y(t_n)$ by using the θ-method:

$$y_{n+1} = y_n + \Delta t[\theta\varphi(t_{n+1}, y_{n+1}) + (1-\theta)\varphi(t_n, y_n)], \quad 0 \le n \le N-1.$$

When $\theta = 0$ it is the explicit Euler method and when $\theta = 1$ is the implicit Euler method. Moreover, the method is of order 1 if $\theta \neq \frac{1}{2}$ and of order 2 if $\theta = \frac{1}{2}$ (Crank-Nicolson method).

To see the stability, we consider the test equation $y' = -\lambda y$, $y(0) = y_0$, $t \ge 0$, where $\lambda > 0$. Applying the method we get

$$y_{n+1} = \frac{1 - (1-\theta)\lambda\Delta t}{1 + \theta\lambda\Delta t} y_n, \quad 0 \le n \le N-1.$$

If we define $r(x) = \frac{1-(1-\theta)x}{1+\theta x}$, then $y_n = [r(\lambda \Delta t)]^n y_0$. Then, the sequence $(y_n)_{n\geq 0}$ is bounded iff $|r(\lambda \Delta t)| \leq 1$. Then

1. If $\theta \geq 1/2$ then $(y_n)_{n\geq 0}$ is bounded $\forall\, \Delta t > 0$ (absolutely stable),
2. If $0 \leq \theta < \frac{1}{2}$ it is bounded if $\lambda \Delta t \leq \frac{2}{1-2\theta}$.

Suppose that $f \in C^0(0, T; H)$. Then $\beta_i : t \mapsto \beta_i(t) = (f(t), \varphi_i)$ is continuous on $[0, T]$. We want to approximate the solution of

$$M\frac{d\xi}{dt}(t) + R\xi(t) = \beta(t), \quad \xi(0) = \xi_0.$$

If ξ^n is the approximate value of $\xi(t_n)$, one has for $0 \leq n \leq N-1$,

$$\frac{1}{\Delta t}M(\xi^{n+1} - \xi^n) + R(\theta\xi^{n+1} + (1-\theta)\xi^n) = \theta\beta(t_{n+1}) + (1-\theta)\beta(t_n),$$

with initial condition $\xi^0 = \xi_0$.

If we define $u_h^n = \sum_{j=1}^{I} \xi_j^n \varphi_j \in V_h$, it is the solution of

$$\forall\, v_h \in V_h, \quad \frac{1}{\Delta t}(u_h^{n+1} - u_h^n, v_h) + a(\theta u_h^{n+1} + (1-\theta) u_h^n, v_h) =$$

$$= (\theta f(t_{n+1}) + (1-\theta) f(t_n), v_h); \quad 0 \le n \le N-1,$$

with initial condition $u_h^0 = u_{0,h}$.

Comment

For each time step we have to solve the following linear system:

$$(M + \theta \Delta t\, R)\xi^{n+1} = \eta^n,$$

where $\eta^n \in \mathbb{R}^I$ is a known vector. As $\theta \geq 0$, $M + \theta\Delta t R$ is symmetric an positive definite. As the matrix does not depend on n, we can perform a unique Choleski factorization, and solve two triangular systems for $n = 0, 1, \ldots, N-1$.
On the other hand, when $\theta = 0$, we can define explicitely ξ^{n+1} from ξ^n only if M is diagonal.

Discretization error

Definition

Suppose that $a(\cdot,\cdot)$ is V-elliptic, that is $\exists\,\alpha > 0$ s.t. $\forall\, v \in V$, $a(v,v) \geq \alpha \|v\|^2$. We define the discretization error as

$$e_h^n = u_h^n - \Pi_h u(t_n) \in V_h, \quad 0 \leq n \leq N,$$

where $\Pi_h \in L(V;V_h)$ is the operator of elliptic projection defined by

$$\forall\, v_h \in V_h, \quad a(\Pi_h u - u, v_h) = 0.$$

Comment

We notice that if $a(\cdot,\cdot)$ is symmetric then $\Pi_h u \in V_h$ is the best approximation to $u \in V$ from V_h with respect to the inner product $a(\cdot,\cdot)$.

Proposition

The error $\{e_h^n \in V_h,\ 0 \leq n \leq N\}$ *is the solution, for* $0 \leq n \leq N-1$, *of*

$$\forall\, v_h \in V_h, \quad \frac{1}{\Delta t}(e_h^{n+1} - e_h^n, v_h) + a(\theta e_h^{n+1} + (1-\theta)e_h^n, v_h) = (\varepsilon_h^n, v_h),$$

where $\varepsilon_h^n \in V_h$ *is defined for* $0 \leq n \leq N-1$ *by*

$$\forall\, v_h \in V_h, \quad (\varepsilon_h^n, v_h) = (\theta f(t_{n+1}) + (1-\theta) f(t_n), v_h) -$$

$$-\frac{1}{\Delta t}(\Pi_h u(t_{n+1}) - \Pi_h u(t_n), v_h) - a(\theta u(t_{n+1}) + (1-\theta) u(t_n), v_h).$$

The proof is immediate, taking into account the definitions of u_h^n and $\Pi_h u$.
Now, we will prove the fundamental result of stability:

Theorem

Suppose that $a(\cdot,\cdot)$ is V-elliptic and symmetric, and the canonical injection from V to H is compact. Then the solution $\{e_h^n \in V_h;\ 0 \le n \le N\}$ of the previous scheme satisfy:

1. *If $\frac{1}{2} < \theta \le 1$, there exist for all $\Delta t_0 > 0$ two constants μ and $C > 0$ depending on λ_1, θ and Δt_0 such that for $\Delta t \le \Delta t_0$*

$$|e_h^n| \le |e_h^0| e^{-\mu t_n} + C\Delta t \sum_{k=0}^{n-1} e^{-\mu(t_n - t_k)} |\varepsilon_h^k|;$$

2. *If $\theta = \frac{1}{2}$, we have*

$$|e_h^n| \le |e_h^0| + \Delta t \sum_{k=0}^{n-1} |\varepsilon_h^k|;$$

3. *If $0 \le \theta < \frac{1}{2}$ the previous inequality holds if*

$$(\Delta t)\lambda_{I,h} \le \frac{2}{1-2\theta}.$$

Comment

$\lambda_{I,h}$ is the greatest eigenvalue of the spectral problem: find λ for which there exists $u_h \in V_h$, $u_h \neq 0$, s.t.

$$\forall v_h \in V_h, \quad a(u_h, v_h) = \lambda(u_h, v_h).$$

Proof.

Let $(w_{i,h})$ an orthonormal basis of eigenvectors as before. Then

$$e_h^n = \sum_{i=1}^{I} e_i^n w_{i,h}, \qquad \varepsilon_h^n = \sum_{i=1}^{I} \varepsilon_i^n w_{i,h}$$

$$|e_h^n| = \left(\sum_{i=1}^{I} (e_i^n)^2\right)^{1/2}, \qquad |\varepsilon_h^n| = \left(\sum_{i=1}^{I} (\varepsilon_i^n)^2\right)^{1/2}.$$

Then

$$\frac{1}{\Delta t}(e_h^{n+1} - e_h^n, v_h) + a(\theta e_h^{n+1} + (1-\theta)e_h^n, v_h) = (\varepsilon_h^n, v_h),$$

is equivalent to

$$\frac{1}{\Delta t}(e_i^{n+1} - e_i^n) + \lambda_{i,h}(\theta e_i^{n+1} + (1-\theta)e_i^n) = \varepsilon_i^n, \qquad 1 \le i \le I,$$

that is,

Proof.

$$e_i^{n+1} = r(\Delta t\lambda_{i,h})e_i^n + \frac{\Delta t}{1+\theta\Delta t\lambda_{i,h}}\varepsilon_i^n, \quad 1 \le i \le I.$$

By induction, we deduce that:

$$e_i^n = [r(\Delta t\lambda_{i,h})]^n e_i^0 + \frac{\Delta t}{1+\theta\Delta t\lambda_{i,h}}\sum_{k=0}^{n-1}[r(\Delta t\lambda_{i,h})]^{n-k-1}\varepsilon_i^k, \quad 1 \le n \le N.$$

Suppose that $|r(\Delta t\lambda_{i,h})| \le 1$. *This is true if* $\Delta t\lambda_{i,h} \le \frac{2}{1-2\theta}$, *if* $0 \le \theta < \frac{1}{2}$, *or for all* Δt *if* $\frac{1}{2} \le \theta \le 1$. *Then*

$$|e_i^n| \le |e_i^0| + \Delta t\sum_{k=0}^{n-1}|\varepsilon_i^k|.$$

As $\lambda_{i,h} \le \lambda_{I,h}$, $1 \le i \le I$, *the inequality is true under the conditions of the theorem. Moreover, by the Minkowski inequality (triangular inequality):*

Proof.

$$|e_h^n| = \left(\sum_{i=1}^{I} (e_i^n)^2\right)^{1/2} \leq \left(\sum_{i=1}^{I} \left(|e_i^0| + \Delta t \sum_{k=0}^{n-1} |\varepsilon_i^k|\right)^2\right)^{1/2} \leq$$

$$\leq \left(\sum_{i=1}^{I} (e_i^0)^2\right)^{\frac{1}{2}} + \Delta t \sum_{k=0}^{n-1} \left(\sum_{i=1}^{I} (\varepsilon_i^k)^2\right)^{1/2} = |e_h^0| + \Delta t \sum_{k=0}^{n-1} |\varepsilon_h^k|.$$

Then, it remains to prove item 1, when $\frac{1}{2} < \theta \leq 1$: *In this case* $r(x) = \frac{1-(1-\theta)x}{1+\theta x}$ *satisfies* $|r(x)| \leq s(x)$, $\forall\, x \geq 0$, *with*

$$s(x) = \begin{cases} r(x) & 0 \leq x \leq x_\theta, \\ \frac{1-\theta}{\theta}, & x \geq x_\theta, \end{cases}$$

where $x_\theta = \frac{2\theta-1}{2\theta(1-\theta)} > 0$ *for* $\frac{1}{2} < \theta < 1$ *and* $x_\theta = +\infty$ *for* $\theta = 1$.

Proof.

As s is decreasing and $\lambda_{i,h} \geq \lambda_1$ we have

$$|r(\Delta t \lambda_{i,h})| \leq s(\Delta t \lambda_1), \quad 1 \leq i \leq I.$$

Fix $\Delta t_0 > 0$ and define μ s.t. $e^{-\mu \Delta t_0} = s(\Delta t_0 \lambda_1)$. Then $\mu > 0$ and it is easy to see (check!)

$$\Delta t \leq \Delta t_0 \Longrightarrow s(\Delta t \lambda_{1,h}) \leq e^{-\mu \Delta t}.$$

Then $|e_i^n| \leq |e_i^0| e^{-\mu n \Delta t} + \frac{\Delta t}{1+\theta \Delta t \lambda_1} \sum_{k=0}^{n-1} e^{-\mu(n-k-1)\Delta t} |\varepsilon_i^k| \leq$

$$\leq |e_i^0| e^{-\mu t_n} + C \Delta t \sum_{k=0}^{n-1} e^{-\mu(t_n - t_k)} |\varepsilon_i^k|,$$

with

$$C = \sup_{\Delta t \leq \Delta t_0} \frac{e^{\mu \Delta t}}{1 + \theta \Delta t \lambda_1}.$$

$\square$

Comment

When $0 \leq \theta < \frac{1}{2}$ *and the condition of stability is not satisfied, then*

$$\lim_{n\to\infty} |r^n(\Delta t \lambda_{i,h})| = +\infty, \text{ if } \Delta t \lambda_{i,h} \geq \frac{2}{1-2\theta}.$$

Therefore, certain components e_i^n *of the discrete error are in general amplified very fast, which give completely wrong numerical results. We say that in this case the scheme*

$$\forall\, v_h \in V_h, \quad \frac{1}{\Delta t}(e_h^{n+1} - e_h^n, v_h) + a(\theta e_h^{n+1} + (1-\theta)e_h^n, v_h) = (\varepsilon_h^n, v_h),$$

is unstable. However, if $\frac{1}{2} < \theta \leq 1$*, the property of stability tell us that the contribution to the total error at time* t_n *of an error made at time* $t_k < t_n$ *decreases exponentially with* $t_n - t_k$*. We say the the method is strongly stable. One can also see that if we fix* λ_1 *and* Δt_0*, the constant* μ *is maximum when* $\theta = 1$ *(implicit Euler method).*

Now we want to find the error of consistence ε_h^n, $0 \leq n \leq N-1$.

Lemma

There exists a constant $C > 0$ depending only on θ such that

1. *If $\theta \neq \frac{1}{2}$ and $u \in C^1(0,T;V) \cap C^2(0,T;H)$, one has*

$$|\varepsilon_h^n| \leq \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \left| (I - \Pi_h) \frac{du}{dt}(s) \right| \, ds + C \int_{t_n}^{t_{n+1}} \left| \frac{d^2 u}{dt^2}(s) \right| \, ds,$$

2. *If $\theta = \frac{1}{2}$ and if $u \in C^1(0,T;V) \cap C^3(0,T;H)$, one has*

$$|\varepsilon_h^n| \leq \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \left| (I - \Pi_h) \frac{du}{dt}(s) \right| \, ds + C \Delta t \int_{t_n}^{t_{n+1}} \left| \frac{d^3 u}{dt^3}(s) \right| \, ds.$$

Proof.

As $\forall\, v \in V$,

$$\frac{d}{dt}(u(t), v) + a(u(t), v) = (f(t), v),$$

and

$$\forall\, v_h \in V_h, \quad (\varepsilon_h^n, v_h) = (\theta f(t_{n+1}) + (1-\theta) f(t_n), v_h) -$$

$$-\frac{1}{\Delta t}(\Pi_h u(t_{n+1}) - \Pi_h u(t_n), v_h) - a(\theta u(t_{n+1}) + (1-\theta) u(t_n), v_h),$$

then

$$(\varepsilon_h^n, v_h) = \left(\theta \frac{du}{dt}(t_{n+1}) + (1-\theta)\frac{du}{dt}(t_n), v_h\right) - \frac{1}{\Delta t}(\Pi_h(u(t_{n+1}) - u(t_n)), v_h),$$

and

$$(\varepsilon_h^n, v_h) = \left(\theta \frac{du}{dt}(t_{n+1}) + (1-\theta)\frac{du}{dt}(t_n) - \frac{1}{\Delta t}(u(t_{n+1}) - u(t_n)), v_h\right) +$$

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \left((I - \Pi_h)\frac{du}{dt}(s), v_h\right) ds.$$

Proof.

Therefore, taking $v_h = \varepsilon_h^n / |\varepsilon_h^n|$, *we get*

$$|\varepsilon_h^n| \leq |\eta(t_n)| + \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \left| (I - \Pi_h) \frac{du}{dt}(s) \right| ds,$$

with

$$\eta(t_n) = \theta \frac{du}{dt}(t_{n+1}) + (1-\theta) \frac{du}{dt}(t_n) - \frac{1}{\Delta t}(u(t_{n+1}) - u(t_n)).$$

If $u \in C^2(0, T; H)$, *we get from the Taylor formula at* t_n *with the integral remainder*

$$\eta(t_n) = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} (s - (1-\theta) t_{n+1} - \theta t_n) \frac{d^2 u}{dt^2}(s) \, ds$$

and

$$|\eta(t_n)| \leq \max(\theta, 1-\theta) \int_{t_n}^{t_{n+1}} \left| \frac{d^2 u}{dt^2}(s) \right| ds.$$

Proof.

Then, we have the first formula, where $C = \max(\theta, 1-\theta)$.
For $\theta = \frac{1}{2}$, and $u \in C^3(0, T; H)$, one has

$$\eta(t_n) = \frac{1}{2\Delta t} \int_{t_n}^{t_{n+1}} (t_{n+1} - s)(t_n - s) \frac{d^3 u}{dt^3}(s)\, ds,$$

and

$$|\eta(t_n)| \leq \frac{\Delta t}{8} \int_{t_n}^{t_{n+1}} \left| \frac{d^3 u}{dt^3}(s) \right| ds.$$

Then, in the second case, we take $C = 1/8$. □

Comment

We recall that if u is of class $m+1$ then

$$u(t_{n+1}) = u(t_n) + \sum_{i=1}^{m} \frac{1}{i!} u^{(i)}(t_n)(\Delta t)^i + \frac{1}{m!} \int_{t_n}^{t_{n+1}} u^{(m+1)}(s)(t_{n+1} - s)^m\, ds.$$

Taking the results of the previous theorem and lemma we obtain:

Theorem

The solution $\{u_h^n \in V_h;\ 0 \leq n \leq N\}$ *satisfies:*

- *If* $\frac{1}{2} < \theta \leq 1$ *and if* $u \in C^1(0,T;V) \cap C^2(0,T;H)$, $\exists$ *for all* $\Delta t_0 > 0$ *two constants* μ *and* $C > 0$ *depending only on* λ_1, θ *and* Δt_0 *s.t. for* $\Delta t \leq \Delta t_0$

$$|u_h^n - u(t_n)| \leq |u_{0,h} - \Pi_h u_0| e^{-\mu t_n} + |(I - \Pi_h) u(t_n)| +$$

$$+ C \int_0^{t_n} \left\{ \left| (I - \Pi_h) \frac{du}{dt}(s) \right| + \Delta t \left| \frac{d^2 u}{dt^2}(s) \right| \right\} e^{-\mu(t_n - s)}\, ds;$$

- *If* $\theta = \frac{1}{2}$ *and if* $u \in C^1(0,T;V) \cap C^3(0,T;H)$, *we have*

$$|u_h^n - u(t_n)| \leq |u_{0,h} - \Pi_h u_0| + |(I - \Pi_h) u(t_n)| +$$

$$+ \int_0^{t_n} \left\{ \left| (I - \Pi_h) \frac{du}{dt}(s) \right| + C \Delta t^2 \left| \frac{d^3 u}{dt^3}(s) \right| \right\} ds,$$

where C *us independent of* h, Δt *and* u;

Theorem

(cont.)

- *If* $0 \leq \theta < \frac{1}{2}$ *and if* $u \in C^1(0, T; V) \cap C^2(0, T; H)$, *under the condition of stability*

$$|u_h^n - u(t_n)| \leq |u_{0,h} - \Pi_h u_0| + |(I - \Pi_h)u(t_n)| +$$

$$+ \int_0^{t_n} \left\{ \left| (I - \Pi_h) \frac{du}{dt}(s) \right| + C\Delta t \left| \frac{d^2 u}{dt^2}(s) \right| \right\} ds,$$

where the constant C *is indepedent of* h, Δt *and* u.

Proof.

We will only prove the case $\frac{1}{2} < \theta \leq 1$. Note that

$$\Delta t \sum_{k=0}^{n-1} e^{-\mu(t_n - t_k)} |\varepsilon_h^k| \leq \sum_{k=0}^{n-1} e^{-\mu(t_n - t_k)} \int_{t_k}^{t_{k+1}} \left| (I - \Pi_h) \frac{du}{dt}(s) \right| ds +$$

$$+ C \Delta t \sum_{k=0}^{n-1} e^{-\mu(t_n - t_k)} \int_{t_k}^{t_{k+1}} \left| \frac{d^2 u}{dt^2}(s) \right| ds,$$

and, therefore,

$$\Delta t \sum_{k=0}^{n-1} e^{-\mu(t_n - t_k)} |\varepsilon_h^k| \leq \int_0^{t_n} \left| (I - \Pi_h) \frac{du}{dt}(s) \right| e^{-\mu(t_n - s)} \, ds +$$

$$+ C \Delta t \int_0^{t_n} \left| \frac{d^2 u}{dt^2}(s) \right| e^{-\mu(t_n - s)} ds.$$

From this inequality and the previous lemma, we obtain the desired result. $\square$

Comment

In the case $0 \le \theta < \frac{1}{2}$*, if the stability condition does not hold then* $\max_{0\le n\le N} |u_h^n - u(t_n)|$ *goes in general to* $+\infty$ *when* h *and* Δt *goes to zero.*
As a conclusion, the Crank-Nicolson method which is absolutely convergent and of order 2 is the most used in practice. However, when T *is large or* u *is not regular, it is better to use the implicit Euler method. A compromise between the request of precision and stability is obtained for an intermediate value of* θ*, for example* $\theta = 2/3$.

The wave equation

Let $\Omega \subset \mathbb{R}^n$ an open and bounded set with piecewise C^1 boundary Γ. Moreover, for all $T > 0$ we define $Q_T = \Omega \times (0, T)$, $\quad \Sigma_T = \Gamma \times (0, T)$, and consider the following problem: Given $u_0, u_1 : \Omega \to \mathbb{R}$ and $f : Q_T \to \mathbb{R}$, find a map $u : (x, t) \in Q_T \mapsto u(x, t) \in \mathbb{R}$ such that

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f \text{ in } Q_T \text{ (wave equation)},$$

$$u = 0 \text{ on } \Sigma_T \text{ (boundary condition) },$$

$$u(\cdot, 0) = u_0, \quad \frac{\partial u}{\partial t}(\cdot, 0) = u_1 \text{ in } \Gamma \text{ (initial conditions)}.$$

If we multiply the equation by a test function $v \in H_0^1(\Omega)$ and integrate over Ω :

$$\int_\Omega \frac{\partial^2 u}{\partial t^2}(x, t)v(x)\, dx - \int_\Omega \Delta u(x.t)v(x)\, dx = \int_\Omega f(t, x)v(x)\, dx,$$

and by the Green formula

$$\frac{d^2}{dt^2}\int_\Omega u(x, t)v(x)\, dx + \sum_{i=1}^n \int_\Omega \frac{\partial u}{\partial x_i}(x, t)\frac{\partial v}{\partial x_i}(x)\, dx = \int_\Omega f(x, t)v(x)\, dx.$$

We introduce $u(t) : x \in \Omega \mapsto u(x,t) \in \mathbb{R}$ and

$$(\varphi, \psi) = \int_\Omega \varphi(x)\psi(x)\, dx,$$

$$a(\varphi, \psi) = \sum_{i=1}^{n} \int_\Omega \frac{\partial \varphi}{\partial x_i}(x) \frac{\partial \psi}{\partial x_i}(x)\, dx.$$

Then we have to find a map $u : t \in [0, T] \mapsto u(t) \in H_0^1(\Omega)$ such that

$$\forall\, v \in H_0^1(\Omega), \quad \frac{d^2}{dt^2}(u(t), v) + a(u(t), v) = (f(t), v),$$

$$u(0) = u_0, \quad \frac{du}{dt}(0) = u_1.$$

We ask the following regularity of u :

$$u \in C^0(0, T; H_0^1(\Omega)) \cap C^1(0, T; L^2(\Omega)).$$

Abstract hyperbolic problems

One introduces:

- Two Hilbert spaces V and H s.t. $V \subset H$ with continuous injection, and V is dense in H.
- A continuous bilinear form $u, v \mapsto a(u, v)$ on $V \times V$.

Moreover, $(\cdot, \cdot)$ us the scalar product in H, $|\cdot|$ the corresponding norm and $\|\cdot\|$ the norm in V.

A general hyperbolic problem is; given $u_0 \in V$, $u_1 \in H$ and $f \in L^2(0, T; H)$, find a map u s.t.

- $u \in C^0(0, T; V) \cap C^1(0, T; H)$,
- $\forall\, v \in V$, $\frac{d^2}{dt^2}(u(t), v) + a(u(t), v) = (f(t), v)$,
- $u(0) = u_0$, $\frac{du}{dt}(0) = u_1$.
- The bilinear form $a(\cdot, \cdot)$ is symmetric;
- The bilinear form $a(\cdot, \cdot)$ is V-elliptic: there exists a constant $\alpha > 0$ s.t. $\forall\, v \in V$, $a(v, v) \geq \alpha \|v\|^2$.
- The canonical injection from V into H is compact.

Theorem

Under the previous hypotheses, the abstract hyperbolic equation has a unique solution given by

$$u(t) = \sum_{i\geq 1} \left\{ (u_0, w_i)\cos(\omega_i t) + \frac{1}{\omega_i}(u_1, w_i)\sin(\omega_i t) + \right.$$

$$\left. + \frac{1}{\omega_i}\int_0^t \sin(\omega_i(t-s))(f(s), w_i)\, ds \right\} w_i,$$

with $(w_i)_i$ an orthonormal basis in H of eigenvectors satisfying

$$\forall\, v \in V, \quad a(w_i, v) = \lambda_i (w_i, v),$$

where $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_i \leq \cdots$ are the eigenvalues, and $\omega_i = \sqrt{\lambda_i}$.

Proof.

We only see how to get the formula. Let u be a solution. Then

$$u(t) = \sum_{i \geq 1} (u(t), w_i) w_i,$$

$$a(u(t), w_i) = \lambda_i (u(t), w_i).$$

From the differential equation applied to $v = w_i$ and defining $\alpha_i(t) = (u(t), w_i)$, we have that $\alpha_i(t)$ is the solution of the linear ode

$$\left\{ \begin{array}{ccc} \frac{d^2}{dt^2}\alpha_i(t) + \lambda_i \alpha_i(t) & = & (f(t), w_i), \\ \alpha_i(0) = (u_0, w_i), & & \frac{d\alpha_i}{dt}(0) = (u_1, w_i). \end{array} \right.$$

with solution

$$\alpha_i(t) = (u_0, w_i) \cos(\omega_i t) + \frac{1}{\omega_i} (u_1, w_i) \sin(\omega_i t) + \frac{1}{\omega_i} \int_0^t \sin(\omega_i (t-s))(f(s), w_i) ds.$$

Then, we obtain the formula for u. $\square$

Semi-discretization method

We introduce again the subspace $V_h \subset V$ of finite dimension $I = I(h)$, and consider the following semi-discrete problem: Given $u_{0,h}$, $u_{1,h} \in V_h$, find a function $u_h : t \in [0, T] \mapsto u_h(t) \in V_h$, solution of the system of ode:

$$\begin{cases} \forall\, v_h \in V_h \quad \frac{d^2}{dt^2}(u_h(t), v_h) + a(u_h(t), v_h) & = & (f(t), v_h), \\ u_h(0) = u_{0,h}, & & \frac{du_h}{dt}(0) = u_{1,h} \end{cases}$$

We recall that there exists a sequence of eigenvalues $0 \leq \lambda_{1,h} \leq \lambda_{2,h} \leq \cdots \leq \lambda_{I,h}$ and an orthonormal hilbertian basis $(w_{i,h})$ of V_h such that

$$\forall\, v_h \in V_h, \quad a(w_{i,h}, v_h) = \lambda_{i,h}(w_{i,h}, v_h).$$

We put $\omega_{i,h} = \sqrt{\lambda_{i,h}}$.

Theorem

The previous problem have a unique solution u_h given by

$$u_h(t) = \sum_{i=1}^{I} \Big\{ (u_{0,h}, w_{i,h}) \cos(w_{i,h} t) + \frac{1}{\omega_{i,h}} (u_{1,h}, w_{i,h}) \sin(w_{i,h} t) +$$

$$+ \frac{1}{\omega_{i,h}} \int_0^t \sin(\omega_{i,h}(t-s)) (f(s), w_{i,h})\, ds \Big\} w_{i,h}.$$

Comment

A way to solve numerically this problem is to use the formula or truncate it.

If we introduce a basis $(\varphi_i)_{1\leq i\leq I}$ of V_h and seek a solution u_h as

$$u_h(t) = \sum_{j=1}^{I} \xi_j(t)\varphi_j,$$

then, writing

$$u_{0,h} = \sum_{j=1}^{I} \xi_{0,j}\varphi_j, \quad u_{1,h} = \sum_{j=1}^{I} \xi_{1,j}\varphi_j$$

and

$$\xi(t) = (\xi_1(t), \ldots, \xi_I(t))^T,$$
$$\chi(t) = (\chi_1(t), \ldots, \chi_I(t))^T, \quad \chi_i(t) = (f(t), \varphi),\ 1 \leq i \leq I,$$

the problem has the form

$$M\frac{d^2\xi}{dt^2}(t) + R\xi(t) = \chi(t),$$

$$\xi(0) = \xi_0, \quad \frac{d\xi}{dt}(0) = \xi_1,$$

where $M = ((\varphi_j, \varphi_i))_{1\leq i,j\leq I}$ and $R = (a(\varphi_j, \varphi_i))_{1\leq i,j\leq I}$, as before.

Total discretization of hyperbolic problems

First we consider the Cauchy problem

$$y''(t) = \varphi(t, y(t), y'(t)), \quad 0 \leq t \leq T,$$

$$y(0) = y_0, \quad y'(0) = z_0,$$

where φ is continuous in $[0, T] \times \mathbb{R} \times \mathbb{R}$ and $y_0, z_0 \in \mathbb{R}$ are given. We define

$$\Delta t = \frac{T}{N}$$

and

$$t_n = n\Delta t, \text{ for } 0 \leq n \leq N.$$

We want to obtain an approximation (y_n, z_n), $n = 1, \ldots, N$, of $(y(t_n), y'(t_n))$. We will use the Newmark method.

Newmark method

We want to obtain the method of Newmark. First, we can write

$$y(t_{n+1}) = y(t_n) + \Delta t\, y'(t_n) + (\Delta t)^2 \left(\beta y''(t_{n+1}) + \left(\frac{1}{2} - \beta\right) y''(t_n)\right) + O((\Delta t)^3)$$

Then:

$$y(t_{n+1}) = y(t_n) + \Delta t\, y'(t_n) + (\Delta t)^2 \left(\beta \varphi(t_{n+1}, y(t_{n+1}), y'(t_{n+1})) + \right.$$

$$\left. + \left(\frac{1}{2} - \beta\right) \varphi(t_n, y(t_n), y'(t_n))\right) + O((\Delta t)^3),$$

and

$$y'(t_{n+1}) = y'(t_n) + \Delta t(\gamma y''(t_{n+1}) + (1-\gamma) y''(t_n)) + O((\Delta t)^2),$$

$$y'(t_{n+1}) = y'(t_n) + \Delta t \left[\gamma \varphi(t_{n+1}, y(t_{n+1}), y'(t_{n+1})) + \right.$$

$$\left. + (1-\gamma)\varphi(t_n, y(t_n), y'(t_n))\right] + O((\Delta t)^2),$$

where β and γ are parameters.

The method of Newmark is:

$$\begin{cases} y_{n+1} &= y_n + \Delta t\, z_n + (\Delta t)^2 \left(\beta\varphi_{n+1} + \left(\frac{1}{2} - \beta\right)\varphi_n\right) \\ z_{n+1} &= z_n + \Delta t(\gamma\varphi_{n+1} + (1-\gamma)\varphi_n),\ n \le 0 \le N-1, \end{cases}$$

with $\varphi_n = \varphi(t_n, y_n, z_n)$. This is an implicit method, except when $\beta = \gamma = 0$.

Comment

If φ does not depend on y', then, we can remove z_n from the equations:

$$y_{n+2} - 2y_{n+1} + y_n = (y_{n+2} - y_{n+1}) - (y_{n+1} - y_n) =$$

$$= \Delta t(z_{n+1} - z_n) + (\Delta t)^2 \left(\beta\varphi_{n+2} + \left(\frac{1}{2} - 2\beta\right)\varphi_{n+1} - \left(\frac{1}{2} - \beta\right)\varphi_n\right).$$

Using the second equation, we obtain:

$$y_{n+2} - 2y_{n+1} + y_n = (\Delta t)^2 \left(\beta\varphi_{n+2} + \left(\frac{1}{2} - 2\beta + \gamma\right)\varphi_{n+1} + \left(\frac{1}{2} + \beta - \gamma\right)\varphi_n\right).$$

The method is implicit, except when $\beta = 0$.

Proposition

The Newmark method is of order 1 if $\gamma \neq 1/2$ and of order 2 if $\gamma = 1/2$.

Proof.

Let $y(t)$ be the solution of the Cauchy problem. Then

$$y(t_{n+1}) = y(t_n) + \Delta t y'(t_n) + (\Delta t)^2 \left(\beta y''(t_{n+1}) + \left(\frac{1}{2} - \beta \right) y''(t_n) \right) +$$

$$+(\Delta t)^3 \left(\frac{1}{6} - \beta \right) y'''(t_n) + O((\Delta t)^4),$$

$$y'(t_{n+1}) = y'(t_n) + \Delta t(\gamma y''(t_{n+1}) + (1-\gamma) y''(t_n)) +$$

$$(\Delta t)^2 \left(\frac{1}{2} - \gamma \right) y'''(t_n) + O((\Delta t)^3),$$

when $\Delta t \to 0$, where we have use the Taylor expansion of $y(t)$ and $y''(t)$ about $t = t_n$.

Proof.

If $y_n = y(t_n)$ and $z_n = y'(t_n)$, then

$$\left\{\begin{array}{rcl} y(t_{n+1}) - y_{n+1} & = & \left(\frac{1}{6} - \beta\right)(\Delta t)^2 y'''(t_n) + O((\Delta t)^3), \\ y'(t_{n+1}) - z_{n+1} & = & \left(\frac{1}{2} - \gamma\right)(\Delta t) y'''(t_n) + O((\Delta t)^2) \end{array}\right.$$

Then the method is of order 1 for $\gamma \neq \frac{1}{2}$ and of order 2 if $\gamma = \frac{1}{2}$. $\square$

Stability of the Newmark method

Consider the test equation

$$y'' + \omega^2 y = 0, \qquad \omega > 0.$$

Let $y(t)$ be a solution and

$$H(t) = (\omega y(t))^2 + y'(t)^2, \text{ (energy)}.$$

It is immediate that $H(t) = H(0)$. In particular,

$$H(t) \text{ is bounded when } t \to +\infty.$$

We want the same property for the Newmark method, that is

$$(\omega y_n)^2 + z_n^2 \text{ bounded when } n \to \infty.$$

If we apply the method

$$\left\{\begin{array}{rcl} y_{n+1} & = & y_n + \Delta t\ z_n + (\Delta t)^2\left(\beta\varphi_{n+1} + \left(\frac{1}{2}-\beta\right)\varphi_n\right) \\ z_{n+1} & = & z_n + \Delta t(\gamma\varphi_{n+1} + (1-\gamma)\varphi_n),\ n \leq 0 \leq N-1, \end{array}\right.$$

to the test function, we get:

$$\left\{\begin{array}{rcl} y_{n+1} & = & y_n + \Delta t z_n - \omega^2(\Delta t)^2\left(\beta y_{n+1} + \left(\frac{1}{2}-\beta\right)y_n\right), \\ z_{n+1} & = & z_n - \omega^2\Delta t(\gamma y_{n+1} + (1-\gamma)y_n). \end{array}\right.$$

If we define $\theta = \omega\Delta t$, then

$$\left\{\begin{array}{rcl} (1+\beta\theta^2)y_{n+1} & = & \left(1+\left(\beta-\frac{1}{2}\right)\theta^2\right)y_n + \Delta t\ z_n, \\ \gamma\theta\omega y_{n+1} + z_{n+1} & = & -(1-\gamma)\theta\omega y_n + z_n. \end{array}\right.$$

From now on, we will suppose that $\beta \geq 0$. If we define

$$\alpha(\theta) = \frac{\theta^2}{1+\beta\theta^2}$$

$$B(\theta) = \begin{pmatrix} 1-\frac{\alpha(\theta)}{2} & \frac{\alpha(\theta)}{\theta} \\ -\theta\left(1-\frac{\gamma\alpha(\theta)}{2}\right) & 1-\gamma\alpha(\theta) \end{pmatrix},$$

then $\begin{pmatrix} \omega y_{n+1} \\ z_{n+1} \end{pmatrix} = B(\theta)\begin{pmatrix} \omega y_n \\ z_n \end{pmatrix}$.

Definition

We say that the Newmark method is stable if $\|B(\theta)^n\|$ is bounded when $n \to \infty$, where $\|\cdot\|$ is the matrix norm associated to the euclidean norm ($\|B(\theta)^n\| = \max_{|\xi|\leq 1} \frac{|B(\theta)^n \xi|}{|\xi|}$).

Proposition

If the Newmark method is stable for $\theta = \omega \Delta t$ *then* $\rho(B(\theta)) \leq 1$.

Proof.

We know that

$$\rho(B(\theta))^n = \rho(B(\theta)^n) \leq \|B(\theta)^n\|,$$

and this implies that $\rho(B(\theta)) \leq 1$. $\square$

We will use the following result for the next proposition:

Lemma

Let b and c be real numbers. The solutions of $x^2 - bx + c = 0$ have modulus less or equal than 1 iff $|c| \leq 1$ and $|b| \leq 1 + c$.

Proposition

Suppose that $\beta \geq 0$. Then $\rho(B(\theta)) \leq 1$ iff

$$\delta = \gamma - \frac{1}{2} \geq 0,$$

$$\theta^2 \leq \begin{cases} \frac{4}{1+2\delta-4\beta} & \text{if} \quad \beta < \frac{1+2\delta}{4} \\ \\ +\infty & \text{if} \quad \beta \geq \frac{1+2\delta}{4} \end{cases}$$

Proof.

The characteristic polynomial of $B(\theta)$ is

$$\det(B(\theta) - \mu I) = \mu^2 - \mu(2 - \alpha(1+\delta)) + 1 - \alpha\delta.$$

Let $b = 2 - \alpha(1+\delta)$ and $c = 1 - \alpha\delta$. If $|c| \leq 1$ then $1 - \alpha\delta \leq 1$, which implies the first condition. Moreover, if $|b| \leq 1 + c$ then $-2 + \alpha\delta \leq 2 - \alpha(1+\delta)$ which implies that $\alpha \leq \frac{4}{1+2\delta}$ and taking into account that

$$\alpha = \frac{\theta^2}{1+\beta\theta^2}$$

then $(1 + 2\delta - 4\beta)\theta^2 \leq 4$ which implies the second condition.

We note that $b = 2 - \alpha(1+\delta) \leq 1 + c = 2 - \alpha\delta$. If the first condition is fulfilled then $c \leq 1$. The second condition implies that $(1 + 2\delta - 4\beta)\theta^2 \leq 4$ and $\alpha \leq \frac{4}{1+2\delta}$, which implies that $b \geq -1 - c$ and $c \geq -1$.

$\square$

Comment

The eigenvalues of $B(\theta)$ are not real or there is a double eigenvalue iff $\alpha \leq \frac{4}{(1+\delta)^2}$. In this case $\rho(B(\theta)) = \sqrt{1-\delta\alpha} \leq 1$ iff $1-\delta\alpha \leq 1$ and

$$\theta^2 \leq \begin{cases} \frac{4}{(1+\delta)^2 - 4\beta} & \textit{if} \quad \beta < \frac{(1+\delta)^2}{4} \\ \\ +\infty & \textit{if} \quad \beta \geq \frac{(1+\delta)^2}{4} \end{cases} \tag{1}$$

Now, we want to see in which measure the necessary conditions of stability are sufficient. We will use the following property without proof.

Lemma

Let A be a real $n \times n$ normal matrix (that is $A^T A = AA^T$). Then $\|A\| = \rho(A)$.

The two-dimensional normal matrices are easy to characterize:

Lemma

A real matrix $A = (a_{ij})_{1 \leq i,j \leq 2}$ is normal iff one of the following conditions is satisfied:

- $a_{11} = a_{22}$, $a_{12} = -a_{21}$,
- $a_{21} = a_{12}$

Proof.

We have

$$A^T A - AA^T = \begin{pmatrix} a_{21}^2 - a_{12}^2 & (a_{11} - a_{22})(a_{12} - a_{21}) \\ (a_{11} - a_{22})(a_{12} - a_{21}) & a_{12}^2 - a_{21}^2 \end{pmatrix},$$

which implies the result. □

Comment

In our case, it is easy to see that $B(\theta)$ *is normal iff* $\beta = 1/4$, *and* $\gamma = 1/2$. *In the general case, if we have a* 2×2 *matrix* $G(\theta)$ *s.t.* $G(\theta)B(\theta)G(\theta)^{-1}$ *is normal, then*

$$B(\theta)^n = G(\theta)^{-1}(G(\theta)B(\theta)G(\theta)^{-1})^n G(\theta),$$

and

$$\|B(\theta)^n\| \leq \|G(\theta)\| \, \|G(\theta)^{-1}\| \, \rho(B(\theta))^n.$$

Theorem

Suppose that $\beta \geq 0$ *and* $\delta = \gamma - \frac{1}{2} \geq 0$. *Then we have*

- *If* $\beta \geq \frac{(1+\delta)^2}{4}$, $\exists$ *a positive, continuous and increasing map* $\theta \mapsto b(\theta)$, *s.t* $b(\theta) \to +\infty$ *as* $\theta \to \infty$, *and*

$$\|B(\theta)^n\| \leq b(\theta);$$

- *if* $\beta < \frac{(1+\delta)^2}{4}$, *and under the stability condition*

$$\theta^2 \leq \frac{4}{(1+\delta)^2 - 4\beta}(1-\epsilon), \quad 0 < \epsilon < 1,$$

  $\exists$ *a constant* $C(\epsilon) > 0$, *s.t.* $C(\epsilon) \to +\infty$ *as* $\epsilon \to 0$, *and*

$$\|B(\theta)^n\| \leq C(\epsilon).$$

Proof.

We consider a lower triangular matrix:

$$G(\theta) = \begin{pmatrix} 1 & 0 \\ s(\theta) & t(\theta) \end{pmatrix},$$

such that $G(\theta)B(\theta)G(\theta)^{-1}$ *is normal. One has*

$$GBG^{-1} = \begin{pmatrix} 1 - \dfrac{\alpha}{2} - \dfrac{\alpha}{\theta}\dfrac{s}{t} & \dfrac{\alpha}{\theta t} \\ \left(\gamma - \dfrac{1}{2}\right)\alpha s - \left(1 - \dfrac{\gamma\alpha}{2}\right)\theta t - \dfrac{\alpha}{\theta}\dfrac{s^2}{t} & 1 - \gamma\alpha + \dfrac{\alpha}{\theta}\dfrac{s}{t} \end{pmatrix}.$$

In order to be normal, we use the first condition of the lemma characterizing the 2×2 *normal matrices. That is*

$$\begin{cases} 1 - \frac{\alpha}{2} - \frac{\alpha}{\theta}\frac{s}{t} &= 1 - \gamma\alpha + \frac{\alpha}{\theta}\frac{s}{t}, \\ \left(\gamma - \frac{1}{2}\right)\alpha s - \left(1 - \frac{\gamma\alpha}{2}\right)\theta t - \frac{\alpha}{\theta}\frac{s^2}{t} &= -\frac{\alpha}{\theta t} \end{cases}$$

Proof.

From the first equation

$$s = \left(\gamma - \frac{1}{2}\right) \frac{\theta t}{2} = \frac{\delta\theta t}{2}.$$

Using this value in the second equation, we obtain

$$\theta^2 t^2 \left(1 - \frac{\alpha}{4}(1+\delta)^2\right) = \alpha.$$

As $\alpha \geq 0$, this is only possible if $\alpha \leq \frac{4}{(1+\delta)^2}$, that is, (1). Under these conditions:

$$t^2 = \frac{\alpha}{\theta^2 \left(1 - \frac{\alpha}{4}(1+\delta)^2\right)}, \quad s = \frac{\delta\theta t}{2},$$

and, taking into account that $\alpha = \theta^2/(1+\beta\theta^2)$:

$$t^2 = \frac{1}{1 + \theta^2 \left(\beta - \frac{(1+\delta)^2}{4}\right)}, \quad s = \frac{\delta\theta t}{2}.$$

Proof.

With these values of s and t and assuming (1), the matrix GBG^{-1} *is normal. As* $\delta \geq 0$ *then also* $\rho(B(\theta)) \leq 1$, *and*

$$\|B(\theta)\|^2 \leq \|G(\theta\| \, \|G(\theta)^{-1}\|.$$

Now we have to bound $\|G\|$ *and* $\|G^{-1}\|$:

$$\|G\|^2 \leq 1 + s^2 + t^2, \qquad \|G^{-1}\|^2 \leq \frac{1 + s^2 + t^2}{t^2},$$

since

$$G^{-1} = \begin{pmatrix} 1 & 0 \\ -\frac{s}{t} & \frac{1}{t} \end{pmatrix},$$

and the euclidean norm of a matrix is less or equal than the Frobenius norm.

Proof.

Suppose that $\beta \geq (1+\delta)^2/4$: *Then* $t^2 \leq 1$, *and*

$$1 + s^2 + t^2 \leq 1 + \left(\frac{\delta^2\theta^2}{4} + 1\right) t^2 \leq 2 + \frac{\delta^2\theta^2}{4},$$

$$\frac{1 + s^2 + t^2}{t^2} \leq \left(1 + \theta^2 \left(\beta - \frac{(1+\delta)^2}{4}\right)\right) \left(2 + \frac{\delta^2\theta^2}{4}\right).$$

Therefore, $\|G(\theta)\| \, \|G(\theta)^{-1}\| \leq b(\theta)$, *with*

$$b(\theta) = \left(1 + \theta^2 \left(\beta - \frac{(1+\delta)^2}{4}\right)\right)^{1/2} \left(2 + \frac{\delta^2\theta^2}{4}\right),$$

which prove the first item of the theorem.

Proof.

Suppose now that $\beta < (1+\delta)^2/4$. We have

$$\|G\|^2 \leq 1 + \left(\frac{\delta^2\theta^2}{4} + 1\right) \frac{1}{1 + \theta^2\left(\beta - \frac{(1+\delta)^2}{4}\right)},$$

and, under the conditon on θ in the second item:

$$\|G\|^2 \leq 1 + \frac{1}{\epsilon}\left(1 + \frac{\delta^2\theta^2}{4}\right).$$

As $\beta < (1+\delta)^2/4$: $\|G^{-1}\|^2 \leq 2 + \frac{\delta^2\theta^2}{4}$, and, $\|G(\theta)\|\,\|G(\theta)^{-1}\| \leq C(\epsilon)$,

$$C(\epsilon) = \left\{\left(1 + \frac{1}{\epsilon}\left(1 + \frac{\delta^2\theta_0^2(\epsilon)}{4}\right)\right)\left(2 + \frac{\delta^2\theta_0^2(\epsilon)}{4}\right)\right\}^{1/2},$$

$$\theta_0^2(\epsilon) = \frac{4(1-\epsilon)}{(1+\delta)^2 - 4\beta}.$$

□

Numerical solution of the abstract hyperbolic problem using the Newmark method.

Recall that we have a subspace $V_h \subset V$ of finite dimension $I = I(h)$, and consider the following problem: Given $u_{0,h}$, $u_{1,h} \in V_h$, find a function $u_h : t \in [0, T] \mapsto u_h(t) \in V_h$, solution of the system of ode:

$$\begin{cases} \forall\, v_h \in V_h \quad \frac{d^2}{dt^2}(u_h(t), v_h) + a(u_h(t), v_h) & = \quad (f(t), v_h), \\ \qquad\qquad\qquad\qquad u_h(0) = u_{0,h}, & \qquad \frac{du_h}{dt}(0) = u_{1,h} \end{cases}$$

We know that there exists a sequence of eigenvalues $0 \leq \lambda_{1,h} \leq \lambda_{2,h} \leq \cdots \leq \lambda_{I,h}$ and an orthonormal hilbertian basis $(w_{i,h})$ of V_h such that

$$\forall\, v_h \in V_h, \quad a(w_{i,h}, v_h) = \lambda_{i,h}(w_{i,h}, v_h).$$

We put $\omega_{i,h} = \sqrt{\lambda_{i,h}}$.

As we saw, we introduce a basis $(\varphi_i)_{1\le i\le I}$ of V_h and seek a solution u_h as

$$u_h(t) = \sum_{j=1}^{I} \xi_j(t)\varphi_j,$$

then, writing

$$u_{0,h} = \sum_{j=1}^{I} \xi_{0,j}\varphi_j, \quad u_{1,h} = \sum_{j=1}^{I} \xi_{1,j}\varphi_j$$

and

$$\xi(t) = (\xi_1(t), \ldots, \xi_I(t))^T,$$
$$\chi(t) = (\chi_1(t), \ldots, \chi_I(t))^T, \quad \chi_i(t) = (f(t), \varphi),\ 1 \le i \le I,$$

the problem has the form

$$M\frac{d^2\xi}{dt^2}(t) + R\xi(t) = \chi(t),$$
$$\xi(0) = \xi_0, \quad \frac{d\xi}{dt}(0) = \xi_1,$$

where $M = ((\varphi_j, \varphi_i))_{1\le i,j\le I}$ and $R = (a(\varphi_j, \varphi_i))_{1\le i,j\le I}$, as before. We assume $f \in C^0(0, T; H)$, which implies $t \in [0, T] \mapsto \chi(t)$ is continuous.

Let ξ^n and σ^n be the approx. values of $\xi(t_n)$ and $\frac{d\xi}{dt}(t_n)$ resp. One has:

$$\frac{1}{(\Delta t)^2} M(\xi^{n+1} - \xi^n - (\Delta t)\sigma^n) + R\left(\beta\xi^{n+1} + \left(\frac{1}{2} - \beta\right)\xi^n\right) =$$

$$= \beta\chi(t_{n+1}) + \left(\frac{1}{2} - \beta\right)\chi(t_n), \quad 0 \leq n \leq N-1,$$

$$\frac{1}{\Delta t} M(\sigma^{n+1} - \sigma^n) + R(\gamma\xi^{n+1} + (1-\gamma)\xi^n) = \gamma\chi(t_{n+1}) + (1-\gamma)\chi(t_n),$$
$$1 \leq n \leq N-1,$$

$$\xi^0 = \xi_0, \quad \sigma^0 = \sigma_0, \quad \xi_0 \text{ and } \sigma_0 \text{ given vectors in } \mathbb{R}^I.$$

Removing σ^n from the equations:

$$\frac{1}{(\Delta t)^2}M(\xi^{n+2}-2\xi^{n+1}+\xi^n)+R\left(\beta\xi^{n+2}+\left(\frac{1}{2}-2\beta+\gamma\right)\xi^{n+1}+\right.$$

$$\left.+\left(\frac{1}{2}+\beta-\gamma\right)\xi^n\right)=\beta\chi(t_{n+2})+\left(\frac{1}{2}-2\beta+\gamma\right)\chi(t_{n+1})+$$

$$+\left(\frac{1}{2}+\beta-\gamma\right)\chi(t_n),\ 0\leq n\leq N-2,$$

and from the first equation with $n=0$:

$$\frac{1}{(\Delta t)^2}M(\xi^1-\xi^0-(\Delta t)\sigma_0)+R\left(\beta\xi^1+\left(\frac{1}{2}-\beta\right)\xi^0\right)=$$

$$=\beta\chi(t_1)+\left(\frac{1}{2}-\beta\right)\chi(t_0)$$

Then, for each time step, we have to solve a linear system

$$(M+\beta(\Delta t)^2R)\xi^{n+1}=\eta^n,$$

with the know vector $\eta^n\in\mathbb{R}^I$. As $\beta\geq 0$, the symmetric matrix $M+\beta(\Delta t)^2R$ is positive definite. We use one Choleski factorization to solve the system for all n. The method is explicit if $\beta=0$ and M diagonal.

Convergence of the approximate solution

Theorem

Suppose that $\beta \geq 0$ *and* $\delta = \gamma - \frac{1}{2} \geq 0$. *Then the solution* $\{(u_h^n, z_h^n) \in V_h \times V_h;\, 0 \leq n \leq N\}$ *satisfies*

- *If* $u \in C^2(0,T;V) \cap C^3(0,T;H)$ *and under the stability condition*

$$(\Delta t)^2 \lambda_{I,h} \leq \begin{cases} L \text{ if } \beta \geq \frac{(1+\delta)^2}{4}, & (\text{useless if } (\beta,\delta) = (1/4,0)) \\ \frac{4}{(1+\delta)^2 - 4\beta}(1-\epsilon) & \text{if } \beta < \frac{(1+\delta)^2}{4}, \end{cases}$$

  Then

$$|u_h^n - u(t_n)| \leq C\{|u_{0,h} - \Pi_h u_0| + |u_{1,h} - \Pi_h u_1| + |(I - \Pi_h)u(t_n)| +$$

$$+ \int_0^{t_n} \left\{ \left| (I - \Pi_h) \frac{d^2 u}{dt^2}(s) \right| + \Delta t \left| \frac{d^3 u}{dt^3}(s) \right| \right\} ds\},$$

  where the constant C *is independent of* h, Δt *and* u *(but depend on* L *or* ϵ*);*

Theorem

(cont.)

- *If $\delta = 0$ and if $u \in C^2(0,T;V) \cap C^4(0,T;H)$ and under the stability condition of the previous item (with $\delta = 0$), we have*

$$|u_h^n - u(t_n)| \leq C\{|u_{0,h} - \Pi_h u_0| + |u_{1,h} - \Pi_h u_1| + |(I - \Pi_h)u(t_n)| +$$

$$+ \int_0^{t_n} \left\{ \left| (I - \Pi_h) \frac{d^2 u}{dt^2}(s) \right| + (\Delta t)^2 \left| \frac{d^4 u}{dt^4}(s) \right| \right\} ds\},$$

  where C is independent of h, Δt and u.

Comment

If $\beta \geq (1+\delta)^2/4$ *with* $\delta > 0$, *using that* $\rho(B(\theta)) = \sqrt{1-\delta\alpha} < 1$ *we can prove, for* $\Delta t \leq \Delta t_0$:

$$|u_h^n - u(t_n)| \leq C\{(|u_{0,h} - \Pi_h u_0| + |u_{1,h} - \Pi_h u_1|)e^{-\mu(\Delta t)t_n} +$$

$$+|(I-\Pi_h)u(t_n)| + \int_0^{t_n} \left\{ \left|(I-\Pi_h)\frac{d^2u}{dt^2}(s)\right| + +\Delta t \left|\frac{d^3u}{dt^3}(s)\right| \right\} e^{-\mu\Delta(t_n-s)}\, ds\},$$

for constants $C = C(\lambda_1, L, \beta, \delta, \Delta t_0)$ *and* $\mu = \mu(\lambda_1, \beta, \delta, \Delta t_0)$.
When $\beta < (1+\delta)^2/4$ *and the condition of stability does not hold, the scheme is unstable and the error* $\max_{0\leq n\leq N} |u_h^n - u(t_n)| \to +\infty$ *as* h *and and* Δt *tends to zero. On the other hand, the condition when* $\beta \geq (1+\delta)^2/4$, *the condition* $(\Delta t)^2\lambda_{1,h} \leq L$ *does not a strong restriction, in practice, because* L *is arbitrary.*

Comment

In conclusion, the Newmark method corresponding to $\beta \geq 1/4$ and $\delta = 0$ is (inconditianally) stable and of order 2 in t. This gives a good approximation of the solution. The most used method is when $\beta = 1/4$ and $\delta = 0$. When the solution u does not smooth enough and we have to solve the problem in a large interval of time $[0, T]$, the approximate solution u_h have parasite oscillations that does not disappear. In this case, it is better to use the method with isuitable $\beta \geq (1+\delta)^2/4$, $\delta > 0$. This is a method of order 1, but the error behaves better.

Finally, the Newmark method corresponding to $\beta = \delta = 0$ is used frequently when M is diagonal. In this case the method is explicit and the condition of stability $(\Delta t)^2 \lambda_{I,h} \leq 4(1-\epsilon)$ give a weaker restriction that in the parabolic case.

For more information:
P. A. Raviart, J.-M. Thomas: *Introduction à l'analyse numérique des équations aux derivées partielles*. Masson (1988).