Numerical methods for Partial Differential Equations

Adriano Festa
Politecnico of Turin, Italy

Undergraduate Lecture at KSU
2022

Table of Contents

Consider the initial-value probem (also known as *Cauchy problem*)

$$\begin{cases} \boldsymbol{u}' = \boldsymbol{F}(\boldsymbol{u}, t) \ , & 0 < t \leq T \ , \\ \boldsymbol{u}(0) = \boldsymbol{u}_0 \ , \end{cases} \tag{62}$$

for a system of n first-order differential equations in n unknowns.

Here $\boldsymbol{u} : [0, T] \to \mathbb{R}^n$ is the vector, depending upon time in a differentiable way, that collects the unknowns, whereas $\boldsymbol{F} : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$ is a globally continuous function, which is *Lipschitz-continuous* in the variable $\boldsymbol{u}$ uniformly in t, i.e., there exists a constant $L > 0$ such that

$$\|\boldsymbol{F}(\boldsymbol{u}, t) - \boldsymbol{F}(\boldsymbol{v}, t)\| \leq L \|\boldsymbol{u} - \boldsymbol{v}\| \ , \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n \ , \quad \forall t \in [0, T] \ .$$

For instance, this happens if

$$\boldsymbol{F}(\boldsymbol{u}, t) = \mathcal{A}\boldsymbol{u} + \boldsymbol{b}(t) \ ,$$

since

$$\|\boldsymbol{F}(\boldsymbol{u}, t) - \boldsymbol{F}(\boldsymbol{v}, t)\| = \|\mathcal{A}(\boldsymbol{u} - \boldsymbol{v})\| \leq \|\mathcal{A}\| \, \|\boldsymbol{u} - \boldsymbol{v}\| \ .$$

Under the assumptions made on $\boldsymbol{F}$, problem (62) admits one and only one solution for any choice of the initial data $\boldsymbol{u}_0 \in \mathbb{R}^n$.

For the numerical approximation of such a solution, the most popular strategy consists in resorting to a *time-advancing scheme*.

This means that one chooses (in various manners) $K > 0$ time instants t_k, with

$$0 = t_0 < t_1 < \cdots < t_k < t_{k+1} < \cdots < t_K = T \ ;$$

for each of them, one recursively defines an approximation

$$\boldsymbol{u}^k \simeq \boldsymbol{u}(t_k)$$

of the exact solution; precisely, at the time t_{k+1}, one defines $\boldsymbol{u}^{k+1}$ using the knowledge of the approximations $\boldsymbol{u}^j$, $j \leq k$ already computed at one or more previous time instants.

The conceptually simplest situation occurs when one advances with a constant time step $\Delta t > 0$ (in this case, one sets $K = T/\Delta t$, assuming this ratio to be integer); the time instants are then defined by

$$t_k = k\Delta t \ , \qquad \text{for } k = 0, 1, \ldots, K \ .$$

Explicit Euler scheme (EE): (*explicit - one step - I order*)

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t \boldsymbol{F}(\boldsymbol{u}^k, t_k) \, , \qquad k \geq 0 \, ;$$

motivated by the *forward incremental quotient* formula

$$\frac{\boldsymbol{u}(t_{k+1}) - \boldsymbol{u}(t_k)}{\Delta t} \simeq \boldsymbol{u}'(t_k) = \boldsymbol{F}(\boldsymbol{u}(t_k), t_k) \, .$$

Implicit Euler scheme (IE): (*implicit - one step - I order*)

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^{k} + \Delta t \boldsymbol{F}(\boldsymbol{u}^{k+1}, t_{k+1}) \,, \qquad k \geq 0 \,;$$

motivated by the *backward incremental quotient* formula

$$\frac{\boldsymbol{u}(t_{k+1}) - \boldsymbol{u}(t_k)}{\Delta t} \simeq \boldsymbol{u}'(t_{k+1}) = \boldsymbol{F}(\boldsymbol{u}(t_{k+1}), t_{k+1})$$

Mid-point scheme (MP): (*explicit - two steps - II order*)

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^{k-1} + 2\Delta t \boldsymbol{F}(\boldsymbol{u}^k, t_k) \,, \qquad k \geq 1 \,;$$

motivated by the *centered incremental quotient* formula

$$\frac{\boldsymbol{u}(t_{k+1}) - \boldsymbol{u}(t_{k-1})}{2\Delta t} \simeq \boldsymbol{u}'(t_k) = \boldsymbol{F}(\boldsymbol{u}(t_k), t_k)$$

Trapezoidal (or Crank-Nicolson) scheme (CN): (*implicit - one step - II order*)

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t \left[\tfrac{1}{2} \boldsymbol{F}(\boldsymbol{u}^k, t_k) + \tfrac{1}{2} \boldsymbol{F}(\boldsymbol{u}^{k+1}, t_{k+1}) \right] , \qquad k \geq 0 .$$

motivated by a *numerical quadrature formula* applied to the relation

$$\boldsymbol{u}(t_{k+1}) - \boldsymbol{u}(t_k) = \int_{t_k}^{t_{k+1}} \boldsymbol{u}'(t)\, dt = \int_{t_k}^{t_{k+1}} \boldsymbol{F}(\boldsymbol{u}(t), t)\, dt ,$$

obtained by integrating the identity $\boldsymbol{u}'(t) = \boldsymbol{F}(\boldsymbol{u}(t), t)$ on the interval $[t_k, t_{k+1}]$. The integral on the right-hand side can be approximated by the *trapezoidal formula*

$$\int_{t_k}^{t_{k+1}} \boldsymbol{F}(\boldsymbol{u}(t), t)\, dt \simeq \Delta t \left[\tfrac{1}{2} \boldsymbol{F}(\boldsymbol{u}(t_k), t_k) + \tfrac{1}{2} \boldsymbol{F}(\boldsymbol{u}(t_{k+1}), t_{k+1}) \right] .$$

Runge-Kutta scheme RK3: (*explicit - one-step - III order*)

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t \left[\tfrac{2}{9}\boldsymbol{F}_1 + \tfrac{1}{3}\boldsymbol{F}_2 + \tfrac{4}{9}\boldsymbol{F}_3 \right] , \qquad k \geq 0 ,$$

where the addends on the right-hand side are recursively defined by

$$\begin{aligned}
\boldsymbol{F}_1 &= \boldsymbol{F}(\boldsymbol{u}^k, t_k) , \\
\boldsymbol{F}_2 &= \boldsymbol{F}(\boldsymbol{u}^k + \tfrac{1}{2}\Delta t \boldsymbol{F}_1, t_k + \tfrac{1}{2}\Delta t) , \\
\boldsymbol{F}_3 &= \boldsymbol{F}(\boldsymbol{u}^k + \tfrac{3}{4}\Delta t \boldsymbol{F}_2, t_k + \tfrac{3}{4}\Delta t) .
\end{aligned}$$

Runge-Kutta scheme RK4: (*explicit - one-step - IV order*)

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t \left[\tfrac{1}{6}\boldsymbol{F}_1 + \tfrac{1}{3}\boldsymbol{F}_2 + \tfrac{1}{3}\boldsymbol{F}_3 + \tfrac{1}{6}\boldsymbol{F}_4\right] , \qquad k \geq 0 ,$$

where the addends on the right-hand side are recursively defined by

$$\begin{aligned}
\boldsymbol{F}_1 &= \boldsymbol{F}(\boldsymbol{u}^k, t_k) , \\
\boldsymbol{F}_2 &= \boldsymbol{F}(\boldsymbol{u}^k + \tfrac{1}{2}\Delta t \boldsymbol{F}_1, t_k + \tfrac{1}{2}\Delta t) , \\
\boldsymbol{F}_3 &= \boldsymbol{F}(\boldsymbol{u}^k + \tfrac{1}{2}\Delta t \boldsymbol{F}_2, t_k + \tfrac{1}{2}\Delta t) , \\
\boldsymbol{F}_4 &= \boldsymbol{F}(\boldsymbol{u}^k + \Delta t \boldsymbol{F}_3, t_k + \Delta t) .
\end{aligned}$$

"Backward Difference Formula" scheme BDF2: (*implicit - two-steps - II order*)

$$\tfrac{3}{2}\boldsymbol{u}^{k+1} - 2\boldsymbol{u}^{k} + \tfrac{1}{2}\boldsymbol{u}^{k-1} = \Delta t \boldsymbol{F}(\boldsymbol{u}^{k+1}, t_{k+1}) \,, \qquad k \geq 1 \,,$$

"Backward Difference Formula" scheme BDF3: (*implicit - three-steps - III order*)

$$\tfrac{11}{6}\boldsymbol{u}^{k+1} - 3\boldsymbol{u}^{k} + \tfrac{3}{2}\boldsymbol{u}^{k-1} - \tfrac{1}{3}\boldsymbol{u}^{k-2} = \Delta t \boldsymbol{F}(\boldsymbol{u}^{k+1}, t_{k+1}) \,, \qquad k \geq 2 \,.$$

An explicit one-step scheme has the general form

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t \boldsymbol{\Phi}(\boldsymbol{u}^k, t_k, \Delta t) \ , \qquad k \geq 0 \ ,$$

whereas an implicit one-step scheme has the general form

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t \boldsymbol{\Psi}(\boldsymbol{u}^k, \boldsymbol{u}^{k+1}, t_k, \Delta t) \ , \qquad k \geq 0 \ .$$

The connection with the differential equation is given by the consistency condition:

Definition

A one-step scheme, defined by one of the previous formulas, is said to be **consistent** if the condition

$$\boldsymbol{\Phi}(\boldsymbol{u}, t, 0) = \boldsymbol{F}(\boldsymbol{u}, t) \qquad \text{or} \qquad \boldsymbol{\Psi}(\boldsymbol{u}, \boldsymbol{u}, t, 0) = \boldsymbol{F}(\boldsymbol{u}, t)$$

is satisfied for any $\boldsymbol{u} \in \mathbb{R}^n$ and $t \in [0, T]$.

All the previously presented one-step schemes are consistent.

Definition

A consistent one-step scheme is said to be **of order** $p > 0$ if, considering the solution $\boldsymbol{u}^1$ produced by the scheme at the first instant $t_1 = \Delta t$, one has

$$\|\boldsymbol{u}(t_1) - \boldsymbol{u}^1\| = O(\Delta t^{p+1}) \qquad \text{for } \Delta t \to 0 \ ,$$

for all solutions $\boldsymbol{u}$ that are $(p+1)$-times differentiable at $t = 0$.

Theorem

If a one-step scheme is consistent and of order p, and if the exact solution $\boldsymbol{u}(t)$ is $(p+1)$-times differentiable in $[0, T]$, one has

$$\max_{1 \le k \le K} \|\boldsymbol{u}(t_k) - \boldsymbol{u}^k\| \le C_{L,T} \Delta t^p \max_{t \in [0,T]} \left\| \frac{d^{p+1}\boldsymbol{u}}{dt^{p+1}}(t) \right\| \ ,$$

where $C_{L,T}$ denotes a constant only depending on L, T and the numerical scheme.

The theorem guarantees the *convergence* of the scheme, i.e., the fact that the discrete solution generated by the numerical scheme converges towards the exact solution as $\Delta t \to 0$; it also predicts the behaviour of the error as the time step decreases.

In applications, it is important that a time-advancing scheme generates discrete solutions $\boldsymbol{u}^k$ which stay bounded as $t_k \to +\infty$, whenever the exact solutions of the initial-value problem (62) stay bounded as $t \to +\infty$.

A quite relevant situation occurs when the differential system is

- *linear* and *autonomous*, i.e., one has $\boldsymbol{F}(\boldsymbol{u}, t) = \boldsymbol{F}(\boldsymbol{u}) = \mathcal{A}\boldsymbol{u} + \boldsymbol{b}$ with $\mathcal{A}$ square matrix of order n and $\boldsymbol{b} \in \mathbb{R}^n$ independent of time, and
- *dissipative*, i.e., $\mathcal{A}$ is diagonalizable with eigenvalues all having real part < 0.

In this case, whichever is the initial datum $\boldsymbol{u}_0$, the solution $\boldsymbol{u}(t)$ of the problem

$$\begin{cases} \boldsymbol{u}' = \mathcal{A}\boldsymbol{u} + \boldsymbol{b} \,, & t > 0 \,, \\ \boldsymbol{u}(0) = \boldsymbol{u}_0 \,, \end{cases} \tag{63}$$

converges, as $t \to +\infty$, towards the *steady state* $\boldsymbol{u}_\infty \in \mathbb{R}^n$, solution of the linear system

$$\mathcal{A}\boldsymbol{u}_\infty + \boldsymbol{b} = \boldsymbol{0} \,;$$

hence, in particular, it stays bounded as $t \to +\infty$.

Indeed, applying the change of dependent variable $\boldsymbol{v}(t) = \boldsymbol{u}(t) - \boldsymbol{u}_\infty$ and setting $\boldsymbol{v}_0 = \boldsymbol{u}_0 - \boldsymbol{u}_\infty$, one immediately checks that $\boldsymbol{v}$ is the solution of the homogeneous Cauchy problem

$$\begin{cases} \boldsymbol{v}' = \mathcal{A}\boldsymbol{v} \ , & t > 0 \ , \\ \boldsymbol{v}(0) = \boldsymbol{v}_0 \ . \end{cases} \tag{64}$$

Let $\mathcal{A} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^{-1}$ be the diagonalization of the matrix $\mathcal{A}$, where $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix collecting the eigenvalues and $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ the matrix collecting the eigenvectors (placed column-wise).
Let us substitute such expression in (64) and let us multiply on the left by $\boldsymbol{W}^{-1}$. Setting $\boldsymbol{z}(t) = \boldsymbol{W}^{-1}\boldsymbol{v}(t)$ and $\boldsymbol{z}_0 = \boldsymbol{W}^{-1}\boldsymbol{v}_0$, problem (64) is equivalent to

$$\begin{cases} \boldsymbol{z}' = \boldsymbol{\Lambda}\boldsymbol{z} \ , & t > 0 \ , \\ \boldsymbol{z}(0) = \boldsymbol{z}_0 \ , \end{cases}$$

namely, to the n independent scalar problems in the components z_p, $1 \le p \le n$, of the vector $\boldsymbol{z}$

$$\begin{cases} z_p' = \lambda_p z_p \ , & t > 0 \ , \\ z_p(0) = z_{0p} \ . \end{cases}$$

The solutions of the n independent scalar problems

$$\begin{cases} z_p' = \lambda_p z_p \ , & t > 0 \ , \\ z_p(0) = z_{0p} \ , \end{cases} \tag{65}$$

are given by

$$z_p(t) = \mathrm{e}^{\lambda_p t} z_{0p}$$

and since by assumption $\mathbb{R}e\, \lambda_p < 0$, one has

$$|z_p(t)| \to 0 \qquad \text{for } t \to +\infty \quad \text{and for each } p \ .$$

As a consequence, for $t \to +\infty$ one has $\|\boldsymbol{z}(t)\| \to 0$, hence, also $\|\boldsymbol{u}(t) - \boldsymbol{u}_\infty\| = \|\boldsymbol{v}(t)\| = \|\boldsymbol{W}\boldsymbol{z}(t)\| \to 0$, which means, as anticipated, $\boldsymbol{u}(t) \to \boldsymbol{u}_\infty$.

Thus, we are led to consider the generic scalar problem introduced above, i.e.,

$$\begin{cases} z' = \lambda z \ , & t > 0 \ , \\ z(0) = z_0 \ , \end{cases} \tag{66}$$

and to ask ourselves under which conditions a time-advancing scheme applied to such a problem produces discrete solutions z^k which stay bounded as $k \to \infty$.

The asymptotic stability of the Explicit Euler scheme (I)

The explicit Euler scheme applied to the equation $z' = \lambda z$ yields

$$z^{k+1} = z^k + \Delta t \lambda z^k = (1 + \Delta t \lambda) z^k \ , \qquad k \geq 0 \ ,$$

hence by recursion, and keeping into account the initial condition, one gets the explicit expression

$$z^k = (1 + \Delta t \lambda)^{\wedge k} z_0 \ ,$$

where the symbol $^{\wedge k}$ means raising the basis to the k-th power.
The condition $|z^k| \to 0$ as $k \to +\infty$ is then equivalent to the condition

$$|1 + \Delta t \lambda| < 1 \ .$$

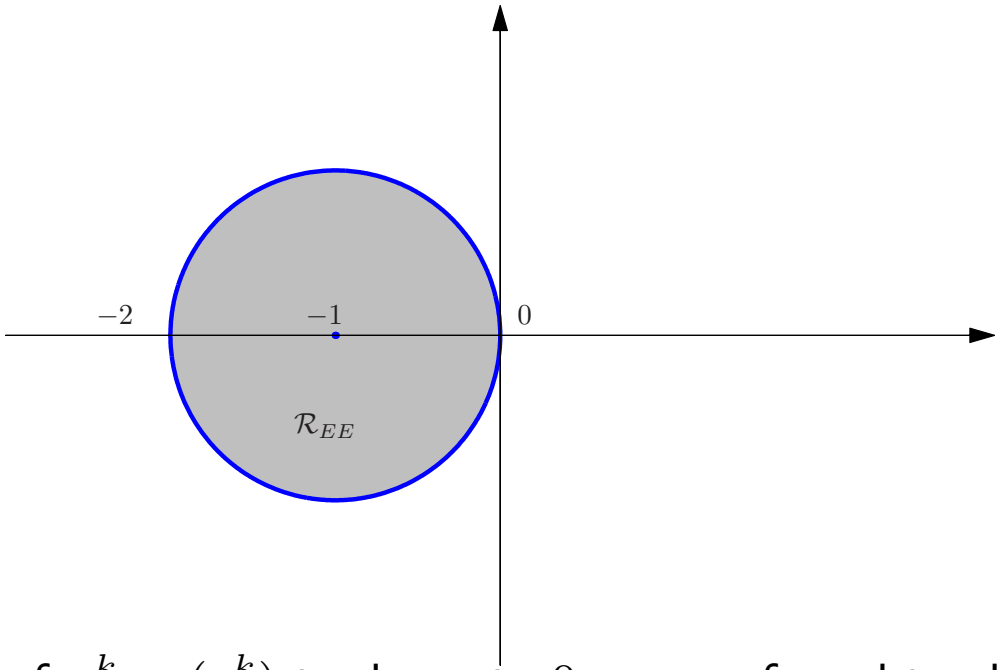
On the other hand, if instead one has $|1 + \Delta t \lambda| = 1$, then the solution stays bounded as $k \to +\infty$, although it does not tend to 0; indeed, one has $|z^k| = |z_0|$ for each $k \geq 0$. At last, if $|1 + \Delta t \lambda| > 1$, then necessarily one has $|z^k| \to +\infty$ as $k \to +\infty$.

In conclusion, setting $\alpha = \Delta t \lambda \in \mathbb{C}$, boundedness of the discrete solutions is equivalent to

$$|1 + \alpha| \leq 1 \ .$$

The asymptotic stability of the Explicit Euler scheme (II)

The inequality $|1+\alpha| \leq 1$ defines, in the complex plane, a circle of center -1 and radius 1. This region is termed the *region of asymptotic stability* of the explicit Euler scheme, and denoted by $\mathcal{R}_{EE}$. Its internal part will be denoted by $\text{int}\,(\mathcal{R}_{EE})$.



Thus, if we want all the components of $\boldsymbol{z}^k = (z_p^k)$ to decay to 0, we are forced to choose Δt in such a way that the condition

$$\Delta t \lambda_p \in \text{int}\,(\mathcal{R}_{EE}) \quad \text{for each } p = 1, \ldots, n\ , \tag{67}$$

is satisfied; this is always possible, thanks to the assumption $\mathbb{R}e\lambda_p < 0$ on the eigenvalues of $\mathcal{A}$.

The explicit Euler scheme is therefore **conditionally asymptotically stable**, i.e., it is necessary to satisfy this condition on Δt, termed the **asymptotic stability condition**, in order to get the desired behaviour of the discrete solutions. It may occur, however, that such a condition is overly restrictive in practice.

The asymptotic stability of the Explicit Euler scheme (III)

In the important case where the matrix $\mathcal{A}$ is symmetric, and consequently all its eigenvalues are real and negative, we have $\Delta t \lambda_p \in \text{int}\,(\mathcal{R}_{EE})$ if and only if

$$-2 < \Delta t \lambda_p < 0\,;$$

the second inequality is always fulfilled, whereas the first one is equivalent to

$$\Delta t < \frac{2}{|\lambda_p|}\,.$$

Thus, the asymptotic stability condition (67) becomes

$$\Delta t < \frac{2}{\max_p |\lambda_p|}\,. \tag{68}$$

If the matrix $\mathcal{A}$ has eigenvalues with orders of magnitude quite different from each other (i.e., if it is *ill-conditioned* - in the language of systems of differential equations one says that the system is *stiff*), we are obliged to advance with a time-step dictated by the eigenvalue of largest absolute value, even if the actual behaviour of the exact solution would not require such a restriction.

(See next example.)

The asymptotic stability of the Implicit Euler scheme (I)

The implicit Euler scheme applied to the equation $z' = \lambda z$ yields

$$z^{k+1} = z^k + \Delta t \lambda z^{k+1} , \qquad k \geq 0 ,$$

i.e.,

$$(1 - \Delta t \lambda) z^{k+1} = z^k , \qquad k \geq 0 .$$

Thus, the scheme generates the sequence

$$z^k = \left(\frac{1}{1 - \Delta t \lambda} \right)^{\wedge k} z_0 , \qquad k \geq 0 ,$$

which converges to 0 as $k \to +\infty$ if and only if the quantity in parenthesis is smaller than 1 in absolute value.

Setting $\alpha = \Delta t \lambda$, this is equivalent to the condition

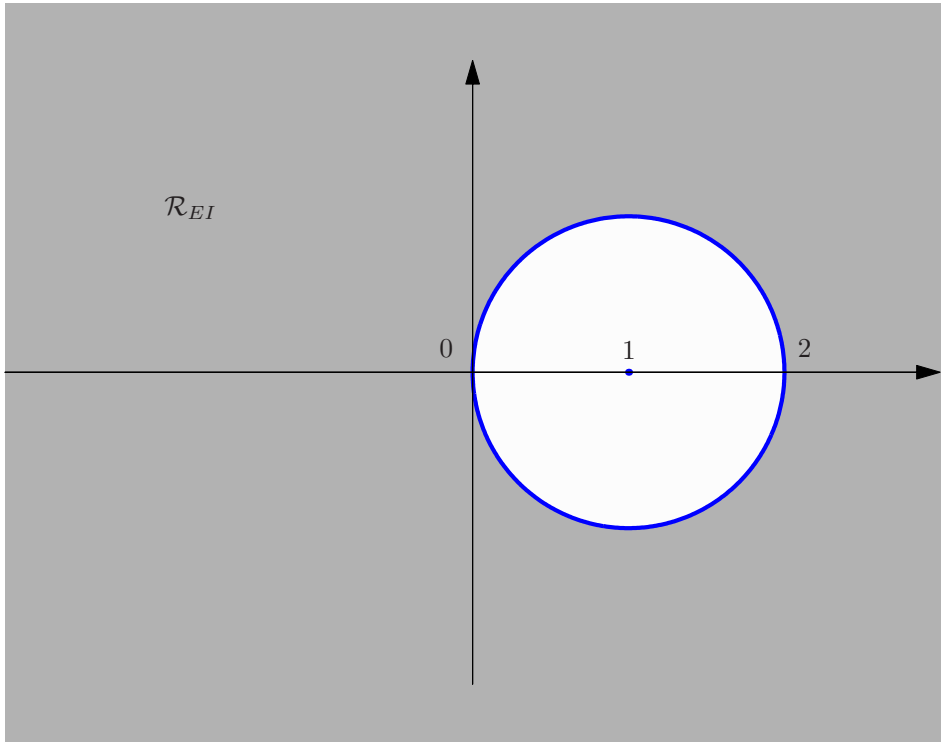
$$\frac{1}{|1 - \alpha|} < 1 ,$$

i.e.,

$$|1 - \alpha| > 1 .$$

The asymptotic stability of the Implicit Euler scheme (II)

Therefore, the *region of asymptotic stability* $\mathcal{R}_{EI}$ of the implicit Euler scheme is defined as the set of all complex numbers α such that $|1-\alpha| \geq 1$.



Since the real part of $\Delta t \lambda_p$ is strictly negative, we have

$$\Delta t \lambda_p \in \text{int}\,(\mathcal{R}_{EI}) \quad \text{for each } p = 1, \ldots, n \quad \text{and } \textit{whichever } \Delta t > 0 \textit{ is}\,.$$

Thus, there are no restrictions on the choice of the time step. We conclude that the implicit Euler scheme is **unconditionally asynptotically stable**.

The asymptotic stability of the trapezoidal scheme (I)

The trapezoidal (or Crank-Nicolson) scheme applied to the equation $z' = \lambda z$ yields

$$z^{k+1} = z^k + \Delta t \lambda \left(\frac{1}{2} z^k + \frac{1}{2} z^{k+1} \right) , \qquad k \geq 0 ,$$

namely,

$$(1 - \frac{\Delta t}{2} \lambda) z^{k+1} = (1 + \frac{\Delta t}{2} \lambda) z^k , \qquad k \geq 0 ,$$

whose exact solution is given by

$$z^k = \left(\frac{1 + \frac{\Delta t}{2} \lambda}{1 - \frac{\Delta t}{2} \lambda} \right)^{\wedge k} z_0 , \qquad k \geq 0 .$$

Setting $\alpha = \Delta t \lambda$, such a solution tends to 0 as $k \to +\infty$ if and only if one has

$$\left| 1 + \frac{\alpha}{2} \right| < \left| 1 - \frac{\alpha}{2} \right| .$$

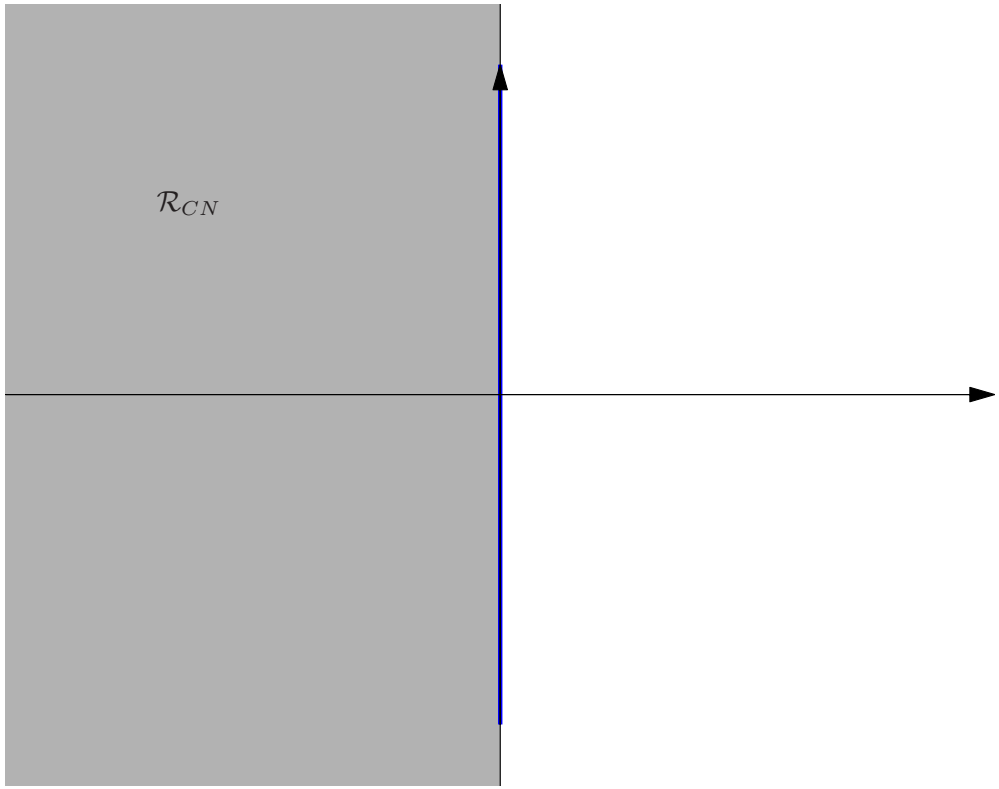
Now, it is easily seen that if α is any complex number, the inequality

$$|1 + \frac{\alpha}{2}| \leq |1 - \frac{\alpha}{2}|$$

is equivalent to $\mathbb{R}e\, \alpha \leq 0$.

The asymptotic stability of the trapezoidal scheme (II)

Therefore, the *region of asymptotic stability* $\mathcal{R}_{CN}$ of the trapezoidal scheme is give by the half-plane on the left of the imaginary axis; such a region coincides indeed with the region in which all the eigenvalues of the matrix $\mathcal{A}$ have to fall in order for the system (64) to have all its solutions bounded as $t \to +\infty$.



In this case, too, one has

$$\Delta t \lambda_p \in \text{int}\,(\mathcal{R}_{CN}) \quad \text{for each } p = 1, \ldots, n \quad \text{and whichever } \Delta t > 0 \text{ is ;}$$

hence, the trapezoidal scheme is **unconditionally asymptotically stable**.

The asymptotic stability of other schemes

- The Runge Kutta schemes presented above are all **conditionally asymptotically stable**. Their regions of asymptotic stability are larger that the one for the Explicit Euler scheme.
  (See plots in the Notes of the Course).

- The BDF schemes presented above have regions of asymptotic stability which contain a corner in the half-plane $\mathbb{R}\mathrm{e}\,\alpha \leq 0$, with center at the origin and symmetrically placed around the real negative semi-axis.
  Therefore, such methods are particularly suited for the discretization of *stiff* differential systems, as they turn out to be **unconditionally asymptotically stable** if the matrix of the system has all real and negative eigenvales.

An example of "stiff" system (I)

Consider the problem

$$\begin{cases} \boldsymbol{v}' = \mathcal{A}\boldsymbol{v} \ , & t > 0 \ , \\ \boldsymbol{v}(0) = \boldsymbol{v}_0 \ , \end{cases}$$

with

$$\mathcal{A} = -\begin{pmatrix} \frac{1001}{2} & \frac{999}{2} \\ \frac{999}{2} & \frac{1001}{2} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{v}_0 = \begin{pmatrix} 2 \\ 0 \end{pmatrix} .$$

The eigenvalues of the matrix are $\lambda_1 = -1$ and $\lambda_2 = -1000$, with corresponding eigenvectors given by $\boldsymbol{w}_1 = (1,\ 1)^T$ and $\boldsymbol{w}_2 = (1,\ -1)^T$. Thus, the exact solution can be written as

$$\begin{pmatrix} v_1(t) \\ v_2(t) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} z_1(t) \\ z_2(t) \end{pmatrix} ,$$

with $z_1(t) = \mathrm{e}^{-t}$ and $z_2(t) = \mathrm{e}^{-1000t}$, i.e., we obtain

$$\begin{cases} v_1(t) = \mathrm{e}^{-t} + \mathrm{e}^{-1000t} \ , \\ v_2(t) = \mathrm{e}^{-t} - \mathrm{e}^{-1000t} \ . \end{cases}$$

We thus have a very small transient, of the order of $1/1000$ seconds, followed by a much slower evolution, in the time scale of seconds.

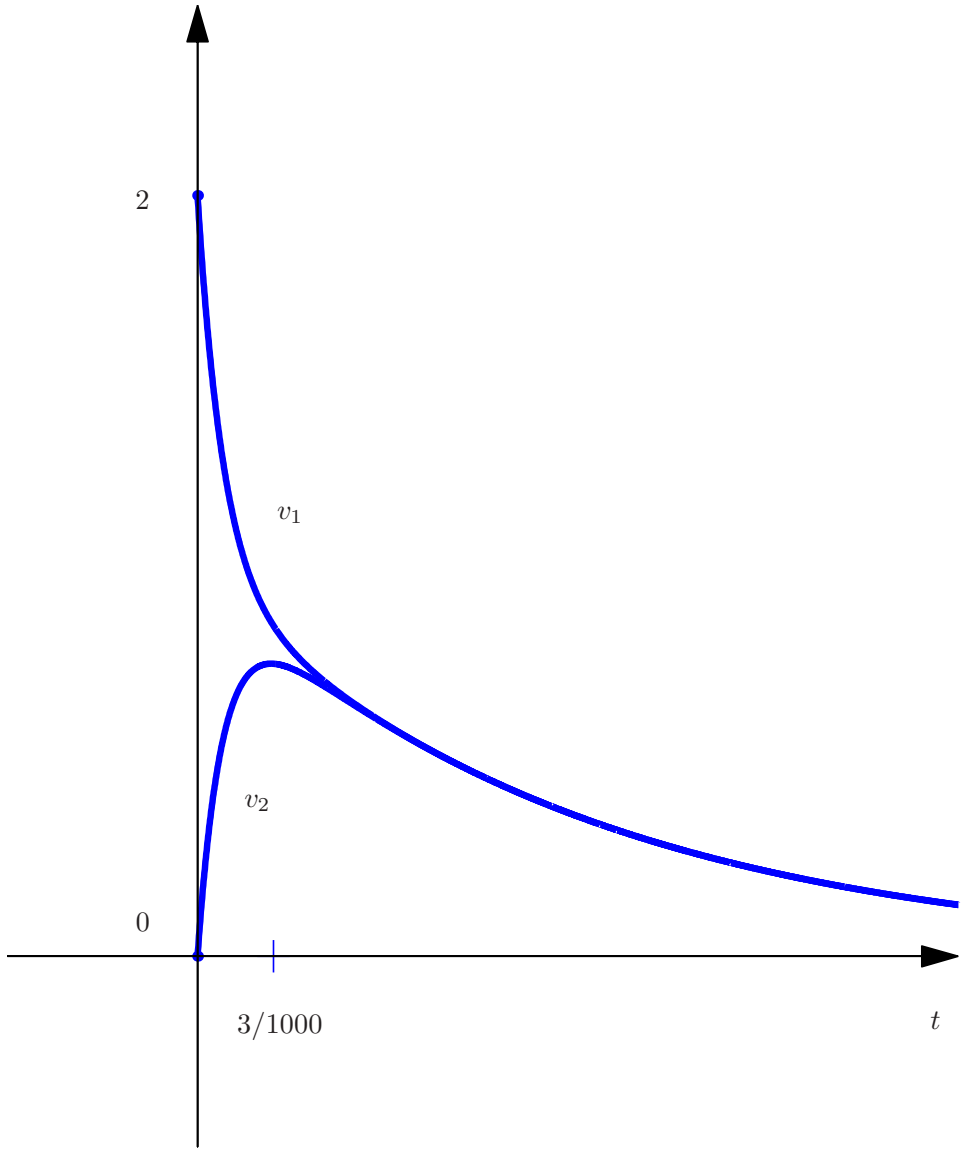
An example of "stiff" system (II)

If we advance in time with the explicit Euler scheme, we are forced to respect the asymptotic stability condition (68), namely

$$\Delta t < \frac{2}{\max(|-1|, |-1000|)} = \frac{1}{500} \ ,$$

not only during the transient period, but throughout the whole simulation, otherwise the computation quickly yields an overflow.

Using instead the implicit Euler scheme, or the more accurate trapezoidal scheme, allows us to vary the time-advancing step, choosing it in the order of $1/500$ seconds during the transient period, but subsequently letting it to increase as the solution goes towards a steady state.



Time discretization of the thermal model

Let us now discretize in time the problem

$$\begin{cases} \boldsymbol{B}\boldsymbol{u}' + \boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}(t) \,, & 0 < t \leq T \,, \\ \boldsymbol{u}(0) = \boldsymbol{u}_0 \,. \end{cases}$$

using one of the previously presented schemes, namely the Explicit Euler scheme, the Implicit Euler scheme and the Trapezoidal (or Crank-Nicolson) scheme.

To this end, it is convenient to write the differential system in the equivalent (normal) form

$$\boldsymbol{u}' = \boldsymbol{F}(\boldsymbol{u}, t) = -\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{u} + \boldsymbol{B}^{-1}\boldsymbol{f}(t) \qquad (= \mathcal{A}\boldsymbol{u} + \boldsymbol{b}(t)) \,.$$

Let us stress that such a transformation is only useful at the conceptual level, in order to apply the abstract time-advancing schemes in our specific setting. At the implemententation level, the multiplication by the inverse of $\boldsymbol{B}$ is almost invariably not efficient, hence, to be avoided; it is surely preferable to leave such matrix on the left-hand side of the equation, as shown in the sequel.

- The *Esplicit Euler* scheme yields

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k - \Delta t \boldsymbol{B}^{-1} \boldsymbol{A} \boldsymbol{u}^k + \Delta t \boldsymbol{B}^{-1} \boldsymbol{f}(t_k) \,, \qquad k \geq 0 \,,$$

which we re-formulate as

$$\boldsymbol{B} \boldsymbol{u}^{k+1} = (\boldsymbol{B} - \Delta t \boldsymbol{A}) \boldsymbol{u}^k + \Delta t \boldsymbol{f}(t_k) \,, \qquad k \geq 0 \,.$$

In the latter form, the scheme is not explicit, since at each iteration it requires the solution of a linear system with matrix $\boldsymbol{B}$. However, if we approximate $\boldsymbol{B}$ with the "lumped" mass matrix $\tilde{\boldsymbol{B}}$, which is diagonal, then the cost of computing $\boldsymbol{u}^{k+1}$ is essentially comparable to that of an explicit method.

In the sequel, we will see that the *asymptotic stability condition* for the Explicit Euler scheme becomes, in this case,

$$\Delta t \leq C \, h^2$$

(with C proportional to $\mu = \kappa / c$); this poses a restriction on the choice of the time-step, which often is not acceptable in practice.

The scheme is first-order accurate in time.

- The *Implicit Euler* scheme yields

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k - \Delta t \boldsymbol{B}^{-1} \boldsymbol{A} \boldsymbol{u}^{k+1} + \Delta t \boldsymbol{B}^{-1} \boldsymbol{f}(t_{k+1}) \,, \qquad k \geq 0 \,,$$

which we re-formulate as

$$(\boldsymbol{B} + \Delta t \boldsymbol{A}) \boldsymbol{u}^{k+1} = \boldsymbol{B} \boldsymbol{u}^k + \Delta t \boldsymbol{f}(t_{k+1}) \,, \qquad k \geq 0 \,.$$

In this case, one has to solve a linear system at each iteration, with matrix $\boldsymbol{B} + \Delta t \boldsymbol{A}$ symmetric and positive definite (since it is the sum of two matrices with these properties).
The gain over Explicit Euler is that the scheme is unconditionally stable.

The scheme is again first-order accurate in time.

- The *Trapezoidal* (or *Crank-Nicolson*) scheme yields, for our problem, the update

$$(\boldsymbol{B} + \tfrac{\Delta t}{2}\boldsymbol{A})\boldsymbol{u}^{k+1} = (\boldsymbol{B} - \tfrac{\Delta t}{2}\boldsymbol{A})\boldsymbol{u}^{k} + \tfrac{\Delta t}{2}(\,\boldsymbol{f}(t_k) + \boldsymbol{f}(t_{k+1})\,)\,, \qquad k \geq 0\,.$$

The scheme is again unconditionally stable, but now second-order accurate in time.

Hence, with a moderate increment in the cost of computing the right-hand side as compared to the Implicit Euler scheme, one significantly gains in precision, as the order of the scheme is increased by one.

If we keep the time-step Δt constant and we use a direct method to solve the linear system, then it is convenient to factorize the matrix

$$\boldsymbol{B} + \tfrac{\Delta t}{2}\boldsymbol{A}$$

(for instance by computing its Choleski factorization) once and for all at the beginning of the time loop: at each time instant, we will only have to perform a forward-backward substitution.

On the contrary, if one uses an iterative method, such as Conjugate Gradient, then the computed solution $\boldsymbol{u}^k$ at the previous time instant will provide the initial guess for the new iteration.

The asymptotic stability condition for the Explicit Euler scheme

Recall that the asymptotic stability condition for the Explicit Euler scheme

$$\Delta t\, \theta_h \in \text{int}\,(\mathcal{R}_{EE}) \quad \text{for any eigenvalue } \theta_h \text{ of the matrix } \mathcal{A} = -\boldsymbol{B}^{-1}\boldsymbol{A}.$$

Let us observe that θ_h is an eigenvalue of $\mathcal{A} = -\boldsymbol{B}^{-1}\boldsymbol{A}$ if and only if there exists a non-zero vector $\boldsymbol{w}$ such that

$$\mathcal{A}\boldsymbol{w} = \theta_h \boldsymbol{w} \ ,$$

i.e.,

$$\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{w} = -\theta_h \boldsymbol{w} \ ,$$

namely, if and only if $\theta_h = -\lambda_h$, with λ_h eigenvalue of the matrix $\boldsymbol{B}^{-1}\boldsymbol{A}$.

In turns, this is true if and only if λ_h is a solution of the *generalized eigenvalue problem*

$$\boldsymbol{A}\boldsymbol{w} = \lambda_h \boldsymbol{B}\boldsymbol{w} \ .$$

We have seen that such a problem arises in the discretization of the *modal analysis problem* for an elastic membrane.

From this relation we derive

$$\boldsymbol{w}^T \boldsymbol{A}\boldsymbol{w} = \lambda_h \boldsymbol{w}^T \boldsymbol{B}\boldsymbol{w} \ ,$$

which provides th expression of λ_h as a **Rayleigh quotient**

$$\lambda_h = \frac{\boldsymbol{w}^T \boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{B}\boldsymbol{w}}.$$

This shows that all the eigenvalues λ_h are *real and strictly positive* (since the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are both symmetric and positive definite), hence, all the eigenvalues θ_h of the matrix $\mathcal{A}$ are real and negative.

Consequently, the asymptotic stability condition of the Explicit Euler scheme

$$\Delta t \leq \frac{2}{\max |\theta_h|}$$

is equivalent to

$$\Delta t \leq \frac{2}{\max \lambda_h} \ .$$

Since we have stated above that $\max \lambda_h \sim c h^{-2}$, the asymptotic stability condition becomes

$$\Delta t \leq C\, h^2,$$

as anticipated.

Time discretization of the elastic model: the Newmark scheme

In order to discretize in time the second-order system

$$\boldsymbol{u}'' = \boldsymbol{F}(\boldsymbol{u}, t) = -\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{u} + \boldsymbol{B}^{-1}\boldsymbol{f}(t) \ ,$$

it is customary to resort to the *Newmark scheme*.
It is written as

$$\boldsymbol{u}^{k+1} - 2\boldsymbol{u}^k + \boldsymbol{u}^{k-1} = \Delta t^2(\beta \boldsymbol{F}(\boldsymbol{u}^{k+1}, t_{k+1}) + (1-2\beta)\boldsymbol{F}(\boldsymbol{u}^k, t_k) + \beta \boldsymbol{F}(\boldsymbol{u}^{k-1}, t_{k-1})) \ ,$$

where $\beta \geq 0$ is a parameter to be chosen. Equivalently, this can be written as

$$\boldsymbol{B}(\boldsymbol{u}^{k+1} - 2\boldsymbol{u}^k + \boldsymbol{u}^{k-1}) + \Delta t^2 \boldsymbol{A}(\beta \boldsymbol{u}^{k+1} + (1-2\beta)\boldsymbol{u}^k + \beta \boldsymbol{u}^{k-1}) =$$
$$= \Delta t^2(\beta \boldsymbol{f}(t_{k+1}) + (1-2\beta)\boldsymbol{f}(t_k) + \beta \boldsymbol{f}(t_{k-1})) \ .$$

The scheme defines the time approximations starting from $k = 1$. Hece, it requires two initial values, $\boldsymbol{u}^0$ and $\boldsymbol{u}^1$. The first value is obviously chosen equal to $\boldsymbol{u}_0$, whereas the Taylor expansion of $\boldsymbol{u}(t)$ at the origin, $\boldsymbol{u}(t_1) = \boldsymbol{u}(0) + \Delta t \boldsymbol{u}'(0) + O(\Delta t^2)$, suggests the choice $\boldsymbol{u}^1 = \boldsymbol{u}_0 + \Delta t \boldsymbol{v}_0$.

Let us remark that if $\beta = 0$ and if the mass matrix $\boldsymbol{B}$ is replaced by the lumped mass matrix $\tilde{\boldsymbol{B}}$, then the previous scheme is explicit. In any other case, the scheme is implicit, and at each time instant $\boldsymbol{u}^{k+1}$ is determined by solving a linear system with matrix $\boldsymbol{B} + \Delta t^2 \beta \boldsymbol{A}$.

As far as accuracy is concerned, the scheme turns out to be *second-order accurate* for each value of β.

On the other hand, as far as asymptotic stability is concerned, the scheme turns out to be

- *unconditionally stable* if $\beta \geq \frac{1}{4}$.
- only *conditionally stable* if $\beta < \frac{1}{4}$.

In the latter case, the asymptotic stability condition is

$$\Delta t^2 \lambda_{h,max} < \frac{4}{1 - 4\beta} \ ,$$

where $\lambda_{h,max}$ denotes the maximum eigenvalue of the matrix $\boldsymbol{B}^{-1}\boldsymbol{A}$. We have already noticed that the order of magnitude of this eigenvalue is $O(h^{-2})$.

As a consequence, the previous conditions enforces a restriction of the type

$$\Delta t \leq C\, h$$

on the choice of the time step. Such a condition is by far less stringent that the stability condition for a conditionally stable scheme applied to the heat equation, which is of the type $\Delta t \leq C\, h^2$.

In most cases, such a condition turns out to be fully acceptable, since it is anyway required by the need of guaranteeing enough time accuracy on the discretization.