MATH0024 – Modeling with PDEs

Finite difference and finite element methods for the heat equation

Maarten Arnst and Romain Boman

November 22, 2017

- Finite difference method:
  - Discretized equations.
  - Consistency, stability, convergence.
  - Method of lines.
  - Review of sampling theory.
  - Von Neumann stability analysis.
- Finite element method.
- Summary and conclusion.
- References.

Finite difference method

Model problem

- Let us consider the numerical approximation of the solution to the initial-boundary value problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0, & 0 < x < 1, \quad 0 < t < \tau, \\ u(0,t) = u(1,t) = 0, & \quad\quad\quad\quad\quad\; 0 < t < \tau, \quad \text{(boundary conditions)}, \\ u(x,0) = g(x), & 0 < x < 1, \quad\quad\quad\quad\quad\quad \text{(initial condition)}. \end{cases}$$

- In the spatial domain, let grid points $x_0,\ x_1,\ x_2,\ \dots,\ x_{\mu_h}$ be introduced as follows:

The grid spacing is denoted by h; thus, $x_j = jh$ for $j = 0, \dots, \mu_h$ with $\mu_h = 1/h$.

- In the time domain, let approximations be computed at successive times $t_0, t_1, t_2, \dots, t_{\nu_k}$. The time step is denoted by k; thus, $t_n = nk$ for $n = 0, \dots, \nu_k$ with $\nu_k = \tau/k$.

- System of notation: numerical solution u_j^n approximates exact solution $u(x_j, t_n)$ at (x_j, t_n).

Centered-in-space forward-in-time method

- The centered-in-space forward-in-time method is obtained by requiring that

$$\begin{cases} \dfrac{u_j^{n+1}-u_j^n}{k} - \dfrac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2} = 0, & 1 \leq j \leq \mu_h - 1, \quad 0 \leq n \leq \nu_k - 1, \\ u_0^n = u_{\mu_h}^n = 0, & \qquad\qquad\qquad\qquad 0 \leq n \leq \nu_k - 1, \\ u_j^0 = g(x_j), & 1 \leq j \leq \mu_h - 1. \end{cases}$$

  This corresponds to replacing $\frac{\partial^2 u}{\partial x^2}(x_j, t_n)$ by its centered difference approximation and $\frac{\partial u}{\partial t}(x_j, t_n)$ by its forward difference approximation in the PDE.

- This system of fully discrete equations can be written equivalently as

$$\begin{cases} u_j^{n+1} = u_j^n + k\dfrac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2}, & 0 \leq j \leq \mu_h - 1, \quad 0 \leq n \leq \nu_k - 1, \\ u_0^n = u_{\mu_h}^n = 0, & \qquad\qquad\qquad\qquad 0 \leq n \leq \nu_k - 1, \\ u_j^0 = g(x_j), & 1 \leq j \leq \mu_h - 1. \end{cases}$$

Centered-in-space forward-in-time method (continued)

- The aforementioned system of fully discrete equations can be written equivalently as

$$
\left\{
\begin{aligned}
&\underbrace{\begin{bmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_{\mu_h-2}^{n+1} \\ u_{\mu_h-1}^{n+1} \end{bmatrix}}_{\boldsymbol{u}^{hk}(t_{n+1})} = \underbrace{\begin{bmatrix} u_1^{n} \\ u_2^{n} \\ \vdots \\ u_{\mu_h-2}^{n} \\ u_{\mu_h-1}^{n} \end{bmatrix}}_{\boldsymbol{u}^{hk}(t_n)} + k\,\frac{1}{h^2} \underbrace{\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}}_{[A]} \underbrace{\begin{bmatrix} u_1^{n} \\ u_2^{n} \\ \vdots \\ u_{\mu_h-2}^{n} \\ u_{\mu_h-1}^{n} \end{bmatrix}}_{\boldsymbol{u}^{hk}(t_n)}, \quad 0 \le n \le \nu_k - 1, \\
&\underbrace{\begin{bmatrix} u_1^{0} \\ u_2^{0} \\ \vdots \\ u_{\mu_h-2}^{0} \\ u_{\mu_h-1}^{0} \end{bmatrix}}_{\boldsymbol{u}^{hk}(t_0)} = \underbrace{\begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_{\mu_h-2}) \\ g(x_{\mu_h-1}) \end{bmatrix}}_{\boldsymbol{g}}.
\end{aligned}
\right.
$$

- Hence, more compactly,

$$
\begin{cases}
\boldsymbol{u}^{hk}(t_{n+1}) = [I + kA]\boldsymbol{u}^{hk}(t_n), & 0 \le n \le \nu_k - 1, \\
\boldsymbol{u}^{hk}(t_0) = \boldsymbol{g}.
\end{cases}
$$

Centered-in-space trapezoidal-in-time method

- The centered-in-space trapezoidal-in-time method is obtained by requiring that

$$\begin{cases} \dfrac{u_j^{n+1}-u_j^n}{k} - \dfrac{1}{2}\left(\dfrac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2} + \dfrac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{h^2}\right) = 0, & 1 \leq j \leq \mu_h - 1, \quad 0 \leq n \leq \nu_k - 1, \\ u_0^n = u_{\mu_h}^n = 0, & 0 \leq n \leq \nu_k - 1, \\ u_j^0 = g(x_j), & 1 \leq j \leq \mu_h - 1, \end{cases}$$

that is, $\left[I - \frac{k}{2}A\right] \boldsymbol{u}^{hk}(t_{n+1}) = \left[I + \frac{k}{2}A\right] \boldsymbol{u}^{hk}(t_n)$, $0 \leq n \leq \nu_k - 1$, with $\boldsymbol{u}^{hk}(t_0) = \boldsymbol{g}$.

Centered-in-space backward-in-time method

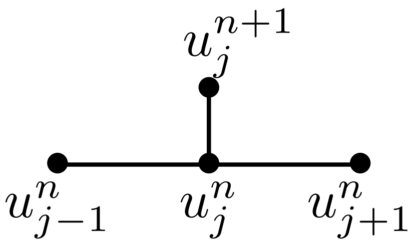
- The centered-in-space backward-in-time method is obtained by requiring that

$$\begin{cases} \dfrac{u_j^{n+1}-u_j^n}{k} - \dfrac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{h^2} = 0, & 1 \leq j \leq \mu_h - 1, \quad 0 \leq n \leq \nu_k - 1, \\ u_0^n = u_{\mu_h}^n = 0, & 0 \leq n \leq \nu_k - 1, \\ u_j^0 = g(x_j), & 1 \leq j \leq \mu_h - 1, \end{cases}$$

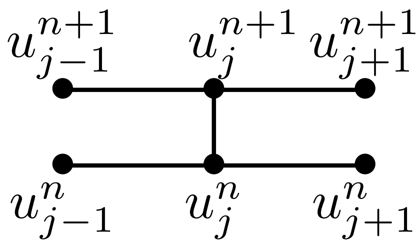
that is, $[I - kA]\boldsymbol{u}^{hk}(t_{n+1}) = \boldsymbol{u}^{hk}(t_n)$, $0 \leq n \leq \nu_k - 1$, with $\boldsymbol{u}^{hk}(t_0) = \boldsymbol{g}$.

Stencils

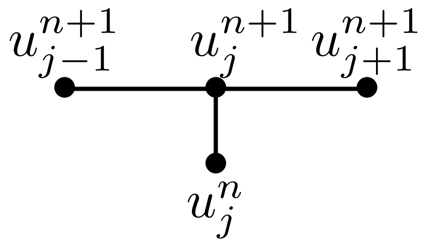
- The aforementioned finite difference methods have the following graphical representations:



Centered in space
forward in time.



Centered in space
trapezoidal in time.



Centered in space
backward in time.

Explicit versus implicit methods

- Because the centered-in-space trapezoidal-in-time and backward-in-time methods give implicit equations that must be solved for $\boldsymbol{u}^{hk}(t_{n+1})$, they are **implicit** methods, whereas the centered-in-space forward-in-time method is **explicit**.

Notions of consistency, stability, and convergence

- Notions of consistency, stability, and convergence are introduced to evaluate how good a finite difference method is in approximating the solution:

$$
\begin{tikzcd}
\textbf{Local truncation error} \arrow[rr, leftrightarrow, bend right=10, "\textbf{stability}"'] \arrow[d, dash, dashed] & & \textbf{Global error} \arrow[d, dash, dashed] \\
\textbf{Consistency} & & \textbf{Convergence}
\end{tikzcd}
$$

- The important point is that for an IBVP, we cannot let the grid spacing h and the time step k go to zero at independent rates and necessarily expect the resulting numerical solution to converge.

  Thus, a key aspect of studying consistency, stability, and convergence of finite difference methods for IBVPs is in understanding whether some **proper relation must hold between h and k.**

  It is often useful to think of such a relation as indicating how to properly **balance approximation errors between the discretization of space and that of time**. Clearly, this is of great practical importance because of the guidance it provides for properly refining a numerical solution.

  Conversely, much effort in numerical mathematics has been expended to conceive numerical methods that allow h and k to be refined at the same rate and/or entirely avoid any restriction.

Consistency of centered-in-space forward-in-time method

- The **local truncation error** τ_j^{n+1} at (x_j, t_{n+1}) is obtained by inserting the exact solution into the finite difference equation and determining by how much it fails to satisfy this equation.

- For the centered-in-space forward-in-time method, τ_j^{n+1} is obtained as

$$\tau_j^{n+1} = \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{k} - \frac{u(x_{j-1}, t_n) - 2u(x_j, t_n) + u(x_{j+1}, t_n)}{h^2}.$$

  If the exact solution is sufficiently regular, we can use the "Big-oh" characterization of the remainder in a Taylor series to obtain "Big-oh" characterizations of the finite difference approximations:

$$\frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{k} = \frac{\partial u}{\partial t}(x_j, t_n) + \frac{1}{2}k\frac{\partial^2 u}{\partial t^2}(x_j, t_n) + O(k^2),$$

$$\frac{u(x_{j-1}, t_n) - 2u(x_j, t_n) + u(x_{j+1}, t_n)}{h^2} = \frac{\partial^2 u}{\partial x^2}(x_j, t_n) + \frac{1}{12}h^2\frac{\partial^4 u}{\partial x^4}(x_j, t_n) + O(h^4).$$

  Inserting these "Big-oh" characterizations of the finite difference approximations in the expression for τ_j^{n+1}, we find that $\partial_t u(x_j, t_n)$ and $\partial_x^2(x_j, t_n)$ drop out because of the PDE, thus leading to

$$\max_{1 \leq n \leq \nu_k} |\tau_j^n| = O(h^2 + k).$$

- As a conclusion, the **centered-in-space forward-in-time method** is **consistent**. It is said to be of **order 2 in space** and **order 1 in time**.

Stability of centered-in-space forward-in-time method

- The centered-in-space forward-in-time method is **stable** in that there exists for every time duration τ, a constant $c > 0$ such that for all grid spacings h and time steps k that satisfy
$$\frac{k}{h^2} \le \frac{1}{2}$$
we have
$$\max_{1 \le n \le \nu_k} \sqrt{h \sum_{j=1}^{\mu_h - 1} \left(u(x_j, t_n) - u_j^n\right)^2} \le c \max_{1 \le n \le \nu_k} \sqrt{h \sum_{j=1}^{\mu_h - 1} \left(\tau_j^n\right)^2},$$
with c independent of h and k.

- Please note that whereas we used in Lecture 3 Part B the maxima $\max_{1 \le j \le \mu_h} |u(x_j) - u_j|$ and $\max_{1 \le j \le \mu_h} |\tau_j|$ to gauge the magnitude of the values $u(x_1) - u_1, \ldots, u(x_{\mu_h - 1}) - u_{\mu_h - 1}$ and $\tau_1, \ldots, \tau_{\mu_h - 1}$ relative to the grid points $x_1, \ldots, x_{\mu_h - 1}$, we use here a 2-norm to gauge the magnitude of the values $u(x_1, t_n) - u_1^n, \ldots, u(x_{\mu_h - 1}, t_n) - u_{\mu_h - 1}^n$ and $\tau_1^n, \ldots, \tau_{\mu_h - 1}^n$.

  In fact, finding an appropriate way of gauging the magnitude of the global error and the local truncation error, in which some form of stability can subsequently be proven, is often a key challenge in analyzing finite difference methods. The approach used to subsequently prove stability will depend on the particular norm that is being considered. Here, we consider a 2-norm, and we will see that that we can prove stability by explicitly computing the eigenvalues of $[A]$.

Consistency, stability, convergence

Proof of stability of centered-in-space forward-in-time method:

- With reference to the definition of the local truncation error, the exact solution satisfies
$$\boldsymbol{u}(t_n) = [I + kA]\boldsymbol{u}(t_{n-1}) + k\boldsymbol{\tau}^{hk}(t_n), \qquad n = 1, \ldots, \nu_k,$$
and the numerical solution satisfies
$$\boldsymbol{u}^{hk}(t_n) = [I + kA]\boldsymbol{u}^{hk}(t_{n-1}), \qquad n = 1, \ldots, \nu_k.$$

- Subtracting these equations from each other, we obtain
$$\boldsymbol{u}(t_n) - \boldsymbol{u}^{hk}(t_n) = [I + kA]\big(\boldsymbol{u}(t_{n-1}) - \boldsymbol{u}^{hk}(t_{n-1})\big) + k\boldsymbol{\tau}^{hk}(t_n).$$
and therefore
$$\boldsymbol{u}(t_n) - \boldsymbol{u}^{hk}(t_n) = \sum_{m=1}^{n} [I + kA]^{n-m} k\boldsymbol{\tau}^{hk}(t_m).$$

- Thus, to ensure that each contribution to the global error can be bounded in terms of its original size, we must ensure that the magnitude of $[I + kA]^{n-m}$ remains bounded as h and k tend to zero.

Consistency, stability, convergence

Proof of stability of centered-in-space forward-in-time method (continued):

- Denoting the 2-norm of a vector $\boldsymbol{v}$ in $\mathbb{R}^{\mu_h-1}$ by $\|\boldsymbol{v}\| = \sqrt{h\sum_{j=1}^{\mu_h-1} v_j^2}$, we have

$$\left\|\boldsymbol{u}(t_n) - \boldsymbol{u}^{hk}(t_n)\right\| = \left\| \sum_{1\le m\le n} [I + kA]^{n-m} k\boldsymbol{\tau}^{hk}(t_m)\right\|.$$

- Let $\|[I + kA]\|$ be the induced matrix norm of $[I + kA]$, that is,

$$\|[I + kA]\| = \max_{\substack{\boldsymbol{v}\in\mathbb{R}^{\mu_h-1}\\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|[I + kA]\boldsymbol{v}\|}{\|\boldsymbol{v}\|};$$

  then, we have

$$\left\|\boldsymbol{u}(t_n) - \boldsymbol{u}^{hk}(t_n)\right\| \le \sum_{1\le m\le n} \|[I + kA]\|^{n-m} k\|\boldsymbol{\tau}^{hk}(t_m)\|.$$

- If $\|[I + kA]\| \le 1$, then we can conclude,

$$\max_{1\le n\le\nu_k} \left\|\boldsymbol{u}(t_n) - \boldsymbol{u}^{hk}(t_n)\right\| \le c \max_{1\le n\le\nu_k} \left\|\boldsymbol{\tau}^{hk}(t_n)\right\| \quad \text{with} \quad c = k\nu_k = \tau,$$

  as asserted.

Proof of stability of centered-in-space forward-in-time method (continued):

- It remains to be shown that if $\frac{k}{h^2} \leq \frac{1}{2}$, then $\|[I + kA]\| \leq 1$. Below, we show this property by explicitly computing the eigenvalues of $[A]$.

- Because $[I + kA]$ is symmetric, $\|[I + kA]\|$ is equal to the spectral radius of $[I + kA]$, that is,
$$\|[I + kA]\| = \max_{1 \leq j \leq \mu_h - 1} |1 + k\lambda_j|, \qquad [A]\boldsymbol{\varphi}_j = \lambda_j \boldsymbol{\varphi}_j, \qquad 1 \leq j \leq \mu_h - 1.$$
With $\boldsymbol{i}_j$ the unit vector with 1 on the j-th row, the eigenvalues λ_j and eigenvectors $\boldsymbol{\varphi}_j$ read as
$$\lambda_j = \frac{2}{h^2}\big(\cos(j\pi h) - 1\big), \quad \boldsymbol{\varphi}_j \cdot \boldsymbol{i}_i = \sin(j\pi i h), \quad i = 1, \ldots, \mu_h - 1, \quad j = 1, \ldots, \mu_h - 1.$$
Indeed, using the aforementioned expression for the matrix $[A]$, we have
$$\begin{aligned}
[A]\boldsymbol{\varphi}_j \cdot \boldsymbol{i}_i &= \frac{1}{h^2}\Big(\sin(j\pi(i-1)h) - 2\sin(j\pi i h) + \sin(j\pi(i+1)h)\Big) \\
&= \frac{1}{h^2}\Big(\sin(j\pi i h)\cos(j\pi h) - \cos(j\pi i h)\sin(j\pi h) - 2\sin(j\pi i h) \\
&\qquad + \sin(j\pi i h)\cos(j\pi h) + \cos(j\pi i h)\sin(j\pi h)\Big) \\
&= \frac{2}{h^2}\Big(\cos(j\pi h) - 1\Big)\sin(j\pi i h).
\end{aligned}$$
Hence, $-\frac{4}{h^2} < \lambda_j < 0$, so that, indeed, if $\frac{k}{h^2} \leq \frac{1}{2}$, then $\|[I + kA]\| \leq 1$, as asserted.

Convergence of centered-in-space forward-in-time method

- The centered-in-space forward-in-time method is **convergent** in that for a sequence of progressively refined grid spacings and time steps for which the relation

$$\frac{k}{h^2} \leq \frac{1}{2}$$

ultimately holds between each pair, we can expect the convergence

$$\lim_{h,k\to 0}\left(\max_{1\leq n\leq \nu_k}\sqrt{h\sum_{j=1}^{\mu_h-1}\big(u(x_j,t_n)-u_j^n\big)^2}\right)=0.$$

Proof of convergence of centered-in-space forward-in-time method:

- It follows from the aforementioned consistency and stability properties that if $\frac{k}{h^2} \leq \frac{1}{2}$, then we have

$$\max_{1\leq n\leq \nu_k}\sqrt{h\sum_{j=1}^{\mu_h-1}\big(u(x_j,t_n)-u_j^n\big)^2}=O(h^2+k).$$

Thus, the method is **convergent** with **order 2 in space** and **order 1 in time**.

Consistency, stability, convergence

Convergence of centered-in-space forward-in-time method (continued)

- We found that we must require $\frac{k}{h^2} \leq \frac{1}{2}$ to ensure convergence. This is a severe restriction: the time step k must decrease at doubly the rate of the grid spacing h as we refine the grid! This restriction will force us to use an excessively small time step k of the order of h^2 when h is small.

- The stability restriction $\frac{k}{h^2} \leq \frac{1}{2}$ and the fact that we might want to take $k = O(h^2)$ anyway just to get the same level of accuracy in both space an time are reasons for not wanting to use the centered-in-space forward-in-time method in practice!

Aforementioned finite difference methods

- Centered-in-space forward-in-time method: explicit, convergent with order 2 in space and order 1 in time under the restriction $\frac{k}{h^2} \leq \frac{1}{2}$ imposed on the grid spacing and time step.

- Centered-in-space trapezoidal-in-time method: implicit, unconditionally convergent with order 2 in space and order 2 in time [Proof left as an excercise].

- Centered-in-space backward-in-time method: implicit, unconditionally convergent with order 2 in space and order 1 in time [Proof left as an excercise].

Outlook to method of lines and Von Neumann stability analysis

- We have seen that the analysis of consistency, stability, and convergence has practical relevance because it provides guidance for choosing and refining a finite difference method. Still, carrying out a comprehensive analysis of consistency, stability, and convergence can sometimes be hard.

- As an alternative to such a comprehensive analysis, there exist also other approaches that can sometimes give stability restrictions (relation that must hold between h and k, ...) more easily:

  - The **method of lines** provides a bridge with time-marching methods for IVPs, thus allowing stability by using theory for time-marching methods for IVPs.

  - **Von Neumann stability analysis** provides a bridge between finite difference methods and sampling theory, which allows stability to be studied by using Fourier analysis.

- In the following, we highlight only the main ideas underlying the method of lines and Von Neumann stability analysis. Please refer to the literature for more details about the exact relationships between consistency, stability, and convergence, the method of lines, and Von Neumann stability analysis (one form of stability being necessary or sufficient or both to ensure another form of stability, ...).

Notion of method of lines

- A **method-of-lines discretization** begins by discretizing in space alone. When a finite difference method is used, this leads to a system of ODEs, often called system of semidiscrete equations, in which each scalar equation is associated with the solution at some grid point. This system of ODEs is then discretized in time by using a time-marching method.

- For example, we may discretize the aforementioned IBVP in space by requiring that

$$\begin{cases} \dfrac{du_j}{dt}(t) - \dfrac{u_{j-1}(t) - 2u_j(t) + u_{j+1}(t)}{h^2} = 0, & 1 \leq j \leq \mu_h - 1, \quad 0 < t < \tau, \\ u_0(t) = u_{\mu_h}(t) = 0, & 0 < t < \tau, \\ u_j(0) = g(x_j), & 1 \leq j \leq \mu_h - 1. \end{cases}$$

  This **system of semidiscrete equations** can be written more compactly as

$$\begin{cases} \dfrac{d\boldsymbol{u}^h}{dt}(t) = [A]\boldsymbol{u}^h(t), \quad 0 < t < \tau, \\ \boldsymbol{u}^h(0) = \boldsymbol{g}. \end{cases}$$

  Using the forward Euler time-marching method, we obtain the **system of fully discrete equations**

$$\begin{cases} \boldsymbol{u}^{hk}(t_{n+1}) = [I + kA]\boldsymbol{u}^{hk}(t_n), & 0 \leq n \leq \nu_k - 1, \\ \boldsymbol{u}^{hk}(t_0) = \boldsymbol{g}, \end{cases}$$

  that is, we recover the centered-in-space forward-in-time method.

Separation of variables

- **Eigenproblem**: We begin by solving the following eigenproblem

$$[A]\boldsymbol{\varphi}_j = \lambda_j \boldsymbol{\varphi}_j, \qquad 1 \leq j \leq \mu_h - 1.$$

Because $[A]$ is a real, symmetric, square $(\mu_h - 1)$-dimensional matrix, the eigenvalues λ_1, ..., λ_{μ_h-1} are real and there exists an orthonormal basis consisting of eigenvectors $\boldsymbol{\varphi}_1$, ..., $\boldsymbol{\varphi}_{\mu_h-1}$.

- **Function series**: Given the eigenvectors $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{\mu_h-1}$, we seek a solution of the following form:

$$\boldsymbol{u}^h(t) = \sum_{j=1}^{\mu_h-1} b_j(t)\boldsymbol{\varphi}_j.$$

- **System of uncoupled equations (“diagonalization”)**: Inserting this into the IVP, we obtain

$$\begin{cases} \dfrac{db_j}{dt}(t) = \lambda_j b_j(t) & \text{for } 0 < t < \tau \\ b_j(0) = g_j = \boldsymbol{\varphi}_j \cdot \boldsymbol{g} & \text{at } t = 0 \end{cases}, \qquad \text{where } 1 \leq j \leq \mu_h - 1.$$

Link with absolute stability of time-marching methods

- We have seen that the centered-in-space forward-in-time method corresponds to a method-of-lines discretization that involves the application of the centered finite difference method for the discretization of space followed by the application of the forward Euler method.

  Now, with reference to the aforementioned system of uncoupled equations, we must require the time step k to satisfy $|1 + k\lambda_j| < 1$ for $1 \leq j \leq \mu_h - 1$ in order for this forward Euler time-marching method to be absolutely stable. For details, please refer to Lecture 3 Part B.

  Because the eigenvalues are given by $\lambda_j = \frac{2}{h^2}(\cos(j\pi h) - 1)$, hence, $-\frac{4}{h^2} < \lambda_j < 0$, we conclude that if $\frac{k}{h^2} \leq \frac{1}{2}$, then the forward Euler time-marching method is absolutely stable.

  We can observe that the restriction $\frac{k}{h^2} \leq \frac{1}{2}$ under which we were able previously to prove stability of the centered-in-space forward-in-time method coincides with that under which the forward Euler method in the method-of-lines discretization is absolutely stable.

- The previous observation suggests that it is sometimes possible to bridge finite difference methods for IBVPs with time-marching methods for IVPs and find stability restrictions for the former (relation that must hold between h and k, ...) by using theory for the latter.

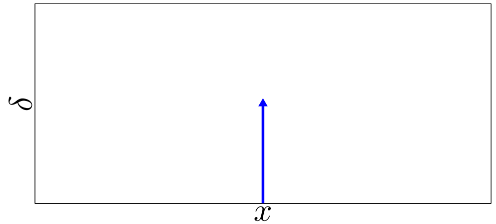
Review of sampling theory

This is not a lecture but rather a summary of key elements of sampling theory. For a more complete treatment of sampling theory, please refer to MATH0007 Analyse Mathématique II (F. Bastin) and SYST0002 Modélisation et analyse des systèmes (R. Sepulchre).

Review of sampling theory

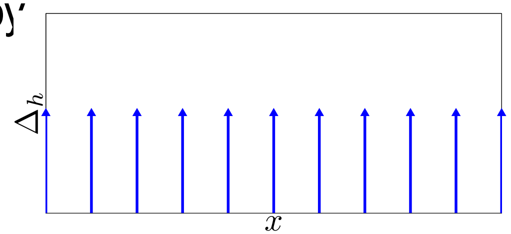
- For any y in $\mathbb{R}$, let $\delta(\cdot - y)$ be the **Dirac impulse** centered at y, which has the property that for any smooth function u from $\mathbb{R}$ into $\mathbb{R}$, we have

$$\delta(\cdot - y) \quad \text{such that} \quad \int_{\mathbb{R}} u(x)\delta(x-y)dx = u(y).$$



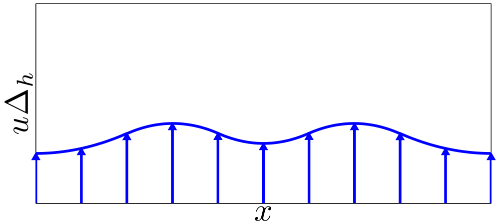
- For any $h > 0$, let Δ_h be the **Dirac comb** with period h defined by

$$\Delta_h = \sum_{j=-\infty}^{+\infty} \delta(\cdot - jh).$$



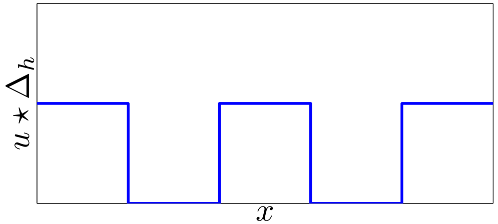
- Let u be a smooth function from $\mathbb{R}$ into $\mathbb{R}$. The **product** of u and the Dirac comb Δ_h provides a representation of the **sampling** of u with period h, that is,

$$u\Delta_h = \sum_{j=-\infty}^{+\infty} u(jh)\delta(\cdot - jh).$$



- Let u have a closed and bounded support. The **convolution** of u and the Dirac comb Δ_h leads to the **periodic repetition** of u with period h, that is,

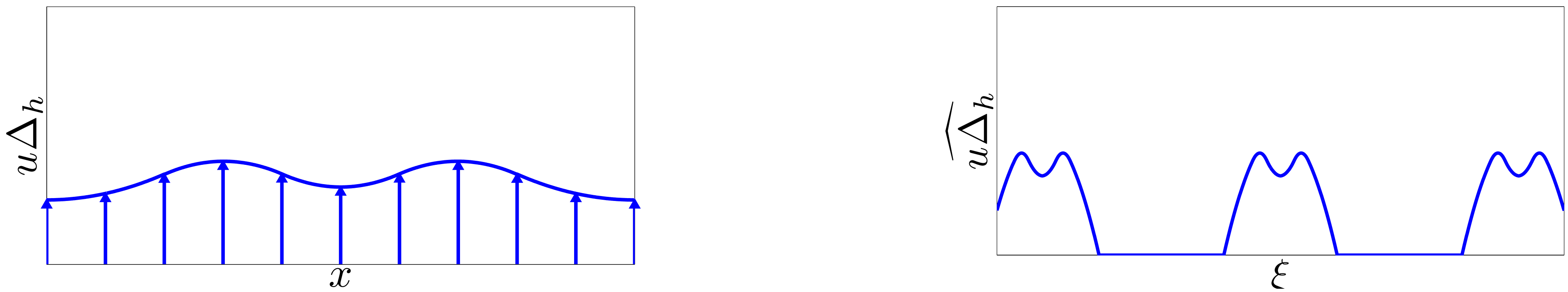
$$(u \star \Delta_h)(x) = \sum_{j=-\infty}^{+\infty} u(x - jh).$$



- The **Fourier transform** of the Dirac impulse and Dirac comb are as follows:

$$\widehat{\delta(\cdot - jh)}(\xi) = \exp(-i\xi jh) \quad \text{and} \quad \widehat{\Delta}_h(\xi) = \frac{2\pi}{h}\Delta_{\frac{2\pi}{h}}(\xi).$$

Sampling with spurious “aliasing” at a rate lower than the **Nyquist rate**: $\frac{1}{h} < \frac{\xi_L}{\pi}$.



“Meaningful” sampling at a rate higher than the **Nyquist rate**: $\frac{1}{h} \geq \frac{\xi_L}{\pi}$.

Poisson formula

- Let u be a smooth function from $\mathbb{R}$ into $\mathbb{R}$ that is **band-limited** in that its Fourier transform, $\hat{u}(\xi) = \int_{\mathbb{R}} u(x)\exp(-i\xi x)dx$, has a closed and bounded support $[-\xi_\mathsf{L}, \xi_\mathsf{L}]$. Then, if we sample u at a rate higher than the **Nyquist rate**, that is, $\frac{1}{h} \geq \frac{\xi_\mathsf{L}}{\pi}$, then we have

$$\hat{u}(\xi) = h \sum_{j=-\infty}^{+\infty} u(jh)\exp(-i\xi jh), \qquad \xi \in [-\xi_\mathsf{L}, \xi_\mathsf{L}]$$

(**Poisson formula**).

$$u(jh) = \frac{1}{2\pi}\int_{-\pi/h}^{\pi/h} \hat{u}(\xi)\exp(i\xi jh)d\xi, \quad -j \in \mathbb{Z}$$

- The Poisson formula can be proven by making rigorous (convergence,. . .) the following scheme:

$$\left(\hat{u} \star \Delta_{\frac{2\pi}{h}}\right) = \sum_{j=-\infty}^{+\infty} \hat{u}\left(\xi - j\frac{2\pi}{h}\right) \qquad = h\sum_{j=-\infty}^{+\infty} u(jh)\exp(-i\xi jh),$$

$$\downarrow_{\mathcal{F}^{-1}} \qquad\qquad\qquad\qquad \uparrow^{\mathcal{F}}$$

$$\mathcal{F}^{-1}\left(\hat{u} \star \Delta_{\frac{2\pi}{uh}}\right) = u\mathcal{F}^{-1}\left(\Delta_{\frac{2\pi}{h}}\right) = uh\Delta_h = h\sum_{j=-\infty}^{+\infty} u(jh)\delta(\cdot - jh)$$

Shannon theorem and Parseval equality

- Let u be a smooth function from $\mathbb{R}$ into $\mathbb{R}$ that is square-integrable and **band-limited** in that its Fourier transform, $\hat{u}(\xi) = \int_{\mathbb{R}} u(x)\exp(-i\xi x)dx$, has a closed and bounded support $[-\xi_{\mathsf{L}}, \xi_{\mathsf{L}}]$. Then, if we sample u at the **Nyquist rate**, that is, $\frac{1}{h} = \frac{\xi_{\mathsf{L}}}{\pi}$, then, we have

$$u(x) = \sum_{j=-\infty}^{+\infty} u(jh)\frac{\sin\left(\xi_{\mathsf{L}}(x-jh)\right)}{\xi_{\mathsf{L}}(x-jh)} \qquad \text{(\textbf{Shannon})},$$

where convergence is in the sense of the norm of the square-integrable functions, and

$$\|u\|^2 = \frac{1}{2\pi}\int_{-\pi/h}^{\pi/h} |\hat{u}(\xi)|^2 d\xi = h\sum_{j=-\infty}^{+\infty} |u(jh)|^2 \qquad \text{(\textbf{Parseval equality})}.$$

- This can be proven by writing the Poisson formula and then making rigorous the following scheme:

$$\begin{array}{ccc} 1_{[-\xi_{\mathsf{L}},\xi_{\mathsf{L}}]}(\xi)\hat{u}(\xi) & = & \sqrt{h}\displaystyle\sum_{j=-\infty}^{+\infty} u(jh)\sqrt{h}\,1_{[-\xi_{\mathsf{L}},\xi_{\mathsf{L}}]}(\xi)\exp(-i\xi jh), \\ \downarrow_{\mathcal{F}^{-1}} & & \downarrow_{\mathcal{F}^{-1}} \\ u(x) & = & \sqrt{h}\displaystyle\sum_{j=-\infty}^{+\infty} u(jh)\frac{\sin\left(\xi_{\mathsf{L}}(x-jh)\right)}{\sqrt{\pi\xi_{\mathsf{L}}}x} \end{array}$$

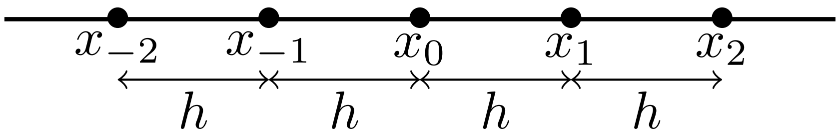
The Parseval equality follows by showing that the functions $\{\sin(\xi_{\mathsf{L}}(x-jh))/(\sqrt{\pi\xi_{\mathsf{L}}}x)\}_{j=-\infty}^{+\infty}$ form an orthonormal basis for $L^2(\mathbb{R})$.

Model problem

- Von Neumann stability analysis is typically applied to problems defined on all of space, as required for Fourier analysis. Hence, let us consider the initial value problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0, & -\infty < x < +\infty, \qquad 0 < t < +\infty, \\ u(x,0) = g(x), & -\infty < x < +\infty, \qquad \text{(initial condition).} \end{cases}$$

- In the spatial domain, let grid points be introduced as follows:



The grid spacing is denoted by h; thus, $x_j = jh$ for j in $\mathbb{Z}$.

- In the time domain, let approximations be computed at successive times times $t_0, t_1, t_2, \ldots$. The time step is denoted by k; thus, $t_n = nk$ for $0 \leq n < +\infty$.

- System of notation: numerical solution u_j^n approximates exact solution $u(x_j, t_n)$ at (x_j, t_n).

Notion of Von Neumann stability analysis

- Provided that the values $\{u_j^n\}_{j=-\infty}^{+\infty}$ can be considered to be a "meaningful" sampling of a function u^n (specifically, provided that they can be considered as the sampling of a smooth band-limited function at a rate higher than the Nyquist rate), we can use the Poisson formula to obtain

$$u_j^n = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} \hat{u}^n(\xi) \exp(i\xi jh) d\xi, \qquad j \in \mathbb{Z},$$

  that is, we obtain a **representation of the numerical solution as a linear combination of Fourier components** $\exp(i\xi jh)$ (which capture spatial dependence through j) with coefficients $\hat{u}^n(\xi)$ (which capture temporal dependence through n) with ξ in the range $[-\frac{\pi}{h}, \frac{\pi}{h}]$.

- The Von Neumann stability condition is obtained by requiring that there may be **no divergent Fourier components**, that is, by requiring that the amplitudes of the coefficients may not grow indefinitely as the time step tends to infinity. Specifically, upon defining the **amplification factor**

$$\gamma(\xi) = \frac{\hat{u}^{n+1}(\xi)}{\hat{u}^n(\xi)}, \qquad -\frac{\pi}{h} \le \xi \le \frac{\pi}{h},$$

  the **Von Neumann stability condition** is obtained by requiring that

$$|\gamma(\xi)| \le 1, \qquad -\frac{\pi}{h} \le \xi \le \frac{\pi}{h}.$$

Von Neumann stability analysis of centered-in-space forward-in-time method

- For the centered-in-space forward-in-time method, we have
$$u_j^{n+1}=\frac{1}{2\pi}\int_{-\pi/h}^{\pi/h}\hat{u}^n(\xi)\exp(i\xi jh)\underbrace{\left(1+\frac{k}{h^2}\Big(\exp(-i\xi h)-2+\exp\left(i\xi h\right)\Big)\right)}_{=\gamma(\xi)}d\xi,\ j\in\mathbb{Z},$$

- For this method, the amplification factor is as follows:
$$\gamma(\xi)=1+2\frac{k}{h^2}\big(\cos(\xi h)-1\big).$$
Hence, we find that if $\frac{k}{h^2}\leq\frac{1}{2}$, then the Von Neumann stability condition is fulfilled.

- We can observe that the restriction $\frac{k}{h^2}\leq\frac{1}{2}$ under which we were able previously to prove stability of the centered-in-space forward-in-time method coincides with that under which the Von Neumann stability condition is fulfilled. This can be explained as follows. By the Parseval equality (if it applies), we have $\|u^{n+1}\|^2=\frac{1}{2\pi}\int_{-\pi/h}^{\pi/h}|\hat{u}^n(\xi)|^2|\gamma(\xi)|^2d\xi$. Hence, if $|\gamma(\xi)|\leq 1$, then $\|u^{n+1}\|\leq\|u^n\|$, a property that resembles the requirement for the spectral radius to be smaller than or equal to one in the previous analysis of the stability of the centered-in-space forward-in-time method.

- The previous observation suggests that it is sometimes possible to bridge finite difference methods with sampling theory and find stability restrictions for the former (relation that must hold between h and k, ...) by using Fourier analysis.

Finite element method

Model problem

- Let us consider the numerical approximation of the solution to the initial-boundary value problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0, & 0 < x < 1, \quad 0 < t < \tau, \\ u(0,t) = u(1,t) = 0, & \qquad\qquad\quad 0 < t < \tau, \quad \text{(boundary conditions)}, \\ u(x,0) = g(x), & 0 < x < 1, \qquad\qquad\quad \text{(initial condition)}. \end{cases}$$

Variational formulation

- A variational formulation of this initial-boundary value problem:

> Given a sufficiently regular space-dependent function g, find a sufficiently regular space- and time-dependent function u with $u(0,t) = u(1,t) = 0$ for $0 < t < \tau$ such that
>
> $$\int_0^1 \frac{\partial u}{\partial t}(x,t) v(x) dx + \int_0^1 \frac{\partial u}{\partial x}(x,t) \frac{dv}{dx}(x) dx = 0, \qquad 0 < t < \tau,$$
>
> for all sufficiently regular space-dependent functions v with $v(0) = 0$ and $v(1) = 0$
>
> and the initial condition $u(x,0) = g(x)$ for $0 < x < 1$ is fulfilled.

Galerkin approximation

- Let a finite number μ_h of basis functions $\varphi_1, \ldots, \varphi_{\mu_h}$ be given. Let each basis function φ_j be a sufficiently regular space-dependent function such that $\varphi_j(0) = \varphi_j(1) = 0$.
- Then, the Galerkin approximation leads to the construction of an approximate solution u^h of the form of a linear combination of the basis functions, that is, $u^h(x,t) = \sum_{j=1}^{\mu_h} u_j(t)\varphi_j(x)$.
- The coefficients $u_1(t), \ldots, u_{\mu_h}(t)$ are determined by requiring the equation in the variational formulation to hold for all test functions that are linear combinations of the basis functions.
- Written compactly, the Galerkin approximate problem thus obtained takes the following form:

$$\begin{cases} [M]\dfrac{d\boldsymbol{u}}{dt}(t) + [K]\boldsymbol{u}(t) = 0 & \text{for } 0 < t < \tau, \\ \boldsymbol{u}(0) = \boldsymbol{g} & \text{at } t = 0, \end{cases}$$

  where $[M]$ and $[K]$ are square μ_h-dimensional matrices with $M_{ij} = \int_0^1 \varphi_i \varphi_j dx$ and $K_{ij} = \int_0^1 \frac{d\varphi_i}{dx}\frac{d\varphi_j}{dx} dx$ and $\boldsymbol{g}$ is a μ_h-dimensional vector whose components g_j are such that $u^h(x,0) = \sum_{j=1}^{\mu_h} g_j \varphi_j(x)$ is an appropriate approximation of $g(x)$ for $0 < x < 1$.
- This initial-value problem can be discretized in time by using any appropriate time-marching method.

Finite element method

- A finite element method is obtained when the basis functions in the Galerkin approximation are constructed, after meshing the domain, as elementwise low-degree polynomials.

Summary and conclusion

- The numerical approximation of the solution to an initial-boundary value problem requires the discretization of space and time.
- The discretization of space and time are not independent of each other. Often, a proper relation must hold between them to ensure convergence as they are refined.
- Consistency, stability, and convergence are analyzed to evaluate how good finite difference methods are in approximating solutions to IBVPs. Such analyses provide guidance for choosing and refining finite difference methods.
- As an alternative to comprehensive analyses of consistency, stability, and convergence, there exist also other approaches that can sometimes provide guidance more easily:
  - The method of lines bridges finite difference methods for IBVPs with time-marching methods for IVPs, thus allowing stability of the former to be studied by using theory for the latter.
  - Von Neumann stability analysis provides a bridge between finite difference methods and sampling theory, which allows stability to be studied by using Fourier analysis.
- Variational formulation. Galerkin approximation. Finite element method.
- Working through numerical examples is very helpful towards understanding this material. Please do not hesitate to come up with examples yourself to try things out using small Matlab codes.

References

Suggested reading material

- P. Olver. Introduction to Partial Differential Equations. Springer, 2014. Section 5.2.

Additional references also consulted to prepare this lecture

- C. Gasquet and P. Witomski. Analyse de Fourier et applications. Masson, 1990.
- R. LeVeque. Finite difference methods for ordinary and partial differential equations. SIAM, 2007.
- P.-A. Raviart and J.-M. Thomas. Introduction à l'analyse numérique des équations aux dérivées partielles. Dunod, 1993.
- C. Soize. Méthodes mathématiques en analyse du signal. Masson, 1993.