ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

SMA

Splitting Methods with Modified Potentials and Application to the Damped Wave Equation

Jonathan ROCHAT
under the supervision of A. Abdulle
and G.Vilmart

Contents

Abstract

In this report we study and compare particular integration methods to solve ordinary differential equations, which are separable in solvable parts. The main source for this work is the article of Blanes and Casas: "On the necessity of negative coefficient for operator splitting schemes of order higher than two", which was published by ELSEVIER in 2004. After a brief introduction and some preliminaries on fundamental aknownledged, namely the flow of a differential equation which will allow to construct splitting schemes, we start the third part of this work with some definitions and fundamental theorems for general splitting schemes. At the end of this section, we will look more carefully on some special schemes, with modified potentials. In the fourth part, we study and compare some of the different methods seen in the third part of this report on an ordinary separable differential equation. In fifth part, we use these splitting schemes on the damped wave equation and look at the conservation of the Energy. Finally, you will find the main MATLAB code in the annexe.

1 Introduction

Splitting methods are particular numerical schemes applied to ordinary and partial differential equations, which are separable in solvable parts. These methods are always studied nowadays. More precisely, we consider the ordinary differential equation (ODE)

$$y(t)' = f(y(t)), \quad y_0 = y(0) \in \mathbb{R}^n,$$

with $f : \mathbb{R}^n \to \mathbb{R}^n$ a smooth function and the associated vector field

$$F = \sum_{i=1}^{n} f_i(y) \frac{\partial}{\partial y_i}.$$

If we assume that $f(y)$ can be written as $f(y) = f_A(y) + f_B(y)$, then the vector field F is split accordingly as $F = F_A + F_B$ and the flows $\varphi_h^{[A]}$ and $\varphi_h^{[B]}$ corresponding to F_A and F_B can be computed. A splitting scheme is then a composition of these flows with fractionnal time steps. More precisely, a general splitting method is

$$\psi_h = \varphi^{[B]}_{b_m h} \circ \varphi^{[A]}_{a_m h} \circ \cdots \circ \varphi^{[B]}_{b_1 h} \circ \varphi^{[A]}_{a_1 h}. \tag{1}$$

We will see that a scheme of order $p \geq 3$ on the form (1) must have some of his coefficients negative [2]. Sometimes it is possible to build higher order methods by modifie potential. These splitting methods have positive coefficients and we can for example obtain a method of order four from second order schemes.

In the second part of this work, we will carefully look at two examples. The first one will be an ordinary differential equation that is non linear but separable. By computing the numerical error of these methods, we will see that we obtain schemes of order four on this example.

A more interesting example will be the damped wave equation, which is a hyperbolic partial differential equation. It has a computable exact solution by using the method of separation of variables. For the numerical approximation, we transforme this PDE into a system of ODE, and then we will be able to apply on it the different splitting schemes studied in this report. Finally, we study the conservation of energy of the system.

2 Preliminaries

In the following section we give main results about Lie Bracket, flows of an ordinay differential equation and the BCH formula, that will be fundamental for the comprehension of splitting schemes.

2.1 Lie Bracket

The Lie Bracket is a very usefull operator that we mainly see in Lie Algebra or in the study of smooth manifolds. But it will be very useful for this analysis context too. We will use it for the paticular case of matrix and vector fields.

Definition 2.1 (Lie Algebra)
Let K be a field and V a vector space of finite dimension on K. A lie Bracket is a bilinear application

$$[\cdot,\cdot] : V \times V \to V,$$

such that the following properies are verified:

1. $[x, y] = -[y, x]$ *for all* $x, y \in V$ *(skew-symmetry)*
2. $[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0$ *for all* $x, y, z \in V$ *(Jacobi identity)*

The vector space V together with $[\cdot,\cdot]$ is called a Lie Algebra and is denoted by $\mathcal{L}(V)$.

Remark 2.2
Skew-symmetry of the Lie Bracket implies directly that

$$[x, x] = 0 \quad \forall x \in V.$$

Example 2.3
If $V = M_n(K)$, for two matrices $A, B \in V$ the Lie Bracket is

$$[A, B] = AB - BA.$$

We easily see that V with $[\cdot,\cdot]$ is a Lie Algebra, denoted by $\boldsymbol{gl}_n(K)$.

Definition 2.4 (Graded Free Lie-Algebra)
A free Lie Algebra over a given field K is a Lie algebra generated by a set X. A graded free Lie Algebra $\mathcal{L}$ is a free lie algebra when X is a set of vector spaces:

$$\mathcal{L} = \bigoplus_{i\in\mathbb{Z}} \mathcal{L}_i,$$

such that the Lie-Bracket respects this gradation:

$$[\mathcal{L}_i, \mathcal{L}_j] \subset \mathcal{L}_{i+j}.$$

Definition 2.5 (Lie Bracket for vector fields)
Let X and $Y \in \chi(\mathbb{R}^n)$, the set of smooth vectors fields on $\mathbb{R}^n$. Given a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, we can apply X to f to obtain another smooth function Xf. In turn, we can apply Y to this function, and yet obtain another smooth function $YX(f) = Y(Xf)$. But the operation $f \mapsto YX(f)$ in general does not satisfies the derivation rules. However, we can also apply the same two vector fields in the

opposite order. Applying both of these operators to f and substrating, we obtain an operator

$$[X, Y] : C^\infty(\mathbb{R}^n) \mapsto C^\infty(\mathbb{R}^n)$$

called the Lie Bracket of X and Y, which is similarly defined as before:

$$[X, Y]f = XYf - YXf.$$

This Lie Bracket verifies skew-symmetry and Jacobi identity, which implies that $\chi(\mathbb{R}^n)$ is a Lie Algebra. Moroever we have for $X, Y \in \chi(\mathbb{R}^n)$ and $f, g \in C^\infty(\mathbb{R}^n)$:

$$[fX, gY] = fg[X, Y] + (fXg)Y - (gYf)X.$$

Example 2.6
Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a smooth function and consider the vector fields

$$F_A \equiv x_2 \frac{\partial}{\partial x_1} \text{ and } F_B \equiv g(x_1) \frac{\partial}{\partial x_2}.$$

Then, the Lie Bracket of this two fields is computing as follows:

$$\begin{aligned}
[F_A, F_B] &= x_2 \frac{\partial}{\partial x_1} \left(g(x_1) \frac{\partial}{\partial x_2} \right) - g(x_1) \frac{\partial}{\partial x_2} \left(x_2 \frac{\partial}{\partial x_1} \right) \\
&= x_2 \left(g'(x_1) \frac{\partial}{\partial x_2} + g(x_1) \frac{\partial^2}{\partial x_1 \partial x_2} \right) - g(x_1) \left(\frac{\partial}{\partial x_1} + x_2 \frac{\partial^2}{\partial x_2 \partial x_1} \right) \\
&= x_2 g'(x_1) \frac{\partial}{\partial x_2} - g(x_1) \frac{\partial}{\partial x_1}.
\end{aligned}$$

where in the last step we used the fact that we always compute Lie Bracket on smooth function, which mixed partial derivatives can be taken in any order.

2.2 Flows

Flow are the most important elements of our work, because they allow to construct Splitting methods. The main result to remember here is the fact that the set of flows is a group for the composition of application.

Definition 2.7 (Flow of an ordinary differential equation (ODE))
Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function. Then, for any initial condition $y_0 \in \mathbb{R}^n$, the ODE

$$y'(t) = f(y(t), t) \tag{2}$$

has a unique solution on $[0, T[$, $T \in]0, \infty[$, by the theorem of Cauchy-Lipschitz. We denote this solution by $y(t)$. Consequently, $y(0) = y_0$ and $y'(t) = f(y(t), t)$. We suppose that $T \to \infty$ and then we define the flow of (2), which is a map

$$\varphi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$$

defined by

$$\varphi(h, y_0) = y(h).$$

We often abuse the terminology and call the map

$$\varphi_h : \mathbb{R}^n \to \mathbb{R}^n$$

defined by

$$\varphi_h(y) = \varphi(h, y)$$

the h-flow of (2).

Theorem 2.8
For any $s, t \in \mathbb{R}$:

1. $\varphi_0(y_0) = y_0$;
2. $\varphi_{t+s}(y) = \varphi_t \circ \varphi_s(y)$;
3. $\varphi_t^{-1}(y) = \varphi_{-t}(y)$.

Proof.

1. $\varphi_0(y_0) = \varphi(0, y_0) = y_0$.
2. First of all, we remark that the composition $\varphi_t \circ \varphi_s(y)$ and the flow $\varphi_{t+s}(y)$ have the same initial condition $y(s)$. In addition,

$$\begin{aligned} \frac{\partial}{\partial t}(\varphi_{t+s}(y_0)) &= f(\varphi_{t+s}(y_0))\frac{\partial(t+s)}{\partial t} = f(\varphi_{t+s}(y_0)), \\ \frac{\partial}{\partial t}(\varphi_t \circ \varphi_s(y_0)) &= f(\varphi_t(\varphi_s(y_0))). \end{aligned}$$

   Consequently, $\varphi_t \circ \varphi_s(y)$ and $\varphi_{t+s}(y)$ are both flows of (2). But they have the same initial condition, so

$$\varphi_t(\varphi_s(y_0)) = \varphi_{t+s}(y_0).$$

   We have then, by the uniqueness of solutions,

$$\varphi_{t+s}(y) = \varphi_t \circ \varphi_s(y).$$

3. As shown in the precedent point, $\varphi_{t-t}(y) = \varphi_t \circ \varphi_{-t}(y) = \varphi_{-t} \circ \varphi_t(y) = y$, which clearly implies that $\varphi_t^{-1}(y) = \varphi_{-t}(y)$.

$\square$

Remark 2.9
The theorem (2.8) *signifies that the set of all h-flow of* (2) *is in fact a group for the composition of map.*

Remark 2.10
Sometimes, we can define the flow of an partial differential equation in a similar way.

The next result will be very useful to define splitting schemes with modified potentials.

Theorem 2.11
Let X and Y be smooth vector fields on $\mathbb{R}^n$, with φ and ψ the respective flows. The following properties are equivalent:

1. $[X, Y] = 0$;
2. $\varphi_t \circ \psi_s = \psi_s \circ \varphi_t$.

Proof. See proposition 18.5 on page 489 of [5]. The proof use some applications defined for the study of Smooth Manifold, called pushforward, that we will not introduce here. $\square$

2.3 Baker-Campell-Hausdorff (BCH) formula

This formula will be very usefull when we will look on the order of a splitting method. For example, two matrix $A, B \in \mathbf{gl}_n(\mathbb{R})$., we have hardly ever $[A, B] = 0$, and then

$$e^A e^B \neq e^{A+B}.$$

The BCH formula will complete the lack.

Definition 2.12
Let $F \in \chi(\mathbb{R}^n)$ a smooth vector field. The application $\exp : \chi(\mathbb{R}^n) \to \chi(\mathbb{R}^n)$ *is the serie*

$$\exp(F) = \sum_{k=0}^{\infty} \frac{F^k}{k!}$$

where

$$F^k = F \circ \cdots \circ F \quad k \textit{ times.}$$

When F is a linear, it can be designed as a matrix A, and we obtain

$$\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

for the product of matrix.

Proposition 2.13
Let $n \in \mathbb{N}$ and $g : \mathbb{R}^n \to \mathbb{R}^n$. For any $\epsilon > 0$ we have

$$e^{g(x)+\epsilon} = e^{g(x)} + \mathcal{O}(\epsilon)$$

Proof. We simply use the Taylor Development of $f(g(x) + \epsilon) = e^{g(x)+\epsilon}$ around $g(x)$ and we have

$$e^{g(x)+\epsilon} = e^{g(x)} + \frac{\epsilon}{2}(g'(x)e^{g(x)} + \frac{\epsilon^2}{6}(e^{g(x)}(g'(x) + g''(x)) + \cdots = e^{g(x)} + \mathcal{O}(\epsilon).$$

□

Proposition 2.14 (Baker-Campell-Hausdorff Formula (BCH))
We consider A, $B \in \boldsymbol{gl}_n(\mathbb{R})$ and $h \in \mathbb{R}$. Then, we have the following approximation up to order three:

$$e^{hA}e^{hB} = \exp\left(h(A+B) + \frac{h^2}{2}[A,B] + \frac{h^3}{12}[A,[A,B]] - \frac{h^3}{12}[B,[A,B]] + \mathcal{O}(h^4)\right). \tag{3}$$

Proof. First, we develop $e^{hA}e^{hB}$ up to order three:

$$\begin{aligned}
&= \left(I + hA + \frac{h^2}{2}A^2 + \frac{h^3}{6}A^3 + \mathcal{O}(h^4)\right)\left(I + hB + \frac{h^2}{2}B^2 + \frac{h^3}{6}B^3 + \mathcal{O}(h^4)\right) \\
&= I + h(A+B) + \frac{h^2}{2}(A^2 + B^2 + 2AB) \\
&+ \frac{h^3}{6}(A^3 + B^3 + 3AB^2 + 3A^2B) + \mathcal{O}(h^4).
\end{aligned}$$

On the other hand, we obtain:

$$
\begin{aligned}
& \exp\left(h(A+B)+\frac{h^2}{2}[A,B]+\frac{h^3}{12}[A,[A,B]]-\frac{h^3}{12}[B,[A,B]]+\mathcal{O}(h^4)\right) \\
= \; & I+h(A+B)+\frac{h^2}{2}[A,B]+\frac{h^3}{12}[A,[A,B]]-\frac{h^3}{12}[B,[A,B]] \\
+ \; & \frac{1}{2}\left(h(A+B)+\frac{h^2}{2}[A,B]+\frac{h^3}{12}[A,[A,B]]-\frac{h^3}{12}[B,[A,B]]\right)^2 \\
+ \; & \frac{1}{6}\left(h(A+B)\right)^3+\mathcal{O}(h^4) \\
= \; & I+h(A+B)+\frac{h^2}{2}(AB-BA+A^2+B^2+AB+BA) \\
+ \; & \frac{h^3}{6}((A^3+B^3+ABA+BA^2+A^2B+AB^2+BAB+B^2A) \\
+ \; & \frac{1}{2}\left([A,[A,B]]-[B,[A,B]]\right))+\frac{3}{2}\left((A+B)[A,B]+[A,B](A+B)\right)+\mathcal{O}(h^4) \\
= \; & I+h(A+B)+\frac{h^2}{2}(A^2+B^2+2AB) \\
+ \; & \frac{h^3}{6}(A^3+B^3+3A^2B+3AB^2)+\mathcal{O}(h^4)
\end{aligned}
$$

As we compare the two developments, the proposition follows. □

3 Splitting Schemes

Splitting methods are numerical schemes constructed by a composition of flows. We first give results when the coefficients can be negative and we will see how to construct method up to order three. Then we will look particulary on symmetric methods and so called "effective order" schemes, which are useful to obtain for example a scheme of order four avoiding too much compute. Finally, when the potential is modified, we can obtain high order method with positive coefficients and we will give examples of schemes of order and effective order four.

3.1 Scheme with Positive and Negative Coefficients

Consider $n \in \mathbb{N}$ and the following separable ODE

$$
\begin{aligned}
y'(t) &= f(y(t),t), \\
y(0) &= y_0
\end{aligned}
\tag{4}
$$

for $y \in \mathbb{R}^n, h \in I \subset \mathbb{R}$ and $f:\mathbb{R}^n \to \mathbb{R}^n$ a function. We can associate to f a vector field F, called the "Lie Operator" which is defined by

$$F=\sum_{i=1}^{n} f_i(x)\frac{\partial}{\partial x_i}.$$

We denote by φ_h the h-flow of the ODE, and the exact solution is given by

$$y(h)=\varphi_h(y_0).$$

Now let assume that $f(y)$ can be written $f(y)=f_A(y)+f_B(y)$ so that the equations $y'=f_A(y)$ and $y'=f_B(y)$ are solvable. Then the vector field F is split as $F=$

$F_A + F_B$ and the h-flows corresponding to F_A and F_B can be exactly computed and we note them $\varphi_h^{[A]}$ and $\varphi_h^{[B]}$.

For the major theorical part of this report, we only consider linear problem. It means that problem (4) becomes

$$\begin{aligned} y'(h) &= My(h), \\ y(0) &= y_0 \end{aligned} \tag{5}$$

for $y \in \mathbb{R}^n$, $h \in I \subset \mathbb{R}$ and $M \in M_n(\mathbb{R})$. In this case, the flow, easy to find, is noted $\varphi_h^{[M]}(y) = ye^{hM}$. When $M = A + B$, it motivates to use splitting methods, which are compositions of the two flows associated to (5). Here, the two flows are

$$\varphi_h^{[A]}(y) = e^{hA}y \quad \text{and} \quad \varphi_h^{[B]}(y) = e^{hB}y.$$

Example 3.1
One of the easiest splitting flow you can do with (5) *is called the "Lie-Trotting scheme" and this is just a composition of the two flows:*

$$\psi_h(y) = \varphi_h^{[B]} \circ \varphi_h^{[A]}(y) = ye^{hA}e^{hB}.$$

Later we will see later more difficult examples.

Definition 3.2 (Order of a scheme)
Let $M = A + B \in M_n(\mathbb{R})$ and $\varphi_h^{[A]}, \varphi_h^{[B]}$ the two flows of the ODE (5) *and consider the following scheme:*

$$\psi_h = \varphi_{b_m h}^{[B]} \circ \varphi_{a_m h}^{[A]} \circ \cdots \circ \varphi_{b_1 h}^{[B]} \circ \varphi_{a_1 h}^{[A]} = e^{ha_1A}e^{hb_1B} \ldots e^{ha_mA}e^{hb_mB}, \tag{6}$$

with $(a_1, b_1, \ldots, a_m, b_m) \in \mathbb{R}^{2m}$. We say that the scheme (6) *is of order p if*

$$\psi_h = \phi_h + \mathcal{O}(h^{p+1})$$

for a proper choice of m and the coefficient a_i, b_i.

Theorem 3.3
[2, p.27] We consider the scheme (6)*:*

$$\psi_h = e^{ha_1A}e^{hb1B} \ldots e^{ha_mA}e^{hb_mB}.$$

This scheme is of order p if, in terms of the coefficient a_i, b_i, it corresponds to the following order conditions:

$$\begin{aligned} &p = 1 : \textstyle\sum_{i=1}^m a_i = 1, \qquad \sum_{i=1}^m b_i = 1; \\ &p = 2 : \textstyle\sum_{i=1}^m b_i \left(\sum_{j=1}^i a_j\right) = \frac{1}{2}; \\ &p = 3 : \textstyle\sum_{i=1}^{m-1} b_i \left(\sum_{j=i+1}^m a_j\right)^2 = \frac{1}{3}, \quad \sum_{i=1}^m a_i \left(\sum_{j=i}^m b_j\right)^2 = \frac{1}{3}. \end{aligned}$$

Proof. As it is a very long compute, we will not show the result for $p = 3$. For the proof of $p = 1, 2$ we must compute the general term: using the BCH formula and proceding by recurrence on m, we find the following expression for the scheme (6)

$$\psi_h = e^{h\left(\sum_{i=1}^m a_iA + \sum_{j=1}^m b_jB\right) + \frac{h^2}{2}\left(\sum_{i=1}^m b_i\left(\sum_{j=1}^i a_i\right) - \sum_{i=1}^{m-1} b_i\left(\sum_{j=i+1}^m a_j\right)\right)[A,B] + \mathcal{O}(h^3)} \tag{7}$$

Now, we could easily see the two conditions: if we want a scheme of order $p = 1$, it means that

$$\psi_h = e^{h(A+B)} + \mathcal{O}(h^2),$$

and this is verified if $\sum_{i=1}^{m} a_i = 1$ and $\sum_{i=1}^{m} b_i = 1$ and by proposition 2.
When $p = 2$, we want to show that

$$\psi_h = e^{h(A+B)} + \mathcal{O}(h^3).$$

In others words, we want that the coefficient of h^2 vanishes:

$$\sum_{i=1}^{m} b_i \left(\sum_{j=1}^{i} a_i \right) - \sum_{i=1}^{m-1} b_i \left(\sum_{j=i+1}^{m} a_j \right) = 0. \tag{8}$$

But, if

$$\sum_{i=1}^{m} b_i \left(\sum_{j=1}^{i} a_j \right) = \frac{1}{2}, \quad \sum_{i=1}^{m} a_i = 1, \quad \sum_{i=1}^{m} b_i = 1$$

then we have

$$\begin{aligned} \sum_{i=1}^{m-1} b_i \left(\sum_{j=i+1}^{m} a_j \right) = \sum_{i=1}^{m-1} b_i \left(1 - \sum_{j=1}^{i} a_j \right) &= \sum_{i=1}^{m-1} b_i - \sum_{i=1}^{m-1} b_i \left(\sum_{j=1}^{i} a_j \right) \\ &= 1 - b_m - \sum_{i=1}^{m-1} b_i \left(\sum_{j=1}^{i} a_j \right) \\ &= 1 - \sum_{i=1}^{m} b_i \left(\sum_{j=1}^{i} a_j \right) \\ &= 1 - \frac{1}{2} = \frac{1}{2} \end{aligned}$$

Consequently, the relation (8) is verified and the scheme (6) is of order 2. □

Example 3.4
The Lie Scheme has only the coefficient $a_1 = b_1 = 1$. The first condition is evident, but the second one is not verified, because $a_1 b_1 = 1 \neq \frac{1}{2}$. So, it is of order 1.

The Strang scheme has the coefficients $a_1 = a_2 = \frac{1}{2}$ and $b_1 = 1, b_2 = 0$ and conditions (1) and (2) are verified:

$$a_1 + a_2 = b_1 + b_2 = 1 \text{ and } b_1 a_1 + b_2(a_1 + a_2) = \frac{1}{2}.$$

This implies that the Strang scheme is of order two. It is not of order three because for example, $b_1(a_2)^2 = \frac{1}{4} \neq \frac{1}{3}$.

Definition 3.5
For a scheme ψ_h, we define his adjoint scheme by $\psi_h^* = \psi_{-h}^{-1}$.

Theorem 3.6
[2, theorem 2, p.27 − 28] If p is a positive integer such as $p \geq 3$ and m any finite positive integer, then, for every pth-order method of the form (6), one has

$$min_{1 \leq i \leq m} a_i < 0 \quad \text{and} \quad min_{1 \leq j \leq m} b_j < 0.$$

Proof. This theorem tells us precisely that a scheme of the form (6) with m any positive integer and all the coefficients a_i, b_i being positive cannot satisfy the equations from theorem (3.3):

$$\sum_{i=1}^{m-1} b_i \left(\sum_{j=i+1}^{m} a_j \right)^2 = \frac{1}{3}, \quad \sum_{i=1}^{m} a_i \left(\sum_{j=i}^{m} b_j \right)^2 = \frac{1}{3}.$$

Consider the first order method $\phi_h = e^{hB}e^{hA}$, its adjoint $\phi_h^* = e^{hA}e^{hB}$ and the splitting method

$$\psi_h = e^{b_m B}e^{a_m A} \ldots e^{a_2 A}e^{b_1 B}e^{a_1 A}. \tag{9}$$

There exists a connection between this three schemes. Indeed, by composing ϕ_h and ϕ_h^* with different time steps we obtain:

$$\psi_h = \phi^*_{\beta_{2m}h} \circ \phi_{\beta_{2m-1}h} \circ \cdots \circ \phi^*_{\beta_2 h} \circ \phi_{\beta_1 h} \circ \phi^*_{\beta_0 h} \tag{10}$$

Now, by inserting the explicit form of $\phi_{\beta_i h}$ and $\phi^*_{\beta_i h}$ in (10) we have

$$\begin{aligned} \psi &= e^{\beta_{2m}hA}e^{\beta_{2m}hB}e^{\beta_{2m-1}hB}e^{\beta_{2m-1}hA} \ldots e^{\beta_{2h}A}e^{\beta_{2h}B}e^{\beta_{1h}B}e^{\beta_{1h}A}e^{\beta_{0h}A}e^{\beta_{0h}B} \\ &= e^{\beta_{2m}hA}e^{(\beta_{2m}+\beta_{2m-1})hB}e^{(\beta_{2m}+\beta_{2m-1})hA} \ldots e^{(\beta_2+\beta_1)hB}e^{(\beta_1+\beta_0)hA}e^{\beta_0 hB} \end{aligned}$$

when we use the property of the group of the exact flow in the last equality. If we take $\beta_{2m} = \beta_0 = 0$, then the last scheme is equivalent to the scheme (19), with

$$a_i = \beta_{2i-1} + \beta_{2i-2} \quad \text{and} \quad b_i = \beta_{2i} + \beta_{2i-1}.$$

Then

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{m} b_j = \sum_{k=0}^{2m} \beta_k$$

and the consistency of both schemes requirs in fact that

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{m} b_j = \sum_{k=0}^{2m} \beta_k = 1.$$

It has been shown that the order conditions for the coefficients a_i, b_j to get a method of order $p \geq 1$ are equivalent to the order conditions for the β_i. In this case, the scheme ψ_h (6) can be expressed as

$$\psi_h = \exp(-X_{-\beta_0 h})\exp(X_{\beta_1 h})\exp(-X_{-\beta_2 h}) \ldots \exp(X_{\beta_{2m-1}h})\exp(-X_{-\beta_{2m}h}).$$

By repeting application of the BCH formula, we have

$$\psi_h = \exp(hf_{1,1}X_1 + h^2 f_{2,1}X_2 + h^3(f_{3,1}X_3 + f_{3,2}[X_1, X_2]) + \mathcal{O}(h^4)) \tag{11}$$

where the coefficients $f_{k,j}$ are polynomials of degree k in the variable β_i and the X_k are the Lie Brackets of the three order BCH formula:

$$X_1 = A + B, X_2 = \frac{1}{2}[A, B], X_3 = \frac{1}{12}([A, [A, B]] - [B, [A, B]]).$$

The first terms are

$$f_{1,1} = \sum_{i=0}^{2m} \beta_i, \quad f_{2,1} = \sum_{i=0}^{2m} (-1)^{i+1}\beta_i^2, \quad f_{3,1} = \sum_{i=0}^{2m} \beta_i^3. \tag{12}$$

Then, $f_{1,1} = 1$ and $f_{i,j} = 0$ for all $i \leq p$ are sufficient for the method to be of order p. From (12) it is clear that

$$f_{3,1} = \sum_{i=0}^{2m} \beta_i^3 = 0$$

is a necessary condition to be satisfied by any method of order $p \geq 3$. We suppose that more than two β_i are differents from zero, because $\beta_1^3 + \beta_2^3 = 0$ together with the consistency condition $\beta_1 + \beta_2 = 1$ have no real solutions. On the other hand, for all positive integers m,

$$\begin{aligned} f_{3,1} &= \sum_{i=0}^{2m} \beta_i^3 = \beta_0^3 + \beta_1^3 + \cdots + \beta_{2m-2}^3 + \beta_{2m-1}^3 + \beta_{2m}^3 \\ &= (\beta_0^3 + \beta_1^3) + \cdots + (\beta_{2m-2}^3 + \beta_{2m-1}^3) + \beta_{2m}^3 = \sum_{i=0}^{m} (\beta_{2i-2}^3 + \beta_{2i-1}^3) + \beta_{2m}^3 \\ &= \sum_{i=0}^{m} (\beta_{2i-2}^3 + \beta_{2i-1}^3), \end{aligned}$$

because $\beta_{2m} = 0$. Consequently, $\beta_{2j-1}^3 + \beta_{2j-2}^3$ has to be negative for some $1 \leq j \leq m$. But we can verify that $\text{sign}(x^3 + y^3) = \text{sign}(x + y)$ for any $x, y \in \mathbb{R}$, which implies that

$$a_j = \beta_{2j-2} + \beta_{2j-1} < 0$$

for some j so that $1 \leq j \leq m$. Similarly, we can also write

$$f_{3,1} = \beta_0^3 + \sum_{i=0}^{m} (\beta_{2i-1}^3 + \beta_{2i}^3) = \sum_{i=0}^{m} (\beta_{2i-1}^3 + \beta_{2i}^3) = 0$$

so that $b_k = \beta_{2k-1} + \beta_{2k} < 0$. □

Remark 3.7
In general, for a splitting method of order $p \geq 3$ of the form (6), the negative coefficients are in consecutive places. See [2] for further informations.

Definition 3.8
We say that a numerical method $y_{n+1} = \psi_h(y_n)$ is symmetric if it satisfies

$$\psi_h \circ \psi_{-h} = Id.$$

Consequently, ψ_h is symmetric if and only if $\psi_h = \psi_h^*$.

Proposition 3.9
Let us consider a symmetric method ϕ_h of odd order p. Then, ϕ_h is of order $p + 1$.

Proof. Because ϕ_h is of order p, we can write

$$\phi_h(y) = \varphi_h(y) + h^{p+1} C(y) + \mathcal{O}(h^{p+2}) \tag{13}$$

where $C(y)$ only depends of y. Now we will show that we can write ϕ_h^* as

$$\phi_h^*(y) = \varphi_h(y) + (-1)^p h^{p+1} C(y) + \mathcal{O}(h^{p+2}).$$

First of all, by replacing h by $-h$ in (13), we obtain

$$\phi_{-h}(y) = \varphi_{-h}(y) + (-1)^{p+1} h^{p+1} C(y) + \mathcal{O}(h^{p+2}).$$

Then, using the group property of the flow, a right composition with φ_h gives

$$\phi_{-h}(\varphi_h(y)) = y + (-1)^{p+1}h^{p+1}C(\varphi_h(y)) + \mathcal{O}(h^{p+2}).$$

We compose now with $\phi_h^* = \phi_{-h}^{-1}$ and we apply its Taylor development:

$$\begin{aligned}
\phi_{-h}^{-1} \circ \phi_{-h} \circ \varphi_h(y) = \varphi_h(y) &= \phi_{-h}^{-1}(y + (-1)^{p+1}h^{p+1}C(\varphi_h(y)) + \mathcal{O}(h^{p+2}) \\
&= \phi_{-h}^{-1}(y) + \phi_{-h}^{'-1}(y)((-1)^{p+1}h^{p+1}C(\varphi_h(y))) + \mathcal{O}(h^{p+2}) \\
&= \phi_{-h}^{-1}(y) - (-1)^{p}h^{p+1}C(\varphi_h(y))\phi_{-h}^{'-1}(y) + \mathcal{O}(h^{p+2}).
\end{aligned}$$

We remark that

$$\varphi_h(y) = y + \mathcal{O}(h)$$

and a small compute gives

$$\phi_{-h}^{'-1}(y) = 1 + \mathcal{O}(h)$$

These two affirmations lead us to

$$C(\varphi_h(y))\phi_{-h}^{'-1}(y) = C(y + \mathcal{O}(h))(1 + \mathcal{O}(h)) = C(y) + \mathcal{O}(h)$$

and consequently to

$$\phi_h^*(y) = \varphi_h(y) + (-1)^{p}h^{p+1}C(y) + \mathcal{O}(h^{p+2}). \tag{14}$$

Comparing (13) and (14) we must have

$$(-1)^p = 1 \quad \text{or } C(y) = 0.$$

But, as p is odd, the first condition is not verified. Consequently, $C(y) = 0$ and then

$$\phi_h(y) = \varphi_h(y) + \mathcal{O}(h^{p+2}).$$

So the method is of order $p + 1$. □

3.2 Composition Methods of Effective Order $p \geq 3$

This is a particular kind of schemes. The advantage is two-fold: these methods are not so difficult to construct and we can obtain high order method without too much compute.

Definition 3.10
We say that the scheme

$$\psi_h = e^{b_m hB} e^{a_m hA} \dots e^{b_1 hB} e^{a_1 hA}$$

is of effective order p if a parametric map $\pi_h : \mathbb{R}^D \to \mathbb{R}^D$ exists for which the method $\hat{\psi}_h = \pi_h \circ \psi_h \circ \pi_h^{-1}$ is of order p, that is

$$\pi_h \circ \psi_h \circ \pi_h^{-1} = \varphi_h + \mathcal{O}(h^{p+1}).$$

The map π_h is called the post-processor.

Example 3.11
Consider problem (5) *with Lie-Trotting scheme $\psi_h = e^{hA}e^{hB}$ and the Strang splitting scheme $\phi_h = e^{\frac{h}{2}A}e^{hB}e^{\frac{h}{2}A}$. Then ψ_h is of effective order two with the post-processor $\pi_h = e^{-\frac{h}{2}A}$ because*

$$\phi_h = e^{-\frac{h}{2}A} \circ \psi_h \circ e^{\frac{h}{2}A}.$$

Remark 3.12
Because of:

$$y_n = \hat{\psi}_h(y_{n-1}) = (\hat{\psi}_h)^n(y_0) = (\pi_h \circ \psi_h \circ \pi_h^{-1})^n(y_0) = \pi_h \circ (\psi_h)^n \circ \pi_h^{-1}(y_0),$$

an implementation of y_n over n steps with constant step size h has the same computational efficiency as $\hat{\psi}_h$. Indeed, the compute of π_h^{-1} has only to be done at the beginning of the integration, and π_h has to be evalued only at output points.

Theorem 3.13
[2, pp 30-31] Consider the scheme

$$\psi_h = e^{b_1 hB} e^{a_1 hA} \dots e^{b_m hB} e^{a_m hA}.$$

At least one of the a_i as well as one of the b_i have to be negative in the scheme if ψ_h is of effective order $p \geq 3$.

Proof. Suppose that ψ_h is a scheme of effective order $p \geq 3$ with all a_i positive. Let us consider a post-processor π_h formally as the exact 1-flow of a vector field P_h. It is natural to choose the vector field as an element of the graded free Lie Algebra generated by A and B. We consider then

$$P_h = h(c_1 A + c_2 B) + \frac{h^2}{2} c_3 [A, B] + \mathcal{O}(h^3). \tag{15}$$

On the other hand, since

$$c_1 A + c_2 B = (c_2 - c_1)B + c_1(A + B) = (c_1 - c_2)A + c_2(A + B),$$

from (15), we can write

$$\pi_h = e^{P_h} = e^{hc_1(A+B)} e^{hcB} e^{\frac{h^2}{2} d_1 [A,B]} + \mathcal{O}(h^3) \tag{16}$$

and

$$\pi_h = e^{P_h} = e^{hc_2(A+B)} e^{-hcA} e^{\frac{h^2}{2} d_2 [A,B]} + \mathcal{O}(h^3), \tag{17}$$

where $c = c_2 - c_1$ and d_1, d_2 are parameters depending on c_1, c_2, c_3. Then, from (16) with $c_1 = 0$, we have

$$\pi_h = e^{P_h} = e^{hcB} e^{\frac{h^2}{2} d_1 [A,B]} + \mathcal{O}(h^3).$$

Consider $\hat{\psi}_h$ the scheme of order 3 such as

$$\hat{\psi}_h = \pi_h \circ \psi_h \circ \pi_h^{-1}.$$

This last relation leads us to

$$e^{-\frac{h^2}{2} d_1 [A,B]} e^{-hcB} \psi_h e^{hcB} e^{\frac{h^2}{2} d_1 [A,B]} = \hat{\psi}_h = e^{h(A+B)} + \mathcal{O}(h^4),$$

or equivalently

$$\bar{\psi}_h = e^{-hcB} \psi_h e^{hcB} = e^{\frac{h^2}{2} d_1 [A,B]} e^{h(A+B)} e^{-\frac{h^2}{2} d_1 [A,B]} + \mathcal{O}(h^4). \tag{18}$$

We can notice that all coefficients a_i are positive, since $\bar{\psi}_h$ is associated with the composition map

$$\bar{\psi}_h = e^{-chB} e^{a_1 hA} e^{b_1 hB} e^{a_2 hA} \dots e^{a_m hA} e^{b_m BhA} e^{chB}$$

which can be written as a composition of the first order method and its adjoint with coefficient $\bar{\beta}_i$.

$$\bar{\psi}_h = e^{\bar{\beta}_0 hA} e^{(\bar{\beta}_0+\bar{\beta}_1)hB} e^{(\bar{\beta}_1+\bar{\beta}_2)hA} \dots e^{(\bar{\beta}_{2k-2}+\bar{\beta}_{2k-1})hB} e^{(\bar{\beta}_{2k-1}+\bar{\beta}_{2k})hA} e^{\bar{\beta}_{2kh} B}$$

with $\bar{\beta}_0 = \bar{\beta_{2k}} = 0$. As previously, we obtained from (11) that a necessary condition for the scheme $\hat{\psi}_h$ to be of order $p \geq 3$ is

$$\hat{f_{3,1}} := \sum_{i=0}^{2k} \hat{\beta}_i^3 = 0$$

But, as we know from the proof of the theorem (3.6), this condition cannot be satisfied with all a_i positive. If we assume that all positive b_i, the same argument leads to the same condradiction. □

3.3 Composition Methods with all Coefficients Being Positive

[2, p.35] In this subsection, we will see that we can construct high order splitting scheme, but with all coefficient begin positive. It it not a contradiction with theorem (3.6) because we will compute flows with modified vector field, by applying lie bracket between them: splitting methods with modified potential.

3.3.1 A Method of Order Four

Consider now the second order differential equation

$$y'' = g(y, y')$$

which can be written in the form (4) by taking $x = (x_1, x_2) = (y, y')$ and

$$f_A(x) = (x_2, 0), \quad f_B(x) = (0, g(x_1, x_2))$$

Equivalently, we can also consider the vector fields

$$F_A \equiv x_2 \frac{\partial}{\partial x_2} \text{ and } F_B \equiv g(x_1, x_2) \frac{\partial}{\partial x_2}.$$

When $g(y, y') = g(y)$, this equation frequently appears in relevant problems arising in classical and quantum mechanics: there the operator F_A is related to the kinetic energy and F_B is associated with the potential energy. It is then possible to compute the flow corresponding to $F_C \equiv [F_B, [F_A, F_B]]$. Moroever, if $[F_B, F_C] = 0$, we know by theorem (2.11) that the corresponding flows commutes. Consequently, it makes sense to compute the flow $\varphi_{bh,ch^3}^{[B,C]}$ associated with the vector field $hbF_B + ch^3 F_C$. A composition of this flow with the standard flows of the problem can give a high order method, which is constructed with positive coefficients. A general method is given by

$$\psi_h = \varphi_{b_m h, c_m h^3}^{[B,C]} \circ \varphi_{a_m}^{[A]} \circ \dots \circ \varphi_{b_1 h, c_1 h^3}^{[B,C]} \circ \varphi_{a_1}^{[A]}, \tag{19}$$

and a particular case is the method

$$\psi_h = \varphi_{h/6}^{[B]} \circ \varphi_{h/2}^{[A]} \circ_{2h/3,h^3/72}^{[B,C]} \circ \varphi_{h/2}^{[A]} \circ \varphi_{h/6}^{[B]}. \tag{20}$$

Using BCH formula, we can see that this scheme is of order three, and it is not very difficult to observe that it is also symmetric. Proposition (3.9) implies then that it is of order four.

3.3.2 A Method of Effective Order Four

Let us consider the following scheme:

$$\psi_h = \varphi_{h/2}^{[A]} \circ \varphi_{h,h^3/24}^{[B,C]} \circ \varphi_{h/2}^{[A]}$$

We still suppose that we stay in a linear problem. It implies that ψ_h is given by

$$\psi_h = e^{\frac{h}{2}A} e^{hB + \frac{h^3}{24}[B,[A,B]]} e^{\frac{h}{2}A}$$

This is a symmetric method of order two, because his development with BCH formula gives:

$$\begin{aligned}
\psi_h &= e^{\dfrac{h}{2}A} e^{hB + \dfrac{h^3}{24}[B,[A,B]]} e^{\dfrac{h}{2}A} \\
&= e^{\dfrac{h}{2}A} e^{hB + \dfrac{h}{2}A + \dfrac{h^2}{4}[B,A] - \dfrac{h^3}{48}[A,[B,A]] + \mathcal{O}(h^4)} \\
&= e^{h(A+B) + \dfrac{h^3}{24}[A,[B,A]] - \dfrac{h^3}{24}[B,[A,B]]} + \mathcal{O}(h^4).
\end{aligned}$$

We want to show that this scheme is of effective order four. Consequently, we must find a post-processor π_h so that the method

$$\Phi_h = \pi_h \circ \psi_h \circ \pi_h^{-1}$$

becomes of order four. In order to avoid too much compute, we take a post processor such that the method Φ_h is symmetric:

$$\pi_h \circ \psi_h \circ \pi_h^{-1} = \Phi_h = \Phi_{-h}^{-1} = (\pi_{-h} \circ \psi_{-h} \circ \pi_{-h}^{-1})^{-1} = \pi_{-h} \circ \psi_h \circ \pi_{-h}^{-1}$$

which implies that the processor verifies $\pi_h = \pi_{-h}$ for all h. Consequently, if we show that Φ_h is a method of order three, then proposition (3.9) implies order four because it is symmetric. Suppose that $\pi_h = \exp\{ah^2[A,B]\}$. We develop now Φ_h with BCH formula:

$$\begin{aligned}
\Phi_h &= \pi_h \circ \psi_h \circ \pi_h^{-1} \\
&= e^{ah^2[A,B]} \circ e^{h(A+B) + \dfrac{h^3}{24}[A,[B,A]] - \dfrac{h^3}{24}[B,[A,B]]} \circ e^{-ah^2[A,B]} \\
&= e^{ah^2[A,B]} \circ e^{h(A+B) - ah^2[A,B] + h^3\left(\dfrac{12a+1}{24}\right)([A,[B,A]] - [B,[A,B]])} \\
&= e^{h(A+B) + h^3\left(\dfrac{24a+1}{24}\right)([A,[B,A]] - [B,[A,B]])} + \mathcal{O}(h^4)
\end{aligned}$$

In order to the coefficient of h^3 vanishes, we must choose $a = -\frac{1}{24}$ and the post processor is finally given by

$$\pi_h = \varphi^{[A,B]}_{\frac{h^2}{24}}$$

However, when in practice we compute the Lie Bracket $[A, B]$, this is often costly. So we must find a good approximation of this post processor, in other words we search for a_1, a_2, a_3, b_1 and for b_2 some constants such that the scheme

$$\kappa_h = e^{a_1 A} e^{b1B} e^{a2A} e^{b2B} e^{a_3 A}$$

approximes the post-processor π_h with order three and so that the method

$$\Psi_h = \kappa_h \circ \psi_h \circ \kappa_h^{-1}$$

stay symmetric up to order four. Using the BCH formula, we see that the coefficients must verify the following conditions:

1. $a_1 + a_2 + a_3 = 0$;

2. $b_1 = -b_2$;

3. $a_1 b_1 + a_2 b_2 - a_3 b_2 + a_1 b_2 - (a_2 + a_3) b_1 = -\frac{1}{24}$;

4. $\dfrac{a_1^2 b_2 a_2^2 b_2 + a_1^2 b_1}{12} + \dfrac{a_1 a_2 b_2 - a_3 a_2 b_2 - 2a_1 a_3 b_2}{4} = 0$;

5. $\dfrac{b_2^2 a_3 - b_2^2 a_2 + b_1(a_2 + a_3)}{12} + \dfrac{b_1 a_2 b_2 - b_1 a_3 b_2 - b_2^2 a_2}{4} = 0.$

This is a system of five equation with five variable. It is possible to solve it with for example MAPPLE or MATHEMATICA, and it gives a method of effective order four, with a post processor that is not so costly for the compute.

4 Application to Separable ODE

We will now study splitting methods and their orders on different separable equation. We begin in this section with a one order problem.

4.1 A Non-Linear Problem

Consider the following equation:

$$\begin{aligned} y'(t) &= -y^2(t) - y(t) \quad t \in [0, 1], \\ y(0) &= -\tfrac{1}{2}. \end{aligned} \tag{21}$$

This is a non linear ordinary separable differential equation and its exact solution is given by

$$y(t) = -\frac{e^{-t}}{1 + e^{-t}}.$$

4.2 Splitting

We split the function f : $[0,1] \times \mathbb{R} \to \mathbb{R}$, f$(t, y(t)) = -y^2(t) - y(t)$ into two functions:

$$\begin{aligned} y'(t) &= f_A(t) + f_B(t), \\ y(0) &= -\tfrac{1}{2}. \end{aligned} \tag{22}$$

where $f_A(t) = -y^2(t)$ and $f_B(t) = -y(t)$. We have then the associated vector fields

$$F_A(y) = -y^2 \frac{\partial}{\partial y} \text{ and } F_B(y) = -y \frac{\partial}{\partial y}.$$

In order to find the flow corresponding to each vector field, we solve the following differential equation

$$y'(t) = -y^2(t),$$

and

$$y'(t) = -y(t).$$

We obtain then the flows

$$\varphi_h^{[A]}(y) = \frac{y}{1+hy},$$

and

$$\varphi_h^{[B]}(y) = ye^{-h}.$$

We see that these two flows are well defined when h become small.

4.3 Implementation of Four Schemes

We will now construct the different schemes studied in the precedent section for the problem (22). Because we know explicit the flow for each solvable part, we will be able to obtain the well known schemes (Lie and Strang) and two high order methods.

4.3.1 Lie

The first order method (Lie-Trotter Splitting) applicated to equation (21) is

$$\psi_h(y) = \varphi_h^{[B]} \circ \varphi_h^{[A]}(y) = \frac{y}{hy+1} e^{-h}. \tag{23}$$

4.3.2 Strang

Similarly, the Strang scheme is written as

$$\psi_h(y) = \varphi_{\frac{h}{2}}^{[A]} \circ \varphi_h^{[B]} \circ \varphi_{\frac{h}{2}}^{[A]}(y). \tag{24}$$

4.3.3 The Method of Order Four

Problem (21) is not a second order differential equation, but we want to try the scheme studied in section 3.3.1 to see of which order it is applied to this problem. Remember, this scheme is given by

$$\psi_h = \varphi_{h/6}^{[B]} \circ \varphi_{h/2}^{[A]} \circ_{2h/3, h^3/72}^{[B,C]} \circ \varphi_{h/2}^{[A]} \circ \varphi_{h/6}^{[B]} \tag{25}$$

where F_A and F_B are two vector fields well chosen such that the flow associated to $F_C = [F_B, [F_A, F_B]]$ is computable and $[F_B, F_C] = 0$. We can consider the precedent splittings:

$$F_A(y) = -y^2 \frac{\partial}{\partial y} \quad F_B(y) = -y \frac{\partial}{\partial y},$$

but this one gives $F_C = [F_A, [F_B, F_A]] = -y^2 \frac{\partial}{\partial y}$ and

$$[F_B, F_C] = \Big[-y \frac{\partial}{\partial y}, -y^2 \frac{\partial}{\partial y} \Big] = y^2 \frac{\partial}{\partial y} \neq 0. \tag{26}$$

Therefore, (25) is not applicable. Nevertheless, if we permute the vector field F_A and F_B,

$$F_A(y) = -y \frac{\partial}{\partial y}, \quad F_B(y) = -y^2 \frac{\partial}{\partial y} \tag{27}$$

we have $F_C = 0$ and consequently $[F_B, F_C] = 0$. In this case, (25) becomes

$$\psi_h = \varphi^{[B]}_{h/6} \circ \varphi^{[A]}_{h/2} \circ \varphi^{[B]}_{2h/3} \circ \varphi^{[A]}_{h/2} \circ \varphi^{[B]}_{h/6}. \tag{28}$$

where now, because of the permutation of the vector field,

$$\varphi^{[B]}_h(y) = \frac{y}{1 + hy}, \quad \varphi^{[A]}_h(y) = ye^{-h}.$$

This method is at least of order two because it satisfies the first and the second order conditions of theorem (3.3). We will study later the real order of this method applied for this problem.

4.3.4 The Method of Effective Order Four

Taking the precedent splitting, we have $F_C = 0$, and then the scheme studied in section 2.3.4 becomes

$$\psi_h = \varphi^{[A]}_{h/2} \circ \varphi^{[B]}_h \circ \varphi^{[A]}_{h/2}.$$

which is nothing else than the Strang Scheme. To see if the Strang method is of effective order four, we will compute the post-processor π_h found in section 3.3.2, given by

$$\pi_h = \varphi^{[A,B]}_{\frac{h^2}{24}}.$$

Taking F_A and F_B define in (27), we find

$$[F_A, F_B] = -F_B = y^2 \frac{\partial}{\partial y}.$$

Consequently, we obtain

$$\varphi^{[A,B]}_{\frac{h^2}{24}}(y) = \varphi^{-[B]}_{\frac{h^2}{24}}(y) = \frac{y}{-y\frac{h^2}{24} + 1}.$$

and then the method is given by

$$\Phi_h = \pi_h \circ \psi_h \circ \pi_h^{-1} = \varphi^{[-B]}_{\frac{h^2}{24}} \circ \varphi^{[A]}_{\frac{h}{2}} \circ \varphi^{[B]}_h \circ \varphi^{[A]}_{\frac{h}{2}} \circ \varphi^{[-B]}_{\frac{-h^2}{24}}.$$

4.4 Numerical Approximation and Order Comparison

We numerically solve the problem (21) with the different methods. By splitting the interval $[0, 1]$ in N parts, the Lie scheme becomes for example the following iterative method:

$$\begin{aligned} y_0 &= -\frac{1}{2} \\ y_{k+1} &= \psi_h(y_k) = \frac{1}{h + \frac{1}{y_k}} e^{-h}. \end{aligned}$$

where $h = \dfrac{1}{N+1}$ is the time step. We resolve the Lie scheme on MATLAB and the result can be seen on figure 1.

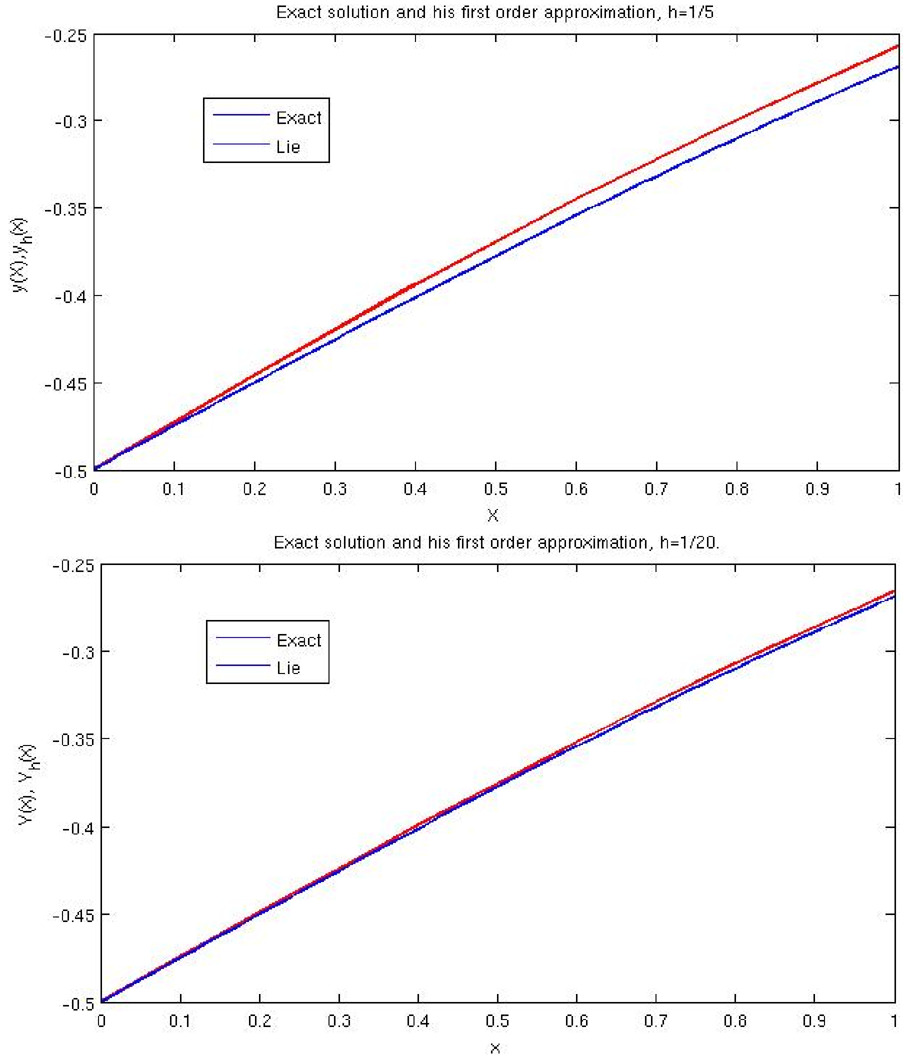

Figure 1: Lie scheme applicated to the separable differential equation $y' = -y - y^2$, with initial condition $y(0) = -\frac{1}{2}$ and two different value for h.

We see that the numerical solution converge quickly to the exact solution when h become smaller, as wanted. Nevertheless, what is the speed of convergence? In other words, what is the numerical order for the Lie method? To look at this interesting point, we compute the error

$$e(N) = y_N - y(1)$$

where $y(1)$ is the value of the exact solution at $t = 1$ and where y_N is the approximation of the solution at this point. For each $N \in \mathbb{N}, N \in [1, 100]$, we will look at the evolution of the error against the time step. On figure 2, we see that the Strang's error is clearly better than the Lie's one:

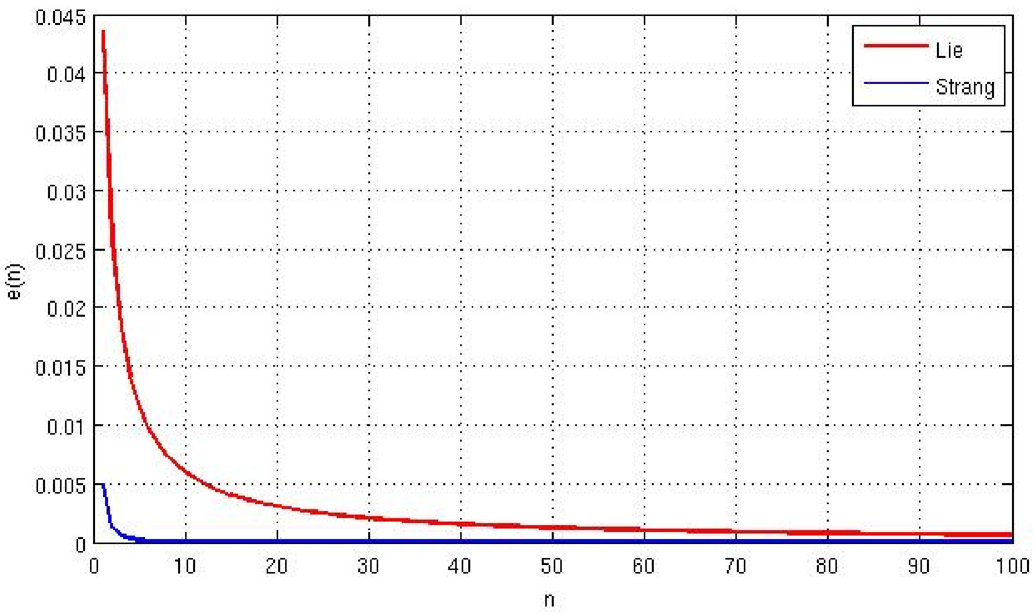

Figure 2: Plot of the the time step against the error for the Lie and Strang Schemes, applicated to the differential equation $y' = -y - y^2$, with initial condition $y(0) = -\frac{1}{2}$.

But actually we are interested in the order of such methods, and this plot is not sufficient to conclude. Remember, a method is of order p if

$$e(N) = y_N - y(1) = \mathcal{O}((1/N)^{p+1}).$$

Taking the logarithm on the two sides, we obtain

$$\log(e(N)) = \mathcal{O}((p+1)(\log(1/N))) \approx \mathrm{p}\log(1/N) + b$$

where the error b verifies

$$\mathrm{b} \approx c\frac{1}{N^p}.$$

Then the plot of n against the logarithm of the error will give a line. The slope of this line will gives us the order of our scheme. In figure 3, we can see the four lines corresponding to each of the studied methods.

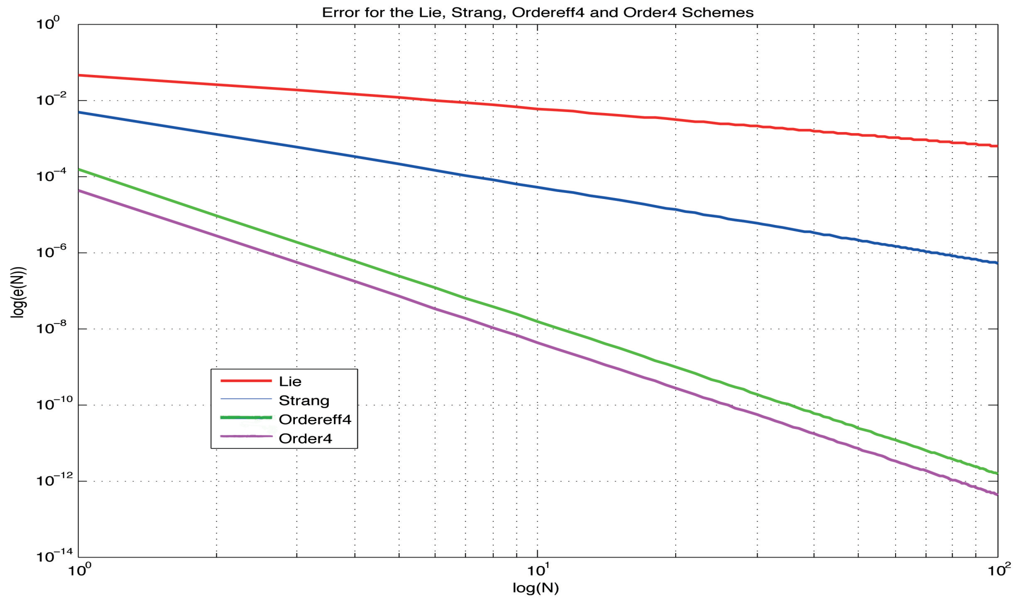

Figure 3: For the time step N increasing between 1 and 100: a logarithm plot of the error $y_N - y(1)$ for the Lie and Strang schemes and two high order methods on the differential equation , $y' = -y - y^2$, with initial condition $y(0) = -\frac{1}{2}$

By a linear regression on all the points, we compute on MATLAB the equations of these lines, which are on the form

$$y = ax + b$$

where a is the slope that represents the order of the method, and b is the error. All the results are summarized in the following array:

Methods	Equation of the line	Order
Lie	$0.9638 \log\left(\frac{1}{n}\right) - 2.9299$	1
Strang	$1.9972 \log\left(\frac{1}{n}\right) - 5.2743$	2
Efforder4	$3.9988 \log\left(\frac{1}{n}\right) - 8.8039$	4
Order4	$3.9985 \log\left(\frac{1}{n}\right) - 10.0568$	4

Figure 4: Equation of the line of the error of the precedent numerical schemes, applied to the differential problem $y' = -y - y^2$, $y(0) = -0.5$

We then see that the Lie, Strang and respectively Efforder4 schemes are of order one, two and respectively four, as wanted. We finally see that method (28), applied to this problem, is of order four too.

5 Application to the Damped Wave Equation

5.1 The Damped Wave Equation

Let us consider the following Partial Differential Equation (PDE): we search $u = u(x,t) : [0,1] \times \mathbb{R}_+ \to \mathbb{R}$ such that, for constant c and λ given,

$$\frac{\partial^2}{\partial t^2}u(x,t) = c^2 \frac{\partial^2}{\partial x^2}u(x,t) - \lambda \frac{\partial}{\partial t}u(x,t) \tag{29}$$

with the boundary condition (Dirichlet conditions),

$$u(0,t) = u(1,t) = 0, \quad t > 0$$

and the initial conditions:

$$\begin{aligned} u(x,0) &= h(x) \\ \frac{\partial}{\partial t}u(x,0) &= 0, \end{aligned}$$

where h is a sufficiently smooth function. This second order hyperbolic partial differential equation can describe the vibration of a string, which is fixed at its extremities and has a null speed at time $t = 0$. The constant c equal to the propagation speed of the wave and parameter λ is the damping coefficient.

5.2 Exact Solution of the Damped Wave Equation

If $\lambda < 4\pi c$ and in order to compute the exact solution of problem (29), we use separation of variable and Fourier series: suppose that u can be written as

$$u(x,t) = f(x)g(t).$$

Inserting this into the PDE we obtain

$$\frac{f''(x)}{f(x)} = -p_n^2 = \frac{g''(t) + \lambda g'(t)}{c^2 g(t)} \tag{30}$$

for some constants p_n. With the boundary conditions $f(0) = f(1) = 0$, by resolving the ODE

$$f''(x) + p_n^2 f(x) = 0$$

we find $p_n = n\pi$ and $f_n(x) = \sin(n\pi x)$. Left part of (30) gives

$$g''(t) + \lambda g'(t) + c^2 n^2 \pi^2 g(t) = 0,$$

which is a second order homogeneous differential equation with constant coefficients. Because $\lambda^2 - 4c^2\pi^2 < 0$, it implies that for all $n \in \mathbb{N}$, $\lambda^2 - 4n^2c^2\pi^2 < 0$ and then the solution is

$$g_n(t) = e^{\frac{-\lambda t}{2}}(a_n \sin(\mu_n t) + b_n \cos(\mu_n t))$$

where

$$\mu_n = \frac{\sqrt{4n^2\pi^2c^2 - \lambda^2}}{2}.$$

Consequently, a solution for problem (29) is given by

$$u_n(x,t) = \sin(n\pi x)e^{\frac{-\lambda t}{2}}(a_n \sin(\mu_n t) + b_n \cos(\mu_n t)).$$

By using the fact that a sum of solutions is still a solution and Fourier series, it can be shown that a candidate is the function

$$u(x,t) = e^{\frac{-\lambda t}{2}} \sum_{n=1}^{\infty} (\sin(n\pi x))(a_n \sin(\mu_n t) + b_n \cos(\mu_n t)) \tag{31}$$

where

1. $b_n = 2\int_0^1 \sin(n\pi x)h(x)dx$;
2. $a_n = \dfrac{\lambda}{\mu_n}\displaystyle\int_0^1 \sin(n\pi x)h(x)dx = \dfrac{b_n}{2\mu_n}$.

In fact, if we suppose that the function $h \in C^3([0,1])$, $h^{(4)}$ is piecewise continuous and $h(0) = h(1) = h''(0) = h''(1) = 0$, then the sequence

$$u_N(x,t) = e^{\frac{-\lambda t}{2}} \sum_{n=1}^{N} (\sin(n\pi x))(a_n \sin(\mu_n t) + b_n \cos(\mu_n t))$$

converges to the solution $u(x,t)$ of the problem for all $(x,t) \in [0,1] \times \mathbb{R}_+$ and

$$\lim_{t \to 0} u(x,t) = h(x)$$

uniformly. Consequently, this stamped wave equation problem is well posed because it has a unique solution given by (31). Moreover, if $\lambda > 0$, we easily see that

$$\lim_{t \to \infty} u(x,t) = 0.$$

We consider now the smooth function $h(x) = \sin(\pi x)$. A small compute gives $a_n = b_n = 0$ for all $n \neq 1$ which implies that the unique solution of the wave problem (29) for this initial periodic condition h is given by

$$u(x,t) = \sin(\pi x)e^{\frac{-\lambda t}{2}} \left(\frac{\lambda}{\sqrt{4\pi^2c^2 - \lambda^2}} \sin\left(\frac{\sqrt{4\pi^2c^2 - \lambda^2}}{2} t \right) + \cos\left(\frac{\sqrt{4\pi^2c^2 - \lambda^2}}{2} t \right) \right).$$

5.3 Numerical Approximation of the Damped Wave Equation

Let us take $u = u(x,t)$ and $v = v(x,t)$ with

$$v = \frac{\partial u}{\partial t}.$$

We write $U = (u,v)$ and define $F : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$F(U) = F(u,v) = \begin{pmatrix} v \\ c^2 \frac{\partial^2}{\partial x^2} u - \lambda v \end{pmatrix}.$$

The problem (29) can be now rewritten as

$$\frac{\partial U}{\partial t} = F(U),$$

in others words:

$$\begin{aligned} \frac{\partial u}{\partial t} &= v \\ \frac{\partial^2 u}{\partial t^2} &= c^2 \frac{\partial^2}{\partial x^2} u - \lambda v. \end{aligned} \tag{32}$$

with the boundary conditions

$$u(0,t) = u(1,t) = v(0,t) = v(1,t) = 0$$

and the initial conditions

$$u(x,0) = \sin(\pi x), v(x,0) = 0.$$

This partial differential equation can be transformed into a system of ordinary differential equation by discretizing the space. Indeed, we split $[0,1]$ into $N_x + 1$ subintervals of length

$$\delta_x = \frac{1}{N_x + 1}.$$

Then, $x_i = \frac{i}{N_x+1}$ and $u(x_i, t) \approx u_i(t), 0 \leq i \leq N_x + 1$. Since $u_0(t)$ and $u_{N_x+1}(t)$ are already known from the boundary conditions, the unknown functions becomes the vectors of size N_x:

$$\mathbf{u}(t) = (u_1(t), \ldots, u_{N_x}(t))^T \text{ and } \frac{\partial \mathbf{u}}{\partial t} = \mathbf{v}(t) = (v_1(t), \ldots, v_{N_x}(t))^T.$$

Moreover, if we use the finite difference for the approximation of the first and second derivative:

$$\begin{aligned} u'(x_i) &\approx \frac{u(x_i + \delta_x) - u(x_i - \delta_x)}{\delta_x} \\ u''(x_i) &\approx \frac{u'(x_i + \frac{\delta_x}{2}) - u'(x_i - \frac{\delta_x}{2})}{\delta_x}, \end{aligned}$$

then

$$\frac{\partial^2}{\partial x^2} u(x_i, t) \approx (A\mathbf{u}(t))_i,$$

where the matrix A of size $N_x \times N_x$ is given by

$$A = \frac{1}{\delta_x^2} \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & 1 & \ddots & \ddots \\ & & \ddots & \end{pmatrix}.$$

Consequently, the numerical approximation of the damped wave equation is given by the following system of ODE:

$$\mathbf{U}'(t) = G(\mathbf{U}(t)) \tag{33}$$

where $\mathbf{U}$'(t)=($\mathbf{u}$'(t),$\mathbf{v}$'(t)) and $G : \mathbb{R}^{2N_x} \to \mathbb{R}^{2N_x}$ is the function defined by

$$G(\mathbf{U}) = G(\mathbf{u}, \mathbf{v}) = \begin{pmatrix} \mathbf{v} \\ c^2 A\mathbf{u} - \lambda \mathbf{v} \end{pmatrix}.$$

5.4 Splitting the Damped Wave Equation

We now split the problem in order to apply the different schemes from the first chapter of this paper. We separe the function G into two function f_A and f_B as follows:

$$G(\mathbf{u},\mathbf{v}) = f_A(\mathbf{u},\mathbf{v}) + f_B(\mathbf{u},\mathbf{v}). \tag{34}$$

where $f_A(\mathbf{u},\mathbf{v}) = \begin{pmatrix} \mathbf{v} \\ -\lambda\mathbf{v}. \end{pmatrix}$ and $f_B(\mathbf{u},\mathbf{v}) = \begin{pmatrix} 0 \\ c^2 A\mathbf{u} \end{pmatrix}$. The first subproblem

$$\mathbf{U}'(h) = f_A(\mathbf{U})$$

has the flow

$$\psi_h^{[A]}(\mathbf{u},\mathbf{v}) = \begin{pmatrix} \mathbf{u} + \mathbf{v}\left(\frac{1-e^{-\lambda h}}{\lambda}\right) \\ \mathbf{v}e^{-\lambda h} \end{pmatrix}.$$

and the one from the second subproblem

$$\mathbf{U}'(t) = f_B(\mathbf{U})$$

is

$$\psi_h^{[B]}(\mathbf{u},\mathbf{v}) = \begin{pmatrix} \mathbf{u} \\ c^2 A\mathbf{u}h + \mathbf{v} \end{pmatrix}.$$

5.5 Implementation of the Different Schemes

Similarly to first order problem, we construct the four schemes studies in the thrid section of this report. Nevertheless, we will see that for high order methods this is a little more difficult.

5.5.1 Lie and Strang methods

The Lie Scheme for problem (33) is given by

$$\psi_h(\mathbf{u},\mathbf{v}) = \psi_h^{[B]} \circ \psi_h^{[A]}(\mathbf{u},\mathbf{v}) = \begin{pmatrix} \mathbf{u} + \mathbf{v}\left(\frac{1-e^{-\lambda h}}{\lambda}\right) \\ (c^2 A\mathbf{u}h + \mathbf{v})e^{-\lambda h} \end{pmatrix}, \tag{35}$$

and the strang method is given by

$$\phi_h(\mathbf{u},\mathbf{v}) = \psi_{\frac{h}{2}}^{[A]} \circ \psi_h^{[B]} \circ \psi_{\frac{h}{2}}^{[A]}(\mathbf{u},\mathbf{v}), \tag{36}$$

which is in fact the Leap-Frog method.

5.5.2 A method of Order Four

Which splitting must we choose to apply the scheme studied in section 3.2 given by

$$\psi_h = \varphi_{h/6}^{[B]} \circ \varphi_{h/2}^{[A]} \circ_{2h/3,h^3/72}^{[B,C]} \circ \varphi_{h/2}^{[A]} \circ \varphi_{h/6}^{[B]}? \tag{37}$$

We take

$$F_A = y_2\frac{\partial}{\partial y_1} - \lambda y_2\frac{\partial}{\partial y_2} \text{ and } F_B = c^2 A y_1\frac{\partial}{\partial y_2}.$$

These vectors fields mean that for all $1 \leq i \leq N_x$

$$\begin{aligned} \frac{\partial u_i}{\partial t} &= v_i \\ \frac{\partial v_i}{\partial t} &= c^2 (Au)_i - \lambda v_i \end{aligned} \tag{38}$$

and then the vector fields F_A and F_B become in composantes:

$$F_{Ai} = y_{2i} \frac{\partial}{\partial y_{1i}} - \lambda y_{2i} \frac{\partial}{\partial y_{2i}} \text{ and } F_{Bi} = c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}}.$$

We compute now the vector field

$$F_{Ci} = [F_{Bi}, [F_{Ai}, F_{Bi}]].$$

Notice that

$$\frac{\partial (Ay_1)_j}{\partial y_{1i}} = \frac{\partial \left(\sum_{k=1}^{N_x} A_{jk} y_{1k} \right)}{\partial y_{1i}} = A_{ji}.$$

Then, the compute gives:

$$\begin{aligned} [F_{Ai}, F_{Bi}] &= \left[y_{2i} \frac{\partial}{\partial y_{1i}} - \lambda y_{2i} \frac{\partial}{\partial y_{2i}}, c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}} \right] \\ &= y_{2i} \frac{\partial}{\partial y_{1i}} \left(c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}} \right) - c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}} \left(y_{2i} \frac{\partial}{\partial y_{1i}} \right) \\ &- \lambda y_{2i} \frac{\partial}{\partial y_{2i}} \left(c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}} \right) + c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}} \left(\lambda y_{2i} \frac{\partial}{\partial y_{2i}} \right) \\ &= y_{2i} c^2 A_{ii} \frac{\partial}{\partial y_{2i}} - c^2 (Ay_1)_i \frac{\partial}{\partial y_{1i}} + \lambda c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}} \\ &= \left(\lambda c^2 (Ay_1)_i + y_{2i} c^2 A_{ii} \right) \frac{\partial}{\partial y_{2i}} - c^2 (Ay_1)_i \frac{\partial}{\partial y_{1i}} \end{aligned}$$

$$\begin{aligned} F_{Ci} &= [F_{Bi}, [F_{Ai}, F_{Bi}]] = \left[c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}}, \left(\lambda c^2 (Ay_1)_i + y_{2i} c^2 A_{ii} \right) \frac{\partial}{\partial y_{2i}} \right] \\ &- \left[c^2 (Ay_1)_i \frac{\partial}{\partial y_{2i}}, c^2 (Ay_1)_i \frac{\partial}{\partial y_{1i}} \right] \\ &= 2c^4 A_{ii} (Ay_1)_i \frac{\partial}{\partial y_{2i}} = 2c^2 A_{ii} F_{B_i} \end{aligned}$$

Consequently $[F_{B_i}, F_{C_i}] = 0$ and then it makes sense to compute the flow associated to the vector field $(\frac{2}{3}h + 2\mathrm{c}^2 \mathrm{diag}(A) \frac{1}{72} h^3) F_B$, which is given by

$$\varphi^{[B,C]}_{\frac{2}{3}h, \frac{h^3}{72}}(\mathbf{u}, \mathbf{v}) = \begin{pmatrix} \mathbf{u} \\ \mathrm{c}^2 A \mathbf{u} (\frac{2h}{3} + 2\mathrm{c}^2 \mathrm{diag}(A) \frac{h^3}{72}) + \mathbf{v} \end{pmatrix}.$$

5.5.3 A Method of Effective Order Four

Let us again have a look on the method we define in section 3.3.2:

$$\psi_h = \varphi^{[A]}_{h/2} \circ \varphi^{[B,C]}_{h,h^3/24} \circ \varphi^{[A]}_{h/2}$$

We have seen that it is of effective order four if we compose it with the post-processor found in the section 3.2 given by

$$\pi_h = \psi^{[A,B]}_{\frac{h^2}{24}}$$

and its inverse

$$\pi_h^{-1} = \psi^{[A,B]}_{-\frac{h^2}{24}}.$$

The precedent compute gives

$$[F_A, F_B] = \Big(\lambda c^2(Ay_1) + y_2 c^2 \mathrm{diag}(A)\Big)\frac{\partial}{\partial y_2} - c^2(Ay_1)\frac{\partial}{\partial y_1}$$

Then, in order to find the flow associated to the vector field $\frac{-h^2}{24}[F_A, F_B]$, we must first solve the following system of ODE:

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} &= -c^2 A\mathbf{u} \\ \frac{\partial \mathbf{v}}{\partial t} &= \lambda c^2 A\mathbf{u} + 2c^2 \mathrm{diag}(A)\mathbf{v}. \end{aligned} \tag{39}$$

This system has the general solution

$$\begin{aligned} \mathbf{u} &= C_1 e^{-c^2 At} \\ \mathbf{v} &= C_2 e^{2c^2 \mathrm{diag}(A)t} + \int^t e^{2c^2\mathrm{diag}(A)x} \lambda c^2 A C_1 e^{-c^2 Ax} dx \end{aligned}$$

We obtain then the following flow:

$$\varphi^{[A,B]}_{\frac{-h^2}{24}}(\mathbf{u},\mathbf{v}) = \begin{pmatrix} e^{\frac{c^2h^2}{24}A}\mathbf{u} \\ 2e^{\frac{-c^2h^2}{24}\mathrm{diag}(A)}\mathbf{v} + \mathbf{u}\int^t e^{2c^2\mathrm{diag}(A)x}\lambda c^2 A e^{-c^2Ax} dx \end{pmatrix}.$$

But in general, it is difficult to compute the exponential of a matrix. On MATLAB, we can make it quite well, but the approximation of the integral is not easy. We do not have time for that, and consequently we stop here for the effective order four method.

5.5.4 Solving the Damped Wave Equation with the Lie Scheme and Different Values for the Damping Coeffficient

We look now at the approximation of the solution with the Lie method on MATLAB. On the following figures, we can see the computed solution on the time interval $[0, 10]$, split in N_t parts, which gives the time step

$$h = \frac{10}{N_t + 1}.$$

To have more precision and good plots, we take a high value for the parameters N_t and N_x. For the physical coefficient, we take $c = 0.8$ and different values for λ. We fortunately see on figure 5,6,7 that the solution is periodic if $\lambda = 0$ and on the other hand, the waves come to zero in a more speeder way when the value of λ increases, as desired. We remark that, for a T fixed, we could not take all the value of N_x and N_t we want. Indeed, the Lie method is of effective order two with the Strang method, which is a Leap Frog method. But these methods are not absolutely stable, and once we choose the value of N_x, we must take a high value for N_t. More precisely, when $T = 1$, we must choose $N_t \geq N_x$, when $T = 10$, the stability condition is about $N_t > 10N_x$. Nevertheless, we will not go into details here.

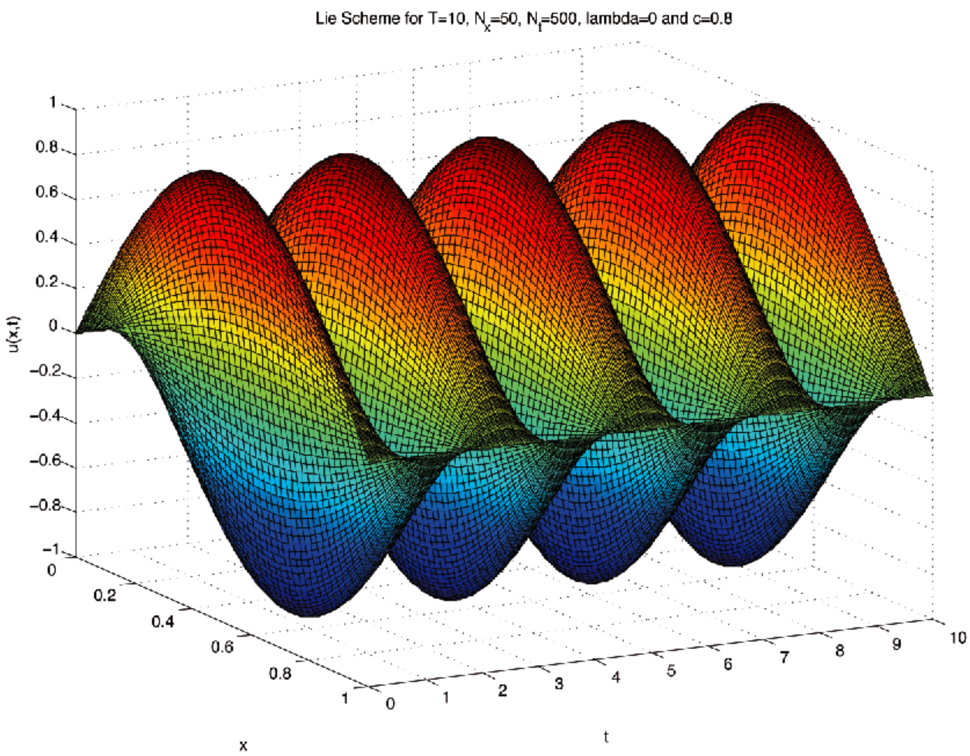

Figure 5: Lie Scheme for the damped wave equation on $[0, 10]$, with $N_x = 50$, $N_t = 500$, $\lambda = 0$ and $c = 0.8$. This is a periodic function because there is no damping effect.

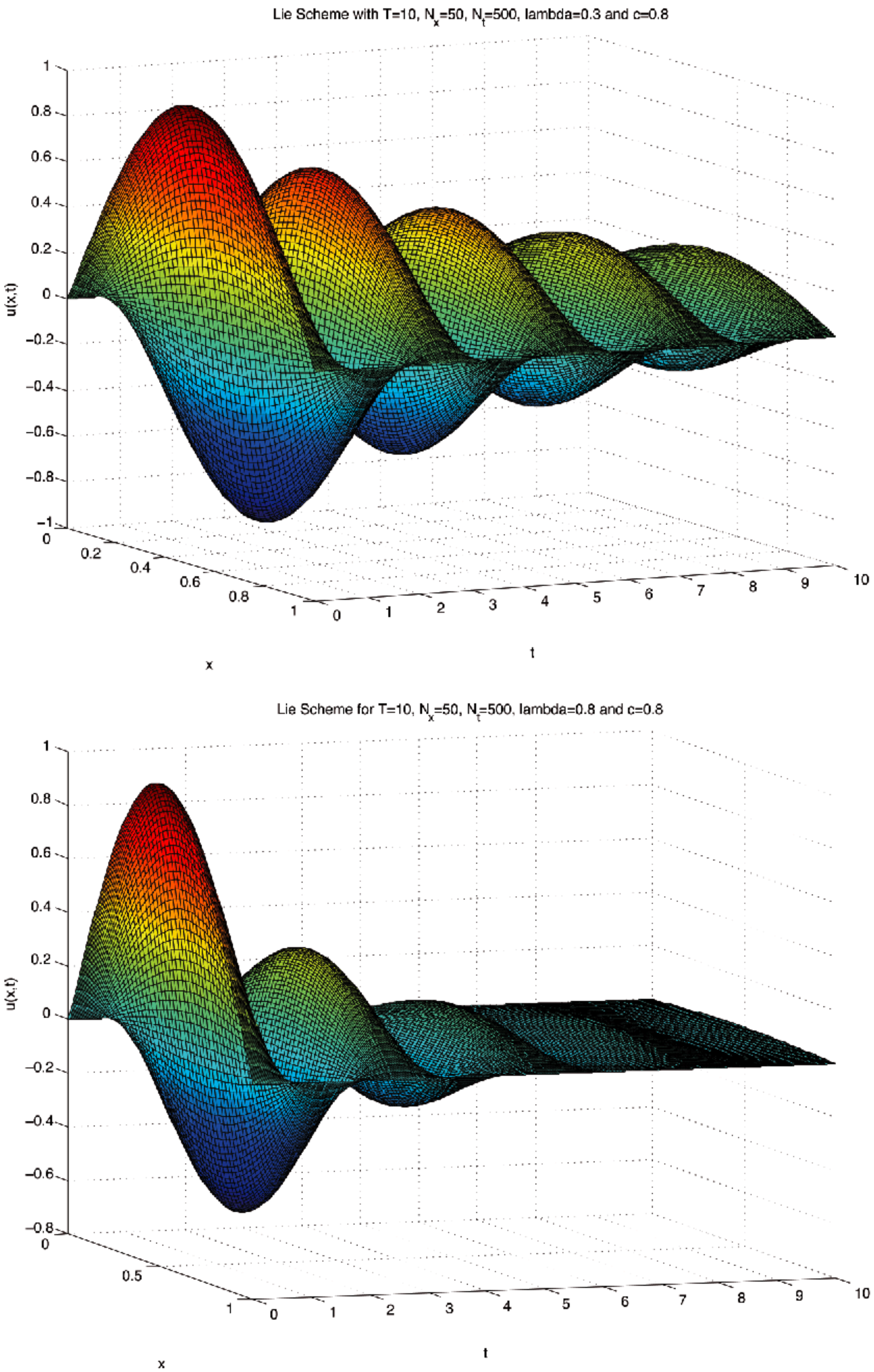

Figure 6: Lie Scheme for the damped wave equation on $[0, 10]$, with $N_x = 50$, $N_t = 500$, and $c = 0.8$. The waves become more smaller and the function goes faster to zero when the value of λ is increasing, here for $\lambda = 0.3$ and 0.8

5.5.5 Error of the schemes

We now look at the errors for the different schemes. We consider two ways. The first one is to compute the error with the $\mathcal{L}^2$ norm: for a space step $n \in \mathbb{N}$ given, we have

$$\mathrm{E}_2(t,n) = \frac{1}{n}\left(\sum_{j=1}^{n}(u(x_j,t) - u_j(x_j,t))^2\right)^{\frac{1}{2}}$$

The second way is to use the $\mathcal{L}^\infty$ norm, which for a space step $n \in \mathbb{N}$ given is defined by

$$\mathrm{E}_\infty(t,n) = \max_{1\le j\le n} |u(x_j,t) - u_j(x_j,t)|$$

We similarly procede to the section 4.4. For all $n_t \in [1, N_t]$, we take $T = 1$ and $n_t = n_x$, because we know that the schemes are stable in this case. Then, we compute for all n_t the $\mathcal{L}^\infty$ error between exact and numerical approximation at time $T = 1$. We plot the logarithm of n_t against

$$\log(\mathrm{E}_\infty(n_x,1)) = \log(\max_{1\le j\le n_x} |u(x_j,1) - u_j(x_j,n_t)|)$$

The result can be seen on figure for the Lie and Strang Scheme. The error is not as good as the precedent example, but nevertheless acceptable.

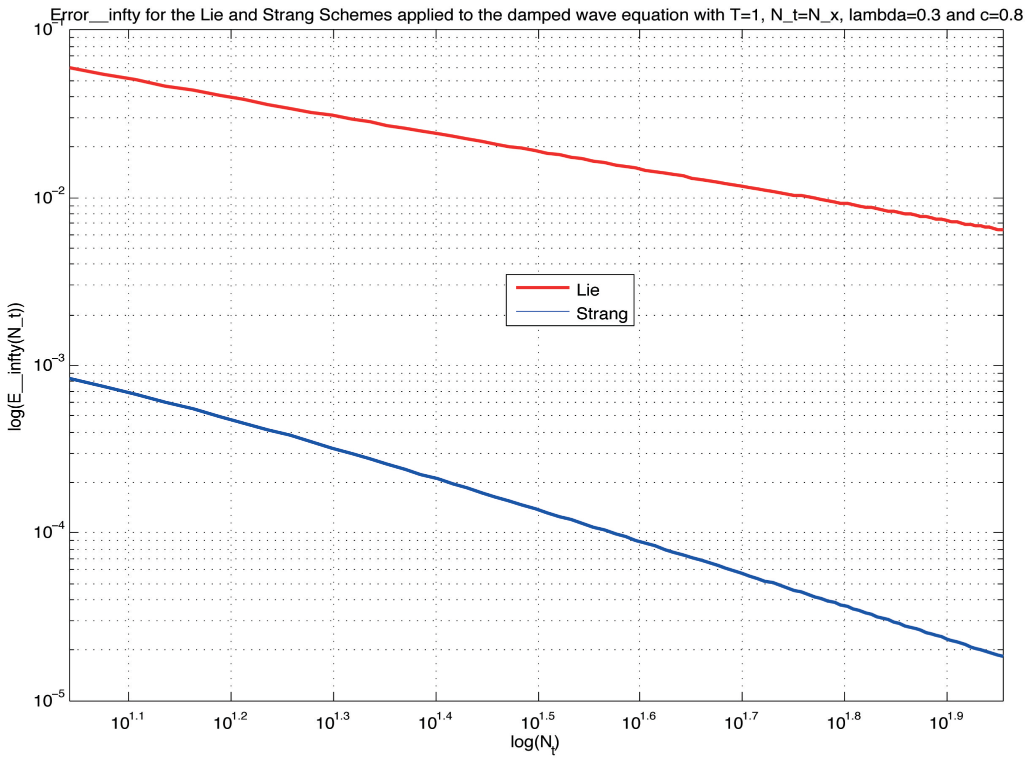

Figure 7: Error for the Strang and Lie Scheme applied to the damped wave equation, with $N_x = N_t$, N_t goes from 11 to 100, $\lambda = 0.3$ and $c = 0.8$

We compute the slope of these lines with help of a linear regression and we obtain

Methods	Equation of the line	Order
Lie	$0.9525\log\left(\frac{1}{n}\right) - 0.3379$	1
Strang	$1.8772\log\left(\frac{1}{n}\right) - 2.4617$	2

Figure 8: Equation of the line of the error of the precedent numerical schemes, applied to the damped wave equation

We obtain the good order for the Lie and Strang method, as desired.

5.5.6 Conservation of Energy

The energy of the damped wave equation is given by

$$E(t) = \frac{1}{2}\Big(\int_\Omega \Big(\Big(\frac{\partial u}{\partial t}\Big)^2 + c^2\Big(\frac{\partial u}{\partial x}\Big)^2\Big)dx\Big)$$

For our problem, we have $\Omega = [0,1]$. When $\lambda = 0$, a compute shows that

$$\frac{d}{dt}E(t) = 0.$$

and then $E(t) = E(0) = \frac{\pi^2 c^2}{4}$.Nevertheless, if we take a positive value for λ, we remark that

$$\lim_{t\to\infty}\frac{\partial u}{\partial t} = \lim_{t\to\infty}\frac{\partial u}{\partial x} = 0$$

because the function $e^{-\frac{\lambda t}{2}}$ from the exact solution that goes to zero and the other one are all bounded. And then, using the fact that the above functions are smooth, we obtain

$$\begin{aligned}\lim_{t\to\infty}E(t) &= \lim_{t\to\infty}\frac{1}{2}\Big(\int_\Omega \Big(\Big(\frac{\partial u}{\partial t}\Big)^2 + c^2\Big(\frac{\partial u}{\partial x}\Big)^2\Big)dx\Big)\\ &= \frac{1}{2}\Big(\int_\Omega \Big(\Big(\lim_{t\to\infty}\frac{\partial u}{\partial t}\Big)^2 + c^2\Big(\lim_{t\to\infty}\frac{\partial u}{\partial x}\Big)^2\Big)dx\Big) = 0\end{aligned}$$

because we can permute limit and integral in this case. Consequently, the energy is conserved when $\lambda = 0$ and is decreasing in the other cases. We will verifie now the conservation on our numerical solutions. In order to compute the energy, we will use the trapezoidal method to integrate the functions

$$\left(\frac{\partial u}{\partial t}\right)^2, \quad \left(\frac{\partial u}{\partial x}\right)^2 : [0,1] \to \mathbb{R}.$$

Recall that the interval $[0,1]$ is split in $N_x + 1$ parts and

$$\delta_x = \frac{1}{N_x + 1}$$

We take the first order interpolation polynome from these functions, with nodes $f(x_i)$, where $x_i = \delta_x i$, $1 \le i \le N_x + 1$. For any fixed $t \in [0,T]$, the nodes for the function $\frac{\partial u}{\partial t}^2$ are given by

$$\mathbf{v}(t) = (v_0(t), v_1(t), \ldots, v_{N_x}(t), v_{N_x+1}(t))$$

which is nothing else than the vector defined in the space discretization of the problem. For the approximation of $\left(\frac{\partial u}{\partial x}\right)^2$, we will use the finite center difference: for all $2 \leq i \leq N_x - 1$

$$\frac{\partial u_i}{\partial x} \approx \frac{u_{i+1} - u_{i-1}}{2\delta_x}.$$

For the missing values (because, for example, we cannot use the above approximation for $\frac{\partial u_{N_x+1}}{\partial x}$ as u_{N_x+2} is not defined), we simply use the first order finite difference approximation:

$$\frac{\partial u_i}{\partial x} \approx \frac{u_{i+1} - u_i}{\delta_x}, i = 0, 1 \quad \text{and} \quad \frac{\partial u_j}{\partial x} \approx \frac{u_j - u_{j-1}}{\delta_x}, j = N_x, N_x + 1.$$

This is a second order approximation. Consequently, we can now compute the integrals with the trapezoidal method:

$$\int_\Omega \Big(\frac{\partial u}{\partial t}\Big)^2 dx \approx \delta_x \left(\frac{1}{2} v_0^2 + v_1^2 + \cdots + v_{N_x}^2 + \frac{1}{2} v_{N_x+1}^2\right)$$
$$\int_\Omega \Big(\frac{\partial u}{\partial x}\Big)^2 dx \approx \frac{1}{\delta_x} \left(\frac{1}{2}(u_1 - u_0)^2 + \frac{(u_3 - u_1)^2}{4} + \dots\right)$$
$$+ \quad \frac{1}{\delta_x} \left(\frac{(u_{N_x} - u_{N_x-2})^2}{4} + \frac{1}{2}(u_{N_x+1} - u_{N_x})^2\right)$$

and obtain then a numerical way to compute the Energy for this system:

$$\begin{aligned} E(t_n) &= \frac{1}{2}\Big(\int_\Omega \Big(\Big(\frac{\partial u(x,t_n)}{\partial t}\Big)^2 + c^2 \Big(\frac{\partial u(x,t_n)}{\partial x}\Big)^2\Big) dx\Big) \\ &\approx \frac{1}{2}\delta_x \left(\frac{1}{2} v_0^2 + v_1^2 + \cdots + v_{N_x}^2 + \frac{1}{2} v_{N_x+1}^2\right) \\ &+ \frac{c^2}{2}\frac{1}{\delta_x} \left(\frac{1}{2}(u_1 - u_0)^2 + \frac{(u_3 - u_1)^2}{4} + \dots\right) \\ &+ \frac{c^2}{2}\frac{1}{\delta_x} \left(\frac{(u_{N_x} - u_{N_x-2})^2}{4} + \frac{1}{2}(u_{N_x+1} - u_{N_x})^2\right). \end{aligned}$$

We can see results on the following figure:

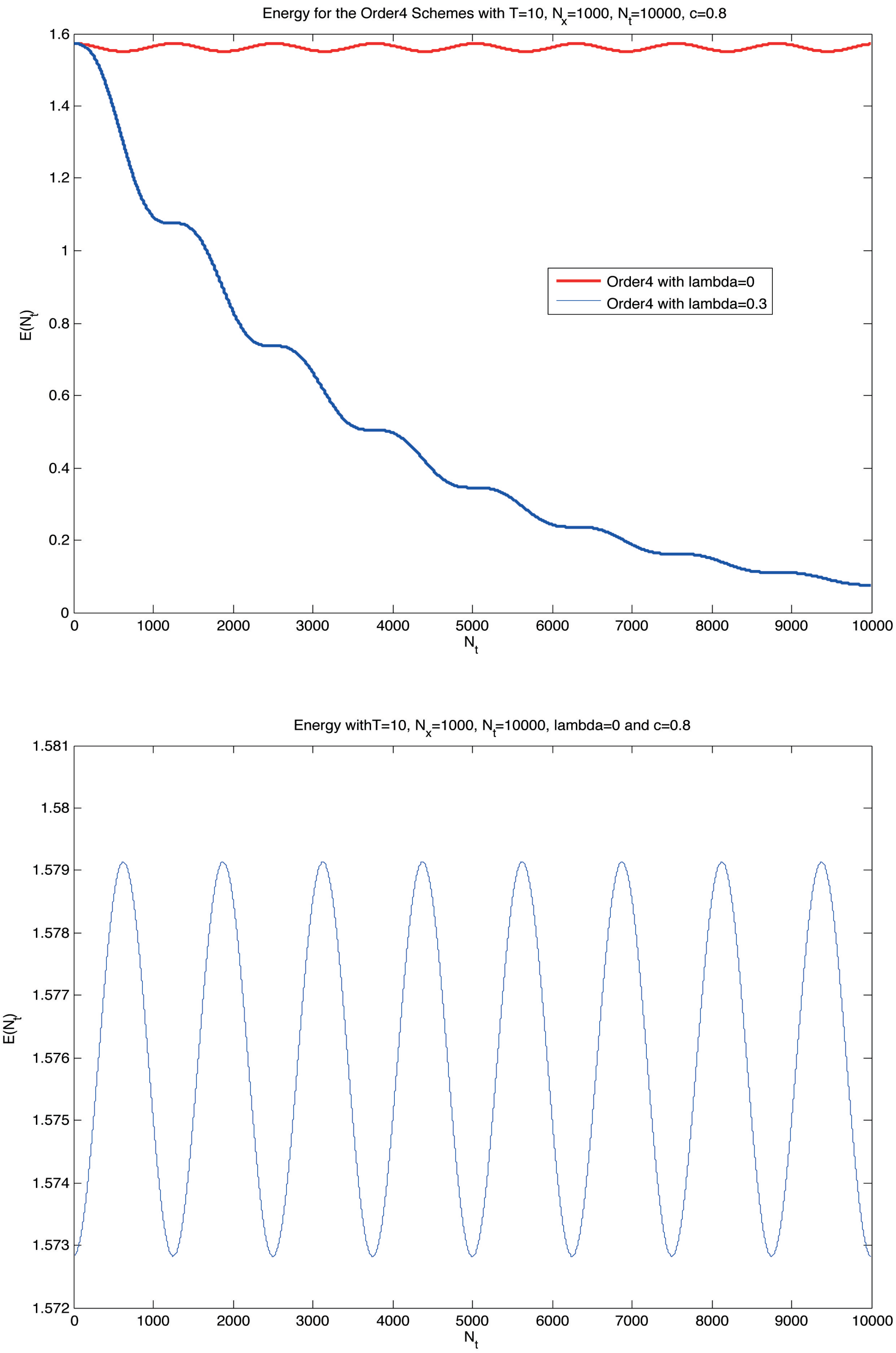

Figure 9: Energy for the damped wave equation, computed by the Order_4 method on $[0, 10]$, with $N_x = 1000$, $N_t = 10000$, $c = 0.8$ and different values of λ. We see that the energy is conserved near the initial value for $\lambda = 0$ and goes to 0 when $\lambda = 0.3$.

We clearly see that we have a conservation of the Energy when $\lambda = 0$. When

$c = 0.8$, the small oscillations that stay around $E(0) = \frac{c^2\pi^2}{4} \approx 1.579$ are the error of the numerical methods, but they are reasonnable. On the other hand, the energy of the system is clearly decreasing when $\lambda > 0$, which is not a surprise because this is the damping coefficient. The value of c will influence the number of waves. On figure 10, we see that if we increase the value of c and λ, we obtain a curve that has less waves than before but goes more faster to zero than the precedent one:

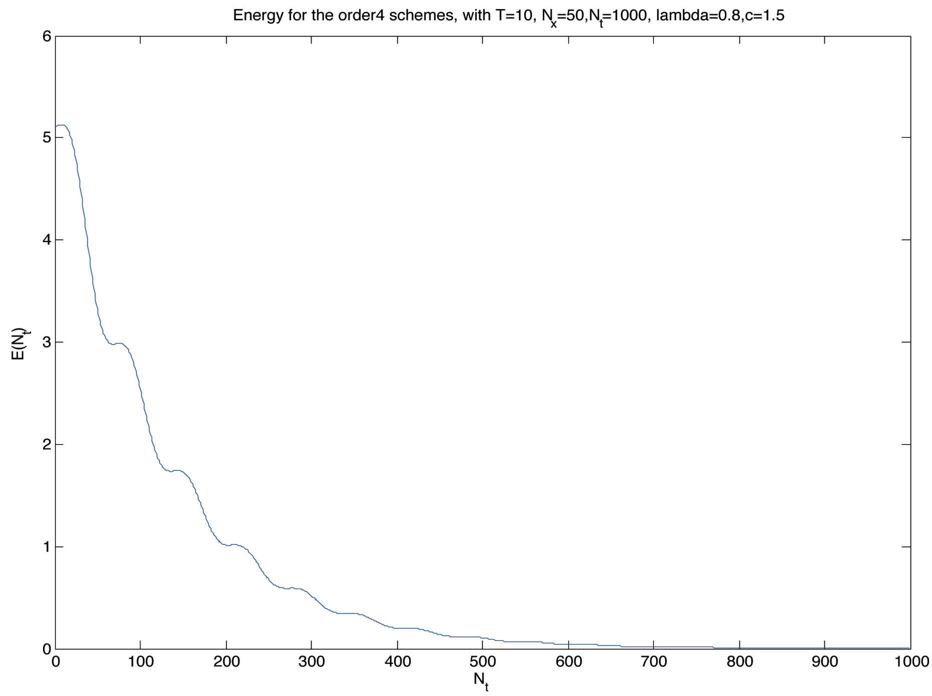

Figure 10: Energy for the damped wave equation, with $N_x = 50$, $N_t = 1000$, $\lambda = 0.8$ and $c = 1.5$. The Energy goes to zero in a more faster way than the precedent examples

6 Conclusion

In these two examples, we have seen that Splitting Methods are very useful schemes to solve numerically ordinary and partial differential equation. For the non linear ordinary equation we have studied, it was not so difficult to obtain method of order four and effective order four. Nevertheless, for the damped wave equation, Lie and Strang Schemes are quickly constructed but high order method are more difficult to develop, because we sometimes need to make a lot of compute. But in all the cases, this is an elegant way of doing numerical analysis!

References

[1] A.Abdulle *Numerical Integration of Dynamical System.* Course MA ma1, EPFL, 2010

[2] S. Blanes, F.Casas *On the necessity of negative coefficient for operator splitting schemes of order higher than two.* Article, Elsevier, 2004.

[3] E.Hairer, C.Lubich, G.Wanner *Geometric Numerical Integration.* Second Edition, Springer Berlin 2006.

[4] E.Hairer, G.Wanner *Solving Ordinary Differential Equation II - Stiff and Differential-Algebraic Problems.* Second Revised Edition, Springer Berlin 1996.

[5] John M. Lee *Introduction to Smooth Manifolds.* Springer, New York, 2003.

[6] A. Quarteroni, Riccardo Sacco, Fausto Saleri *Numerical Mathematics.* Second Edition, Springer Berlin 2007.

7 MATLAB code for tests on the Damped Wave Equation

The following code is portion of what you can see in the fifth part of this work. All the text preceded by "%" explain what made the code (functions, plots) with reference to the third part of this work for the construction of the Splitting Methods.

7.1 The different flows used to construct Splitting Methods

```
%Compute the flow associated to the vector field
%F_A= (v,-lambda v) for the
%damped wave equation
function [fu,fv]=flowA(u,v,lambda,h)

for i=1:length(u)
     if lambda > 0
         %well defined because the limit exist when
         %lambda become smaller.
     fu(i)=u(i)+v(i)*h*(1-exp(-lambda*h))./(lambda*h);
     fv(i)=v(i)*exp(-lambda*h);
     else
         %Lambda=0 case
     fu(i)=u(i)+v(i)*h;
     fv(i)=v(i)*exp(-lambda*h);
     end
end

%Compute the flow associated to the vector field
%F_B= (0,c^2Au) for the
%damped wave equation
function [fu,fv]=flowB(u,v,c,h)
n=length(u);
%Definition of the Laplacien matrix
A=(n+1)^2*(-1*2*diag(ones(n,1),0)+diag(ones(n-1,1),1)
+ diag(ones(n-1,1),-1));
fu=zeros(1,n);
fv=zeros(1,n);
    fu=u';
    fv=(c*c)*h*A*(u')+v';

    %Compute the flow associated to the vector field
    %[F_B,F_C]= (2/3h+2c^2
%diag(A) 1/72 h^3)F_B for the
%damped wave equation
function [fu,fv]=flowBC(u,v,c,h1,h2)
n=length(u);
%Laplacien Matrix
A=((n+1)^2)*(-1*2*diag(ones(n,1),0)+diag(ones(n-1,1),1)
+ diag(ones(n-1,1),-1));
```

```
fu=zeros(1,n);
fv=zeros(1,n);
for i=1:n
    fu(i)=u(i);
    fv(i)=(c*c)*A(i,1:n)*u'*h1
                 +2*c^4*h2*A(i,i)*A(i,1:n)*u'+v(i);
end
```

7.2 Implementation of Order$_4$ Scheme

```
%This function returns two matrix computed with the
%Order_4 Scheme.
%The first is the approximation of
%the solution and the second one is the approximation
%of the derivative of the
%solution by t.
function [vecu,vecv]=solexact(T,Nx,Nt,lambda,c)
%Because we split [0,1] in N_x+1 parts, we have
%N_x+2 nodes for the space
n=Nx+2;
%Because we split [0,T] in N_t parts, we have
%N_t+2 nodes for the time
m=Nt+1;
%Time Step
h=T./(Nt);

%Definition of the two matrix
Mv=zeros(n,m); % Approximation of the Solution
Mu=zeros(n,m); % Approximation of the derivative of the
%solution by t

%Initial conditions for u: u(0,t)=u(1,t)=0
Mu(1,1:m)=0;
Mu(n,1:m)=0;

%Initial conditions for du/dt: du/dt(0,t)=du/dt(1,t)=0
Mv(1,1:m)=0;
Mv(n,1:m)=0;

%Condition. u(x,0)= sin(pi*x) and du/dt(x,0)=0;
for i=1:n
Mu(i,1)=sin(pi*(i-1)./(Nx+1));
Mv(i,1)=0;
%The first columns of the two matrix Mu and Mv
%correspond to the initialconditions. The second
%columns correspond to the solution at time t=h. The
%third correspond to the solution at time t=2h, etc.
%We compute each column of these matrix
%with the Order_4 Scheme.
for j=1:m-1
```

```
    [fuB1,fvB1]=flowB(Mu(2:n-1,j)',Mv(2:n-1,j)',c,h./6);
    [fuA1,fvA1]=flowA(fuB1,fvB1,lambda,h./2);
    [fuBC,fvBC]=flowBC(fuA1,fvA1,c,2*h/3,(h^3)./72);
    [fuA2,fvA2]=flowA(fuBC,fvBC,lambda,h./2);
    [Mu(2:n-1,j+1),Mv(2:n-1,j+1)]=flowB(fuA2,fvA2,c,h./6);
end
end
%Defines the matrix with nodes of [0,1] times [0,T].
x=linspace(0,1,Nx+2);
t=linspace(0,T,Nt+1);
Mx=repmat(x',1,Nt+1);
Mt=repmat(t,Nx+2,1);
x1=linspace(0,1,102);
t1=linspace(0,T,101);
Mx1=repmat(x1',1,101);
Mt1=repmat(t1,102,1);
%We evalue the exact solution on these nodes.
Mexact=solexact(c,lambda,x1,t1);
%Plot of the approximation and the exact solution.
surf(Mx1,Mt1,Mexact);
hold on
surf(Mx,Mt,Mu);
```

7.3 Compute of the Error for the Strang Method

```
%This function returns the L^2 error of the Strang
%Scheme when computed on the damped wave equation.
function[e]=ErrorL2Strang(N,M,T,lambda,c)
%We compute for the chosen time and
%space steps N respectively M.
[Mu,Mv]=strangwave(T,N,M,lambda,c);
%We define the nodes to compute
%the exact solution on them.
x1=linspace(0,1,N+2);
t1=linspace(0,T,M+1);
%Compute of the exact solution
[fu]=solexact(c,lambda,x1,t1);
%The error vector
e1=zeros(1,N+1);
%We compute the error with the (L2 norm)^2
for i=1:N+2
e1(1,i)=1/(N+2)*(Mu(i,M+1)-fu(i,M+1))^2;
end
%After having additioned all the terms of e1,
%we finally take the root square
e2=sum(e1);
e=sqrt(e2);

%This function returns the L_inftyerror of the Strang
%Scheme when computed on the damped wave equation.
```

```
function[e]=ErrorLinfStrang(N,lambda,c)
%Compute of the exact solution on some nodes
x=linspace(0,1,N+2);
t=linspace(0,1,N+1);
[fu]=solexact(c,lambda,x,t);
%Approximation of the solution with Strang Scheme
[Mu,Mv]=strangwave(1,N,N,lambda,c);
%Definition of the Error
e1=zeros(1,N+2);
%We take the L^infty norm
for i=1:N+2
e1(i)=sqrt((Mu(i,N+1)-fu(i,N+1))^2);
end
double e;
e=e1(1);
%It choose the biggest element of e1
for i=1:N+2
if e1(i)>e
    e=e1(i);
end
end

% We compute the L^2 or L^{infty} Error
%for the three Schemes
%for a time step N_t \in [11,M].
function ErrorPDE(T,M,lambda,c)
%Definition of the three error vectors
e1=zeros(1,M);
e2=zeros(1,M);
e3=zeros(1,M);
%COmpute of the error with the different norm
%and the different schemes
for m=1:M
    % e1(1,m)=ErrorL2Lie(m,m,T,lambda,c);
    %e2(1,m)=ErrorL2Strang(m,m,T,lambda,c);
    %e3(1,m)=ErrorL2order_4(m,m,T,lambda,c);
     e1(1,m)=ErrorLinfLie(m,lambda,c);
     e2(1,m)=ErrorLinfStrang(m,lambda,c);
     e3(1,m)=ErrorLinfOrder_4(m,lambda,c);
end;
%Define the nodes between [11,M].
b=linspace(11,M-10,M-10)
%Plot of the logarithm of the Error for each methods.
loglog(b,e1(11:M));
set(loglog(b,e1(11:M)),'Color','red','LineWidth',2)
hold on
loglog(b,e2(11:M));
set(loglog(b,e2(11:M)),'Color','blue','LineWidth',2)
hold on
loglog(b,e3(11:M));
```

```
set(loglog(b,e3(11:M)),'Color','green','LineWidth',2)
legend('Lie','Strang')
%Put a grid on the plot
grid on

%COmpute the slope of each line with a linear regression
polyfit(log(b),log(e1(11:M)),1)
polyfit(log(b),log(e2(11:M)),1)
polyfit(log(b),log(e3(11:M)),1)
```

7.4 Conservation of the Energy

```
%This function returns a vector of Energy, where
%elements are the energy at the fixed t.
function[Energ]=Energy(N,lambda,c)
%Definition of the vector Energy
Energ=zeros(1,10*N-1);
%Vector to approxime part of the first integral
int1_1=zeros(1,N-2);
%Approximation of the first integral
int1=zeros(1,10*N-1);
%Vector to approxime part of the second integral
int2_1=zeros(1,N-2);
%Approximation of the second integral
int2=zeros(1,10*N-1);
%Computation of a numerical solution with the order_4
%method and a time steps about ten times the space step.
[Mu,Mv]=order_4(10,N,10*N-1,lambda,c);

% For each j, we compute the nodes of the functions
%(du/dt)^2 and (du/dx)^2 as seen in section 5.5.6,
%and then compute an approximation of
% the integral with trapezoidal method.
for j=1:10*N-1
for i=2:N-1
int1_1(i)=Mv(i,j)^2;
int2_2(i)=(Mu(i+1,j)-Mu(i-1,j))^2;
end
int1(j)=sum(int1_1)+0.5*Mv(1,j)^2
+Mv(N+1,j)^2+0.5*Mv(N+2,j)^2;
int2(j)=0.25*sum(int2_2)
+0.5*(Mu(2,j)-Mu(1,j))^2+0.5*(Mu(N+2,j)-Mu(N+1,j))^2;
%Compute the Energy for each j
Energ(j)=0.5*c*c*(N+1)*int2(j)+0.5*(1/(N+1))*int1(j);
end
%Plot of the Energy against the time
b=linspace(1,10*N-1,10*N-1);
plot(b,Energ);

N=1000;
```

```
%We compute the Energy with lambda=0
[A]=Energy(N,0,0.8);
%We compute the Energy with lambda>0
[B]=Energy(N,0.3,0.8);
%We split [1,10000]
b=linspace(1,10*N-1,10*N-1);
%Plot of the two Energy against the time steps
plot(b,A);
set(plot(b,A),'Color','red','LineWidth',2)
hold on
plot(b,B);
set(plot(b,B),'Color','blue','LineWidth',2)
legend('Order4 with lambda=0','Order4 with lambda=0.3')
```