Computational Mathematics

José Augusto Ferreira

Department of Mathematics

University of Coimbra

2009-2010

Contents

1-Numerical Methods for ODEs

1.1 Some Analytical Results: Existence, Uniqueness, Stability

Ordinary differential equations are often used for mathematically model problems in many branches of sciences, engineering and economy. Such equations are frequently complemented by the state of the dependent variables at some initial time arising, naturally, the so called initial value problems- IVP.

The general formulation of an IVP for a system of ODEs is

$$u'(t) = F(t, u(t)), t > t_0,\ u(t_0) = u_0, \tag{1.1.1}$$

with $F : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ and $u_0 \in \mathbb{R}^m$.

Existence and Uniqueness Results:

We start by establishing some classical results that enable us to conclude the existence and uniqueness of a solution of the IVP (1.1.1).

For the scalar case we have:

Theorem 1.1.1 *[Picard's Theorem] Suppose that F is a continuous function in $\mathcal{R} = \{(t, u) \in \mathbb{R}^2 : t_0 \le t \le T, |u - u_0| \le \delta\}$. Suppose also that F has the Lipschitz constant L with respect to the second argument in $\mathcal{R}$. Finally, letting*

$$M = \max_{\mathcal{R}} |F|,$$

suppose that $M(T - t_0) \le \delta$. Then, there exists a unique continuously differentiable function u defined on the interval $[t_0, T]$ which satisfies (1.1.1).

The essence of the proof of the Picard's Theorem is to consider the sequence (u_n) defined by

$$u_0(t) = u_0,\ u_n(t) = u_0(t) + \int_{t_0}^{t} F(s, u_{n-1}(s))\, ds\, n \in \mathbb{N},\ t \in [t_0, T].$$

As $u_n \in C[t_0, T]$, showing that (u_n) converges uniformly on $[t_0, T]$ to u defined by

$$u(t) = u_0 + \int_{t_0}^{t} F(s, u(s))ds, t \in [t_0, T],$$

we conclude that u is continuously differentiable and is solution of (1.1.1).

The idea of the proof will be crucial for the construction of numerical methods for the IVP (1.1.1).

The Picard's Theorem has a natural extension to systems of ODEs. For this extension the modulus $|.|$ is naturally replaced by the Euclidian norm $\|.\|_2$ defined on $\mathbb{R}^m$ by $\|x\|_2 = \big(\sum_{i=1}^{m} x_i^2\big)^{1/2}$.

Theorem 1.1.2 *[Picard's Theorem] Suppose that F is a continuous function in $\mathcal{R} = \{(t,v) \in \mathbb{R}^{m+1} : t_0 \le t \le T, \|v-u_0\|_2 \le \delta\}$. Suppose also that F has the Lipschitz constant L with respect to the second argument in $\mathcal{R}$*

$$\|F(t,w) - F(t,v)\|_2 \le L\|u-v\|_2, (t,w), (t,v) \in \mathcal{R}. \tag{1.1.2}$$

Finally, letting

$$M = \max_{(t,v)\in\mathcal{R}} \|F(t,v)\|_2,$$

suppose that $M(T-t_0) \le \delta$. Then, there exists a unique continuously differentiable function u defined on the closed interval $[t_0,T]$ which satisfies (1.1.1).

The proof of this results follows the proof of the Theorem 1.1.1 being the sequence (u_n) defined analogously. Both proofs can be seen, for example, in the classical book [2].

If the Jabonian matrix of F, $\dfrac{\partial F}{\partial v}(t,v) = [\dfrac{\partial F_i}{\partial v_j}(t,v)]$ satisfies

$$\|\frac{\partial F}{\partial v}(t,v)\|_2 \le L, (t,v) \in \mathcal{R}, \tag{1.1.3}$$

then F satisfies the Lipschitz condition (1.1.2). In (1.1.3), the norm $\|.\|_2$ is the matrix norm subordinated to the Euclidian vector norm defined.

Let us now consider the linear systems

$$u'(t) = Au(t) + g(t),\ t > t_0,\ u(t_0) = u_0. \tag{1.1.4}$$

The unique solution of the IVP (1.1.4) admits the representation

$$u(t) = e^{(t-t_0)A}u_0 + \int_{t_0}^{t} e^{(t-s)A}g(s)ds,\ t \ge t_0. \tag{1.1.5}$$

In the representation (1.1.5) the exponential of the matrix $(t-t_0)A$ represents the sum

$$\sum_{n=0}^{\infty} \frac{(t-t_0)^n A^n}{n!}.$$

We remark that the last exponential is defined considering a matrix norm. Since the individual terms in the power series are bounded by $\dfrac{(t-t_0)^n\|A\|^n}{n!}$, it follows that the power series converges and $\|e^{(t-t_0)A}\| \le e^{(t-t_0)\|A\|}$.

Stability Results:

A crucial concept on the analysis of the behaviour of the solution of (1.1.1) is the concept of stability regarding to perturbations of the initial condition. Such concept means that if two solutions of the same problem starts close enough, then they stay close enough in some interval.

A solution u of the IVP (1.1.1) is said to be stable on the interval $[t_0, T]$ if, for every $\epsilon > 0$, there exists $\eta > 0$ such that for all v_0 satisfying $\|u_0 - v_0\| < \eta$, the solution of the IVP

$$v'(t) = F(t, v(t)),\ t > t_0,\ v(t_0) = v_0,$$

is defined on $[t_0, T]$ and satisfies

$$\|u(t) - v(t)\| < \epsilon,\ t \in [t_0, T].$$

If u is stable on $[t_0, \infty)$ (i.e. u is stable on $[t_0, T]$ for all $T > t_0$ with η independent of T), then u is said to be stable in the sense of Lyapunov. Moreover, if

$$\lim_{t\to\infty} \|u(t) - v(t)\| = 0$$

then u is called asymptotically stable.

The concept of stability depends on the norm used. In the definition of stability, $\|.\|$ represents any norm in $\mathbb{R}^m$.

On the assumptions of Picard's Theorem, the solution of (1.1.1) is stable. In fact, we have the following result:

Theorem 1.1.3 *Under the assumptions of the Theorem 1.1.2, the solution u of (1.1.1) is stable in* $[t_0, T]$.

Proof: As we have

$$u(t) = u_0 + \int_{t_0}^{t} F(s, u(s))\, ds, v(t) = v_0 + \int_{t_0}^{t} F(s, v(s))\, ds,$$

then

$$\|u(t) - v(t)\|_2 \le \|u_0 - v_0\|_2 + L \int_{t_0}^{t} \|u(s) - v(s)\|_2\, ds, t \in [t_0, T]. \tag{1.1.6}$$

From the inequality (1.1.6) we get

$$\frac{d}{dt}\Big(e^{-Lt} \int_{t_0}^{t} \|u(s) - v(s)\|_2\, ds\Big) \le e^{-Lt}\|u_0 - v_0\|_2,$$

which is equivalent to

$$\frac{d}{dt}\Big(e^{-Lt} \int_{t_0}^{t} \|u(s) - v(s)\|_2\, ds + \frac{1}{L}e^{-Lt}\|u_0 - v_0\|_2\Big) \le 0. \tag{1.1.7}$$

The inequality (1.1.7) shows that $e^{-Lt} \displaystyle\int_{t_0}^{t} \|u(s) - v(s)\|_2\, ds + \frac{1}{L}e^{-Lt}\|u_0 - v_0\|_2$ is a non-increasing function on $[t_0, T]$. Consequently,

$$L \int_{t_0}^{t} \|u(s) - v(s)\|_2\, ds \le \|u_0 - v_0\|_2\big(e^{L(t-t_0)} - 1\big). \tag{1.1.8}$$

Taking into account in (1.1.6) the upper bound (1.1.8) we deduce

$$\|u(t) - v(t)\|_2 \leq \|u_0 - v_0\|_2 e^{L(t-t_0)}\,, t \in [t_0, T],$$

which conclude the proof.

■

Let us consider now the linear system (1.1.4). If u and v are solutions of the previous system with initial conditions u_0, v_0, respectively, then we obtain

$$\|u(t) - v(t)\| \leq \|e^{(t-t_0)A}\|\|u_0 - v_0\|.$$

As we have

$$\|e^{\mu A}\| \leq e^{\mu\|A\|}, \mu > 0, \tag{1.1.9}$$

we conclude

$$\|u(t) - v(t)\| \leq e^{(t-t_0)\|A\|}\|u_0 - v_0\|. \tag{1.1.10}$$

Generally, the last inequality does not give useful information because the inequality (1.1.9) gives a large over-estimation. In fact, for example, for the scalar case with $A = -\lambda, \lambda >> 1$, we have $\|e^{\mu A}\| = e^{-\mu\lambda} << e^{\mu\|A\|} = e^{\mu\lambda}$ and

$$|u(t) - v(t)| = e^{-\lambda(t-t_0)}|u_0 - v_0|.$$

It is desirable to have the estimate

$$\|e^{\mu A}\| \leq Ke^{\mu\omega}, \mu > 0, \tag{1.1.11}$$

with constant $K > 0$ and $\omega \in \mathbb{R}$. In this case we obtain the stability estimate

$$\|u(t) - v(t)\| \leq Ke^{(t-t_0)\omega}\|u_0 - v_0\|. \tag{1.1.12}$$

If $\omega < 0$ we conclude the asymptotic stability of the solution of the IVP (1.1.4).

The natural question that we should answer is the following: In what conditions holds (1.1.12)? If A is diagonalizable, $A = MDM^{-1}$ and then

$$\|e^{\mu A}\| \leq \|M\|\|e^{tD}\|\|M^{-1}\| = cond(M)\|e^{tD}\|,$$

where $cond(M)$ denotes the condition number of M. Immediately we obtain

$$\|e^{\mu A}\| \leq cond(M) \max_{i=1,\ldots,m} |e^{t\lambda_i}| = cond(M) \max_{i=1,\ldots,m} |e^{tRe(\lambda_i)}|$$

provided that

$$\|e^{tD}\| \leq \max_{i=1,\ldots,m} |e^{t\lambda_i}|.$$

Then, holds (1.1.12) with

$$\omega = \max_{i=1,\ldots,m} Re(\lambda_i).$$

Nevertheless, an estimate to $cond(M)$ should be obtained. In particular, if A is a normal matrix (i.e. $AA^* = A^*A$ where $A^* = \bar{A}^T$), then A has a complete set of orthogonal eigenvectors. Consequently, $A = MDM^{-1}$, where M is a unitary matrix, and we conclude that

$$\|e^{\mu A}\|_2 \leq \max_{i=1,\ldots,n} |e^{t\lambda_i}|.$$

We introduce in what follows a more general convenient concept to obtain bounds for $\|e^{tA}\|$.

The Logarithmic Norm of Matrices:

Let A in $\mathbb{R}^m \times \mathbb{R}^m$ or in $\mathbb{C}^m \times \mathbb{C}^m$. The logarithmic norm of A is defined by

$$\mu[A] = \lim_{\tau \to 0} \frac{\|I + \tau A\| - 1}{\tau}, \ \tau > 0. \tag{1.1.13}$$

Lemma 1 *The limit in (1.1.13) exists for all matrix norm $\|.\|$ and for all matrices A provided that the matrix norm $\|.\|$ satisfies $\|I\| = 1$.*

Proof: Let $\theta \in (0, 1)$. We have

$$\frac{\|I + \theta\tau A\| - 1}{\theta\tau} \leq \frac{\theta\|I + \tau A\| + (1-\theta)\|I\| - 1}{\theta\tau} \leq \frac{\|I + \tau A\| - 1}{\tau},$$

for $\theta \in (0, 1)$. From the last inequality we conclude that the ratio appearing in (1.1.13) is monotonically non-decreasing function on τ. As

$$-\|A\| \leq \frac{\|I + \tau A\| - 1}{\tau} \leq \|A\|$$

holds, we conclude that it has finite limit.

■

Some properties of the logarithmic norm of matrices are presented in the next result.

Proposition 1 *For A, B in $\mathbb{R}^m \times \mathbb{R}^m$ or in $\mathbb{C}^m \times \mathbb{C}^m$ we have the following:*

1. $\mu[\gamma A] = \gamma\mu[A]$, $\gamma \geq 0$,
2. $\mu[sI + \gamma A] = s + \gamma\mu[A], s \in \mathbb{R}, \gamma \geq 0$;
3. $\mu[A + B] \leq \mu[A] + \mu[B]$;
4. *The logarithmic norm is a continuous function, i.e.*

$$|\mu[A] - \mu[B]| \leq \|A - B\|;$$

5. *For the matrix norm induced by inner product $< ., . >$ holds the following*

$$\mu[A] = \max_{v \neq 0} \frac{Re < Av, v >}{\|v\|^2}; \tag{1.1.14}$$

6. If $A \in \mathbb{R}^m \times \mathbb{R}^m$, then

(a)

$$\mu_2[A] = \lambda_{max}(\frac{A + A^T}{2}), \tag{1.1.15}$$

where $\lambda_{max}(\frac{A + A^T}{2})$ denotes the largest eigenvalue of $\frac{A + A^T}{2}$;

(b) $\mu_1[A] = \max_j \Big(a_{jj} + \sum_{i \neq j} |a_{ij}|\Big)$;

(c) $\mu_\infty[A] = \max_i \Big(a_{ii} + \sum_{j \neq i} |a_{ij}|\Big)$;

Proof: We have

$$\begin{aligned}
\frac{\|I + \tau A\| - 1}{\tau} &= \max_{v \neq 0} \frac{\|v + \tau Av\| - \|v\|}{\tau \|v\|} \\
&= \max_{v \neq 0} \frac{\|v + \tau Av\|^2 - \|v\|^2}{\tau \|v\| (\|v + \tau Av\| + \|v\|)} \\
&= \max_{v \neq 0} \frac{2\tau Re < Av, v > + \tau^2 \|A\|^2 \|v\|^2}{\tau \|v\| (\|v + \tau Av\| + \|v\|)} \\
&= \max_{v \neq 0} \frac{\tau Re < Av, v > + \frac{\tau^2}{2} \|A\|^2 \|v\|^2}{\tau \|v\|^2 + \frac{\tau}{2} \|v\| (\|v + \tau Av\| - \|v\|)}
\end{aligned}$$

and then we conclude (1.1.14).

From Property 5, we easily obtain

$$< Av, v > = < \frac{A + A^T}{2} v, v > \leq \lambda_{max}(\frac{A + A^T}{2}) \|v\|^2.$$

If v is the eigenvector corresponding to $\lambda_{max}(\frac{A + A^T}{2})$, then the last inequality holds as equality.

As $\mu_1[A] = \max_j \Big(|\frac{1}{\tau} + a_{jj}| + \sum_{i \neq j} |a_{ij}|\Big) - \frac{1}{\tau}$ and when $\tau \to 0^+$ we have $|\frac{1}{\tau} + a_{jj}| = \frac{1}{\tau} + a_{jj}$ we conclude 6*b*.

The proof of 6*c* is similar to the proof of 6*b*.

■

The importance of the logarithmic norm lies in the following result.

Theorem 1.1.4 *Let $A \in \mathbb{C}^m \times \mathbb{C}^m$ and $\omega \in \mathbb{R}$. We have*

$$\mu[A] \leq \omega \Longleftrightarrow \|e^{\eta A}\| \leq e^{\eta \omega}, \forall \eta \geq 0.$$

Proof: Suppose that $\mu[A] \leq \omega$. We start by noting that

$$e^{\tau A} = I + \tau A + O(\tau^2).$$

and then

$$e^{n\tau A} = e^{\eta A} = \lim_{\tau \to 0} (I + \tau A)^n, \tag{1.1.16}$$

for n and τ such that $n\tau = \eta$. Otherwise, as $\mu[A] \leq \omega$, we have

$$\|I + \tau A\| \leq 1 + \omega\tau + O(\tau^2). \tag{1.1.17}$$

As a consequence

$$\|(I + \tau A)^n\| \leq \left(1 + \omega\tau + O(\tau^2)\right)^n \to e^{\eta\omega}, \tag{1.1.18}$$

for n and τ as before. From (1.1.16) and (1.1.18) we get $\|e^{\eta A}\| \leq e^{\omega\eta}$.

On the other hand, suppose that $\|e^{\eta A}\| \leq e^{\omega\eta}$ for all $\eta > 0$. Since $I + \tau A = e^{\tau A} + O(\tau^2)$ it follows the inequality (1.1.17) from which we obtain $\mu[A] \leq \omega$.

■

Using the last characterization result, we can establish asymptotic stability of the solution of (1.1.4)

- with respect to norm $\|.\|_2$ for real matrices such that $< Av, v > \leq 0$;
- with respect to norm $\|.\|_\infty$ if A has negative diagonal entries and A is row-wise diagonally dominant;
- with respect to norm $\|.\|_1$ if A has negative diagonal entries and A is column-wise diagonally dominant.

The Stability of Nonlinear IVP

We intent to study the stability of the solution (1.1.1). Let u and w be two solutions with initial conditions u_0 and w_0, respectively. From the Mean Value Theorem for vectorial functions for $Z(t) = u(t) - w(t)$ holds the following representation

$$Z'(t) = \int_0^1 \frac{\partial F}{\partial v}(t, \sigma u(t) + (1-\sigma)w(t))d\sigma Z(t) = M(t)Z(t) \tag{1.1.19}$$

Then,

$$\begin{aligned} \frac{d}{dt}\|Z(t)\| &= \lim_{\tau\to 0} \frac{\|Z(t+\tau)\| - \|Z(t)\|}{\tau} \\ &= \lim_{\tau\to 0} \frac{\|Z(t) + (Z(t+\tau) - Z(t))\| - \|Z(t)\|}{\tau} \\ &= \lim_{\tau\to 0} \frac{\|Z(t) + \tau M(t)Z(t))\| - \|Z(t)\|}{\tau} \\ &= \lim_{\tau\to 0} \frac{\|I + \tau M(t)\| - 1}{\tau}\|Z(t)\|. \end{aligned} \tag{1.1.20}$$

If

$$\mu[M(t)] \leq \omega, t \geq t_0, \tag{1.1.21}$$

then we get

$$\frac{d}{dt}\|Z(t)\| \leq \omega\|Z(t)\|.$$

As from the last inequality we obtain

$$\|Z(t)\| \leq e^{\omega(t-t_0)}\|Z(t_0)\|,$$

we conclude the stability of u on $[t_0, T]$ for some T.

In what conditions the inequality (1.1.21) holds?

Theorem 1.1.5 *If the Jacobian of F,* $JF = \dfrac{\partial F}{\partial v}$, *satisfies*

$$\mu[\frac{\partial F}{\partial v}(t, \sigma u + (1-\sigma)v)] \leq \omega, \forall \sigma \in [0,1], \tag{1.1.22}$$

then

$$\mu[\int_0^1 \frac{\partial F}{\partial v}(t, \sigma u + (1-\sigma)v)\, d\sigma] \leq \omega. \tag{1.1.23}$$

Proof: Let $M(t)$ be defined by $M(t) = \displaystyle\int_0^1 \frac{\partial F}{\partial v}(t, \sigma u + (1-\sigma)v)\, d\sigma$. Then

$$\|I + \tau M(t)\| \leq \max_{\xi = \sigma u + (1-\sigma)v, \sigma \in [0,1]} \|I + \tau \frac{\partial F}{\partial v}(t, \xi)\|.$$

Consequently

$$\begin{aligned} \mu[M(t)] &\leq \lim_{\tau \to 0} \max_{\xi} \frac{\|I + \tau \frac{\partial F}{\partial v}(t, \xi)\| - 1}{\tau} \\ &\leq \max_{\xi} \mu[\frac{\partial F}{\partial v}(t, \xi)] \\ &\leq \omega. \end{aligned}$$

In order to justify the second inequality we point out that, with

$$\mu(\tau)[A] = \frac{\|I + \tau A\| - 1}{\tau}, \tau > 0,$$

for any matrix A, we have

$$\mu(\tau)[\beta A_1 + (1-\beta)A_2] \leq \beta \mu(\tau)[A_1] + (1-\beta)\mu(\tau)[A_2],\ \beta \in [0,1].$$

Hence, $\mu(\tau)$ is convex and, due to this fact, μ is continuous in A.This implies that μ is also convex and thus continuous in A being the limit of a convergent sequence of convex functions. Furthermore, the sequence is monotone. We conclude that the convergence of $\mu(\tau)[A]$ to $\mu[A]$ is uniform on a bounded close matrix set which implies the second inequality.

■

Finally we establish the stability result for the solution of (1.1.1).

Theorem 1.1.6 *If the Jacobian of F,* $\dfrac{\partial F}{\partial v}$, *satisfies (1.1.22), then the solution u of (1.1.1) is stable on* $[t_0, T]$.

1.2 Discretizations for ODEs

The θ-Method

We introduce below a family of numerical methods, firstly for the scalar case - the $\theta-$methods. We define in $[t_0, T]$ the grid $\{t_n, n = 0, \ldots, N\}$ with $t_{n+1} = t_n + \Delta t$, with the step size $\Delta t = \dfrac{T - t_0}{N}$. We seek a numerical approximation u_n to the solution of (1.1.1). Integrating (1.1.1) between to consecutive mesh points t_n and t_{n+1} we deduce

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} F(s, u(s))\, ds, n = 0, \ldots, N-1. \tag{1.2.1}$$

Considering the one-parameter family of integration rules of the form

$$\int_{t_n}^{t_{n+1}} g(s)\, ds \simeq \Delta t\Big((1-\theta)g(t_n) + \theta g(t_{n+1})\Big), \tag{1.2.2}$$

with the parameter θ in $[0, 1]$, and applying (1.2.2) with $g(s) = F(s, u(s))$ we find the following one-parameter family of methods : given u_0, we define

$$u_{n+1} = u_n + \Delta t\Big((1-\theta)F(t_n, u_n) + \theta F(t_{n+1}, u_{n+1})\Big), n = 0, \ldots, N-1, \tag{1.2.3}$$

parameterised by $\theta \in [0, 1]$. When $\theta = 0$ we obtain the explicit Euler's method, being the implicit Euler's method defined when $\theta = 1$. The trapezium rule method is obtained for $\theta = \dfrac{1}{2}$.

Example 1 *Let consider the IVP*

$$u'(t) = -50(u - cos(t)), t \in (0, 1],\ u(0) = 0. \tag{1.2.4}$$

In Figure 1 we plot the numerical solutions obtained with the θ-methods for $\theta = 0, 1, 0.5$.

The numerical solution obtained with the explicit Euler's method presents some oscillations near to $t = 0$ for $\Delta t \geq \dfrac{1}{44}$. We point out that the solution of the IVP (1.2.4) is given by

$$u(t) = \frac{50^2}{50^2+1}\big(cos(t) - e^{-50t}\big) + \frac{sin(t)}{50^2+1},\ t \in [0, 1].$$

■

Example 2 *The explicit Euler's method applied to the IVP (1.1.1) with $m = 2$ is given by*

$$\begin{cases} u_{1,n+1} = u_{1,n} + \Delta t F_1(t_n, u_{1,n}, u_{2,n}) \\ \\ u_{2,n+1} = u_{2,n} + \Delta t F_2(t_n, u_{1,n}, u_{2,n}),\ n = 0, \ldots, N \end{cases} \tag{1.2.5}$$

with $u_{1,0} = u_1(t_0), u_{2,0} = u_2(t_0)$. We consider

$$F_1(t, u_1(t), u_2(t)) = d u_1(t) + \frac{1}{\epsilon} u_2(t)$$

and

$$F_2(t, u_1(t), u_2(t)) = -\frac{1}{\epsilon} u_2(t).$$

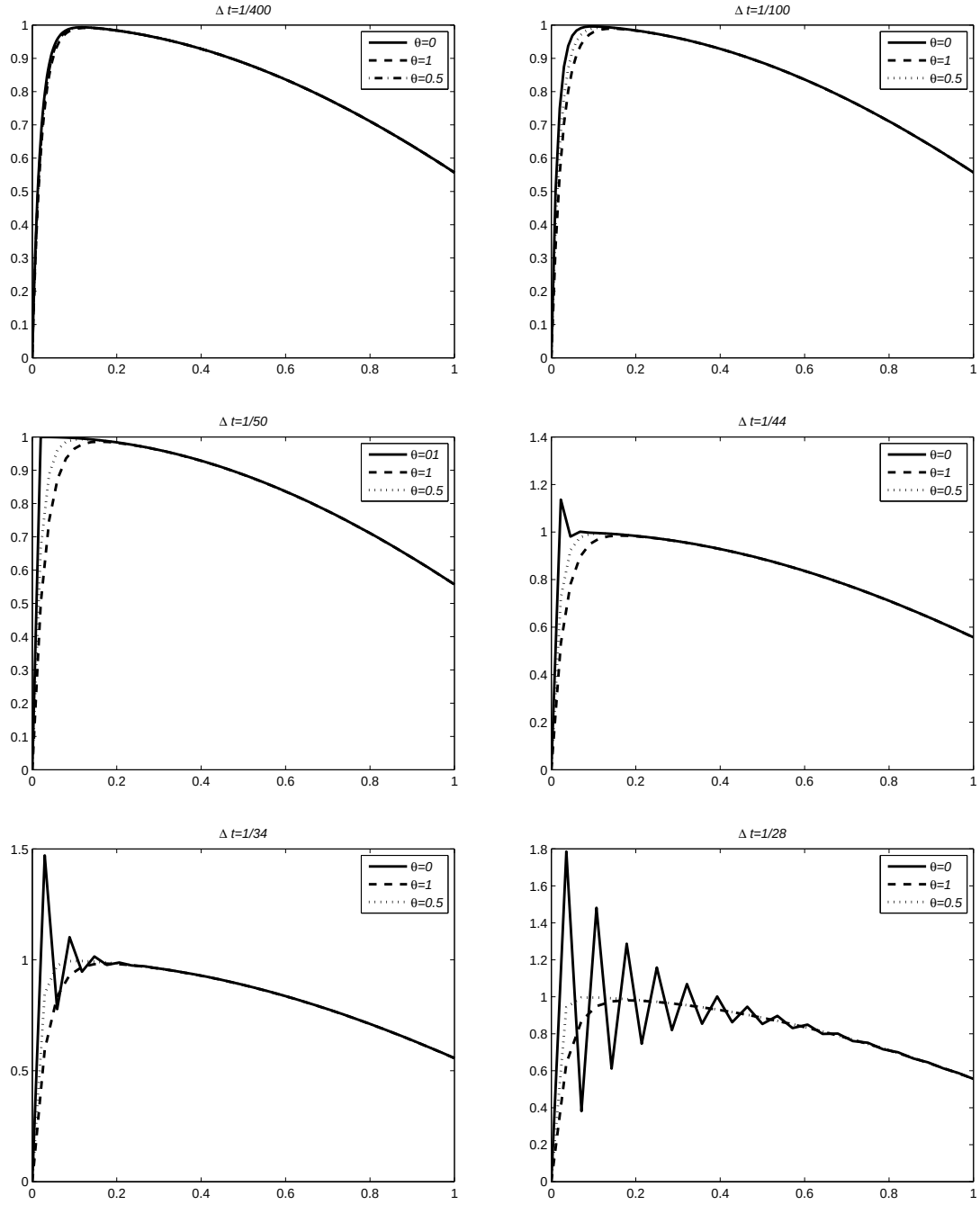

Figure 1: Numerical results obtained with the explicit Euler, implicit Euler and trapezium rule methods for the IVP (1.2.4).

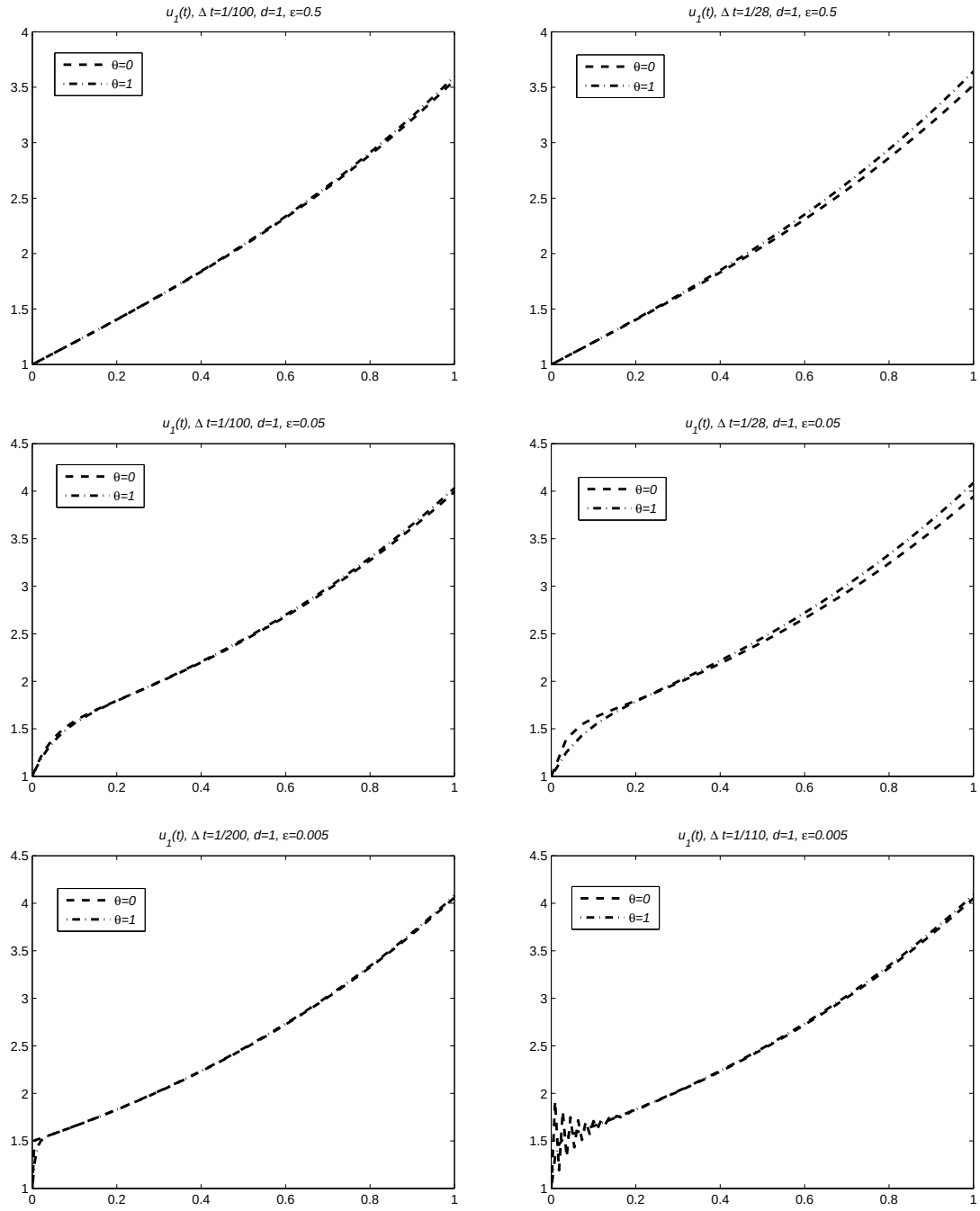

Figure 2: Numerical approximations for u_1 obtained with the explicit Euler and implicit Euler methods

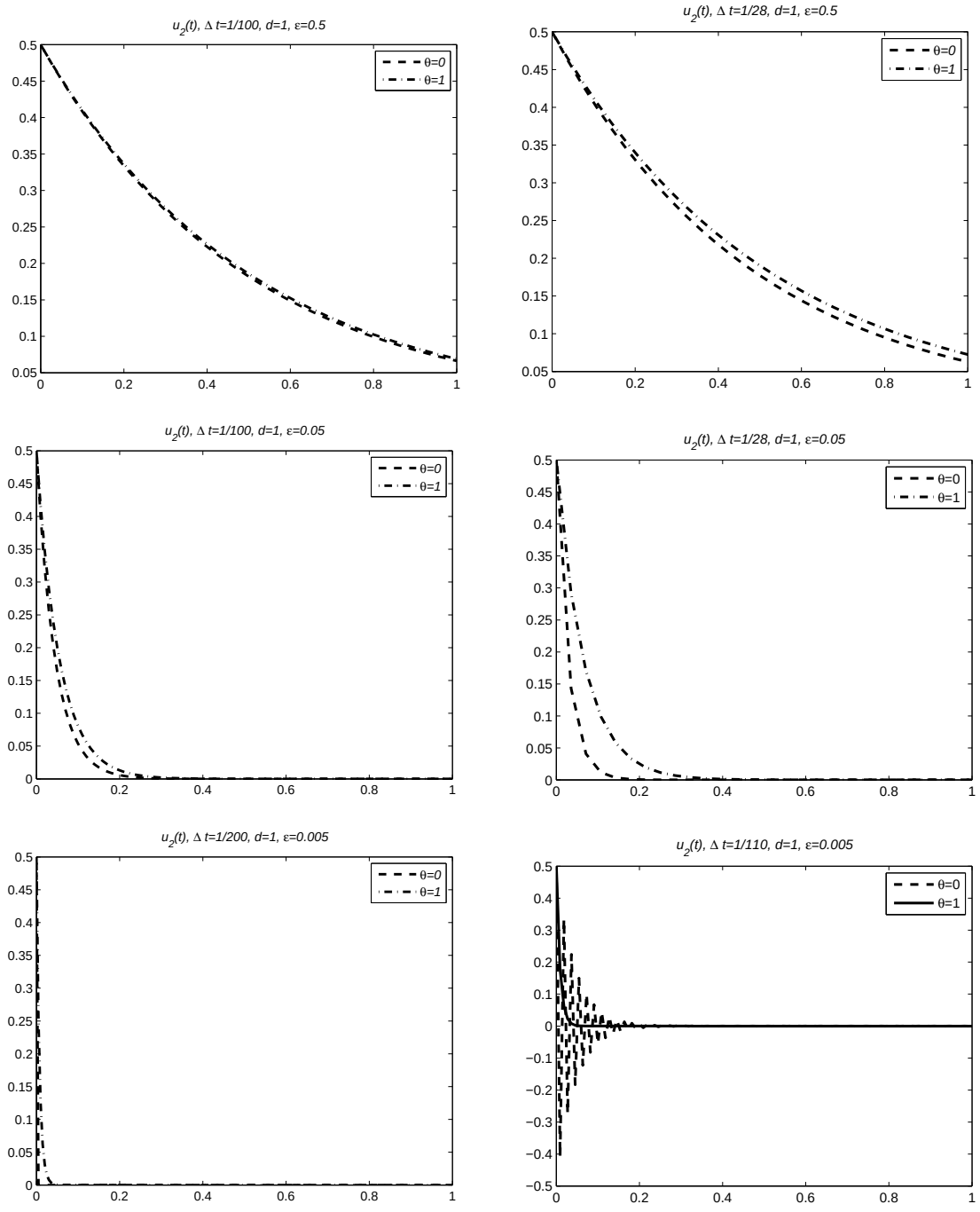

Figure 3: Numerical approximations for u_2 obtained with the explicit Euler and implicit Euler methods

In the numerical experiments we took $u_1(0) = 1, u_2(0) = 0.5$ *and* $d = 1$. *The numerical results plotted in Figures 2 and 9 were obtained for several values of* ϵ *and* Δt.

We point out that for $\epsilon \to 0$ *the explicit Euler's methods requires a very small step size.*

■

The Runge-Kutta Methods:

Another class of methods often used on the numerical computations is the Runge-Kutta methods. On the evaluation of the numerical approximation at time level t_{n+1}, u_{n+1}, the methods of this class only use the numerical approximation at time level t_n.

We consider in what follows the class of s-stage Runge-Kutta methods defined by

$$u_{n+1} = u_n + \Delta t \sum_{r=1}^{s} c_r k_r, \tag{1.2.6}$$

with

$$k_r = F(t_n + a_r \Delta t, u_n + \Delta t \sum_{i=1}^{s} b_{ri} k_i), r = 2, \ldots, s. \tag{1.2.7}$$

A convention for (1.2.6)-(1.2.7) frequently used is

$$a_r = \sum_{j=1}^{s} b_{rj}. \tag{1.2.8}$$

This formula is natural since it implies that the Runge-Kutta method gives the same approximation values for the non-autonomous system $w'(t) = F(w, t)$ as for the augmented system

$$\begin{pmatrix} \text{w'} \\ \\ \text{t'} \end{pmatrix} = \begin{pmatrix} F(w,t) \\ \\ 1 \end{pmatrix}.$$

The coefficients of the R-stage Runge-Kutta methods can be condensed in the Butcher table:

a_1	b_{11}	b_{12}	b_{13}	$\ldots$	b_{1s-1}	b_{1s}
a_2	b_{21}	b_{22}	b_{23}	$\ldots$	b_{2s-1}	b_{2s}
a_3	b_{31}	b_{32}	b_{33}	$\ldots$	b_{3R-1}	b_{3R}
$\ldots$	$\ldots$					
a_s	b_{s1}	b_{s2}	b_{s3}	$\ldots$	b_{ss-1}	b_{ss}
	c_1	c_2	c_3	$\ldots$	c_{s-1}	c_s

The method (1.2.6) is called explicit if $b_{ij} = 0$ for $j \geq i, i, j = 1, \ldots, s$, since then the internal approximations k_i can be computed one after another from an explicit relation. Otherwise the method is called implicit due to the fact that $k_i, i = 1, \ldots, s$, must be obtained from a system of linear or nonlinear equations.

The explicit s-stage Runge-Kutta methods can be represented by the following Butcher-table

a_2	b_{21}					
a_3	b_{31}	b_{32}				
$\dots$	$\dots$					
a_s	b_{s1}	b_{s2}	b_{s3}	$\dots$	b_{ss-1}	
	c_1	c_2	c_3	$\dots$	c_{s-1}	c_s

The computational cost increases when we consider an implicit method. Let us consider the application of a θ-method. If F is linear in the second argument then, in each time step, we need to solve a linear system $Au_{n+1} = b$. This system can be solved using a direct method like Gaussian elimination with pivoting strategies or a stationary iterative method

$$v^{m+1} = Bv^m + c.$$

Those methods define a sequence (v^m) which should converge to u_{n+1} and we can consider

- Jacobi method :
$$B = D^{-1}(L + U),\ c = D^{-1}b,$$
where D, $-L$ and $-U$ represent the diagonal, strictly lower triangular and strictly upper triangular parts of A, respectively;

- Gauss-Seidel method :
$$B = (D - L)^{-1}U,\ c = (D - L)^{-1}b;$$

- successive overrelaxation method (SOR) :
$$B = (D - \omega L)^{-1}\Big(\omega U + (1 - \omega)D\Big),\ c = \omega\Big(D - \omega L\Big)^{-1}b,\ \omega \in [0, 1].$$
(See the convergence of the previous iterative methods).

Another class of methods can be used, the so called non-stationary methods like step descendent methods.
However, if F is nonlinear in the second argument then we need to solve in each time level a non linear system $G(u_{n+1}) = 0$. The most popular method to solve this problem is the Newton's method which defines a sequence (v^m) as follows

$$v^{m+1} = v^m - \frac{G(v^m)}{G'(v^m)},$$

for the scalar case. In the vectorial case we have

$$v^{n+1} = v^m + C^m$$

where C^m is the solution of the linear system

$$JG(v^m)C^m = -G(v^m).$$

If we consider an implicit Runge-Kutta method and F is linear in the second argument then in the computation of the parameters $k_r, r = 1, \dots, s$, a linear system with $s \times m$ equations should be solved. Otherwise if F is nonlinear in the second argument the computation of the mentioned parameters requires the computation of the solution of a nonlinear system with $s \times m$ equations.

1.3 The one-step methods

1.3.1 Consistency

The methods considered before belong to the class of one-step methods. Such methods are characterized by the use of u_n on the computation of u_{n+1}. The family of one-step methods admits the representation

$$u_{n+1} = u_n + \Delta t \phi(t_n, u_n, u_{n+1}, \Delta t), n = 0, \ldots, N-1, u_0 = u(t_0). \tag{1.3.1}$$

For the particular case of the explicit methods we have the representation

$$u_{n+1} = u_n + \Delta t \phi(t_n, u_n, \Delta t), n = 0, \ldots, N-1, u_0 = u(t_0). \tag{1.3.2}$$

In (1.3.1) and (1.3.2), $u_n \in \mathbb{R}^m$ and $\phi : [t_0, T] \times \mathbb{R}^{2m} \times [0, \Delta t_0] \to \mathbb{R}^m$, $\phi : [t_0, T] \times \mathbb{R}^m \times (0, \Delta t_0] \to \mathbb{R}^m$, respectively, for some Δt_0.

Let us replace in (1.3.1) u_n by $u(t_n)$. We define the truncation error T_n by

$$T_n = \frac{u(t_{n+1}) - u(t_n)}{\Delta t} - \phi(t_n, u(t_n), u(t_{n+1}), \Delta t). \tag{1.3.3}$$

The one-step method (1.3.1) is said consistent with the equation $u'(t) = F(t, u(t))$, if

$$\lim_{\Delta t \to 0, n \to \infty} T_n = 0, \ n\Delta t \leq T - t_0.$$

As we have

$$\lim_{\Delta t \to 0} T_n = \lim_{\Delta t \to 0} u'(t_n) + O(\Delta t) - \phi(t_n, u(t_n), u(t_n) + O(\Delta t), \Delta t),$$

we conclude that the one-step method (1.3.1) is consistent if and only if

$$F(t, u) = \phi(t, u, u, 0),$$

provided that ϕ is a continuous function. Furthermore, if the order of the truncation error is p, which means that

$$\|T_n\| \leq C \Delta t^p, n = 0, \ldots, N-1, \tag{1.3.4}$$

where C is Δt independent, and p is the largest positive integer satisfying the last inequality, then the one-stem method (1.3.1) is said with consistency order p.

Example 3 *The truncation error of the explicit Euler's method is given by*

$$T_n = \frac{1}{2} \Delta t u''(\bar{t}), \bar{t} \in [t_n, t_{n+1}].$$

For truncation error the implicit method holds

$$T_n = -\frac{1}{2} \Delta t u''(t^*), t^* \in [t_n, t_{n+1}].$$

The consistency order of the Euler's method is equal to one provided that u has bounded second derivative.

Example 4 *The consistency order of the trapezium rule method is equal to two. In fact, we have*

$$T_n = \frac{\Delta t}{2}u''(t_n) + O(\Delta t^2) - \frac{\Delta t}{2}F_u(t, u(t_n))u'(t_n) = O(\Delta t^2).$$

Example 5 *The explicit Euler's method applied to the IVP (1.1.1) with $m = 2$ is given by*

$$\begin{cases} u_{1,n+1} = u_{1,n} + \Delta t F_1(t_n, u_{1,n}, u_{2,n}) \\ \\ u_{2,n+1} = u_{2,n} + \Delta t F_2(t_n, u_{1,n}, u_{2,n}),\ n = 0, \ldots, N \end{cases} \tag{1.3.5}$$

with $u_{1,0} = u_1(t_0), u_{2,0} = u_2(t_0)$. Such method has the following truncation error

$$T_n = \frac{\Delta t}{2}\begin{pmatrix} u_1'(\bar{t}) \\ \\ u_1'(t^*) \end{pmatrix}$$

with $\bar{t}, t^* \in [t_n, t_{n+1}]$. For the implicit version of the method (1.3.5) we have an analogous truncation error.

Example 6 *The explicit two-stage R-K method is second-order consistent provided that*

$$c_1 + c_2 = 1,\ a_2c_2 = b_{21}c_2 = \frac{1}{2}.$$

Then $a_2 = b_{21}, c_2 = \dfrac{1}{2a_2}, c_1 = 1 - \dfrac{1}{2a_2}$.

For $a_2 = \dfrac{1}{2}$, we obtain the modified Euler's method

$$u_{n+1} = u_n + \Delta t F(t_n + \frac{\Delta t}{2}, u_n + \frac{\Delta t}{2}F(t_n, u_n)) \tag{1.3.6}$$

The improved Euler's method is obtained for $a_2 = 1$,

$$u_{n+1} = u_n + \frac{\Delta t}{2}\left(F(t_n, u_n) + F(t_{n+1}, u_n + \Delta t F(t_n, u_n))\right). \tag{1.3.7}$$

Example 7 *The explicit tree-stage R-K method is third-order consistent provided that*

$$\begin{cases} c_1 + c_2 + c_3 = 1, \\ \\ c_2a_2 + c_3a_3 = \dfrac{1}{2}, \\ \\ c_2a_2^2 + c_3a_3^2 = \dfrac{1}{3}, \\ \\ \qquad c_3a_2b_{32} = \dfrac{1}{6}. \end{cases} \tag{1.3.8}$$

The system (1.3.8) defines a two-parameter family of 3-stage R-K methods. Notables examples of this family are

1. the Heun method

$$\begin{aligned} u_{n+1} &= u_n + \frac{1}{4}\Delta t\big(k_1 + 3k_3\big) \\ k_1 &= F(t_n, u_n) \\ k_2 &= F(t_n + \frac{1}{2}\Delta t, u_n + \frac{1}{3}\Delta t k_1) \\ k_3 &= F(t_n + \frac{2}{3}\Delta t, u_n + \frac{2}{3}\Delta t k_2), \end{aligned} \tag{1.3.9}$$

with the Butcher-table

$$\begin{array}{c|ccc} \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

2. the standard third-order R-K method

$$\begin{aligned} u_{n+1} &= u_n + \frac{1}{6}\Delta t\big(k_1 + 4k_2 + k_3\big) \\ k_1 &= F(t_n, u_n) \\ k_2 &= F(t_n + \frac{1}{2}\Delta t, u_n + \frac{1}{2}\Delta t k_1) \\ k_3 &= F(t_{n+1}, u_n - \Delta t k_1 + 2\Delta t k_2), \end{aligned} \tag{1.3.10}$$

with the Butcher-table

$$\begin{array}{c|ccc} \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array}$$

1.3.2 Convergence

Let $e_n = u(t_n) - u_n, n = 0, \ldots, N$, be the global error of the approximation u_n. The one-step method is said to be convergent if

$$e_n \to 0, \Delta t \to 0, n \to \infty,\ n\Delta t \leq T - t_0.$$

If

$$\|e_n\| \leq C\Delta t^q$$

with C time independent and $\Delta t \in (0, \Delta t]$, where Δt_0 is an upper bound for the time stepsize, being q the largest positive number with the above property, then the one-step method is said with order q.

For the one-step method (1.3.2), the truncation error T_n and the global error e_n are related by the following equality

$$e_{n+1} = e_n + \Delta t\big(\phi(t_n, u(t_n), \Delta t) - \phi(t_n, u_n, \Delta t)\big) + \Delta t T_n, n = 0, \ldots, N-1. \tag{1.3.11}$$

Then, if ϕ has a Lipschitz constant L with respect to the second argument, we deduce that

$$\|e_{n+1}\| \leq (1 + \Delta t L)\|e_n\| + \Delta t\|T\|, n = 0, \ldots, N-1, \tag{1.3.12}$$

with $\|T\| = \max\limits_{j=0,\ldots,N-1} \|T_j\|$.

$$\|e_n\| \leq (1 + \Delta t L)^n \|e_0\| + \|T\| \frac{(1 + \Delta t L)^n - 1}{L}. \tag{1.3.13}$$

We proved the following convergence result:

Theorem 1.3.1 *Let $u_n, n = 0, \ldots, N$, be the numerical approximation to the solution of (1.1.1) defined by the explicit method (1.3.2). Let us suppose that ϕ is continuous and satisfies a Lipschitz condition with respect to its second argument in $\mathcal{R}$($\mathcal{R}$ is defined in Picard's theorem) with Lipschitz constant L. If $\|u_n - u_0\| \leq \delta, n = 1 \ldots, N$, then*

$$\|e_n\| \leq e^{n\Delta t L}\|e_0\| + \|T\| \frac{e^{n\Delta t L} - 1}{L}, n = 1, \ldots, N, \tag{1.3.14}$$

where $\|T\| = \max\limits_{j=0,\ldots,N-1} \|T_j\|$.

As a corollary of the Theorem 1.3.1, we immediately conclude that, under the assumption of the Theorem 1.3.1, if the one-step method (1.3.2) is consistent, then it is also convergent. Furthermore, if the order of the truncation error is p, then the order or the global error is at least also p.

The quality of the estimate (1.3.14) strongly depends on $e^{n\Delta t L}\|e_0\|$. This last quantity is related with the propagation in time of $\|e_0\|$. If the initial error is very small, the previous quantity should remains bounded and also small.

The proved result can be extended to the one-step method (1.3.1). In fact, if we assume that

$$\|\phi(t_n, u_n, u_{n+1}, \Delta t) - \phi(t_n, \tilde{u}_n, \tilde{u}_{n+1}, \Delta t)\| \leq L\big(\|u_n - \tilde{u}_n\| + \|u_{n+1} - \tilde{u}_{n+1}\|\big),$$

then we obtain for the error the estimate

$$\|e_{n+1}\| \leq \frac{1 + \Delta t L}{1 - \Delta t L}\|e_n\| + \frac{\Delta t}{1 - \Delta t L}\|T_n\|, \tag{1.3.15}$$

provided that $1 - \Delta t L > 0$.

The inequality (1.3.15) implies

$$\|e_{n+1}\| \leq e^{2(n+1)\Delta t \frac{L}{1-\Delta t_0 L}}\|e_0\| + \max_{0 \leq i \leq N-1} \|T_i\| \frac{1}{2L}\Big(e^{2n\Delta t \frac{L}{1-\Delta t_0 L}} - 1\Big), \tag{1.3.16}$$

for $\Delta t \in (0, \Delta t_0]$ with $1 - L\Delta t_0 > 0$. From (1.3.16) the convergence estimate

$$\|e_{n+1}\| \leq \max_{0 \leq i \leq N-1} \|T_i\| \frac{1}{2L}\Big(e^{2n\Delta t \frac{L}{1-\Delta t_0 L}} - 1\Big), \tag{1.3.17}$$

is deduced.

1.3.3 Stability

In the analysis of numerical methods for IVP the term stability, like in the context of the theory of IVP, is a sort of collective noun for properties about the perturbation sensitivity during the evolution in time. In the widest sense of the word, stability should mean that the difference between any two solutions defined by (1.3.1) for the same step size remains bounded in same suitable defined way. A stronger concept is contractivity which means that the mentioned difference will not increase in time. Stability in the aforementioned sense allows an increase in this difference but not beyond any bound. Clearly that contractivity implies stability.

The one-step method (1.3.1) is called C-stable if real numbers C and Δt_0 exist such that

$$\|u_{n+1} - \tilde{u}_{n+1}\| \leq (1 + \Delta t C)\|u_n - \tilde{u}_n\|, \forall \Delta t \in (0, \Delta t_0] \tag{1.3.18}$$

where u_{n+1} and $\tilde{u}_{n+1}$ are defined by (1.3.1).

In (1.3.18) the constant C is Δt independent. The increase on the initial perturbations $\|u_0 - \tilde{u}_0\|$ remains bonded by $e^{(T-t_0)C}\|u_0 - \tilde{u}_0\|$. The magnitude of the bound depends on C.

In the particular case that the upper bound (1.3.18) holds with a positive constant $C_c \leq 1$ replacing $1 + \Delta t C$ then the method (1.3.1) is called contractive. If $C_c < 1$, then we have the so called strictly contractive.

In the definition of the previous stability properties, the step size is restricted by Δt_0. If there isn't any restriction to the step size then we say that the method (1.3.1) is unconditionally C-stable or unconditionally contractive.

Another very important concept in the context on numerical methods for IVP is the absolute stability. This concept is introduced considering the test equation

$$u'(t) = \lambda u(t),\ \lambda \in \mathbb{C},\ u(t_0) = u_0. \tag{1.3.19}$$

Even though this equation is very simple, it is used as a model to predict the stability behaviour of numerical methods for general nonlinear systems.

Let us consider (1.3.1) applied to (1.3.19). We get

$$u_{n+1} = R(z)u_n,\ z = \Delta t\lambda, \tag{1.3.20}$$

where $R : \mathbb{C} \to \mathbb{C}$ is a polynomial or rational function. R is called stability function of the method (1.3.1). If R is such that $|R(z)| \leq 1$, then we say that the method is absolutely stable at z. Of course that, if the method is absolutely stable at z, then $|u_{n+1}| \leq |u_n|$ for any pair $(\Delta t, \lambda)$ such that $\Delta t\lambda = z$.

The set $\mathcal{S} = \{z \in \mathbb{C} : |R(z)| \leq 1\}$ is called the region of absolute stability. If $\mathcal{S} \subseteq \mathbb{C}_- = \{z \in \mathbb{C} : Re z \leq 0\}$, then we have unconditional absolute stability of the method when applied to the test equation (1.3.19), which means that we have absolute stability without any condition on the step size Δt. In this case the method is said to be A-stable.

Example 8 *The stability function of the θ-method is given by*

1. $\theta = 0$

$$R(z) = 1 + z;$$

2. $\theta = 1$

$$R(z) = \frac{1}{1-z};$$

3. $\theta \in (0,1)$

$$R(z) = \frac{1+(1-\theta)z}{1-\theta z}.$$

Consequently, the stability region are

1. $\theta = 0$: *the circle with center* $(-1,0)$ *and radius* 1,
2. $\theta = 1$: *the complement of the open circle with center* $(0,1)$ *and radius* 1*;*
3. $\theta = \frac{1}{2}$*: the semi-plan* $Rez \leq 0$.

1.3.4 The θ-Method

Stability Analysis

Let us consider θ-method applied to the linear IVP

$$u'(t) = Au(t) + g(t), t > 0, u(t_0) = u_0. \tag{1.3.21}$$

The application of the θ-method to the last problem enable us to obtain

$$u_{n+1} = R(\Delta tA)u_n + (I - \theta\Delta tA)^{-1}\Delta t g_{n+\theta}, \tag{1.3.22}$$

with

$$R(\Delta tA) = (I - \theta\Delta tA)^{-1}(I + (1-\theta)\Delta tA),$$

and

$$g_{n+\theta} = (1-\theta)g(t_n) + \theta g(t_{n+1}).$$

If we consider two numerical approximations defined by the θ-method with different initial conditions u_0 and $\tilde{u}_0$, we get for $w_{n+1} = u_{n+1} - \tilde{u}_{n+1}$ the following equation

$$w_{n+1} = R(\theta A)^{n+1} w_0. \tag{1.3.23}$$

Hence, the power $R(\theta A)^{n+1}$ determines the growth of the initial error w_0. We note that

$$(I - \theta\Delta tA)^{-1} = \sum_{j=0}^{\infty} (\theta\Delta tA)^j$$

provided that $\theta\Delta\|A\|$ is sufficiently small. Then

$$R(\Delta tA) = I + \Delta tA + O(\Delta t^2),$$

and thus

$$\|R(\Delta tA)\| \leq (1 + C\Delta t), \tag{1.3.24}$$

for $\Delta t\|A\|$ small enough and C depending on $\|A\|$. From (1.3.23) and (1.3.24) we conclude the C-stability of the θ-method.

As $\|A\|$ can be very large, the estimate (1.3.24) is then useless. Better bounds can be deduced by invoking the stability region.

Theorem 1.3.2 *Suppose that $\|.\|$ is an absolutely vector norm and $A = MDM^{-1}$ where $cond(M) \leq k$ and D is a diagonal matrix, $D = diag(\lambda_j)$. If $\Delta t\lambda_j \in \mathcal{S}$, for all j, then*

$$\|R(\Delta tA)^n\| \leq k, \ \forall n. \tag{1.3.25}$$

Proof: From the fact $A = MDM^{-1}$ it can be easily seen that

$$R(\Delta tA) = MR(\Delta tD)M^{-1}$$

and, therefore,

$$R(\Delta tA) = MR(\Delta tD)M^{-1}.$$

As $R(\Delta tA) = Diag(R(\Delta t\lambda_j)$ we conclude the proof using the fact that the vector norm is absolute.[1]

■

Theorem 1.3.2 enable us to establish C stability. In fact, if $k \leq 1 + \Delta tC$ for some constant C, then the θ-method is C-sable.

Considering normal matrices in the last result we obtain the corollary below mentioned.

Corollary 1 *Suppose that A is a normal matrix. If $\Delta t\lambda_j \in \mathcal{S}$, for all j, then*

$$\|R(\Delta tA)\|_2 \leq 1.$$

■

For a large number of applications, Theorem 1.3.2 gives a sufficient condition for stability. However, for non diagonalizable matrices or diagonalizable matrices such that $cond(M)$ is large the mentioned result does not allow to conclude stability. In what follows we establish a result based on the logarithmic norms, which can be very helpful.

Theorem 1.3.3 *Suppose that the vectorial norm is induced by an inner product $< ., . >$. If*

$$Re < Av, v > \leq \omega\|v\|^2, \ \forall v \in \mathbb{C}^m, \tag{1.3.26}$$

then

$$\|R(\Delta tA)\| \leq \sup_{Rez \leq \Delta t\omega} |R(z)| \leq \max\big(|R(\Delta t\omega)|, |R(\infty)|\big). \tag{1.3.27}$$

provided that

$$1 - \omega\theta\Delta t > 0. \tag{1.3.28}$$

Proof: Let $Z = \Delta tA$ and consider $w_1 = R(Z)w_0$ which can be rewritten as

$$w_1 = (I + (1-\theta)Z)(I - \theta Z)^{-1}w_0 = u + (1 - \theta Z)u$$

with $u = (I - \theta Z)^{-1}w_0$. We also have $w_0 = u - \theta Zu$. Let $v = \dfrac{u}{\|u\|}$. It is easy to show that

$$\frac{\|w_1\|^2}{\|w_0\|^2} = \frac{1 + 2(1-\theta)Re < Zv, v > +(1-\theta)^2\|Zv\|^2}{1 - 2\theta Re < Zv, v > +\theta^2\|Zv\|^2}. \tag{1.3.29}$$

[1] A norm $\|.\|$ in $\mathbb{R}^n$ is said absolute norm if for any two vectors u, such that $|u_i| = |v_i|$, $\|u\| = \|v\|$. In this case the norm of a diagonal matrix is the maximum of the absolute value of the diagonal components.

The quotient (1.3.29) admits the representation

$$\frac{\|w_1\|^2}{\|w_0\|^2} = |R(\xi)|^2,\ \xi = Re < Zv, v > +i\sqrt{\|Zv\|^2 - Re < Zv, v >^2}. \tag{1.3.30}$$

Since $Re\xi = Re < Zv, v > \leq \Delta t\omega$ it follows that $\|R(Z)\|$ is bounded by $\sup\{|R(z)| : Rez \leq \Delta t\omega\}$. Using the Theorem of Maximum Modulus[2], we obtain

$$\|R(Z)\| \leq max\{|R(\Delta t\omega), |1 - \frac{1}{\theta}|\}.$$

provided that $1 - \Delta t\theta\omega > 0$.

■

The condition (1.3.26) is equivalent to $\mu[A] \leq \omega$. As a consequence, using the upper bound for the logarithmic norm, we establish an upper bound to $\|R(\Delta tA)\|$. Fixing Δt_0, such that

$$\max\left(|R(\Delta t\omega)|, |1 - \frac{1}{\theta}|\right) \leq 1,\ \Delta t \in (0, \Delta t_0]$$

we conclude that the θ-method is contractive. Furthermore, if

$$\max\left(|R(\Delta t\omega)|, |1 - \frac{1}{\theta}|\right) \leq 1 + C\Delta t,\ \Delta t \in (0, \Delta t_0]$$

we conclude the C-stability of the θ-method.

Corollary 2 *If* $\mu[A] \leq 0$ *and* $\theta \geq \frac{1}{2}$ *then*

$$\|R(\Delta tA)\| \leq 1.$$

Proof: We only point out that we have $\|R(\Delta tA)\| \leq |1 - \frac{1}{\theta}| \leq 1$.

■

Immediately, if $\theta \geq \frac{1}{2}$, we conclude that the θ-method is unconditionally contractive. For $\theta < \frac{1}{2}$ we need to impose a restriction on the step size Δt in order to get stability.

Theorem 1.3.3 is valid just for norms induced by inner products. For a general norm we have the following characterization valid for implicit Euler's method only.

Theorem 1.3.4 *Let* A *in* $\mathbb{C}^m \times \mathbb{C}^m$ *and* $\omega \in \mathbb{R}$. *Then,*

$$\mu[A] \leq \omega \ \textit{if and only if}\ \|(I - \Delta tA)^{-1}\| \leq \frac{1}{1 - \omega\Delta t}$$

provided that $1 - \omega\Delta t > 0$.

[2]Theorem of the Maximum Modulus: Let ϕ be a non-constant complex function which is analytic on a set $\mathcal{D} \subset \mathbb{C}$ and continuous on $\overline{\mathcal{D}}$. Then

$$\max_{\overline{\mathcal{D}}} |\phi(z)| = \max_{\partial\mathcal{D}} |\phi(z)|,$$

where $\partial\mathcal{D}$ denotes the boundary of $\mathcal{D}$.

Proof: Suppose that $\mu[A] \leq \omega$. As we have

$$\mu[B] \geq -\frac{\|Bv\|}{\|v\|},\ v \neq 0,$$

considering $w_1 = (I - \Delta tA)^{-1}w_0$, $w_0 = (I - \Delta tA)w_1$, and then, taking $B = \Delta tA - I$, we obtain

$$\|w_0\| \geq -\mu[\Delta tA - I]\|w_1\| \geq (1 - \Delta t\omega)\|w_1\|.$$

Thus, if $1 - \Delta t\omega > 0$, we deduce that $I - \Delta tA$ is nonsingular and

$$\|(I - \Delta tA)^{-1}\| \leq (1 - \Delta t\omega)^{-1}.$$

On the other hand, assuming that the latter inequality holds for Δt small enough, then using the series expansion $(I - \Delta tA)^{-1} = \sum_{j=0}^{\infty}(\Delta tA)^j$ which holds if $\Delta t\|A\| < 1$, it follows that

$$\|I + \Delta tA\| \leq \|(I - \Delta tA)^{-1}\| + O(\Delta t^2) \leq \frac{1}{1 - \Delta t\omega} + O(\Delta t^2).$$

Consequently, we obtain $\mu[A] \leq \omega$.

■

Theorem 1.3.4 stands that, if $\mu[A] \leq \omega$ and Δt_0 is such that $1 - \Delta t_0\omega > 0$, then we obain

$$\|u_{n+1} - \tilde{u}_{n+1}\| \leq \big(1 + \frac{\omega}{1 - \Delta t_0\omega}\Delta t\big)\|u_n - \tilde{u}_n\|,$$

which means that the implicit Euler's method is C-stable.

For nonlinear problems we have the following extension:

Theorem 1.3.5 *Let $\|.\|$ be a given norm. Suppose that*

$$\mu[\frac{\partial F}{\partial v}(t_{n+1}, \xi)] \leq \omega.$$

Then for any two numerical approximations for the solution of (1.1.1) defined by the implicit Euler's method we have

$$\|u_{n+1} - \tilde{u}_{n+1}\| \leq \frac{1}{1 - \Delta t\omega}\|u_n - \tilde{u}_n\| \tag{1.3.31}$$

provided that $1 - \omega\Delta t > 0$.

Proof: By the Mean Value Theorem, we have, for $w_{n+1} = u_{n+1} - \tilde{u}_{n+1}$,

$$\Big(I - \Delta t \int_0^1 \frac{\partial J}{\partial v}(t_{n+1}, \sigma u_{n+1} + (1 - \sigma)\tilde{u}_{n+1})\, d\sigma\Big)w_{n+1} = w_n. \tag{1.3.32}$$

Let $M(t_{n+1})$ denotes $\displaystyle\int_0^1 \frac{\partial J}{\partial v}(t_{n+1}, \sigma u_{n+1} + (1 - \sigma)\tilde{u}_{n+1})\, d\sigma$.

As

$$\frac{\|w_n\|}{\|w_{n+1}\|} \geq -\mu[-I + \Delta tM(t_{n+1})],$$

and, by Proposition 1 and Theorem 1.1.5, we get

$$\mu[-I + \Delta t M(t_{n+1}) \leq -1 + \Delta\mu[M(t_{n+1})] \leq -1 + \Delta t\omega.$$

So, we deduce

$$\frac{\|w_n\|}{\|w_{n+1}\|} \geq (1 - \Delta t\omega)$$

which allow us to conclude (1.3.31).

■

Convergence

The Linear Case

Let us start by considering the θ method applied to (1.3.21). Let T_n be the truncation error. Then for the error e_n holds the representation

$$e_{n+1} = e_n + \Delta t(1-\theta)Ae_n + \theta\Delta tAe_{n+1} + \Delta tT_n.$$

It follows that

$$e_{n+1} = R(\Delta tA)e_n + (I - \Delta t\theta)A)^{-1}\Delta tT_n. \tag{1.3.33}$$

Supposing that

$$\|R(\Delta tA)^n\| \leq k, n\Delta t \leq T - t_0, \tag{1.3.34}$$

from (1.3.33) we obtain

$$\|e_n\| \leq k\|e_0\| + k\Delta t \sum_{j=0}^{n-1} \|(I - \theta\Delta tA)^{-1}\|\|T_j\|. \tag{1.3.35}$$

Using the definition of R, we have

$$(I - \Delta t\theta A)^{-1} = \theta R(\Delta tA) + (1-\theta)I$$

[3] and then $(I - \Delta t\theta A)^{-1}$ is bounded if we can bound $R(\Delta tA)$. As we are assuming that (1.3.34) holds, we deduce that exists $(I - \Delta t\theta A)^{-1}$ and its norm is bounded by some constant C. Consequently, from (1.3.35), we establish

$$\|e_n\| \leq k\|e_0\| + kC(T - t_0)\|T\|, \tag{1.3.36}$$

where $\|T\|$ represents the maximum of the truncation error. For $\theta = 1/2$, $\|T\| \leq C'\Delta t^2$ and $\|T\| \leq C'\Delta t$ in the other cases. Thus, using the fact $e_0 = 0$, we obtain

$$\|e_n\| \leq C^*\Delta t^p,\ n\Delta t \leq T - t_0, \tag{1.3.37}$$

where C^* stands for the product of constants that arise above.

[3] $\theta R(\Delta tA) = \theta I + \theta(I - \theta\Delta tA)^{-1}\Delta tA$
$= \theta I - (I - \theta\Delta tA)^{-1}(I - \theta\Delta tA) + (I - \theta\Delta tA)^{-1}$

The estimate (1.3.37) establishes the convergence of the θ-method when applied to the IVP (1.3.21). On the proof of this convergence, the stability inequality (1.3.34) has an important role. Obviously, if $\|R(\Delta tA)\| \leq \hat{k}$, then $k := \hat{k}^n$. So, in $\hat{k} >> 1$ then k is very large and then the estimate (1.3.37) does not give helpful information. If $\hat{k} = 1 + \Delta tC$ for some C, we have C-stability, and then in (1.3.37) $k = e^{C(T-t_0)}$ which is bounded.

The Nonlinear Case

We consider in what follows the application of θ-method to the IVP (1.1.1). Let T_n be the truncation error and e_n the global error. For these two errors we have

$$\begin{aligned} e_{n+1} &= e_n + (1-\theta)\Delta t\big(F(t_n, u(t_n)) - F(t_n, u_n)\big) \\ &\quad +\theta\Delta t\big(F(t_{n+1}, u(t_{n+1})) - F(t_{n+1}, u_{n+1})\big) + \Delta t T_n, \end{aligned}$$

and, using the Mean Value Theorem, we obtain

$$\begin{aligned} e_{n+1} &= e_n + (1-\theta)\Delta t \int_0^1 \frac{\partial F}{\partial v}(t_n, \sigma u(t_n) + (1-\sigma)u_n)\, d\sigma (u(t_n) - u_n) \\ &\quad +\theta\Delta t \int_0^1 \frac{\partial F}{\partial v}(t_{n+1}, \sigma u(t_{n+1}) + (1-\sigma)u_{n+1})\, d\sigma (u(t_{n+1}) - u_{n+1}) + \Delta t T_n. \end{aligned}$$

Considering the notation

$$M(t_n) := \int_0^1 \frac{\partial F}{\partial v}(t_n, \sigma u(t_n) + (1-\sigma)u_n)\, d\sigma,$$

we get for the error e_{n+1}, e_n the following equality

$$(I - \theta\Delta t M(t_{n+1}))e_{n+1} = (I + (1-\theta)\Delta t M(t_n))e_n + \Delta t T_n, \tag{1.3.38}$$

which can be rewritten in the following form

$$\tilde{e}_{n+1} = (I + (1-\theta)\Delta t M(t_n))(I - \theta\Delta t M(t_n))^{-1}\tilde{e}_n + \Delta t T_n, \tag{1.3.39}$$

where

$$\tilde{e}_{n+1} = (I - \theta\Delta t M(t_{n+1}))e_{n+1}, \tilde{e}_n = (I - \theta\Delta t M(t_n))e_n.$$

If we assume that

$$\mu[\frac{\partial F}{\partial v}(t, v)] \leq 0, \forall t, v,$$

then, for vectorial norms induced by inner products and for $\theta \geq \frac{1}{2}$, we have

$$\|R(\Delta t M(t_n))\| \leq 1,$$

which implies

$$\|\tilde{e}_{n+1}\| \leq \|\tilde{e}_n\| + \Delta t\|T_n\|. \tag{1.3.40}$$

From (1.3.40) we get

$$\|\tilde{e}_{n+1}\| \leq \|\tilde{e}_0\| + \sum_{j=0}^{n} \Delta t \|T_j\| \leq \|\tilde{e}_0\| + n\Delta t \|T\|, \tag{1.3.41}$$

where $\|T\|$ is defined as before.

Taking into account that $(I - \theta \Delta t M(t_n))^{-1} = \theta R(\Delta t M(t_n)) + (1-\theta)I$ and $\|R(\Delta t M(t_n))\| \leq 1$, we deduce that

$$\|(I - \Delta t \theta M(t_n))^{-1}\| \leq 1.$$

This upper bound implies that

$$\|e_{n+1}\| = \|(I - \Delta t \theta M(t_n))^{-1} \tilde{e}_{n+1}\| \leq \|\tilde{e}_{n+1}\|,$$

and then, from (1.3.41), the following estimate

$$\|e_{n+1}\| \leq \|(I - \theta \Delta t M((t_0))e_0\| + n\Delta t \|T\| \tag{1.3.42}$$

can be established. Finally, another upper bound can be obtained if we take the estimate $\|M(t_0)\| \leq L$, where L represents the Lipschitz constant

$$\|e_{n+1}\| \leq (1 + L\Delta t)\|e_0\| + n\Delta t \|T\|. \tag{1.3.43}$$

1.4 Stiff systems

The problems called stiff are diverse and it is rather cumbersome to give a rigorous mathematic definition of stiffness. Consequently, in the literature, there are various definitions. Hairer and Wanner, in their book [14], wrote that *While the intuitive meaning of stiff is clear for all specialists, much controversy is going on about its correct mathematical definition.* They agree that the most pragmatical opinion is also historically the first one: *stiff equations are equations where certain implicit methods perform better, usually tremendously better, than the explicit ones.* This idea of stiff equation is based on the use of numerical methods.

We will introduce in what follows the concept of stiff equations, trying not to use the performance of some numerical methods. The essence of stiffness is given by Dekker and Verwer in their book [4]: *The essence of stiffness is that the solution to be computed is slowly varying but that perturbations exist which are rapidly damped.*

Let us consider some illustrative examples of stiffness.

Example 9 *Let F be a slowly varying smooth function and λ be a parameter such that $\lambda << 0$. Let u be the solution of the IVP*

$$u'(t) = \lambda u(t) + F'(t) - \lambda F(t),\ t > 0, u(0) = u_0. \tag{1.4.1}$$

The solution of the IVP (1.4.1) is given by

$$u(t) = F(t) + e^{\lambda t}(u_0 - F(0)).$$

As $\lambda << 0$, after a very short time distance, the behaviour of the term $e^{\lambda t}(u_0 - F(0))$ does not influence the behaviour of u. Nevertheless, for short time distance, u is determined by the

mentioned term. Such term is called transient term, stiff component of u or strongly varying component of u. The term $F(t)$ is called nontransient, smooth component or slowly varying component of u.

Let us now consider the integration of (1.4.1) over the time interval $[t_n, t_{n+1}]$ of length Δt

$$u(t_{n+1}) = e^{\lambda \Delta t}(u(t_n) - F(t_n)) + F(t_{n+1}).$$

This expression show that if there is an perturbation of the smooth component, such perturbation is rapidly damped.

Let us consider now the Explicit and the Implicit Euler's methods

$$u_{n+1} = (1 + \lambda \Delta t)(u_n - F(t_n)) + F(t_n) + \Delta t F'(t_n),$$

$$u_{n+1} = (1 - \lambda \Delta t)^{-1}(u_n - F(t_n)) + (1 - \lambda \Delta t)^{-1}(F(t_n) + \Delta t F'(t_{n+1}) - \lambda \Delta t F(t_{n+1})).$$

If we take u_n as a perturbation of $F(t_n)$, then, in the Explicit method, the perturbation $u_n - F(t_n)$ is damped if $\Delta t \in (0, \frac{2}{-\lambda})$, which implies a severe restriction on the time step size. Otherwise, the implicit method simulates the behaviour of the continuous model for all Δt.

Regarding the approximation of $F(t_{n+1})$ by the corresponding terms of both methods, for the explicit method the term $F(t_n) + \Delta t F(t_n)$ is acceptable approximation for $F(_{n+1})$ with Δt larger than the imposed by the stability behaviour. This situation is typical when explicit method is applied to a stiff problem. Of course that for the implicit method we have $(1 - \lambda \Delta t)^{-1}(F(t_n) + \Delta t F'(t_{n+1}) - \lambda \Delta t F(t_{n+1})) - F(t_{n+1}) \to 0$.

In Figures 4 and 5 we plot the numerical solutions obtained with the explicit and implicit Euler methods. The implicit method performs tremendously better than the explicit one. When λ decreases drastically then the restriction for the time stepsize is in fact very severe as we can see in Figure 10

■

Example 10 *Let us consider again the IVP defined in Example 2. The solution of such problem is given by*

$$u(t) = \begin{bmatrix} e^{dt} & \dfrac{e^{dt} - e^{-\epsilon^{-1}t}}{1 + d\epsilon} \\ 0 & e^{-\epsilon^{-1}t} \end{bmatrix} u_0.$$

The second component $e^{-\epsilon^{-1}t}$ of the solution dies after a short period. This component, the transient one, determines the solution only for small times. After the transient time, the solution is determined by the smooth component e^{dt}. This problem is considered also stiff. We point out that the problem is considered stiff only on the nontransient phase.

We consider now the Explicit and the Implicit Euler's methods defined by

$$u_{n+1} = \begin{bmatrix} 1 + \Delta t d & \Delta t \epsilon^{-1} \\ 0 & 1 - \Delta t \epsilon^{-1} \end{bmatrix} u_n,$$

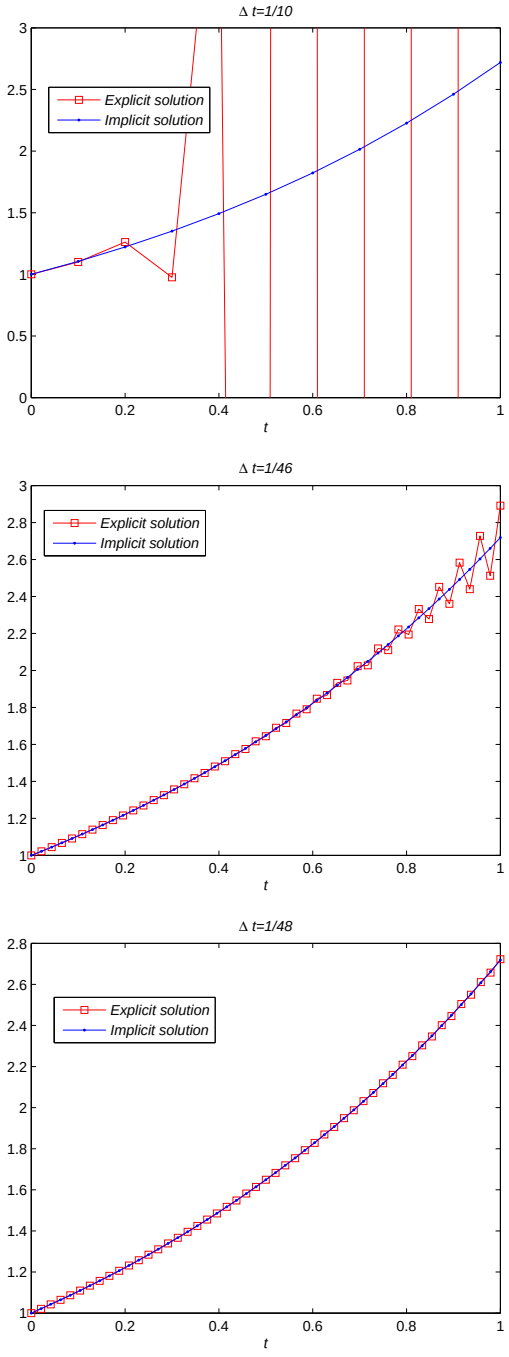

Figure 4: Numerical approximations obtained with the explicit Euler and implicit Euler methods for $F(t) = e^t$ and $\lambda = -100$.

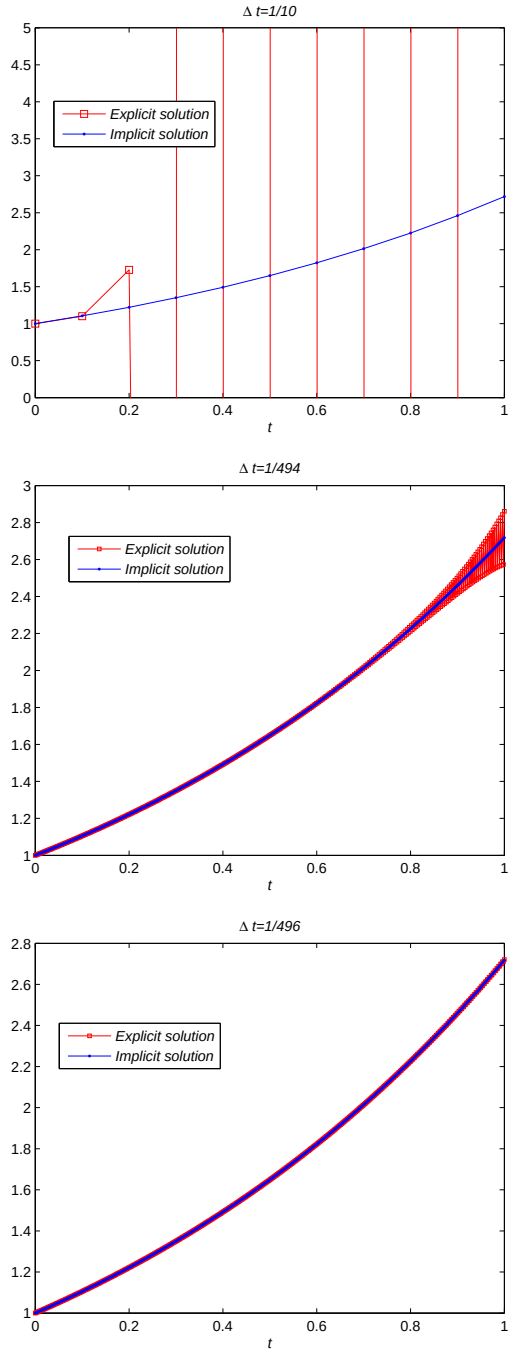

Figure 5: Numerical approximations obtained with the explicit Euler and implicit Euler methods for $F(t) = e^t$ and $\lambda = -1000$.

$$u_{n+1} = \begin{bmatrix} \dfrac{1}{1-\Delta t d} & \dfrac{\Delta t \epsilon^{-1}}{(1-\Delta t d)(1+\Delta \epsilon^{-1}} \\ 0 & \dfrac{1}{1+\Delta t \epsilon^{-1}} \end{bmatrix} u_n.$$

On the transient phase both methods perform satisfactorily because they are computing an approximation to $e^{-\Delta t \epsilon^1}$ *using* $1-\Delta t\epsilon^{-1}$ *and* $1-\dfrac{\Delta t}{\epsilon+\Delta t}$. *After this phase, the transient component died and the behaviour of the solution is determined by* e^{dt}. *Obviously, in this phase, the explicit method becomes inefficient .*

■

We next discuss the concept of stifness for the general linear problem

$$u'(t) = Au(t) + r(t), t > 0, u(t_0) = u_0, \tag{1.4.2}$$

where $A \in \mathbb{R}^m \times \mathbb{R}^m$ and r denotes a source smooth term. The obvious way to define stiffness for the linear system is by using the nature of the eigenvalues $\lambda_i, i = 1, \ldots, m$. The linear IVP (1.4.2) is stiff if

1.
$$\exists \lambda_i : Re\lambda_i << 0, \tag{1.4.3}$$

2.
$$\exists \lambda_i : |\lambda_i| \text{is small when compared with the modulus of the eigenvalues satisfying the first requirement,} \tag{1.4.4}$$

3.
$$\nexists \lambda_i : Re\lambda_i >> 0, \tag{1.4.5}$$

4.
$$\nexists \lambda_i : Im\lambda_i >> 0 \text{ unless } Re\lambda_i << 0. \tag{1.4.6}$$

Of course that if (1.4.2) is stiff according to the last definition then it is also stiff in the sense introduced before.

The previous concept of stiffness allows now the introduction the stiffness for the nonlinear IVP (1.1.1). If the eigenvalues of the Jacobian of F, JF at $u(t)$ for $t = \tilde{t}$, satisfies (1.4.3)-(1.4.6), then we say that (1.1.1) is stiff at $\tilde{t}$. Let us suppose that we perturb $u(\tilde{t})$ to $\tilde{u}(\tilde{t})$. As, for $t > \tilde{t}$, $w(t) = \tilde{u}(t) - u(t)$ satisfies

$$w'(t) = M(t)w(t), t > \tilde{t},\ w(\tilde{t}) = \tilde{u}(\tilde{t}) - u(\tilde{t}),$$

where $M(t) = \displaystyle\int_0^1 JF(t, \sigma\tilde{u}(t) + (1-\sigma u(t)))\, d\sigma$, then, if u varies slowly for $t > \tilde{t}$, and $w(t)$ contains rapidly damped components solutions, or, even the hole solution $w(t)$ is rapidly damped, we say that the IVP (1.1.1) is stiff for $t \geq \tilde{t}$.

It should be stressed that there is not a satisfactory mathematical definition of stiffness for a nonlinear problem. Nevertheless the stiff problems from practice are well recognized. In

fact, any physical problem modelled by the IVP (1.1.1) with physical components with greatly differing time constant leads to a stiff problem. The physical components with the smallest time constants show a very rapid change and make the problem stiff. The slowly varying solution of a stiff problem is determined by the latter components.

It is well assumed that a property of the stiff problems is the presence of a large Lipschitz constant. Hence, the error estimates for the one-step methods established in Section 1.4 are not helpful for this kind of problems. Error estimates obtained using the logarithmic norms can be more convenient for stiff problems.

1.5 The Runge-Kutta Methods

The class of s stage Runge-Kutta methods was introduced in Section 1.3. We present in what follows some results on the Runge-Kutta methods

a_1	b_{11}	b_{12}	b_{13}	$\dots$	b_{1s-1}	b_{1s}
a_2	b_{21}	b_{22}	b_{23}	$\dots$	b_{2s-1}	b_{2s}
a_3	b_{31}	b_{32}	b_{33}	$\dots$	b_{3s-1}	b_{3s}
$\dots$	$\dots$					
a_s	b_{s1}	b_{s2}	b_{s3}	$\dots$	b_{ss-1}	b_{ss}
	c_1	c_2	c_3	$\dots$	c_{s-1}	c_s

1.5.1 The Order Conditions

As introduced before, the s-stage R-K methods is consistent if the truncation error T_n, defined by

$$\Delta t T_n = u(t_{n+1}) - u(t_n) - \Delta t \sum_{i=1}^{s} c_i k_i$$

with

$$k_i = F(t_n + \Delta t a_i, u(t_n) + \Delta t \sum_{j=1}^{s} b_{ij} k_j),$$

satisfies

$$\|T_n\| = O(\Delta t),\ \forall n : \Delta t \leq T - t_0.$$

If

$$\|\Delta t T_n\| = \|u(t_{n+1}) - u(t_n) - \Delta t \sum_{i=1}^{s} c_i k_i\| = O(\Delta t^{p+1}),$$

then the s-stage R-K method is consistent with order equal to p.

As we have

$$\begin{aligned} \Delta t T_n = &\sum_{m=1}^{p} \frac{1}{m!} \Delta t^m u^{(m)}(t_n) + \frac{1}{(p+1)!} \Delta t^{p+1} u^{(p+1)}(t_n + \sigma \Delta t) \\ &- \Delta t \sum_{j=1}^{s} c_i \Big(\sum_{m=1}^{p-1} \frac{1}{m!} \Delta t^m k_i^{(m)}(0) + \frac{1}{p!} \Delta t^p k_i^{(p)}(\sigma_i \Delta t) \Big), \end{aligned} \tag{1.5.1}$$

with $\sigma, \sigma_i \in [0, 1]$, the conditions for the consistency order can be established for explicit Runge-Kutta methods.

Example 11 *Let us consider the explicit* 4*-stage R-K method. Using (1.5.1) it is a tedious task to compute the conditions for the coefficients of the method such that the consistency order is* 4*. Considering that* $a_i = \sum_{j=1}^{i-1} b_{ij}$ *such conditions can be reduced to*

$$\begin{aligned}
&c_1 + c_2 + c_3 + c_4 = 1\\
&c_2a_2 + c_3a_3 + c_4a_4 = \frac{1}{2}\\
&c_2a_2^2 + c_3a_3^2 + c_4a_4^2 = \frac{1}{3}\\
&c_2a_2^3 + c_3a_3^3 + c_4a_4^3 = \frac{1}{4}\\
&c_3a_3b_{32}a_2 + c_4a_4(b_{42}a_2 + b_{43}a_3) = \frac{1}{8}\\
&c_3b_{32} + c_4b_{42} = c_2(1 - a_2)\\
&c_4b_{43} = c_3(1 - a_3)\\
&c_4(1 - a_4) = 0
\end{aligned} \tag{1.5.2}$$

(see [14], pg 133-136). Examples of fourth consistency order are given in the following Butcher tables

$$\begin{array}{c|cccc}
\frac{1}{2} & \frac{1}{2} & & & \\
\frac{1}{2} & 0 & \frac{1}{2} & & \\
1 & 0 & 0 & 1 & \\
\hline
 & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6}
\end{array}
\qquad
\begin{array}{c|cccc}
\frac{1}{3} & \frac{1}{3} & & & \\
\frac{2}{3} & -\frac{1}{3} & 1 & & \\
1 & 1 & -1 & 1 & \\
\hline
 & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8}
\end{array}$$

■

We presented until explicit R-K methods with p stages and p consistency order for $p \leq 4$. Is it possible to construct an explicit s stage R-K method with s consistency order? The answer to this question was given by several authors independently. For instance, Butcher proved the following result which can be seen in [14] (pg-173).

Theorem 1.5.1 *For* $p \geq 5$ *no explicit R-K method exists of order* p *with* $s = p$ *stages.*

■

As far as the existence of implicit R-K methods is concerned, we remark that we should compute, at each iteration, the solution of nonlinear system

$$k_i = F(t_n + a_i\Delta t, u_n + \Delta t \sum_{j=1}^{s} b_{ij}k_j),\ i = 1, \ldots, s.$$

The computation of the consistency order was made by assuming that k_i is differentiable with respect to the time step size. A sufficient condition guarantying the legitimation to compute k_i and $k_i'(0)$ is given in what follows.

Theorem 1.5.2 *Let F be continuous in the first argument and Lipschitz with constant L with respect to the second argument. If*

$$\Delta t < \frac{1}{L \max_{i=1,\ldots,s} \sum_{j=1}^{s} |b_{ij}|}, \tag{1.5.3}$$

then there exists a unique solution defined by the implicit s-stage Runge-Kutta method. Moreover, if F is p times continuously differentiable, then $k_i, i = 1, \ldots, s$, are in C^p with respect to the time step size.

Proof: Let $\mathbb{F} : \mathbb{R}^{sm} \to \mathbb{R}^{sm}$ be defined by

$$\mathbb{F}(K) = (\mathbb{F}_i(K)) = (F_i(t_n + \Delta t a_i, u_n + \Delta t \sum_{j=1}^{s} b_{ij} k_j))$$

with $K = (k_1, \ldots, k_s)$.
In $\mathbb{R}^{sm}$ we consider the norm

$$\|K\| = \max_{i=1,\ldots,s} \|k_i\|.$$

As $\mathbb{F}$ satisfies de Lipschitz condition, we have

$$\|\mathbb{F}(K_1) - \mathbb{F}(K_2)\| \leq \Delta t \max_{i=1,\ldots,s} \sum_{j=1}^{s} |b_{ij}| \|K_1 - K_2\| < \|K_1 - K_2\|,$$

and, using (1.5.3), we conclude that $\mathbb{F}$ is a contraction with respect to the last norm.

The differentiability of k_i is a consequence of the Implicit Function Theorem for

$$K - \mathbb{F}(K) = 0.$$

■

We point out that the application of the last result implies the use of a very small step size for large Lipschitz constants. Nevertheless, we apply this result in the last context.

Example 12 *The implicit 2-stage R-K method with the coefficients satisfying*

$$\begin{aligned} & c_1 + c_2 = 1 \\ & c_1 a_1 + c_2 2a_2 = \frac{1}{2} \\ & c_1 a_1^2 + c_2 a_2^2 = \frac{1}{3} \\ & c_1(b_{11}a_1 + b_{12}a_2) + c_2(b_{21}a_1 + b_{22}a_2) = \frac{1}{6} \end{aligned} \tag{1.5.4}$$

has consistency order equal to 3. An example of a third order consistency R-K method is given in the following Butcher table

$$\begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \gamma = \frac{3 \pm \sqrt{3}}{6}.$$

■

Butcher established a sufficient condition for the consistency order of a general s-stage R-K method ([4]).

Theorem 1.5.3 *If*

$$\sum_{i=1}^{s} c_i a_i^{m-1} = \frac{1}{m}, m = 1, \ldots, p, \tag{1.5.5}$$

$$\sum_{j=1}^{s} b_{ij} a_j^{m-1} = \frac{a_i^m}{m}, i = 1, \ldots, s, m = 1, \ldots, q, \tag{1.5.6}$$

and

$$\sum_{i=1}^{s} c_i a_i^{m-1} b_{ij} = \frac{c_j}{m}(1 - a_j^m), j = 1, \ldots, s, m = 1, \ldots, \ell, \tag{1.5.7}$$

with $p \leq q + \ell + 1, p \leq 2q + 2$, *then the consistency order order of the method is equal to* p.

■

Let us look to the previous conditions. The R-K methods are constructed, replacing in the Picard's sequence,

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} F(t, u(t))\, dt = u(t_n) + \Delta t \int_0^1 F(t_n + \sigma \Delta t, u(t_n + \sigma \Delta t))\, d\sigma,$$

the term $F(t_n + \sigma \Delta t, u(t_n + \sigma \Delta t))\, d\sigma$ by an approximation defined by

$$\sum_{i=1}^{s} c_i F(t_n + \Delta t a_i, u(t_n) + \Delta t \sum_{j=1}^{s} b_{ij} k_j),$$

with k_i defined above. This approximation is a particular case of the approximation rule

$$\int_0^1 g(\sigma)\, d\sigma \simeq \sum_{i=1}^{s} c_i g(a_i).$$

If this integration rule is exact for polynomials with degree less or equal to $p - 1$, we have

$$\int_0^1 \sigma^{m-1} d\sigma = \sum_{i=1}^{s} c_i a_i^{m-1}$$

and then the equality (1.5.6) holds.

Let us consider now the integral

$$\int_0^{a_i} F(t_n + \Delta t \sigma, u(t_n + \Delta t \sigma))\, d\sigma$$

approximated by

$$\sum_{j=1}^{s} b_{ij} k_j.$$

This approximation is defined using the integration rule

$$\int_0^{a_i} g(\sigma)\, d\sigma \simeq \sum_{j=1}^{s} b_{ij} g(a_j).$$

If the last approximation is of order p, i.e, the integration rule is exact for polynomials of degree less or equal to $p-1$ we have (1.5.6).

Example 13 *An example of an implicit R-K method with* 3*-stages with order* 6 *is the so called Kuntzmann-Butcher method given by the following Butcher table*

$$\begin{array}{c|cccc}
\frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\
\frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\
\frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\
\hline
 & \frac{5}{18} & \frac{4}{9} & \frac{5}{18}
\end{array}$$

■

Finally, we point out that, as the R-K methods are one-step methods and the upper bounds for the global error were established using the truncation error, at leat for IVP with a Lipschitz function F, we conclude that if the consistency order is p, then the convergence order is at least p.

1.5.2 Stability

The stability function of the general s-stage Runge-Kutta method is defined when such method is applied to the test equation (1.3.19). In this case we get

$$u_{n+1} = u_n + \Delta t [c_i]^t [k_i], \tag{1.5.8}$$

where

$$k_i = \lambda u_n + \Delta t \lambda [b_{ij}]_i [k_i]. \tag{1.5.9}$$

Then

$$(I - z[b_{ij}])[k_i] = \lambda \mathbb{1} u_n,$$

where $z = \Delta t \lambda$ and $\mathbb{1}$ denotes the vector with all components equal to 1. The last equality implies

$$[k_i] = (I - z[b_{ij}])^{-1} \mathbb{1} \lambda u_n,$$

and then

$$u_{n+1} = u_n + \Delta t \lambda [c_i]^t (I - z[b_{ij}])^{-1} \mathbb{1} u_n,$$

which is equivalent to

$$u_{n+1} = R(z)u_n$$

with the stability function $R(z)$ given by

$$R(z) = 1 + z[c_i]^t(I - z[b_{ij}])^{-1}\mathbb{1}. \tag{1.5.10}$$

We determine, in what follows, a new representation for the stability function of the s-stage Runge-Kutta method (1.5.10).

Applying the s-stage Runge-Kutta method to the test equation (1.3.19), we get (1.5.8) and (1.5.9). The numerical approximation u_{n+1} can be computed using the linear system

$$\begin{bmatrix} 1-zb_{11} & -zb_{12} & \dots & -zb_{1s} & 0 \\ -zb_{21} & 1-zb_{22} & \dots & -zb_{2s} & 0 \\ \dots & \dots & \dots & \dots & 0 \\ -zbs1 & -zb_{s2} & \dots & 1-zb_{ss} & 0 \\ -\Delta tc_1 & -\Delta tc_2 & \dots & -\Delta tc_s & 1 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \dots \\ k_s \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} \lambda u_n \\ \lambda u_n \\ \dots \\ \lambda u_n \\ u_n \end{bmatrix}. \tag{1.5.11}$$

By Cramer's rule we get

$$u_{n+1} = \frac{det \begin{bmatrix} 1-zb_{11} & -zb_{12} & \dots & -zb_{1s} & \lambda u_n \\ -zb_{21} & 1-zb_{22} & \dots & -zb_{2s} & \lambda u_n \\ \dots & \dots & \dots & \dots & \dots \\ -zbs1 & -zb_{s2} & \dots & 1-zb_{ss} & \lambda u_n \\ -\Delta tc_1 & -\Delta tc_2 & \dots & -\Delta tc_s & u_n \end{bmatrix}}{det \begin{bmatrix} 1-zb_{11} & -zb_{12} & \dots & -zb_{1s} & 0 \\ -zb_{21} & 1-zb_{22} & \dots & -zb_{2s} & 0 \\ \dots & \dots & \dots & \dots & 0 \\ -zbs1 & -zb_{s2} & \dots & 1-zb_{ss} & 0 \\ -\Delta tc_1 & -\Delta tc_2 & \dots & -\Delta tc_s & 1 \end{bmatrix}}, \tag{1.5.12}$$

which admits the representation

$$u_{n+1} = \frac{det \begin{bmatrix} 1-zb_{11} & -zb_{12} & \dots & -zb_{1s} & u_n \\ -zb_{21} & 1-zb_{22} & \dots & -zb_{2s} & u_n \\ \dots & \dots & \dots & \dots & \dots \\ -zbs1 & -zb_{s2} & \dots & 1-zb_{ss} & u_n \\ -zc_1 & -zc_2 & \dots & -zc_s & u_n \end{bmatrix}}{det(I - z[b_{ij}])} \tag{1.5.13}$$

From (1.5.13), we deduce

$$u_{n+1} = \frac{det \begin{bmatrix} 1-zb_{11}+zc_1 & -zb_{12}+zc_2 & \dots & -zb_{1s}+zc_3 & 0 \\ -zb_{21}+zc_1 & 1-zb_{22}+zc_2 & \dots & -zb_{2s}+zc_3 & 0 \\ \dots & \dots & \dots & \dots & 0 \\ -zb_{s1}+zc_1 & -zb_{s2}+zc_2 & \dots & 1-zb_{ss}+zc_3 & 0 \\ -zc_1 & -zc_2 & \dots & -zc_s & 1 \end{bmatrix}}{det(I - z[b_{ij}])} u_n.$$

We conclude that the stability function $R(z)$ defined by (1.5.10) is also given by

$$R(z) = \frac{det(I - z[b_{ij}] + z\mathbb{1}[c_i]^t)}{det(I - z[b_{ij}])}. \tag{1.5.14}$$

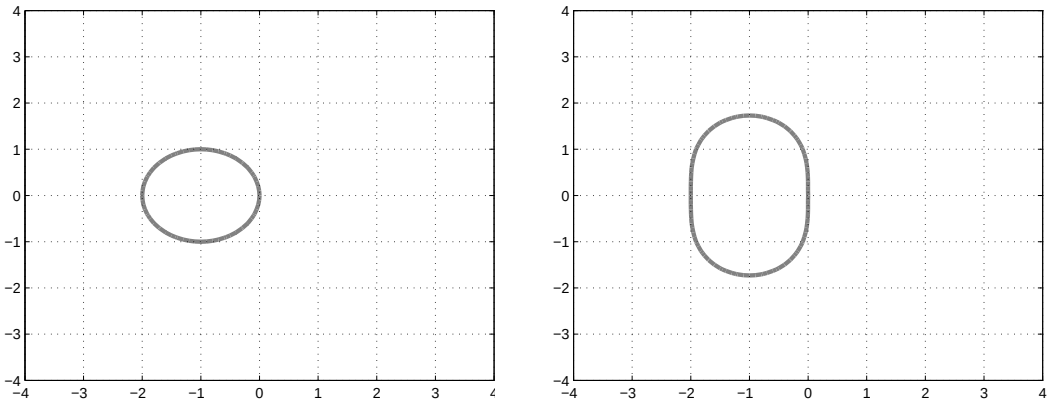

Figure 6: Boundaries of the stability regions of the Euler's method and modified Euler's method.

The last expression for the stability function of the s-stage Runge-Kutta method has, for the explicit case, an immediate consequence. In fact, if the Runge-Kutta method is explicit, then its stability function is a z polynomial of degree less or equal to s. Otherwise, the stability function is rational function with the degree of both numerator and denominator less or equal to s.

The stability region of the s-Runge-Kutta method is given by $\mathcal{S} = \{z \in \mathbf{C} : |R(z)| \leq 1\}$ with R given by (1.5.14).

Example 14 *The modified Euler's method has the stability function*

$$R(z) = 1 + z + \frac{z^2}{2}$$

and the stability region $\mathcal{S} = \{z \in \mathbf{C} : |2 + 2z + z^2| \leq 2\}$. *In Figure 6 we plot the boundaries of the stability regions of the explicit Euler's and the modified Euler's methods.*

■

Example 15 *The stability function of the explicit* 3*-stage Runge-Kutta method is given by*

$$R(z) = 1 + z\sum_{i=1}^{3} c_i + z^2(c_3(b_{31} + b_{32}) + c_2 a_2) + z^3 b_{21} b_{32} c_3.$$

If the method has third consistency order, then, using the order conditions, we obtain

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3.$$

■

Example 16 *The implicit methods*

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

respectively, the implicit midpoint and the implicit trapezoidal methods, share the stability function

$$R(z) = \frac{1+\frac{z}{2}}{1-\frac{z}{2}}.$$

As the stability region is $\mathbb{C}_-$*, we conclude that both methods are A-stables.*

■

Example 17 *The implicit method*

$$\begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

with third consistency order for $\gamma = \dfrac{3 \pm \sqrt{3}}{6}$*, has the stability function*

$$R(z) = \frac{1 + (1-2\gamma)z + (\frac{1}{2} - 2\gamma + \gamma^2)z^2}{(1-\gamma z)^2}.$$

For $\gamma \geq \dfrac{1}{4}$ *the R-K method is A-stable.*

■

The stability region can be used to compute bounds to the time step size. In fact, considering the stability region $\mathcal{S}$ and its intersection with the straight line $Imz = 0$, (R_i, R_s), which is usually called interval of absolute stability, we have absolutely stability if and only if $\lambda\Delta t \in (R_i, R_s)$. Using the interval (R_i, R_s), we know how the magnitude of the time step size should be in order to guarantee stability. For example, we have

- $(R_i, R_s) = (-2, 0)$ for the 2-stage R-K methods,
- $(R_i, R_s) = (-2.51, 0)$ for the 3-stage R-K methods,
- $(R_i, R_s) = (-2.78, 0)$ for the 3-stage R-K methods.

The stability function can be related to the consistency order. In fact, for the solution of the test equation, we may obtain

$$u(t_{n+1}) = e^z u(t_n),\ z = \lambda\Delta t. \tag{1.5.15}$$

Otherwise, if we apply a R-K method to the test equation, then

$$u_{n+1} = R(z)u_n.$$

Replacing in the last identity the approximated solution by the continuous one, we get

$$u(t_{n+1}) = R(z)u(t_n) + \hat{T}_n, \tag{1.5.16}$$

where $\hat{T}_n$ depends on the truncation error. From (1.5.15) and (1.5.16), we conclude the following result:

Theorem 1.5.4 *If the R-K method is of consistency order p then*

$$e^z = R(z) + O(z^{(p+1)}.$$

■

As a corollary we deduce:

Corollary 3 *If the explicit R-K method has order p, then*

$$R(z) = 1 + z + \frac{z^2}{2} + \cdots + \frac{z^p}{p!} + O(z^{p+1}).$$

Proof: From Theorem 1.5.4, $R(z)$ is an approximation of e^z with order $p+1$.

■

Our aim now is to establish sufficient conditions for the stability the Runge-Kutta methods: contractivity or C-stability, for linear system of ODEs. We follow the analysis of θ-method stability. Let $R(z)$ be the rational function

$$R(z) = \frac{p_0 + p_1 z + \cdots + p_s z^s}{q_0 + q_1 s + \cdots + q_s z^s}.$$

For a $m \times m$ Z matrix we define $R(Z)$ by

$$R(Z) = (p_0 I + p_1 Z + \cdots + p_s Z^s)(q_0 I + q_1 Z + \cdots + q_s Z^s)^{-1}. \tag{1.5.17}$$

Applying the s-stage Runge-Kutta method to

$$u'(t) = Au(t), u(t_0) = u_0$$

where A is a $m \times m$ matrix, we obtain

$$u_{n+1} = R(\Delta t A) u_n,$$

where $R(z)$ denotes de stability function of the R-K method. In fact, as in the scalar case, we have

$$u_{n+1} = u_n + [c_i]^t [k_i]_{i=1,\ldots,s},$$

where $[k_i]_{i=1,\ldots,s}$ is a s column vector with the i component equal to the m column vector k_i, $[c_i]^t [k_i]_{i=1,\ldots,s}$ denotes $\sum_{i=1}^{s} c_i k_i$, where $[k_i]_{i=1,\ldots,s}$ is defined by

$$[k_i]_{i=1,\ldots,s} = (I - \Delta t A[b_{ij}])^{-1} A \mathbb{1} u_n,$$

being $A[b_{ij}]$ a $s \times s$ block matrix with entries Ab_{ij}, and $\mathbb{1}Au_n$ a s column vector with components Au_n. According to the introduced notations we obtain

$$R(\Delta t A) = I + \Delta t A [c_i]^t (I - \Delta t A[b_{ij}])^{-1} \mathbb{1},$$

where $[c_i]^T$ represents a block vector whose entries are $c_i I$, $i = 1, \ldots, s$, and $\mathbb{1}$ denotes now the column block vector with s blocks being each block the m identity matrix. We remark that the last representation can be obtained formally from (1.5.10) with the convenient modifications.

Of course that, for $w_n = u_n - \tilde{u}_n$, where $u_n, \tilde{u}_n$ are defined by the initial conditions $u_0, \tilde{u}_0$, respectively, we obtain

$$\|w_{n+1}\| \leq \|R(\Delta t A)\|^{n+1} \|\|w_0\|. \tag{1.5.18}$$

The stability behaviour of the Runge-Kutta method, when applied to the linear problem, depends on the behaviour of $\|R(\Delta t A)\|^n$ when n increases.

Let us suppose that $R(z)$ is defined by (1.5.17) and A is a diagonalizable matrix, $A = MDM^{-1}$ with $D = diag(\lambda_i)$. Then

$$\|R(\Delta t A)\| \leq cond(M) \|diag(R(\Delta t \lambda_i))\|. \tag{1.5.19}$$

If $\Delta t \lambda_i \in \mathcal{S}$, then $|R(\Delta t \lambda_i)| \leq 1$, and consequently

$$\|R(\Delta t A)\| \leq cond(M). \tag{1.5.20}$$

We proved the next result:

Theorem 1.5.5 *Let us suppose that A is diagonalizable, $A = MDM^{-1}$ with $D = diag(\lambda_i)$. If $\Delta t \lambda_i \in \mathcal{S}$, then holds (1.5.20).*

■

For normal matrices we conclude contractivity.

In what follows we extend the previous result to more general matrices but just for methods with a bounded stability function $R(z)$ for $Re z \leq 0$. We assume that

$$Re < u, Au > \leq 0, \ \forall u \in \mathbb{C}^m, \tag{1.5.21}$$

where $< ., . >$ denotes the Euclidian inner product. This condition enable us to conclude that, for the continuous problem $u' = Au$, holds the following

$$\frac{d}{dt}\|u(t)\|^2 = 2Re < u(t), u'(t) > = 2Re < u(t), Au(t) > \leq 0$$

and then $\|u(t)\|^2 \leq \|u_0\|$.

Theorem 1.5.6 *(see [14], pg 168-169) If $R(z)$ is bounded for $Re z \leq 0$ and A satisfies (1.5.21) then*

$$\|R(A)\| \leq \sup_{Re z \leq 0} |R(z)|. \tag{1.5.22}$$

■

An immediate consequence of the previous result is that

$$\|R(\Delta t A)\| \leq \sup_{Re z \leq 0} |R(z)|.$$

Obviously, if the R-K method is A-stable then the R-K method is contractive with respect to the Euclidian norm.

Example 18 *Let us consider again the class of implicit methods defined in Example 17. Those methods have the stability function*

$$R(z) = \frac{1 + (1-2\gamma)z + (\frac{1}{2} - 2\gamma + \gamma^2)z^2}{(1-\gamma z)^2}.$$

For $\gamma \geq \frac{1}{4}$, *these methods are A-stables. Then, for linear problems satisfying* $< Au, u > \leq 0$, *we immediately conclude contractivity.*

■

The condition (1.5.21) can be replaced by

$$Re < u, Au > \leq \omega \|u\|^2, \ \forall u \in \mathbf{C}^m. \tag{1.5.23}$$

In fact, (1.5.23) implies

$$Re < u, (A - \omega I)u > \leq 0.$$

Consequently, taking in Theorem 1.5.6 A replaced by $\tilde{A} = A - \omega I$ and $R(z)$ replaced by $\tilde{R}(z) = R(z + \omega)$, we have

$$\|R(A)\| = \|\tilde{R}(\tilde{A})\| \leq \sup_{Rez \leq 0} |R(z + \omega)\| \leq \sup_{Rez \leq \omega} |R(z)|.$$

From (1.5.23) we deduce

$$Re < u, \Delta t A > \leq \Delta t \omega \|u\|^2, u \in \mathbf{C}^m,$$

which implies

$$\|R(\Delta t A)\| \leq \sup_{Rez \leq \Delta t \omega} |R(z)|. \tag{1.5.24}$$

Let us consider a R-K method with a stability function R. Applying this method to a linear problem such that (1.5.23) holds, then, using (1.5.24), we easily get an estimate for $\|R(\Delta t A)\|$.

We point out that the stability analysis for nonlinear case will not be considered here but can be seen for example in [4] and in [14].

1.6 Linear Multistep Methods

1.6.1 Some Examples

The Runge-Kutta methods studied in the last section are a natural improvement of the Euler's method in terms of accuracy. However, to increase the accuracy implies an increasing on the computational effort, which is measured in function of the evaluation of F at each step. The high computational cost of the Runge-Kutta methods can be avoided using more than two time level at each step. More precisely, let us consider the Picard's sequence defined using the interval $[t_{n-1}, t_{n+1}]$

$$u(t_{n+1}) = u(t_{n-1}) + \int_{t_{n-1}}^{t_{n+1}} F(t, u(t))\, dt. \tag{1.6.1}$$

If we replace the integral term by the Simpson's rule, we obtain the numerical method

$$u_{n+1} = u_n + \frac{\Delta t}{3}\big(F(t_{n-1}, u_{n-1}) + 4F(t_n, u_n) + F(t_{n+1}, u_{n+1})\big), n = 1, \ldots, N-1. \quad (1.6.2)$$

As u_{n+1} depends on u_{n-1}, u_n, this method does not belong to the class of the one-step methods and it is called 2-step method. The computation of a numerical approximation to the solution of the IVP, using the method (1.6.2), needs an approximation to u_1 which should be computed with another method.

Our aim, in the following sections, is to study the class of methods that includes the method (1.6.2).

A numerical method such that u_{n+1} depends on $u_{n+1-q}, \ldots, u_n$ is called q-step method. In this section we study linear multistep methods defined by

$$\sum_{j=0}^{q} \alpha_j u_{n+j} = \Delta t \sum_{j=0}^{q} \beta_j F_{n+j},\ n = 0, \ldots, N-q, \quad (1.6.3)$$

with $F_{n+j} = F(t_{n+j}, u_{n+j})$ and $N\Delta t = T - t_0$. The method is identified by the coefficients $\alpha_j, \beta_j, j = 0, \ldots, q$. If $\alpha_q \neq 0$ and $\beta_q = 0$ then the method is explicit. Otherwise, the method is implicit.

The computational advantage of a linear q-step method over a one s-stage R-K method can be observed for explicit and implicit methods. For the first methods only one F evaluation is needed while the R-K method needs s F evaluations. When implicit methods are used only one nonlinear system has to be solved. However, the initial values $u_1, \ldots, u_{q-1}$ needed the method q-step method (1.6.3) should be computed with a one-step R-K method.

Some classes of linear multistep methods are: the Adams methods and the Backward Differentiating Formulae (BDF). We present now these classes of methods.

1. The Adams methods: This class of methods is obtained takin in (1.6.3)

   $$\alpha_0 = \cdots = \alpha_{q-2} = 0, \alpha_{q-1} = -1, \alpha_q = 1.$$

   The Adams method is characterized by the expression

   $$u_{n+q} = u_{n+q-1} + \Delta t \sum_{j=0}^{q} \beta_j F_{n+j}.$$

   If $\beta_q = 0$, then the method is explicit, else is implicit. The explicit Adams methods are usually called Adams-Bashforth methods while the implicit ones are called Adams-Moulton methods.

2. Backward Differentiating Formulae (BDF): The methods belonging to this class are characterized by

   $$\beta_q = 1,\ \beta_{q-1} = \cdots = \beta_0 = 0$$

   which means that they are defined by

   $$\sum_{j=0}^{q} \alpha_j u_{n+j} = \Delta t F_{n+q}.$$

1.6.2 Consistency

We define the truncation error of the q-step method (1.6.3) by

$$\Delta t T_{n+q-1} = \sum_{j=0}^{q} \alpha_j u(t_{n+j}) - \Delta t \sum_{j=0}^{q} F(t_{n+j}, u(t_{n+j}). \tag{1.6.4}$$

Analogously to the one-step methods, the quantity $\Delta t T_{n+1}$ is the residual generated at t_{n+1} when the exact solution is considered in the numerical scheme. If

$$\|T_{n+q-1}\| \to 0, \Delta t \to 0, n \to \infty, n\Delta t \leq T - t_0,$$

then the method (1.6.3) is consistent with the equation $u'(t) = F(t, u(t))$. If $\|T_{n+q-1}\| \leq Const.\Delta t^p$, then the method is said to have consistency order equal to p.

We establish in what follows the conditions that imply the consistence of the q-step method (1.6.3) and the required conditions for a prescribed consistency order.

For the truncation error holds the representation

$$\Delta t T_{n+q-1} = C_0 u(t_n) + \Delta t C_1 u'(t_n) + \Delta t^2 C_2 u''(t_n) + \dots, \tag{1.6.5}$$

with

$$C_0 = \sum_{j=0}^{q} \alpha_j,\ C_i = \frac{1}{i!} \sum_{j=0}^{q} \left(\alpha_j j^i - i\beta_j j^{i-1}\right) \tag{1.6.6}$$

provided that the solution u is smooth enough. Then the method (1.6.3) has p consistency order provided that

$$\sum_{j=0}^{q} \alpha_j = 0, \quad \sum_{j=0}^{q} \alpha_j j^i = i \sum_{j=0}^{q} \beta_j j^{i-1},\ i = 1, \dots, p. \tag{1.6.7}$$

Example 19 *The 2-step method*

$$u_{n+2} = u_n + 2\Delta t F_{n+1}, n = 0, \dots, N-2,$$

has 2 consistency order. ■

Example 20 *The Adams-Bashforth methods are consistent with $u'(t) = F(t, u(t))$ and they are characterized by*

$$\alpha_q = 1, \alpha_{q-1} = -1,\ \alpha_j = \beta_q = 0, j = 0, \dots, q-2.$$

The coefficients $\beta_j, j = 0, \dots, q-1$, should be computed such that the order is optimal.

The 2-step method with 2 consistency order is defined by

$$u_{n+2} - u_{n+1} = \frac{\Delta t}{2}\big(-F_n + 3F_{n+1}\big),\ n = 0, \dots, N-2,$$

while the 3-step method with consistency order equal to 3 is defined by

$$u_{n+3} - u_{n+2} = \frac{\Delta t}{12}\big(5F_n - 16F_{n+1} + 23F_{n+2}\big), n = 0, \dots, N-3.$$

■

Example 21 *The q-step Adams-Moulton methods are characterized by*

$$\alpha_q = 1, \alpha_{q-1} = -1,\ \alpha_j = 0, j = 0, \ldots, q-2.$$

The coefficients $\beta_j, j = 0, \ldots, q$, *should be computed in such way that the method has* $q+1$ *consistency order.*

The 2*-step method*

$$u_{n+2} - u_{n+1} = \frac{\Delta t}{12}\big(- F_n + 8F_{n+1} + 5F_{n+2}\big),\ n = 0, \ldots, N-2,$$

has consistency order equal to 3 *while the* 3*-step method defined by*

$$u_{n+3} - u_{n+2} = \frac{\Delta t}{24}\big(F_n - 5F_{n+1} + 19F_{n+2} + 9F_{n+3}\big), n = 0, \ldots, N-3,$$

has consistency order equal to 4. ■

Example 22 *The BDF methods are characterized by*

$$\beta_q = 1, \beta_j = 0, j = 0, \ldots, q-1,$$

and the coefficients $\alpha_j, j = 0, \ldots, q$ *should be chosen such that the order is optimal. The* 2 *step method*

$$\frac{3}{2}u_{n+2} - 2u_{n+1} + \frac{1}{2}u_n = \Delta t F_{n+2},\ n = 0, \ldots, N-2$$

is of order 2.

■

1.6.3 Stability

The use of the method (1.6.3) requires the computation of the initial values $u_1, \ldots, u_{q-1}$ because only u_0 is given. Such values are computed using, for example, an one-step method. As those values contain numerical errors, it is very important to now how these error affects further approximations $u_n, n \geq q$. The stability behaviour of the multistep method will be considered with respect to small perturbations in the starting values.

Let u_n and $\tilde{u}_n$ be defined by the q-step method (1.6.3) with the initial values $u_i, i = 0, \ldots, q-1$ and $\tilde{u}_i, i = 0, \ldots, q-1$. The q-step method (1.6.3) is said zero-stable if

$$\|u_n - \tilde{u}_n\| \leq C \max_{i=0,\ldots,q-1} \|u_i - \tilde{u}_i\|. \tag{1.6.8}$$

We will show that the zero-stability of a multistep method can be deduced using the test equation $u' = 0$. The designation zero-stability is due to the use of $F = 0$.

The two polynomials

$$\rho(\xi) = \sum_{j=0}^{q} \alpha_j \xi^j,\ \sigma(\xi) = \sum_{j=0}^{q} \beta_j \xi^j$$

are associated with the q-step method (1.6.3) and they are called first and second characteristic polynomials.

The q-step method (1.6.3) satisfies the rout condition if the roots ξ_i of $\rho(\xi) = 0$ satisfy

$$|\xi_i| \leq 1, \forall i, |\xi_i| < 1 \text{ if } \xi_i \text{ is not simple.} \tag{1.6.9}$$

In the following result we establish, for q-step method (1.6.3), the equivalence between the root condition and the zero stability of the method. This equivalence leads to the definition of zero-stability using the root condition.

Theorem 1.6.1 *The q-step method (1.6.3) is zero-stable for any IVP $u'(t) = F(t, u(t)), t > t_0, u(t_0) = u_0$, where F satisfies the Lipschitz condition with respect to the second argument if and only if it satisfies the root condition.*

Proof: Let us suppose that the root condition is violated. We prove in what follows that the q-step method (1.6.3) is not zero-stable.

Consider the q-step method (1.6.3) applied to the IVP with $F = 0$

$$\sum_{j=0}^{q} \alpha_j u_{n+j} = 0. \tag{1.6.10}$$

Let ξ_i and ξ_ℓ be solutions of $\rho(\xi) = 0$ with multiplicity 1 and m_ℓ, respectively. Consequently, ξ_i^n and ξ_ℓ^n, $n\xi_\ell^n$, ..., $n^{m_\ell - 1}\xi_\ell^n$, are solutions of (1.6.10).[4] Then, any combination of the last solution still be a solution of (1.6.10). Lets us consider that the solution of (1.6.10) admits the representation

$$u_n = \sum_i \gamma_i \xi_i^n + \sum_\ell \xi_\ell^n \sum_j^{m_\ell - 1} \gamma_{\ell,j} n^j,$$

where the coefficients are determined by the initial conditions. We can assume that the solution of (1.6.10) is given by

$$u_n = \sum_i p_i(n) \xi_i^n, \tag{1.6.11}$$

where ξ_i is a zero of $\rho(\xi) = 0$ and p_s is a polynomial of degree one less than the multiplicity of ξ_i.[5]

If $|\xi_i| > 1$, then there are starting values for which the corresponding solution grows like $|\xi_i|^n$. If $|\xi_i| = 1$ and its multiplicity is m_i, then there are solutions growing like $n^{m_i - 1}$.

[4] If ξ_i is such that $\rho(\xi_i) = 0$ then

$$\sum_{j=0}^{q} \alpha_j \xi_i^{n+j} = \xi_i^n \rho(\xi_i) = 0.$$

Otherwise let ξ_ℓ be a zero of the first characteristic polynomial with multiplicity m_ℓ. Then

$$\sum_{j=0}^{q} \alpha_j (n+j) \xi_\ell^{n+j} = \xi_\ell^n \Big(n\rho(\xi_\ell) + \rho'(\xi_\ell) \Big) = 0.$$

Using the same procedure it can be shown that

$$\sum_{j=0}^{q} \alpha_j (n+j)^k \xi_\ell^{n+j} = 0, k = 2, \ldots, m_\ell - 1.$$

[5] Let us consider the case that the first characteristic polynomial has the roots $\xi_i, i = 1, \ldots, q$ simple. Then

Let us consider the initial values $u_0, \ldots, u_{q-1}$, which induces the unbounded solution u_n and $u_0 = 0, \ldots, u_{q-1} = 0$, which induces the null solution $\tilde{u}_n$. Then for $u_n - \tilde{u}_n$ does not hold the inequality (1.6.8).

Let us prove that the root condition is sufficient for the q-step method (1.6.3) to be zero-stable. Let u_n and $\tilde{u}_n$ be the sequences defined by the previous methods for the initial conditions $u_i, i = 0, \ldots, q-1, \tilde{u}_i, i = 0, \ldots, q-1$, respectively, and let w_n be the difference between the two defined solutions. We have[6]

---

the set of fundamental solutions $\{\xi_i^n, n = 0, 1, \ldots\}, i = 1, \ldots, q$, is such that

$$u_n = \sum_{j=1}^{q} \gamma_j \xi_j^n.$$

We introduce the new set $\{\phi_i^{(n)}, n = 0, \ldots, \}, i = 0, \ldots, q-1$ such that

$$\phi_i^{(j)} = \delta_{ij}, i, j = 0, \ldots, q-1.$$

As $\phi_i^{(n)} = \sum_{j=1}^{q} \gamma_{i,j} \xi_j^n, i = 0, \ldots, q-1,$, we deduce for the coefficient $\gamma_{i,j}, j = 1, \ldots, q$, the system

$$\sum_{j=1}^{q} \gamma_{i,j} \xi_j^\ell = \delta_{i,\ell}, \ell = 0, \ldots, q-1,$$

which is equivalent to

$$R\gamma_i = e_i, i = 0, \ldots, q-1,$$

with $R = (\xi_j^\ell)$, $\gamma_i = (\gamma_{i,1}, \ldots, \gamma_{i,q}$ and e_j denotes the unitary vector of $\mathbb{R}^q$.

Consequently we obtain for u_n the representation

$$u_n = \sum_{i=0}^{q-1} u_j \psi_j^{(n)}.$$

The procedure presented can be followed when some of the roots of the first characteristic polynomial have multiplicity greater than two.

[6]We established that for the homogeneous equation

$$\sum_{j=0}^{q} \alpha_j u_{n+j} = 0$$

holds the following

$$u_n = \sum_{j=0}^{q-1} u_j \psi_j^{(n)}.$$

The solution of the non homogeneous equation

$$\sum_{j=0}^{q} \alpha_j u_{n+j} = \phi_{n+q}$$

is established using the solution of the corresponding homogeneous equation and a particular solution of the corresponding non homogeneous one. In this case it can be show that

$$u_n = \sum_{j=0}^{q-1} u_j \psi_j^{(n)} + \sum_{j=q}^{n} \phi_j \psi_{q-1}^{n-j+q-1}, n = 0, 1, \ldots$$

where $\psi_{q-1}^{(i)} =$ for all $i < 0$ and $\phi_j = 0$ for $j < q$.

$$\sum_{j=0}^{q} \alpha_j w_{n+j} = \psi_{n+q}, n = 0, \ldots, N-q, \tag{1.6.12}$$

where

$$\psi_{n+q} = \Delta t \sum_{j=0}^{q} \beta_j \big(F(t_{n+j}, u_{n+j}) - F(t_{n+j}, \tilde{u}_{n+j})\big).$$

The solution of the difference equation (1.6.12) is given by

$$w_n = \sum_{j=0}^{q-1} w_j \psi_j^{(n)} + \sum_{j=q}^{n} \psi_{q-1}^{(n-j+q-1)} \phi_j \ n = q, \ldots, \tag{1.6.13}$$

where $\{\psi_j^{(n)}, j = 0, \ldots, q-1\}$ is the set of fundamental solutions of the homogeneous difference equation associated to (1.6.12). It can be shown that the fundamental solutions are uniformly bounded if and only if the root condition is satisfied (see [9], Theorem 6.3.2). As a consequence,

$$\|\psi_j^{(n)}\| \leq M, \|\psi_{q-1}^{(n-j+q-1)}\| \leq M.$$

Considering the last upper bounds in (1.6.13) we get

$$\|w_n\| \leq M\Big(q\|w_j\|_{max} + \sum_{j=q}^{n} \|\psi_j\|\Big), n = q, \ldots, N-q, \tag{1.6.14}$$

where

$$\max_{j=0,\ldots,q-1} \|w_j\| = \|w_j\|_{max}.$$

As F satisfies the Lipschitz condition, we obtain, for ψ_{n+q}, the upper bound

$$\|\psi_{n+q}\| \leq \Delta t L \|\beta_j\|_{max} \sum_{j=0}^{q} \|w_{n+j}\|,$$

which implies

$$\|\psi_\ell\| \leq \Delta t L |\beta_j|_{max} \sum_{j=0}^{q} \|w_{\ell-q+j}\|, \tag{1.6.15}$$

with

$$|\beta_j|_{max} = \max_{j=0,\ldots,q} |\beta_j|.$$

Taking in (1.6.14) the estimate (1.6.25) we conclude

$$\|w_n\| \leq M\Big(q\|w_j\|_{max} + \Delta t L |\beta_j|_{max} \sum_{\ell=q}^{n} \sum_{j=0}^{q} \|w_{\ell-q+j}\|\Big), n = q, \ldots, N-q. \tag{1.6.16}$$

From inequality (1.6.16) we also have

$$\big(1 - \Delta t M L |\beta_j|_{max}\big)\|w_n\| \leq Mq\|w_j\|_{max} + \Delta t M L |\beta_j|_{max} q \sum_{m=0}^{n-1} \|w_m\|\Big) \tag{1.6.17}$$

Assuming that

$$1 - \Delta t ML|\beta_j|_{max} > 0, \tag{1.6.18}$$

we should obtain an upper bound for the sequence

$$c_n \leq g_0 + k \sum_{j=0}^{n-1} c_j$$

with

$$c_n = \|w_n\|, \;\; g_0 = \max\{\frac{Mq\|w_j\|_{max}}{1 - \Delta t ML|\beta_j|_{max}}, \|w_0\|\}$$

and

$$k = \frac{\Delta t MqL|\beta_j|_{max}}{1 - \Delta t ML|\beta_j|_{max}}.$$

If $c_0 \leq g_0$ the its easy to show that

$$c_n \leq g_0(1+k)^n \leq g_0 e^{nk}.$$

Applying the last estimate we get

$$\|w_n\| \leq \frac{Mq\|w_j\|_{max}}{1 - \Delta t ML|\beta_j|_{max}} e^{n\frac{\Delta t MqL|\beta_j|_{max}}{1-\Delta t ML|\beta_j|_{max}}}.$$

The last estimate enable us to conclude the proof because implies

$$\|w_n\| \leq \frac{Mq\|w_j\|_{max}}{1 - \Delta t_0 ML|\beta_j|_{max}} e^{(T-t_0)\frac{MqL|\beta_j|_{max}}{1-\Delta t_0 ML|\beta_j|_{max}}}$$

for $\Delta t \in (0, \Delta t_0]$ with Δt_0 satisfying (1.6.18).

■

Example 23 *The Adams methods are zero-stables because* $\rho(\xi) = \xi^2 - \xi$.

■

We introduce in what follows a new concept of stability induced by the behaviour of the multistep method when applied to the scalar test equation considered on the context of A-stability of the one-step methods.

Let us define another polynomial associated to the q-step method (1.6.3): the characteristic polynomial

$$\pi(\xi) = \rho(\xi) - \Delta t \sigma(\xi) = \sum_{j=0}^{q} (\alpha_j - \Delta t \beta_j)\xi^j, \tag{1.6.19}$$

where $\rho(\xi)$ and $\sigma(\xi)$ are the first and the second characteristic polynomials associated to the multistep method (1.6.3).

When we apply the q-step method (1.6.3) to the scalar test equation we obtain

$$\sum_{j=0}^{q} (\alpha_j - z\beta_j) u_{n+j} = 0, \;\; z = \lambda \Delta t. \tag{1.6.20}$$

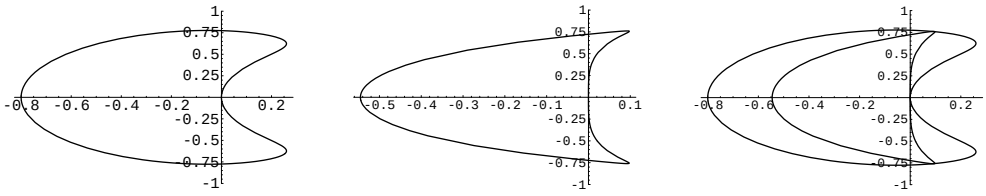

Figure 7: Stability regions of the Adams-Bashford methods with 2 and 3 steps respectively.

We associate to the last recursion the polynomial

$$\pi_z(\xi) = \rho(\xi) - z\sigma(\xi).$$

We say that the q-step method (1.6.3) satisfies the root condition at z if $\xi_i(z), i = 0, \ldots, q-1$, satisfy the condition (1.6.9).

Let $\xi_i, i \in I_1$ be the set of simple roots of the polynomial $\pi_z(\xi)$ and $\xi_i, i \in I_2$, be the set of roots of $\pi_z(\xi)$ with multiplicity $m_i, i \in I_2$. Taking into account that the solution of (1.6.20) takes the form

$$u_n = \sum_{j \in I_1} \gamma_i \xi_i(z)^n + \sum_{j \in I_2} \Big(\sum_{i=0}^{m_j - 1} \gamma_{ij} n^i \Big) \xi_j(z)^n,$$

the root condition arises as a natural requirement for the boundness of the sequence u_n.

We define the stability region of the q-step method (1.6.3), $\mathcal{S} \subset \mathbf{C}$ by

$$\mathcal{S} = \{ z \in \mathbf{C} : \pi_z \, \text{satisfies} \quad \text{the} \quad \text{root condition at} \, z \}.$$

If the stability region of q-method contains $\mathbf{C}_-$, then the method is said to be A-stable.

Let $\delta\mathcal{S}$ be the boundary of the stability region $\mathcal{S}$. Let $\xi(x)$ be a root of the polynomial $\pi_z(\xi)$. Then

$$z = \frac{\rho(\xi)}{\sigma(\xi)}.$$

The root condition is satisfied if $|\xi| < 1$ (ξ is a simple or has multiplicity greater than two) and $|\xi| \leq 1$ (is a simple root). Then $\delta\mathcal{S}$ is obtained when $\xi = e^{i\theta}$ with $\theta \in [0, 2\pi]$.

Example 24 The stability region of the Adams-Bashford methods

$$u_{n+2} - u_{n+1} = \frac{\Delta t}{2} \big(- F_n + 3F_{n+1} \big), \; n = 0, \ldots, N-2,$$

$$u_{n+3} - u_{n+2} = \frac{\Delta t}{12} \big(5F_n - 16F_{n+1} + 23F_{n+2} \big), n = 0, \ldots, N-3,$$

are plotted in Figure7

The 3-Adams-Bashford method presents a smaller stability region but a higher order (3).

Example 25 The stability region of the Adams-Moulton methods

$$u_{n+2} - u_{n+1} = \frac{\Delta t}{12} \big(- F_n + 8F_{n+1} + 5F_{n+2} \big), \; n = 0, \ldots, N-2,$$

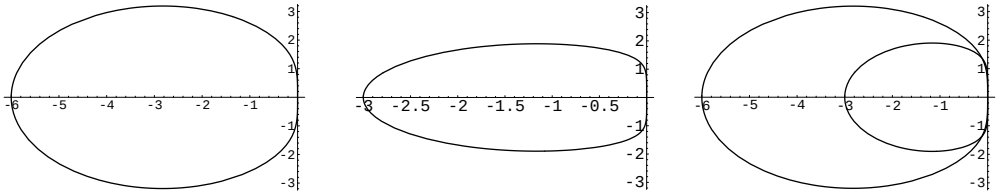

Figure 8: Stability regions of the Adams-Moulton methods with 2 and 3 steps respectively.

$$u_{n+3} - u_{n+2} = \frac{\Delta t}{24}\big(F_n - 5F_{n+1} + 19F_{n+2} + 9F_{n+3}\big), n = 0, \ldots, N-3,$$

are plotted in Figure 8

The 3-Adams-Moulton method presents a smaller stability region but a higher order (3).

It is possible to characterize the stability of the q-step method following the analysis of the one-step methods. In fact, this can be done because the equation (1.6.20) can be rewritten in an one-step form. We note that (1.6.20) is equivalent to

$$u_{n+q} = -\sum_{j=0}^{q-1} \frac{\alpha_j - z\beta_j}{\alpha_q - z\beta_q} u_{n+j},$$

which admits the following one-step form representation

$$U_{n+1} = R(z)U_n \tag{1.6.21}$$

with $U_n = (u_{n+q-1}, \ldots, u_n)^t$ and

$$R(z) = \begin{bmatrix} r_1(z) & r_2(z) & \ldots & r_{q-1}(z) & r_q(z) \\ 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \\ o & 0 & \ldots & 1 & 0 \end{bmatrix}, \quad r_i(z) = -\frac{\alpha_{q-i} - z\beta_{q-i}}{\alpha_q - z\beta_q}.$$

The matrix $R(z)$ is called companion matrix of the multistep method and we have

$$U_n = R(z)^n U_0$$

being U_0 the vectors of the initial approximations. Finally we get $z \in \mathcal{S}$ if and only if $R(z)$ is power bounded.

As final remark of this section we point out that similar results established in the one-step methods context can be also established when we apply the multistep methods to linear ODEs.

1.6.4 Convergence

The convergence of the multistep method (1.6.3) is defined considering the global error $e_n = u(t_n) - u_n$. We say that the method (1.6.3) is convergent if

$$\|e_n\| \to 0, \Delta t \to 0, n \to \infty, n\Delta t \leq T - t_0$$

provided that

$$\|e_j\| \to 0 \,\text{as}\, \Delta t \to 0, j = 0, \ldots, q-1.$$

As for one-step methods, if $\|e_n\| = O(\Delta t^p)$, then we say that the multistep method has p convergence order or is of order p.

The global error is solution of the difference equation

$$\sum_{j=0}^{q} \alpha_j e_{n+j} = \Delta t \sum_{j=0}^{q} \beta_j \Big(F(t_{n+j}, u(t_{n+j})) - F(t_{n+j}, u_{n+j}) \Big) + \Delta t T_{n+q-1}, \tag{1.6.22}$$

with the initial values

$$e_0 = 0, e_j = u(t_j) - u_j, j = 1, \ldots, q-1,$$

where the values $u_j, j = 1, \ldots, q-1$, were obtained using another method like an one-step method.

It is clear that the convergence of the multistep method depends on the behaviour of the initial values $u_j, j = 1, \ldots, q-1$.

Theorem 1.6.2 *Let us suppose that F satisfies the Lipschitz condition with respect to the second argument. A consistent multistep method is convergent if and only if it satisfies the root condition and the initial data tends to zero as the time step size goes to zero. Moreover, if the consistency order is p equal to the order of the initial errors then the multistep method is of order p.*

Proof: Suppose that the multistep method is consistent and convergent. By contradiction it can be shown that the method satisfies the root condition. In this proof, the IVP with $F = 0$ and $u(0) = 0$ should be considered and the fact $u_n \to 0$ for all set of initial values $u_i, i = 0, \ldots, q-1$, converging to zero should be used.

Let us suppose now that the method satisfies the root condition and it is consistent. As the error e_n is solution of (1.6.22), following the proof of Theorem 1.6.1, we get

$$e_n = \sum_{j=0}^{q-1} e_j \psi_j^{(n)} + \sum_{j=q}^{n} \psi_{q-1}^{(n-j+q-1)} \psi_j \,,\, n = q, \ldots \tag{1.6.23}$$

where

$$\psi_{n+q} = \Delta t \sum_{j=0}^{q} \beta_j \big(F(t_{n+j}, u(t_{n+j})) - F(t_{n+j}, u_{n+j}) \big) + \Delta t T_{n+q-1}.$$

Hence, e_n satisfies

$$\|e_n\| \le M \Big(q \|e_j\|_{max} + \sum_{j=q}^{n} \|\psi_j\| \Big) \,, n = q, \ldots, N-q, \tag{1.6.24}$$

As F is a Lipschitz function with respect to the second argument, the upper bound for ψ_{n+q}

$$\|\psi_{n+q}\| \le \Delta t L |\beta_j|_{max} \sum_{j=0}^{q} \|e_{n+j}\| + (T - t_0) \max_i \|T_i\|$$

can be established. The last estimate implies

$$\|\psi_\ell\| \leq \Delta t L |\beta_j|_{max} \sum_{j=0}^{q} \|e_{\ell-q+j}\| + (T - t_0) \max_i \|T_i\|. \tag{1.6.25}$$

Following the proof of Theorem 1.6.1, it can be shown that for e_n holds the estimate

$$\|e_n\| \leq C \max\{\max_{j=0,\ldots,q-1} \|e_j\|, \max_{j=q,\ldots,N} \|T_j\|\}$$

where C is a positive constant, time independent, and $\Delta t \in (0, \Delta t_0]$ with Δt_0 satisfying (1.6.18).

■

2-Numerical Methods for PDEs

2.1 Some Analytical Results

2.1.1 Some Mathematical Models

1. **The transport equation**

   Let us consider a tube with gas. Our aim is to establish a mathematical model which allow us to characterize the density and the speed of the gas particles at each point of the tube at each time. We introduce the reference system defining the x-axis as the line passing by the center of the tube. The origin is some point in this line. The final objective of the problem is to define the density $\rho(x, y, z, t)$ and the speed $v(x, y, z, t)$. In order to simplify the model we assume some realistic assumption on the physical model. We suppose that each transversal section has unitary area and, in each point of each transversal section, the gas has the same properties. Then, we have

   $$\rho(x, y, z, t) = \rho(x, 0, 0, t) := \rho(x, t),\ v(x, y, z, t) = v(x, 0, 0, t) := v(x, t).$$

   We establish now a mathematical law for $\rho(x, t)$ and for $v(x, t)$. Let $M(t)$ be the gas mass in the circular sector defined by $x_1 < x_2$ at time t,

   $$M(t) = \int_{x_1}^{x_2} \rho(x, t)dx.$$

   We assume that the wall tube is impermeable and the gas evolution only depends on the transport phenomenon. In this case the flux at x point and at time t, $J(x, t)$, is given by

   $$J(x, t) = v(x, t)\rho(x, t).$$

   Considering the gas mass in the tube sector and the flux at x_1 and x_2 we can establish the mass variation at time t. In fact, we have

   $$M'(t) = \int_{x_1}^{x_2} \frac{\partial \rho}{\partial t}(x, t)dx,$$

   and

   $$M'(t) = J(x_1, t) - J_2(x, t) = -\int_{x_1}^{x_2} \frac{\partial}{\partial x}(\rho(x, t)v(x, t))dx,$$

which implies

$$\int_{x_1}^{x_2} \frac{\partial \rho}{\partial t}(x,t)dx + \frac{\partial}{\partial x}(\rho v)(x,t)\,dx = 0, \tag{2.1.1}$$

provided that the density ρ and the speed v are smooth enough. From (2.1.1) we obtain the following PDEs

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v) = 0,\ x \in \mathbb{R}, t > 0, \tag{2.1.2}$$

usually called mass conservation equation. This equation is complemented by the two equations

$$\frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho v^2 + p) = 0, \tag{2.1.3}$$

$$\frac{\partial E}{\partial t} + \frac{\partial}{\partial x}(v(E+p)) = 0, \tag{2.1.4}$$

where p denotes the pressure and E represents the energy.

If we defined

$$u = \begin{bmatrix} \rho \\ \rho v \\ E \end{bmatrix},$$

then (2.1.2), (2.1.3), (2.1.4) are rewritten in the equivalent form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, \tag{2.1.5}$$

where

$$f(u) = \begin{bmatrix} \rho v \\ \rho v^2 + p \\ v(E+p) \end{bmatrix} = \begin{bmatrix} u_2 \\ \frac{u_2^2}{u_1} + p \\ u_2(u_3+p)/u_1 \end{bmatrix}.$$

In the particular case $f(u) = cu$, the established equation is known as transport equation.

If the speed v is known, then we only should compute the density ρ. In this case, if the initial particle distribution is known, which is translated specifying $\rho(x,0)$, the problem is defined by

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + v\dfrac{\partial \rho}{\partial x} = 0,\ x \in \mathbb{R}, t > 0, \\ \\ \rho(x,0) = \rho_0(x),\ x \in \mathbb{R}. \end{cases} \tag{2.1.6}$$

The problem (2.1.6), known as Initial Value Problem (IVP) or Cauchy problem, has the following solution

$$\rho(x,t) = \rho_0(x - vt), x \in \mathbb{R}, t \geq 0,$$

and the behaviour of ρ is completely determined by the initial condition ρ_0. For each time t, $\rho(x,t)$ is obtained from ρ_0 moving its graph from the left to the right if $v > 0$ and from the right to the left if $v < 0$.

2. **Diffusion equation** Let us consider a finite tube, with length ℓ, containing a solvent and a solute. Our aim is to compute the concentration of the solute in each point of the tube and at each time t. We introduce a reference system as in the previous model. However

we will now consider the origin coinciding with a tube end. Let $c(x, y, z, t)$ be the solute concentration at the point (x, y, z) at time t. If we assume that each transversal section has unitary area and in all points of each section we have equal concentration, then

$$c(x, y, z, t) = c(x, 0, 0, t) := c(x, t).$$

In order to establish a PDEs for the concentration, we suppose that the wall of the tube is impermeable and their ends are isolated. Let $x_1, x_2 \in (0, \ell)$, $x_1 < x_2$, and $M(t)$ the total mass in the tube sector defined by x_1, x_2,

$$M(t) = \int_{x_1}^{x_2} c(x, t)\, dx.$$

Then the instantaneously time mass variation is given by

$$M'(t) = \int_{x_1}^{x_2} \frac{\partial c}{\partial t}(x, t)\, dx.$$

Otherwise, $M'(t)$ can be computed considering the particles flux at the ends of the tube sector assuming that the flux $J(x, t)$ is defined by the Fick's law

$$J(x, t) = -D\nabla c(x, t), \tag{2.1.7}$$

where D is the diffusion coefficient related to the capacity of the solute particles to cross the solvent. The particles flux is a consequence of the molecular shocks being the particles movement from regions of high concentration to regions of low concentration.

Considering the Fick's law for the flux we have

$$\begin{aligned} M'(t) = J(x_1, t) - J(x_2, t) &= D(-\frac{\partial c}{\partial x}(x_1, t) + \frac{\partial c}{\partial x}(x_2, t)) \\ &= D\int_{x_1}^{x_2} \frac{\partial^2 c}{\partial x^2} dx. \end{aligned}$$

Then

$$\int_{x_1}^{x_2} \frac{\partial c}{\partial t} dx = D\int_{x_1}^{x_2} \frac{\partial^2 c}{\partial x^2} dx,$$

which implies

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2}, \tag{2.1.8}$$

for $x \in (0, \ell), t > 0$.

We assumed that the tube ends are isolated, which means that there is not any flux at $x = 0$ and at $x = \ell$,

$$\frac{\partial c}{\partial x}(0, t) = \frac{\partial c}{\partial x}(\ell, t) = 0, t > 0. \tag{2.1.9}$$

These two conditions are known as Neumann boundary conditions. Of course that we can assume that the initial solute concentration distribution is known by given

$$c(x, 0) = c_0(x), x \in (0, \ell). \tag{2.1.10}$$

We obtained the following initial boundary value problem

$$\begin{cases} \dfrac{\partial c}{\partial t} = D\dfrac{\partial^2 c}{\partial x^2},\ x \in (0,\ell), t > 0, \\ \dfrac{\partial c}{\partial x}(0,t) = \dfrac{\partial c}{\partial x}(\ell,t) = 0, t > 0, \\ c(x,0) = c_0(x), x \in (0,\ell). \end{cases} \tag{2.1.11}$$

If we assume that the solute is a fluid with movement, then the flux has Fickian and transport contributions. In this case $J(x,t)$ is given by

$$J(x,t) = -D\frac{\partial c}{\partial x}(x,t) + vc(x,t),$$

and for the concentration we obtain the following IBVP

$$\begin{cases} \dfrac{\partial c}{\partial t} = D\dfrac{\partial^2 c}{\partial x^2} - v\dfrac{\partial c}{\partial x},\ x \in (0,\ell), t > 0, \\ \dfrac{\partial c}{\partial x}(0,t) = \dfrac{\partial c}{\partial x}(\ell,t) = 0, t > 0, \\ c(x,0) = c_0(x), x \in (0,\ell). \end{cases} \tag{2.1.12}$$

If the solute and the solvent react, then on the definition of the instantaneously time mass variation using the particles flux, we should consider another term: $\displaystyle\int_0^\ell r(c(x,t),x,t)\,dx$, and the IBVP (2.1.12) is replaced by

$$\begin{cases} \dfrac{\partial c}{\partial t} = D\dfrac{\partial^2 c}{\partial x^2} - v\dfrac{\partial c}{\partial x} + r(c),\ x \in (0,\ell), t > 0, \\ \dfrac{\partial c}{\partial x}(0,t) = \dfrac{\partial c}{\partial x}(\ell,t) = 0, t > 0, \\ c(x,0) = c_0(x), x \in (0,\ell). \end{cases} \tag{2.1.13}$$

The Neumann boundary conditions can be replaced if we prescribe the solute concentration at the tube ends defining the Dirihlet boundary conditions

$$c(0,t) = c_e(t),\ c(\ell,t) = c_\ell(t), t > 0.$$

We remark that the Fick's law for the particles solute flux is formally equivalent to the Fourier law for the heat flux. Using this fact, the diffusion equation is also used to describe heat conduction phenomena being known as heat equation.

Let us consider two bodies with different constant temperatures in contact at initial time. The initial temperature distribution is a discontinuous function. Nevertheless, after the initial time, intuitively, the temperature distribution is very smooth. Independently of the smoothness of the initial condition, it seems that the solution of this kind of problems

are very smooth. In fact, it can be shown that the solution after the initial time is C^∞ provided that the initial condition is only bounded.

The high dimension problem correspondent to (2.1.13) can be written in the following form

$$\begin{cases} \dfrac{\partial c}{\partial t} = Lc, \ x \in \Omega, t > 0, \\ B_\eta c = g, x \in \partial\Omega, \ t > 0, \\ c(x,0) = c_0(x), x \in \Omega, \end{cases} \tag{2.1.14}$$

where Ω is an open subset of $\mathbb{R}^n$ with boundary $\partial\Omega$. In (2.1.14), L is an operator (linear or nonlinear), only presenting partial derivatives with respect to the x-components, defined between two function spaces. The boundary condition is defined by B_η which can be one of the following types:

- Dirichlet: $B_\eta c = c$;
- Neumann: $B_\eta c = \dfrac{\partial c}{\partial \eta}$, where η denotes the exterior unitary normal to Ω,
- Robin: $B_\eta c = \alpha \dfrac{\partial c}{\partial \eta} + \beta u$.

3. **The wave equation**: Let us consider a string with length ℓ with fixed ends. Suppose that at an initial time the string has some position and after that time it starts to move. Our aim is to describe the string movement.We assume that the motion takes place at a plan where a reference system $0xy$ with the origin at one end of the string was introduced. If $(x, u(x,t))$ is the position of a point of the string at time t, in what follows we establish a PDEs for u. In order to do that we should make some assumptions on the string motion:

   - the points only present vertical displacement,
   - on the string the tension force acts with the tangential direction,
   - the gravitational force is not considered.

   By ρ we denote the density of the string which is assumed time independent.

   Let PQ be an arc of the string with length Δs defined by x and $x + \Delta x$, where Δx is infinitesimal quantity. Let α and β be the angles of the tension vectors $T(P)$ and $T(Q)$ with $-e_1$ and e_1, respectively. As the string only has vertical movement, the horizontal components of the tension vectors acting on the arc PQ should be canceled, which means that

$$cos(\beta)\|T(Q)\| = cos(\alpha)\|T(P)\| = \|T\|. \tag{2.1.15}$$

   Consequently, on the arc PQ the force

$$F = (sen(\beta)\|T(Q)\| - sen(\alpha)\|T(P)\|)e_2 \tag{2.1.16}$$

   acts. By the second Newton's law, the force acting on the arc PQ can be computed using the mass of the arc and its acceleration. We have

$$M = \int_{PQ} \rho(s)\, ds,$$

and assuming smoothness on ρ we get

$$M = \rho(\xi(x))\Delta s,$$

with $\xi(x) \in (x, x+\Delta x)$. As we can assume that $\Delta s \simeq \Delta x$, and the acceleration of the arc can be given by $\dfrac{\partial^2 u}{\partial t^2}(\eta(x), t)$ with $\eta(x) \in (x, x+\Delta x)$, we deduce

$$sen(\beta)\|T(Q)\| - sen(\alpha)\|T(P)\| = \rho(\xi(x))\Delta x \frac{\partial^2 u}{\partial t^2}(\eta(x), t). \tag{2.1.17}$$

From (2.1.15) and (2.1.17), we get

$$tg(\beta) - tg(\alpha) = \frac{\rho(\xi(x))\Delta x}{T}\frac{\partial^2 u}{\partial t^2}(\eta(x), t), \tag{2.1.18}$$

which is equivalent to

$$\frac{1}{\Delta x}\left(\frac{\partial u}{\partial x}(x+\Delta x, t) - \frac{\partial u}{\partial x}(x,t)\right) = \frac{\rho(\xi(x))}{T}\frac{\partial^2 u}{\partial t^2}(\eta(x), t). \tag{2.1.19}$$

Takin in (2.1.19) $\Delta x \to 0$, we conclude for u the following PDEs

$$c^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2}(x,t),\ x \in (0,\ell), t > 0, \tag{2.1.20}$$

with $c^2 = \dfrac{T}{\rho(x)}$.

The equation (2.1.20) is known as wave equation and it is complemented with boundary and initial conditions: as the ends of the string are fixed on the x-axis we have

$$u(0,t) = u(\ell, t) = 0, t \geq 0,$$

the position of the string is known for the initial time $t = 0$

$$u(x,0) = \phi(x), x \in [0, \ell],$$

and the initial velocity is also known

$$\frac{\partial u}{\partial t}(x,0) = \psi(x), x \in [0, \ell].$$

Finally, for the displacement u we get the following IBVP

$$\begin{cases} c^2\dfrac{\partial^2 u}{\partial x^2} = \dfrac{\partial^2 u}{\partial t^2}(x,t), \quad x \in (0,\ell), t > 0, \\ u(0,t) = u(\ell,t) = 0, \quad t \geq 0, \\ u(x,0) = \phi(x), \quad x \in [0,\ell], \\ \dfrac{\partial u}{\partial t}(x,0) = \psi(x), \quad x \in [0,\ell]. \end{cases} \tag{2.1.21}$$

The high dimension of the IBVP (2.1.21) admits the representation

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = Lu,\ x \in \Omega, t > 0, \\ B_\eta u = g, x \in \partial\Omega,\ t > 0, \\ \dfrac{\partial u}{\partial t}(x,0) = \psi(x),\ x \in \Omega, \\ u(x,0) = \phi(x), x \in \Omega, \end{cases} \tag{2.1.22}$$

where Ω is an open subset of $\mathbb{R}^n$ with boundary $\partial\Omega$. In (2.1.23), L represents an operator between two function spaces only defined by the partial derivatives with respect to the x-components and B_η defines the boundary conditions.

4. **Stationary equations** Let us consider the diffusion IBVP (2.1.14) or the wave IBVP (2.1.21) when the solution is time independent. In this case we get the BVP

$$\begin{cases} Lu = f \ \text{ in } \Omega, \\ B_\eta u = g \ \text{ on } \partial\Omega, \end{cases} \tag{2.1.23}$$

where L, as before, presents only partial derivatives with respect x-components.

2.1.2 Some Solutions

The models presented in the last section are well-known examples of the use of PDEs on the mathematical modeling of physical problems. The second order PDEs are the most common on the applications and they are divided in three groups, depending on the behaviour of their coefficients.

The PDEs

$$A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x \partial yt} + C\frac{\partial^2 u}{\partial t^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial t} + Fu = G\,. \tag{2.1.24}$$

is called

1. elliptic if $B^2 - 4AC < 0$,
2. parabolic if $B^2 - 4AC = 0$,
3. hyperbolic if $B^2 - 4AC > 0$.

Then the linear diffusion equation is of parabolic type while the linear wave equation is hyperbolic. The Poisson equation

$$\Delta u = f$$

is elliptic.

The given classification can be extended to PDEs in higher dimensions. The equation

$$\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}(a_{ij}\frac{\partial u}{\partial x_i}) + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + a_0 u = g, \tag{2.1.25}$$

with $a_{ij} = a_{ji}$ is said parabolic if at least one of the eigenvalues of the matrix $[a_{ij}]$ is zero. If all eigenvalues have the same signal, then the equation (2.1.25) is elliptic. On the other hand the previous equation is hyperbolic if one eigenvalue has signal different from the others. Otherwise, it is said ultra-hyperbolic. Of course that if the coefficients are x-dependent, then the given classification depends on the point considered.

In what follows we present some results about the solutions of the problems presented before.

1. **The diffusion equation:** We start by considering the following IVP

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\dfrac{\partial^2 u}{\partial x^2} \; x \in \mathbb{R}, \; t > 0, \\ \\ u(x,0) = \phi(x), \; x \in \mathbb{R}. \end{cases} \tag{2.1.26}$$

Theorem 2.1.1 *If ϕ is continuous and bounded in $\mathbb{R}$, then*

$$\int_{\mathbb{R}} S(x-y,t)\phi(y)\,dy \tag{2.1.27}$$

is in $C^\infty(\mathbb{R}\times(0,+\infty))$,

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} \quad in \quad \mathbb{R}\times(0,+\infty)$$

and

$$\lim_{t\to 0^+} u(x,t) = \phi(x), \; x \in \mathbb{R},$$

where $S : \mathbb{R}\times(0,+\infty) \to \mathbb{R}$,

$$S(x,t) = \frac{1}{2\sqrt{D\pi t}} e^{-\frac{x^2}{4Dt}}, \; (x,t) \in \mathbb{R}\times(0,+\infty),$$

is the Green's function.

■

The maximum principle is one of the main properties of the solution of the diffusion equation which is established in the following result:

Theorem 2.1.2 *If u is continuous in $[x_1,x_2]\times[t_1,t_2]$ and $\dfrac{\partial u}{\partial t} = D\dfrac{\partial^2 u}{\partial x^2}$ in $(x_1,x_2)\times(t_1,t_2]$, then*

$$\max_{[x_1,x_2]\times[t_1,t_2]} u = \max_{\ell_1\cup\ell_2\cup\ell_3} u$$

with

$$\begin{aligned} \ell_1 &= \{(x_1,t),\, t\in[t_1,t_2]\}, \\ \ell_2 &= \{(x_2,t),\, t\in[t_1,t_2]\}, \\ \ell_3 &= \{(x,t_1) : x\in[x_1,x_2]\}. \end{aligned}$$

■

The solution of the IBVP

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\dfrac{\partial^2 u}{\partial x^2}\,,\ x>0,\ t>0, \\ \\ u(x,0)=\phi(x),\ x>0, \\ \\ u(0,t)=0,\ t>0. \end{cases} \tag{2.1.28}$$

is determined extending the initial condition ϕ to an odd function $\tilde{\phi}$ and computing the solution of the IBVP

$$\begin{cases} \dfrac{\partial \tilde{u}}{\partial t} = D\dfrac{\partial^2 \tilde{u}}{\partial x^2}\,,\ x\in \mathbb{R},\ t>0, \\ \\ \tilde{u}(x,0)=\tilde{\phi}(x),\ x\in \mathbb{R}. \end{cases} \tag{2.1.29}$$

We get, for $x \geq 0, t > 0$,

$$u(x,t)=\tilde{u}(x,t) \;=\; \frac{1}{2\sqrt{D\pi t}}\int_0^{+\infty}\left(e^{-\frac{(x-y)^2}{4Dt}}-e^{-\frac{(x+y)^2}{4Dt}}\right)\phi(y)\,dy.$$

If the homogeneous Neumann boundary condition is considered at $x = 0$, an even extension of the initial condition should be taken. Following the procedure considered in the Dirichlet case, we obtain

$$u(x,t)=\frac{1}{2\sqrt{D\pi t}}\int_0^{+\infty}\left(e^{-\frac{(x-y)^2}{Dkt}}+e^{-\frac{(x+y)^2}{Dkt}}\right)\phi(y)\,dy,\ x\geq 0,\ t>0.$$

The method of separation of variables allow us to compute the solution of the diffusion problem

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\dfrac{\partial^2 u}{\partial x^2}\ x\in(0,\ell),\ t>0, \\ \\ u(x,0)=\phi(x),\ x\in(0,\ell), \\ \\ \dfrac{\partial u}{\partial x}(0,t)=\dfrac{\partial u}{\partial x}(\ell,t)=0,\ t>0, \end{cases} \tag{2.1.30}$$

and the following result can be proved:

Theorem 2.1.3 *If ϕ is continuous in $[0,\ell]$, $\phi' \in L^2[0,\ell]$, then the Fourier's series*

$$u(x,t)=\frac{a_0}{2}+\sum_{n=1}^{+\infty}a_n e^{-D(\frac{n\pi}{\ell})^2 t}cos(\frac{n\pi}{\ell}x),$$

with

$$a_0=\frac{2}{\ell}\int_0^{\ell}\phi(x)\,dx\,,\ a_n=\frac{2}{\ell}\int_0^{\ell}\phi(x)cos(\frac{n\pi}{\ell}x)\,dx,$$

is such that

(a) u is continuous in $[0,\ell]\times[0,+\infty)$,

(b) there exists $\dfrac{\partial u}{\partial x}$ in $[0,\ell]\times(0,+\infty)$ and it is continuous in the previous domain,

(c) $\dfrac{\partial u}{\partial t} = D\dfrac{\partial^2 u}{\partial x^2}$ *in* $(0,\ell)\times(0,+\infty)$
(d) $\dfrac{\partial u}{\partial x}(0,t) = \dfrac{\partial u}{\partial u}(\ell,t) = 0,\ t>0,$
(e) $u(x,0) = \phi(x)$ *in* $[0,\ell]$.

■

If the Neumann boundary conditions are replaced by the Dirichlet boundary condition, a similar result can be established.

2. **The wave equation:** A class of wave problems have solutions with an explicit form depending explicitly on the data. For the following IVP

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = c^2\dfrac{\partial^2 u}{\partial x^2},\ x\in\mathbb{R}, t>0, \\ \dfrac{\partial u}{\partial t}(x,0) = \psi(x),\ x\in\mathbb{R}, \\ u(x,0) = \phi(x), x\in\mathbb{R}, \end{cases} \tag{2.1.31}$$

holds the following result:

Theorem 2.1.4 *If* $\phi\in C^2(\mathbb{R})$ *and* $\psi\in C^1(\mathbb{R})$, *then*

$$u(x,t) = \frac{\phi(x+ct)+\phi(x-ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct}\psi(s)\,ds,\ x\in\mathbb{R}, t\geq 0, \tag{2.1.32}$$

is solution of (2.1.31).

If ϕ *and* ψ *have compact support, then (2.1.32) is the unique solution of the IVP (2.1.4).*

■

If we consider (2.1.31) with $\mathbb{R}$ replaced by $\mathbb{R}_+$, that is, if we consider the IBVP

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = c^2\dfrac{\partial^2 u}{\partial x^2}, \quad x\in(0,+\infty), t>0, \\ u(x,0) = \phi(x),\ x\in[0,+\infty), \\ \dfrac{\partial u}{\partial t}(x,0) = \psi(x),\ x\in[0,+\infty), \\ u(0,t) = h(t),\ t\geq 0. \end{cases} \tag{2.1.33}$$

then, under compatibility conditions on the data, we have

$$u(x,t) = \begin{cases} \dfrac{1}{2}\left(\phi(x+ct)+\phi(x-ct)\right) + \dfrac{1}{2c}\displaystyle\int_{x-ct}^{x+ct}\psi(s)\,ds \\ \qquad\qquad (x,t):\ x-ct\geq 0, \\ \dfrac{1}{2}\left(\phi(x+ct)-\phi(-x+ct)\right) + \dfrac{1}{2c}\displaystyle\int_{-x+ct}^{x+ct}\psi(s)\,ds + h(-\dfrac{x-ct}{c}) \\ \qquad\qquad (x,t):\ x-ct<0 \quad . \end{cases} \tag{2.1.34}$$

If the spatial domain is bounded, the IBVP (2.1.21) has boundary conditions on both boundary points. In this case the solution of such problem can be computed using the method of separation of variables. For Dirichlet boundary conditions we have the following result.

Theorem 2.1.5 *Let $\phi, \psi : [0, \ell] \to \mathbb{R}$ be such that*

(a) $\phi, \phi', \phi'', \psi, \psi'$ *are continuous in* $[0, \ell]$,

(b) ϕ''', ψ'' *are piecewise continuous in* $[0, \ell]$,

(c) $\phi(0) = \phi(\ell) = \phi''(0) = \phi''(\ell) = 0$,

(d) $\psi(0) = \psi(\ell) = 0$.

Then

$$\sum_{n=1}^{+\infty} \left(A_n cos(\frac{cn\pi}{\ell} t) + B_n sen(\frac{cn\pi}{\ell} t) \right) sen(\frac{n\pi}{\ell} x), \tag{2.1.35}$$

with

$$A_n = \frac{2}{\ell} \int_0^\ell \phi(x) sen(\frac{n\pi}{\ell} x)\, dx, \;\; B_n = \frac{2}{cn\pi} \int_0^\ell \psi(x) sen(\frac{n\pi}{\ell} x)\, dx, \; n \in \mathbb{N}, \tag{2.1.36}$$

defines a function u such that

(a) u *is continuous in* $[0, \ell] \times [0, +\infty)$

(b) $\dfrac{\partial u}{\partial t}$ *is continuous in* $[0, \ell] \times [0, +\infty)$,

(c) $u \in C^2((0, \ell) \times (0, \infty))$[7]

(d)

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},$$
$$u(0, t) = u(\ell, t) = 0,$$
$$u(x, 0) = \phi(x)$$

and

$$\frac{\partial u}{\partial t}(x, 0) = \psi(x).$$

■

In all the results presented before for the wave equation, the smoothness of the data have a crucial role on the construction of the solutions.

[7] If Ω is an open set in $\mathbb{R}^n$ and $k \in \mathbb{N}$, by $C^m(\Omega)$ we denote the set of all continuous functions such that $\dfrac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}$ is continuous in Ω. We used the notations $\alpha = (\alpha_1, \dots, \alpha_n)$ and $|\alpha| = \sum_{i=1}^n \alpha_i$. By $C_0^m(\Omega)$ we represent the set of all functions in $C^k(\Omega)$ whose support is a bounded set of Ω. By $C_0^\infty(\Omega)$ we represent the set $\cap_{m \geq 0} C_0^m(\Omega)$.

3. **The Laplace equation** Let Ω be an open subset of $\mathbb{R}^n$ with a smooth boundary $\partial\Omega$. We characterize in what follows the solution of the problem

$$\begin{cases} \Delta u = f \ \text{ in } \Omega, \\ \\ B_\eta u = g \ \text{ on } \ \partial\Omega, \end{cases} \tag{2.1.37}$$

where B_η defines the Dirichlet or the Neumann boundary conditions. On this characterization the Green's identities

(a)

$$\int_\Omega v\Delta u\, dx = -\sum_{i=1}^{n} \int_\Omega \frac{\partial v}{\partial x_i}\frac{\partial u}{\partial x_i} + \int_{\partial\Omega} v\frac{\partial u}{\partial \eta} ds, \tag{2.1.38}$$

(b)

$$\int_\Omega v\Delta u\, dx = \int_\Omega u\Delta v + \int_{\partial\Omega} \left(v\frac{\partial u}{\partial \eta} - u\frac{\partial v}{\partial \eta}\right) ds, \tag{2.1.39}$$

(c)

$$\int_\Omega \Delta u\, dx = \int_{\partial\Omega} \frac{\partial u}{\partial \eta}\, ds, \tag{2.1.40}$$

where $u, v \in C^2(\overline{\Omega})$,[8] have an important role.

It can be shown that for $B_\eta = id$ or $B_\eta = \dfrac{\partial}{\partial \eta}$, the BVP (2.1.37) has at most one solution in $C^2(\overline{\Omega})$. In the Neumann case, we should assume the compatibility condition between f and g

$$\int_\Omega f\, dx = \int_{\partial\Omega} g\, ds. \tag{2.1.41}$$

In the last case, if u_1 and u_2 are two solutions, then $u_1 = u_2 + C$, for some constant C.

Theorem 2.1.6 *If $u \in C^2(\overline{\Omega})$, then, for $\xi \in \Omega$, holds the representation*

$$u(\xi) = \int_\Omega K(x,\xi)\Delta u\, dx - \int_{\partial\Omega} \left(K(x,\xi)\frac{\partial u}{\partial \eta} - u\frac{\partial K}{\partial \eta}(x,\xi)\right) ds, \tag{2.1.42}$$

where

$$K(x,\xi) = \begin{cases} \dfrac{r^{2-n}}{(2-n)\omega_n}, & n > 2, \\ \\ \dfrac{1}{\omega_n} log(r), & n = 2, \end{cases} \tag{2.1.43}$$

$r = \|x - \xi\|$ and ω_n denotes the area of the unitary ball of $\mathbb{R}^n$.

■

[8] If Ω is a bounded set of $\mathbb{R}^n$, by $C^m(\overline{\Omega})$ we denote the set of all functions $u \in C^m(\Omega)$ such that D^α can be extended from Ω to a continuous function on $\overline{\Omega}$, for all α such that $|\alpha| \leq m$. $C^m(\overline{\Omega})$ can be equipped with the norm

$$\|u\|_{C^m(\overline{\Omega})} = \sum_{|\alpha|\leq m} \sup_{x\in\Omega} |D^\alpha u(x)|.$$

Corollary 4 *If $u \in C^2(\overline{\Omega})$, then, for $\xi \in \Omega$,*

$$u(\xi) = \int_\Omega G(x,\xi)\Delta u\, dx - \int_{\partial\Omega} \left(G(x,\xi)\frac{\partial u}{\partial \eta} - u\frac{\partial G}{\partial \eta}(x,\xi) \right) ds, \tag{2.1.44}$$

where $G(x,\xi) = K(x,\xi) + w(x)$, $x \in \overline{\Omega}$, $\xi \in \Omega, x \neq \xi$, and $w \in C^2(\overline{\Omega})$ is a harmonic function in Ω.

■

Remark 1 *If the Green function G is such that $G = 0$ on $\partial\Omega$, then*

$$u(\xi) = \int_\Omega G(x,\xi)\Delta u\, dx + \int_{\partial\Omega} u\frac{\partial G}{\partial \eta}\, ds. \tag{2.1.45}$$

The solution of the boundary problem

$$\Delta u = f \, in \, \Omega, \; u = g \, on \, \partial\Omega,$$

satisfies

$$u(\xi) = \int_\Omega G(x,\xi) f(x)\, dx + \int_{\partial\Omega} g\frac{\partial G}{\partial \eta}\, ds. \tag{2.1.46}$$

If the Green function G is such that $\dfrac{\partial G}{\partial \eta} = 0$, then, for the solution of the boundary value problem

$$\Delta u = f \, in \, \Omega, \quad \frac{\partial u}{\partial \eta} = g \, on \, \partial\Omega,$$

holds the following representation

$$u(\xi) = \int_\Omega G(x,\xi) f(x)\, dx - \int_{\partial\Omega} gG(x,\xi)\, ds. \tag{2.1.47}$$

The fundamental question on the construction of the solution of the Poisson equation, with the mentioned boundary conditions, is the computation of the Green's function with the specified requirements.

As the next result holds for harmonic functions, we can establish an explicit form for the solution of the Laplace equation with Dirichlet boundary condition defined on a ball of radius ρ.

Theorem 2.1.7 *If $u \in C^2(\overline{B_\rho(\xi)})$ and u is harmonic in $B_\rho(\xi)$, then*

$$u(\xi) = \frac{1}{\omega_n \rho^{n-1}} \int_{S_\rho(\xi)} u(x)\, ds. \tag{2.1.48}$$

■

We can consider for the Green's function definition the following extension

$$G(x,\xi) = K(x,\xi) + w(x,\xi),$$

where $w \in C^2$, $\Delta_x w = 0, x \in \Omega$, $x \neq \xi \in \Omega$. Using this definition, it can be constructed a Green's function G, null at $S_a(0)$. Consequently, the next result can be proved.

Theorem 2.1.8 *If $u \in C^2(\overline{B_a(0)})$, u is harmonic in $B_a(0)$, then, for $\xi \in B_a(0)$,*

$$u(\xi) = \int_{S_a(0)} H(x,\xi)u(x)\,ds, \tag{2.1.49}$$

where the Poisson kernel is given by

$$H(x,\xi) = \frac{1}{a\omega_n}\frac{a^2 - \|\xi\|^2}{\|x-\xi\|^n}, x \in S_a(0),\ \xi \in B_a(0).$$

■

Analogously to the solution of the diffusion equation, for harmonic functions hold the maximum principle:

Theorem 2.1.9 *Let Ω be a connected open bounded set of $\mathbb{R}^n$. If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ and u is harmonic in Ω, then*

$$\max_{x\in\overline{\Omega}} u(x) = \max_{x\in\partial\Omega} u(x)$$

■

An short overview on some results for some well known equations: diffusion equation, wave equation and Poisson's equation, was given. The expressions of the solutions of the previous problems with homogeneous equations and homogeneous boundary conditions were presented. Nevertheless, we can also construct the solutions of some non homogeneous equations and for non homogeneous boundary conditions. It should be stressed that only for a very small number of cases the obtained expressions are mathematically manipulated. The existence of the solution and the study of its regularity properties, for more general BVPs and IBVP, can be seen for instance in [3], [5], [19], [32], [33], [36].

2.2 Finite Difference Methods for Elliptic Equations

2.2.1 Introduction: the One-Dimensional BVP

In what follows we introduce the finite difference methods (FDMs) for elliptic equations considering an one-dimensional BVP. We point out that FDMs were considered before for IVP where the new approximation at the new time level is obtained using the approximations at the previous points. For BVP the finite difference equation should be consider for all grid points leading to a linear or nonlinear system of equations where the unknowns are the approximations at the grid points.

Let us consider the BVP

$$-u''(x) = f(x),\ x \in (a,b),\ u(a) = u_a, u(b) = u_b. \tag{2.2.1}$$

We fixe the step size $h > 0$ and in $[a,b]$ we introduce the spatial grid $I_h = \{x_i, i = 0, \ldots, n\}$ with $x_i - x_{i-1} = h, i = 1, \ldots, n$, $x_0 = a$ and $x_n = b$.

Let $V_h(I_h)$ and $V_h(I_h')$ be vector spaces of grid functions defined in I_h and in $I_h' = I_h - \{x_i, i = 0, n\}$, respectively. We represent by D_{-x}, D_x, D_c and D_2 the finite difference operators

$$D_{-x}u_h(x_i) = \frac{u_h(x_i) - u_h(x_{i-1})}{h}, i = 1, \ldots, n,$$

$$D_x u_h(x_i) = \frac{u_h(x_{i+1}) - u_h(x_i)}{h}, i = 0, \ldots, n-1,$$

$$D_c u_h(x_i) = \frac{u_h(x_{i+1}) - u_h(x_{i-1})}{2h}, i = 1, \ldots, n-1,$$

$$D_2 u_h(x_i) = \frac{u_h(x_{i+1}) - 2u_h(x_i) + u_h(x_{i-1})}{h^2}, i = 1, \ldots, n-1,$$

defined for $u_h \in V_h(I_h)$.

We remark that if $u \in C^2(a,b)$, then

$$D_{-x}u(x_i) = u'(x_i) - \frac{h}{2}u''(\xi_i), \xi_i \in (x_{i-1}, x_i), i = 1, \ldots, n,$$

$$D_x u(x_i) = u'(x_i) + \frac{h}{2}u''(\xi_i), \xi_i \in (x_i, x_{i+1}), i = 0, \ldots, n-1.$$

If $u \in C^3(a,b)$, then

$$D_c u(x_i) = u'(x_i) + \frac{h^2}{12}(u'''(\xi_i) + u'''(\eta_i)), \xi_i, \eta_i \in (x_{i-1}, x_{i+1}), i = 1, \ldots, n-1,$$

$$D_2 u(x_i) = u''(x_i) + \frac{h^2}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)),\ \xi_i, \eta_i \in (x_{i-1}, x_{i+1}), i = 1, \ldots, n-1. \tag{2.2.2}$$

Consider in the equation (2.2.1), $x = x_i \in (a,b)$. As (2.2.2) holds, we obtain

$$-D_2 u(x_i) - \frac{h^3}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)) = f(x_i), i = 1, \ldots, n-1,$$

that allow us to define the following system

$$-D_2 u_h(x_i) = f(x_i), i = 1, \ldots, n-1,\ u_h(x_0) = u_a, u_h(x_n) = u_b, \tag{2.2.3}$$

where $u_h(x_i)$ represents the numerical approximation for $u(x_i)$. If we replace $f(x_i)$ by its approximation $f_h(x_i)$, then we get

$$-D_2 u_h(x_i) = f_h(x_i), i = 1, \ldots, n-1,\ u_h(x_0) = u_a, u_h(x_n) = u_b. \tag{2.2.4}$$

This last approach is usually followed when $f(x_i)$ is computationally difficult to evaluate.

In order to simplify our analysis we rewrite (2.2.3) ((2.2.4)) in the condensed form

$$L_h u_h = \tilde{f}_h,$$

where $L_h u_h$ and f_h denote the grid functions

$$L_h u_h(x_i) = \begin{cases} -\dfrac{u_h(x_2) - 2u_h(x_1)}{h^2}, & , i = 1, \\ -D_2 u_h(x_i) & , i = 2, \ldots, n-1, \\ -\dfrac{2u_h(x_{n-1}) + u_h(x_{n-2})}{h^2}, & , i = n-1, \end{cases}$$

and

$$\tilde{f}_h = (f_h(x_1) + \frac{u_a}{h^2}, f_h(x_2), \ldots, f_h(x_{n-1}) + \frac{u_b}{h^2}).$$

The error presented in the grid function u_h defined by (2.2.4) is studied now. In order to do that, we consider the general BVP

$$Lu = f \text{ in } (a,b), u(a) = u_a, u(b) = u_b, \tag{2.2.5}$$

where L is a second order linear or nonlinear differential operator. The solution of the BVP (2.2.6) is approximated by the solution of the finite difference problem

$$L_h u_h = \tilde{f}_h, \tag{2.2.6}$$

where $u_h \in V_h(I_h')$ and L_h represents a finite difference operator. The global error and the truncation error associated with the method (2.2.6) are defined as in the context of the IVPs. The global error e_h is defined by

$$e_h(x_i) = u(x_i) - u_h(x_i), i = 1, \ldots, n-1, \; e_h(x_0) = e_h(x_n) = 0.$$

Nevertheless, the global error $e_h(x_i), i = 1, \ldots, n-1$, can be rewritten in the following from $e_h = R_h u - u_h$, where $R_h : C^2(a,b) \cap C[a,b] \to V_h(I_h')$ represents the restriction operator. Analogously, as the truncation error has $n-1$ components, such error admits the representation $T_h = L_h(R_h u) - \tilde{R}_h(Lu)$, where $\tilde{R}_h$ is a restriction operator analogous to R_h. The convergence and the consistency of the method (2.2.6) is defined considering the convergence to zero of the previous errors. In this case, such convergence should be considered with respect to a norm $\|.\|_h$ defined in $V_h(I_h')$. The global and the truncation errors are related by

$$L_h e_h = T_h.$$

In fact

$$L_h e_h = L_h(R_h u) - L_h u_h = L_h(R_h u) - f_h = L_h(R_h u) - \tilde{R}_h Lu = T_h.$$

The consistency and convergence orders are defined as in the IVP context.

The concept of stability has here a different meaning. Such concept is introduced considering perturbations in the second member of (2.2.6) and analysing the behaviour of the difference between the solution and its perturbation. Let $F_h : V_h(I_h') \to V_h(I_h')$ be a general finite difference operator(linear or nonlinear). If $\|F_h(u_h) - F_h(v_h)\|_h \to 0$, then $\|u_h - v_h\|_h \to 0$, we say that F_h is stable.

The convergence of the method (2.2.6) can be deduced, at least for the linear case, from its stability and its consistency. In fact, by consistency we have $\|T_h\|_h \to 0, h \to 0$, which implies that $\|L_h e_h\|_h \to 0, h \to 0$. From the stability of L_h we get $\|e_h\|_h \to 0$.

For the linear case, a sufficient condition for stability can be easily established:

Theorem 2.2.1 *If the finite difference operator $L_h : V_h(I_h') \to V_h(I_h')$ is injective and there exists a positive constant C, h-independent, such that $\|L_h^{-1}\|_h \leq C$, for $h \leq h_0$, then L_h is stable, for $h \leq h_0$.*

■

Under the assumptions of the Theorem 2.2.1, from the consistency we conclude the convergence of the method and we also deduce that the convergence order is at least equal to the consistency order.

The convergence properties of the FDMs are established with respect to specified norms. The most common norms are: $\|.\|_\infty$ and $\|.\|_2$being the last one defined by

$$\|u_h\|_2 = \left(\sum_{j=1}^{n-1} h u_h(x_j)^2 \right)^{1/2}, \; u_h \in V_h(I_h').$$

For linear case the stability can be deduced analysing the properties of its associated matrix. In the considered example, we immediately conclude the existence of L_h^{-1} because the matrix is diagonally dominant and it is strictly dominant in the first and in the last rows. The operator L_h has the eigenvalues $\lambda_m = \frac{4}{h^2} sen^2(\frac{m\pi h}{2})$, $m = 1, \ldots, n-1$, and the correspondent eigenvectors $\mu_m(x) = sen(m\pi x)$, $m = 1, \ldots, n-1$. We also have $\|L_h^{-1}\|_\infty \leq \frac{1}{8}$ (this result can be proved by using the results presented in the next section). An estimate for the error induced by the method (2.2.3) with respect to the infinity norm is immediately established provided that $C^4[a, b]$.

If in (2.2.1), we replace the Dirichlet boundary conditions by Neumann boundary conditions we can consider two approaches:

- If the finite difference equation is considered for $i = 1, \ldots, n-1$, then we should discretize the boundary conditions using the forward and backward finite difference operators;
- If the finite difference equation is considered for $i = 0, \ldots, n$, then the boundary conditions are discretized using the centered finite difference operator and, in order to do that, we need to introduce two fictitious points $x_{-1} = a - h, x_{n+1} = b + h$. Consequently, the space $V(I_h')$ is replaced by $V(I_h)$ and the norms used in the convergence analysis should include in their definition the boundary points.

Let us give now some details when nonlinear BVP

$$F(u) = f \text{ in } (a, b),\; u(a) = u_a,\; u(b) = u_b,$$

are considered. Let us discretize the previous problem by

$$F_h u_h = \tilde{f}_h,$$

where$\tilde{f}_h = \tilde{R}_h f$. Then for the truncation error

$$T_h = F_h(R_h u) - \tilde{R}_h F(u)$$

holds the representation

$$T_h = F_h(R_h u) - F_h u_h.$$

Obviously, if F_h is stable and consistent, then $\|e_h\|_h \to 0$. In order to get an estimate for $\|e_h\|_h$, we observe that

$$T_h = F_h(R_h u) - F_h u_h = F_h'(R_h u + \theta e_h)e_h.$$

Consequently, if

$$\|F_h'(R_h u + \theta e_h)^{-1}\|_h \leq C,$$

we obtain

$$\|e_h\|_h \leq C\|T_h\|_h.$$

The previous FDMs were introduced for uniform grids. Nevertheless, if the solution of the BVP has high gradients zones, the computation of an accurate solution requires the use of a huge number of points increasing the computational cost. In order to avoid the high computational cost we should use non dense nonuniform grids where the grid points are concentrated only in those zones. In this case, the finite difference operators, previously introduced for uniform grids, are defined by

$$D_{-x}u_h(x_i) = \frac{u_h(x_i) - u_h(x_{i-1})}{h_i}, i = 1, \ldots, n,$$

$$D_x u_h(x_i) = \frac{u_h(x_{i+1}) - u_h(x_i)}{h_{i+1}}, i = 0, \ldots, n-1,$$

$$D_c u_h(x_i) = \frac{u_h(x_{i+1}) - u_h(x_{i-1})}{h_i + h_{i+1}}, i = 1, \ldots, n-1,$$

$$D_2 u_h(x_i) = \frac{h_i u_h(x_{i+1}) - (h_i + h_{i+1})u_h(x_i) + h_{i+1}u_h(x_{i-1})}{h_i h_{i+1}(h_i + h_{i+1})/2}, i = 1, \ldots, n-1,$$

where $h_i = x_i - x_{i-1}, i = 1, \ldots, n$.

The convergence analysis of the FDMs defined on nonuniform grids follows the steps used on uniform case. Consistency and stability of the method should imply convergence. As far as consistency is concerned, the order of the truncation error decreases when nonuniform grids are used. Consequently, the convergence order for nonuniform grids is apparently lower than the correspondent convergence order for uniform grids. This convergence order can be deduced from the estimate for the global error established using the truncation error. However, numerical experiments shown that this convergence order is in fact apparent. Numerically was observed that for nonuniform grids the order of the global error is greater than the order of the truncation error. Since the 80s several authors shown analytically this property refining stability inequalities or deducing a second order expression for the global error. This phenomenon was called supraconvergence and was studied for instance in [7], [8], [10], [15], [22], [25], [38].

The properties of the finite difference operators, or equivalently the properties of the associated matrices, have an important role in the convergence study. In the next section we present an overview on several results of matrix analysis.

2.2.2 Some Matrix Results

Let $A = [a_{ij}]$ and $B = [b_{ij}]$, $i = 1, \ldots, m, j = 1, \ldots, m$, be square matrices. If $a_{ij} \geq b_{ij}$, we write $A \geq B$. Analogously we define the inequalities $\leq, <, >$.

The matrix $A = [a_{ij}]$ is called an M-matrix if

$$a_{ii} > 0, \forall i,\ a_{ij} \leq 0,\ i \neq j,$$

A is nonsingular and $A^{-1} \geq 0$.

We associate to $A = [a_{ij}]$ a graph defined in what follows. Let us consider the indexes $i, j \in \{1, \ldots, n\}$. The index i is said directly connected with the index j if $a_{ij} \neq 0$. We say that the index i is connected with j if there exists a connection (chain of direct connections) $\alpha_0 = i, \alpha_1, \ldots, \alpha_k = j$ such that $a_{\alpha_{\ell-1}\alpha_\ell} \neq 0$. The graph of A is defined by the set $\{1, \ldots, n\}$ with the direct connections.

A square matrix A is said to be irreducible if every $i \in \{1, \ldots, n\}$ is connected with every $j \in \{1, \ldots, n\}$. A is said irreducibly diagonally dominant if A is irreducible and A is diagonally dominant

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|,\ i = 1, \ldots, n,$$

being the last inequality strictly satisfied for some i.

The Gershgorin theorem can be used analyse if a matrix is nonsingular.

Theorem 2.2.2 *If $A = [a_{ij}]$ is a real square matrix of order n and λ is an eigenvalue of A, then*

$$\lambda \in \bigcup_{i \in \{1,\ldots,n\}} \overline{B}_{r_i}(a_{ii}).$$

If A is irreducible, then

$$\lambda \in \cup_{i \in \{1,\ldots,n\}} B_{r_i}(a_{ii}) \cup \cap_{i \in \{1,\ldots,n\}} S_{r_i}(a_{ii}),$$

where $r_i = \sum_{j \neq i} |a_{ij}|$.

9

9

Proof: Let x be an eigenvector associated with the eigenvalue λ and i such that $|x_i| = \|x\|_\infty$. Immediately we have

$$|a_{ii} - \lambda| \leq \sum_{j \neq i} |a_{ij}| \frac{|x_j|}{|x_i|} \leq r_j.$$

If A irreducible and $\lambda \in \cup_{i \in \{1,\ldots,n\}} B_{r_i}(a_{ii})$, the proof is concluded. If $\lambda \notin \cup_{i \in \{1,\ldots,n\}} B_{r_i}(a_{ii})$, then we should prove that $\lambda \in \cap_{i \in \{1,\ldots,n\}} S_{r_i}(a_{ii})$.

Let us suppose that $a_{ij} \neq 0$, that is, i is connected with j, and $|x_i| = 1$. We start by proving that

$$|\lambda - a_{ii}| = r_i \Longrightarrow |x_j| = 1, |\lambda - a_{jj}| = r_j \tag{2.2.7}$$

holds. As $|\lambda - a_{ii}| = r_i$, we deduce

$$\sum_{\ell \neq i} |a_{i\ell}||x_\ell| = \sum_{\ell \neq i} |a_{i\ell}|,$$

and $|a_{i\ell}||x_\ell| = |a_{i\ell}|$ because $|x_\ell| \leq \|x\|_\infty$. Then $|x_\ell| = 1$ and $|\lambda - a_{\ell\ell}| \leq r_\ell$. Finally using the fact $\lambda \notin \cup_{j \in \{1,\ldots,n\}} B_{r_j}(a_{jj})$, we get $|\lambda - a_{\ell\ell}| = r_\ell$, that is, we conclude the proof of (2.2.7).

If x is an eigenvector, we can assume that there exists i such that $\|x\|_\infty = |x_i| = 1$ and $|\lambda - a_{ii}| \leq r_i$. Then $\lambda \in \cup_{i \in \{1,\ldots,n\}} S_{r_i}(a_{ii})$ and thus $|\lambda - a_{ii}| = r_i$. As A is irreducible, for $j \in \{1, \ldots, n\}$, there exists a index sequence $\alpha_1, \ldots \alpha_k$ such that $a_{\alpha_{\ell-1}\alpha_\ell} \neq 0$ and

$$|x_{\alpha_\ell}| = 1,\ |\lambda - a_{\alpha_\ell\alpha_\ell}| = r_{\alpha_\ell}.$$

As a particular case, we have $\lambda \in S_{r_j}(a_{jj})$. As j is arbitrary, we conclude that $\lambda \in \cap_{j \in \{1,\ldots,n\}} S_{r_j}(a_{jj})$.

■

We split $A = [a_{ij}]$ into

$$A = D - B$$

where D is the diagonal part of A and $B = D - A$ is the off-diagonal part of A

$$b_{ii} = 0, b_{ij} = -a_{ij}.$$

Let C be defined by $C = D^{-1}B$. We characterize in what follows the spectral radius of C.

Theorem 2.2.3 *If A is strictly diagonally dominant or irreducibly diagonally dominant, then*

$$\rho(C) < 1. \tag{2.2.8}$$

Proof: Let λ be an eigenvalue of C. By the Gershgorin Theorem, $\lambda \in \cup_{i=1,\ldots,n} \overline{B}_{r_i}(0)$, with $r_i = \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|}$. If A is strictly diagonally dominant, then $r_i < 1$. Otherwise, if A is irreducibly diagonally dominant, then $\lambda \in \cup_{i=1,\ldots,n} B_{r_i}(0) \cup \cap_{i=1,\ldots,n} S_{r_i}(0)$. As for some $i \in \{1, \ldots, n\}$, we get $r_i < 1$, and then

$$r_j = r_i, \forall j, \text{ or } \exists j \in \{1, \ldots, n\} : r_j \neq r_i.$$

Let us only consider the second case. We have $S_{r_i}(0) \cap S_{r_j}(0) = \emptyset$ and thus $\lambda \in \cup_{i=1,\ldots,n} B_{r_i}(0)$, that is, $\rho(C) < 1$.

■

Using the previous result we establish a necessary and sufficient condition for M-matrices.

Theorem 2.2.4 *The matrix $A = [a_{ij}]$ such that $a_{ii} > 0, a_{ij} \leq 0$, is a M-matrix if and only if $\rho(C) < 1$.*

Proof:

- If $\rho(C) < 1$, then $S = \sum_{j=0} C^j = (I - C)^{-1}$. Furthermore

  $$I = S(I - C) = S(I - D^{-1}B) = SD^{-1}(D - B) = SD^{-1}A,$$

  which allow us to conclude that A is nonsingular. As $C^m \geq 0$, we deduce that $S \geq 0$. Using the fact $D^{-1} \geq 0$, we conclude that $A^{-1} \geq 0$.

- Let us suppose now that A is a M-matrix. We prove in what follows that $\rho(C) < 1$.

  Let λ be an eigenvalue of C and x the correspondent eigenvector. Let $|x|$ be the vector whose components are the absolute value of the components of x. We have

  $$|\lambda||x| = |\lambda x| = |D^{-1}Bx| \leq D^{-1}B|x|,$$

  and consequently $|\lambda|D|x| \leq B|x|$. As $A^{-1} \geq 0$, from the last inequality we get $|\lambda|A^{-1}D|x| \leq A^{-1}B|x|$ which enable us to conclude the following

  $$-|\lambda|A^{-1}D|x| \geq -A^{-1}B|x|.$$

Using the last inequality, an estimate for $|x|$ is obtained as follows

$$\begin{aligned} |x| &= A^{-1}A|x| \\ &= A^{-1}(D-B)|x| \\ &= A^{-1}D|x| - A^{-1}B|x| \\ &\leq A^{-1}D(|x| - |\lambda||x|). \end{aligned}$$

Finally, if $|\lambda| \geq 1$ then $|x| = 0$ which conclude the proof.

■

The next result is a corollary of Theorem 2.2.4.

Corollary 5 *Let A be a matrix such that $a_{ii} > 0, a_{ij} \leq 0, i \neq j$. If A is strictly diagonally dominant or irreducibly diagonally dominant, then A is a M-matrix.*

On the construction of global error estimates for the solution obtained using a FDM defined by a matrix A, the estimates for $\|A^{-1}\|$ have an important role. If A is a M-matrices, we can obtain such estimates without the evaluation of its inverse.

Theorem 2.2.5 *If A is a M-matrix and w is such that $Aw \geq \mathbb{1}$, then $\|A^{-1}\|_\infty \leq \|w\|_\infty$.*

Proof: For $x \in \mathbb{R}^n$ we have

$$|x| \leq \|x\|_\infty \mathbb{1} \leq \|x\|_\infty Aw.$$

As $A^{-1} \geq 0$ obtain

$$A^{-1}|x| \leq \|x\|_\infty w,$$

which implies

$$\|A^{-1}|x|\|_\infty \leq \|x\|_\infty \|w\|_\infty.$$

Finally, as

$$\|A^{-1}x\|_\infty \leq \|x\|_\infty \|w\|_\infty,$$

we conclude

$$\|A^{-1}\|_\infty \leq \|w\|_\infty.$$

■

An upper bound for $\|A\|_2$ can be obtained by using the spectral radius of A^tA. We remak that

$$\begin{aligned} \|A\|_2 &= \sup_{0 \neq x \in \mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2} \\ &= \sup_{0 \neq x \in \mathbb{R}^n} \frac{(x^tA^tAx)^{1/2}}{\|x\|_2}. \end{aligned}$$

If A is nonsingular, then A^tA is a symmetric positive definite matrix ($x^tA^tAx > 0, x \neq 0$). Such properties enable us to deduce that the eigenvalues of A^tA are positive.[10] As A^tA is a

[10]

diagonalizable matrix, we have

$$\|Ax\|_2^2 = (Q^t x)^t Q^t A^t A Q (Q^t x) = y^t Diag[\lambda_i] y,$$

where Q is a orthogonal matrix. Thus

$$\|Ax\|_2^2 = \sum_i \lambda_i y_i^2 \leq \rho(A^t A)\|x\|_2^2$$

which implies

$$\|A\|_2 \leq \rho(A^t A)^{1/2}.$$

Otherwise, we also have

$$\|A\|_2^2 = \sup_{0 \neq x \in \mathbb{R}^n} \frac{\|Ax\|_2^2}{\|x\|_2^2} \geq \frac{u^t A^t A u}{\|u\|_2^2} \geq \rho(A^t A),$$

where u denotes an arbitrary eigenvector of $A^t A$ correspondent to the eigenvalue λ.

Combining the two estimates for $\|A\|_2$, we conclude the following identity

$$\|A\|_2 = \rho(A^t A)^{1/2}.$$

We proved the following the result.

Theorem 2.2.8 *If A is a square matrix, then $\|A\|_2 = \rho(A^t A)^{1/2}$. Otherwise, if A is symmetric, then $\|A\|_2 = \rho(A)$.*

■

As a consequence of this result we have:

$$\|A\|_2 = \lambda_{max}, \|A^{-1}\|_2 = \frac{1}{\lambda_{min}},$$

provided that A is symmetric positive definite matrix.

It is easy to show that a symmetric matrix is positive definite if and only if all eigenvalues are positive. A criterion to test if a symmetric matrix is positive definite can be established using the Gershgorin Theorem.

Theorem 2.2.6 *If A is a real symmetric positive definite matrix, then the eigenvalues of A are positive.*

Proof: For λ and x, respectively, the eigenvalue and the correspondent eigenvector of A, we have

$$0 < x^t A x = \lambda x^t x \Longrightarrow \lambda > 0.$$

■

Theorem 2.2.7 *If A is a real symmetric positive definite matrix, then A is nonsingular and A^{-1} is positive definite.*

Proof: As A is a real symmetric positive definite matrix, then $A = Q^t Diag[\lambda_i] Q$ and $A^{-1} = Q^t Diag[1/\lambda_i] Q$, where Q is an orthogonal matrix.

If $\lambda > 0$ is an eigenvalue of A, then λ^{-1} is an eigenvalue of A^{-1}. Consequently, A^{-1} has positive eigenvalues. For $y \in \mathbb{R}^n$, $y \neq 0$, we have

$$y^t A^{-1} y = (Q^t y)^t Diag[1/\lambda_i](Q^t y) = \sum_i \frac{1}{\lambda_i} y_i^2 > 0.$$

■

Theorem 2.2.9 *If A is a real symmetric matrix with $a_{ii} > 0$, strictly diagonally dominant or irreducibly diagonally dominant, then A is positive definite.*

Proof: If A is strictly diagonally dominant and λ its eigenvalue, then $\lambda > 0$ because

$$\lambda \in \cup_{i=1,\ldots,n} \overline{B}_{r_i}(a_{ii}).$$

If A is irreducibly diagonally dominant, then

$$\lambda \in \cup_{i=1,\ldots,n} B_{r_i}(a_{ii}) \cup \cap_{i=1,\ldots,n} S_{r_i}(a_{ii}).$$

As for some i, $r_i < a_{ii}$, we get

$$0 \notin \cap_{j=1,\ldots,n} S_{r_j}(a_{jj}),$$

and then $\lambda > 0$.

■

2.2.3 The Poisson Equation: the Five-Point Formula - Qualitative and Quantitative Analysis

Let Ω be the two-dimensional rectangle $\Omega = (0, a) \times (0, b)$ with boundary $\partial\Omega$. We introduce in what follows a finite difference discretization of the Poisson equation with Dirichlet boundary conditions

$$\begin{cases} -\Delta u = f \text{ in } \Omega, \\ \\ u = g \text{ on } \partial\Omega. \end{cases} \tag{2.2.9}$$

On $\overline{\Omega}$ we define the grid

$$\overline{\Omega}_H = \{(x_i, y_j), i = 0, \ldots, n, j = 0, \ldots, m, x_0 = y_0 = 0, x_n = a, y_m = b\}$$

where

$$H = (h, k), x_i - x_{i-1} = h, y_j - y_{j-1} = k.$$

By $\partial\Omega_H$ we denote the set of grid points on the boundary $\partial\Omega$, that is,

$$\partial\Omega_H = \overline{\Omega}_H \cap \partial\Omega, \Omega_H = \overline{\Omega}_H \cap \Omega.$$

By $W_H(\overline{\Omega}_H)$, $W_H(\Omega_H)$ and $W_H(\partial\Omega_H)$ we represent the grid spaces defined, respectively, in $\overline{\Omega}_H$, Ω_H and $\partial\Omega_H$.

By Δ_H we denote the finite difference operator

$$\Delta_H : W_H(\overline{\Omega}_h) \to W_H(\Omega_H),$$

defined by

$$\Delta u_H(x_i, y_j) = D_{2,x} u_H(x_i, y_j) + D_{2,y} u_H(x_i, y_j),\ (x_i, y_j) \in \Omega_H,$$

for $u_H \in W_H(\overline{\Omega}_H)$ and where $D_{2,x}, D_{2,y}$ are the second order centered finite difference operators in x and y directions, respectively. Usually, this operator is called five-point formula for the

Laplace operator. For the particular case $h = k$, we associate with the finite difference operator Δ_h the following matrix

$$\frac{1}{h^2}\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}.$$

If we consider grid functions $f_H \in W_H(\Omega_H)$ and $g_H \in W_H(\partial\Omega_H)$, approximations to f and g, respectively, an approximation u_h for the solution of the BVP (2.2.9), $u_H \in W_H(\overline{\Omega}_h)$, can be computed using the FDM

$$\begin{cases} -\Delta_H u_H = f_H \text{ in } \Omega_H, \\ \\ u_H = g_H \text{ on } \partial\Omega_H. \end{cases} \tag{2.2.10}$$

The FDM can be rewritten in matrix form

$$L_h u_H = \tilde{f}_H \tag{2.2.11}$$

For example, if we take $h = k$ and $a = b$, and we introduce in the grid points an enumeration: from the bottom to the top and from the left to the right, then

$$L_H := \frac{1}{h^2}\begin{bmatrix} T & -I & 0 & \dots & 0 & 0 \\ -I & T & -I & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -I & T \end{bmatrix} \tag{2.2.12}$$

where I denotes the $n-1$ identity matrix and T is defined by

$$T = \begin{bmatrix} 4 & -1 & 0 & \dots & 0 & 0 \\ -1 & 4 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 4 \end{bmatrix}.$$

In (2.2.11) $\tilde{f}_H$ is given by

$$\tilde{f}_h = \begin{cases} \begin{cases} f_H(x_1,y_1) + \frac{1}{h^2}(g_H(x_1,y_0) + g_H(x_0,y_1), \\ f_H(x_i,y_1) + \frac{1}{h^2}g_H(x_i,y_0), i = 2,\dots,n-2, \\ f_H(x_{n-1},y_1) + \frac{1}{h^2}(g_H(x_{n-1},y_0) + g_H(x_n,y_1),) \end{cases} \\ \begin{cases} f_H(x_1,y_j) + \frac{1}{h^2}g_H(x_0,y_j), \\ f_H(x_i,y_j),\ i = 2,\dots,n-2, \\ f_H(x_{n-1},y_j) + \frac{1}{h^2}g_H(x_n,y_j), \qquad\qquad j = 2,\dots,n-2, \end{cases} \\ \begin{cases} f_H(x_1,y_{n-1}) + \frac{1}{h^2}(g_H(x_1,y_n) + g_H(x_0,y_{n-1}), \\ f_H(x_i,y_{n-1}) + \frac{1}{h^2}(g_H(x_i,y_n), i = 2,\dots,n-2, \\ f_H(x_{n-1},y_{n-1}) + \frac{1}{h^2}(g_H(x_{n-1},y_n) + g_H(x_n,y_{n-1})). \end{cases} \end{cases} \tag{2.2.13}$$

The existence and the uniqueness of the solution of (2.2.10) is consequence of the properties of L_H presented in the following result.

Theorem 2.2.10 *If L_H is defined by (2.2.12), then*

1. *L_H is a M-matrix,*

2. L_H *is positive definite,*

3. $\|L_H\|_\infty \leq \frac{8}{h^2}$, $\|L_H^{-1}\|_\infty \leq \frac{1}{8}$,

4. $\|L_H\|_2 \leq \frac{8}{h^2} cos^2(\frac{\pi h}{2})$, $\|L_H^{-1}\|_2 \leq \frac{h^2}{8} cosec^2(\frac{\pi h}{2})$.

Proof:

1. We observe that the diagonal entries of L_H are positive and the off-diagonal entries are negative. As L_H is irreducibly diagonally dominant, we conclude that L_H is a M -matrix;

2. L_H is a symmetric irreducibly diagonally dominant which implies that the eigenvalues of L_H are positive. Consequently, L_H is positive definite.

3. $\|L_H\|_\infty \leq \frac{8}{h^2}$ is trivial. Let us consider the $w_H \in W_H(\Omega_H)$ defined by $w_H(x,y) = \frac{x}{2}(1-x), (x,y) \in \Omega_H$. This grid function satisfies $L_H w_H(x,y) \geq \mathbb{1}$. Then $\|L_H^{-1}\|_\infty \leq \|w_H\|_\infty = \frac{1}{8}$ because L_H is a M-matrix.

4. As the eigenvalues of L_H and the corresponding eigenvectors are defined by

$$\lambda_{ij} = \frac{4}{h^2}\left(sen^2(\frac{i\pi h}{2}) + sen^2(\frac{j\pi h}{2})\right),$$

$$\mu_{ij}(x,y) = sen(i\pi x) sen(j\pi y), (x,y) \in \Omega_H,$$

for $i,j = 1,\ldots,n-1$, we conclude the proof.

■

Qualitative properties: the mean value theorem, the maximum principle

We study now some qualitative properties of the solution of the discrete Poisson equation (2.2.10) for $h = k$. We establish a discrete version of the Mean Value Theorem - Theorem 2.1.7 - for harmonic functions and a discrete version of the Maximum Principle - Theorem 2.1.9.

Let us consider $f = 0$ in Ω. If $u_H \in W_H(\overline{\Omega}_H)$ is solution of the finite difference problem (2.2.10) with $f_H = 0$ we say that u_H is discretely harmonic.

The next result is consequence of the definition of Δ_H. For $P \in \Omega_H$ we denote by $\mathcal{V}(P)$ the following grid set

$$\mathcal{V}(P) = \{P \pm he_1, P \pm he_2\}.$$

Theorem 2.2.11 *If $u_H \in W_H(\overline{\Omega}_H)$ is discretely harmonic and $P \in \Omega_H$, then*

$$u_H(P) = \frac{1}{4}\sum_{Q \in \mathcal{V}(P)} u_H(Q).$$

■

For discretely harmonic functions we have the following discrete maximum principle.

Theorem 2.2.12 *If $u_H \in W_H(\overline{\Omega}_H)$ is discretely harmonic, then*

$$\max_{\overline{\Omega}_H} u_H = \max_{\partial\Omega_H} u_H, \qquad \min_{\overline{\Omega}_H} u_H = \min_{\partial\Omega_H} u_H.$$

Proof: If u_H is a constant function then the result holds. Let u_H be a discretely harmonic function in $\overline{\Omega}_H$ which has its maximum value at $P \in \Omega_H$. As

$$u_H(P) = \frac{1}{4} \sum_{Q \in \mathcal{V}_H(P)} u_H(Q),$$

$u_h(P)$ satisfies

$$u_H(P) = u_H(Q),\ Q \in \mathcal{V}_H(P).$$

Following this procedure we can prove that u_H is constant in $\overline{\Omega}_H$. This conclusion contradicts the assumption on u_H.

■

An upper bound to the norm of a discretely harmonic function can be obtained as a consequence of the discrete maximum principle.

Corollary 6 *If $u_H \in W_H(\overline{\Omega}_H)$ is discretely harmonic in Ω_H and $u_H = g_H$ on $\partial\Omega_H$, then*

$$\|u_H\|_\infty \leq \|g_H\|_\infty.$$

■

We study now the stability of the FDM (2.2.12). Let $u_H^{(i)}, i = 1, 2$, be grid functions in $W_H(\overline{\Omega}_H)$, defined by (2.2.12) for different boundary conditions

$$\begin{cases} -\Delta_H u_H^{(i)} = f_H \text{ em } \Omega_H, \\ u_H^{(i)} = g_H^{(i)} \text{ em } \partial\Omega_H. \end{cases}$$

Then $u_H^{(1)} - u_H^{(2)}$ is discretely harmonic and, by the discrete maximum principle, we have

$$\|u_H^{(1)} - u_H^{(2)}\|_\infty \leq \|g_H^{(1)} - g_H^{(2)}\|_\infty.$$

Furthermore, as $L_H^{-1} \geq 0$, if

$$g_H^{(1)} \geq g_H^{(2)} \text{ on } \partial\Omega_H,$$

we obtain

$$u_H^{(1)} - u_H^{(2)} \geq 0 \text{ in } \overline{\Omega}_H.$$

We proved the next corollary:

Corollary 7 *Let $u_H^{(i)}, i = 1, 2$, be grid function in $W_H(\overline{\Omega}_H)$, defined by*

$$\begin{cases} -\Delta_H u_H^{(i)} = f_H \text{ in } \Omega_H \\ u_H^{(i)} = g_H^{(i)} \text{ on } \partial\Omega_H\,. \end{cases}$$

If $g_H^{(1)} \geq g_H^{(2)}$ on $\partial\Omega$, then

$$\|u_H^{(1)} - u_H^{(2)}\|_\infty \leq \|g_H^{(1)} - g_H^{(2)}\|_\infty,$$

and $u_H^{(1)} \geq u_H^{(2)} \geq 0$ in $\overline{\Omega}_H$.

■

An upper bound for $\|u_H\|_\infty$, where u_H is the solution of the discrete Poisson equation, is now obtained using $\|f_H\|_\infty$ and $\|g_H\|_\infty$.

Theorem 2.2.13 *If u_H in ${}_H(\overline{\Omega}_H)$ is solution of (2.2.10), then*

$$\|u_H\|_\infty \leq \frac{1}{8}\|f\|_\infty + \|g_H\|_\infty. \tag{2.2.14}$$

Proof: The grid function $\tilde{f}_H$, defined by (2.2.10), admits the representation $\tilde{f}_H = f_H + \tilde{g}_H$ for a convenient $\tilde{g}_h$.

We introduce now two grid functions: $u_H^{(1)} \in W_H(\overline{\Omega}_H)$ is solution of the discrete Poisson equation with f_H as a second member and with homogeneous boundary conditions, $u_H^{(2)}$ is solution of the discrete Laplace equation with g_H as a Dirichlet boundary condition. We have

$$\|u_H^{(1)}\|_\infty \leq \frac{1}{8}\|f_H\|_\infty.$$

Otherwise, by Corollary 6, we also have

$$\|u_H^{(2)}\|_\infty \leq \|g_H\|_\infty.$$

As $u_H = u_H^{(1)} + u_H^{(2)}$, from the two last estimates, we conclude the proof of the estimate(2.2.14). ■

Quantitative properties

The behaviour of the finite difference solutions when the step size sequence converges to zero is now studied. The concepts of consistency, convergence and stability were introduced before for FDMs in one-dimensional context. We formalize now the same definitions for FDMs for two-dimensional problems. The correspondent definitions can be easily given for high dimensions.

Let Λ be a sequence of vectors $H = (h, k)$ such that $h \in \Lambda_1, k \in \Lambda_2$, and Λ_i converges to zero, $i = 1, 2$. As for one-dimensional problem, the finite difference problem can be seen as a boundary finite difference problem on Ω_H or on $\overline{\Omega}_H$ and we denote this set by Ω_H^*.

Let u_H be a grid function in $W_H(\overline{\Omega}_H)$ defined by

$$\begin{cases} \mathcal{A}_H u_H = f_H \text{ in } \Omega_H^* \\ \\ B_H u_H = g_H \text{ on } \partial\Omega_H, \end{cases} \tag{2.2.15}$$

which approximates the solution $u \in \mathcal{U}$ of the elliptic BVP

$$\begin{cases} \mathcal{A}u = f \text{ in } \Omega, \\ \\ Bu = g \text{ on } \partial\Omega, \end{cases} \tag{2.2.16}$$

where B denotes the boundary operator and B_H its discretization.

By R_H we represent the restriction operator $R_H : \mathcal{U} \to W_H(\overline{\Omega}_H)$, where $\mathcal{U}$ is a vector space containing the solution u. Let $\|.\|_H$ be a norm in $W_H(\overline{\Omega}_H)$.

If

$$\|R_H u - u_H\|_H \to 0, H \to 0,$$

then the FDM (2.2.15) is said convergent.

As $f_H = \tilde{R}_H f$, for the error $e_H = R_H u - u_H$ we have

$$\begin{aligned} \mathcal{A}_H(R_H u - u_H) &= \mathcal{A}_H(R_H u) - f_H \\ &= \mathcal{A}_H(R_H u) - \tilde{R}_H f \\ &= \mathcal{A}_H(R_H u) - \tilde{R}_H \mathcal{A} u. \end{aligned}$$

As far as the error e_H on the boundary points is concerned we establish

$$\begin{aligned} B_H(R_H u - u_H) &= B_H(R_H u) - g_H \\ &= B_H(R_H u) - R_{H,\partial\Omega} g \\ &= B_H(R_H u) - R_{H,\partial\Omega} B u, \end{aligned}$$

where $R_{H,\partial\Omega}$ denotes the restriction operator for functions defined on $\partial\Omega$.

The grid function $T_H \in W_H(\overline{\Omega}_H)$ given by

$$T_H = \mathcal{A}_H(R_H u) - \tilde{R}_H \mathcal{A} u$$

in Ω_H^* and

$$T_H = B_H(R_H u) - R_{H,\partial\Omega} B u$$

on $\partial\Omega_H$, is called truncation error of the FDM (2.2.15). If $\|T_H\|_H \to 0$, then this method is said consistent with (2.2.16). Furthermore, if $\|T_H\|_H = O(H_{max}^p)$, then the FDM (2.2.15) is said with consistency order equal to p.

If

$$\|B_H v_h\|_{\partial\Omega_H} \to 0,\ \|\mathcal{A}_H v_H\|_H \to 0$$

then

$$\|v_H\|_H \to 0,$$

the FDM (2.2.15) is said stable.

A sufficient condition for the convergence of a FDM can be easily proved using consistency and stability.

Theorem 2.2.14 *If the FDM (2.2.15) is stable and consistent, then it is convergent.*

■

We analyse in what follows the convergence properties of the five-point formula (2.2.10) with

$$f_H(P) = f(P), P \in \Omega_H,\ g_H(P) = g(P), P \in \partial\Omega_H.$$

It is easy to show

$$-\Delta_H R_H u(P) = -\Delta u(P) - \frac{h^2}{24}\left(\frac{\partial^4 u}{\partial x^4}(P_1) + \frac{\partial^4 u}{\partial x^4}(P_2) + \frac{\partial^4 u}{\partial y^4}(P_3) + \frac{\partial^4 u}{\partial y^4}(P_4)\right),$$

where $P_i \in (x_P - h, x_P + h) \times (y_P - h, y_P + h), i = 1, 2, 3, 4.$

If $u \in C^4(\overline{\Omega})$, then

$$\|T_H\|_\infty \leq \frac{h^2}{6}\|u\|_{C^4}.$$

Using Theorem 2.2.10, we obtain

$$\|u_H - R_H u\|_\infty \leq \frac{h^2}{48}\|u\|_{C^4}. \tag{2.2.17}$$

The error estimate (2.2.17) was established assuming that $u \in C^4(\overline{\Omega})$. The last smoothness requirement can be avoided. In fact, for $P = (x_i, y_j)$ we have

$$\begin{aligned} -D_{2,x}u(P) &= -\frac{\partial^2 u}{\partial x^2}(P) - \frac{1}{6h^2}\int_{x_i}^{x_{i+1}} \frac{\partial^4 u}{\partial x^4}(s, y_j)(x_{i+1} - s)^3\, ds \\ &- \frac{1}{6h^2}\int_{x_i}^{x_{i-1}} \frac{\partial^4 u}{\partial x^4}(s, y_j)(x_{i-1} - s)^3\, ds \end{aligned}$$

where

$$\begin{aligned} \frac{1}{6h^2}\int_{x_i}^{x_{i+1}} \frac{\partial^4 u}{\partial x^4}(s, y_j)(x_{i+1} - s)^3\, ds &= \frac{1}{2h^2}\int_{x_i}^{x_{i+1}} (\frac{\partial^3 u}{\partial x^3}(x_i, y_j) - \frac{\partial^3 u}{\partial x^3}(s, y_j))(x_{i+1} - s)^2\, ds \\ &\leq \frac{h^2}{24}\|u\|_{C^{3,1}(\overline{\Omega})}, \end{aligned}$$

[11]

$$\frac{1}{6h^2}\int_{x_i}^{x_{i-1}} \frac{\partial^4 u}{\partial x^4}(s, y_j)(x_{i-1} - s)^3\, ds \leq \frac{h^2}{24}\|u\|_{C^{3,1}(\overline{\Omega})}.$$

Consequently

$$\|T_H\|_\infty \leq \frac{h^2}{6}\|u\|_{C^{3,1}(\overline{\Omega})},$$

and therefore

$$\|u_H - R_H u\|_\infty \leq \frac{h^2}{48}\|u\|_{C^{3,1}(\overline{\Omega})}.$$

2.2.4 FDMs of High Order

The FDM for the Poisson equation with Dirichlet boundary conditions studied in the last section has second convergence order. In what follows we define a new DFM for the same problem with higher convergence order.

We start by considering the one-dimensional problem. Let D_h be defined by

$$D_h u_h(x) = \sum_{j=-k}^{k} c_j u_h(x + jh).$$

If we replace u_h by a function u smooth enough, we obtain

$$D_h u(x) = \sum_{m=0}^{2k} a_m h^m u^{(m)}(x) + O(h^{2k+1}), \;\; a_m = \frac{1}{m!}\sum_{j=-k}^{k} c_j j^m.$$

[11] $C^{k,1}(\overline{\Omega})$ is the set of all functions u in $C^k(\overline{\Omega})$ whose derivatives $D^\alpha u$ are Lipschits continuous for $|\alpha| \leq k$.

Thus, if

$$a_2 = \frac{1}{h^2}, a_j = 0, j = 0, 1, 3 \ldots, 2k,$$

we get a finite difference approximation for the second derivative. For $k = 1$ we obtain the centered finite difference operator D_2. For $k = 2$ we obtain the finite difference operator

$$\begin{aligned} D_h u_h(x) = \quad & \frac{-1}{12h^2}\left(u_h(x-2h) + u_h(x+2h)\right) \\ & + \frac{4}{3h^2}\left(u_h(x-h) + u_h(x+h)\right) - \frac{5}{2h^2}u_h(x). \end{aligned} \tag{2.2.18}$$

This finite difference operator is fourth order consistent. We remark that for the grid function u_h defined on $\overline{\Omega}_h$, $D_h u_h$ is only defined for x such that $x - 2h, x + 2h \in \overline{\Omega}_h$.

Let us consider now the two-dimensional case with $H = (h, h)$. Applying the operator D_h in both directions x and y, we easily get a finite difference discretization of the Laplace operator $-\Delta$ which can be represented by the matrix

$$\frac{1}{12h^2}\begin{bmatrix} & & 1 & & \\ & & -16 & & \\ 1 & -16 & 60 & -16 & 1 \\ & & -16 & & \\ & & 1 & & \end{bmatrix}.$$

This FDM is fourth order consistent but presents some difficulties near to the boundary points. For example, if we use the previous formula at $(h, h) \in \Omega_H$, we need $u_h(-h, h)$, $u(h, -h)$ outside of $\overline{\Omega}_H$. This difficulty can be avoided if the five-point formula is used at the points near to the boundary. Nevertheless, this approach leads to a decrease of the consistency order.

In order to overcome the weakness of the last approach we construct FDM of the following type

$$D_H u_H(x, y) = \sum_{i,j=-k}^{k} c_{ij} u_H(x + ih, y + jh).$$

Replacing u_H by a smooth function u we obtain

$$D_H u(x, y) = \sum_{n,m=0} \frac{\partial^{n+m} u}{\partial x^n \partial y^m}(x, y) h^{n+m} a_{nm}$$

with

$$a_{nm} = \sum_{ij} c_{ij} \frac{1}{n!m!} i^n j^m,$$

and $H = (h, h)$.

For $k = 1$, a nine-point formula is deduced. If we compute the nine coefficients $c_{ij}, i, j = -1, 0, 1$, such that $D_H u(x, y) = -\Delta u(x, y) + O(h^p)$, the finite difference formula $-\tilde{\Delta}_H$ represented by the matrix

$$\frac{1}{6h^2}\begin{bmatrix} -1 & -4 & -1 \\ -4 & 20 & -4 \\ -1 & -4 & -1 \end{bmatrix},$$

is obtained with $p = 2$. This nine-point formula and the five-point formula have the same consistency order being the first one computationally inefficient. Nevertheless, the nine-point formula can be used to define a FDM with higher consistency order defining conveniently f_H. In fact, let us consider

$$-\tilde{\Delta}_H u_H = f_H \text{ in } \Omega_H.$$

We have

$$-\tilde{\Delta}_H u(x,y) = -\Delta u - \frac{h^2}{12}\Delta^2 u - \frac{h^4}{360}\left(\frac{\partial^4}{\partial x^4} + 4\frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4}\right)\Delta u + O(h^6), \tag{2.2.19}$$

provided that $u \in C^6(\overline{\Omega})$. In (2.2.19), Δ^2 represents the biharmonic operator

$$\Delta^2 u = \Delta\Delta u = \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}.$$

Let f_H be defined by

$$\begin{aligned} f_H(x,y) &= \frac{1}{12}\big(f(x-h,y) + f(x+h,y) \\ &\quad + f(x,y-h) + f(x,y+h) + 8f(x,y)\big). \end{aligned}$$

As we have

$$\tilde{f}_H(x,y) = f(x,y) + \frac{h^2}{12}\Delta f(x,y) + O(h^4),$$

provided $f \in C^4(\overline{\Omega})$, we deduce that

$$T_H = -\tilde{\Delta}_H u - f_H = O(h^4),$$

provided that $u \in C^6(\overline{\Omega})$.

It can be shown that the last finite difference approximation is fourth order convergent to the solution of the Poisson equation with Dirichlet boundary condition provided that $u \in C^6(\overline{\Omega})$.

2.2.5 FDMs for the Poisson Equation with Neumann Boundary Conditions

We consider the Poisson equation defined on $\Omega = (0,1)\times(0,1)$ with the Neumann boundary condition

$$\frac{\partial u}{\partial \eta} = g \text{ on } \partial\Omega.$$

This condition is meaning less for the boundary points $V = \{(0,0),(1,0),(0,1),(1,1)\}$, and we define the discretization of the normal derivative, at points in $\Omega_H - V$, by

$$B_\eta u_H(P) = \frac{u_H(P) - u_H(P - h\eta)}{h},\ P \in \partial\Omega_H - V.$$

Let $\overline{\Omega}_H^*$ be given by $\overline{\Omega}_H - V$ and let u_H in $W_H(\overline{\Omega}_H^*)$ be such that

$$\begin{cases} -\Delta_H u_H = f_H \text{ in } \Omega_H, \\ B_\eta u_H = g_H, \text{ on } \partial\Omega_H - V. \end{cases} \tag{2.2.20}$$

The FD problem (2.2.20) can be rewritten in the equivalent form

$$L_H u_H = \tilde{f}_H \text{ em } \Omega_H, \tag{2.2.21}$$

where L_H is the $(n-1)^2$ square matrix

$$L_H = \frac{1}{h^2}\begin{bmatrix} T_1 & -I & 0 & \dots & 0 & 0 \\ -I & T & -I & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -I & T_1 \end{bmatrix}, \tag{2.2.22}$$

where

$$T_1 = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 & 0 \\ -1 & 3 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 2 \end{bmatrix}, \; T = \begin{bmatrix} 3 & -1 & 0 & \dots & 0 & 0 \\ -1 & 4 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 3 \end{bmatrix},$$

and

$$\tilde{f}_h = \begin{cases} \begin{cases} f_H(x_1,y_1) + \frac{1}{h}(g_H(x_1,y_0) + g_H(x_0,y_1)), \\ f_H(x_i,y_1) + \frac{1}{h}g_H(x_i,y_0), i = 2,\dots,n-2, \\ f_H(x_{n-1},y_1) + \frac{1}{h}(g_H(x_{n-1},y_0) + g_H(x_n,y_1)), \end{cases} \\ \begin{cases} f_H(x_1,y_j) + \frac{1}{h}g_H(x_0,y_j) \\ f_H(x_i,y_j),\ i = 2,\dots,n-2, \\ f_H(x_{n-1},y_j) + \frac{1}{h}g_H(x_n,y_j), \qquad\qquad j = 2,\dots,n-2, \end{cases} \\ \begin{cases} f_H(x_1,y_{n-1}) + \frac{1}{h}(g_H(x_1,y_n) + g_H(x_0,y_{n-1})), \\ f_H(x_i,y_{n-1}) + \frac{1}{h}g_H(x_i,y_{n-1}), i = 2,\dots,n-2, \\ f_H(x_{n-1},y_{n-1}) + \frac{1}{h}(g_H(x_{n-1},y_n) + g_H(x_n,y_{n-1})). \end{cases} \end{cases} \tag{2.2.23}$$

As in the continuous case, in general the problem (2.2.21) is not solvable. In fact, while L_H is irreducible, it is not irreducibly diagonally dominant. Moreover, as $L_H \mathbb{1} = 0$, where $\mathbb{1}$ represents here the $(n-1)^2$ vector with unitary components, L_H is singular. Otherwise, if we eliminate in this matrix a row and the correspondent column, then L_H is irreducibly diagonally dominant. This fact leads to

$$car(L_H) = (n-1)^2 - 1 \text{ and } \mathcal{N}(L_H) = \mathcal{L}\{\mathbb{1}\}.$$

Problem (2.2.21) has a solution u_H in $W_H(\overline{\Omega}_H^*)$ if and only if $\tilde{f}_H \in \mathcal{C}(L_H)$ if and only if $\tilde{f}_H$ is orthogonal to $\mathcal{N}(L_H)$, that is

$$0 = < \tilde{f}_H, \mathbb{1} > = \sum_{P \in \Omega_H} \tilde{f}_H.$$

Using the definition of $\tilde{f}_H$, the last equality is equivalent to

$$h^2 \sum_{P \in \Omega_H} f_H(P) = -h \sum_{P \in \partial\Omega_H^*} g_H(P). \tag{2.2.24}$$

Furthermore, if u_H, v_H are solutions of (2.2.21), then $u_H - v_H \in \mathcal{N}(L_H)$, or equivalently

$$u_H - v_H = c\mathbb{1}.$$

We proved the following existence result:

Theorem 2.2.15 *The finite difference problem (2.2.20) has a solution in $W_H(\overline{\Omega}_H)$ if and only if the compatibility condition (2.2.24) holds. Any two solution of (2.2.20) can only differ by a constant.*

■

Let us suppose now that the compatibility condition (2.2.24) holds and let Q be a fixed grid point in Ω_H. Then there exists a unique solution of the finite difference problem (2.2.20) such that $u_H(Q) = 0$. This solution can be computed using (2.2.21) where the row and the column associated with the grid point Q were deleted.

Another approach to solve (2.2.21) can considered replacing this problem by

$$\overline{L}_H \overline{u}_H = \overline{\tilde{f}}_H, \tag{2.2.25}$$

where

$$\overline{L}_H = \begin{bmatrix} L_H & \mathbb{1} \\ \mathbb{1}^t & 0 \end{bmatrix}, \ \bar{u}_H = \begin{bmatrix} u_H \\ \lambda \end{bmatrix}, \ \overline{\tilde{f}}_H = \begin{bmatrix} \tilde{f}_H \\ \sigma \end{bmatrix},$$

and σ can be prescribed arbitrarily.

As $\mathbb{1}$ and the columns of L_H are linearly independent, we get $rank([L_H \mathbb{1}]) + 1 = (n-1)^2$. Furthermore, as $(\mathbb{1}^t, 0)$ and the rows of $[L_H \mathbb{1}]$ are linearly independent, we conclude $rank(\overline{L}_H) = (n-1)^2 + 1$, which means that (2.2.25) has a unique solution $\overline{u}_H$. If $\overline{u}_H$ is such that $\lambda = 0$, then the compatibility condition (2.2.24) holds. Thus, u_H is solution of (2.2.21) with $\sigma = \mathbb{1} u_H$. Otherwise, if $\lambda \neq 0$, then u_H is solution of the modified problem

$$L_H u_H = \tilde{f}_H - \lambda \mathbb{1}.$$

The last problem is associated with the FDM

$$\begin{cases} -\Delta_H u_H = f_H - \lambda \text{ in } \Omega_H, \\ \\ B_\eta u_H = g_H \text{ on } \partial\Omega_H^*. \end{cases} \tag{2.2.26}$$

The truncation error induced by (2.2.26) satisfies the following relations

$$-\Delta_H (R_H u - u_H) = -\tilde{R}_H \Delta u + \lambda - \Delta_H R_H u = T_H^{(1)} + \lambda,$$

$$B_\eta (R_H u - u_H) = B_\eta R_H u - g_H = B_\eta R_H u - R_{H,\partial\Omega} \frac{\partial u}{\partial \eta} = T_H^{(2)},$$

where

$$\|T_H^{(1)}\|_\infty \leq C h^2 \|u\|_{C^{3,1}(\overline{\Omega})}, \ \|T_H^{(2)}\|_\infty \leq C h \|u\|_{C^{1,1}(\overline{\Omega})}.$$

We establish now an estimate for λ. We start by noting that

$$\mathbb{1}\lambda = -L_H u_H + \tilde{f}_H.$$

As $< \mathbb{1}, L_H u_H >= 0$, we get

$$\begin{aligned} \lambda &= \frac{\mathbb{1}^t \tilde{f}_H}{\mathbb{1}^t \mathbb{1}} \\ &= \frac{1}{(n-1)^2} \left(\sum_{P \in \Omega_H} f_H(P) + \frac{1}{h} \sum_{P \in \partial\Omega_H^*} g_H(P) \right) \\ &= \frac{1}{h^2 (n-1)^2} \left(h^2 \sum_{P \in \Omega_H} f_H(P) + h \sum_{P \in \partial\Omega_H^*} g_H(P) \right). \end{aligned}$$

We also have

$$\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} \int_{y-\frac{h}{2}}^{y+\frac{h}{2}} f\,dy\,dx = h^2 f(x,y) + I_H,$$

where

$$|I_H| \leq Ch^3 \|f\|_{C^{0,1}(\overline{\Omega})}.$$

Then we obtain

$$\int_{\Omega} f\,dx\,dy = h^2 \sum_{P\in\Omega_H} f(P) + I_\Omega,$$

with

$$|I_\Omega| \leq Ch \|f\|_{C^{0,1}(\overline{\Omega})}.$$

Analogously, it can be shown that

$$\int_{\partial\Omega} g\,ds = h \sum_{P\in\partial\Omega'_H} g(P) + I_{\partial\Omega},$$

where

$$|I_{\partial\Omega}| \leq Ch \|g\|_{C^{0,1}(\partial\Omega)}.$$

As the compatibility condition in the continuous context

$$\int_{\Omega} f\,dxdy + \int_{\partial\Omega} g\,ds = 0$$

holds, we get the desired estimate for λ

$$|\lambda| \leq Ch\big(\|f\|_{C^{0,1}(\overline{\Omega})} + \|g\|_{C^{0,1}(\partial\Omega)}\big). \tag{2.2.27}$$

Let E_H be defined by $E_H = R_H u - u_H$. This error satisfies

$$\begin{bmatrix} L_H & \mathbb{1} \\ \mathbb{1}^t & 0 \end{bmatrix} \begin{bmatrix} E_H \\ -\lambda \end{bmatrix} = \begin{bmatrix} T_H^{(1)} + \psi(T_H^{(2)}) \\ \sigma \end{bmatrix},$$

where $\psi(T_H^{(2)})$ is a certain function of $T_H^{(2)}$ and σ is an arbitrary constant. It is easy to establish that

$$\begin{bmatrix} L_H & \mathbb{1} \\ \mathbb{1}^T & 0 \end{bmatrix} \begin{bmatrix} E_H - d\mathbb{1} \\ -\lambda \end{bmatrix} = \begin{bmatrix} T_H^{(1)} + \psi(T_H^{(2)}) \\ \mathbb{1}^t E_H - d\mathbb{1}^t\mathbb{1} \end{bmatrix}.$$

If we take $\sigma = 0$, or equivalently

$$c = \frac{1}{\mathbb{1}^t\mathbb{1}} \mathbb{1}^t E_H,$$

we obtain

$$\begin{bmatrix} L_H & \mathbb{1} \\ \mathbb{1}^t & 0 \end{bmatrix} \begin{bmatrix} E_H - d\mathbb{1} \\ 0 \end{bmatrix} = \begin{bmatrix} T_H^{(1)} + \psi(T_H^{(2)}) \\ 0 \end{bmatrix},$$

which is induced by the FDM

$$\begin{cases} -\Delta_H(R_H u - u_H - d\mathbb{1}) = T_H^{(1)} \text{ in } \Omega_H, \\ B_\eta(R_H u - u_H) = T_H^{(2)} \text{ on } \partial\Omega_H - V. \end{cases}$$

Using the stability of the FDM (see [12], Section 4.7.4) we obtain

$$\|u_H - R_H u - d\|_\infty \leq C\left(h\|u\|_{C^{1,1}(\overline{\Omega})} + h^2\|u\|_{C^{3,1}(\overline{\Omega})} + |\lambda|\right). \tag{2.2.28}$$

We study in what follows another discretization of the Poisson equation with Neumann boundary conditions with a symmetric discretization of the boundary conditions. In order to define the boundary discretization we introduce the fictitious points:

$$(x_{-1}, y_j) = (-h, y_j),\ (x_{n+1}, y_j) = (1+h, y_j), j = 0, \ldots, n,$$

$$(x_i, y_{-1}) = (x_i, -h),\ (x_i, y_{n+1}) = (x_i, 1+h), i = 0, \ldots, n.$$

The previous auxiliary points enable us to use the five-point formula in $\overline{\Omega}_H$.

We consider the FDM

$$\begin{cases} -\Delta_H u_H = f_H \text{ in} \overline{\Omega}_H, \\ B_\eta u_H = g_H \text{ on } \partial\Omega_H, \end{cases} \tag{2.2.29}$$

where

$$B_\eta u_H(P) = \frac{u_H(P+h\eta) - u_H(P-h\eta)}{2h},\ P \in \partial\Omega_H.$$

At any point in V we should consider two normals with respect to the normals of both sides of $\partial\Omega$.

The FDM (2.2.29) induces the linear system

$$L_H u_H = \tilde{f}_H, \tag{2.2.30}$$

where

$$L_H = \frac{1}{h^2}\begin{bmatrix} T & -2I & 0 & \ldots & 0 & 0 \\ -I & T & -I & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -2I & T \end{bmatrix}, \tag{2.2.31}$$

$$T = \begin{bmatrix} 4 & -2 & 0 & \ldots & 0 & 0 \\ -1 & 4 & -1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -2 & 4 \end{bmatrix}$$

and

$$\tilde{f}_h = \begin{cases} \begin{cases} f_H(x_0, y_0) + \frac{4}{h} g_H(x_0, y_0), \\ f_H(x_i, y_0) + \frac{2}{h} g_H(x_i, y_0), i = 2, \ldots, n-1, \\ f_H(x_n, y_0) + \frac{4}{h} g_H(x_n, y_0), \end{cases} \\ \begin{cases} f_H(x_0, y_j) + \frac{2}{h} g_H(x_0, y_j), \\ f_H(x_i, y_j),\ i = 1, \ldots, n-1, \\ f_H(x_n, y_j) + \frac{2}{h} g_H(x_n, y_j), \qquad j = 1, \ldots, n-1, \end{cases} \\ \begin{cases} f_H(x_0, y_n) + \frac{4}{h} g_H(x_0, y_n), \\ f_H(x_i, y_n) + \frac{2}{h} g_H(x_i, y_n), i = 1, \ldots, n-1, \\ f_H(x_n, y_n) + \frac{4}{h} g_H(x_n, y_n). \end{cases} \end{cases} \tag{2.2.32}$$

The nonsymmetric structure of L_H requires the use of a linear transformation D_H such that $D_H L_H$ is a symmetric matrix. Let D_H be defined by

$$D_H = \begin{bmatrix} D & 0 & 0 & \dots & 0 & 0 \\ 0 & D_1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & D \end{bmatrix}, \; D = \begin{bmatrix} \frac{1}{4} & 0 & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{4} \end{bmatrix}$$

and

$$D_1 = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{2} \end{bmatrix}.$$

Using the transformation D_H, the FDM (2.2.30) can be rewritten in the equivalent form

$$D_H L_H u_H = D_H \tilde{f}_H,$$

where

$$\tilde{L}_H := D_H L_H = \frac{1}{h^2} \begin{bmatrix} \hat{T}_1 & \hat{T}_2 & 0 & \dots & 0 & 0 \\ \hat{T}_2 & \hat{T} & \hat{T}_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \hat{T}_2 & \hat{T}_1 \end{bmatrix},$$

with

$$\hat{T}_1 = \begin{bmatrix} 1 & -\frac{1}{2} & 0 & \dots & 0 & 0 \\ -\frac{1}{2} & 2 & -\frac{1}{2} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\frac{1}{2} & 1 \end{bmatrix},$$

$$\hat{T} = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 & 0 \\ -1 & 4 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 2 \end{bmatrix}$$

and

$$\hat{T}_2 = \begin{bmatrix} -\frac{1}{2} & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & -\frac{1}{2} \end{bmatrix}.$$

We observe that $\tilde{L}_H$ is singular because $\tilde{L}_H \mathbb{1} = 0$. From $\tilde{L}_H$ a irreducibly diagonally dominant matrix can be defined eliminating a row and the correspondent column. As $car(\tilde{L}_H) = (n+1)^2 - 1$ and $\mathcal{N}(\tilde{L}_H) = \mathcal{L}(\mathbb{1})$, we deduce that there exists a unique solution of the FDM (2.2.29) if and only if

$$\mathbb{1}^t D_H \tilde{f}_H = 0.$$

Moreover, any two solutions of the boundary value problem (2.2.29) may differ by a constant.

The solution u_h can be computed solving the linear system

$$\begin{bmatrix} \tilde{L}_H & \mathbb{1} \\ \mathbb{1}^t & 0 \end{bmatrix} \begin{bmatrix} u_H \\ \lambda \end{bmatrix} = \begin{bmatrix} D_H \tilde{f}_H \\ \sigma \end{bmatrix}.$$

If we obtain $\lambda = 0$ for $\sigma = \mathbb{I}^t u_H$, then u_H is solution of the initial finite difference discretization. Otherwise, u_H is solution of a perturbed problem similar to the perturbed problem (2.2.26). Following the construction of the estimates (2.2.27), (2.2.28), estimates for λ and for the error $u_H - r_H u - d\mathbb{I}$ can be established.

2.2.6 Convergence Analysis with Respect to Discrete Sobolev Norms

It was shown that the five-point formula enable us to obtain a second order approximation u_H for the solution u of the Poisson equation with Dirichlet boundary conditions, that is

$$\|R_H u - u_H\|_\infty \leq Ch^2.$$

We intent to define a new norm which can be seen as a discretization of a continuous one and such that u_H is also a second order approximation with respect to this new norm.

The grid function $e_H = R_H u - u_H$ is defined on $\overline{\Omega}_H$ and it is null on $\partial\Omega_H$. Let $W_0(\overline{\Omega}_H)$ be the set of grids functions defined $\overline{\Omega}_H$ and null on $\partial\Omega_H$. In this space we introduce the norm

$$\|w_H\|_1^2 = \|w_H\|^2 + \sum_{\Omega_{H,l}} h(D_{-x}w_H(P))^2 + \sum_{\Omega_{H,t}} h(D_{-y}w_H(P))^2. \tag{2.2.33}$$

In (2.2.33), the notations

$$\Omega_{H,l} = \overline{\Omega}_H - \{(x_0, y_j) \in \partial\Omega_H\}, \Omega_{H,t} = \overline{\Omega}_H - \{(x_i, y_0) \in \partial\Omega_H\}$$

and

$$\|w_H\|^2 = h^2 \sum_{\Omega_H} w_H(P)^2 \tag{2.2.34}$$

were used. We point out that the norm (2.2.34) is induced by the inner product

$$(w_H, v_H) = h^2 \sum_{\Omega_H} w_H(P)v_H(P), w_H, w_H \in W_0(\overline{\Omega}_H).$$

As

$$w_H(x_i, y_j) = \sum_{\ell=0}^{i} hD_{-x}w_H(x_\ell, y_j),$$

we have

$$\sum_{\Omega_H} h^2 w_H(P)^2 \leq C \sum_{\Omega_{H,l}} h^2 (D_{-x}w_H(P))^2,$$

where C is a positive H-independent constant. From the last inequality we deduce that

$$\|w_H\|^2 \leq C \left(\sum_{\Omega_{H,l}} h^2 (D_{-x}w_H(P))^2 + \sum_{\Omega_{H,t}} h^2 (D_{-y}w_H(P))^2 \right), \forall w_H \in W_0(\overline{\Omega}_H). \tag{2.2.35}$$

For $w_H \in W_0(\overline{\Omega}_H)$, $-\Delta_H w_h$ can be identified with a linear functional in the dual of $W_0(\overline{\Omega}_H)$.[12]

This remark gives sense to the next result:

Theorem 2.2.17 *There exists a positive constant C, H-independent, such that*

$$\| -\Delta_H w_H\|_{-1} \geq C\|w_H\|_1, \forall w_H \in W_0(\overline{\Omega}_H). \tag{2.2.36}$$

Proof: To prove (2.2.36) we note that

$$\begin{aligned}
\| -\Delta w_H\|_{-1} &= \sup_{0\neq v_H\in W_0(\overline{\Omega}_H)} \frac{|-\Delta w_H(v_H)|}{\|v_H\|_1} \\
&= \sup_{0\neq v_H\in W_0(\overline{\Omega}_H)} \frac{|(-\Delta w_H, v_H)|}{\|v_H\|_1} \\
&= \sup_{0\neq v_H\in W_0(\overline{\Omega}_H)} \frac{|\sum_{\Omega_{H,l}} h^2 D_{-x}w_H(P)D_{-x}v_H(P) + \sum_{\Omega_{H,t}} h^2 D_{-y}w_H(P)D_{-y}v_H(P)|}{\|v_H\|_1} \\
&\geq \frac{|\sum_{\Omega_{H,l}} h^2 (D_{-x}w_H(P))^2 + \sum_{\Omega_{H,t}} h^2 (D_{-y}w_H(P))^2|}{\|w_H\|_1} \\
&\geq C\|w_H\|_1.
\end{aligned}$$

■

From Theorem 2.2.17, $-\Delta_H$ is injective. Taking in (2.2.36), w_H replaced e_H, we obtain

$$\|T_H\|_{-1} \geq C\|e_H\|_1.$$

As we have

$$\|T_H\|_{-1} = \sup_{0\neq v_H\in W_0(\overline{\Omega}_H)} \frac{|T_H(v_H)|}{\|v_H\|_1} = \sup_{0\neq v_H\in W_0(\overline{\Omega}_H)} \frac{|(T_H, v_H)|}{\|v_H\|_1} \leq Ch^2\|u\|_{C^4(\overline{\Omega})},$$

we conclude that

$$\|e_H\|_1 \leq Ch^2\|u\|_{C^4(\overline{\Omega})}$$

provided that $u \in C^4(\overline{\Omega})$.

It can be also shown that

$$\|e_H\|_1 \leq Ch^2\|u\|_{C^{3,1}(\overline{\Omega})},$$

provided that $u \in C^{3,1}(\overline{\Omega})$.

[12]Let V be a Hilbert space. By V' we denote its dual, that is the space of all bounded linear mappings of V onto $\mathbb{R}$. V' is a Banach space with respect to the dual norm

$$\|\ell\|_{-1} = \sup_{u\in V, u\neq 0} \frac{|\ell(u)|}{\|u\|_V}.$$

The identification between a Hilbert space and its dual is based on the Riesz representation theorem.

Theorem 2.2.16 *Let V be a Hilbert space and $\ell \in V'$. Then there exists a unique $u_\ell \in V$ such that*

$$\ell(u) = (u, u_\ell)_V, \forall u \in V,\ \|\ell\|_{-1} = \|u_\ell\|_V.$$

2.3 Tools of Functional Analysis

Space of Integrable Functions We introduce the a class of spaces that consists of (Lebesgue)-integrable functions. Let p be a real number, $p \geq 1$, and let Ω be a open subset of $\mathbb{R}^n$. By $L^p(\Omega), p \geq 1$, we denote the set of all functions such that

$$\int_\Omega |u(x)|^p \, dx < \infty.$$

Any two functions which are equal almost everywhere on Ω are identified with each other. In $L^p(\Omega)$ we consider the norm

$$\|u\|_{L^p(\Omega)} = \Big(\int_\Omega |u(x)|^p \, dx \Big)^{1/p}.$$

The particular case $p = 2$ has an important role in the sequel. In this case, the norm $\|.\|_{L^2(\Omega)}$ is induced by the inner product

$$(u, v) = \int_\Omega u(x)v(x) \, dx.$$

By $L^\infty(\Omega)$ we represent the set of all functions u defined on Ω such that $|u|$ has finite essential supremum over Ω (there exists $M > 0$ such that $u \leq M$ in Ω all most everywhere (in $\Omega - \Omega^*$ where $meas(\Omega^*) = 0$) and the smallest M is called essential supremum of u and it is denoted by $ess.supu$). In $L^\infty(\Omega)$ we consider the norm

$$\|u\|_{L^\infty(\Omega)} = ess.sup|u|.$$

For $p \in [1, \infty]$, the space $L^p(\Omega)$ is a Banach space[13] For $p = 2$, $L^2(\Omega)$ is a Hilbert space.

Sobolev spaces Let Ω be a open subset of $\mathbb{R}^n$ and $u \in C^m(\Omega)$. If $v \in C_0^\infty$, then

$$\int_\Omega D^\alpha uv dx = (-1)^{|\alpha|} \int_\Omega u D^\alpha v dx,$$

for $|\alpha| \leq m$.

Consider $\Omega = \mathbb{R}$ and $u(x) = (1 - |x|)_+$. This function satisfies

$$\int_{\mathbb{R}} u(x)v'(x)dx = -\int_{\mathbb{R}} w(x)v(x)dx, \forall v \in C_0^\infty(\mathbb{R}),$$

where

$$w(x) = \begin{cases} 0, \, x < -1, \\ 1, \, x \in (-1, 0), \\ -1, \, x \in (0, 1), \\ 0, \, x > 1. \end{cases}$$

The function w can not be seen as the usual derivative of u, but can be interpreted as a "weak" derivative of the given function. This example motivate the introduction of the concept of weak derivative.

[13] A normed linear space B is called a Banach space if all Cauchy sequences in B converge in B.

Let Ω be an open set of $\mathbb{R}^n$ and u be locally integrable on Ω (u is integrable on every bounded ω subset of Ω with $\overline{\omega} \subset \Omega$). If there exists a function w_α, locally integrable on Ω such that

$$\int_\Omega w_\alpha v dx = (-1)^{|\alpha|} \int_\Omega u D^\alpha v dx,$$

for all $v \in C_0^\infty(\Omega)$, $|\alpha| \leq m$, then we say that w_α is the weak derivative of the function u of order $|\alpha|$.

If u, locally integrable on Ω, has two weak derivatives of order $|\alpha|$, w_α, w_α^*, then

$$\int_\Omega (w_\alpha - w_\alpha^*) v \, dx = 0 \, \forall v \in C_0^\infty(\Omega),$$

and consequently $w_\alpha = w_\alpha^*$.

The base for the definition of the Sobolev spaces is the concept of the weak derivative. Let m be non negative integer and $p \in [1, \infty]$. The Sobolev space of order m is given by

$$W^{m,p}(\Omega) = \{u \in L^p(\Omega) : \, D^\alpha \in L^p(\Omega), \, |\alpha| \leq m\},$$

equipped with the norm

$$\|u\|_{W^{m,p}(\Omega)} = \Big(\sum_{|\alpha| \leq m} \|D^\alpha u\|^p_{L^p(\Omega)} \Big)^{1/p},$$

if $p \in [1, \infty)$ and

$$\|u\|_{W^{m,\infty}(\Omega)} = \sum_{|\alpha| \leq m} \|D^\alpha u\|_{L^\infty(\Omega)},$$

for $p = \infty$.

The previous norms can be given by:

- for $p \in [1, \infty)$

$$\|u\|_{W^{m,p}(\Omega)} = \Big(\sum_{j=0}^{m} |u|^p_{W^{j,p}(\Omega)} \Big)^{1/p},$$

  with

$$|u|_{W^{j,p}(\Omega)} = \Big(\sum_{|\alpha| = j} \|D^\alpha u\|^p_{L^p(\Omega)} \Big)^{1/p},$$

- $p = \infty$

$$\|u\|_{W^{m,\infty}(\Omega)} = \sum_{j=0}^{m} |u|^p_{W^{j,\infty}(\Omega)},$$

  with

$$|u|_{W^{j,\infty}(\Omega)} = \sum_{|\alpha| = j} \|D^\alpha u\|_{L^\infty(\Omega)}.$$

The particular case $p = 2$ is very useful in a huge number of applications. In this case, $W^{m,2}(\Omega)$ is a Hilbert space with respect to the inner product

$$(u, v)_{W^{m,2}(\Omega)} = \sum_{|\alpha| \leq m} (D^\alpha u, D^\alpha v),$$

and this space is denoted by $H^m(\Omega)$. It can be shown that $C^\infty(\Omega) \cap H^m(\Omega)$ is dense in $H^m(\Omega)$.

For the particular choice $m = 1$, we introduce the subset of all $u \in H^1(\Omega)$ which are the limit of a sequence in $C_0^\infty(\Omega)$, that is, the closure of $C_0^\infty(\Omega)$. We denote this space by $H_0^1(\Omega)$. If the boundary $\partial\Omega$ is smooth (for instance if Ω is a polygonal domain of $\mathbb{R}^2$ or a polyhedron in $\mathbb{R}^3$)

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : u = 0 \,\text{on}\, \partial\Omega\}.$$

[14] We remark that $H_0^1(\Omega)$ is a Hilbert space with the same norm and inner product as $H^1(\Omega)$.

The Poincaré-Friedrichs inequality

$$\|u\|_{L^2(\Omega)} \leq C(\Omega)\Big(\sum_{i=1}^{n} \|\frac{\partial u}{\partial x_i}\|^2_{L^2(\Omega)}\Big)^{1/2}$$

holds for $u \in H_0^1(\Omega)$, provided that Ω is bounded. The proof can be considered for $u \in C_0^\infty(\Omega)$ and the result holds for $u \in H_0^1(\Omega)$ because $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$.

2.4 Weak Solutions for Elliptic Problems

2.4.1 Variational Problems for Elliptic BVP

Let Ω be a bounded open set of $\mathbb{R}^n$, and we consider the second order differential equation

$$Au = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\big(a_{ij}(x)\frac{\partial u}{\partial x_i}\big) + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x),\ x \in \Omega, \tag{2.4.1}$$

where

$$a_{ij} \in C^1(\overline{\Omega}), b_i, c, f \in C(\overline{\Omega}).$$

We assume that

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \tilde{C}\sum_{i=1}^{n}\xi_i^2,\ \xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n,\ x \in \overline{\Omega}. \tag{2.4.2}$$

which is usually referred to as uniform ellipticity. The condition (2.4.2) implies that $[a_{ij}(x)]$ has positive eigenvalues and then (2.4.1) is an elliptic equation.

Let us consider the boundary value problem (2.4.1) with homogeneous Dirichlet boundary conditions. In many applications where non smooth data are presented, there isn't a classical solution of this boundary value problem, that is a function u in $C^2(\Omega) \cap C(\overline{\Omega})$ satisfying the PDEs (2.4.1) and $u = 0$ on $\partial\Omega$.

In order to overcome the limitation of the classical theory and to be able to deal with PDEs with non smooth data, we generalise the notion of solution by weakening the differentiability requirements and introducing the variational problems induced by the PDEs.

Let u be the classical solution of the introduced boundary value problem. From (2.4.1), for $v \in C_0^1(\Omega)$, we obtain

$$\sum_{i,j=1}^{n}(a_{ij}\frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_j}) + \sum_{i=1}^{n}(b_i\frac{\partial u}{\partial x_i}, v) + (cu, v) = (f, v). \tag{2.4.3}$$

[14] This characterization holds for a domain Ω with boundary C^1.

In order to this equality makes sense we do not need to assume that $u \in C^2(\Omega)$. It is sufficient to suppose that $u \in L^2(\Omega)$ and $\dfrac{\partial u}{\partial x_i} \in L^2(\Omega), i = 1, \ldots, n$. As $u = 0$ on $\partial\Omega$, it is natural to seek u in $H_0^1(\Omega)$. Furthermore, as $C_0^1(\Omega) \subset H_0^1(\Omega)$ the equality (2.4.3) has sense for $v \in H_0^1(\Omega)$. Therefore, we replace the computation of u in $C^2(\Omega) \cap C(\overline{\Omega})$ such that $u = 0$ on $\partial\Omega$ by the following problem:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u, v) = \ell(v),\ \forall v \in H_0^1(\Omega), \tag{2.4.4}$$

where

$$a(w, v) = \sum_{i,j=1}^{n} (a_{ij} \frac{\partial w}{\partial x_i}, \frac{\partial w}{\partial x_j}) + \sum_{i=1}^{n} (b_i \frac{\partial w}{\partial x_i}, v) + (cw, v),\ w, v \in H_0^1(\Omega), \tag{2.4.5}$$

and

$$\ell(v) = (f, v),\ v \in H_0^1(\Omega). \tag{2.4.6}$$

The smoothness requirements on a_{ij}, b_i, c presented before can be weakened considering that these coefficients belong to $L^\infty(\Omega)$.

The solution of the problem (2.4.4) is called weak solution of the equation (2.4.1) complemented with homogeneous Dirichlet boundary conditions. It is clear that if u is a classical solution of (2.4.4) such that $u = 0$ on $\partial\Omega$, then u is also weak solution. Nevertheless, if $u \in H_0^1(\Omega)$ is weak solution of this problem, then u is classical solution of the same problem if u is smooth enough. In fact, if $u \in C^2(\Omega)$ and $u = 0$ on $\partial\Omega$, then

$$(Au - f, v) = 0,\ \forall v \in C_0^\infty(\Omega).$$

Consequently, $Au = f$ almost everywhere in Ω.

In order to study the existence of the solution of the problem (2.4.4), usually called variational problem because it is related with the computation of the solution of a minimization problem, we introduce in what follows some concepts and results associated with general variational problems.

2.4.2 General Variational Problems

Let V be a Hilbert space with the inner product $(.,.)$ and let $a(.,.) : V \times V \to \mathbb{R}$ be a bilinear form,that is, $a(.,.)$ is linear in each argument. Let ℓ be in V'.

We consider in what follows the general variational problem:

$$\text{find } u \in V \text{such that } a(u, v) = \ell(v),\ \forall v \in V. \tag{2.4.7}$$

The existence and uniqueness of the solution of the variational problem (2.4.7) are guaranteed imposing some requirements on the bilinear form $a(.,.)$.

If there exists a positive constant C such that

$$|a(u, v)| \leq C\|u\|_V \|v\|_V,$$

then $a(.,.)$ is said bounded (or continuous). If

$$a(u, u) \geq C_e \|u\|_V^2,\ \forall u, v \in V,$$

for some positive constant, then we say that $a(.,.)$ is $V-$ elliptic.

Lemma 2 (Lax-Milgram Lemma) *If $a(.,.)$ is a continuous V-elliptic bilinear form, then the problem (2.4.7) has a unique solution u in V and the operator $P : V' \to V$ defined by*

$$P\ell = u,\ \ell \in V'$$

is continuous.

Proof: Attending that $\ell \in V'$, by the Riesz Representation Theorem, there exists $P\ell \in V$ such that

$$(P\ell, v)_V = \ell(v)\, \forall v \in V.$$

Let u be fixed in V. The linear functional $a(u,.) : V \to \mathbb{R}$ belongs to V'. By the Riesz Representation Theorem, there exists $\mathcal{A}u$ in V such that

$$(\mathcal{A}u, v) = a(u,v),\ \forall v \in V. \tag{2.4.8}$$

Let $\mathcal{A} : V \to V$ be defined by (2.4.8). This operator is linear and satisfies

$$\|\mathcal{A}u\|_V = \sup_{v\in V, v\neq 0} \frac{|a(u,v)|}{\|v\|_V} \leq C\|u\|_V.$$

Then $\mathcal{A}$ is continuous.

Using the two operators P and $\mathcal{A}$, the variational problem (2.4.7) can be rewritten in the equivalent for:

$$\text{find}\, u \in V \,\text{such that}\, \mathcal{A}u = P\ell. \tag{2.4.9}$$

We prove that $\mathcal{A}$ is bijective.

- $\mathcal{A}$ is injective:

  As $a(.,.)$ is V-elliptic, we have

  $$C_e\|v\|_V^2 \leq a(v,v) = (\mathcal{A}v, v) \leq \|\mathcal{A}v\|_V\|v\|_V,\ \forall v \in V,$$

  and then

  $$C_e\|v\|_V \leq \|\mathcal{A}v\|_V\, \forall v \in V. \tag{2.4.10}$$

  This inequality implies the injectivity of $\mathcal{A}$.

- $\mathcal{A}$ satisfies $\mathcal{A}V = V$:

  We prove that $\mathcal{A}(V) = \{\mathcal{A}v, v \in V\}$ is closed in V and $\mathcal{A}(V)^\perp = \{0\}$.

  Let w be in $\overline{\mathcal{A}(V)}$ and let $(\mathcal{A}v_n)$ be a sequence, in $\mathcal{R}(\mathcal{A})$, that converges to w. As

  $$\|\mathcal{A}v_n - \mathcal{A}v_m\|_V \geq \|v_n - v_m\|_V,$$

  (v_n) is a Cauchy sequence in V. So exists v in V such that $v_n \to v$ because V is an Hilbert space. Furthermore, by continuity,

  $$\mathcal{A}v_n \to \mathcal{A}v.$$

  Finally, as $\mathcal{A}v_n \to w$ we conclude that $w = \mathcal{A}v \in \mathcal{R}(\mathcal{A})$.

  Let v_0 be in $\mathcal{A}(V)^\perp$. As

  $$C_e\|v_0\|_V \leq (\mathcal{A}v_0, v_0) = 0$$

  we get $v_0 = 0$ and consequently $\mathcal{A}(V)^\perp = \{0\}$.

As problem (2.4.9) is equivalent to the variational problem (2.4.7), we conclude that there exists a unique solution u of the last problem. Moreover, from (2.4.10) we obtain

$$\|\mathcal{A}^{-1}w\|_V \leq \frac{1}{C_e}\|w\|_V,\ w \in V,$$

that is $\mathcal{A}^{-1}$ is continuous and

$$\|u\|_V = \|\mathcal{A}^{-1}P\ell\|_V \leq \frac{1}{C_e}\|P\ell\|_V = \frac{1}{C_e}\|\ell\|_{-1}.$$

■

The variational problem (2.4.7) is associated with a minimization problem when symmetric variational forms are considered. The bilinear form $a(.,.)$ is said symmetric if $a(u,v) = a(v,u)$ for $u, v \in V$.

Let ℓ be fixed in V' and let $J : V \to \mathbb{R}$ be defined by

$$J(v) = a(v,v) - 2\ell(v),\ v \in V.$$

The solution of the variational problem (2.4.7) is related with the solution of the minimization problem

$$\text{find } u \in V \text{such that } J(u) = \min_{v \in V} J(v). \tag{2.4.11}$$

Theorem 2.4.1 *Let $a(.,.)$ be V-elliptic and symmetric. If $\ell \in V'$ then the solution, then the solution of the variational problem (2.4.7) is the unique solution of the minimization problem (2.4.11).*

Proof: Let u be the solution of the variational problem (2.4.7) and let v be in V. Then, for $z = u - v$, we have

$$\begin{aligned} J(v) = J(z+u) &= J(u) + a(z,z) + 2(a(u,z) - \ell(z)) \\ &= J(u) + a(z,z) \\ &\geq J(u) + C_e\|z\|_V = J(u) + C_e\|u - v\|_V. \end{aligned}$$

Consequently, $J(v) \geq J(u)$ for $v \neq u$.

■

The concept of V-ellipticity seems to indicate that elliptic boundary value problems correspond V-elliptic bilinear forms. In general elliptic boundary value problems are associated with V-coercive bilinear forms. The definition of V-coercivity requires another space U such that $V \subset U \subset V'$. By the Riesz Representation Theorem the last inclusions has sense. The Hilbert space U is such that $\overline{V} = U$ and the identity operator $i : V \to U$ is continuous, that is $\|v\|_U \leq C\|v\|_V, v \in V$. In this context, we say that a bilinear form $a(.,.) : V \times V \to \mathbb{R}$ is V-coercive if

$$a(v,v) \geq C_e\|v\|_V^2 - C_c\|v\|_U^2,\ v \in V, \tag{2.4.12}$$

where $C_e > 0$.

The existence and uniqueness of the solution of a variational problem with a V-coercive bilinear form is established by using the Riesz-Schauder theory.

2.4.3 Again Variational Problems for Elliptic Equations

Dirichlet homogeneous boundary conditions: We return to the variational problem (2.4.4) where the bilinear form $a(.,.) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is defined by (2.4.5). Under the regularity assumptions $a_{ij}, b_i, c \in L^\infty(\Omega)$, we have

$$\begin{aligned} |a(w,v)| &\leq \hat{C}\Big(\sum_{ij}\int_\Omega |\frac{\partial w}{\partial x_i}\frac{\partial v}{\partial x_j}|dx + \sum_i \int_\Omega |\frac{\partial w}{\partial x_i}v|dx + \int_\Omega |wv|dx\Big) \\ &\leq 2n\hat{C}\|w\|_{H_0^1(\Omega)}\|v\|_{H_0^1(\Omega)} \end{aligned} \tag{2.4.13}$$

for $w, v \in H_0^1(\Omega)$, where we used the notation

$$\hat{C} = \max\{\max_{ij}\max_{x\in\overline{\Omega}}|a_{ij}(x)|, \max_i \max_{x\in\overline{\Omega}}|b_i(x)|, \max_{x\in\overline{\Omega}}|c(x)|\}.$$

Consequently, $a(.,.)$ is continuous on $H_0^1(\Omega) \times H_0^1(\Omega)$.

The $H_0^1(\Omega)$-ellipticity of $a(.,.)$ can be deduced from the uniform ellipticity of the operator A. In fact, from the condition (2.4.2) we easily obtain

$$\begin{aligned} a(u,u) &\geq \tilde{C}\sum_{i=1}^n \int_\Omega |\frac{\partial u}{\partial x_i}|^2\,dx + \sum_{i=1}^b \int_\Omega b_i(x)\frac{1}{2}\frac{\partial}{\partial x_i}(u^2)\,dx + \int_\Omega cu^2\,dx \\ &= \tilde{C}\sum_{i=1}^n \int_\Omega |\frac{\partial u}{\partial x_i}|^2\,dx + \int_\Omega \Big(c - \frac{1}{2}\sum_{i=1}^n \frac{\partial b_i}{\partial x_i}\Big)u^2\,dx. \end{aligned}$$

If we suppose that the coefficient functions c and b_i satisfy

$$c - \frac{1}{2}\sum_{i=1}^n \frac{\partial b_i}{\partial x_i} \geq 0,\ x \in \overline{\Omega}, \tag{2.4.14}$$

then

$$a(u,u) \geq \tilde{C}\sum_{i=1}^n \int_\Omega |\frac{\partial u}{\partial x_i}|^2,$$

and, by the Poincaré-Friedrichs inequality, we conclude that $a(.,.)$ is $H_0^1(\Omega)$-elliptic.

As for linear functional ℓ holds the following

$$\ell(v) = (f,v) \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_{H_0^1(\Omega)},\ v \in H_0^1(\Omega)$$

the Lax-Milgram Lemma allow us to conclude the next result:

Theorem 2.4.2 *If the coefficient functions are such that*

$$a_{ij}, b_i, c \in L^\infty(\Omega)$$

and the conditions (2.4.2), (2.4.14) hold, then the variational problem (2.4.4) has a unique weak solution in $H_0^1(\Omega)$. Moreover

$$\|u\|_{H_0^1(\Omega)} \leq \frac{1}{C_e}\|f\|_{L^2(\Omega)}. \tag{2.4.15}$$

Proof: We only prove that the estimate (2.4.15) holds. As

$$C_e\|u\|^2_{H^1_0(\Omega)} \leq a(u,u) = (f,u) \leq \|f\|_{L^2(\Omega)}\|u\|_{H^1_0(\Omega)},$$

we conclude the proof.

■

Theorem 2.4.2 can be established under weaker smoothness assumptions on $b_i, i = 1,\ldots,n$, than those consider before. In fact, if we only require that $b_i \in L^\infty(\Omega)$, then we can prove that $a(.,.)$ is $H^1_0(\Omega)$-elliptic. We have

$$a(v,v) \geq \tilde{C}|v|^2_{H^1_0(\Omega)} - \max_{i=1,\ldots,n}\|b_i\|_{L^\infty(\Omega)}\Big|v|_{H^1_0(\Omega)}\|v\|_{L^2(\Omega)} + (cv,v).$$

As

$$\max_{i=1,\ldots,n}\|b_i\|_{L^\infty(\Omega)}|v|_{H^1_0(\Omega)}\|v\|_{L^2(\Omega)} \leq \epsilon^2|v|_{H^1_0(\Omega)} + \frac{1}{4\epsilon^2}\max_{i=1,\ldots,n}\|b_i\|_{L^\infty(\Omega)}\|v\|^2_{L^2(\Omega)},$$

we obtain

$$a(v,v) \geq \frac{\tilde{C}}{2}|v|^2_{H^1_0(\Omega)} + \big(\frac{\tilde{C}}{2} - \epsilon^2\big)|v|^2_{H^1_0(\Omega)} - \frac{1}{4\epsilon^2}\|v\|^2_{L^2(\Omega)}\max_{i=1,\ldots,n}\|b_i\|^2_{L^\infty(\Omega)} + (cv,v).$$

Taking $\epsilon^2 = \dfrac{\tilde{C}}{2}$ we get

$$a(v,v) \geq \frac{\tilde{C}}{2}|v|^2_{H^1_0(\Omega)} + \int_\Omega \Big(c - \frac{1}{2\tilde{C}}\max_{i=1,\ldots,n}\|b_i\|^2_{L^\infty(\Omega)}\Big)v(x)^2\,dx.$$

If

$$c - \frac{1}{2\tilde{C}}\max_{i=1,\ldots,n}\|b_i\|^2_{L^\infty(\Omega)} \geq 0,$$

then

$$a(v,v) \geq \frac{\tilde{C}}{2}|v|^2_{H^1_0(\Omega)}$$

and, by the Poincaré-Friedrichs inequality, we conclude that $a(.,.)$ is $H^1_0(\Omega)$-elliptic.

An immediate corollary of the Theorem 2.4.2 is the stability of the solution of the variational problem (2.4.4) with respect to perturbations of f. In fact, let $u_i, i = 1,2$, be solutions in $H^1_0(\Omega)$ of the variational problem (2.4.4) for $f_i \in L^2(\Omega), i = 1,2$, respectively. As $f = f_1 - f_2$ is in $L^2(\Omega)$, then, by Theorem 2.4.2, we have

$$\|u_1 - u_2\|_{H^1_0(\Omega)} \leq \frac{1}{C_e}\|f_1 - h_2\|_{L^2(\Omega)}.$$

Thus, if $\|f_1 - h_2\|_{L^2(\Omega)}$ is small, $\|u_1 - u_2\|_{H^1_0(\Omega)}$ remains small.

Non homogeneous boundary conditions:

Let us consider the PDEs (2.4.1) with the non homogeneous Dirichlet boundary condition

$$u = g \ \text{ on } \partial\Omega. \tag{2.4.16}$$

From the PDEs (2.4.1), with $v \in C_0^\infty(\Omega)$, we obtain the variational problem (2.4.4). However, its solution does not satisfy the prescribed boundary condition. This fact leads to the definition of the variational problem

$$\text{find } u \in H^1(\Omega) \text{ such that } u = g \text{ on } \partial\Omega \text{ and } a(u,v) = \ell(v),\ \forall v \in H_0^1(\Omega). \tag{2.4.17}$$

It can be shown that if $u \in H^1(\Omega)$ is solution of the variational problem (2.4.17), and u is smooth enough, then u satisfies (2.4.1) and $u = g$ on the boundary $\partial\Omega$.

In order to compute a solution of the variational problem (2.4.17) we start by fixing $u_0 \in H^1(\Omega)$ such that $u_0|_{\partial\Omega} = g$. Let w be in $H_0^1(\Omega)$ and given by $w = u - u_0$. This function is solution of the variational problem (2.4.4) with

$$\ell(v) = (f,v) - a(u_0,v), v \in H_0^1(\Omega). \tag{2.4.18}$$

As

$$|\ell(v)| \leq (\|f\|_{L^2(\Omega)} + 2n\hat{C}\|u_0\|_{H_0^1(\Omega)})\|v\|_{H_0^1(\Omega)}, v \in H_0^1(\Omega),$$

the problem (2.4.4) with ℓ given by (2.4.18) has a unique solution in $H_0^1(\Omega)$, provided that the coefficient functions satisfy

$$a_{ij}, b_i, c \in L^\infty(\Omega)$$

and the conditions (2.4.2), (2.4.14) hold. Finally, taking

$$u = w + u_0,$$

a solution of the variational problem (2.4.17) is obtained.

Poisson's equation with Neumann boundary conditions:

We introduce now a variational problem associated with the BVP

$$\begin{cases} -\Delta u + a_0 u = f \text{ in } \Omega, \\ \frac{\partial u}{\partial \eta} = g \text{ on } \Omega. \end{cases} \tag{2.4.19}$$

From Poisson equation with $v \in C^\infty(\Omega) \cap H^1(\Omega)$, we get

$$\sum_{i=1}^n \int_\Omega \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\,dx + \int_\Omega a_0 uv\,dx = \int_\Omega f\phi\,dx\,dy + \int_{\partial\Omega} \frac{\partial u}{\partial \eta} v ds.$$

Let $a(.,.) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ be defined by

$$a(v,w) = \sum_{i=1}^n \int_\Omega \frac{\partial v}{\partial x_i}\frac{\partial w}{\partial x_i}\,dx + \int_\Omega a_0 vw\,dxv, w \in H^1(\Omega).$$

We introduce the variational problem:

$$\text{find } u \in H^1(\Omega) \text{ such that } a(u,v) = \ell(v), v \in H^1(\Omega), \tag{2.4.20}$$

where $\ell : H^1(\Omega) \to \mathbb{R}$ is given by

$$\ell(v) = \int_\Omega fv\,dx\,dy + \int_{\partial\Omega} gv\,ds,\ v \in H^1(\Omega).$$

A solution of problem (2.4.20) is called weak solution of the boundary value problem (2.4.19). If the weak solution u is smooth enough, then

$$a(u,v) = \ell_1(v),\ \forall v \in C_0^\infty(\Omega), \tag{2.4.21}$$

with

$$\ell_1(v) = (f,v), v \in H_0^1(\Omega).$$

From (2.4.21), we obtain

$$(-\Delta u + a_0 u - f, v) = 0, \forall v \in C_0^\infty(\Omega).$$

Thus the PDEs of (2.4.19) holds in $L^2(\Omega)$.

As

$$-\Delta u + a_0 u = f$$

holds in $L^2(\Omega)$, we get

$$(\frac{\partial u}{\partial \eta} - g, v)_{L^2(\partial\Omega)} = 0,\ \forall v \in H^1(\Omega).$$

Consequently, $\dfrac{\partial u}{\partial \eta} = g$ in $L^2(\partial\Omega)$.

If we assume that a_0 is positive, bounded and $a_0 \geq \alpha_1 > 0$ in $\overline{\Omega}$, then $a(.,.)$ is $H^1(\Omega)$-elliptic. For ℓ we get

$$\begin{aligned} |\ell(v)| &\leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\|v|_{\partial\Omega}\|_{L^2(\partial\Omega)} \\ &\leq C(\|f\|_{L^2} + C\|g\|_{L^2(\partial\Omega)})\|v\|_{H^1(\Omega)},\ v \in H^1(\Omega), \end{aligned}$$

which leads to $\ell \in H^1(\Omega)'$, provided that $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$.

Under the previous assumptions, there exists a unique solution of the variational problem (2.4.20), in $H^1(\Omega)$, such that

$$\|u\|_{H^1(\Omega)} \leq \left(\|f\|_{L^2(\Omega)} + C\|g\|_{L^2(\partial\Omega)}\right).$$

The Stokes Equations

Let Ω be an open bounded domain. For $f = (f_1, \ldots, f_n) \in (L^2(\Omega))^n$[15] let $u = (u_1, \ldots, u_n)$ and p be defined in Ω and such that

$$\begin{cases} -\mu\Delta u_i + \dfrac{\partial p}{\partial x_i} = f_i\,, \text{ in } \Omega, i = 1, \ldots, n, \\ \nabla.u = 0\,, \text{ in } \Omega, \\ u_i = 0 \text{ on } \partial\Omega,\ i = 1, \ldots, n, \end{cases} \tag{2.4.22}$$

[15] $(L^2(\Omega))^n$ represents the space of vector functions $v = (v_1, \ldots, v_n) : \Omega \to \mathbb{R}$ with $v_i \in L^2(\Omega), i = 1, \ldots, n$.

where μ denotes a positive constant. In fluid mechanics the Stokes equations (2.4.22) describe the flow of an incompressible medium with viscosity μ and exterior force f. In (2.4.22), u represents the velocity field (u_i is the velocity of the medium in x_i direction) and p the pressure. The homogeneous Dirichlet boundary condition means that the flow vanishes at the boundary.

Let v be in $(C_0^\infty(\Omega))^n$. From the first equation of (2.4.22)[16] we obtain

$$\mu \sum_{i,j=1}^{n} \int_\Omega \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\,dx - \sum_{i=1}^{n}\int_\Omega p\frac{\partial v_i}{\partial x_j}\,dx = \sum_{i=1}^{n}\int_\Omega f_i v_i\,dx, i=1,\ldots,n.$$

Then for v such that $\nabla.v = 0$ we get

$$\mu \sum_{i,j=1}^{n} \int_\Omega \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\,dx = \sum_{i=1}^{n}\int_\Omega f_i v_i\,dx, i=1,\ldots,n. \tag{2.4.23}$$

This fact induces the introduction of the following space

$$V = \{v \in (H_0^1(\Omega))^n \,:\, \nabla.v = 0\}.$$

[17] V is closed in $(H_0^1(\Omega))^n$ and it is a Hilbert space with respect to the inner product defined in $(H_0^1(\Omega))^n$.

Let $a(.,.) : V \times V \to \mathbb{R}$ be the bilinear form

$$a(w,v) = \mu \sum_{i,j=1}^{n} \int_\Omega \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\,dx,\ w,v \in V.$$

Using the Poincaré-Friedrichs inequality, $a(.,.)$ is V-elliptic. As $f \in (L^2(\Omega))^n$,

$$\ell(v) = \sum_{i=1}^{n}\int_\Omega f_i v_i\,dx, v \in V,$$

belongs to V'. The Lax-Milgram Lemma allow us to conclude that there exists a unique solution of the variational problem:

$$\text{find}\, u \in V \,:\, a(u,v) = \ell(v),\ v \in V. \tag{2.4.24}$$

As $(C_0^\infty(\Omega))^n$ is not contained in V, it is not possible to give a direct interpretation of the last variational problem. In order to avoid this difficulty, we define a new variational problem equivalent to the last one using the next result.

[16] The inner product in $(L^2(\Omega))^n$ is defined by

$$(w,v)_{(L^2(\Omega))^n} = \sum_{i=1}^{n}(w_i,v_i)_{L^2(\Omega)}, w,v \in (L^2(\Omega))^n.$$

[17] By $(H_0^1(\Omega))^n$ we represents the space of vector functions $v = (v_1,\ldots,v_n) : \Omega \to \mathbb{R}$ with $v_i \in H_0^1(\Omega), i = 1,\ldots,n$. In this space we consider the inner product

$$(w,v)_{(H_0^1(\Omega))^n} = \sum_{i=1}^{n}(w_i,v_i)_{H_0^1(\Omega)}, w,v \in (H_0^1(\Omega))^n,$$

and the norm induced by this inner product. The space $(H_0^1(\Omega))^n$ is a Hilbert space with respect to the inner product $(.,.)_{(H_0^1(\Omega))^n}$.

Theorem 2.4.3 *Let Ω be a convex bounded open set with boundary $\partial\Omega$ smooth enough ($\partial\Omega$ piecewise C^1) and let $\hat{\ell}$ be in $[(H_0^1(\Omega))^n]'$. Then $\hat{\ell}$ in null on V if and only if there exists $\phi \in L^2(\Omega)$ such that*

$$\hat{\ell}(v) = \int_\Omega \phi \nabla.v \, dx \, \forall v \in (H_0^1(\Omega))^n. \tag{2.4.25}$$

Two any function ϕ_1, ϕ_2 differ by a constant.

If $\hat{\ell}$ is defined by (2.4.25) with $\phi \in L^2(\Omega)$, then $\hat{\ell} \in [(H_0^1(\Omega))^n]'$ and $\hat{\ell}$ is null on V.

The crucial point in the proof of the Theorem 2.4.3 is the existence of a function ϕ, in $L^2(\Omega)$, such that (2.4.25) holds provide that $\hat{\ell}$ is null on V. If ϕ_1 and ϕ_2 satisfy (2.4.25), then

$$\forall v \in (H_0^1(\Omega))^n \quad \int_\Omega (\phi_1 - \phi_2) \nabla.v \, dx = -\sum_{i=1}^n \int_\Omega v \frac{\partial}{\partial x_i}(\phi_1 - \phi_2) \, dx = 0.$$

Thus

$$\forall v \in (C_0^\infty(\Omega))^n \quad \sum_{i=1}^n \int_\Omega v \frac{\partial}{\partial x_i}(\phi_1 - \phi_2) \, dx = 0,$$

which leads to

$$\frac{\partial}{\partial x_i}(\phi_1 - \phi_2) = 0, i = 1, \ldots, n.$$

Finally, from the last equality we deduce $\phi_1 - \phi_2 = const.$ in Ω.

In order to use the Theorem 2.4.3, we define

$$\hat{\ell}(v) = a(u, v) - \ell(v), v \in (H_0^1(\Omega))^n, \tag{2.4.26}$$

where u is the solution of the variational problem (2.4.24). We have $\hat{\ell} \in [(H_0^1(\Omega))^n]'$ and $\hat{\ell}(v) = 0$, for $v \in V$. Then, by the Theorem 2.4.3, there exists a function $p \in L^2(\Omega)$ such that

$$\int_\Omega pv \, dx = a(u, v) - \ell(v), \; v \in (H_0^1(\Omega))^n,$$

that is

$$a(u, v) - \int_\Omega pv \, dx = \sum_{i=1}^n \int_\Omega f v_i \, dx, \; v \in (H_0^1(\Omega))^n. \tag{2.4.27}$$

We proved that there exists a unique $u \in V$ and $p \in L^2(\Omega)$ such that (2.4.27) holds. Finally, if $(u, p) \in (H_0^1(\Omega))^n \times L^2(\Omega)$ is solution of the variational problem (2.4.27), then u is solution of the variational problem (2.4.24).

2.5 The Ritz-Galerkin Method

2.5.1 The Finite Element Method

Let V_H be a subspace of V with dimension N_H. The variational problem (2.4.7) defined in V, can be considered in V_H,

$$\text{find } u_H \in V_H \text{ such that } a(u_H, v_H) = \ell(v_H), \; \forall v_H \in V_H. \tag{2.5.1}$$

As $V_H \subset V$, then V_H equipped with the norm $\|.\|_V$ still is a Banach space. Moreover, $a(.,.)$ on $V_H \times V_H$ has exactly the same properties that it has when defined on $V \times V$. As $\ell \in V'$ then $\ell \in V_H'$, and so the variational problem (2.5.1) is well defined.

In certain sense, the solution of the finite dimensional variational problem (2.5.1) is an approximation for the solution of the problem (2.4.7) being the error defined by $e_H = u - u_H$. The solution of this new variational problem is called Ritz-Galerkin solution and the designed method is called Ritz-Galerkin method.

Let $\{\phi_j, j = 1, \ldots, N_H\}$ be a basis of V_H. Then $u_H = \sum_{j=1}^{N_H} \alpha_j \phi_j \in V_H$ is the Ritz-Galerkin solution of (2.5.1) if and only if

$$\sum_{j=1}^{N_h} a(\phi_j, \phi_i)\alpha_j = \ell(\phi_i)\, i = 1, \ldots, N_H,$$

if and only if the coefficients $\alpha_j, j = 1, \ldots, N_H$, satisfy

$$A\alpha = F,\ A = [a(\phi_j, \phi_i)], \alpha = [\alpha_j], F = [f(\phi_i)]. \tag{2.5.2}$$

The matrix A, usually called stiffness matrix, is symmetric if and only if $a(.,.)$ is symmetric.

The existence and uniqueness of the Ritz-Galerkin solution are characterized in the following result:

Theorem 2.5.1 *Let $\{\phi_j, j = 1, \ldots, N_H\}$ be a basis of V_H. There exists a unique Ritz-Galerkin solution of the variational problem (2.5.1) if and only if the linear system (2.5.2) has a unique solution.*

If $a(.,.)$ is V-elliptic, then $[a(\phi_i, \phi_j)]$ is nonsingular

Proof: Let us suppose that $[a(\phi_j, \phi_i)]$ is singular. Then there exists $[\alpha_i] \neq 0$, such that

$$[a(\phi_j, \phi_i)][\alpha_i] = 0,$$

which implies

$$a(\sum_j \alpha_j \phi_j, \sum_i \alpha_i \phi_i) = 0.$$

As $a(.,.)$ is V-elliptic, we deduce that $\sum_j \alpha_j \phi_j = 0$, and consequently $\alpha_j = 0$ for all j.

■

Example 26 *Let us consider the one dimensional problem*

$$Lu(x) = -(p(x)u'(x))' + q(x)u(x) = f, x \in (a,b), u(a) = u(b) = 0,$$

with $p \in C^1(a,b), q \in C(a,b)$.

The variational problem is defined by

$$a(u,v) = \int_a^b (pu'v' + quv)\, dx, u, v \in H_0^1(a,b), \tag{2.5.3}$$

and

$$\ell(v) = \int_a^b f v \, dx \in (H_0^1(a,b))', \tag{2.5.4}$$

where $f \in L^2(a,b)$.

Let $\{x_i, i = 0, \ldots, n\}$ *be a nonuniform grid in* $[a,b]$, *where* $x_0 = a, x_b = b$ *and* $x_i - x_{i-1} = h_i$. *Let* V_H *be the space of piecewise linear functions. The finite dimensional variational problem is: find* $u_H \in V_H$ *such that* $a(u_H, v_H) = \ell(v_H)$ *for all* $v_H \in V_H$. *In order to deduce the linear system which defines the Ritz-Galerkin solution, we fixe in* V_H *the basis* $\{\phi_i, i = 1, \ldots, n-1\}$,

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i} & x \in [x_{i-1}, x_i], \\ -\dfrac{x - x_{i+1}}{h_{i+1}} & x \in (x_i, x_{i+1}], \qquad i = 1, \ldots, n-1. \\ 0 & x \in [a, x_{i-1}) \cup (x_{i+1}, b], \end{cases}$$

It is clear that $V_H = \mathcal{L}\{\phi_i, i = 1, \ldots, n-1\}$ *and* $V_H \subset H_0^1(a,b)$. *The coefficients of the Ritz-Galerkin solution* $u_H(x) = \sum_{i=1}^{n-1} \alpha_i \phi_i(x), x \in [a,b]$, *satisfy*

$$A[\alpha_i] = [\int_a^b f(x)\phi_i(x)\, dx],$$

with

$$a_{ii} = \int_{x_{i-1}}^{x_i} \left(p(x)\frac{1}{h_i^2} + q(x)\frac{(x - x_i)^2}{h_i^2} \right) dx + \int_{x_i}^{x_{i+1}} \left(p(x)\frac{1}{h_{i+1}^2} + q(x)\frac{(x - x_{i+1})^2}{h_{i+1}^2} \right) dx,$$

$$a_{i-1i} = \int_{x_{i-1}}^{x_i} \left(p(x)(-\frac{1}{h_i^2}) - q(x)\frac{(x - x_{i-1})(x - x_i)}{h_i^2} \right) dx,$$

$$a_{ii+1} = \int_{x_i}^{x_{i+1}} \left(p(x)(-\frac{1}{h_{i+1}^2}) - q(x)\frac{(x - x_i)(x - x_{i+1})}{h_{i+1}^2} \right) dx.$$

Example 27 *Let us consider the Poisson equation in the unitary square* $\Omega = (0,1) \times (0,1)$ *with homogeneous Dirichlet boundary conditions. The weak formulation of this BVP is given by (2.4.7) with* $V = H_0^1(\Omega)$,

$$a(u,v) = \int_\Omega \nabla u . \nabla v \, dx, \; \ell(v) = \int_\Omega f v \, dx \, .$$

Let V_H *be the finite dimensional space of* V *spanned by the functions*

$$\phi_1(x_1, x_2) = \sin(\pi x_1)\sin(\pi x_2)\,, \; \phi_2(x_1, x_2) = \sin(3\pi x_1)\sin(\pi x_2),$$

$$\phi_3(x_1, x_2) = \sin(\pi x_1)\sin(3\pi x_2)\,, \; \phi_4(x_1, x_2) = \sin(3\pi x_1)\sin(3\pi x_2).$$

The matrix of the linear system for the coefficients of the Ritz-Galerkin solution is a diagonal matrix where

$$a_{11} = \frac{\pi^2}{2}, a_{22} = a_{33} = \frac{5\pi^2}{2}, a_{44} = \frac{9\pi^2}{2}.$$

As for $f = 1$,

$$\ell(\phi_1) = \frac{4}{\pi^2},\ \ell(\phi_2) = \ell(\phi_3) = \frac{4}{3\pi^2}, \ell(\phi_4) = \frac{4}{9\pi^2},$$

we get

$$u_H(x_1, x_2) = \frac{8}{\pi^4}(x_1, x_2) \quad + \frac{8}{15\pi^4}(\phi_2(x_1, x_2) + \phi_3(x_1, x_2)) + \frac{8}{81\pi^4}\phi_4(x_1, x_2),\ (x_1, x_2) \in [0, 1] \times [0, 1].$$

If the basis of the finite dimensional space is not fixed according to the previous examples, the matrix of the Ritz-Galerkin solution is in general full, that is, $a_{ij} = a(\phi_i, \phi_j) \neq 0$ for almost every i and j. Therefore, the computational cost of the Ritz-Galerkin solution increases drastically when the dimension of V_H increases. This behaviour can be avoided if the choice of the basis follows the basic principle followed in the choice of the basis of the previous examples: the support of ϕ_i has a nonempty intersection with the support of ϕ_j just for few j. Such property induces a sparse structure in the matrix of the Ritz-Galerkin solution. For instance, in Example 26, the basis $\{\phi_i\}$ was defined considering the sets $[a_i, b_i]$, $i = 0, \ldots, n$, such that

$$[a, b] = \cup_{i=0}^{n}[a_i, b_i], (a_i, b_i) \cap (a_j, b_j) = \emptyset, i \neq j,$$

and, for each i, the set of all j such that

$$supp(\phi_i) \cap supp(\phi_j) \neq \emptyset$$

is very "small".

The domain Ω is partitioned into small pieces, the so called finite elements, and the basis functions are defined in such a way that their supports are composed by a collection of finite elements. In this case, the Ritz-Galerkin method is usually called Finite Element method.

The weak formulations of the second order elliptic equations requires that $V = H_0^1(\Omega)$ or $V = H^1(\Omega)$. For a polygonal domain of $\mathbb{R}^2$, we introduce the finite dimensional subspace V_H of V fixing its basis with the previous requirements.

Let Ω be an open polygonal domain of $\mathbb{R}^2$. The partition of Ω, $\{\Omega_i, i = 1, \ldots, P\}$, is called an admissible partition if the following conditions are fulfilled:

P_0) $\Omega_i, i = 1, \ldots, P$, are open sets,

P_1) $\Omega_i \cap \Omega_j = \emptyset, i \neq j$,

P_2) $\cup_i \overline{\Omega}_i = \overline{\Omega}$,

P_3) $\overline{\Omega}_i \cap \overline{\Omega}_j$ is either empty or a common side or a common edge.

The pieces $\Omega_i, i = 1, \ldots, P$, are the finite elements.

1. Linear elements for $\Omega \subset \mathbb{R}^2$

   Let us consider a triangulation $\mathcal{T}_H$ of Ω with the triangles $T_1, \ldots, T_t$. This triangulation is admissible if $\{T_i, i = 1, \ldots, t\}$ satisfies the conditions P_0, P_1 , P_2 and P_3. The edges of

the triangles of $\mathcal{T}_H$ define the nodes of the partition. These nodes can be interior nodes or boundary nodes. Let N_H the number of interior nodes. We use the notation

$$V_H := \{v_H \in C^0(\overline{\Omega}) : v_h = 0 \text{ on } \partial\Omega, \\ v_h(x_1, yx_2) = a_0 + a_1x_1 + a_2x_2,\ (x_1, x_2) \in T,\ T \in \mathcal{T}_H\}.$$

We point out that if $v_H \in V_H$, then, in $T \in \mathcal{T}_H$, v_H is uniquely determined by the values $v_H(x^{(1)}), v_H(x^{(2)})$ and $v_H(x^{(3)})$, where $x^{(i)}, i = 1, 2, 3$, are the edges of T. Moreover, v_H is continuous in $\overline{\Omega}$, and the partial derivatives of v_H are constant on each triangle presenting, eventually, jumps on the common sides of the triangles. Thus $V_H \subset H_0^1$.

Let $x^{(i)}, i = 1, \ldots, N_H$, be the interior nodes of the triangulation $\mathcal{T}_H$. Let ϕ_i be associated with the vertex $x^{(i)}$ such that

$$\phi_i(x^{(i)}) = 1,\ \phi_i(x^{(j)}) = 0,\ j \neq i.$$

For instance, if T has vertices $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$, $x^{(j)} = (x_1^{(j)}, x_2^{(j)})$ and $x^{(m)} = (x_1^{(m)}, x_2^{(m)})$, then

$$\phi_i(x_1, x_2) = \frac{(x_1 - x_1^{(ij)})(x_2^{(m)} - x_2^{(j)}) - (x_2 - x_2^{(j)})(x_1^{(m)} - x_1^{(j)})}{(x_1^{(i)} - x_1^{(j)})(x_2^{(m)} - x_2^{(j)}) - (x_2^{(i)} - x_2^{(j)})(x_1^{(m)} - x_1^{(j)})},\ (x_1, x_2) \in T,$$

and $\phi_i = 0$ on all triangles that do not have $x^{(i)}$ as a vertex.

The functions $\phi_i, i = 1, \ldots, N_H$, have the properties:

(a) $\{\phi_i, i = 1, \ldots, N_H\}$ is a basis of V_H.

(b) $supp(\phi_i) = \cup\{\overline{T} \in \mathcal{T}_H : T \text{ has } x^{(i)} \text{ as a vertex }\}$.

(c) If the corners $x^{(i)}$ and $x^{(j)}$ are connected by a side, then $\Omega_i \cap \Omega_j = \emptyset$.

As a consequence of the definition of $\mathcal{T}_H$ and of $\{\phi_i, i = 1, \ldots, N_H\}$, the stiffness matrix $[a(\phi_i, \phi_j)]$ is sparse. In each row i, the entries of the matrix, eventually, not null are in the j columns where j is such that $x^{(i)}$ and $x^{(ij)}$ are connected by a side. For instance, for the bilinear form (2.4.5) with $b_i = 0, i = 1, \ldots, n$, we have

$$\begin{aligned} a(\phi_i, \phi_j) &= \sum_{\ell,k} \int_\Omega a_{\ell k} \frac{\partial \phi_i}{\partial x_\ell} \frac{\partial \phi_j}{\partial x_k}\, dx + \int_\Omega a_0 \phi_i \phi_j\, dx \\ &= \sum_{T \in \mathcal{T}_h} \sum_{\ell,k} \int_T a_{\ell k} \frac{\partial \phi_i}{\partial x_\ell} \frac{\partial \phi_j}{\partial x_k}\, dx + \int_T a_0 \phi_i \phi_j\, dx \\ &= \sum_{m \in I} \Big(\sum_{\ell,k} \int_{T_m} a_{\ell k} \frac{\partial \phi_i}{\partial x_\ell} \frac{\partial \phi_j}{\partial x_k}\, dx + \int_{T_m} a_0 \phi_i \phi_j\, dx \Big), \end{aligned}$$

where $T_m, m \in I$, is the set of all triangles with $x^{(i)}$ as a vertex.

The integration $\displaystyle\int_{T_m} dx_1\, dx_2$ seems a difficulty of the computation of the finite element solution. However, for each m, we can express $\displaystyle\int_{T_m} dx_1\, dx_2$ as an integral over the reference

triangle Δ defined by the points $(0,0),(0,1)$ and $(1,0)$. In order to show that, let us suppose that T is an arbitrary triangle of $\mathcal{T}_H$ with the vertices: $x^{(1)}$, $x^{(2)}$, $x^{(3)}$. Let Ψ_T be the transformation

$$\Psi : \Delta \to T, \quad \Psi(\xi,\eta) = x^{(1)} + \xi(x^{(2)} - x^{(1)}) + \eta(x^{(3)} - x^{(1)}),\ (\xi,\eta) \in \Delta. \tag{2.5.5}$$

If ϕ is defined on T, then

$$\phi(x,y) = \phi(\Psi(\xi,\eta) = \hat{\phi}(\xi,\eta).$$

The partial derivatives of ϕ are now given in function of the partial derivatives of $\hat{\phi}$ with respect to ξ and η. We have

$$\frac{\partial \phi}{\partial x_1} = \frac{\partial \hat{\phi}}{\partial \xi}\frac{\partial \xi}{\partial x_1} + \frac{\partial \hat{\phi}}{\partial \eta}\frac{\partial \eta}{\partial x_1}.$$

For $\dfrac{\partial \phi}{\partial x_2}$ holds an analogous representation. Thus we get

$$\nabla_{x_1,x_2}\phi = J(\Psi^{-1})^t \nabla_{\xi,\eta}\hat{\phi},$$

where

$$J(\Psi^{-1}) = \begin{bmatrix} \dfrac{\partial \xi}{\partial x_1} & \dfrac{\partial \xi}{\partial x_2} \\ \dfrac{\partial \eta}{\partial x_1} & \dfrac{\partial \eta}{\partial x_2} \end{bmatrix}.$$

Moreover, we can compute ϕ_i using the basis functions on the reference triangle Δ. In fact, we have

$$\begin{array}{l} \phi_1 = \hat{\phi}_{0,0}\left(\Psi^{-1}\right) \text{ in } T,\ \hat{\phi}_{0,0}(\xi,\eta) = 1-\xi-\eta,\ (\xi,\eta) \in \Delta, \\ \phi_2 = \hat{\phi}_{1,0}\left(\Psi^{-1}\right) \text{ in } T,\ \hat{\phi}_{1,0}(\xi,\eta) = \eta,\ (\xi,\eta) \in \Delta, \\ \phi_3 = \hat{\phi}_{0,1}\left(\Psi^{-1}\right) \text{ in } T,\ \hat{\phi}_{0,1}(\xi,\eta) = \xi,\ (\xi,\eta) \in \Delta, \end{array}$$

and

$$\nabla_{x_1,x_2}\phi_i = \mathrm{J}(\Psi^{-1})\nabla_{\xi,\eta}\hat{\phi}_i. \tag{2.5.6}$$

For instance let us consider

$$a(u,v) = (\nabla u, \nabla v),\ u,v \in H_0^1(\Omega).$$

On the evaluation of $a(\phi_i,\phi_j)$ we do not need to know explicitly ϕ_i and ϕ_j as functions on the triangles of $\mathcal{T}_H$. In fact, we have

$$\begin{aligned} a(\phi_i,\phi_j) &= \int_\Omega \nabla_{x_1,x_2}\phi_i \nabla_{x_1,x_2}\phi_j dx_1 dx_2 \\ &= \sum_{T\in\mathcal{T}_h} \int_T \nabla_{x_1,x_2}\phi_i \nabla_{x_1,x_2}\phi_j dx_1 dx_2, \end{aligned}$$

where

$$\begin{aligned}
&\int_T \nabla_{x_1,x_2}\phi_i \nabla_{x_1,x_2}\phi_j dx_1 dx_2 \\
&= \int_{T_{\xi,\eta}} \nabla_{x_1,x_2}\phi_i(\Psi(\xi,\eta))\nabla_{x_1,x_2}\phi_j(\Psi(\xi,\eta))|J(\psi)|\, d\xi d\eta \\
&= \int_{T_{\xi,\eta}} J(\Psi^{-1})^t \nabla_{\xi,\eta}\hat{\phi}_i(\xi,\eta)) J(\Psi^{-1})^t \nabla_{\xi,\eta}\hat{\phi}_j(\xi,\eta)|J(\psi)|\, d\xi d\eta.
\end{aligned}$$

In contrast to the finite difference methods, finite element discretization offers us the possibility to use, locally, triangles of different sizes.

2. Bilinear elements for $\Omega \subset \mathbb{R}^2$

Let Ω be a rectangle or an union of rectangles. Let $\mathcal{R}_H$ be the set $\{R_1, \ldots, R_{N_H}\}$ of rectangles of Ω. If the conditions P_i, $i = 0, 1, 2, 3$, hold with $\Omega_i = R_i$, then $\mathcal{R}_H$ is admissible partition of Ω. We point out that this partition can be induced by the two grids $\{x_{1,i}\}$ and $\{x_{2,j}\}$.

If homogeneous Dirichlet boundary conditions are considered in the differential problem, then we introduce the following space of bilinear functions

$$\begin{aligned}
V_H := \big\{ v_h \in C^0(\overline{\Omega}) :\ & v_h = 0 \,\text{on}\, \partial\Omega \\
& v_h(x_1,x_2) = (a_0 + a_1 x_1)(b_0 + b_1 x_2),\ (x_1,x_2) \in R,\ R \in \mathcal{R}_h \big\}.
\end{aligned}$$

Let v_H be a function in V_H. In $R \in \mathcal{R}_H$, v_H is univocally determined by the values $v_H(x_{1,i}, x_{2,j})$, $v_H(x_{1,i+1}, x_{2,j})$, $v_H(x_i, x_{2,j+1})$ and $v_H(x_{1,i+1}, x_{2,j+1})$,

$$\begin{aligned}
v_H(x_1,x_2) = {} & v_H(x_{1,i},x_{2,j})\Phi_{i,j}(x_1,x_2) + v_H(x_{1,i},x_{2,j+1})\Phi_{i,j+1}(x_1,x_2) \\
& + v_H(x_{1,i+1},x_{2,j})\Phi_{i+1,j}(x_1,x_2) + v_H(x_{1,i+1},x_{2,j+1})\Phi_{i+1,j+1}(x_1,x_2),
\end{aligned}$$

where

$$\Phi_{p,q}(x_1,x_2) = \phi_q(x_1)\phi_q(x_2), p = i, i+1, q = j, j+1,$$

and $\phi_p(x_1)$, $\phi_q(x_2)$ are the "hat" functions for $x_{1,p}$ and $x_{2,q}$, respectively.

We summarize the properties of V_H.

(a) $V_H \subset H_0^1(\Omega)$

(b) $\{\Phi_{i,j},\ (i,j) : (x_{1,i}, x_{2,j})\,\text{is an interior node}\}$ is a basis of V_H,

(c) $supp(\Phi_{i,j}) = \overline{\Omega} \cap [x_{i-1}, x_{i+1}] \times [y_{j-1}, y_{j+1}]$,

(d) If $(x_{1,i}, x_{2,j})$ and $(x_{1,\ell}, x_{2,p})$ are vertices of the same rectangle, then $supp(\Phi_{i,j}) \cap supp(\Phi_{\ell,p}) \neq \emptyset$. Else $supp(\Phi_{i,j}) \cap supp(\Phi_{\ell,p}) = \emptyset$.

A rectangular partition of a domain Ω requires that Ω is a rectangle or an union of rectangles. If Ω is a polygonal domain such that at least one side is not parallel to the axis, then the rectangles should be replaced by parallelograms.

A collection $\mathcal{P}_H$ of parallelogram Π_i is an admissible partition of Ω if the conditions P_i, $i = 0, 1, 2, 3$, hold with $\Omega_i = \Pi_i$. We suppose that $\mathcal{P}_H$ has N_H interior nodes.

We extended now the concept of bilinear function defined on a rectangle to a parallelogram. Let $x^{(1)}$, $x^{(2)}$, $x^{(3)}$ and $x^{(4)}$ be the vertices of the parallelogram Π and let Ψ be defined as follows

$$\begin{aligned}\Psi : [0,1]\times[0,1] &\longrightarrow \Pi\\ (\xi,\eta) &\to x^{(1)} + \xi(x^{(2)} - x^{(1)}) + \eta(x^{(4)} - x^{(1)}).\end{aligned}$$

We have $\Psi(0,0) = x^{(1)}$, $\Psi(1,0) = x^{(2)}, \Psi(0,1) = x^{(4)}, \Psi(1,1) = x^{(3)}$.

We define a bilinear function ϕ in Π by

$$\phi(x_1, x_2) = \hat{\phi}(\Psi^{-1}(x_1, x_2)),$$

where

$$\hat{\phi}(\xi,\eta) = (\alpha + \beta\xi)(\gamma + \sigma\eta).$$

Let $\phi_i, i = 1,\ldots,4$, be the bilinear functions in Π such that $\phi_i(x^{(i)}) = 1$ and $\phi_i(x^{(j)}) = 0$ for $j \neq i$. Then

$$\begin{aligned}
\phi_1 &= \hat{\phi}_{0,0}\left(\Psi^{-1}\right) \text{ in } \Pi, \hat{\phi}_{0,0}(\xi,\eta) = (1-\xi)(1-\eta),\\
\phi_2 &= \hat{\phi}_{1,0}\left(\Psi^{-1}\right) \text{ in } \Pi, \hat{\phi}_{1,0}(\xi,\eta) = \xi(1-\eta),\\
\phi_3 &= \hat{\phi}_{1,1}\left(\Psi^{-1}\right) \text{ in } \Pi, \hat{\phi}_{1,1}(\xi,\eta) = \xi\eta,\\
\phi_4 &= \hat{\phi}_{0,1}\left(\Psi^{-1}\right) \text{ in } \Pi, \hat{\phi}_{0,1}(\xi,\eta) = (1-\xi)\eta,
\end{aligned}$$

for $(\xi,\eta) \in [0,1]\times[0,1]$.

We introduce now the space of the bilinear functions based on the partition $\mathcal{P}_h$:

$$\begin{aligned}V_H := \big\{ v_H \in C^0(\overline{\Omega}) :\ & v_H = 0 \text{ on } \partial\Omega\\ & v_H \text{ is bilinear in } \Pi,\ \Pi \in \mathcal{P}_h \big\}.\end{aligned}$$

Then

(a) $V_H \subset H_0^1(\Omega)$,

(b) $\{\phi_i, i = 1,\ldots,N_H\}$, is a basis of V_H,

(c) $supp(\phi_i)$ is the union of all parallelograms that have $x^{(i)}$ as a vertex,

(d) $supp(\phi_i) \cap supp(\phi_j) \neq \emptyset$ if and only if $x^{(i)}$ and $x^{(j)}$ are vertices of the same parallelogram.

2.5.2 Error Estimates

Let u be the solution of the variational problem (2.4.7) and let u_H its Ritz-Galerkin approximation given by problem (2.5.1). We study now the discretization error $e_H = u - u_H$.

Theorem 2.5.2 [Céa's Theorem] *Let V be a Hilbert space and let ℓ be in V'. Let u be the solution of the variational problem (2.4.7) and u_H be the Ritz-Galerkin solution defined by (2.5.1). If $a(.,.)$ is continuous and V-elliptic, then*

$$\|u - u_H\|_V \leq (1 + \frac{C_c}{C_e}) dist(u, V_h), \tag{2.5.7}$$

where $dist(u, V_h) = \inf_{v_h \in V_h} \|u - v_h\|_V$.

Proof: For $w_H \in V_H$ we have

$$\|u - u_H\|_V \leq \|u - w_H\|_V + \|w_H - v_H\|_V. \tag{2.5.8}$$

As

$$a(u, v_H) = \ell(v_H), v_H \in V_H,\ a(u_H, v_H) = \ell(v_H), v_H \in V_H,$$

we get

$$a(u - u_H, v_H) = 0, v_H \in V_H.$$

Thus

$$a(w_H - u_H, v_H) = a(w_H - u, v_H) + a(u - u_H, v_H) = a(w_H - u, v_H).$$

Let v_H be defined by $v_H = \dfrac{1}{\|w_H - u_H\|_V}(w_H - u_H)$. As $a(.,.)$ is continuous, we deduce

$$a(w_H - u, v_H) \leq C_c \|w_H - u\|_H.$$

Otherwise, we also have

$$a(w_H - u_H, \frac{w_H - u_H}{\|w_H - u_H\|_V}) = \frac{1}{\|w_H - u_H\|} a(w_H - u_H, w_H - u_H) \geq C_e \|w_H - u_H\|,$$

because $a(.,.)$ is V-elliptic. This last inequality enable us to get the upper bound

$$\|w_H - u_H\|_V \leq \frac{C_c}{C_e} \|u - w_H\|_V. \tag{2.5.9}$$

Finally, from (2.5.8) and (2.5.9), we obtain (2.5.7).

■

Error estimates for piecewise linear finite element solution: one dimensional case

We apply Céa's Theorem to establish an upper bound for the error of the Ritz-Galerkin solution defined in Example 26. We have

$$dist(u, V_h) \leq \|u - u_I\|_{H_0^1(a,b)},$$

where u_I represents the piecewise linear interpolator of u. If $u \in C^2(a, b)$, then

$$|u(x) - u_I(x)| \leq \frac{1}{2} \max_{x \in [x_i, x_{i+1}]} (x - x_i)(x_{i+1} - x) \|u''\|_\infty = \frac{h_{i+1}^2}{8} \|u''\|_\infty,\ x \in [x_i, x_{i+1}].$$

The last estimate does not allow us to establish an estimate for $dist(u, V_h)$. In order to get an estimate for $\|u - u_I\|_{H_0^1(a,b)}$ we study $\|(u - u_I)'\|_{L^2(a,b)}$. We have

$$\begin{aligned} \|(u - u_I)'\|_{L^2(a,b)}^2 &= \int_a^b (u - u_I)'(x)\, dx \\ &= \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} (u - u_I)'(x)^2\, dx \\ &= \sum_{j=0}^{n-1} \frac{1}{h_{j+1}} \int_0^1 v'(\xi)^2\, d\xi, \end{aligned} \tag{2.5.10}$$

where $x = x_j + \xi h_{j+1}$, and $v(\xi) = (u - u_I)(x_j + \xi h_{j+1})$.

We prove now that

$$\int_0^1 v'(\xi)^2 \, d\xi \leq \int_0^1 v''(\xi)^2 \, d\xi, \tag{2.5.11}$$

holds. In order to prove this inequality, we point out that $v(x) = u(x) - u_I(x), x \in [a, b]$, satisfies

$$v'(y) = \int_\eta^y v''(\gamma) \, d\gamma, \tag{2.5.12}$$

where $\eta \in (a, b)$ is such that $v'(\eta) = 0$. The representation (2.5.12) leads to

$$|v'(y)| \leq |y - \eta|^{\frac{1}{2}} \left(\int_0^1 v''(y)^2 \, dy \right)^{\frac{1}{2}},$$

which allow us to conclude (2.5.10).

Combining (2.5.10) and (2.5.11) we get

$$\begin{aligned} \|(u - u_I)'\|_{L^2}^2 &= \sum_{j=1}^{n-1} \frac{1}{h_{j+1}} \int_0^1 v''(\xi)^2 \, d\xi \\ &= \sum_{j=1}^{n-1} h_{j+1}^2 \int_{x_j}^{x_{j+1}} (u - u_I)''(x)^2 \, dx \\ &\leq \frac{h^2}{2} \|u''\|_{L^2}^2, \end{aligned}$$

where $h = \max_i h_i$.

From the Poincaré-Friedrichs inequality we conclude that

$$\|u - u_I\|_{H_0^1(a,b)} \leq \frac{h}{2} \|u''\|_{L^2},$$

and then, applying Céa's Theorem, we obtain the error estimate

$$\|u - u_H\|_{H_0^1(a,b)} \leq h \|u''\|_{L^2}, \tag{2.5.13}$$

provided that $u \in H^2(a, b)$.

We proved the following convergence result:

Theorem 2.5.3 *Let u_h be the piecewise linear finite element solution defined by (2.5.3) and (2.5.4), where $0 < \alpha_0 \leq p \leq \alpha_1$, $q \geq 0$. Then the error $u - u_h$ satisfies (2.5.13) provided that $u \in H^2(a, b) \cap H_0^1(a, b)$.*

For the particular case $p = 1, q = 0$, it can be shown that

$$\|u - u_h\|_{L^2} \leq C h^2 \|u''\|_{L^2}. \tag{2.5.14}$$

In fact, let w be the solution of the auxiliary problem

$$-w'' = u - u_h \text{ in } (0, 1), \; w(0) = w(1) = 0.$$

We have

$$\begin{aligned}
\|u-u_H\|_{L^2}^2 &= (u-u_h, u-u_h) \\
&= (u-u_h, -w'') \\
&= \int_0^1 (u-u_h)' w'\, dx \\
&= a(u-u_h, w) \\
&= a(u-u_h, w-v)\,, \forall v \in V_h \\
&\leq \|u-u_h\|_{H_0^1(a,b)} \|w-v\|_{H_0^1(a,b)}\,, \forall v \in V_h.
\end{aligned}$$

Thus

$$\|u-u_h\|_{L^2} \leq \|u-u_h\|_1 \frac{\|w-v\|_{H_0^1(a,b)}}{\|w''\|_{L^2}} \ \forall v \in V_h.$$

If we consider $v = w_h$ as the piecewise linear finite element solution we have

$$\|w - w_h h\|_{H_0^1(a,b)} \leq h \|w''\|_{L^2},$$

which implies

$$\|u-u_h\|_{L^2} \leq h \|u-u_h\|_{H_0^1(a,b)}.$$

We finally conclude (2.5.41) using the estimate

$$\|u-u_h\|_{H_0^1(a,b)} \leq Ch \|u''\|_{L^2}.$$

Error estimates for piecewise linear finite element solution: two dimensional case

Let us consider the Poisson equation defined on the unitary square $\Omega = (0,1) \times (0,1)$ with homogeneous Dirichlet boundary condition. We recall that the weak formulation of this problem is

$$\begin{aligned}
&\text{find}\, u \in H_0^1(\Omega) : a(u,v) = \ell(v),\ \forall v \in H_0^1(\Omega), \\
&a(w,v) = \int_\Omega \nabla w . \nabla v\, dx,\ w, v \in H_0^1(\Omega), \ell(v) = \int_\Omega f v\, dx, v \in H_0^1(\Omega).
\end{aligned} \tag{2.5.15}$$

In order to construct the finite element approximation we define the finite element solution, we triangulate the the domain Ω considering the rectangular grid $\{(x_{1,i}, x_{2,j}),\ i,j = 0, \ldots, N\}$ with $x_{1,0} = x_{2,0} = 0$, $x_{1,N} = x_{2,N} = 1$, $x_{1,i} - x_{2,i-1} = x_{2,j} - x_{2,j-1} = h$(see Figure 9).

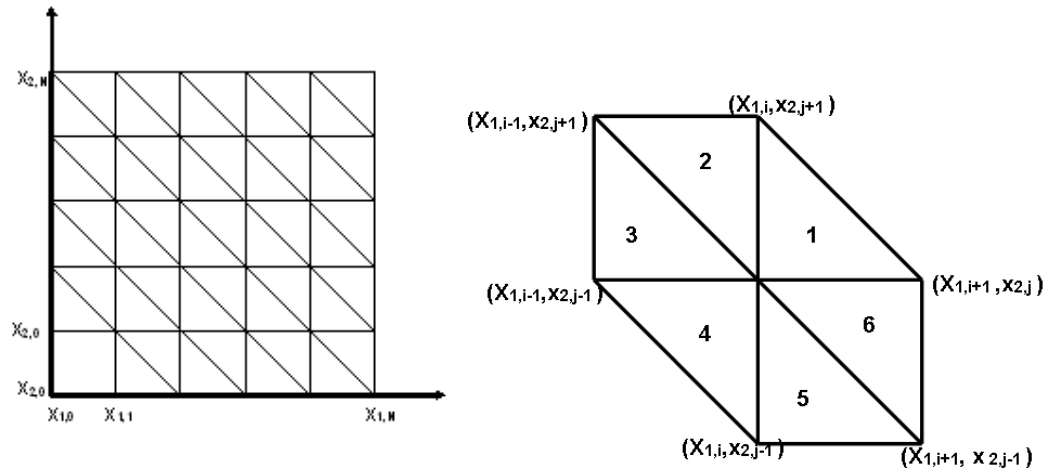

Figure 9: Triangulation induced by a rectangular grid.

For each $(x_{1,i}, x_{2,j})$ in Ω we associate a basis function ϕ_{ij} defined by

$$\phi_{ij}(x_1,x_2) = \begin{cases} 1 - \dfrac{x_1 - x_{1,i}}{h} - \dfrac{x_2 - x_{2,j}}{h},\ (x_1,x_2) \in 1, \\ 1 - \dfrac{x_2 - x_{2,j}}{h},\ (x_1,x_2) \in 2, \\ 1 - \dfrac{x_1 - x_{1,i}}{h},\ (x_1,x_2) \in 3, \\ 1 + \dfrac{x_1 - x_{1,i}}{h} + \dfrac{x_2 - x_{2,j}}{h},\ (x_1,x_2) \in 4, \\ 1 + \dfrac{x_2 - x_{2,j}}{h},\ (x_1,x_2) \in 5, \\ 1 - \dfrac{x_1 - x_{1,i}}{h},\ (x_1,x_2) \in 6, \\ 0 \qquad \text{otherwise}, \end{cases}$$

(see Figure 9).

For $V_H = \mathcal{L}\{\phi_{ij}, i,j = 1,\ldots,N-1\}$ we defined the Ritz-Galerkin solution

$$\begin{aligned} &\text{find}\, u_H \in V_H : a(u_H, v_H) = \ell(v_H),\ \forall v_H \in V_H, \\ &a(w_H, v_H) = \int_\Omega \nabla w_H . \nabla v_H \, dx,\ w_H, v_H \in V_H, \ell(v) = \int_\Omega f v_H \, dx, v_H \in V_H. \end{aligned} \tag{2.5.16}$$

According to Céa's Lemma

$$\|u - u_H\|_{H_0^1(\Omega)} \leq \|u - u_I\|_{H_0^1(\Omega)}, \tag{2.5.17}$$

where u_I denotes the continuous piecewise linear interpolant of the function u on the set $\overline{\Omega}$ given by

$$u_I(x_1,x_2) = \sum_{i,j=1}^{N-1} u(x_{1,i}, x_{2,j}) \phi_{ij}(x_1,x_2),\ (x_1,x_2) \in \overline{\Omega}.$$

We estimate in what follows

$$|u-u_I|^2_{H^1_0(\Omega)} = |e_I|^2_{H^1_0(\Omega)} = \sum_T \left(\int_T \left(\frac{\partial e_I}{\partial x_1} \right)^2 dx + \int_T \left(\frac{\partial e_I}{\partial x_2} \right)^2 dx \right). \tag{2.5.18}$$

Let us suppose that

$$T = \{(x_1, x_2) : x_{1,i} \le x_1 \le x_{1,i+1}, x_{2,j} \le x_2 \le x_{2,j+1} + x_{1,i} - x_1\}$$

and we define the transformation from the reference triangle $\Delta = \{(\xi, \eta) : 0 \le \xi \le 1, 0 \le \eta \le 1 - \xi\}$ on T by

$$x_1 = x_{1,i} + \xi h,\ x_2 = x_{2,j} + \eta h,\ 0 \le \xi, \eta \le 1.$$

Thus

$$\begin{aligned}
\int_T \left(\frac{\partial e_I}{\partial x_1} \right)^2 dx &= \int_\Delta |\frac{\partial \hat{u}}{\partial \xi} - \hat{u}(1,0) + \hat{u}(0,0)|^2 \, d\xi d\eta \\
&= \int_0^1 \int_0^{1-\xi} |\frac{\partial \hat{u}}{\partial \xi} - \int_0^1 \frac{\partial \hat{u}}{\partial \xi} \, d\sigma|^2 \, d\eta d\xi \\
&= \int_0^1 \int_0^{1-\xi} | \int_0^1 \left(\frac{\partial \hat{u}}{\partial \xi}(\xi, \eta) - \frac{\partial \hat{u}}{\partial \xi}(\sigma, \eta) \right) d\sigma d\eta d\xi \\
&+ \int_0^1 \int_0^{1-\xi} \int_0^1 \left(\frac{\partial \hat{u}}{\partial \xi}(\sigma, \eta) - \frac{\partial \hat{u}}{\partial \xi}(\sigma, 0) \right) d\sigma|^2 d\eta d\xi \\
&= \int_0^1 \int_0^{1-\xi} | \int_0^1 \int_\sigma^\xi \frac{\partial^2 \hat{u}}{\partial \xi^2}(\theta, \eta) d\theta \, d\sigma + \int_0^1 \int_0^\eta \frac{\partial^2 \hat{u}}{\partial \eta \partial \xi}(\sigma, \mu) d\sigma d\mu|^2 d\xi d\eta \\
&\le 2 \int_0^1 \int_0^{1-\xi} \int_0^1 \int_\sigma^\xi |\frac{\partial^2 \hat{u}}{\partial \xi^2}(\theta, \eta)|^2 d\theta \, d\sigma d\eta d\xi \\
&+ 2 \int_0^1 \int_0^{1-\xi} \int_0^1 \int_0^\eta |\frac{\partial^2 \hat{u}}{\partial \eta \partial \xi}(\sigma, \mu)|^2 d\mu d\sigma d\eta d\xi \\
&\le 2 \int_0^1 \int_0^1 |\frac{\partial^2 \hat{u}}{\partial \xi^2}(\theta, \eta)|^2 d\theta d\eta + 2 \int_0^1 \int_0^1 |\frac{\partial^2 \hat{u}}{\partial \eta \partial \xi}(\sigma, \mu)|^2 d\sigma d\mu \\
&= 2 \int_{x_{1,i}}^{x_{1,i+1}} \int_{x_{2,j}}^{x_{2,j+1}} |\frac{\partial^2 u}{\partial x_1^2}|^2 (h^2)^2 h^{-2} dx + 2 \int_{x_{1,i}}^{x_{1,i+1}} \int_{x_{2,j}}^{x_{2,j+1}} |\frac{\partial^2 u}{\partial x_2 \partial x_1}|^2 (h^2)^2 h^{-2} dx.
\end{aligned}$$

Therefore

$$\int_T \left(\frac{\partial e_I}{\partial x_1} \right)^2 dx \le 2h^2 \int_{x_{1,i}}^{x_{1,i+1}} \int_{x_{2,j}}^{x_{2,j+1}} \left(|\frac{\partial^2 u}{\partial x_1^2}|^2 + |\frac{\partial^2 u}{\partial x_2 \partial x_1}|^2 \right) dx. \tag{2.5.19}$$

Similarly

$$\int_T \left(\frac{\partial e_I}{\partial x_2} \right)^2 dx \le 2h^2 \int_{x_{1,i}}^{x_{1,i+1}} \int_{x_{2,j}}^{x_{2,j+1}} \left(|\frac{\partial^2 u}{\partial x_2^2}|^2 + |\frac{\partial^2 u}{\partial x_1 \partial x_2}|^2 \right) dx. \tag{2.5.20}$$

Substituting (2.5.19) and (2.5.20) into (2.5.18) we obtain

$$|u-u_I|^2_{H^1_0(\Omega)} \le 2h^2 \int_\Omega \left(|\frac{\partial^2 u}{\partial x_1^2}|^2 + |\frac{\partial^2 u}{\partial x_2 \partial x_1}|^2 + |\frac{\partial^2 u}{\partial x_1 \partial x_2}|^2 \frac{\partial^2 u}{\partial x_2^2}|^2 \right) dx. \tag{2.5.21}$$

We proved the next result:

Theorem 2.5.4 *Let u be the weak solution of the Poisson equation with homogeneous Dirichlet boundary condition defined by (2.5.15) and let u_H be the piecewise linear finite element solution defined by (2.5.16). Suppose that $u \in H^2(\Omega) \cap H_0^1(\Omega)$, then*

$$|u - u_H|_{H_0^1(\Omega)} \leq \sqrt{2} h |u|_{H^2(\Omega)}. \tag{2.5.22}$$

■

Since $u \in H_0^1(\Omega)$ and $u_I \in H_0^1(\Omega)$, by the Poincaré-Friedrichs inequality we have

$$\|u - u_I\|^2_{L^2(\Omega)} \leq \frac{1}{4} \|u - u_I\|^2_{H_0^1(\Omega)}, \tag{2.5.23}$$

which implies

$$\|u - u_I\|^2_{H_0^1(\Omega)} \leq \frac{5}{4} \|u - u_I\|^2_{H_0^1(\Omega)}. \tag{2.5.24}$$

Taking into account the estimates (2.5.21), (2.5.24) and (2.5.17), we conclude the proof of the next result:

Corollary 8 *Let u be the weak solution of the Poisson equation with homogeneous Dirichlet boundary condition defined by (2.5.15) and let u_H be the piecewise linear finite element solution defined by (2.5.16). If $u \in H^2(\Omega) \cap H_0^1(\Omega)$, then*

$$\|u - u_H\|_{H_0^1(\Omega)} \leq \sqrt{\frac{5}{2}} h |u|_{H^2(\Omega)}. \tag{2.5.25}$$

■

The error estimate (2.5.25) indicates that the error in L^2 norm between u and its piecewise linear finite element solution is of the size $O(h)$. As in one dimensional case, we prove in what follows that in fact we have

$$\|u - u_H\|_{L^2(\Omega)} \leq 2h^2 |u|_{H^2(\Omega)}. \tag{2.5.26}$$

It is obvious that if $w \in H^2(\Omega) \cap H_0^1(\Omega)$, then

$$\|\Delta w\|^2_{L^2(\Omega)} = \int_\Omega (\frac{\partial^2 w}{\partial x_1^2})^2 dx + 2 \int_\Omega (\frac{\partial^2 w}{\partial x_1^2} \frac{\partial^2 w}{\partial x_2})^2 dx + \int_\Omega (\frac{\partial^2 w}{\partial x_2^2})^2 dx.$$

As $w = 0$ on $\partial\Omega$, we have

$$\int_\Omega (\frac{\partial^2 w}{\partial x_1^2} \frac{\partial^2 w}{\partial x_2}^2)^2 dx = \int_\Omega \frac{\partial^2 w}{\partial x_1 \partial x_2} \frac{\partial^2 w}{\partial x_1 \partial x_2} = \int_\Omega |\frac{\partial^2 w}{\partial x_1 \partial x_2}|^2 \, dx,$$

and then

$$\begin{aligned} \|\Delta w\|^2_{L^2(\Omega)} &= \int_\Omega \Big(|\frac{\partial^2 w}{\partial x_1^2}|^2 + 2 |\frac{\partial^2 w}{\partial x_1 \partial x_2}|^2 + |\frac{\partial^2 w}{\partial x_2^2}|^2 \\ &= |w|^2_{H^2(\Omega)}. \end{aligned}$$

Given $g \in L^2(\Omega)$ let $w_g \in H_0^1(\Omega)$ be the weak solution of the boundary value problem

$$\begin{cases} -\Delta w_g = g \text{ in } \Omega, \\ w_g = 0 \text{ on } \partial\Omega \,. \end{cases} \tag{2.5.27}$$

Then $w_H \in H^2(\Omega) \cap H_0^1(\Omega)$ and

$$|w_g|_{H^2(\Omega)} = \|\Delta w_g\|_{L^2(\Omega)} = \|g\|_{L^2(\Omega)}.$$

Let g be in $L^2(\Omega)$ and let u_H be the piecewise linear finite element solution introduced before. As consequence of the Cauchy-Schwarz inequality we have

$$(u - u_H, g) \leq \|g\|_{L^2(\Omega)} \|u - u_H\|_{L^2(\Omega)}.$$

Therefore

$$\|u - u_H\|_{L^2(\Omega)} = \sup_{g \in L^2(\Omega)} \frac{(u - u_H, g)}{\|g\|_{L^2(\Omega)}}. \tag{2.5.28}$$

If we consider g fixed in $L^2(\Omega)$, then the weak solution w_g of the differential problem (2.5.27), is defined by

$$\begin{aligned} &a(w_g, v) = \ell_g(v),\ \forall v \in H_0^1(\Omega), \\ &a(w_g, v) = \int_\Omega \nabla w_g . \nabla v \, dx, \ell_g(v) = \int_\Omega g v \, dx,\ v \in H_0^1(\Omega) \end{aligned} \tag{2.5.29}$$

belongs to $H_0^1(\Omega)$. Let $w_{g,H}$ be the piecewise linear finite element approximation defined by

$$a(w_{g,H}, v_H) = \ell_g(v_H),\ \forall v_H \in V_H. \tag{2.5.30}$$

For the error $w_g - w_{g,H}$ holds the following

$$|w_g - w_{g,H}|_{H_0^1(\Omega)} \leq \sqrt{2} h |w_g|_{H^2(\Omega)},$$

and therefore

$$|w_g - w_{g,H}|_{H_0^1(\Omega)} \leq \sqrt{2} h |g|_{H^2(\Omega)}. \tag{2.5.31}$$

As $w_{g,H} \in V_H$, then

$$a(u - u_H, w_{g,H}) = 0,$$

and we get

$$\begin{aligned} (u - u_H, g) &= (g, u - u_H) \\ &= \ell_g(u - u_H) \\ &= a(w_g, u - u_H) \\ &= a(u - u_H, w_g) \\ &= a(u - u_H, w_g - w_{g,H}) \\ &\leq |u - u_H|_{H_0^1(\Omega)} |w_g - w_{g,H}|_{H_0^1(\Omega)}. \end{aligned} \tag{2.5.32}$$

Considering the estimates (2.5.22), (2.5.31), we deduce

$$(u - u_H, g) \leq 2h^2 |u|_{H^2(\Omega)} \|g\|_{L^2(\Omega)}. \tag{2.5.33}$$

Substituting (2.5.33) into the right-hand side of (2.5.28) we obtain the desired estimate

$$\|u - u_H\|_{L^2(\Omega)} \leq 2h^2 |u|_{H^2(\Omega)}. \tag{2.5.34}$$

The proof presented above is called Aubin-Nitsche duality arguments.

The piecewise linear finite element solution based on the triangulation illustrated in Figure 9 was studied. We prove in what follows that the same estimates hold for a general polygonal domain with a more general triangulation.

Theorem 2.5.5 *Let $T = \{(\xi, \eta) : \xi, \eta \geq 0, \xi + \eta \leq 1\}$. If $u \in H^2(T)$, then*

$$\|u\|^2_{H^2(T)} \leq C \left(\sum_{x \in V(T)} u(x)^2 + \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T)} \right), \tag{2.5.35}$$

where $V(T)$ denotes the set of vertices of T.

Proof:

- Let us consider the bilinear form $a(.,.) : H^2(\Omega) \times H^2(\Omega) \to \mathbb{R}$ defined by

$$a(u, v) = \sum_{x \in V(T)} u(x)v(x) + \sum_{|\alpha|=2} (D^\alpha u, D^\alpha v)_{L^2(T)}.$$

As the identity operator $i_d : H^2(T) \to C^0(T)$ is continuous, then $a(.,.)$ is continuous. In fact, using the continuity of the identity operator we have $|w(x)| \leq C\|w\|_{H^2(T)}$ for $w = u, v$, which implies

$$|a(u, v)| \leq (1 + 3C)\|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}.$$

The bilinear form $a(.,.)$ is coercive, that is

$$a(u, u) \geq C_1\|u\|^2_{H^2(T)} - C_2\|u\|^2_{L^2(T)}, u \in H^2(T), \tag{2.5.36}$$

where C_1 denotes a positive constant.

In order to prove (2.5.36) we start by point out that

$$a(u, u) \geq \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T)} = \|u\|^2_{H^2(T)} - \|u\|^2_{H^1(T)}.$$

As $i_d : H^2(T) \to H^1(T)$ is compact[18], for $\epsilon > 0$ there exists γ such that [19]

$$\|u\|^2_{H^1(T)} \leq \epsilon\|u\|^2_{H^2(T)} + \gamma\|u\|^2_{L^2(T)}.$$

Then, for ϵ fixed such that $\epsilon < 1$, we obtain (2.5.36) with $C_1 = 1 - \epsilon$.

[18] Every bounded sequence in $H^2(\Omega)$ has a subsequence converging in $H^1(\Omega)$

[19] Theorem: Let $U \subset V \subset W$ be Banach spaces such that the operators $i_d : U \to V$ $i_d : V \to W$ are continuous being the first one compact. Then, for $\epsilon > 0$, there exists $C_\epsilon > 0$ such that $\|u\|_V \leq \epsilon\|u\|_U + C_\epsilon\|u\|_W$. (Lemma 6.5.18, [12]).

- Let $A : H^2(T) \to H^2(T)'$ be defined by $Aw = a(w,.), w \in H^2(T)$. As $a(.,.)$ satisfies (2.5.36) and it is continuous, then A has inverse or $\lambda = 0$ is eigenvalue of A.[20]

  Let us suppose that $\lambda = 0$ is an eigenvalue of A and let $0 \neq e \in H^2(T)$ be an eigenfunction. As $a(e,e) = 0$ then $\sum_{x \in V(T)} e(x)^2 = 0$ and $\sum_{|\alpha|=2} \|D^\alpha e\|^2_{L^2(T)} = 0$. From the last equality, e is linear in T which leads to $e = 0$ on T, because $e(x) = 0$ for $x \in V(T)$.

  As $\lambda = 0$ is not an eigenvalue of A, we conclude that A has inverse and then $a(.,.)$ is $H^2(\Omega)$-elliptic.[21]

■

The previous theorem is extended in the following result for the triangle $T_h = hT$.

Theorem 2.5.6 *If $u \in H^2(T_h)$, then, for $|\beta| \leq 2$,*

$$\|D^\beta u\|^2_{L^2(T_h)} \leq C \left(h^{2-2|\beta|} \sum_{x \in V(T_h)} u(x)^2 + h^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T)} \right), \tag{2.5.37}$$

holds.

Proof: Let u be in $H^2(T_h)$ and let us consider the transformation $\psi : T \to T_h$ defined by $(x_1, x_2) = \psi(\xi, \eta) = (h\xi, h\eta)$ and by $v(\xi, \eta)$ we represent $u(h\xi, h\eta)$. We have $v \in H^2(T)$ and

$$\begin{aligned} \|D^\beta u\|^2_{L^2(T_h)} &= \int_{T_h} |D^\beta u|^2 \, dx \\ &= |det(J(\psi))| \int_T |D^\beta_{\xi,\eta} u|^2 \, d\xi d\eta \\ &= h^2 \int_T h^{-2|\beta|} |D^\beta v|^2 \, d\xi d\eta \\ &= h^{2-2|\beta|} \|D^\beta v\|^2_{L^2(T)}. \end{aligned}$$

From Theorem 2.5.5 we get

$$\|D^\beta v\|^2_{L^2(T)} \leq C \left(\sum_{x \in V(T)} v(x)^2 + \sum_{|\alpha|=2} \|D^\alpha v\|^2_{L^2(T)} \right),$$

and the proof of (2.5.37) is concluded because

$$\|D^\alpha v\|^2_{L^2(T)} = h^2 \|D^\alpha u\|^2_{L^2(T_h)},$$

for $|\alpha| = 2$. ■

The generalization of the last result for an arbitrary triangle is the aim of the next theorem.

[20] Theorem: Let V and U be Hilbert spaces such that $id : V \to U$ is continuous and compact. Let $a(.,.)$ be a continuous bilinear form such that

$$a(u,u) \geq C_1 \|u\|^2_V - C_2 \|u\|_U, u \in V,$$

and let A be the operator $A : V \to V'$ such that $Aw = a(w,.)$, $w \in V$. Then $\lambda = 0$ is eigenvalue of A or $A^{-1} \in \mathcal{L}(V', V)$.(Theorem 6.5.15, [12]).

[21] Let $A : V \to V'$ be defined by $Aw = a(w,.)$, $w \in V$. If $a(.,.)$ is continuous, symmetric, nonnegative and $A^{-1} \in \mathcal{L}(V', V)$, then $a(.,)$ is V-elliptic (Lemma6.5.2, Exercise 6.5.6 c) [12]).

Theorem 2.5.7 *Let $\tilde{T}$ be a triangle with the side lengths less or equal to h_{max} and with the interior angles greater or equal to $\alpha_0 > 0$. For $u \in H^2(\tilde{T})$, and $|\beta| \leq 2$,*

$$\|D^\beta u\|^2_{L^2(\tilde{T})} \leq C(\alpha_0) \left(h_{max}^{2-2|\beta|} \sum_{x \in V(\tilde{T})} u(x)^2 + h_{max}^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(\tilde{T})} \right), \tag{2.5.38}$$

holds.

Proof: Let $x^{(1)}, x^{(2)}$ and $x^{(3)}$ be the vertices of the triangle $\tilde{T}$. We consider the triangle $T_{h_{max}}$ defined as in Theorem 2.5.6 and the transformation $\psi : T_{h_{max}} \to \tilde{T}$ given by

$$\psi(\xi, \eta) = x^{(1)} + \frac{\xi}{h_{max}}(x^{(2)} - x^{(1)}) + \frac{\eta}{h_{max}}(x^{(3)} - x^{(1)}).$$

For $u \in H^2(\tilde{T})$, the function $v(\xi, \eta) = u(\psi^{-1}(\xi, \eta))$ belongs to $H^2(T_{h_{max}})$. We also have

$$\|D^\beta u\|^2_{L^2(\tilde{T})} = |det(J(\psi))| \int_{T_{h_{max}}} |D_x^\beta u(\psi^{-1}(\xi, \eta)|^2 \, d\xi d\eta,$$

where

$$|det(J(\psi))| = |\frac{(x_1^{(2)} - x_1^{(1)})(x_2^{(3)} - x_2^{(1)}) - (x_1^{(3)} - x_1^{(1)})(x_2^{(2)} - x_2^{(1)})}{h_{max}^2}|.$$

As $|det(J(\psi))| \in [\frac{1}{k(\alpha)}, k(\alpha_0)]$ we get

$$\|D^\beta u\|^2_{L^2(\tilde{T})} \leq C_1(\alpha_0) \sum_{|\beta'|=|\beta|} \|D^{\beta'}_{\xi,\eta} v\|^2_{L^2(T_{h_{max}})}.$$

Finally, applying Theorem 2.5.6 we obtain

$$\|D^{\beta'}_{\xi,\eta} v\|^2_{L^2(T_{h_{max}})} \leq C \left(h_{max}^{2-2|\beta|} \sum_{x \in V(T_{h_{max}})} v(x)^2 + h_{max}^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha_{\xi,\eta} v\|^2_{L^2(T_{h_{max}})} \right).$$

We conclude the proof of (2.5.38) using the inequality

$$\sum_{|\alpha|=2} \|D^\alpha v\|^2_{L^2(T_{h_{max}})} \leq C_2(\alpha_0) \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(\tilde{T})}.$$

■

Theorem 2.5.7 has a central role on the establishment of an upper bound for $dist(u, V_h)$.

Theorem 2.5.8 *Let $\mathcal{T}_H$ be an admissible triangulation of the polygonal domain Ω. We suppose that the side lengths of all triangles of $\mathcal{T}_H$ is less or equal to h_{max} and the interior angles of all triangles of $\mathcal{T}_H$ are great or equal to α_0. Let V_H be defined by*

$$V_H = \{v_H \in C^0(\overline{\Omega}) : v_H|_{\partial\Omega} = 0, \, v_H(x_1, x_2) = a + bx_1 + cx_2, (x_1, x_2) \in T, T \in \mathcal{T}_H\}$$

(or

$$V_H = \{v_H \in C^0(\overline{\Omega}) : v_H(x_1, x_2) = a + bx_1 + cx_2, (x_1, x_2) \in T, T \in \mathcal{T}_H\}$$

), then

$$\inf_{v_H \in V_H} \|u - v_H\|_{H^s(\Omega)} \leq C(\alpha_0) h_{max}^{2-s} |u|_{H^2(\Omega)}$$

for all $u \in H^2(\Omega) \cap V$, with $V = H_0^1(\Omega)$ (or $V = H^1(\Omega)$).

Proof: Let u be in $H^2(\Omega)$ and let u_I be the interpolater defined by

$$u_I = \sum_{x \in V(\mathcal{T}_H)} u(x)\phi_x$$

where ϕ_x is a basis function such that $\phi_x(x) = 1$ and $\phi_x(\bar{x}) = 0, \bar{x} \neq x$. The error function $e_H = u - u_I$ belongs to $H^2(\Omega)$ and satisfies

$$dist(u, V_H) \leq \|e_H\|_{H^s(\Omega)}.$$

Theorem 2.5.7 enable us to conclude that

$$\|D^\beta e_H\|^2_{L^2(\Omega)} \leq C(\alpha_0) \sum_{T \in \mathcal{T}_H} \left(h_{max}^{2-2|\beta|} \sum_{x \in V(T)} e_H(x)^2 + h_{max}^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T)} \right),$$

for $|\beta| \leq s$. As $e_H = 0$ at the vertices of each triangle, we finally obtain

$$\|D^\beta e_H\|^2_{L^2(\Omega)} \leq C(\alpha_0) h_{max}^{4-2|\beta|} \sum_{T \in \mathcal{T}_H} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T)} \leq C(\alpha_0) h_{max}^{4-2|\beta|} |u|^2_{H^2(\Omega)}.$$

■

The last theorem implies

1. for $s = 0$
$$\|u - u_I\|_{L^2(\Omega)} \leq C(\alpha_0) h^2_{max} |u|_{H^2(\Omega)},$$
2. for $s = 1$
$$\|u - u_I\|_{H^1(\Omega)} \leq C(\alpha_0) h_{max} |u|_{H^2(\Omega)}.$$

The following convergence result is established combining the Theorem 2.5.8 with Céa's Theorem.

Theorem 2.5.9 *Let $\mathcal{T}_H$ be an admissible triangulation of the polygonal domain Ω. We suppose that the side length of all triangles of $\mathcal{T}_H$ is less or equal to h_{max} and the interior angles of all triangles of $\mathcal{T}_H$ are great or equal to α_0. Let V_H be defined by*

$$V_H = \{v_H \in C^0(\overline{\Omega}) : v_H|_{\partial\Omega} = 0,\ v_H(x_1, x_2) = a + bx_1 + cx_2, (x_1, x_2) \in T, T \in \mathcal{T}_H\}$$

(or

$$V_H = \{v_H \in C^0(\overline{\Omega}) : v_H(x_1, x_2) = a + bx_1 + cx_2, (x_1, x_2) \in T, T \in \mathcal{T}_H\}$$

).

If $a(.,.) : V \times V \to \mathbb{R}$ is a V-elliptic continuous bilinear form ($V = H^1(\Omega)$ or $V = H^1_0(\Omega)$) and $\ell \in V'$, then there exist a unique weak solution u in V, a unique finite element solution u_H in V_H, such that

$$a(u, v) = \ell(v), \forall v \in V,$$
$$a(u_H, v_H) = \ell(v_H), \forall v_H \in V_H.$$

Moreover, if $u \in V \cap H^2(\Omega)$, then

$$\|u - u_H\|_{H^1(\Omega)} \leq C(\alpha_0) h_{max} |u|_{H^2(\Omega)}, \tag{2.5.39}$$

where $V = H^1(\Omega)$ (or$V = H^1_0(\Omega)$).

■

The convergence order with respect to the L^2 norm can be improved? The answer is positive and its is based on the Aubin-Nitsche duality arguments. Let us suppose that $a(.,.)$ is symmetric and let w be the solution of the variational problem

$$a(w,v) = (u - u_H, v), \forall v \in V. \tag{2.5.40}$$

As

$$\|u - u_H\|^2_{L^2(\Omega)} = a(w, u - u_H)$$

holds, we get

$$\|u - u_H\|^2_{L^2(\Omega)} = a(w - w_H, u - u_H), \tag{2.5.41}$$

where w_H is the piecewise linear finite element approximation for the weak solution w. In fact, we have (2.5.41) because

$$0 = a(u - u_H, v_H) = a(v_H, u - u_H),$$

holds for $v_H \in V_H$. Then, the estimate (2.5.41), is obtained taking $v_H = w_H$.

From (2.5.41) we can deduce the estimate

$$\|u - u_H\|^2_{L^2(\Omega)} \leq C_c \|w - w_H\|_{H^1(\Omega)} \|u - u_H\|_{H^1(\Omega)},$$

where C_c is the continuity constant of $a(.,.)$. From the last inequality, using Theorem 2.5.9, we get

$$\begin{aligned} \|u - u_H\|_{L^2(\Omega)} &\leq C_c \frac{\|w - w_H\|_{H^1(\Omega)}}{\|u - u_H\|_{L^2(\Omega)}} C(\alpha) h_{max} |u|_{H^2(\Omega)} \\ &\leq C_c h^2_{max} \frac{|w|_{H^2(\Omega)}}{\|u - u_H\|_{L^2(\Omega)}} |u|_{H^2(\Omega)}. \end{aligned} \tag{2.5.42}$$

If we suppose that, for each $f \in L^2(\Omega)$, the variational problem has a solution u in $H^2(\Omega) \cap V$ such that $|u|_{H^2} \leq \|f\|_{L^2(\Omega)}$, we have

$$|w|_{H^2(\Omega)} \leq C \|u - u_H\|_{L^2(\Omega)}.$$

From (2.5.42) we obtain the second convergence order for the piecewise linear finite element solution

$$\|u - u_H\|_{L^2(\Omega)} \leq C(\alpha) h^2_{max} |u|_{H^2(\Omega)}.$$

The finite element problem is solved considering a fixed partition of the domain Ω. However we should know that the decreasing of the diameter of the finite elements implies an decreasing on the error of the finite element solution. For the piecewise linear finite element solution, the estimates established until now depend on the smaller interior angle of the triangles of the admissible triangulation $\mathcal{T}_H$. In order to avoid such dependence, the triangulations should be carefully constructed.

Let h_T be the longest side of $T \in \mathcal{T}_H$, then

$$h_{max} = max\{h_T, T \in \mathcal{T}_H\}$$

is the longest side length which occurs in $\mathcal{T}_H$ for $H \in \Lambda$. By ρ_T we denote the radius of the inscribed circle in $T \in \mathcal{T}_H$. The ratio h_T/ρ_T tends to infinity exactly when the smallest interior angle tends to zero. If

$$\max_{T\in\mathcal{T}_H} \frac{h_T}{\rho_T}$$

is bounded for a family of triangulations then we call the family quasi-uniform. Otherwise, if the family satisfies the stronger requirement

$$\frac{h_{max}}{\min_{T\in\mathcal{T}_H} \rho_T} \leq Const,$$

then the family is said uniform.

For quasi-uniform triangulations we can improve the quality of the estimates for the piecewise linear finite element solution obtained before. For instance, if we consider Theorem 2.5.9 for a family of quasi-uniform triangulations, then there exists a positive constant such that

$$\|u - u_H\|_{H^1(\Omega)} \leq C h_{max} |u|_{H^2(\Omega)}.$$

Moreover, by the Aubin-Nitsche duality arguments, we also have

$$\|u - u_H\|_{L^2(\Omega)} \leq C h^2 |u|_{H^1(\Omega)}.$$

2.5.3 A Neumann Problem

Let us consider now the Poisson problem with the Neumann boundary condition $\dfrac{\partial u}{\partial \eta} = g$ on $\partial\Omega$, where Ω is the unitary square of $\mathbb{R}^2$. The weak solution of this problem is defined by

$$\begin{aligned} &\text{find } u \in H^1(\Omega) \,:\, a(u,v) = \ell(v),\ \forall v \in H^1(\Omega), \\ &a(w,v) = \int_\Omega \nabla w.\nabla v\, dx,\ \ell(v) = \int_\Omega f v dx + \int_{\partial\Omega} g v\, ds, w, v \in H^1(\Omega). \end{aligned} \tag{2.5.43}$$

The bilinear form $a(.,.)$ is not $H^1(\Omega)$-elliptic. In order to define a variational problem with a unique solution, we consider the Friedrichs inequality

$$\|v - \bar{v}\|_{H^1(\Omega)} \leq C|v|_{H^1(\Omega)}, v \in H^1(\Omega),$$

for $\bar{v} = \dfrac{1}{|\Omega|}\displaystyle\int_\Omega v(x)\, dx$. If V is defined by

$$V = \{v \in H^1(\Omega) : \int_\Omega v\, dx = 0\}.$$

Then, the bilinear form $a(.,.)$ is V-elliptic. In fact,

$$a(v,v) = |v|_{H^1(\Omega)} \geq C_e \|v\|_{H^1(\Omega)},\ v \in V.$$

Consequently, the variational problem

$$\text{find } u \in V \,:\, a(u,v) = \ell(v),\ \forall v \in V, \tag{2.5.44}$$

has a unique solution.

Let $\mathcal{T}_H$ be an admissible triangulation for Ω. Let V_H be the space of piecewise linear functions induced by $\mathcal{T}_H$. In this space we consider the subset of functions such that $\int_\Omega v_H dx = 0$, that is

$$V_H = \{v_H \in C^0(\overline{\Omega}) : v_H(x,y) = a + bx_1 + cx_2, (x_1,x_2) \in T, T \in \mathcal{T}_H, \int_\Omega v_H \, dx = 0\}.$$

Let u_H be the finite element approximation for the solution of the Poisson equation with Neumann boundary condition. This solution is unique and the error $u - u_H$ satisfies (2.5.39).

2.5.4 Superapproximation in Mesh-Dependent Norms

In this section we establish that the error of the piecewise linear finite element solution is an $O(h_{max}^2)$ when a certain mesh-dependent norm is consider.

Let u_H be the piecewise linear finite element solution defined by (2.5.15) when the triangulation $\mathcal{T}_H$ plotted in Figure 9 is considered. The triangulation can be seen induced by the uniform partition in both axis with step size h. As u_H admits the representation

$$u_H(x_1,x_2) = \sum_{i,j=1}^{N-1} u_H(x_{1,i},x_{2,j})\phi_{ij}(x_1,x_2)$$

the node values $u_{i,j} = u_H(x_{1,i},x_{2,j})$ are computed solving the linear system

$$\begin{cases} -\Delta_H u_{i,j} = \hat{f}_{i,j}, i,j = 1,\ldots,N-1, \\ u_{i,j} = 0, \; i = 0 \vee i = N \vee j = 0 \vee j = N, \end{cases} \tag{2.5.45}$$

where

$$\hat{f}_{i,j} = \frac{1}{|supp(\phi_{ij})|} \int_{supp(\phi_{ij})} f\phi_{ij} \, dx.$$

The linear system (2.5.45) is analogous to the one considered in section 2.2.6. Nevertheless, here the second member is $\hat{f}_{i,j}$ while, in the system considered before, the second member was $f(x_{1,i},x_{2,j})$.

As for the truncation error

$$\|T_H\|_{-1} \leq Ch^2,$$

holds, provided that $u \in C^4(\overline{\Omega})$, we conclude that

$$\|u - u_H\|_1 \leq Ch^2\|u\|_{C^4(\overline{\Omega})}$$

that is,

$$\|R_H u - R_H u_H\|_1 \leq Ch^2\|u\|_{C^4(\overline{\Omega})},$$

where $\|.\|_1$ is defined by (2.2.33). We proved that with respect to this discrete version of the $H^1(\Omega)$-norm, the piecewise liner finite element solution is second order convergent.

Finally we point out that similar results can be obtained if nonuniform meshes are considered.

2.6 The Ritz-Galerkin Method for Time-Dependent PDEs

2.6.1 The RG Solution

In what follows we introduce the Ritz-Galerkin approximation for the solution of the problem

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} + f \text{ in} (a,b) \times (0,T], \\ u(x,0) = u_0(x), x \in (a,b), \\ u(a,t) = u(b,t) = 0, t \in (0,T], \end{cases} \tag{2.6.1}$$

where the reaction term f can be x and t dependent. In order to do that we define the weak formulation of the IBVP (2.6.1) and we approximate this problem by a new variational problem on a finite dimensional space with respect to the space variable. An ordinary differential system is deduced whose solution is called Ritz-Galerkin solution.

The Weak Solution

Let us consider $v \in C_0^\infty(a,b)$. From the PDEs of the IBVP (2.6.1) we get

$$\int_a^b \frac{\partial u}{\partial t} v \, dx = -\int_a^b \frac{\partial u}{\partial x} v' \, dx + \int_a^b f v \, dx, \forall v \in C_0^\infty(a,b).$$

We introduce the problem

$$\begin{aligned} &\text{find } u \in L^2(0,T,H_0^1(a,b)) : \frac{\partial u}{\partial t} \in L^2(0,T,L^2(a,b)), u(x,0) = u_0(x),\ x \in (a,b), \\ &(\frac{\partial u}{\partial t}, v) + a(u(t),v) = (f,v),\ \forall v \in H_0^1(0,a), \\ &a(w,v) = \int_a^b w'v' \, dx, w, v \in H_0^1(a,b). \end{aligned} \tag{2.6.2}$$

By $L^2(0,T,H_0^1(a,b))$ we denote the space of functions $v(x,t)$ such that, for each $t \in (0,T)$, $v(.,t) \in H_0^1(a,b)$, that is $v(t) \in H_0^1(a,b)$, and $\displaystyle\int_a^b \|v(s)\|^2_{H^1(a,b)} \, ds < \infty$.

The solution of the variational IVP (2.6.2) is called weak solution of the IBVP (2.6.1). It is clear that if u is a classical solution of the BVP (2.6.1), then u is also a weak solution. Otherwise, if u is a weak solution and it is smooth enough, then, from (2.6.2), we get

$$(\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - f, v) = 0, \forall v \in C_0^\infty(a,b).$$

Consequently,

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - f = 0$$

in $L^2(a,b)$. If $\dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2}$ is continuous, the last equality holds in (a,b).

Let us consider, in the variational equation, $v = u(t)$. Then we get

$$(\frac{\partial u}{\partial t}, u(t)) + a(u(t),u(t)) = (f, u(t)).$$

As

$$(\frac{\partial u}{\partial t}, u(t)) = \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2},\ \int_a^b (\frac{\partial u}{\partial x})^2\, dx \geq \frac{2}{(b-a)^2}\|u(t)\|^2_{L^2}$$

and

$$(f, u(t)) \leq \|f\|_{L^2}\|u(t)\|_{L^2} \leq \frac{1}{4\epsilon^2}\|f\|^2_{L^2} + \epsilon^2\|u\|^2_{L^2},$$

we establish the differential inequality

$$\frac{d}{dt}\|u(t)\|^2_{L^2} + (\frac{4}{(b-a)^2} - 2\epsilon^2)\|u(t)\|^2_{L^2} \leq \frac{1}{2\epsilon^2}\|f\|_{L^2}, \tag{2.6.3}$$

which can be rewritten in the equivalent form

$$\frac{d}{dt}\left(\|u(t)\|^2_{L^2} e^{(\frac{4}{(b-a)^2} - 2\epsilon^2)t} - \frac{1}{2\epsilon^2}\int_0^t e^{(\frac{4}{(b-a)^2} - 2\epsilon^2)s}\|f\|^2_{L^2}\, ds\right) \leq 0.$$

Consequently, $\|u(t)\|^2_{L^2} e^{(\frac{4}{(b-a)^2} - 2\epsilon^2)t} - \frac{1}{2\epsilon^2}\int_0^t e^{(\frac{4}{(b-a)^2} - 2\epsilon^2)s}\|f\|^2_{L^2}\, ds$ decreases in time. Due to this fact we deduce the following upper bound

$$\|u(t)\|^2_{L^2} \leq e^{-(\frac{4}{(b-a)^2} - 2\epsilon^2)t}\|u_0\|^2_{L^2} + \frac{1}{2\epsilon^2}\int_0^t e^{(\frac{4}{(b-a)^2} - 2\epsilon^2)(s-t)}\|f\|^2_{L^2}\, ds,\ t \geq 0. \tag{2.6.4}$$

If, in (2.6.4), we fix ϵ such that $\frac{4}{(b-a)^2} - 2\epsilon^2 > 0$, we conclude the proof of the following result:

Theorem 2.6.1 *The variational problem (2.6.2) has at most one solution which satisfies (2.6.4).*

■

The estimate (2.6.4) can be modified in order to get some information for the behaviour of the $\|\frac{\partial u}{\partial x}(t)\|_{L^2}$. In fact, from the variational problem we obtain

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2} + \frac{d}{dt}\int_0^t \|\frac{\partial u}{\partial x}(s)\|^2_{L^2}\, ds \leq \frac{1}{2}\|f\|^2_{L^2} + \frac{1}{2}\|u(t)\|^2_{L^2}, \tag{2.6.5}$$

which leads to

$$\frac{d}{dt}\Big(\|u(t)\|^2_{L^2} + 2\int_0^t \|\frac{\partial u}{\partial x}(s)\|^2_{L^2}\, ds\Big) \leq \|u(t)\|^2_{L^2} + 2\int_0^t \|\frac{\partial u}{\partial x}(s)\|^2_{L^2}\, ds + \|f\|^2_{L^2}. \tag{2.6.6}$$

From last inequality we get

$$\frac{d}{dt}\Big(e^{-t}\Big(\|u(t)\|^2_{L^2} + 2\int_0^t \|\frac{\partial u}{\partial x}(s)\|^2_{L^2}\, ds\Big) - \int_0^t e^{-s}\|f\|^2_{L^2}\, ds\Big) \leq 0. \tag{2.6.7}$$

From (2.6.7) we easily conclude the following result:

Theorem 2.6.2 *There exists at most one solution of the variational problem (2.6.2) such that*

$$\|u(t)\|^2_{L^2} + 2\int_0^t \|\frac{\partial u}{\partial x}(s)\|^2_{L^2}\, ds \leq e^t\|u_0\|^2_{L^2} + \int_0^t e^{(t-s)}\|f\|^2_{L^2}\, ds,\ t \in [0, T]. \tag{2.6.8}$$

■

Comparing the estimates (2.6.4), (2.6.8), from the second one we conclude that

$$\int_0^t \|\frac{\partial u}{\partial x}(s)\|_{L^2}^2 \, ds$$

is bounded in bounded time intervals while, from the first one, we only obtain information on the behaviour of

$$\|u(t)\|_{L^2}^2.$$

If we have not reaction term, then $\|u(t)\|_{L^2} \to 0, t \to \infty$. This asymptotic behaviour is deduced from (2.6.4) provided that $\epsilon^2 < \frac{2}{(b-a)^2}$.

Theorem 2.6.2 enable us to conclude the stability of the weak solution of the initial boundary value problem (2.6.1) with respect to perturbations of the initial condition. In fact, if u and $\tilde{u}$ are weak solutions of the IBVP (2.6.1), then $w = u - \tilde{u}$ is a weak solution of this IBVP with $f = 0$ and with the initial condition $u_0 - \tilde{u}_0$. Then

$$\|w(t)\|_{L^2}^2 + 2\int_0^t \|\frac{\partial w}{\partial x}(s)\|_{L^2}^2 \, ds = \|u_0 - \tilde{u}_0\|_{L^2}^2, t \in [0,T], \tag{2.6.9}$$

holds. Consequently

$$\|w(t)\|_{L^2}^2 \leq \|u_0 - \tilde{u}_0\|_{L^2}^2, t \in [0,T],$$

and

$$\int_0^t \|\frac{\partial w}{\partial x}(s)\|_{L^2}^2 \, ds \leq \|u_0 - \tilde{u}_0\|_{L^2}^2, t \in [0,T].$$

Considering the estimate (2.6.4) we also have

$$\|w(t)\|_{L^2} \leq e^{-\frac{2}{(b-a)^2}t}\|u_0 - \tilde{u}_0\|_{L^2} \to 0, t \to \infty.$$

The RG Solution

Let V_H be a subspace of $H_0^1(a,b)$ with $dim V_H = N_H$. Let u_H be defined in $[a,b] \times [0,T]$, such that, for each $t \in [0,T]$, $u_H(.,t) \in V_H$, and

$$(\frac{\partial u_H}{\partial t}, v_H) + a(u_H(t), v_H) = (f, v_H), \forall v_H \in V_H, \tag{2.6.10}$$

and

$$u_H(0) = u_{0,H}, \tag{2.6.11}$$

where $u_{0,H} \in V_H$ is an approximation for u_0. The solution of the variational problem (2.6.10), (2.6.11) is called the Ritz-Galerkin approximation for the solution of the IBVP (2.6.1).

As in the Ritz-Galerkin method, the RG solution is computed considering in V_H a basis. Let $\{\phi_i, i = 1, \ldots, N_H\}$ be such basis. Then

$$u_H(x,t) = \sum_{j=1}^{N_h} \alpha_j(t)\phi_j(x), x \in [a,b], t \geq 0,$$

where the coefficients satisfy

$$\sum_{i=1}^{N_h} \alpha_i'(t)(\phi_i, \phi_j) + \sum_{i=1}^{N_h} \alpha_i(t) a(\phi_i, \phi_j) = (f, \phi_j),\ j = 1, \ldots, N_h,$$

which can be rewritten in equivalent form

$$[(\phi_i, \phi_j)]\alpha'(t) + [a(\phi_i, \phi_j)]\alpha(t) = F,\ t \in (0, T], \tag{2.6.12}$$

where $\alpha(t) = (\alpha_i(t))$ and $F_i = (f, \phi_i)$. The initial condition for the ordinary differential system (2.6.12) is obtained from the initial condition

$$u_{0,H}(x) = \sum_{j=1}^{N_H} \alpha_j(0)\phi_j(x), x \in [a, b].$$

For the particular case of the finite element method, the linear system to be solved for the computation of the components of the RG solution is characterized by sparse matrices. Moreover, the solution of the ordinary differential system is the vector of the finite element solution in the nodes of the partition. In fact, let $\{x_j, x_0 = a, x_{N_H-1} = b, x_{j+1} - x_j = h_j\}$ be a partition of $[a, b]$ and let $\{\phi_j\}$ be a basis of V_H such that $\phi_j(x_i) = \delta_{ij}$. Then

$$u_H(x, t) = \sum_{j=1}^{N_h} u_H(x_j, t)\phi_j(x), x \in [a, b], t \geq 0,$$

where the coefficients $u_H(x_j, t), j = 1, \ldots, N_H$, are defined by the linear system (2.6.12).

Let $e_H(t) = u(t) - u_H(t)$ be the error of the RG solution $u_H(t), t \in [0, T]$. This error is solution of the variational problem

$$\begin{cases} (\dfrac{\partial e_H}{\partial t}, v_H) + a(e_H(t), v_H) = 0,\ \forall v_H \in V_H, \\ e_H(0) = u_0 - u_{0,H}. \end{cases} \tag{2.6.13}$$

In what follows we establish an estimate for the solution of the initial value problem (2.6.13) with respect to the norm $\|.\|_{L^2}$. In order to do that, we introduce the auxiliary function $\tilde{u}_H(t)$ defined by

$$a(\tilde{u}_H(t), v_H) = -(\frac{\partial u}{\partial t}, v_H) + (f, v_H),\ \forall v_H \in V_H. \tag{2.6.14}$$

We split the error $e_H(t)$

$$e_H(t) = \rho(t) + \theta(t), \tag{2.6.15}$$

where

$$\rho(t) = u(t) - \tilde{u}_H(t),\ \theta(t) = \tilde{u}_H(t) - u_H(t).$$

An estimate for $e_H(t)$ is obtained estimating separately $\theta(t)$ and $\rho(t)$.

As the error $\rho(t)$ can be estimated from the results for the time independent Ritz-Galerkim solution, an estimate for the error $e_H(t)$ is obtained estimating $\theta(t)$. In the estimation procedure arises $\dfrac{\partial \tilde{u}_H}{\partial t}$ which is solution of the variational equation

$$a(\frac{\partial \tilde{u}_H}{\partial t}, v_H) = -(\frac{\partial^2 u}{\partial t^2}, v_H) + (\frac{\partial f}{\partial t}, v_H),\ \forall v_H \in V_H. \tag{2.6.16}$$

Its existence depends on the regularity of the weak solution u, more precisely, on the existence of the derivatives $\frac{\partial^2 u}{\partial t^2} \in L^2(a,b)$, $\frac{\partial f}{\partial t} \in L^2(a,b)$. Under these assumptions, $\frac{\partial \theta}{\partial t}$ satisfies

$$(\frac{\partial \theta}{\partial t}, v_H) + a(\theta(t), v_H) = (f, v_H) - (\frac{\partial \tilde{u}_H}{\partial t}, v_H) - a(\tilde{u}_H(t), v_H).$$

Considering that $\tilde{u}_H$ is solution of (2.6.14), we obtain

$$(\frac{\partial \theta}{\partial t}, v_H) + a(\theta(t), v_H) = (\frac{\partial u}{\partial t} - \frac{\partial \tilde{u}_H}{\partial t}, v_H)$$

which is equivalent to

$$(\frac{\partial \theta}{\partial t}, v_H) + a(\theta(t), v_H) = -(\frac{\partial \rho}{\partial t}, v_H),\ v_H \in V_H. \tag{2.6.17}$$

As estimate for $\frac{\partial \rho}{\partial t}$ can be established noting that $\frac{\partial \tilde{u}_H}{\partial t}$ satisfies (2.6.16) and it is the Ritz-Galerkin solution which approximates the solution of the variational problem

$$a(\frac{\partial u}{\partial t}, v) = -(\frac{\partial^2 u}{\partial t^2}, v) + (\frac{\partial f}{\partial t}, v),\ \forall v \in H_0^1(a,b). \tag{2.6.18}$$

Let us consider, in (2.6.17), $v_H = \theta(t)$. As

$$\|\theta(t)\|_{L^2} \frac{d}{dt}\|\theta(t)\|_{L^2} = (\frac{\partial \theta}{\partial t}, \theta(t)),$$

we obtain

$$\|\theta(t)\|_{L^2} \frac{d}{dt}\|\theta(t)\|_{L^2} + a(\theta(t), \theta(t)) = -(\frac{\partial \rho}{\partial t}, \theta(t)).$$

By the Poincaré-Friedrichs inequality we deduce the following differential inequality

$$\frac{d}{dt}\|\theta(t)\|_{L^2} + \frac{2}{(b-a)^2}\|\theta(t)\|_{L^2} \leq \|\frac{\partial \rho}{\partial t}\|_{L^2}. \tag{2.6.19}$$

which is equivalent to

$$\frac{d}{dt}\left(e^{\frac{2}{(b-a)^2}t}\|\theta(t)\|_{L^2} - \int_0^t e^{\frac{2}{(b-a)^2}s}\|\frac{\partial \rho}{\partial t}\|_{L^2}\, ds \right) \leq 0. \tag{2.6.20}$$

The next inequality

$$\|\theta(t)\|_{L^2} \leq e^{-\frac{2}{(b-a)^2}t}\left(\int_0^t e^{\frac{2}{(b-a)^2}s}\|\frac{\partial \rho}{\partial t}\|_{L^2}\, ds + \|\theta(0)\|_{L^2} \right) \tag{2.6.21}$$

is easily deduced from (2.6.20) because $e^{\frac{2}{(b-a)^2}t}\|\theta(t)\|_{L^2} - \int_0^t e^{\frac{2}{(b-a)^2}s}\|\frac{\partial \rho}{\partial t}\|_{L^2}\, ds$ is not increasing function.

Taking into account, in the error decomposition (2.6.15), the estimate (2.6.21), we finally obtain

$$\|u(t) - u_H(t)\|_{L^2} \leq \|\rho(t)\|_{L^2} + e^{-\frac{2}{(b-a)^2}t}\left(\int_0^t e^{\frac{2}{(b-a)^2}s}\|\frac{\partial \rho}{\partial t}\|_{L^2}\, ds + \|\theta(0)\|_{L^2} \right). \tag{2.6.22}$$

We proved the next result:

Theorem 2.6.3 *Let $u(t)$ be the weak solution of the IBVP (2.6.1) defined by (2.6.2) and let $u_H(t)$ be its Ritz-Galerkin approximation defined by (2.6.10), (2.6.11). If $\frac{\partial^2 u}{\partial t^2} \in L^2(a,b)$, $\frac{\partial f}{\partial t} \in L^2(a,b)$, then for $e_H(t), t \in [0,T]$, holds (2.6.22), where $\theta(t) = \tilde{u}_H(t) - u_H(t), \rho(t) = u - \tilde{u}_H(t)$ and $\tilde{u}_H(t)$ is defined by (2.6.14).*

■

We remak that $\theta(0) = \tilde{u}_H(0) - u_{0,H}$, where $\tilde{u}_H(0)$ satisfies

$$a(\tilde{u}_H(0), v_H) = -(\frac{\partial u}{\partial t}(0), v_H) + (f(0), v_H),\ \forall v_H \in V_H.$$

If we consider $u_{0,H} = \tilde{u}_H(0)$ then $\theta(0) = 0$. Otherwise this quantity should be estimated.

Let us particularize now the previous result for the piecewise linear finite element method. In this case, using the Aubin-Nitsche duality arguments, it can be shown that

$$\|\rho(t)\|_{L^2} \leq Ch^2 |u(t)|_{H^2(a,b)}, \tag{2.6.23}$$

provided that $u(t) \in H^2(a,b)$. Analogously, if $\frac{\partial^2 u}{\partial t^2}(t) \in L^2(a,b)$ and $\frac{\partial u}{\partial t}(t) \in H^2(a,b)$ then

$$\|\frac{\partial \rho}{\partial t}\|_{L^2} \leq Ch^2 |\frac{\partial u}{\partial t}|_{H^2(a,b)}. \tag{2.6.24}$$

Considering the estimates (2.6.23), (2.6.24) in Theorem 2.6.3 we conclude for the picewise linear RG solution $u_H(t)$ the following estimate

$$\|u(t) - u_H(t)\|_{L^2} \leq Ch^2 \big(|u(t)|_{H^2(a,b)} + \int_0^t |\frac{\partial u}{\partial t}|_{H^2(a,b)}\, ds\big),\ t \in [0,T].$$

A General Parabolic Problem

Let Ω be a bounded open set of $\mathbb{R}^n$ with boundary $\partial\Omega$ and let $T > 0$. We introduce in what follows the weak solution and the Ritz-Galerkin solution for the following IBVP

$$\begin{cases} \dfrac{\partial u}{\partial t} = \displaystyle\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}(a_{ij}\frac{\partial u}{\partial x_i}) - \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} - cu + f \text{ em } \Omega \times (0,T], \\ u(x,0) = u_0(x), x \in \Omega, \\ u(x,t) = 0, x \in \partial\Omega,\ t \in (0,T], \end{cases} \tag{2.6.25}$$

where $a_{ij} = a_{ji}$. We suppose that the coefficient functions are bounded in $\overline{\Omega} \times [0,T]$ and c, b satisfy

$$c - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i} \geq 0,\ \forall x \in \overline{\Omega}, \forall t \geq 0. \tag{2.6.26}$$

We also assume that there exists a positive constant α_0 such that

$$\xi^t [a_{ij}] \xi \geq \alpha_0 \|\xi\|^2,\ \forall \xi \in \mathbb{R}^n, \forall x \in \overline{\Omega}, \forall t \geq 0. \tag{2.6.27}$$

The Weak Solution

By $L^2(0,T,H_0^1(\Omega))$ we denote the space of functions v defined in $[0,T]\times\overline{\Omega}$ such that, for $t\in(0,T]$, $v(t)\in H_0^1(\Omega)$ and $\displaystyle\int_0^T \|v(s)\|^2_{H^1(\Omega)}\,ds<\infty$. Let $\phi\in C_0^\infty(\Omega)$. From the PDEs of the IBVP we easily obtain

$$(\frac{\partial u}{\partial t},\phi)+\sum_{ij=1}^{n}\int_\Omega a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial \phi}{\partial x_j}\,dx+\sum_{i=1}^{n}\int_\Omega b_i\frac{\partial u}{\partial x_i}\phi\,dx+\int_\Omega cu\phi\,dx=\int_\Omega f\phi dx.$$

Then

$$(\frac{\partial u}{\partial t},v)+a(u(t),v)=(f,v)\ \ \forall v\in H_0^1(\Omega),\ \forall t>0, u(0)=u_0, \tag{2.6.28}$$

where

$$a(.,.):H_0^1(\Omega)\times H_0^1(\Omega)\to \mathbb{R}$$
$$a(w,v)=\sum_{i,j=1}^{n}\int_\Omega a_{ij}\frac{\partial w}{\partial x_j}\frac{\partial v}{\partial x_i}\,dx+\sum_{i=1}^{n}\int_\Omega b_i\frac{\partial w}{\partial x_i}v\,dx+\int_\Omega cwv\,dx\,,\ \ w,v\in H_0^1(\Omega). \tag{2.6.29}$$

The weak solution of the IBVP (2.6.25) is the function $u\in L^2(0,T,H_0^1(\Omega))$ such that $\dfrac{\partial u}{\partial t}\in L^2(0,T,L^2(\Omega))$ and u satisfies (2.6.28). It is easy to show that if u is solution of the IVBP (2.6.28), then u is weak solution of this problem. Otherwise, if u is a weak solution of the IBVP (2.6.28) and it is smooth enough, then u is also solution of the IBVP (2.6.28).

We study in what follows the behabiour of the weak solution. Let us consider, in (2.6.28), $v=u(t)$. Taking into account the assumptions (2.6.27) and (2.6.26) for the coefficient functions we have

$$\begin{aligned} a(u(t),u(t)) &\geq \alpha_0\int_\Omega \nabla u(t)^2\,dx+\sum_{i=1}^{n}\int_\Omega b_i\frac{\partial u}{\partial x_i}u(t)+cu^2\,dx\\ &= \alpha_0\|\nabla u(t)\|^2_{L^2}+\int_\Omega (c-\frac{1}{2}\sum_{i=1}^{n}\frac{\partial b_i}{\partial x_i})u^2\,dx\\ &\geq \alpha_0\|\nabla u(t)\|^2_{L^2}. \end{aligned}$$

Considering now the Poincaré - Friedrichs inequality we deduce

$$a(u(t),u(t))\geq C\|u(t)\|^2_{L^2}. \tag{2.6.30}$$

Combining (2.6.28), for $v=u(t)$, with (2.6.30) the following inequality

$$\frac{1}{2}\frac{d}{dt}\|u\|^2_{L^2}+(C-\epsilon^2)\|u(t)\|^2_{L^2}\leq\frac{1}{4\epsilon^2}\|f\|^2_{L^2}, \tag{2.6.31}$$

can be established for an arbitrary nonzero constant ϵ.

Integrating the differential inequality (2.6.31) we finally obtain

$$\|u(t)\|^2_{L^2}\leq e^{-2(C-\epsilon^2)t}\|u(0)\|^2_{L^2}+\int_0^t e^{2(C-\epsilon^2)(s-t)}\frac{1}{2\epsilon^2}\|f\|^2_{L^2}\,ds. \tag{2.6.32}$$

Theorem 2.6.4 *Under the assumptions (2.6.27), (2.6.26), if the variational problem (2.6.28) has a solution $u\in L^2(0,T,H_0^1(\Omega))$such that $\dfrac{\partial u}{\partial t}\in L^2(\Omega)$, then u is unique. Moreover, such solution satisfies (2.6.32) and it is stable with respect to perturbations of the initial condition.*

■

The estimate (2.6.32) gives information on the behaviour of u. Nevertheless, we can establish a new estimate which allow us to get some information on the gradient of u. It is easy to prove that

$$\frac{d}{dt}\big(\|u(t)\|^2_{L^2} + 2\alpha_0 \int_0^t \|\nabla u(s)\|^2_{L^2}\, ds\big) \leq \|f\|^2_{L^2} + \|u(t)\|^2_{L^2}.$$

Then we also have

$$\frac{d}{dt}\Big(\|u(t)\|^2_{L^2} + 2\alpha_0 \int_0^t \|\nabla u(s)\|^2_{L^2}\, ds\Big) \leq \|u(t)\|^2_{L^2} + 2\alpha_0 \int_0^t \|\nabla u(s)\|^2_{L^2}\, ds + \|f\|^2_{L^2},$$

which implies

$$\|u(t)\|^2_{L^2} + 2\alpha_0 \int_0^t \|\nabla u(s)\|^2_{L^2}\, ds \leq e^t \|u_0\|^2_{L^2} + \int_0^t e^{(t-s)} \|f\|^2_{L^2}\, ds. \tag{2.6.33}$$

The RG Solution

Let us consider (2.6.28) with $H^1_0(\Omega)$ replaced by $V_H \subset H^1_0(\Omega)$ with $dim V_H < \infty$. The RG solution u_H is such that, for each $t \in [0,T]$, $u_H(t) \in V_H$ and

$$\begin{cases} (\dfrac{\partial u_H}{\partial t}, v_H) + a(u_H(t), v_H) = (f, v_H),\ \forall v_H \in V_H,\ \forall t > 0, \\ \\ u_H(0) = u_{0,H}, \end{cases} \tag{2.6.34}$$

where $u_{0,H} \in V_H$ is an approximation for u_0.

We remark that the qualitative properties of the weak solution can be considered for the RG solution. For instance, it is easy to prove the following

$$\|u_H(t)\|_{L^2} \leq e^{-Ct} \|u_H(0)\|_{L^2} + \int_0^t e^{C(s-t)} \|f\|_{L^2}\, ds,\ t \geq 0.$$

Similarly, we also have

$$\|u_H(t)\|^2_{L^2} + 2\alpha_0 \int_0^t \|\nabla u_H(s)\|^2_{L^2}\, ds \leq e^t \|u_H(0)\|^2_{L^2} + \int_0^t e^{(t-s)} \|f\|^2_{L^2}\, ds. \tag{2.6.35}$$

The RG solution is easily computed if we fix in V_H a basis. In fact, if $\{\phi_i, i = 1, \ldots, N_h\}$ is a basis of V_H, then for the coefficients $\alpha_i(t), i = 1, \ldots, N_h$, such that

$$u_H(x,t) = \sum_i \alpha_i(t) \phi_i(x),$$

we obtain the ordinary differential system

$$[(\phi_i, \phi_j)]\alpha'(t) + [a(\phi_i, \phi_j)]\alpha(t) = F,\ t \in (0,T]. \tag{2.6.36}$$

Let $e_H(t) = u(t) - u_H(t)$ be the error for the RG approximation. The study of this error follows the procedures used for the one-dimensional introductory example.

Let $\tilde{u}_H(t) \in V_H$ be defined by

$$a(\tilde{u}_H(t), v_h) = (f, v_H) - (\frac{\partial u}{\partial t}, v_H), v_H \in V_H. \tag{2.6.37}$$

Under the assumptions for the coefficient functions, $a(.,.)$ is continuous, $H_0^1(\Omega)$-elliptic and the functional

$$\ell(v_H) = (f, v_H) - (\frac{\partial u}{\partial t}, v_H), v_H \in V_H,$$

is continuous. Then, by the Lax-Milgram Lemma, we conclude the existence of $\tilde{u}_H(t) \in V_H$. Furthermore, if $\frac{\partial f}{\partial t}, \frac{\partial^2 u}{\partial t^2} \in L^2(\Omega)$, then there also exists the solution of the new problem

$$a(\frac{\partial \tilde{u}_H}{\partial t}(t), v_H) = (\frac{\partial f}{\partial t}, v_H) - (\frac{\partial u^2}{\partial t^2}, v_H), v_H \in V_H. \tag{2.6.38}$$

As the error e_H can be decomposed as the sum between $\rho(t) = u(t) - \tilde{u}_h(t)$ and $\theta(t) = \tilde{u}_H(t) - u_H(t)$, an estimate for e_H is obtained estimating $\theta(t)$ and $\rho(t)$. As $\tilde{u}_h(t)$ is the Ritz-Galerkin approximation for the solution of the variational problem

$$a(w, v) = -(\frac{\partial u}{\partial t}, v) + (f, v), v \in H_0^1(\Omega),$$

then the estimates for $\rho(t)$ are obtained by using this fact. For the term $\theta(t)$ it can be shown the following inequality

$$\|\theta(t)\|_{L^2} \leq e^{-Ct}\|\theta(0)\|_{L^2} + \int_0^t e^{C(s-t)}\|\frac{\partial \rho}{\partial t}\|_{L^2}\, ds, \tag{2.6.39}$$

because $\frac{\partial \theta}{\partial t} \in L^2(\Omega)$ and

$$(\frac{\partial \theta}{\partial t}, \theta(t)) + a(\theta(t), \theta(t)) = -(\frac{\partial \rho}{\partial t}, \theta(t)).$$

Considering now the decomposition of the error e_H, we obtain

$$\|u_H(t) - u(t)\|_{L^2} \leq \|\rho(t)\|_{L^2} + e^{-Ct}\|\theta(0)\|_{L^2} + \int_0^t e^{C(s-t)}\|\frac{\partial \rho}{\partial t}\|_{L^2}\, ds. \tag{2.6.40}$$

In what concerns the term $\|\theta(0)\|_{L^2}$ we remark that using the definition of $\theta(t)$, we have $\theta(0) = \tilde{u}_H(0) - u_{0,H}$. If we choose, as in the introductory example, satisfying $u_{0,H} = \tilde{u}_H(0)$ where $\tilde{u}_H(0)$ that is

$$a(u_{0,H}, v_H) = -(\frac{\partial u}{\partial t}(0), v_H) + (f(0), v_H), \forall v_H \in V_H, \tag{2.6.41}$$

then

$$\|\theta(0)\|_{L^2} = 0. \tag{2.6.42}$$

We proved the following:

Theorem 2.6.5 *Let $u(t)$ be the weak solution of the IBVP (2.6.25) defined by (2.6.34) and let $u_H(t)$ be its Ritz-Galerkin approximation defined by (2.6.34). If $\frac{\partial^2 u}{\partial t^2} \in L^2(a,b)$, $\frac{\partial f}{\partial t} \in L^2(a,b)$, then for $e_H(t), t \in [0,T]$, holds (2.6.40), provided that the coefficient function a_{ij}, b_i and c satisfy (2.6.27), (2.6.26).*

■

Let Ω be a polygonal domain of $\mathbb{R}^2$ and let $\mathcal{T}_H$ be an admissible triangulation of Ω. Let V_H be the space of piecewise linear functions, induced by $\mathcal{T}_H$, which are null on the boundary. Let us suppose that the side lengths of the triangles in $\mathcal{T}_H$ are less or equal to h and the interior angles of the triangles in $\mathcal{T}_H$ are greater or equal to $\beta_0 > 0$. For the particular case of the heat equation, by the Aubin-Nitsche duality arguments was shown that

$$\|\rho_(t)\|_{L^2} \leq Ch^2|u(t)|_{H^2}, \tag{2.6.43}$$

provided that $u(t) \in H^2(\Omega) \cap H_0^1(\Omega)$, and

$$\|\frac{\partial \rho}{\partial t}\|_{L^2} \leq Ch^2|\frac{\partial u}{\partial t}|_{H^2} \tag{2.6.44}$$

provided that $\dfrac{\partial u}{\partial t} \in H^2(\Omega) \cap H_0^1(\Omega)$. Considering, in (2.6.40), the estimates (2.6.43) and (2.6.44), we conclude the following upper bound

$$\|u_h(t) - u(t)\|_{L^2} \leq Ch^2\Big(|u(t)|_{H^2} + \int_0^t |\frac{\partial u}{\partial t}|_{H^2}\, ds\Big)$$

In the following result we summarize the previous considerations:

Theorem 2.6.6 *Let us suppose that Ω is a bounded open polygonal set of $\mathbb{R}^2$ and the coefficient functions of the IBVP (2.6.25), with $n = 2$, satisfy (2.6.27) and (2.6.26). For each $t \in (0, T]$, let $u(t)$ be solution of the variational problem*

$$(\frac{\partial u}{\partial t}, v) + a(u(t), v) = (f, v),\ v \in H_0^1(0, a),$$

where $a(w, v) = (\nabla w, \nabla v), w, v \in H_0^1(\Omega)$. Let $\mathcal{T}_H$ be an admissible triangulation for Ω such that the side lengths are less or equal to h and the interior angles are greater or equal to $\beta_0 > 0$. Let $V_H \subset H_0^1(\Omega)$ be defined by

$$V_H := \big\{v_H \in C^0(\overline{\Omega}) : v_H = 0 \text{ on } \partial\Omega, \quad v_H(x_1, yx_2) = a_0 + a_1x_1 + a_2x_2,\ (x_1, x_2) \in T,\ T \in \mathcal{T}_H\big\}.$$

and let $u_H(t)$ be the piecewise linear finite element solution in V_H defined by

$$(\frac{\partial u_H}{\partial t}, v_H) + a(u_H(t), v_H) = (f, v_H),\ v_H \in V_H,$$

$$u_H(0) = u_{0,H} \in V_h,$$

where $u_{0,H}$ is fixed according (2.6.42).

If $\dfrac{\partial u}{\partial t} \in H^2(\Omega) \cap H_0^1(\Omega), \dfrac{\partial^2 u}{\partial t^2} \in L^2(\Omega)$, then

$$\|u(t) - u_H(t)\|_{L^2} \leq Ch^2\Big(|u|_{H^2(\Omega)} + \int_0^t |\frac{\partial u}{\partial t}|_{H^2(\Omega)}\, ds\Big),\ t \in [0, T].$$

■

2.6.2 The Time-discrete RG Solution

Let us consider a time integration method for the ordinary differential system (2.6.36) defined in the mesh $\{t_m, t_0 = 0, t_M = T, t_m - t_{m-1} = \Delta t\}$. For instance, if we consider the implicit Euler method, we obtain

$$[(\phi_i, \phi_j)]\frac{\alpha^m - \alpha^{m-1}}{\Delta t} + [a(\phi_i, \phi_j)]\alpha^m = F,\ m = 1, \ldots .$$

which is equivalent to consider the previous time integration method for the RG solution. In fact, from the last equality we have

$$\sum_{j=1}^{N_H}(\frac{\alpha_j^{m+1} - \alpha_j^m}{\Delta t}\phi_j, \phi_i) + \sum_{j=1}^{N_H} a(\alpha_j^m\phi_j, \phi_i) = (f, \phi_i), i = 1, \ldots, N_H,$$

which is equivalent to

$$(\frac{u_H^m - u_H^{m-1}}{\Delta t}, v_H) + a(u_H^m, v_H) + (f, v_h),\ \forall v_H \in V_H, m = 1, \ldots, M, \tag{2.6.45}$$

where

$$u_H^j(x) = \sum_{i=1}^{N_H} \alpha_i^j\phi_i(x),$$

is an approximation for the RG solution $u_H(x, t_j)$.

Let $e_H^m = u(t_m) - u_H^m$ be the error of the time-discrete RG solution u_H^m. This error should converge to zero when the time stepsize converges to zero. Let us consider $n = 2$ and let $u_H(t)$ be the piecewise linear finite element solution. Let us suppose that V_H is induced by a family of triangulations $\mathcal{T}_H, H \in \Lambda$, where the maximum of the side lengths of all triangles in each triangulation $\mathcal{T}_H$ converges to zero when $H \in \Lambda$. The convergence

$$\lim_{\Delta t\to 0, h\to 0} \|u(t_m) - u_H^m\|_{L^2} = 0, \tag{2.6.46}$$

should be verified.

Our aim in what follows is to establish the conditions which allow us to conclude (2.6.46). We start by remarking that

$$\|u(t_m) - u_H^m\|_{L^2} \leq \|u(t_m) - u_H(t_m)\|_{L^2} + \|u_H(t_m) - u_H^m\|_{L^2}, \tag{2.6.47}$$

holds. As in the previous section, an estimate for $\|u(t_m) - u_H(t_m)\|_{L^2}$ can be easily established. We study now $\|\hat{e}_H^m\|_{L^2}$ where $\hat{e}_H^m := u_H(t_m) - u_H^m$. This error satisfies the following

$$(\frac{\hat{e}_H^m - \hat{e}_H^{m-1}}{\Delta t}, v_H) + a(\hat{e}_H^m, v_H) = (T_H^m, v_H), \forall v_H \in V_H, m = 1, \ldots, M. \tag{2.6.48}$$

where

$$T_H^m = \frac{\Delta t}{2}\frac{\partial^2 u_H}{\partial t^2}(t^*), t^* \in (t_{m-1}, t_m).$$

Taking, in (2.6.48), $v_H = \hat{e}_H^m$, we obtain

$$\|\hat{e}_H^m\|_{L^2}^2 + \Delta t a(\hat{e}_H^m, \hat{e}_h^m) = (\hat{e}_H^m, \hat{e}_H^{m-1}) + \Delta t(T_H^m, \hat{e}_H^m).$$

Thus

$$\|\hat{e}_H^m\|_{L^2}^2 + \Delta t a(\hat{e}_H^m, \hat{e}_H^m) \leq \|\hat{e}_H^m\|_{L^2}\|\hat{e}_H^{m-1}\|_{L^2} + \Delta t\|T_H^m\|_{L^2}\|\hat{e}_H^m\|_{L^2}.$$

As

$$a(\hat{e}_H^m, \hat{e}_H^m) \geq \alpha_0\|\nabla\hat{e}_H^m\|_{L^2}^2 \geq C\|\hat{e}_H^m\|_{L^2}^2,$$

we also have

$$\|\hat{e}_H^m\|_{L^2}(1 + C\Delta t) \leq \|\hat{e}_H^{m-1}\|_{L^2} + \Delta t\|T_H^m\|_{L^2}, m = 1, \ldots. \tag{2.6.49}$$

Inequality (2.6.49) implies that

$$\|\hat{e}_H^m\|_{L^2} \leq \Delta t \sum_{j=1}^{m} \frac{1}{(1 + C\Delta t)^{m+1-j}}\|T_H^j\|_{L^2}, m = 1, \ldots, \tag{2.6.50}$$

which induces the upper bound

$$\|\hat{e}_H^m\|_{L^2} \leq \frac{1}{C}\Big(1 - \frac{1}{(1 + C\Delta t)^m}\Big) \max_{j=1,\ldots,m}\|T_H^j\|_{L^2}, m = 1, \ldots, \tag{2.6.51}$$

If we assume that $u_H(t)$ has bounded second order time derivative, then, from (2.6.51), we get

$$\lim_{m\to+\infty}\|\hat{e}_H^m\|_{L^2} = 0. \tag{2.6.52}$$

From the last convergence we finally conclude

$$\lim_{\Delta t\to 0, h\to 0}\max_{m=1,\ldots,M}\|u(t_m) - u_H^m\|_{L^2} = 0, \tag{2.6.53}$$

provided that

$$\lim_{h\to 0}\|u(t_m) - u_H(t_m)\|_{L^2} = 0, \forall m.$$

The smoothness assumption for the RG solution $u_H(t)$, namely that $\dfrac{\partial^2 u_H}{\partial t^2}$ is bounded, had a central role in the proof of the convergence (2.6.53). In what follows we prove the same convergence result avoiding the smoothness assumption previous considered. The procedure that we use is an adaptation of the procedure used in the last section when $\|u(t) - u_H(t)\|_{L^2}$ was estimated.

Let $\tilde{u}_H^m$ be defined by (2.6.37) with $t = t_m$. As

$$u(t_m) - u_H^m = u(t_m) - \tilde{u}_H^m + \tilde{u}_H^m - u_H^m := \rho_H^m + \theta_H^m, \tag{2.6.54}$$

an estimate for e_H^m is obtained estimating separately ρ_H^m and θ_H^m. The term ρ_h^m is the error of the Ritz-Galerkin solution $\tilde{u}_H^m$ which approximates the weak solution $u(t_m) \in H_0^1(\Omega)$. As this error was previously estimated, we only need to estimate θ_H^m. For this last term it is easy to show that

$$(D_{-t}\theta_H^m, \theta_H^m) + a(\theta_H^m, \theta_H^m) = (D_{-t}\tilde{u}_H^m - \frac{\partial u}{\partial t}(t_m), \theta_H^m) \tag{2.6.55}$$

holds, where D_{-t} denotes the backward finite difference operator. Manipulating the expressions in (2.6.55), we deduce

$$\|\theta_H^m\|_{L^2}(1 + C\Delta t) \leq \|\theta_H^{m-1}\|_{L^2} + \Delta t\|D_{-t}\tilde{u}_H^m - \frac{\partial u}{\partial t}(t_m)\|_{L^2}, \tag{2.6.56}$$

which implies

$$\|\theta_h^m\|_{L^2} \leq \frac{1}{(1+C\Delta t)^m}\|\theta_H^0\|_{L^2} + \Delta t \sum_{j=1}^{m} \frac{1}{(1+C\Delta t)^{m-j+1}} \|\frac{\partial u}{\partial t}(t_j) - D_{-t}\tilde{u}_h^j\|_{L^2}. \tag{2.6.57}$$

As

$$D_{-t}u(t_j) = \frac{\partial u}{\partial t}(t_j) - \frac{1}{\Delta t}\int_{t_{j-1}}^{t_j} \frac{\partial^2 u}{\partial t^2}(s)(s-t_{j-1})ds,$$

that is

$$\|D_{-t}u(t_j) - \frac{\partial u}{\partial t}(t_j)\|_{L^2} \leq \int_{t_{j-1}}^{t_j} \|\frac{\partial^2 u}{\partial t^2}(s)\|_{L^2} ds,$$

from (2.6.54) and (2.6.57), we obtain

$$\begin{aligned} \|u_H^m - u(t_m)\|_{L^2} &\leq \|\rho_H^m\|_{L^2} + \frac{1}{(1+C\Delta t)^m}\|\theta_H^0\|_{L^2} \\ &+\Delta t \sum_{j=1}^{m} \frac{1}{(1+C\Delta t)^{m-j+1}} \|D_{-t}\rho_H(t_j)\|_{L^2} \\ &+\Delta t \sum_{j=1}^{m} \frac{1}{(1+C\Delta t)^{m-j+1}} \int_{t_{j-1}}^{t_j} \|\frac{\partial^2 u}{\partial t^2}(s)\|_{L^2} ds. \end{aligned} \tag{2.6.58}$$

Particularizing the previous estimate for $n = 2$ and for the space of piecewise linear functions defined induced by a quasi-uniform admissible triangulation $\mathcal{T}_H$, we establish

$$\|D_{-t}\rho_H^j\|_{L^2} \leq Ch^2 |D_{-t}u(t_j)|_{H^2} \leq Ch^2 \Big(|\frac{\partial u}{\partial t}(t)|_{H^2} + \int_{t_{j-1}}^{t_j} |\frac{\partial^2 u}{\partial t^2}(s)|_{H^2}\, ds \Big),$$

when the heat equation is considered. Then

$$\begin{aligned} \|u_H^m - u(t_m)\|_{L^2} &\leq Ch^2 |u(t_m)|_{H^2(\Omega)} + \frac{1}{(1+C\Delta t)^m}\|\theta_H^0\|_{L^2} \\ &+C\Delta t h^2 \sum_{j=1}^{m} \frac{1}{(1+C\Delta t)^{m-j+1}} \Big(|\frac{\partial u}{\partial t}(t)|_{H^2} + \int_{t_{j-1}}^{t_j} |\frac{\partial^2 u}{\partial t^2}(s)|_{H^2}\, ds \Big) \\ &+\Delta t \sum_{j=1}^{m} \frac{1}{(1+C\Delta t)^{m-j+1}} \int_{t_{j-1}}^{t_j} \|\frac{\partial^2 u}{\partial t^2}(s)\|_{L^2} ds, \end{aligned} \tag{2.6.59}$$

which depends on the regularity of u and on the accuracy of the approximated initial condition $u_{0,H}$. If we consider $u_{0,H} = \tilde{u}_H^0$ then $\theta_H^0 = 0$ then (2.6.59) takes the form

$$\begin{aligned} &\|u_H^m - u(t_m)\|_{L^2} \leq Ch^2 |u(t_m)|_{H^2(\Omega)} \\ &+C\Delta t h^2 \sum_{j=1}^{m} \frac{1}{(1+C\Delta t)^{m-j+1}} \Big(|\frac{\partial u}{\partial t}(t)|_{H^2} + \int_{t_{j-1}}^{t_j} |\frac{\partial^2 u}{\partial t^2}(s)|_{H^2}\, ds \Big) \\ &+\Delta t \sum_{j=1}^{m} \frac{1}{(1+C\Delta t)^{m-j+1}} \int_{t_{j-1}}^{t_j} \|\frac{\partial^2 u}{\partial t^2}(s)\|_{L^2} ds. \end{aligned}$$

2.7 FDM for Time-Dependent PDES

2.7.1 The Method of Lines

The method of lines defines a new approach to solve PDEs where the spatial discretization defined by finite difference operators is combined with a time integration method. In the first step of MOL approach, an ODE is obtained. This ODE is numerically integrated using a specialized time integration method studied before, and a fully discrete numerical approximation for the solution of the PDEs is computed.

This approach offers a grand advantage: it allow the use highly valuable methods in the field of numerical ODEs, some of which were presented in the first chapter. These methods can be of practical use for solving time-dependent PDEs. Another attractive practical point is that there exist nowadays many well developed ODES methods and for these methods sophisticated software is freely available.

For some time dependent PDEs, if we apply a standard ODE method to the ODE problem obtained in the first step of the MOL approach, some information of the underlying PDEs problem might be neglected. Namely, for advection problems where the so called characteristics can be combined with a space-time integration to obtain a more efficient numerical method.

Let us consider a time dependent PDEs defined on a space domain Ω. By Ω_H we denote, as before, a spatial grid depending on a parameter H. Discretizing the spatial derivatives using finite difference operators, we obtain a semi-discrete system (the spatial variable is discrete and the time variable is continuous)

$$\begin{cases} u'_H(t) = F_H(t, u_H(t)), \ t \in (0, T], \\ \\ u_H(0) = u_{0,H}, \end{cases} \tag{2.7.1}$$

where $u_H(t) = (u_{H,j}(t)) \in \mathbb{R}^m$, is called semi-discrete approximation for u and m is proportional to the number of grid points in space. The discretization of the boundary conditions are supposed to be contained in F_H. According to MOL approach, a fully discrete approximation $u_j^n \simeq u(x_i, t_n)$, for the time levels $t_n = n\Delta t, n = 1, \ldots,$ is now obtained by applying some suitable ODE method. As a standard example, we consider the θ-method (1.2.3) studied in chapter 1

$$u_H^{n+1} = u_H^n + \Delta t\Big((1-\theta)F_H(t_n, u_H^n) + \theta F_H(t_{n+1}, u_H^{n+1})\Big),$$

where $u_H^p = (u_{H,i}^p), p = n, n+1$, denotes the vector containing the fully discrete numerical solution at time level $t = t_p$.

The properties of the semi-discrete solution $u_H(t)$ - solution of the initial value problem (2.7.1)- have a central role on the properties fully discrete approximation. Due to this fact, we study in what follows some spatial discretizations for the advection equation and for the diffusion equation.

2.7.2 The Spatial Discretization:Some Qualitative Properties

The scalar advection-diffusion equation with periodic boundary conditions

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = D\dfrac{\partial^2 u}{\partial x^2},\ x \in \mathbb{R}, t \in (0,T], \\ u(x \pm 1, t) = u(x,t),\ x \in \mathbb{R}, t \in [0,T], \\ u(x,0) = u_0(x),\ x \in \mathbb{R}, \end{cases} \tag{2.7.2}$$

where $a \in \mathbb{R}, D \geq 0$, is an important test model for numerical schemes. It is sufficient to consider u on the spatial interval $[0,1]$.

In $\Omega = [0,1]$ we introduce the grid $\overline{\Omega}_H = \{x_i = x_{i-1} + h, i = 1, \ldots, m, x_0 = 0, x_m = 1$. On this space grid, the approximation $u_H(x_i,t)$ for $u(x_j,t)$ is founded replacing, in (2.7.2), the spatial derivatives $\dfrac{\partial}{\partial x}, \dfrac{\partial^2}{\partial x^2}$ by difference operators. We obtain a ODE for $u_H(x_i,t)$

$$u'_H(x_i,t) = \sum_k a_k u_H(x_{i+k},t),$$

which can be rewritten in the vectorial form

$$u'_H(t) = A_H u_H(t), \tag{2.7.3}$$

where A is the square matrix of order m

$$A_H = \begin{bmatrix} a_0 & a_1 & a_2 & . & . & a_{m-1} \\ a_{m-1} & a_0 & a_1 & a_2 & . & a_{m-2} \\ a_{m-2} & a_{m-1} & a_0 & a_1 & . & a_{m-3} \\ . & . & . & . & . & . \\ a_1 & a_2 & a_3 & . & a_{m-1} & a_0 \end{bmatrix}.$$

For example, if we discretize the first and second order spatial derivatives with the finite difference operators D_c and D_2, respectively, we obtain

$$a_0 = -\frac{2D}{h^2}, a_1 = \frac{D}{h^2} - \frac{a}{2h},\ a_2 = 0, \ldots, a_{m-2} = 0, a_{m-1} = \frac{a}{2h} + \frac{D}{h^2}.$$

The matrix A_H is called circulant matrix. We point out that circulant matrices arise in the discretization of periodic PDEs with constant coefficients.

We study now the stability properties of the semi-discrete approximation $u_H(t)$. In the stability analysis we use the Fourier modes

$$\phi_k(x) = e^{2\pi i k x}, k \in \mathbb{Z}.$$

The set $\{\phi_k, k \in \mathbb{Z}\}$ is a orthonormal "basis" of $L^2(0,1)$ and if $v \in L^2(0,1)$, then

$$v = \sum_{k \in \mathbb{Z}} (v, \phi_k)\phi_k,$$

where $(.,.)$ is the usual inner product in $L^2(0,1)$.

By $L^2(\overline{\Omega}_H - \{0\})$ we represent the space of grid functions defined on $\overline{\Omega}_H - \{0\}$, where the discrete L^2 inner product

$$(v_H, w_H)_H = h \sum_{x \in \overline{\Omega}_H - \{0\}} \overline{v_H(x)} w_H(x)$$

is considered. By $\|.\|_{L^2(\overline{\Omega}_H - \{0\})}$ we denote the norm induced by the previous inner product. Let $R_H \phi_k$ be discrete Fourier mode, that is, the restriction of the Fourier mode ϕ_k to the grid $\overline{\Omega}_H - \{0\}$.

The set $\{R_H \phi_k, k = 1, \ldots, m\}$ is a orthonormal basis of $L^2(\overline{\Omega}_H - \{0\})$. In fact, this space can be identified with $\mathbb{C}^m$ and

$$(R_H \phi_k, R_H \phi_\ell)_H = h \sum_{\overline{\Omega}_H - \{0\}} e^{2\pi i(\ell - k)x} = h \sum_{j=1}^{m} e^{2\pi i(\ell-k)jh} = h \sum_{j=1}^{m} \rho^j = \begin{cases} 1, \ \ell = k \\ 0, \ \ell \neq k \end{cases}$$

where $\rho = e^{2\pi i(\ell - k)h}$.

As $\{R_H \phi_k, k = 1, \ldots, m\}$ is a orthonormal basis of $L^2(\overline{\Omega}_H - \{0\})$, if $v_H \in L^2(\overline{\Omega}_H - \{0\})$, then

$$v_H = \sum_{\ell=1}^{m} (v_H, R_H \phi_\ell)_H R_H \phi_\ell.$$

Moreover, a discrete version of the Parseval identity holds

$$\|v_H\|^2_{L^2(\overline{\Omega}_H - \{0\})} = (v_H, v_H)_H = \sum_{\ell} |(v_H, R_H \phi_\ell)|^2.$$

A special property of the circulant matrix A_H is that every discrete Fourier model $R_H \phi_k$ is an eigenvector associated with the eigenvalue

$$\lambda_k = \sum_{j=1}^{m} a_j e^{2\pi i k x_j}.$$

It is easy to show that the solution of (2.7.3) admits the representation

$$u_H(t) = \sum_{k=1}^{m} (u_H(0), R_H \phi_k)_H e^{\lambda_k t} R_H \phi_k.$$

Usually we deal with circulant matrices where all λ_k have a non-positive real part. In this case, we have

$$\begin{aligned} \|u_H(t)\|^2_{L^2(\overline{\Omega}_H - \{0\})} &= \sum_{k=1}^{m} |(u_H(0), R_H \phi_k)_H e^{\lambda_k t}|^2 \\ &\leq \sum_{k=1}^{m} |(u_H(0), R_H \phi_k)_H|^2 \\ &= \|u_H(0)\|^2_{L^2(\overline{\Omega}_H - \{0\})}. \end{aligned}$$

Consequently,

$$\|e^{tA}\|^2_{L^2(\overline{\Omega}_H - \{0\})} \leq 1, \qquad t \geq 0,$$

which shows that (2.7.3) is stable with respect to the norm $\|.\|_{L^2(\overline{\Omega}_H-\{0\})}$.

We rewrite the previous considerations in terms of matrices as we done in the first chapter. Let Q_H be the following matrix $Q_H = \sqrt{h}[R_K\phi_1\, R_H\phi_2 \ldots R_H\phi_m]$ and let $\mathcal{D}$ be the diagonal matrix with entries λ_k. As Q_H is an unitary matrix and $A = Q_H D Q_H^{-1}$, we obtain

$$\|e^{tA}\|^2_{L^2(\overline{\Omega}_H-\{0\})} = \|V_H e^{t\mathcal{D}} Q_H^{-1}\|^2_{L^2(\overline{\Omega}_H-\{0\})} = \max_{k=1,\ldots,m} |e^{\lambda_k t}|.$$

We detail some of the previous conclusions for some particular cases of (2.7.2).

The Advection Equation:

Let us take, in (2.7.2), $D = 0$. We consider the forward finite difference operator D_{-x} when $a > 0$ and D_x when $a < 0$, obtaining the upwind schemes

$$u'_H(x_j,t) = \frac{a}{h}(u_H(x_{j-1},t) - u_H(x_j,t)), j = 1,\ldots,m, u_H(x_0,t) = u_H(x_m,t), \tag{2.7.4}$$

$$u'_H(x_j,t) = \frac{a}{h}(u_H(x_j,t) - u_H(x_{j+1},t)), j = 1,\ldots,m, u_H(x_{m+1},t) = u_H(x_1,t), \tag{2.7.5}$$

respectively.

The upwind scheme (2.7.4) can be rewritten in the equivalent form (2.7.3) with

$$A_H = \frac{a}{h}\begin{bmatrix} -1 & 0 & 0 & . & 0 & 0 & 1 \\ 1 & -1 & 0 & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . \\ . & . & . & 0 & 1 & -1 & 0 \\ 0 & 0 & . & 0 & 0 & 1 & -1 \end{bmatrix}.$$

If we use the finite difference operator D_c then we obtain

$$u'_H(x_j,t) = \frac{a}{2h}(u_H(x_{j-1},t) - u_H(x_{j+1},t)), j = 1,\ldots,m, u_H(x_{m+1},t) = u_H(x_1,t), \tag{2.7.6}$$

which induces the ODE equation $u'_H(t) = A_H u_H(t)$ with

$$A_H = \frac{a}{2h}\begin{bmatrix} 0 & -1 & 0 & . & 0 & 0 & 1 \\ 1 & 0 & -1 & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & 1 & 0 & -1 \\ -1 & 0 & . & . & 0 & 1 & 0 \end{bmatrix}.$$

When the solution of the initial value problem is considered in (2.7.3) and Taylor's expansion is used, we get for the forward finite difference discretization

$$\frac{\partial u}{\partial t}(x_i,t) + a\frac{\partial u}{\partial x}(x_i,t) = T_H(x_i,t),$$

with $T_H(x_i,t) = O(h)$, while in the second case we get exactly the same expression with $T_H(x_i,t) = O(h^2)$. The term T_H is called spatial truncation error. Due to the behaviour of the spatial truncation error, the scheme (2.7.3) obtained with D_x and $D = 0$ is called first order upwind scheme while the scheme obtained with D_c is called second order central scheme.

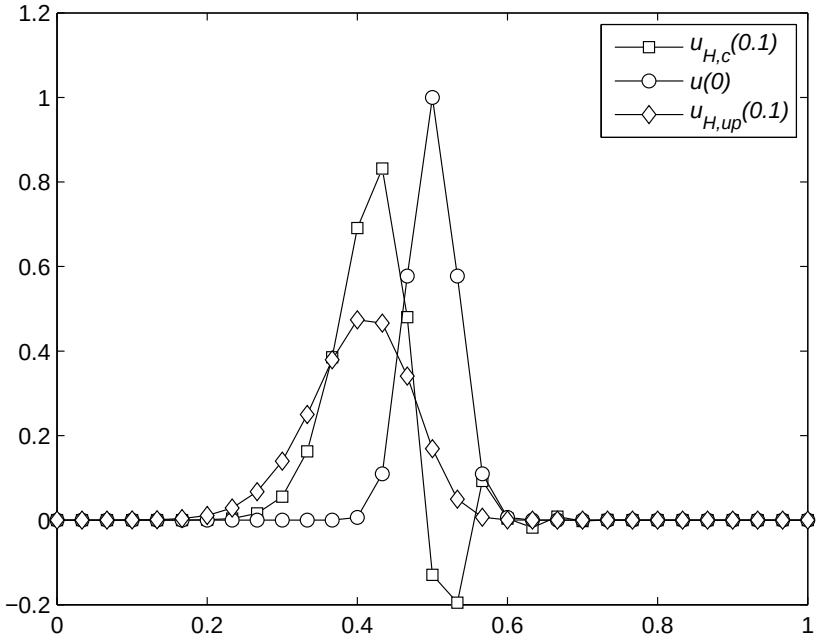

Figure 10: ($a > 0$) Numerical solutions obtained with the upwind scheme ($u_{H,up}$) and with the central scheme ($u_{H,c}$).

It can be observed by experimental results that the first order upwind scheme is not accurate while the second order central scheme induces numerical oscillations (see Figure 10 for $u(x,0) = (sem(\pi x))^{100}, h = 1/50$). We will justify in what follows the previous qualitative behaviour. In order to do that we use the so called modified equation.

For the upwind scheme it can be shown that $T_H(x_i,t) = \frac{1}{2}ah\frac{\partial^2 u}{\partial x^2} + O(h^2)$. Such fact indicates that the solution obtained by this scheme is closer to the solution of the modified equation

$$\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x} = \frac{1}{2}ah\frac{\partial^2 \tilde{u}}{\partial x^2}. \tag{2.7.7}$$

This explains the diffusive behaviour of the first order upwind scheme: although we are computing a solution to the advection equation, we are actually generating a solution that is close to an advection-diffusion equation with diffusion coefficient $\frac{1}{2}ah$. The advection-diffusion equation (2.7.7) is called modified equation for the first order upwind scheme.

As for the second order central scheme $T_H(x_i,t) = -\frac{1}{6}ah^2\frac{\partial^3 \tilde{u}}{\partial x^3} + O(h^4)$, then

$$\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x} = -\frac{1}{6}ah^2\frac{\partial^3 \tilde{u}}{\partial x^3}, \tag{2.7.8}$$

is the modified equation of this scheme. Thus, the solution given by this scheme is a fourth order approximation for the solution of its modified equation. In order to justify the behaviour of the numerical solution defined by the second order central scheme, we look to the behaviour of the solution of the initial value problem (2.7.8) with the initial condition $\tilde{u}(x,0) = e^{2\pi ikx}$. As

such solution is given by

$$\tilde{u}(x,t) = e^{2\pi kix + a_k t} = e^{2\pi ik(x - a(1-\frac{2}{3}\pi^2 k^2 h^2))},$$

all Fourier modes move with different speeds. Consequently, the fine-tuning is lost and the oscillations will occur.

The stability of the semi-discrete approximation depends on the eigenvalues of A_H. It is easy to show that the matrices of the upwind and central schemes have the following eigenvalues

$$\lambda_k = \frac{a}{h}\big(cos(2\pi kh) - 1\big) - i\frac{a}{h}sin(2\pi kh), k = 1, \ldots, m,$$

when $a > 0$, and

$$\lambda_k = -i\frac{a}{h}sen(2\pi kh), k = 1, \ldots, m,$$

respectively. If $a < 0$, then

$$\lambda_k = \frac{-a}{h}\big(cos(2\pi kh) - 1\big) + i\frac{a}{h}sin(2\pi kh), k = 1, \ldots, m.$$

As the eigenvalues λ_k have real non-positive real part ($a > 0$), then e^{tA_H} satisfies $\|e^{tA_H}\| \leq 1$. Hence both schemes are stable. As the eigenvalues for the second order central scheme have null real part then $\|u_H(t)\| = \|u_H(0)\|$. If we consider the upwind scheme (2.7.4) with $a < 0$, then the eigenvalues would be in the right half-plan of the complex plan with real part as large as $-\frac{a}{h}$, and thus this scheme became unstable when $h \to 0$. The eigenvalues of the second order central scheme are in the imaginary axis.

The semi-discrete system $u'_H(t) = A_H u_H(t)$ has the solution $u_H(t) = e^{\lambda_k t}\phi_k$ if the initial condition is $u_H(0) = \phi_k$, where λ_k is the eigenvalue of A_H corresponding to ϕ_k. Component-wise this reads

$$u_H(x_j,t) = e^{\lambda_k t}e^{2\pi ikx_j} = e^{Re\lambda_k t}e^{2\pi ik(x_j - a_k)},\ a_k = -\frac{1}{2\pi k}Im\lambda_k.$$

The correspondent exact solution is given by

$$u(x_j,t) = e^{2\pi ik(x-at)}.$$

The first order upwind scheme has eigenvalues λ_k such that $Re\lambda_k < 0$. This fact implies that

$$|u_H(x_j,t)| = |e^{Re\lambda_k t}| \to 0, t \to \infty.$$

As the $Re\lambda_k = 0$ for the second order central scheme, in this case we have

$$|u_H(x_j,t)| = 1.$$

The factor $e^{Re\lambda_k t}$ determines the amount of numerical damping or dissipation for the k Fourier mode. If $Re\lambda_k < 0$ for all $k \neq m$, then $e^{Re\lambda_k t} \to 0, t \to \infty$, and, consequently, the scheme is said dissipative. Otherwise, if $Re\lambda_k = 0$, then the scheme is said non-dissipative. Obviously, the second order central scheme is non-dissipative whereas the first order upwind scheme is dissipative. The velocity a_k of the kth Fourier mode is called the numerical phase velocity.When the phase velocity differs from a this will leads to a phase error. If they are different from each other we have dispersion. Dispersion may give rise to oscillations.

The first order upwind and the second order central scheme have the same velocity $a_k \neq a$. So the upwind scheme is also dispersive, but oscillations will not show up in actual calculations because the damping factor $e^{Re\lambda_k t}$ suppresses all these Fourier modes.

It is obvious that both schemes (2.7.4) and the central scheme (2.7.6) have drawbacks, the first being too dissipative and the second one too dispersive. Increasing the order of the discretization we can obtain semi-discrete schemes where the dispersion and the dissipation is diminished.

The Diffusion Equation:

Considering, in (2.7.2) with $a = 0, D > 0$, the finite difference operator D_2 we obtain

$$\begin{cases} u'_H(x_j,t) = \dfrac{D}{h^2}\big(u_H(x_{j-1},t) - 2u_H(x_j,t) + u_H(x_{j+1},t)\big), j = 1,\ldots,m, \\ \\ u_H(x_0,t) = u_H(x_m,t), u_H(x_{m+1},t) = u_H(x_1,t), \end{cases} \tag{2.7.9}$$

which is equivalent to $u'_H(t) = A_H u_H(t)$ with

$$A_H = \frac{D}{h^2}\begin{bmatrix} -2 & 1 & 0 & . & 0 & 0 & 1 \\ 1 & -2 & 1 & . & 0 & 0 & 0 \\ . & . & . & . & . & 0 & 0 \\ 0 & 0 & 0 & . & 1 & -2 & 1 \\ 1 & 0 & 0 & . & 0 & 1 & -2 \end{bmatrix}.$$

As for the advection discretizations we can look for the modified equation of the scheme (2.7.9). As

$$D_2 u(x_j,t) = \frac{\partial^2 u}{\partial x^2}(x_j,t) + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(x_j,t) + O(h^4),$$

the modified equation of the scheme (2.7.9) is given by

$$\frac{\partial \tilde{u}}{\partial t} = D\frac{\partial^2 \tilde{u}}{\partial x^2} + \frac{Dh^2}{12}\frac{\partial^4 \tilde{u}}{\partial x^4}. \tag{2.7.10}$$

Then the scheme (2.7.9) defines a fourth order approximation for the solution of the modified equation (2.7.10). Nevertheless, this equation is unstable. In fact, if we compute the solution of the modified equation with initial condition $\tilde{u}(x,0) = e^{2\pi ikx}$, we obtain

$$\tilde{u}(x,t) = e^{-4D\pi^2k^2(1-\frac{1}{3}h^2k^2)t}e^{2\pi ikx}, \tag{2.7.11}$$

which grows for $h^2k^2 > \dfrac{1}{3}$. This instability is an artefact in the sense that the diffusion equation admits solutions composed of Fourier modes

$$\phi_k(x,t) = e^{2\pi ikx}e^{-Dk^2t}.$$

One could include another term into the modified equation, for example, leading to

$$\frac{\partial \tilde{u}}{\partial t} = D\frac{\partial^2 \tilde{u}}{\partial x^2} + \frac{Dh^2}{12}\frac{\partial^4 \tilde{u}}{\partial x^4} + \frac{Dh^4}{360}\frac{\partial^6 \tilde{u}}{\partial x^6}. \tag{2.7.12}$$

The scheme (2.7.9) defines a six order approximation for the solution of the stable new modified equation (2.7.12). In this case (2.7.11) is replaced by

$$\tilde{u}(x,t) = e^{4D\pi^2k^2\left(-1+\frac{h^2\pi^2k^2}{12}\left(1-\frac{4\pi^2h^2}{30}\right)\right)t}e^{2\pi ikx}.$$

The eigenvalues of A_H are real and negative

$$\lambda_k = \frac{2D}{h^2}(cos(2\pi kh)-1) = -\frac{4D}{h^2}sen(\pi kh)^2 \in [-\frac{4D}{h^2},0], k=1,\ldots,m,$$

showing the stability of the discretization. As when $h \to 0$ $\max_k |\lambda_k| \to \infty$, the semi-discrete problem (2.7.9) is usually stiff.

The the semi-discrete solution with initial condition $u_H(0) = \phi_k$ is given by

$$u_H(x_j,t) = e^{\lambda_k t}e^{2\pi ikx_j},$$

while the correspondent solution of the diffusion problem admits the representation

$$u(x_j,t) = e^{-4D\pi^2k^2t}e^{2\pi ikx_j}.$$

As $h \to 0$, $\lambda_k \to -4D\pi^2k^2$, we get

$$e^{\lambda_k t} \to e^{-4D\pi^2k^2t},$$

which means that, for h small enough, u_H and u have the same behaviour. However, as for fixed h, most of the λ_k are not closed to their continuous counterpart. Nevertheless, this discrepancy does not implies a wrong behaviour of the semi-discrete solution.

Higher-Order Schemes

The schemes studied for the advection equation and for the diffusion equation suffer some pathologies. However, increasing the order of the the finite difference schemes the adverse effects can be diminished.

Advection Equation: A general finite difference scheme for the periodic advection problem can be rewritten as

$$u'_H(x_j,t) = \frac{a}{h}\sum_{k=-r}^{s}\gamma_k u_H(x_{j+k},t), j=1,\ldots,m, \tag{2.7.13}$$

where $u_H(x_i,t) = u_H(x_{i+m},t)$ to impose the periodicity condition. Replacing, in (2.7.13), u_H by the solution of the advection equation and by using Taylor's expansion, we obtain

$$\begin{aligned}
&\frac{\partial u}{\partial t}(x_j,t) - \frac{a}{h}\sum_k \gamma_k u(x_{j+k},t)\\
&= -a\frac{\partial u}{\partial x}(x_j,t) - \frac{a}{h}\sum_k \gamma_k\big(u(x_j,t) + kh\frac{\partial u}{\partial x}(x_j,t) + \frac{1}{2}k^2h^2\frac{\partial^2 u}{\partial x^2}(x_j,t) + \ldots\big)\\
&= -\frac{a}{h}\sum_k \gamma_h u(x_j,t) - a\Big(1+\sum_k k\gamma_k\Big)\frac{\partial u}{\partial x}(x_j,t) - \frac{a}{2}h\Big(\sum_k k^2\gamma_k\Big)\frac{\partial^2 u}{\partial x^2}(x_j,t) + \ldots
\end{aligned}$$

If we assume that

$$\sum_k \gamma_k = 0,\ \sum_k k\gamma_k = -1,\ \sum_k k^2\gamma_k = 0\ \ldots\ \sum_k k^q\gamma_k = 0, \tag{2.7.14}$$

then the residual error is of order q. The conditions (2.7.13), usually called order conditions, leads to the coefficients $\gamma_{-r}, \ldots, \gamma_s$. The linear system for such coefficients is characterized by a Vandermonde type matrix which leads to $\gamma_{-r}, \ldots, \gamma_s$ for $q \leq r+s$. If $q = r+s$ then the scheme is unique which satisfies the following result due to Iserls and Strang (1983)([17])

Theorem 2.7.1 *If $a > 0$ and $q = r + s$ with $s \leq r \leq s + 2$, then the scheme (2.7.13) is L^2 stable. Otherwise (2.7.13) is unstable.*

A similar result can be established for $a < 0$ with $r = s, s+1, s+2$.(see [18]).
For $a > 0$ and $r = 2, s = 1$, we obtain the third order upwind advection scheme

$$u'_H(x_j,t) = \frac{a}{h}(-\frac{1}{6}u_H(x_{j-2},t) + u_H(x_{j-1},t) - \frac{1}{2}u_H(x_j,t) - \frac{1}{3}u_H(x_{j+1},t)).$$

For $a < 0$ we have

$$u'_H(x_j,t) = \frac{a}{h}(\frac{1}{3}u_H(x_{j-1},t) + \frac{1}{2}u_H(x_j,t) - u_H(x_{j+1},t) + \frac{1}{6}u_H(x_{j+2},t)).$$

For $r = s = 2$ we get the fourth order central advection scheme

$$u'_H(x_j,t) = \frac{a}{h}(-\frac{1}{12}u_H(x_{j-2},t) + \frac{2}{3}u_H(x_{j-1},t) - \frac{2}{3}u_H(x_{j+1}(t)) + \frac{1}{12}u_H(x_{j+2},t)).$$

Diffusion Equation:A general finite difference scheme for the periodic diffusion problem can be rewritten as

$$u'_H(x_j,t) = \frac{D}{h^2}\sum_{k=-r}^{s}\gamma_k u_H(x_{j+k},t), j = 1,\ldots,m, \tag{2.7.15}$$

where $u_H(x_i,t) = u_H(x_{i+m},t)$ to impose the periodicity condition. We assume that $r = s$ and $\gamma_{-k} = \gamma_k$, that is the symmetry in space. Replacing, in (2.7.15), u_H by the solution of the diffusion equation and by using Taylor's expansion we obtain

$$\begin{aligned}
&\frac{\partial u}{\partial t}(x_j,t) - \frac{D}{h^2}\sum_k \gamma_k u(x_{j+k},t)\\
&= D\frac{\partial^2 u}{\partial x^2}(x_j,t) - \frac{D}{h^2}\sum_k \gamma_k\big(u(x_j,t) + kh\frac{\partial u}{\partial x}(x_j,t) + \frac{1}{2}k^2h^2\frac{\partial^2 u}{\partial x^2}(x_j,t) + \ldots\big)\\
&= -\frac{D}{h^2}\sum_k \gamma_h u(x_j,t) + D\Big(1 - \frac{1}{2}\sum_k k^2\gamma_k\Big)\frac{\partial^2 u}{\partial x^2}(x_j,t) - \frac{D}{4!}h^2\Big(\sum_k k^4\gamma_k\Big)\frac{\partial^4 u}{\partial x^4}(x_j,t) + \ldots
\end{aligned}$$

If we assume

$$\sum_k \gamma_k = 0,\ \sum_k k^2\gamma_k = 2,\ \sum_k k^4\gamma_k = 0\ \ldots\ \sum_k k^q\gamma_k = 0, \tag{2.7.16}$$

then the residual error is of order q. The order conditions (2.7.16) can be satisfied for $q \leq 2s$. For instance, for $s = 2$ we obtain the fourth central upwind discretization

$$u'_H(x_j,t) = \frac{D}{h^2}\Big(-\frac{1}{12}u_H(x_{j-2},t) + \frac{4}{3}u_H(x_{j-1},t) - \frac{5}{2}u_H(x_j,t) + \frac{4}{3}u_H(x_{j+1},t) - \frac{1}{12}u_H(x_{j+2},t)\Big).$$

2.7.3 Convergence of the Spatial Discretization

The study of the numerical methods for ODEs was based on the concepts of stability and consistency. For elliptic equations the same concepts were used to establish the convergence. The main ingredients in the study of the convergence properties of the semi-discrete approximation will be, as in the previous section for the advection equations and diffusion equations with periodic boundary conditions, the concepts of stability and consistency.

Let Λ be a sequence of positive vectors converging to zero. If Ω is a subset of $\mathbb{R}^n$ then Λ is a sequence of positive vectors of $\mathbb{R}^n$ when uniform meshes are used. Thus, for $H \in \Lambda$, the semi-discrete solution $u_H(t)$ is solution of (2.7.1). As before, the discretization of the boundary conditions are supposed included in F_H.

The spatial discretization error $e_H(t)$ is defined by

$$e_H(t) = R_H u(t) - u_H(t).$$

Let $T_H(t)$ be the spatial truncation error

$$T_H(t) = u'(t) - F_H(t, R_H u(t)),$$

which is the residual obtained substituting the solution of the PDE into the difference scheme. A bound for $\|T_H(t)\|$ is obtained by Taylor expansion provided that the solution u is smooth enough. The concept of consistency is introduced as before analyzing the behaviour of the truncation error. If

$$\|T_H(t)\| = O(H_{max}^q) \text{ for } t \in [0, T], \tag{2.7.17}$$

then the semi-discrertization is called consistent of order q. In (2.7.17)

$$H_{max} = \max\{h_i, H = (h_1, \ldots, h_n) \in \Lambda\}.$$

The spatial discretization is said convergent with order p if

$$\|e_H(t)\| = O(H_{max}^p) \text{ for } [0, T]. \tag{2.7.18}$$

The stability concept has in the context of the semi-discretizations a convenient adaptation. As we are dealing with the solution of an ordinary initial value problem, the semi-discretization (2.7.1) is said stable if its solution $u_H(t)$ is stable in the sense of ordinary differential problems but for $H_{max} \to 0$. More precisely, if $u_H(t)$ and $\tilde{u}_H(t)$ are solutions of (2.7.1) with initial conditions $u_H(0)$ and $\tilde{u}_H(0)$ such that

$$\lim_{H_{max} \to 0} \|u_H(0) - \tilde{u}_H(0)\| = 0,$$

then

$$\lim_{H_{max} \to 0} \|u_H(t) - \tilde{u}_H(t)\| = 0 \text{ for } [0, T].$$

Of course that when (2.7.1) is linear, that is

$$F_H(t, u_H(t)) = A_H u_H(t) + g_H(t), \tag{2.7.19}$$

where $g_H(t)$ represents a discretization of a source term or arises from the discretrization of the boundary conditions, a sufficient condition is

$$\|e^{tA_H}\| \leq Ke^{\omega t},\ t \in [0,T], \tag{2.7.20}$$

where K and ω are H independent. Under this sufficient condition is easy to establish an upper bound for the error $\|e_H(t)\|$ at leat for the linear case. In fact, from the definitions of $T_H(t)$ and $e_H(t)$, we have

$$e'_H(t) = A_H e_H(t) + T_H(t),$$

and thus

$$e_H(t) = e^{tA_H} e_H(0) + \int_0^t e^{(t-s)A_H} T_H(s)\, ds.$$

Consequently,

$$\|e_H(t)\| \leq \|e^{tA_H}\|\|e_H(0)\| + \int_0^t \|e^{(t-s)A_H}\|\|T_H(s)\|\, ds. \tag{2.7.21}$$

Applying the sufficient condition (2.7.20) in (2.7.21) the upper bound

$$\|e_H(t)\| \leq Ke^{\omega t}\|e_H(0)\| + \frac{K}{\omega}(e^{\omega t} - 1) \max_{s\in[0,t]} \|T_H(s)\|, \tag{2.7.22}$$

is easily established. This error estimate leads to the following convergence result:

Theorem 2.7.2 *Consider the semi-discrete system (2.7.1) with F_H given by (2.7.19). Suppose that the condition (2.7.20) holds and $\|T_H\| \leq CH^q_{max}$ for $t \in [0,T]$, $\|e_H(0)\| \leq C_0 H^q_{max}$, with C, C_0 H-independent. Then*

$$\|e_H(t)\| \leq C_0 K e^{\omega t} H^q_{max} + \frac{CK}{\omega}(e^{\omega t} - 1)H^q_{max},\ t \in [0,T], \tag{2.7.23}$$

provided that $\omega \neq 0$, and

$$\|e_H(t)\| \leq C_0 K H^q_{max} + CKtH^q_{max},\ t \in [0,T], \tag{2.7.24}$$

when $\omega = 0$.

■

Example 28 *The first order upwind scheme (2.7.4) is convergent with order 1 for*

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0$$

and it is convergent with order 2 for

$$\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x} = \frac{1}{2}ah\frac{\partial^2 \tilde{u}}{\partial x^2}.$$

The second order central scheme (2.7.9) is convergent with order 2 for

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}$$

and it is convergent with order 6 for

$$\frac{\partial \tilde{u}}{\partial t} = D\frac{\partial^2 \tilde{u}}{\partial x^2} + \frac{Dh^2}{12}\frac{\partial^4 \tilde{u}}{\partial x^4} + \frac{Dh^4}{360}\frac{\partial^6 \tilde{u}}{\partial x^6}.$$

2.7.4 Semi-Discretization in Conservative Form

Advection Equation: Let us consider the advection-diffusion equation (2.7.2) with $D = 0$ in the equivalent form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(a(x)u(x)) = 0, \tag{2.7.25}$$

with periodic conditions $u(x \pm 1, t) = u(x, t)$. The velocity $a(x)$ is assumed to be also 1-periodic and differentiable. We say that the equation (2.7.25) is in the conservative form in the sense that leads to

$$\begin{aligned} M'(t) &= \frac{d}{dt}\int_0^1 u(x,t)\,dx \\ &= \int_0^1 \frac{\partial u}{\partial t}(x,t)\,dx \\ &= -a(1)u(1,t) + a(0)u(0,t) \\ &= 0, \end{aligned}$$

that is the mass

$$M(t) = const,\ t \in [0, T]. \tag{2.7.26}$$

In $[0, 1]$ we introduce the uniform mesh $\{x_j\}$ with step size h and we define the auxiliary points $x_{j\pm1/2} = x_j \pm \frac{h}{2}$. Further we consider the cell $I_j = [x_{j-1/2}, x_{j+1/2}]$ and the cell average

$$\bar{u}(x_j, t) = \frac{1}{h}\int_{I_j} u(x,t)\,dx.$$

Then

$$h\frac{d}{dt}\bar{u}(x_j, t) = a(x_{j-\frac{1}{2}})u(x_{j-\frac{1}{2}}, t) - a(x_{j+\frac{1}{2}})u(x_{j+\frac{1}{2}}, t). \tag{2.7.27}$$

This equation tell us that the rate of change of mass over I_j is equal to the difference in-going and out-going fluxes over the cell boundaries.

It is natural to define a semi-discretization that mimics (2.7.27). We consider the semi-discrete approximation defined by

$$u'_H(x_j, t) = \frac{1}{h}\left(a(x_{j-\frac{1}{2}})u_H(x_{j-\frac{1}{2}}, t) - a(x_{j+\frac{1}{2}})u_H(x_{j+\frac{1}{2}}, t)\right), j = 1, \ldots, m, \tag{2.7.28}$$

where $u_H(x_{j\pm\frac{1}{2}}, t)$ are approximate values at the cell boundaries that should be defined in terms of neighbouring points $u_H(x_i, t)$ at the grid points. We remark that (2.7.28) mimics (2.7.26). In fact,

$$\begin{aligned} \frac{d}{dt}h\sum_{j=1}^{m} u_H(x_j, t) &= \sum_{j=1}^{m} a(x_{j-\frac{1}{2}})u_H(x_{j-\frac{1}{2}}, t) - a(x_{j+\frac{1}{2}})u_H(x_{j+\frac{1}{2}}, t) \\ &= a(x_{\frac{1}{2}})u_H(x_{\frac{1}{2}}, t) - a(x_{m+\frac{1}{2}})u_H(x_{m+\frac{1}{2}}, t) \\ &= 0, \end{aligned}$$

because by periodicity we have $a(x_{\frac{1}{2}})u_H(x_{\frac{1}{2}}, t) = a(x_{m+\frac{1}{2}})u_H(x_{m+\frac{1}{2}}, t)$.

In what follows we consider $a > 0$ and $u_H(x_{j+1/2},t) = u_H(x_j,t)$ and $u_H(x_{j-1/2},t) = u_H(x_{j-1},t)$, that is the upwing difference scheme in flux form

$$u'_H(x_j,t) = \frac{1}{h}\Big(a(x_{j-\frac{1}{2}})u_H(x_{j-1},t) - a(x_{j+\frac{1}{2}})u_H(x_j,t)\Big)\,, j = 1,\ldots,m \tag{2.7.29}$$

with $u_H(x_m,t) = u_H(x_0,t)$.

It is easy to establish the consistency of (2.7.29). In what concerns the stability, we rewrite (2.7.29) in equivalent form $u'_H(t) = A_H u_H(t)$, where

$$A_H = \frac{1}{h}\begin{bmatrix} -a(x_{3/2}) & 0 & 0 & . & 0 & a(x_{1/2}) \\ a(x_{3/2}) & -a(x_{5/2}) & 0 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & . & . & a(x_{m-1/2}) & -a(x_{1/2}) \end{bmatrix}.$$

As $\mu_1[A_H] = 0$ and

$$\mu_\infty[A_H] = \max_i \frac{1}{h}\big(-a(x_{i+1/2}) + a(x_{i-1/2})\big) \le \omega,$$

where ω is an upper bound to a', we conclude that

$$\|e^{tA_H}\|_1 \le 1,\ \|e^{tA_H}\|_\infty \le e^{\omega t},\ t \ge 0.$$

Otherwise, as we also have the Hölder inequality for matrices

$$\|e^{tA_H}\|_2 \le \sqrt{\|e^{tA_H}\|_1 \|e^{tA_H}\|_\infty},$$

we deduce

$$\|e^{tA_H}\|_2 \le e^{\frac{\omega}{2}t}, t \ge 0.$$

Taking into consideration the established stability inequalities, we conclude the stability of the semi-discrete scheme (2.7.29) with respect to the norms $\|.\|_p$ for $p = 1, 2, \infty$. Hence the scheme is convergent with respect to the norm $\|.\|_p$ for $p = 1, 2, \infty$.

The convergence order can be improved if we replace (2.7.29) by the new upwind scheme

$$\begin{aligned} u'_H(x_j,t) = \frac{1}{h}\Big(&a(x_{j-\frac{1}{2}})\frac{u_H(x_{j-1},t) + u_H(x_j,t)}{2} \\ &-a(x_{j+\frac{1}{2}})\frac{u_H(x_j,t) + u_H(x_{j+1},t)}{2}\Big)\,, j = 1,\ldots,m, \end{aligned} \tag{2.7.30}$$

with $u_H(x_m,t) = u_H(x_0,t)$ and $u_H(x_{m+1},t) = u_H(x_1,t)$.
Writing the semi-discrete scheme (2.7.30) as $u'_H(t) = A_H u_H(t)$ it can be shown that

$$(A_H v, v)_2 \le \frac{1}{2}\omega\|v\|_2^2,\ v \in \mathbb{R}^m.$$

Consequently,

$$\|e^{tA_H}\|_2 \le e^{\frac{\omega}{2}t},\ t \ge 0$$

holds establishing the stability and the second order convergence with respect to the L^2-norm on finite intervals $[0,T]$.

Diffusion Equation: Let us consider now the diffusion equation with a variable diffusion coefficient, a source term and Dirichlet boundary conditions

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial}{\partial x}\Big(D(x)\dfrac{\partial u}{\partial x}\Big) + s(x,t),\ t>0, x\in(0,1), \\ u(0,t)=b_0(t),\ u(1,t)=b_1(t), t>0, \\ u(x,0)=u_0(x),\ x\in(0,1), \end{cases} \tag{2.7.31}$$

where $D(x) \geq D_0 > 0$. Discretizing the second order derivative using the conservative central scheme we obtain

$$u'_H(x_j,t) = \frac{1}{h^2}\Big(D(x_{j+1/2})(u_H(x_{j+1},t)-u_H(x_j,t)) \\ -D(x_{j-1/2})(u_H(x_j,t)-u_H(x_{j-1},t))\Big) + s(x_j,t),$$

for $j = 1,\ldots,m-1$, where

$$u_H(x_0,t) = b_0(t), u_H(x_m,t) = b_1(t), t>0,$$

$$u_H(x_i,0) = u_0(x_i), i = 1,\ldots,m-1.$$

This finite difference scheme is equivalent to the ODE

$$u'_H(t) = A_H u_H(t) + g_H(t),$$

where

$$A_H = \frac{1}{h^2}\begin{bmatrix} a_1 & c_1 & 0 & . & . & 0 & 0 \\ c_1 & a_2 & c_2 & . & . & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & c_{m-2} & a_{m-1} \end{bmatrix},$$

$$g_H(t) = \begin{bmatrix} s_1 \\ s_2 \\ . \\ s_{m-1} \end{bmatrix} + \frac{1}{h^2}\begin{bmatrix} b_0(t)c_0 \\ 0 \\ . \\ b_1(t)c_{m-1} \end{bmatrix},$$

and

$$a_j = -\frac{1}{2}(D(x_{j+1/2}) + D(x_{j-1/2})),\ c_j = D(x_{j+1/2}),\ s_j = s(x_j,t).$$

Assuming smoothness of $D(x)$ and $s(x)$, it is easy to prove second order consistency. Furthermore, as

$$\mu_1[A_H] \leq 0,\ \mu_\infty[A_H] \leq 0,$$

we have

$$\|e^{tA_H}\|_p \leq 1, t \geq 0, p = 1,\infty,$$

which implies

$$\|e^{tA_H}\|_2 \leq 1, t \geq 0.$$

The previous estimate can be refined if we use energy method. For the discrete L^2 norm induced by the inner product $(.,.)_H$ defined before but here for grid functions null on the boundary points, we have

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|u_H(t)\|^2_{L^2} &= (A_H u_H, u_H)_H \\
&= \frac{1}{h}\sum_{j=1}^{m-1} D(x_{j+1/2})u_H(x_j,t)(u_H(x_{j+1},t)-u_H(x_j,t)) \\
&\quad -\frac{1}{h}\sum_{j=0}^{m-2} D(x_{j+1/2})u_H(x_{j+1},t)(u_H(x_j,t)-u_H(x_{j-1},t)) \\
&= -h\sum_{j=1}^{m} D(x_{j-1/2})\Big(D_{-x}u_H(x_j,t)\Big)^2 \\
&\leq -D_0 h\sum_{j=1}^{m}\Big(D_{-x}u_H(x_j,t)\Big)^2.
\end{aligned}$$

As

$$\|u_H(t)\|^2_{L^2} \leq \sum_{j=1}^{m}\Big(D_{-x}u_H(x_j,t)\Big)^2,$$

we conclude

$$\|u_H(t)\|_{L^2} \leq e^{-D_0 t}\|u_0\|_{L^2}.$$

Some Remarks: For advection-diffusion equations with variable coefficients or advection-diffusion-reaction equations with a nonlinear reaction term, the consistency it is easily verified. Nevertheless, in what concerns the stability, even for the method (2.7.28) with

$$u_H(x_{j+\frac{1}{2}},t) = \begin{cases} \frac{1}{6}\big(-u_H(x_{j-1},t)+5u_H(x_j,t)+2u_H(x_{j+1},t)\big), \text{ if } a(x_{j+1/2})>0 \\ \frac{1}{6}\big(2u_H(x_j,t)+5u_H(x_{j+1},t)-u_H(x_{j+2},t)\big), \text{ if } a(x_{j+1/2})<0, \end{cases}$$

for advection equation, simple estimates are not available.

Singularly perturbed problems - problems with the diffusion coefficient very small when compared with the advection coefficient - are usually characterized by a boundary layer. If a uniform mesh is used to solve such problems, then a very huge number of grid points should be considered. Such approach is computationally inefficient. A remedy that can avoid the inefficiency of the uniform meshes is to use nonuniform meshes well adapted to the layer . Such remedy increases the difficulties on the stability analysis (see [31]).

2.7.5 Refined Global Estimates

In the previous section we studied the spatial discretizations of a diffusion equation with Dirichlet boundary conditions. The presence of boundary can complicate the numerical treatment. As in the stationary case, if we consider Neumann boundary conditions, the discretizations is made with a different discretization. As we saw this has an adverse effect on the global accuracy. However this effect is often not as large as expected in the sense that global order of convergence p can be greater than the order of consistency q.

Let us consider the linear semi-discrete system

$$u'_H(t) = A_H u_H(t) + g_H(t),$$

where the discretization of the boundary conditions and (or) of the source term are included in the semi-discrete source term g_H. Let us suppose that the spatial discretization is stable, for instance $\|e^{tA_H}\| \leq Ke^{\omega t}$ holds, and the the spatial truncation error $T_H(t)$ with order q admits the representation

$$T_H(t) = A_H T_H^{(1)}(t) + T_H^{(2)}(t), \tag{2.7.32}$$

where

$$\|T_H^{(1)}(t)\| \leq Ch^r, \|T_H^{(1)'}(t)\| \leq Ch^r, \|T_H^{(2)}(t)\| \leq Ch^r. \tag{2.7.33}$$

Considering in the error equation

$$e'_H(t) = A_H e_H(t) + T_H(t),$$

the representation (2.7.32) we get

$$e'_H(t) = A_H\Big(e_H(t) + T_H^{(1)}(t)\Big) + T_H^{(2)}(t),$$

thus

$$e'_H(t) + T_H^{(1)'}(t) = A_H(e_H(t) + T_H^{(1)}(t)) + T_H^{(2)}(t) + T_H^{(1)'}(t). \tag{2.7.34}$$

As in the proof of Theorem 2.7.2, (2.7.34) leads to

$$\begin{aligned} \|e_H(t) + T_H^{(1)}(t)\| &\leq Ke^{\omega t}\|e_H(0) + T_H^{(1)}(0)\| \\ &+\frac{K}{\omega}\big(e^{\omega t} - 1\big) \max_{s\in[0,t]} \|T_H^{(2)}(s) + T_H^{(1)'}(s)\|, t \geq 0. \end{aligned} \tag{2.7.35}$$

Furthermore, if we consider in (2.7.35) the assumption (2.7.33) we establish

$$\|e_H(t)\| \leq Ch^r + Ke^{\omega t}\|e_H(0)\| + 2Ch^r\frac{K}{\omega}\big(e^{\omega t} - 1\big), t \geq 0. \tag{2.7.36}$$

We proved the next refined result:

Theorem 2.7.3 *Consider the linear system $u'_H(t) = A_H u_H(t)+g_H(t)$ and assume that $\|e^{tA_H}\| \leq Ke^{\omega t}, t \geq 0$. Suppose that the truncation error $T_H(t)$ satisfies (2.7.32) and (2.7.33) and suppose that $\|e_H(0)\| \leq Ch^r$. Then*

$$\|e_H(t)\| \leq Ch^r\Big(1 + Ke^{\omega t} + 2\frac{K}{\omega}\big(e^{\omega t} - 1\big)\Big), t \geq 0. \tag{2.7.37}$$

■

We apply the last result to the diffusion model

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\dfrac{\partial^2 u}{\partial x^2}\ , x \in (0,1), t > 0, \\ u(0,t) = b_0(t),\ \dfrac{\partial u}{\partial x}(1,t) = 0, t > 0, \\ u(x,0) = u_0(x),\ x \in (0,1). \end{cases} \tag{2.7.38}$$

We consider in $[0,1]$ a uniform mesh $x_j = jh, j = 0,\ldots,m, x_0 = 0, x_m = 1$ and the auxiliary point $x_{m+1} = x_m + h$ for the discretization of the Neumann boundary condition. Discretizing the second order derivative using the conservative central scheme we obtain

$$u'_H(x_j,t) = \frac{D}{h^2}\Big(u_H(x_{j+1},t) - 2u_H(x_j,t) + u_H(x_{j-1},t)\Big),$$

for $j = 1,\ldots,m$, where

$$u_H(x_0,t) = b_0(t), u_H(x_{m+1},t) = \theta u_H(x_m,t) + (1-\theta)u_H(x_{m-1},t),\ \theta \in \{0,1\}.$$

When $\theta = 0$, we discretize $\dfrac{\partial u}{\partial x}(1,t)$ with the finite difference operator D_c and for $\theta = 1$ the discretization of such term is made by the finite difference operator D_x.

The above finite difference scheme is equivalent to the ODE system

$$u'_H(t) = A_H u_H(t) + g_H(t),$$

where

$$A_H = \frac{D}{h^2}\begin{bmatrix} -2 & 1 & 0 & . & . & 0 & 0 \\ 1 & -2 & 1 & . & . & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & 2-\theta & \theta-2 \end{bmatrix}, \qquad g_H(t) = \frac{1}{h^2}\begin{bmatrix} b_0(t) \\ 0 \\ . \\ 0 \end{bmatrix}.$$

As

$$\mu_1[A_H] \le 0,\ \mu_\infty[A_H] \le 0,$$

we have

$$\|e^{tA_H}\|_p \le 1, t \ge 0, p = 1,\infty,$$

which implies

$$\|e^{tA_H}\|_2 \le 1, t \ge 0.$$

For $\theta = 0$ the truncation error has order 2 while it is inconsistent for $\theta = 1$. Nevertheless, we show in what follows that in the last case the scheme is convergent. In order to do that we compute the decomposition of the truncation error

$$T_H(t) = A_H T_H^{(1)}(t) + T_H^{(2)}(t).$$

Let $T_H^{(2)}$ be such that $T_H^{(2)}(x_i,t) = O(h^2), i = 1,\ldots,m-1, T_H^{(2)}(x_m,t) = O(h)$, and let $T_H^{(1)}$ be the solution of the difference equation

$$A_H T_H^{(1)}(t) = \xi_H(t), \tag{2.7.39}$$

where $\xi_H(x_i,t) = 0, i = 1,\ldots,m-1$, and $\xi_H(x_m,t) = -\dfrac{\theta}{2}\dfrac{\partial^2 u}{\partial x^2}(x_m,t)$. If we fix $T_H^{(1)}(x_m,t) = 0$, then $T_H^{(1)}(x_j,t)$ is given by

$$T_H^{(1)}(x_j,t) = j\frac{-h^2\theta}{2(2-\theta)}\frac{\partial^2 u}{\partial x^2}(x_j,t), j = 1,\ldots,m-1.$$

It follows that $\|T_H^{(1)}(t)\| = O(h)$ and by Theorem 2.7.3, we conclude the convergence with respect to the norms: $\|.\|_p, p = 1,2,\infty$.

2.7.6 Fully Discrete FDM: MOL Approach, Direct Discretizations

MOL Approach: So far we have studied the spatial discretizations of some time dependent PDES, that is the ODEs obtained discretizing the spatial derivatives of the PDEs defined by finite difference operators. In the proof of the stability and convergence results the stability analysis for ordinary differential problems had a central role. The aim of this section is to study some fully discrete schemes which can be obtained by the MOL approach, that is, integrating numerically the semi-discrete problem with a numerical method for ODEs. It should be emphasis that the method of lines is not a method in the numerical sense but an approach to construct numerical methods for time dependent problems.

Let us consider the ODE (2.7.1) numerically integrated in time, for example, with the θ-method

$$u_H^{n+1} = u_H^n + \Delta t\Big((1-\theta)F_H(t_n, u_H^n) + \theta F_H(t_{n+1}, u_H^{n+1})\Big), n = 0, \ldots,$$

where $u_H^0 \simeq u_H(0)$. Then $u_H^n(x)$ defines an approximation for $u(x, t_n)$ for x in the spatial grid $\overline{\Omega}_H$.

The error of the numerical approximation u_H^n is given by $e_H^n(x) = u(x, t_n) - u_H^n(x), x \in \overline{\Omega}_H$. As

$$\|e_H^n\| \leq \|R_H u(t_n) - u_H(t_n)\| + \|u_H(t_n) - u_H^n\|,$$

an estimate for $\|e_H^n\|$ is obtained estimating the two errors

$$\|R_H u(t_n) - u_H(t_n)\|, \|u_H(t_n) - u_H^n\|.$$

The first one was studied in the previous sections and the second one was studied in the first chapter. For instance, if the semi-discretization is of order p_1 and the time integration method is of order p_2, then

$$\|e_H^n\| \leq C_1 h^{p_1} + C_2 \Delta t^{p_2}. \tag{2.7.40}$$

As the ordinary differential problem is in fact a family of ordinary differential problems depending on the space step size, in the convergence estimate

$$\|u_H(t_n) - u_H^n\| \leq C_2 \Delta t^{p_2}$$

for the time integration error, we should have C_2 and p_2 independent on the space step size, that is, the previous convergence estimate should be uniform with respect to the space step size. Stability and consistency of the ODE method should thus be verified for all time step size. This sentence implies an eventually restriction on the space and time step sizes.

Direct Discretization: Even a scheme can be seen as a combination between the spatial discretizatian followed by the time integration, it can be advantageous to consider space and time errors simultaneously.

We consider in what follows FDMs for time dependent problems which admit the representation

$$B_0 u_H^{n+1} = B_1 u_H^n + G(t_n, t_{n+1}), n = 0, \ldots, \tag{2.7.41}$$

where the matrices $B_0, B_1 \in \mathbb{R}^{m^2}$, and $G(t_n, t_{m+1}) \in \mathbb{R}^m$ depends on the space and time step sizes. Their dimensions depend on the number of points on the spatial domain $\overline{\Omega}$. Of course that if the scheme is explicit, then $B_0 = I$.

Example 29 *If we apply the θ-method to the semi-discrete problem $u'_H(t) = A_H u_H(t) + g_H(t)$ we obtain the two-level scheme*

$$u_H^{n+1} = u_H^n + \Delta t\Big((1-\theta)(A_H u_H^n + g(t_n)) + \theta A_H u_H^{n+1}\Big)$$

$$+\Delta t\Big((1-\theta)g_H(t_n) + \theta g_H(t_{n+1})\Big),$$

which admits the representation (2.7.41) with

$$B_0 = I - \Delta t\theta A_H,\ B_1 = I + \Delta t(1-\theta)A_H$$

and

$$G(t_n, t_{n+1}) = \Delta t((1-\theta)g_H(t_n) + \theta g_H(t_{n+1})).$$

Example 30 Courant-Isaacson-Rees Scheme *Discretizing the advection equation*

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0, \\ \\ u(x \pm 1, t) = u(x, t), \end{cases}$$

with $a > 0$, by the upwind scheme and the explicit Euler's method we obtain the fully discrete scheme

$$\begin{cases} u_H^{n+1}(x_j) = u_H^n(x_j) + \dfrac{a\Delta t}{h}(u_H^n(x_{j-1}) - u_H^n(x_j)),\ j = 1, \ldots, m, \\ \\ u_H^n(x_0) = u_H^n(x_m), n = 0, \ldots \end{cases} \tag{2.7.42}$$

This scheme, known as Courant-Isaacson-Rees scheme, can be rewritten in the matrix form (2.7.41) with

$$B_0 = I, B_1 = I + \Delta t A_H$$

where A_H is the upwind matrix introduced before.

The convergence analysis of the Courant-Isaacson-Rees scheme can be performed by using the MOL approach. In this case we have (2.7.40) with $p_1 = p_2 = 1$ but under appropriate restrictions for the space and time step sizes. Nevertheless, the convergence analysis can be considered directly. Replacing, in (2.7.42), $u_H^n(x_j)$ by the true solution $u(x_j, t_n)$, we obtain

$$u(x_j, t_{n+1}) = u(x_j, t_n) + \frac{a\Delta t}{h}(u(x_{j-1}, t_n) - u(x_j, t_n)) + \Delta t T_H^n(x_j),$$

where the truncation error is given by

$$T_H^n(x_j) = -\frac{1}{2}ah(1 - \frac{a\Delta t}{h})\frac{\partial^2 u}{\partial x^2}(x_j, t_n) + O(h^2) + O(\Delta t^2).$$

For the error e_H^n *we have*

$$e_H^{n+1} = B_1 e_H^n + \Delta t T_H^n.$$

If we assume that

$$\frac{a\Delta t}{h} \leq 1, \tag{2.7.43}$$

then $\|B_1\|_p \leq 1$ *for* $p = 1, 2, \infty$. *Under this assumption we obtain*

$$\|e_H^{n+1}\| \leq \|e_H^n\| + \Delta t \|T_H^n\|,$$

which implies

$$\|e_H^n\| \leq t_n \frac{1}{2} ah(1 - \frac{a\Delta t}{h}) \max_{x,t} \|\frac{\partial^2 u}{\partial x^2}\| + O(h^2) + O(\Delta t^2),$$

provided that $e_H^0 = 0$.

The stability condition (2.7.43) is usually called Courant-Friedrichs-Levy condition (CFL).

An explicit scheme for the advection equation can be rewritten as

$$u_H^{n+1}(x_j) = \sum_{k=-r}^{s} \gamma_k u_H^n(x_{j+k})$$

with the coefficients γ_k *depending on* $\nu = \dfrac{a\Delta t}{h}$. *Then* $u_H^n(x_j)$ *depends on the initial condition on the grid points* $x_i, i = j - nr, \ldots, j + ns$. *If we consider* $h, \Delta t \to 0$ *with ratio constant* ν *then* $x_{j-nr} \to x - (r/\nu)at$ *and* $x_{j+ns} \to x + (s/\nu)at$, *and thus the numerical approximations for* $u(x,t)$ *are determined by the initial data in the interval*

$$[x - \frac{r}{\nu}at, x + \frac{s}{\nu}at].$$

This interval is called domain of dependence.

Example 31 Lax-Wendrof Scheme *As we mention before, there are fully discrete numerical methods for time dependent problems which can not be obtained from the MOL approach. An example is the so called Lax-Wendrof scheme*

$$\begin{aligned} u_H^{n+1}(x_j) &= u_H^n(x_j) + \frac{a\Delta t}{2h}(u_H^n(x_{j-1}) - u_H^n(x_{j+1})) \\ &+ \frac{1}{2}\Big(\frac{a\Delta t}{h}\Big)^2 \big(u_H^n(x_{j-1}) - 2u_H^n(x_j) + u_H^n(x_{j+1})\big). \end{aligned} \tag{2.7.44}$$

This scheme can be obtained replacing, in the Taylor expansion

$$u(x_j, t_{n+1}) = u(x_j, t_n) + \Delta t \frac{\partial u}{\partial t}(x_j, t_n) + \frac{1}{2}\Delta t^2 \frac{\partial^2 u}{\partial t^2}(x_j, t_n) + O(\Delta t^3),$$

$\dfrac{\partial u}{\partial t}$ *by* $-a\dfrac{\partial u}{\partial x}$ *and* $\dfrac{\partial^2 u}{\partial t^2}$ *by* $a^2\dfrac{\partial^2 u}{\partial x^2}$ *and considering the central formulas in the last derivatives.*

The Lax-Wendrof scheme has the truncation error

$$T_H^n(x_i) = \frac{1}{6}ah^2\big(1 - \big(\frac{a\Delta t}{h}\big)^2\big)\frac{\partial^3 u}{\partial x^3}(x_j, t_n) + O(\Delta t^3).$$

2.7.7 Stability, Consistency and Convergence

The convergence analysis of the Courant-Isaacson-Rees scheme in Example 30 uses the concepts of stability and convergence. In what follows we formalize the previous analysis for the two-level scheme (2.7.41).

The truncation error for (2.7.41) at time level t_n, T_H^n is defined by

$$B_0 R_H u(t_{n+1}) = B_1 R_H u(t_n) + G(t_n, t_{n+1}) + \Delta t T_H^n. \tag{2.7.45}$$

For the error $e_H^n = R_H u(t_n) - u_H^n$ is solution of the following problem

$$B_0 e_H^{n+1} = B_1 e_H^n + \Delta t T_H^n, \tag{2.7.46}$$

which can be rewritten in the equivalent form

$$e_H^{n+1} = B_0^{-1} B_1 e_H^n + \delta_H^n, \tag{2.7.47}$$

$$B = B_0^{-1} B_1,\ \delta_H^n = \Delta t B_0^{-1} T_H^n.$$

Let Λ be a sequence of space step sizes. If $\Omega \subset \mathbb{R}^n$, then Λ is a sequence of n vectors such that $H_{max} = \max_i h_i \to 0$, with $H = (h_1, \ldots, h_n)$.

The concept of stability, that is the sensitivity of the solution defined by (2.7.41) to perturbations of the initial condition, has here a natural formalization: the two-time level scheme (2.7.41) is stable if for any initial conditions $u_H^0, \tilde{u}_H^0$ such that

$$\lim_{H_{max}, \Delta t \to 0} \|u_H^0 - \tilde{u}_H^0\| = 0,$$

the correspondent numerical solutions $u_H^n, \tilde{u}_H^n$ satisfy

$$\lim_{H_{max}, \Delta t \to 0} \|u_H^n - \tilde{u}_H^n\| = 0, \forall n : n\Delta t \leq T,$$

where $[0, T]$ denotes the time interval.

A sufficient condition for stability is now immediate.

Theorem 2.7.4 *If*

$$\|B^n\| \leq K, n\Delta t \leq T, \tag{2.7.48}$$

where K is independent of Δt and H, then the two-level scheme (2.7.41) is stable.

■

The condition (2.7.48) in general holds when some restriction is imposed on H and on Δt. Usually the concept of stability is replaced by its sufficient condition (2.7.48).

Using the sufficient condition (2.7.41), it is easy to prove the convergence of (2.7.41), that is

$$\lim_{H_{max}, \Delta t \to 0} \|e_H^n\| = 0, \forall n : n\Delta t \leq T, \tag{2.7.49}$$

provided that the finite difference scheme is consistent. In fact, from the error equation (2.7.47) we have

$$\|e_H^n\| \leq \|B^n\| \|R_H u_0 - u_H^0\| + \|B^n\| \|B_0^{-1}\| t_n \max_j \|T_H^j\|.$$

Consequently, under the consistency of (2.7.41), we conclude that $\|e_H^n\| \to 0$ as $H_{max}, \Delta t \to 0$.

Implicitly we supposed in the convergence proof that $\|B_0^{-1}\| \leq C$, where C is H and Δt independent. This is a natural condition which means that the finite difference scheme (2.7.41) is well defined.

We have shown that the two-level scheme is convergent provided that is stable and consistent. The Lax Theorem establishes that for a consistent finite difference scheme stability is also a necessary condition. This result can be seen in [30].

2.7.8 Stability for MOL

The stability analysis of the finite difference scheme defined by the MOL approach is based on the stability analysis of the numerical methods for ODEs. Let us consider any one step method for the liner semi-discrete system

$$u'_H(t) = A_H u_H(t) + g_H(t).$$

The stability of the fully discrete solution u_H^n is studied analyzing the behaviour of the discrete scheme

$$w_H^{n+1} = R(\Delta t A_H) w_H^n, \tag{2.7.50}$$

where R is the stability function. The stability restrictions on Δt in terms of the space step size is obtained from the stability region and from the properties of A_H.

We consider now the θ-method studied in chapter 1 whose stability function is given by

$$R(z) = \frac{1 + (1-\theta)z}{1 - \theta z},$$

for $\theta \in [0, 1]$. In this case we established results for the behaviour of $\|R(\Delta t A_H)^n\|$.

1. **Theorem 1.3.2**: Let us suppose that $A_H = MDM^{-1}$ where $cond(M) \leq k$ and $D = diag(\lambda_j)$. If $\Delta t \lambda_j \in \mathcal{S}$, for all j, then

   $$\|R(\Delta t A_H)^n\| \leq k, \ \forall n.$$

   ■

   In this result $\mathcal{S}$ denotes the stability region and we should consider n such that $n\Delta t \leq T$.

2. **Corollary 1**: Suppose that A_H is a normal matrix. If $\Delta t \lambda_j \in \mathcal{S}$, for all j, then

   $$\|R(\Delta t A_H)\|_2 \leq 1.$$

   ■

3. **Theorem 1.3.3**: Suppose that the vectorial norm is induced by an inner product $< ., . >$. If

   $$Re < A_H v, v > \leq \omega \|v\|^2, \ \forall v \in \mathbb{C}^m,$$

   then

   $$\|R(\Delta t A_H)\| \leq \max\left(|R(\Delta t \omega)|, |R(\infty)|\right),$$

   provided that $1 - \omega\theta\Delta t > 0$.

   ■

4. **Corollary 2**: If $\mu[A_H] \leq 0$ and $\theta \geq \frac{1}{2}$, then

$$\|R(\Delta t A)\| \leq 1.$$

■

This corollary establishes the unconditional stability of the θ-method for $\theta \in [\frac{1}{2}, 1]$ when $\mu[A_H] \leq 0$.

Let us suppose that A_H is a normal matrix. In this case

$$\|R(\Delta t A_H)^n\|_2 = \max_{j=1,\ldots,m} |R(\Delta t \lambda_j)^n|,$$

where $\lambda_j, j = 1, \ldots, m$, are the eigenvalues of A_H. If we consider the step sizes such that $\Delta t \lambda_j \in \mathcal{S}$ then

$$\|R(\Delta t A_H)^n\|_2 \leq 1.$$

A sufficient condition for stability is

$$|R(\Delta t \lambda_j)| \leq 1 + k' \Delta t$$

where k' is independent on the step sizes. In this case

$$\|R(\Delta t A_H)^n\|_2 \leq e^{k' n \Delta t} \leq e^{k' T}.$$

We remark that, for $\theta \in [0, \frac{1}{2})$, the stability region of the θ-method is given by

$$\mathcal{S} = \{z \in \mathbb{C} : |z + \alpha| \leq \alpha\},$$

with

$$\alpha = \frac{1}{1 - 2\theta}.$$

Example 32 *Let us consider the Courant-Isaacson-Rees scheme considered in Example 30 for the advection equation with $a > 0$. As $\mu[A_H] \leq 0$, the θ-method is unconditionally stable for $\theta \in [\frac{1}{2}, 1]$. We conclude stability of the Courant-Isaacson-Rees scheme without any restriction on the CFL number when $\theta \in [\frac{1}{2}, 1]$. As the eigenvalues of A_H are given by*

$$\lambda_k = \frac{a}{h}\big(cos(2\pi k h) - 1\big) - i\frac{a}{h} sin(2\pi k h), k = 1, \ldots, m,$$

for $\theta \in [0, \frac{1}{2})$, we have stability provided the CFL number satisfies

$$\frac{a\Delta t}{h} \in (0, \frac{1}{1 - 2\theta}]. \tag{2.7.51}$$

Example 33 *The central scheme for the advection equation is unconditionally stable for* $\theta \in [\frac{1}{2}, 1]$. *As the eigenvalues of* A_H *are in the imaginary axis*

$$\lambda_k = -i\frac{a}{h} sen(2\pi kh), k = 1, \ldots, m,$$

and as $\mathcal{S}$ *has no intersection with this axis, the fully discrete scheme obtained integrating numerically the central scheme with the* θ*-method is unstable for* $\theta \in [0, \frac{1}{2})$.

Example 34 *The fully discrete scheme for the diffusion equation defined by (2.7.9) and by the* θ*-method, is unconditionally stable for* $\theta \in [\frac{1}{2}, 1]$. *As the eigenvalues of* A_H

$$\lambda_k = \frac{2D}{h^2}(cos(2\pi kh) - 1) = -\frac{4D}{h^2} sen(\pi kh)^2 \in [-\frac{4D}{h^2}, 0], k = 1, \ldots, m,$$

then the fully discrete scheme is stable for $\theta \in [0, \frac{1}{2})$, *provided that*

$$\frac{D\Delta t}{h^2} \in (0, \frac{1}{2 - 4\theta}].$$

Example 35 *The explicit Euler's method is unstable for the central advection discretization. This behaviour is avoided if some diffusion is added, that is if we consider*

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = D\frac{\partial^2 u}{\partial x^2}$$

with the boundary conditions $u(x \pm 1, t = u(x, t)$. *In fact, the fully discrete scheme*

$$u_H^{n+1}(x_j) = (\mu + \frac{1}{2}\nu)u_H^n(x_{j-1}) + (1 - 2\mu)u_H^n(x_j) + (\mu - \frac{1}{2}\nu)u_H^n(x_{j+1}), j = 1, \ldots, m, \quad (2.7.52)$$

with $u_H^n(x_0) = u_H^n(x_m), u_H^n(x_{m+1}) = u_H^n(x_1)$, $x_j = jh, h = \frac{1}{m}$ *and* $\mu = \frac{D\Delta t}{h^2}$, $\nu = \frac{a\Delta t}{h}$, *is such that the eigenvalues* λ_j *of the matrix* A_H *admit the representation*

$$\Delta t\lambda_j = 2\mu(cos(2\pi jh) - 1) - i\nu sen(2\pi jh), j = 1, \ldots, m.$$

Then

$$\nu^2 \leq 2\mu \leq 1$$

is a necessary and sufficient condition for stability.

Let us suppose now that A_H is a non-normal diagonalizable matrix

$$A = MDM^{-1}, \mathcal{D} = diag(\lambda_j).$$

Then

$$\|R(\Delta t A_H)^n\| \leq cond(M) \max_j |R(\Delta t\lambda_j)^n|.$$

If the condition number does not grow as $h \to 0$ and takes a moderate size, then the eigenvalue criterion followed for the normal matrices can still be applied in the stability study. Nevertheless, for non-normal matrices the eigenvalue criterion can leads to wrong conclusions.

Example 36 *The Courant-Isaacson-Rees scheme for $a = 1$ and with $u(0,t) = 0$ leads to the matrix*

$$A_H = \frac{1}{h}\begin{bmatrix} -1 & 0 & . & . & 0 \\ 1 & -1 & . & . & 0 \\ . & . & . & . & . \\ 0 & . & . & 1 & -1 \end{bmatrix},$$

which as only the eigenvalue $\lambda = -\frac{1}{h}$. Then $\Delta t\lambda \in \mathcal{S}$ if and only if $\frac{\Delta t}{h} \leq 2$. This conclusion contrast with the previous one for the advection equation with periodic boundary conditions $\frac{\Delta t}{h} \leq 1$ indicating that for $1 < \frac{\Delta t}{h} \leq 2$ we get wrong results.

2.7.9 Von Neumann Stability Analysis

The von Neumann analysis is based on Fourier discrete modes and it is applied only to problems without boundary conditions and constant coefficients. In practice, however, this type of analysis is also used when these conditions are not verified leading to a reliable step criteria. The von Neumann analysis can be used in multiple dimensions, for systems of PDEs and with all kind of time integrations formulas.

We exemplify the von Neumann analysis on the finite difference scheme

$$u_H^{n+1}(x_j) = u_H^n(x_j) + \frac{D\Delta t}{h^2}\big(u_H^n(x_{j-1}) - 2u_H^n(x_j) + u_H^n(x_{j+1})\big),\ j = 1, \ldots, m$$

with $u_H^m(x_0) = u_H^n(x_m), u_H^m(x_1) = u_H^n(x_{m+1})$.

Considering the space $L^2(0,1)$ and the Fourier modes as introduced before, we have for each time level, the numerical approximation given by

$$u_H^n = \sum_{k=1}^{m} \alpha_k(t_n) R_H \phi_k$$

where ϕ_k denotes the Fourier mode. Assuming that $\alpha_k(t_n) = \alpha_k r_k^n$ we have

$$r_k^{n+1} e^{2\pi i k x_j} = r_k^n e^{2\pi i k x_j} + \frac{D\Delta t}{h^2}\big(r_k^n e^{2\pi i k x_{j-1}} - 2r_k^n e^{2\pi i k x_j} + r_k^n e^{2\pi i k x_{j-1}}\big).$$

Consequently

$$r_k = 1 + \frac{D\Delta t}{h^2}\big(e^{-2\pi i k h} - 2 + e^{2\pi i k h}\big) = 1 - \frac{4Dr}{h^2} sen^2(\pi h k). \tag{2.7.53}$$

Supposing that $u_H^0 = \sum_{k=1}^{m} \alpha_k R_H \phi_k$ we get $u_H^n = \sum_{k=1}^{m} \alpha_k r_k^n R_H \phi_k$ and by the Parseval identity

$$\|u_H^n\|_2^2 = \sum_{k=1}^{m} |\alpha_k|^2 |r_k|^{2n} \leq \sum_{k=1}^{m} |\alpha_k|^2 = \|u_H^0\|_2^2,$$

provided that

$$|r_k| \leq 1, \tag{2.7.54}$$

that is the amplification factor is less or equal to one. Applying this condition with r defined by (2.7.53) we obtain

$$\frac{D\Delta t}{h^2} \leq \frac{1}{2}.$$

We consider now the homogeneous version of the two-level scheme (2.7.41). Considering $u_H^n = \sum_{k=1}^{m} \alpha_k r_k^n R_H \phi_k$ with $u_H^0 = \sum_{k=1}^{m} \alpha_k R_H \phi_k$, the amplification factor r_k is the eigenvalue of B associated with the eigenvector $R_H \phi_k$. Then with respect to the discrete L^2-norm we have

$$\|B\|_2 = \max_{k=1,\ldots,m} |r_k|.$$

The stability condition $\|B^n\|_2 \leq k$ for n such that $n\Delta t \leq T$, holds provided that the step sizes Δt and h are such that

$$|r_k| \leq 1 + O(\Delta t). \tag{2.7.55}$$

This conditions is called von Neumann condition and it is often replaced by the strict von Neumann condition (2.7.54). However the strict condition is some time restrictive. For instance if we consider the diffusion equation with a source

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + cu,$$

with $c > 0, D > 0$, the source term leads to an exponential growth of the solution. The numerical solution should mimic this growth and hence we cannot require the amplification factor to be bounded by one in modulus.

3-Computational Projects

1. Consider IVP defined by the reaction of Robertson

$$\begin{cases} u_1'(t) = -0.04u_1 + 10^4 u_2 u_3 \\ u_2' = 0.04u_1 - 10^4 u_2 u_3 - 3 \times 10^7 u_2^2 \\ u_3' = 3 \times 10^7 u_2^2, \end{cases} \tag{1.0.1}$$

for $t \in (0, 40]$, with the initial condition

$$\begin{cases} u_1(0) = 1 \\ u_2(0) = 0 \\ u_3(0) = 0\,. \end{cases} \tag{1.0.2}$$

Integrate the IVP (1.0.1), (1.0.2) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method,

(c) The Gauss method defined by the following Butcher table

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2}. \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

2. Consider IVP defined by the reaction of Robertson

$$\begin{cases} u_1'(t) = -0.04u_1 + 10^4 u_2 u_3 \\ u_2' = 0.04u_1 - 10^4 u_2 u_3 - 3 \times 10^7 u_2^2 \\ u_3' = 3 \times 10^7 u_2^2, \end{cases} \tag{1.0.3}$$

for $t \in (0, 40]$, with the initial condition

$$\begin{cases} u_1(0) = 1 \\ u_2(0) = 0 \\ u_3(0) = 0\,. \end{cases} \tag{1.0.4}$$

Integrate the IVP (1.0.3), (1.0.4) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method,

(c) The Radau method defined by the following Butcher table

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4}. \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

3. Consider IVP defined by the reaction of Robertson

$$\begin{cases} u_1'(t) = -0.04u_1 + 10^4 u_2 u_3 \\ u_2' = 0.04u_1 - 10^4 u_2 u_3 - 3 \times 10^7 u_2^2 \\ u_3' = 3 \times 10^7 u_2^2, \end{cases} \tag{1.0.5}$$

for $t \in (0, 40]$, with the initial condition

$$\begin{cases} u_1(0) = 1 \\ u_2(0) = 0 \\ u_3(0) = 0\,. \end{cases} \tag{1.0.6}$$

Integrate the IVP (1.0.5), (1.0.6) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method,

(c) The Radau method defined by the following Butcher table

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4}. \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

4. Consider IVP defined by the reaction of Belusov-Zhabotinskii

$$\begin{cases} u_1'(t) = 77.27(u_2 + u_1(1 - 8.375 \times 10^{-6}u_1 - u_2)) \\ u_2' = \dfrac{1}{77.27}(u_3 - u_2(1 + u_1)) \\ u_3' = 0.161(u_1 - u_3), \end{cases} \tag{1.0.7}$$

for $t \in (0, 50]$, with the initial condition

$$\begin{cases} u_1(0) = 1 \\ u_2(0) = 2 \\ u_3(0) = 3\,. \end{cases} \tag{1.0.8}$$

Integrate the IVP (1.0.7), (1.0.8) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method

(c) The Gauss method defined by the following Butcher table

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2}. \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

5. Consider IVP defined by the reaction of Belusov-Zhabotinskii

$$\begin{cases} u_1'(t) = 77.27(u_2 + u_1(1 - 8.375 \times 10^{-6}u_1 - u_2)) \\ u_2' = \dfrac{1}{77.27}(u_3 - u_2(1 + u_1)) \\ u_3' = 0.161(u_1 - u_3), \end{cases} \tag{1.0.9}$$

for $t \in (0, 50]$, with the initial condition

$$\begin{cases} u_1(0) = 1 \\ u_2(0) = 2 \\ u_3(0) = 3\,. \end{cases} \tag{1.0.10}$$

Integrate the IVP (1.0.9), (1.0.10) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method,

(c) The Radau method defined by the following Butcher table

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4}. \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

6. Consider IVP defined by the reaction of Belusov-Zhabotinskii

$$\begin{cases} u_1'(t) = 77.27(u_2 + u_1(1 - 8.375 \times 10^{-6}u_1 - u_2)) \\ u_2' = \dfrac{1}{77.27}(u_3 - u_2(1 + u_1)) \\ u_3' = 0.161(u_1 - u_3), \end{cases} \tag{1.0.11}$$

for $t \in (0, 50]$, with the initial condition

$$\begin{cases} u_1(0) = 1 \\ u_2(0) = 2 \\ u_3(0) = 3\,. \end{cases} \tag{1.0.12}$$

Integrate the IVP (1.0.11), (1.0.12) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method

(c) The Radau method defined by the following Butcher table

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4}. \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

7. Consider the chemical reaction problem

$$\begin{cases} u_1'(t) = -Au_1 - Bu_1u_2 \\ u_2' = Au_1 - MCu_2u_3 \\ u_3' = Au_1 - Bu_1u_3 - MCu_2u_3 + Cu_4 \\ u_4' = Bu_1u_3 - Cu_4, \end{cases} \tag{1.0.13}$$

where $A = 7.89 \times 10^{-10}, B = 1.1 \times 10^7, C = 1.13 \times 10^3$ and $M = 10^6$, for $t \in (0, 1000]$, with the initial condition

$$\begin{cases} u_1(0) = 1.76 \times 10^{-3} \\ u_2(0) = 0 \\ u_3(0) = 0 \\ u_4(0) = 0\,. \end{cases} \tag{1.0.14}$$

Integrate the IVP (1.0.13), (1.0.14) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method

(c) The Radau method defined by the following Butcher table

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4}. \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

8. Consider the chemical reaction problem

$$\begin{cases} u_1'(t) = -Au_1 - Bu_1u_2 \\ u_2' = Au_1 - MCu_2u_3 \\ u_3' = Au_1 - Bu_1u_3 - MCu_2u_3 + Cu_4 \\ u_4' = Bu_1u_3 - Cu_4, \end{cases} \tag{1.0.15}$$

where $A = 7.89 \times 10^{-10}, B = 1.1 \times 10^7, C = 1.13 \times 10^3$ and $M = 10^6$, for $t \in (0, 1000]$, with the initial condition

$$\begin{cases} u_1(0) = 1.76 \times 10^{-3} \\ u_2(0) = 0 \\ u_3(0) = 0 \\ u_4(0) = 0\,. \end{cases} \tag{1.0.16}$$

Integrate the IVP (1.0.15), (1.0.16) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method

(c) The Lobato method defined by the following Butcher table

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

9. Consider the chemical reaction problem

$$\begin{cases} u_1'(t) = -Au_1 - Bu_1u_2 \\ u_2' = Au_1 - MCu_2u_3 \\ u_3' = Au_1 - Bu_1u_3 - MCu_2u_3 + Cu_4 \\ u_4' = Bu_1u_3 - Cu_4 \end{cases} \tag{1.0.17}$$

where $A = 7.89 \times 10^{-10}, B = 1.1 \times 10^7, C = 1.13 \times 10^3$ and $M = 10^6$, for $t \in (0, 1000]$, with the initial condition

$$\begin{cases} u_1(0) = 1.76 \times 10^{-3} \\ u_2(0) = 0 \\ u_3(0) = 0 \\ u_4(0) = 0\,. \end{cases} \tag{1.0.18}$$

Integrate the IVP (1.0.17), (1.0.18) by using the following methods

(a) Explicit Euler's method,

(b) Implicit Euler's method

(c) The Gauss method defined by the following Butcher table

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2}. \end{array}$$

Compare the previous methods taking into account the accuracy and the stability.

References

[1] P. Ciarlet, The Finite Element Methods for Elliptic Problems, Noth Holland, 1978.

[2] E.A. Coddington, N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955.

[3] R. Dautry, J. L. Lions, Mathematical Analysis and Numerical Methods for Science and Technology, 1- Physical Origins and Classical Mathods, Springer Verlag, 1990.

[4] K. Dekker, J. Verwer, Stability of Runge-Kutta methods for stiff Nonlinear Differential Equations, CWI Monograpfs 2, North-Holland, 1984.

[5] E. DiBenedetto, Partial Differential Equations, Birkh*ä*user, 1995.

[6] D. Euvrard, Résolution Numérique des Équations aux Dérivées Partielles - Différences Finies, Element Finis, Masson, 1988.

[7] J.A. Ferreira, R.D. Grigorieff, On the supraconvergence of elliptic finitte difference schemes, Applied Numerical Mathematics, 28, 275-292, 1999.

[8] P. A. Forsyth, P. H. Sammon, Quadratic convergence for cell-centered grids, Applied NUmerical Mathematics, 4, 377-394, 1988.

[9] W. Gautschi, Numerical Analysis. An Introduction. Birkh*ä*user, Berlin, 1987.

[10] R.D. Grigorieff, Some stability inequalities for compact finite difference operators, Mathematic Nachter, 135, 93-101, 1986.

[11] B. Gustafsson, H-O Kreiss, J. Oliger, Time Dependent Problems and Difference Methods, John Wiley & Sons, 1995.

[12] W. Hackbush, Elliptic Differential Equations: Theory and Numerical Treatment, Springer, 1992.

[13] E. Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations II- Stiff and Differential Equations-Algebraic Problems, Second Edition, Springer-Verlag, Berlim, 1996.

[14] E. Hairer, G. Wanner, Solving Ordinary Differential Equations I- Nonstiff Problems, Springer-Verlag, Berlim, 1987.

[15] F. de Hoog, D. Jacket, On the rate of convergence of finite difference schemes on nonuniform grids, Journal of Australian Mathematical Society Sr. B, 247-256, 1985.

[16] W. Hunsdorfer, J.G. Verwer, Numerical Solution of Time-Dependent Advection-Diffusion Reaction Equation. Springer-Verlag, Berlin Heidelberg, 2003.

[17] A. Iserles, G. Strang,The optimal accuracy of difference schemes, Translations of the American Mathematal Society, 277, 779-803. 1983.

[18] A. Iserles, S.P. Nørsett, Order Stars. Applied Mathematics and Mathematical Computation, 2, Chapman & Hall, London, 1991.

[19] F. John, Partial Differential Equations, Springer-Verlag, Fourth Edition, 1982.

[20] C. Johnson, Numerical Solution of Partial Differential Equations by Finite Element Method, Cambridge University, 1990.

[21] H.B. Keller, Numerical Methods for Two-point Boundary Value Problems, SIAM, Philadelphia, 1976.

[22] H. O. Kreiss, T.A. Manteuffel, B. Swartz, B. Wendrof, A. B. White Jr. Supraconvergent schemes on irregular grids, Mathematics of Computations, 47, 537-554, 1986.

[23] J.D. Lambert, Computational Methods for Ordinary Differential Equations, Wiley, Chichester, 1991.

[24] K.W. Morton, Stability of finite difference approximation to a diffusion-convection equation, International Journal of Numerical Methods for Engineering,15, 677-683, 1980.

[25] M. A. Marletta, Supraconvergence of discretization methods on nonuniform meshes, Master Sciences Thesis, Oxford University, 1988.

[26] J.T. Odden, J.N. Reddy, An Introduction to the Mathematical Theory of Finite Elements, John Wiley & Sons, 1976.

[27] A. Quarteroni, A. Valli, Numerical Approximation of Partial Differential Equations, Springer, 1997.

[28] A. Quarteroni, R. Saco, F. Saleri, Numerical Mathematics, Springer, New York, 2000.

[29] P.A. Raviart, J.M. Thomas Introduction à l'Analyse Numérique des Équations aux Dérivées Partielles, Masson, 1983.

[30] R.D. Rychtmyer, K. W. Morton, Difference Methods for Initial-Value Problems. Second Edition, John Wiley & Sons, Interscience Publishers, New York. 1976.

[31] H.G. Roos, M. Stynes, L. Tobiska, Numerical Methods for Singularly Perturbed Differential Equations. Springer Series in Computational Mathematics, 24, Springer, Berlin, 1996.

[32] I. Rubinstein, L. Rubinstein, Partial Differential Equations in Classical Mathematical Physics, Cambridge University Press, 1993.

[33] W. S. Strauss, Partial Differential Equations - An Introduction, John Wiley & Sons, 1992.

[34] J.W. Thomas, Numerical Partial Differential Equations: Finite Difference Methods, Springer, 1995.

[35] V. Thomée, Finite Element Method for Parabolic Problems, Springer, 1984.

[36] D.W. Trim, Applied Partial Differential Eq.uations, PWS-Kent Publishing Company, 1990.

[37] G. A. Sod, Numerical Methods in Fluid Dynamics, Cambridge University Press, 1988.

[38] A. Weiser, M. F. Wheeler, On convergence of block-centered finite differences for elliptic problems, SIAM Journal of Numerical Analysis, 25, 351-375, 1988.