Numerical Methods
in Engineering and Applied Science

Lecture 13. Part 1. Geometric Integration of ODEs.

Generally, a numerical method is required to give a 'good' approximation to the exact solution of the problem. It is quite natural to fix the initial condition, fix the time t_N and ask that the numerical solution u_N be close to the exact solution $u(t_N)$. This idea gave us the notions of *global error* and *order of convergence*. We have also introduced *the absolute stability* to characterize the behavior of the solution in the long term. In this lecture, we are going to consider some other desirable properties in the case where the equation we are solving is a conservation law, i.e., certain quantities are conserved along the trajectory. Here we consider only autonomous systems,

$$\boldsymbol{u}'(t) = \boldsymbol{f}(\boldsymbol{u}). \tag{1}$$

with $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{f} \ : \ \mathbb{R}^m \to \mathbb{R}^m$.

Definition. A non-constant function $F : \mathbb{R}^m \to \mathbb{R}$ is called a *first integral* of (1) if

$$(\boldsymbol{\nabla} F) \cdot \boldsymbol{f} = 0 \quad \text{for all } \boldsymbol{v} \in \mathbb{R}^m. \tag{2}$$

It is also called an *invariant*.
This definition implies that, if $\boldsymbol{u}$ is the solution of (1) with $\boldsymbol{u}(0) = \boldsymbol{u}_0$, it satisfies $F(\boldsymbol{u}) = F(\boldsymbol{u}_0) = const.$

Example (total energy conservation). Hamilton's equations

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial H(p,q)}{\partial q}, \quad \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial H(p,q)}{\partial p}, \tag{3}$$

have the Hamiltonian H as their first integral:

$$\left(-\frac{\partial H(p,q)}{\partial q}\right)\frac{\partial H(p,q)}{\partial p} + \left(\frac{\partial H(p,q)}{\partial p}\right)\frac{\partial H(p,q)}{\partial q} = 0. \tag{4}$$

Linear invariants. We say that there is a linear invariant if a linear combination of the components of the solution is preserved,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(c_1 u_1(t) + c_2 u_2(t) + ... + c_m u_m(t)\right) = 0, \tag{5}$$

where c_1, c_2, ..., c_3 are constants. We can rewrite (5) in vector form,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{c}^T\boldsymbol{u}(t) = 0, \tag{6}$$

where $\boldsymbol{c} \in \mathbb{R}^m$. Hence, $F(\boldsymbol{u}) = \boldsymbol{c}^T\boldsymbol{u}$. Using the ODE (1), we obtain

$$\boldsymbol{c}^T\boldsymbol{f}(\boldsymbol{u}) = 0 \quad \text{for all} \quad \boldsymbol{u} \in \mathbb{R}^m, \tag{7}$$

i.e., $\boldsymbol{f}$ is orthogonal to c.
If we employ the Euler method, $\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + h\boldsymbol{f}(\boldsymbol{u}_n)$, we obtain

$$\boldsymbol{c}^T\boldsymbol{u}_{n+1} = \boldsymbol{c}^T\boldsymbol{u}_n + h\boldsymbol{c}^T\boldsymbol{f}(\boldsymbol{u}_n) = \boldsymbol{c}^T\boldsymbol{u}_n, \tag{8}$$

and we see that the linear invariant is conserved.

In the same way, we show that any multi-step or Runge–Kutta method, if it is consistent, preserves linear invariants.

One finds linear invariants in many applications, for example

- chemical kinetics, where molecules of X_1 and X_2 produce molecules of X_3, but mass is conserved;
- mechanics, which conserve mass, momentum, etc.
- of stochastic models, where the ODE describes the evolution of the probabilities and the sum of the probabilities must be equal to 1.
- models of social phenomena, where the population size can be constant.

Classical numerical methods preserve these properties.

Quadratic invariants . We say that the system $\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u})$ has a quadratic invariant if

$$F(\boldsymbol{u}) = \boldsymbol{u}^T C \boldsymbol{u}, \tag{9}$$

where $C \in \mathbb{R}^{m\times m}$ is a symmetric matrix.

Example. The components of the angular momentum of a free rigid body, with its center of mass at the origin, satisfy

$$\begin{cases} u_1'(t) = a_1 u_2(t) u_3(t) \\ u_2'(t) = a_2 u_3(t) u_1(t) \\ u_3'(t) = a_3 u_1(t) u_2(t) \end{cases} \tag{10}$$

The coefficients a_1, a_2 et a_3 are related to the principal moments of inertia,

$$a_1 = \frac{I_2 - I_3}{I_2 I_3}, \quad a_2 = \frac{I_3 - I_1}{I_3 I_1}, \quad a_3 = \frac{I_1 - I_2}{I_1 I_2} \tag{11}$$

Note that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(u_1^2+u_2^2+u_3^2) &= 2u_1u'_1+2u_2u'_2+2u_3u'_3\\ &= 2u_1a_1u_2u_3+2u_2a_2u_3u_1+2u_3a_3u_1u_2\\ &= 2u_1u_2u_3(a_1+a_2+a_3)=0\end{aligned} \tag{12}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\left(\frac{u_1^2}{I_1}+\frac{u_2^2}{I_2}+\frac{u_3^2}{I_3}\right)\right) &= \frac{u_1u'_1}{I_1}+\frac{u_2u'_2}{I_2}+\frac{u_3u'_3}{I_3}\\ &= \frac{u_1a_1u_2u_3}{I_1}+\frac{u_2a_2u_3u_1}{I_2}+\frac{u_3a_3u_1u_2}{I_3}\\ &= u_1u_2u_3(\frac{a_1}{I_1}+\frac{a_2}{I_2}+\frac{a_3}{I_3})\\ &= \frac{u_1u_2u_3}{I_1I_2I_3}(I_2-I_3+I_3-I_1+I_1-I_2)=0\end{aligned} \tag{13}$$

The solution lies at the intersection of a sphere and an ellipsoid.

There are two invariants: $\boldsymbol{u}^T C_1 \boldsymbol{u}$ and $\boldsymbol{u}^T C_2 \boldsymbol{u}$ with

$$C_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{et} \quad C_2 = \frac{1}{2} \begin{bmatrix} 1/I_1 & 0 & 0 \\ 0 & 1/I_2 & 0 \\ 0 & 0 & 1/I_3 \end{bmatrix} \tag{14}$$

To generalize, we consider the equation

$$U' = A(U)U, \tag{15}$$

where A is an antisymmetric matrix, $A^T = -A$, and U can also be a matrix. We see that $F(U) = U^T U$ is an invariant:

$$(\nabla F) \cdot f = U^T (AU) + (AU)^T U = U^T AU + U^T A^T U = 0. \tag{16}$$

To know if a method preserves the quadratic invariants, we use $C = C^T$ and we note that

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{u}^T C \boldsymbol{u}) = \boldsymbol{u}'^T C \boldsymbol{u} + \boldsymbol{u}^T C \boldsymbol{u}' = 2\boldsymbol{u}^T C \boldsymbol{u}' = 2\boldsymbol{u}^T C \boldsymbol{f}(\boldsymbol{u}). \tag{17}$$

There are methods that ensure $\boldsymbol{u}^T C \boldsymbol{f}(\boldsymbol{u}) = const$ for all $\boldsymbol{u} \in \mathbb{R}^m$.
For example, *the implicit midpoint method*,

$$\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + h\boldsymbol{f}\left(\frac{\boldsymbol{u}_n + \boldsymbol{u}_{n+1}}{2}\right). \tag{18}$$

To prove this, let us define $\boldsymbol{f}_n^{mid} := \boldsymbol{f}\left((\boldsymbol{u}_n + \boldsymbol{u}_{n+1})/2\right)$.

$$\begin{aligned} \boldsymbol{u}_{n+1}^T C \boldsymbol{u}_{n+1} &= (\boldsymbol{u}_n + h\boldsymbol{f}_n^{mid})^T C (\boldsymbol{u}_n + h\boldsymbol{f}_n^{mid}) \\ &= \boldsymbol{u}_n^T C \boldsymbol{u}_n + 2h\boldsymbol{u}_n^T C \boldsymbol{f}_n^{mid} + h^2 (\boldsymbol{f}_n^{mid})^T C \boldsymbol{f}_n^{mid}. \end{aligned} \tag{19}$$

We can rewrite the method (18) as

$$\boldsymbol{u}_n = \frac{\boldsymbol{u}_n + \boldsymbol{u}_{n+1}}{2} - \frac{1}{2} h \boldsymbol{f}_n^{mid}, \tag{20}$$

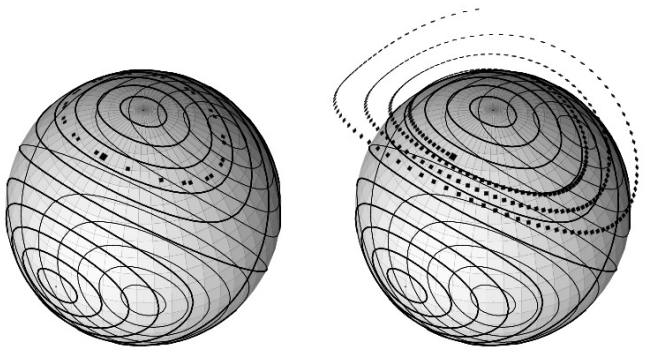
and we obtain

$$\begin{aligned} &\boldsymbol{u}_{n+1}^T C \boldsymbol{u}_{n+1} \\ &= \boldsymbol{u}_n^T C \boldsymbol{u}_n + 2h \left(\frac{\boldsymbol{u}_n + \boldsymbol{u}_{n+1}}{2} - \frac{1}{2} h \boldsymbol{f}_n^{mid} \right)^T C \boldsymbol{f}_n^{mid} + h^2 (\boldsymbol{f}_n^{mid})^T C \boldsymbol{f}_n^{mid} \\ &= \boldsymbol{u}_n^T C \boldsymbol{u}_n + 2h \left(\frac{\boldsymbol{u}_n + \boldsymbol{u}_{n+1}}{2} \right)^T C \boldsymbol{f}_n^{mid} \\ &= \boldsymbol{u}_n^T C \boldsymbol{u}_n + 2h \left(\frac{\boldsymbol{u}_n + \boldsymbol{u}_{n+1}}{2} \right)^T C \boldsymbol{f} \left(\frac{\boldsymbol{u}_n + \boldsymbol{u}_{n+1}}{2} \right) = \boldsymbol{u}_n^T C \boldsymbol{u}_n. \end{aligned}$$

In the same way, we see that the Euler method $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + h\boldsymbol{f}_n$, in general éral, does not preserve quadratic invariants.

$$\begin{aligned}\boldsymbol{u}_{n+1}^T C \boldsymbol{u}_{n+1} &= (\boldsymbol{u}_n + h\boldsymbol{f}_n)^T C (\boldsymbol{u}_n + h\boldsymbol{f}_n) \\ &= \boldsymbol{u}_n^T C \boldsymbol{u}_n + 2h\boldsymbol{u}_n^T C \boldsymbol{f}_n + h^2 (\boldsymbol{f}_n)^T C \boldsymbol{f}_n = \boldsymbol{u}_n^T C \boldsymbol{u}_n + h^2 (\boldsymbol{f}_n)^T C \boldsymbol{f}_n.\end{aligned}$$

In general, $(\boldsymbol{f}_n)^T C \boldsymbol{f}_n \neq 0$. In the example of the rotating solid body, we note that $(\boldsymbol{f}_n)^T C_1 \boldsymbol{f}_n > 0$ and $(\boldsymbol{f}_n)^T C_2 \boldsymbol{f}_n > 0$ for all $\boldsymbol{f}_n \neq 0$. A comparison of midpoint (left) and Euler (right) methods [Hairer, Lubich & Wanner. Geometric numerical integration]:



Theorem (Cooper 1987). If the coefficients of a Runge–Kutta method satisfy

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \quad \text{for all } i, j = 1, ..., s, \tag{21}$$

it preserves quadratic invariants.

Proof. By definition, $\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i$, and we obtain

$$\boldsymbol{u}_{n+1}^T C \boldsymbol{u}_{n+1} = \boldsymbol{u}_n^T C \boldsymbol{u}_n + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i^T C \boldsymbol{u}_n + h \sum_{j=1}^{s} b_j \boldsymbol{u}_n^T C \boldsymbol{k}_j + h^2 \sum_{i,j=1}^{s} b_i b_j \boldsymbol{k}_i^T C \boldsymbol{k}_j, \tag{22}$$

where $\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{U}_i)$ with $\boldsymbol{U}_i = \boldsymbol{u}_n + h \sum_{j=1}^{s} a_{ij} \boldsymbol{k}_j$. We solve this last formula for with respect to $\boldsymbol{u}_n$ and substitute the result into (22).

$$\boldsymbol{u}_{n+1}^T C \boldsymbol{u}_{n+1} = \boldsymbol{u}_n^T C \boldsymbol{u}_n + 2h \sum_{i=1}^{s} b_i \boldsymbol{U}_i^T C \boldsymbol{f}(\boldsymbol{U}_i) + h^2 \sum_{i,j=1}^{s} (b_i b_j - b_i a_{ij} - b_j a_{ji}) \boldsymbol{k}_i^T C \boldsymbol{k}_j.$$

The second term is zero and the third gives the condition of the theorem. □

The criterion (21) is very restrictive. Explicit methods cannot verify this and there are few implicit methods that do. Gaussian methods satisfy this criterion.

Projection methods.

Suppose we have a submanifold of dimension $m - \mu$ of $\mathbb{R}^m$,

$$\mathcal{M} = \{\boldsymbol{u}; \boldsymbol{g}(\boldsymbol{u}) = 0\}\,, \tag{23}$$

($\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^\mu$), and the equation $\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u})$ which has the following property:

$$\boldsymbol{u}_0 \in \mathcal{M} \quad \Rightarrow \quad \boldsymbol{u}(t) \in \mathcal{M} \quad \text{at any } t. \tag{24}$$

This condition is equivalent to $(\boldsymbol{\nabla g})\boldsymbol{f} = \boldsymbol{0}$ for $\boldsymbol{u} \in \mathcal{M}$. This is weaker than the condition introduced in the invariant definition. We say that $\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u})$ is a differential equation on a submanifold.

Example. The pendulum equation in the Cartesian coordinate system,

$$\begin{aligned} q_1' &= p_1 \qquad p_1' = -q_1\lambda \\ q_2' &= p_2 \qquad p_2' = -1 - q_2\lambda \end{aligned} \tag{25}$$

where $\lambda = (p_1^2 + p_2^2 - q_2)/(q_1^2 + q_2^2)$. We see that $F = q_1p_1 + q_2p_2$ is invariant, i.e., this quantity satisfies (2):

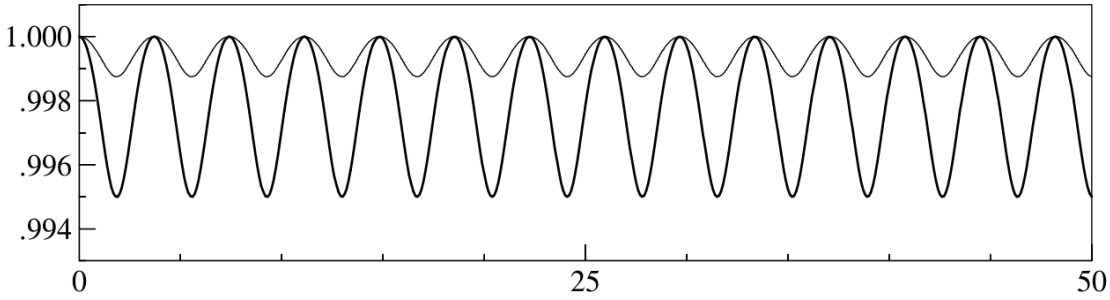
$$\frac{\partial F}{\partial q_1} = p_1, \quad \frac{\partial F}{\partial q_2} = p_2, \quad \frac{\partial F}{\partial p_1} = q_1, \quad \frac{\partial F}{\partial p_2} = q_2,$$

and we obtain

$$p_1p_1 + p_2p_2 - q_1q_1\lambda - q_2(1 + q_2\lambda) = 0. \tag{26}$$

This corresponds to the orthogonality of the position vector and the velocity. The length of the pendulum is also constant, but $q_1^2+q_2^2$ is only a weak invariant.

The figure shows the time evolution of $q_1^2+q_2^2$ obtained by performing numerical calculations. The two curves correspond to the results of two calculations with the midpoint method with two values of h. We see that $q_1^2+q_2^2$ is not conserved. In general, if a method can compute $\boldsymbol{f}(\boldsymbol{u})$ for $\boldsymbol{u} \neq \mathcal{M}$, it cannot not preserve weak invariants.



Algorithm (projection method). Suppose that $\boldsymbol{u}_n \in \mathcal{M}$. A step $\boldsymbol{u}_n \longmapsto \boldsymbol{u}_{n+1}$ is decomposed in two stages:

- Calculate $\tilde{\boldsymbol{u}}_{n+1} = \boldsymbol{\Phi}_h(\boldsymbol{u}_n)$, where $\boldsymbol{\Phi}_h$ describes any method applied to $\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u})$.
- Project $\tilde{\boldsymbol{u}}_{n+1}$ on the manifold $\mathcal{M}$ to obtain $\boldsymbol{u}_{n+1} \in \mathcal{M}$.

The distance from $\tilde{\boldsymbol{u}}_{n+1}$ to $\mathcal{M}$ is of the same order of magnitude as the local truncation error, $\mathcal{O}(h^{p+1})$. Consequently, the projection does not deteriorate the order of convergence of the method.
To get $\tilde{\boldsymbol{u}}_{n+1}$, we solve the constrained minimization problem,

$$||\boldsymbol{u}_{n+1} - \tilde{\boldsymbol{u}}_{n+1}|| \to \min \quad \text{and} \quad \boldsymbol{g}(\boldsymbol{u}_{n+1}) = 0. \tag{27}$$

Let us introduce the Lagrange multipliers, $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_\mu)$. Let us consider the function $\mathcal{L}(\boldsymbol{u}_{n+1}, \boldsymbol{\lambda}) = ||\boldsymbol{u}_{n+1} - \tilde{\boldsymbol{u}}_{n+1}||/2 - \boldsymbol{g}(\boldsymbol{u}_{n+1})^T\boldsymbol{\lambda}$. The condition $\partial\mathcal{L}/\partial\boldsymbol{u}_{n+1} = 0$ gives the following system:

$$\begin{cases} \boldsymbol{u}_{n+1} &= \tilde{\boldsymbol{u}}_{n+1} + \boldsymbol{g}'(\boldsymbol{u}_{n+1})^T \boldsymbol{\lambda} \\ \boldsymbol{0} &= \boldsymbol{g}(\boldsymbol{u}_{n+1}). \end{cases} \tag{28}$$

We use the Newton–Raphson method to calculate $\boldsymbol{\lambda}$,

$$\begin{aligned} &\boldsymbol{\lambda}^{[l+1]} = \boldsymbol{\lambda}^{[l]} + \delta\boldsymbol{\lambda}^{[l]}, \quad \text{where} \\ &\delta\boldsymbol{\lambda}^{[l]} = -\left(\boldsymbol{g}'(\tilde{\boldsymbol{u}}_{n+1})\boldsymbol{g}'(\tilde{\boldsymbol{u}}_{n+1})^T\right)^{-1} \boldsymbol{g}\left(\tilde{\boldsymbol{u}}_{n+1} + \boldsymbol{g}'(\tilde{\boldsymbol{u}}_{n+1})^T \boldsymbol{\lambda}^{[l]}\right). \end{aligned} \tag{29}$$

If $\boldsymbol{\lambda}^{[0]}$, we have $\delta\boldsymbol{\lambda}^{[0]} = \mathcal{O}(h^{p+1})$.

Symplectic integrators.

Consider two vectors, $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^2$. We can define the oriented area of the corresponding parallelogram as

$$a(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T J \boldsymbol{v}, \quad \text{where} \quad J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{30}$$

Consider a matrix $P \in \mathbb{R}^{2\times 2}$ and a mapping $\boldsymbol{u} \longmapsto P\boldsymbol{u}$, $\boldsymbol{v} \longmapsto P\boldsymbol{v}$. This mapping does not change the area iff

$$P^T J P = J. \tag{31}$$

A similar condition for a nonlinear map $\boldsymbol{g} : \mathbb{R}^2 \to \mathbb{R}^2$ is

$$\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}\right)^T J \left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}\right) = J. \tag{32}$$

The maps that satisfy this condition are *symplectic*.

Consider a Hamiltonian system,

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial H(p,q)}{\partial q}, \quad \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial H(p,q)}{\partial p}. \tag{33}$$

Let us introduce a map

$$\psi_t \left(\begin{bmatrix} p_0 \\ q_0 \end{bmatrix} \right) = \begin{bmatrix} p(t) \\ q(t) \end{bmatrix}. \tag{34}$$

This map conserves the oriented area and

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(P(t)^T J P(t) \right) = 0, \tag{35}$$

where

$$P(t) = \frac{\partial \psi_t}{\partial \boldsymbol{u}_0}, \quad \boldsymbol{u}_0 = [p_0, q_0]^T. \tag{36}$$

The property (35) of the Hamiltonian system is preserved by *the symplectic Euler method*, which is a combination of the methods forward and backward:

$$\begin{aligned} p_{n+1} &= p_n - h\frac{\partial H}{\partial q}(p_{n+1}, q_n), \\ q_{n+1} &= q_n + h\frac{\partial H}{\partial p}(p_{n+1}, q_n). \end{aligned} \tag{37}$$

It turns out that a Runge–Kutta method is symplectic if it satisfies the same condition (21) which ensures the conservation of quadratic invariants.