

So we have two classes: **explicit methods**, and **implicit methods**.

In all cases we want the sequence $\{y_0, y_1, \dots, y_N\}$ to converge to the sequence $\{y_0, y(t_1), \dots, y(T)\}$.

If, given a method, we can prove that

$$\exists C > 0 \text{ such that } \max_n |y_n - y(t_n)| \leq C \Delta t^p$$

with C independent of Δt and $p > 0$, then we say that *the method is convergent*, and *the order of convergence is p* (the bigger p , the faster the convergence). Consistency and stability will be defined separately. For one-step methods, a Lipschitz estimate for the step map gives the following convergence result.

One-step convergence on a fixed time interval

Write $y_{n+1} = \Phi_h(t_n, y_n)$, $h = (T - t_0)/N$, and assume

$$\|\Phi_h(t, u) - \Phi_h(t, v)\| \leq (1 + Lh)\|u - v\|,$$

with $L \geq 0$ independent of h , and a local defect

$$\|y(t_{n+1}) - \Phi_h(t_n, y(t_n))\| \leq Ch^{p+1}.$$

For $e_n = y(t_n) - y_n$ this implies

$$\|e_{n+1}\| \leq (1 + Lh)\|e_n\| + Ch^{p+1}.$$

Iteration gives, for $t_n \leq T$,

$$\|e_n\| \leq e^{L(T-t_0)}(\|e_0\| + C(T - t_0)h^p).$$

With exact (or order- p) initial data the method converges with order p . This finite-time stability estimate differs from absolute stability on the test equation.

Intuitive explanation of *consistency* and *stability*

- the *consistency error* at a given time step measures the error which is created at that step; it is defined as *the error made when the scheme is applied to the exact solution of the problem*. It can be seen that the consistency error measures how much the discrete scheme resembles the differential problem.
- *stability* measures how the error, created and accumulated during the previous steps, goes to the next step (is it amplified? does it decay?...)

Consistency of a scheme

Consistency is a measure of how much the discrete scheme resembles the differential problem: the *consistency error* is *the error made when the scheme is applied to the exact solution of the problem*.

The *consistency error* of a given scheme is defined as

$$\tau = \max_{n=0,\dots,N-1} |\tau_n|$$

where τ_n is the *local truncation error* at the step n , that is defined as:

$$\tau_n = \frac{y(t_{n+1}) - \tilde{y}_{n+1}}{\Delta t}$$

where $\tilde{y}_{n+1}$ is obtained applying one step of the method to the exact solution at the previous time instant $y(t_n)$. We say that the scheme is consistent of order p if:

$$\exists C > 0, \text{ independent of } \Delta t, \text{ s. t. } \tau \leq C \Delta t^p$$

We observe that if the numerical scheme is consistent, then $\tau \rightarrow 0$ for $\Delta t \rightarrow 0$.

Consistency of Explicit Euler

The exact solution fulfils: $y'(t_n) - f(t_n, y(t_n)) = 0, \quad n = 0, 1, 2, \dots$

If we apply the Explicit Euler scheme to the exact solution we will have:

$$\tilde{y}_{n+1} = y(t_n) + \Delta t f(t_n, y(t_n))$$

Since $y'(t_n) = f(t_n, y(t_n))$, by Taylor expansion centered in t_n , there exists $z \in [t_n, t_{n+1}]$ such that

$$y(t_{n+1}) = y(t_n) + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(z) = y(t_n) + \Delta t f(t_n, y(t_n)) + \frac{\Delta t^2}{2} y''(z)$$

Thus, the local truncation error becomes

$$\tau_n = \frac{y(t_{n+1}) - \tilde{y}_{n+1}}{\Delta t} = \frac{y(t_{n+1}) - y(t_n)}{\Delta t} - f(t_n, y(t_n)) = \frac{\Delta t}{2} y''(z),$$

and the consistency error is then

$$\tau = \max |\tau_n| \leq \frac{\Delta t}{2} \max_{t \in [t_0, T]} |y''(t)| = C \Delta t.$$

Thus, Explicit Euler scheme is consistent with **order of consistency 1**.

Consistency of the Heun method ****NOT FOR THE EXAM****

We apply the Heun scheme again to the exact solution $\tilde{y}_n = y(t_n)$:

$$\tilde{y}_{n+1} = y(t_n) + \frac{\Delta t}{2} (f(t_n, y(t_n)) + f(t_{n+1}, y(t_n) + \Delta t f(t_n, y(t_n))))$$

Since $f(t_n, y(t_n)) = y'(t_n)$, by Taylor expansion centered in t_n , there exists $z_1 \in [t_n, t_{n+1}]$ such that

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + (t_{n+1} - t_n) y'(t_n) + \frac{\Delta t^2}{2} y''(z_1) \\ &= y(t_n) + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(z_1) \end{aligned} \tag{1}$$

and, by Taylor expansion centered in t_{n+1} , there exists $z_2 \in [t_n, t_{n+1}]$ such that

$$\begin{aligned} y(t_n) &= y(t_{n+1}) + (t_n - t_{n+1}) y'(t_{n+1}) + \frac{\Delta t^2}{2} y''(z_2) \\ &= y(t_{n+1}) - \Delta t y'(t_{n+1}) + \frac{\Delta t^2}{2} y''(z_2) \end{aligned} \tag{2}$$

The local truncation error, setting $\hat{y}_{n+1} = y(t_n) + \Delta t f(t_n, y(t_n))$ and using again that $y'(t_{n+1}) = f(t_{n+1}, y(t_{n+1}))$ is:

$$\begin{aligned}
 \tau_n &= \frac{y(t_{n+1}) - \tilde{y}_{n+1}}{\Delta t} \\
 &= \frac{(y(t_{n+1}) - y(t_n) - \frac{\Delta t}{2} y'(t_n) - \frac{\Delta t}{2} f(t_{n+1}, \hat{y}_{n+1})))}{\Delta t} \\
 &= \frac{(y(t_{n+1}) - y(t_n) - \frac{\Delta t}{2} y'(t_n) - \frac{\Delta t}{2} y'(t_{n+1}) + \frac{\Delta t}{2} f(t_{n+1}, y(t_{n+1})) - \frac{\Delta t}{2} f(t_{n+1}, \hat{y}_{n+1})))}{\Delta t} \\
 &= \frac{(y(t_{n+1}) - y(t_n) - \frac{\Delta t}{2} y'(t_n) - \frac{\Delta t}{2} y'(t_{n+1}))}{\Delta t} + \frac{1}{2} (f(t_{n+1}, y(t_{n+1})) - f(t_{n+1}, \hat{y}_{n+1})) \\
 &= A + B,
 \end{aligned}$$

where

$$A = \frac{(y(t_{n+1}) - y(t_n) - \frac{\Delta t}{2} y'(t_n) - \frac{\Delta t}{2} y'(t_{n+1}))}{\Delta t}$$

and

$$B = \frac{1}{2} (f(t_{n+1}, y(t_{n+1})) - f(t_{n+1}, \hat{y}_{n+1}))$$

For the first term, Taylor's formula at the two endpoints gives

$$A = \frac{h}{4}(y''(z_1) - y''(z_2)), \quad h = \Delta t, \quad z_1, z_2 \in [t_n, t_{n+1}].$$

Assume $y \in C^3([t_0, T])$. Since

$$|y''(z_1) - y''(z_2)| \leq |z_1 - z_2| \|y^{(3)}\|_\infty \leq h \|y^{(3)}\|_\infty,$$

we obtain

$$|A| \leq \frac{h^2}{4} \|y^{(3)}\|_\infty.$$

This estimate also holds when $z_1 = z_2$; no division by their difference is needed.

For the second term we have

$$\begin{aligned}|B| &= \frac{1}{2} |f(t_{n+1}, y(t_{n+1})) - f(t_{n+1}, \hat{y}_{n+1})| \\ &\leq \frac{L}{2} |y(t_{n+1}) - \hat{y}_{n+1}| \\ &\leq \frac{L}{2} \frac{\Delta t^2}{2} |y''(z_1)|\end{aligned}$$

where we have assumed and used the Lipschitzianity of f with respect to its second argument, that is, $|f(t, \eta_1) - f(t, \eta_2)| \leq L|\eta_1 - \eta_2|$, and then (1) again. In conclusion

$$\tau = \max |\tau_n| \leq \frac{\Delta t^2}{4} (\|y'''\|_\infty + L\|y''\|_\infty) = C\Delta t^2$$

Thus, the Heun scheme is consistent with [order of consistency 2](#).

Polynomial solutions and consistency

Explicit Euler is exact for affine solutions. However, an order-2 Runge–Kutta method need not reproduce every quadratic solution exactly. For example,

$$y' = 2t + y - t^2, \quad y(0) = 0, \quad y(t) = t^2.$$

One Heun step gives $K_1 = 0$, $K_2 = 2h - h^2$, hence

$$y_1 = \frac{h}{2}(K_1 + K_2) = h^2 - \frac{h^3}{2} \neq y(h).$$

Polynomial exactness characterizes the residual consistency conditions of linear multistep formulas. General Runge–Kutta methods require their own order conditions.

Consistency of implicit methods: review of explicit Euler

$$\frac{y_{n+1} - y_n}{\Delta t} - f(t_n, y_n) = 0 \quad \forall n$$

Applying this scheme to the exact solution of the Cauchy Problem means considering

$$\frac{y(t_{n+1}) - y(t_n)}{\Delta t} - y'(t_n)$$

which is in general $\neq 0$. Is the order of consistency 1? For this we should have the above $= 0$ when the solution of the Cauchy Problem is 1, t .

$$\text{When } y(t) = 1, y' = 0 = f, \implies \frac{1 - 1}{\Delta t} - 0 = 0;$$

$$\text{When } y(t) = t, y' = 1 = f, \implies \frac{t_{n+1} - t_n}{\Delta t} - 1 = 1 - 1 = 0;$$

Hence, the order of consistency of EE method is 1. Note that

$$\text{When } y(t) = t^2, y' = 2t = f, \implies \frac{t_{n+1}^2 - t_n^2}{\Delta t} - 2t_n = t_{n+1} - t_n \neq 0$$

Consistency of implicit methods: implicit Euler

$$\frac{y_{n+1} - y_n}{\Delta t} - f(t_{n+1}, y_{n+1}) = 0 \quad \forall n$$

Applying this scheme to the exact solution of the Cauchy Problem means considering

$$\frac{y(t_{n+1}) - y(t_n)}{\Delta t} - y'(t_{n+1})$$

which is in general $\neq 0$. Is the order of consistency 1? For this we should have the above $= 0$ when the solution of the Cauchy Problem is 1, t .

$$\text{When } y(t) = 1, y' = 0 = f, \implies \frac{1 - 1}{\Delta t} - 0 = 0;$$

$$\text{When } y(t) = t, y' = 1 = f, \implies \frac{t_{n+1} - t_n}{\Delta t} - 1 = 1 - 1 = 0;$$

Hence, the order of consistency of IE method is 1. Note that

$$\text{When } y(t) = t^2, y' = 2t = f, \implies \frac{t_{n+1}^2 - t_n^2}{\Delta t} - 2t_{n+1} = t_n - t_{n+1} \neq 0$$

Consistency of implicit methods: Crank-Nicolson

$$\frac{y_{n+1} - y_n}{\Delta t} - \frac{1}{2} \left(f(t_n, y_n) + f(t_{n+1}, y_{n+1}) \right) = 0 \quad \forall n$$

Applying this scheme to the exact solution means considering

$$\frac{y(t_{n+1}) - y(t_n)}{\Delta t} - \frac{1}{2} \left(y'(t_n) + y'(t_{n+1}) \right).$$

Let's check the above expression when the solution of the Cauchy Problem is $1, t, t^2$:

$$\text{When } y(t) = 1, y' = 0 = f, \implies \frac{1 - 1}{\Delta t} - \frac{1}{2}(0 + 0) = 0;$$

$$\text{When } y(t) = t, y' = 1 = f, \implies \frac{t_{n+1} - t_n}{\Delta t} - \frac{1}{2}(1 + 1) = 1 - 1 = 0;$$

$$\text{When } y(t) = t^2, y' = 2t = f, \implies \frac{t_{n+1}^2 - t_n^2}{\Delta t} - \frac{1}{2}(2t_n + 2t_{n+1}) = 0$$

Hence, the order of consistency of Crank-Nicolson is 2.

Absolute Stability

The concept of *stability* is another very important and useful concept whose precise definition has to be made precise at various occurrences. Roughly speaking, stability is what guarantees that the errors generated during a numerical procedure do not grow too much.

With Ode stability is a delicate issue, especially when the phenomenon under study has to be followed for a long time. To better see what happens, let us consider a simple model problem, for which we know the exact solution:

$$\begin{cases} y'(t) = ay(t) & t > 0 \\ y(0) = y_0 \end{cases} \quad a \in \mathbb{C}$$

which exact solution is

$$y(t) = y_0 e^{(\operatorname{Re} a)t} (\cos((\operatorname{Im} a)t) + i \sin((\operatorname{Im} a)t))$$

If $\operatorname{Re}(a) > 0$ the exact solution *grows* exponentially. We cannot expect (and we do not want!) that the discrete scheme remains bounded, and it is not even the case to discuss “stability”.

Instead, if $\operatorname{Re}(a) < 0$ the exact solution not only is bounded, but *decays* exponentially:

$$\operatorname{Re}(a) < 0 \longrightarrow |y(t)| \leq |y_0| \quad \text{and} \quad \lim_{t \rightarrow \infty} |y(t)| = 0.$$

In this case we need to analyse the discrete schemes, and see whether the discrete solution decays too, and behaves like the exact solution. Hence, let $\operatorname{Re}(a) < 0$, and let $\{y_n\}$ be the sequence generated by a numerical scheme. Does $\{y_n\}$ satisfy the following relation?

$$a \in \mathbb{C} \text{ with } \operatorname{Re} a < 0 \longrightarrow |y_n| \leq |y_0| \quad \text{and} \quad \lim_{n \rightarrow \infty} |y_n| = 0?$$

The scheme is absolutely stable at $z = a\Delta t$ if its amplification factor satisfies $|R(z)| < 1$. It is A-stable if this holds for every $\operatorname{Re}(z) < 0$.

Checking absolute stability for Explicit Euler

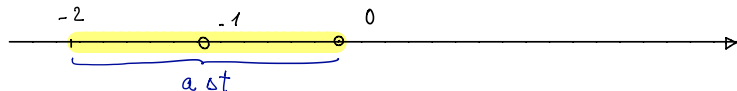
By applying **Explicit Euler** method to the model problem, we get
($y_{n+1} = y_n + \Delta t f(t_n, y_n)$ with $f(t_n, y_n) = ay_n$)

$$y_{n+1} = (1 + a\Delta t)y_n \quad n = 0, 1, \dots \implies y_n = y_0(1 + a\Delta t)^n.$$

The exact solution decays exponentially from the initial value y_0 , while the growth-decay factor for the discrete scheme is $G = 1 + a\Delta t$.

For having $\lim_{n \rightarrow \infty} |y_n| = 0$ we need $|G| < 1$. We observe that $|1 + a\Delta t|$ is the distance between -1 and $a\Delta t$, so $|G| < 1$ if and only if $a\Delta t$ is inside the unit disk centered at -1 .

Stability for Explicit Euler: the real setting

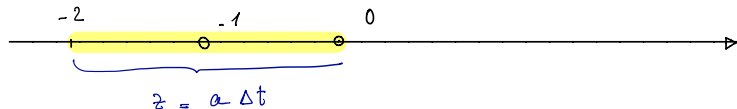


$$EE \text{ is stable} \iff |1 + a \Delta t| < 1$$

$$\iff -2 < a \Delta t < 0$$

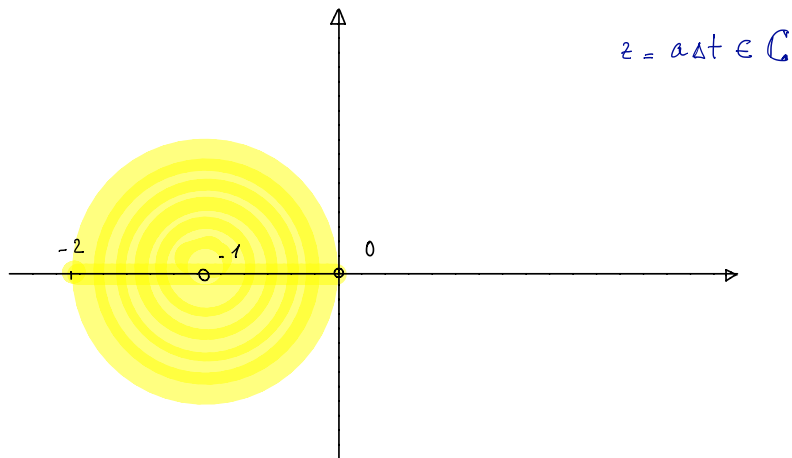
$$\iff -\frac{2}{a} > \Delta t > 0 \quad (\text{since } a < 0)$$

Stability for Explicit Euler: the real setting



$$\begin{aligned} \text{EE is stable} &\iff |1 + a \Delta t| < 1 \\ &\iff -2 < \underbrace{a \Delta t}_z < 0 \end{aligned}$$

Stability for Explicit Euler: the complex setting



$$\begin{aligned} \text{EE is stable} &\iff |1 + \overbrace{a\Delta t}^z| < 1 \\ &\iff z \in \text{circle with radius } 1 \text{ and center } -1 \end{aligned}$$

It holds:

$$|1 + a\Delta t| < 1 \iff |1 + a\Delta t|^2 < 1$$

but we have

$$|1 + a\Delta t|^2 = (1 + a\Delta t)(1 + \bar{a}\Delta t) = 1 + 2\operatorname{Re}(a)\Delta t + |a|^2\Delta t^2$$

Thus

$$|1 + a\Delta t| < 1 \iff 0 < \Delta t < -\frac{2\operatorname{Re}(a)}{|a|^2} \quad (\operatorname{Re}(a) < 0)$$

This is the drawback of Explicit Euler scheme, and of explicit Runge–Kutta methods: for small enough time steps the stability condition is satisfied, but when $\operatorname{Re}(a)$ is strongly negative (exactly the case of rapid decay in the true solution) we are compelled to keep Δt small.

Two local error conventions

For a one-step map, the local defect is

$$d_{n+1} = y(t_{n+1}) - \Phi_h(t_n, y(t_n)), \quad \tau_n = d_{n+1}/h.$$

For an implicit formula, it is often simpler to insert the exact solution into its equation. For implicit Euler this residual is

$$r_n = \frac{y(t_{n+1}) - y(t_n)}{h} - f(t_{n+1}, y(t_{n+1})).$$

It is not exactly the same quantity as τ_n . If f is Lipschitz in y with constant L and $hL < 1$, then

$$\|d_{n+1}\| \leq \frac{h\|r_n\|}{1 - hL}.$$

Thus a residual of order h^p yields a defect of order h^{p+1} for sufficiently small h .