Lecture Notes - Course 5 - M. Massot - L. Séries

MAP 551 - Systèmes dynamiques pour la modélisation et la simulation des milieux réactifs multi-échelles (2018-2019)

Chapter 1 refers to time operator splitting techniques to numerically integrate time dependent PDEs. The review on these schemes is not exhaustive but aims at giving sufficient information on the theoretical characterization of splitting methods and some important issues often encountered in the numerical solution of stiff problems. The reader may refer to the book of Hundsdorfer & Verwer (Hundsdorfer and Verwer 2003) for further details on different types of splitting technique.

Chapter 2 deals with the time integration of stiff ODEs by one-step Runge-Kutta schemes. This description complements the previous chapter and gives a more detailed insight into the numerical solution of stiff problems. In particular, we focus on Runge-Kutta methods given by implicit and stabilized explicit techniques. A complete information can be found in the book of Hairer & Wanner (Hairer and Wanner 1996).

Chapter 1

Time Operator Splitting for Multi-Scale Evolutionary PDEs

We are concerned with the numerical solution of time dependent PDEs involving reactive terms and transport operators such as diffusion or both, issued from the mathematical modeling of general multi-scale phenomena. This kind of problem is rather common in many applications so that efficient solution schemes are of the utmost importance. In this chapter, our attention will be focused on the so-called time operator splitting methods for the numerical solution of such problems. A time operator splitting procedure allows us to consider dedicated solvers for the reaction part which is numerically decoupled from the other physical phenomena like convection, diffusion or others, for which there also exist dedicated numerical methods. A completely independent optimization of the solution of each subsystem might be hence pursued in practice. These methods have been used for a long time and there exists a large literature showing their efficiency for time dependent problems, as we will briefly detail in the following. We will then describe the general configuration of such methods and the classical first and second order, Lie and Strang, splitting schemes. A mathematical characterization of the splitting approximation errors will be also provided for both linear and nonlinear operators. In the second part of this chapter, we will introduce some mathematical tools and previous theoretical results concerning the numerical behavior of such methods for the solution of time and space multi-scale PDEs, illustrated in the context of reaction-diffusion systems. All of these descriptions rely on the theoretical background of the PhD of M. Duarte (**?**). A detailed survey and mathematical characterization of different types of splitting method can also be found in the book of Hundsdorfer & Verwer (Hundsdorfer and Verwer 2003). Let us remark that throughout this chapter we will describe the numerical solutions issued from splitting techniques and the resulting splitting errors, considering neither time nor space discretization issues in the time integration of the inner subproblems. The latter matters will be discussed in the forthcoming chapter.

1.1 Time Operator Splitting

Operator splitting techniques (Marchuk 1968; Strang 1968; Marchuk 1975; Marchuk 1990), also called fractional steps methods (Témam 1969a; Témam 1969b; Yanenko 1971), were first introduced in the late sixties with the main objective of reducing computational resources. In this context, a complex and potentially large problem can be split into smaller parts with an important reduction of the algorithmic complexity as well as the computational requirements. The latter characteristics were largely exploited over the past years to carry out numerical simulations in several domains, going, for instance, from electrocardiology simulations (Bernus, Wilders, Zemlin, Verschelde, and Panfilovz 2002; Trangenstein and Kim 2004), to combustion (Oran and Boris 2001; Schwer, Lu, Green, and Semião 2003) or air pollution modeling (Ostromsky, Owczarz, and Zlatev 2001; Sportisse 2007) applications. These methods can be thus considered as a standard approach in numerical applications and continue to be widely used mainly because of their simplicity of implementation and their high degree of liberty in terms of choice of dedicated numerical solvers for the split subproblems. Other advantages of these methods are given by the possibility of time stepping for the various subproblems since each one of them is independently evolved in time. Additionally, the global numerical stability of the splitting scheme is guaranteed as long as each of the inner numerical solvers ensures stability, and the mathematical formulation remains valid.

In the context of stiff problems, a particular care must be addressed to choose adequate methods that properly handle and damp out fast transients introduced by the splitting procedure in the split subproblems, for instance, in the reaction (Verwer, Blom, van Loon, and Spee 1996; Spee, Verwer, de Zeeuw, Blom, and Hundsdorfer 1998; Verwer, Spee, Blom, and Hundsdorfer 1999) or diffusion (Ropp and Shadid 2005; Ropp and Shadid 2009) terms.

In most applications, first and second order splitting schemes are implemented, for which a general mathematical background is available (see*, e.g.,* (Hairer, Lubich, and Wanner 2006) for ODEs and (Hundsdorfer and Verwer 2003) for PDEs). Even though higher order schemes are theoretically feasible, they are usually not suitable for the solution of PDEs and moreover stiff PDEs (Hundsdorfer and Verwer 2003), which constitutes a natural drawback to these schemes. On the other hand, the separate time evolution of each subproblem during a given splitting time step introduces naturally the so-called *splitting errors* into the numerical solutions. In the context of PDEs, Lanser & Verwer conducted in (Lanser and Verwer 1999) a fine analysis on the splitting errors in the solution of reaction-diffusion-convection systems, and defined the particular configurations for which splitting errors arising from the numerical separation of convection, diffusion and reaction subproblems, can be avoided. This type of study gave new insights into the use of splitting techniques for PDE problems and furthermore, complemented the classical theoretical basis.

Nevertheless, for general problems that do not display the particular characteristics defined in (Lanser and Verwer 1999), the splitting errors will likely remain throughout the numerical time integration. On the other hand, it was shown that for more complex problems involving multi-scale features, the classical mathematical characterization based on asymptotic analysis*, i.e.,* sufficiently small time steps, fails often because of time scales much faster than the considered splitting time step. Actually, the same kind of order reduction that appears in the context of time integration of stiff ODEs (see*, e.g.,* (Hairer, Lubich, and Roche 1988; Hairer and Wanner 1996)), arise similarly when considering splitting techniques for stiff problems. For PDEs, this stiffness is usually induced by highly time/space multi-scale features which furthermore are very common in the mentioned applications. All these numerical observations motivated more rigorous studies on the splitting errors, specially for the solution of stiff problems, as we will present in the second part of this chapter.

1.1.1 General Setting

Let us first consider a general linear initial value problem:

$$\left.\begin{aligned} &\mathrm{d}_t U = AU + BU, \quad t > 0, \\ &U(0) = U_0, \end{aligned}\right\} \tag{1.1}$$

with linear operators A, $B \in \mathcal{M}_m(\mathbb{R})$, where $\mathcal{M}_m(\mathbb{R})$ is the set of real square matrices of size m, $U_0 \in \mathbb{R}^m$ and $U : \mathbb{R} \to \mathbb{R}^m$, for which the exact solution is given by

$$U(t) = \mathrm{e}^{t(A+B)} U_0, \qquad t \geq 0. \tag{1.2}$$

A time operator splitting technique consists in successively solving the evolutionary problems associated with each time operator in an independent way. For system (1.1) this amounts to separately solve problems:

$$\mathrm{d}_t U = AU, \quad t > 0, \tag{1.3}$$

and

$$\mathrm{d}_t U = BU, \quad t > 0, \tag{1.4}$$

with appropriate initial conditions for each subproblem. Then, for a time discretization given by $t_0 = 0 < t_1 < \ldots < t_N$, the associated time steps or *splitting time steps* are defined as $\Delta t_n = t_{n+1} - t_n$ for $n = 0, 1, \ldots, N-1$.

Starting from the initial condition of (1.1): $U_0 = U(0)$, the splitting numerical approximation U_{n+1} of the exact values $U(t_{n+1})$ is computed from the previous U_n for $n = 0, 1, \ldots, N-1$, by means of a composition of $s \geq 1$ independent solutions of (1.3) and (1.4) with the recurrence relation:

$$U_{n+1} = \mathrm{e}^{\beta_s \Delta t_n B} \mathrm{e}^{\alpha_s \Delta t_n A} \ldots \mathrm{e}^{\beta_2 \Delta t_n B} \mathrm{e}^{\alpha_2 \Delta t_n A} \mathrm{e}^{\beta_1 \Delta t_n B} \mathrm{e}^{\alpha_1 \Delta t_n A} U_n, \tag{1.5}$$

where $\mathrm{e}^{tA} U_0$ and $\mathrm{e}^{tB} U_0$ are, respectively, the exact solutions of (1.3) and (1.4) for $t \geq 0$ from initial condition U_0. The values of the real or complex coefficients of the scheme: $(\alpha_i, \beta_i)_{i=1}^s$ such that $\sum_i \alpha_i = \sum_i \beta_i = 1$, will then define the order of approximation of the method. These splitting schemes can be seen as composition methods for which the general order conditions are well known (see (Hairer, Lubich, and Wanner 2006)).

1.1.2 First and Second Order Splitting Schemes

Taking into account the Taylor series expansion of the exact solution $U(\Delta t)$ after time Δt, if the corresponding numerical approximation U_1 is of order p, then the *local error* is given by

$$U(\Delta t) - U_1 = \mathcal{O}(\Delta t^{p+1}). \tag{1.6}$$

For system (1.1), the exact solution is given by $U(\Delta t) = \mathrm{e}^{\Delta t (A+B)} U_0$, whereas U_1 is the numerical solution at Δt, both computed from the initial value U_0.

Keeping this in mind for the splitting schemes, we introduce the first order *Lie* (or *Lie-Trotter* (Trotter 1959)) splitting formulae, for which $p = 1$ and

$$s = 1, \quad \alpha_1 = \beta_1 = 1, \tag{1.7}$$

or alternatively,

$$s = 2, \quad \alpha_1 = \beta_2 = 0, \quad \alpha_2 = \beta_1 = 1, \tag{1.8}$$

into (1.5). From a practical point of view and considering problem (1.1), the first scheme (1.7) is performed by first considering the initial value problem:

$$\left.\begin{aligned} \mathrm{d}_t U &= AU, \\ U(0) &= U_0, \end{aligned}\right\} \tag{1.9}$$

during a splitting time step Δt, which yields $U(\Delta t) = \mathrm{e}^{\Delta t A} U_0$. And then, problem:

$$\left.\begin{aligned} \mathrm{d}_t U &= BU, \\ U(0) &= \mathrm{e}^{\Delta t A} U_0, \end{aligned}\right\} \tag{1.10}$$

also during Δt, that yields finally the numerical solution:

$$U_1 = \mathcal{L}_1^{\Delta t} U_0 = \mathrm{e}^{\Delta t B} \mathrm{e}^{\Delta t A} U_0, \tag{1.11}$$

according to (1.5) with coefficients given by (1.7). Alternatively, the second Lie scheme (1.8) considers first problem (1.10), and then (1.9), so that

$$U_1 = \mathcal{L}_2^{\Delta t} U_0 = \mathrm{e}^{\Delta t A} \mathrm{e}^{\Delta t B} U_0. \tag{1.12}$$

Considering both Lie approximations, we can see that one corresponds to the *adjoint method* of the other. That is, $\mathcal{L}_1^{\Delta t}$ (resp., $\mathcal{L}_2^{\Delta t}$) is the inverse map of $\mathcal{L}_2^{\Delta t}$ (resp., $\mathcal{L}_1^{\Delta t}$) with reversed time step Δt:

$$\mathcal{L}_1^{-\Delta t}\mathcal{L}_2^{\Delta t}U_0 = \mathrm{e}^{-\Delta tB}\mathrm{e}^{-\Delta tA}\mathrm{e}^{\Delta tA}\mathrm{e}^{\Delta tB}U_0 = U_0. \tag{1.13}$$

In general it can be shown that composing one-step methods of order p yields a composition method of at least order $p+1$ (Hairer, Lubich, and Wanner 2006). In particular, composing with half-sized steps one method of odd order p with its adjoint, yields a symmetric $p+1$ method. In this way, we can obtain a symmetric second order splitting scheme known as the *Strang* (or *Marchuk* (Marchuk 1968)) splitting formulae (Strang 1963; Strang 1968) by composing $\mathcal{L}_1^{\Delta t/2}$ (resp., $\mathcal{L}_2^{\Delta t/2}$) with its adjoint method $\mathcal{L}_2^{\Delta t/2}$ (resp., $\mathcal{L}_1^{\Delta t/2}$):

$$\mathcal{S}_1^{\Delta t} = \mathcal{L}_1^{\Delta t/2}\mathcal{L}_2^{\Delta t/2}, \tag{1.14}$$

or alternatively,

$$\mathcal{S}_2^{\Delta t} = \mathcal{L}_2^{\Delta t/2}\mathcal{L}_1^{\Delta t/2}. \tag{1.15}$$

Symmetry is guaranteed because $\mathcal{S}_1^{\Delta t}$ is equal to its adjoint (the same follows for $\mathcal{S}_2^{\Delta t}$)*, i.e.,*

$$\mathcal{S}_1^{-\Delta t}\mathcal{S}_1^{\Delta t} = \mathcal{L}_2^{-\Delta t/2}\mathcal{L}_1^{-\Delta t/2}\mathcal{L}_1^{\Delta t/2}\mathcal{L}_2^{\Delta t/2} = \mathrm{Id}. \tag{1.16}$$

Coming back to problem (1.1), we have thus the numerical solutions:

$$U_1 = \mathcal{S}_1^{\Delta t}U_0 = \mathrm{e}^{\Delta tB/2}\mathrm{e}^{\Delta tA}\mathrm{e}^{\Delta tB/2}U_0, \tag{1.17}$$

or

$$U_1 = \mathcal{S}_2^{\Delta t}U_0 = \mathrm{e}^{\Delta tA/2}\mathrm{e}^{\Delta tB}\mathrm{e}^{\Delta tA/2}U_0, \tag{1.18}$$

for which $p = 2$, and, respectively,

$$s = 2, \quad \alpha_1 = 0, \quad \alpha_2 = 1, \quad \beta_1 = \beta_2 = \frac{1}{2}, \tag{1.19}$$

or

$$s = 2, \quad \alpha_1 = \alpha_2 = \frac{1}{2}, \quad \beta_1 = 1, \quad \beta_2 = 0, \tag{1.20}$$

into (1.5).

Higher order splitting schemes are also possible. Nevertheless, the order conditions for such composition methods state that either negative or complex coefficients $(\alpha_i, \beta_i)_{i=1}^s$ in (1.5) are necessary (see*, e.g.,* (Hairer, Lubich, and Wanner 2006)). Several higher order schemes of this type were already proposed (see*, e.g.,* (Yoshida 1990; Descombes 2001; McLachlan and Quispel 2002; Schatzman 2002; Thalhammer 2008; Castella, Chartier, Descombes, and Vilmart 2009; Hansen and Ostermann 2009; Descombes and Thalhammer 2010)). The former implies usually important stability restrictions and more sophisticated numerical implementations in terms of algorithmic complexity with respect to less accurate but much simpler first and second order splitting schemes. In the particular case of negative time steps, they are completely undesirable for PDEs that are ill-posed for negative time progression like parabolic equations or very stiff terms issued, for instance, from detailed chemical kinetics (Hundsdorfer and Verwer 2003).

1.1.3 Classical Numerical Analysis for Splitting Schemes

In this section, we will introduce some classical mathematical tools used for the numerical analysis of splitting schemes that are going to be used throughout this work. In a first step, we will describe the *Baker-Campbell-Hausdorff* (BCH) formula on composition of exponentials.

For the linear operators A and B, for which their exponentials e^{tA} and e^{tB} can be understood as a formal series expansion[1], we define the commutator:

$$[A,B] = AB - BA, \tag{1.21}$$

that we will also denote as[2]

$$\partial_A B = [A,B]. \tag{1.22}$$

The main idea is then to find $C(t)$ such that we can write

$$\mathrm{e}^{tA}\mathrm{e}^{tB} = \mathrm{e}^{C(t)}. \tag{1.23}$$

This exponential representation is known as the BCH formula for which it was demonstrated that $C(t)$ is the solution of the differential equation:

$$\mathrm{d}_t C = A + B + \frac{1}{2}[A-B,C] + \sum_{i\geq 2} \frac{B_i}{i!}\partial_C^i (A+B), \tag{1.24}$$

with initial value $C(0) = 0$ (Varadarajan 1974), where B_i are the *Bernoulli numbers* given by[3]

$$\sum_{i\geq 0} \frac{B_i}{i!}x^i = \frac{x}{\mathrm{e}^x - 1}. \tag{1.25}$$

Taking into account the series expansions performed in the left-hand side of (1.23), we can infer that for sufficiently small t, $C(t)$ can be also written as

$$C(t) = tC_1 + t^2C_2 + t^3C_3 + t^4C_4 + \ldots \tag{1.26}$$

which should naturally satisfy (1.23):

$$\mathrm{e}^{tA}\mathrm{e}^{tB} = \mathrm{e}^{tC_1 + t^2C_2 + t^3C_3 + t^4C_4 + \ldots}. \tag{1.27}$$

Therefore, in order to explicitly determine the coefficients of the series of $C(t)$, we insert the expansion (1.26) into (1.24), and compare like powers of t which yields

$$\left.\begin{array}{rcl}
C_1 & = & A + B, \\
C_2 & = & \dfrac{1}{4}[A-B,C_1] = \dfrac{1}{4}[A-B,A+B] = \dfrac{1}{2}[A,B], \\
C_3 & = & \dfrac{1}{6}[A-B,C_2] + \dfrac{B_2}{6}\partial_{C_1}^2(A+B) = \dfrac{1}{12}\Big[A-B,[A,B]\Big] \\
 & = & \dfrac{1}{12}\Big[A,[A,B]\Big] + \dfrac{1}{12}\Big[B,[B,A]\Big], \\
C_4 & = & \ldots = \dfrac{1}{24}\Big[A,\Big[B,[B,A]\Big]\Big].
\end{array}\right\} \tag{1.28}$$

[1] That is, $\mathrm{e}^{tA} = \left(\sum_{n=0}^{+\infty} \frac{t^n}{n!}A^n\right)$.

[2] Notice that for fixed A, the operator $\partial_A\cdot$ defines also a linear operator $B \mapsto [A,B]$ which is also called the *adjoint operator* (Varadarajan 1974).

[3] See (Hairer, Lubich, and Wanner 2006) for more details.

Using the BCH formula (1.23) and the coefficients (1.28) for $C(t)$, it is straightforward to see that the first order Lie formulae (1.11) and (1.12) verify, respectively,

$$U(\Delta t) - \mathcal{L}_1^{\Delta t} U_0 = \mathrm{e}^{\Delta t(A+B)} U_0 - \mathrm{e}^{\Delta t B} \mathrm{e}^{\Delta t A} U_0 = -\frac{\Delta t^2}{2} [B, A] U_0 + \mathcal{O}(\Delta t^3), \tag{1.29}$$

and

$$U(\Delta t) - \mathcal{L}_2^{\Delta t} U_0 = \mathrm{e}^{\Delta t(A+B)} U_0 - \mathrm{e}^{\Delta t A} \mathrm{e}^{\Delta t B} U_0 = -\frac{\Delta t^2}{2} [A, B] U_0 + \mathcal{O}(\Delta t^3). \tag{1.30}$$

It is important to notice that if the linear operators commute: $[A, B] = 0$, all the coefficients in the series of $C(t)$ are zero in (1.28) except for $C_1 = A + B$, and both Lie operators $\mathcal{L}_1^{\Delta t}$ and $\mathcal{L}_2^{\Delta t}$ act as the flow $\mathrm{e}^{\Delta t(A+B)}$ of the coupled system (1.1), according to (1.27).

Applying this time the BCH formula (1.23) to

$$\mathrm{e}^{tA/2} \mathrm{e}^{tB/2} = \mathrm{e}^{C(t)}, \tag{1.31}$$

and taking into account that

$$\mathrm{e}^{tB/2} \mathrm{e}^{tA/2} = \mathrm{e}^{-C(-t)}, \tag{1.32}$$

we can apply a second time the BCH formula (1.23) to

$$\mathrm{e}^{C(t)} \mathrm{e}^{-C(-t)} = \mathrm{e}^{tA/2} \mathrm{e}^{tB} \mathrm{e}^{tA/2} = \mathrm{e}^{S(t)}, \tag{1.33}$$

in order to obtain $S(t)$:

$$S(t) = tS_1 + t^3 S_3 + t^5 S_5 + \dots, \tag{1.34}$$

with

$$\begin{aligned} S_1 &= A + B, \\ S_3 &= -\frac{1}{24} \Big[A, [A, B]\Big] + \frac{1}{12} \Big[B, [B, A]\Big]. \end{aligned} \tag{1.35}$$

Notice that only odd powers of t are present in (1.34) since the adjoint method of the symmetric scheme $\mathrm{e}^{tA/2} \mathrm{e}^{tB} \mathrm{e}^{tA/2}$ is obtained by just changing the sign of t and therefore of $\mathrm{e}^{S(t)}$, according to (1.33). In this case, $\mathrm{e}^{S(t)}$ is not other than the Strang scheme $\mathcal{S}_2^t$ according to (1.18), and we see that the local errors can be written as

$$\begin{aligned} U(\Delta t) - \mathcal{S}_1^{\Delta t} U_0 &= \mathrm{e}^{\Delta t(A+B)} U_0 - \mathrm{e}^{\Delta t B/2} \mathrm{e}^{\Delta t A} \mathrm{e}^{\Delta t B/2} U_0 \\ &= \frac{\Delta t^3}{24} \Big[B, [B, A]\Big] U_0 - \frac{\Delta t^3}{12} \Big[A, [A, B]\Big] U_0 + \mathcal{O}(\Delta t^4), \end{aligned} \tag{1.36}$$

and

$$\begin{aligned} U(\Delta t) - \mathcal{S}_2^{\Delta t} U_0 &= \mathrm{e}^{\Delta t(A+B)} U_0 - \mathrm{e}^{\Delta t A/2} \mathrm{e}^{\Delta t B} \mathrm{e}^{\Delta t A/2} U_0 \\ &= \frac{\Delta t^3}{24} \Big[A, [A, B]\Big] U_0 - \frac{\Delta t^3}{12} \Big[B, [B, A]\Big] U_0 + \mathcal{O}(\Delta t^4). \end{aligned} \tag{1.37}$$

In this way, we can formally represent the local errors of both Lie and Strang schemes. We remark that for both cases no splitting error is introduced for commuting operators. Furthermore, the latter error expressions can be easily extended to an arbitrary number of linear operators. However, it is important to notice that these estimates are asymptotically verified for sufficiently small splitting time steps Δt, since they are based on Taylor series expansions. Extension to general nonlinear configurations is straightforward using a Lie operator formalism (Sanz-Serna and Calvo 1994), in which case the same previous estimates remain valid with linear operators defined by the Lie derivatives associated with the various nonlinear operators, as we will show in what follows.

1.1.4 An example of order reduction for splitting methods

We are considering a very simple case of an ordinary differential equation in $\mathbb{R}^3$, let A, P and D be the following matrices:

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad D = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{1.38}$$

with $A = I_3 + D$, and Q is defined by $I_3 = P + Q$.

The system that we consider is the following:

$$\begin{cases} \mathrm{d}_t U_\varepsilon = AU_\varepsilon - \frac{P}{\varepsilon} U_\varepsilon \\ U_\varepsilon(0) = U_0, \end{cases} \tag{1.39}$$

with exact solution

$$U_\varepsilon(t) = e^{t(A-P/\varepsilon)} U_0. \tag{1.40}$$

The Lie and Strang approximations of the previous system are defined by:

$$\begin{aligned} L_{1\varepsilon}(t) =& e^{tA} e^{-tP/\varepsilon} U_0, \\ L_{2\varepsilon}(t) =& e^{-tP/\varepsilon} e^{tA} U_0, \\ S_{1\varepsilon}(t) =& e^{tA/2} e^{-tP/\varepsilon} e^{tA/2} U_0, \\ S_{2\varepsilon}(t) =& e^{-tP/2\varepsilon} e^{tA} e^{-tP/2\varepsilon} U_0. \end{aligned}$$

Let us insist on the fact that A and P do not commute since $[A,P] = AP - PA = D$.

The various operator involved can be evaluated exactly through the following lemma.

Lemma 1.1. *We have the following expressions for the two Lie formulae and the exact flow:*

$$e^{tA} e^{-tP/\varepsilon} = e^t \left(e^{-t/\varepsilon} P + t\, D + Q \right), \tag{1.41}$$

$$e^{-tP/\varepsilon} e^{tA} = e^t \left(e^{-t/\varepsilon} P + t e^{-t/\varepsilon}\, D + Q \right), \tag{1.42}$$

$$e^{A-tP/\varepsilon} = e^t \left(e^{-t/\varepsilon} P + \varepsilon(1 - e^{-t/\varepsilon})\, D + Q \right). \tag{1.43}$$

$$\tag{1.44}$$

Proof. We need a few expressions, which can easily be verified:

$$A = \begin{pmatrix} 1 & n & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad D^2 = 0, \quad DQ = D, \quad QD = 0, \quad DP = 0.$$

From there it is easy to that

$$e^{tA} = e^t I_3 + \alpha\, D, \quad \alpha = t + 2\frac{t^2}{2!} + 3\frac{t^3}{3!} + \ldots + n\frac{t^n}{n!} + \ldots = te^t.$$

Besides, the projection matrix has a the straightforward flow:

$$e^{-tP/\varepsilon} = e^{-t/\varepsilon} P + Q.$$

The first two equalities of 1.41 are then obtained. Evaluating the exact flow requires a bit more algebra. Let us denote $C = A - P/\varepsilon$, then:

$$C^2 = \begin{pmatrix} (1-\frac{1}{\varepsilon})^2 & 2-\frac{1}{\varepsilon} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \left(1-\frac{1}{\varepsilon}\right)^2 P + \left(2-\frac{1}{\varepsilon}\right) D + Q.$$

The generic term of C^n in P is $(1-\frac{1}{\varepsilon})^n$ and the one in Q is still 1. If α_n is the generic one in D, then $\alpha_2 = 2 - \frac{1}{\varepsilon} = 1 + (1-\frac{1}{\varepsilon})^1$, and by recurrence, $\alpha_n = \sum_{j=0}^{n-1}(1-\frac{1}{\varepsilon})^j$, so that finally:

$$\alpha_n = \frac{1-(1-\frac{1}{\varepsilon})^n}{1-(1-\frac{1}{\varepsilon})} = \varepsilon\left(1-\left(1-\frac{1}{\varepsilon}\right)^n\right).$$

The series $\sum_{j=0}^{+\infty}\alpha_n\frac{t^n}{n!}$ in front of D thus can be easily evaluated:

$$\sum_{j=0}^{+\infty}\alpha_n\frac{t^n}{n!} = \varepsilon\left(e^t - 1 - (e^{t-t/\varepsilon}-1)\right),$$

and the last and third estimate is obtained. □

Consequently, we have the following result :

Proposition 1.2. *For our special system, we can then have exact error estimates*

$$U_\varepsilon(t) - L_{1\varepsilon}(t) = e^t\left(\varepsilon(1-e^{-t/\varepsilon}) - t\right)D, \tag{1.45}$$

$$U_\varepsilon(t) - L_{2\varepsilon}(t) = e^t\left(\varepsilon(1-e^{-t/\varepsilon}) - te^{-t/\varepsilon}\right)D. \tag{1.46}$$

$$\tag{1.47}$$

Let $t > 0$, for ε sufficiently small satisfying $\varepsilon << t$, we have the following estimates

$$\lim_{\varepsilon\to 0} e^{tA}e^{-tP/\varepsilon} = e^tQ, \tag{1.48}$$

$$\lim_{\varepsilon\to 0} e^{-tP/\varepsilon}e^{tA} = e^t(Q + t\,D), \tag{1.49}$$

$$\lim_{\varepsilon\to 0} e^{t(A-P/\varepsilon)} = e^tQ. \tag{1.50}$$

$$\tag{1.51}$$

This proposition shows that in the presence of stiffness, Lie formula ending with the stiff part is the best candidate since it does not suffer from order reduction. In fact the other Lie formula ends up with a local order of 1 and a global order of 0, which means that we are evaluating something, but we are not even sure of convergence of the scheme. Another way of seing this is to consider the limit of the various flows when $\varepsilon \to 0$. in order to recover in the limit of small ε the exact flow. Even if the two lie formulae are symmetric and equivalent for non-stiff operators, they sensibly differ when one operator is introducing strong stiffness into the system.

The proof of order reduction is based on the study of the difference of the two reduces problems and in fact follows the diagram

$$\begin{tikzcd}
\textbf{System} \arrow[r, "O(\exp(-t/\varepsilon))", "\textit{Singular Perturbation}"'] \arrow[d] & \textbf{Reduced System} \arrow[d] \\
\textbf{Splitting} \arrow[r, "O(\exp(-t/\varepsilon))", "\textit{Singular Perturbation}"'] & \textbf{Reduced Splitting}
\end{tikzcd}$$

The key point of this proof is the singular perturbation hypothesis (as in (Massot 2002)) : the reduced problems are well-posed and the differences between all the problems and their reduced problems behave like $O(\exp(-t/\varepsilon))$.

1.1.5 The Lie Operator Formalism

We introduce the Lie operator formalism in order to generalize the use of exponentials of linear operators in the context of nonlinear operators. Let X be a Banach space, $T > 0$, and an unbounded nonlinear operator F from $D(F) \subset X$ to X, we consider the general autonomous equation:

$$\left.\begin{aligned} &\mathrm{d}_t U = F(U(t)), && 0 < t < T, \\ &U(0) = U_0, && t = 0. \end{aligned}\right\} \tag{1.52}$$

The exact solution of this evolutionary equation is formally given by

$$U(t) = T^t U_0, \quad 0 \leq t \leq T, \tag{1.53}$$

where T^t is the semiflow associated with (1.52). The Lie operator D_F associated with F is then a linear operator acting on the space of operators defined in X (see*, e.g.,* (Sanz-Serna and Calvo 1994; Hairer, Lubich, and Wanner 2006; Descombes and Thalhammer 2011)). More precisely, for any unbounded nonlinear operator G from $D(G) \subset X$ to X with Fréchet derivative G', D_F maps G into a new operator $D_F G$, such that for any v in X:

$$(D_F G)(v) = G'(v)F(v). \tag{1.54}$$

Using the chain rule for the solution $U(t)$ of (1.52), we have that

$$\partial_t G(U(t)) = (D_F G)(U(t)), \tag{1.55}$$

and hence applying the Lie operator iteratively, we obtain

$$\partial_t^n G(U(t)) = (D_F^n G)(U(t)). \tag{1.56}$$

A formal Taylor expansion yields[4]

$$G(U(t)) = \sum_{n=0}^{+\infty} \frac{t^n}{n!} \left(\partial_t^n G(U(t))\right)\Big|_{t=0} = \left(\sum_{n=0}^{+\infty} \frac{t^n}{n!} D_F^n G\right) U_0 = \left(\mathrm{e}^{tD_F} G\right) U_0. \tag{1.57}$$

If we now assume that G is the identity operator Id, we finally get

$$U(t) = T^t U_0 = \left(\mathrm{e}^{tD_F} \mathrm{Id}\right) U_0. \tag{1.58}$$

Therefore, the Lie operator is indeed a way of writing the solution of a nonlinear ODE in terms of a linear but differential operator.

Following (1.57), an important result obtained by Gröbner in (Gröbner 1967) considers the composition of two semiflows T_1^t and T_2^s associated with F_1 and F_2 for any v in X:

$$T_1^t T_2^s v = \left(\mathrm{e}^{sD_{F_2}} T_1^t\right) v = \left(\mathrm{e}^{sD_{F_2}} \mathrm{e}^{tD_{F_1}} \mathrm{Id}\right) v. \tag{1.59}$$

[4] We remark that if $F(U(t))$ is not an analytic function in (1.57), but $F \in \mathcal{C}^N(\mathbb{R})$, then the series has to be truncated and a $\mathcal{O}(t^N)$ remainder must be included.

Notice that the order of the operators to the left and right are permuted for the equivalent representations in (1.59). The latter result can naturally be extended to more than two semiflows $T_1^t, T_2^s, \dots, T_m^r$ associated with $F_1, F_2, \dots, F_m$:

$$T_1^t T_2^s \cdots T_m^r v = \left(\mathrm{e}^{rD_{F_m}} \cdots \mathrm{e}^{sD_{F_2}} \mathrm{e}^{tD_{F_1}} \mathrm{Id}\right) v. \tag{1.60}$$

The same analysis previously detailed to estimate the splitting errors can be analogously performed by applying the Baker-Campbell-Hausdorff formula (1.23) to (1.59):

$$\mathrm{e}^{sD_{F_2}} \mathrm{e}^{tD_{F_1}} = \mathrm{e}^{D(s,t)}, \tag{1.61}$$

where the differential operator $D(s,t)$ is given by

$$\begin{aligned} D(s,t) \quad = \quad & sD_{F_2} + tD_{F_1} + \frac{st}{2}[D_{F_2}, D_{F_1}] + \frac{s^2 t}{12}\Big[D_{F_2}, [D_{F_2}, D_{F_1}]\Big] \\ & + \frac{st^2}{12}\Big[D_{F_1}, [D_{F_1}, D_{F_2}]\Big] + \frac{s^2 t^2}{24}\Big[D_{F_2}, \Big[D_{F_1}, [D_{F_1}, D_{F_2}]\Big]\Big] + \dots \end{aligned} \tag{1.62}$$

according to (1.28). The *Lie bracket* for differential operators is defined exactly as for linear operators (1.21):

$$[D_{F_1}, D_{F_2}] = D_{F_1} D_{F_2} - D_{F_2} D_{F_1}, \tag{1.63}$$

and acts again as a linear differential operator:

$$[D_{F_1}, D_{F_2}] = \left(F_2' F_1 - F_1' F_2\right) \partial_v, \tag{1.64}$$

for any v in X according to (1.54).

In this way, considering a general system of nonlinear ODEs

$$\left.\begin{aligned} &\mathrm{d}_t U = F_1(U(t)) + F_2(U(t)), \quad t > 0, \\ &U(0) = U_0, \end{aligned}\right\} \tag{1.65}$$

with $U_0 \in \mathbb{R}^m$, $U : \mathbb{R} \to \mathbb{R}^m$, and $F_1, F_2 : \mathbb{R}^m \to \mathbb{R}^m$, the same asymptotic expressions for the local error estimates for the Lie and Strang formulae (1.29) and (1.30), and (1.36) and (1.37), can be recast with the linear operators A and B replaced by the Lie operators D_{F_1} and D_{F_2}. The same follows for an arbitrary number of operators. Furthermore, splitting order conditions can be then deduced by using this Lie formalism for general nonlinear operators (Yoshida 1990; Hairer, Lubich, and Wanner 2006). In particular, it was with this representation that the commuting conditions for nonlinear or linear operators, yielding no splitting errors, were introduced in (Lanser and Verwer 1999) for the splitting solution of reaction-convection-diffusion systems (see (Hundsdorfer and Verwer 2003) for more details). Exact splitting error representations introduced in (Descombes and Schatzman 2002) can be also analyzed in this framework for general nonlinear PDEs (Descombes, Duarte, Dumont, Laurent, Louvet, and Massot 2014).

1.2 Splitting Errors for Time/Space Multi-Scale PDEs

In this second part, we will present some theoretical results previously introduced in the literature, to characterize the numerical behavior of splitting techniques for the solution of multi-scale PDEs. These multi-scale features might arise in time because of the presence of different numerical or physical evolution rates within a rather broad range, or in space because of the presence of steep gradients or large higher order spatial derivatives within the computational domain. More likely, they are coupled both in time and space throughout the numerical integration. As a consequence, there might be some perturbing effects in the accuracy of the numerical approximations of the governing equations, traduced usually by an order reduction of the splitting method. This kind of numerical difficulty might be theoretically characterized as a direct result of the stiffness of the time dependent equations as we will discuss in the next chapter, and generally speaking we can say that we are dealing with the numerical solution of stiff PDEs.

In what follows we detail some elements to describe the numerical behavior of splitting schemes faced with the mentioned stiffness, in the case of reaction-diffusion systems. The study of this kind of problem allows us to illustrate the numerical difficulties encountered in general, and the resulting conclusions might be partially extended to more complex configurations. Nevertheless, there is a continuous research in this field and more detailed mathematical descriptions are always needed to further understand these issues.

1.2.1 Mathematical Framework: Reaction-Diffusion Systems

We focus on a class of multi-scale phenomena that can be modeled by general reaction-diffusion systems of type:

$$\left.\begin{array}{ll}\partial_t U - \partial_{\mathrm{x}} \cdot (D(U)\partial_{\mathrm{x}} U) = F(U), & \mathrm{x} \in \mathbb{R}^d,\ t > 0,\\ U(0,\mathrm{x}) = U_0(\mathrm{x}) & \mathrm{x} \in \mathbb{R}^d,\end{array}\right\} \tag{1.66}$$

where $F : \mathbb{R}^m \to \mathbb{R}^m$, $U_0 : \mathbb{R}^d \to \mathbb{R}^m$ and $U : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^m$, with the diffusion matrix $D(U)$, which is a tensor of order $d \times d \times m$. In case we are only considering linear diagonal diffusion, the elements of the diffusion matrix are written as $D_{i_1 i_2 i_3}(U) = D_{i_3}\delta_{i_1 i_2}$ with indices i_1, i_2, $i_3 = 1, \ldots, m$, so that the diffusion operator reduces to the heat operator with scalar diffusion coefficient D_{i_3} for component $u^{(i_3)}$ of U, and the system (1.66) becomes

$$\left.\begin{array}{ll}\partial_t U - D\,\partial_{\mathrm{x}}^2 U = F(U), & \mathrm{x} \in \mathbb{R}^d,\ t > 0,\\ U(0,\mathrm{x}) = U_0(\mathrm{x}) & \mathrm{x} \in \mathbb{R}^d.\end{array}\right\} \tag{1.67}$$

In general, the source term F into (1.66) and (1.67) models reactive chemical mechanisms with a broad time scale spectrum. On the other hand, complementary stiffness results from the potentially fast scales introduced in the numerical solution when applying the diffusion operator to localized steep spatial gradients or highly inhomogeneous distributions, as it is usually the case in physical phenomena characterized by the presence of fronts or irregular space multi-scale configurations. In this way, the associated stiffness will surely have an effect on the numerical behavior of the splitting schemes as we will briefly describe in the following.

1.2.2 Splitting Order Reduction for Time Multi-Scale Systems

Even though splitting schemes are usually quite efficient for the solution of time dependent equations, several works showed that the standard numerical analysis of splitting schemes fails in presence of scales much faster than the splitting time step (Goyal, Paul, Mukunda, and Deshpande 1988; D'Angelo 1994; D'Angelo and Larrouturou 1995; Yang and Pope 1998; Verwer and Sportisse 1998; Sportisse, Bencteux, and Plion 2000), and that an order reduction of the methods is numerically observed. In particular, a first major step towards a rigorous study of such cases was conducted by Sportisse in (Sportisse 2000) in the framework of a linear system of ODEs, issued from a reaction-diffusion system with a linear source term and diagonal diffusion. In this work, a fast characteristic time was associated with the source term by means of a multiplying factor ϵ^{-1}, with small ϵ, to split the original system into a stiff and a non stiff subproblem. In this context, a local order reduction of the splitting schemes was mathematically described based on singular perturbation theory, whereas splitting methods ending with the stiffest operator were also shown to be more accurate than the others. Similar conclusions were obtained by Kozlov *et al.* in (Kozlov, rnø, and Owren 2004) for nonlinear systems of ODEs, split also into stiff and non stiff parts, using singular perturbation elements as well. In this framework, Descombes & Massot introduced in (Descombes and Massot 2004) a general theoretical approach for nonlinear reaction-diffusion systems with time multi-scale features issued from more realistic physical configurations. We will briefly describe in the following a few results coming from (Descombes and Massot 2004).

Supposing that the system (1.67) shows a well partitioned structure such that $U = (u^\epsilon, v^\epsilon)^T$ and thus $F(U) = (f(u^\epsilon, v^\epsilon), g(u^\epsilon, v^\epsilon)/\epsilon)^T$, where $u^\epsilon \in \mathbb{R}^{m^{\text{slow}}}$ and $v^\epsilon \in \mathbb{R}^{m^{\text{fast}}}$ stand, respectively, for the slow and fast variables of the dynamical system associated with (1.67), and $m = m^{\text{slow}} + m^{\text{fast}}$; we consider the following reaction-diffusion system:

$$\left.\begin{aligned}
&\partial_t u^\epsilon - \partial_{\mathrm{x}}^2 u^\epsilon = f(u^\epsilon, v^\epsilon), && \mathrm{x} \in \mathbb{R}^d,\ t > 0,\\
&\partial_t v^\epsilon - \partial_{\mathrm{x}}^2 v^\epsilon = \frac{g(u^\epsilon, v^\epsilon)}{\epsilon}, && \mathrm{x} \in \mathbb{R}^d,\ t > 0,\\
&u^\epsilon(0,\mathrm{x}) = u_0(\mathrm{x}), && \mathrm{x} \in \mathbb{R}^d,\\
&v^\epsilon(0,\mathrm{x}) = v_0(\mathrm{x}), && \mathrm{x} \in \mathbb{R}^d,
\end{aligned}\right\} \tag{1.68}$$

for a small parameter ϵ and the identity in $\mathcal{M}_m(\mathbb{R})$, as diffusion matrix. For the sake of brevity, we will only consider this diagonal case, even though a quasi-linear non-diagonal diffusion was also analyzed in (Descombes and Massot 2004). We denote by $(u^\epsilon(t), v^\epsilon(t)) = T^t_\epsilon(u_0, v_0)$ the solution of (1.68) at some time t.

In order to settle an appropriate mathematical framework, we assume that this system admits an entropic structure (Massot 2002) so that the source term admits a well partitioned Tikhonov normal form (Tikhonov, Vasil'eva, and Sveshnikov 1985). Therefore, there is a partial equilibrium manifold where the fast time scales have been relaxed, which is globally stable. In particular, the entropy is a global Lyapounov function and we can thus perform a singular perturbation analysis with asymptotic expansions (Massot 2002). In this context, we can consider the singular perturbation analysis for the finite dimensional dynamical system:

$$\left.\begin{aligned}
&\mathrm{d}_t \bar{u}^\epsilon = f(\bar{u}^\epsilon, \bar{v}^\epsilon), && t > 0,\\
&\mathrm{d}_t \bar{v}^\epsilon = \frac{g(\bar{u}^\epsilon, \bar{v}^\epsilon)}{\epsilon}, && t > 0,\\
&\bar{u}^\epsilon(0) = \bar{u}_0,\\
&\bar{v}^\epsilon(0) = \bar{v}_0,
\end{aligned}\right\} \tag{1.69}$$

which corresponds to a homogeneous system without diffusion. The corresponding reduced system can thus be written as

$$\left.\begin{aligned}
&\mathrm{d}_t \bar{u} = G(\bar{u}), && t > 0,\\
&\bar{u}(0) = \bar{u}_0,\\
&\bar{v}(t) = h(\bar{u}(t)), && t \geq 0,
\end{aligned}\right\} \tag{1.70}$$

where $G(\bar{u}) = f(\bar{u}, h(\bar{u}))$, and $g(\bar{u}, \bar{v}) = \bar{v} - h(\bar{u}) = 0$. The inner boundary layer, because of the well-partitioned structure of the dynamical system, can be considered as a projection step in an affine manifold onto the partial equilibrium $h(\bar{u}_0)$ in the $\bar{v}$ variable. Denoting by $\Pi_0\bar{v}$ the associated variable centered at $h(\bar{u}_0)$, the boundary layer, parametrized by the spatial coordinate x, can be described by the following differential equation:

$$\left.\begin{aligned}
&\mathrm{d}_\tau \Pi_0\bar{v} = g(u_0, h(u_0) + \Pi_0\bar{v}), && \tau > 0,\\
&\Pi_0\bar{v}(0) = v_0 - h(u_0),
\end{aligned}\right\} \tag{1.71}$$

for a time scale defined by $\tau = t/\epsilon$.

Assuming that there exists a convex compact set K which contains the initial condition $(\bar{u}_0, \bar{v}_0) \in K$, and which is invariant by (1.68), (1.70) and (1.71), it has been proved in (Massot 2002) that for ϵ sufficiently small, we have for $t \in [0, +\infty)$:

$$\bar{v}^\epsilon(t, \epsilon) = \Pi_0 \bar{v}(t/\epsilon) + \bar{v}(t) + \mathcal{O}(\epsilon), \tag{1.72}$$

$$\bar{u}^\epsilon(t, \epsilon) = \bar{u}(t) + \mathcal{O}(\epsilon), \tag{1.73}$$

and for some $\kappa > 0$, we obtain an estimate for the inner boundary layer

$$\Pi_0 \bar{v}(t/\epsilon) = \mathcal{O}\left(\mathrm{e}^{\left(-\kappa t/\epsilon\right)}\right). \tag{1.74}$$

Considering now the reduced problem associated with the complete system (1.68):

$$\left.\begin{aligned} &\partial_t u - \partial_{\mathrm{x}}^2 u = G(u), && \mathrm{x} \in \mathbb{R}^d,\ t > 0, \\ &u(0, \mathrm{x}) = u_0(\mathrm{x}), && \mathrm{x} \in \mathbb{R}^d, \\ &v(t, \mathrm{x}) = h(u(t, \mathrm{x})), && \mathrm{x} \in \mathbb{R}^d, t \geq 0, \end{aligned}\right\} \tag{1.75}$$

and based on the previous singular perturbation analysis as detailed in (Descombes and Massot 2004), if we assume that $(\bar{u}_0(\mathrm{x}), \bar{v}_0(\mathrm{x})) \in K$ for $\mathrm{x} \in \mathbb{R}^d$ and that the solution $T^t u_0 = (u(t), h(u(t))$ of (1.75) leaves also K invariant, for ϵ sufficiently small, we have for $t \in [0, +\infty)$:

$$\|u^\epsilon(t, \epsilon) - u(t)\|_{L^2} = \mathcal{O}(\epsilon), \tag{1.76}$$

$$\|v^\epsilon(t, \epsilon) - \Pi_0 \bar{v}(t/\epsilon) - h(u(t))\|_{L^2} = \mathcal{O}(\epsilon), \tag{1.77}$$

and the corresponding estimate for the inner boundary layer:

$$\|\Pi_0 \bar{v}(t/\epsilon)\|_{L^2} = \mathcal{O}\left(\mathrm{e}^{\left(-\kappa t/\epsilon\right)}\right). \tag{1.78}$$

With this framework, we introduce the standard decoupling of the diffusion and reaction problems for system (1.68). Let us then denote by $X^t(u_0, v_0)$ the solution of the diffusion problem:

$$\left.\begin{aligned} &\partial_t u_D - \partial_{\mathrm{x}}^2 u_D = 0, && \mathrm{x} \in \mathbb{R}^d,\ t > 0, \\ &\partial_t v_D - \partial_{\mathrm{x}}^2 v_D = 0, && \mathrm{x} \in \mathbb{R}^d,\ t > 0, \end{aligned}\right\} \tag{1.79}$$

for some initial data $u_D(0, \cdot) = u_0(\cdot)$ and $v_D(0, \cdot) = v_0(\cdot)$; and by $Y_\epsilon^t(u_0, v_0)$ the solution of the reaction problem:

$$\left.\begin{aligned} &\partial_t u_R^\epsilon = f(u_R^\epsilon, v_R^\epsilon), && \mathrm{x} \in \mathbb{R}^d, t > 0, \\ &\partial_t v_R^\epsilon = \frac{g(u_R^\epsilon, v_R^\epsilon)}{\epsilon}, && \mathrm{x} \in \mathbb{R}^d, t > 0, \end{aligned}\right\} \tag{1.80}$$

with initial data $u_R^\epsilon(0, \cdot) = u_0(\cdot)$ and $v_R^\epsilon(0, \cdot) = v_0(\cdot)$, where the spatial coordinate x can be considered as a parameter. The Lie and Strang splitting formulae associated with (1.68) are given by:

$$\mathcal{L}_{1,\epsilon}^t(u_0, v_0) = X^t Y_\epsilon^t(u_0, v_0), \tag{1.81}$$

$$\mathcal{L}_{2,\epsilon}^t(u_0, v_0) = Y_\epsilon^t X^t(u_0, v_0), \tag{1.82}$$

$$\mathcal{S}_{1,\epsilon}^t(u_0, v_0) = X^{t/2} Y_\epsilon^t X^{t/2}(u_0, v_0), \tag{1.83}$$

$$\mathcal{S}_{2,\epsilon}^t(u_0, v_0) = Y_\epsilon^{t/2} X^t Y_\epsilon^{t/2}(u_0, v_0). \tag{1.84}$$

If we consider now the reduced problem of (1.80) when ϵ tends to zero:

$$\left.\begin{aligned} &\partial_t u_R = f(u_R, h(u_R)) = G(u_R), && \mathrm{x} \in \mathbb{R}^d, t > 0,\\ &u_R(0,\mathrm{x}) = u_0(\mathrm{x}), && \mathrm{x} \in \mathbb{R}^d,\\ &v_R(t,\mathrm{x}) = h(u_R(t,\mathrm{x})), && \mathrm{x} \in \mathbb{R}^d, t \geq 0, \end{aligned}\right\} \tag{1.85}$$

with solution given by $(u_R(t), h(u_R(t)) = Y^t u_0$ as for (1.70), we define the corresponding reduced splitting schemes:

$$\mathcal{L}_1^t u_0 = X^t Y^t u_0, \tag{1.86}$$

$$\mathcal{L}_2^t (u_0, v_0) = Y^t X^t (u_0, v_0), \tag{1.87}$$

$$\mathcal{S}_1^t (u_0, v_0) = X^{t/2} Y^t X^{t/2} (u_0, v_0), \tag{1.88}$$

$$\mathcal{S}_2^t u_0 = Y^{t/2} X^t Y^{t/2} u_0, \tag{1.89}$$

where the fast scales have been previously relaxed in the reaction part by considering the reduced problem (1.85).

To study the order of approximation of the exact solution T_ϵ^t of the coupled problem (1.68) by the splitting schemes (1.81)-(1.84), we investigate the order of approximation of T^t associated with the reduced problem (1.75) by the reduced splitting schemes (1.86)-(1.89). Defining the corresponding local errors:

$$\left.\begin{aligned} (u_{\mathrm{err}1}, v_{\mathrm{err}1}) &= T^t u_0 - \mathcal{L}_1^t u_0,\\ (u_{\mathrm{err}2}, v_{\mathrm{err}2}) &= T^t u_0 - \mathcal{L}_2^t (u_0, v_0),\\ (u_{\mathrm{err}3}, v_{\mathrm{err}3}) &= T^t u_0 - \mathcal{S}_1^t (u_0, v_0),\\ (u_{\mathrm{err}4}, v_{\mathrm{err}4}) &= T^t u_0 - \mathcal{S}_2^t u_0, \end{aligned}\right\} \tag{1.90}$$

it was demonstrated in (Descombes and Massot 2004) that the local error for the slow and fast variables of the various splitting schemes satisfies

$$\|u_{\mathrm{err}1}\|_{L^2} = \mathcal{O}(t^2), \qquad \|v_{\mathrm{err}1}\|_{L^2} = \mathcal{O}(t), \tag{1.91}$$

$$\|u_{\mathrm{err}2}\|_{L^2} = \mathcal{O}(t^2), \qquad \|v_{\mathrm{err}2}\|_{L^2} = \mathcal{O}(t^2), \tag{1.92}$$

$$\|u_{\mathrm{err}3}\|_{L^2} = \mathcal{O}(t^3), \qquad \|v_{\mathrm{err}3}\|_{L^2} = \mathcal{O}(t), \tag{1.93}$$

$$\|u_{\mathrm{err}4}\|_{L^2} = \mathcal{O}(t^3), \qquad \|v_{\mathrm{err}4}\|_{L^2} = \mathcal{O}(t^3). \tag{1.94}$$

Taking into account that, for instance, for $\mathcal{L}_{1,\epsilon}^t (u_0, v_0)$ the error of approximation with respect to $T_\epsilon^t (u_0, v_0)$ is given by

$$\begin{aligned} T_\epsilon^t (u_0, v_0) - \mathcal{L}_{1,\epsilon}^t (u_0, v_0) &= T_\epsilon^t (u_0, v_0) - T^t u_0 + T^t u_0 - \mathcal{L}_1^t u_0\\ &\quad + \mathcal{L}_1^t u_0 - \mathcal{L}_{1,\epsilon}^t (u_0, v_0), \end{aligned} \tag{1.95}$$

and that

$$\begin{aligned} \|T_\epsilon^t (u_0, v_0) - \mathcal{L}_{1,\epsilon}^t (u_0, v_0)\|_{L^2} &\leq \|T_\epsilon^t (u_0, v_0) - T^t u_0\|_{L^2} + \|T^t u_0 - \mathcal{L}_1^t u_0\|_{L^2}\\ &\quad + \|\mathcal{L}_1^t u_0 - \mathcal{L}_{1,\epsilon}^t (u_0, v_0)\|_{L^2}, \end{aligned} \tag{1.96}$$

for ϵ sufficiently small and for $t \geq 0$ sufficiently small, the local errors admit the following asymptotic expansions (Descombes and Massot 2004):

$$\|T_\epsilon^t(u_0, v_0) - \mathcal{L}_{1,\epsilon}^t(u_0, v_0)\|_{L^2} = \mathcal{O}(t) + \mathcal{O}\left(\mathrm{e}^{\left(-\kappa t/\epsilon\right)}\right) + \mathcal{O}(\epsilon), \tag{1.97}$$

$$\|T_\epsilon^t(u_0, v_0) - \mathcal{S}_{1,\epsilon}^t(u_0, v_0)\|_{L^2} = \mathcal{O}(t) + \mathcal{O}\left(\mathrm{e}^{\left(-\kappa t/\epsilon\right)}\right) + \mathcal{O}(\epsilon), \tag{1.98}$$

and

$$\|T_\epsilon^t(u_0, v_0) - \mathcal{L}_{2,\epsilon}^t(u_0, v_0)\|_{L^2} = \mathcal{O}(t^2) + \mathcal{O}\left(\mathrm{e}^{\left(-\kappa t/\epsilon\right)}\right) + \mathcal{O}(\epsilon), \tag{1.99}$$

$$\|T_\epsilon^t(u_0, v_0) - \mathcal{S}_{2,\epsilon}^t(u_0, v_0)\|_{L^2} = \mathcal{O}(t^3) + \mathcal{O}\left(\mathrm{e}^{\left(-\kappa t/\epsilon\right)}\right) + \mathcal{O}(\epsilon), \tag{1.100}$$

considering estimates (1.91)-(1.94) for the second term of the right hand side of (1.96), and (1.76)-(1.78) for the other two terms.

Through this mathematical model and the corresponding numerical analysis, we can conclude that no order reduction of the splitting schemes is expected for the slow variables whenever we consider splitting time steps much larger than the fastest scales present in the problem: $t > \epsilon$, following (Descombes and Massot 2004). On the other hand, for a linear diagonal diffusion, if we use splitting schemes ending with the reaction operator which includes the fastest scales, then there is no reason to expect order reductions not even for the fast variables. In particular, in the configuration of a partial equilibrium manifold with non zero curvature, a situation which can only be obtained with a nonlinear reaction source term, the splitting schemes ending with the diffusion operator encounter an order reduction related to the Lie bracket between the Laplacian operator and the h function defining the partial equilibrium manifold (Descombes and Massot 2004). Finally, let us recall that in practical implementations of splitting techniques, dedicated solvers must be considered to properly handle the fast transients associated with the inner boundary layers given by (1.74), as previously remarked (Verwer, Blom, van Loon, and Spee 1996; Spee, Verwer, de Zeeuw, Blom, and Hundsdorfer 1998; Verwer, Spee, Blom, and Hundsdorfer 1999)[5], and also to ensure the mathematical framework detailed in this section in which the split reaction and diffusion subproblems were exactly solved for estimates (1.97)-(1.100).

1.2.3 Splitting Errors with High Spatial Gradients

We have seen in the previous study that the classical error representations of splitting schemes are not always enough to describe more precisely some important features related to the modeling equations. Therefore, more rigorous studies were performed and in particular an exact representation of the local errors of splitting schemes was achieved by Descombes & Schatzman in (Descombes and Schatzman 2002) for general linear problems like (1.1). Once again, extension to nonlinear operators is straightforward using a Lie operator formalism as shown in (Descombes, Duarte, Dumont, Laurent, Louvet, and Massot 2014). These results led to many further mathematical studies on splitting errors (see*, e.g.,* (Descombes and Thalhammer 2010; Descombes and Thalhammer 2011)), and such a precise error representation showed to be mandatory to better analyze some particular issues like the influence of high spatial gradients on the solution of reaction-diffusion systems solved by splitting techniques (Descombes, Dumont, Louvet, and Massot 2007; Duarte, Descombes, and Massot 2011; Descombes, Duarte, Dumont, Laurent, Louvet, and Massot 2014). In this way, it

[5] The same remark is valid for the numerical integration of stiff ODEs (Hairer, Lubich, and Roche 1988; Hairer and Wanner 1996).

is possible to better depict some potential numerical difficulties issued this time from the space multi-scale character of some physical phenomena modeled by the governing equations, *e.g.,* (1.66), as previously remarked and as analyzed, for instance, in (Ropp, Shadid, and Ober 2004; Ropp and Shadid 2005).

Let us recall the initial value problem (1.1), for some linear operators A, $B \in \mathcal{M}_m(\mathbb{R})$, $U_0 \in \mathbb{R}^m$, $U : \mathbb{R} \to \mathbb{R}^m$:

$$\left.\begin{aligned} &\mathrm{d}_t U + AU + BU = 0, \quad t > 0, \\ &U(0) = U_0, \end{aligned}\right\} \tag{1.101}$$

for which the exact solution is given by

$$U(t) = \mathrm{e}^{-t(A+B)} U_0, \qquad t \geq 0. \tag{1.102}$$

The first order Lie and the second order Strang splitting formulae are given, for instance, by

$$\mathcal{L}_2^t U_0 = \mathrm{e}^{-tA}\mathrm{e}^{-tB} U_0, \tag{1.103}$$

and

$$\mathcal{S}_2^t U_0 = \mathrm{e}^{-tA/2}\mathrm{e}^{-tB}\mathrm{e}^{-tA/2} U_0. \tag{1.104}$$

In this context, it was proved in (Descombes and Schatzman 2002) that the following identities hold:

$$\mathcal{L}_2^t \;=\; \mathrm{e}^{-t(A+B)} + \int_0^t \int_0^s \mathrm{e}^{-(t-s)(A+B)} \mathrm{e}^{-(s-r)A} \big(\partial_A B\big) \mathrm{e}^{-rA} \mathrm{e}^{-sB} \, \mathrm{d}r \, \mathrm{d}s, \tag{1.105}$$

$$\begin{aligned} \mathcal{S}_2^t \;=\; & \mathrm{e}^{-t(A+B)} + \\ & \frac{1}{4} \int_0^t \int_0^s (s-r) \mathrm{e}^{-(t-s)(A+B)} \mathrm{e}^{-(s-r)A/2} \big(\partial_A^2 B\big) \mathrm{e}^{-rA/2} \mathrm{e}^{-sB} \mathrm{e}^{-sA/2} \, \mathrm{d}r \, \mathrm{d}s \\ & -\frac{1}{2} \int_0^t \int_0^s (s-r) \mathrm{e}^{-(t-s)(A+B)} \mathrm{e}^{-sA/2} \mathrm{e}^{-rB} \big(\partial_B^2 A\big) \mathrm{e}^{-(s-r)B} \mathrm{e}^{-sA/2} \, \mathrm{d}r \, \mathrm{d}s. \end{aligned} \tag{1.106}$$

These new estimates provide then an exact representation of the local errors, comparing with previous estimates for $\mathcal{L}_2^t$ (1.30) and $\mathcal{S}_2^t$ (1.37). It follows the same for $\mathcal{L}_1^t$ and $\mathcal{S}_1^t$.

In order to illustrate the influence of space multi-space phenomena given, for instance, by high spatial gradients in the solutions of the PDEs, we will consider a simplified scalar reaction-diffusion system coming from (1.67), with $m = 1$ and $d = 1$:

$$\left.\begin{aligned} &\partial_t u - \partial_x^2 u + V(x) u = 0 \quad && x \in \mathbb{R}, t > 0, \\ &u(x,0) = u_0(x) && x \in \mathbb{R}, \end{aligned}\right\} \tag{1.107}$$

where $V : \mathbb{R} \to \mathbb{R}$ is supposed to be a positive and bounded function of class $\mathcal{C}^\infty(\mathbb{R})$ with all bounded derivatives, and the L^2-norm of the derivative of the smooth initial condition u_0 is assumed to be very high. Similar systems were considered in (Descombes, Dumont, Louvet, and Massot 2007; Duarte, Descombes, and Massot 2011; Descombes, Duarte, Dumont, Laurent, Louvet, and Massot 2014) where in particular V can be seen as coming from the linearizion of the corresponding scalar reaction term $f(u)$ in (1.67). Considering that the linear operator A in (1.101) corresponds to the multiplication by V and that $B = -\partial_x^2$ (minus the second partial derivative with respect to x in one dimension), their commutator (1.21) is given by

$$\partial_A B = [A, B] = (\partial_x^2 V) + 2(\partial_x V)\partial_x. \tag{1.108}$$

If we now define

$$\mathrm{E}_{\mathcal{L}_2}^t = \mathrm{e}^{t(\partial_x^2 - V)} - \mathrm{e}^{-tV}\mathrm{e}^{t\partial_x^2}, \tag{1.109}$$

and consider (1.105), we can write the local error associated with the $\mathcal{L}_2^t$ scheme for system (1.107) as

$$\mathrm{E}^t_{\mathcal{L}_2} u_0 = -\int_0^t \int_0^s \mathrm{e}^{-(t-s)(\partial_x^2 - V)} \mathrm{e}^{-(s-r)V} \big(\partial_A B\big) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 \,\mathrm{d}r\,\mathrm{d}s, \tag{1.110}$$

with commutator $\partial_A B$ given by (1.108). Taking norms, we have that in $L^2(\mathbb{R})$:

$$\begin{aligned} \big\|\mathrm{E}^t_{\mathcal{L}_2} u_0\big\|_{L^2} &\leq \int_0^t \int_0^s \Big\| \mathrm{e}^{-(t-s)(\partial_x^2 - V)} \mathrm{e}^{-(s-r)V} \big(\partial_A B\big) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 \Big\|_{L^2} \,\mathrm{d}r\,\mathrm{d}s \\ &\leq \int_0^t \int_0^s \Big\| \big(\partial_A B\big) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 \Big\|_{L^2} \,\mathrm{d}r\,\mathrm{d}s. \end{aligned} \tag{1.111}$$

Since

$$\begin{aligned} \big(\partial_A B\big) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 &= (\partial_x^2 V) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 + 2(\partial_x V)\partial_x \left(\mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0\right) \\ &= (\partial_x^2 V) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 - 2(\partial_x V) r (\partial_x V) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 \\ &\quad + 2(\partial_x V) \mathrm{e}^{-rV} \partial_x \left(\mathrm{e}^{s\partial_x^2} u_0\right) \\ &= (\partial_x^2 V) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 - 2(\partial_x V) r (\partial_x V) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 \\ &\quad + 2(\partial_x V) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} \partial_x u_0, \end{aligned} \tag{1.112}$$

the integration of (1.111) yields

$$\big\|\mathrm{E}^t_{\mathcal{L}_2} u_0\big\|_{L^2} \leq \left(\frac{t^2}{2} \|\partial_x^2 V\|_\infty + \frac{t^3}{3} \|\partial_x V\|_\infty^2 \right) \|u_0\|_{L^2} + t^2 \|\partial_x V\|_\infty \|\partial_x u_0\|_{L^2}. \tag{1.113}$$

Nevertheless, we have supposed that the L^2-norm of $\partial_x u_0$ is very high, therefore the latter error bound is only interesting if the splitting time step t is sufficiently small. It is then specially relevant in this stiff configuration to obtain alternative error estimates which do not involve the derivative of the initial condition (Descombes, Dumont, Louvet, and Massot 2007). Thanks to the regularizing effect of the Laplacian, we can demonstrate through a Fourier transform of the diffusion operator, that for all $u_0 \in L^2$ and for $t > 0$:

$$\|\partial_x \mathrm{e}^{t\partial_x^2} u_0\|_{L^2} \leq \frac{1}{\sqrt{2\mathrm{e}t}} \|u_0\|_{L^2}. \tag{1.114}$$

Therefore, taking into account that

$$\begin{aligned} \big(\partial_A B\big) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 &= (\partial_x^2 V) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 - 2(\partial_x V) r (\partial_x V) \mathrm{e}^{-rV} \mathrm{e}^{s\partial_x^2} u_0 \\ &\quad + 2(\partial_x V) \mathrm{e}^{-rV} \partial_x \left(\mathrm{e}^{s\partial_x^2} u_0\right), \end{aligned} \tag{1.115}$$

into (1.111), its integration now yields

$$\big\|\mathrm{E}^t_{\mathcal{L}_2} u_0\big\|_{L^2} \leq \left(\frac{4}{3} t\sqrt{t} \frac{\|\partial_x V\|_\infty}{\sqrt{2\mathrm{e}}} + \frac{t^2}{2} \|\partial_x^2 V\|_\infty + \frac{t^3}{3} \|\partial_x V\|_\infty^2 \right) \|u_0\|_{L^2}. \tag{1.116}$$

An order reduction is thus shown to appear in the local error estimate (Descombes, Dumont, Louvet, and Massot 2007). Similar conclusions are drawn considering the $\mathcal{L}_1^t$-Lie scheme, explicit computations of the estimates can be found in (Duarte, Descombes, and Massot 2011). Estimates (1.113) and (1.116) describe then the behavior of the local errors, and we see that for $t > 0$:

$$\left\|\mathrm{E}^t_{\mathcal{L}_2} u_0\right\|_{L^2} \propto \left(\|\partial_x u_0\|_{L^2} t^2, \|u_0\|_{L^2} t^{1.5}\right). \tag{1.117}$$

The first term is more relevant when t is sufficiently small, whereas the second one when t is not small enough and $\|\partial_x u_0\|_{L^2}$ is very high. More precisely, there exists some constant $\theta > 0$ such that for $t \leq \theta$, $\|\mathrm{E}^t_{\mathcal{L}_2} u_0\|_{L^2}$ behaves like t^2 and for $t \geq \theta$, $\|\mathrm{E}^t_{\mathcal{L}_2} u_0\|_{L^2}$ behaves like $t^{1.5}$ (Descombes, Dumont, Louvet, and Massot 2007; Duarte, Descombes, and Massot 2011; Descombes, Duarte, Dumont, Laurent, Louvet, and Massot 2014).

In the same way, defining for the $\mathcal{S}^t_2$-Strang scheme

$$\mathrm{E}^t_{\mathcal{S}_2} = \mathrm{e}^{t(\partial_x^2 - V)} - \mathrm{e}^{-tV/2}\mathrm{e}^{t\partial_x^2}\mathrm{e}^{-tV/2}, \tag{1.118}$$

and considering (1.106), we can also write the local error associated with the $\mathcal{S}^t_2$ scheme for system (1.107). An order reduction can be once again detected and estimated for these stiff configurations. The explicit computations are shown in (Duarte, Descombes, and Massot 2011), that finally yield

$$\left\|\mathrm{E}^t_{\mathcal{S}_2} u_0\right\|_{L^2} \propto \left(\|\partial_x u_0\|_{L^2} t^3, \|u_0\|_{L^2} t^2\right), \tag{1.119}$$

so that the local error $\|\mathrm{E}^t_{\mathcal{S}_2} u_0\|_{L^2}$ behaves either like t^3 for small splitting time steps or like t^2 with a consequent order reduction of the scheme.

It can thus be seen through these theoretical illustrations that an order reduction may arise for both Lie and Strang schemes whenever the solution features high spatial gradients. On the other hand, the hypothesis of a linear source term in (1.107) have just allowed us to simplify the computations and to better target the analysis on the effects of the diffusion operator on the solution. These theoretical estimates were validated through some numerical tests presented in (Descombes, Dumont, Louvet, and Massot 2007; Duarte, Descombes, and Massot 2011; Descombes, Duarte, Dumont, Laurent, Louvet, and Massot 2014) for stiff problems coming from nonlinear chemical dynamics. Taking into account that in the numerical applications envisioned in this work some of them are characterized by propagating fronts with potentially steep spatial gradients, an influence of the formers may be observed in the accuracy order of the splitting schemes. More precisely, an order reduction will likely arise for both Lie and Strang formulae for sufficiently large splitting time steps Δt. Nevertheless, the mathematical description introduced in these studies confirms that from a practical point of view the splitting errors are still set by the splitting time step even for this type of stiff configuration, whereas on the other hand a more precise theoretical understanding of the splitting errors for non asymptotic regimes was achieved. Finally, as in the previous mathematical descriptions, the numerical solvers implemented in practice should solve correctly the time evolution associated with each operator. For instance, Ropp & Shadid showed in (Ropp and Shadid 2005; Ropp and Shadid 2009) that better results are obtained when using an L-stable method for the numerical solution of the diffusion in, respectively, reaction-diffusion and reaction-diffusion-convection problems[6].

[6] We will see in the following chapter that L-stability allows us to rapidly damp out fast numerical transients associated in this particular case with high frequencies or wave numbers arising when the discretized Laplacian operator is applied to a given solution (see, *e.g.,* (Hairer and Wanner 1996; Hundsdorfer and Verwer 2003)).

Chapter 2

Runge-Kutta Methods for Time Integration of Stiff ODEs

In the last chapter, we have first considered splitting techniques for the solution of linear systems of ODEs of type (1.1), with a general mathematical description on the numerical errors of such methods. A formal extension to general nonlinear systems was also detailed by means of the Lie operator formalism. We have then discussed the numerical solution by splitting methods of stiff PDEs for reaction-diffusion systems like (1.66), modeling potentially multi-scale phenomena. A theoretical characterization of the splitting errors was thus presented in the context of time and space stiff reaction-diffusion problems, which has introduced a few criteria to take into account, even for more complex PDEs. Even though the latter studies have led to the description of some numerical difficulties issued from the modeling PDEs, we have not given any detail on the solution of the split subproblems. Actually, throughout all these analyses we have assumed that the subsystems of equations were exactly solved in order to characterize only numerical errors coming from the splitting scheme. In this way, we have not considered yet either the time or space discretizations, or the numerical time integration of the associated subproblems. Nevertheless, it is quite natural to expect that the same numerical features of these modeling equations that influence the splitting accuracy, will also be present during the numerical solution of each split subproblem.

We have seen that in the context of splitting techniques we aim at solving independently and successively different time dependent systems of equations, starting from the immediately previous numerical solution. Hence, several initial value problems or Cauchy problems for PDEs are to be considered within each splitting time step. Therefore, in this chapter we will focus on the so-called *one-step integration methods* which contrarily to *multi-step methods*, do not require initial lower order approximations to build the numerical solution of each initial value problem. In this way, in this chapter we will first characterize some numerical difficulties associated with the solution of the ODEs issued from the previous problems to then describe some one-step Runge-Kutta methods that were developed in the past years to efficiently cope with these matters. In particular, we will concentrate on implicit and stabilized explicit Runge-Kutta schemes that have shown to be very efficient for the numerical solution of, respectively, reaction and diffusion problems, as an illustration of proper selection criteria of time integration solvers for the split subproblems issued from a splitting technique. For further details, an exhaustive mathematical description and analysis on the numerical solution of stiff systems of ODEs can be found in the book of Hairer & Wanner (Hairer and Wanner 1996).

2.1 Characterization of Stiffness

Let us consider for $t > 0$, the scalar initial value problem:

$$\left.\begin{aligned} \mathrm{d}_t u &= f(t, u(t)), \\ u(0) &= u_0, \end{aligned}\right\} \tag{2.1}$$

with some $u_0 \in \mathbb{R}$ and $u : \mathbb{R} \to \mathbb{R}$, $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. We aim at obtaining a numerical approximation u_n of the exact solution $u(t_n)$ of (2.1) for a time discretization given by $t_0 = 0 < t_1 < \ldots < t_n < \ldots$, and $n = 0, 1, \ldots$.

Nevertheless, we assume, and therefore we must take into account, that (2.1) is a stiff problem for which a precise and simple notion of stiffness is given in (Hairer and Wanner 1996):

"Stiff equations are problems for which explicit methods don't work."

In order to illustrate this, we will first approximate the solution of (2.1) at some $t_1 = t_0 + \delta t$

$$u(\delta t) = u_0 + \int_{t_0}^{t_0+\delta t} f(t, u(t))\, \mathrm{d}t, \tag{2.2}$$

by

$$u_1 = u_0 + \delta t f(t_0, u_0), \tag{2.3}$$

which implies an explicit time discretization solution of (2.2) and it is known as the *explicit Euler* method, where δt is defined as the integration time step. It is straightforward to see that this is a first order method according to (1.6).

Taking a very simple case for (2.1), given by

$$\left.\begin{aligned} \mathrm{d}_t u &= -100\, u, \\ u(0) &= u_0, \end{aligned}\right\} \tag{2.4}$$

with exact solution $u(\delta t) = \mathrm{e}^{-100\,\delta t} u_0$ at $t_1 = \delta t$. We have that u_1 computed by (2.3) is given by

$$u_1 = u_0 - 100\,\delta t u_0. \tag{2.5}$$

If we set, for instance, an initial condition $u_0 = 1$, and a relatively small time step of $\delta t = 0.5$ compared with 100, the exact and numerical solutions give, respectively, $u(0.5) = \mathrm{e}^{-50} \approx 1.9 \times 10^{-22}$ and $u_1 = -49$. And integrating over another time step δt: $u(1) = \mathrm{e}^{-100} \approx 3.7 \times 10^{-44}$ and $u_2 = 2401$. It follows then that the explicit time discretization given by (2.3) is not capable of reproducing the right dynamics given by the exact solution. However, since this solution models a rapid transition from u_0 towards a final equilibrium value, we can easily identify the associated time scale $\tau = 1/100 = 0.01$ of the transient phase and therefore, we can expect that integration time steps δt of the order or smaller than τ will be capable to track the right dynamics. For instance, for $\delta t = 0.001$, we have $u(0.001) = \mathrm{e}^{-0.1} \approx 0.904837418$ and $u_1 = 0.9$, and $u(0.002) = \mathrm{e}^{-0.2} \approx 0.818730753$ and $u_2 = 0.81$. These rapid variations or transients associated with fast scales are typical of stiff equations, but they are neither sufficient nor necessary to qualify them as stiff. Actually, an initial condition u_0 close enough to the equilibrium manifold of the solution will not develop such fast transients, and thus stiff features may not be observed.

As a first conclusion, we can deduce that an explicit time discretization scheme to solve (2.4) will generally fail to approach the right dynamics, unless we consider integration time steps smaller than the time scales disclosed by the equations. This may seem natural. Nevertheless, if we consider the counter-part of (2.3), *i.e.,* an *implicit Euler* method, also of order 1:

$$u_1 = u_0 + \delta t f(t_1, u_1), \tag{2.6}$$

and the previous $\delta t = 0.5$, we obtain the numerical approximations $u_1 = 0.019607843$ and $u_2 = 0.000384468$. Therefore, although solutions are not quite accurate, they show convergence towards the right solution with a time step several times the associated time scale. As a second conclusion, we can then add that both explicit and implicit schemes are of the same order, and would therefore yield results of the same accuracy for sufficiently small time steps. From a time step larger than a given value, the explicit method will not deliver any valid result.

2.1.1 Some Typical Stiff Configurations

If we now consider a general nonlinear system

$$\mathrm{d}_t U = F(U) \tag{2.7}$$

with $U : \mathbb{R} \to \mathbb{R}^m$, $F : \mathbb{R}^m \to \mathbb{R}^m$ and define a solution $\varphi(t) \in \mathbb{R}^m$ such that $\mathrm{d}_t\varphi(t) = F(\varphi(t))$, we can linearize F in its neighborhood:

$$\mathrm{d}_t U = F(\varphi(t)) + \partial_U F(\varphi(t))\,(U(t) - \varphi(t)) + \mathcal{O}\left((U(t) - \varphi(t))^2\right), \tag{2.8}$$

to obtain

$$\mathrm{d}_t\overline{U} = J\overline{U}, \tag{2.9}$$

where higher order terms in $\overline{U}(t) := U(t) - \varphi(t)$ are neglected, and with the Jacobian: $J(U) = \partial_U F(U)$. Supposing a constant Jacobian that is moreover diagonalizable, we can write the i-th component $\overline{u}^{(i)}(t)$ of $\overline{U}(t)$, solution of (2.9), as

$$\overline{u}^{(i)}(t) = \sum_{i=1}^{m} c_i \mathrm{e}^{\lambda_i t}\overline{u}_0^{(i)}, \tag{2.10}$$

for some initial condition $\overline{U}_0 \in \mathbb{R}^m$ and constants c_i, where the λ_i are the corresponding eigenvalues associated with J. Therefore, we can see that the solutions $\overline{u}^{(i)}(t)$ of (2.9) are clearly reproduced by a linear combination of

$$\left(\mathrm{e}^{\lambda_i t}\overline{u}_0^{(i)}\right)_{i=1,2,\ldots,m}, \tag{2.11}$$

that is, solutions of the same type as for the previous linear problem (2.4), and thus the latter simpler case mimics somehow the dynamics of more general nonlinear problems. We can then expect the same behavior previously described for explicit and implicit schemes, depending in this case on the spectrum of the Jacobian J and the set of initial conditions $\overline{u}_0^{(i)}$, $i = 1, 2, \ldots, m$.

As a consequence, if (2.7) is a stiff system of ODEs, then it is very likely that some λ_i with large negative real part $\operatorname{Re}\lambda_i \leq 0$, will take a leading role in the transient phase of the solution, whenever the initial solution does not belong to a partial equilibrium manifold where the fast scales are already relaxed. In particular, not only large eigenvalues will generate the fast variations previously discussed, but also an important dispersion of the eigenvalues in the spectrum of J will certainly induce multi-scale dynamics issued from the composition of the various time scales (or eigenvalues) included in (2.10). This is a typical situation for example in the context of chemical reaction systems modeling a set of reactions with different reaction scales, and hence time scales for which fast projection of some species onto equilibrium manifolds are usually developed (see*, e.g.,* (Maas and Pope 1992)). These systems are usually very stiff and moreover, the stiffness increases with the precision and the detail of description of the mathematical model.

Another classical example of a stiff problem, where stiffness is not necessarily related to the presence of fast variables, is given by the system:

$$\mathrm{d}_t U = AU, \tag{2.12}$$

with $U : \mathbb{R} \to \mathbb{R}^m$, $A \in \mathcal{M}_m(\mathbb{R})$:

$$A = \frac{1}{\Delta x^2}\begin{pmatrix} -1 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -1 \end{pmatrix} \tag{2.13}$$

and $\Delta x = 1/(N_x + 1)$, issued from the spatial discretization on a grid of $N_x = m$ points with second order centered finite differences for the heat equation:

$$\partial_t u - \partial_x^2 u = 0, \tag{2.14}$$

for $x \in [0,1]$ and homogeneous Neumann conditions.

As previously seen, the solution of (2.14) in infinite dimension is given by

$$u(x,t) = \mathrm{e}^{t\partial_x^2} u_0(x), \tag{2.15}$$

for some initial condition $u_0(x)$, where the associated spectrum of the differential operator is given by the whole set of numbers in the negative real axis. Furthermore, performing a Fourier transform in the x direction

$$\hat{f}(k) = \mathcal{F}(f) := \int_{-\infty}^{\infty} \mathrm{e}^{-ikx} f(x)\,\mathrm{d}x, \tag{2.16}$$

of the heat equation (2.14) yields as solution:

$$\hat{u}(k,t) = \mathrm{e}^{-k^2 t}\hat{u}_0(k). \tag{2.17}$$

Hence, a strong damping of the highest frequencies given by the frequency parameter[1] k will arise and will smoothen the initial condition. This is typical of diffusion problems. The analogy with the previous linear case (2.4) can be then established by this analysis for which in particular, we see that the frequency composition of the initial condition u_0, will or will not activate these fast decays, similar to (2.4). As a consequence, we can directly relate the stiffness associated with equation (2.14) to the presence of high gradients or discontinuities in x in $u_0(x)$. For instance, if we consider an extreme case for which $u_0(x) = \delta(x)$, *i.e.,* the *Dirac delta function*, all the frequency spectrum will appear on (2.17) with fast decays, since $\hat{u}_0(k) = 1$.

Coming back to the discretized problem (2.12) which is the one that will be numerically integrated, we can infer that the discretized counter-part mimics the previous theoretical analysis. This is reflected, for instance, by the spectrum of the matrix A:

$$\lambda_j = -\frac{4}{\Delta x^2}\sin^2\left(\frac{\pi j \Delta x}{2}\right), \quad j = 1, \dots, N_x, \tag{2.18}$$

or alternatively,

$$\lambda_j = -4(N_x+1)^2 \sin^2\left(\frac{\pi j}{2(N_x+1)}\right), \quad j = 1, \dots, N_x, \tag{2.19}$$

for which we can identify potentially large eigenvalues increasing quadratically with the number of discretization points N_x with a maximum dispersion between $-4(N_x+1)^2$ and 0, which explains the spurious patterns found in some numerical approximations for this kind of problem (see some illustrations in(Hairer and Wanner 1996)). In particular, we see that finer discretizations that introduce naturally more resolution scales, result also in broader spectra to represent them. Once again, these large eigenvalues will arise in the global solution depending on the distribution of the initial conditions. Stiffer behavior will then take place for discontinuous or large variations within the initial distributions.

With this brief introduction and illustrations, we introduce in the following the so-called Runge-Kutta (RK) time integrations method, as well as some dedicated RK schemes conceived to handle stiff systems of ODEs.

2.2 Runge-Kutta Time Integration Methods

We have previously considered the explicit Euler method given by (2.3). This kind of method is called *one-step integration method* because we aim at recursively approximating the exact solution

[1] Also referred to as the wave number.

(2.2) after one time step, based on the previous one. The Euler scheme is of first order but by approximating the integral in (2.2) by a higher order quadrature formula, we can build higher order one-step methods. A second order scheme can be constructed, for instance, by using the mid-point approximation:

$$u_1 = u_0 + \delta t f\left(t_0 + \frac{\delta t}{2}, u\left(t_0 + \frac{\delta t}{2}\right)\right), \tag{2.20}$$

and the Euler method, which leads to the *Runge method*:

$$u_1 = u_0 + \delta t f\left(t_0 + \frac{\delta t}{2}, u_0 + \frac{\delta t}{2} f(u_0)\right). \tag{2.21}$$

Generalizing this idea with higher order quadrature formulae leads to define the so-called s-stage *Runge-Kutta methods*:

$$\left.\begin{aligned} g_i &= u_0 + \delta t \sum_{j=1}^{s} a_{ij} f\left(t_0 + c_j \delta t, g_j\right), \quad i = 1, \dots, s; \\ u_1 &= u_0 + \delta t \sum_{j=1}^{s} b_j f\left(t_0 + c_j \delta t, g_j\right), \end{aligned}\right\} \tag{2.22}$$

for which the arrays b, $c \in \mathbb{R}^s$ gather the various coefficients $b = (b_1, \dots, b_s)^T$ and $c = (c_1, \dots, c_s)^T$, and $A \in \mathcal{M}_s(\mathbb{R})$ such that $A = (a_{ij})_{1 \leq i,j \leq s}$. These coefficients are usually arranged in a mnemonic device, known as a Butcher tableau:

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

For instance, for the Runge method (2.21), we have

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

When $a_{ij} = 0$ for $j \geq i$, the scheme is explicit in time (*Explicit RK methods*, ERK) with

$$g_i = u_0 + \delta t \sum_{j=1}^{i-1} a_{ij} f\left(t_0 + c_j \delta t, g_j\right), \quad i = 1, \dots, s, \tag{2.23}$$

in (2.22), whereas the case for which $a_{ij} = 0$ for $j > i$ and at least one of the diagonal coefficients is non-zero, $a_{ii} \neq 0$, is defined as a *Diagonal Implicit RK method* (DIRK). Otherwise, we are considering *Implicit RK methods* (IRK). We will further describe these schemes in the following, but first, we will introduce some basic theoretical and numerical properties for general RK methods: the order and stability features, as well as the choice of the time steps of integration.

2.2.1 Order and Stability of Runge-Kutta Schemes

We now consider the *Dahlquist test equation* (Dahlquist 1963):

$$\left.\begin{aligned} \mathrm{d}_t u &= \lambda u, \\ u(0) &= 1, \end{aligned}\right\} \tag{2.24}$$

with $\lambda \in \mathbb{C}$ (a particular case was given by (2.4)), and we can successively compute the g_j of the explicit RK method (2.23) for problem (2.24). We obtain

$$u_1 = R(z) u_0, \quad z = \delta t \lambda, \tag{2.25}$$

where

$$R(z) = 1 + z \sum_j b_j + z^2 \sum_{j,k} b_j a_{j,k} + \ldots, \tag{2.26}$$

is a polynomial of degree $\leq s$. If the RK method is of order p we know that $u_1 = R(z)u_0$ must satisfy

$$\mathrm{e}^z - R(z) = \mathcal{O}(\delta t^{p+1}) = \mathcal{O}(z^{p+1}), \tag{2.27}$$

where $\mathrm{e}^z u_0$ is the exact solution of (2.24), and thus $R(z)$ is given by

$$R(z) = 1 + z + \frac{z^2}{2!} + \ldots + \frac{z^p}{p!} + \mathcal{O}(z^{p+1}). \tag{2.28}$$

In particular, for all explicit RK methods of order p with $s = p$ intermediate stages, we have

$$R(z) = 1 + z + \frac{z^2}{2!} + \ldots + \frac{z^s}{s!}. \tag{2.29}$$

A classical analysis based on the Dahlquist test equation (2.24) allows us to define $R : \mathbb{C} \to \mathbb{C}$ given in general by (2.25), as the *stability function* of a given method. That is, $R(z)$ is the numerical solution of (2.24) given by the method itself after one time step δt. Furthermore, the numerical solution recursively computed can be written as

$$u_n = \left(R(z)\right)^n u_0 \tag{2.30}$$

which allows us to define the *stability domain* of the method given by the set of z for which u_n remains bounded for $n \to \infty$, *i.e.,*

$$S := \left\{ z \in \mathbb{C} \text{ s.t. } |R(z)| \leq 1 \right\}. \tag{2.31}$$

For instance, considering the explicit Euler method (2.3) for which

$$R(z) = 1 + z, \tag{2.32}$$

according to (2.28), its stability domain S is given by all $z \in \mathbb{C}$ such that

$$|1 + z| = |z - (-1)| \leq 1, \tag{2.33}$$

which is the circle of radius 1 and center -1 in the complex plane. Coming back to the previous example (2.4) with $\lambda = -100$, we can see that an explicit Euler method will remain stable as long as $z = \delta t \lambda \in S$, *i.e.,* $0 \leq \delta t \leq 2/100$, which explains the previous bad results for $\delta t = 0.5$. Alternatively, considering the implicit Euler method (2.6) yields

$$R(z) = \frac{1}{1 - z}, \tag{2.34}$$

as stability function, with stability domain given by all $z \in \mathbb{C}$ such that

$$\left|\frac{1}{1-z}\right| \leq 1 \;\Rightarrow\; |z-1| \geq 1, \tag{2.35}$$

that is, the exterior of the circle with radius 1 and center +1 in the complex plane. For problem (2.4), we can then see that $R(z)$ will remain bounded for any time step $\delta t > 0$, as it is shown by $(R(z = -100\,\delta t))^n = (1 + 100\,\delta t)^{-n}$ into (2.30). This better performance of an implicit discretization for large negative λ into (2.24), characteristic of stiff ODEs, leads us to give more details on these schemes in a forthcoming section. In particular, it was demonstrated that for $p \geq 5$ there is no explicit RK method of order p with $s = p$ stages (Butcher 1964c; Butcher 1964d). This and other order constraints for explicit RK schemes are known as the *Butcher Barriers* (see more details in (Hairer, Nørsett, and Wanner 1987)). Finally, it is important to recall that in a general case, we can perform the same analysis on the linearized problem (2.9), similar to the Dahlquist test equation, taking into account the complex eigenvalues $\lambda_i, i = 1, \cdots, m$, of the associated Jacobian J.

2.2.2 Time Step Selection

Whether the time discretization schemes are explicit or implicit, or if the orders of approximations are high or low, a key question for a numerical time integration method is the choice of the time step of integration. We have seen, for instance, that for stiff problems, explicit methods should consider rather small time steps to guarantee the stability of computations. However, for a given problem if we suppose that we are only considering time steps contained in the stability domain, the former ones must be chosen such that the numerical solutions yield approximations within a desired accuracy. In this case, a constant time step might be sufficient for some kind of problem to efficiently solve the corresponding dynamics. In a more general context, more sophisticated techniques must be consider to dynamically select these time steps in order to render computations efficient or even possible in practice. In any of both cases, the main goal is to choose a time step δt such that the local error verifies

$$\|u(\delta t) - u_1\| = C\delta t^{p+1} \leq Tol, \tag{2.36}$$

where *Tol* is the desired accuracy requested to the numerical computations. It is straightforward to see that higher order methods would satisfy (2.36) with larger time steps. Furthermore, for a given scheme the expression (2.36) might be satisfied with time steps evolving in time. For problems describing different dynamics, having an adaptive time step strategy would then involve important savings of numerical work. In this context, a lot of research has been conducted to develop time step control or adaptive time stepping techniques. A review of some explicit solvers with automatic time step selection can be found in (Hairer, Nørsett, and Wanner 1987) for non stiff problems. A complementary idea developed for explicit schemes was to use these control techniques to automatically detect stiffness (see*, e.g.,* (Shampine 1977; Shampine and Hiebert 1977; Hairer and Wanner 1996)) in order to automatically switch to a more suitable method.

One of the most standard ways of time stepping is based on computing a numerical approximation: *err*, of the exact local error in (2.36), by considering a solution $\hat{u}_1$ computed by a lower order method of order $\hat{p} < p$ (Hairer, Nørsett, and Wanner 1987), such that

$$\|u(\delta t) - u_1\| \lessapprox err = \|u_1 - \hat{u}_1\|. \tag{2.37}$$

Since

$$u_1 - \hat{u}_1 = (u_1 - u(\delta t)) - (\hat{u}_1 - u(\delta t)) = \mathcal{O}(\delta t^{p+1}) + \mathcal{O}(\delta t^{\hat{p}+1}) \approx \mathcal{O}(\delta t^{\hat{p}+1}), \tag{2.38}$$

and thus,

$$err \approx \tilde{C}\delta t^{\hat{p}+1}, \tag{2.39}$$

we can suppose that the optimal time step δt_{opt} such that $err \approx Tol$:

$$Tol \approx \tilde{C}\delta t_{\text{opt}}^{\hat{p}+1}, \tag{2.40}$$

is given by

$$\delta t_{\text{opt}} = fac \cdot \delta t \left(\frac{Tol}{err}\right)^{1/\hat{p}+1}, \tag{2.41}$$

where *fac* is a safety factor usually close to 1.

In this way, we can compute the time step needed to integrate problem (2.1) with a local accuracy given by *Tol*, where the $\hat{p}$-order method should be embedded into the p-order method in order to minimize the required number of operations. Additionally, we can use the expression (2.41) to dynamically compute the time steps in time. In this case, we use the computations at the n-th step to predict the error at the next step:

$$err_{n+1} = \|u_n - \hat{u}_n\| \approx \tilde{C}_n \delta t_n^{\hat{p}+1}, \tag{2.42}$$

which yields as new time step:

$$\delta t_{\text{new}} = fac \cdot \delta t_n \left(\frac{Tol}{err_{n+1}}\right)^{1/\hat{p}+1}, \tag{2.43}$$

by assuming $\tilde{C}_{n+1} \approx \tilde{C}_n$ into

$$Tol \approx \tilde{C}_{n+1}\delta t_{\text{new}}^{\hat{p}+1}. \tag{2.44}$$

The next step δt_{n+1} will be then given by δt_{new} if $err_{n+1} \leq Tol$. Alternatively, the current n-th time step will be rejected if $err_{n+1} > Tol$, and in this case the procedure works as an *a posteriori* verification where the same n-th step will be integrated again with the new time step δt_{new}.

Based on the same ideas and on more rigorous theoretical studies carried out by Gustafsson (Gustafsson 1994), a better procedure assumes that $\log C_n$ is a linear function of n, and thus $\log C_{n+1} - \log C_n$ is constant or, equivalently (Hairer and Wanner 1996)

$$\frac{C_{n+1}}{C_n} \approx \frac{C_n}{C_{n-1}}, \tag{2.45}$$

which finally yields

$$\delta t_{\text{new}} = fac \cdot \delta t_n \left(\frac{Tol}{err_{n+1}}\right)^{1/\hat{p}+1} \frac{\delta t_n}{\delta t_{n-1}} \left(\frac{err_n}{err_{n+1}}\right)^{1/\hat{p}+1}. \tag{2.46}$$

This technique is also known as the *step size strategy with memory* of Watts (Watts 1984) and Gustafsson (Gustafsson 1994), and usually shows better performances than the standard technique (2.43). In particular, it allows us fast reduction of time steps without rejection in the context of stiff problems (Hairer and Wanner 1996). There are other step size control techniques to numerically estimate or predict local errors and therefore, to guarantee a given accuracy of computations according to (2.36). We mention, for instance, time step computations using extrapolation techniques (Deuflhard 1983; Shampine 1987), or theoretical or numerical estimates of the leading term of the local error expansion (Hindmarsh 1980; Sommeijer, Shampine, and Verwer 1997).

2.3 Implicit Runge-Kutta Methods

Let us consider now the implicit RK scheme (2.22). We apply it to the Dahlquist test equation (2.24), and we obtain

$$\left.\begin{aligned} g &= u_0\mathbb{1} + \delta t\lambda A g, \\ u_1 &= u_0 + \delta t\lambda b^T g, \end{aligned}\right\} \tag{2.47}$$

with $g = (g_1, \dots, g_s)^T$ and $\mathbb{1} = (1, \dots, 1)^T$. The linear system for $g_1, \dots, g_s$ gives

$$g = (\mathrm{Id} - \lambda\delta t A)^{-1} u_0 \mathbb{1}, \tag{2.48}$$

and the corresponding stability function may be written as

$$R(z) = 1 + z b^T (\mathrm{Id} - zA)^{-1}\mathbb{1}. \tag{2.49}$$

However, a better representation might be obtained by considering the solution of (2.47):

$$\begin{pmatrix} \mathrm{Id} - zA & 0 \\ -zb^T & 1 \end{pmatrix} \begin{pmatrix} g \\ u_1 \end{pmatrix} = u_0 \begin{pmatrix} \mathbb{1} \\ 1 \end{pmatrix}, \tag{2.50}$$

using the Cramer's rule:

$$u_1 = \frac{\det \begin{pmatrix} \mathrm{Id} - zA & u_0\mathbb{1} \\ -zb^T & u_0 \end{pmatrix}}{\det \begin{pmatrix} \mathrm{Id} - zA & 0 \\ -zb^T & 1 \end{pmatrix}}, \tag{2.51}$$

and taking into account that

$$\det \begin{pmatrix} \mathrm{Id} - zA & \mathbb{1} \\ -zb^T & 1 \end{pmatrix} = \det \begin{pmatrix} \mathrm{Id} - zA + z\mathbb{1}b^T & 0 \\ -zb^T & 1 \end{pmatrix} = \det\left(\mathrm{Id} - zA + z\mathbb{1}b^T\right). \tag{2.52}$$

This yields

$$R(z) = \frac{P(z)}{Q(z)} = \frac{\det\left(\mathrm{Id} - zA + z\mathbb{1}b^T\right)}{\det(\mathrm{Id} - zA)}, \tag{2.53}$$

so we can see that for implicit RK schemes, the stability function $R(z)$ becomes a rational function with polynomial numerator $P(z)$ and denominator $Q(z)$ of degree less than or equal to s.

A direct consequence of this rational stability function as seen for the implicit Euler method (2.6), is that the associated schemes can be stable on the entire left-half plane $\mathbb{C}^-$. This set of z corresponds precisely to eigenvalues of negative real part for which the exact solutions are bounded in time $|e^z| \le 1$ and for which we have seen before, the numerical method should preserve this stability property.

A method is then called *A-stable* if its stability domain satisfies (Dahlquist 1963)

$$S \supset \{z \in \mathbb{C} \text{ s.t. } \mathrm{Re}\, z \le 0\}. \tag{2.54}$$

For instance, the implicit Euler method (2.6) is A-stable. Even though this is a desirable and necessary stability property to properly handle stiff problems, it is not sufficient for very stiff problems. For eigenvalues with very large real part, the stability function $R(z)$ of an A-stable method will surely keep the numerical approximations bounded during the fast transients. Nevertheless, only a $R(z)$ much smaller that 1, can guarantee that the numerical solutions will rapidly approach the exact solution, damping out the numerical transients phases. Numerical methods with such a property are known as *L-stable* (Ehle 1969).

Taking into account that for rational functions

$$\lim_{z\to\infty} R(z) = \lim_{z\to-\infty} R(z), \tag{2.55}$$

a method is called L-stable if it is A-stable and if in addition

$$\lim_{z\to\infty} R(z) = 0. \tag{2.56}$$

Considering that for an implicit RK method we have that

$$R(\infty) = 1 - b^T A^{-1}\mathbb{1}, \tag{2.57}$$

according to (2.49), it follows that if an A-stable implicit RK method with nonsingular A satisfies one of the following conditions:

$$a_{sj} = b_j, \quad j = 1, \dots, s; \tag{2.58}$$

$$a_{i1} = b_1, \quad i = 1, \dots, s, \tag{2.59}$$

then $R(\infty) = 0$ in (2.57), and the method is also L-stable. In particular, methods satisfying (2.58) are called *stiffly accurate* (Prothero and Robinson 1974) and are particularly important for the solution of singular perturbation problems and for differential-algebraic equations (Hairer and Wanner 1996).

Finally, there are some implicit schemes with large stability domains that are not A-stable. In order to characterize these methods, $A(\alpha)$*-stability* constitutes another stability property for which a method is said to be $A(\alpha)$-stable if a sector α is contained in the stability region (Widlund 1967):

$$S_\alpha = \{z \in \mathbb{C} \text{ s.t. } |\arg(-z)| < \alpha,\, z \neq 0\}. \tag{2.60}$$

In this work, we consider only one-step integration methods. Nevertheless, dedicated *multi-step integration methods* for the resolution of stiff problems were also developed. These schemes consider several time steps in order to reconstruct the numerical solution that satisfies the differential equations at each considered time step. Moreover, the *Second Dahlquist Barrier* states that an A-stable multi-step method must be of order $p \leq 2$ (Dahlquist 1963). Nevertheless, there are many multi-step schemes performing good $A(\alpha)$-stability properties for high orders, and L-stability for lower ones, which can be efficiently used to solve stiff problems. Some examples are the *LSODE* (Hindmarsh 1980; Hindmarsh 1983) (Livermore Solver for ODEs) or the *VODE* solver (Brown, Byrne, and Hindmarsh 1989) (Variable-coefficient ODE solver), both based on a variable-order (up to fifth) *Backward Differentiation Formulae* developed by Gear (Gear 1971) (see (Hairer and Wanner 1996) for more details on dedicated multi-step methods for stiff problems).

2.3.1 Construction of Implicit Runge-Kutta Methods

As previously detailed for the explicit case, an implicit RK method is of order p if condition (2.27) is satisfied, in which case we see that $R(z)$ is this time a rational approximation to e^z according to (2.53). In this context, the construction of fully implicit RK methods relies heavily on the following conditions (Hairer and Wanner 1996):

$$\left.\begin{array}{lll} B(p): & \displaystyle\sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q}, & q = 1, \dots, p; \\ C(\eta): & \displaystyle\sum_{j=1}^{s} a_{ij} c_j^{q-1} = \frac{c_i^q}{q}, & i = 1, \dots, s, \quad q = 1, \dots, \eta; \\ D(\zeta): & \displaystyle\sum_{i=1}^{s} b_i c_i^{q-1} a_{ij} = \frac{b_j}{q}(1 - c_j^q), & j = 1, \dots, s, \quad q = 1, \dots, \zeta. \end{array}\right\} \tag{2.61}$$

The first condition $B(p)$ states that the quadrature formula $(b_i, c_i)_{i=1}^s$ is of order p, whereas it was proved by Butcher (Butcher 1964a) that if the coefficients b_i, c_i, a_{ij} of a RK method satisfy $B(p)$, $C(\eta)$, $D(\zeta)$ with $p \leq \eta + \zeta + 1$ and $p \leq 2\eta + 2$, then the method is of order p.

With these tools, one way of building these RK schemes considers *collocation methods* based on quadrature formulae. The main goal is to find a polynomial $p(t)$ of degree s such that $p(t_n) = u_n$, and that for a set of *collocation points* $0 \leq c_1 < \ldots < c_s \leq 1$, it verifies

$$\mathrm{d}_t p(t_n + c_i \delta t) = f\left(p(t_n + c_i \delta t)\right), \quad i = 1, \dots, s; \tag{2.62}$$

such that $u(t_{n+1}) = u(t_n + \delta t)$ will be approximated by $u_{n+1} = p(t_n + \delta t)$ (Guillon and Soulé 1969; Wright 1971). We can then determine the collocation points based on the quadrature formulae used to numerically approximate

$$\int_{t_0}^{t_0+\delta t} f(t)\,\mathrm{d}t \approx \delta t \sum_{i=1}^{s} b_i f(t_0 + c_i \delta t). \tag{2.63}$$

If the quadrature method yields approximations of order p, an important mathematical result is that the collocation method will also yield approximations of order p for the differential problem (2.62) (Guillon and Soulé 1969).

In this way, Butcher (Butcher 1964b) introduced RK methods based on Radau quadrature formulae (Radau 1880), for which the collocation points $c_1, \dots, c_s$, are the zeros of the polynomials

$$\mathrm{I}: \quad \mathrm{d}_x^{s-1}\left(x^s (x-1)^{s-1}\right), \tag{2.64}$$

$$\mathrm{II}: \quad \mathrm{d}_x^{s-1}\left(x^{s-1} (x-1)^{s}\right), \tag{2.65}$$

and the weights $b_1, \dots, b_s$, are computed in order to verify $B(s)$ for the quadrature formula $(b_i, c_i)_{i=1}^s$ into (2.61). Finally, we have that $B(2s-1)$ since $p = 2s - 1$ for a Radau quadrature formula. Both polynomials have positive zeros with $c_1 = 0$ and $c_i < 1$, $i = 2, \dots, s$ for (2.64), and $c_i > 0$, $i = 1, \dots, s-1$ and $c_s = 1$ for (2.65), whereas the remaining coefficients are computed based on the order conditions (2.61). These first schemes were not A-stable but based on these ideas, Ehle (Ehle 1969) constructed some A- and L-stable schemes which gave birth to the families of formulae called RadauIA and RadauIIA, depending on the used quadrature formula (2.64) or (2.65). Tables 2.1 and 2.2 show, respectively, the corresponding coefficients for RadauIA and RadauIIA of order $p = 5$ with $s = 3$ stages. L-stability can be retrieved in this case for $p = 5$, by verifying, respectively, conditions (2.59) and (2.58).

Alternatively, other schemes were derived based on other quadrature formulae. For instance, a family of s-stage *Gauss methods* were constructed this time from Gaussian quadrature formulae, and perform A-stability properties with the maximum possible order: $p = 2s$ (Butcher 1964a; Ehle 1968). Nevertheless, these schemes are usually not L-stable. Another large group considers Lobatto quadrature formulae which yields some A- and L-stable schemes of order $p = 2s - 2$

Table 2.1 – *RadauIA method of order 5.*

0	$\frac{1}{9}$	$\frac{-1-\sqrt{6}}{18}$	$\frac{-1+\sqrt{6}}{18}$
$\frac{6-\sqrt{6}}{10}$	$\frac{1}{9}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{88-43\sqrt{6}}{360}$
$\frac{6+\sqrt{6}}{10}$	$\frac{1}{9}$	$\frac{88+43\sqrt{6}}{360}$	$\frac{88-7\sqrt{6}}{360}$
	$\frac{1}{9}$	$\frac{16+\sqrt{6}}{36}$	$\frac{16-\sqrt{6}}{36}$

Table 2.2 – *RadauIIA method of order 5.*

$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$	$\frac{296-169\sqrt{6}}{1800}$	$\frac{-2+3\sqrt{6}}{225}$
$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\sqrt{6}}{1800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$
1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$

(Butcher 1964a; Ehle 1968; Chipman 1971; Axelsson 1972). In what follows, we will recall some of the previous concepts and give some insights into the practical implementation of these implicit RK methods by considering the *Radau5* solver developed by Hairer & Wanner (Hairer and Wanner 1996).

2.3.2 The Radau5 Solver

Let us recall the general nonlinear problem (2.1), this time of dimension m, that is, $u_0 \in \mathbb{R}^m$, $u : \mathbb{R} \to \mathbb{R}^m$, and $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$, to keep the previous notations:

$$\left.\begin{aligned} \mathrm{d}_t u &= f(t, u(t)), \\ u(0) &= u_0. \end{aligned}\right\} \tag{2.66}$$

The solution of this problem by a s-stage fully implicit RK method (2.22) will lead to the solution of a nonlinear system of equations of size $m \times s$ in order to determine the unknowns $g_1, \dots, g_s$. In order to avoid solving these large systems, a family of diagonally implicit RK schemes called *SDIRK* (Singly Diagonally Implicit RK) were developed, that considers a less expensive alternative by solving s successive stages with only m-dimensional systems to be solved at each stage. Nevertheless, more stages than the previously seen for fully implicit RK schemes are usually needed to build A- or stiffly accurate L-stable methods, for instance, $p = s + 1$ or $p = s$. A further simplification considered the linearization of DIRK schemes in order to replace the nonlinear systems by a sequence of linear problems. These methods are usually called *linearly implicit RK methods* or simply *Rosenbrock methods*, and show good $A(\alpha)$-stability properties. A survey and analysis of these and other methods can be found in (Hairer and Wanner 1996).

As a consequence, we can infer that an efficient solution of large nonlinear systems is mandatory for practical purposes and constitutes the main difficulty in the implementation of a fully implicit RK method (Hairer and Wanner 1996). In this context, Hairer & Wanner developed the Radau5 solver for which they had to introduce a few performing tools to overcome the many numerical difficulties associated with the practical implementation of implicit RK schemes. All of these issues are discussed in details in their book (Hairer and Wanner 1996), but we will present here some of them that are usually common to various implicit RK solvers, for the sake of completeness of this work.

Radau5 implements the fifth order, 3-stage Ehle's method RadauIIA, given in Table 2.2. This is a high order, A- and L-stable scheme, very suitable for highly stiff problems. The solver considers RadauIIA because among other reasons, this is a stiffly accurate scheme given by condition (2.58). From a practical point of view and for very stiff problems such as singularly perturbed problems, condition (2.58) implies that the numerical solution becomes also an internal stage in the solution of the $g_1, \ldots, g_s$ ($c_3 = 1$ in Table 2.2). Therefore, we can expect that fast transients in the exact solution will be better reproduced by numerically considering the relaxed fast variables after one time step δt (Hairer and Wanner 1996).

Considering the general implicit RK scheme (2.22), we define a new set o variables $z_1, \ldots, z_s$, for the computation of the $g_1, \ldots, g_s$:

$$z_i = g_i - u_0, \tag{2.67}$$

in order to reduce the influence of round-off errors (Hairer and Wanner 1996). This yields

$$\left.\begin{aligned} z_i &= \delta t \sum_{j=1}^{s} a_{ij} f(t_0 + c_j \delta t, u_0 + z_j), \qquad i = 1, \ldots, s; \\ u_1 &= u_0 + \delta t \sum_{j=1}^{s} b_j f(t_0 + c_j \delta t, u_0 + z_j). \end{aligned}\right\} \tag{2.68}$$

Therefore, knowing the solution $z_1, \ldots, z_s$ implies an explicit formula for u_1, for which s additional function evaluations are required. These extra computation can nevertheless be avoided if the matrix $A = (a_{ij})$ is nonsingular, which is the case for RadauIIA. Actually, considering that

$$\begin{pmatrix} z_1 \\ \vdots \\ z_s \end{pmatrix} = A \begin{pmatrix} \delta t f(t_0 + c_1 \delta t, u_0 + z_1) \\ \vdots \\ \delta t f(t_0 + c_s \delta t, u_0 + z_s) \end{pmatrix}, \tag{2.69}$$

the computation of u_1 is equivalent to

$$u_1 = u_0 + \sum_{i=1}^{s} d_i z_i, \tag{2.70}$$

where

$$(d_1, \ldots, d_s) = (b_1, \ldots, b_s) A^{-1}. \tag{2.71}$$

Taking into account the coefficients in Table 2.2, we see that for RadauIIA: $d = (0, 0, 1)$, since $b_i = a_{si}$ for all i according to (2.58).

To solve the nonlinear system (2.69), Radau5 considers an iterative Newton's method. This amounts to solve at each iteration a linear system with the matrix:

$$\begin{pmatrix} \mathrm{Id} - \delta t a_{11} \partial_u f(t_0 + c_1 \delta t, u_0 + z_1) & \ldots & -\delta t a_{1s} \partial_u f(t_0 + c_s \delta t, u_0 + z_s) \\ \vdots & \ddots & \vdots \\ -\delta t a_{s1} \partial_u f(t_0 + c_1 \delta t, u_0 + z_1) & \ldots & \mathrm{Id} - \delta t a_{ss} \partial_u f(t_0 + c_s \delta t, u_0 + z_s) \end{pmatrix}. \tag{2.72}$$

If we approximate all Jacobians $\partial_u f(t_0 + c_i \delta t, u_0 + z_i)$ by

$$J \approx \partial_u f(t_0, u_0), \tag{2.73}$$

we consider a simplified Newton's method for

$$G(Z) = Z - (\mathrm{Id} - \delta t A \otimes J)F(Z) = 0, \tag{2.74}$$

where $Z = (z_1, \ldots, z_s)^T$, and $F(Z) = (f(u_0 + c_1\delta t, u_0 + z_1), \ldots, f(t_0 + c_s\delta t, u_0 + z_s))^T$, so that the $(k+1)$-th approximation of the solution Z is recursively computed by

$$\left.\begin{array}{rcl} (\mathrm{Id} - \delta t A \otimes J)\Delta Z^k & = & -Z^k + \delta t(A \otimes \mathrm{Id})F(Z^k), \\ Z^{k+1} & = & Z^k + \Delta Z^k. \end{array}\right\} \tag{2.75}$$

Each iteration requires then s evaluations of f to compute $F(Z^k)$, and the solution of a $m \times s$ linear system to compute the increments $\Delta Z^k = (\Delta z_1^k, \ldots, \Delta z_s^k)^T$. Fortunately, the matrix $(\mathrm{Id} - \delta t A \otimes J)$ is the same for all iterations with the approximated Jacobians (2.73), and its inversion by an LU-decomposition, usually quite expensive, is done only once. Furthermore, exploiting the special structure of the matrix $(\mathrm{Id} - \delta t A \otimes J)$, a decomposition of the linear system into two subsystems following a procedure introduced by Butcher (Butcher 1976), leads to an important reduction of the number of operations, which is also implemented in the Radau5 solver (Hairer and Wanner 1996). If no analytical expression is available, the Jacobians can always be numerically approximated by

$$J_{ij} \approx \frac{f^{(i)}(t_0, u^{(j)} + \delta u^{(j)}) - f^{(i)}(t_0, u^{(j)})}{\delta u^{(j)}}, \qquad i, j = 1, \ldots, m, \tag{2.76}$$

for relatively small, positive perturbations: $\delta u = (\delta u^{(1)}, \ldots, \delta u^{(m)})$. Finally, Hairer & Wanner defined also dedicated stopping criteria for the iterative method as well as appropriate starting values Z^0 for the Newton iterations (Hairer and Wanner 1996).

In order to select the time step and guarantee a prescribed accuracy, Radau5 uses a lower order embedded method to numerically estimate the local error in the same spirit of section § 2.2.2. We illustrate this procedure for this particular case. A lower order approximation of the solution $\hat{u}_1$ according to (2.37) is computed by

$$\hat{u}_1 = u_0 + \delta t \hat{b}_0 f(t_0, y_0) + \delta t \sum_{i=1}^{3} \hat{b}_i f(t_0 + c_i \delta t, g_i), \tag{2.77}$$

using the same collocation points c_1, c_2, c_3 of RadauIIA (see Table 2.2), and thus the same evaluations of f. An extra evaluation of f is needed at t_0, whereas $\hat{b}_0 = \hat{\gamma}_0$, where $\hat{\gamma}_0^{-1}$ is a real eigenvalue of A^{-1} previously computed. In order to set the new weights $\hat{b}_1$, $\hat{b}_2$, $\hat{b}_3$ we consider the difference:

$$\hat{u}_1 - u_1 = \delta t \hat{\gamma}_0 f(t_0, y_0) + \delta t \sum_{i=1}^{3} (\hat{b}_i - b_i) f(t_0 + c_i \delta t, g_i), \tag{2.78}$$

into (2.61) for $B(3)$ such that $\hat{u}_1 - u_1 = \mathcal{O}(\delta t^4)$. Considering the representation (2.70), this yields finally

$$\hat{u}_1 - u_1 = \delta t \hat{\gamma}_0 f(t_0, y_0) + \sum_{i=1}^{3} \hat{d}_i z_3, \tag{2.79}$$

where

$$(\hat{d}_1, \hat{d}_2, \hat{d}_3) = \frac{\hat{\gamma}_0}{3}(-13 - 7\sqrt{6}, -13 + 7\sqrt{6}, -1). \tag{2.80}$$

With these solutions, Radau5 computes the approximation:

$$err = (\mathrm{Id} - \delta t \hat{\gamma}_0 J)^{-1}(\hat{u}_1 - u_1), \tag{2.81}$$

as error estimate in order to simultaneously guarantee that the difference (2.79) is bounded for $\delta t \to 0$ and $\delta t\lambda \to \infty$ (if $f(u) = \lambda u$ and $J = \lambda$), for stiff problems (Hairer and Wanner 1996).

The time steps are then computed by taking the minimum of

$$\delta t_{\text{new}} = fac \cdot \delta t_n \left(\frac{1}{\|err_{n+1}\|} \right)^{1/4}, \tag{2.82}$$

and

$$\delta t_{\text{new}} = fac \cdot \delta t_n \left(\frac{1}{\|err_{n+1}\|} \right)^{1/4} \frac{\delta t_n}{\delta t_{n-1}} \left(\frac{\|err_n\|}{\|err_{n+1}\|} \right)^{1/4}, \tag{2.83}$$

based, respectively, on (2.43) and (2.46) with

$$\|err\| = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\frac{err^{(i)}}{sc_i} \right)^2}, \tag{2.84}$$

with $err^{(i)} = (\text{Id} - \delta t \hat{\gamma}_0 J)^{-1} \left(\hat{u}_1^{(i)} - u_1^{(i)} \right)$, and $sc_i = Atol_i + \max(|u_0^{(i)}|, |u_1^{(i)}|) \cdot Rtol_i$, where *Atol* and *Rtol* are defined as absolute and relative accuracy tolerances (Hairer and Wanner 1996). With the definition of the error estimate given by (2.84), the current time step is accepted if $\|err\| \leq 1$, otherwise it is rejected. In this case as well as for the first step, Radau5 uses a second error estimate instead of (2.81):

$$\widetilde{err} = (\text{Id} - \delta t \hat{\gamma}_0 J)^{-1} \left(\delta t \hat{\gamma}_0 f(t_0, y_0 + err) + \sum_{i=1}^{3} \hat{d}_i z_3 \right), \tag{2.85}$$

which implies an additional evaluation of f, but we have that $\widetilde{err} \to 0$ is satisfied for $\delta t\lambda \to \infty$, in the same way as the numerical solution u_1 does.

2.4 Stabilized Explicit Runge-Kutta Methods

In many cases, there are stiff problems for which A-stable methods are not necessarily required. Some remarkable examples come from the discretization of parabolic PDEs which lead to stiff problems with a Jacobian matrix involving (possibly large) eigenvalues close to the real negative axis. This is the particular case of the discretized heat equation (2.14) in § 2.1.1, for which the real negative eigenvalues (2.18) increase with finer spatial discretizations. Therefore, instead of A-stable but time consuming implicit procedures, *stabilized explicit RK methods* should be preferred. These explicit methods avoid the solution of algebraic systems, while featuring an extended stability domain along the negative real axis, very appropriate for this type of problem. A detailed survey on these schemes can be found in (Verwer 1996), and in the book of Hundsdorfer & Verwer (Hundsdorfer and Verwer 2003).

The main goal is to construct methods of order p with a family of stability polynomial R_s of degree s:

$$R_s(z) = 1 + z + \cdots + \frac{z^p}{p!} + \sum_{p+1}^{s} \alpha_{i,s} z^i, \tag{2.86}$$

with $s \geq p + 1$, and $\alpha_{i,s} \in \mathbb{C}$, such that $R_s(z)$ remains bounded as much as possible along the real negative axis, *i.e.,*

$$|R_s(z)| \leq 1, \qquad z \in [-\ell_s, 0], \tag{2.87}$$

with ℓ_s as large as possible. One way of building such stability polynomials considers the family of Chebyshev polynomials:

$$T_s(\cos(z)) = \cos(s\,z), \tag{2.88}$$

defined also by the recurrence relation:

$$T_0(z) = 1, \quad T_1(z) = z, \quad T_s(z) = 2zT_{s-1}(z) - T_{s-2}(z), \tag{2.89}$$

which remain bounded between 1 and -1 for $z \in [-1, 1]$, and in particular yield boundaries ℓ_s proportional to s^2.

These schemes are usually called *Runge-Kutta-Chebyshev* methods, and feature extended real stability intervals proportional to s^2, a good property inherited from Chebyshev-type polynomials. For instance, for $p = 1$, the optimal polynomials that satisfies (2.86) are directly the shifted Chebyshev polynomials:

$$R_s(z) = T_s\left(1 + \frac{z}{s^2}\right), \tag{2.90}$$

which are shown to yield the optimal $\ell_s = 2s^2$. However, in the points where $R_s(z) = \pm 1$ for $z \in \mathbb{R}^-$, the stability domain has zero width and therefore, there is no damping at all of high frequencies. The standard way to overcome this difficulty considers a small parameter $\varepsilon > 0$ in order to build *damped Chebyshev stability functions* (Guillou and Lago 1961):

$$R_s(z) = \frac{1}{T_s(w)} T_s(w_0 + w_1 z), \quad w_0 = 1 + \frac{\varepsilon}{s^2}, \quad w_1 = \frac{T_s(w_0)}{T_s'(w_0)}. \tag{2.91}$$

As a consequence, the stability domains are reduced by approximatively ε: $|R_s(z)| \leq 1 - \varepsilon$, while the stability length is shortened by approximatively $(4\varepsilon/3)s^2$; nevertheless, the order of the scheme is preserved and a safe distance from the real axis is guaranteed (Hairer and Wanner 1996).

Based on these ideas, a first family of method called *Lebedev-type methods* (Lebedev 1989; Lebedev 1994), aims at building RK schemes based on the optimal stability polynomials that satisfy (2.86) for a given p. For $p = 1$ we have seen that these polynomials are the shifted Chebyshev polynomials (2.90), so the idea is to write them as (Saul'ev 1960; Guillou and Lago 1961):

$$R_s(z) = \prod_{i=1}^{s} (1 + \delta_i z), \quad \delta_i = -\frac{1}{z_i}, \tag{2.92}$$

where z_i are the roots of $R_s(z)$, and to represent the RK scheme as a composition of explicit Euler steps:

$$\left.\begin{aligned} g_0 &= u_0, \\ g_i &= g_{i-1} + \delta t \delta_i f(g_{i-1}), \quad i = 1, \ldots, s, \\ u_1 &= g_s. \end{aligned}\right\} \tag{2.93}$$

The main difficulty constitutes finding the best sequence of integration of the Euler steps to ensure stability properties of the scheme (Lebedev 1993a; Lebedev 1993b). Formulae of order up to four were also achieved even though there is no analytical expression for the optimal stability polynomials of order $p \geq 2$ (Lebedev and Medovikov 1998; Medovikov 1998). The computations of these polynomials are therefore performed numerically and yield, for instance, second order schemes with practically optimal $\ell_s \approx 0.82 \cdot s^2$ for $s \gg 1$. These results have been implemented in the *DUMKA* code (Lebedev 1994; Lebedev 2000).

Based on numerical approximations of the optimal boundaries ℓ_s (van der Houwen 1977), and knowing that among all polynomials of order p and degree s satisfying (2.86), the optimal one satisfies the so-called *equal ripple property* which states that there exist $s-p+1$ points $z_0 < z_1 < \cdots < z_{s-p} < 0$, with $z_0 = -\ell_s$, such that

$$\left.\begin{aligned} R(z_i) &= -R(z_{i+1}), & i &= 0,\ldots,s-p-1, \\ |R(z_i)| &= 1, & i &= 0,\ldots,s-p; \end{aligned}\right\} \tag{2.94}$$

another approach known as the *Van der Houwen-Sommeijer methods* (van der Houwen and Sommeijer 1980), constructs the RK schemes based on a linear combination of scaled and shifted Chebyshev polynomials that aim at approximating the optimal polynomial by verifying (2.94), and generates about 80 % of the optimal interval ℓ_s. First and second order schemes known as *RKC* methods were built with these approximated optimal polynomials using the three-term recurrence formula (2.89):

$$\left.\begin{aligned} g_0 &= u_0, \\ g_1 &= g_0 + \tilde{\mu}_1 \delta t f(g_0), \\ g_i &= (1-\mu_i-\nu_i) g_0 + \mu_i g_{i-1} + \nu_i g_{i-2} \\ &\quad + \tilde{\mu}_i \delta t f(g_{i-1}) + \tilde{\gamma}_i \delta t f(g_0), \qquad i = 2,\ldots,s, \\ u_1 &= g_s, \end{aligned}\right\} \tag{2.95}$$

where all the coefficients $(\tilde{\mu}_i, \mu_i, \nu_i, \tilde{\gamma}_i)$ are available in analytical form for arbitrary $s \geq 2$ (Sommeijer and Verwer 1980). In this way, an efficient second order solver known simply as *RKC* proposed by Sommeijer *et al.* in (Sommeijer, Shampine, and Verwer 1997), gained notorious reputation over the last years. The RKC solver also features local error control, with variable step sizes, computed on an approximation of the leading term of the local error expansion, theoretically derived from a detailed stability and convergence analysis presented in (Verwer, Hundsdorfer, and Sommeijer 1990). The stability bound is given by $\ell_s \approx 0.653 \cdot s^2$ for the second order RKC scheme, and hence for a given time step computed according to a prescribed accuracy tolerance, an adequate number of stages s is chosen in order to ensure the stability of the method.

2.4.1 The ROCK Method

A third approach that combined the previous ones by searching practically optimal stability bounds ℓ_s, and by using a three-term recurrence relation, gave birth to the *ROCK* methods (for Orthogonal-Runge-Kutta-Chebyshev) (Abdulle and Medovikov 2001; Abdulle 2002). A preliminary important result of Abdulle (Abdulle 2000) was that the optimal stability polynomials satisfying (2.86) for a given p and the equal ripple property (2.94), possess exactly p complex roots if p is even and exactly $p-1$ complex roots if p is odd. Therefore, if p is even, we can then split the stability function in the following form:

$$R_s(z) = w_p(z) P_{s-p}(z), \tag{2.96}$$

where w_p retains the p complex roots and P_{s-p}, the remaining $(s-p)$ real roots. The idea developed by Medovikov & Abdulle in (Abdulle and Medovikov 2001) for $p=2$, and then extended to $p=4$ by Abdulle in (Abdulle 2002), was to approximate $R_s(z)$ by

$$\tilde{R}_s(z) = \tilde{w}_p(z) \tilde{P}_{s-p}(z), \tag{2.97}$$

with the orthogonal polynomials $\tilde{P}_{s-p}$, associated with the weight function $\tilde{w}_p^2(z)/\sqrt{1-z^2}$, such that $\tilde{R}_s(z)$ results in a p-order stability polynomial which remains bounded as much as possible along the negative real axis, taking also into account some damping. The techniques to compute the orthogonal polynomials and the weight function are given in (Abdulle and Medovikov 2001) and (Abdulle 2002).

Once the stability functions have been computed, a three-term recurrence relation:

$$\tilde{P}_0(z) = 1, \quad \tilde{P}_1(z) = 1 + \mu_1 z, \quad \tilde{P}_i(z) = (\mu_i z - \nu_i)\tilde{P}_{i-1}(z) - \kappa_i \tilde{P}_{i-2}(z), \tag{2.98}$$

with $i = 2, \dots, s-p$, satisfied by the orthogonal polynomials, is used to define the internal stages of the RK method following the idea of (van der Houwen and Sommeijer 1980):

$$\left.\begin{aligned} g_0 &= u_0, \\ g_1 &= g_0 + \tilde{\mu}_1 \delta t f(g_0), \\ g_i &= \tilde{\mu}_i \delta t f(g_{i-1}) - \nu_i g_{i-1} - \kappa_i g_{i-2}, \quad i = 2, \dots, s-p. \end{aligned}\right\} \tag{2.99}$$

Considering $\mathrm{d}_t u = \lambda u$ and $z = \lambda \delta t$, the resulting $\tilde{P}_{s-p}(z)$ is the stability function associated with (2.99): $g_{s-p} = \tilde{P}_{s-p}(z) u_0$. The coefficients (μ_i, ν_i, κ_i) are computed by a procedure introduced in (Abdulle and Medovikov 2001).

The case $p = 2$ yields thus the second order ROCK2 method (Abdulle and Medovikov 2001) for which $\tilde{w}_2(z)$ is a two-stage finishing procedure applied to $g_{s-2} = \tilde{P}_{s-2}(z) u_0$. For $\mathrm{d}_t u = \lambda u$ and $z = \lambda \delta t$, this implies

$$u_1 = \tilde{w}_2(z) g_{s-2} = \tilde{w}_2(z) \tilde{P}_{s-2}(z) u_0 = \tilde{R}_s(z) u_0. \tag{2.100}$$

The order conditions for $p = 2$ are classical to explicit RK schemes and allow us to compute the coefficients of the final stages. In particular for second order, the order conditions are the same for both linear and nonlinear problems. A solution $\hat{u}_1$ of order $\hat{p} = 1$, is computed embedded at the final step $\tilde{w}_2(z)$, and an estimate of the local error $err = (\hat{u}_1 - u_1)$, is computed for the step size selection, according to the same criteria used by Radau5 (Hairer and Wanner 1996) with expressions (2.82) and (2.83). The nearly optimal stability interval is given by $\tilde{\ell}_s \approx 0.81 \cdot s^2$ (the optimal ratio is about 0.82 (van der Houwen 1977)). Therefore, with the time step fixed by the prescribed accuracy (*Atol* and *Rtol*), the number of stages needed to guarantee stability is computed by

$$\delta t \rho \left(\partial_u f(u)\right) \leq 0.81 \cdot s^2, \tag{2.101}$$

where ρ is the spectral radius of the Jacobian of the system of ODEs. A dynamic computation of this spectral radius is provided by ROCK2 using a non-linear power method which is a slight modification of the algorithm proposed in (Sommeijer, Shampine, and Verwer 1997) for the RKC code.

Just like before, for the fourth order ROCK4 ($p = 4$) the coefficients of the weight function $\tilde{w}_4(z)$ must be computed such that the order conditions of order 4 are satisfied. As in (Medovikov 1998), a theory of composition of methods (the "Butcher group") is applied to achieve a fourth order method denoted WP, where the first method, denoted by P is given by the three-term recurrence relation in (2.99) this time with $p = 4$, whereas the coefficients of the four stages method W associated with $\tilde{w}_4(z)$ are computed such that the "composite" method WP is of order 4 as shown in (Abdulle 2002). As in the previous second order case, an embedded method $\hat{W}$ is built embedded into W in order to keep the same recurrence formulae (2.98) for both the fourth order and embedded methods. A third order embedded RK scheme is thus constructed by adding a new stage to $\tilde{w}_4(z)$, and the coefficients are computed with the same composition technique such that the "composite" method $\hat{W}P$ is of order 3, and that the stability polynomials of the embedded methods are bounded in the same interval as the ones of the ROCK4 scheme. The latter feature is indispensable to guarantee stability of the lower order method, and to obtain thus reliable error estimates.

The practically optimal stability interval is this time given by $\tilde{\ell}_s \approx 0.35 \cdot s^2$ (the optimal ratio for fourth order is about 0.34 in (van der Houwen 1977) and 0.35 in (Medovikov 1998)). The ROCK4 solver implements the same tools as ROCK2 for time step selection in terms of estimates (2.82) and (2.83), as well as the numerical computation of the spectral radius. For a given time step δt, computed based on the prescribed accuracy ($Atol$ and $Rtol$), the number of stages that ensures stability of computations is now given by

$$\delta t \rho \left(\partial_u f(u) \right) \leq 0.35 \cdot s^2. \tag{2.102}$$

A notorious advantage of the three-term recurrence formulae used by the RKC (2.95) and ROCK (2.99) methods, is that even though an arbitrary number of stages s might be required to guarantee stability, only the current three arrays in the recurrence relations need to be saved. Considering the two-stage $\tilde{w}_2(z)$ for the second order ROCK2, five solution arrays need thus to be saved to perform all the computations. The same follows for ROCK4 for which seven arrays shall be required. Notice that the construction of the ROCK schemes through (2.96) involves at least $s = 3$ and $s = 5$ internal stages, respectively, for ROCK2 and ROCK4 schemes. The main advantage of the ROCK schemes compared with previous stabilized RK schemes is that it combines the best features of both Lebedev- and Van der Houwen-Sommeijer-type methods by using the three-term recurrence formulae with practically optimal stability polynomials. The latter implies larger stability domains in the practical implementations considering that ℓ_s is approximated by $0.81 \cdot s^2$ for ROCK2 compared with $0.65 \cdot s^2$ for the also second order RKC solver (Sommeijer, Shampine, and Verwer 1997). In particular, a higher order, stabilized explicit scheme of easy implementation with an optimal stability interval, was achieved with the ROCK4 solver. In this way, the stability domains of explicit RK methods are extended without altering the orders of the numerical approximations, and furthermore without requiring excessive supplementary memory space with respect to a standard explicit RK scheme.

References

Abdulle, A. (2000). On roots and error constants of optimal stability polynomials. *BIT Numer. Math. 40*(1), 177–182.

Abdulle, A. (2002). Fourth order Chebyshev methods with recurrence relation. *SIAM J. Sci. Comput. 23*(6), 2041–2054.

Abdulle, A. and A. Medovikov (2001). Second order Chebyshev methods based on orthogonal polynomials. *Numer. Math. 90*(1), 1–18.

Axelsson, O. (1972). A note on a class of strongly A-stable methods. *BIT Numer. Math. 12*, 1–4.

Bernus, O., R. Wilders, C. Zemlin, H. Verschelde, and A. Panfilovz (2002). A computationally efficient electrophysiological model of human ventricular cells. *Am. J. Physiol. Heart Circ. Physiol. 282*(6), H2296–H2308.

Brown, P., G. Byrne, and A. Hindmarsh (1989). VODE: A variable-coefficient ODE solver. *SIAM J. Sci. Stat. Comput. 10*, 1038–1051.

Butcher, J. (1964a). Implicit Runge-Kutta processes. *Math. Comp. 18*, 50–64.

Butcher, J. (1964b). Integration processes based on Radau quadrature formulas. *Math. Comp. 18*, 233–244.

Butcher, J. (1964c). On Runge-Kutta processes of high order. *J. Austral. Math. Soc. 4*, 179–194.

Butcher, J. (1964d). On the attainable order of Runge-Kutta methods. *Math. Comp. 19*, 408–417.

Butcher, J. (1976). On the implementation of implicit Runge-Kutta methods. *BIT Numer. Math. 6*, 237–240.

Castella, F., P. Chartier, S. Descombes, and G. Vilmart (2009). Splitting methods with complex times for parabolic equations. *BIT Numer. Math. 49*, 487–508.

Chipman, F. (1971). A-stable Runge-Kutta processes. *BIT Numer. Math. 11*, 384–388.

Dahlquist, G. (1963). A special stability problem for linear multistep methods. *Nordisk Tidskr. Informations-Behandling 3*, 27–43.

D'Angelo, Y. (1994). *Analyse et simulation numérique de phénomènes liés à la combustion supersonique.* Ph. D. thesis, Ecole Nationale des Ponts et Chaussées.

D'Angelo, Y. and B. Larrouturou (1995). Comparison and analysis of some numerical schemes for stiff complex chemistry problems. *RAIRO Modél. Math. Anal. Numér. 29*(3), 259–301.

Descombes, S. (2001). Convergence of a splitting method of high order for reaction-diffusion systems. *Math. Comp. 70*(236), 1481–1501.

Descombes, S., M. Duarte, T. Dumont, F. Laurent, V. Louvet, and M. Massot (2014). Analysis of operator splitting in the nonasymptotic regime for nonlinear reaction-diffusion equations. Application to the dynamics of premixed flames. *SIAM J. Numer. Anal. 52*(3), 1311–1334.

Descombes, S., T. Dumont, V. Louvet, and M. Massot (2007). On the local and global errors of splitting approximations of reaction-diffusion equations with high spatial gradients. *Int. J. of Computer Mathematics 84*(6), 749–765.

Descombes, S. and M. Massot (2004). Operator splitting for nonlinear reaction-diffusion systems with an entropic structure: Singular perturbation and order reduction. *Numer. Math. 97*(4), 667–698.

Descombes, S. and M. Schatzman (2002). Strang's formula for holomorphic semi-groups. *J. Math. Pures Appl. 81*(1), 93–114.

Descombes, S. and M. Thalhammer (2010). An exact local error representation of exponential operator splitting methods for evolutionary problems and applications to linear Schrödinger equations in the semi-classical regime. *BIT Numer. Math. 50*, 729–749.

Descombes, S. and M. Thalhammer (2011). The Lie-Trotter splitting method for nonlinear evolutionary problems involving critical parameters. An exact local error representation and application to nonlinear Schrödinger equations in the semi-classical regime. *Preprint, available at HAL (http://hal.archives-ouvertes.fr/hal-00557593).*

Deuflhard, P. (1983). Order and stepsize control in extrapolation methods. *Numer. Math. 41*, 399–422.

Duarte, M. (2011). *Méthodes numériques adaptatives pour la simulation de la dynamique de fronts de réaction multi-échelles en temps et en espace.* Ph. D. thesis, Ecole Centrale Paris, France.

Duarte, M., S. Descombes, and M. Massot (2011). Parareal operator splitting techniques for multi-scale reaction waves: Numerical analysis and strategies. *ESAIM: Math. Model. Numer. Anal. 45*, 825–852.

Ehle, B. (1968). High order A-stable methods for the numerical solution of systems of DEs. *BIT Numer. Math. 8*, 276–278.

Ehle, B. (1969). On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems. *Research Report CSRR 2010*.

Gear, C. (1971). *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice-Hall series in automatic computation. Englewood Cliffs, NJ: Prentice-Hall.

Goyal, G., P. Paul, H. Mukunda, and S. Deshpande (1988). Time dependent operator-split and unsplit schemes for one dimensional premixed flames. *Combust. Sci. Technol. 60*, 167–189.

Gröbner, W. (1967). *Die Liereihen und ihre Anwendungen.* Berlin 1960: VEB Deutscher Verlag der Wiss. 2nd Edition.

Guillon, A. and F. Soulé (1969). La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation. *RAIRO Anal. Numér. Ser. Rouge, v. R-3*, 17–44.

Guillou, A. and B. Lago (1961). Domaine de stabilité associé aux formules d'intégration numérique d'équations différentielles à pas séparés et à pas liés. Recherche de formules à grands rayons de stabilité. *1er Cong. Assoc. Fran. Calcul, AFCAL, Grenoble*, 43–56.

Gustafsson, K. (1994). Control-theoretic techniques for stepsize selection in implicit Runge-Kutta methods. *ACM Trans. Math. Softw. 20*, 496–517.

Hairer, E., C. Lubich, and M. Roche (1988). Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations. *BIT Numer. Math. 28*, 678–700.

Hairer, E., C. Lubich, and G. Wanner (2006). *Geometric Numerical Integration* (2nd ed.). Berlin: Springer-Verlag. Structure-Preserving Algorithms for Ordinary Differential Equations.

Hairer, E., S. P. Nørsett, and G. Wanner (1987). *Solving Ordinary Differential Equations I.* Berlin: Springer-Verlag. Nonstiff Problems.

Hairer, E. and G. Wanner (1996). *Solving Ordinary Differential Equations II* (2nd ed.). Berlin: Springer-Verlag. Stiff and Differential-Algebraic Problems.

Hansen, E. and A. Ostermann (2009). High order splitting methods for analytic semigroups exist. *BIT Numer. Math. 49*, 527–542.

Hindmarsh, A. (1980). LSODE and LSODI, two new initial value ordinary differential equation solvers. *SIGNUM Newsl. 15*, 10–11.

Hindmarsh, A. (1983). ODEPACK, a systematized collection of ODE solvers. In *Scientific computing (Montreal, Que., 1982)*, pp. 55–64. New Brunswick, NJ: IMACS.

Hundsdorfer, W. and J. Verwer (2003). *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations.* Berlin: Springer-Verlag.

Kozlov, R., A. K. rnø, and B. Owren (2004). The behaviour of the local error in splitting methods applied to stiff problems. *J. Comput. Phys. 195*(2), 576–593.

Lanser, D. and J. Verwer (1999). Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling. *J. Comput. Appl. Math. 111*(1-2), 201–216.

Lebedev, V. (1989). Explicit difference schemes with time-variable steps for solving stiff systems of equations. *Sov. J. Numer. Anal. Math. Modelling 4*, 111–135.

Lebedev, V. (1993a). A new method for determining the zeros of polynomials of least deviation on a segment with weight and subject to additional conditions. part I. *Russian J. Numer. Anal. Math. Modelling 8*, 195–222.

Lebedev, V. (1993b). A new method for determining the zeros of polynomials of least deviation on a segment with weight and subject to additional conditions. part II. *Russian J. Numer. Anal. Math. Modelling 8*, 397–426.

Lebedev, V. (1994). How to solve stiff systems of differential equations by explicit methods. In *Numerical Methods and Applications*, pp. 45–80. Boca Raton:CRC Press.

Lebedev, V. (2000). Explicit difference schemes for solving stiff problems with a complex or separable spectrum. *Comput. Math. and Math. Phys. 40*(12), 1729–1740.

Lebedev, V. and A. Medovikov (1998). An explicit method of the second order of accuracy for solving stiff systems of ordinary differential equations. *Russian Izv. Vyssh. Uchebn. Zaved. Mat. 9*, 55–63.

Maas, U. and S. Pope (1992). Simplifying chemical kinetics: Intrinsic low-dimensional manifolds in composition space. *Combust. and Flame 88*(3-4), 239–264.

Marchuk, G. (1968). Some application of splitting-up methods to the solution of mathematical physics problems. *Applications of Mathematics 13*(2), 103–132.

Marchuk, G. (1975). *Methods of Numerical Mathematics*. Appl. Math. New York, NY: Springer. Trans. from the Russian.

Marchuk, G. (1990). Splitting and alternating direction methods. In *Handbook of Numerical Analysis, Vol. I*, pp. 197–462. Amsterdam: North-Holland.

Massot, M. (2002). Singular perturbation analysis for the reduction of complex chemistry in gaseous mixtures using the entropic structure. *Discrete Contin. Dyn. Syst. Ser. B 2*(3), 433–456.

McLachlan, R. and R. Quispel (2002). Splitting methods. *Acta Numerica 11*, 341–434.

Medovikov, A. (1998). High order explicit methods for parabolic equations. *BIT Numer. Math. 38*, 372–390.

Oran, E. and J. Boris (2001). *Numerical Simulation of Reacting Flows*. Cambridge University Press. Second Edition.

Ostromsky, T., W. Owczarz, and Z. Zlatev (2001). Computational challenges in large-scale air pollution modelling. In *Proceedings of the 15th International Conference on Supercomputing*, ICS '01, pp. 407–418.

Prothero, A. and A. Robinson (1974). On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Math. Comp. 28*(125), 145–162.

Radau, R. (1880). Étude sur les formules d'approximation qui servent à calculer la valeur numérique d'une intégrale definie. *J. Math. Pures Appl. 6*, 283–336.

Ropp, D. and J. Shadid (2005). Stability of operator splitting methods for systems with indefinite operators: Reaction-diffusion systems. *J. Comput. Phys. 203*(2), 449–466.

Ropp, D. and J. Shadid (2009). Stability of operator splitting methods for systems with indefinite operators: Advection-diffusion-reaction systems. *J. Comput. Phys. 228*(9), 3508–3516.

Ropp, D., J. Shadid, and C. Ober (2004). Studies of the accuracy of time integration methods for reaction-diffusion equations. *J. Comput. Phys. 194*(2), 544–574.

Sanz-Serna, J. and M. Calvo (1994). *Numerical Hamiltonian Problems*. London: Chapman & Hall.

Saul'ev, V. (1960). Integration of parabolic type equations with the method of nets. *Moscow,* *Fizmatgiz*. In Russian.

Schatzman, M. (2002). Toward non commutative numerical analysis: High order integration in time. *J. Scientific Computing 17*(1-3), 107–125.

Schwer, D., P. Lu, W. Green, and V. Semião (2003). A consistent-splitting approach to computing stiff steady-state reacting flows with adaptive chemistry. *Combust. Theory Modelling 7*(2), 383–399.

Shampine, L. (1977). Stiffness and nonstiff differential equation solvers, II: Detecting stiffness with Runge-Kutta methods. *ACM Trans. Math. Softw. 3*, 44–53.

Shampine, L. (1987). Control of step size and order in extrapolation codes. *J. Comput. Appl. Math. 18*(1), 3–16.

Shampine, L. and K. Hiebert (1977). Detecting stiffness with the Fehlberg (4,5) formulas. *Computers & Mathematics with Applications 3*(1), 41–46.

Sommeijer, B., L. Shampine, and J. Verwer (1997). RKC: An explicit solver for parabolic PDEs. *J. Comput. Appl. Math. 88*(2), 315–326.

Sommeijer, B. and J. Verwer (1980). *A Performance Evaluation of a Class of Runge-Kutta-Chebyshev Methods for Solving Semidiscrete Parabolic Differential Equations*. Afdeling Numerieke Wiskunde [Department of Numerical Mathematics], 91. Amsterdam: Mathematisch Centrum.

Spee, E., J. Verwer, P. de Zeeuw, J. Blom, and W. Hundsdorfer (1998). A numerical study for global atmospheric transport-chemistry problems. *Mathematics and Computers in Simulation 48*(2), 177–204.

Sportisse, B. (2000). An analysis of operator splitting techniques in the stiff case. *J. Comput. Phys. 161*(1), 140–168.

Sportisse, B. (2007). A review of current issues in air pollution modeling and simulation. *Computational Geosciences 11*, 159–181.

Sportisse, B., G. Bencteux, and P. Plion (2000). Method of Lines versus Operator Splitting for reaction-diffusion systems with fast chemistry. *Environmental Modelling & Software 15*(6-7), 673–679.

Strang, G. (1963). Accurate partial difference methods. I. Linear Cauchy problems. *Arch. Ration. Mech. Anal. 12*, 392–402.

Strang, G. (1968). On the construction and comparison of difference schemes. *SIAM J. Numer. Anal. 5*, 506–517.

Témam, R. (1969a). Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires. I. *Arch. Rational Mech. Anal. 32*, 135–153.

Témam, R. (1969b). Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires. II. *Arch. Rational Mech. Anal. 33*, 377–385.

Thalhammer, M. (2008). High-order exponential operator splitting methods for time-dependent Schrödinger equations. *SIAM J. Numer. Anal. 46*(4), 2022–2038.

Tikhonov, A., A. Vasil'eva, and A. Sveshnikov (1985). *Differential Equations*. Berlin: Springer-Verlag.

Trangenstein, J. and C. Kim (2004). Operator splitting and adaptive mesh refinement for the Luo-Rudy I model. *J. Comput. Phys. 196*(2), 645–679.

Trotter, H. (1959). On the product of semi-groups of operators. *Proc. Am. Math. Soc. 10*, 545–551.

van der Houwen, P. (1977). *Construction of Integration Formulas for Initial Value Problems*. North-Holland Pub.Co.

van der Houwen, P. and B. Sommeijer (1980). On the internal stability of explicit, m-stage Runge-Kutta methods for large m-values. *Z. Angew. Math. Mech. 60*(10), 479–485.

Varadarajan, V. (1974). *Lie Groups, Lie Algebras and their Representations*. Englewood Cliffs, New Jersey: Prentice-Hall.

Verwer, J. (1996). Explicit Runge-Kutta methods for parabolic partial differential equations.

Appl. Numer. Math. 22(1-3), 359–379.

Verwer, J., J. Blom, M. van Loon, and E. Spee (1996). A comparison of stiff ODE solvers for atmospheric chemistry problems. *Atmospheric Environment 30*(1), 49–58.

Verwer, J., W. Hundsdorfer, and B. Sommeijer (1990). Convergence properties of the Runge-Kutta-Chebyshev method. *Numer. Math. 57*, 157–178.

Verwer, J., E. Spee, J. Blom, and W. Hundsdorfer (1999). A second-order Rosenbrock method applied to photochemical dispersion problems. *SIAM J. Sci. Comput. 20*, 1456–1480.

Verwer, J. and B. Sportisse (1998). Note on operator splitting in a stiff linear case. *Rep. MAS-R9830*.

Watts, H. (1984). Step size control in ordinary differential equation solvers. *Trans. Soc. Computer Simulation 1*, 15–25.

Widlund, O. (1967). A note on unconditionally stable linear multistep methods. *BIT Numer. Math. 7*, 65–70.

Wright, K. (1971). Some relationships between implicit Runge-Kutta, collocation and Lanczos τ methods, and their stability properties. *BIT Numer. Math. 10*, 217–227.

Yanenko, N. (1971). *The Method of Fractional Steps. The Solution of Problems of Mathematical Physics in Several Variables.* New York: Springer-Verlag.

Yang, B. and S. Pope (1998). An investigation of the accuracy of manifold methods and splitting schemes in the computational implementation of combustion chemistry. *Combust. and Flame 112*(1-2), 16–32.

Yoshida, H. (1990). Construction of higher order symplectic integrators. *Physics Letters A 150*(5-7), 262–268.