Introduction to numerical methods for Ordinary Differential Equations

C. -E. Bréhier

November 14, 2016

Abstract

The aims of these lecture notes are the following.

- We introduce Euler numerical schemes, and prove existence and uniqueness of solutions of ODEs passing to the limit in a discretized version.
- We provide a general theory of one-step integrators, based on the notions of stability and consistency.
- We propose some procedures which lead to construction of higher-order methods.
- We discuss splitting and composition methods.
- We finally focus on qualitative properties of Hamiltonian systems and of their discretizations.

Good references, in particular for integration of Hamiltonian systems, are the monographs **Geometric Numerical Integration**, by E. Hairer, C. Lubich and G. Wanner, and **Molecular Dynamics. With deterministic and stochastic numerical methods.**, by B. Leimkuhler and C. Matthews.

Contents

1 Ordinary Differential Equations

The main objective of these lecture notes is to present simulatable approximations of solutions of **Ordinary Differential Equations (ODEs)** of the form

$$x' = F(x) \tag{1}$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is a vector field, and $d \in \mathbb{N}^\star$. The minimal assumption on F is continuity; however well-posedness requires stronger conditions, in terms of (local) **Lipschitz continuity**.

We recall what we mean by a solution of (1).

Definition 1.1. *Let $T \in (0, +\infty)$; a solution of (1) on the interval $[0, T]$ is a mapping $x : [0, T] \to \mathbb{R}^d$, of class $\mathcal{C}^1$, such that for every $t \in [0, T]$, $x'(t) = F\big(x(t)\big)$.*

We also recall the definition of a Lipschitz continuous function.

Definition 1.2. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$.*

F is globally Lipschitz continuous if there exists $\mathrm{L}_F \in (0, +\infty)$ such that for every $x_1, x_2 \in \mathbb{R}^d$

$$\|F(x_2) - F(x_1)\| \leq \mathrm{L}_F \|x_2 - x_1\|.$$

F is locally Lipschitz continuous if, for every $R \in (0, +\infty)$, there exists $\mathrm{L}_F(R) \in (0, +\infty)$ such that for every $x_1, x_2 \in \mathbb{R}^d$, with $\max\big(\|x_1\|, \|x_2\|\big) \leq R$,

$$\|F(x_2) - F(x_1)\| \leq \mathrm{L}_F(R) \|x_2 - x_1\|.$$

1.1 Well-posedness theory

In order to deal with solutions of (1), we first need to study well-posedness of the **Cauchy problem**:

$$x' = F(x) \quad , \quad x(0) = x_0, \tag{2}$$

where $x_0 \in \mathbb{R}^d$ is an initial condition.

The Cauchy problem is globally **well-posed in the sense of Hadamard** if, firstly, for every $T \in (0, +\infty)$, there exists a unique solution x of (1) on $[0, T]$, such that $x(0) = x_0$; and if, secondly, the solution depends continuously on the initial condition x_0.

Assuming the global Lipschitz continuity of F provides global well-posedness: this is the celebrated **Cauchy-Lipschitz theorem**. If one only assumes that F is locally Lipschitz continuous, well-posedness is only local: the existence time T may depend on the initial condition x_0. However, there are criteria ensuring global existence of solutions. Finally, if F is only assumed to be continuous, uniqueness may not be satisfied; however existence may be proven (this is the Cauchy-Peano theorem), using the strategy presented below (and an appropriate compactness argument). Such situations are not considered in these notes.

Theorem 1.3 (Cauchy-Lipschitz)**.** *Assume that F is (globally) Lipschitz continuous. Then the Cauchy problem is well-posed in the sense of Hadamard: for every $T \in (0,+\infty)$ and every $x_0 \in \mathbb{R}^d$, there exists a unique x solution of (2).*

Then let $\phi : [0,+\infty) \times \mathbb{R}^d \to \mathbb{R}^d$, defined by $\phi(t,x_0) = x(t)$. There exists $C \in (0,+\infty)$, such that, for every $T \in (0,+\infty)$, and every $x_1, x_2 \in \mathbb{R}^d$

$$\sup_{t\in[0,T]} \|\phi(t,x_1) - \phi(t,x_2)\| \leq \exp\big(CT\big)\|x_1 - x_2\|.$$

The mapping ϕ is called the **flow** of the ODE. The following notation is often used: $\phi_t = \phi(t,\cdot)$, for every $t \geq 0$. Thanks to the Cauchy-Lipschitz Theorem 1.3, for every $t, s \geq 0$,

$$\phi_{t+s} = \phi_t \circ \phi_s. \tag{3}$$

This important equality is the flow property (or the semigroup property).

Note that if F is of class $\mathcal{C}^k$ for some $k \in \mathbb{N}^\star$, then the flow is also of class $\mathcal{C}^k$, *i.e.* Φ_t is of class $\mathcal{C}^k$ for every $t \geq 0$. We do not prove this result in these notes.

The key tool to prove all the results in Theorem 1.3 (as well as many other results) is the celebrated **Gronwall's Lemma**.

Lemma 1.4 (Gronwall's Lemma)**.** *Let $T \in (0,+\infty)$, and $\theta : [0,T] \to [0,+\infty)$ a continuous, nonnegative, function.*

Assume that there exists $\alpha \in [0,+\infty)$ and $\beta \in (0,+\infty)$ such that for every $t \in [0,T]$

$$\theta(t) \leq \alpha + \beta \int_0^t \theta(s)ds.$$

Then $\theta(t) \leq \alpha \exp(\beta T)$ for every $t \in [0,T]$.

Proof of Gronwall's Lemma. The mapping $t \in [0,T] \mapsto \frac{\alpha+\beta\int_0^t \theta(s)ds}{\alpha e^{\beta t}}$ is of class $\mathcal{C}^1$, non-increasing (its derivative is nonpositive), and is equal to 1 at $t = 0$. □

We are now in position to prove Theorem (1.3). To simplify the presentation, in addition to the global Lipschitz condition, we assume that F is bounded: there exists $M \in (0,+\infty)$ such that $\|F(x)\| \leq M$ for every $x \in \mathbb{R}^d$.

The standard proof of the Cauchy-Lipschitz theorem exploits the Picard iteration procedure, or the Banach fixed point theorem. Here we present a different proof, where we construct approximate solutions by means of a numerical scheme; this strategy can be generalized to prove the Cauchy-Peano Theorem, assuming only that F is continuous. The proof presented here also supports the fact that numerical approximation may also be a nice approach in theoretical problems.

Proof of Cauchy-Lipschitz theorem. 1. Uniqueness: let x_1, x_2 denote two solutions on $[0,T]$. Then for every $t \in [0,T]$

$$\|x_1(t) - x_2(t)\| = \Big\| \int_0^t \big(F(x_1(s)) - F(x_2(s))\big)ds\Big\| \leq \mathrm{L}_F \int_0^t \|x_1(s) - x_2(s)\|ds,$$

where we have used $x_1(0) = x_0 = x_2(0)$, and Lipschitz continuity of F.

The conclusion follows from Gronwall's Lemma: $\|x_1(t) - x_2(t)\| \leq 0$ for every $t \in [0, T]$.

2. Existence: we construct a sequence of approximate solutions by means of a numerical scheme; we then prove (uniform) convergence of the sequence, and prove that the limit is solution of (1).

   Let $N \in \mathbb{N}^\star$, and denote $h = \frac{T}{2^N}$, and $t_n = nh$ for every $n \in \left\{0, 1, \ldots, 2^N\right\}$.

   Define

$$\begin{aligned} x^N(t_0) &= x_0, \\ x^N(t_{n+1}) &= x^N(t_n) + hF(x^N(t_n)), \quad n \in \left\{0, 1, \ldots, 2^N - 1\right\}, \\ x^N(t) &= \frac{(t_{n+1} - t)}{t_{n+1} - t_n} x^N(t_n) + \frac{(t - t_n)}{t_{n+1} - t_n} x^N(t_{n+1}), \quad t \in [t_n, t_{n+1}], n \in \left\{0, 1, \ldots, 2^N - 1\right\}. \end{aligned} \tag{4}$$

   The initial condition for x^N is x_0 and does not depend on N. The second line defines $x^N(t_n)$ recursively, for $n \in \left\{0, 1, \ldots, 2^N - 1\right\}$. Finally, the third line defines a continuous function $x^N : [0, T] \to \mathbb{R}^d$ by linear interpolation. Note that x^N is continuous, and is moreover differentiable, except at points t_n (where left and right derivatives exist).

   Let $n \in \left\{0, 1, \ldots, 2^N - 1\right\}$. Then

$$\|(x^N)'(t)\| = \frac{\|x^N(t_{n+1}) - x^N(t_n)\|}{h} = \|F(x^N(t_n))\| \leq M$$

   for every $t \in (t_n, t_{n+1})$, and thus $\|x^N(t) - x^N(t_n)\| \leq M(t - t_n) \leq Mh$.

   This yields $\|(x^N)'(t) - F(x^N(t))\| = \|F(x^N(t)) - F(x^N(t_n))\| \leq \mathrm{L}_F Mh$, for every $t \in (t_n, t_{n+1})$. By subdividing the interval $[0, t]$ as $\bigcup_{0 \leq k \leq K-1} [t_k, t_{k+1}) \cup [t_K, t]$, with K such that $t_K \leq t < t_{K+1}$, and integrating, for every $t \in [0, T]$

$$\|x^N(t) - x_0 - \int_0^t F(x^N(s))ds\| \leq \mathrm{L}_F Mth. \tag{5}$$

   Now let $z^N = x^{N+1} - x^N$. Then, using the triangle inequality,

$$\begin{aligned} \|z^N(t)\| &= \|x^{N+1}(t) - x^N(t)\| \\ &\leq \| \int_0^t F(x^{N+1}(s))ds - \int_0^t F(x^N(s))ds\| \\ &+ \|x^{N+1}(t) - x_0 - \int_0^t F(x^{N+1}(s))ds\| + \|x^N(t) - x_0 - \int_0^t F(x^N(s))ds\| \\ &\leq \mathrm{L}_F \int_0^t \|z^N(s)\|ds + 2\mathrm{L}_F M \frac{T^2}{2^N}. \end{aligned}$$

Thanks to Gronwall's Lemma, $\sup_{0\leq t\leq T}\|x^{N+1}(t)-x^N(t)\|\leq 2\mathrm{L}_F M\frac{T^2}{2^N}$.

We deduce that for every $t\in[0,T]$, the series $\sum_{N\in\mathbb{N}}\big(x^{N+1}(t)-x^N(t)\big)$ converges; moreover, the convergence is uniform. Define $x(t)=\sum_{N=0}^{+\infty}\big(x^{N+1}(t)-x^N(t)\big)+x_0(t)$.

Note that $x:[0,T]\to\mathbb{R}^d$ is continuous, as the uniform limit of continuous functions. Moreover, $x(t)=\lim_{N\to+\infty}x^N(t)$ (telescoping sum argument).

Since F is continuous, we can pass to the limit $N\to+\infty$ in the left-hand side of (5); since $h=T2^{-N}$, we get the equality

$$x(t)=x_0+\int_0^t F(x(s))ds$$

for every $t\in[0,T]$, which is equivalent to (1). This concludes the proof of the existence result.

3. Lipschitz continuity of the flow: if $x_1,x_2\in\mathbb{R}^d$, by the triangle inequality

$$\|\phi_t(x_1)-\phi_t(x_2)\|\leq\|x_1-x_2\|+\int_0^t\|F(\phi_s(x_1))-F(\phi_s(x_2))\|ds,$$

for every $t\in[0,T]$. Thanks to Lipschitz continuity of F and Gronwall's Lemma, we get

$$\sup_{0\leq t\leq T}\|\phi_t(x_1)-\phi_t(x_2)\|\leq e^{\mathrm{L}_F T}\|x_1-x_2\|.$$

□

1.2 Some important examples

1.2.1 Linear ODEs

If $F(x)=Ax$ for every $x\in\mathbb{R}^d$, where A is a $d\times d$ real-valued matrix, then for every $t\geq 0$ and every $x_0\in\mathbb{R}^d$, one has

$$\phi_t(x_0)=e^{tA}x_0,$$

where $e^{tA}=\sum_{n=0}^{+\infty}\frac{t^n}{n!}A^n$ is the exponential of the matrix tA.

Properties of theses ODEs can be studied by choosing an appropriate representative B, such that $B=P^{-1}AP$ for some invertible matrix P.

For instance, if A can be diagonalized, it is straightforward to provide explicit solutions of the ODE, in terms of the eigenvalues and of the eigenvectors.

1.2.2 Gradient dynamics

Let $V:\mathbb{R}^d\to\mathbb{R}$ be of class $\mathcal{C}^2$. V is called the **potential energy function**.

The **gradient dynamics** corresponds to the choice $F(x) = -\nabla V(x)$:

$$x' = -\nabla V(x).$$

Without further assumptions on the growth of V, solutions are *a priori* only local, but they satisfy the inequality $V\big(x(t)\big) \leq V\big(x(0)\big)$: indeed,

$$\frac{dV\big(x(t)\big)}{dt} = -\|\nabla V\big(x(t)\big)\|^2 \leq 0.$$

If for every $r \in \mathbb{R}$, the level set $\big\{x \in \mathbb{R}^d \ ; \ V(x) \leq r\big\}$ is compact, then solutions are global. The compact level set condition is satisfied for instance if there exists $c, R \in (0, +\infty)$ such that $\|V(x)\| \geq c\|x\|^2$ when $\|x\| \geq R$.

Note that a linear ODE, $F(x) = Ax$, corresponds to a gradient dynamics if and only if A is symmetric (Schwarz relations). In that case, $A = -D^2V(x)$ (the Hessian matrix) for every $x \in \mathbb{R}^d$, and $V(x) = -\frac{1}{2}\langle x, Ax\rangle$. The level sets are compact if and only if $-A$ only has nonnegative eigenvalues.

1.2.3 Hamiltonian dynamics

Let $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be of class $\mathcal{C}^2$. H is called the **Hamiltonian function**.

The **Hamiltonian dynamics** is given by the ODE

$$\begin{cases} \dot{q} = \nabla_p H(q,p) \\ \dot{p} = -\nabla_q H(q,p) \end{cases}$$

where we used the standard notation $\dot{q}$ and $\dot{p}$ to represent the time derivative.

The typical example from mechanics is given by the total energy function $H(q,p) = \frac{\|p\|^2}{2} + V(q)$, where $q \in \mathbb{R}^d$ represents the position, $p \in \mathbb{R}^d$ represents the momentum. The total enery function is the sum of the kinetic energy $\frac{\|p\|^2}{2}$ and of the potential energy $V(q)$. The Hamiltonian dynamics is then given by

$$\dot{q} = p \quad , \quad \dot{p} = -\nabla V(q)$$

which can be rewritten as a second-order ODE: $\ddot{q} = -\nabla V(q)$.

One of the remarkable properties of Hamiltonian dynamics is the preservation of the Hamiltonian function: if $t \mapsto \big(q(t), p(t)\big)$ is a local solution, then $H\big(q(t), p(t)\big) = H\big(q(0), p(0)\big)$. If for every $H_0 \in \mathbb{R}$, the level set $\big\{(q,p) \in \mathbb{R}^{2d} \ ; \ H(q,p) = H_0\big\}$ is compact, then solutions are global.

1.2.4 Examples of potential energy functions

The first example of potential function corresponds to the so-called harmonic oscillator (when considering the associated Hamiltonian system): $V(q) = \frac{\omega^2}{2}q^2$. The equation is linear, this gives a good toy model to test properties of numerical methods.

The harmonic oscillator models a linear pendulum: a nonlinear example is given by $V(q) = \omega^2(1 - \cos(q))$, which gives $-V'(q) = -\omega^2 \sin(q)$.

Even if the case of harmonic oscillators might seem very simple, chains of oscillators might be challenging in presence of slow and fast oscillations.

For instance, for ω large, consider a chain of $2m+2$ oscillators with Hamiltonian function

$$H\big(q_1, \dots, q_{2m+1}, p_1, \dots, p_{2m+1}\big) = \sum_{i=1}^{2m} p_i^2 + \sum_{i=1}^{m} \frac{\omega^2}{2}(q_{2i} - q_{2i-1})^2 + \sum_{i=0}^{m} \frac{1}{2}(q_{2i+1} - q_{2i})^2,$$

with $q_0 = q_{2m+1} = 0$ by convention, where stiff (frequency ω) and nonstiff (frequency 1) springs are alternated, and $q_1, \dots, q_{2m}$ indicate displacements with respect to equilibrium positions of springs. This is an example called the **Fermi-Pasta-Ulam model**.

In general, systems of ODEs with multiple time scales (stiff springs evolve at a fast time scale and nonstiff springs evolve at a slow time scale) require the use of specific multiscale methods. This is still an active research area.

When studying systems with a large number of particles (**molecular dynamics**) or planets (**celestial mechanics**), it is natural to consider potential energy functions which are the sums of internal and pairwise interaction terms:

$$V(q) = \sum_{i=1}^{L} V_i(q_i) + \sum_{1 \le i < j \le L} V_{ij}(q_i, q_j).$$

Typically, homogeneous interactions are considered: $V_{ij}(q_i, q_j)$ is a function $V_{ij}(q_j - q_i)$ of $q_j - q_i$.

In celestial mechanics, with $d = 3$, typically $V_i(q) = -\frac{GMm_i}{\|q\|}$ and $V_{ij}(q_i, q_j) = -\frac{Gm_i m_j}{\|q_i - q_j\|}$, where m_i is the mass of planet i and M is the (large) mass of another planet or star, and $G > 0$ is a physical parameter.

In molecular dynamics, a popular example is given by the Lennard-Jones potential

$$V_{ij}(q_i, q_j) = 4\epsilon_{ij}\Big(\frac{\sigma_{ij}^{12}}{\|q_j - q_i\|^{12}} - \frac{\sigma_{ij}^{6}}{\|q_j - q_i\|^{6}}\Big).$$

The force is repulsive for short distances, and attractive for sufficiently large distances.

Both the examples in celestial mechanics and the molecular dynamics are challenging, in particular due to the singularity of the potential functions at 0, and to the chaotic behavior they induce.

1.2.5 The Lotka-Volterra system

The **Lotka-Volterra** system is a popular predator-prey model: if $u(t)$, resp. $v(t)$, is the size of the population of predators, resp. of preys, at time t, the ODE system is given by

$$\dot{u} = u(v - 2) \quad , \quad \dot{v} = v(1 - u),$$

where the values 2 and 1 are chosen arbitrarily.

A first integral of the system is given by $I(u,v) = \ln(u) - u + 2\ln(v) - v$: this means that I is invariant by the flow, equivalently $\frac{d}{dt} I(u(t), v(t)) = 0$.

This property is important, since it implies that the trajectories lie on level sets of I, and that in fact solutions of the ODE are periodic.

It would be desirable to have numerical methods which do not destroy too much this nice property.

The conservation of I is not surprising: setting $p = \ln(u)$ and $q = \ln(v)$, and $H(q,p) = I(e^q, e^p) = p - e^p + 2q - e^q$, the Lotka-Volterra system is transformed into an Hamiltonian system. In other words, the flow of the Lotka-Volterra system is conjugated (via the change of variables) to an Hamiltonian flow.

We will see later on nice integrators for Hamiltonian dynamics: they naturally yield nice integrators for the Lotka-Volterra system.

2 Euler schemes

The proof of the existence part of Theorem 1.3 used the construction of approximate solutions, by means of a numerical method, see (4).

The principle was as follows: if $t \mapsto x(t)$ is differentiable at time t_0, then $x'(t) = \lim_{h\to 0} \frac{x(t+h)-x(t)}{h}$. In other words, $x(t+h) = x(t) + hx'(t) + \mathrm{o}(h)$. If x is solution of (1), then $x'(t) = F(x(t))$.

Neglecting the $\mathrm{o}(h)$ term leads to the definition of a recursion of the form

$$x_{n+1}^h = x_n^h + hF(x_n^h).$$

This sequence is well-defined provided that an initial condition x_0^h is given. Moreover, the condition to have a simulatable sequence is that $F(x)$ can be computed for every $x \in \mathbb{R}^d$.

It is natural to assume that $h = \frac{T}{N}$, where $N \in \mathbb{N}^\star$ is an integer. For $n \in \{0, \dots, N\}$, x_n^h is an approximation of $x(nh)$, the value of the exact solution at time $t_n = nh$.

Note that in the proof of Theorem 1.3, we have considered the case $h = T2^{-N}$, in order to easily set up the telescoping sum argument.

2.1 The explicit Euler scheme

The **explicit Euler scheme** (also called the forward Euler scheme), with time-step size h, associated with the Cauchy problem (2) is given by

$$x_0^h = x_0 \quad , \quad x_{n+1}^h = x_n^h + hF(x_n^h). \tag{6}$$

We prove the following result:

Theorem 2.1. *Let $T \in (0, +\infty)$, and consider the explicit Euler scheme (6) with time step size $h = \frac{T}{N}$, with $N \in \mathbb{N}^\star$.*

Assume that the vector field F is bounded and Lipschitz continuous.

Then there exist $c, C \in (0, +\infty)$ such that for every $N \in \mathbb{N}$,

$$\sup_{0\le n\le N} \|x(t_n) - x_n^h\| \le e^{cT} h,$$

where x is the unique solution of (2).

We give two different detailed proofs, even if we give a more general statement below which includes the case of the explicit Euler scheme. The first proof is similar to the computations in the proof of Theorem 1.3. The second proof allows us to exhibit the two fundamental properties of numerical schemes which will be studied below: stability and consistency.

First proof of Theorem 2.1. Let $t_n = nh$, for $n \in \mathbb{N}$.

Introduce the right-continuous, piecewise linear, respectively the piecewise constant, interpolations of the sequence $\big(x_n^h\big)_{n\geq 0}$: define for $t\in[t_n,t_{n+1})$, $n\in\mathbb{N}$,

$$x^h(t)=\frac{t_{n+1}-t}{t_{n+1}-t_n}x_n^h+\frac{t-t_n}{t_{n+1}-t_n}x_{n+1}^h$$
$$\overline{x}^h(t)=x_n^h.$$

Then $x^h(t)=x_0+\int_0^t F\big(\overline{x}^h(s)\big)ds$.

Note also that, for every $n\in\mathbb{N}$ and $t\in[t_n,t_{n+1}]$, $\|x^h(t)-\overline{x}^h(t)\|\leq\|x_{n+1}^h-x_n^h\|\leq Mh$.

Let $\varepsilon^h(t)=x(t)-x^h(t)$, where x is the solution of (2). Then for every $t\in[0,T]$,

$$\begin{aligned}\|\varepsilon^h(t)\|&\leq\int_0^t\|F\big(x(s)\big)-F\big(\overline{x}^h(s)\big)\|ds\\&\leq\int_0^t\|F\big(x^h(s)\big)-F\big(\overline{x}^h(s)\big)\|ds+\int_0^t\|F\big(x(s)\big)-F\big(x^h(s)\big)\|ds\\&\leq \mathrm{L}_F Mh+\mathrm{L}_F\int_0^t\|\varepsilon^h(s)\|ds.\end{aligned}$$

By Gronwall's Lemma, $\sup_{t\in[0,T]}\|\varepsilon^h(t)\|\leq \mathrm{L}_F Mhe^{L_F T}$.

□

Second proof of Theorem 2.1. For every $n\in\mathbb{N}$, define $e_n^h=x(t_{n+1})-x(t_n)-hF\big(x(t_n)\big)$.

Note that for $t\in[t_n,t_{n+1}]$, $\|x(t)-x(t_n)\|\leq\int_{t_n}^t\|F\big(x(s)\big)\|ds\leq Mh$. As a consequence,

$$\|e_n^h\|=\|\int_{t_n}^{t_{n+1}}\big(F\big(x(s)\big)-F(x_n^h)\big)ds\|\leq \mathrm{L}_F Mh^2.$$

Let also $\varepsilon_n^h=x(t_n)-x_n^h$. Then

$$\begin{aligned}\|\varepsilon_{n+1}^h\|&=\|x(t_{n+1})-x_{n+1}^h\|\\&=\|x(t_n)+hF\big(x(t_n)\big)+e_n^h-x_n^h-hF\big(x_n^h\big)\|\\&\leq\|e_n^h\|+\big(1+\mathrm{L}_F h\big)\|\varepsilon_n^h\|\\&\leq \mathrm{L}_F Mh^2+\big(1+\mathrm{L}_F h\big)\|\varepsilon_n^h\|.\end{aligned}$$

Dividing both sides by $(1+\mathrm{L}_F h)^{n+1}$ gives

$$\frac{\|\varepsilon_{n+1}^h\|}{(1+\mathrm{L}_F h)^{n+1}}\leq\frac{\|\varepsilon_n^h\|}{(1+\mathrm{L}_F h)^n}+\frac{\mathrm{L}_F Mh^2}{(1+\mathrm{L}_F h)^{n+1}},$$

which yields by a telescoping sum argument (since $\varepsilon_0^h=0$)

$$\begin{aligned}\|\varepsilon_n^h\|&\leq \mathrm{L}_F Mh^2\sum_{k=0}^{n-1}(1+\mathrm{L}_F h)^{n-1-k}\\&\leq \mathrm{L}_F Mh^2\frac{(1+\mathrm{L}_F h)^n-1}{\mathrm{L}_F h}\\&\leq e^{\mathrm{L}_F T}Mh.\end{aligned}$$

□

2.2 The implicit Euler scheme

Before we focus on the general analysis of one-step numerical methods, we introduce the implicit Euler scheme, as an alternative to the explicit Euler scheme.

Let $\lambda > 0$, and consider the linear ODE on $\mathbb{R}$

$$\dot{x} = -\lambda x.$$

Given $x_0 \in \mathbb{R}$, obviously $\phi_t(x_0) = e^{-\lambda t} x_0$. In particular:

- if $x_0 > 0$, then $\phi_t(x_0) > 0$ for every $t > 0$;
- $\phi_t(x_0) \underset{t \to +\infty}{\to} 0$, for every $x_0 \in \mathbb{R}$.

Are these properties satisfied by the explicit Euler scheme with time-step h, when applied in this simple linear situation?

Thanks to (6), one easily sees that $x_n^h = \big(1 - \lambda h\big)^n x_0$.

On the one hand, the positivity property is thus only satisfied if $\lambda h < 1$: thus the time-step h needs to be sufficiently small. On the other hand, $\big(x_n^h\big)_{n \in \mathbb{N}}$ is bounded if and only if $|1 - \lambda h| \leq 1$, *i.e.* $\lambda h \leq 2$; and $x_n^h \underset{n \to +\infty}{\to} 0$ if and only if $\lambda h < 2$.

The conditions $h < \lambda^{-1}$ and $h < 2\lambda^{-1}$ are very restrictive when λ is large.

The alternative to recover stability properties is to use the backward Euler approximation. It is based on writing the Taylor expansion $x(t) = x(t-h) + hx'(t) + \mathrm{o}(h)$, which leads to the following scheme in the linear case:

$$x_{n+1}^h = x_n^h - h\lambda x_{n+1}^h.$$

The solution writes $x_n^h = \frac{1}{(1+\lambda h)^n} x_0$, and both positivity and asymptotic convergence properties of the exact solution are preserved by the numerical scheme, whatever $h > 0$.

In the general case, the **implicit Euler scheme**, also called the backward Euler scheme, with time-step size h is defined by

$$x_{n+1}^h = x_n^h + hF\big(x_{n+1}^h\big). \tag{7}$$

Since at each iteration the unknown x_{n+1}^h appears on both sides of the equation (7), the scheme is implicit, and often the equation cannot be solved explicitly. For instance, one may use the Newton method to compute approximate values.

Moreover, the fact that there exists a unique solution y to the equation $y = x + hF(y)$, for every $x \in \mathbb{R}^d$, may not be guaranteed in general. However, we have the following criterion, which requires again h to be small enough.

Proposition 2.2. *Assume that $h\mathrm{L}_F < 1$. Then, for every $x \in \mathbb{R}^d$, there exists a unique solution $y(x)$ to the equation $y = x + hF(y)$.*

Moreover, $x \mapsto y(x)$ is Lipchitz continuous, and $\|y(x_2) - y(x_1)\| \leq \frac{1}{1 - h\mathrm{L}_F} \|x_2 - x_1\|$.

Proof.

- Uniqueness: if $y = x + hF(y)$ and $z = x + hF(z)$ are two solutions, then

$$\|y - z\| = h\|F(y) - F(z)\| \leq h\mathrm{L}_F\|y - z\|,$$

which implies $\|y - z\|$, and thus $y = z$.

- Existence: let $x \in \mathbb{R}^d$, and consider an arbitrary $y_0 \in \mathbb{R}^d$. Define $y_{k+1} = x + hF(y_k)$.

Then $\|y_{k+1} - y_k\| \leq h\mathrm{L}_F\|y_k - y_{k-1}\| \leq \ldots \leq \big(h\mathrm{L}_F\big)^k\|y_1 - y_0\|$, for every $k \in \mathbb{N}$.

Since $\mathbb{R}^d$ is a complete metric space, we can define $y_\infty = y_0 + \sum_{k=0}^{+\infty}\big(y_{k+1} - y_k\big) = \lim\limits_{k\to+\infty} y_k$.

Passing to the limit $k \to +\infty$ in the equation $y_{k+1} = x + hF(y_k)$, we get $y_\infty = x + hF(y_\infty)$. This gives the existence of a solution y_∞.

- Lipschitz continuity: let $x_2, x_1 \in \mathbb{R}^d$. Then

$$\big(1-h\mathrm{L}_F\big)\|y(x_2)-y(x_1)\| \leq \Big(\|x_2-x_1\|+h\mathrm{L}_F\|y(x_2)-y(x_1)\|\Big)-h\mathrm{L}_F\|y(x_2)-y(x_1)\| \leq \|x_2-x_1\|,$$

and one concludes using $1 - h\mathrm{L}_F > 0$.

□

In the case of the linear equation $\dot{x} = -\lambda x$, with $\lambda > 0$, no such condition is required: positivity of λ guarantees the well-posedness of the scheme.

An implicit scheme may then provide nice stability properties, but may be difficult to be applied in practice for nonlinear problems. It is often a good strategy to provide semi-implicit schemes: a linear part is treated implicitly, and a nonlinear part is treated explicitly, see the example at the beginning of Section 5

2.3 The θ-scheme

The explicit and the implicit Euler schemes belong to the family of θ-method: let $\theta \in [0, 1]$, then the θ-scheme is defined as follows:

$$x_{n+1}^{\theta,h} = x_n^{\theta,h} + (1 - \theta)hF\big(x_n^{\theta,h}\big) + \theta hF\big(x_{n+1}^{\theta,h}\big). \tag{8}$$

The case $\theta = 0$ gives the explicit Euler scheme, whereas the case $\theta = 1$ gives the implicit Euler scheme.

As soon as $\theta > 0$, the θ-scheme (8) is not explicit.

Choosing $\theta = 1/2$, gives a popular method: the **Crank-Nicolson scheme**

$$x_{n+1}^{h} = x_n^{h} + \frac{h}{2}\big(F\big(x_n^{h}\big) + F\big(x_{n+1}^{h}\big)\big). \tag{9}$$

3 General analysis of one-step integrators

3.1 One-step integrators

Let h denote a time-step size.

We consider numerical methods which have the form

$$x_{n+1}^h = \Phi_h(x_n^h). \tag{10}$$

Such schemes are referred to as **one-step integrators**, or simply as integrators.

The terminology "one-step integrator" refers to the fact that the computation of the position x_{n+1}^h, at time $n+1$, only requires to know the position x_n^h, at time n. Higher-order recursions $x_{n+1}^h = \Phi_h(x_n^h, x_{n-1}^h, x_{n-m+1}^h)$ are referred to as "multi-step methods" in the literature.

In the sequel, the following notation is useful: for $h > 0$, let $\Psi_h(x) = \frac{1}{h}\big(\Phi_h(x) - x\big)$. We assume that $\Psi_0(x) = \lim_{h\to 0} \Psi_h(x)$ is also well-defined, for every $x \in \mathbb{R}^d$. We will also sometimes assume that Φ_h is also defined for $h < 0$.

Note that $x_n = \Phi_h^n(x_0)$, where Φ_h^n denotes the composition $\Phi_h \circ \Phi_h^{n-1}$. The sequence $\big(\Phi_h^n\big)_{0\le n\le N}$ is often called the **numerical flow**: indeed, a semi-group property $\Phi_h^{n+m} = \Phi_h^n \circ \Phi_h^m$ is also satisfied, with integers n and m.

3.2 Stability

The first notion is the stability of a numerical scheme.

Definition 3.1. *Let $T \in (0, +\infty)$, and consider an integrator Φ_h.*

The scheme (10) is **stable** *on the interval $[0, T]$ if there exists $h^\star > 0$ and $C \in (0, +\infty)$, such that if $h = \frac{T}{N}$, with $N \ge N^\star > \frac{T}{h^\star}$, then for any sequence $\big(\rho_n\big)_{0\le n\le N}$ in $\mathbb{R}^d$, the sequence $\big(y_n\big)_{0\le n\le N}$ defined by the recursion,*

$$y_{n+1} = \Phi_h(y_n) + \rho_n, \tag{11}$$

satisfies the following inequality: for every $n \in \{0, \dots, N\}$,

$$\|y_n - x_n^h\| \le C\Big(\|y_0 - x_0^h\| + \sum_{m=0}^{n-1} \|\rho_m\|\Big). \tag{12}$$

The following sufficient condition for stability holds true.

Proposition 3.2. *Assume that there exists $\mathrm{L} \in (0, +\infty)$ and $h^\star > 0$ such that, for every $h \in (0, h^\star)$,*

- *$\Phi_h(x) = x + h\Psi_h(x)$ for every $x \in \mathbb{R}^d$;*
- *$\|\Psi_h(x_2) - \Psi_h(x_1)\| \le \mathrm{L}\|x_2 - x_1\|$ for every $x_1, x_2 \in \mathbb{R}^d$.*

Then the scheme (10) *is stable on any interval* $[0,T]$*, with* $C = e^{\mathrm{L}T}$.

Note that $C = e^{\mathrm{L}T} \underset{T\to+\infty}{\to} +\infty$: if the size of the interval is not fixed, the result of the proposition is not sufficient to establish a relationship between the asymptotic behaviors of the solution of the ODE ($T \to +\infty$) and of its numerical approximation ($N \to +\infty$). For instance, this is the situation described by the linear example $\dot{x} = -\lambda x$, with $\lambda > 0$.

Proof. The proof follows from arguments similar to the second proof of Theorem 2.1.

Indeed,

$$\begin{aligned}\|y_{n+1} - x_{n+1}^h\| &\leq \|\rho_n\| + \|y_n - x_n^h\| + h\|\Psi_h(y_n) - \Psi_h(x_n^h)\| \\ &\leq \|\rho_n\| + \big(1 + \mathrm{L}h\big)\|y_n - x_n^h\|.\end{aligned}$$

Dividing each side of the inequality above by $\big(1+\mathrm{L}h\big)^{n+1}$, and using a telescoping sum argument, we then obtain

$$\|y_n - x_n^h\| \leq \big(1+\mathrm{L}h\big)^n\|y_0 - x_0^h\| + \sum_{m=0}^{n-1}\|\rho_m\|\big(1+\mathrm{L}h\big)^{n-m-1}.$$

We conclude using the inequality $\big(1+\mathrm{L}h\big)^n \leq e^{\mathrm{L}nh} \leq e^{\mathrm{L}T}$, for every $n \in \{0,\dots,N\}$.

□

The stability of the Euler schemes is obtained as a straightforward consequence of the previous proposition.

Corollary 3.3. *The explicit Euler scheme* (6) *is unconditionally stable on bounded intervals* $[0,T]$ *(there is no condition on the time-step size* $h > 0$*).*

The implicit Euler scheme (7) *is stable, on bounded intervals* $[0,T]$*, under the condition* $h < h^\star < \mathrm{L}_F^{-1}$.

Proof.

- Explicit Euler scheme: we apply the previous Proposition, with $\Psi_h(x) = F(x)$, for every $x \in \mathbb{R}^d$, and $h > 0$.
- Implicit Euler scheme: the map Φ_h is such that $y = \Phi_h(x) = x + hF(y)$; thus $\Psi_h(x) = F(y(x))$ where $y(x)$ is the unique solution of the equation $y = x + hF(y)$, under the condition $h\mathrm{L}_F < 1$.

  Then $\|\Psi_h(x_2) - \Psi_h(x_1)\| \leq \frac{\mathrm{L}_F}{1-h\mathrm{L}_F}\|x_2 - x_1\| \leq \frac{\mathrm{L}_F}{1-h^\star\mathrm{L}_F}\|x_2 - x_1\|$.

□

Finally, note that the question of the stability of the scheme (10) is completely independent of the differential equation (1).

In a heuristic way, a stable integrator only has moderate oscillations on bounded intervals; moreover, small changes in the initial condition or in the vector field doe not have a dramatic influence on the numerical solution on bounded intervals.

As we will see below, the inequality (12) is related to a discretized version of the Gronwall's Lemma, which was the main tool to prove similar stability properties for the flow.

3.3 Consistency

Contrary to the notion of stability which was not linked with the differential equation (1), we now introduce another important property of numerical methods which explicitly depends on the differential equation.

Definition 3.4. *Let $T \in (0, +\infty)$, $x_0 \in \mathbb{R}^d$, and $x : [0, T] \to \mathbb{R}^d$ denote the unique solution of (2).*

The **consistency error** *at time $t_n n$ associated with the numerical integrator (10), with time-step size h, is defined by*

$$e_n^h = x(t_{n+1}) - \Phi_h\big(x(t_n)\big) = x(t_{n+1}) - x(t_n) - h\Psi_h\big(x(t_n)\big). \tag{13}$$

The scheme is **consistent with** (1) *if*

$$\sum_{n=0}^{N-1} \|e_n^h\| \underset{h\to 0}{\to} 0. \tag{14}$$

Let $p \in \mathbb{N}^\star$. The scheme is **consistent at order p with** (1)*, if there exists $C \in (0, +\infty)$ and $h^\star > 0$ such that for every $h \in (0, h^\star)$*

$$\sup_{0\le n\le N} \|e_n^h\| \le Ch^{p+1}. \tag{15}$$

First, note that the consistency error e_n^h, defined by (13), depends on the exact solution, not on the numerical solution.

Second, note that the power in the right-hand side of (15) is $p+1$: indeed, the consistency error is defined as a local error, *i.e.* on an interval (t_n, t_{n+1}). When dealing with the error on $[0, T]$, we will consider a global error, which consists on the accumulation of $N = \frac{T}{h}$ local errors: the global error will thus be of order p when the local error is of order $p + 1$.

The following necessary and sufficient conditions for consistency hold true.

Proposition 3.5. *The integrator Φ_h given by (10) is consistent with (1) if and only if $\Psi_0(x) = F(x)$ for every $x \in \mathbb{R}^d$.*

More generally, the integrator Φ_h given by (10) is consistent with (1) at order $p \in \mathbb{N}^\star$ if, assuming that F is of class $\mathcal{C}^p$, one has

$$\frac{\partial^k \Psi_h(x)}{\partial h^k}\Big|_{h=0} = \frac{1}{k+1} F^{[k]}(x) \quad , \quad k \in \{0, \dots, p-1\}, \tag{16}$$

where $\Phi_h(x) = x + h\Psi_h(x)$, with $h \mapsto \Psi_h(x)$ of class $\mathcal{C}^p$, and $F^{[0]}, \dots, F^{[p-1]}$ are functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ defined recursively by

$$\begin{aligned} F^{[0]}(x) &= F(x) \\ F^{[k+1]}(x) &= DF^{[k]}(x).F(x). \end{aligned} \tag{17}$$

The equations (16) are called **order conditions**: indeed the largest p such that theyr are satisfied determines the order of convergence of the consistency error (and of the global error as we will see below).

We only give a sketch of proof of this result. It is based on the use of Taylor expansions.

Proof. We assume that F is of class $\mathcal{C}^p$, with $p \in \mathbb{N}^\star$, and that F and its derivatives of any order are bounded.

Then, for every $x \in \mathbb{R}^d$, it is straightforward to prove recursively that $t \mapsto \phi_t(x)$ is of class $\mathcal{C}^{p+1}$, and that for $k \in \{0, \dots, p\}$, $\eta^{(k)} : t \mapsto \frac{\partial^k \phi_t(x)}{\partial t^k}$ is solution of

$$\eta^{(k)}(t) = F^{[k]}\big(x(t)\big). \tag{18}$$

The proof then follows by comparison of Taylor expansions of $x(t_{n+1}) - x(t_n)$, and of $\Psi^h(x(t_n))$. Indeed, we get

$$x(t_{n+1}) - x(t_n) = \sum_{k=1}^{p} \frac{h^k}{k!} F^{[k-1]}\big(x(t_n)\big) + \mathrm{O}(h^{p+1}),$$
$$h\Psi_h\big(x(t_n)\big) = h \sum_{\ell=0}^{p-1} \frac{h^\ell}{\ell!} \frac{\partial^\ell \Psi_h\big(x(t_n)\big)}{\partial h^\ell}\Big|_{h=0} + \mathrm{O}(h^{p+1}).$$

Thus (with the change of index $k = \ell + 1$)

$$e_n^h = h \sum_{\ell=0}^{p-1} \frac{h^\ell}{\ell!} \Big(\frac{F^{[\ell]}\big(x(t_n)\big)}{\ell+1} - \frac{\partial^\ell \Psi_h\big(x(t_n)\big)}{\partial h^\ell}\Big|_{h=0} \Big) + \mathrm{O}(h^{p+1}).$$

First, note that (using the case $p = 1$), a Riemann sum argument yields

$$\sum_{n=0}^{N-1} \|e_n^h\| = \int_0^T \|F\big(x(t)\big) - \Psi_0\big(x(t)\big)\| dt + \mathrm{o}(1).$$

Then (14) is satisfied if and only if $\int_0^T \|F\big(x(t)\big) - \Psi_0\big(x(t)\big)\| dt$, *i.e.* $F\big(x(t)\big) = \Psi_0\big(x(t)\big)$ for all $t \in [0, T]$. Evaluating at $t = 0$, this is equivalent to $F(x) = \Psi_0(x)$, for all $x \in \mathbb{R}^d$.

Second, we easily see that the condition (16) is sufficient for (15). It is also necessary, evaluating at $n = 0$, for every initial condition $x \in \mathbb{R}^d$. □

We can now study the consistency order of the Euler schemes.

Corollary 3.6. *The explicit and the implicit Euler schemes, given by* (6) *and* (7) *respectively, are consistent with order* 1*.*

More generally, the θ-scheme is consistent with order 1 *if $\theta \neq 1/2$, and it is consistent with order* 2 *if $\theta = 1/2$.*

Proof.

- Explicit Euler scheme: we have $\Psi_h(x) = F(x)$ for every $h \geq 0$ and $x \in \mathbb{R}^d$. Thus consistency with order 1 holds true. Since $\frac{\partial \Psi_h(x)}{\partial h} = 0 \neq F^{[1]}(x)$ in general, the scheme is not consistent with order 2.
- Implicit Euler scheme: it is also easy to check that $\Psi_0(x) = F(x)$: indeed, if $y^h(x)$ is the unique solution of $y = x + hF(y)$, then $y^h(x) \underset{h\to 0}{\to} x$, and $\Psi_h(x) = \frac{x+hF(y^h(x))-x}{h} = F(y^h(x)) \underset{h\to 0}{\to} F(x)$. This proves consistency with order 1. However, one checks that $\frac{\partial \Psi_h(x)}{\partial h}\Big|_{h=0} = F^{[1]}(x) \neq \frac{1}{2}F^{[1]}(x)$: the scheme is not consistent with order 2.
- The θ-scheme, with $\theta \neq 1/2$: left as an exercice.
- Crank-Nicolson scheme ($\theta = 1/2$): one checks that (16) is satisfied with $p = 2$.

□

3.4 Convergence

We are now in position to prove the main theoretical result on the convergence of numerical methods for ODEs.

Definition 3.7. *Let $T \in (0, +\infty)$, $x_0 \in \mathbb{R}^d$, and $x : [0, T] \to \mathbb{R}^d$ the unique solution of (2).*

The **global error** *associated with the numerical integrator (10), with time-step size h, is*

$$\mathcal{E}(h) = \sup_{0\leq n\leq N} \|x(t_n) - x_n^h\|. \tag{19}$$

The scheme is said to be **convergent** *if the global error goes to 0 when $h \to 0$: $\mathcal{E}(h) \underset{h\to 0}{\to} 0$.*

The scheme is said to be **convergent with order** *p if there exists $C \in (0, +\infty)$ and $h^\star > 0$ such that $\big|\mathcal{E}(h)\big| \leq Ch^p$ for every $h \in (0, h^\star)$.*

Theorem 3.8. *A scheme which is stable and consistent is convergent.*

A scheme which is stable and consistent with order p is convergent with order p.

The proof of this result is similar to the second proof of Theorem 2.1.

Proof. Observe that the sequence $(y_n)_{0\leq n\leq N} = \big(x(t_n)\big)_{0\leq n\leq N}$ satisfies

$$y_{n+1} = x(t_{n+1}) = e_n^h + \Phi_h\big(x(t_n)\big) = e_n^h + \Phi_h(y_n),$$

thanks to the definition (13) of the consistency error. Thus $(y_n)_{0\leq n\leq N}$ satisfies (11) with $\rho_n = e_n^h$. Moreover, $x(0) = x_0 = x_0^h$.

Since the scheme is assumed to be stable, we thus have (12), which gives

$$\sup_{0\leq n\leq N} \|x(t_n) - x_n^h\| \leq C \sum_{n=0}^{N-1} \|e_n^h\|.$$

The conclusion then follows using (14) and (15).

□

3.5 Long-time behavior: A-stability

We have seen that both the explicit and the implicit Euler scheme are stable on bounded intervals (provided that the implicit scheme is well-defined).

However, we have seen that in the case of a linear ODE $\cdot x = -\lambda x$, with $\lambda > 0$, using the explicit scheme when time increases requires a condition on the time-step h. The implicit scheme in this case does not require such a condition.

The associated notion of stability is called A-stability.

In this section, we only consider linear scalar ODEs $\cdot x = kx$, and approximations of the type $x_{n+1} = \Phi(hk)x_n$, where $\Phi : \mathbb{C} \to \mathbb{C}$ is the stability function.

Here are a few examples:

- explicit Euler scheme: $\Phi(z) = 1 + z$;
- implicit Euler scheme: $\Phi(z) = \frac{1}{1-z}$;
- θ-scheme: $\Phi(z) = \frac{1+(1-\theta)z}{1-\theta z}$.

Definition 3.9. *The absolute stability region of the numerical method is*

$$\mathcal{S}(\Phi) = \{z \in \mathbb{C} \ ; \ |\Phi(z)| < 1\}.$$

The method is called A-stable if

$$\mathcal{S}(\Phi) \supset \{z \in \mathbb{C} \ ; \ \mathrm{Re}(z) < 0\}.$$

Like the notion of stability, the notion of A-stability only refers to the numerical method, and is independent of the ODE it is applied to.

Assume that the method is A-stable, and consider the ODE $\cdot x = -\lambda x$, with $\lambda > 0$. Then $-\lambda h \in \mathcal{S}(\Phi)$ for every $h > 0$, which means that $x_n \underset{n\to+\infty}{\to} 0$ (for every initial condition of the numerical scheme).

Thus A-stability is the appropriate notion of stability to deal with the asymptotic behavior of numerical solutions of linear ODEs.

We conclude this section with the study of the A-stability of the θ-scheme.

Proposition 3.10. *The θ-scheme is A-stable if and only if $\theta \geq 1/2$.*

Note in particular that the implicit Euler scheme and the Crank-Nicolson scheme are both A-stable, whereas the explicit Euler scheme is not A-stable.

4 Some constructions of higher-order methods

We work in the framework of Theorem 3.8: we assume that a numerical integrator (10) is given, and the integrator is assumed to be stable and consistent, and thus convergent.

The aim of this section is to present several approaches to define higher-order numerical methods: we wish to increase the order of convergence p of the global error to 0. This amounts to define methods such that the consistency error is of order $p+1$. Thanks to Proposition 3.5, one needs to enforce the order conditions (16).

The list of the constructions we propose is not exhaustive, and we limit ourselves in practice to constructions of order 2 methods from order 1 methods, in order to keep the presentation simple.

4.1 Taylor expansions

Let $p \in \mathbb{N}$.

For every $x \in \mathbb{R}^d$ and $h > 0$, set $\Phi_h^{(p)}(x) = x + h\Psi_h^{(p)}(x)$,

$$\Psi_h^{(p)}(x) = \sum_{k=0}^{p-1} \frac{h^k}{(k+1)!} F^{[k]}(x),$$

with the functions $F^{[k]}$ given by (17). Since $F^{[k]}(x) = \left.\frac{d^{k+1}\phi_t(x)}{dt^{k+1}}\right|_{t=0}$, $\Phi_h^{(p)}(x)$ is the Taylor polynomial of order p at time 0 of the flow, starting at x.

For instance, when $p = 1$, $\Psi_h^{(1)}(x) = F^{[0]}(x) = F(x)$: we get the explicit Euler scheme.

When $p = 2$, we get

$$\Phi_h^{(2)}(x) = x + hF(x) + \frac{h^2}{2} DF(x).F(x).$$

Clearly, by construction $\left.\frac{\partial^k \Psi_h^{(p)}(x)}{\partial h^k}\right|_{h=0} = \frac{1}{k+1} F^{[k]}(x)$, for every $0 \leq k \leq p-1$: the order conditions (16) are satisfied, and the method is consistent of order p. Note that it is not consistent of order $p+1$ in general.

We are thus able to provide, using truncated Taylor expansions at an arbitrary order, numerical integrators which are consistent, with an arbitrary order. Nonetheless, the concrete implementation of such schemes requires to compute $F^{[k]}(x)$ for every $0 \leq k \leq p-1$: this condition can be a severe limitation in practice, especially when dimension d increases. This aspect will be seen below, when we discuss questions of cost of the integrators.

Obviously, the constructions above are possible only if the vector field F is sufficiently regular, since derivatives of order $0, \ldots, p-1$ are required.

4.2 Quadrature methods for integrals

We now make a connection between the numerical approximations of solutions of ODEs and of integrals. Numerical methods for the approximations of integrals are referred to as **quadrature methods** in the literature.

Observe first that x is solution of the ODE (2), with initial condition x_0, if and only it is solution of the integral equation

$$x(t) = x_0 + \int_0^t F\big(x(s)\big)ds \quad , \quad t \in [0,T]. \tag{20}$$

This equivalent formulation is in fact the starting point of the Picard iteration/Banach fixed point theorem approach to the Cauchy-Lipschitz Theorem.

Now let $h > 0$ denote a time-step size. Our aim is to provide integrators for ODEs (1) based on quadrature methods for integrals applied to the equation (20). As usual, we provide approximations x_n of x_n^h at times $t_n = nh$, with $0 \leq n \leq N$ and $T = Nh$.

Thanks to (20),

$$x(t_{n+1}) = x(t_n + h) = x(t_n) + \int_{t_n}^{t_n+h} F\big(x(s)\big)ds.$$

Using the approximation $hF\big(x(t_n)\big)$ (resp. $hF\big(x(t_{n+1})\big)$) of the integral $\int_{t_n}^{t_n+h} F\big(x(s)\big)ds$ then provides the numerical scheme $x_{n+1} = x_n + hF(x_n)$ (resp. $x_{n+1} = x_n + hF(x_{n+1})$): we recover the explicit (resp. implicit) Euler scheme.

The associated quadrature methods are called the **left** (resp. the **right**) **point rule**: more generally if $\varphi : [0,T] \to \mathbb{R}^d$ is a continuous function, we may approximate the integral

$$I(\varphi) = \int_0^T \varphi(s)ds$$

with the **Riemann sum** $I_N^{\text{left}} = h\sum_{n=0}^{N-1} \varphi(nh)$ (resp. $I_N^{\text{right}} = h\sum_{n=1}^{N} \varphi(nh)$).

These two methods are known to be of order 1: if φ is of class $\mathcal{C}^1$, there exists $C(\varphi) \in (0,+\infty)$ such that for every $N \in \mathbb{N}$

$$\big|I(\varphi) - I_N^{\text{left}}(\varphi)\big| \leq \frac{C(\varphi)}{N} \quad , \quad \big|I(\varphi) - I_N^{\text{left}}(\varphi)\big| \leq \frac{C(\varphi)}{N},$$

with $h = T/N$. The order of convergence 1 is optimal, as can be seen by taking $\varphi(s) = s$. In fact, finding the order 1 both for the quadrature method and for the associated numerical integrator is not suprising, due to (20): the two approximation problems are deeply linked.

It is possible to construct higher-order quadrature methods: for instance, the **trapezoidal rule**

$$I_N^{\text{trap}}(\varphi) = h\sum_{n=0}^{N-1} \frac{\varphi(nh) + \varphi((n+1)h)}{2}$$

is a quadrature method of order 2: if φ is of class $\mathcal{C}^2$, there exists $C(\varphi) \in (0,+\infty)$ such that for every $N \in \mathbb{N}$

$$\big|I(\varphi) - I_N^{\text{trap}}(\varphi)\big| \leq \frac{C(\varphi)}{N^2}.$$

Note that we have increased the regularity requirement on the integrand φ to obtain an improved order of convergence. If φ is less regular, only order 1 can be achieved in general. The order 2 above is optimal, as can be seen by choosing $\varphi(s) = s^2$; note also that if $\varphi(s) = \alpha s + \beta$, then $I_N(\varphi) = I(\varphi)$, which means that the trapezoidal rule quadrature method is exact for polynomials of degree less than 1.

More generally, a quadrature method which is exact for polynomials of degree less than p is of order $p + 1$. The proof is out of the scope of these lecture notes.

Using the trapezoidal rule yields the following numerical scheme:

$$x_{n+1} = x_n + \frac{h}{2}F(x_n) + \frac{h}{2}F(x_{n+1}),$$

which is the Crank-Nicolson scheme (9). This link with the trapezoidal rule partly explains why the Crank-Nicolson scheme is popular and efficient.

4.3 A few recipes

4.3.1 Removing derivatives

The integrator $\Phi_h^{(2)}$ obtained above using the second-order Taylor expansion may not always be practically implementable, due to the presence of the derivative $DF(x)$.

It is however possible to design another, simpler, method with order 2, by setting

$$\Phi_h(x) = x + hF\big(x + \frac{h}{2}F(x)\big). \tag{21}$$

Indeed, $F\big(x+\frac{h}{2}F(x)\big) = F(x)+\frac{h}{2}DF(x).F(x)+\mathrm{O}(h^2)$, which means that $\Phi_h(x) = \Psi_h^{(2)}(x)+\mathrm{O}(h^3)$. One can also directly check the order conditions.

The integrator (21) is called the **explicit midpoint rule**, and is of order 2. It only involves computations of F. Note also that the integrator is explicit; however two computations of F are required, instead of one for the explicit Euler scheme. Moreover, we can write the scheme as follows:

$$\begin{aligned} x_{n+1/2} &= x_n + \frac{h}{2}F(x_n), \\ x_{n+1} &= x_n + hF(x_{n+1/2}). \end{aligned}$$

The notation $x_{n+1/2}$ is used since this quantity is an approximation of $x(t_n + \frac{h}{2})$.

This scheme is an example of a **predictor-corrector** scheme. Indeed, the first step corresponds to a prediction $x_{n+1/2}$ of the update, and the second step to a correction, giving the true update x_{n+1}. The prediction is an approximation of the solution at the midpoint $t_{n+1/2} = nh + \frac{h}{2}$ of the interval $[t_n, t_{n+1}]$.

A direct way of understanding the construction of the explicit midpoint method, in terms of Taylor expansions, is the following formula:

$$x(t + h) = x(t) + hx'(t + \frac{h}{2}) + \mathrm{O}(h^3).$$

To conclude this section, we also introduce the **implicit midpoint rule**:

$$x_{n+1} = x_n + hF(\frac{x_n + x_{n+1}}{2}). \tag{22}$$

4.3.2 Removing implicitness

We have seen that among the family of θ-schemes, the Crank-Nicolson ($\theta = 1/2$) is the only one to be of order 2. However, the practical implementation may lead to disappointing results since solving an implicit nonlinear equation is required at each iteration.

Using again a predictor-corrector approach leads to define the following explicit integrator of order 2:

$$\begin{aligned} y_{n+1} &= x_n + hF(x_n), \\ x_{n+1} &= x_n + \frac{h}{2}\big(F(x_n) + F(y_{n+1})\big), \end{aligned} \tag{23}$$

called the **Heun's method**. The corrector step is similar to the Crank-Nicolson scheme (9), using the predicted value computed using an explicit Euler approximation.

4.4 Runge-Kutta methods

For completeness, we present the popular **Runge-Kutta methods**. They are based on quadrature formulas, and they encompass several of the integrators defined above.

A s-stage Runge Kutta method for a (non-autonomous) ODE $x'(t = F\big(t, x(t)\big)$ is given by

$$x_{n+1} = x_n + h_n \sum_{i=1}^{s} b_i k_i, \quad n \geq 0$$

where the k_i, $1 \leq i \leq s$ are given by

$$k_i = f\big(t_n + c_i h_n, x_n + h_n \sum_{j=1}^{s} a_{i,j} k_j\big).$$

We have used the notation $h_n = t_{n+1} - t_n$ (the time-step may also be non-constant).

The integrator depends on two vectors $\big(b_i\big)_{1\leq i\leq s}$ and $\big(c_i\big)_{1\leq i\leq s}$, such that $b_i, c_i \in [0, 1]$; and on a matrix $\big(a_{i,j}\big)_{1\leq i,j\leq s}$. The standard notation for Runge-Kutta methods is given by the associated **Butcher's tableau**:

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \dots & a_{1,s} \\ c_2 & a_{2,1} & \dots & a_{2,s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s,1} & \dots & a_{s,s} \\ \hline & b_1 & \cdots & b_s \end{array}$$

Note that the method is explicit if $a_{i,j} = 0$ when $j \geq i$: indeed, one can compute successively and explicitly $k_1, k_2, \ldots, k_s$. In the case of an autonomous equation, the parameters c_i do not play a role in the computation.

It is possible to derive general order conditions for consistency of order p, depending only on the b_i, c_i and $a_{i,j}$. We refrain from giving them and their proof. We only stress out the consistency condition is $\sum_{i=1}^{s} b_i = 1$. It is also possible to derive a stability condition.

To conclude this section, note that the explicit and implicit Euler integrators are examples of Runge-Kutta methods. The corresponding Butcher's tableaux are:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad\qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Nevertheless, the Crank-Nicolson integrator is not a Runge-Kutta method.

Note also that the explicit midpoint (predictor-corrector) and Heun integrators are 2 stages Runge-Kutta methods, with Butcher's tableaux

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array} \qquad\qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}$$

A popular Runge-Kutta method is given by the following Butcher's tableau:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

The method is explicit and has order 4. It is called the RK4 method.

4.5 Order of convergence and cost of a method

Exhibiting an high-order of convergence (typically, $p \geq 2$) is not the only important aspect of a numerical integrator. Indeed, the total cost of the computation may also depend a lot on the integrator map.

Indeed, denote by $c_{\text{it}}(h)$ the cost of the computation of one iteration of the scheme, *i.e.* of computing $x_{n+1}^h = \Phi^h\big(x_n^h\big)$ assuming that x_n^h is known from the previous iteration. The total cost of the numerical method, on an interval $[0, T]$, is then equal to $\text{Cost}(h) = N c_{\text{it}}(h) = T\frac{c_{\text{it}}(h)}{h}$, since N successive iterations of the integrator are computed.

For instance, one iteration of the explicit Euler scheme requires one evaluation of F, one multiplication by h and one addition; typically, only the evaluation of F may be computationally expensive: thus $c_{\text{it}}(h) = c(F)$, where $c(F)$ is the cost of one evaluation of F.

The cost of one iteration of the implicit Euler scheme is typically larger: one uses an iteration scheme (such as Newton's method), where each iteration requires one evaluation of F. Thus $c_{\text{it}}(h) = Mc(F)$, where M is the number of iterations to give an accurate approximation of the solution of the nonlinear implicit equation.

Finally, methods based on higher-order Taylor expansions may have a prohibitve cost due to the necessity of evaluating the derivatives of F.

To simplify the presentation, we consider that the cost of one iteration of the integrator is bounded from above, uniformly on h: $c_{\mathrm{it}}(h) \leq c_{\mathrm{it}}$.

Let us consider a convergent scheme of order p, and let $\epsilon > 0$. Then the global error is controlled by ϵ, *i.e.* $\mathcal{E}(h) \leq \epsilon$, if the time-step h is chosen such that $Ch^p \leq \epsilon$: this gives the condition $h \leq \left(\epsilon/C\right)^{1/p}$.

The minimal cost to have a global error less than ϵ is thus equal to

$$C^{1/p}\epsilon^{-1/p}c_{\mathrm{it}}.$$

In this expression, we thus see the influence on the cost of:

- the order of convergence p: if p increases, the cost decreases;
- the constant C, such that $|\mathcal{E}(h)| \leq Ch^p$: if p increases, the cost increases;
- the cost of one iteration c_{it}.

Note that the restriction $h < h^\star$ may also cause trouble, if $h^\star$ has to be chosen very small, due for instance to stability issues. Thus the stability also plays a role on the cost of an integrator.

In some cases, the parameters C, c_{it} and $h^\star$ may have a huge influence on the cost of the method. Due to these considerations, a scheme with high-order of convergence may be inefficient in practice.

Other considerations also need to be taken into account for the choice of an integrator: stability and preservation of qualitative properties are often more important than trying to increase the order of convergence.

5 Splitting and composition methods

The aim of this section is to present other procedures which lead to the construction of new methods from basic ones. We also discuss elements of their qualitative behavior.

To illustrate the abstract constructions of this section, consider the following example of ODE on $\mathbb{R}^d$:

$$\dot{x} = -\frac{1}{\epsilon}Ax + F(x), \tag{24}$$

where $\epsilon > 0$, A is a positive self-adjoint matrix, and F is a globally Lipschitz function.

Such a system occurs for instance as the discretization of semilinear parabolic PDEs (such as the reaction-diffusion equations). In the regime $\epsilon \to 0$, the ODE contains a fast and a slow parts, and this is a typically difficult situation for numerical methods.

On the one hand, assume first that $F = 0$. We have seen that an explicit Euler scheme then is not unconditionally stable on $[0, +\infty)$, contrary to the implicit Euler and the Crank-Nicolson schemes. The stability condition on the time-step size h of the explicit Euler scheme is $\frac{h\lambda_{\max}(A)}{\epsilon} < 1$, where $\lambda_{\max}(A)$ is the largest eigenvalue of $-A$; when ϵ is small, this is a severe restriction.

On the other hand, when F is non zero, using an implicit scheme may be too costly, in the case that evaluations of F are expensive, or if the iteration method used to solve the nonlinear equation at each step of the scheme does not converge fast enough.

We thus need to deal with the two constraints below:

- we need an implicit scheme to treat the linear part;
- we can only use an explicit scheme to treat the nonlinear part.

One possibility is to consider the following **semi-implicit** scheme:

$$x_{n+1} = x_n - \frac{h}{\epsilon}Ax_{n+1} + hF(x_n),$$

which can be rewritten in a well-posed, explicit, form $x_{n+1} = \big(I + \frac{h}{\epsilon}A\big)^{-1}\big(x_n + hF(x_n)\big)$.

This is a nice solution; we also present other more general constructions below, which possess nice qualitative properties and will be very useful to construct and study integrators for Hamiltonian dynamics.

5.1 Splitting methods

Splitting methods are designed to treat ODEs $\dot{x} = F(x)$, where the vector field can be decomposed as a sum $F = F^1 + F^2$, where the vector fields F^1 and F^2 satisfy the following properties:

- F^1 and F^2 are globally Lipschitz continuous (to simplify the presentation),
- if F is of class $\mathcal{C}^k$, then F^1, F^2 are also of class $\mathcal{C}^k$,

- the flow ϕ^i of the ODE $\dot{x} = F^i(x)$ is exactly known, for $i = 1$ and $i = 2$.

In the example (24), one has $F^1(x) = -\frac{1}{\epsilon}Ax$, with exact flow $\phi_t^1(x) = \exp\big(-\frac{t}{\epsilon}A\big)x$, and $F^2(x) = F(x)$ – for which the exact flow is not known in general.

Obviously, the most restrictive assumption in practice is the third one. Composition methods below are a way to construct integrators, in the spirit of splitting methods, without that assumption being satisfied.

The basic splitting methods are given by the following integrators: given $h > 0$, define

$$\Phi_h = \phi_h^2 \circ \phi_h^1 \quad , \quad \Phi_h = \phi_h^1 \circ \phi_h^2. \tag{25}$$

Those integrators consist of one step of the exact flow associated with one of the vector fields, followed by one step of the exact flow associated with the other vector field. Interpreting the vector fields F, F^1 and F^2 as forces, this means that to update the position from time nh to $(n+1)h$, the contributions of the forces F^1 and F^2 are treated successively.

It is easy to check the consistency at order 1 of the splitting integrators introduced above. In general, these integrators are not consistent at order 2. Indeed, consider the case of a linear ODE

$$\dot{x} = Bx + Cx$$

and set $F^1(x) = Bx$, $F^2(x) = Cx$. The exact flows are given by $\phi_t^1(x) = e^{tB}x = \sum_{k=0}^{+\infty} \frac{t^k B^k}{k!}x$, and $\phi_t^2(x) = e^{tC}x$. Consider the splitting scheme given by $\Phi_h(x) = e^{hC}e^{hB}x$.

In that example, the consistency error is thus equal to

$$\begin{aligned}
e_n^h &= x(t_{n+1}) - x(t_n) - e^{hC}e^{hB}x(t_n) \\
&= \big(e^{h(B+C)} - e^{hB}e^{hC}\big)e^{t_n(B+C)}x_0 \\
&= \big(I + hB + hC + \frac{h^2}{2}(B+C)^2 + \mathrm{O}(h^3)\big)e^{t_n(B+C)}x_0 \\
&\quad - \big(I + hB + hC + hCB + \frac{h^2}{2}(B^2+C^2) + \mathrm{O}(h^3)\big)e^{t_n(B+C)}x_0 \\
&= \big(\frac{h^2}{2}[B,C] + \mathrm{O}(h^3)\big)e^{t_n(B+C)}x_0,
\end{aligned}$$

with the **commutator** $[B,C] = BC - CB$.

The local consistency error e_n^h is thus in general only of order $2 = 1 + 1$, except if $[B,C] = 0$. This condition is equivalent to $BC = CB$, *i.e.* B and C commute: in that case, it is well-known that $e^{tB}e^{tC} = e^{t(B+C)}$, so the splitting method is exact. On the contrary, when B and C do not commute, the splitting integrator is of order 1.

In the general, nonlinear case, a similar formula is satisfied:

$$\Phi_h(x) - \phi_h^2 \circ \phi_h^1(x) = \frac{h^2}{2}[f,g](x) + \mathrm{O}(h^3),$$

where $[f,g](x) = \frac{\partial f}{\partial x}g(x) - \frac{\partial g}{\partial x}f(x)$.

The integrators (25) are called **Lie-Trotter splitting** methods. It is possible to construct integrators of order 2, using the so-called **Strang splitting** methods:

$$\Phi_h = \phi^2_{h/2} \circ \phi^1_h \circ \phi^2_{h/2} \quad , \quad \Phi_h = \phi^1_{h/2} \circ \phi^2_h \circ \phi^1_{h/2}. \tag{26}$$

The better properties of the Strang splitting are due to some symmetry properties. To explain these properties, we introduce below the necessary tools in a general context.

As an exercise, one can check that the Strang splitting methods are of order 2 in the linear case.

5.2 The adjoint of an integrator

Consider the ODE (1). Even if we have only considered positive times t in the proof of the Cauchy-Lipschitz theorem, the flow ϕ is in fact defined on $\mathbb{R}$, and the semi-group property $\phi_{t+s} = \phi_t \circ \phi_s$, see (3) is satisfied for every $t, s \in \mathbb{R}$.

In particular, for every $t \in \mathbb{R}$, the mapping ϕ_t is invertible and $\phi_t^{-1} = \phi_{-t}$. This property can be interpreted as a time-reversibility of the ODE: for $T > 0$, one has $x(0) = \phi_{-T}(x(T))$, and ϕ_{-t} is the flow associated with the ODE $\dot{x} = -F(x)$, where the sign of the vector field is changed.

The time-reversibility property is not satisfied in general by numerical integrators. It is convenient to introduce the notion of the adjoint of an integrator. Below, we assume that all quantities are well-defined.

Definition 5.1. *The* **adjoint** *of an integrator Φ_h is defined by*

$$\Phi^\star_h = \Phi^{-1}_{-h}. \tag{27}$$

If $\Phi^\star_h = \Phi_h$, the integrator is called **self-adjoint** *(or symmetric).*

Note that $\left(\Phi^\star_h\right)^\star = \Phi_h$. Moreover, $\left(\Phi_h \circ \Psi_h\right)^\star = \Psi^\star_h \Phi^\star_h$.

Here are a few examples:

- if $\Phi_h = \phi_h$ is the exact flow at time h, then Φ_h is self-adjoint.
- the adjoint of the explicit Euler scheme is the implicit Euler scheme.
- the implicit midpoint and the Crank-Nicolson schemes are self-adjoint.
- Strang splitting integrators are self-adjoint.

Given an integrator Φ_h, of order 1, introduce

$$\Psi_h = \Phi_{h/2} \Phi^\star_{h/2}.$$

By construction, Ψ_h is self-adjoint. Moreover, one checks that if Φ_h is of order 1, then Ψ_h is of order 2. More generally, a self-adjoint integrator is necessarily of even order: if it is consistent at order 1, it is thus of order 2.

Two interesting examples are given below:

- If Φ_h is the implicit Euler scheme, then Ψ_h is the Crank-Nicolson scheme.
- If $\Phi_h = \phi_h^2 \circ \phi_h^1$ is a Lie-Trotter splitting scheme, then $\Psi_h = \phi_{h/2}^2 \circ \phi_h^1 \circ \phi_{h/2}^2$ is the associated Strang splitting scheme.

The notion of symmetry of an integrator is a way to study how one qualitative property of the exact flows, namely the time-reversibility, is preserved or not by numerical integrators. We will see other examples of such notions in the following, especially in the case of Hamiltonian dynamics.

5.3 Composition methods

More generally, considering several integrators $\Phi_h^1, \dots, \Phi_h^k$, integrators of the form

$$\Psi_h = \Phi_{\alpha_1 h}^1 \circ \dots \circ \Phi_{\alpha_k h}^k,$$

with $\alpha_1, \dots, \alpha_k \in [0,1]$, are called **composition methods**.

For instance, splitting methods are composition methods based on using the exact flows associated with each of the vector fields. In the case of the Lie-Trotter splitting, one has $\alpha_1 = \alpha_2 = 1$; in the case of the Strang splitting, one has $\alpha_1 = \alpha_3 = 1/2$ and $\alpha_2 = 1$.

Replacing the exact flows in a splitting method with appropriate numerical integrators thus gives a composition method. For instance, one can construct an integrator of the form $\Phi_{h/2}^2 \circ \Phi_h^1 \circ \Phi_{h/2}^2$, based on the Strang splitting. However, in general, note that this integrator is not self-adjoint anymore.

For instance, for the ODE (24), one can use an implicit Euler or Crank-Nicolson scheme (which avoids computing the exponential of the matrix) to treat the linear part, and the explicit Euler scheme to treat the nonlinear part.

Finally, the integrator $\Psi_h = \Phi_{h/2} \Phi_{h/2}^\star$, using the adjoint, is also a composition method.

5.4 Conjugate methods, effective order, processing

To conclude this section, we introduce further notions, which are best illustrated again with the example of splitting methods.

Consider $\Phi_h^{\mathrm{Lie}} = \phi_h^2 \circ \phi_h^1$ and $\Phi_h^{\mathrm{Strang}} = \phi_{h/2}^2 \circ \phi_h^1 \circ \phi_{h/2}^2$, respectively the Lie-Trotter and the Strang spitting schemes (using the exact flows). Then observe that $\Phi_h^{\mathrm{Strang}} = \phi_{h/2}^2 \Phi_h^{\mathrm{Lie}} \phi_{-h/2}^2 = \phi_{h/2}^2 \Phi_h^{\mathrm{Lie}} \big(\phi_{h/2}^2\big)^{-1}$.

If the integrators are applied N times, we also have $\big(\Phi_h^{\mathrm{Strang}}\big)^N = \phi_{h/2}^2 \big(\Phi_h^{\mathrm{Lie}}\big)^N \big(\phi_{h/2}^2\big)^{-1}$. Applying the Lie-Trotter scheme $x_{n+1} = \Phi_h^{\mathrm{Lie}}(x_n)$ or the Strang scheme $y_{n+1} = \Phi_h^{\mathrm{Strang}}(y_n)$ is thus equivalent, when $y_n = \phi_{h/2}^2(x_n)$; the mapping $\phi_{h/2}^2$ and its inverse $\phi_{-h/2}^2$ play the role of a change of variables. However, we have seen that the orders of convergence of these splitting schemes are different, so the observation above is non trivial.

More generally, we have the following notion of conjugacy and of effective order for general integrators.

Definition 5.2. *Let Φ_h and Ψ_h denote two integrators.*

They are **conjugated** *if there exists a bijective mapping χ_h such that $\Psi_h = \chi_h \circ \Phi_h \circ \chi_h^{-1}$.*

If Φ_h is of order p, and is conjugated with Ψ_h of order $q > p$, then we say that Φ_h has **effective order** *q.*

The Lie-Trotter scheme is thus conjugated to the Strang scheme, and has order 1 and effective order 2.

The mapping χ_h is called the **processing** mapping, and plays the role of a change of variables. Indeed, we can write $y = \chi_h(x)$, and consider that the integrator Φ_h acts on the x variable, and that the integrator Ψ_h acts on the y variable.

What we know, is that $y_N = \left(\Psi_h\right)^N(y_0)$ is an approximation of $\phi_T(y_0)$ at order q. To take advantage of this result, in practice if we want to apply only the integrator Φ_h, we need to do the following. The initial conditions satisfy $x_0 = \chi_h^{-1}(y_0)$: we have to **pre-process** the initial condition. We then compute $x_N = \left(\Phi_h\right)^N(x_0)$, iterating the integrator. At the end, we need to compute $y_N = \chi_h(x_N)$: this is a **post-processing**. Note that the mapping χ_h and its inverse are only computed once.

A processing may thus improve the accuracy of the approximation, giving an effective order which is larger than the order of the method. In relation with qualitative properties, a processing may also be helpful: two conjugate methods do not possess the same qualitative behavior in general. For instance, two conjugate methods are not necessarily both self-adjoint; but one of them being self-adjoint shows that up to action of the bijective mapping, the other one possesses in turn an approximate time-reversibility property.

6 Integrators for Hamiltonian dynamics

In this section, we construct and study integrators which are well adapted for Hamiltonian systems,

$$\begin{cases} \dot{q} = \nabla_p H(q,p), \\ \dot{p} = -\nabla_q H(q,p). \end{cases} \tag{28}$$

We first propose a method, in the case of a separable Hamiltonian $H(q,p) = \frac{1}{2}\|p\|^2 + V(q)$ with quadratic kinetic energy, and potential energy function V.

We then provide important qualitative properties of Hamiltonian flows: conservation of H, and area preservation. We also introduce the notion of symplectic mapping.

6.1 The Störmer-Verlet method

Assume that the Hamiltonian function is $H(q,p) = \frac{1}{2}\|p\|^2 + V(q)$; then the ODE (28) is written

$$\dot{q} = p \quad , \quad \dot{p} = -\nabla V(q),$$

and is thus equivalent to the second-order ODE

$$\ddot{q} = -\nabla V(q).$$

One way of discretizing the second-order derivative $\ddot{q}(t_n) = q''(t_n)$ is with a central second order difference quotient:

$$q''(t_n) = \frac{q(t_{n+1}) - 2q(t_n) + q(t_{n-1})}{h^2} + \mathrm{o}(1),$$

using a Taylor's expansion.

This leads to the **Störmer-Verlet scheme** (also called the **leapfrog scheme**):

$$q_{n+1} - 2q_n + q_{n-1} = -h^2 \nabla V(q_n). \tag{29}$$

Geometrically, the interpretation is as follows, for $d = 1$: the parabola on each interval $[q_{n-1}, q_{n+1}]$ which interpolates the values of the scheme at times t_{n-1}, t_n, t_{n+1} has the constant second-order derivative $-V'(q_n)$ – contrary to first-order ODEs (1), approximated by Euler schemes, for which the first-order derivative of a piecewise linear interpolation on (t_n, t_{n+1}) is given by $f(x_n)$ or $f(x_{n+1})$.

Below we propose a derivation of the scheme based on a (discretized) variational formulation of the second-order ODE. We then go back to the formulation (28) and provide two versions of the Störmer-Verlet scheme for general Hamiltonian functions H.

6.1.1 A derivation of the Störmer-Verlet scheme

Introduce the **Lagrangian function** $\mathcal{L}(q,v) = \frac{1}{2}\|v\|^2 - V(q)$, and for any $T \in (0,+\infty)$, and any function $q : [0,T] \to \mathbb{R}$ of class $\mathcal{C}^1$, define the **action**

$$\mathcal{A}_T(q) = \int_0^T \mathcal{L}\big(q(t), q'(t)\big)dt.$$

We use different notations v and p: in fact these variables are conjugated via the so-called Legendre transform:

$$H(q,p) = \sup_{v\in\mathbb{R}} \big(\langle p, v\rangle - \mathcal{L}(q,v)\big).$$

The appropriate terminology is to call v the velocity and p the momentum. Indeed, in the definition of the action $v = q'(t)$ is the velocity at time t of the curve.

Consider initial and terminal conditions q_0 and q_T. Then the unique (under appropriate conditions) minimizer of $\mathcal{A}_T(q)$, satisfying $q(0) = q_0$ and $q(T) = q_T$ is solution of the so-called **Euler-Lagrange equations**

$$\frac{d}{dt}\frac{\partial L}{\partial \dot q}\big(q(t), \dot q(t)\big) = \frac{\partial L}{\partial q}\big(q(t), \dot q(t)\big),$$

which here is equivalent to $\ddot q(t) = \frac{d}{dt}\dot q(t) = -\nabla_q V(q(t))$.

The link between Lagrangian and Hamiltonian formulations, using the Legendre transform, is more general, but this is out of the scope of these lectures.

Consider the following discretized version of the action:

$$\hat A_T = h\sum_{n=0}^{N-1} \mathcal{L}\big(q_n, \frac{q_{n+1}-q_n}{h}\big) = h\sum_{n=0}^{N-1}\Big[\frac{\|q_{n+1}-q_n\|^2}{2h^2} - V(q_n)\Big],$$

with $Nh = T$, using an approximation $v_n = \frac{q_{n+1}-q_n}{h}$ of the velocity.

With given initial and terminal conditions q_0 and q_N respectively, we want to minimize $\hat A_T(q_1, \ldots, q_{N-1})$, as a function of the unknown positions at times $h, \ldots, (N-1)h$. We thus need to solve

$$\frac{\partial \hat A_T}{\partial q_n} = 0 \quad , \quad n \in \{1, \ldots, N-1\}\,;$$

this is equivalent to $\frac{(q_n - q_{n+1}) + (q_n - q_{n-1})}{h^2} - \nabla_q V(q_n) = 0$, which is the definition (29) of the Störmer-Verlet scheme.

6.1.2 Formulations for general Hamiltonian functions

Note that the scheme (29) only deals with the positions q_n. To treat the case of general Hamiltonian functions, it is necessary to incorporate velocities/momenta.

There are two versions, each using an approximation of the value at time $t_{n+1/2}$ of either p or q:

$$\begin{cases} p_{n+1/2} = p_n - \frac{h}{2}\nabla_q H(q_n, p_{n+1/2}) \\ q_{n+1} = q_n + \frac{h}{2}\big(\nabla_p H(q_n, p_{n+1/2}) + \nabla_p H(q_{n+1}, p_{n+1/2})\big) \\ p_{n+1} = p_{n+1/2} - \frac{h}{2}\nabla_q H(q_{n+1}, p_{n+1/2}) \end{cases} \qquad (30)$$

$$\begin{cases} q_{n+1/2} = q_n + \frac{h}{2}\nabla_p H(q_{n+1/2}, p_n) \\ p_{n+1} = p_n - \frac{h}{2}\big(\nabla_q H(q_{n+1/2}, p_n) + \nabla_q H(q_{n+1/2}, p_{n+1})\big) \\ q_{n+1} = q_{n+1/2} + \frac{h}{2}\nabla_p H(q_{n+1/2}, p_n). \end{cases}$$

In each scheme, the first two steps are in general implicit, whereas the third step is explicit. The first step is based on the implicit Euler scheme, the second one on the Crank-Nicolson scheme, and the third one on the explicit Euler scheme. Note also that for Euler schemes the time step size is $h/2$, since we compute approximations of p or q at times $t_n, t_{n+1/2}, t_{n+1}$. The central step uses the time step size h.

As an exercise, one can check that in the case $H(q,p) = \frac{\|p\|^2}{2} + V(q)$, one can eliminate the variables p from the scheme and recover (29).

6.2 Conservation of the Hamiltonian

We recall that H is preserved by the Hamiltonian flow: $H(q(t), p(t)) = H(q(0), p(0))$. This property is extremely important.

For instance, with $d = 1$, consider the case of a double well-potential $V(q) = \frac{q^4}{4} - \frac{q^2}{2}$. Since $V'(q) = q^3 - q = q(q-1)(q+1)$, we see that V admits three stationary points. More precisely, $q = -1$ and $q = +1$ are global minima, whereas $q = 0$ is a local maximum.

If the initial velocity $p_0 = 0$ is equal to 0, and if $q_0 \in \{-1, 0, +1\}$ is equal to one of the stationary points, then for all $t \geq 0$, one has $q_t = q_0$ and $p_t = 0$.

If the initial velocity $p_0 \neq 0$ is not equal to 0, the position q evolves. A natural question is then the following: starting with $p_0 \neq 0$ and $q_0 = -1$, are there times T such that $q_T = -1$? If yes, since there must exists t such that $q_t = 0$ and $p_t \neq 0$, using conservation of H we have

$$\frac{1}{2}p_0^2 + V(-1) = \frac{1}{2}p_t^2 + V(0),$$

and thus $\frac{1}{2}p_0^2 > V(0) - V(-1)$. The initial velocity needs to be sufficiently large to allow for transitions between the two wells. It is not difficult to check that the condition above is sufficient.

It is thus natural to try to construct integrators which conserve the Hamiltonian. However, we will see that the popular methods introduced below preserve other qualitative properties, but do not in general preserve the Hamiltonian.

6.3 Symplectic mappings

Introduce the following matrix: $J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$. Then the Hamiltonian ODE (28) can be rewritten in the equivalent form

$$\dot{z} = J\nabla H(z), \tag{31}$$

with $z = (q,p) \in \mathbb{R}^{2d}$ and $\nabla H(q,p) = \begin{pmatrix} \nabla_q H(q,p) \\ \nabla_p H(q,p) \end{pmatrix}$.

For any $\eta, \xi \in \mathbb{R}^{2d}$, define $\omega(\xi,\eta) = \xi^\star J\eta$.

Definition 6.1. *A real-valued, square matrix $A = \big(a_{i,j}\big)_{1\le i,j\le d}$, is called* **symplectic** *if it satisfies*

$$\omega(A\xi, A\eta) = \omega(\xi,\eta)$$

for every $\xi,\eta \in \mathbb{R}^{2d}$.

Equivalently, A is symplectic if and only if $A^\star JA = J$.

A mapping $\Phi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is symplectic if the Jacobian matrix $D\Phi(z)$ is symplectic for every $z \in \mathbb{R}^{2d}$.

Let us note the following nice properties of symplectic matrices: if A is symplectic, then $\det(A)^2 = 1$, thus $\det(A) \in \{-1,1\}$.

The following result is fundamental.

Theorem 6.2. *Hamiltonian flow maps are symplectic: for every $t \ge 0$, $\phi_t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is symplectic.*

Moreover, $\det(\phi_t) = 1$, and thus Hamiltonian flows preserve the volume.

Proof. To prove the second property, use that $\det(\phi_t) \in \{-1,1\}$, that $t \mapsto \det(\phi_t)$ is continuous, and that $\phi_0 = I$.

It remains to prove the first property. Note that $W_t^z h = D\phi_t(z).h$ is the solution of

$$\dot{W} = JD^2H(\phi_t(z))W,$$

where $S(t) = D^2H(\phi_t(z))$ is a symmetric mapping. Then

$$\begin{aligned} \frac{d}{dt}\big(D\phi_t(z)^\star JD\phi_t(z)\big) &= \frac{d}{dt}\big(W_t^\star JW_t\big) \\ &= \big(JS(t)W_t\big)^\star JW_t + W_t^\star J\big(JS(t)W_t\big) \\ &= W_t^\star\big(-S(t)J^2 + J^2S(t)\big)W_t, \end{aligned}$$

using $J^\star = -J$. It is straightforward to check that $\big(-S(t)J^2 + J^2S(t)\big)^\star = -\big(-S(t)J^2 + J^2S(t)\big)$, *i.e.* that $\big(-S(t)J^2 + J^2S(t)\big)$ is skew-symmetric. Thus $\frac{d}{dt}\big(D\phi_t(z)^\star JD\phi_t(z)\big) = 0$, and this yields

$$D\phi_t(z)^\star JD\phi_t(z) = D\phi_0(z)^\star JD\phi_0(z) = J,$$

using $\phi_0 = I$. This concludes the proof. □

Good integrators for Hamiltonian systems will be symplectic. We give several examples below.

In connection with the splitting and composition methods, the following property is very useful.

Proposition 6.3. *If ϕ^1 and ϕ^2 are symplectic mappings, then $\phi^2 \circ \phi^1$ is also a symplectic mapping.*

Moreover, the inverse of a symplectic mapping is symplectic.

Proof. First, note that if A and B are symplectic matrices, then

$$(AB)^\star J(AB) = B^\star\big(A^\star JA\big)B = B^\star JB = J,$$

thus the product AB is symplectic.

The result in the general case follows from the chain rule. □

6.4 Symplectic integrators

Definition 6.4. *An integrator Φ^h for an Hamiltonian system is symplectic if Φ_h is a symplectic mapping.*

We first give examples of non-symplectic integrators: the explicit and the implicit Euler schemes. To do so, assume that $d = 1$, and that $H(q,p) = \frac{1}{2}p^2 + V(q)$ with quadratic potential energy function $V(q) = \frac{\alpha}{2}q^2$.

In this example, the explicit Euler scheme is given by

$$q_{n+1} = q_n + hp_n \quad , \quad p_{n+1} = p_n - h\alpha q_n,$$

equivalently $z_{n+1} = \begin{pmatrix} 1 & h \\ -\alpha h & 1 \end{pmatrix}$. The integrator is thus a linear mapping A_h.

Since $\det(A_h) = 1 + \alpha h^2 > 1$, volume is not preserved, and A_h is not symplectic.

The implicit Euler scheme is given by

$$q_{n+1} = q_n + hp_{n+1} \quad , \quad p_{n+1} = p_n - h\alpha q_{n+1},$$

equivalently $z_{n+1} = \begin{pmatrix} 1 & -h \\ \alpha h & 1 \end{pmatrix}^{-1} z_n$. The integrator is again a linear mapping B_h, with $\det(B_h) = \frac{1}{1+\alpha h^2} < 1$: B_h is not symplectic.

Symplectic Euler schemes are defined as follows:

$$\begin{cases} q_{n+1} = q_n + h\nabla_p H(p_{n+1}, q_n) \\ p_{n+1} = p_n - h\nabla_q H(p_{n+1}, q_n) \end{cases} \quad , \quad \begin{cases} q_{n+1} = q_n + h\nabla_p H(p_n, q_{n+1}) \\ p_{n+1} = p_n - h\nabla_q H(p_n, q_{n+1}) \end{cases} \tag{32}$$

Note that in each version, the scheme is implicit in one of the variables and explicit in the other one. However, in the case $H(q,p) = \frac{1}{2}p^2 + V(q)$, both are explicit.

These schemes are in fact constructed using a Lie-Trotter splitting technique, in the case $H(q,p) = \frac{1}{2}p^2 + V(q)$. Indeed, define $H^1(q,p) = \frac{1}{2}p^2$ and $H^2(q,p) = V(q)$. Let ϕ^1 and ϕ^2 denote the associated flows maps. Then it is straightforward to check that

$$\phi_t^1(q,p) = \begin{pmatrix} q + tp \\ p \end{pmatrix} \quad , \quad \phi_t^2(q,p) = \begin{pmatrix} q \\ p - t\nabla_q V(q) \end{pmatrix}$$

Since ϕ_t^1 and ϕ_t^2 are the flows at time t of Hamiltonian systems, they are symplectic mappings. Composing them thus also yields a symplectic mapping. By straightforward computations,

$$\phi_h^1 \circ \phi_h^2(q,p) = \begin{pmatrix} q + h(p - h\nabla_q V(q)) \\ p - h\nabla_q V(q) \end{pmatrix} \quad , \quad \phi_h^2 \circ \phi_h^1(q,p) = \begin{pmatrix} q + hp \\ p - h\nabla_q V(q + hp) \end{pmatrix},$$

which are equivalent to the general formulations of symplectic Euler schemes.

The Störmer-Verlet schemes given by (30) are also symplectic integrators. Indeed, each scheme is the composition of one of the symplectic Euler methods with the other one. Moreover, they are methods of order 2: indeed, one checks that the symplectic Euler methods are the adjoints of each other. Note also that the Störmer -Verlet schemes are in fact obtained by a Strang splitting technique.

The last example of symplectic integrator is the implicit midpoint rule, $z_{n+1} = z_n + hJ\nabla H\left(\frac{z_n + z_{n+1}}{2}\right)$. It is also an order 2 method.

6.5 Preservation of a modified Hamiltonian for the Störmer-Verlet scheme

As will be seen in Section 8, when applied to the harmonic oscillator with $V(q) = \frac{\omega^2}{2}q^2$, the Hamiltonian function H is not preserved by the numerical flow of the Störmer-Verlet integrator.

However, with straightforward computations, the following **modified Hamiltonian** function H_h is preserved (in this particular case):

$$H_h(q,p) = \frac{1}{2}(1 - \frac{h^2\omega^2}{4})\omega^2 q^2 + \frac{1}{2}p^2.$$

This means that $H_h(q_{n+1}, p_{n+1}) = H_h(q_n, p_n)$. Note also that $H_h(q,p) = H(q,p) + \mathrm{O}(h^2)$.

This conservation of a modified Hamiltonian function is interesting for the following reason: we can write

$$H(q_N, p_N) = \mathrm{O}(h^2) + H_h(q_N, p_N) = \mathrm{O}(h^r) + H_h(q_0, p_0) = \mathrm{O}(h^2) + H(q_0, p_0),$$

on intervals $[0, T]$, not depending on h for simplicity. Even if H is not conserved by the numerical flow, it remains close the the level set of H determined by the initial condition.

For general Hamiltonian functions and symplectic integrators, it is possible to construct modified Hamiltonian functions, using truncated series expansions, which are preserved up to an error of order h^p which is arbitrarily large, and which are close to the exact Hamiltonian function at order h^r. Precise statements and proofs are out of the scope of these lecture notes.

7 Conclusion

We have introduced general one-step integrators for first-order ODEs (1), and studied in a general framework properties in terms of stability and consistency, which lead to convergence.

We have seen that getting high-order methods is not necessarily the only aim when constructing a numerical scheme: in addition to stability, (approximately) preserving qualitative properties of the flow may be both interesting on its own and to study the properties of the method.

Such qualitative properties can be convergence when time goes to infinity, preservation of first integral, or the symplectic property of the flow, in the important case of Hamiltonian dynamics.

This is not the end of the story: we have not dealt with multiscale or highly oscillatory problems, neither with the discretization of Partial Differential Equations.

Another important field of research is given by stochastic dynamics: when random terms are added. The case of Gaussian white noise/Brownian Motion is the most popular example, and leads to Stochastic Differential and Partial Differential Equations.

There are now many results on these topics, but still there are many challenges.

8 Numerical illustration: the harmonic oscillator

We illustrate the qualitative properties of the integrators introduced above in the case of the harmonic oscillator in dimension 1: $H(q,p) = \frac{p^2}{2} + \frac{\omega^2 q^2}{2}$.

The Hamiltonian system has a nice formulation when considering the complex variable $z = p + i\omega q$: $\dot{z} = i\omega z$, and thus $z(t) = z(0)e^{i\omega t}$. If $z(0) = r_0 e^{i\varphi_0}$, then $p(t) = r_0 \cos(\omega t + \varphi_0)$ and $q(t) = \frac{r_0}{\omega}\sin(\omega t + \varphi_0)$. Some of the integrators can be written in a similar way.

Moreover, the Hamiltonian function is symply given by $\frac{1}{2}|z|^2$.

Indeed, the explicit Euler scheme, the implicit Euler scheme and the Crank-Nicolson scheme can be written respectively

$$\begin{aligned} z_{n+1} &= (1 + ih\omega)z_n, \\ z_{n+1} &= \frac{1}{1 - ih\omega} z_n, \\ z_{n+1} &= \frac{2 + ih\omega}{2 - ih\omega} z_n. \end{aligned} \tag{33}$$

In particular for the explicit (resp. the implicit) Euler scheme, one sees that $|z_n| \underset{n\to+\infty}{\to} +\infty$ (resp. $z_n \underset{n\to+\infty}{\to} 0$), for every $h > 0$. On the contrary, for the Crank-Nicolson scheme $\frac{1}{2}|z_n|^2 = \frac{1}{2}|z_0|^2$.

In the numerical simulations, we take $\varphi_0 = 0$, $r_0 = 1$. We first take $T = 10$ and $h = 0.01$, and second $T = 100$ and $h = 0.1$.

Let us state some observations from the figures:

- Figure 1: the solution is periodic, and the conservation of the Hamiltonian implies that $\big(q(t), p(t)\big)$ belongs to an ellipse.
- Figure 2: the numerical solution for the explicit Euler scheme does not belong to an ellipse, and energy increases.
- Figure 3: the numerical solution for the explicit Euler scheme does not belong to an ellipse, and energy decreases.
- Figure 4: in the case of the explicit Euler scheme, the energy becomes extremely large (of the order 10^{49}); in the case of the implicit Euler scheme, the enery tends to 0.
- Figure 5: the solution takes values in the same ellipse as the exact solution.
- Figures 6 and 7: the solution remains bounded and periodic, it seems to take values in another ellipse. The energy remains bounded and oscillates.
- Figures 8 and 9: the solution remains bounded and periodic, it seems to take values in another ellipse. The energy remains bounded and oscillates.

- Qualitatively and quantitatively the Störmer-Verlet scheme is better than the symplectic Euler scheme: the energy oscillates on a much smaller interval, and the ellipse is closer to the initial one.

- The near conservation of the energy in Figures 8 and 9 is well-explained by the conservation of the modified energy by the numerical flow, with $H_h(q,p) = \frac{1}{2}(1-\frac{h^2\omega^2}{4})\omega^2 q^2 + \frac{1}{2}p^2$. In the case $h = 0.01$, H_h is equal to 12.49; in the case $h = 0.1$, H_h is equal to 11.72, to be compared with the value $H = 12.5$.

Figure 1: Exact solution, $T = 10$.

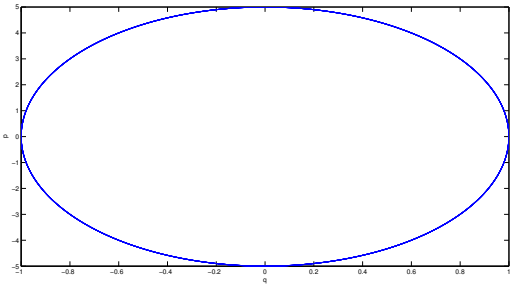


Figure 2: Explicit Euler scheme, $T = 10$ and $h = 0.01$. Upper figure: solution in the (q, p) plane. Lower figure: Evolution of the energy.

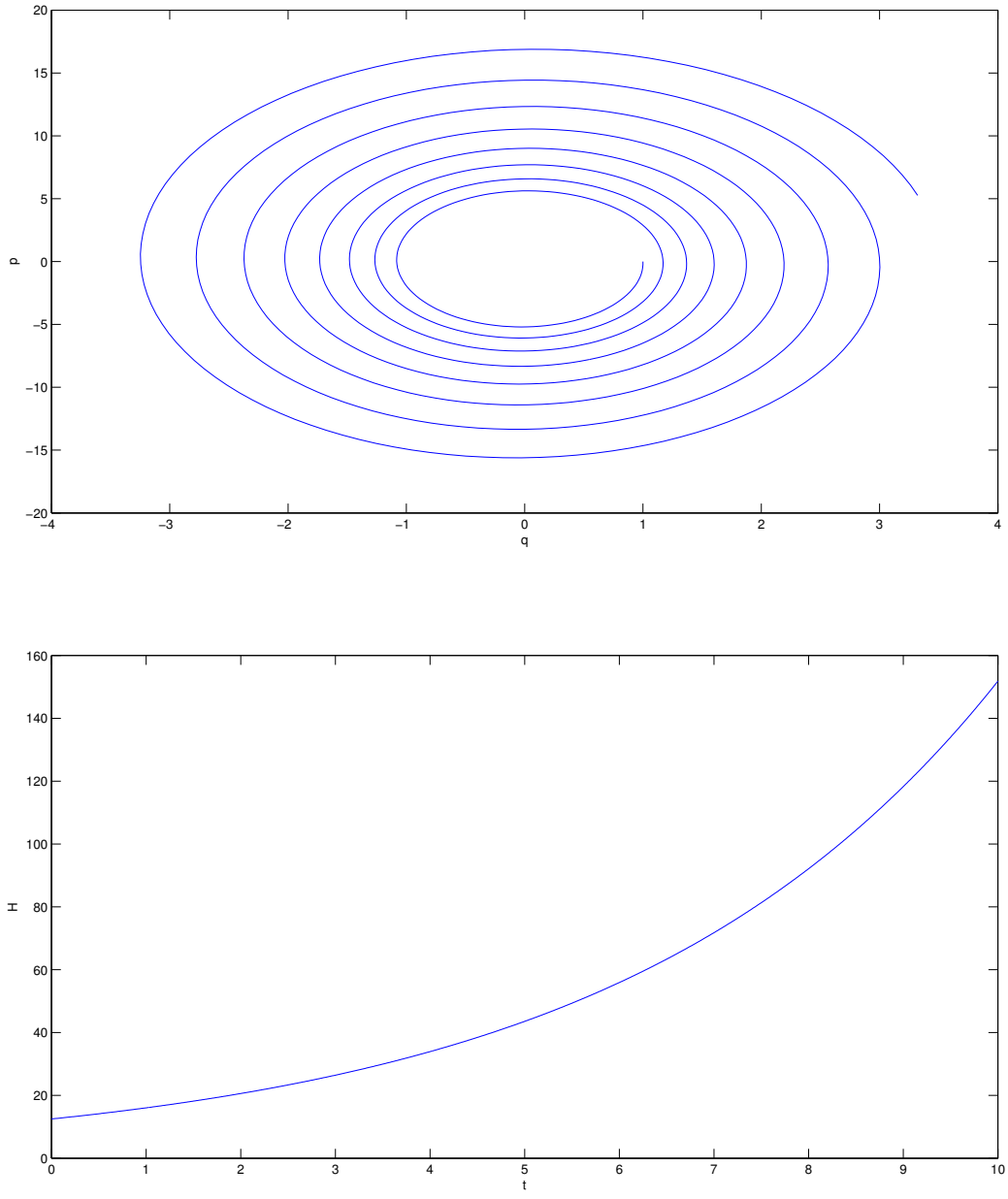


Figure 3: Implicit Euler scheme, $T = 10$ and $h = 0.01$. Upper figure: solution in the (q, p) plane. Lower figure: Evolution of the energy.

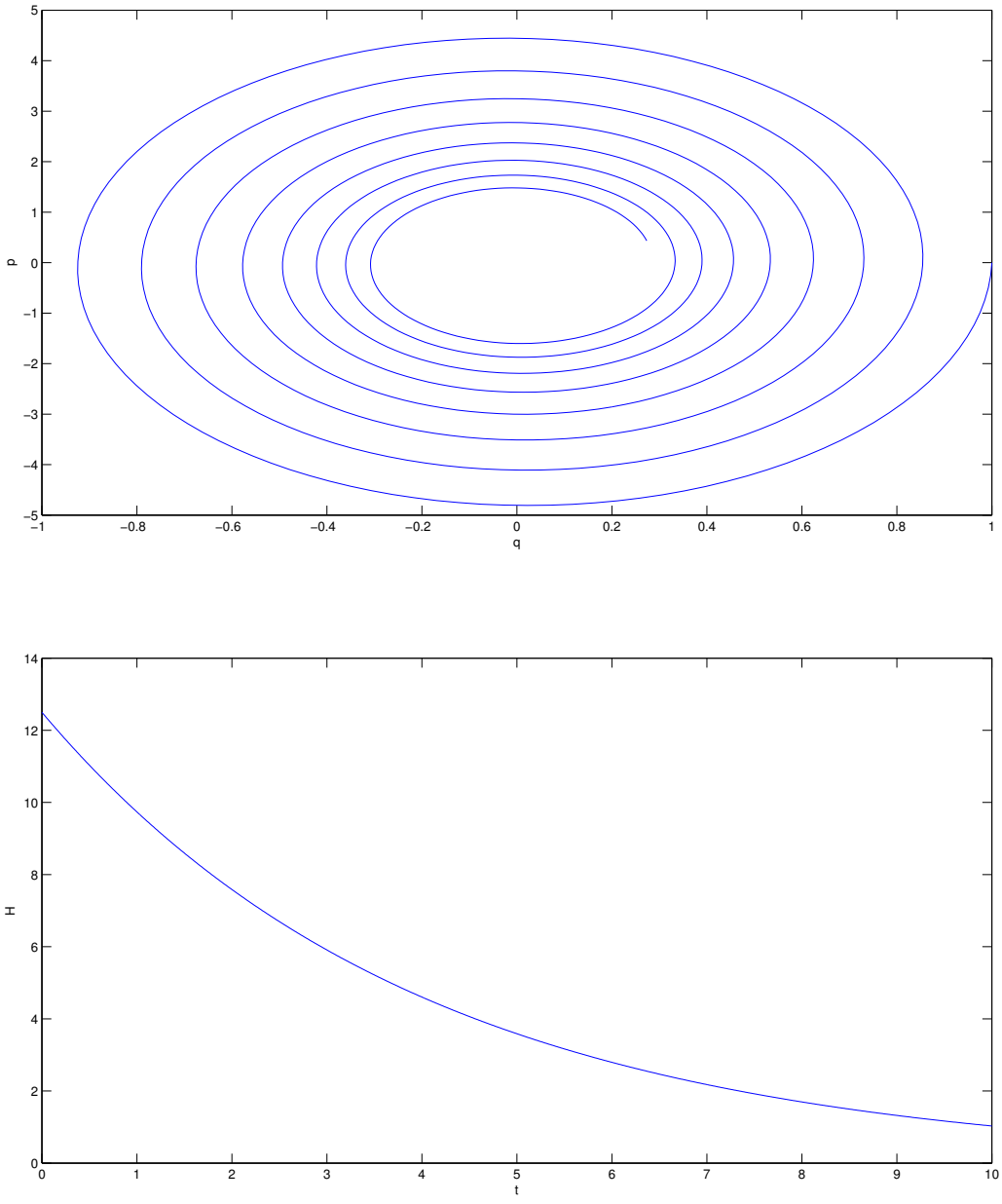
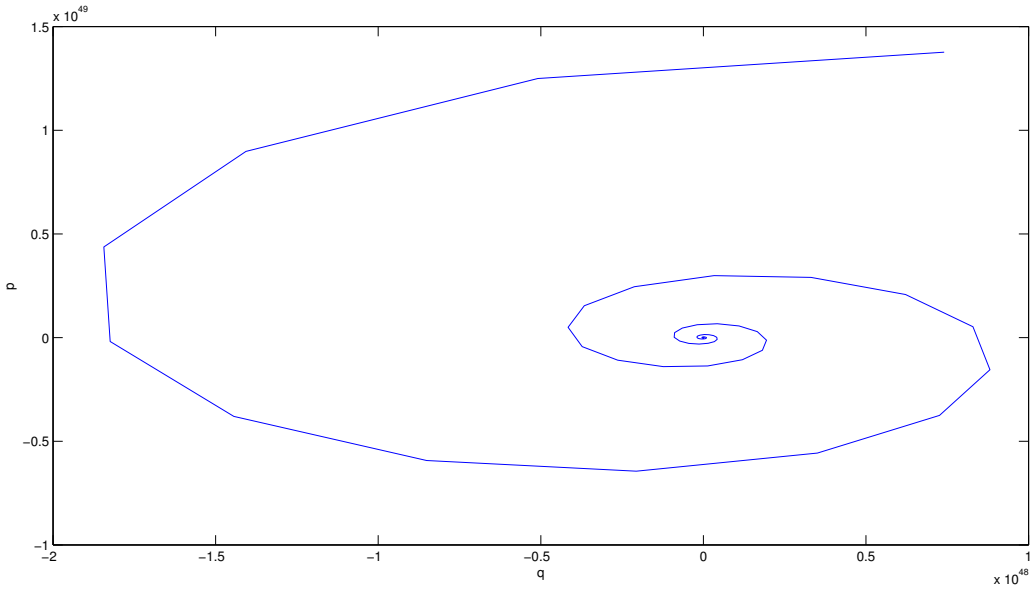


Figure 4: Euler schemes, $T = 100$ and $h = 0.1$. Upper figure: explicit scheme. Lower figure: implicit scheme.



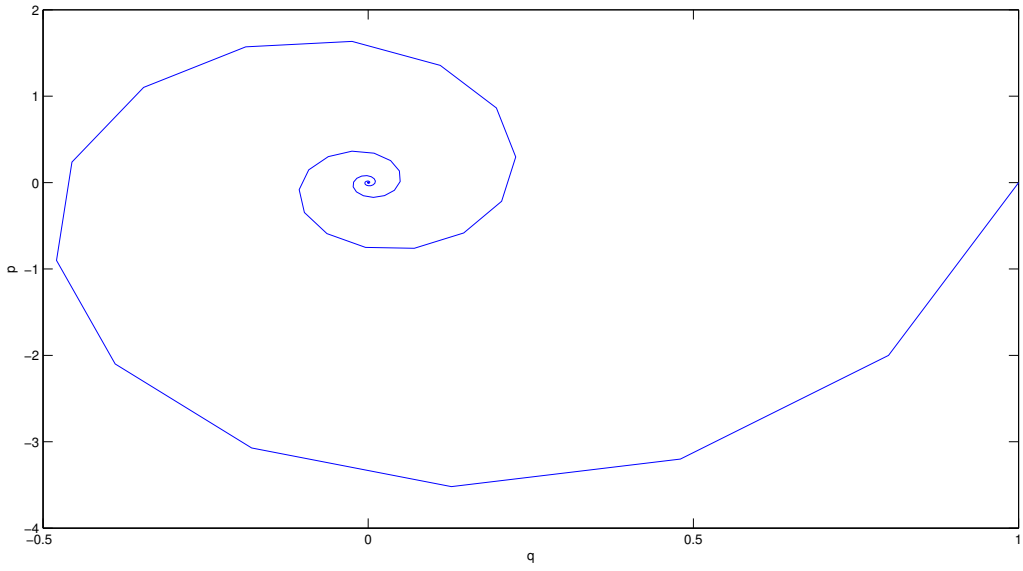


Figure 5: Crank-Nicolson scheme. Upper figure: $T = 10$, $h = 0.01$. Lower figure: $T = 100$, $h = 0.1$.

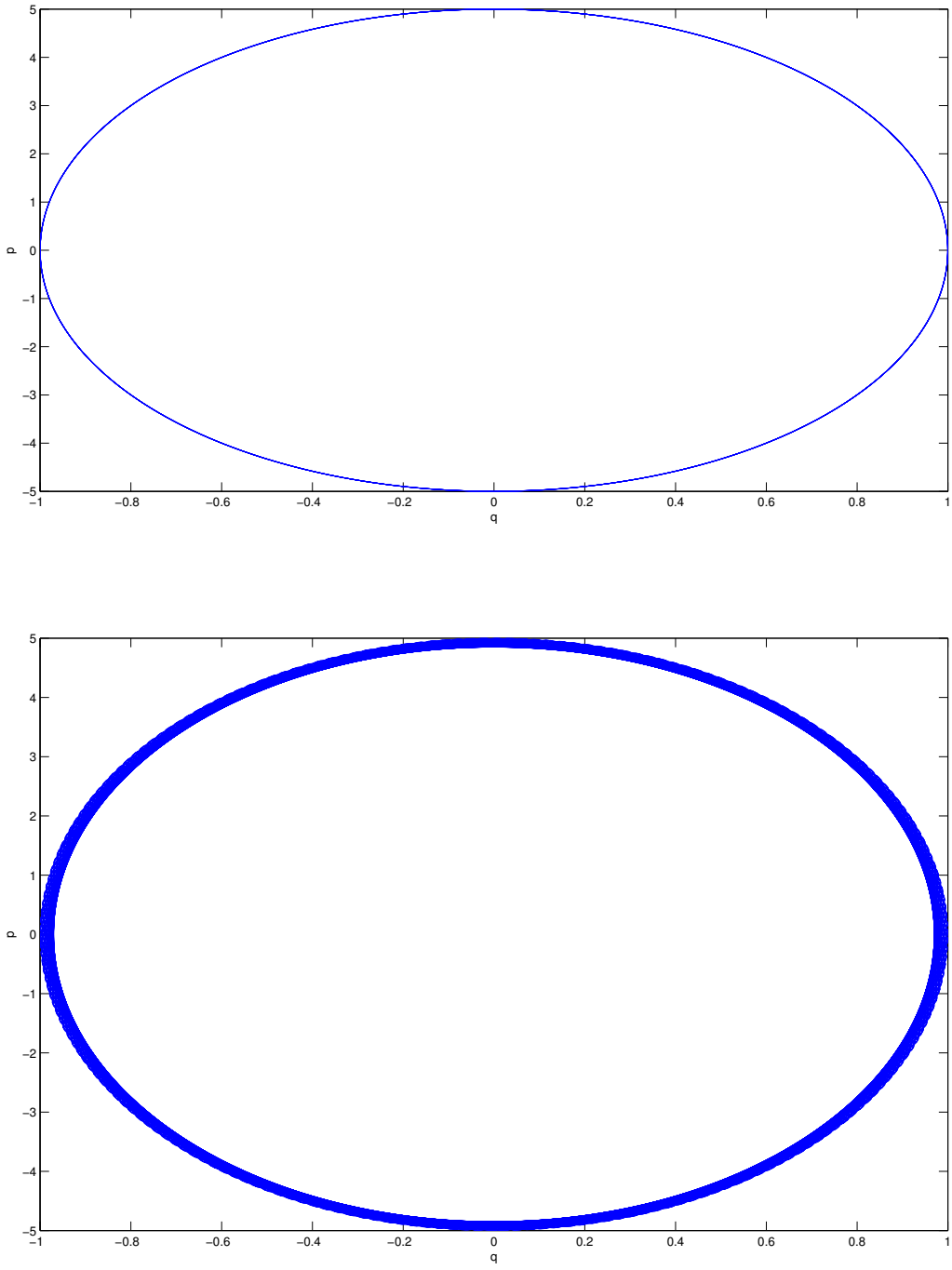


Figure 6: Symplectic Euler scheme, $T = 10$, $h = 0.01$. Upper figure: solution in the (q, p) plane. Lower figure: Evolution of the energy.

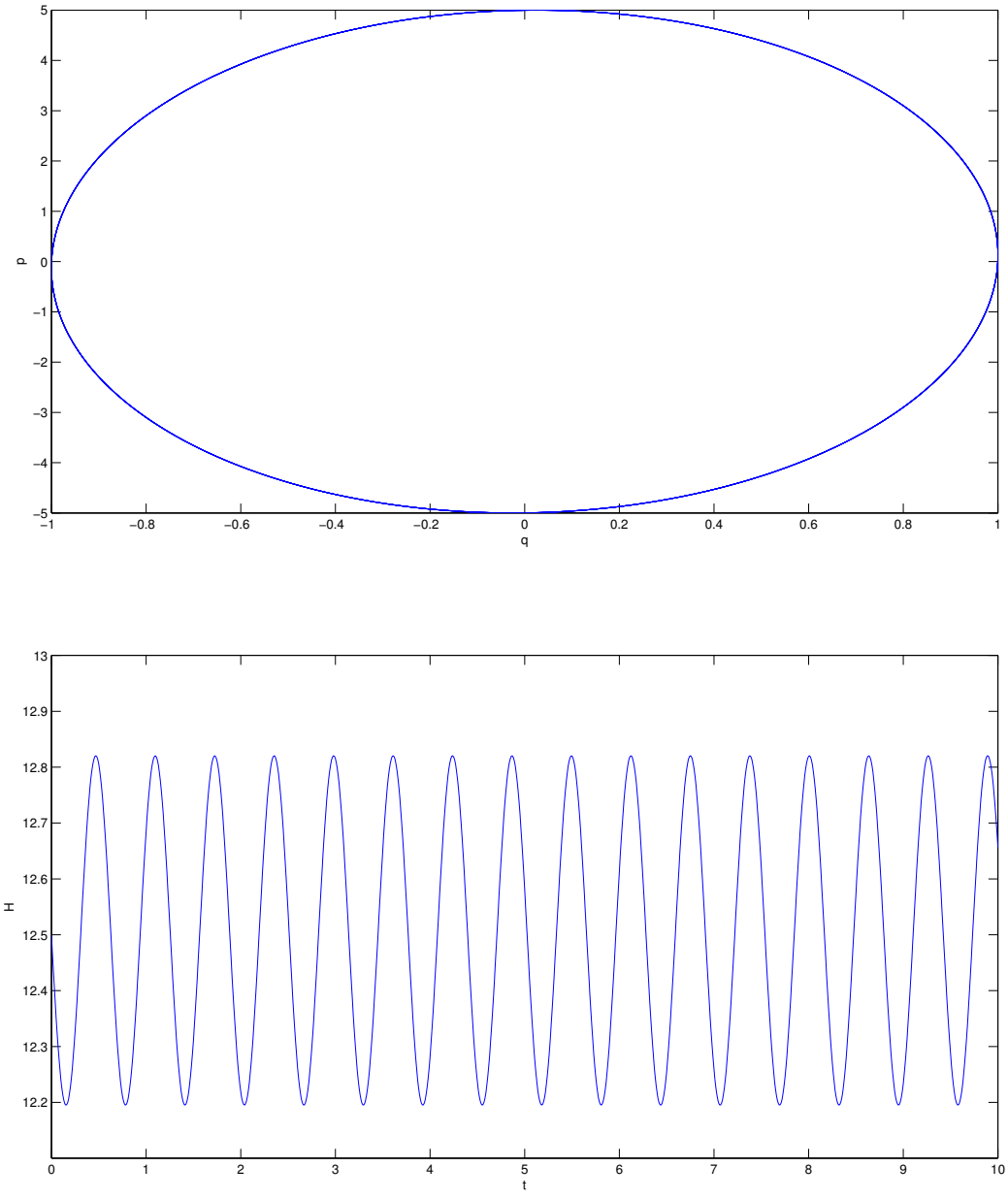
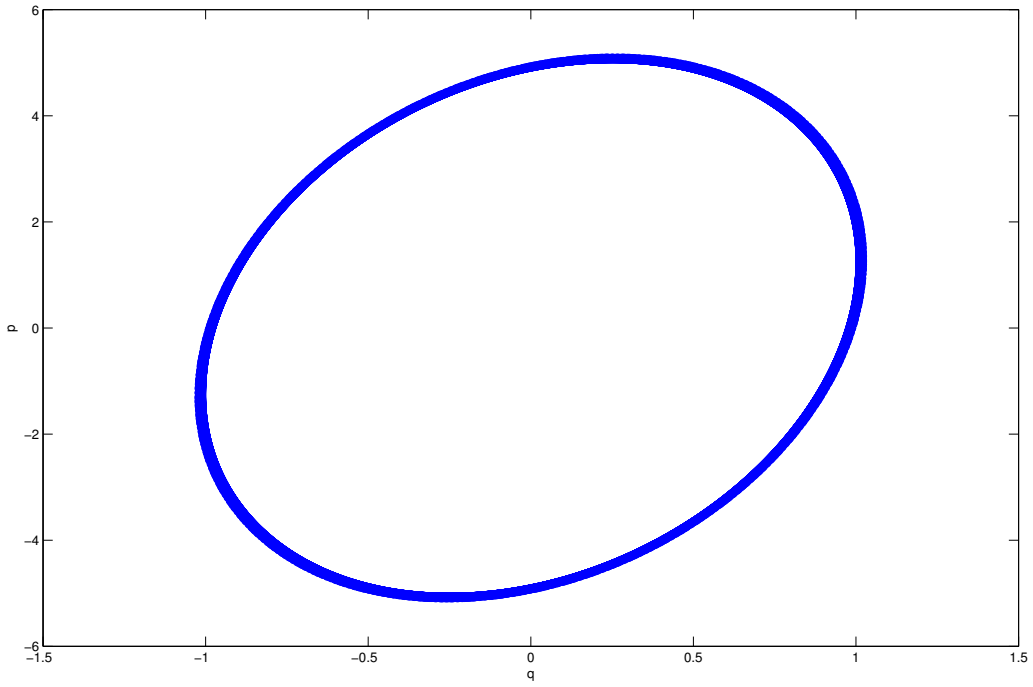


Figure 7: Symplectic Euler scheme, $T = 100$, $h = 0.1$. Upper figure: solution in the (q, p) plane. Lower figure: Evolution of the energy.



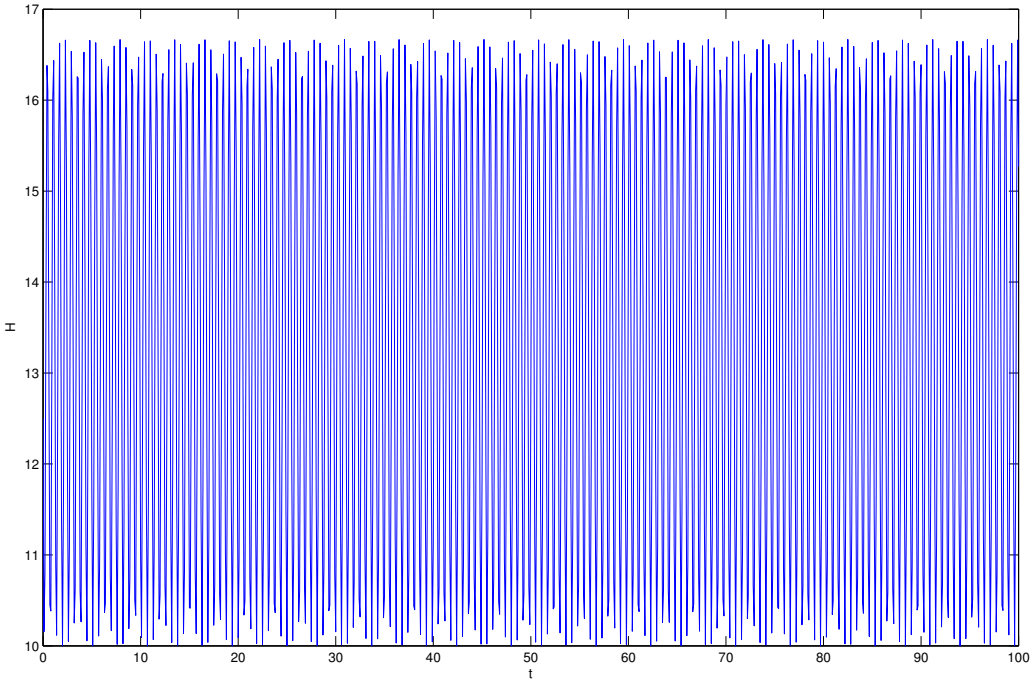


Figure 8: Störmer-Verlet scheme, $T = 10$, $h = 0.01$. Upper figure: solution in the (q, p) plane. Lower figure: Evolution of the energy.

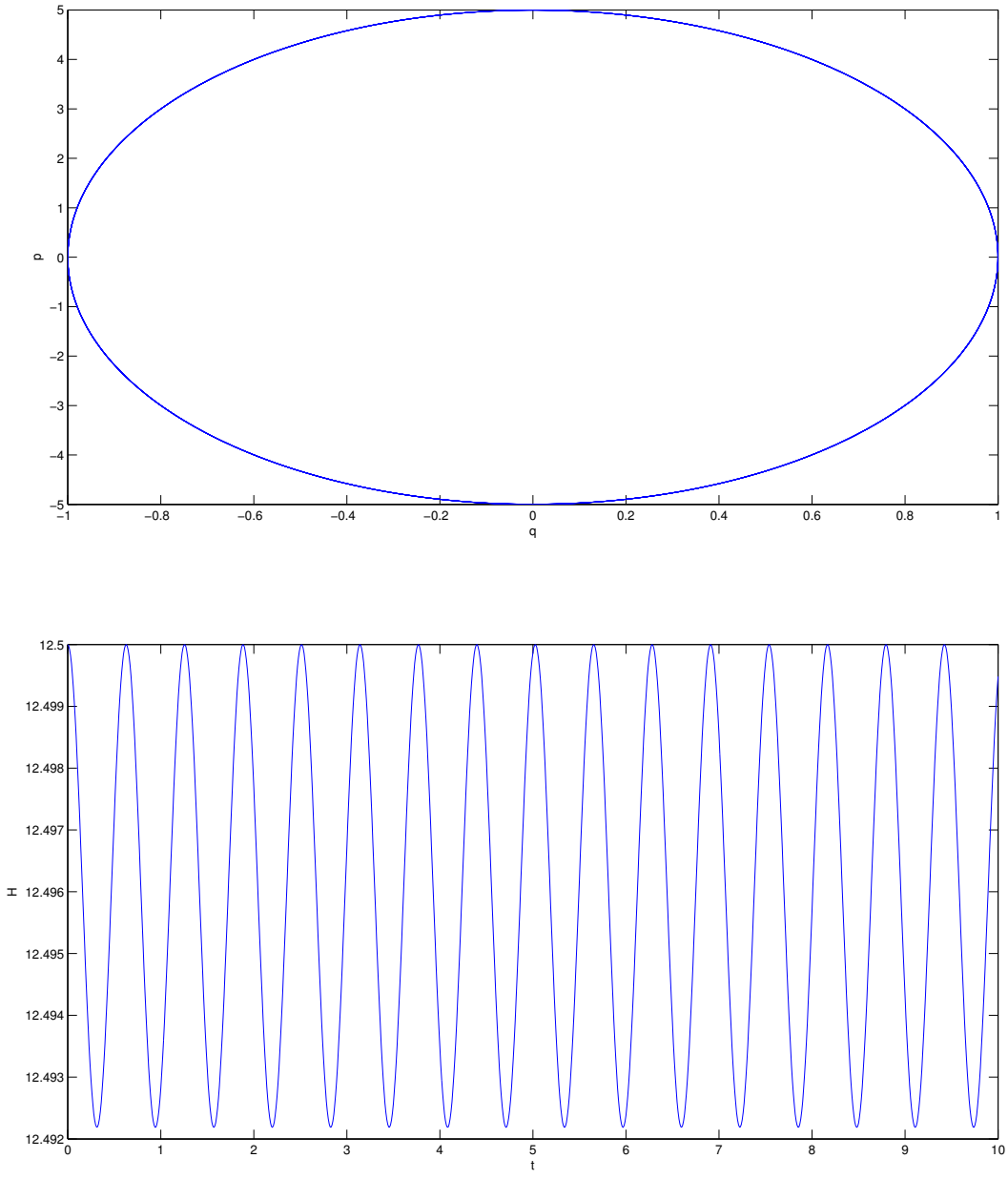


Figure 9: Störmer-Verlet scheme, $T = 100$, $h = 0.1$. Upper figure: solution in the (q, p) plane. Lower figure: Evolution of the energy.