I Introduction to Runge-Kutta methods

1. Examples and first definitions

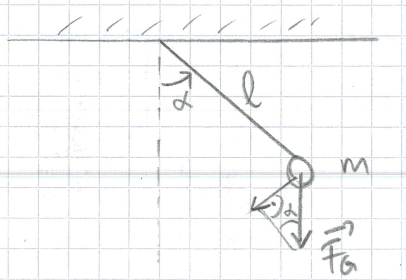
General goal of NIDS:

Given $f: \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$

$$\begin{cases} y' = f(t, y(t)) \\ y(0) = y_0 \end{cases} \quad \text{ODE}$$

Find numerically $y(t)$, $0 \leq t \leq T$ $\to$ in the most efficient way - "f dependent"

Example 1: Pendulum



(mathematical pendulum)

(α: angle in radiant)

Newton's law 2: $F = m \cdot a$

$\Rightarrow -\cancel{m} g \sin \alpha = \cancel{m} \cdot l \cdot \ddot{\alpha}$

$\Rightarrow \boxed{\ddot{\alpha} = -\frac{g}{l} \cdot \sin \alpha}$

[redacted]

WLOG, take $g = l = 1$

$\Rightarrow \boxed{\ddot{\alpha}(t) = -\sin \alpha(t)}$ 2nd order ODE

Coordinate transform: $q(t) = \alpha(t)$ (position)
$p(t) = \dot{\alpha}(t)$ (angular velocity)

$$\Rightarrow \begin{cases} \dot{q}(t) = p(t) \\ \dot{p}(t) = -\sin(q(t)) \end{cases} \quad (*)$$

System of two 1st order ODE

Consider $H(p,q) = \underbrace{\frac{1}{2}p^2}_{\text{kinetic energy}} - \underbrace{\cos q}_{\text{potential energy}}$ (Hamiltonian fct)

This total energy $H(p,q)$ is conserved by the system.

Observe that $\begin{cases} \dot{q}(t) = \frac{\partial H}{\partial p}(p,q) \\ \dot{p}(t) = -\frac{\partial H}{\partial q}(p,q) \end{cases}$, i.e. we have rewritten $(*)$.

Defn: (Hamiltonian system)

Suppose that $H: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a smooth function. Then the dynamical system $t \to (p(t), q(t))$ given by

$$\begin{cases} \dot{p}_k(t) = -\frac{\partial H}{\partial q_k}(p,q) \\ \dot{q}_k(t) = \frac{\partial H}{\partial p_k}(p,q) \end{cases} \quad k = 1,2,\dots,d \qquad (1)$$

is a Hamiltonian system, and H is the Hamiltonian energy.

Rem.: The value of d denotes the degree of freedom of the system.

Back to the example: let us compute

$$\frac{d}{dt} H(p(t),q(t)) = \frac{\partial H}{\partial p}(p,q) \cdot \dot{p}(t) + \frac{\partial H}{\partial q}(p,q) \cdot \dot{q}(t)$$
$$\overset{(*)}{=} p(t) \cdot (-\sin(q(t))) + \sin(q(t)) \cdot p(t)$$
$$= \underline{\underline{0}}$$

$\Rightarrow H(p(t), q(t)) = \text{constant}$

$\Rightarrow$ This quantity is conserved along the trajectories.

In general:

<u>Lemma:</u> H is conserved along the solution of (1).

<u>Proof:</u>
$$\frac{d H}{dt}(p(t),q(t)) \overset{\text{chain rule}}{=} \sum_{k=1}^{d} \frac{\partial H}{\partial p_k}(p(t),q(t))\,\dot p_k(t) + \sum_{k=1}^{d} \frac{\partial H}{\partial q_k}(p(t),q(t))\,\dot q_k(t)$$

If $(p(t),q(t))$ solves (1)

$$\Rightarrow \quad = \sum_{k=1}^{d} \frac{\partial H}{\partial p_k}(p(t),q(t))\left(-\frac{\partial H}{\partial q_k}(p(t),q(t))\right) + \sum_{k=1}^{d} \frac{\partial H}{\partial q_k}(p(t),q(t))\cdot\left(\frac{\partial H}{\partial p_k}(p(t),q(t))\right)$$

$$= \underline{\underline{0}}$$

$$\Rightarrow H(p(t),q(t)) = H(p(0),q(0)) = \text{constant}.$$

#

More in general, consider $f: \mathbb{R}^d \to \mathbb{R}^d$ and the ODE

$$\begin{cases} y' = f(y) \\ y(t_0) = y_0 \end{cases} \qquad (2)$$

<u>Defn.:</u> (First integral)
A non-constant differentiable function $I: \mathbb{R}^d \to \mathbb{R}$ is a <u>first integral</u> of (2) if and only if

$$\underbrace{I'(y)}_{\substack{\text{Jacobian}\\ \text{matrix } d\times d}}\ \underbrace{f(y)}_{\substack{\text{vector}\\ d\times 1}} = \underbrace{0}_{\substack{\text{vector}\\ d\times 1}} \quad \forall\, y \in \mathbb{R}^d.$$

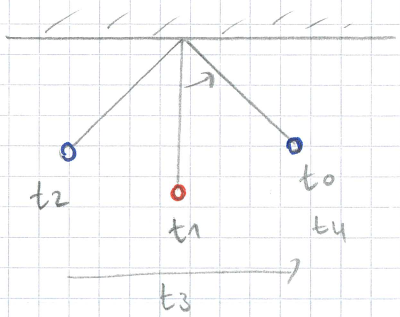
<u>Rem.:</u>
$$\frac{d}{dt} I(y(t)) = I'(y(t))\cdot y'(t) = I'(y(t))\,f(y(t)) = 0$$

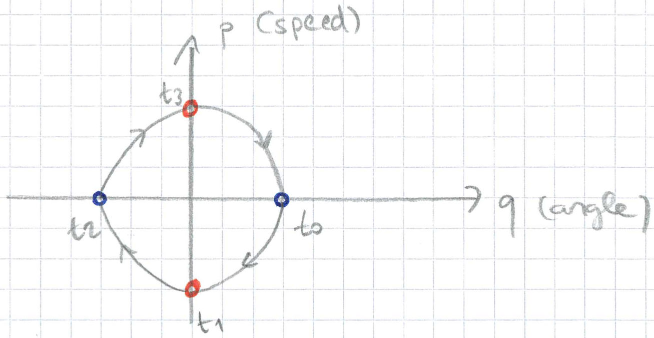
$$\Rightarrow I(y(t)) = \underline{\underline{\text{constant}}}.$$ ($I(y(t))$ is constant along the trajectories)

Rem: H is a first integral of (1).

Rem: Unless stated otherwise, f is assumed to be smooth enough.
The fct I needs not to be defined on all $\mathbb{R}^d$, but just on a local region around $f(y)$.

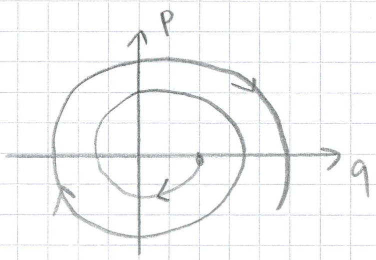
Numerical integration of the pendulum (example 1)





ODE: $\begin{cases} \dot{p} = -\sin q = f(p,q) \\ \dot{q} = p = g(p,q) \end{cases}$, time step $h > 0$.

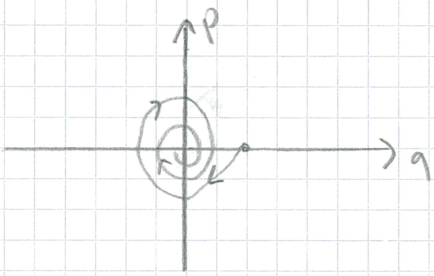
Explicit Euler method: $\begin{cases} p_{n+1} = p_n \underbrace{- h \sin q_n}_{= h f(p_n,q_n)} \\ q_{n+1} = q_n + \underbrace{h p_n}_{= h g(p_n,q_n)} \end{cases}$ $\left(\begin{matrix} p_0 = 0 \\ q_0 = 0{,}5 \end{matrix} \right)$



X

(creates energy that comes from nowhere)

Implicit Euler method: $\begin{cases} p_{n+1} = p_n - h \sin q_{n+1} \\ q_{n+1} = q_n + h p_{n+1} \end{cases}$

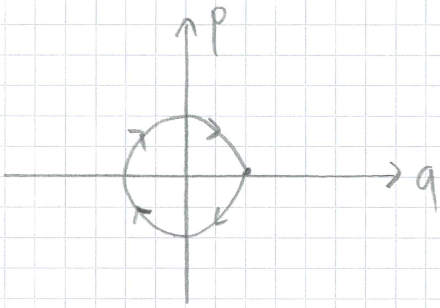


X

(system shrinks the orbit)

Symplectic Euler method:

$$\begin{cases} p_{n+1} = p_n \overbrace{- h \sin q_n}^{= f(p_{n+1}, q_n)} \\ q_{n+1} = q_n + \underbrace{h\, p_{n+1}}_{= g(p_{n+1}, q_n)} \end{cases}$$

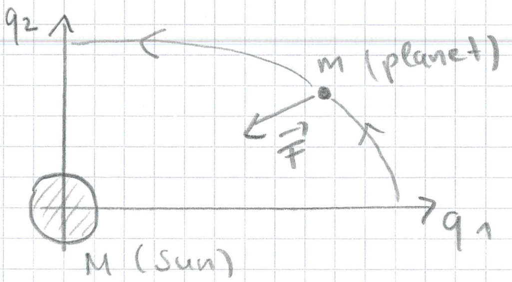


(qualitatively correct, same cost as explicit Euler but better quality)

Rem: For the pendulum this method is in fact explicit (because $H(p,q) = T(p) + U(q)$).
In general, the symplectic Euler method is implicit, but becomes explicit for separable systems, i.e.

$$H(p,q) = T(p) + U(q).$$

Example 2: The Kepler problem (two-body-problem)



Newton 2: $F = m a$

Gravitational force: $F = -G \dfrac{M m\, q}{\|q\|^3}$

$$\Rightarrow \cancel{m} \cdot \ddot{q} = -G \frac{M \cancel{m}\, q}{\|q\|^3} \quad \text{with } q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}, \ \|q\| = \sqrt{q_1^2 + q_2^2}$$

Introduce $\dot{q} = p$ (velocity) and WLOG $G = M = 1$

$$\Rightarrow \begin{cases} \dot{p}_i(t) = - \dfrac{q_i(t)}{\|q(t)\|^3} \\ \dot{q}_i(t) = p_i(t) \end{cases} \quad i = 1,2 \qquad (3)$$

[redacted]

Hamiltonian system for the function } to verify (ex.)

$$H(p,q) = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}}.$$

conserved quantity of the system

Another conserved quantity (first integral) for (3) is the angular momentum:

$$L(p,q) = q_1 p_2 - q_2 p_1$$

<u>Proof:</u> (ex.)

[redacted]

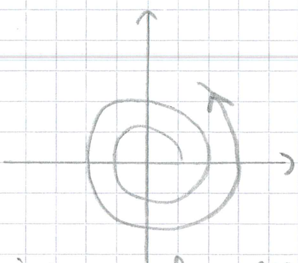
(some numerical methods preserve this exactly)

<u>Numerical experiments:</u>

$$\dot{p} = f(p,q) = \begin{pmatrix} -\frac{q_1}{\|q\|^3} \\ -\frac{q_2}{\|q\|^3} \end{pmatrix}$$

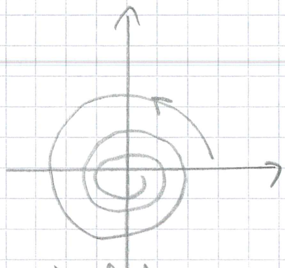
$$\dot{q} = g(p,q) = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$$

<u>Explicit Euler</u>



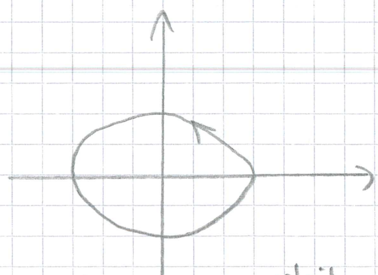
increase of energy

<u>Implicit Euler</u>



implodes

(or implicit midpoint)
<u>Symplectic Euler</u>



remains on orbit

<u>Implicit midpoint rule:</u>

$$\begin{pmatrix} p_{i+1} \\ q_{i+1} \end{pmatrix} = \begin{pmatrix} p_i \\ q_i \end{pmatrix} + h \begin{pmatrix} f\left(\frac{p_i + p_{i+1}}{2}, \frac{q_i + q_{i+1}}{2}\right) \\ g\left(\frac{p_i + p_{i+1}}{2}, \frac{q_i + q_{i+1}}{2}\right) \end{pmatrix}$$

Qualitative behaviour:

Method	Error in $H(p,q)$ (total energy)	Error in $L(p,q)$	Total error $\|q_n - q(t_n)\|$
Explicit Euler	$O(th)$	$O(th)$	$O(t^2h)$
Implicit Euler	$O(th)$	$O(th)$	$O(t^2h)$
Symplectic Euler	$O(h)$	0	$O(th)$
Implicit Midpoint rule	$O(h^2)$	0	$O(th^2)$

linear drift

nothing to do with accuracy, due to bad method for this problem

Symp. E. 1st order method
IMR 2nd order method

exactly conserved (no error at all)

(more accurate results for higher order methods can be expected)

2. Runge-Kutta methods

Consider the ODE

$$(*)\quad \begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = y_0 \end{cases}, \quad f: \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$$

Fundamental theorem of calculus: for $h > 0$

$$y(t_0+h) = y_0 + \int_{t_0}^{t_0+h} y'(t)\,dt \qquad \text{(integral form of the ODE)}$$

$$= y_0 + \underbrace{\int_{t_0}^{t_0+h} f(t, y(t))\,dt}_{=: \, J(t_0)} \quad \text{to be approximated}$$

1) Explicit Euler: (Euler 1768)

$$J(t_0) \approx h f(t_0, y(t_0)) \implies y_1 = y_0 + h f(t_0, y_0)$$

2) Implicit Euler: (Cauchy 1824)

$$J(t_0) \approx h f(t_0+h, y(t_0+h)) \implies y_1 = y_0 + h f(t_0+h, y_1)$$

Accuracy of RK methods

Defn: (order p)

A one-step method has order p for sufficiently smooth problems $(*)$ if the local error after one step satisfies:

$$\| y_1 - y(t_0+h) \| = \mathcal{O}(h^{p+1}) \quad \text{as } h \to 0.$$
$$\leq C h^{p+1}$$

(the one-step method starts at the exact value y_0)

Other approximations of $J(t_0)$:

3) i) $\begin{cases} y_{1/2} = y_0 + \frac{1}{2} h f(t_0, y_0) \\ y_1 = y_0 + h f(t_0 + \frac{1}{2}h, y_{1/2}) \end{cases}$ (explicit Euler step for intermediate value)

written more elegantly:

$$\begin{cases} K_1 = f(t_0, y_0) \\ K_2 = f(t_0 + \frac{1}{2}h, \overbrace{y_0 + \frac{1}{2} h K_1}^{= y_{1/2}}) \\ y_1 = y_0 + h K_2 \end{cases}$$

explicit midpoint rule (Runge 1895)

ii) $\begin{cases} y_{1/2} = y_0 + \frac{1}{2} h f(t_0 + \frac{1}{2}h, y_{1/2}) \\ y_1 = y_0 + h f(t_0 + \frac{1}{2}h, y_{1/2}) \end{cases}$ (implicit Euler step)

idem:

$$\begin{cases} K_1 = f(t_0 + \frac{1}{2}h, y_0 + \frac{1}{2} h K_1) \\ y_1 = y_0 + h K_1 \end{cases}$$

implicit midpoint rule

Rem: Both methods 3) have second order (see exercises).
(Runge - Kutta is a algebraic study of numerical methods).
$\Rightarrow$ Generalisation: Runge-Kutta methods.

Defn: (Runge-Kutta method, Runge 1895, Kutta 1901, Heun 1901)
Let $s \in \mathbb{N}^*$ (stage number) and $b_i, a_{ij}, c_i \in \mathbb{R}$, $1 \leq i, j \leq s$
A s-stage Runge-Kutta method is given by

$$\begin{cases} k_i = f(t_0 + c_i h, y_0 + h \sum_{j=1}^{s} a_{ij} k_j), \quad i = 1, 2, \dots, s \\ y_1 = y_0 + h \sum_{i=1}^{s} b_i k_i \end{cases}$$

Rem: i) Usually take $c_i = \sum_{j=1}^{s} a_{ij}$

this ensures that $f(t_0+c_ih, y(t_0+c_ih)) - k_i = O(h^2)$

[redacted]

ii) If $a_{ij}=0$ for $j \geq i \Rightarrow$ RK is explicit, otherwise implicit.

Notation: Butcher table

$$
\begin{array}{c|cccc}
c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\
c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\
\vdots & \vdots & & & \\
c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\
\hline
 & b_1 & b_2 & \cdots & b_s
\end{array}
$$

Examples: Explicit Euler: $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$

Implicit Euler: $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$

Explicit midpoint rule: $\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$

Implicit midpoint rule: $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$

[redacted]

Rem.: One can take variable time steps:

$$y_1 = y_0 + h_0 \sum b_i k_i(h_0)$$
$$y_2 = y_1 + h_1 \sum b_i k_i(h_1)$$
$$\vdots$$

How does the error propagate?

From local to global convergence.

Recall: $(*)\ \begin{cases} y' = f(t, y) \\ y(t_0) = y_0 \end{cases}$

Theorem 1: let $y(t)$ be the solution of $(*)$.

Assume i) for any $t_k \in [t_0, T]$, $h \leq h_{max}$

$$\| y_{k+1} - y(t_k + h) \| \leq C h^{p+1} \quad \text{(RK method with local order } p\text{)}$$

ii) f is Lipschitz continuous:

$$\| f(t, y) - f(t, z) \| \leq L \| y - z \| \quad \text{(in a neighborhood of the exact soln)}$$

Then

$$\max_{t_0 \leq t_n \leq T} \| y_n - y(t_n) \| \leq C h^p, \quad \text{where } h = \max_i h_i .$$

Proof: See Hairer-Wanner-Norsett Vol 1 II.3.

[redacted]