MATH0024 – Modeling with PDEs

Time-marching methods for ODEs
Finite difference method for Laplace/Poisson equation

Maarten Arnst and Romain Boman

October 11, 2017

Outline

- Review of ordinary differential equations (ODEs):
  - Fundamental results.
  - Linear ODEs.
- Time-marching methods for ODEs:
  - One-step time-marching methods.
  - Consistency, zero-stability, convergence.
  - Absolute stability.
  - Advanced time-marching methods.
- Finite difference method for Laplace/Poisson equation:
  - Finite difference method.
  - Consistency, stability, convergence.
- Summary and conclusion.
- References.

Review of ordinary differential equations (ODEs)

This is not a lecture but rather a summary of key elements of ODEs. For a more complete treatment of ODEs, please refer to MATH0002 “Analyse Mathématique” (E. Delhez).

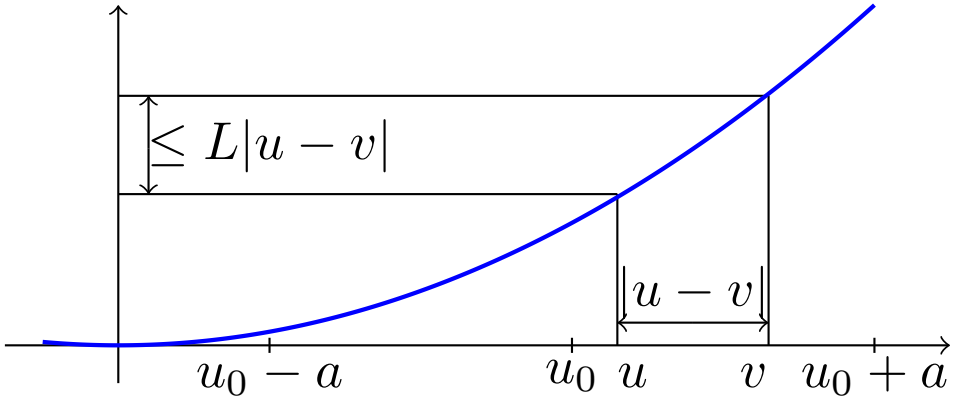
Notion of Lipschitz continuity

- Much of the theory of ODEs and many numerical methods for ODEs exploit in an essential manner the notion of Lipschitz continuity. We will therefore begin by taking a closer look at it.
- A function $\boldsymbol{f}$ from $\mathbb{R}^n$ into $\mathbb{R}^n$ is **Lipschitz continuous** over a domain

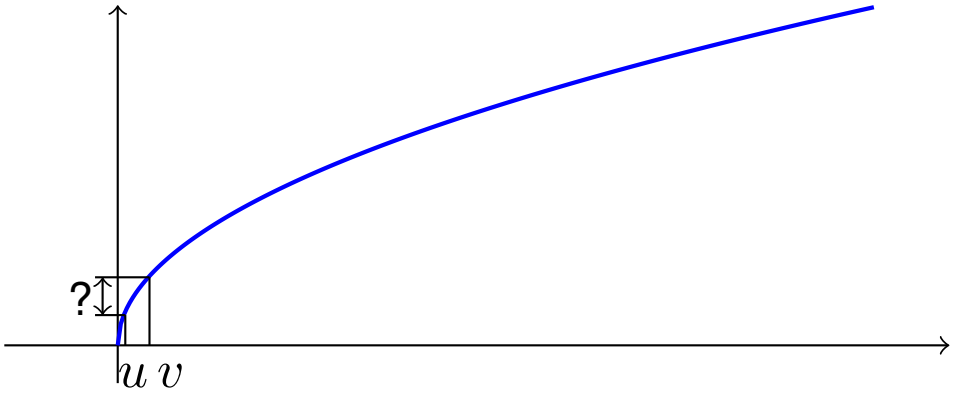
$$\mathcal{D} = \{\boldsymbol{u} \in \mathbb{R}^n \ : \ \|\boldsymbol{u} - \boldsymbol{u}_0\| \leq a\}$$

  if there exists a constant $L \geq 0$ such that

$$\|\boldsymbol{f}(\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{v})\| \leq L\|\boldsymbol{u} - \boldsymbol{v}\|, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{D}.$$

- 

The function $f(u) = u^2$ is Lipschitz continuous over any finite interval $|u - u_0| \leq a$, with $L = 2(u_0 + a)$ if $u_0 \geq 0$.



The function $f(u) = \sqrt{u}$ is not Lipschitz continuous near $u_0 = 0$ because $\partial_u f(u) = 1/(2\sqrt{u}) \to +\infty$ as $u \to 0$.

Notion of initial-value problem (IVP)

- An **initial-value problem** (IVP) is an ODE completed by an initial condition specified at an initial time. An IVP takes in general the form

$$\begin{cases} \dfrac{d\boldsymbol{u}}{dt}(t) = \boldsymbol{f}\big(t, \boldsymbol{u}(t)\big) & \text{for } t > t_0, \\ \boldsymbol{u}(t_0) = \boldsymbol{u}_0 & \text{at } t = t_0, \end{cases}$$

  where it is often assumed that $t_0 = 0$ for the sake of simplicity.

Well-posedness of IVPs

- A local existence, uniqueness, and stability of a solution to an IVP can be established provided that the function $\boldsymbol{f}(t, \boldsymbol{u})$ is continuous with respect to its first argument t and Lipschitz continuous with respect to its second argument $\boldsymbol{u}$. The details and the proofs are outside the scope of this course.

- As an example of ill-posedness that may arise if the function on the right-hand side is not Lipschitz continuous, consider the IVP

$$\begin{cases} \dfrac{du}{dt}(t) = \sqrt{u(t)} & \text{for } t > 0, \\ u(0) = 0 & \text{at } t = 0. \end{cases}$$

  The function f is not Lipschitz continuous near $u = 0$. As a result, the IVP need not have a unique solution. In fact, it has two solutions $u(t) = 0$ and $u(t) = \frac{1}{4}t^2$.

IVP involving n-th order linear ODE with constant coefficients

- First, we consider an IVP involving an n-th order linear ODE with constant coefficients:

$$\begin{cases} a_n \dfrac{d^n u}{dt^n}(t) + \ldots + a_1 \dfrac{du}{dt}(t) + a_0 u = b(t), & \text{for } t > t_0, \quad a_n, \ldots, a_0 \in \mathbb{C}, \\ \dfrac{d^{n-1} u}{dt^{n-1}}(t_0) = u_{n-1}, \ldots, \dfrac{du}{dt}(t_0) = u_1, u(t_0) = u_0, & \text{at } t = t_0. \end{cases}$$

Equivalent system of first-order linear ODEs with constant coefficients

- The ODE involved in the aforementioned IVP can be written equivalently as a system of first-order linear ODEs with constant constant coefficients as follows:

$$\frac{d\boldsymbol{u}}{dt}(t) = [A]\boldsymbol{u}(t) + \boldsymbol{b}(t),$$

in which $\boldsymbol{u}(t)$, $[A]$, and $\boldsymbol{b}(t)$ are given by

$$\boldsymbol{u}(t) = \begin{bmatrix} u(t) \\ \frac{du}{dt}(t) \\ \vdots \\ \frac{d^{n-2}u}{dt^{n-2}}(t) \\ \frac{d^{n-1}u}{dt^{n-1}}(t) \end{bmatrix}, \quad [A] = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ -\frac{a_0}{a_n} & -\frac{a_1}{a_n} & -\frac{a_2}{a_n} & \ldots & -\frac{a_{n-1}}{a_n} \end{bmatrix}, \quad \boldsymbol{b}(t) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{b(t)}{a_n} \end{bmatrix}.$$

Notion of homogeneous solution

- A so-called **homogeneous solution** u_h to the ODE involved in the aforementioned IVP is a solution to this ODE when the right-hand side is set to zero, that is,

$$a_n \frac{d^n u_\mathsf{h}}{dt^n}(t) + \ldots + a_1 \frac{du_\mathsf{h}}{dt}(t) + a_0 u_\mathsf{h} = 0, \quad \text{for } t > t_0, \quad a_n, \ldots, a_0 \in \mathbb{C}.$$

- Owing to the aforementioned results for well-posedness of IVPs, this homogeneous solution is not unique; rather, it can be represented in general as a linear combination of n elementary solutions.
- These elementary solutions can be determined from the **characteristic polynomial**

$$p(\lambda) = a_n \lambda^n + \ldots + a_1 \lambda + a_0,$$

which, owing to the "fundamental theorem of algebra," can be factored into linear factors as

$$p(\lambda) = a_n \prod_{j=1}^{s} (\lambda - \lambda_j)^{n_j}.$$

where $n_1, \ldots, n_s$ are the multiplicities of the distinct roots $\lambda_1, \ldots, \lambda_s$ with $n_1 + \ldots + n_s = n$.

- The homogeneous solution u_h can then be represented as a linear combination

$$u_\mathsf{h}(t) = \sum_{j=1}^{s} \sum_{q=1}^{n_j} \alpha_{j,q} u_{j,q}(t), \quad \alpha_{j,q} \in \mathbb{C},$$

of the linearly independent **elementary solutions**

$$u_{j,q}(t) = t^{q-1} \exp(\lambda_j t), \quad 1 \leq q \leq n_j, \quad 1 \leq j \leq s.$$

Notion of homogeneous solution (continued)

- If the coefficients are real, that is, $a_n, \ldots, a_0 \in \mathbb{R}$, complex roots must occur in complex conjugate pairs so that the elementary solutions $t^{q-1}\exp((\lambda_\mathsf{R} \pm \lambda_\mathsf{I} i)t)$ for $\lambda = \lambda_\mathsf{R} \pm \lambda_\mathsf{I} i$ can be replaced with the equivalent elementary solutions $t^{q-1}\exp(\lambda_\mathsf{R} t)\cos(\lambda_\mathsf{I} t)$ and $t^{q-1}\exp(\lambda_\mathsf{R} t)\sin(\lambda_\mathsf{I} t)$.

Example: linear oscillator

- Let us consider the linear oscillator with mass m, damping c, and stiffness k:

$$m\frac{d^2u}{dt^2}(t) + c\frac{du}{dt}(t) + ku(t) = f(t), \quad \text{for } t > t_0, \quad m, c, k \in \mathbb{R}_0^+.$$

- The characteristic polynomial reads as

$$p(\lambda) = m\lambda^2 + c\lambda + k.$$

- If $c^2 - 4mk > 0$, there are two distinct real roots $\lambda_{1,2} = \frac{-c \pm \sqrt{c^2-4mk}}{2m}$; hence,

$$u_\mathsf{h}(t) = \alpha_1 \exp(\lambda_1 t) + \alpha_2 \exp(\lambda_2 t).$$

- If $c^2 - 4mk = 0$, there is a repeated real root $\lambda = \frac{-c}{2m}$; hence,

$$u_\mathsf{h}(t) = \alpha_{1,0} \exp(\lambda t) + \alpha_{1,1} t \exp(\lambda t).$$

- If $c^2 - 4mk < 0$, there are two complex conjugate roots $\lambda_\mathsf{R} \pm \lambda_\mathsf{I} i = \frac{-c}{2m} \pm \frac{\sqrt{4mk-c^2}}{2m} i$; hence,

$$u_\mathsf{h}(t) = \alpha_1 \exp(\lambda_\mathsf{R} t)\cos(\lambda_\mathsf{I} t) + \alpha_2 \exp(\lambda_\mathsf{R} t)\sin(\lambda_\mathsf{I} t).$$

Notion of particular solution

- A so-called **particular solution** u_{p} to the ODE involved in the aforementioned IVP is a solution to this ODE which may, but need not, satisfy the initial conditions, that is,

$$a_m \frac{d^n u_{\mathrm{p}}}{dt^n}(t) + \ldots + a_1 \frac{du_{\mathrm{p}}}{dt}(t) + a_0 u_{\mathrm{p}} = b(t), \quad \text{for } t > t_0, \quad a_m, \ldots, a_0 \in \mathbb{C}.$$

- Such a particular solution u_{p} can be determined by looking for a solution of the form

$$u_{\mathrm{p}}(t) = \sum_{j=1}^{s} \sum_{q=1}^{n_j} \beta_{j,q}(t) u_{j,q}(t), \qquad \text{(\textbf{variation of constants}).}$$

- The coefficients $\beta_{j,q}(t)$ can be determined by solving the following system for their derivatives:

$$\begin{cases} \displaystyle\sum_{j=1}^{s} \sum_{q=1}^{n_j} \frac{d\beta_{j,q}}{dt}(t) u_{j,q}(t) = 0 \\ \ldots \\ \displaystyle\sum_{j=1}^{s} \sum_{q=1}^{n_j} \frac{d\beta_{j,q}}{dt}(t) \frac{d^{n-2} u_{j,q}}{dt^{n-2}}(t) = 0 \\ \displaystyle\sum_{j=1}^{s} \sum_{q=1}^{n_j} \frac{d\beta_{j,q}}{dt}(t) \frac{d^{n-1} u_{j,q}}{dt^{n-1}}(t) = \frac{1}{a_n} b(t) \end{cases},$$

  and then determining the coefficients $\beta_{j,q}(t)$ therefrom by integration.

Notion of particular solution (continued)

- Indeed, we then have:

$$\frac{du_\mathsf{p}}{dt}(t) = \underbrace{\sum_{j=1}^{s}\sum_{q=1}^{n_j}\frac{d\beta_{j,q}}{dt}(t)u_{j,q}(t)}_{=0} + \sum_{j=1}^{s}\sum_{q=1}^{n_j}\beta_{j,q}(t)\frac{du_{j,q}}{dt}(t),$$

$$\ldots$$

$$\frac{d^n u_\mathsf{p}}{dt^n}(t) = \underbrace{\sum_{j=1}^{s}\sum_{q=1}^{n_j}\frac{d\beta_{j,q}}{dt}(t)\frac{d^{n-1}u_{j,q}}{dt^{n-1}}(t)}_{=\frac{1}{a_n}b(t)} + \underbrace{\sum_{j=1}^{s}\sum_{q=1}^{n_j}\beta_{j,q}(t)\frac{d^n u_{j,q}}{dt^n}(t)}_{=\left(-\frac{a_{n-1}}{a_n}\right)\frac{d^{n-1}u_\mathsf{p}}{dt^{n-1}}(t)+\ldots+\left(-\frac{a_0}{a_n}\right)u_\mathsf{p}(t)}$$

- We will encounter these equations again, although written in a more easily understandable form, when reviewing the Duhamel formula later.

Solution to IVP involving n-th order linear ODE with constant coefficients

- The solution to the aforementioned IVP can be written as

$$u(t) = u_\mathsf{p}(t) + \sum_{j=1}^{s}\sum_{q=1}^{n_j}\alpha_{j,q}u_{j,q}(t),$$

where the coefficients $\alpha_{j,q}$ must be determined such that the initial conditions are fulfilled.

IVP involving system of linear ODEs with constant coefficients

- Next, we consider an IVP involving a system of linear ODEs with constant coefficients:

$$\begin{cases} \dfrac{d\boldsymbol{u}}{dt}(t) = [A]\boldsymbol{u}(t) + \boldsymbol{b}(t), \quad \text{for } t > t_0, \quad [A] \in \mathsf{M}_n(\mathbb{R}), \\ \boldsymbol{u}(t_0) = \boldsymbol{u}_0. \end{cases}$$

Matrix exponential

- The **matrix exponential** of a square n-dimensional matrix $[A]$ is the matrix $\exp([A])$ such that

$$\exp([A]) = \sum_{j=0}^{+\infty} \frac{1}{j!}[A]^j.$$

- The matrix exponential has the property that $\frac{d}{dt}\exp([A]t) = [A]\exp([A]t) = \exp([A]t)[A]$.
- If $[A]$ can be diagonalised by using eigendecomposition, that is, if $[A]$ can be written as $[A] = [V][D][V]^{-1}$, where $[V] = [\boldsymbol{v}_1|\ldots|\boldsymbol{v}_n]$ and $[D] = [\mathsf{Diag}(\lambda_1, \ldots, \lambda_n)]$ collect the eigenvectors and eigenvalues of $[A]\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j$, $j = 1, \ldots, n$, the matrix exponential $\exp([A])$ is $\exp([A]) = [V]\exp([D])[V]^{-1}$ with $\exp([D]) = [\mathsf{Diag}(\exp(\lambda_1), \ldots, \exp(\lambda_n))]$.
- One way of generalizing this method of obtaining the matrix exponential to nondiagonalisable matrices involves resorting to the Jordan decomposition.

Solution to IVP involving homogeneous system of linear ODEs with constant coefficients

- If $\boldsymbol{b} = \boldsymbol{0}$, then the ODE involved in the aforementioned IVP is homogeneous. The solution to the homogeneous system $\frac{d\boldsymbol{u}}{dt}(t) = [A]\boldsymbol{u}(t)$ with initial condition $\boldsymbol{u}(t_0) = \boldsymbol{u}_0$ is
$$\boldsymbol{u}(t) = \exp\big([A](t - t_0)\big)\boldsymbol{u}_0.$$

Solution to IVP with inhomogeneous system of linear ODEs with constant coefficients

- If $\boldsymbol{b}$ is not identically zero, then the solution to the inhomogeneous system $\frac{d\boldsymbol{u}}{dt}(t) = [A]\boldsymbol{u}(t) + \boldsymbol{b}(t)$ with initial condition $\boldsymbol{u}(t_0) = \boldsymbol{u}_0$ can be written as
$$\boldsymbol{u}(t) = \exp\big([A](t - t_0)\big)\boldsymbol{u}_0 + \int_{t_0}^{t} \exp\big([A](t - s)\big)\boldsymbol{b}(s)ds.$$
This is known as **Duhamel's formula**.

- Indeed, if a particular solution is sought of the form $\boldsymbol{u}_\mathrm{p}(t) = \exp([A]t)\boldsymbol{v}(t)$ (var. of constants), $\frac{d\boldsymbol{u}_\mathrm{p}}{dt}(t) = [A]\boldsymbol{u}_\mathrm{p}(t) + \exp([A]t)\frac{d\boldsymbol{v}}{dt}(t)$; hence, $\exp([A]t)\frac{d\boldsymbol{v}}{dt}(t) = \boldsymbol{b}(t)$; thus, $\boldsymbol{v}(t) = \int_{t_0}^{t} \exp(-[A]s)\boldsymbol{b}(s)ds$.

Time-marching methods for ODEs

Model IVP

- Let us consider numerical methods for the approximation of the solution to the IVP

$$\begin{cases} \dfrac{du}{dt}(t) = f\big(t, u(t)\big) & \text{for } t \in]t_0, t_0 + \tau[, \\ u(t_0) = u_0 & \text{at } t = t_0. \end{cases}$$

Forward Euler method

- There are two pieces of information: the value $u(t_0) = u_0$ of the solution at the initial time and, given the function f, the slope of the solution at the initial time from the ODE.
- The purpose of the time-marching method being to approximate the value of the solution at a later time, say t_1, the most elementary approach is to use linear extrapolation, that is,

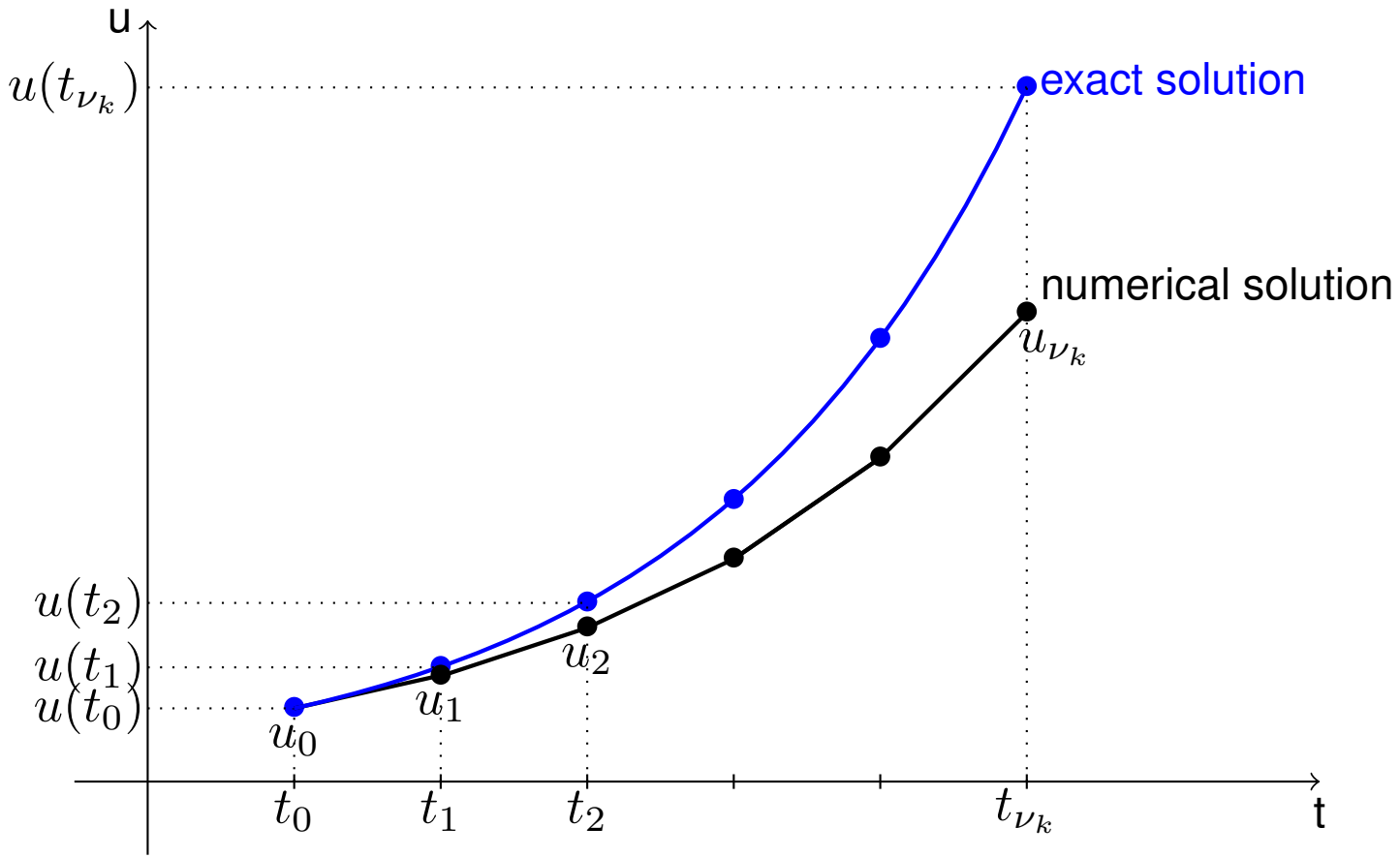
$$u(t_1) = u(t_0) + \int_{t_0}^{t_1} f\big(s, u(s)\big) ds \approx u_0 + \underbrace{(t_1 - t_0)}_{k \text{ = time step}} f\big(t_0, u(t_0)\big).$$

- The **forward Euler method** uses this linear extrapolation to march forward in time, computing approximations $u_0,\ u_1,\ u_2,\ \ldots,\ u_{\nu_k}$ at successive times $t_0,\ t_1,\ t_2,\ \ldots,\ t_{\nu_k}$ as follows:

$$u_{n+1} = u_n + k\, f(t_n, u_n).$$

  The time step is denoted by k; thus, $t_n = nk$ for $n = 0, \ldots, \nu_k$ with $\nu_k = \tau/k$.
- System of notation: numerical solution u_n approximates exact solution $u(t_n)$ at t_n.

Forward Euler method (continued)



Model IVP

$$\begin{cases} \dfrac{du}{dt}(t) = f\big(t, u(t)\big) & \text{for } t \in]t_0, t_0 + \tau[, \\ u(t_0) = u_0 & \text{at } t = t_0. \end{cases}$$

Forward Euler method

$$\begin{cases} u_{n+1} = u_n + k\, f(t_n, u_n) & \text{for } t_n = t_0, \ldots, t_{\nu_k}, \\ u_0 = u_0 & \text{at } t_0. \end{cases}$$

Trapezoidal method (Crank-Nicolson)

- The linear extrapolation method may not be very good, and it may make more sense to make the approximation of the derivative equal to the average of its values at the endpoints:

$$u(t_1) = u(t_0) + \int_{t_0}^{t_1} f\big(s, u(s)\big) ds \approx u_0 + (t_1 - t_0)\frac{1}{2}\Big(f\big(t_0, u(t_0)\big) + f\big(t_1, u(t_1)\big)\Big).$$

- This is the motivation behind the trapezoidal method (Crank-Nicolson):

$$u_{n+1} = u_n + \frac{k}{2}\big(f(t_n, u_n) + f(t_{n+1}, u_{n+1})\big).$$

Backward Euler method

- The backward Euler method uses

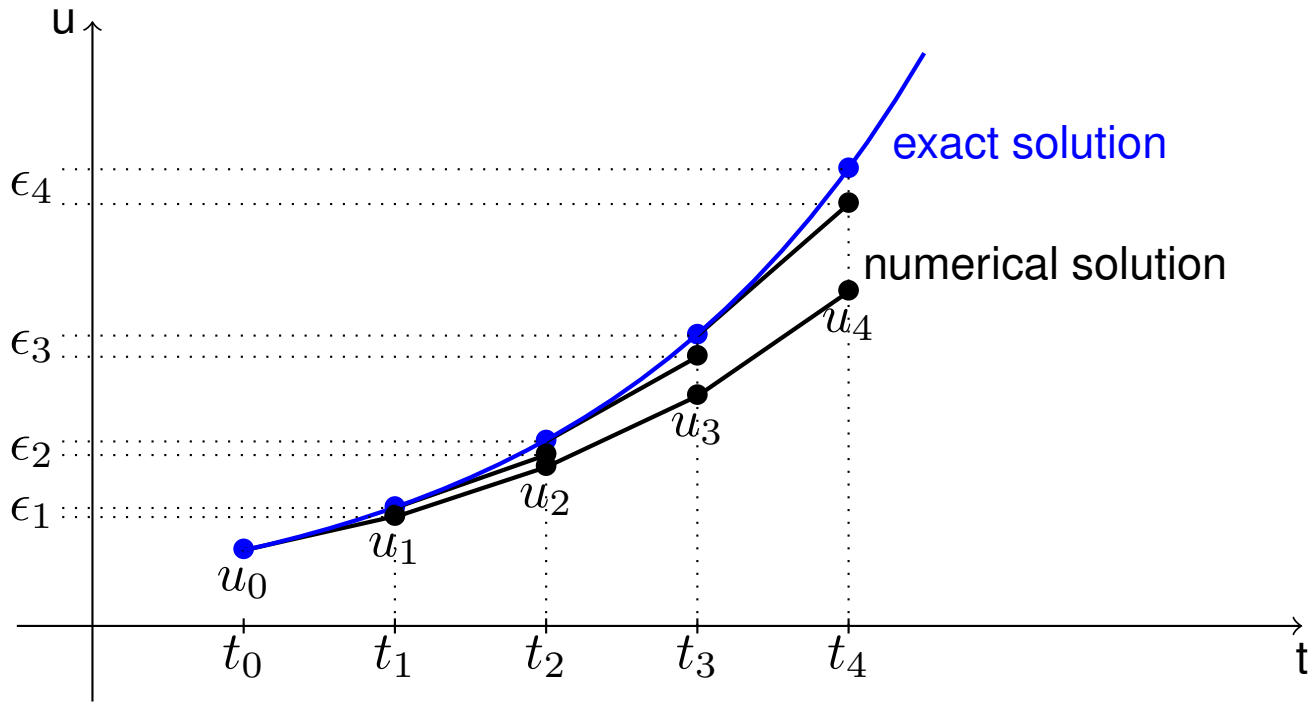
$$u_{n+1} = u_n + k\, f(t_{n+1}, u_{n+1}).$$

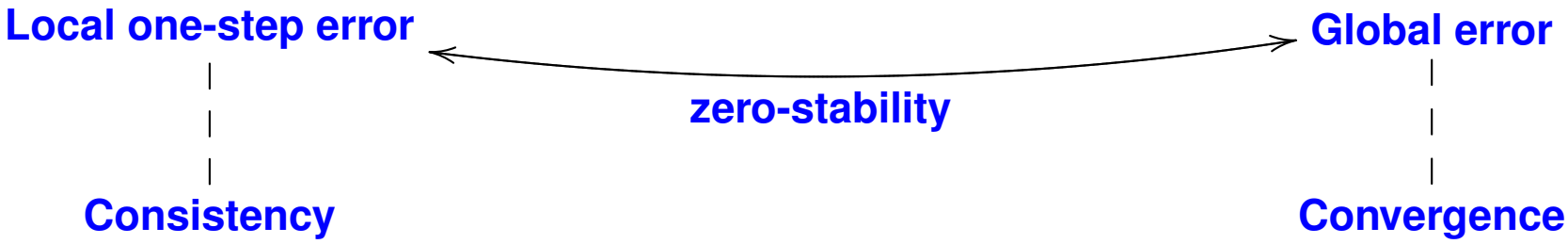
Notion of explicit and implicit time-marching methods

- Because the trapezoidal and the backward Euler method give an implicit equation that must be solved for u_{n+1}, they are **implicit** methods, whereas the forward Euler method is **explicit**.

Introduction to consistency, zero-stability, and convergence

- The notions of consistency, zero-stability, and convergence are concerned with how good a time-marching method is in approximating the solution to an IVP.



- 

Consistency, zero-stability, convergence

Notions of consistency and convergence

- The **local one-step error** ϵ_n is the error committed in the n-th step of the time-marching method assuming that no errors were made in the previous steps.
- A time-marching method is **consistent** if

$$\max_{1 \le n \le \nu_k} |\epsilon_n| = o(k).$$

- A time-marching method is of **order** p if

$$\max_{1 \le n \le \nu_k} |\epsilon_n| = O(k^{p+1}).$$

- A time-marching method is **convergent** if

$$\lim_{k \to 0} \left(\max_{1 \le n \le \nu_k} |u(t_n) - u_n| \right) = 0.$$

- A time-marching method is **convergent with order** p if

$$\max_{1 \le n \le \nu_k} |u(t_n) - u_n| = O(k^p).$$

- Whereas the notion of consistency is relevant to the local one-step error, the notion of convergence is relevant to the global error between the exact and the numerical solution.
- Consistency is typically established by using Taylor series, and the connection between consistency and convergence is typically made by using zero-stability, as described next.

Taylor series

- The Taylor series of a sufficiently regular function g from $\mathbb{R}$ into $\mathbb{R}$ about $\bar{t}$ in $\mathbb{R}$ is the power series

$$g(\bar{t}+k) = g(\bar{t}) + k\,\frac{dg}{dt}(\bar{t}) + \frac{1}{2}\,k^2\,\frac{d^2g}{dt^2}(\bar{t}) + \frac{1}{6}\,k^3\,\frac{d^3g}{dt^3}(\bar{t}) + \dots$$

  that is, more compactly,

$$g(\bar{t}+k) = \sum_{j=0}^{+\infty} \frac{1}{j!}\,k^j\,\frac{d^jg}{dt^j}(\,\bar{t}\,).$$

- If g is sufficiently regular, the remainder in the Taylor series truncated at degree p satisfies

$$g(\bar{t}+k) = \sum_{j=0}^{p} \frac{1}{j!}\,k^j\,\frac{d^jg}{dt^j}(\,\bar{t}\,) + \underbrace{O(k^{p+1})}_{\text{remainder}}.$$

Consistency and order of forward Euler method

- The forward Euler method is consistent and of order 1.

Proof of consistency and order of forward Euler method:

- Assuming that no errors were made in the previous steps, the one-step error ϵ_{n+1} made in the $(n+1)$-th step of the forward Euler method is obtained as

$$\epsilon_{n+1} = u(t_{n+1}) - \Big(u(t_n) + k\, f\big(t_n, u(t_n)\big)\Big).$$

- Because the exact solution satisfies the ODE $\frac{du}{dt}(t) = f(t, u(t))$, we obtain

$$\epsilon_{n+1} = u(t_{n+1}) - \Big(u(t_n) + k\frac{du}{dt}(t_n)\Big).$$

- If the exact solution u is sufficiently regular, the remainder in its degree-1 Taylor series satisfies

$$u(t_{n+1}) = u(t_n + k) = u(t_n) + k\,\frac{du}{dt}(t_n) + O(k^2).$$

- Hence,

$$\epsilon_{n+1} = O(k^2).$$

  As a conclusion, the forward Euler method is consistent and of order 1, as asserted.

Consistency, zero-stability, convergence

Notion of zero-stability

- Loosely speaking, a time-marching method is **zero-stable** if the global error incurred by this time-marching method can be bounded in terms of the sizes of the local one-step errors.

Zero-stability of forward Euler method

- If the function f (the right-hand side in the ODE in the aforementioned IVP) is globally Lipschitz continuous with respect to its second argument,

$$\exists L > 0,\ \forall (t,u),(t,v) \in [t_0, t_0+\tau] \times \mathbb{R}\ :\ \big|f(t,u) - f(t,v)\big| \leq L|u-v|,$$

  then the forward Euler method is zero-stable in that there exists a time step k_0 and a constant $c > 0$ such that for all time steps k smaller than k_0, with c independent of k, we have

$$\max_{1 \leq n \leq \nu_k} |u(t_n) - u_n| \leq c \sum_{n=1}^{\nu_k} |\epsilon_n|.$$

Proof of zero-stability of forward Euler method:

- With reference to the definition of the local one-step error, the exact solution satisfies

$$u(t_n) = u(t_{n-1}) + k\, f\big(t_{n-1}, u(t_{n-1})\big) + \epsilon_n, \quad n = 1, \ldots \nu_k,$$

  and the numerical solution satisfies

$$u_n = u_{n-1} + k\, f(t_{n-1}, u_{n-1}), \quad n = 1, \ldots \nu_k.$$

Consistency, zero-stability, convergence

Proof of zero-stability of forward Euler method (continued):

- Subtracting these equations from each other, we obtain

$$u(t_n) - u_n = u(t_{n-1}) - u_{n-1} + k\, f\big(t_{n-1}, u(t_{n-1})\big) - k\, f(t_{n-1}, u_{n-1}) + \epsilon_n.$$

- Using the triangle inequality and the global Lipschitz continuity of f, we obtain

$$\begin{aligned} |u(t_n) - u_n| &\leq |u(t_{n-1}) - u_{n-1}| + k \Big| f\big(t_{n-1}, u(t_{n-1})\big) - f(t_{n-1}, u_{n-1}) \Big| + |\epsilon_n| \\ &\leq (1 + kL)|u(t_{n-1}) - u_{n-1}| + |\epsilon_n| \\ &\leq \sum_{1 \leq j \leq n} (1 + kL)^{n-j} |\epsilon_j| \end{aligned}$$

- Because $(1 + kL) \leq \exp(kL)$, we obtain

$$\big|u(t_n) - u_n\big| \leq \sum_{1 \leq j \leq n} \exp\big((n-j)kL\big)|\epsilon_j|.$$

- Hence, with $\nu_k = \tau/k$, we conclude that

$$\max_{1 \leq n \leq \nu_k} \big|u(t_n) - u_n\big| \leq c \sum_{n=1}^{\nu_k} |\epsilon_n| \qquad \text{with} \qquad c = \exp(\tau L).$$

- The local one-step-error $\epsilon_n$contributes the term $(1 + kL)^{\nu_k - n}|\epsilon_n|$ to the (majorant of the) global error. Because $(1 + kL)^{\nu_k - n} \leq \exp(\tau L)$ remains bounded as $k \to 0$, each contribution to the (majorant of the) global error can be bounded in terms of its original size.

Convergence of forward Euler method

- If the function $f(t,u)$ (the right-hand side in the ODE in the IVP) is globally Lipschitz continuous w.r.t. its second argument u, the forward Euler method is convergent with order 1.

Proof of convergence of forward Euler method:

- Using the zero-stability of the forward Euler method, we obtain

$$\max_{1\leq n\leq \nu_k} |u(t_n)-u_n| \leq \exp(\tau L) \sum_{n=1}^{\nu_k} |\epsilon_n| \leq \exp(\tau L) \underbrace{\nu_k\, k}_{=\tau} \max_{1\leq n\leq \nu_k} \frac{|\epsilon_n|}{k}.$$

- Using the consistency of the forward Euler method, we obtain

$$\lim_{k\to 0} \left(\max_{1\leq n\leq \nu_k} |u(t_n)-u_n| \right) = 0.$$

  As a conclusion, the forward Euler method is convergent, as asserted.

- Using the fact that the forward Euler method is of order 1, we obtain

$$\max_{1\leq n\leq \nu_k} |u(t_n)-u_n| = O(k).$$

  As a conclusion, the forward Euler method is convergent with order 1, as asserted.

- Observe that the properties of **consistency** and **zero-stability** collectively imply **convergence**.

Absolute stability

Introduction

- Previously, we dealt with behavior over bounded intervals in the limit as k tended to zero.
- By contrast, the notion of absolute stability is concerned with asymptotic behavior of the solution as the independent variable tends to infinity for noninfinitesimal time steps.

Model IVP used to define the notion of absolute stability

- Absolute stability is concerned with the approximation of the solution to the IVP

$$\begin{cases} \dfrac{du}{dt}(t) = \lambda\, u(t) & \text{for } t > 0, \qquad \lambda \in \mathbb{C}, \\ u(0) = 1 & \text{at } t = 0. \end{cases}$$

- The asymptotic behavior of the exact solution, that is, $u(t) = \exp(\lambda t)$, is such that

$$\lim_{t \to +\infty} |u(t)| = 0 \quad \text{if} \quad \mathrm{Re}(\lambda) < 0.$$

Notion of absolute stability

- A time-marching method is **absolutely stable** for a time step k if its application to this model IVP leads for this time step k to a numerical solution with the same asymptotic behavior, that is,

$$\lim_{n \to +\infty} |u_n| = 0 \quad \text{if} \quad \mathrm{Re}(\lambda) < 0.$$

Notion of region of absolute stability

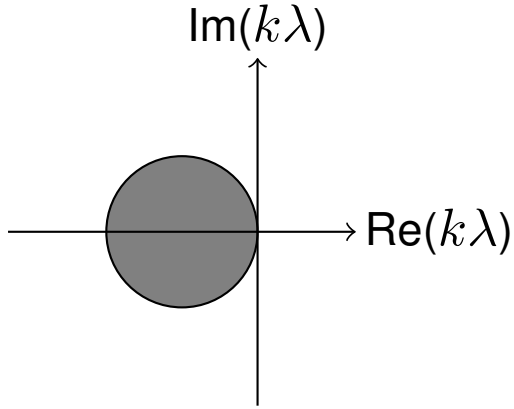
- If the application of a time-marching method to the aforementioned IVP with parameter λ leads for a time step k to a numerical solution that decays to zero, then the product $k\lambda$ is said to lie in the **region of absolute stability** for this time-marching method:

$$\mathcal{A} = \Big\{ k\lambda \in \mathbb{C} \ : \ \lim_{n \to +\infty} |u_n| = 0 \Big\}.$$

A time-marching method is said to be **unconditionally stable** if its region of absolute stability contains the entire left halfplane, that is, $\mathbb{C}^- \subset \mathcal{A}$, and it is **conditionally stable** otherwise.

Region of absolute stability of forward Euler method

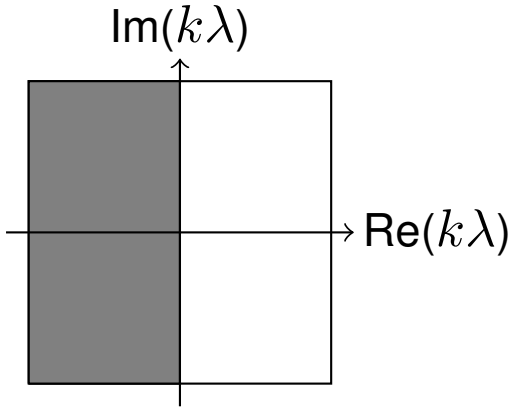
- The forward Euler method uses $u_{n+1} = u_n + k\lambda u_n$.
- Hence, with $r(k\lambda) = \frac{u_{n+1}}{u_n} = (1 + k\lambda)$, the region of absolute stability is



$$\mathcal{A} = \Big\{ k\lambda \in \mathbb{C} \ : \ |r(k\lambda)| < 1 \Big\} = \Big\{ k\lambda \in \mathbb{C} \ : \ |1 + k\lambda| < 1 \Big\}.$$

As a conclusion, the forward Euler method is conditionally stable. The time step k must be small enough for $k\lambda$ to lie within the unit circle with center at $(-1, 0)$ for the forward Euler method to furnish a numerical solution with the same asymptotic behavior as the exact solution.

Region of absolute stability of trapezoidal method

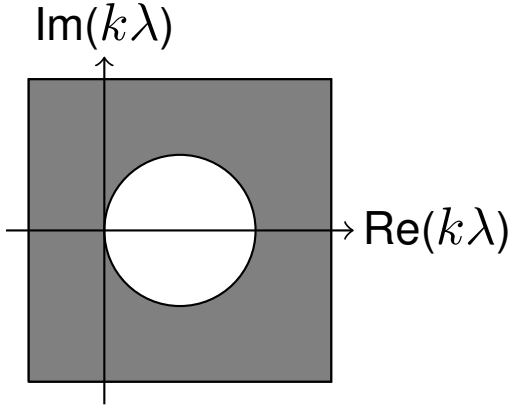


- The trapezoidal method uses $u_{n+1} = u_n + \frac{k}{2}(\lambda u_n + \lambda u_{n+1})$.
- Hence, with $r(k\lambda) = \frac{u_{n+1}}{u_n} = \frac{1+\frac{k\lambda}{2}}{1-\frac{k\lambda}{2}}$, the region of absolute stability is

$$\mathcal{A} = \left\{ k\lambda \in \mathbb{C} \ : \ |r(k\lambda)| < 1 \right\} = \mathbb{C}^-.$$

As a conclusion, because the region of absolute stability contains the entire left halfplane, the trapezoidal method is unconditionally stable.

Region of absolute stability of backward Euler method



- The backward Euler method uses $u_{n+1} = u_n + k\,\lambda u_{n+1}$.
- Hence, with $r(k\lambda) = \frac{u_{n+1}}{u_n} = \frac{1}{1-k\lambda}$, the region of absolute stability is

$$\mathcal{A} = \left\{ k\lambda \in \mathbb{C} \ : \ |r(k\lambda)| < 1 \right\} = \left\{ k\lambda \in \mathbb{C} \ : \ |1 - k\lambda| > 1 \right\}.$$

As a conclusion, because the region of absolute stability contains the entire left halfplane, the backward Euler method is unconditionally stable.

Advanced time-marching methods

Aforementioned time-marching methods

- Forward Euler method: explicit, convergent with order 1, conditionally stable.
- Trapezoidal method: implicit, convergent with order 2, unconditionally stable.
- Backward Euler method: implicit, convergent with order 1, unconditionally stable.

Advanced time-marching methods

- Improved time-marching methods (higher order, better stability properties,...) can be obtained by using more function evaluations (e.g. Runge-Kutta) or by using information from previous time steps (e.g. Adams-Bashforth and Adams-Moulton multistep methods).

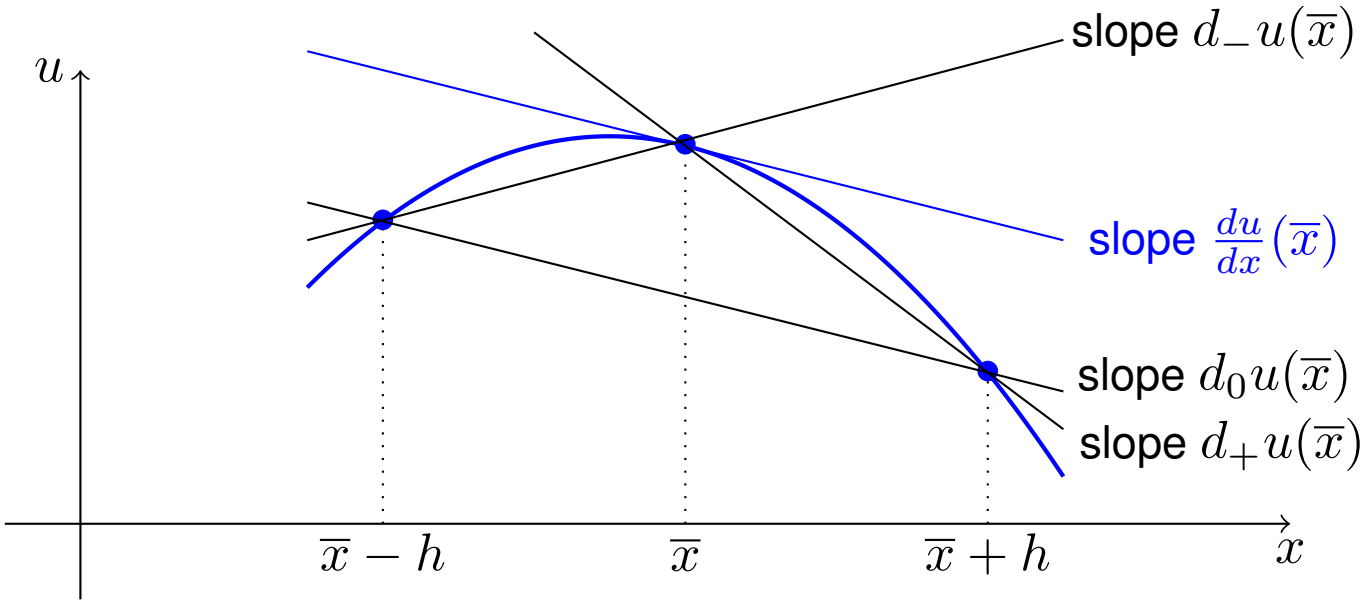
Stiff ODEs

- Stiff ODE problems involve slow and fast time scales simultaneously, and they may therefore risk requiring excessively small time steps over long time intervals.
- Stiff ODE problems require special attention, and dedicated time-marching methods must be used. Implicit methods are often used because they are often unconditionally stable.

Finite difference method for Laplace/Poisson equation

Notion of finite difference approximation

- Several finite difference approximations of $\frac{du}{dx}(\overline{x})$:



$$d_+u(\overline{x}) = \frac{u(\overline{x}+h)-u(\overline{x})}{h}, \quad d_-u(\overline{x}) = \frac{u(\overline{x})-u(\overline{x}-h)}{h}, \quad d_0u(\overline{x}) = \frac{u(\overline{x}+h)-u(\overline{x}-h)}{2h}.$$

- Similar finite difference approximations can be defined for higher order derivatives, for example,

$$\frac{d^2u}{dx^2}(\overline{x}) \approx d_0^2u(\overline{x}) = \frac{u(\overline{x}-h)-2u(\overline{x})+u(\overline{x}+h)}{h^2}.$$

- Similar finite difference approximations can be defined for partial derivatives in PDEs.

A simple finite difference method for Laplace/Poisson equation

- Let us consider the Dirichlet problem

$$\begin{cases} \dfrac{d^2u}{dx^2}(x) = f(x) & \text{for } 0 < x < 1, \\ u(0) = u(1) = 0 & \text{at } x = 0 \text{ and } x = 1. \end{cases}$$

- Let grid points $x_0,\ x_1,\ x_2,\ \ldots,\ x_{\mu_h}$ be introduced as follows:

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

- A simple **finite difference method** is then obtained by computing approximations $u_0,\ \ldots,\ u_{\mu_h}$ at the grid points $x_0,\ \ldots,\ x_{\mu_h}$ by requiring that

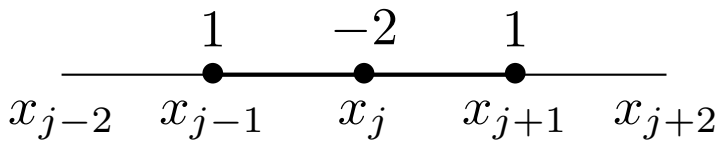
$$\begin{cases} \dfrac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = f(x_j) & \text{for } j = 1, \ldots, \mu_h - 1, \\ u_0 = u_{\mu_h} = 0 & \text{at } x_0 = 0 \text{ and } x_{\mu_h} = 1. \end{cases}$$

  This corresponds to replacing $\frac{d^2u}{dx^2}(x_j)$ by its finite difference approximation $d_0^2 u(x_j)$ in the PDE.

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

Notion of stencil

- The **stencil** is a graphical representation of the finite difference approximation being used.
- For the aforementioned finite difference method, we have the stencil



Linear problem defined by aforementioned finite difference method

- The linear problem provided by the aforementioned finite difference method can be written as

$$\underbrace{\frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}}_{[A]} \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{\mu_h-2} \\ u_{\mu_h-1} \end{bmatrix}}_{\boldsymbol{u}^h} = \underbrace{\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{\mu_h-2}) \\ f(x_{\mu_h-1}) \end{bmatrix}}_{\boldsymbol{f}};$$

hence, more compactly,

$$[A]\boldsymbol{u}^h = \boldsymbol{f}.$$

Properties of linear problem that must be solved in aforementioned finite difference method

- The matrix $[A]$ obtained in the aforementioned finite difference method is **negative definite**,

$$\begin{aligned}\boldsymbol{v}^h \cdot [A]\boldsymbol{v}^h &= \frac{1}{h^2}\left(\sum_{j=1}^{\mu_h-1}(-2)v_j^2 + 2\sum_{j=2}^{\mu_h-1} v_j v_{j-1}\right)\\ &= \frac{1}{h^2}\left(-\sum_{j=2}^{\mu_h-1}(v_j - v_{j-1})^2 - v_1^2 - v_{\mu_h-1}^2\right)\\ &< 0 \quad \text{for all } \boldsymbol{v}^h \neq \mathbf{0},\end{aligned}$$

  so that $[A]$ is invertible; hence, the numerical solution exists and is unique.

- The matrix $[A]$ is **sparse** and **symmetric**. See MATH0471 “Multiphysics integrated computational project” (R. Boman and C. Geuzaine) for details on appropriate storage and solution algorithms.

- The matrix $[A]$ can be **large** and **ill-conditioned**, especially if h is small. See INFO0939 “High-performance scientific computing” (C. Geuzaine) for details on appropriate solution algorithms.

Notions of consistency and convergence

- The **local truncation error** τ_j at x_j is obtained by inserting the exact solution into the finite difference equation and determining by how much it fails to satisfy this finite difference equation.
- A finite difference method is **consistent** if
$$\lim_{h\to 0}\left(\max_{1\le j\le \mu_h-1}\left|\tau_j\right|\right)=0.$$
- A finite difference method is **convergent** if
$$\lim_{h\to 0}\left(\max_{1\le j\le \mu_h-1}\left|u(x_j)-u_j\right|\right)=0.$$
- A finite difference method is **convergent with order** p if
$$\max_{1\le j\le \mu_h-1}\left|u(x_j)-u_j\right|=O(h^p).$$
- We note that other ways of gauging the magnitude of the local truncation and the global error can also be considered. For example, instead of using the "max-norm" $\|\boldsymbol{v}^h\|_\infty=\max_{1\le j\le \mu_h-1}|v_j|$, consistency and convergence can also be defined by using the "1-norm" or the "2-norm."
- As in the case of ODEs, consistency is relevant to the local truncation error, and convergence is relevant to the global error between the exact and the numerical solution.
- As in the case of ODEs, consistency is typically established by using Taylor series, and the connection between consistency and convergence is typically made by using some form of stability.

Consistency of aforementioned finite difference method

- The aforementioned finite difference method is consistent.

Proof of consistency of aforementioned finite difference method:

- The local truncation error τ_j at x_j is obtained as

$$\tau_j = \frac{1}{h^2}\big(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\big) - f(x_j).$$

- Because the exact solution satisfies the PDE $\frac{d^2u}{dx^2}(x) = f(x)$, we obtain

$$\tau_j = \frac{1}{h^2}\big(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\big) - \frac{d^2u}{dx^2}(x_j).$$

- If the exact solution u is sufficiently regular, the remainder in its degree-3 Taylor series satisfies

$$u(x_{j+1}) = u(x_j + h) = u(x_j) + h\,\frac{du}{dx}(x_j) + \frac{1}{2}\,h^2\,\frac{d^2u}{dx^2}(x_j) + \frac{1}{6}\,h^3\,\frac{d^3u}{dx^3}(x_j) + O(h^4),$$

$$u(x_{j-1}) = u(x_j - h) = u(x_j) - h\,\frac{du}{dx}(x_j) + \frac{1}{2}\,h^2\,\frac{d^2u}{dx^2}(x_j) - \frac{1}{6}\,h^3\,\frac{d^3u}{dx^3}(x_j) + O(h^4),$$

- Hence, by combining these results, we find

$$\tau_j = O(h^2).$$

As a conclusion, the aforementioned finite difference method is consistent, as asserted.

Stability of aforementioned finite difference method

- The aforementioned finite difference method is stable in that there exists a grid spacing h_0 and a constant $c > 0$ such that for all grid spacings h smaller than h_0, with c independent of h, we have

$$\max_{1 \le j \le \mu_h - 1} \left| u(x_j) - u_j \right| \le c \max_{1 \le j \le \mu_h - 1} \left| \tau_j \right|.$$

Proof of stability of aforementioned finite difference method:

- With reference to the definition of the local truncation error, the exact solution satisfies

$$\frac{1}{h^2}\big(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\big) = f(x_j) + \tau_j, \quad 1 \le j \le \mu_h - 1,$$

and the numerical solution satisfies

$$\frac{1}{h^2}(u_{j-1} - 2u_j + u_{j+1}) = f(x_j), \quad 1 \le j \le \mu_h - 1.$$

- Subtracting these equations from each other, we obtain

$$\frac{1}{h^2}\Big(\big(u(x_{j-1}) - u_{j-1}\big) - 2\big(u(x_j) - u_j\big) + \big(u(x_{j+1}) - u_{j+1}\big)\Big) = \tau_j, \quad 1 \le j \le \mu_h - 1.$$

We can observe that the global error satisfies a system of equations that has exactly the same form as the original system except that the right-hand side is given by the local truncation error.

Proof of stability of aforementioned finite difference method (continued):

- In fact, the previous system of equations can be written equivalently as

$$\frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \begin{bmatrix} u(x_1)-u_1 \\ u(x_2)-u_2 \\ \vdots \\ u(x_{\mu_h-2})-u_{\mu_h-2} \\ u(x_{\mu_h-1})-u_{\mu_h-1} \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{\mu_h-2} \\ \tau_{\mu_h-1} \end{bmatrix}$$

- After solving

$$\frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \begin{bmatrix} g_1^1 \\ g_2^1 \\ \vdots \\ g_{\mu_h-2}^1 \\ g_{\mu_h-1}^1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \ldots, \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \begin{bmatrix} g_1^{\mu_h-1} \\ g_2^{\mu_h-1} \\ \vdots \\ g_{\mu_h-2}^{\mu_h-1} \\ g_{\mu_h-1}^{\mu_h-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

we can establish the superposition formula

$$\begin{bmatrix} u(x_1)-u_1 \\ u(x_2)-u_2 \\ \vdots \\ u(x_{\mu_h-2})-u_{\mu_h-2} \\ u(x_{\mu_h-1})-u_{\mu_h-1} \end{bmatrix} = \tau_1 \begin{bmatrix} g_1^1 \\ g_2^1 \\ \vdots \\ g_{\mu_h-2}^1 \\ g_{\mu_h-1}^1 \end{bmatrix} + \ldots + \tau_{\mu_h-1} \begin{bmatrix} g_1^{\mu_h-1} \\ g_2^{\mu_h-1} \\ \vdots \\ g_{\mu_h-2}^{\mu_h-1} \\ g_{\mu_h-1}^{\mu_h-1} \end{bmatrix}.$$

Proof of stability of aforementioned finite difference method (continued):

- It can be readily verified that

$$g_j^i = h\,G(x_j, x_i), \quad \text{where} \quad G(x_j, x_i) = \begin{cases} x_j(x_i - 1) & \text{if } j \le i, \\ x_i(x_j - 1) & \text{if } j \ge i, \end{cases}$$

  where G is the Green's function for the Dirichlet problem on $]0, 1[$ introduced in Lecture 2.

- Using the aforementioned superposition formula, we thus obtain

$$\begin{aligned} |u(x_j) - u_j| &= \sum_{i=1}^{\mu_h - 1} \tau_i\, h\, G(x_j, x_i) \\ &\le \Big(\max_{1 \le i \le \mu_h - 1} |\tau_i| \Big) h \sum_{i=1}^{\mu_h - 1} |G(x_j, x_i)| \\ &\le \Big(\max_{1 \le i \le \mu_h - 1} |\tau_i| \Big) \underbrace{h\mu_h}_{=1} \underbrace{\frac{x_j(1 - x_j)}{2}}_{\le \frac{1}{8} \text{ for } 0 \le x_j \le 1} \end{aligned}$$

  Hence, with $c = \frac{1}{8}$, we have $\max_{1 \le j \le \mu_h - 1} \big|u(x_j) - u_j\big| \le c \max_{1 \le j \le \mu_h - 1} \big|\tau_j\big|$, as asserted.

Proof of stability of aforementioned finite difference method (continued):

- In going from the second to the third line, we used the partial sum formula $\sum_{m=1}^{n} m = \frac{n(n+1)}{2}$ as

$$\sum_{i=1}^{\mu_h-1} |G(x_j,x_i)| = \underbrace{\sum_{i=1}^{j} |G(x_j,x_i)|}_{\substack{=\sum_{i=1}^{j} x_i(1-x_j) \\ =\frac{(j+1)x_j}{2}(1-x_j)}} + \underbrace{\sum_{i=j+1}^{\mu_h-1} |G(x_j,x_i)|}_{\substack{=\sum_{i=j+1}^{\mu_h-1} x_j(1-x_i) \\ =\frac{(\mu_h-(j+1))x_j}{2}(1-x_j)}} .$$

Convergence of aforementioned finite difference method

- The aforementioned finite difference method is convergent with order 2.

Proof of convergence of aforementioned finite difference method:

- It follows from $\tau_j = O(h^2)$, $1 \leq j \leq \mu_h - 1$ and the stability property that

$$\max_{1\leq j\leq \mu_h-1} \left|u(x_j) - u_j\right| = O(h^2).$$

- Observe that the properties of **consistency** and **stability** collectively imply **convergence**.

Conclusion

- Key properties of a numerical method:
  - Convergence.
  - It may also be desirable that certain properties (existence, uniqueness, maximum property,...) of the exact solution are preserved in the numerical solution.

- A finite difference method is obtained by replacing one or more partial derivatives in a PDE by finite difference approximations. By doing this for a grid of points, a system of equations is obtained, whose solution provides an approximation to the exact solution at the grid points.

- A convergence analysis of a finite difference method typically begins by examining the local one-step/truncation error by using Taylor series. Then, some form of stability is used to relate the local one-step/truncation error to the global error and establish convergence.

- 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. Illustrative numerical examples will also be included in the homework.

References

Suggested reading material:

- P. Olver. Introduction to Partial Differential Equations. Springer, 2014. Sections 5.1 and 5.5.

References also consulted to prepare this lecture:

- S. Benzoni-Gavage. Calcul différentiel et équations différentielles. Dunod, 2010.
- H. Cartan. Cours de calcul différentiel. Editions Hermann, 2007.
- J.-P. Demailly. Analyse numérique et équations différentielles. Collections Grenoble Sciences, 2006.
- D. Euvrard. Résolution numérique des équations aux dérivées partielles. Masson, 1994.
- A. Iserles. A first course in the numerical analysis of differential equations. Cambridge, 2009.
- R. LeVeque. Finite difference methods for ordinary and partial differential equations. SIAM, 2007.
- A. Quarteroni, R. Sacco, and F. Saleri. Numerical Mathematics, Springer, 2010.