1MAE011 – Scientific Computing

Michel Fournié


INSTITUT DE MATHÉMATIQUES
de TOULOUSE

michel.fournie@math.univ-toulouse.fr or Michel.FOURNIE@ext.isae-supaero.fr

1.1.1 (1/5) – Contents of the course, Notion of PDE

- mathematical model : representation or abstract interpretation of physical reality using partial differential equations (PDEs)
  $\rightsquigarrow$ mathematical analysis of models
- numerical simulation : calculate the solution of these models on computers (to simulate physical reality)
  $\rightsquigarrow$ numerical analysis

Definition

PDEs are Differential equations in several variables
(time and space for example)

Remark

Sometimes, numerical calculation of the solutions has some unpleasant surprises which can be explained by understanding their mathematical properties only

1.1.1 (2/5) – BVP and IVP

- Most of the problems and applications are fundamentally **nonlinear** but for simplicity we consider only **linear** problems
- We consider deterministic problems (no stochastic components)

Definition

A Boundary Value Problem (BVP) is a problem formed by one or several P.D.E. with boundary conditions given on the totality of the boundary

Definition

A Cauchy's problem (or pure initial value problem IVP) is a time depending problem with an initial data in time

1.1.1 (3/5) – Notations

To write PDE we use some notations that are independent on the space dimension ($N = 1,\ 2,\ 3$) of the problem

For some vectors u and v in $\mathbb{R}^N$ we denote

- $u.v = (u, v)_N = \sum_{i=1}^{N} u_i v_i$ the scalar product
- $\|u\|_N = ((u, u)_N)^{\frac{1}{2}} = \left(\sum_{i=1}^{N} u_i^2 \right)^{\frac{1}{2}}$ the euclidian norm
- $u \bigotimes v \in \mathbb{R}^{N,N}$ where $(u \bigotimes v)_{ij} = u_i v_j,\ 1 \leq i, j \leq N$

For some matrices σ and τ in $\mathbb{R}^{N,N}$

- $\sigma : \tau = \sum_{i,j=1}^{N} \sigma_{ij} \tau_{ij}$

1.1.1 (4/5) – Differential operators

For a function u from $\mathbb{R}^N$ into $\mathbb{R}$

Definition

Gradient : $\text{grad}(u) \in \mathbb{R}^N$

$$\text{grad}(u) = \nabla u = (\partial_i u)_{1 \leq i \leq N} = \left(\frac{\partial u}{\partial x_1}, \cdots, \frac{\partial u}{\partial x_N} \right)$$

Definition

Laplacian : $\nabla^2 u \in \mathbb{R}$

$$\nabla^2 u = \Delta u = \sum_{i=1}^{N} \frac{\partial^2 u}{\partial x_i^2}$$

Remark

These operators must be considered for space variables (not for time)

1.1.1 (5/5) – Differential operators

For a function u from $\mathbb{R}^N$ into $\mathbb{R}^N$

Definition

Gradient : $\text{grad}(u) \in \mathbb{R}^{N,N}$ is defined by

$$\text{grad}(u) = \nabla u = (\partial_j u_i)_{1 \le i \le N} = \begin{pmatrix} \partial_1 u_1 & \cdots & \partial_N u_1 \\ \vdots & & \vdots \\ \partial_1 u_N & \cdots & \partial_N u_N \end{pmatrix}$$

Definition

Divergence : $\text{div}(u) \in \mathbb{R}$ is defined by

$$\text{div}(u) = \nabla . u = \sum_{i=1}^{N} \frac{\partial u_i}{\partial x_i} \quad \text{with} \quad u = (u_1, \cdots, u_N)$$

Definition

Laplacian $\Delta u \in \mathbb{R}^N$ is $\Delta u = \left(\sum_{j=1}^{N} \frac{\partial^2 u_i}{\partial x_j^2} \right)_{1 \le i \le N} = \text{div}\left(\text{grad}(u)\right) = \Delta u$

1.3 – Classification of PDE, Well posed problem

Definition

We consider (only) the general second order PDE of the specific form

(the variables are x and y and can represent space and time variables)

$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g$$

We compute $\delta = b^2 - 4ac$ and we said that this equation is

elliptic if $\delta < 0$, parabolic if $\delta = 0$, hyperbolic if $\delta > 0$

Remark

This definition doesn't include all P.D.E.

For other P.D.E., technical algebraic manipulations must be done

Definition

The problem $A(u) = f$ is well posed if for all data f,

- there exists a unique solution u and
- the solution depends continuously on the data f

1.4.1 – Poisson equation (elliptic 2nd order)

- The steady state in heat conduction (no time)
- The displacement of a structure subjects to some forces (Newton's law)

  gives a PDE of the form

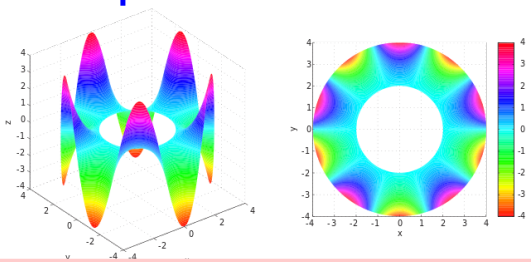
$$-\Delta u = f,\ x \in \Omega$$

- If the source term $f = 0$

  the PDE is called Laplace equation or potential equation

  then the PDE is reduced to

$$\Delta u = 0,\ x \in \Omega$$



Exercice

Prove that the nature of the PDE is Elliptic

1.4.2 – Transport or advection equation (hyperbolic 1st order)

- To describe the transport / convection of polluting substance along a channel where u is the concentration of this substance and c is the stream speed

$$u_t + c.\nabla u = 0,\ x \in \mathbb{R}^N,\ t > 0,\ c \in \mathbb{R}^N$$

- In 1D, the PDE is reduced for $u = u(x,t)$ to

$$u_t + cu_x = 0,\ x \in \mathbb{R},\ t > 0,\ c \in \mathbb{R}$$

Exercice

Prove that the nature of the PDE is Hyperbolic

1.4.3–Waves equation (hyperbolic 2nd order)

- To describe the propagation of waves in vibrating string where u is the wave amplitude and c is the propagation speed

$$\boxed{u_{tt} - c^2 \Delta u = 0}$$

- In 1D, the PDE is reduced for $u = u(x,t)$ to

$$u_{tt} - c^2 u_{xx} = 0,\ c > 0$$

Exercice

Prove that the nature of the PDE is Hyperbolic

1.4.4 – Diffusion or heat equation (parabolic 2nd order)

- To represent the heat conduction, where u is the time dependent temperature and κ is the heat conductivity

$$u_t - \kappa \Delta u = 0$$

- In 1D, the PDE is reduced for $u = u(x,t)$ to $u_t - \kappa u_{xx} = 0,\ \kappa > 0$
- to represent the **advection (convection)** coupled with **diffusion**

$$u_t - \Delta u + c.\nabla u = 0$$

- In 1D, the PDE is reduced for $u = u(x,t)$ to $u_t - u'' + cu' = 0,\ \kappa > 0$

Exercice

Prove that the nature of the PDE is Parabolic

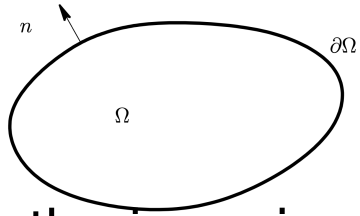
1.4.5 – Boundary Conditions (BC) : Heat Equation

- To find **particular solution** we consider **Boundary Condition**
- The type of boundary conditions depends on the physical context
- If the domain is surrounded by a region of constant temperature, then (after rescaling) the temperature satisfies the

Dirichlet boundary condition $\boxed{T(x) = 0 \text{ for all } x \in \partial\Omega}$

- If the domain is assumed to be adiabatic or thermally isolated from the exterior, then the heat flux across the boundary is zero and the temperature satisfies the **Neumann boundary condition**

$$\boxed{\frac{\partial T}{\partial \boldsymbol{n}}(x) := \boldsymbol{n}(x).\nabla T(x) = 0 \text{ for all } x \in \partial\Omega}$$



- If the heat flux across the boundary is proportional to the jump in the temperature from the exterior to the interior, the temperature satisfies **the Fourier (or Robin) boundary condition**

$$\boxed{\frac{\partial T}{\partial \boldsymbol{n}}(x) + \alpha T(x) = 0 \text{ for all } x \in \partial\Omega, \ (\alpha > 0 \text{ constant})}$$

2.1 – Elliptic PDE

- Thermodynamic : computation of the temperature
- Electromagnetism : computation of the electrical potential
- Elasticity : equilibrium of an elastic solid under constraints
- Fluid dynamic : potential flows
- Quantum physic : Schrödinger solutions

Poisson equation in 1D

- We consider the Poisson problem

$$\begin{cases} -u''(x) = f(x) \quad \text{for} \quad x \in]0,1[\\ u(0) = u(1) = 0 \end{cases}$$

- Denote the Green function by $G(x,y)$ defined by

$$G(x,y) = \begin{cases} y(1-x) & \text{if } 0 \leq y \leq x \\ x(1-y) & \text{if } x \leq y \leq 1 \end{cases}$$

- The solution of the problem is

$$u(x) = \int_0^1 G(x,y) f(y) \mathrm{d}y$$

Qualitative properties

- The problem is **well posed** (existence, uniqueness)
- G is positive (G is equal to zero on the boundary of $[0,1]^2$)
- Then the **maximum principle** is satisfied

  If $f \geq 0$ then $u \geq 0$

  If $f > 0$ then $u > 0$ on $]0,1[$
- For a right hand side f localized in a small zone (for example $f = \chi_{[\zeta,\zeta+\epsilon]}$ with ϵ very small) the solution u is not equal to 0 all over the domain $]0,1[$
- If we modify locally the right hand side f (over a small interval of length ϵ), the solution u is modified **globally** (all over $]0,1[$)
- **Regularization** : If $f \in C^0([0,1])$ then $u \in C^2([0,1])$

Laplace equation on infinite domain

Theorem

The problem defined on all the space

$$\begin{cases} -\Delta u = f \quad \text{for} \quad x \in \mathbb{R}^3 \\ u \longrightarrow 0, \ |x| \longrightarrow \infty \end{cases}$$

has a unique solution given by

$$u(x) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{f(x)}{|x-y|} \mathrm{d}y$$

Remark

The study of this problem is done using the theory of the Distribution and of the Fourier transform

2.2 – Parabolic PDE

- Heat diffusion
- Brownien process
- Fluid dynamics : viscosity phenomena (Navier-Stokes equations)

Consider the initial/boundary value problem (P) on an interval $I \subset \mathbb{R}$

$$(P) \quad \begin{cases} u_t - \kappa u_{xx} = 0, & x \in I, \ t > 0 \\ u(x,0) = \phi(x), & x \in I \\ u \text{ satisfies some Boundary Conditions} \end{cases}$$

The most common boundary conditions are the following :

- **Dirichlet** $(I = (0,l))$: $u(0,t) = 0 = u(l,t)$
- **Neumann** $(I = (0,l))$: $u_x(0,t) = 0 = u_x(l,t)$
- **Robin** $(I = (0,l))$:

  $u_x(0,t) - a_0 u(0,t) = 0$ and $u_x(l,t) + a_l u(l,t) = 0$
- **Periodic** $(I = (-l,l))$:

  $u(-l,t) = u(l,t)$ and $u_x(-l,t) = u_x(l,t)$

Solution of the heat problem with Dirichlet BCs

Theorem

The solution of the heat equation on $I = [0, l]$ with Dirichlet BCs is

$$u(x,t) = \sum_{n=1}^{\infty} A_n sin\left(\frac{n\pi}{l}x\right) e^{-\kappa(\frac{n\pi}{l})^2 t}$$

where

$$A_n = \frac{2}{l}\int_0^l sin\left(\frac{n\pi x}{l}\right)\phi(x)\mathrm{d}x$$

Remark

The function $\sum_{n=1}^{\infty} A_n sin\left(\frac{n\pi}{l}x\right)$ is the **Fourier sine series** of ϕ

Other Boundary conditions

Theorem

The solution of the Heat equation on $I = [-l, l]$ with Periodic Boundary conditions is given by

$$u(x,t) = A_0 + \sum_{n=1}^{\infty} \left[A_n cos\left(\frac{n\pi}{l}x\right) + B_n sin\left(\frac{n\pi}{l}x\right) \right] e^{-\kappa(\frac{n\pi}{l})^2 t}$$

where

$$A_0 = \frac{1}{2} l \int_{-l}^{l} \phi(x)\mathrm{d}x,\ A_n = \frac{1}{l}\int_{-l}^{l} cos\left(\frac{n\pi}{l}x\right)\phi(x)\mathrm{d}x,\ B_n = \frac{1}{l}\int_{-l}^{l} sin\left(\frac{n\pi}{l}x\right)\phi(x)\mathrm{d}x$$

Remark

$A_0 + \sum_{n=1}^{\infty} \left[A_n cos\left(\frac{n\pi}{l}x\right) + B_n sin\left(\frac{n\pi}{l}x\right) \right]$ is the **full Fourier series** of ϕ

Solution of the Heat Equation in $\mathbb{R}$

Theorem

The solution of the Initial Value Problem (IVP) for the heat equation is

$$u(x,t) = \frac{1}{\sqrt{4\kappa\pi t}} \int_{-\infty}^{+\infty} \phi(y) e^{-\frac{(x-y)^2}{4\kappa t}} \mathrm{d}y, \text{ for } t > 0$$

Remark

- For $t = 0$, we consider $\lim_{t \longrightarrow 0^+} u(x,t) = \phi(x)$
- This result is obtained using Fourier analysis

Extension in $\mathbb{R}^N$

- We consider he problem in higher dimension $\mathbb{R}^N$ given by

$$\begin{cases} u_t = \kappa \Delta u, \quad u \in \mathbb{R}^N, \ t > 0, \qquad \boldsymbol{\kappa > 0} \\ u(x,0) = \phi(x) \end{cases}$$

- Using the Fourier transform as in 1-D we obtain the solution

$$u(x,t) = \frac{1}{(4\kappa\pi t)^{\frac{N}{2}}} \int_{\mathbb{R}^n} \phi(y) e^{-\frac{|x-y|^2}{4\kappa t}} \mathrm{d}y, \text{ for } t > 0,$$

$$= [H(t) * \phi](x) = \int_{\mathbb{R}^n} H(x-y,t)\phi(y)\mathrm{d}y$$

where $H(x,t)$ is the **fundamental solution** of the heat equation

$$H(x,t) = \begin{cases} \frac{1}{(4\pi\kappa t)^{\frac{N}{2}}} e^{-\frac{|x|^2}{4\kappa t}}, & t > 0 \\ 0, & t < 0 \end{cases}$$

Parabolic regularization

Theorem

Assuming $\phi \in C(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ then u solution of the heat equation satisfies

$$u \in C^\infty(\mathbb{R}^n \times (0,\infty)) \text{ and } \lim_{\substack{(x,t) \longrightarrow (x_0,0) \\ x_0, x \in \mathbb{R}^n, t>0}} u(x,t) = \phi(x_0)$$

Remark

This is a **regularization phenomena**

This is due to the regularity of $H(.,t) : x \longrightarrow \frac{1}{(4\pi\kappa t)^{\frac{n}{2}}} e^{-\frac{|x|^2}{4\kappa t}}$

which is a $C^\infty(\mathbb{R}^n)$ function for $t > 0$

Qualitative properties

Definition

Asymptotic behavior :

The solution converges to 0 (into $L^2(\Omega)$) when t tends to $+\infty$

Definition

We remark that for $\phi > 0$ then $u(x,\varepsilon)>0,\ \forall x \in \Omega$ and $\varepsilon > 0$

Physical Interpretation :

For an initial point which is cold ($\phi(x)$ very small) and very far from the heat source, it becomes immediately hot ($u(x,\varepsilon) > 0$)

We call that we have a **propagation of the heat with infinite speed**

This is not a realistic model ! ! !

Conservation of the "mass" and Irreversibility in time

Theorem

$$\int_{\mathbb{R}^n} u(x,t)\mathrm{d}x = \int_{\mathbb{R}^n} \phi(x)\mathrm{d}x$$

Theorem

For $t \leq 0$ the problem is **not well posed**

We can not obtain a solution

Remark : $H(x,t) \longrightarrow +\infty$ when $t \longrightarrow 0^-$

There is not a continuous extension of H for $t < 0$.

The corresponding situation is modelized by

the **Backward heat equation** (not well posed)

$$\begin{cases} u_t + \kappa \Delta u = 0 \\ u(x,0) = \phi(x) \end{cases}$$

2.3 – Transport equation

We consider the transport equation :

$$u_t + c u_x = 0, x \in \mathbb{R}, t > 0, c \in \mathbb{R}$$

Definition

We associate the **characteristic curves** $x = x(t)$ such that $\frac{dx}{dt} = c$ so

$$x(t) = ct + x(0)$$

If we differentiate u(x, t) along the characteristic curves, we find that

$$\frac{d}{dt} u(x(t), t) = u_t + u_x \frac{dx}{dt} = u_t + c u_x = 0$$

Hence u is constant along these characteristics

The Cauchy Problem

We consider the pure Initial Value Problem in 1-dimension

$$(\mathscr{P}_1)\begin{cases} u_t + c\,u_x = 0, x \in \mathbb{R}, t > 0, c \in \mathbb{R} \\ u(x,0) = u_0(x),\ x \in \mathbb{R} \end{cases}$$

Using the fact that u is constant along the characteristics, we have :

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

Theorem

The solution of the problem $(\mathscr{P}_1)$ is

$$u(x,t) = u_0(x - ct)$$

Theorem

The solution of the general Cauchy problem in N-dimensions :

$$(\mathscr{P}_1)' \begin{cases} u_t + c.\nabla u = f(x,t), x \in \mathbb{R}^N, t > 0, c \in \mathbb{R}^N, \\ u(x,0) = u_0(x),\ x \in \mathbb{R}^N, \end{cases}$$

is given by

$$u(x,t) = u_0(x - tc) + \int_0^t f(x + (s-t)c, s)\,ds$$

The Initial Boundary Value Problem (IBV)

Theorem

The solution of the problem ($f = 0$)

$$(\mathscr{P}_2)\begin{cases} u_t + cu_x = 0, x > 0, t > 0, c > 0, \\ u(x,0) = u_0(x),\ x \in \mathbb{R}, \\ u(0,t) = u_1(t), t > 0. \end{cases}$$

is equal to (issue from the characteristics method)

$$u(x,t) = \begin{cases} u_0(x-ct) \ \ if\ x > ct, \\ \\ u_1(t - \dfrac{x}{c}) \ \ if\ x < ct. \end{cases}$$

The IBV with a source term

Theorem

The solution of the problem ($f \neq 0$)

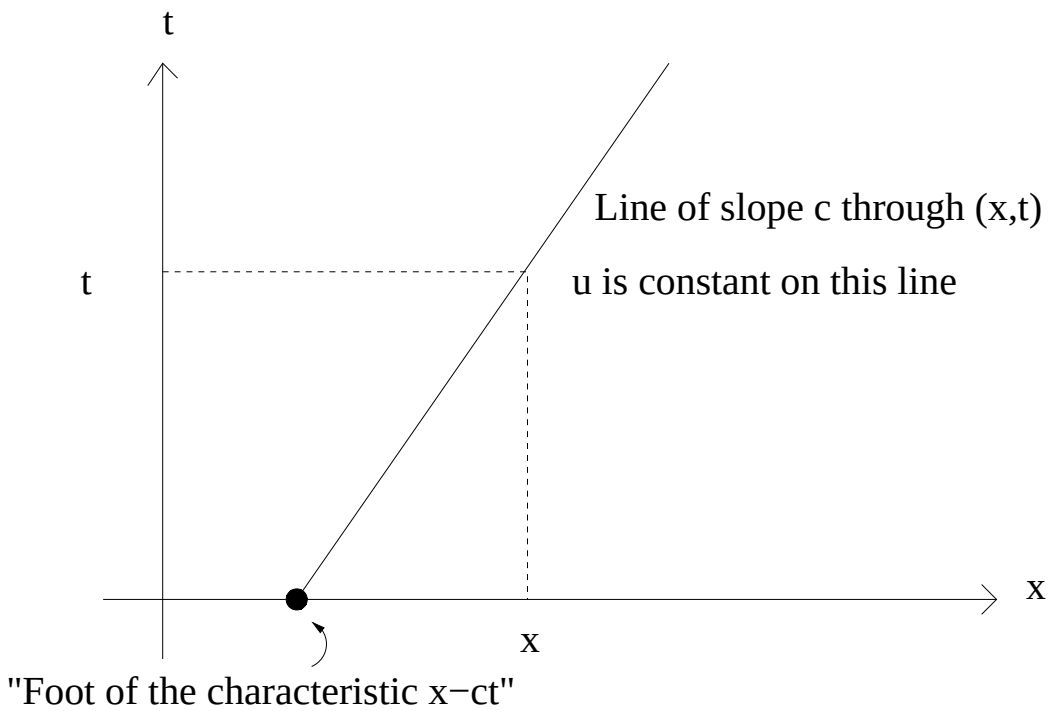
$$(\mathscr{P}_2)\begin{cases} u_t + cu_x = f, x > 0, t > 0, c > 0, \\ u(x,0) = u_0(x),\ x \in \mathbb{R}, \\ u(0,t) = u_1(t), t > 0. \end{cases}$$

is equal to

$$u(x,t) = \begin{cases} u_0(x-ct) + \int_0^t f(x+(s-t)c, s)ds \ \ if\ x > ct, \\ u_1(t-\frac{x}{c}) + \frac{1}{c}\int_0^x f(x-s, t-\frac{s}{c})ds \ \ if\ x < ct. \end{cases}$$

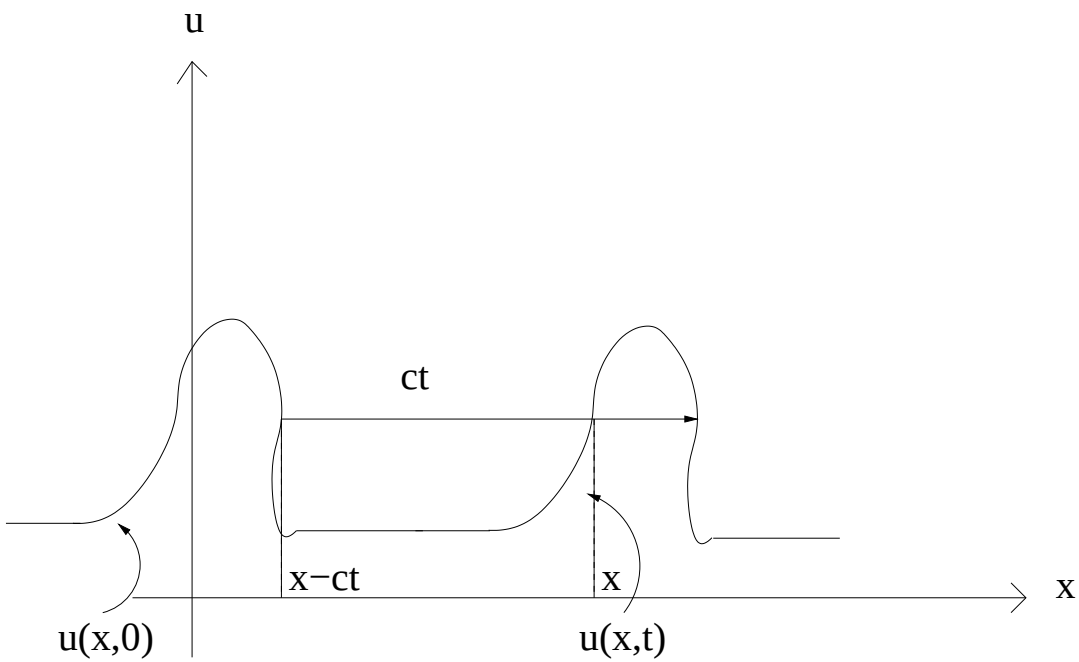
Interpretation

- We have proved that the solution of the transport equation is **constant along the characteristics**
- To obtain the value of $u(x,t)$, it is sufficient to **"come back"** on the characteristics untill $t=0$
- We have the following illustration



Propagation with a finite speed

- From the form $u(x,t) = u_0(x - ct)$, we deduce that the solution corresponds to **translate** u_0 with the quantity ct
- We have the following illustration



Regularisation and Conservation properties

- From $u(x,t) = u_0(x - ct)$, we show that $\forall\, t > 0$ $u(.,t)$ has the same regularity as u_0
- So, there is **not an instantaneous regularisation** like for parabolic problem
- There is a propagation of a singularity

Theorem

We have a conservation of the "mass"

$$\int_{\mathbb{R}} u(x,t)\mathrm{d}x = \int_{\mathbb{R}} u_0(x-ct)\mathrm{d}x = \int_{\mathbb{R}} u_0(\zeta)\mathrm{d}\zeta$$

Theorem

We have conservation of the infinite norm

$$\max_{x\in\mathbb{R}} |u(x,t)| = \max_{x\in\mathbb{R}} |u_0(x)|$$

Reversibility in time

- For $t<0$, the formula $u(x,t)=u_0(x-ct)$ is still satisfied
  so we have a solution of the transport problem for negative time
- The changement $t \longleftarrow -t$ transforms the Cauchy's problem into

$$\begin{cases} u_t - cu_x = 0, \quad x \in \mathbb{R}, \ t>0, \ c>0 \\ u(x,0) = u_0(x), \quad x \in \mathbb{R} \end{cases}$$

  which is again a well posed problem
- The problem is associated to a propagation with a speed $-c$

Introduction

- Generally we can't find exact solution of a PDE
  (the solution is search into an infinite space)
- So we search approximate solution. More precisely, we search approximate values of the solution in some points
  (the solution is search into a finite space)
- This point of vue is called the discretization ("for discret values")
  - $\Longrightarrow$ Finite Difference Method
  - $\Longrightarrow$ Finite Element Method
  - $\Longrightarrow$ Finite Volume Method $\cdots$
- After discretization, the problem is transformed into a Large Linear System to solve using computers
- The nature of the system depends on the PDE and specific algorithms can be choice to have optimal computations (few time consuming with high accuracy $\cdots$)

The discretization – The mesh (decomposition of Ω)

- How to transform an infinite problem into finite problem ?
- We introduce discret points x_j ("discretization") and we search approximate values of the solution u_j at these points
- A mesh grid in 1D for example on $\Omega = [0,1]$ is given by $(x_j)_{0 \le j \le N+1}$

$$x_0 = 0 < x_1 < \cdots < x_N < x_{N+1} = 1$$

Definition

We say that the mesh is uniform if the points x_j (or nodes) are equidistant

$$x_j = j\Delta x \quad \text{with } \Delta x = \frac{1}{N+1}, \quad 0 \le j \le N+1, \qquad (\Delta x \text{ can be written } h)$$

We denote

$$\boxed{u_j \approx u(x_j)}$$

Discretization for time dependent problem

- Same approach is used for the time with time step Δt
- To simplify we consider $t \in [0, T]$ and define

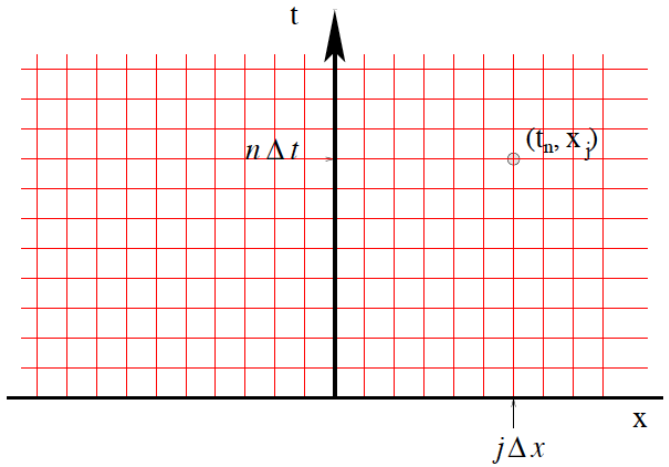
$$u_j^n \approx u(t^n, x_j)$$

- We construct a discretization in time for $t^j = j\Delta t$

$$t^0 = 0 < t^1 < \cdots < t^n < t^{n+1} = T$$

generally we choose $t^{j+1} - t^j = \Delta t$

(Δt and t^n can be written k and t_n)

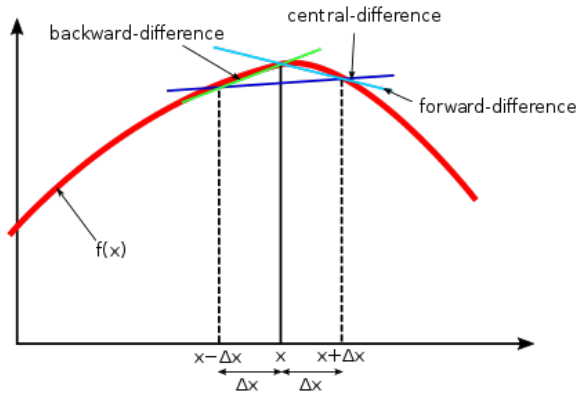


Finite Difference Formula based on **Taylor expansion**

- Finite difference formulas of order 2 (central difference)

$$\left(\frac{\partial u}{\partial x}\right)(x_j) \approx \frac{u_{j+1} - u_{j-1}}{2\Delta x}$$

$$\left(\frac{\partial^2 u}{\partial x^2}\right)(x_j) \approx \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} \dots$$



- Finite difference formula of order 1 (forward and backward difference)

$$\left(\frac{\partial u}{\partial x}\right)(x_j) \approx \frac{u_{j+1} - u_j}{\Delta x} \qquad \left(\frac{\partial u}{\partial x}\right)(x_j) \approx \frac{u_{j-1} - u_j}{\Delta x} \dots$$

Exercice

Prove the order of the finite difference formulae

FDS for Laplace problem (space only)

- We consider x_j for $j = 0, \cdots, N+1$.
  The points x_0 and x_{N+1} are on the boundary
- For Dirichlet boundary problem, (u_0 and u_{N+1} are known)

$$\boxed{\text{we search } u_j \text{ for } j = 1, \cdots, N}$$

- For exemple for the Laplacian problem
  We consider these equation at each interior point

$$\boxed{-\frac{\partial^2 u}{\partial x^2} = f} \Longrightarrow \boxed{-\left(\frac{\partial^2 u}{\partial x^2}\right)_j = f_j, \quad \text{for} \quad j = 1, \cdots, N \quad (f_j \approx f(x_j))}$$

  that is transformed using finite difference formula into

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} = f_j, \quad \text{for} \quad j = 1, \cdots, N$$

- We have N equations with N unknowns

$$\Longrightarrow \text{Linear System } \boxed{AU = b}$$

Linear System (FDS for Laplace problem) : $AU = b$

- The unknown vector (the solution) is given by

$$U = (u_1, u_2, \ldots, u_{N-1}, u_N)^T$$

- The linear system to solve is given by

$$\frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} f_1 - \frac{u_0}{\Delta x^2} \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N - \frac{u_{N+1}}{\Delta x^2} \end{pmatrix}$$

MAPLE : fds.mws

- Be careful to the influence of the B.C.

Maximum principle and Convergence

Theorem

The matrix A is tridiagonal, symmetric and inversible

Theorem

If all components of b are negative or equal to 0
then all components of the solution U are negative or equal to 0

Theorem

For $u \in C^4([0,1])$ the solution of the continuous problem and U the approximated solution given by the linear system, we have

$$\max_{1 \leq i \leq N} |u(x_i) - U_i| \leq \frac{\Delta x^2}{96} \max_{x \in [0,1]} |u^{(4)}(x)|$$

Remark

The convergence is obtained thanks to consistency and stability

Demonstration (1/3)

- For $u \in C^4([0,1])$, using Taylor expansion we have

$$\left| \frac{-u(x+\Delta x)+2u(x)-u(x-\Delta x)}{\Delta x^2} + u''(x) \right| \leq \frac{\Delta x^2}{12} \max_{x\in[0,1]} |u^{(4)}(x)|$$

  - **Remark :** Notions of consistency and accuracy order is 2

- For $E_i = U_i - u(x_i)$ ($1 \leq i \leq N$ and E the corresponding vector)

$$\left| \frac{-E_{i-1}+2E_i-E_{i+1}}{\Delta x^2} \right| \leq \frac{\Delta x^2}{12} \max_{x\in[0,1]} |u^{(4)}(x)|$$

  - **Remark :** We estimate the vector AE, the goal is to estimate $\|E\|_\infty$

- The matrix A is inversible, the coefficients α_{ij} of A^{-1} are ≥ 0 and

$$0 < \sum_{j=1}^{N} \alpha_{ij} \leq \frac{1}{8}$$

main difficulty of the proof

- For $V_i = \frac{1}{\Delta x^2}(-E_{i-1}+2E_i-E_{i+1})$ we have $E_i = \sum_{j=1}^N \alpha_{ij} V_j$

  Then $|E_i| \leq \left(\sum_{j=1}^N \alpha_{ij}\right) \max_{1\leq k\leq N} |V_k|$ and we conclude the proof

  In fact, the solution is bounded, we say that the scheme is stable

  because $U_i = \sum_{j=1}^N \alpha_{ij} F_j \implies |U_i| \leq \frac{1}{8} \sup_{1\leq j\leq N} |F_j|$

Demonstration (2/3)

- The main difficulty of the proof is based on the Discret Maximum Theorem (proof next slide) difficult to prove in general

  if all the components of $V = AU$ are ≤ 0 then

$$U_i \leq 0 \quad \text{for} \quad 1 \leq i \leq N$$

- Then A is inversible and all coefficients α_{ij} of A^{-1} are ≥ 0 (see next slide)

  Remark : For $U_i = \sum_{j=1}^{N} \alpha_{ij} F_j$ we have $|U_i| \leq \frac{1}{8} \sup_{1 \leq j \leq N} |F_j|$
- For v defined by $v(x) = \frac{1}{2} x(1-x)$ for $x \in [0,1]$, we have

$$-\frac{v(x_{i+1}) - 2v(x_i) + v(x_{i-1})}{h^2} = 1, \quad 1 \leq i \leq N$$

- Then $V = (v(x_1), \cdots, v(x_N))$ and $U = (1, \cdots, 1)$ satisfy $AV = U$
- We can deduce that for $1 \leq i \leq N$ we have

$$0 < \sum_{j=1}^{N} \alpha_{ij} = v(x_i) \leq \max_{x \in [0,1]} v(x) = \frac{1}{8}$$

Demonstration (3/3)

Max Principia : ($\Delta x = h$) We fix i such that $U_i = \max_{1\leq j\leq N} U_j$, we suppose that $U_i > 0$ and we prove that is impossible.

• For $i = 1$ then due to the fact that $AU = V$ has all its components ≤ 0, we obtain that $2U_1 - U_2 = h^2 V_1 \leq 0$. However U_1 is the maximum so $U_1 \geq U_2 \Longrightarrow 2U_1 - U_2 \geq U_1 > 0$. So $U_1 \leq 0$ which is in opposition with the hypothesis $U_1 > 0$.
In conclusion $U_1 \leq 0$ for all i.

• For $i = N$ then due to the fact that $AU = V$ has all its components ≤ 0, we obtain that $-U_{N-1} + 2U_N = h^2 V_N \leq 0$. However U_N is the maximum so $U_N \geq U_{N-1} \Longrightarrow -U_{N-1} - 2U_N \geq U_N > 0$. So $U_N \leq 0$ which is in opposition with the hypothesis $U_N > 0$.
In conclusion $U_i \leq 0$ for all i.

• For $2 \leq i \leq N$ then due to the fact that $AU = V$ has all its components ≤ 0, we obtain that
$-U_{i-1} + 2U_i - U_{i+1} = \underbrace{(U_i - U_{i-1})}_{a} + \underbrace{(U_i - U_{i+1})}_{b} = h^2 V_i \leq 0$. However U_i is the maximum so $a = U_i - U_{i-1} \geq 0$ and $b = U_i - U_{i+1} \geq 0$
so $0 \leq a + b \leq 0$ so we have a sum of two positive values equal to 0 which implies that $a = 0$ and $b = 0$. We conclude that $U_{i+1} = U_i = U_{i-1}$. This relation is true for all i (if U_i is a maximum then the previous and next solutions are some maximum too). For exemple for $i = 2$ we have $U_1 = U_2 = U_3$ who are maximum values supposed > 0 then $U_1 > 0$ is a maximum but we have already study this case and conclude (using U_1 we conclude for $i = 2$ so for U_2, using U_2 we conclude for U_3 and so on). So we contradict the hypothesis that the maximum value $U_i > 0$ for all i.

Inversibility of A : We verify that (if $AU = 0$ then $U = 0$ is the unique solution).
Using Max Principe we have $AU = V = 0$ then $U_i \leq 0$ for all i, so we can conclude as soon as we prove that $U_i < 0$ is impossible.
We fix i such that $U_i = \min_{1\leq j\leq N} U_j$, we suppose that this minimum $U_i < 0$ and we prove that is impossible.

• For $i = 1$ then due to the fact that $AU = 0$ we have $2U_1 - U_2 = 0$. However U_1 is the minimum so $U_1 \leq U_2 \Longrightarrow 2U_1 - U_2 \leq U_1 < 0$. We conclude that $0 = 2U_1 - U_2 < 0$ which is impossible. By recurrence (same technic used to prove Max principle), we can conclude.

FDS and time dependent problem

Classical Finite Difference Formulae

- Central approximation

$$\left(\frac{\partial u}{\partial t}\right)(t^n, x_j) \approx \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t}$$

- Backward approximation

$$\left(\frac{\partial u}{\partial t}\right)(t^n, x_j) \approx \frac{u_j^{n} - u_j^{n-1}}{\Delta t}$$

- Forward approximation

$$\left(\frac{\partial u}{\partial t}\right)(t^n, x_j) \approx \frac{u_j^{n+1} - u_j^{n}}{\Delta t}$$

FDS for Heat equation (space and time)

We consider the heat equation

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \dfrac{\partial^2 u}{\partial x^2} = 0, & \text{for } x(x,t) \in (0,1) \times \mathbb{R}_*^+ \\ u(0,x) = u_0(x), & \text{for } x \in (0,1) \quad \text{(Initial Condition I.C.)} \\ u(t,0) = u(t,1) = 0 & \text{for } t \in \mathbb{R}_*^+ \quad \text{(Boundary Conditions B.C.)} \end{cases}$$

- Initialization $u_j^0 = u_0(x_j)$ where $u_0(x)$ is the Initial londition
- We take $u_0^n = u_{N+1}^n = 0$, for $n > 0$ due to the Boundary Conditions
- Several discretizations exist

  What is the best one ? Why ?

- In practice, we test and we try to justify the behavior of the scheme when Δx and Δt tend to 0

Numerical illustration

- We will present classical schemes used to solve the Heat PDE
- We fix $\nu = 1$
- The initial condition we will consider is

$$u_0(x) = \max(1 - x^2, 0)$$

- The domain will be $\Omega = [-10; 10]$
  - **Dirichlet B.C.**
  - We fix $\boldsymbol{\Delta x = 0.05}$ (so we search 401 values $(u_j^n)_{-200 \leq j \leq 200}$)
  - Several choices for Δt will be discuss

FDS for Heat equation : Classical Schemes (2 levels)

Iterative process :

We consider that **we know the values** u_j^n at the time level n and **we search the unknowns** u_j^{n+1} at the time level $n+1$

- **Explicit scheme (or Forward Time Centred Space FTCS)**

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}-\nu\frac{u_{j-1}^n-2u_j^n+u_{j+1}^n}{\Delta x^2}=0$$

Explicit $\longleftrightarrow$ it is straightforward to compute u^{n+1} from u^n

- **Implicit scheme**

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}-\nu\frac{u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1}}{\Delta x^2}=0$$

Implicit $\longleftrightarrow$ linear system to solve to find u^{n+1} from u^n

FDS for Heat equation : Classical Schemes (3 levels)

- **DuFort-Frankel scheme** which is a **Centred scheme** (called **Richardson** scheme if we consider convection term)

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} - \nu \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{\Delta x^2} = 0$$

- **Gear scheme**

$$\frac{3u_j^{n+1} - 4u_j^n + u_j^{n-1}}{2\Delta t} - \nu \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{\Delta x^2} = 0$$

FDS for Heat equation : Classical Schemes (2 levels)

- **θ–scheme** for $0 \leq \theta \leq 1$

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \nu\theta \frac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{\Delta x^2} - \nu(1-\theta) \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{\Delta x^2} = 0$$

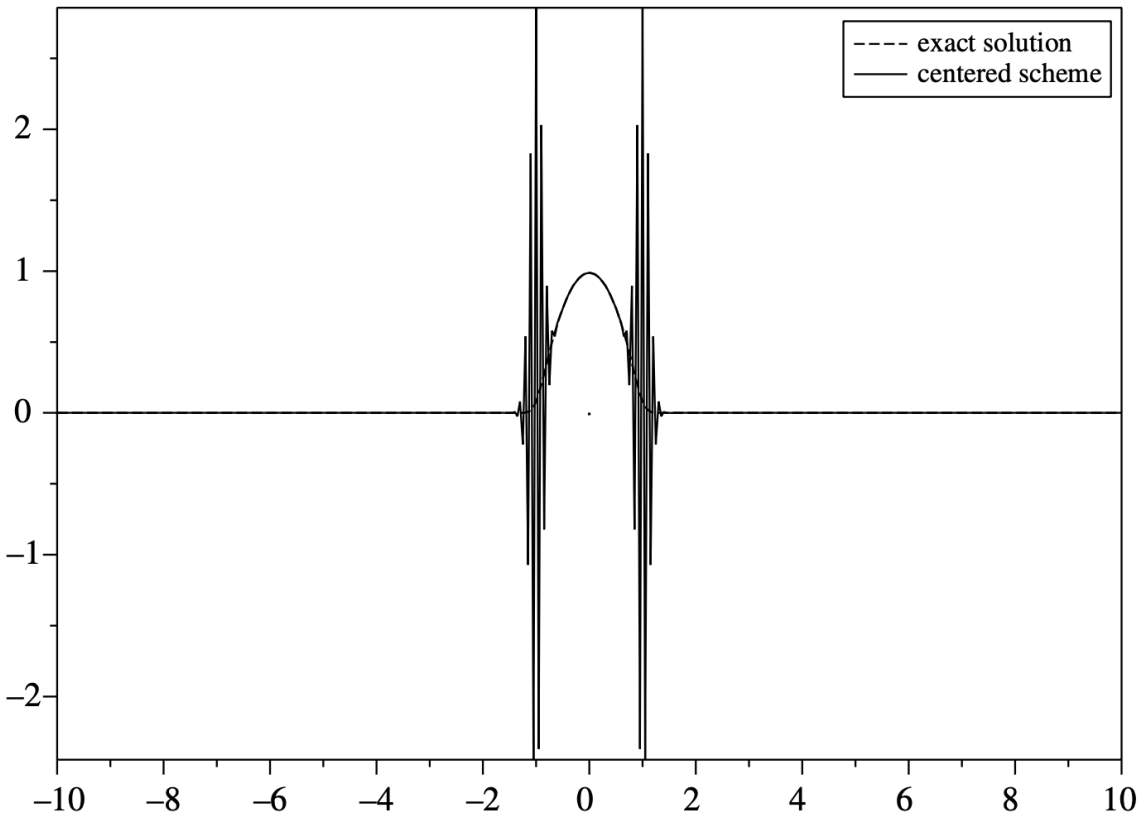
- If $\boxed{\theta = 0}$ then it corresponds to the **Explicit scheme**
- If $\boxed{\theta = 1}$ then it corresponds to the **Implicit scheme**

(if $\theta \neq 0$ then the scheme is implicit)

- If $\boxed{\theta = \frac{1}{2}}$ we obtain the **Crank-Nicolson scheme**

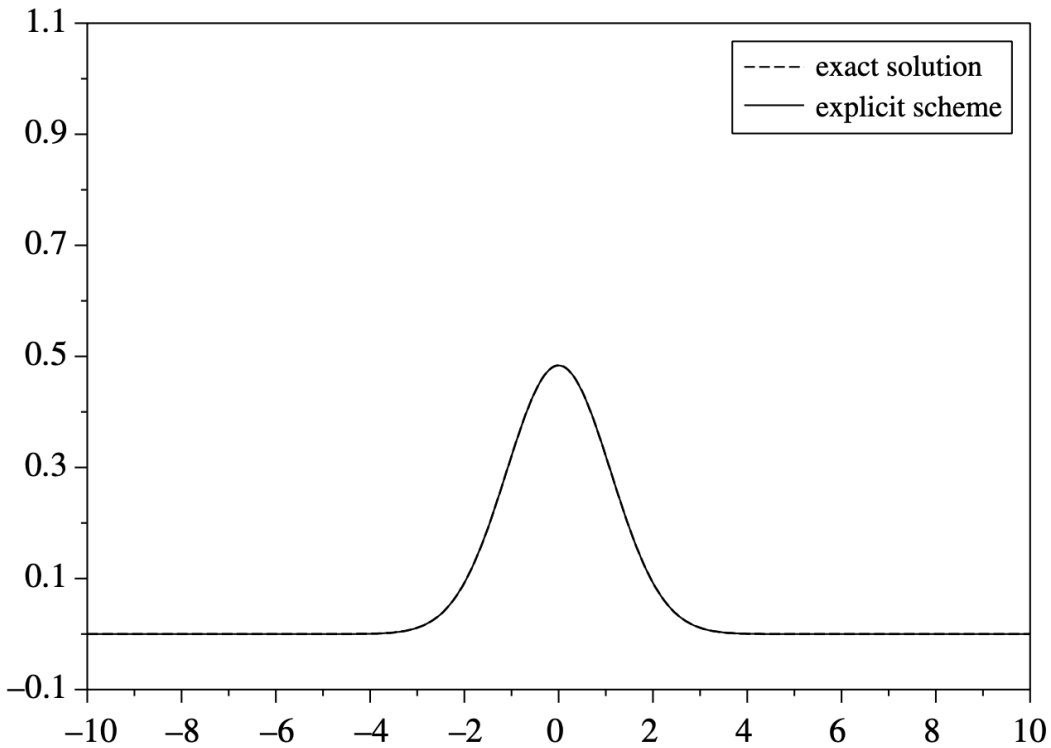
(one of the most popular scheme)

Heat equation, centred scheme, CFL=$\frac{\nu \Delta t}{\Delta x^2} = 0.1$
25 Time steps



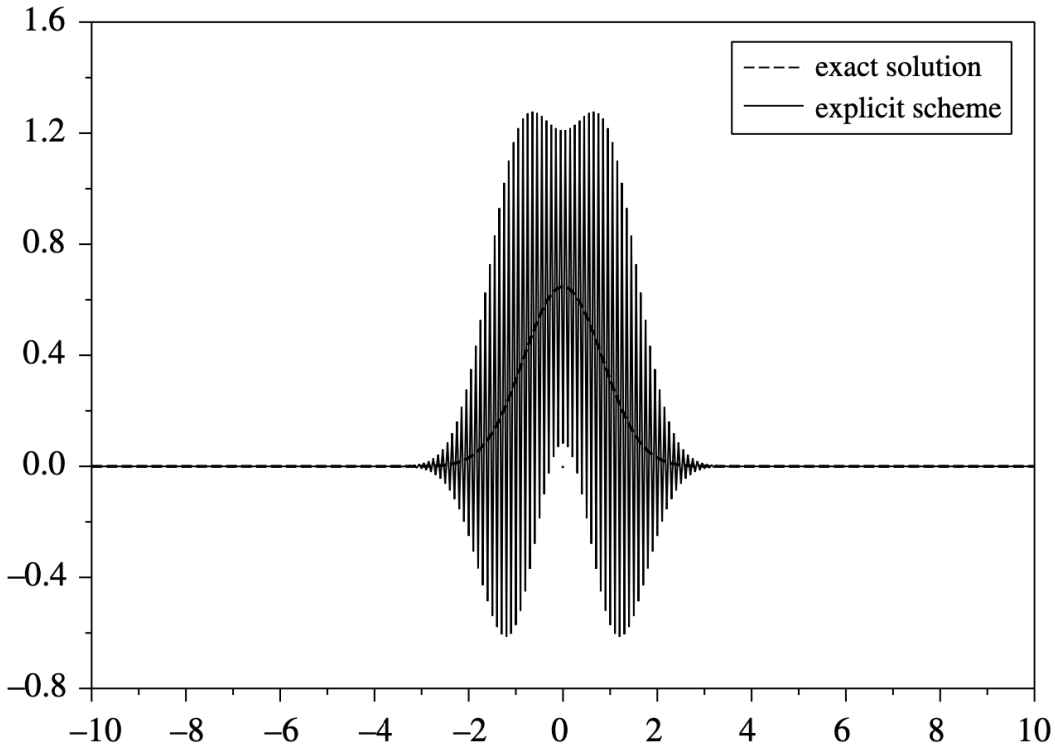
Unstable centred scheme with $\nu \Delta t = 0.1 \Delta x^2$

Heat equation, explicit scheme, CFL=$\frac{\nu \Delta t}{\Delta x^2} = 0.4$ 500 Time steps



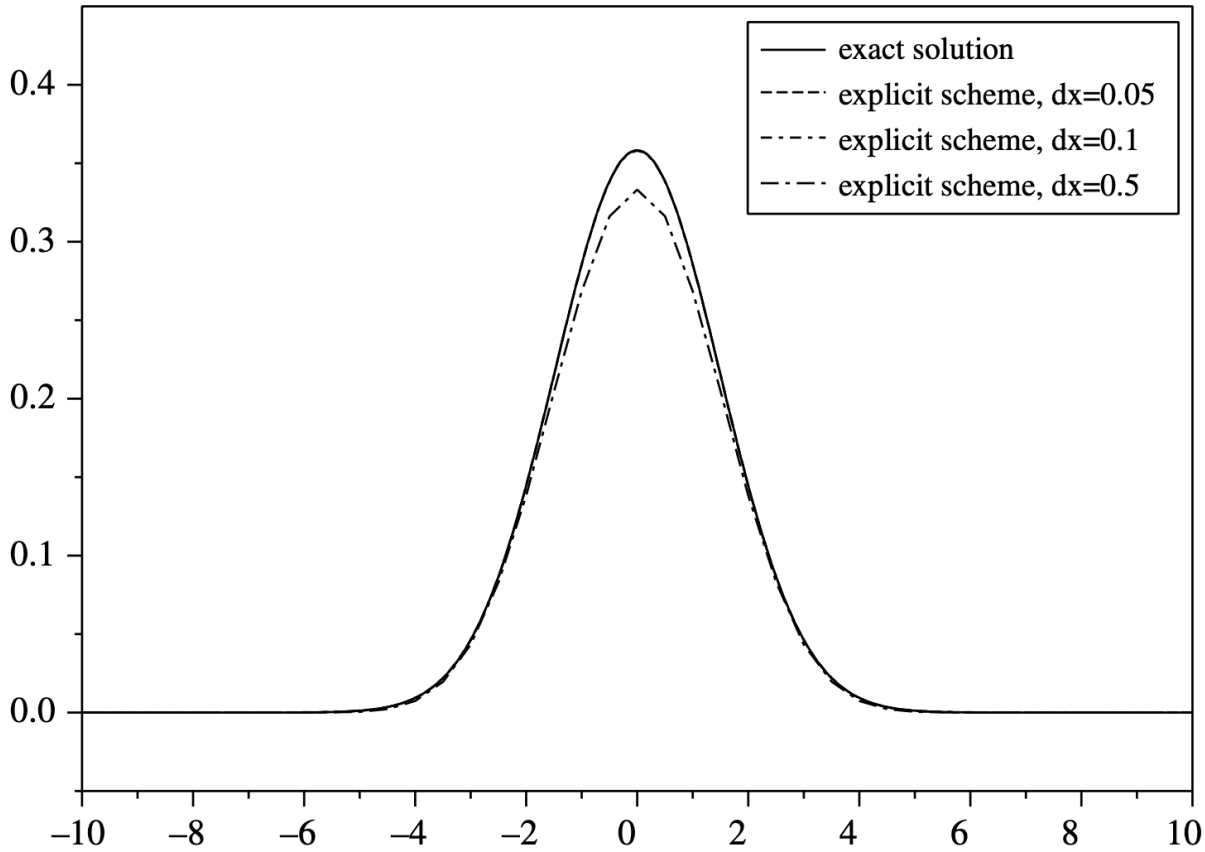
Explicit scheme with $\nu \Delta t = 0.4 \Delta x^2$

Heat equation, explicit scheme, CFL=$\frac{\nu \Delta t}{\Delta x^2} = 0.51$
180 Time steps



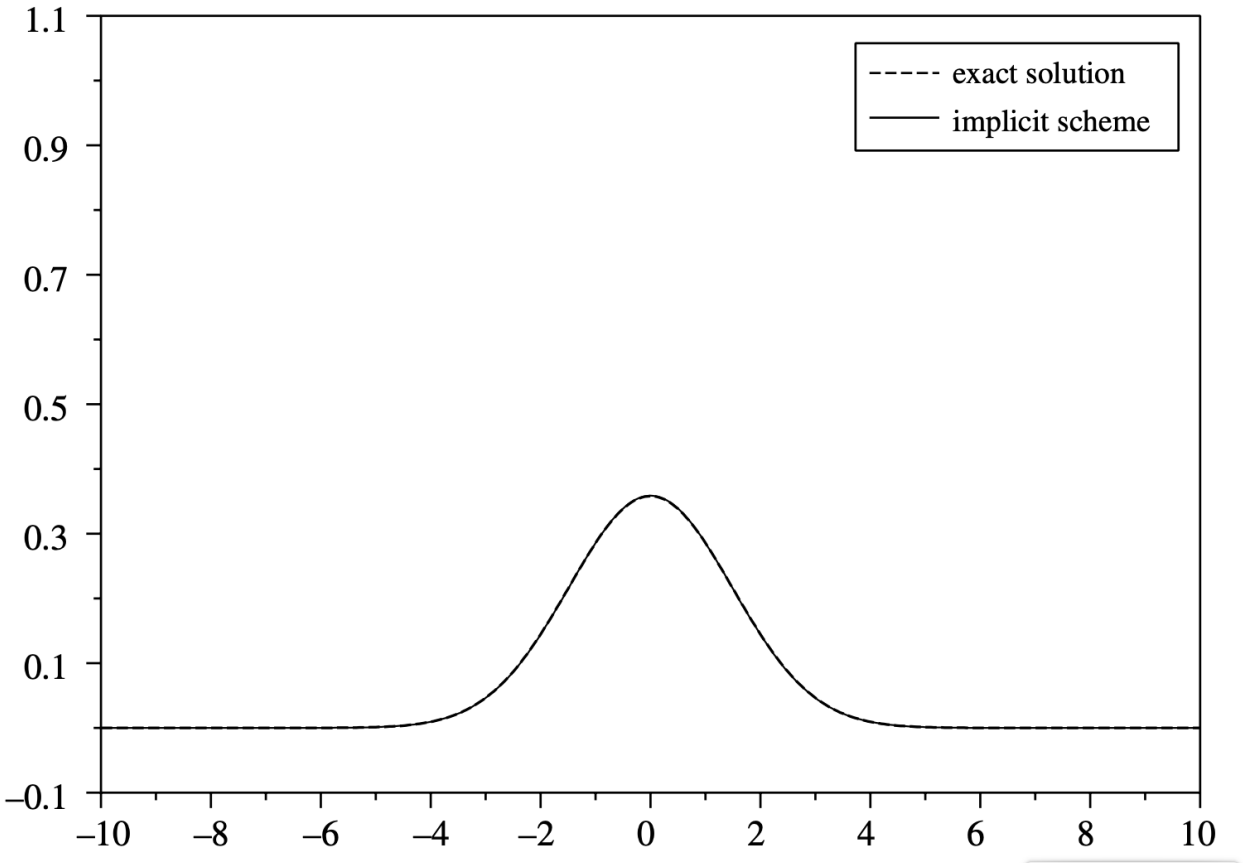
Explicit scheme with $\nu \Delta t = 0.51 \Delta x^2$

Heat equation, explicit scheme, CFL=$\frac{\nu \Delta t}{\Delta x^2} = 0.4$ Final time T=1



Explicit scheme with $\nu \Delta t = 0.4 \Delta x^2$ for various Δx

Heat equation, implicit scheme, CFL=$\frac{\nu \Delta t}{\Delta x^2} = 2$ 200 Time steps



Implicit scheme with $\nu \Delta t = 2\Delta x^2$

Some Definitions

Definition

A general finite difference scheme is formally defined by

$$F_{\Delta t,\Delta x}\left(\left\{u_{j+k}^{n+m}\right\}_{m^-\leq m\leq m^+,k^-\leq k\leq k^+}\right)=0$$

where the integers m^-, m^+, k^-, k^+ define the **width of the stencil**

Remark

To u_j^n we can associate au couple (n,j), called point. The set of the couples (n,j) which appear in the discrete equation at the point (n,j) is called the **stencil of the scheme**

Definition

A FDS is linear if $F_{\Delta t,\Delta x}\left(\left\{u_{j+k}^{n+m}\right\}\right)=0$ is linear with respect to u_{j+k}^{n+m}

Consistency

Definition

The FDS is **consistent** with the PDE defined by $F(u) = 0$
if for every sufficiently regular solution $u(t,x)$ of this equation,
the **truncation error** of the scheme, defined by

$$F_{\Delta t,\Delta x}\left(\{u(t+m\Delta t, x+k\Delta x)\}_{m^-\le m\le m^+, k^-\le k\le k^+}\right)$$

tends to zero, uniformly with respect to (t,x), as Δt and Δx tend to zero independently

Remark

In practice, we calculate the truncation error of a scheme by
replacing u_{j+k}^{n+m} in the scheme by $u(t+m\Delta t, x+k\Delta x)$

Accuracy

Definition

The scheme has an

accuracy of order p in space and of order q in time

if the truncation error tends to zero as $\mathcal{O}\left((\Delta x)^p + (\Delta t)^q\right)$ when Δt and Δx tend to zero

Exercice

- The **explicit scheme** is consistent, accurate with an order 1 in time and an order 2 in space
- If we keep the ratio $\nu \frac{\Delta t}{\Delta x^2} = \frac{1}{6}$ constant, then this scheme is accurate with an order 2 in time and 4 in space

Summary

The notion of **stability** will be presented later, but traduce the fact that the numerical solution has a **good behavior** like

no oscillation, bounded values $\cdots$

Scheme	Truncation Error	Stability
Explicit	$\mathcal{O}(\Delta t + \Delta x^2)$	Stable in L^2 and L^∞ if $2\nu\Delta t \leq \Delta x^2$
Implicit	$\mathcal{O}(\Delta t + \Delta x^2)$	Stable in L^2 and L^∞
Crank-Nicolson ($\theta = \frac{1}{2}$)	$\mathcal{O}(\Delta t^2 + \Delta x^2)$	Stable in L^2 Stable L^∞ if $\nu\Delta t \leq \Delta x^2$
θ-scheme ($\theta \neq \frac{1}{2}$)	$\mathcal{O}(\Delta t + \Delta x^2)$	St.L^2 if $2(1-2\theta)\nu\Delta t \leq \Delta x^2$
DuFort-Frankel	$\mathcal{O}(\frac{\Delta t}{\Delta x^2} + \Delta x^2)$	St.L^2 if $\Delta t/\Delta x^2$ bounded Stable L^∞ if $2\nu\Delta t \leq \Delta x^2$

L^p Norm

Definition

We consider the norm on $\mathbb{R}^N$ for the numerical solution

$$\|u^n\|_p = \left(\sum_{j=1}^{N} \Delta x |u_j^n|^p \right)^{\frac{1}{p}} \quad \text{for } 1 \leq p \leq +\infty$$

When $p = +\infty$ we consider that it is equal to $\|u^n\|_\infty = \max_{1 \leq j \leq N} |u_j^n|$

Remark

- This norm is indentical to the $L^p(0,1)$ norm for piecewise constant functions over the subintervals $[x_j, x_{j+1}[\subset [0,1] \rightsquigarrow \left(\int_0^1 u(x)^2 dx \right)^{\frac{1}{2}}$
- In practice we use $p = 2$ and $p = +\infty$

Stability and Fourier analysis

Definition

- A FDS is stable for the norm $\|.\|$ if there exists a constant $K > 0$ independent on Δt and Δx such that for arbitrary initial data u^0

$$\|u^n\| \leq K\|u^0\| \quad \text{for all } n \geq 0$$

- If this relation only hold for steps Δt and Δx defined under certain inequalities, we say that the scheme is

conditionally stable

- If we use $\|.\|_2$ we call it L^2 stability
  If we use $\|.\|_\infty$ we call it L^∞ stability

Stability of linear scheme

- A two level linear scheme can be written under the form

$$u^{n+1} = Au^n$$

where A is a linear operator (the **iteration matrix**) from $\mathbb{R}^N$ into $\mathbb{R}^N$

- So $u^n = A^n u^0$ and the **stability** is

$$\boxed{\|A^n u^0\| \leq K\|u^0\| \quad \forall n \geq 0,\ \forall u^0 \in \mathbb{R}^N}$$

- If we introduce the norm of the matrix defined by

$$\|M\| = \sup_{u\in\mathbb{R}^N, u\neq 0} \frac{\|Mu\|}{\|u\|}$$

the stability is reduced to

$$\|A^n\| \leq K \quad \forall n \geq 0$$

Stability in L^∞ norm

Definition

A FDS satisfies a **discret maximum principle** if $\forall n \geq 0$ and $1 \leq j \leq N$

$$\min_{0 \leq j \leq N+1} u_j^0 \leq u_j^n \leq \max_{0 \leq j \leq N+1} u_j^0 \quad \text{for arbitrary initial data } u^0$$

Exercice

- The **explicit scheme** satisfies a discret maximum principle
- The explicit scheme is **stable** in the L^∞ norm **if and only if**

  the condition $\boxed{2\nu\Delta t \leq \Delta x^2}$ is satisfied
- This condition is called **CFL condition**
  CFL for "Courant Friedrichs Lewy" (famous searchers)
- The **implicit scheme** is **stable** in the L^∞ norm for any time step Δt and any space step Δx. We say : it is **unconditionally stable**

Exercice

1. Discret maximum principle : the explicit scheme can be written under the form

$$u_j^{n+1} = \frac{\nu \Delta t}{\Delta x^2} u_{j-1}^n + \left(1 - 2\frac{\nu \Delta t}{\Delta x^2}\right) u_j^n + \frac{\nu \Delta t}{\Delta x^2} u_{j+1}^n$$

– So under the CFL condition, u_j^{n+1} is a convex linear combination of u_{j-1}^n, u_j^n, u_{j+1}^n (all terms **are** $> \mathbf{0}$ (CFL) and the sum $= 1$) in time t^n
– In particular by recurrence we can prove that

$$m \leq u_j^0 \leq M, \ \forall j \in \mathbb{Z} \Longrightarrow m \leq u_j^n \leq M, \ \forall j \in \mathbb{Z} \text{ and } n \geq 0$$

2. L^∞ stability : straightforward from the previous question
The CFL condition is required and is **optimal**.

Indeed, without CFL and for $u_j^0 = (-1)^j$ then $u_j^n = (-1)^j \left(\underbrace{1 - 4\frac{\nu \Delta t}{\Delta x^2}}_{<-1} \right)^n$ so $\lim_{n \longrightarrow \infty} |u_j^n| = \infty$ and the scheme is unstable

Stability in L^2 norm

- Many schemes do not satisfy the discrete maximum principle but are nevertheless "good" schemes $\rightsquigarrow$ **L^2 stability notion**
- We assume that the boundary conditions for the heat equation are periodic : $u(t, x+1) = u(t, x)$ for all $x \in [0,1]$ and $t \geq 0$. Then

$$u_j^n = u_{N+1+j}^n \quad \text{for all } n \geq 0$$

- We have to calculate $N+1$ values u_j^n
- For each vector $u_n = (u_j^n)_{0 \leq j \leq N}$ we associate a function $u^n(x)$, piecewise constant, periodic with period 1, defined on $[0,1]$ by

$$u^n(x) = u_j^n \quad \text{if } x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}}$$

with $x_{-\frac{1}{2}} = 0$, $x_{j+\frac{1}{2}} = (j+\frac{1}{2})\Delta x$ for $0 \leq j \leq N$ and $x_{N+1+\frac{1}{2}} = 1$

- Then $u^n(x) \in L^2(0,1)$ and we can use **Fourier analysis**

Stability in L^2 norm

- From Fourier analysis we have

$$u^n(x) = \sum_{k\in\mathbb{Z}} \hat{u}^n(k) exp(2i\pi kx)$$

with $\hat{u}^n(k) = \int_0^1 u^n(x) exp(-2i\pi kx) dx$ and the Plancherel formula

$$\int_0^1 |u^n(x)|^2 dx = \sum_{k\in\mathbb{Z}} |\hat{u}^n(k)|^2$$

- If we denote by $v^n(x) = u^n(x + \Delta x)$ then

$$\boxed{\hat{v}^n(k) = \hat{u}^n(k) exp(2i\pi k\Delta x)}$$

(by changement of variable into the integral applied to periodic function)

Stability in L^2 for Explicit scheme

- We consider the scheme

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}-\nu\frac{u_{j-1}^n-2u_j^n+u_{j+1}^n}{\Delta x^2}=0$$

- We apply Fourier transform to the scheme that becomes

$$\frac{\hat{u}^{n+1}(k)-\hat{u}^n(k)}{\Delta t}-\nu\frac{\hat{u}^n(k)e^{-2i\pi k\Delta x}-2\hat{u}^n(k)+\hat{u}^n(k)e^{2i\pi k\Delta x}}{\Delta x^2}=0$$

$$\Longleftrightarrow \hat{u}^{n+1}=\left(1+\frac{\nu\Delta t}{\Delta x^2}(e^{-2i\pi k\Delta x}-2+e^{2i\pi k\Delta x})\right)\hat{u}^n(k)$$

- $\hat{u}^{n+1}=A(k)\hat{u}^n(k)=A(k)^{n+1}\hat{u}^0(k)$ with $A(k)=1-\frac{4\nu\Delta t}{\Delta x^2}(sin(\pi k\Delta x))^2$
- For $k\in\mathbb{Z}$, the Fourier coefficient $\hat{u}^n(k)$ is bounded as n tends to the infinity if and only if the **amplification factor** satisfies $|A(k)|\leq 1$

$$2\nu\Delta t(\sin(\pi k\Delta x))^2\leq(\Delta x)^2$$

Stability in L^2 for Explicit scheme

- If the CFL given by $2\nu\Delta t \leq \Delta x^2$ is satisfied then the inequality is true for every **Fourier mode** $k \in \mathbb{Z}$ and then by Plancherel formula

$$\|u^n\|_2^2 = \int_0^1 |u^n(x)|^2 dx = \sum_{k\in\mathbb{Z}} |\hat{u}^n(k)|^2 \leq \sum_{k\in\mathbb{Z}} |\hat{u}^0(k)|^2 = \int_0^1 |u^0(x)|^2 dx = \|u^0\|_2^2$$

  which is the L^2 stability of the explicit scheme
- If the CFL is not satisfied, the scheme is **unstable**

Remark

- Practical point of view : To prove L^2 stability of a scheme, we put Fourier modes into the scheme

$$u_j^n = A(k)^n exp(i2\pi k x_j) \quad \text{with } x_j = j\Delta x$$

  and we deduce the value of the amplification factor $A(k)$
- The inequality $|A(k)| \leq 1$ for all modes $k \in \mathbb{Z}$ is called the **Von Neumann stability condition**

Stability in L^2 for Implicit scheme

- We consider the scheme

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}-\nu\frac{u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1}}{\Delta x^2}=0$$

- We apply Fourier transform to the scheme that becomes

$$\hat{u}^{n+1}\left(1+\frac{\nu\Delta t}{\Delta x^2}(-exp(-2i\pi k\Delta x)+2-exp(2i\pi k\Delta x))\right)=\hat{u}^n(k)$$

- $\hat{u}^{n+1}=A(k)\hat{u}^n(k)=A(k)^{n+1}\hat{u}^0(k)$ with $A(k)=\left(1+\frac{4\nu\Delta t}{\Delta x^2}(sin(\pi k\Delta x))^2\right)^{-1}$
- For $k\in\mathbb{Z}$, the Fourier coefficient $\hat{u}^n(k)$ is bounded as n tends to the infinity because the **amplification factor** satisfies $|A(k)|\leq 1$

Theorem

The implicit scheme is **unconditionally stable** in L^2 norm

Stability in L^2 for θ-scheme

Exercice

Prove the properties of the θ-scheme given by

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\theta\nu-\frac{u_{j-1}^{n+1}+2u_j^{n+1}-u_{j+1}^{n+1}}{\Delta x^2}+(1-\theta)\nu\frac{-u_{j-1}^{n+1}+2u_j^{n+1}-u_{j+1}^{n+1}}{\Delta x^2}=0$$

If $0\leq\theta<\frac{1}{2}$

the scheme is stable in the L^2 norm under the CFL condition

$$2(1-2\theta)\leq\Delta x^2$$

If $\frac{1}{2}\leq\theta\leq 1$

the scheme is unconditionnally stable in the L^2 norm

Convergence of the scheme : Lax theorem

Theorem

- We assume that the scheme is **linear**, **two level**, **consistent**, and **stable** for a norm $\|.\|$
- Then the scheme is **convergent** in the sense that

$$\forall T > 0, \lim_{\Delta t, \Delta x \longrightarrow 0} \left(\sup_{t_n \leq T} \|e^n\| \right) = 0$$

for e^n the **error vector** defined by $e_j^n = u_j^n - u(t^n, x_j)$

- If the scheme has accuracy of order p in space and order q in time, then for all time $T > 0$ there exists a constant $C_T > 0$ such that

$$\sup_{t^n \leq T} \|e^n\| \leq C_T \left((\Delta x)^p + (\Delta t)^q \right)$$

Remark : The idea of the proof is similar to the proof of the convergence for the scheme applied to Laplace equation

The Transport equation

- We consider $\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x},\ u(x,0) = u_0(x),\ x \in \mathbb{R},\ t \geq 0$
- We know the solution $u(x,t) = u_0(x - ct)$
- Some properties to preserve
  - the solution satisfies the maximum principle

$$\inf_{y\in\mathbb{R}} u_0(y) \leq u(x,t) \leq \sup_{y\in\mathbb{R}} u_0(y),\ \forall x \in \mathbb{R},\ t \geq 0$$

  - the Total Variation (TV) of the solution is constant

$$TV(v(.,t)) = \int_{\mathbb{R}} \left|\frac{\partial u}{\partial x}(x,t)\right| dx = \int_{\mathbb{R}} |u_0'(x)| dx = TV(u_0) = \text{Constant}$$

  - For nonlinear PDE the Total Variation is not constant but Decreases we talk about (TVD) property

Numerical illustration

- We will present classical schemes to solve the transport PDE
- We compute the solution for the final time $T = 0.2s$
  - We consider 200 points for the mesh on $\Omega = [0;1]$
    - ★ with **Neumann B.C.** $u_0^n = v_1^n, \quad v_{N+1}^n = v_N^n$
      (that corresponds to have possible waves that go outside the domain)
  - We fix $c = 1$ and the **CFL=**$\dfrac{c\Delta t}{\Delta x} = 0.6$ (becareful here CFL $\neq \frac{\nu \Delta t}{\Delta x^2}$)

Test 1 smooth (regular) initial condition (C^2)

$$u^0(x) = \begin{cases} 16\sin\left(4\pi \frac{x-0.375}{0.25}\right) \frac{(x-0.375)^2(0.625-x)^2}{0.25^2} & \text{for } x \in [0.375;0.65] \\ 0 \text{ elsewhere} \end{cases}$$

Test 2 discontinuous initial condition : $u^0(x) = \begin{cases} 1 \text{ for } x < 0.5 \\ 0 \text{ for } x > 0.5 \end{cases}$

The explicit scheme

Theorem

- We consider the **explicit scheme** (centred in space)

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0$$

- The scheme is **unstable** for the L^2 norm
- It is consistent, of order 1 in space and 1 in time

Proof

- The amplification factor is $A(k) = 1 - ic\dfrac{\Delta t}{\Delta x}\sin(2\pi k\Delta x) \in \mathbb{C}$
- $|A(k)| > 1$ as soon as $\sin(2\pi k\Delta x) \neq 0$

Crank-Nicholson scheme (explicit scheme) second order

Exercice

- We consider the **Crank-Nicholson scheme** (centred in space)

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}+\frac{c\Delta t}{\Delta x}\left(\frac{u_{i+1}^n+u_{i+1}^{n+1}}{2}-\frac{u_{i-1}^n+u_{i-1}^{n+1}}{2}\right)=0$$

- Study the stability of the scheme ?
- Study the accuracy of the scheme ?

The implicit scheme

Theorem

- We consider the **implicit centred scheme**

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_{i+1}^{n+1} - u_{i-1}^{n+1}}{2\Delta x} = 0$$

- The scheme is **stable** for the L^2 norm
- It is consistent, of order 2 in space and 1 in time

Remark

Considering implicit scheme instead of explicit has transformed unstable scheme into stable scheme

Explicit but Uncentred scheme

Theorem

- We consider the **explicit uncentred scheme** when $c > 0$

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_i^n - u_{i-1}^n}{\Delta x} = 0$$

- The scheme is **stable** for the L^2 norm under the CFL

$$c\frac{\Delta t}{\Delta x} \leq 1$$

- It is L^∞ stable if $c > 0$ and under the same CFL (convex linear combination)
- It is consistent, of order 1 in space and 1 in time

Remark

We select the **Backward** uncentred scheme as soon as $c > 0$

Backward/Forward scheme

Physical remarks :

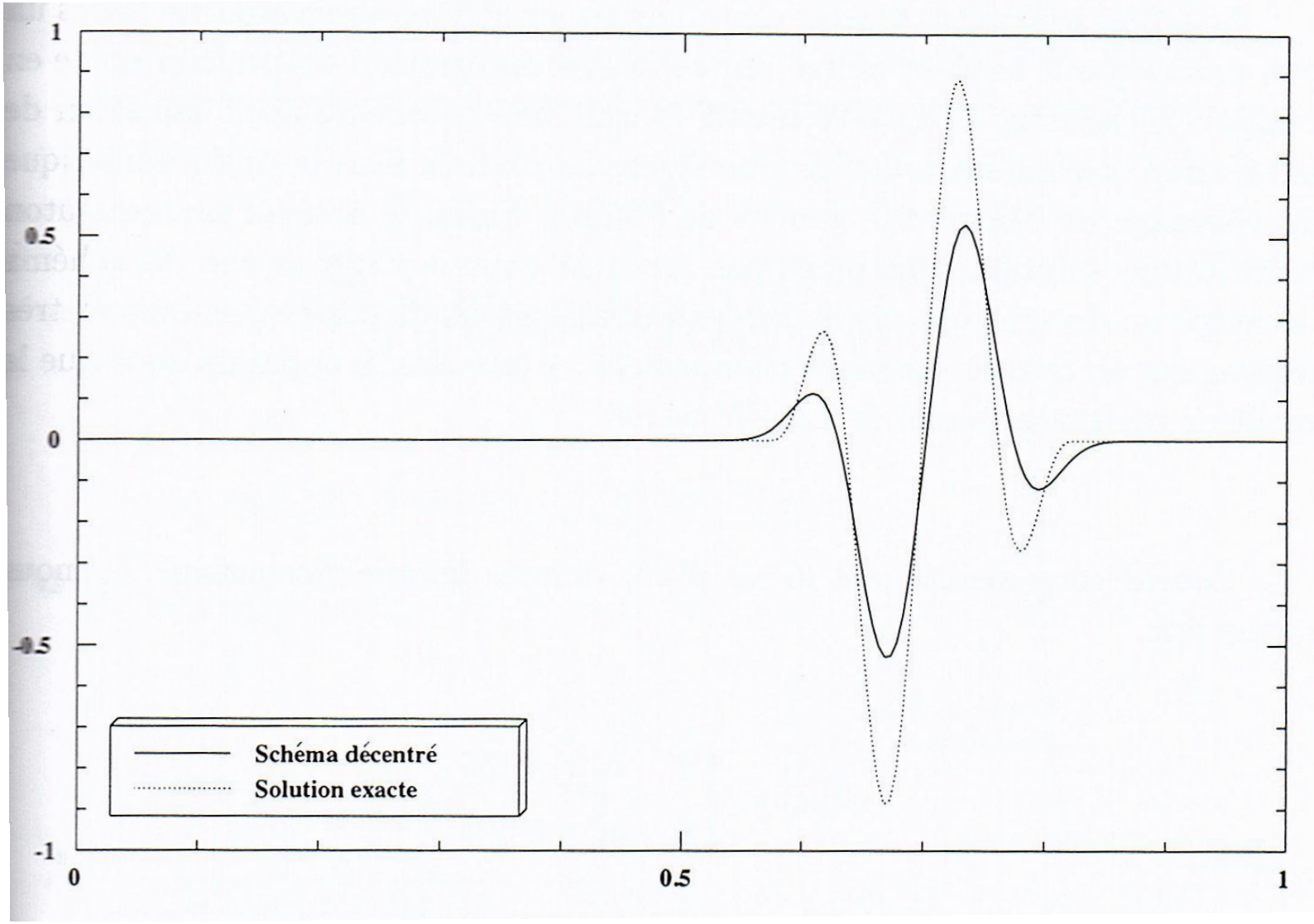
- The stability condition suggests that u_i^{n+1} must be computed by convex linear combination of u_i^n and u_{i-1}^n
  - the information is **picked up from its origine**
    (for $c > 0$: the solution moves from the left to the right – see ch.2)
  - the numerical velocity of the propagation $\frac{\Delta x}{\Delta t}$ must be greater or equal to the physical velocity of the propagation given by c

Remark

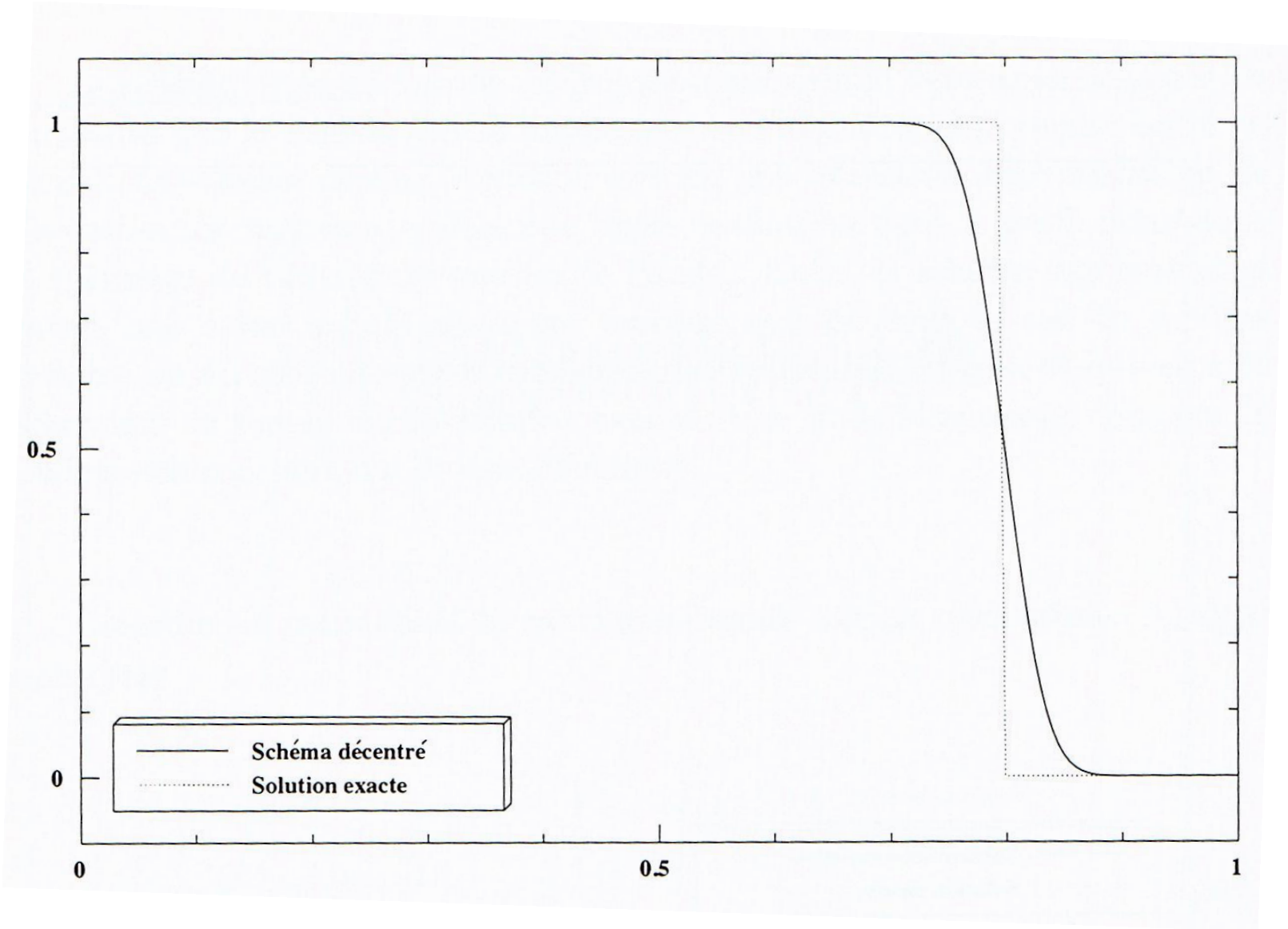
We select the **Forward** uncentred scheme as soon as $c < 0$

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_{i+1}^n - u_i^n}{\Delta x} = 0$$

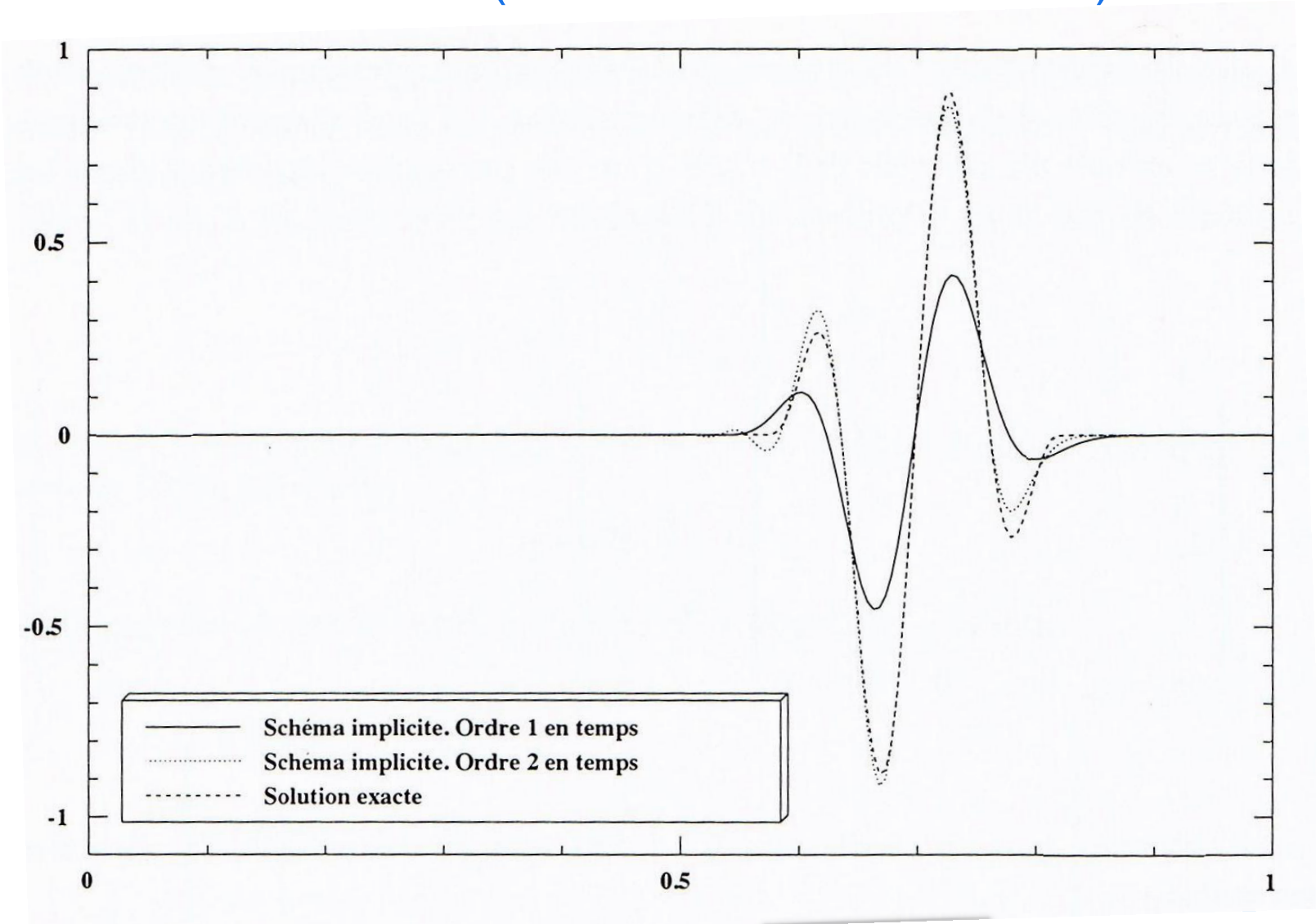
Uncentred scheme (Test 1)



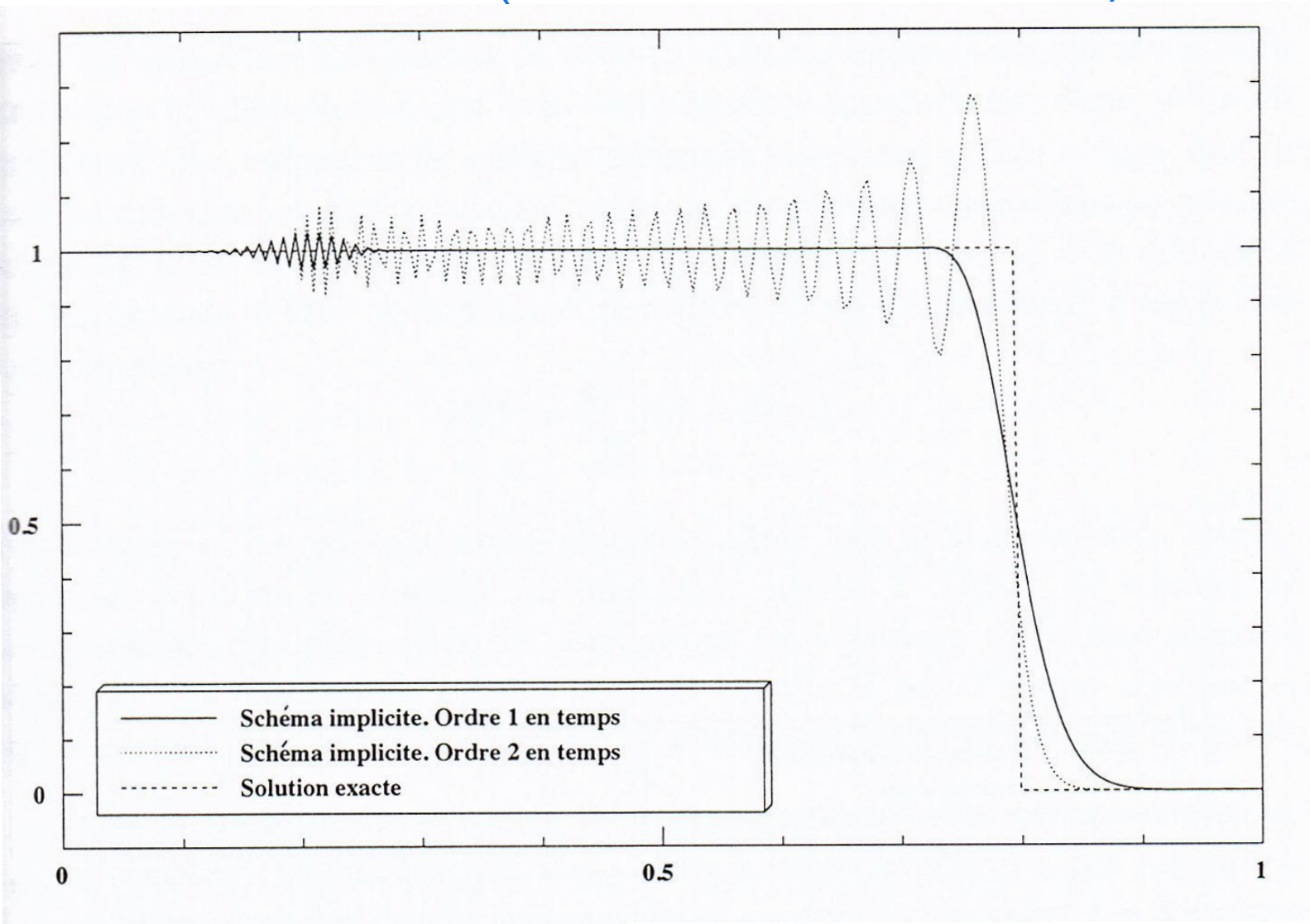
Uncentred scheme (Test 2)



(Test 1) – Implicit (first order in time) Crank-Nicholson (second order in time) schemes



(Test 2) – Implicit (first order in time) Crank-Nicholson (second order in time) schemes



Leep-Frog-scheme

Theorem

- We consider the **Leep-Frog scheme**

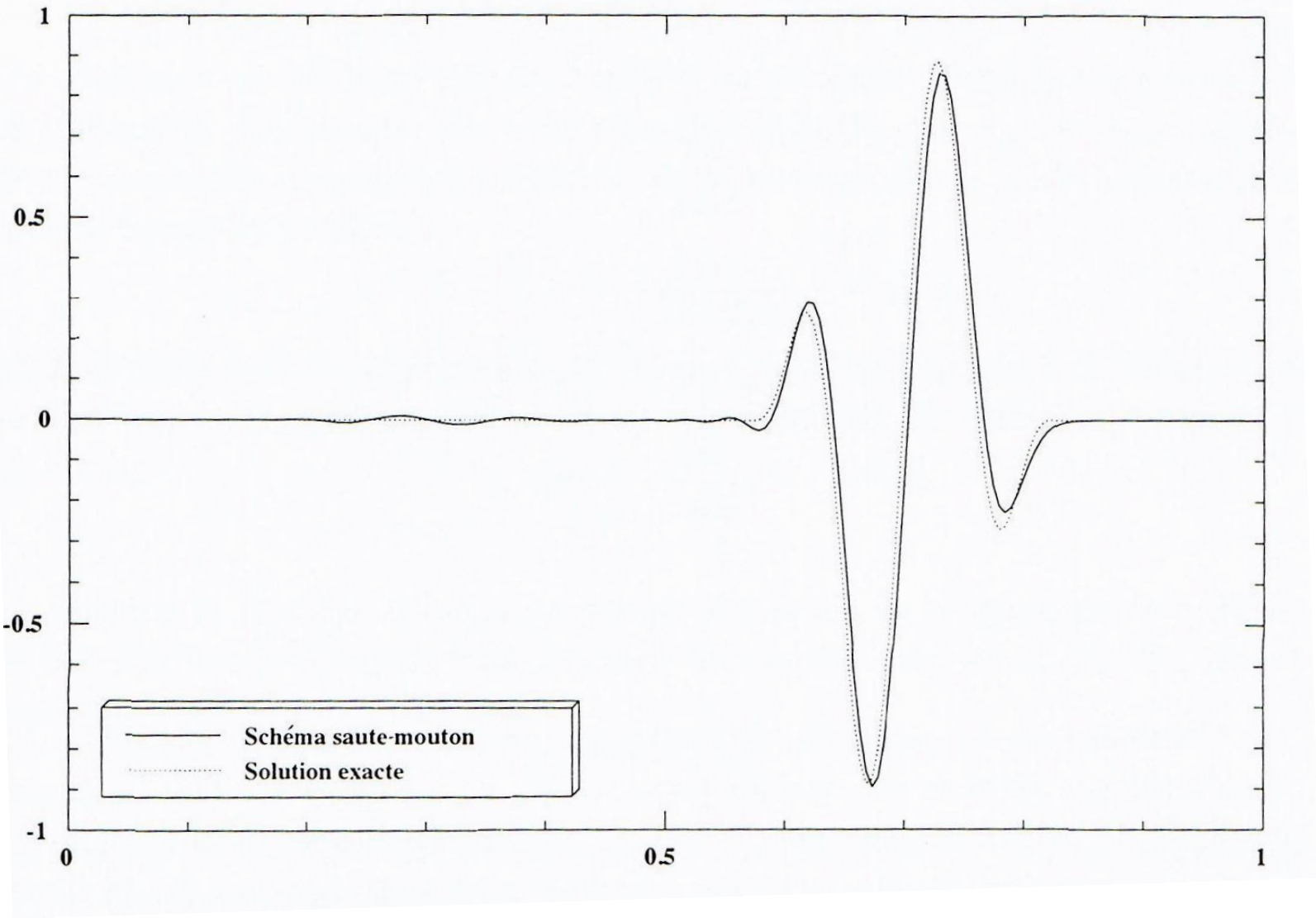
$$\frac{u_i^{n+1} - u_i^{n-1}}{2\Delta t} + c\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0$$

- The scheme is **stable** for the L^2 norm under the CFL

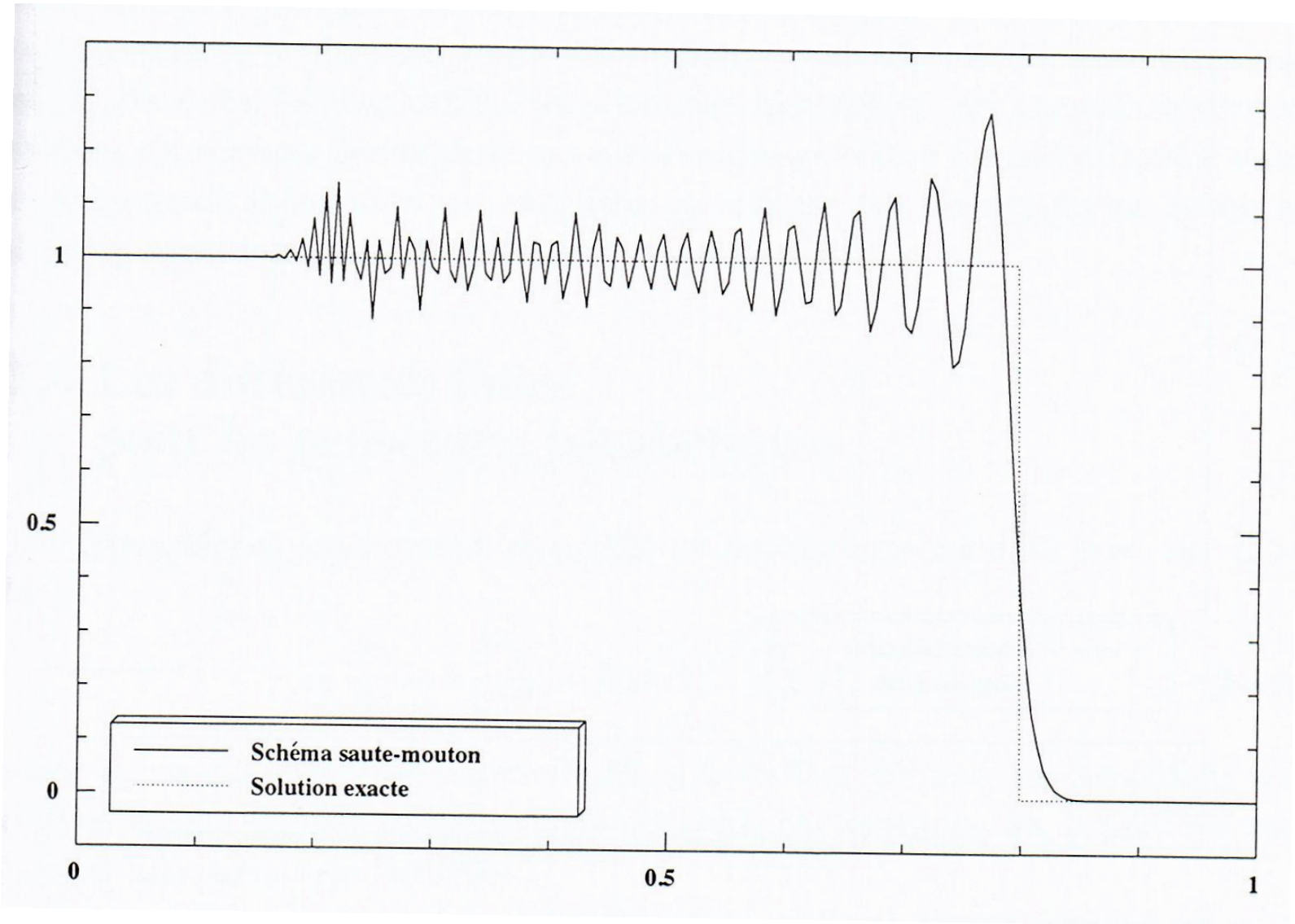
$$c\frac{\Delta t}{\Delta x} \leq 1$$

- It is consistent, of **order 2 in space and 2 in time**

Leep-Frog-scheme (Test 1)



Leep-Frog-scheme (Test 2)



Lax-Friedrichs scheme

Theorem

- We consider the **Lax-Friedrichs scheme**

$$\frac{u_i^{n+1} - \frac{u_{i-1}^n + u_{i+1}^n}{2}}{\Delta t} + c\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0$$

- The scheme is **stable** for the L^2 norm under the CFL

$$c\frac{\Delta t}{\Delta x} \leq 1$$

- It is consistent, of order 2 in space and 1 in time
- It satisfies the discret maximum principle
- $TV^n = \sum_{j\in\mathbb{Z}} |u_j^n - u_{j-1}^n|$ decreases $TV^{n+1} \leq TV^n$

Lax-Wendroff scheme

Theorem

- We consider the **Lax-Wendroff scheme**

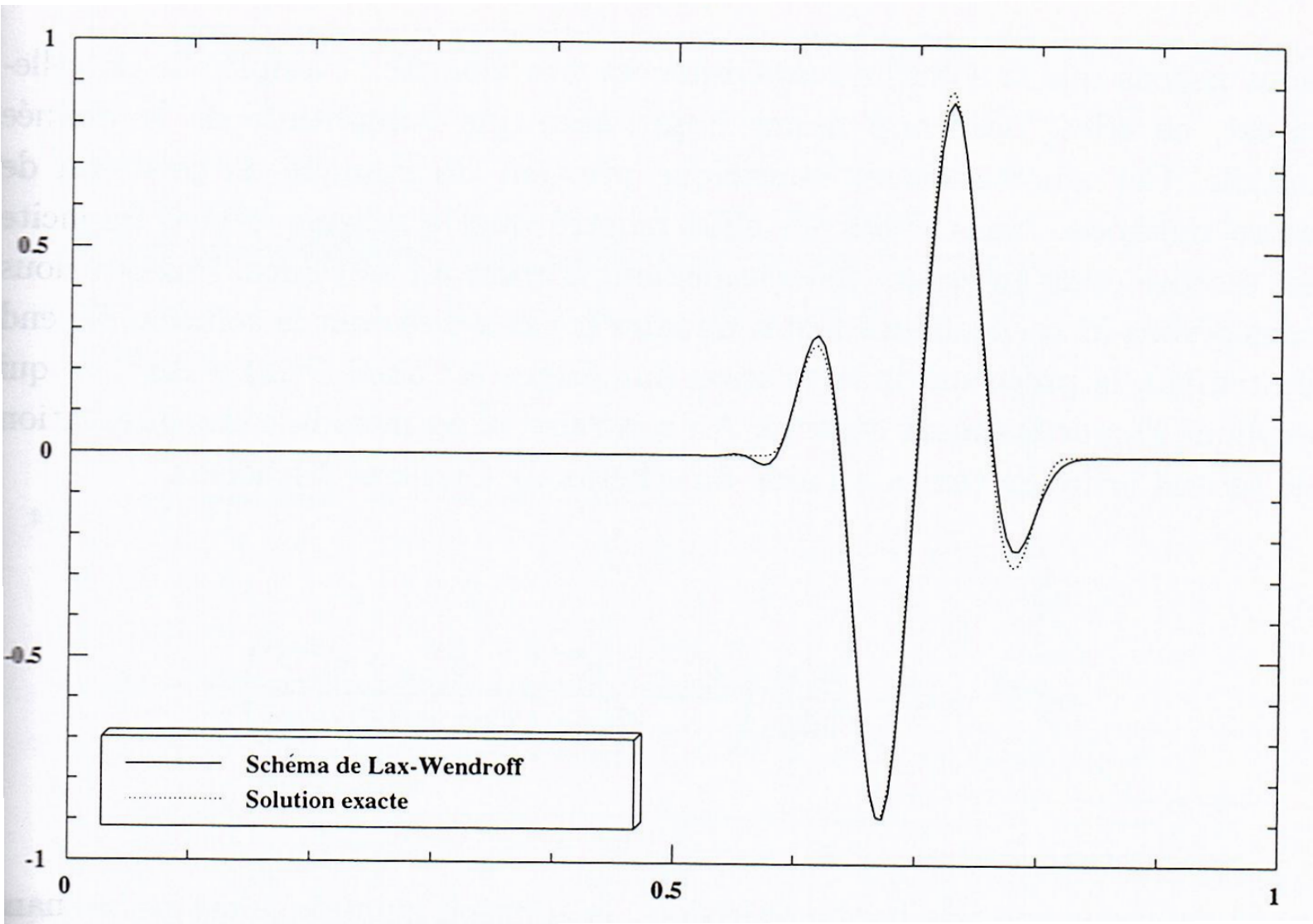
$$\frac{u_i^{n+1}-u_i^n}{\Delta t}+c\frac{u_{i+1}^n-u_{i-1}^n}{2\Delta x}-c^2\frac{\Delta t}{2}\frac{u_{i+1}^n-2u_i^n+u_{i-1}^n}{\Delta x^2}=0$$

- The scheme is **stable** for the L^2 norm under the CFL

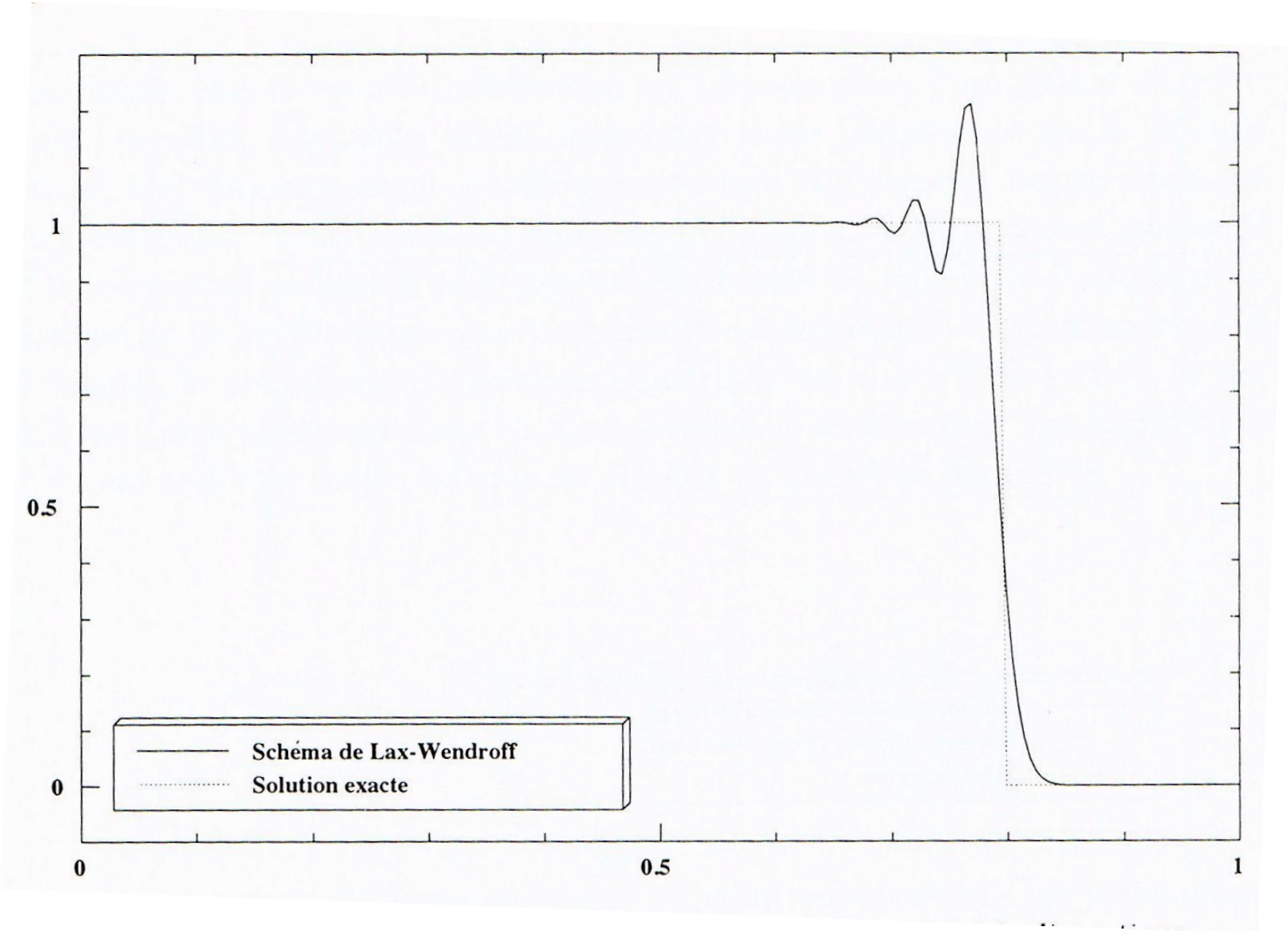
$$c\frac{\Delta t}{\Delta x}\leq 1$$

- It is consistent, of order 2 in space and 2 in time
- The Total Variation is not preserve

Lax-Wendroff scheme (Test 1)



Lax-Wendroff scheme (Test 2)



Boundary Condition (order 1)

- We consider the Elliptic case : Laplace PDE

$$\begin{cases} -u''(x) = f(x), \quad \text{for} \quad x \in]0;1[\\ u(0) = 0 \quad \text{and} \quad u'(1) = 0 \end{cases} \longrightarrow \begin{cases} \dfrac{-u_{i+1} + 2u_i - u_{i-1}}{\Delta x^2} = f(x_i) \\ \text{for} \quad 1 \leq i \leq N+1 \end{cases}$$

- For $i = 0$ We fix $u_0 = 0$ so u_0 is not an unknown
- For $i = 1$ the scheme becomes $\dfrac{2u_1 - u_2}{\Delta x^2} = f(x_1)$
- For $i = N+1$ From $u'(1) = 0$ ($\frac{u_{N+2} - u_{N+1}}{\Delta x} = 0$) we eliminate u_{N+2} into the scheme using $u_{N+2} = u_{N+1}$ then $\dfrac{u_{N+1} - u_N}{\Delta x^2} = f(1)$
- Final scheme 1 :

$$\begin{cases} \dfrac{2u_1 - u_2}{\Delta x^2} & = & f_1 \\ \dfrac{-u_{i+1} + 2u_i - u_{i-1}}{\Delta x^2} & = & f_i, \ 2 \leq i \leq N \\ \dfrac{u_{N+1} - u_N}{\Delta x^2} & = & f_{N+1} \end{cases}$$

Boundary Condition (order 2)

- The **scheme 1** is based on the the approximation of $u'(1) = 0$ by

$$\frac{u_{N+2} - u_{N+1}}{\Delta x} = 0$$

that introduce an error of order 1

- If we replace this approximation by

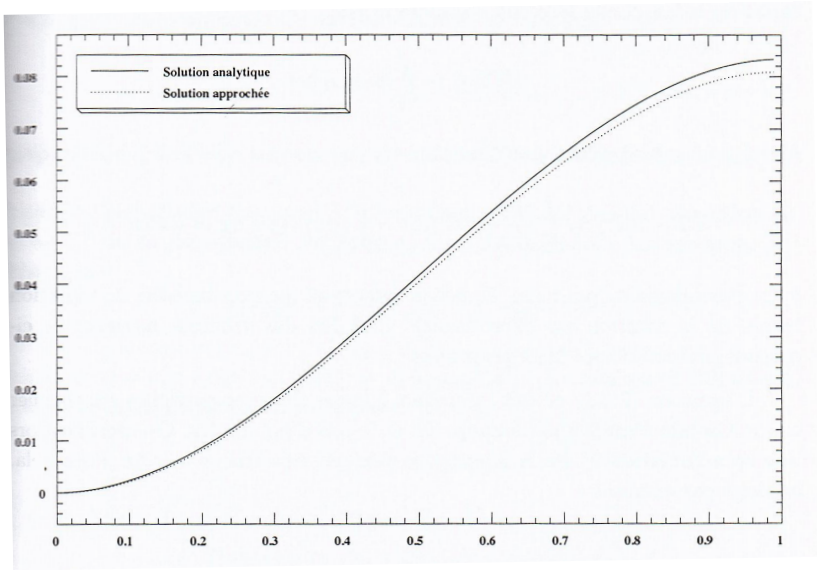
$$u_{N+2} = u_{N+1} + \frac{\Delta x^2}{2} u''(1) = u_{N+1} - \frac{\Delta x^2}{2} f(1)$$

which is an approximation of order 2 supposing that u is solution of the PDE we obtain a new scheme called **scheme 2**

  - the right hand side of the system is modified

Numerical illustration

- We consider $f(x) = x$ and then $u_{exact}(x) = -x^3/6 + x/2$
- Comparison between scheme 1 and exact solution ($N = 50$)



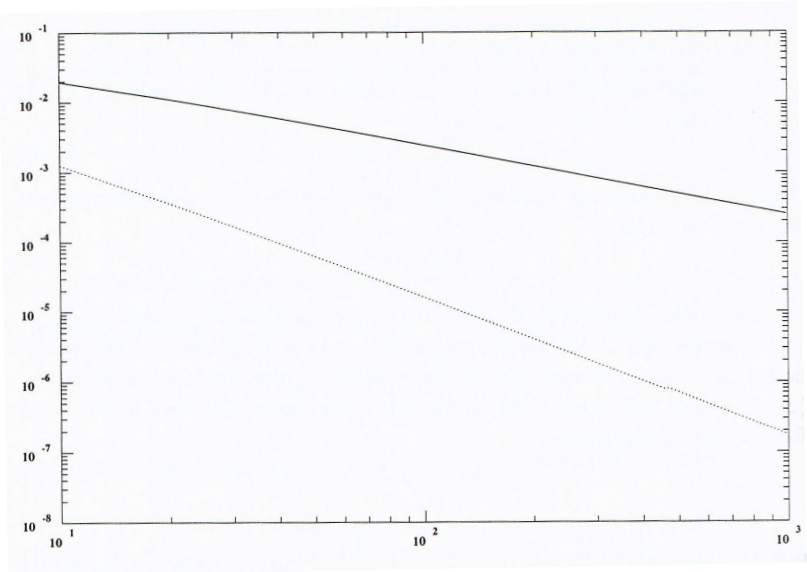
- the error coming from B.C. is propagated all over the domain

Remark

Even if the scheme is of order 2, approximation of order 1 at the boundary implies a global first order

Numerical illustration

- Comparison between scheme 1 and scheme 2
- We plot log(error) versus log(N) for different number of nodes N (corresponding to different step size Δx)



- For the scheme 1 : the slope is equal to $-0.97 \longrightarrow$ order 1
- For the scheme 2 : the slope is equal to $-1.9 \longrightarrow$ order 2

Theoretical results

Theorem

For $E = U - U_{exact}$ the error vector, we can prove that

- For scheme 1 : $|E|_\infty \leq C\Delta x$, so the scheme is of order 1
- For scheme 2 : $|E|_\infty \leq C\Delta x^2$, so the scheme is of order 2

where C is a constant depending only on f, u and the derivates of u

Remark

The scheme 1 is not consistent.
Indeed, the truncation error near the right boundary condition doesn't tend to 0 when Δx tends to 0

Extension to the multidimensional case

- We consider $\Omega = [0,1] \times [0,1]$ and the Heat Equation in 2D

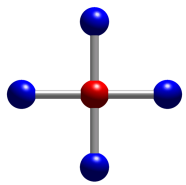
$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \left(\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} \right) = 0, & \text{for } (x,y,t) \in \Omega \times \mathbb{R}_*^+ \\ u(t=0,x,y) = u_0(x,y), & \text{for } (x,y) \in \Omega \\ u(t,x,y) = 0, & \text{for } t \in \mathbb{R}_*^+, \quad (x,y) \in \partial\Omega \end{cases}$$

- Natural notations (as in 1D) : Δt, Δx, Δy, $u_{j,k}^n \approx u(t^n, x_j, y_k)$

$$\frac{u_{j,k}^{n+1} - u_{j,k}^n}{\Delta t} - \nu \frac{u_{j+1,k}^n - 2u_{j,k}^n + u_{j-1,k}^n}{\Delta x^2} - \nu \frac{u_{j,k+1}^n - 2u_{j,k}^n + u_{j,k-1}^n}{\Delta y^2} = 0$$

for $n \geq 0$, $1 \leq j \leq N_x$ $1 \leq k \leq N_y$

- 5 points stencil



Extension to the multidimensional case

Theorem

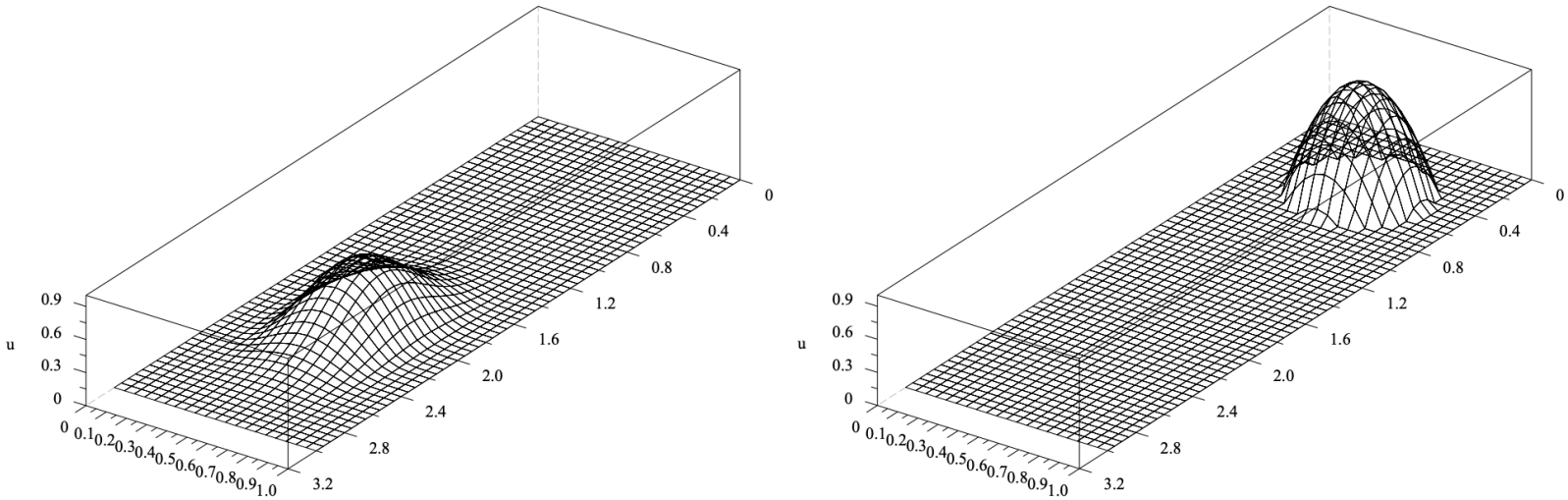
The explicit scheme in 2D is L^∞ stable under the CFL condition

$$\frac{\nu \Delta t}{\Delta x^2} + \frac{\nu \Delta t}{\Delta y^2} \leq \frac{1}{2}$$

Remark

- The CFL condition in 2D is more restrictive than the CFL in 1D
- We can consider extension for
  - the dimension 1D, 2D, 3D
  - for the modelization (PDE) : convection-diffusion $\cdots$

Simulation of the explicit scheme adding a convection term



Convection-Diffusion PDE for $\nu = 0.01$, $C = (1;0)$ convection coefficient

Initial condition (left) Final Time = 1.5 after 74 time steps (right)

Equivalent equation

Definition

The **equivalent equation** of a scheme is the equation obtained by adding the principal part of the truncation error to the model studied (that is, the term with dominant order)

Exercice

Lax-Friedrichs : the principle part of its truncation error is

$$-\frac{\Delta x^2}{2\Delta t}\left(1-\frac{(c\Delta t)^2}{2\Delta x^2}\right)u_{xx}$$

then the equivalent equation of the Lax-Friedrichs scheme is

$$\frac{\partial u}{\partial t}+c\frac{\partial u}{\partial x}-\nu\frac{\partial^2 u}{\partial x^2}=0 \quad \text{with} \quad \nu=-\frac{\Delta x^2}{2\Delta t}\left(1-\frac{(c\Delta t)^2}{2\Delta x^2}\right)$$

Diffusive scheme

Theorem

- Lax-Friedrichs scheme is a good approximation (of order 2) of the equivalent equation where the coefficient of diffusion ν is small
- If Δt is very small, ν may be very large and the scheme is bad as it is too weighted to the diffusion
- The coefficient of diffusion ν of the equivalent equation is called

  numerical diffusion

  If it is large, we say that the **scheme is diffusive (or dissipative)**
- The **typical behavior** of a diffusive scheme is its tendency to artificially spread out the initial data in the course of time
  The schemes which are too diffusive are therefore **bad schemes**

Dispersive scheme

Remark

If we add the principal part of the truncation error of a scheme to the equation, then this scheme is not only consistent with this new **equivalent equation**, but is also strictly more accurate for this equivalent equation.

Theorem

- The equivalent equation of the Lax-Wendroff scheme does not contain a diffusion term but third order term, called **dispersive**
- The **typical behavior** of a dispersive scheme is that it produces oscillations when the solution is discontinuous

Remark

- the coefficient of this dispersive term can be neglected comparing to the coefficient of diffusion for the diffusive schemes

Linear system - Sparse matrix - storage

- When a matrix has a lot of zeros it is a sparse matrix

- There are different ways for its storage
  (CSR) Compressed Spare Row
  (CSC) Compressed Sparse Column
  (replace the rows by columns columns in CSR)
  (MSR) Modified Sparse Row
  (base on the diagonal - not presented)

- We use 1 real vectors (for the coefficients of the matrix) and 2 integer vectors (for position)

CSR format

- We consider the matrix

$$\begin{pmatrix} 1. & 0 & 0 & 2 & 0 \\ 3. & 4. & 0 & 5. & 0 \\ 6. & 0 & 7. & 8. & 9. \\ 0 & 0 & 10. & 11. & 0 \\ 0 & 0 & 0 & 0 & 12. \end{pmatrix}$$

- We save it using

AA =

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.

JA =

1	4	1	2	4	1	3	4	5	3	4	5

IA =

1	3	6	10	12	13

(last value = number of non-zero in the matrix + 1)

  - AA : coefficient a_{ij} of the matrix stored row by row
  - JA : column indices of the element a_{ij} stored in AA
  - IA : pointers of the beginning of each row in AA and JA

Introduction to the numerical linear algebra

Solve $AX = b$

Condition number

- Small variations on A or b can imply big variations on x
- We compute a criteria $C(A)$ named condition number (depending on eigenvalues of A)

$\Downarrow$

$C(A) \approx 1$: Good
$C(A)$ big : Wrong

Direct methods

- Gauss's algorithm
- We isolate an unknown
- And substitute in the system and so on
- Order in the elimination Notion of Pivot

Carl Friedrich Gauss (1777-1855)

Iterative Methods

- We write $A = M - N$

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$$

- For $x^{(0)}$ given we compute $x^{(1)}$ and so on
- When there is convergence $x^{(k+1)}$ approaches $x^{(k)}$ we obtained x

$$x^{(k+1)} \longrightarrow x$$

- Choice of M and N

Condition number

- We consider the matrix

$$\underbrace{\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}}_{b}$$

- This system has a unique solution ($\det(A) = 1$) equals to

$$x = {}^t(1, 1, 1, 1)$$

Condition number

- If b is replaced by $b' = {}^t(32.1, 22.9, 33.1, 30.9)$
  The solution $x = {}^t(1,1,1,1)$ becomes ${}^t(9.2, -12.6, 4.5, -1.1)$

  a small variation of $b \Longrightarrow$ a big variation of x

- If A is replaced by $A' = \begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.91 & 9 & 9.98 \end{pmatrix}$
  the solution $x = {}^t(1,1,1,1)$ becomes ${}^t(-81, 107, -34, 22)$

  a small variation of $A \Longrightarrow$ a big variation of x

Condition number : matrix norm

Definition

For a given matrix $\in M_n(\mathbb{R})$, we can compute its norm by

$$|||A|||_2 = |||A||| = \left(\sum_{i=1}^{n} |a_{ij}| \right)^{\frac{1}{2}}$$

Remark

There exist different matrix norm, for example

$$|||A|||_1 = \max_{j=1,\cdots,n} \sum_{i=1}^{n} |a_{ij}|$$

Condition number

Definition

The condition number of A denoted $C(A)$ is defined by

$$C(A) = |||A|||.|||A^{-1}|||$$

Theorem

- A is well conditionned if $C(A)$ is closed to 1
- A is not well conditionned if $C(A)$ is big

$$A \text{ non inversible} \longleftrightarrow C(A) = \infty$$

Theorem

If a matrix A is Hermitian ($A = A^* = {}^t\overline{A}$) and inversible then

$$C(A) = \frac{\text{Biggest eigenvalue of } A}{\text{Smallest eigenvalue of } A}$$

Preconditionner

Remark

For the example $C(A) \approx 4488$

Theorem

- To solve a system which is not well conditionned ($C(A)$ big) We search a matrix M such that $C(MA) < C(A)$
- So instead of considering $Ax = b$, we study

$$\boxed{MAx = Mb}$$

This is the preconditionning of the matrix M

Direct solvers

- Idea : We transform the system into a triangular system
- So, having A which is an upper triangular of the form

$$\begin{cases} a_{11}x_1 + \cdots \quad \cdots \quad \cdots + a_{1,n}x_n = b_1 \\ \vdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ a_{nn}x_n = b_n \end{cases}$$

- A inversible $\Longrightarrow (a_{ii} \neq 0, \forall i) \Longrightarrow x_n = b_n / a_{nn}$
  And we substitute into the others equations and so on

$$\Longrightarrow x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=i+1}^{n} a_{ij} x_j \right)$$

- The number of operations is equal to $\frac{n(n+1)}{2}$

LU decomposition

- In practice, we call some algorithms to factorize the matrix A under the form

$$\boxed{A = LU}$$

where
L : lower unit triangular matrix (1 on the diagonal)
U : upper triangular matrix

$$\boxed{Ax = b \Longrightarrow LUx = b \Longrightarrow \begin{cases} Ly = b \\ Ux = y \end{cases}}$$

- This system is easy to solve (Gauss elimination)
- One famous package is developped in Toulouse : MUMPS

Iterative solvers

- When n is big (≈ 1000) $\Longrightarrow$ in general, direct methods are not efficient $\Longrightarrow$ we prefer iterative methods
- For $Ax = b$ (A invertible), we construct a sequence of vectors starting from $x^{(0)}$ (initial vector) by using

$$x^{(k+1)} = Bx^{(k)} + C \text{ where } B \in M_n(\mathbb{R}), C \in \mathbb{R}^n$$

- Their efficiencies depend on $C(A)$
- and sometimes they do not converge (theorems)

How to define B and C ?

Classical iterative methods

- We can always decompose A into the form $A = M - N$ (M inversible) then to solve $Ax = b$ we write the iterative algorithm

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b \text{ where , } x^{(0)} \text{ given}$$

- Generally, we decompose A under the form $A = D - E - F$ where
  - D : diagonal matrix formed by the diagonal elements of A
  - E : lower tridiagonal matrix (lower elements of A)
  - F : upper tridiagonal matrix (upper elements of A)
- Under these hypothesis the sequence

$$(x^{(k)}) \longrightarrow x$$

where x is the solution of $Ax = b$

Jacobi- Gauss Seidel-Relaxation

- Jacobi's matrix

$M = D$ et $N = E + F$

$$x^{(k+1)} = D^{-1}(E+F)x^{(k)} + D^{-1}b$$

- Gauss-Seidel's matrix

$M = D - E$ et $N = F$

$$x^{(k+1)} = (D-E)^{-1}Fx^{(k)} + (D-E)^{-1}b$$

- Relaxation's method

$M = \frac{1}{w}(D - wE)$ and $N = \frac{1-w}{w}D + F$ for $w \in]0,2[$

$$x^{(k+1)} = (D-wE)^{-1}[(1-w)D + wF]x^{(k)} + w(D-wE)^{-1}b$$

Jacobi - Gauss Seidel - Relaxation

Definition

- A is strict diagonal dominant (SDD)

$$\forall i = 1, \cdots, n, \ |a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \ (\Longrightarrow A \text{ invertible})$$

- A is Symmetric Definite Positive (SDP)

$$^tXAX > 0 \quad \forall X \neq 0 \ (\Longrightarrow A \text{ invertible})$$

Theorem

- If A is SDD then Jacobi and Gauss-Seidel methods converge
- If A is SDP then Gauss-Seidel converges

Solvers for A symmetric definite positive

Even if **Gauss-Seidel**'s method converges, generally we prefer the **famous Conjugate Gradient**'s method (C.G.) which is based on the following property

Theorem

If A is symmetric definite positive then the solution of the system $AX = b$ is the minimum of

$$J(X) = \frac{1}{2} X.AX - b.X$$

Remark

- **Idea :** $x^{(k+1)} = x^{(k)} - \rho_k J'(x^{(k)})$ and there is different approaches to choose ρ_k who satisfies $J(x^{(k+1)}) < J(x^{(k)})$: fixed or optimal step
- One of the most famous iterative algorithm is GMRES (Krylov method)

A famous Iterative solver : Conjugate Gradient (C.G.)

Demonstration :

- Prove that $AX = b \Longrightarrow J(Y) \geq J(X)\ \forall Y$
  For all Y, $J(Y) = \frac{1}{2} YAY - b.X$
  $= J(X) + (Y - X).(AX - b) + \frac{1}{2}(Y - X).A(Y - X) \Longrightarrow \cdots$
- To prove that if X is the minimum of $J \Longrightarrow AX = b$
  - We have $J(X + \varepsilon Y) \geq J(X),\ \forall \varepsilon,\ \forall Y$
  - We expand then $J(X) + \varepsilon Y.(AX - b) + \frac{\varepsilon^2}{2} Y.AY \geq J(X)$
  - So $\varepsilon Y.(AX - b) + \frac{\varepsilon^2}{2} Y.AY \geq 0$
  - We choose $\varepsilon > 0$ and divide by ε so $\varepsilon Y.(AX - b) + \frac{\varepsilon}{2} Y.AY \geq 0$
  - When $\varepsilon \longrightarrow 0$ we have $Y.(AX - b) \geq 0,\ \forall Y$
    This inequality is again satisfied by $-Y$
  - So $Y.(AX - b) = 0,\ \forall Y$

And we can conclude

References

- G. Allaire and A. Craig, Numerical Analysis and Optimization. An Introduction to Mathematical Modelling and Numerical Simulation. *Oxford Science Publications, 2007*
- W. Strauss, Partial Differential Equations : an Introduction *John Wiley and Sons, 1992*
- L.C. Evans, Partial Differential Equations, *AMS, 1998*
- J.D. Logan, Applied Partial Differential Equations, *Springer, 1998*