Universidad de Valladolid

esDUVa
Escuela de Doctorado Universidad de Valladolid

PROGRAMA DE DOCTORADO EN MATEMÁTICAS

TESIS DOCTORAL

RATIONAL METHODS WITHOUT ORDER REDUCTION FOR ABSTRACT EVOLUTION PROBLEMS

Presentada por Carlos Arranz Simón para optar
al grado de Doctor por la Universidad de Valladolid

Dirigida por:
Dra. Begoña Cano Urdiales
Dr. César Palencia de Lara

Agradecimientos

Han pasado diez años desde que acudí a mi primera clase en la Facultad de Ciencias hasta el día en que termina esta tesis doctoral. Aunque sepa que no puedo recordar a todas las personas que me apoyaron en este camino, me gustaría aprovechar para acordarme de algunas de ellas.

A los directores de esta tesis. A César le estaré agradecido por transmitirme su visión creativa y estética de las matemáticas, y por enseñarme sus conocimientos en esta frontera entre el análisis funcional y el análisis numérico. A Begoña por toda la confianza depositada en mí y el trabajo que ha hecho para que esta tesis pueda salir adelante, desde sus valiosas aportaciones hasta las trabas burocráticas. También al resto de profesores del Departamento de Matemática Aplicada, especialmente a Mari Paz Calvo, por hacerme ver bonito el análisis numérico, como profesora, y el apoyo que me ha brindado durante estos cuatro años, como compañera. Y a Jesús Dueñas, por todas las conversaciones sobre matemáticas que me han abierto la mente y motivado a seguir aprendiendo. Al profesor Alexander Ostermann, por sus valiosas lecciones y tratarme como a un miembro más de su equipo en la Universidad de Innsbruck.

A mi madre, mi padre y mis hermanos, por todo el cariño y apoyo incondicional que me han dado siempre, tanto en las buenas como en las malas. Gracias por creer siempre que sería capaz de hacerlo. Y a mi abuelo, por valorarlo aunque no lo entienda.

A todos mis amigos y amigas que se dieron una vuelta al lago del Miguel Delibes conmigo. Parece una felicitación genérica, pero tengo la suerte de que deja a muy poca gente fuera. Gracias por apoyarme y levantarme el ánimo cuando hace falta.

A Raquel, por hacerme volver a mirar el mundo con más optimismo y por apoyarme siempre que lo necesito.

A mis compañeras y compañeros, por recodarme día a día nuestro lugar en el mundo y mantenerme atado a la realidad. A fin de cuentas, *la ciencia no puede establecer fines e, incluso menos, inculcarlos en los seres humanos; la ciencia puede proveer los medios con los que lograr ciertos fines. Pero los fines por sí mismos son concebidos por personas con altos ideales éticos y –si estos fines no son endebles, sino vitales y vigorosos– son adoptados y llevados adelante por muchos seres humanos quienes, de forma semi-inconsciente, determinan la evolución lenta de la sociedad.*

Es ist wahr, ein Mathematiker, der nicht etwas Poet ist, wird nimmer ein vollkommener Mathematiker sein

---

Sofia Kovalévskaya.

Contents

Introduction

Evolutionary Partial Differential Equations (PDEs) constitute a fundamental mathematical framework for modelling a broad spectrum of time-dependent phenomena governed by underlying physical laws. These equations model everything from classic phenomena such as heat and chemical diffusion, atmospheric and fluid dynamics, and the propagation of vibrational phenomena to complex phase transitions in materials science, the spatio-temporal spread of pandemics, and even large-scale dynamics in the social sciences. Nevertheless, analytical solutions are available only in a very limited number of specific cases. This limitation makes the development of efficient and stable numerical methods essential for bridging these models with their practical applications. In this context, the present thesis is situated within the field of the numerical analysis of partial differential equations.

Partial Differential Equations can be studied through a wide range of approaches. In general, the diversity of their properties and qualitative behaviours makes a unified analysis challenging, if not impossible. The theory of semigroups of operators, however, allows a broad class of evolutionary PDEs to be formulated as abstract evolution equations in a common framework of Banach spaces X. This perspective is particularly advantageous for numerical analysis, and especially for the design of time-integration methods, as it provides a unified framework to study the stability and convergence of numerical schemes across a wide variety of PDEs. This abstract formulation does not, however, resolve all difficulties. The gain in generality shifts the challenge to verify whether each equation under consideration fits within the framework, and in what precise manner. For this theory, we will rely on the following standard references [31, 32, 39, 53, 66, 77].

Once a PDE is formulated, approximating its solutions typically involves two main steps. First, the equation must be discretized in space, which transforms the PDE into a system of differential equations whose integration often leads to a stiff problem. The main challenges at this stage stem from the discretization of the differential operators while accounting for the geometry of the domain and the boundary conditions. Common approaches include finite difference, finite element, and spectral methods. In turn, the resulting semi-discrete system must be integrated in time using an appropriate scheme, with multi-step and Runge–Kutta methods being among the most widely employed.

Our work focuses on the time integration of PDEs after spatial discretization. Multistep methods, particularly BDF schemes, are widely used due to their versatility and ease of implementation. However, incorporating variable step size strategies in this framework is considerably more involved. A more restrictive limitation arises from the second Dahlquist barrier, which establishes that no explicit linear multistep method can be A-stable. Moreover, the maximal order of an implicit A-stable linear multistep method is two. BDF methods, which are A(ϑ)-stable up to order six, see their stability regions shrink as the order increases, rendering them impractical beyond order six. While these methods remain suitable in many contexts, the pursuit of stable numerical schemes with higher orders of convergence does not end with them.

Alternatively, Runge–Kutta methods constitute another major class of time integration schemes. Among its advantages are ease of implementation, including the possibility of easily integrating variable-step strategies, as well as the large number of scenarios in which they can be used. However, the biggest drawback of these methods in the integration of PDEs comes from the phenomenon of order reduction [71, 76]. It happens that for a Runge–Kutta method of order p, applied to approximate a solution $u \in \mathcal{C}([0,\infty), X)$, exhibits an order of convergence $0 \leq \mu \leq p$, which is related to the stage order of the method $q \leq p$ rather than to p itself. In the context of classical PDEs, μ is fractional, no matter how regular the solution u is (in space and time). In [8, 59, 60], optimal orders of these methods are shown. Because of this, several techniques in the literature have been devised to avoid it.

Some of them are based on considering additional restrictions on the coefficients of the methods so that they not only satisfy the classical order conditions, but also some stiff order ones [48, 52] or, more recently, weak stage order conditions [16, 14, 15]. That implies less freedom in the choice of coefficients, so that the error constants of the methods cannot be minimized in the same way, and also the number of stages which is required to obtain a certain order of accuracy may increase, and thus the computational cost of the method. A similar, interesting approach was recently introduced in [68], where the authors add new nodes to the Runge–Kutta tableau, leading to the need of adding some extra evaluations of f.

Another technique for linear problems was suggested in [21], which consists of converting the problem, through the solution of several elliptic problems, to one for which order reduction is not observed. This procedure has the advantage to be valid for any method, but the solution of the corresponding elliptic problems also means a non-negligible computational cost. Moreover, the generalization of this technique to nonlinear problems has just been performed in [20], where the non-natural hypothesis that $f(t, u)$ vanishes for nul u must be made.

A third procedure is based on modifying the boundary values for the stages which are in some way predetermined [1, 2, 3, 4, 5, 6, 28, 65]. That means very little computational cost because the number of nodes on the boundary is negligible with respect to the number of nodes on the whole domain. For linear problems, the expressions for the modified boundaries depend on spatial and time derivatives of data [2] (more particularly, the boundary condition and the source term). If analytic expressions are known for data, that is not a problem. However, in many practical problems, an analytic expression is not known, but just the values at some instants of time. Due to that, numerical differentiation is required to approximate the required modified boundary values for the stages, and it is well known that numerical differentiation is unstable when the grid is refined [70]. For nonlinear problems, there exists the need to resort to numerical differentiation, even if analytic expressions of data are known, when the required order is high enough [6]. We attempt to avoid this issue.

We also note that order reduction is not exclusive to Runge–Kutta methods; it also manifests in other families, for instance, in exponential and deferred correction methods. The exponential methods introduced in [42, 43, 44] are designed to satisfy some stiff order conditions thereby avoiding order reduction for homogeneous boundary conditions. However, the problem persists for time-dependent boundary conditions. In a series of papers [23, 24, 25, 26, 27], modifications to the exponential Runge–Kutta methods are proposed to alleviate or avoid the order reduction, while other approaches consider a correction of the problem itself [9]. Similarly, this issue has been studied for spectral deferred methods in various papers [58, 73]. Multistep methods do not suffer from this phenomenon, but they have other practical limitations such as the already mentioned second Dahlquist barrier.

The main objective of this PhD dissertation is to design a new family of rational methods that overcome the order reduction phenomenon. We build upon an A-stable rational function of order p that approximates the exponential, which may coincide with the rational stability function of a Runge–Kutta method (see [38]). On this basis, we propose and analyze a family of numerical schemes that, on the one hand, inherit the order of convergence p of the rational approximation, and, on the other hand, need only evaluations of the source term and boundary data. In doing so, they avoid the need for numerical differentiation that, as we commented, is one of the main drawbacks of most existing methods designed to avoid order reduction. In the works that give rise to this thesis [11, 12, 10], and throughout this report, we consider different versions of the abstract evolution problem

$$\begin{cases} u'(t) = Au(t) + f(t, u(t)), \quad 0 < t < T, \\ u(0) = u_0, \\ \partial u(t) = g(t), \quad 0 < t < T, \end{cases} \tag{1}$$

so that ideas for designing methods that avoid order reduction are progressively developed through increasingly complex versions of the problem.

In Chapter 1, we have summarized some of the results necessary for the development of the theory we present. These are not new results, but we believe it is appropriate to include them in this report for two reasons: (i) to present them in an organized manner that facilitates clarity of exposition, and (ii) to make the report more accessible to researchers in applied fields who may not be fully familiar with some of the concepts involved. Accordingly, in Section 1.1 we recall the basic notions of semigroup theory, including the definitions of semigroups of operators and analytic semigroups, the Hille–Yosida theorem, and some illustrative examples. Since much of our analysis and convergence results require the use of an operational calculus—that is, a framework for constructing and studying operators of the form $f(A)$, where f is a function and A a (generally unbounded) operator in a Banach space—we review typical constructions of this kind in Section 1.2. Finally, we conclude the chapter by recalling the fundamental properties of Runge–Kutta methods. This includes, on the one hand, a review of the rational approximation of semigroups by rational functions of their generators and the corresponding properties (see [18, 17, 29, 36, 40, 49, 61, 62, 63]), and, on the other hand, a discussion of the origin of the order reduction phenomenon in the setting of Banach spaces.

Once we have reviewed these concepts, we address the construction of the announced methods. Following [10], we consider first the linear, nonhomogeneous version of (1) with homogeneous boundary conditions. The idea to construct them is based on the remark that the homogeneous version of (1) (i.e., $f = 0$) can be discretized by using an A-acceptable approximation $r(z)$ to e^z. To this end, it is enough to implement the *rational approximation*, defined by the recurrence

$$u_{n+1} = r(\tau A)u_n, \qquad n \geq 0, \tag{2}$$

with initial value $u_0 \in X$ and constant step size $\tau > 0$. In this situation there is no order reduction [18]. A step in the recurrence needs solving s linear systems involving A, where s is the number of poles (accounted along their multiplicities) of $r(z)$. Furthermore, when $f = 0$, an A-stable RK method applied to such a homogeneous problem becomes the rational method based on its own stability function. Notice that in this situation the abcissa $\mathbf{c}$ of the Runge–Kutta tableau are not required whatsover.

The main idea is just to cast a non homogeneous IVP into an enlarged, homogeneous problem which is then discretized by a rational method. Essentially, this is achieved by treating f as a new unknown (see Section 2.2), in the line of the approach used in [33] for equations with memory. The resulting discretization is in principle theoretical, but can be implemented within the optimal order just by using auxiliary evaluations of f. To this end, some discrete time grid is required and it turns out that sensible choices of such grids lead to procedures that, per step, require (i) just a new evaluation of f and (ii) solving a number s of linear systems. Thus, we propose a procedure that avoids the order reduction phenomenon. When $r(z)$ is the stability function of a RK method, the new approach maintains the same number s of linear systems per step, as in the Runge–Kutta case, but now only one new evaluation of f is needed. In Section 2.3, we discuss some implementation issues and show numerical illustrations in simple PDEs.

In Chapter 3, we address the extension of these methods to the semilinear case, still under homogeneous boundary conditions. The basic idea for adapting the scheme to this new framework is straightforward. The freedom to select the time grid in the linear case is now exploited in order to evaluate the nonlinear source term $f(t, u(t))$ at time points where the numerical approximations u_n can be used to replace $u(t_n)$, leading to $f(t_n, u(t_n)) \approx f(t_n, u_n)$. However, the nonlinear setting requires a more refined analysis, distinguishing between different types of nonlinearities and incorporating sharp regularization estimates. A new version of the discrete Gronwall inequality is also required for the analysis. Finally, we conclude the chapter by discussing some implementation aspects and presenting numerical examples that illustrate the theory.

This work concludes by extending rational methods to initial boundary value problems. In these problems, order reduction is more pronounced: the convergence order typically decreases by one compared to homogeneous boundary conditions. By a standard transformation, such problems can be reformulated as abstract evolution equations in a Banach space, with boundary conditions appearing as a source term involving a derivative. Using the techniques introduced earlier, we construct a numerical scheme that recovers the optimal order p of the rational function without requiring numerical differentiation of the data.

In this chapter, we also study the full discretization of the method, including the effects of spatial discretization in the error bounds. To illustrate this, the examples employ a higher-order spatial discretization, which also leads to a mass matrix similar to that arising in finite element methods. This demonstrates the computational versatility of the proposed rational methods, making them suitable for realistic applications requiring complex spatial discretizations (meteorology, fluid mechanics, etc).

Finally, the chapter closes by synthesizing the obtained results along the two parallel directions pursued after Chapter 2: the extension of the methods to semilinear problems, on the one hand, and to initial boundary value problems, on the other. With the results established up to this point, it becomes straightforward to analyze the methods for semilinear problems with boundary conditions, with which this thesis comes to its conclusion.

Chapter 1

Preliminaries

> From a philosophical perspective, the exponential function may be viewed as a link between the seemingly contradictory positions of Heraclitus on the one side and Parmenides on the other,. . . . While the time dependent function $t \mapsto T(t)$—the *semigroup*—reflects the aspect of permanent change in a deterministic autonomous system, its *generator* A stands for the eternal, timeless principle behind the system. The exponential functions ties both aspects together through the formula $T(t) = \exp(tA)$.
>
> Tanja Hahn and Carla Perazzolli.

As it is explained in the Introduction, the framework adopted in our study of Partial Differential Equations relies on various techniques from functional analysis. Given that both the analytical treatment of PDEs and their numerical analysis can be approached in many different ways, we find it useful to summarize the main results that we regard as essential for understanding the construction, analysis, and implementation of the methods that form the core of this work. While most of these results are well-established and familiar to specialists in the field, they are included here for the sake of completeness and to assist readers who may not be acquainted with this particular perspective.

In Section 1.1 we introduce the basic concepts of the theory of semigroups of bounded linear operators, that allows us to formulate the Partial Differential Equations we consider as abstract evolution equations in Banach spaces. This abstract framework is useful because we can consider simultaneously a wide variety of PDEs within the same framework. We briefly present the basic definitions, generation results and describe some archetypal examples that will appear in our theory or numerical experiments. Since most of our analysis of convergence and stability results are based on the rational approximation of these semigruoups, we consider convenient to explain how different operators of the form $f(A)$, where f is in a certain class of funtions and A is the generator of the semigroup, are constructed. This is the content of Section 1.2. Finally, in Section 1.3 we review some aspects of the Runge–Kutta methods in relation to the time integration of PDEs. On the one hand, we recall basic definitions of these methods, with special emphasis on the rational stablity functions, and then we enumerate the fundamental stability and approximation results for semigroups obtained via these rational functions, which constitute the basis for the analysis of our methods. On the other hand, we conclude with some remarks on the implementations of these methods and the causes of the order reduction phenomenon, which is precisely the issue we aim to avoid through the design of our methods.

1.1 Semigroups of linear operators

1.1.1 Definitions and first properties

Partial Differential Equations (PDEs) are those who govern a wide variety of models in natural and social science. The study of these equations may be considered from many different approaches that contemplate levels of abstraction and generality. To our purposes, we find it very convenient to consider those PDEs that can be formulated under the framework of semigroups of operators in Banach spaces developed in [31, 32, 53, 66, 77]. This framework covers many of the equations that are commonly encountered in practice and it allows us to work with them in an unitary way.

Definition 1.1 (Semigroup of operators)**.** Let X be a Banach space. A one parameter family $\{T(t)\}_{t\geq 0}$ of bounded linear operators from X into X is a *semigroup of bounded linear operators on* X if

(a) $T(0) = I$, the identity operator on X.

(b) $T(t+s) = T(t)T(s)$ for every $t, s \geq 0$ (the semigroup property).

One of these semigroups is said to be *strongly continuous* if

$$\lim_{t\to 0} T(t)\, x = x \quad \text{for every } x \in X.$$

A strongly continuous semigroup will be called a $\mathcal{C}_0$ semigroup.

Definition 1.2 (Infinitesimal generator of a semigroup)**.** The linear operator A defined in the domain

$$D\left(A\right) = \left\{ x \in X : \lim_{t\to 0} \frac{T(t)x - x}{t} \in X \right\}$$

by

$$Ax = \lim_{t\to 0} \frac{T(t)x - x}{t}, \quad \text{for } x \in D\left(A\right), \tag{1.1}$$

is the *infinitesimal generator* of the semigroup $\{T(t)\}_{t\geq 0}$. The semigroups of linear operators are usually written as $\left\{\mathrm{e}^{tA}\right\}_{t\geq 0}$ due to the analogy between their properties and those of the exponential function. We will adopt this notation in what follows.

For every semigroup $\left\{\mathrm{e}^{tA}\right\}_{t\geq 0}$ of class $\mathcal{C}_0$ there exist constants $\omega \in \mathbb{R}$ and $M \geq 1$ such that

$$\|\mathrm{e}^{tA}\| \leq M\mathrm{e}^{\omega t}. \tag{1.2}$$

If $\omega \leq 0$, $\left\{\mathrm{e}^{tA}\right\}_{t\geq 0}$ is called *uniformly bounded* and, if additionally $M = 1$, it is called a $\mathcal{C}_0$ *semigroup of contractions*. For every $M \geq 1$, $\omega \in \mathbb{R}$ we denote by $\mathcal{G}\left(X, M, \omega\right)$ the set of all the infinitesimal generators of $\mathcal{C}_0$ semigroups of operators in X satisfying the bound (1.2).

Recall that if A is a linear, not necessarily bounded operator in X, the resolvent set $\rho\,(A)$ of A is the open set of all complex numbers λ for which $\lambda I - A$ is invertible, that is, $(\lambda I - A)^{-1}$ is a bounded linear operator in X. The family $R(\lambda) = (\lambda I - A)^{-1}$, $\lambda \in \rho\,(A)$, of bounded linear operators is called the resolvent of A. The following theorem, whose first version was proved independently by Einar Hille and Kōsaku Yosida in 1948, establishes sufficient and necessary conditions on the resolvent of linear operators A to be the infinitesimal generator of a semigroup of operators.

Theorem 1.3 (Hille–Yosida)**.** A linear (unbounded) operator A is the infinitesimal generator of a $\mathcal{C}_0$ semigroup of operators $\left\{\mathrm{e}^{tA}\right\}_{t\geq 0}$, that is, $A \in \mathcal{G}(X, M, \omega)$ for some $M \geq 1$ and $\omega \in \mathbb{R}$, if and only if

(a) A is closed and $D\,(A)$ is dense in X.

(b) The resolvent set $\rho\,(A)$ of A contains the half plane $\{\lambda \in \mathbb{C} : \mathrm{Re}(\lambda) > \omega\}$ and, for every $\mathrm{Re}(\lambda) > \omega$,

$$\|\,(\lambda I - A)^{-n}\,\| \leq \frac{M}{(\mathrm{Re}\,\lambda - \omega)^n} \qquad n = 1, 2, \ldots .$$

One of the consequences of the previous theorem is the fact that the resolvent of A is the Laplace transform of the corresponding semigroup. In fact, if $A \in \mathcal{G}(X, M, \omega)$, then for every $\lambda \in \mathbb{C}$ on the half plane $\mathrm{Re}\,\lambda \geq \omega$, it is true that, for every $u \in X$,

$$(\lambda I - A)^{-1}\; u = \int_0^\infty \mathrm{e}^{-\lambda t}\,\mathrm{e}^{tA}\,u\,dt. \tag{1.3}$$

Notice that the previous integral is absolutely convergent due to (1.2).

Now, we are in conditions to formulate the simplest case of abstract evolution Cauchy problem in the Banach space X: the homogeneous initial value problem (see, e.g., Theorem 4.1.3 in [66]). As we will show in the examples, this is an abstract framework for various differential equations problems.

Theorem 1.4. A function $u : [0, \infty) \to X$ is a classical solution of the initial value problem

$$\begin{cases} u'(t) = Au(t) & t > 0, \\ u(0) = u_0, \end{cases} \tag{1.4}$$

if it is continuous for $t \geq 0$, continuously differentiable and $u(t) \in D\,(A)$ for $t > 0$ and (1.4) is satisfied. The initial value problem (1.4) has a unique classical solution $u(t)$, which is continuously differentiable on $[0, \infty)$, for every initial value $u_0 \in D\,(A)$, if and only if A is the infinitesimal generator of a $\mathcal{C}_0$ semigroup.

1.1.2 Analytic semigroups

The semigroups related to parabolic evolution problems typically may be extended from the ray $[0,\infty)$ to a sector on the complex plane. These are the semigroups whose infinitesimal generator A is *sectorial*, that is, such that the following resolvent bound

$$\| (\lambda I - A)^{-1} \| \leq \frac{M}{|\lambda - \omega|}, \tag{1.5}$$

where $M \geq 1$ is satisfied on the sector $\{\lambda \in \mathbb{C} : 0 \leq |\arg(\lambda - \omega)| \leq \pi - \theta, \lambda \neq \omega\}$, for a certain angle $0 < \theta < \pi/2$. The semigroup of operators generated by one of these generators results to be defined in a sector of the form $S_\theta = \{z \in \mathbb{C} : |\arg(z)| < \pi/2 - \theta\}$. Notice that the angles defining both sectors are complementary.

Definition 1.5 (Analytic semigroup)**.** Let S_θ, $0 < \theta < \pi/2$, be a right-hand sector on the complex plane and $T(z)$ be a bounded linear operator for every $z \in S_\theta$. We say that $\{T(z)\}_{z \in S_\theta}$ is *analytic semigroup in* S_θ if

1. $z \mapsto T(z)$ is analytic in S_θ,
2. $T(0) = I$ and $\lim\limits_{\substack{z \to 0 \\ z \in S_\theta}} T(z)x = x$, for every $x \in X$.
3. $T(z_1 + z_2) = T(z_1)T(z_2)$, for $z_1, z_2 \in S_\theta$.

A semigroup is called analytic if it is analytic in some sector S_θ containing the real axis. We may denote it as $\left\{\mathrm{e}^{zA}\right\}_{z \in S_\theta}$.

Clearly, the restriction of an analytic semigroup to the real positive axis is a $\mathcal{C}_0$ semigroup. The following theorem (see, e.g., Theorem 2.5.2 in [66]) gives necessary and sufficient conditions for a $\mathcal{C}_0$ semigroup to be extended to a sector.

Theorem 1.6. Let $A \in \mathcal{G}(X, M, \omega)$ be the infinitesimal generator of a linear semigroup of operators $\left\{\mathrm{e}^{tA}\right\}_{t \geq 0}$, and fix $0 < \theta < \pi/2$. The following statements are equivalent:

1. $\left\{\mathrm{e}^{tA}\right\}_{t \geq 0}$ can be extended to an analytic semigroup in S_θ and
$$\|T(z)\| \leq M_{\theta'} \, \mathrm{e}^{\omega \operatorname{Re} z} \tag{1.6}$$
for every closed subset $\bar{S}_{\theta'}$, $\theta' > \theta$, of S_θ.
2. There exists a constant C such that for every $\lambda = \sigma + \tau i$ with $\sigma > \omega$, $\tau \neq 0$,
$$\| (\lambda I - A)^{-1} \| \leq \frac{C}{|\tau|}. \tag{1.7}$$
3. There exists $M \geq 1$ such that
$$\| (\lambda I - A)^{-1} \| \leq \frac{M}{|\lambda - \omega|} \tag{1.8}$$
on the sector $\{\lambda \in \mathbb{C} : 0 \leq |\arg(\lambda - \omega)| \leq \pi - \theta, \lambda \neq \omega\}$.

4. e^{tA} is differentiable for $t > 0$ and there exists a constant C such that
$$\|A\,\mathrm{e}^{tA}\| \leq \frac{C}{t}\,\mathrm{e}^{\omega t} \quad \text{for } t > 0. \tag{1.9}$$

We denote by $\mathcal{G}(X, M, \omega, \theta)$ the set of all the infinitesimal generators of analytic semigroups in the sector S_θ that satisfy these equivalent conditions.

Notice that the last condition in the previous theorem implies that for every $t > 0$, the operator $\mathrm{e}^{tA} : X \to D\,(A)$ is bounded, this property is typically known as *parabolic smoothing*. One consequence of this property is that the hypothesis $u_0 \in D(A)$ in Theorem 1.4, required for the initial value problem (1.4) to admit a unique solution, can be weakened to $u_0 \in X$. This relaxes the regularity requirements on the initial data, as the following theorem shows (see, e.g., Theorem 4.1.4 in [66]).

Theorem 1.7. If A is the infinitesimal generator of an analytic semigroup, then for every $u_0 \in X$ the initial value problem
$$\begin{cases} u'(t) = Au(t) \quad t > 0, \\ u(0) = u_0, \end{cases}$$
has a unique solution.

Even for more complex initial value problems, weaker assumptions on the initial data are possible when the semigroup is analytic.

1.1.3 Some examples

In the following, we explain some paradigmatic examples with a double purpose: to clarify the theory and to present some of the cases we will use in the numerical examples.

The matrix exponential

Let X be the euclidean space $\mathbb{R}^d$ and A be the matrix of a linear operator $A : X \to X$. It is well known (e.g., [67]) that the solution of the linear system of differential equations
$$\begin{cases} u'(t) = Au(t) \quad t > 0, \\ u(0) = u_0 \in \mathbb{R}^d, \end{cases} \tag{1.10}$$
has a unique solution given by the matrix exponential
$$\mathrm{e}^{tA} = \sum_{n=0}^{\infty} \frac{(tA)^n}{n!}$$
by means of $u(t) = \mathrm{e}^{tA}u_0$. Notice that A has a finite spectrum and it clearly satisfies the hypothesis of Hille–Yosida theorem, $D(A) = X$ and the properties of the matrix exponential guarantee that $\left\{\mathrm{e}^{tA}\right\}_{t\geq 0}$ is in fact a semigroup of linear operators with generator A. Moreover, if every $\lambda \in \sigma(A)$ is such that $\operatorname{Re}\lambda < \omega$, then
$$\|\mathrm{e}^{tA}\| \leq M\mathrm{e}^{t\omega},$$

for a constant $M \geq 1$, that is, $A \in \mathcal{G}\left(\mathbb{R}^d, M, \omega\right)$. The fact that $\sigma(A)$ is finite, and then bounded, implies that $-A$ is also the generator of a semigroup and then $\left\{\mathrm{e}^{tA}\right\}_{t\in\mathbb{R}}$ is in fact a group. It can also be proved that A generates an analytic semigroup, since it is straightforward to find a sector that contains the bounded set $\sigma(A)$.

The translation semigroup

Let $X = \mathcal{C}_{ub}\left([0,\infty), \mathbb{C}\right)$ be the space of all bounded, uniformly continuous functions on $[0,\infty)$ endowed with the supremum norm $\|\cdot\|_\infty$ (see [32] for different choices of X). The operators $T(t) : X \to X$, $t \geq 0$, defined by

$$\left(T(t)f\right)(s) = f(t+s), \quad t, s \in [0,\infty) \tag{1.11}$$

are called the left translations by t. We claim that $\{T(t)\}_{t\geq 0}$ is a $\mathcal{C}_0$ semigroup of operators in X since:

1. $T(0)f = f$, $T(0)$ is the identity operator in X.
2. $\left(T(t)T(s)f\right)(r) = f(t+s+r) = \left(T(t+s)f\right)(r)$, the semigroup property holds,
3. $\lim_{t\to 0} \|T(t)f - f\|_\infty = 0$ since the functions are uniformly continuous.

Moreover, $\|T(t)f\|_\infty \leq \|f\|_\infty$, so in this case we can take $M = 1$ and $\omega = 0$; it is in fact a $\mathcal{C}_0$ semigroup of contractions. A direct calculation shows that, if f is differentiable,

$$\lim_{t\to 0} \frac{\left(T(t)f\right)(s) - f(s)}{t} = \lim_{t\to 0} \frac{f(t+s) - f(s)}{t} = f'(s),$$

so the infinitesimal generator A of the semigroup is defined by $Af = f'$ for every $f \in D(A)$ in the domain $D(A) = \{f \in X : f' \in X\}$. However, this semigroup is not analytic unless we impose further conditions on the functions of the space X.

The diffusion semigroup

The details of this construction can be found, for instance, in [19, 66]. Let $\Omega \subset \mathbb{R}^d$ be a regular domain with smooth boundary $\partial\Omega$ and the space of complex-valued functions $X = L^2(\Omega)$. We take $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$ and define a realisation of the Laplacian operator $A : D(A) \to X$ by $Af = \Delta f$, for every $f \in D(A)$. In this case, the abstract evolution problem (1.4) stands for the heat diffusion equation in Ω,

$$\begin{cases} u'(t) = \Delta u(t), \quad t > 0, \\ u(0) = u_0 \in X, \\ u(t)|_{\partial\Omega} = 0, \quad t > 0, \end{cases} \tag{1.12}$$

for a given initial data. Notice that the homogeneous Dirichlet boundary conditions have been incorporated in the definition of the domain A, since $u|_{\partial\Omega} = 0$ for every $u \in H_0^1(\Omega)$. One possibility is to write the solution of (1.12) taking into account that there exists a basis of the Hilbert space $L^2(\Omega)$ formed by eigenfunctions of Δ (with zero Dirichlet boundary conditions), i.e.,

$$\begin{cases} \Delta e_n = \lambda_n e_n, & \text{in}\, \Omega, \\ e_n = 0, & \text{in}\, \partial\Omega, \end{cases} \tag{1.13}$$

with $\lambda_n < 0$, for $n = 1, 2, \dots$. Given an initial data $u_0 \in X$,

$$u_0 = \sum_{n=1}^{\infty} \langle u_0, e_n \rangle \, e_n,$$

and the solution is given by the function

$$u(t) = \sum_{n=1}^{\infty} \langle u_0, e_n \rangle \, \mathrm{e}^{\lambda_n t} \, e_n$$

It can then be shown that the operator $T(t) : L^2(\Omega) \to L^2(\Omega)$ given by

$$u_0 = \sum_{n=1}^{\infty} \langle u_0, e_n \rangle \, e_n \mapsto T(t) u_0 = \sum_{n=1}^{\infty} \langle u_0, e_n \rangle \, \mathrm{e}^{\lambda_n t} \, e_n, \quad t > 0,$$

is a semigroup of bounded, linear operators with infinitesimal generator $A = \Delta$. Moreover, the operator A generates an analytic semigroup that can be extended to every sector $S_\theta = \{z \in \mathbb{C} : |\arg(z)| < \theta\}$, $0 < \theta < \pi/2$. As a consequence, the initial value problem (1.12) has a solution for every $u_0 \in X$ and the semigroup regularises the solution for any positive time, that is, $\mathrm{e}^{tA} u_0 \in D(A)$ for every $t > 0$.

Within the framework of semigroups of operators, more complicated variants of this problem can also be formulated. For example, one can prove that a parabolic operator of the form

$$A(x, D) u = \sum_{|\alpha| \leq 2m} a_\alpha(x) D^\alpha u \tag{1.14}$$

in the space $L^p(\Omega)$, $1 < p < \infty$, generates an analytic semigroup whenever the coefficients a_α are sufficiently smooth functions and the operator satisfies the *strong ellipticity condition*, that is, when there exists a constant $c > 0$ such that

$$\mathrm{Re}(-1)^m \sum_{|\alpha| = 2m} a_\alpha(x) \, \xi^\alpha \geq c \, |\xi|^{2m},$$

for every $x \in \bar{\Omega}$ and $\xi \in \mathbb{R}^d$.

Another typical example where an explicit formula for the semigroup can be given is the case where $\Omega = \mathbb{R}^d$. Here the Laplacian operator $A = \Delta$ is defined on the domain $D(A) = H^2(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$. The solution of the diffusion problem

$$\begin{cases} u_t(t, x) = \Delta u(t, x), & t > 0, \quad x \in \mathbb{R}^d, \\ u(0, x) = u_0(x) \in L^2(\mathbb{R}^d), & x \in \mathbb{R}^d, \end{cases} \tag{1.15}$$

is given explicitly by the convolution with the *Gaussian heat kernel*:

$$K(t,x) = \frac{1}{(4\pi t)^{d/2}} \exp\left(-\frac{|x|^2}{4t}\right), \quad t > 0, \quad x \in \mathbb{R}^d. \tag{1.16}$$

The solution operator takes the form

$$u(t,x) = (T(t)u_0)(x) = (K_t * u_0)(x) = \int_{\mathbb{R}^d} K(t,x-y)u_0(y)dy. \tag{1.17}$$

This defines again an analytic semigroup with infinitesimal generator A. This construction extends to $L^p(\mathbb{R}^d)$ for $1 \le p < \infty$. The Gaussian kernel representation is fundamental in parabolic theory, connecting semigroup methods with classical solutions.

1.2 Operational calculus

In the context of closed operators $A \in \mathcal{G}(X, M, \omega)$, most of the results presented in this text require defining the operator $f(A)$ for a certain class of functions f. This can be done in different ways to work with various types of functions and operators, although always in such a way that these constructions coincide for functions that admit both definitions. Throughout this work, it will be useful to employ two approaches: the functional calculus of Hille–Phillips [41] and that of Dunford–Taylor [30]. For instance, the results of convergence and stability in [18, 40] use the Hille–Phillips approach while [8, 36, 49, 61, 62, 63] require techniques similar to Dunford–Taylor calculus. After presenting both definitions and their most important properties, we explain why they are equivalent and we show some examples.

The easiest case to define the function of an operator is to consider polynomials p of a certain degree. In this case, the operators $p(A)$ can be defined in a recursive way. We set

$$D\left(A^n\right) = \left\{x \in D\left(A^{n-1}\right) : A^{n-1}x \in D\left(A\right)\right\} \quad \text{and } A^n x = A\left(A^{n-1}x\right),$$

and for $p(z) = a_0 + a_1\, z + \cdots + a_n\, z^n$, we have the unbounded operator $p(A) : D\left(A^n\right) \to X$ defined by

$$p(A) = a_0\, I + a_1\, A + \cdots + a_n\, A^n.$$

It turns out that $p(A)$ is also a closed operator and the spectral mapping property $\sigma(p(A)) = p(\sigma(A))$ is satisfied (see [30]). When f is a general analytic function, it is not immediately clear how to define the operator $f(A)$. Nonetheless, we seek a definition that satisfies certain properties which lend meaning to the concept. In particular, we require that if $Au = \lambda u$ for some $u \in X$, then $f(A)\,u = f(\lambda)\,u$. In fact, this may be the basis for defining $f(A)$ when $X = \mathbb{R}^d$ and A is a diagonalizable matrix. In the infinite-dimensional case that concerns us, this property will follow as a consequence of the definition.

The operational calculus of Dunford–Taylor

In [69], in the context of Banach algebras, the elements f(x) are defined for every element x of the Banach algebra when f is analytic in a neighbourhood of the spectrum $\sigma(x)$. In our context of linear operators in Banach spaces, this guarantees that if A is a bounded, linear operator with spectrum $\sigma(A)$ (that is also bounded), f is analytic in a neighbourhood of $\sigma(A)$ and Γ is a simple contour positively oriented surrounding $\sigma(A)$ inside that neighbourhood, then the formula

$$f(A) = \frac{1}{2\pi i}\int_\Gamma f(\lambda)\,(\lambda I - A)^{-1}\;d\lambda$$

defines an operator that satisfy the properties we expect. The key to extend this definition is in [30]. We assume that $f : U \to \mathbb{C}$ is analytic in U and at infinity, that is, $g(z) = f(1/z)$ is analytic in $z = 0$. As a consequence, the following limit exists

$$\lim_{z\to\infty} f(z) = \lim_{z\to 0} g(z) = f(\infty) \in \mathbb{C}.$$

Since f is analytic in $z = \infty$, its poles f are contained in a disk $D(0, M)$. Then, for $M < r < R$, the Cauchy theorem and the fact that f is analytic guarantee that

$$f(z) = \frac{1}{2\pi i}\int_{\Gamma_R} f(\lambda)\,(\lambda I - A)^{-1}\;d\lambda - \frac{1}{2\pi i}\int_{\Gamma_r} f(\lambda)\,(\lambda I - A)^{-1}\;d\lambda, \quad \text{if}\, r < |z| < R,$$

with Γ_R, Γ_r circumferences of radius R, r. Taking limits when $R \to \infty$ and using again the Cauchy theorem,

$$f(z) = f(\infty) + \frac{1}{2\pi i}\int_\Gamma f(\lambda)\,(\lambda I - A)^{-1}\;d\lambda, \tag{1.18}$$

where Γ is a simple contour whose bounded component contains the poles of f and is positively oriented.

To consider bounded, closed operators A whose spectrum $\sigma(A)$ is, in general, unbounded, we set a conformal mapping of the form $\Phi(\mu) = (\mu - \alpha)^{-1}$, with $\alpha \in \rho(A)$, to define

$$f(A) = f\left(\Phi^{-1}\,(A - \alpha I)^{-1}\right),$$

with the formula (1.18) and the bounded operator $(A - \alpha I)^{-1}$. The following result summarises the properties of these operators (see Theorems 7.4.4-7.4.10 in [30]).

Proposition 1.8. Let $\mathcal{F}(A)$ be the set of analytic functions on a neighbourhood of $\sigma(A)$ and at infinity. Let U be an open set that contains $\sigma(A)$ and whose boundary Γ consists of a finite number of Jordan curves and such that f is analytic on $U \cup \Gamma$. Then, for $f \in \mathcal{F}(A)$,

$$f(A) = f(\infty)\,I + \frac{1}{2\pi i}\int_\Gamma f(\lambda)\,(\lambda I - A)^{-1}\;d\lambda. \tag{1.19}$$

In addition, for $f, g, h \in \mathcal{F}(A)$, h with a zero of order m, $0 \le m \le \infty$ at infinity, p polynomial of degree n, the following statements hold true:

1. $(f + g)(A) = f(A) + g(A)$,

2. $(fg)(A) = f(A)g(A)$

3. $\sigma(f(A)) = f(\sigma(A) \cup \{\infty\})$, and $\sigma(p(A)) = p(\sigma(A))$.

4. If $g \in \mathcal{F}(f(A))$, $F(z) = g(f(z))$, then $F \in \mathcal{F}(A)$ and $F(A) = g(f(A))$.

5. If $Au = \lambda u$ for $\lambda \in \sigma(A)$, then $f(A)u = f(\lambda)u$.

6. If $u \in D(A^n)$, then $h(A)u \in D(A^{n+m})$, where $m + n = \infty$ if $m = \infty$, and $p(A)f(A)u = f(A)p(A)u$.

7. If $0 \leq n \leq m$ and $h_0(z) = p(z)h(z)$, then $h_0 \in \mathcal{F}(A)$ and $h_0(A) = p(A)h(A)$.

The operational calculus of Hille–Phillips

The framework presented in [41] allows to define $f(A)$ for closed operators $A \in \mathcal{G}\,(X, M, \omega)$ that generate a $\mathcal{C}_0$ semigroup e^{tA} and functions f that are Laplace transform of bounded Borel measures. If μ is a bounded Borel measure on $\mathbb{R}$, with $\operatorname{supp}\mu \subset \mathbb{R}_+$, and $\omega \geq 0$ is such that

$$\int_0^\infty \mathrm{e}^{\omega t}\, d\,|\mu|\,(t) < \infty, \tag{1.20}$$

then we denote by $\widetilde{M}_\omega$ the set of Laplace transforms

$$f(z) = \int_0^\infty \mathrm{e}^{zt}\, d\mu(t), \quad \text{for } \mathrm{Re}(z) \leq \omega, \tag{1.21}$$

of this measures. The idea is to define $f(A)$, for $f \in \widetilde{M}_\omega$ and $A \in \mathcal{G}(X, M, \omega)$ via the Bochner integral

$$f(A) = \int_0^\infty \mathrm{e}^{tA}\, d\mu(t).$$

The properties of this operator are stated in the following result. The first part of the lemma provides an useful criterion to verify if a function f belongs to $\tilde{M}_\omega$, while the second guarantees the correct definition of $f(A)$ and shows a bound for the norm.

Proposition 1.9. Let $\omega \geq 0$ and $A \in \mathcal{G}\,(X, M, \omega)$.

1. Let f be a bounded, analytic function in the half-plane $\{z \in \mathbb{C} : \mathrm{Re}(z) \leq \omega\}$ and $f_{(\omega)}(t) = f(\omega + it)$ for $t \in \mathbb{R}$. Then, if $f_{(\omega)} \in L^2(\mathbb{R})$ and $f'_{(\omega)} \in L^2(\mathbb{R})$, then there exists a bounded measure μ with $\operatorname{supp}\mu \subset \mathbb{R}_+$ and satisfying (1.20) such that

$$f(z) = \int_0^\infty e^{zt} d\mu(t), \text{ for } \mathrm{Re}(z) \leq \omega, \quad \text{and} \quad \int_0^\infty e^{\omega t} d|\mu|(t) \leq \sqrt{2}\|f_{(\omega)}\|_2^{1/2}\|f'_{(\omega)}\|_2^{1/2}, \tag{1.22}$$

that is to say, $f \in \widetilde{M}_\omega$.

2. If f is the Laplace transform of μ, then the integral

$$f(A) = \int_0^\infty \mathrm{e}^{tA}\, d\mu(t), \quad \text{for } \mathrm{Re}(z) \leq \omega,$$

defines a bounded operator in X. The mapping $f \mapsto f(A)$ is an homomorphism from $\widetilde{M}_\omega$ to the algebra of bounded linear operators in X. In addition, we have

$$\|f(A)\| \leq M \int_0^\infty e^{\omega t}\, d|\mu|(t). \tag{1.23}$$

3. If $f, g \in \widetilde{M}_\omega$ and $f(z) = z^k\, g(z)$ for $z \in \mathbb{C}_{\leq \omega}$, $k > 0$, then

$$f(A) = g(A)A^k u, \quad \text{for } k \in D(A^k).$$

   Notice that in this definition it is also required that f is analytic on a neighbourhood of the spectrum of A.

4. If $f \in \widetilde{M}_0$ is the Laplace transform of a bounded Borel measure μ with $\operatorname{supp}\mu \subset \mathbb{R}_+$ and $h(z) = f(\tau z)$ for $\tau > 0$, then $h \in \widetilde{M}_0$ and it is the Laplace transform of another Borel measure ν with $\operatorname{supp}\nu \subset \mathbb{R}_+$ satisfying

$$\int_0^\infty d\,|\mu|\,(t) = \int_0^\infty d\,|\nu|\,(t).$$

   In particular, if $A \in \mathcal{G}(X, M, 0)$, then

$$\|f(\tau A)\| \leq M \int_0^\infty d\,|\mu|\,(t), \quad \text{for } \tau > 0. \tag{1.24}$$

Notice that, under the hypotheses of the previous proposition, if $u \in D(A)$ is such that $Au = \lambda u$, then

$$f(A)u = \int_0^\infty \mathrm{e}^{tA} u\, d\mu(t) = \int_0^\infty \mathrm{e}^{t\lambda} u\, d\mu(t) = f(\lambda)u.$$

Equivalence of the two approaches

The Hille–Phillips construction is more demanding for the operator A, which must be the generator of a $\mathcal{C}_0$ semigroup of operators, not just a closed operator. The Dunford–Taylor construction, however, requires working with analytic functions at infinity, which excludes working with functions such as the exponential defined on a half-plane. Notice that if $f \in \widetilde{M}_\omega$ is the Laplace transform of a measure μ, then it can be easily checked that f is analytic at infinity and $f(\infty) = \mu(\{0\})$. The fact is that the bounds of $f(A)$ (1.22)-(1.23) in terms of the function f are useful in some contexts (see, e.g., [18, 40]), specially working with $\mathcal{C}_0$ semigroups, while in the analytic case it is sometimes more fruitful bounding directly the Cauchy integral (1.19) (see, e.g., [36, 49, 61, 62, 63]).

If for a certain analytic function f and a closed operator A both definitions allows to define the operators $f_{DT}(A)$, $f_{HP}(A)$, then both operators coincide. In fact, if we assume that $A \in \mathcal{G}(X, M, \omega)$ (so that Hille–Phillips makes sense) and that f is analytic on $\mathbb{C}_{\leq\omega}$ and at infinity (so that Dunford–Taylor makes sense) and that it is the Laplace transform of a measure μ, then, for $u \in D(A^2)$, $\gamma > \omega$,

$$\begin{aligned}
f_{HP}(A)\,u &= \int_0^\infty \mathrm{e}^{tA} u\, d\mu(t) = f(\infty)\,u + \int_0^\infty \mathrm{e}^{tA} u\, \chi_{\{t>0\}}\, d\mu(t) \\
&= f(\infty)\,u + \int_0^\infty \left(\frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} \mathrm{e}^{\lambda t}\,(\lambda I - A)^{-1}\, u\, d\lambda\right) \chi_{\{t>0\}}\, d\mu(t) \\
&= f(\infty)\,u + \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} (\lambda I - A)^{-1}\, u \int_0^\infty \mathrm{e}^{\lambda t}\, \chi_{\{t>0\}}\, d\mu(t)\, d\lambda \\
&= f(\infty)\,u + \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} (f(\lambda) - f(\infty))\,(\lambda I - A)^{-1}\, u\, d\lambda \\
&= f(\infty)\,u + (f_{DT}(A) - f(\infty))\,u = f_{DT}(A)\,u,
\end{aligned}$$

where in the third equality the inversion formula for the Laplace transform of $\mathcal{C}_0$ semigroups was used, see [66, Corollary 1.7.5]. Since $D(A^2)$ is dense in X, the bounded operators $f_{DT}(A)$, $f_{HP}(A)$ coincide.

Examples

To conclude the section, we describe some examples of operators of the form $f(A)$, for $A \in \mathcal{G}(X, M, \omega)$, that are then required in the text.

(a) **Polynomials.** Polynomials and power of the operator A were defined at the beginning of the section. They define unbounded closed operators $p(A)$ for every polynomial.

(b) **Rational functions.** Since $A \in \mathcal{G}(X, M, \omega)$ generates a semigroup, the Hille–Yosida condition (1.3) guarantees that the spectrum $\sigma(A)$ is contained in the half-plane $\{z \in \mathbb{C} : \mathrm{Re}(z) \le \omega\}$. If r is a rational function with $|r(z)| \le 1$, for $\mathrm{Re}(z) \le 0$, it can be developed into simple fractions like

$$r(z) = r_\infty + \sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} \frac{r_{j,\ell}}{(1 - z w_\ell)^j}, \qquad \mathrm{Re}(w_\ell) > 0, \quad 1 \le l \le k. \tag{1.25}$$

Set

$$\tau_0(r,\omega) = \begin{cases} +\infty, & \text{if} \quad \omega \le 0, \\ \min_{1\le \ell\le k} \mathrm{Re}(1/w_\ell)/\omega, & \text{if} \quad \omega > 0, \end{cases} \tag{1.26}$$

so that, in view of (1.3), for $0 < \tau < \tau\,(r,\omega)$ (notice that for $\omega \le 0$, there is not upper restriction on τ), it makes sense to define the linear, bounded operator in X

$$r(\tau A) = r_\infty I + \sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} r_{\ell j}(I - \tau w_\ell A)^{-j}. \tag{1.27}$$

Notice that although in this case the operator has been directly defined, it can also be fit into the two constructions presented in the section.

On the one hand, for the Hille–Phillips construction, it suffices to note that $r(\infty) = r_\infty$ is the Laplace transform of the Dirac-delta measure $d\mu(t) = r_\infty \,\delta_0(t)$ while the simple fractions $r_{\ell,j}(z) = r_{\ell,j}(z) = r_{\ell,j}(1 - w_\ell z)^{-j}$ satisfy the conditions in Proposition 1.9-(a), so $r(\tau z) \in \widetilde{M_\omega}$ and the Hille–Phillips calculus may be applied. In fact, for $r_0(z) = (z_0 - z)^{-1}$, with $\mathrm{Re} z_0 > \omega$, we have

$$(z_0 - z)^{-1} = \int_0^\infty \mathrm{e}^{tz}\, \mathrm{e}^{-tz_0}\, dt.$$

Hence in this case, by formula (1.3),

$$r_0(A) = \int_0^\infty \mathrm{e}^{tA}\, \mathrm{e}^{-tz_0}\, dt = (z_0 I - A)^{-1},$$

and the homomorphy of Proposition 1.9 guarantees that this definition coincides with that of (1.27).

On the other hand, r is clearly analytic at infinity and on a open set containing $\sigma(\tau A)$, for $0 < \tau < \tau_0$, so $r(\tau A)$ can be defined via Dunford–Taylor. We have already proved that this definition gives the same operator than the Hille–Phillips one and therefore that of (1.27). In addition, one can extend this definition to the more general case where A is sectorial and r is defined in an neighbourhood of its spectrum.

In this case of rational functions, both constructions are commonly used in different contexts to derive bounds related to the convergence and stability of rational methods, as we will see in the next section.

(c) **Fractional powers of the operator**. The fractional powers of an infinitesimal generator are of special interest when considering evolution problems with nonlinearities. They are introduced, for instance, in [66, 78]. Their properties will be crucial to analyze the methods for semilinear problems that we introduce in Chapter 3. In addition, these operators are also interesting in themselves as they model some natural phenomena under recent investigation. To cite an example, there are anomalous diffusion phenomena that fit with an evolution governed by fractional powers of the Laplacian [57, 56, 72].

First of all, we define the fractional powers $(-A)^\alpha$, $\alpha > 0$, for $A \in \mathcal{G}(X, M, \omega)$. We assume first that $\omega < 0$, so that $0 \notin \rho(A)$.

Then, we can define the operator by an integral path

$$(-A)^{-\alpha} = -\frac{1}{2\pi i} \int_\Gamma \lambda^{-\alpha} \left(\lambda I + A\right)^{-1} d\lambda, \tag{1.28}$$

where the path Γ runs in the resolvent set of $-A$ from $\infty e^{-i\vartheta}$ to $\infty e^{i\vartheta}$, where $0 < \vartheta < \pi/2$ if e^{tA} is a $\mathcal{C}_0$ semigroup and $0 < \vartheta < \pi/2 + \theta$ if e^{tA} is an analytic semigroup defined in a sector S_θ. Alternatively, it can be defined via a Laplace transform by

$$(-A)^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} \mathrm{e}^{tA}\, dt. \tag{1.29}$$

Both definitions are equivalent and it can be checked that they coincide with the (integer) powers $(-A)^{-n}$ for $n = 1, 2, \ldots$. These operators turn out to be one-to-one, so we can define the positive powers by inverting the last negatives. For $D((-A)^{\alpha}) = R((-A)^{-\alpha})$ the range of $A^{-\alpha}$, we define

$$(-A)^{\alpha} = ((-A)^{-\alpha})^{-1}.$$

The fractional powers satisfy the following properties:

1. $(-A)^{\alpha}$ is a bounded operator for $\alpha < 0$ and a closed operator with domain $D(A^{\alpha})$ for $\alpha > 0$.
2. $0 < \beta \leq \alpha$ implies $D((-A)^{\alpha}) \subset D(A^{\beta})$
3. If α, β are real then $(-A)^{\alpha+\beta}u = (-A)^{\alpha}\,(-A)^{\beta}u$ for $u \in D((-A)^{\gamma}$ where $\gamma = \max\{\alpha, \beta, \alpha+\beta\}$.
4. For $\alpha > 0$, the operator $(-A)^{\alpha}$ is the generator of an analytic semigroup.
5. For $t > 0$, it occurs that $\|(-A)^{\alpha} t^{\alpha} \mathrm{e}^{tA}\| \leq M\mathrm{e}^{\omega t}$.

If $A \in \mathcal{G}(X, M, \omega)$, with $\omega \geq 0$, one can set $\omega^* > \omega$ and define the operators $(\omega^* I - A)^{\alpha}$ in an analogous way.

(d) **The exponential function.** As already anticipated in Section 1.1., the semigroup of operators $\left\{e^{tA}\right\}_{t\geq 0}$ is a generalization of the exponential function or the exponential of a matrix to the case of unbounded closed operators. This is consistent with the Dunford–Taylor and Hille–Phillips constructions. For the first case, if $A \in \mathcal{G}(X, M, \omega)$ generates a $\mathcal{C}_0$ semigroup, one has the integral representation, for $u \in D(A^2)$,

$$\mathrm{e}^{tA}u = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} \mathrm{e}^{t\lambda}(\lambda I - A)^{-1}u \, d\lambda, \tag{1.30}$$

with a real $\gamma > \omega$. If, in addition, it is true that $A \in \mathcal{G}(X, M, \omega, \theta)$ generates an analytic semigruoup, then we have the stronger result

$$\mathrm{e}^{tA} = \frac{1}{2\pi i}\int_{\Gamma} \mathrm{e}^{t\lambda}(\lambda I - A)^{-1} \, d\lambda, \tag{1.31}$$

where Γ is a smooth curve in $\rho(A)$ running from $\omega + \infty e^{-i\vartheta}$ to $\omega + \infty e^{i\vartheta}$ for $\theta < \vartheta < \pi/2$, and the integral converges in the strong operator topology. Notice that in this case the exponential function is analytic in a neighbourhood of $\sigma(A)$ and at infinity, so it fits the conditions of the Dunford–Taylor calculus.

For the Hille–Phillips approach, notice that the exponential function $f(z) = \mathrm{e}^{tz}$ is the Laplace transform a point mass at t, so $f(A)$ is the semigroup again.

1.3 Runge–Kutta methods and order reduction

1.3.1 Basic properties of Runge–Kutta methods

Although we have general results for the existence and uniqueness of some PDEs, it is not possible in general to have an explicit formula for the solution that may be used in practical applications. Then, the need arises to develop numerical methods to efficiently approximate the solution of differential equations in order to calculate them and use them for practical purposes.

Among all the methods for integrating a time evolution problem, Runge–Kutta methods have a prominent place due to their efficiency, versatility in implementation and the possibility of being used together with different types of spatial discretizations of the PDEs. The development and study of these methods is the result of a large collective effort, carried out with particular intensity during the 1960s and 1970s. The cornerstone references for this topic are [37, 38], which offer a comprehensive and systematic synthesis of decades of research on numerical methods for differential equations. Since the aim of this thesis is to design a new family of rational methods, intimately related to the Runge–Kutta ones, its main properties are summarized in the following.

We denote by

$$\begin{array}{c|c} \mathbf{c} & W \\ \hline & \mathbf{b}^T \end{array}, \qquad \mathbf{b}, \mathbf{c} \in \mathbb{R}^s, W \in \mathbb{R}^{s\times s},$$

the Butcher tableau of a given Runge–Kutta method of order $p \geq 1$. The stability function of the method

$$r(z) = 1 + z\mathbf{b}^T \left(I - zW\right)^{-1} \mathbf{e}, \quad \mathbf{e} = [1, \ldots, 1]^T \in \mathbb{R}^s, \tag{1.32}$$

plays a fundamental role in our research, since it will be the starting point. It arises when considering the numerical solution of the test problem

$$\begin{cases} u'(t) = \lambda u(t) & t > 0, \\ u(0) = u_0. \end{cases} \tag{1.33}$$

The numerical approximation obtained after n steps of the Runge–Kutta method with time-step length $\tau > 0$ to the latter can be written as

$$u_{n+1} = r(\tau\lambda)\, u_n, \text{ for } n \geq 0, \quad \text{that is,} \quad u_n = r^n\left(\tau\lambda\right) u_0, \text{ for } n \geq 1.$$

The function contains information about the convergence and stability properties of the method. If the method has order of convergence p, then the stability function approximates the exponential e^z with the same order p, that is,

$$r(z) - \mathrm{e}^z = O\left(z^{p+1}\right) \quad \text{as}\, z \to 0. \tag{1.34}$$

Notice that $\mathrm{e}^{\lambda t}$ is precisely the exact solution of the test problem (1.33), that must be approximated with order p by the method. However, the qualitative behaviour of the exact solution for $\operatorname{Re}\lambda < 0$,

$$\lim_{t\to 0} \mathrm{e}^{\lambda t}\, u_0 = 0, \tag{1.35}$$

related to the stability of the numerical approximations, is not reproduced by the numerical solution unless the condition $|r(z)| < 1$ is satisfied. That is the fact that motivates the following definition.

Definition 1.10. A Runge–Kutta method and, by extension, its corresponding rational function is called

1. *A-stable*, if $|r(z)| \leq 1$ for $\operatorname{Re} z \leq 0$,

2. *$A(\vartheta)$-stable*, if $|r(z)| \leq 1$ for every $z \in \{\lambda \in \mathbb{C} : |\arg(-\lambda)| < \vartheta\}$.

3. *Strongly A-stable (resp. strongly $A(\vartheta)$-stable)* if r is A-stable (resp A(ϑ)-stable) and $r_\infty = r(\infty)$ is such that $|r_\infty| < 1$.

Runge–Kutta methods are not only used to solve scalar equations, but they can also be used to solve systems of differential equations or evolutionary equations of the form (1.4). Let us assume that one of the two following cases is true:

(a) $A \in \mathcal{G}(X, M, \omega)$ generates a $\mathcal{C}_0$ semigroup of linear operators and r is an A-stable rational function.

(b) $A \in \mathcal{G}(X, M, \omega, \theta)$ generates an analytic semigroup of linear operators and r is a strongly $A(\vartheta)$-stable rational function, with $\theta < \vartheta$.

In particular, A could be the matrix of a linear system of differential equations. In both cases above, according to Section 1.2, there exist an upper time-step length τ_0 such that $r(\tau A)$ is a well defined bounded, linear operator for every $0 < \tau < \tau_0$. Moreover, due to the stability property of r, it is true that $|r(z)| \leq 1$ for every $z \in \sigma(\tau A)$, for $0 < \tau < \tau_0$. Then, the corresponding recurrence to approximate (1.4) with time-step length $0 < \tau < \tau_0$ is

$$u_{n+1} = r(\tau A) u_n, \text{ for } n \geq 0, \quad \text{that is,} \quad u_n = r^n(\tau A)\, u_0, \text{ for } n \geq 1. \tag{1.36}$$

It is not obvious if a Runge–Kutta method of order p will show again order of convergence p when applied to the abstract problem (1.4), that is, if the error $\| \left(e^{tA} - r^n(\tau A)\right) u_0 \|$ is $O(\tau^p)$ when τ tends to 0. We are also interested in up to what extent the stability property $|r^n(\tau z)| \leq 1$ is conserved when substituting z by A. The convergence question was answered in 1979 by Hersch and Kato [40], who also proved a weak bound for stability, $\||r^n(\tau A)\| = O(n)$, and conjectured that $\|r^n(\tau A)\| = O(\sqrt{n})$. In the same year, Brenner and Thomée [18] proved the stability conjecture and refined the convergence proof using the Hille–Phillips calculus. Further improvements can be found in [49, 61]. We summarize these results since they are essential to our work.

Theorem 1.11. Let $A \in \mathcal{G}(X, M, \omega)$ and r be an A-stable rational function that approximates the exponential with order p. There exist three constants,

$$C_e = C_e(r) > 0, \quad C_s(n) = C_s(r, n) > 0, \quad \kappa = \kappa(r) \geq 1,$$

such that the following hold for $0 < \tau < \tau_0(r, \omega)$, and $\omega^+ = \max\{0, \omega\}$:

(a) *The stability bound,*

$$\|r^n(\tau A)\| \leq M\, C_s(n)\, \mathrm{e}^{\omega^+ \kappa t_n}, \quad t_n = n\tau, \quad C_s(n) = O(\sqrt{n}). \tag{1.37}$$

(b) *The optimal convergence estimate*

$$\|r^n(\tau A) u_0 - \mathrm{e}^{t_n A} u_0\| \leq C_e\, M\, t_n\, \tau^{p'}\, \mathrm{e}^{\omega^+ \kappa t_n}\, \|A^{p'+1} u_0\|, \quad n \geq 1, \tag{1.38}$$

valid for $u_0 \in D(A^{p'+1})$ with $1 \leq p' \leq p$.

If we also assume that A generates an analytic semigroup and that r is strongly A-stable, then the following hold

(c) *The optimal parabolic error estimate*

$$\|r^n(\tau A)u_0 - \mathrm{e}^{t_n A}u_0\| \leq C_e\, M\, \tau^{p'}\, \mathrm{e}^{\omega^+ \kappa t_n}\, \|A^{p'}u_0\|, \quad n \geq 1, \tag{1.39}$$

valid for $u_0 \in D(A^{p'})$ with $1 \leq p' \leq p$.

(d) *The bad initial data error estimate*

$$\|r^n(\tau A)u_0 - \mathrm{e}^{t_n A}u_0\| \leq C_e\, M\, n^{-p}\, \mathrm{e}^{\omega^+ \kappa t_n}\, \|u_0\|, \quad n \geq 1, \tag{1.40}$$

valid for every $u_0 \in X$.

Remark 1.12. Some clarifications and comments on the previous theorem.

(a) The weak stability (1.37) is optimal in general; it is sharp for $A = d/dx$ in the maximum-norm [17], and it can be improved depending on the behaviour of $r(z)$ (Theorem 2 of [18]) and on the nature of the operator A. For instance, the term $C_s(n)$ becomes $O(1)$ in the following cases:

(a1) for $r(z) = 1/(1-z)$, which corresponds to the implicit Euler method,

(a2) for $r_{m,n}(z) = P_m/Q_n$ the Padé approximant of e^z with degree of $P_m = m$ and degree of $Q_n = n$, whenever $n = m - 1$. This is the case of Radau methods.

(a3) when X is a Hilbert space and A an ω-dissipative operator,

(a4) when $r(z)$ is $A(\vartheta)$-stable and $A \in \mathcal{G}(X, M, \omega, \theta)$ generates an analytic semigroup with $\theta < \vartheta$.

(b) The convergence estimate (1.38) shows that with sufficient regularity the Runge–Kutta method maintains the order of convergence p when integrating homogeneous problems in Banach spaces. Observe that $C_s(n)$ does not appear in (1.38), so that the convergence is optimal even in cases of weak stability. However, stability affects the treatment of the nonhomogeneous problems as well as the analysis of full discretizations.

(c) If $r(z)$ is $A(\vartheta)$-stable and $A \in \mathcal{G}(X, M, \omega, \theta)$ generates an analytic semigroup with $\theta < \vartheta$, the stability and the mentioned error estimates (1.37-1.40) remain valid.

Remark 1.13. In some cases in this work, we deal with semigroups whose growth bound has the form (1.2), but includes an additional linear factor:

$$\|\mathrm{e}^{tA}\| \leq M(1+\alpha t)\mathrm{e}^{\omega t}.$$

In this case, the formulas (1.37)–(1.40) remain valid upon adding a linear factor of the form $(1+\alpha t_n)$ to the bound. This holds because, for such a semigroup, it is true that

$$\|\mathrm{e}^{tA}\| \leq M_{\tilde{\omega}}\mathrm{e}^{\tilde{\omega} t}, \quad \text{where} \quad M_{\tilde{\omega}} = \frac{\alpha}{\tilde{\omega} - \omega}\mathrm{e}^{(\tilde{\omega}-\omega)/\alpha - 1}\, M,$$

for every $\tilde{\omega} > \omega$. In particular, we have $A \in \mathcal{G}(X, M_{\tilde{\omega}}, \tilde{\omega})$, so we can apply (1.37)–(1.40) with these constants. As a result, we obtain a term of the form $M_{\tilde{\omega}} \mathrm{e}^{\tilde{\omega} t_n}$ on the right-hand side, which we can minimize over $\tilde{\omega} > \omega$ to obtain

$$M_{\tilde{\omega}_{\min}} \mathrm{e}^{\tilde{\omega}_{\min} t_n} = (1 + \alpha t_n) \mathrm{e}^{\omega t_n},$$

as desired.

When A generates an analytic semigroup, there is a discrete analogue of the parabolic smoothing formula (1.9). The following results (Theorem 1.1. and Lemma 2.2. in [36]) will be useful to deal with nonlinear problems in Chapter 4.

Theorem 1.14. Let $A \in \mathcal{G}(X, M, \omega, \theta)$ and $r(z)$ be a strongly $A(\vartheta)$-stable rational function with $\theta < \vartheta$. Then, there exist $K > 0$ and $\bar{\tau} > 0$ depending on M, θ and the rational function such that for all $n \geq 1$ integer and $0 < \tau < \bar{\tau}$, $t_n = n\tau$, the following estimate hold:

$$\| A \left(r^n(\tau A) - r_\infty^n I \right) \| \leq \frac{K \mathrm{e}^{\bar{\omega}\, t_n}}{t_n}, \tag{1.41}$$

where $\bar{\omega} = 3\omega/2$. Moreover, for $\alpha \in (0, 1)$, it is true that

$$\| \left(r^n \left(\tau A \right) - r_\infty^n I \right) x \|_\alpha \leq \frac{K \mathrm{e}^{\bar{\omega}\, t_n}}{t_n^\alpha} \|x\|, \quad x \in X. \tag{1.42}$$

1.3.2 Efficient implementation of the Runge–Kutta methods in the linear case

We briefly describe in this section the way in which we implement the Runge–Kutta methods to solve linear problems of the form

$$\begin{cases} u'(t) = Au(t) + f(t) \quad t > 0, \\ u(0) = u_0, \end{cases} \tag{1.43}$$

The corresponding adaptations of the proposed algorithms to deal with initial boundary value problems are not described exhaustively, since the key ideas are covered in the linear case.

We consider a Runge–Kutta method of s stages and order p with Butcher tableau defined by the vectors $\mathbf{b}, \mathbf{c} \in \mathbb{R}^s$ and a matrix $W \in \mathbb{R}^{s \times s}$, which we assume to be either diagonalizable or lower triangular. The stability function of the method is then given by formula (1.27). Our goal is to show how a step of the method can be done by solving s linear systems (resolvents of A) and avoiding products of the form Au, that are numerically unstable.

Let $\tau > 0$ be the time step size and N the total number of steps, we denote $t_n = \tau n$, $0 \leq n \leq N$. The Kronecker product (of a matrix and an operator) is denoted by $\otimes$. We describe the method in the Banach space X, but notice that in practice we work with an spatial discretization of X, A, f.

Once we have computed the numerical approximation u_n for some integer $0 \leq n \leq N-1$, the internal stages $U_n = \{U_n^i\}_{i=1}^s \in X^s$ are defined by

$$U_n = (\mathbf{e} \otimes I)\, u_n + \tau\, (W \otimes A)\, U_n + \tau\, (W \otimes I)\, F_n, \tag{1.44}$$

where $\mathbf{e} = [1, \dots, 1]^T \in X^s$ and $F_n = \{f(t_n + \tau c_i)\}_{i=1}^s$. The stages are then the solution of the system of s equations in $D(A)$

$$(I - \tau\, (W \otimes A))\, U_n = (\mathbf{e} \otimes I)\, u_n + \tau\, (W \otimes I)\, F_n. \tag{1.45}$$

Let $\boldsymbol{\xi} \in \mathbb{R}^s$ be such that $\mathbf{b}^T = \boldsymbol{\xi}^T W$, which is always possible under our assumptions on W. Notice that

$$r_\infty = \lim_{z\to\infty} \left(1 + z\mathbf{b}^T\, (1 - zW)^{-1}\, \mathbf{e}\right) = 1 - \mathbf{b}^T W^{-1}\mathbf{e} = 1 - \boldsymbol{\xi}^T\mathbf{e}. \tag{1.46}$$

Then, once we have computed the stages $U_n \in X^s$, a step of the method is done by

$$\begin{aligned}
u_{n+1} &= u_n + \tau\left(\mathbf{b}^T \otimes A\right) U_n + \tau\left(\mathbf{b}^T \otimes I\right) F_n \\
&= u_n + \left(\mathbf{b}^T \otimes A\right) (I - \tau\, (W \otimes A))^{-1}\, ((\mathbf{e} \otimes I)\, u_n + \tau\, (W \otimes I)\, F_n) + \tau\left(\mathbf{b}^T \otimes I\right) F_n \\
&= u_n - \boldsymbol{\xi}^T\, (I - \tau\, (W \otimes A) - I)\, (I - \tau\, (W \otimes A))^{-1}\, ((\mathbf{e} \otimes I)\, u_n + \tau\, (W \otimes I)\, F_n) \\
&\qquad + \tau\boldsymbol{\xi}^T\, (W \otimes I)\, F_n \\
&= u_n + \boldsymbol{\xi}^T U_n + \boldsymbol{\xi}^T\, (\tau\, (W \otimes I)\, F_n - (\mathbf{e} \otimes I)\, u_n + \tau\, (W \otimes I)\, F_n) \\
&= r_\infty u_n + \boldsymbol{\xi}^T U_n,
\end{aligned}$$

which is a linear combination of the previous step and the stages. It turns out that we avoid evaluating the unbounded operator A.

To conclude, we explain how to solve the linear systems (1.45) depending on the structure of the matrix.

(a) $W \in \mathbb{R}^{s\times s}$ is invertible. There exists an invertible matrix $P \in \mathbb{R}^{s\times s}$ such that $W = P\Lambda P^{-1}$, where Λ is a diagonal matrix with eigenvalues $\lambda_1, \dots, \lambda_s \in \mathbb{C}$. In this case the system (1.45) is equivalent to

$$P\ (I - \tau\, (W \otimes A))\ P^{-1} U_n = (\mathbf{e} \otimes I)\, u_n + \tau\, (W \otimes I)\, F_n, \tag{1.47}$$

so we can solve it by doing

(1) $\eta_n = P^{-1}\, ((\mathbf{e} \otimes I)\, u_n + \tau\, (W \otimes I)\, F_n) \in X^s$,

(2) solve $(P^{-1}U_n)^k = (I - \tau\lambda_j A)^{-1}\, \eta_n^k$, for $k = 1, \dots, s$,

(3) $u_n = P\, (P^1 U_n)$.

In this case, the linearity of the problem and the fact that the matrix is diagonalizable allow the computation of the stages to be decoupled.

(b) $W \in \mathbb{R}^{s\times s}$ is lower triangular, that is, $W = \Lambda + T$ where Λ is a diagonal matrix with eigenvalues $\lambda_1, \dots, \lambda_s \in \mathbb{R}$ and T is nilpotent, that is, $T^s = 0$. In this case

$$\begin{aligned}(I - \tau (W \otimes A))^{-1} &= (I - \tau (\Lambda \otimes A) - \tau (T \otimes A))^{-1} \\ &= (I - \tau (D \otimes A))^{-1} \left(I - \tau (I - \tau (D \otimes A))^{-1} (T \otimes A)\right)^{-1} \\ &= \sum_{k=0}^{s-1} \left[(I - \tau (D \otimes A))^{-1} (T \otimes A)\right]^k (I - \tau (D \otimes A))^{-1}, \quad (1.48)\end{aligned}$$

since the terms of this geometric series are zero for $k \geq s$. In fact, T is lower triangular and for $\{V^i\}_{i=1}^s \in X^s$ such that $V^i = 0$, for $i = 1, \dots, l$, then $[(T \otimes A)]^i = 0$ for $i = 1, \dots, l+1$, and so is $\left[(I - \tau (D \otimes A))^{-1} (T \otimes A) V\right]^i = 0$, due to the fact that D is diagonal. Moreover, the identity

$$\begin{aligned}(T \otimes A) (I - \tau (D \otimes A))^{-1} &= D^{-1} T (I - \tau (D \otimes A) - I) (I - \tau (D \otimes A))^{-1} \\ &= D^{-1} T - D^{-1} T (I - \tau (D \otimes A))^{-1}\end{aligned}$$

allows us not to evaluate the unbounded operator A.

By carefully analyzing formula (1.48), taking into account the diagonal structure of D and the lower triangular structure of T, and by leveraging the formula for products of resolvents with T, one can apply a Horner-like algorithm to solve the system in $D(A)^s$ in (1.45) solving only s linear systems and avoiding matrix-vector products with A. The verification is left to the reader, but is done by computing the following steps. We part form $\eta_n = (\mathbf{e} \otimes I) u_n + \tau (W \otimes I) F_n \in X^s$. During the iteration, we update the components of the vector η_n, which will eventually contain the values of U_n. For $k = 1, \dots, s$,

(1) solve $(I - \tau \lambda_k A)^{-1} \eta_n^k \in X$,

(2) for $j = k+1, \dots, s$, store

$$\eta_n^j \leftarrow \eta_n^j - \frac{T_{jk}}{d_k} \left(\eta_n^k - (I - \tau \lambda_k A)^{-1} \eta_n^k\right),$$

(3) and then store

$$\eta_n^j \leftarrow (I - \tau \lambda_k A)^{-1} \eta_n^k.$$

After the iteration, we get $U_n \leftarrow \eta_n$.

1.3.3 Interpolation spaces and order reduction

We end this chapter with a brief explanation of why order reduction occurs and what orders of convergence can be expected in the problems we will discuss. To this end, we start this section by commenting on some results concerning interpolation spaces, which are essential to understand the order reduction phenomenon. The analysis of order reduction in different scenarios can be found in [8, 59, 60, 71], whereas for the interpolation spaces the classic reference is [75] and for further results we suggest [13, 50, 54].

Interpolation spaces and boundary conditions

Let us assume that $A \in \mathcal{G}(X, M, \omega)$ is an infinitesimal generator of a $\mathcal{C}_0$ semigroup, fix $\omega^* > \omega$, and for $\nu \geq 0$, set $X_\nu = D\left((\omega^* I - A)^\nu\right)$. The space X_ν is endowed with the graph norm $\|\cdot\|_\nu$ of $(\omega^* I - A)^\nu$. It is well known that X_ν is independent of $\omega^* > \omega$ and that changing $\omega^* > \omega$ results in an equivalent norm.

On the other hand, the real interpolation method [75] provides different intermediate spaces $X_{\nu,p} = [X_0, X_1]_{\nu,p}$ with norms $\|\cdot\|_{\nu,p}$, $0 \leq \nu < 1$, $1 \leq p \leq \infty$. It is important to notice that (see Theorem 4.17 of [54])

$$X_{\nu+\epsilon,p} \hookrightarrow X_\nu \hookrightarrow X_{\nu-\epsilon,p}, \quad 0 \leq p, q \leq \infty, \quad 0 \leq \nu - \epsilon < \nu + \epsilon \leq 1$$

with continuous embeddings. As a consequence, for $0 < \nu^* < 1$, $u \in X$, it is true that

$$u \notin X_{\nu,p} \text{ for } \nu < \nu^* \leq 1 \Leftrightarrow u \notin X_\nu \text{ for } \nu < \nu^* \leq 1. \tag{1.49}$$

Now we illustrate the previous concepts in the context of typical evolutionary PDEs in an L^p, $p \geq 1$, framework. Let us consider $X = L^p(\Omega)$, $p \geq 1$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain with regular boundary Γ. Moreover, we are given two linear partial differential operators P, Q on Ω of orders m and $k \leq m - 1/p$, respectively, with smooth coefficients. Typically, the operator A acts on

$$D(A) = \left\{\phi \in W^{m,p}(\Omega) \,/\, Q\phi|_\Gamma = 0\right\}, \tag{1.50}$$

and $A\phi = P\phi$ for $\phi \in D(A)$. Set $\nu^* = (k + 1/p)/m$. Notice that the values of a function $u \in L^p$ on Ω or Γ may not be defined, since u is not necessarily a continuous function in Ω. By *Sobolev embedding theorems* (see, e.g., [19, 77]), functions in $W^{m,p}(\Omega)$ exhibit improved regularity depending on the balance between m, p and the space dimension d. In particular, whenever $m > d/p$ one has the continuous embedding

$$W^{m,p}(\Omega) \hookrightarrow C(\overline{\Omega}),$$

so that boundary values are well defined in the classical sense. Moreover, there exists a constant M depending on d, m, p such that

$$\|u\|_\infty \leq M\|u\|_{m,p} \text{ for } u \in W^{m,p}(\Omega). \tag{1.51}$$

However, when this condition is not fulfilled, continuity up to the boundary is no longer guaranteed, and the notion of trace has to be defined in a weaker sense. Moreover, the *trace theorem* (see e.g. [75, 77]) guarantees that the trace of Q, i.e., the operator $\partial : \phi \mapsto Q\phi|_\Gamma$, can be understood as a linear, bounded operator from $W^{\nu m,p}(\Omega)$ to $L^p(\Omega)$ whenever $\nu > \nu^*$. The remarkable result in [50] states that for the p-real interpolation method, there holds that

$$X_{\nu,p} = [X_0, X_1]_{\nu,p} = \left\{ \begin{array}{lll} W^{\nu m,p}(\Omega), & \text{if} & \nu < \nu^*, \\ W^{\nu m,p}(\Omega) \cap \ker\partial, & \text{if} & \nu > \nu^*. \end{array} \right. \tag{1.52}$$

This means that when interpolating, the boundary condition does not need to be imposed when the trace operator does not make sense. Therefore, if the trace of a smooth mapping $\phi : \Omega \to \mathbb{C}$ is different from 0 on Γ, then ϕ cannot belong to $X_{\nu,p}$ for $\nu > \nu^*$. Notice that this is also true for the domains of the fractional powers of X_ν according to (1.49). This fact is what governs order reduction.

Order reduction for RK methods

As we commented in a previous section, when a Runge–Kutta method used in its rational form (1.36) is applied to a homogeneous problem (1.4), it achieves its classic order of convergence p whenever the initial data has sufficient regularity; this is the content of the convergence estimate (1.38). However, for nonhomogeneous problems of the form

$$\begin{cases} u'(t) = Au(t) + f(t) \quad t > 0, \\ u(0) = u_0, \end{cases} \tag{1.53}$$

the order of convergence p is not achieved in general no matter the time regularity of its solution. Actually, let $u \in \mathcal{C}^p\left([0,\infty), X\right)$ be the solution of (1.53). The local error of the Runge–Kutta method applied to this problem is given by the expression [8, 59, 60]

$$\epsilon_n = \sum_{l=q+1}^{p} \frac{\tau^l}{l!} r_l\left(\tau A\right) u^{(l)}(t_n) + O\left(\tau^{p+1}\right), \tag{1.54}$$

where

$$r_l(z) = z\mathbf{b}^T\left(1 - zW\right)^{-1}\left(\mathbf{c}^l - l\mathbf{c}^{l-1}\right), \qquad \text{for}\, q+1 \le l \le p. \tag{1.55}$$

Moreover, the order conditions of the method guarantee that

$$r_l(z) = z^{p-q} r_l^*(z), \qquad \text{for}\, q+1 \le l \le p, \tag{1.56}$$

for certain rational mappings $r_l^*(z)$, $q+1 \le l \le p$. Notice that since $r_l(z)$ and $r_l^*(z)$, $q+1 \le l \le p$, possess no poles on the half-plane $\mathrm{Re}(z) \le 0$, we can argue as in (1.27) and see that the operators $r_l(\tau A)$ and $r_l^*(\tau A)$, $q+1 \le l \le p$, are bounded for $0 < \tau < \tau_0\left(r, \omega\right)$.

For $x \in X_{p-q}$ we also have (Theorem 1.9.3) that

$$\|r_l(\tau A)x\| \le C_e\, M\, \tau^{p-q}\|u\|_{p-q}, \quad q+1 \le l \le p,$$

and, by interpolation, we deduce that for $x \in X_\nu$, $0 \le \nu \le p-q$,

$$\|r_l(\tau A)x\| \le C_e\, M\, \tau^{\nu}\|x\|_{\nu}, \quad q+1 \le l \le p.$$

Thus, in view of (1.54), for $u \in \mathcal{C}^{p+1}\left([0,\infty), X_\nu\right)$, we get

$$\|\epsilon_n\| = O(\tau^{q+1+\nu}) \sup_{0 \le t \le n\tau} \|u^{(p+1)}(t)\|_\nu.$$

Therefore, only under the stronger assumption $u \in \mathcal{C}^{p+1}\left([0,\infty), X_{p-q}\right)$ we reach the optimal local order $p+1$.

The origin of the order reduction phenomenon relies in the fact that, as we mentioned in (1.52), all we can expect in the context of standard PDEs is that $u \in \mathcal{C}^{p+1}\left([0,\infty), X_{\nu^*}\right)$, for some well defined value $0 < \nu^* < 1$. Setting $\nu = \nu^*$ (or $\nu = \min(\nu^* + 1, p-q)$ when $r(\infty) \ne 1$), we easily get the error estimate

$$\|u(t_n) - u_n\| = C_s(r,n)\tau^{q+\nu} \sup_{0 \le t \le n\tau} \|u^{p+1}(t)\|_\nu, \qquad n \ge 1,$$

where $C_s(r, n)$ stands for the stability bound of the recurrence.

These ideas can be extended to the situation of variable step sizes. Besides, for constant step sizes and in case $r(\infty) \neq 1$, the clever summation-by-parts argument in [59, 60] (extended in [8] to general semigroups) leads to the improvement

$$\|u(t_n) - u_n\| = C_s(r, n)\tau^{q+\mu+1} \sup_{0 \leq t \leq n\tau} \|u^{p+1}(t)\|_\nu, \qquad n \geq 1,$$

where $\mu = \min(\nu, p-q-1)$. It is worth noticing that such a fractional order of convergence is the one occurring in practical computation.

Chapter 2

Rational methods for abstract, linear, nonhomogeneous problems without order reduction

One had to journey through the icy wasteland of abstraction, in order to definitively arrive at concrete philosophizing.

Theodor Adorno

2.1 Introduction

In this chapter we are concerned with the numerical time integration of abstract, linear, nonhomogeneous initial value problems (IVP) of the form

$$\begin{cases} u'(t) = Au(t) + f(t), \quad t > 0, \\ u(0) = u_0, \end{cases} \tag{2.1}$$

where $A \in \mathcal{G}(X, M, \omega)$ is the infinitesimal generator of a semigroup of operators, $u_0 \in X$ and $f : (0, \infty) \to X$ is a source term. We begin with this class of problems, as they represent the simplest setting in which order reduction occurs. Starting from a rational A-stable approximation $r(z)$ to the exponential of order p, we construct a family of stable methods of the same order that, in fact, avoid order reduction. These methods constitute the foundation for the approaches developed in subsequent chapters to address more complex problems.

We recall some basic results related to the problem (2.1). Any classical solution of (2.1) can be written using the variation-of-constants formula

$$u(t) = \mathrm{e}^{tA} u_0 + \int_0^t \mathrm{e}^{(t-s)A} f(s)\, ds. \tag{2.2}$$

Different assumptions on the regularity of the source term f lead to different possible assumptions on the regularity of the initial data. It is known [31, 66] that for

$f \in L^1((0,\infty,X)$, the problem (2.1) has a unique *mild solution* for every initial data $u_0 \in X$, that is, a function that satisfies (2.2). If in addition further one of the following hypothesis:

(a) $f \in L^1((0,\infty,X)$ is continuously differentiable on $(0,\infty)$,

(b) $f \in L^1((0,\infty,X)$ is continuous on $(0,\infty)$, $f(t) \in D(A)$ for $t > 0$ and $Af \in L^1((0,\infty,X)$,

is satisfied, then the IVP has a unique classical solution for every $u_0 \in D(A)$. Under the further assumption that $A \in \mathcal{G}(X, M, \omega, \theta)$ generates an analytic semigroup and that $f \in L^1((0,\infty,X)$ is locally Hölder continuous, then the existence and uniqueness of a classical solution requires only $u_0 \in X$. Finally, we point out that optimal regularity results for the solution u in terms of that of f is studied, for instance, in [53, 77].

Our goal is then to numerically approximate the solutions of (2.1). As previously anticipated in the introduction, our strategy to design a method without order reduction is to cast a nonhomogeneous IVP into an enlarged homogeneous problem. Essentially, this is achieved by treating f as a new unknown, in the line of the approach used in [33] for equations with memory. Then, the fact that a rational method applied to a homogeneous problem achieves the classic order of the method p (1.38) allows us to integrate the enlarged system without order reduction if the initial condition is regular enough. Finally, we use some evaluations of f to approximate this theoretical discretization within the adequate order of convergence. We want to emphasize that this abstract construction is useful to design these methods but, as we will show, the resulting schemes have computational requirements similar to those of Runge–Kutta ones. As the epigraph reflects, this abstract framework is essential to a complete understanding of this family of methods, which prove to be both simple and efficient. Once a rational function $r(z)$ is fixed, we still have freedom in choosing the set of nodes $\mathbf{c}_n$ at which the function f will be evaluated at each step to completely determine the scheme we will use. We will discuss the benefits and disadvantages of each possible choice. These ideas are developed throughout Section 2.2.1, whereas the final analysis of the convergence of the methods under different hypotheses is carried out in Section 2.2.2. There, we prove Theorem 2.6 that is the main result of the chapter. In Section 2.2.3, we explain in detail how an efficient variable-step version of the proposed methods would be implemented.

Finally, in Section 2.3, we present numerical illustrations demonstrating the performance of the proposed methods for the time integration of various prototypical PDEs.

2.2 Derivation and analysis of the method

2.2.1 Motivation of the scheme and previous results

We introduce some notation we require on functional Banach spaces. Given a complex Banach space X and $m \geq 0$, let $\mathcal{C}^m_{ub}([0,\infty), X)$ denote the space formed by all the mappings $h : [0,\infty) \to X$ such that $h^{(j)}$, $0 \leq j \leq m$, are bounded and uniformly continuous on $[0,\infty)$. Set, for $m \geq 0$ and $0 \leq t \leq \infty$,

$$\|h\|_{m,t} = \max_{0\leq j\leq m} \sup_{0\leq s\leq t} \|h^{(j)}(s)\|. \tag{2.3}$$

The space $\mathcal{C}_{ub}^m([0,\infty),X)$ endowed with the norm $\|\cdot\|_{m,\infty}$, is a Banach space. We now study the *semigroup of translations* on $\mathcal{C}_{ub}([0,\infty),X)$, whose properties are summarized in the following lemma.

Proposition 2.1. Let X be a Banach space and $\mathcal{C}_{ub}\left([0,\infty),X\right)$ endowed with the supremum norm. Then, the semigroup of translations $T_B(t):\mathcal{C}_{ub}\left([0,\infty),X\right)\to\mathcal{C}_{ub}\left([0,\infty),X\right)$, $t\geq 0$, defined by

$$[T_B(t)\,v]\,(s)=v(t+s),\quad v\in\mathcal{C}_{ub}\left([0,\infty),X\right),s\geq 0,$$

is a $\mathcal{C}_0$ semigroup. The operator $B:D(B)\subset\mathcal{C}_{ub}\left([0,\infty),X\right)\to\mathcal{C}_{ub}\left([0,\infty),X\right)$ defined in

$$D(B)=\{v\in\mathcal{C}_{ub}\left([0,\infty),X\right):v'\in\mathcal{C}_{ub}\left([0,\infty),X\right)\}=\mathcal{C}_{ub}^1([0,\infty),X)\tag{2.4}$$

by $Bv=v'$, is the infinitesimal generator of the semigroup and we can write $T_B(t)=\mathrm{e}^{Bt}$, for $t\geq 0$. In fact, it is true that $D(B^m)=\mathcal{C}_{ub}^m([0,\infty),X)$ for $m=1,2,\ldots$. Moreover, the bound

$$\|\mathrm{e}^{tB}v\|_\infty\leq\|v\|_\infty$$

holds for every $v\in\mathcal{C}_{ub}\left([0,\infty),X\right)$.

Proof. It is clear from the definition that $\{T_B(t)\}_{t\geq 0}$ satisfies the properties (a), (b) in Definition 1.1. To prove the strong continuity, notice that since every $v\in\mathcal{C}_{ub}\left([0,\infty),X\right)$ is uniformly continuous, the difference

$$\|\,[T_B(t)\,v]\,(s)-v(s)\|_\infty=\|v(t+s)-v(s)\|_\infty\tag{2.5}$$

tends to 0 when $t\to 0$ uniformly on $s\in[0,\infty)$, due to the uniform continuity of the elements of Y, so $\{T_B(t)\}_{t\geq 0}$ is a $\mathcal{C}_0$ semigroup. If $v\in D(B)$, by formula (1.1), the following limit exists

$$[B\,v](s)=\lim_{t\to 0}\frac{v(t+s)-v(s)}{t}=v'(s),$$

so $Bv=v'\in\mathcal{C}_{ub}\left([0,\infty),X\right)$. On the other hand, if $v,v'\in\mathcal{C}_{ub}\left([0,\infty),X\right)$, then the fundamental theorem of calculus implies that

$$\left\|\frac{v(t+s)-v(s)}{t}-v'(s)\right\|_\infty\leq\max_{s\leq u\leq t+s}\|v'(u)-v'(s)\|,$$

and the fact that v' is uniformly continuous implies that $v\in D(B)$ and $Bv=v'$. Finally, notice that

$$\|\mathrm{e}^{tB}\,v\|_\infty=\|v(t+\cdot)\|_\infty\leq\|v\|_\infty.$$

□

We also define the operator $L:\mathcal{C}_{ub}\left([0,\infty),X\right)\to X$ defined by $L\,v=v(0)$, for $v\in\mathcal{C}_{ub}\left([0,\infty),X\right)$, the Banach product space $Z=X\times\mathcal{C}_{ub}\left([0,\infty),X\right)$ with the norm $\|(u,v)\|_Z=\|u\|+\|v\|_\infty$ and the operator

$$G\begin{pmatrix}u\\v\end{pmatrix}=\begin{pmatrix}A&L\\0&B\end{pmatrix}\begin{pmatrix}u\\v\end{pmatrix},\quad\text{for }(u,v)^T\in D\left(G\right):=D\left(A\right)\times D\left(B\right).\tag{2.6}$$

The following proposition states some properties of the operator G, which turns out to be the infinitesimal generator of a semigroup in the space Z. Remind that we were trying to cast the IVP (2.1) into an enlarged homogeneous problem with the same solution. The following result establishes that the homogeneous problem reated to G is in fact the enlarged system we were looking for. We also provide a way to compute the resolvents of G in terms of the resolvents of A and B.

Proposition 2.2. Let $A \in \mathcal{G}(X, M, \omega)$ and G be the operator defined in (2.6). Then,

(a) G is the infinitesimal generator of a $\mathcal{C}_0$ semigroup of bounded, linear operators $\left\{ \mathrm{e}^{tG} \right\}_{t \geq 0}$ in the space Z. More precisely, the semigroup is

$$\mathrm{e}^{tG} = \begin{pmatrix} \mathrm{e}^{tA} & \int_0^t \mathrm{e}^{(t-s)A} L \cdot \, ds \\ 0 & \mathrm{e}^{tB} \end{pmatrix} \tag{2.7}$$

and it has growth

$$\| \mathrm{e}^{tG} \| \leq M(1+t)\, \mathrm{e}^{\omega^+ t}. \tag{2.8}$$

In particular, $G \in \mathcal{G}(Z, M_{\tilde{\omega}}, \tilde{\omega})$ for every $\tilde{\omega} > \omega^+$ and an adequate $M_{\tilde{\omega}} > M$.

(b) Let $u_0 \in D(A)$, $f \in D(B)$. Then, the solution of the IVP

$$\left\{ \begin{array}{lcl} U'(t) & = & G\, U(t), \quad t \geq 0, \\ U(t_0) & = & (u_0, f)^T, \end{array} \right. \tag{2.9}$$

is given by $U(t) = (u(t), v(t))^T = \mathrm{e}^{tG} (u_0, f)^T$, $t \geq 0$, so even the generalised solutions of (2.1) are provided by the first component of the solutions of (2.9).

(c) The k-th powers of the resolvent of the operator G can be expressed by computing k separate resolvents of the operators A and B. In fact, for every $(u_0, v_0)^T \in X \times \mathcal{C}_{ub}([0, \infty), X)$, we can obtain $U_k = (u_k, v_k)^T = (\lambda I - G)^{-k} (u_0, v_0)^T$ via the following recurrence

$$v_j = (\lambda I - B)^{-1} v_{j-1}, \qquad u_j = (\lambda I - A)^{-1} (u_{j-1} + L v_j), \quad \text{for } j = 1, \ldots, k. \tag{2.10}$$

Proof. The system (2.9) is

$$u'(t) = A\, u(t) + L\, v(t), \qquad v'(t) = B\, v(t).$$

Taking into account Proposition 2.1 and the definition of L,

$$v(t) = f(t + \cdot), \qquad Lv(t) = f(t),$$

so we obtain

$$\left\{ \begin{array}{lcl} u'(t) & = & A\, u(t) + f(t), \quad t \geq 0, \\ u(0) & = & u_0, \end{array} \right. \tag{2.11}$$

By continuity, we conclude that even the generalised solutions of (2.11) are provided by the first component of the solutions of (2.9). The variation-of-constants formula (2.2) implies that

$$\|u(t)\| \leq M\,\mathrm{e}^{\omega t}\|u_0\| + tM\,\mathrm{e}^{\omega t}\|f\|_\infty \leq M\,(1+t)\;\mathrm{e}^{\omega^+ t}\|(u_0,f)^T\|_Z,$$

and together with

$$\|v(t)\|_\infty \leq \|f\|_\infty \leq \|(u_0,f)^T\|_Z,$$

leads to

$$\|U(t)\|_Z \leq M\,(1+t)\,\mathrm{e}^{\omega^+ t}\,\|(u_0,f)^T\|_Z.$$

To prove (c), notice that the identity $U_k = (\lambda I - G)^{-1}U_{k-1}$ leads to the equation

$$\begin{pmatrix} \lambda I - A & -L \\ 0 & \lambda I - B \end{pmatrix} \begin{pmatrix} u_k \\ v_k \end{pmatrix} = \begin{pmatrix} u_{k-1} \\ v_{k-1} \end{pmatrix},$$

that can be immediately expressed as

$$v_k = (\lambda I - B)^{-1} v_{k-1} = (\lambda I - B)^{-k} v_0,$$

and

$$u_k = (\lambda I - A)^{-1}\left(u_{k-1} + Lv_k\right) = (\lambda I - A)^{-1}\left(u_{k-1} + L(\lambda I - B)^{-k}v_0\right). \tag{2.12}$$

□

At this point, it seems natural to apply a rational method to (2.9) and retain the first components to approximate the solutions of (2.1). In this way, it is clear by (1.38) that no order reduction occurs if $(u_0,f)^T \in D(G^{p+1})$. This leads to the recurrence

$$\bar{U}_{n+1} = r\,(\tau G)\begin{pmatrix} \bar{u}_n \\ \bar{v}_n \end{pmatrix} = r_\infty \begin{pmatrix} \bar{u}_n \\ \bar{v}_n \end{pmatrix} + \sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} r_{\ell j}\,(I - \tau w_\ell G)^{-j}\begin{pmatrix} \bar{u}_n \\ \bar{v}_n \end{pmatrix}. \tag{2.13}$$

Thus, assuming that $u, f \in \mathcal{C}^{p+1}\left([0,\infty), X\right)$, it is clear that $U \in \mathcal{C}^{p+1}\left([0,\infty), Z\right)$, and since (2.9) is a homogeneous problem, $U \in \mathcal{C}\left([0,\infty), D\left(G^{p+1}\right)\right)$. Notice that $D\left(G^{p+1}\right)$ may be different from $D\left(A^{p+1}\right) \times D\left(B^{p+1}\right)$, so u may not lie in $D\left(A^{p+1}\right)$. Under this assumption, the convergence result (1.38) together with Remark 1.13 applied to G and initial data u_0 and $v_0 = f$ guarantees that

$$\|u(t_n) - \bar{u}_n\| \leq C_e M(1+t_n)\,t_n\,\tau^p \mathrm{e}^{\omega^+\kappa t_n}\left(\|u^{(p+1)}\|_\infty + \|f^{(p+1)}\|_\infty\right), \quad n \geq 1, \tag{2.14}$$

whereas the same result with generator B and initial data v_0 leads to

$$\|f(t_n + \cdot) - \bar{v}_n\|_\infty = \|\mathrm{e}^{t_n B}f - r^n(\tau B)f\|_\infty \leq C_e t_n \tau^p \|f^{(p+1)}\|_\infty, \quad n \geq 1, \tag{2.15}$$

a bound that guarantees that we can use the exact values of the source term $f(t_n + s)$ instead of the terms of the recurrence $\bar{v}_n(s)$, since they are similar within the order of convergence that we deserve.

We need an explicit isolate expression for the recurrence of $\bar{u}_n$ where the resolvents of the operator G does not appear, but only resolvents of A and B. To this end, we use part (c) in Proposition 2.2.

Proposition 2.3. The recurrence for $\bar{u}_n$ in (2.13) can be written as

$$\bar{u}_{n+1} = r\left(\tau A\right)\bar{u}_n + \tau E(\tau)\bar{v}_n, \qquad \bar{v}_{n+1} = r\left(\tau B\right)\bar{v}_n, \quad n \geq 1, \tag{2.16}$$

where $E(\tau) : \mathcal{C}_{ub}\left([0,\infty), X\right) \to X$ is the linear operator given by

$$E(\tau)v = \sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} r_{\ell,j} w_\ell \sum_{i=1}^{j} \left(I - \tau w_\ell A\right)^{-j+i-1} L\left(I - \tau w_\ell B\right)^{-i} v, \tag{2.17}$$

for $v \in \mathcal{C}_{ub}\left([0,\infty), X\right)$, which is bounded for $0 < \tau < \tau_0$.

Proof. Taking into account the discrete variation-of-constants formula and the recurrence (2.12) with $\lambda = 1/\tau w_\ell$, the first component of the resolvent

$$\left(I - \tau w_\ell G\right)^{-j}\begin{pmatrix}\bar{u}_n\\ \bar{v}_n\end{pmatrix} \tag{2.18}$$

is given by the expression

$$\left(I - \tau w_\ell A\right)^{-j}\bar{u}_n + \tau w_\ell \sum_{i=1}^{j}\left(I - \tau w_\ell A\right)^{-j+i-1} L\left(I - \tau w_\ell B\right)^{-i}\,\bar{v}_n. \tag{2.19}$$

Thus, the fist component of (2.13) can be developed as

$$\begin{aligned}\bar{u}_{n+1} &= r_\infty\,\bar{u}_n + \sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} r_{\ell,j}\left(I - \tau w_\ell A\right)^{-j}\bar{u}_n\\ &\qquad + \tau\sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} r_{\ell,j} w_\ell \sum_{i=1}^{j}\left(I - \tau w_\ell A\right)^{-j+i-1} L\left(I - \tau w_\ell B\right)^{-i}\,\bar{v}_n\\ &= r(\tau A)\bar{u}_n + \tau E(\tau)\,\bar{v}_n.\end{aligned}$$

The fact that the operator $E(\tau)$ is bounded for $0 < \tau < \tau_0$ follows from the fact that the resolvents $\left(I - \tau w_\ell A\right)^{-j}$ are bounded for $0 < \tau < \tau_0$ and that the resolvents $\left(I - \tau w_\ell B\right)^{-j}$ are bounded for every $\tau > 0$. □

However, even thought the semigroup of translations $\left\{\mathrm{e}^{tB}\right\}_{t\geq 0}$ is trivial, the resolvents of B cannot be computed in a direct way by using evaluations of the argument f along a discrete mesh. However, for our purposes, it will be sufficient to approximate the resolvent with a suitable order. Recalling from basic theory of semigroups that (1.3), then

$$\left(\lambda I - B\right)^{-1} = \int_0^\infty \mathrm{e}^{-\lambda s}\mathrm{e}^{sB}\,ds,$$

so we see that

$$\left[\left(\lambda I - B\right)^{-1} v\right](t) = \int_0^\infty \mathrm{e}^{-\lambda s} v(t+s)\,ds, \tag{2.20}$$

for $t \geq 0$ and $v \in \mathcal{C}_{ub}([0,\infty), X)$. Though in principle it makes sense to approximate (2.20) by some adequate quadrature formula, our approach is based on Lemma 2.4 below that also allows us to approximate $F(\tau B)$ for more general functions F, in particular, for the powers of the resolvent of B. In what follows, for $v \in \mathcal{C}_{ub}([0,\infty), X)$, $\mathbf{c} \in \mathbb{R}^p$ and $t, \tau > 0$ such that $t + \tau\mathbf{c} \geq 0$, $v(t+\tau\mathbf{c})$ denotes $[v(t+\tau c_1), \ldots, v(t+\tau c_p)]^T$. Moreover, for a vector $\boldsymbol{\gamma} \in \mathbb{C}^q$, we set $\boldsymbol{\gamma}^T \cdot v(t+\tau\boldsymbol{c}) = \sum_{i=1} \gamma_i v(t+\tau c_i) \in X$.

Lemma 2.4. Let F be a rational mapping with no poles on the half-plane $\mathrm{Re}(z) \leq 0$ and bounded at infinity and $\boldsymbol{c} \in \mathbb{R}^p$ with $c_k \neq c_j$ for $k \neq l$. Then, there exists a unique $\boldsymbol{\gamma} = \{\gamma_k\}_{k=1}^p \in \mathbb{C}^p$ and $C > 0$ such that for t, $\tau > 0$, $t + \tau\boldsymbol{c} \geq 0$,

$$\|F(\tau B)\, v(t+\cdot) - \boldsymbol{\gamma}^T \cdot v(t+\tau\boldsymbol{c}+\cdot)\| \leq C\tau^p \|B^p v\|, \qquad v \in D(B^p) \tag{2.21}$$

Proof. Let us consider the Taylor expansions

$$\begin{aligned} F(z) &= f_0 + f_1 z + \cdots + f_{p-1} z^{p-1} + O(z^p), \\ \mathrm{e}^{c_k z} &= 1 + c_k z + \cdots + \frac{c_k^{p-1}}{(p-1)!} z^{p-1} + O(z^p), \quad 1 \leq k \leq p, \end{aligned}$$

and try to find $\boldsymbol{\gamma} = \{\gamma_k\}_{k=1}^p$ such that

$$H(z) = F(z) - \sum_{k=1}^p \gamma_k \mathrm{e}^{c_k z} = O(z^p).$$

This leads to the Vandermonde system

$$c_1^j \gamma_1 + \cdots + c_p^j \gamma_p = j!\, f_j, \qquad 0 \leq j \leq p-1, \tag{2.22}$$

which, since $c_k \neq c_l$ for $k \neq l$, has a unique solution. We denote by $V(\mathbf{c})$ the Vandermonde matrix of the system and by $\tilde{\mathbf{f}} = \{j!\, f_j\}_{j=0}^{p-1}$, so the previous system is $V(\mathbf{c})\,\gamma = \tilde{\mathbf{f}}$.

Set $P_{p-1}(z) = \sum_{k=0}^{p-1} f_k z^k$. For $v \in D(B^p)$, the above calculation implies that, for $t > 0$, $\tau > 0$, with $t + \tau c_k \geq 0$, $1 \leq k \leq p$,

$$\begin{aligned} \sum_{k=0}^{p-1} f_k \tau^k v^{(k)}(t) - \sum_{j=1}^p \gamma_j v(t+\tau c_j) &= \sum_{k=0}^{p-1} \tau^k v^{(k)}(t) \left(f_k - \sum_{j=1}^p \frac{\gamma_j c_j^k}{k!} \right) - \sum_{j=1}^p \gamma_j\, R_j \\ &= -\sum_{j=1}^p \gamma_j\, R_j, \end{aligned} \tag{2.23}$$

where R_j is the integral remainder

$$R_j = \frac{1}{(p-1)!} \int_0^{\tau\, c_j} (\tau c_j - s)^{p-1}\, v^{(p)}(t+s)\, ds, \quad 1 \leq j \leq p,$$

which is in fact bounded by

$$\|R_j\| \leq \frac{(\tau\, |c_j|)^p}{p!} \|B^p v\|, \quad 1 \leq j \leq p. \tag{2.24}$$

Then, by (2.23) and (2.24),

$$\left\| \sum_{k=0}^{p-1} f_k \tau^k v^{(k)}(t) - \sum_{j=1}^{p} \gamma_j v(t+\tau c_j) \right\| \leq \left(\frac{1}{(p-1)!} \|\tilde{\mathbf{f}}\|_\infty \|V(\mathbf{c})^{-1}\|_\infty \max_{1\leq k\leq p} |c_k|^p \right) \tau^p \|B^p\, v\| \tag{2.25}$$

Since this is also valid for $t+s$, with $s \geq 0$, we can claim that there exists $C_1 > 0$ (depending only on F and $\mathbf{c}$) such that

$$\|P_{p-1}(\tau B) S_B(t) v - \boldsymbol{\gamma}^T \cdot v(t+\tau c_k + \cdot)\| \leq C_1 \tau^p \|B^p v\|, . \tag{2.26}$$

Let us now consider the rational mapping $F_0(z) = (F(z) - P_{p-1}(z))/z^p$, that is bounded on the half plane $\mathrm{Re}(z) \leq 0$. Given that $(F - F(\infty))$, F', F_0 and F_0' are square integrable along the imaginary axis, Proposition 1.9.1 implies that $F(z), F_0(z) \in \widetilde{M}_0$. Moreover, by Proposition 1.9.3, we have

$$F(\tau B)v - P_{p-1}(\tau B) w = \tau^p F_0(\tau B) B^p w, \qquad w \in D(B^p),$$

and, by Proposition 1.9.4, the norms $\|F_0(\tau B)\|$ are uniformly bounded, for $\tau > 0$, by the total variation of the original measure of F_0. Therefore, there exists $C_2 > 0$ such that

$$\|F(\tau B) w - P_{p-1}(\tau B) w\| \leq C_2 \tau^p \|B^p w\|, \qquad w \in D(B^p),$$

an estimate that, applied to $w = S_B(t)v$, leads to

$$\|F(\tau B) S_B(t) v - P_{p-1}(\tau B) S_B(t) v\| \leq C_2 \tau^p \|B^p v\|. \tag{2.27}$$

To conclude the proof, we just write

$$F(\tau B) S_B(t) v - \boldsymbol{\gamma}^T \cdot v(t + \tau\mathbf{c}) = (I) + (II),$$

where

$$\begin{aligned} (I) &= F(\tau B) S_B(t) v - P_{p-1}(\tau B) S_B(t) v, \\ (II) &= P_{p-1}(\tau B) S_B(t) v - \boldsymbol{\gamma}^T \cdot v(t + \tau\mathbf{c}), \end{aligned}$$

and recall (2.27) and (2.26). □

Remark 2.5. The proof of Lemma 2.4 shows that the constant C in (2.21) increases linearly with the terms $\max_{1\leq k\leq p} |c_k|^p$ and $\|V(\mathbf{c})^{-1}\|_\infty$. This means that although it is impossible to approximate the resolvent with order p using any nodes, some node choices will give larger errors due to the increase on the constant. We briefly describe how these terms influence the error:

- The factor $\|V(\mathbf{c})^{-1}\|_\infty$ is related to the *interpolation properties of the chosen set of nodes* (see e.g. [34, 35]), and we expect it to grow with the Lebesgue constant. To reduce the number of function evaluations, it is interesting to consider equispaced nodes. It can be checked that the norm ρ of the matrix of the system (2.22) (with respect to the maximum norm) is minimal among the equispaced nodes when we use the p centered nodes (2.41). With such a choice, the values of ρ , for $3 \leq p \leq 9$, are approximately [2, 5, 9.5, 16.7, 37.4, 133.5, 356], moderate enough for practical use at least for $p \leq 7$.

However, other nodes, such as the Chebyshev nodes, may also be of interest, since they minimize the Lebesgue constant, although they have the drawback that p function evaluations per step are required.

- The factor $\max_{1\leq k\leq p} |c_k|^p$ shows that nodes with *larger absolute value* result in larger errors. Therefore, among all the equispaced nodes, the most central ones are those that contribute to a bigger reduction of this factor. Chebyshev nodes have also good properties in this point.

We finally propose the following method. Set $\mathcal{D} = \{\boldsymbol{c} \in \mathbb{R}^p / c_i \neq c_j, i \neq j\}$ and choose a sequence $\boldsymbol{c_n}$, $n \geq 1$, in $\mathcal{D}$. Lemma 2.4 applied to $F_{\ell,i}(z) = (1 - w_\ell z)^{-i}$, $1 \leq \ell \leq k$, $1 \leq i \leq m_\ell$, and a vector $\boldsymbol{c_n}$, provides $\boldsymbol{\gamma^n_{\ell,i}} \in \mathbb{R}^p$ that leds to an approximation of order p (in the sense of (2.21))

$$L\,(I - \tau w_\ell B)^{-i}\, v \approx \boldsymbol{\gamma^n_{\ell,i}}^T \cdot v\,(\tau \boldsymbol{c_n})\,. \tag{2.28}$$

We then adopt

$$\begin{aligned} u_{n+1} &= r_\infty\, u_n \\ &\quad + \sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} r_{\ell,j}\,(I - \tau w_\ell A)^{-j}\left(u_n + \tau w_\ell \sum_{i=1}^{j} (I - \tau w_\ell A)^{i-1}\, \boldsymbol{\gamma}^{n\ T}_{\ell,i} \cdot f\,(t_n + \tau \boldsymbol{c}_n)\right) \\ &= r\,(\tau A)\, u_n + \tau E_n(\tau) f\,(t_n + \tau \boldsymbol{c_n})\,, \end{aligned} \tag{2.29}$$

where $E_n(\tau) : X^p \to X$, $0 < \tau < \tau_0$, is the bounded, linear operator defined by

$$E_n(\tau) f\,(t_n + \tau \boldsymbol{c_n}) = \sum_{\ell=1}^{k}\sum_{j=1}^{m_\ell} r_{\ell,j} w_\ell \sum_{i=1}^{j} (I - \tau w_\ell A)^{-j+i-1}\, \boldsymbol{\gamma}^{n\ T}_{\ell,i} \cdot f\,(t_n + \tau \boldsymbol{c_n})\,, \tag{2.30}$$

whose dependence on n is only due to the possibility of choosing different nodes at each step. Notice that the vector of abscissa $\boldsymbol{c}_{RK}$ of the Runge–Kutta method is neither used in (2.29) nor related to $\boldsymbol{c_n}$. Note that the formula in the first lines in (2.29) is written in a more explicit way (that shows how the method can be implemented), while the last line is more compact and will be useful in the analysis of the method. Recalling (2.17) and in view of (2.28) there holds

$$\|E_n(\tau) f(t_n + \tau \boldsymbol{c_n}) - E(\tau) f(t_n + \cdot)\| \leq K_n \tau^p \|f^{(p)}\|_\infty, \tag{2.31}$$

for some $K_n = K_n(\boldsymbol{c_n}) > 0$. One step in (2.29) requires solving s resolvents of A, as in the homogeneous case. In fact, it is a straightforward verification that (2.29) is equivalent to computing the s resolvents

$$U^n_{\ell,0} = u_n, \qquad U^n_{\ell,j} = (I - \tau w_\ell A)^{-1}\left(U^n_{\ell,j-1} + \tau w_\ell\, \boldsymbol{\gamma}^{n\ T}_{\ell,j} \cdot f\,(t_n + \tau \boldsymbol{c_n})\right), \tag{2.32}$$

for $1 \le \ell \le k$, $1 \le j \le m_\ell$, and then taking the linear combination

$$u_{n+1} = r_\infty \, u_n + \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell,j} \, U_{\ell,j}. \tag{2.33}$$

Moreover, for arbitrary $\boldsymbol{c_n}$, it also requires p evaluations of the function f. However, the vectors $\boldsymbol{c_n}$ can be chosen in such a way that only one evaluation per step is done for $n \ge 2$ (see Section 2.3 for details). To prove convergence we also require that all the $\boldsymbol{c_n}$ lie in a compact set $\mathcal{K} \subset D$, in such a way that $K_n(\boldsymbol{c_n}) \le K$ when $n \ge 1$.

2.2.2 Analysis of convergence

We state the main result of the chapter, that assures that the method (2.29) converges to the solution of (2.1) without order reduction. Anyway, the optimal order p could be reduced in case of weak stability (1.37).

Theorem 2.6. Let $u : [0, \infty) \to X$ be the solution of (2.1) to be approximated in the interval $[0, T]$ with constant step size $0 < \tau = T/N < \tau_0$. Assume that $u \in \mathcal{C}^{p+1}([0, \infty), X)$, $f \in \mathcal{C}^{p+1}([0, \infty), X)$. Let u_n be the numerical approximation to $u(t_n)$ obtained by the modified rational method (2.29) with nodes $\boldsymbol{c_n}$ in some compact set $\mathcal{K} \subset \mathcal{D}$. Then, there exists a constant $K = K(\mathcal{K}) > 0$ such that, for $0 \le n \le N$,

$$\|u(t_n) - u_n\| \le K C_e C_s(n)(1 + t_n) t_n M \mathrm{e}^{\omega^+ \kappa t_n} \tau^p \left(\|u^{(p+1)}\|_\infty + \|f^{(p)}\|_\infty + \|f^{(p+1)}\|_\infty \right). \tag{2.34}$$

Proof. First, notice that (2.16) can be written as

$$\bar{u}_{n+1} = r\left(\tau A\right) \bar{u}_n + \tau E(\tau) \, r^n \left(\tau B\right) f,$$

and subtracting this expression from (2.29),

$$\begin{aligned} u_{n+1} - \bar{u}_{n+1} &= r\left(\tau A\right)\left(u_n - \bar{u}_n\right) + \tau \left(E_n(\tau) f(t_n + \tau \boldsymbol{c_n}) - E(\tau) r^n \left(\tau B\right) f\right) \\ &= r\left(\tau A\right)\left(u_n - \bar{u}_n\right) + \tau \left(E_n(\tau) f(t_n + \tau \boldsymbol{c_n}) - E(\tau) f(t_n + \cdot)\right) \\ &\quad + \ \tau \left(E(\tau) \left(f(t_n + \cdot) - r^n \left(\tau B\right) f\right)\right), \end{aligned}$$

with $u_0 = \bar{u}_0$. Then, by the variation-of-constants formula, the error can be bounded by three terms

$$\|u(t_n) - u_n\| \le (I) + (II) + (III), \tag{2.35}$$

where

$$\begin{aligned} (I) &= \|u(t_n) - \bar{u}_n\|, \\ (II) &= \tau \sum_{k=0}^{n-1} \|r^{n-k}\left(\tau A\right)\| \|E_k(\tau) f(t_k + \tau \boldsymbol{c_k}) - E(\tau) f(t_k + \cdot)\|, \\ (III) &= \tau \sum_{k=0}^{n-1} \|r^{n-k}\left(\tau A\right)\| \|E(\tau) \left(f(t_k + \cdot) - r^k \left(\tau B\right) f \right) \|. \end{aligned}$$

To bound these terms, we proceed as follows. A bound for (I) is given by (2.14). Moreover, taking into account the compactness of $\mathcal{K}$, (1.37) and (2.31) we get

$$(II) \leq K\, M\, C_s(n)\, t_n\, \mathrm{e}^{\omega^+ \kappa t_n}\, \tau^p \|f^{(p)}\|_\infty, \qquad K = K(\mathcal{K}).$$

Finally, the fact that $E(\tau)$ is bounded together with (1.37) and (2.15) assures that

$$(III) \leq M\, C_e\, C_s(n)\, t_n^2\, \mathrm{e}^{\omega^+ \kappa t_n}\, \tau^p \|f^{(p+1)}\|_\infty,$$

and the proof concludes combining the three estimates. □

When both the semigroup $\left\{\mathrm{e}^{tA}\right\}_{t\geq 0}$ and the source term $f(t)$ admit analytic continuations to some sector

$$S_\theta = \left\{z \in \mathbb{C} \,/\, |\arg(z)| \leq \theta\right\}, \qquad 0 < \theta < \pi/2,$$

it is possible to reformulate Theorem 2.6 in the line of (1.39). Moreover, as we mentioned, the stability constant $C_s(n)$ can be dropped so the order p is actually recovered and further results concerning bad initial values and a variable step size version of the method can be considered.

We introduce the space $\mathcal{A}_{ub}\left(S_\theta, X\right)$ of all bounded, uniformly continuous, analytic functions f from S_θ to X endowed with the supremum norm, that turns out to be a Banach space. This is in fact equivalent to the Laplace transform F of f having an analytic extension to a sector of the form $\{\lambda \in \mathbb{C} : 0 \leq |\arg(\lambda)| \leq \pi - \theta, \lambda \neq \omega\}$.

There we define the analytic semigroup of translations $S_{B_\theta} : \mathcal{A}_{ub}\left(S_\theta, X\right) \to \mathcal{A}_{ub}\left(S_\theta, X\right)$ given by

$$\left[S_{B_\theta}(z)\, v\right](w) = v(z + w), \quad \text{for} \quad v \in \mathcal{A}_{ub}\left(S_\theta, X\right), w \in S_\theta.$$

We can take $D\left(B_\theta\right) = \mathcal{A}_{ub}\left(S_\theta, X\right)$, the corresponding generator is $B_\theta : Y_\theta \to Y_\theta$ with $B_\theta v = v'$ for $v \in \mathcal{A}_{ub}\left(S_\theta, X\right)$. A similar proof to that in Proposition 2.1 guarantees that this is a $\mathcal{C}_0$ semigroup of operators. To see that it is in fact an analytic semigroup, we can check condition 4 in Theorem 1.6, which is a direct consequence of the fact that, for $f \in \mathcal{A}_{ub}\left(S_\theta, X\right)$, the Cauchy theorem implies that

$$|f'(t)| \leq \frac{\|f\|_\infty}{|t \sin\theta|},$$

for $t > 0$.

Then, the semigroup generated by $G_\theta = X \times \mathcal{A}_{ub}\left(S_\theta, X\right)$ becomes analytic too (see, e.g., section 2.3 in [77]) and we can state the analytic version of Theorem 2.6.

Theorem 2.7. Let $u : [0, \infty) \to X$ be the solution of (2.1) to be approximated on the interval $[0, T]$ with constant step size $0 < \tau = T/N < \tau_0$. Assume that $A \in \mathcal{G}(X, M, \omega, \theta)$ generates an analytic semigroup of linear operators and r is a strongly $A(\vartheta)$-stable rational function, with $\theta < \vartheta$. Assume also that $u \in \mathcal{C}^p\left([0, T], X\right)$, $f \in \mathcal{A}_{ub}\left(S_\theta, X\right)$. Let u_n be the numerical approximation to $u(t_n)$ obtained by the modified rational method (2.29) with nodes $\boldsymbol{c_n}$ in some compact set $\mathcal{K} \subset \mathcal{D}$. Then, there exists a constant $K = K(\mathcal{K}) > 0$ such that

$$\|u(t_n) - u_n\| \leq K\, M\, C_e\, (1 + t_n)\, \mathrm{e}^{\omega^+ \kappa t_n}\, \tau^p \left(\|u^{(p)}\|_\infty + \|f^{(p)}\|_\infty\right), \quad 0 \leq n \leq N.$$

Proof. The proof is similar to that of Theorem 2.6, but in this case we take advantage of the optimal parabolic estimate (1.39) instead of (1.38). We split the local error in the same three terms as in (2.35). Then, the estimate (1.39) applied to G_θ and initial data u_0 and f leads to

$$(I) \leq M\, C_e\, (1+t_n)\; \mathrm{e}^{\omega^+ \kappa t_n} \tau^p \, \left(\|u^{(p)}\|_\infty + \|f^{(p)}\|_\infty \right).$$

Moreover, as we mentioned in Section 3, in the analytic case $C_s(n)$ is $O(1)$, so that

$$(II) \leq K\, C_s\, M\, t_n\, \mathrm{e}^{\omega^+ \kappa t_n}\, \tau^p \, \|f^{(p)}\|_\infty, \qquad K = K(\mathcal{K}).$$

To conclude, the application of (1.39) to B_θ with initial data f and (1.37) gives

$$(III) \leq M\, C_s\, C_e\, t_n \mathrm{e}^{\omega^+ \kappa t_n}\, \tau^p \|f^{(p)}\|_\infty,$$

thus completing the proof. □

2.2.3 Variable step size version of the methods

The numerical scheme (2.29) turns out to be efficient in terms of computational cost required per step. However, solving evolution equations of the form (2.1) in practical scenarios naturally leads to the need for a variable step size strategy compatible with the method. In this section, we describe how to adapt the typical variable step size implementations of the Runge–Kutta methods to the case of the rational schemes that we propose.

The variable step size implementations of the Runge–Kutta methods is based on the construction of *embedded Runge–Kutta pairs*. These are methods that share the same matrix $W \in \mathbb{R}^{s\times s}$ and node vector $\mathbf{c} \in \mathbb{R}^s$, but they have different weights $\mathbf{b}, \hat{\mathbf{b}} \in \mathbb{R}^s$, in such a way that they have orders of convergence p and $p-1$, respectively. The Butcher tableau is typically written in the form

$$\begin{array}{c|c} \mathbf{c} & W \\ \hline & \mathbf{b}^T \\ & \hat{\mathbf{b}}^T \end{array}, \qquad \mathbf{b}, \mathbf{c} \in \mathbb{R}^s, W \in \mathbb{R}^{s\times s},$$

This embedded pair allows to calculate an efficient estimate of the local error by subtracting the numerical approximation to the solution produced by the two methods. This estimate of the local error is the key for using a method with variable step size, since it allows to decide to increase the step size when the errors are small and to decrease it when the errors are large. In this way, a balance between efficiency and error control is maintained.

To adapt this framework to our rational method schemes, notice that the rational stability function of the methods are given by

$$r_1(z) = 1 + z\mathbf{b}^T\,(I - zW)^{-1}\,\mathbf{e}, \quad r_2(z) = 1 + z\hat{\mathbf{b}}^T\,(I - zW)^{-1}\,\mathbf{e}, \quad \mathbf{e} = [1,\ldots,1]^T \in \mathbb{R}^s. \tag{2.36}$$

This formula shows that the poles of both rational mappings are determined by the common matrix W and does not depend on $\mathbf{b}$, $\hat{\mathbf{b}}$ whatsoever. They can be developed into simple fractions like

$$r_1(z) = r_\infty + \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} \frac{r_{j,\ell}}{(1 - zw_\ell)^j}, \qquad r_2(z) = \hat{r}_\infty + \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} \frac{\hat{r}_{j,\ell}}{(1 - zw_\ell)^j}.$$

We can then consider two different schemes of orders $p-1$, p of the form (2.29) corresponding to each rational function with the same nodes $\mathbf{c}_n \in \mathbb{R}^p$ in each step. For the scheme of order $p-1$, it is not necessary to use p nodes instead of $p-1$, but since both rational functions have the same poles, after this choice of nodes we will have to solve exactly *the same linear systems* for both methods. An efficient estimate of the local error may be constructed in this way after solving the linear systems. In fact, after solving the s linear systems in (2.32), we can take

$$EST_{n+1} = (r_\infty - \hat{r}_\infty)\, u_n + \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} (r_{\ell,j} - \hat{r}_{\ell,j})\, U_{\ell,j}. \tag{2.37}$$

as an estimate of the local error that allows us to decide the variation of the step size in the following step. Then a step of the method can be done by taking

$$u_{n+1} = r(\tau_n A)u_n + \tau_n E(\tau_n) f(t_n + \tau_n \mathbf{c}_n).$$

The convergence analysis for the variable step size case follows closely the arguments given in the proof of Theorem 2.6. The analysis of the local error is essentially the same. For the stability analysis, one needs to investigate the growth of the product

$$\prod_{j=0}^{n} \|r(\tau_j A)\|,$$

for which (1.37) provides the corresponding bound in the constant step size case. For general semigroups, the techniques of the Hille–Phillips functional calculus (as previously described in Section 1 and in [18]) can be employed to show that the latter shows a similar bound to that in (1.37) whenever $a \le \tau_n/\tau_m \le b$, for $n, m \ge 0$, and some constants $0 < a < b$. Moreover, for analytic semigroups $A \in \mathcal{G}(X, M, \omega, \theta)$ and $A(\vartheta)$-stable rational functions with $\vartheta > \theta$, it was proved in [62] that the estimate

$$\prod_{j=0}^{n} \|r(\tau_j A)\| \le C_s$$

holds for every sequence of positive step sizes.

2.3 Practical Implementation and Results

2.3.1 Efficient implementation of the methods

In this section we focus on the practical aspects of the rational methods introduced in the previous one. First, we show how these methods can be implemented efficiently. Then, we present several numerical results to show the convergence behaviour of the proposed method.

We deal with simple partial differential equations which are integrated by the method of lines. The spatial discretization is accomplished by standard finite differences. If $h > 0$ stands for the space-discretization parameter, we are led to systems of ordinary differential equations

$$\left\{ \begin{array}{rcl} u_h'(t) & = & A_h u_h(t) + f_h(t), \quad t \geq 0, \\ u_h(0) & = & u_{0,h}. \end{array} \right. \tag{2.38}$$

To focus on the error due to time integration we proceed as follows:

1. We start from a known solution $u(t,x)$ of (2.1), corresponding to some source term f, so that u takes values in the intermediate space X_ν, with $\nu < \nu^*$ (ν^* given in Section 1.3).

2. We adjust f_h in such a way that the restriction of u to the discrete mesh is the exact solution of (2.38).

3. We accept as reasonable that the order reduction of the RK method applied to (2.38) is close to the one determined by ν^* [21].

4. We compare a RK method with its rational version (2.29).

The illustrations use a constant step size $\tau > 0$ and the auxiliary vectors $\boldsymbol{c_n}$, $n \geq 0$, are chosen to be either the *explicit nodes*

$$\boldsymbol{c}_n^E = \left\{ \begin{array}{lll} [-n, -n+1, \dots, p-1-n] & \text{for} & n = 0, \dots, p-2 \\ [-p+1, -p+2, \dots, 0] & \text{for} & n > p-2, \end{array} \right. \tag{2.39}$$

the *implicit nodes*

$$\boldsymbol{c}_n^I = \left\{ \begin{array}{lll} [-n, -n+1, \dots, p-1-n] & \text{for} & n = 0, \dots, p-3 \\ [-p+2, -p+2, \dots, 1] & \text{for} & n > p-3, \end{array} \right. \tag{2.40}$$

the *centered nodes*

$$\boldsymbol{c}_n^C = \left\{ \begin{array}{lll} [-n, -n+1, \dots, p-1-n] & \text{for} & n = 0, \dots, p-3 \\ [-\hat{p}, -\hat{p}+1, \dots, \hat{p}] & \text{for} & n > p-3, \end{array} \right. \tag{2.41}$$

where $\hat{p} = (p-1)/2$, or the *Chebyshev nodes*

$$\boldsymbol{c}_n^T = \left\{ \cos\left(\frac{\pi}{2} \frac{2k-1}{p} \right) \right\}_{k=1}^{p}, \quad \text{for} \quad n \geq 0. \tag{2.42}$$

Notice that with the choices of nodes $\boldsymbol{c}_n^E$, $\boldsymbol{c}_n^I$, $\boldsymbol{c}_n^C$ only one function evaluation per step is required for $n \geq 1$. Although we call them explicit-implicit-centered, in the linear case discussed in this chapter all sets of nodes operate in the same way and with the same computational cost. The use of Chebyshev nodes requires p evaluations of f per step, since the previous evaluations cannot be used again in the following step. However, they are optimal for Lagrange interpolation and Remark 2.5 suggests its use which, as we shall see, turns out to be more efficient in some cases. We consider the Runge–Kutta and the rational methods with $\boldsymbol{c}_n^E$, $\boldsymbol{c}_n^I$ in all the cases, since we want to compare our methods with the Runge–Kutta and $\boldsymbol{c}_n^E$, $\boldsymbol{c}_n^I$ are the basis to integrate nonlinear problems. We alternate the use of $\boldsymbol{c}_n^C$ and $\boldsymbol{c}_n^T$ to show different behaviours.

Concerning the implementation, first of all, we need linear solvers for the involved systems

$$(I - \tau w_\ell A_h)x_h = y_h \in X_h, \quad 1 \leq l \leq k. \tag{2.43}$$

In practice, the spatial discretization is included in the scheme via the particular form that these solvers take. Since we use a fixed time step for the integration, the linear systems to be solved at each time step share the same matrices. Therefore, when the solver allows it, an LU factorization of each matrix can be precomputed and reused throughout the integration, significantly reducing the computational cost by avoiding redundant operations in the repeated resolution of these systems.

To get the $s := \sum_{\ell=1}^{k} m_\ell$ vectors $\boldsymbol{\gamma}_{\boldsymbol{\ell},\boldsymbol{j}}^{\boldsymbol{n}} \in \mathbb{R}^p$, $1 \leq \ell \leq k$, $1 \leq j \leq m_\ell$, required in (2.30), we consider the expansions

$$(1 - w_\ell\, z)^{-j} = \sum_{q=0}^{p-1} \binom{-j}{q} (-1)^q w_\ell^q z^q + O(z^p)$$

and solve the s corresponding Vandermonde systems (2.22), for $F(z) = (1 - w_\ell z)^{-j}$, $1 \leq l \leq k$, $1 \leq j \leq m_\ell$ and $\mathbf{c}_n$, $n \geq 0$. It is clear that, in our context, we can restrict the task to solve such systems for $0 \leq n \leq p$.

Finally, once we know the coefficients, (2.30) is implemented in a Horner like algorithm, by using the available linear solvers. We describe how to implement this for the specific methods we used in the experiments. The rational function of the first method has several simple poles, while the second has a single multiple pole. For a general rational function with several multiple poles, it would suffice to combine the development of these two specific cases, grouping each multiple pole into blocks.

Notice that the rational function $r(z)$ may have a pair of complex conjugate numbers $w_\ell, \bar{w}_\ell$ as poles. When this is the case, it is straightforward to verify that the factors $r_\ell, \bar{r}_\ell$ and the corresponding vectors provided by Lemma 2.4 are also conjugates $\boldsymbol{\gamma}_{\ell,j}, \bar{\boldsymbol{\gamma}}_{\ell,j}$, for every rational mapping F. As a consequence, the sum

$$r_\ell(I - \tau w_\ell A)^{-1}\left(u_n + \tau \boldsymbol{\gamma}_{\ell,j} \cdot f(t_n + \tau \boldsymbol{c}_n)\right) + \bar{r}_\ell(I - \tau \bar{w}_\ell A)^{-1}\left(u_n + \tau \bar{\boldsymbol{\gamma}}_{\ell,j} \cdot f(t_n + \tau \boldsymbol{c}_n)\right)$$

is real whenever $u(t), f(t) \in \mathbb{R}$, for every $t \geq 0$, and one expects a real approximation u_n. Furthermore, complex arithmetic can be avoided by working directly with the real and imaginary parts of the involved quantities and rearranging the systems, if necessary.

We consider rational methods based on two classical Runge–Kutta schemes: the three-stage Gauss method and the three-stage SDIRK method. Our aim is to illustrate the theorems and present different implementations, since the poles of the Gauss method are simple whereas those of SDIRK are multiple. For a detailed discussion of the efficiency of Runge–Kutta methods in practical scenarios, see for instance [46, 47]. Nevertheless, many other Runge–Kutta schemes can be adapted to our framework in a similar way, even in a variable step size formulation.

The 3-stages Gauss method (see, e.g., [38]) is the Runge–Kutta method with the following Butcher tableau

$$\begin{array}{c|ccc}
\frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\
\frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\
\frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline
 & \frac{5}{18} & \frac{4}{9} & \frac{5}{18}
\end{array}.$$

The method has $s=3$ stages and order of convergence $p=6$, it is A-stable and it has a rational stability function that can be written in the form

$$r(z)=\frac{r_1}{1-w_1 z}+\frac{r_2}{1-w_2 z}+\frac{r_3}{1-w_3 z}-1, \tag{2.44}$$

with $r_\ell, w_\ell \in \mathbb{C}$, $\mathrm{Re}(w_\ell)>0$, for $\ell=1,2,3$. As the method has order $p=6$, in each step we fix $\mathbf{c_n}\in\mathbb{R}^6$ and obtain three vectors $\boldsymbol{\gamma_\ell^n}\in\mathbb{R}^6$ by solving the mentioned linear systems. Then, a step of the method is done by computing

$$u_{n+1}=\sum_{\ell=1}^{3} r_\ell\,(I-\tau w_\ell A)^{-1}\left(u_n+\tau\, w_\ell\,\boldsymbol{\gamma_\ell^n}\cdot f\left(t_n+\tau\boldsymbol{c_n}\right)\right)-u_n.$$

The 3-stages SDIRK method (see, e.g., [38]) is defined by the tableau

$$\begin{array}{c|ccc}
\gamma & \gamma & & \\
\frac{1}{2} & \frac{1}{2}-\gamma & \gamma & \\
1-\gamma & 2\gamma & 1-4\gamma & \gamma \\ \hline
 & \delta & 1-2\delta & \delta
\end{array}
\qquad \gamma=\frac{1}{\sqrt{3}}\cos\left(\frac{\pi}{18}\right)+\frac{1}{2}, \qquad \delta=\frac{1}{6(2\gamma-1)^2}.$$

This method has $s=3$ stages, order of convergence $p=4$, it is A-stable and it has a stability function of the form

$$r(z)=\sum_{j=1}^{3} r_j\,(1-\gamma z)^{-j}. \tag{2.45}$$

Since $r_\infty=0$, the method is in fact strongly A-stable. Once again, after choosing nodes $\mathbf{c}_n\in\mathbb{R}^4$ and solving the corresponding linear systems to obtain $\boldsymbol{\gamma_j^n}\in\mathbb{R}^4$, the scheme is

$$u_{n+1}=\sum_{j=1}^{3} r_j\,(I-\tau\gamma A)^{-j}\;u_n+\tau\sum_{j=1}^{3} r_j\,\gamma\sum_{i=1}^{3}(I-\tau\gamma A)^{-j+i-1}\,\boldsymbol{\gamma_j^n}\cdot f\left(t_n+\tau\boldsymbol{c_n}\right). \tag{2.46}$$

This formula can be computed in practice by using the Horner-like form of the algorithm. In fact, if we define recursively the values

$$U_0^n=u_n, \quad U_j^n=(I-\tau\gamma A)^{-1}\left(U_{j-1}^n+\tau\,\gamma\,\boldsymbol{\gamma_j^n}\cdot f\left(t_n+\tau\boldsymbol{c_n}\right)\right),$$

for $j = 1, 2, 3$, then (2.46) is equivalent to

$$u_{n+1} = \sum_{j=1}^{3} r_j \,(I - \tau\gamma A)^{-j} \left(U_{j-1}^n + \tau\,\gamma\, \boldsymbol{\gamma_j^n} \cdot f\,(t_n + \tau \boldsymbol{c_n}) \right) = \sum_{j=1}^{3} r_j \, U_j^n. \qquad (2.47)$$

This formula recalls that a step of the scheme can be computed by solving s linear systems even in the case of multiple poles.

2.3.2 Numerical experiments

Hyperbolic problem

We consider a hyperbolic problem in the unit interval with homogeneous boundary conditions,

$$\left\{ \begin{array}{lcll} u_t(t,x) & = & -u_x(t,x) + f(t,x), & 0 \le t \le 1, \quad 0 \le x \le 1, \\ u(0,x) & = & u_0(x), & 0 \le x \le 1, \\ u(t,0) & = & 0, & 0 \le t \le 1, \end{array} \right. \qquad (2.48)$$

where $f : [0,1] \times [0,1] \to \mathbb{C}$, $u_0 : [0,1] \to \mathbb{C}$. In order to fit the problem in our framework, we take $X = L^2[0,1]$, $A = -d/dx$, $D(A) = \{u \in H^1[0,1] : u(0+) = 0\}$. The operator A satisfies (1.2) with $\omega = 0$ and $M = 1$. We adjust the data u_0 and f in such a way that $u(t,x) = x\mathrm{e}^t$, $0 \le x \le 1$, $0 \le t \le 1$, is the solution of the problem (2.48). According to the results in Chapter 1, it is straightforward to prove that $u(t,\cdot) \in X_\nu$, $0 \le t \le t_f$, for every $0 < \nu < 1.5$.

Table 2.1: Errors and orders for the hyperbolic example (2.48) solved with RK and rational SDIRK3 method with h=1/100, $tf = 1$.

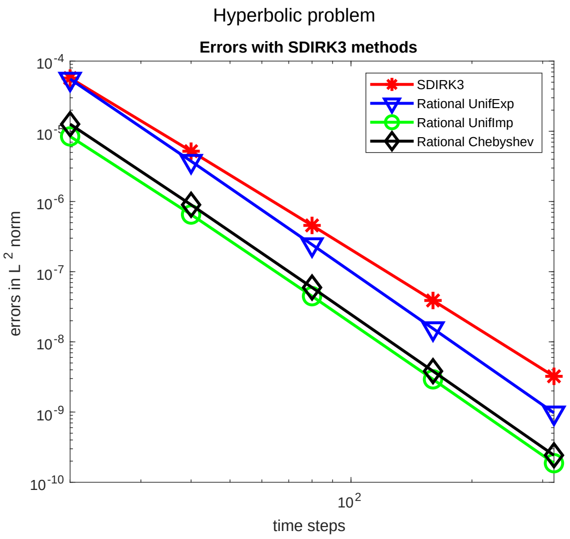
	Runge–Kutta		Explicit		Implicit		Chebyshev	
step size	error	order	error	order	error	order	error	order
5.000e-02	5.735e-05	–	5.550e-05	–	8.452e-06	–	1.264e-05	–
2.500e-02	5.207e-06	3.46	3.748e-06	3.89	6.517e-07	3.70	8.978e-07	3.82
1.250e-02	4.562e-07	3.51	2.430e-07	3.95	4.475e-08	3.86	5.951e-08	3.92
6.250e-03	3.887e-08	3.55	1.546e-08	3.97	2.925e-09	3.94	3.825e-09	3.96
4.167e-03	9.101e-09	3.58	3.072e-09	3.99	5.864e-10	3.96	7.627e-10	3.98

We discretize (2.48) by the method of lines, combining upwind finite diference for the discretization in space and the 3-stage SDIRK method ($p = 4$, $q = 1$) for the integration in time (see, e.g., [38]). The matrix of the semidiscrete system (2.38) of this spatial discretization on the uniform grid $x_i = ih$, for $i = 1, \ldots, N$, with $h = 1/N$, is

$$A_h = \frac{1}{h} \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}. \qquad (2.49)$$

It is clear that the sparsity of the matrix allows us to solve the systems (2.43) with a Thomas-like algorithm with $O(N)$ operations and memory usage. The SDIRK3 method suffers from order reduction, and according to the main result in [8], the reduced order turns out to be $p^* = q + \nu + 1 = 3.5$. The method is implemented with the different nodes $\mathbf{c}_n$ proposed (2.39-2.42), leading to the results shown in Table 2.1. The order p is achieved with the rational methods. In this problem, the nodes that minimize the error are $\mathbf{c}_n^I$.

Figure 2.1. Errors of the different implementations for the hyperbolic example (2.48) solved with RK and rational SDIRK3 method with h=1/100, $tf = 1$.



Parabolic problem in 1D

We consider the one-dimensional heat equation with homogeneous boundary conditions

$$\left\{ \begin{array}{rcll} u_t(t,x) & = & u_{xx}(t,x) + f(t,x), & 0 \le t \le t_f, \quad 0 \le x \le 1, \\ u(0,x) & = & u_0(x), & 0 \le x \le 1, \\ u(t,0) & = & 0, & 0 \le t \le t_f, \\ u(t,1) & = & 0, & 0 \le t \le t_f, \end{array} \right. \tag{2.50}$$

where $f : [0,1] \times [0,1] \to \mathbb{C}$, $u_0 : [0,1] \to \mathbb{C}$. In this case we consider $X = L^2[0,1]$, $A = d^2/dx^2$, $D(A) = H^2[0,1] \cap H_0^1[0,1]$. It is known that under these considerations the operator A satisfies (1.2) with $\omega = 0$ and $M = 1$. We adjust the data u_0 and f in such a way that $u(t,x) = (1-x)\sin(tx)\mathrm{e}^{t^2 x}$, $0 \le t \le t_f$, $0 \le x \le 1$, is the solution of problem (2.50). In this case, the results in Chapter 1 prove that $u(t,\cdot) \in X_\nu$, $0 \le t \le t_f$, for every $0 < \nu < 1.25$.

The problem (2.50) is discretized combining centered finite difference for the discretization in space and either the 3-stages Gauss method ($p = 6$, $q = 3$) or the 3-stage SDIRK method ($p = 4$, $q = 1$) for the integration in time. In this case the matrix of the semidiscrete system (2.38) of this spatial discretization on the uniform grid $x_i = ih$, for

Table 2.2: Errors and orders for the parabolic example (2.50) solved with RK and rational SDIRK3 method with h=1/100, $tf = 1$.

	Runge–Kutta		Explicit		Implicit		Chebyshev	
step size	error	order	error	order	error	order	error	order
5.000e-02	5.910e-05	–	2.156e-05	–	8.418e-07	–	1.481e-06	–
2.500e-02	9.256e-06	2.67	1.478e-06	3.87	8.838e-08	3.25	1.802e-07	3.04
1.250e-02	1.248e-06	2.89	9.899e-08	3.90	9.943e-09	3.15	1.699e-08	3.41
6.250e-03	1.486e-07	3.07	6.591e-09	3.91	9.249e-10	3.43	1.395e-09	3.61
3.125e-03	1.661e-08	3.16	4.343e-10	3.92	7.540e-11	3.62	1.044e-10	3.74
1.563e-03	1.802e-09	3.20	2.819e-11	3.95	5.559e-12	3.76	7.301e-12	3.84

$i = 1, \dots, N$, with $h = 1/(N+1)$, is

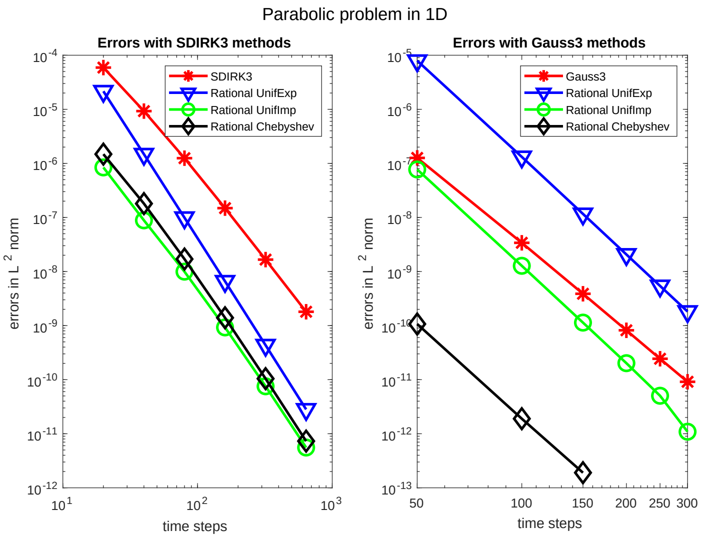
$$A_h = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & -2 \end{pmatrix}. \tag{2.51}$$

The linear systems (2.43) can be solved using the Thomas algorithm with $O(N)$ operations and memory usage. Regarding the time integrators, according to [8], the reduced orders are $p^* = q + \nu + 1 = 5.25$ for the Gauss3 and $p^* = q + \nu + 1 = 3.25$ for the SDIRK3. Tables 2.2 and 2.3 show the numerical orders obtained, which are in good agreement with the expected orders. Notice that figure 2.2 shows that the rational methods based on SDIRK3 are more efficient than the Runge–Kutta if the computational cost is just measured in terms of number of steps (we recall that the number of linear systems to be solved at each step is the same in all cases but the number of function evaluations may differ significantly). Conversely, in the Gauss3 case the Runge–Kutta is more efficient than the rational method with nodes $\boldsymbol{c}^E$ despite of the fact that the rational one has better order of convergence. Remark 2.5 explains this behaviour. The other choices of nodes do improve efficiency over Runge–Kutta methods. It is also consistent with the content of this remark the fact that Chevyshev nodes (2.42) are better for higher orders p than the uniform nodes.

Table 2.3: Errors and orders for the parabolic example (2.50) solved with RK and rational Gauss3 method with h=1/100, $tf = 2$.

	Runge–Kutta		Explicit		Implicit		Chebyshev	
step size	error	order	error	order	error	order	error	order
4.000e-02	1.261e-07	–	7.920e-06	–	7.720e-08	–	1.069e-10	–
2.000e-02	3.395e-09	5.22	1.318e-07	5.91	1.276e-09	5.92	1.897e-12	5.82
1.333e-02	3.877e-10	5.35	1.162e-08	5.99	1.135e-10	5.97	1.904e-13	5.67
1.000e-02	8.183e-11	5.41	2.067e-09	6.00	2.017e-11	6.00	**	**
8.000e-03	2.436e-11	5.43	5.415e-10	6.00	5.039e-12	6.22	**	**
6.667e-03	9.133e-12	5.38	1.819e-10	5.98	1.089e-12	8.40	**	**

Figure 2.2. Errors of the different implementations for the parabolic example (2.50) solved with RK and rational SDIRK3, $tf = 1$, and Gauss3, $t_f = 2$, with h=1/100.



Parabolic problem in 2D

Finally, we study the two dimensional problem in the square domain $\Omega = (0,1) \times (0,1)$, again with homogeneous Dirichlet boundary conditions,

$$\left\{ \begin{array}{rcll} u_t(t,x,y) & = & \Delta u(t,x,y) + f(t,x,y), & 0 \le t \le t_f, \quad (x,y) \in \Omega, \\ u(0,x,y) & = & u_0(x,y), & (x,y) \in \Omega, \\ u(t,x,y) & = & 0, & 0 \le t \le t_f, \quad (x,y) \in \partial\Omega, \end{array} \right. \tag{2.52}$$

where $f : [0,1] \times \bar{\Omega} \to \mathbb{C}$, $u_0 : \bar{\Omega} \to \mathbb{C}$. We consider $X = L^2(\Omega)$, $A = \Delta$, $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$. The bound (1.2) holds with $\omega = 0$ and $M = 1$. We adjust the data u_0 and f in such a way that $u(t,x,y) = x^3 y (x-1)(y-1)^3 \mathrm{e}^t$, $0 \le t \le t_f$, $(x,y) \in \bar{\Omega}$, is the solution of the problem (2.52).

Table 2.4: Errors and orders for the parabolic example (2.52) solved with RK and rational SDIRK3 method with h=1/100, $tf = 1$.

	Runge–Kutta		Explicit		Implicit		Centered	
step size	error	order	error	order	error	order	error	order
5.000e-02	9.612e-07	–	2.904e-08	–	1.311e-09	–	7.333e-10	–
2.500e-02	1.645e-07	2.55	1.745e-09	4.06	6.841e-11	4.26	1.007e-10	2.86
1.250e-02	2.417e-08	2.77	1.081e-10	4.01	7.403e-12	3.21	1.337e-11	2.91
6.250e-03	3.038e-09	2.99	7.008e-12	3.95	8.825e-13	3.07	1.332e-12	3.33
3.125e-03	3.372e-10	3.17	4.638e-13	3.92	8.320e-14	3.41	1.118e-13	3.57
1.563e-03	3.474e-11	3.28	3.118e-14	3.89	7.372e-15	3.50	9.126e-15	3.61

We discretize in space with centered finite differences on the uniform grid $(x_i, y_j) = (ih, jh)$, for $i, j = 1, \ldots, N$, with $h = 1/(N+1)$. If we consider lexicographic order to order the grid points, the matrix A_h is the $N^2 \times N^2$ tridiagonal-block matrix

$$A_h = \frac{1}{h^2}\begin{pmatrix} D & J & & \\ J & D & J & \\ & \ddots & \ddots & \ddots \\ & & J & D \end{pmatrix}, \tag{2.53}$$

where D and J are the $N \times N$ matrices

$$D = \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & -2 \end{pmatrix}, \qquad J = \begin{pmatrix} 0 & 1 & & \\ 1 & 0 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & 0 \end{pmatrix}.$$

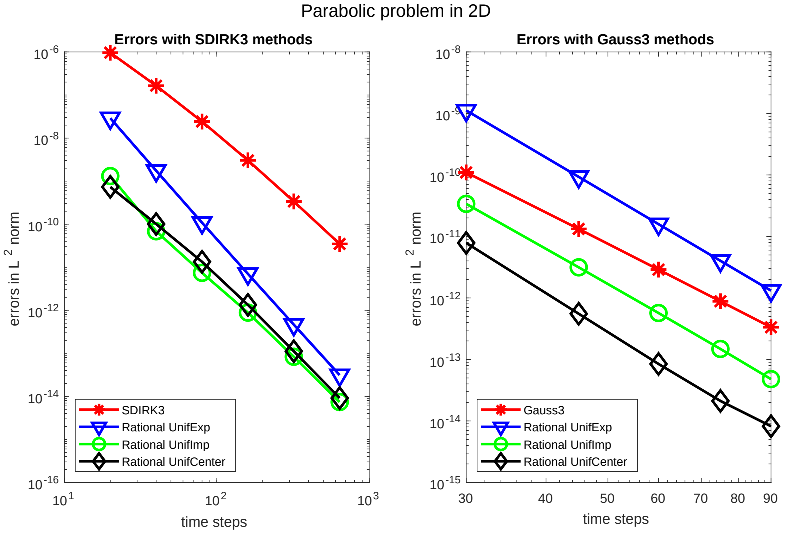
In this case, we solve the systems (2.43) using an iterative algorithm: the gradient conjugate method. This method avoids explicitly constructing the matrix A_h, so that the function values in the mesh are stored in a matrix that mimics the structure of the domain. To solve the system, only a subroutine is needed to apply the system matrix on the function values, which is very efficient due to the sparsity of A_h. We consider again SDIRK3 and Gauss3 for the time integration and, by the same reasoning, the main result in [50] guarantees that $u(t,\cdot,\cdot) \in X_\nu$, $0 \le t \le t_f$, for every $0 < \nu < 1.25$, so the expected orders of convergence of the Gauss3 and SDIRK3 methods are the same as in the previous one-dimensional case. We now consider also the centered nodes (2.41).

Table 2.5: Errors and orders for the parabolic example (2.52) solved with RK and rational Gauss3 method with $h = 1/100$, $t_f = 2$.

	Runge–Kutta		Explicit		Implicit		Centered	
step size	error	order	error	order	error	order	error	order
1.667e-01	1.100e-10	–	1.119e-09	–	3.391e-11	–	7.852e-12	–
1.111e-01	1.325e-11	5.22	9.316e-11	6.13	3.149e-12	5.86	5.538e-13	6.54
8.333e-02	2.893e-12	5.29	1.586e-11	6.15	5.694e-13	5.94	8.454e-14	6.53
6.667e-02	8.831e-13	5.32	4.021e-12	6.15	1.483e-13	6.03	2.117e-14	6.21
5.556e-02	3.344e-13	5.33	1.315e-12	6.13	4.791e-14	6.20	8.214e-15	5.19

We achieve order p with the rational methods as it is shown in Tables 2.4 and 2.5. The centered nodes turn out to have a similar behaviour than the implicit nodes for $p = 4$ but turn out more accurate when $p = 6$, as Remark 2.5 predicts. The rational method with explicit nodes are once again inadequate against Runge–Kutta when $p = 6$ in terms of error against time step size (see also figure 2.3).

Figure 2.3. Errors of the different implementations for the parabolic example (2.50) solved with RK and rational SDIRK3, $tf = 1$, and Gauss3, $t_f = 2$, with h=1/100.



Chapter 3

Rational methods for abstract, semilinear problems without order reduction

> True Laws of Nature cannot be linear.
>
> Albert Einstein

3.1 Introduction

After designing the family of methods (2.29) to time-integrate linear problems, we aim to extend them to nonlinear problems. In particular, in this chapter we consider their extension to semilinear problems of the form

$$\begin{cases} u'(t) = Au(t) + f(t, u(t)), \quad 0 < t < T, \\ u(0) = u_0. \end{cases} \tag{3.1}$$

The key point in fitting these problems into the semigroup framework [31, 39, 53, 66] is the domain of definition of the source term f, which determines the class of admissible nonlinearities. We briefly motivate the necessity of this choice. Typically, one assumes that the source term f is defined on a fractional domain, $f : [0, \infty) \times X_\alpha \to X, \quad 0 \le \alpha < 1$, and satisfies a (global or local) Lipschitz condition of the form

$$\|f(t, u) - f(t, v)\| \le L\|u - v\|_\alpha, \quad \text{for } u, v \in X_\alpha.$$

When this is the case, there are at least two main reasons for requiring f to be defined on a restricted space X_α, rather than on X, when dealing with typical PDEs in L^p spaces. The first, more obvious, reason is that the source term may involve partial derivatives of u. In our context, this requires that the domain of f is a suitable Sobolev space where such derivatives are well defined. The second reason is that, for a function to satisfy a Lipschitz condition in a nontrivial case, it is usually necessary for the functions to be bounded; that is, the domain of f must be a subset of L^∞. According to Sobolev embeddings (1.51), this can be achieved by imposing additional regularity on the considered Sobolev spaces. For this reason, we must adopt different assumptions on the nature of the semigroup and the space X_α depending on the particular nonlinearity under consideration. In Section 3.3, we present some specific cases that illustrate these facts.

Before describing the numerical approximation, it is useful to recall that problem (3.1) admits a unique continuous mild solution $u : [0, T] \to X_\alpha$ under such a suitable Lipschitz condition on f and $u_0 \in X_\alpha$, that is, a function satisfying the variation-of-constants formula

$$u(t) = \mathrm{e}^{tA} u_0 + \int_0^t \mathrm{e}^{(t-s)A} f(s, u(s))\, ds, \quad 0 < t < T. \tag{3.2}$$

The mild solution need not be classical. However, additional regularity of f (for instance, if f is continuously differentiable) ensures that u is a classical solution of (3.1). See the main references [31, 39, 53, 66] for further details.

We now outline the content of this chapter. In Section 3.2.1, we first explain the main hypotheses on problem (3.1), which are closely related to the discussion above on the framework for nonlinearities, and we propose the numerical scheme for their time integration, based on that of Chapter 2. In Section 3.2.2, we state some preliminary results needed to prove the main result of the chapter, Theorem 3.7, which is established in Section 3.2.3. The method for semilinear problems requires a process of precomputing some numerical approximations to the first time steps, similar to what is done in multistep methods; we explain how to carry this out within our framework in Section 3.2.4. Finally, in Section 3.3, we present numerical experiments and discuss practical implementation issues.

3.2 Derivation and analysis of the method

3.2.1 Main assumptions and the proposed method

The goal of this section is then to extend the scheme (2.29) for linear problems (2.1) to semilinear problems of the form (3.1). The construction of the new numerical scheme and the analysis of its convergence require familiarity with the material presented in Chapter 2.

Throughout the chapter, we want to distinguish between the hyperbolic and parabolic case, since for the latter some results can be improved. For reasons that will become clear in the next paragraph, we will refer to the weaker hypothesis of the first case with $\alpha = 0$ and with $\alpha > 0$ to the stronger hypothesis of the second case.

Hypothesis H1.

- If $\alpha = 0$, let $A : D(A) \subset X \to X$ be a densely defined and closed linear operator on X satisfying the resolvent condition

$$\| (\lambda I - A)^{-n} \| \le \frac{M}{(\operatorname{Re}\lambda - \omega)^n} \tag{3.3}$$

on the plane $\{\lambda \in \mathbb{C} : \operatorname{Re}\lambda > \omega\}$ for $M \ge 1$, $\omega \in \mathbb{R}$. That is, $A \in \mathcal{G}(X, M, \omega)$.

- If $\alpha > 0$, let $A : D(A) \subset X \to X$ be a densely defined and closed linear operator on X satisfying the resolvent condition

  $$\| (\lambda I - A)^{-1} \| \leq \frac{M}{|\lambda - \omega|} \tag{3.4}$$

  on the sector $\{\lambda \in \mathbb{C} : 0 \leq |\arg(\lambda - \omega)| \leq \pi - \theta, \lambda \neq \omega\}$ for $M \geq 1$, $\omega \in \mathbb{R}$ and sectorial angle $0 < \theta < \pi/2$, that is, $A \in \mathcal{G}(X, M, \omega, \theta)$.

  Under this assumption, the operator A is the infinitesimal generator of an analytic semigroup $\{\mathrm{e}^{tA}\}_{t\geq 0}$. Fixed $\omega^* > \omega$, the fractional powers of $\tilde{A} = \omega^* I - A$ are well defined. We set $X_\alpha = D(\tilde{A}^\alpha)$ endowed with the graph norm $\| \cdot \|_\alpha$ of $\tilde{A}$. It is well known that X_α is independent of $\omega^* > \omega$ and that changing $\omega^* > \omega$ results in an equivalent norm. In addition to (1.2), we now also have the estimate

  $$\|t^\alpha \tilde{A}^\alpha \mathrm{e}^{tA}\| \leq M\mathrm{e}^{t\omega}. \tag{3.5}$$

The class of nonlinearities f allowed in this setting depends on the nature of the semigroup $\{\mathrm{e}^{tA}\}_{t\geq 0}$.

Hypothesis H2.

- If $\alpha = 0$ and $\{\mathrm{e}^{tA}\}_{t\geq 0}$ is just a $\mathcal{C}_0$ semigroup, we assume that $f : [0,T] \times X \to X$ is locally Lipschitz continuous. Thus, there exists a real number L such that

  $$\|f(t,\xi) - f(t,\eta)\| \leq L\|\xi - \eta\| \tag{3.6}$$

  for all $t \in [0,T]$ and $\max(\|\xi\|, \|\eta\|) \leq R$.

- If $\alpha > 0$ and $\{\mathrm{e}^{tA}\}_{t\geq 0}$ is analytic, we can afford stronger nonlinearities and we assume that $f : [0,T] \times X_\alpha \to X$ is locally Lipschitz. Thus, there exists a real number L such that

  $$\|f(t,\xi) - f(t,\eta)\| \leq L\|\xi - \eta\|_\alpha \tag{3.7}$$

  for all $t \in [0,T]$ and $\max(\|\xi\|_\alpha, \|\eta\|_\alpha) \leq R$.

We note that for the convergence proofs in the chapter, it is sufficient that (3.6) and (3.7) holds in a strip along the exact solution. Although, by simplicity, we assume that f is locally Lipschitz.

The methods designed in [10] part from a rational mapping

$$r(z) = r_\infty + \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell,j} (1 - w_\ell z)^{-j}, \tag{3.8}$$

that may be the stability function of a Runge–Kutta method of order p. Again, when the semigroup is analytic, we can consider a wider class of rational mappings as a starting point.

Let $r(z)$ be a rational function that approximates the exponential e^z with order $p \geq 1$, that is, it satisfies (1.34).

Hypothesis H3.

- If $\alpha = 0$, we assume that r is A-acceptable.
- If $\alpha > 0$ and $A \in \mathcal{G}\left(X, M, \omega, \theta\right)$, we assume that r is strongly A(ϑ)-acceptable with $\vartheta > \theta$.

Under these assumptions, the theory in Chapter 1 guarantees that there exists a number $\tau_0 > 0$ such that the operator

$$r(\tau A) = r_\infty \, I + \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell,j} \left(I - \tau w_\ell A\right)^{-j} \tag{3.9}$$

is well defined for every $0 < \tau < \tau_0$.

We part from the scheme for linear problems (2.29), that requires evaluating the source term f in the times $t_n + \tau \boldsymbol{c_n} \geq 0$. However, when integrating a semilinear problem we cannot dispose directly of the values of f, since they depend on the solution u that we are trying to approximate. We then propose to choose integer nodes $\boldsymbol{c_n} \in \mathbb{Z}^p$, so that the times $t_n + \tau \boldsymbol{c_n}$ fall on the time grid and we can approximate the source term using the approximate values of the function $f(t_n, u(t_n)) \approx f(t_n, u_n)$.

In what follows we will consider the use of the nodes $\boldsymbol{c}_n = [-p+1, \ldots, 0] \in \mathbb{Z}^p$ or $\boldsymbol{c}_n = [-p+2, \ldots, 1] \in \mathbb{Z}^p$. The first choice requires the use of the previous values $\boldsymbol{U}_n = [u_{n-p+1}, \ldots, u_n]$ to compute u_{n+1}, so it is explicit; whereas the second choice requires $\boldsymbol{U}_n = [u_{n-p+2}, \ldots, u_{n+1}]$, and an implicit scheme turns up.

The proposed scheme can be written in a form which is analogous to (2.29),

$$u_{n+1} = r\left(\tau A\right) u_n + \tau E_n(\tau) f\left(t_n + \tau \boldsymbol{c}_n, \boldsymbol{U}_n\right), \quad n \geq p-1. \tag{3.10}$$

Starting values $u_0, u_1, \cdots, u_{p-1}$, must be provided. In Section 4.2.5 we explain how to compute the first values within this framework. In Section 4.3 we discuss in depth the consequences of choosing each of the node possibilities.

3.2.2 Preliminaries: discrete inequalities and regularisation

In this section we state some results which are required to prove the convergence of the scheme (3.10) in the following ones.

The first lemmas are aimed at proving a variant of the discrete Gronwall lemma which is necessary for the proof of the main result of the chapter. The following lemma collects some bounds whose proof is elementary, but which are stated together for the sake of clarity.

Lemma 3.1. Let $1 \leq k \leq n-1$, $p \geq 1$, $m^+(k) = \inf(n-1, k+p-1)$ and $\delta, \alpha \in (0,1)$. Assume that $\tau > 0$ and that $t_m = m\tau$, for $0 \leq m \leq n$. Then the following inequalities hold:

$$\sum_{m=k}^{m^+(k)} t_{n-m}^{-\alpha} \le p^{1+\alpha}\, t_{n-k}^{-\alpha}, \tag{3.11}$$

$$\tau \sum_{m=0}^{n-1} t_{n-m}^{-\alpha} \le \frac{t_n^{1-\alpha}}{1-\alpha}, \tag{3.12}$$

$$\sum_{m=k}^{m^+(k)} \delta^{n-m-1} \le p\, \delta^{1-p} \delta^{n-k-1}, \tag{3.13}$$

$$\sum_{m=1}^{n-1} \delta^{n-m-1} \le \frac{1}{1-\delta}. \tag{3.14}$$

Proof. To prove the first inequality notice that, for $k \le m \le m^+(k)$,

$$\frac{t_{n-m}^{-\alpha}}{t_{n-k}^{-\alpha}} = \frac{(n-k)^\alpha}{(n-m)^\alpha} = \left(1 + \frac{m-k}{n-m}\right)^\alpha \le (1+p-1)^\alpha = p^\alpha,$$

which proves (3.11), since the sum has at most p terms. For the second inequality, notice that

$$\tau \sum_{m=0}^{n-1} t_{n-m}^{-\alpha} = \tau \sum_{m=1}^{n} t_m^{-\alpha} \le \tau^{1-\alpha} \int_0^n \frac{ds}{s^\alpha} \le \frac{t_n^{1-\alpha}}{1-\alpha}.$$

The third one is true since, for $m \le k \le m^+(k)$,

$$\delta^{n-m-1} = \delta^{k-m} \delta^{n-k-1} \le \delta^{1-p} \delta^{n-k-1},$$

and again the sum has at most p terms. The last inequality is just the sum of a geometric series. □

We now state the following lemma, which is a variant of the discrete Gronwall lemma introduced in [55].

Lemma 3.2. Let $\tau > 0$, $N \ge 1$ and $t_n = n\tau$, $0 \le n \le N$. Let ξ_n be a sequence of real positive numbers with $\xi_0 = 0$ and

$$(a)\ \xi_n^p = \sum_{k=n-p+1}^{n} \xi_k \quad \text{or} \quad (b)\ \xi_n^p = \sum_{k=n-p+2}^{n+1} \xi_k, \quad \text{for} \quad p-1 \le n \le N-1. \tag{3.15}$$

Assume that there exist $\alpha \in (0,1)$, $\delta \in [0,1)$ and $K_0, K_1 \ge 0$ such that

$$\max_{0 \le k \le p-1} \xi_k \le K_0 \tag{3.16}$$

and that

$$\xi_{n+p-1} \le K_0 + K_1 \sum_{m=0}^{n-1} \left(\tau\, t_{n-m}^{-\alpha} + \tau^{1-\alpha} \delta^{n-m-1}\right) \xi_{m+p-1}^p, \quad n \ge 0. \tag{3.17}$$

Then, there exists a constant $K \geq 0$ depending on $\gamma, \alpha, T = N\tau, K_1, p, \delta$ such that

$$\xi_n \leq K\, K_0 \qquad \text{for} \quad n \geq p-1. \tag{3.18}$$

Proof. We first assume case (a) in (3.15). Notice that

$$\begin{aligned}\xi_{n+p-1} &\leq K_0 + K_1 \sum_{m=0}^{n-1} \left(\tau\, t_{n-m}^{-\alpha} + \tau^{1-\alpha}\delta^{n-m-1}\right) \sum_{k=m-p+1}^{m} \xi_{k+p-1} \\ &= K_0 + K_1 \sum_{k=-p+1}^{n-1} \sum_{m=m^-(k)}^{m^+(k)} \left(\tau t_{n-m}^{-\alpha} + \tau^{1-\alpha}\delta^{n-m-1}\right) \xi_{k+p-1},\end{aligned}$$

where $m^-(k) = \max\{k, 0\}$ and $m^+(k) = \min\{k+p-1, n-1\}$ are the values that allow the previous sum to be reordered. Now, we use the estimates in Lemma 3.1. For $-p+1 \leq k \leq 0$, $m^-(k) = 0$, and formulas (3.12), (3.14) imply that, for a constant $K > 0$ depending on T, δ, α, it is true that

$$\sum_{m=0}^{m^+(k)} \left(\tau\, t_{n-m}^{-\alpha} + \tau^{1-\alpha}\, \delta^{n-m-1}\right) \xi_{k+p-1} \leq K\, \xi_{k+p-1}.$$

On the other hand, for $1 \leq k \leq n-1$, $m^-(k) = k$, and taking into account (3.11), (3.13),

$$\sum_{m=k}^{m^+(k)} \left(\tau\, t_{n-m}^{-\alpha} + \tau^{1-\alpha}\, \delta^{n-m-1}\right) \xi_{k+p-1} \leq K \left(\tau\, t_{n-k}^{-\alpha} + \tau^{1-\alpha}\delta^{n-k-1}\right) \xi_{k+p-1}.$$

Then, we combine the latter and (3.16) to get that, for $n \geq 0$,

$$\xi_{n+p-1} \leq K\, K_0 + K K_1 \sum_{k=1}^{n-1} \left(\tau\, t_{n-k}^{-\alpha} + \tau^{1-\alpha}\delta^{n-k-1}\right) \xi_{k+p-1}, \tag{3.19}$$

for another constant K. If we consider case (b), we obtain an additional term

$$K\, K_1\, \tau^{1-\alpha}\xi_{n+p-1}$$

in the right hand side. It is clear that, for small enough τ, case (b) may be reduced to formula (3.19). The proof concludes applying Lemma 2.1. in [36]. □

Hereafter, the letter K denotes general positive constants that may depend on the semigroup $(M,\, \omega,\, \alpha)$, the rational method $(C,\, \gamma)$ or the interval $[0, T]$ of integration, but that does not depend on any considered particular solution u, source term f or step size τ.

When A generates an analytic semigroup and we work with functions $u : [0, T] \to X_\alpha$, we expect the numerical approximations to the solution to be in the space X_α, not just in X. Since the nonlinearity f takes values in X, the linear part of the numerical scheme, governed by the operator constructed from the rational function $r\,(\tau A)$, must have some regularisation property that guarantees that the numerical solutions are in X_α. This is what motivates the results with which this section ends.

First of all, we have to recall the estimate (1.42). In accordance with the remarks after Lemma 3.2, we use it in the form

$$\|\,(r\,(\tau A) - r_\infty I)\,x\| \le K\|x\|, \quad x \in X,$$

where K is understood that depends on ω, T. This shows that the linear part of the numerical sheme regularises the solution after any amount of time steps. Moreover, we state in the following lemma that the operators $E(\tau)$ and $E_n(\tau)$ satisfy a similar property.

Lemma 3.3. Let $\alpha \in (0,1)$, $A \in \mathcal{G}(X, M, \omega, \theta)$ and $E(\tau)$, $E_n(\tau)$ be the operators defined in (2.16) and (2.29), respectively. For $0 \le \beta \le \alpha$, it is true that

$$\|E(\tau)\,v\|_\beta \le K\,\tau^{-\beta}\|v\|_\infty, \quad \text{for } v \in \mathcal{C}_{ub}\,([0,\infty), X)\,, \tag{3.20}$$

$$\|E_n(\tau)\,\mathbf{v}\|_\beta \le K\,\tau^{-\beta}\|\mathbf{v}\|_{X^p}, \quad \text{for } \mathbf{v} \in X^p, \tag{3.21}$$

where $\|\cdot\|_{X^p}$ corresponds to the maximum of the norm of each component in X.

Proof. For $A \in \mathcal{G}(X, M, \omega, \theta)$ and $0 < \tau < \tau_0$, formula (1.8) implies that there exists a constant K such that

$$\|\,(I - \tau w_\ell A)^{-n}\,\| \le \frac{M}{|1 - \tau w_\ell\,\omega|^n} \le K.$$

Moreover,

$$A\,(I - \tau w_\ell A)^{-n} = \frac{1}{\tau w_\ell}\left((I - \tau w_\ell A)^{-1} - I\right)(I - \tau w_\ell A)^{-n+1}\,,$$

so we also have that

$$\|A\,(I - \tau w_\ell A)^{-n}\,\| \le \frac{K}{\tau}.$$

These formulas are particular cases of the boundedness of $r(\tau A)$ and (1.41) for the functions $r(z) = (1 - w_\ell\, z)^{-n}$, $1 \le \ell \le k$. The same reasoning proves that

$$\|\,(I - \tau w_\ell\, B)^{-n}\,\| \le 1.$$

Then, the proof concludes since it is true that for the interpolation space $X_\beta = [X, D(A)]_\beta$, $0 \le \beta \le \alpha$,

$$\|\,(I - \tau w_\ell A)^{-1}\,\|_\beta \le \frac{K}{\tau^\beta},$$

and $E(\tau)$, $E_n(\tau)$ are linear combinations of resolvents of A. □

To conclude, we state a lemma which is based on these results that will be useful in the proof of the main theorem.

Lemma 3.4. Let $0 < \alpha < 1$. Under hypotheses H1 and H3, let $\xi_m \in X_\alpha$, $0 \le m \le n$. Then, there exists a positive constant K (that may be different in each case) such that the following estimates hold

$$\left\| \tau \sum_{m=0}^{n-1} r^{n-m-1}(\tau A)\, \xi_m \right\|_\alpha \le K\tau \sum_{m=0}^{n-2} \left(\frac{\|\xi_m\|}{t_{n-m-1}^\alpha} + \gamma^{n-m-1} \|\xi_m\|_\alpha \right) + \tau \|\xi_{n-1}\|_\alpha, \tag{3.22}$$

$$\left\| \tau \sum_{m=0}^{n-1} r^{n-m-1}(\tau A)\, \xi_m \right\|_\alpha \le K \left(\max_{0\le m\le n-2} \|\xi_m\| + \tau \max_{0\le m\le n-1} \|\xi_m\|_\alpha \right). \tag{3.23}$$

Proof. Taking into account the regularization estimate (1.42), the left hand side in (3.22) and (3.23) is bounded by

$$\begin{aligned} &\left\| \tau \sum_{m=0}^{n-1} \left(r^{n-m-1}(\tau A) - r_\infty^{n-m-1} \right) \xi_m \right\|_\alpha + \left\| \tau \sum_{m=0}^{n-1} r_\infty^{n-m-1} \xi_m \right\|_\alpha \\ &\le \tau \sum_{m=0}^{n-2} \frac{K}{t_{n-m-1}^\alpha} \|\xi_m\| + \tau \|\xi_{n-1}\|_\alpha + \tau \sum_{m=0}^{n-2} \gamma^{n-m-1} \|\xi_m\|_\alpha, \end{aligned}$$

which proves (3.22). To prove (3.23), notice that

$$\begin{aligned} \tau \sum_{m=0}^{n-2} \frac{\|\xi_m\|}{t_{n-m-1}^\alpha} &\le \tau^{1-\alpha} \left(\int_0^n \frac{1}{s^\alpha}\, ds \right) \max_{0\le m\le n-2} \|\xi_m\| \\ &\le \frac{\tau^{1-\alpha} n^{1-\alpha}}{1-\alpha} \max_{0\le m\le n-2} \|\xi_m\| \le \frac{T^{1-\alpha}}{1-\alpha} \max_{0\le m\le n-2} \|\xi_m\|, \end{aligned}$$

and

$$\tau \sum_{m=0}^{n-1} \gamma^{n-m-1} \|\xi_m\|_\alpha \le \tau \left(\sum_{m=0}^{\infty} \gamma^m \right) \max_{0\le m\le n-1} \|\xi_m\|_\alpha \le \tau \frac{1}{1-\gamma} \max_{0\le m\le n-1} \|\xi_m\|_\alpha.$$

□

In the case $\alpha = 0$, instead of the above lemma it will be sufficient to use the direct bound

$$\left\| \tau \sum_{m=0}^{n-1} r^{n-m-1}(\tau A)\, \xi_m \right\| \le \tau\, C_s(n) \sum_{m=0}^{n-1} \|\xi_m\|, \tag{3.24}$$

for which no hypothesis on $|r_\infty|$ is required.

3.2.3 Analysis of convergence

Before stating the main theorem of the chapter, we have to prove a previous result. This theorem simply generalises the convergence result of the method for linear problems, Theorem 2.6, to the case where we deal with spaces X_α. The proof is similar to that of that theorem, but now taking into account the regularisation results.

For the rest of the section, assume hypotheses H1, H2 and H3, and let $h : [0, T] \to X$ to be $h(t) = f(t, u(t))$. The linear problem

$$\begin{cases} v'(t) = Av(t) + h(t), \quad 0 < t < T, \\ v(0) = u_0, \end{cases} \tag{3.25}$$

has u as a solution and is now discretized by means of the recurrence

$$v_{n+1} = r(\tau A)\, v_n + \tau E_n(\tau) h(t_n + \tau \boldsymbol{c}_n), \quad n \geq 1, \tag{3.26}$$

for some sequence $\{\boldsymbol{c}_n\}_{n=0}^{N-1}$. We firstly state the version of Proposition 2.2 for the space X_α that we require in the theorem.

Proposition 3.5. Let $A \in \mathcal{G}(X, M, \omega, \theta)$ and G be the operator defined in (2.6). Then, G is the infinitesimal generator of a $\mathcal{C}_0$ semigroup of bounded, linear operators $\left\{\mathrm{e}^{tG}\right\}_{t \geq 0}$ in the space $Z_\alpha = X_\alpha \times \mathcal{C}_{ub}([0, \infty), X)$. The semigroup is the one given by (2.7) and it has growth

$$\|\mathrm{e}^{tG}\|_\alpha \leq M(1 + t^\alpha)\, \mathrm{e}^{\omega^+ t}. \tag{3.27}$$

In particular, $G \in \mathcal{G}(Z_\alpha, M_{\tilde{\omega}}, \tilde{\omega})$ for every $\tilde{\omega} > \omega^+$ and an adequate $M_{\tilde{\omega}} > M$.

Proof. The proof is a direct consequence of the variation-of-constants formula and the regularity estimate (3.5). □

Theorem 3.6. Under the hypotheses of Lemma 3.4, let $u : [0, T] \to X_\alpha$ be the solution of (3.25) to be approximated on the interval $[0, T]$. Assume also that $u \in \mathcal{C}^{p+1}([0, T], X_\alpha)$, $h \in \mathcal{C}^{p+1}([0, T], X)$. If v_n is the numerical approximation to $u(t_n)$ given by (3.26) with constant step size $0 < \tau = T/N < \tau_0$,

$$\|u(t_n) - v_n\|_\alpha \leq K\, \tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty \right), \qquad 0 \leq n \leq N. \tag{3.28}$$

Proof. Since the previous proposition states that S_G is a semigroup in $X_\alpha \times Y$ and the convergence estimate (1.38) guarantees that, for the abstract scheme

$$\bar{v}_{n+1} = r(\tau A)\, \bar{v}_n + \tau E(\tau)\, r^n(\tau B)\, h, \quad n \geq 1, \tag{3.29}$$

we get global error of order p,

$$\|u(t_n) - \bar{v}_n\|_\alpha \leq C\, \tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p+1)}\|_\infty \right). \tag{3.30}$$

Then subtracting (3.29) from (3.26),

$$\begin{aligned} v_{n+1} - \bar{v}_{n+1} &= r(\tau A)(v_n - \bar{v}_n) + \tau \left(E_n(\tau)\, h(t_n + \tau \boldsymbol{c}_n) - E(\tau)\, r^n(\tau B)\, h \right) \\ &= r(\tau A)(v_n - \bar{v}_n) + \tau \left(E_n(\tau)\, h(t_n + \tau \boldsymbol{c}_n) - E(\tau)\, h(t_n + \cdot) \right) \\ &\quad + \tau \left(E(\tau) \left(h(t_n + \cdot) - r^n(\tau B)\, h \right) \right), \quad \text{for } n \geq 0, \end{aligned}$$

with $v_0 = \bar{v}_0$. Then, by the variation-of-constants formula, the error $\|v_n - \bar{v}_n\|_\alpha$ is bounded by the sum of the two terms

$$\begin{aligned}(I) &= \left\| \tau \sum_{m=0}^{n-1} r^{n-m-1}(\tau A)\left(E_m(\tau)\, h(t_m + \tau \boldsymbol{c_m}) - E(\tau)\, h(t_m + \cdot)\right)\right\|_\alpha ,\\ (II) &= \left\| \tau \sum_{m=0}^{n-1} r^{n-m-1}(\tau A)\, E(\tau)\left(h(t_m + \cdot) - r^m(\tau B)\, h\right)\right\|_\alpha .\end{aligned}$$

The proof is concluded taking into account (3.23), the regularisation estimates (3.21) and (3.20), and the approximation estimate (2.21). Notice that in this case $C_s(n) = O(1)$, because the semigroup is analytic. □

Now, we are in position to state and prove the main result.

Theorem 3.7. For $0 \le \alpha < 1$, let $u : [0,T] \to X_\alpha$ be the solution of (3.1) to be approximated in the interval $[0,T]$. Let us assume hypotheses H1, H2 and H3 and also that $u \in \mathcal{C}^{p+1}([0,T], X_\alpha)$ and $h \in \mathcal{C}^{p+1}([0,T], X)$. If u_n is the numerical approximation to $u(t_n)$ given by (3.10) with constant step size $0 < \tau = T/N < \tau_0$, and $u_0, \cdots, u_{p-1} \in X_\alpha$ are starting values satisfying

$$\|u(t_n) - u_n\|_\alpha \le C_0\, \tau^p, \qquad 0 \le n \le p-1, \tag{3.31}$$

then,

$$\|u(t_n) - u_n\|_\alpha \le K\, C_s(n)\, \tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty\right), \qquad 0 \le n \le N. \tag{3.32}$$

Proof. Along the proof we denote $\mathbf{f}_n = f(t_n + \tau \boldsymbol{c}_n, \mathbf{U}_n)$, $\mathbf{h}_n = h(t_n + \tau \boldsymbol{c}_n)$ and $e_n = \|u(t_n) - u_n\|_\alpha$, for $0 \le n \le N$, and

$$e_n^p = \begin{cases} \displaystyle\sum_{k=n-p+1}^{n} e_k, & \text{if } \boldsymbol{c}_n = [-p+1, \cdots, 0], \\ \displaystyle\sum_{k=n-p+2}^{n+1} e_k, & \text{if } \boldsymbol{c}_n = [-p+2, \cdots, 1], \end{cases} \tag{3.33}$$

for $p-1 \le n \le N$. We recall (3.10) and (3.26) to get, for $0 \le n \le N$,

$$u_{n+p} - v_{n+p} = r(\tau A)\left(u_{n+p-1} - v_{n+p-1}\right) + \tau E_{n+p-1}(\tau)\left(\mathbf{f}_{n+p-1} - \mathbf{h}_{n+p-1}\right). \tag{3.34}$$

By the discrete variation-of-constants formula,

$$\begin{aligned} u_{n+p-1} - v_{n+p-1} =& r(\tau A)^n \left(u_{p-1} - v_{p-1}\right) \\ &+ \tau \sum_{m=0}^{n-1} r(\tau A)^{n-m-1} E_{m+p-1}(\tau)\left(\mathbf{f}_{m+p-1} - \mathbf{h}_{m+p-1}\right), \end{aligned} \tag{3.35}$$

for $0 \leq n \leq N-p+1$. We use (3.21) and the Lipschitz property of f to get

$$\|E_{m+p-1}(\tau)\,(\mathbf{f}_{m+p-1}-\mathbf{h}_{m+p-1})\,\| \leq K\,L\,e^p_{m+p-1}, \qquad 0 \leq m \leq N-p,$$

and

$$\|E_{m+p-1}(\tau)\,(\mathbf{f}_{m+p-1}-\mathbf{h}_{m+p-1})\,\|_\alpha \leq \tau^{-\alpha}\,K\,L\,e^p_{m+p-1}, \qquad 0 \leq m \leq N-p.$$

Then, we bound the sum in (3.35) combining the previous estimates together with (3.22). On the other hand, the first term in (3.35) is bounded using (1.37), (3.28) and (3.31), giving rise to

$$\begin{aligned}\|u_{n+p-1}-v_{n+p-1}\|_\alpha \leq M\,C\,C_s(n)\,\tau^p\left(\|u^{(p+1)}\|_{\alpha,\infty}+\|h^{(p)}\|_\infty+\|h^{(p+1)}\|_\infty\right)\\ +K\,L\,C_s(n)\sum_{m=0}^{n-1}\left(\tau\,t^{-\alpha}_{n-m}+\tau^{1-\alpha}\,\gamma^{n-m-1}\right)\,e^p_{m+p-1},\end{aligned}$$

for some other constants C and K. Notice that, due to (3.24), if $\alpha = 0$ the term $\tau^{1-\alpha}\,\gamma^{n-m-1}$ is unnecessary whereas if $\alpha > 0$ then $C_s(n) = O(1)$. Finally, to bound the global error we combine the above estimate with Theorem 3.6 to get

$$\begin{aligned}e_{n+p-1} \leq & M\,C\,C_s(n)\,\tau^p\,\left(\|u^{(p+1)}\|_{\alpha,\infty}+\|h^{(p)}\|_\infty+\|h^{(p+1)}\|_\infty\right)\\ &+K\,L\,C_s(n)\sum_{m=0}^{n-1}\left(\tau\,t^{-\alpha}_{n-m}+\tau^{1-\alpha}\,\gamma^{n-m-1}\right)\,e^p_{m+p-1},\end{aligned}$$

and the proof concludes by using the version of discrete Gronwall lemma in Lemma 3.2. Notice that the hypothesis (3.31) is fully taken into account in this step. □

3.2.4 Starting values

The scheme which has been presented requires evaluating the source term at each time step. To do so, previously computed approximated values u_n can be used to evaluate f at $t = t_n$. Even so, we require some starting values $u_0, u_1, \ldots, u_{p-1}$ to compute the first step and start the recurrence process.

One first possibility is just to use an auxiliary method to compute the starting values. However, in this context it is natural to look for values $u_0, \ldots, u_{p-1}$ that satisfy the implicit scheme

$$u_{n+1} = r\,(\tau A)\,u_n + \tau E_n(\tau)\,f\,(t_n+\tau\boldsymbol{c}_n, \mathbf{U}_n)\,, \quad 0 \leq n \leq p-2, \tag{3.36}$$

where in the first steps $\boldsymbol{c}_n$ are such that $t_n + \tau\boldsymbol{c}_n = [0, \tau, \ldots, (p-1)\tau]^T$ and $\mathbf{U}_n = [u_0, \ldots, u_{p-1}]^T$ for $0 \leq n \leq p-2$. It is then necessary to show that the system (3.36) has a unique solution that approximates the values $u(t_n)$, $1 \leq n \leq p-1$, within the adequate order. To see this, we rewrite the system as a fixed point equation in X_α^{p-1}

$$U^* = \mathcal{N}\,(U^*)\,, \tag{3.37}$$

where $\mathcal{N} : X_\alpha^{p-1} \to X_\alpha^{p-1}$ is the function defined by

$$\mathbf{U} = \begin{pmatrix} u_1 \\ \vdots \\ \\ u_{p-1} \end{pmatrix} \mapsto \mathcal{N}(\mathbf{U}) = \begin{pmatrix} \tilde{u}_1 \\ \vdots \\ \\ \tilde{u}_{p-1} \end{pmatrix}$$

$$= \begin{pmatrix} r\,(\tau A)\, u_0 + \tau E_n(\tau)\, f\left(t_0 + \tau \boldsymbol{c_0}, [u_0, \mathbf{U}^T]^T\right) \\ r\,(\tau A)\, \tilde{u}_1 + \tau E_n(\tau)\, f\left(t_1 + \tau \boldsymbol{c_1}, [u_0, \mathbf{U}^T]^T\right) \\ \vdots \\ r\,(\tau A)\, \tilde{u}_{p-2} + \tau E_n(\tau)\, f\left(t_{p-2} + \tau \boldsymbol{c_{p-2}}, [u_0, \mathbf{U}^T]^T\right) \end{pmatrix}.$$

To show that (3.37) has a unique solution it suffices to see that $\mathcal{N}$ is a contractive mapping for sufficiently small τ. In fact, if $\mathbf{U}, \mathbf{V} \in X_\alpha^{p-1}$, the first component of $\mathcal{N}(\mathbf{U}) - \mathcal{N}(\mathbf{V})$ is

$$\tilde{u}_1 - \tilde{v}_1 = \tau E_n(\tau)\, \left(f\left(t_0 + \tau \boldsymbol{c_0}, [u_0, \mathbf{U}^T]^T\right) - f\left(t_0 + \tau \boldsymbol{c_0}, [u_0, \mathbf{V}^T]^T\right)\right), \tag{3.38}$$

so using (3.21) and the Lipschitz property of f

$$\|\tilde{u}_1 - \tilde{v}_1\|_\alpha \leq K L \tau^{1-\alpha} \|\mathbf{U} - \mathbf{V}\|_\alpha, \tag{3.39}$$

where $\|\mathbf{U}\|_\alpha$ for a vector $\mathbf{U} \in X_\alpha^p$ denotes the maximum of the $\|\cdot\|_\alpha$-norm of its components. Then we assume that $\|\tilde{u}_k - \tilde{v}_k\|_\alpha \leq k\, M^k\, L \tau^{1-\alpha} \|\mathbf{U} - \mathbf{V}\|_\alpha$ and proceed by induction for $1 \leq k \leq p-2$,

$$\begin{aligned} \|\tilde{u}_{k+1} - \tilde{v}_{k+1}\|_\alpha &\leq K \|\tilde{u}_k - \tilde{v}_k\|_\alpha + K L \tau^{1-\alpha} \|\mathbf{U} - \mathbf{V}\|_\alpha \\ &\leq (k+1)\, K^{k+1}\, L\, \tau^{1-\alpha} \|\mathbf{U} - \mathbf{V}\|_\alpha. \end{aligned}$$

Therefore,

$$\|\mathcal{N}(\mathbf{U}) - \mathcal{N}(\mathbf{V})\|_\alpha \leq p\, K^p\, L\, \tau^{1-\alpha} \|\mathbf{U} - \mathbf{V}\|_\alpha, \tag{3.40}$$

and the mapping is contractive for sufficiently small τ. In that case, the contractive mapping theorem guarantees that (3.37) has a unique solution $\mathbf{U} = [u_1 \ldots, u_{p-1}]^T$. To conclude, we show that this fixed point approximates the solution with order p. We set $\mathbf{U}(t_n) = [u_0, u(t_1), \ldots, u(t_{p-1})]^T$ for $0 \leq n \leq p-2$. Under the assumptions of Theorem 3.6, the scheme

$$\bar{u}_{n+1} = r\,(\tau A)\, \bar{u}_n + \tau E_n(\tau) f\left(t_n + \tau \boldsymbol{c}_n, \mathbf{U}(t_n)\right), \quad 0 \leq n \leq p-2, \tag{3.41}$$

is such that

$$\|\bar{u}_n - u(t_n)\|_\alpha \leq C\, K\, \tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty\right), \quad 0 \leq n \leq p-1. \tag{3.42}$$

Then we have, for $0 \leq n \leq p-2$,

$$u_{n+1} - \bar{u}_{n+1} = r\,(\tau A)\,(u_n - \bar{u}_n) + \tau E_n(\tau)\,(f(t_n + \tau \boldsymbol{c}_n, \mathbf{U}_n) - f(t_n + \tau \boldsymbol{c}_n, \mathbf{U}(t_n))),$$

and by the discrete variation-of-constants formula and (3.21), we get that $\|u_n - \bar{u}_n\|_\alpha$ is bounded by

$$\tau \left\| \sum_{m=0}^{n-1} r\,(\tau A)^{n-m-1}\, E_m(\tau A)\,(f(t_m + \tau \boldsymbol{c_m}, \mathbf{U}_m) - f(t_m + \tau \boldsymbol{c_m}, \mathbf{U}(t_m))) \right\|_\alpha \\ \leq pKL\tau^{1-\alpha} \sup_{1\leq k\leq p-1} \|u(t_k) - u_k\|_\alpha .$$

By the triangle inequality, the previous bound and (3.42)

$$\|u_n - u(t_n)\|_\alpha \leq K\,\tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty\right) \\ + pKL\,\tau^{1-\alpha} \sup_{1\leq k\leq p-1} \|u(t_k) - u_k\|_\alpha . \tag{3.43}$$

We can take the supremum in the left hand side and, for small enough τ, and another constant K,

$$\sup_{1\leq k\leq p-1} \|u(t_k) - u_k\|_\alpha \leq K\,\tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty\right). \tag{3.44}$$

In practice, we can compute the initial approximations with an auxiliary method, which in our experiments is Euler implicit method, and then iterate the function $\mathcal{N}$ to obtain initial values within the adequate order.

3.3 Practical implementations and Results

3.3.1 Efficient implementation of the methods

In this section we show numerical results which are obtained with the numerical scheme (3.10) and different choices of the nodes. For the spatial discretization, we use finite difference methods. We take $h > 0$ as discretization parameter and we are led to systems of ODEs

$$\begin{cases} u_h'(t) = A_h u_h(t) + f_h(t, u_h(t)), & t \geq 0, \\ u_h(0) = u_{0,h}. \end{cases} \tag{3.45}$$

The implementation of the scheme (3.10) applied to the latter requires evaluating the operator $r\,(\tau A_h)$. In practice, this means dealing with systems of equations of the form

$$(I - \tau\, w_\ell\, A_h)\, x = y,$$

so it is sufficient to have a routine that solves systems of this form and there is no need to calculate the inverse of the matrix $(I - \tau\, w_\ell\, A_h)$, which tipically has a sparse structure.

After the spatial discretization, scheme (3.10) is applied to the PDE in the discretized form (3.45). Recall the discussion on Section 2.3, where accurate details on the implementation issues of the linear version (2.29) can be found. Such matters are the same in this case by adding the dependence of the source term on the function u and choosing the appropriate nodes. In all examples, the scheme is implemented in three ways:

1. *Explicit mode.* We choose $\boldsymbol{c}_n = [-p+1, \dots, 0]^T \in \mathbb{R}^p$, so $u_{n-p+1}, \dots, u_n$ are used to compute u_{n+1}.

2. *Semiexplicit mode.* First, we take a step with the explicit scheme to have an approximation $\tilde{u}_{n+1}$ to $u(t_{n+1})$. Then, we correct this approximation by taking $\boldsymbol{c}_n = [-p+2, \dots, 1]$ and using $u_{n-p+2}, \dots, \tilde{u}_{n+1}$ to compute u_{n+1}.

3. *Implicit mode.* We set a tolerance TOL and iterate the previous process until two successive iterates $\tilde{u}_{n+1}^{[k]}$ and $\tilde{u}_{n+1}^{[k+1]}$ are such that
$$\|\tilde{u}_{n+1}^{[k]} - \tilde{u}_{n+1}^{[k+1]}\|_\alpha \leq TOL.$$

In all the numerical experiments below, the tolerance to calculate implicitly the starting values or for the iteration in the implicit mode has been $TOL = 10^{-14}$.

In this case we consider the 3-stages SDIRK method and RadauIA3. The 3-stages Gauss method is not strongly A-stable, since its stability function is such that $r_\infty = -1$, so for problems that require $\alpha > 0$ it does not fit in our framework. However, we find that the numerical experiments also work with this method, but we do not include them.

The 3-stages RadauIA method (see, e.g., [38]) is the Runge–Kutta method with the following Butcher tableau

$$\begin{array}{c|ccc}
0 & \frac{1}{9} & \frac{-1-\sqrt{6}}{18} & \frac{-1+\sqrt{6}}{18} \\[2ex]
\frac{6-\sqrt{6}}{10} & \frac{1}{9} & \frac{88+7\sqrt{6}}{360} & \frac{88-43\sqrt{6}}{360} \\[2ex]
\frac{6+\sqrt{6}}{10} & \frac{1}{9} & \frac{88+43\sqrt{6}}{360} & \frac{88-7\sqrt{6}}{360} \\[2ex]
\hline
 & \frac{1}{9} & \frac{16+\sqrt{6}}{360} & \frac{16+\sqrt{6}}{360}
\end{array}.$$

The method has $s = 3$ stages and order of convergence $p = 5$, it is A-stable and it has a rational stability function that can be written in the form
$$r(z) = \frac{r_1}{1 - w_1 z} + \frac{r_2}{1 - w_2 z} + \frac{r_3}{1 - w_3 z},$$
with $r_\ell, w_\ell \in \mathbb{C}$, $\text{Re}(w_\ell) > 0$, for $\ell = 1, 2, 3$. In particular, the method is strongly A-stable. As the method has order $p = 5$, in each step we must fix $\mathbf{c_n} \in \mathbb{R}^5$ and obtain three vectors $\boldsymbol{\gamma_\ell^n} \in \mathbb{R}^5$ by solving the linear systems (2.22). Then, a step of the method is done by computing
$$u_{n+1} = \sum_{\ell=1}^{3} r_\ell \left(I - \tau w_\ell A\right)^{-1} \left(u_n + \tau\, w_\ell\, \boldsymbol{\gamma_\ell^n} \cdot f\left(t_n + \tau \boldsymbol{c_n}, \mathbf{U}_n\right)\right).$$

3.3.2 Numerical experiments

Parabolic problem in 1D

We consider a semilinear parabolic problem in the unit interval with homogeneous boundary conditions,

$$\left\{\begin{array}{rcll} u_t(t,x) & = & u_{xx}(t,x) + \lambda \left(\displaystyle\int_0^1 u(t,x)\,dx \right) u_x + f(t,x), & 0 \le t \le 1, \quad 0 \le x \le 1, \\ u(0,x) & = & u_0(x), & 0 \le x \le 1, \\ u(t,0) & = & 0, & 0 \le t \le 1, \\ u(t,1) & = & 0, & 0 \le t \le 1. \end{array}\right. \tag{3.46}$$

where $f : [0,1] \times X_\alpha \to \mathbb{C}$, $u_0 : [0,1] \to X_\alpha$. In order to fit the problem in our framework, we take $X = L^2[0,1]$ and $A = d^2/dx^2$ with $D(A) = H^2[0,1] \cap H_0^1[0,1]$. We set $\alpha = 1/2$, so that according to (1.52), $X_\alpha = H_0^1[0,1]$ and $\|\cdot\|_\alpha = \|\cdot\|_{H^1(0,1)}$. We adjust the data u_0 and f in such a way that $u(t,x) = x(1-x)\,\mathrm{e}^t$, $0 \le t, x \le 1$, is the solution of the problem. Notice that the source term satisfies the Lipschitz condition, since the Sobolev embedding (1.51) guarantees that $H_0^1[0,1] \hookrightarrow L^\infty[0,1]$. We consider various values of the parameter λ to test how the different implementations of the scheme behave with respect to the stiffness of the source term f. We discretize the problem in space by means of finite differences. To this end, we fix a number J of uniformly distributed nodes $x_j = jh, 1 \le j \le J$, in $(0,1)$, with $h = 1/(J+1)$, the matrix of the semidiscrete system of the spatial discretization is again (2.51), and the corresponding linear systems are solved using the Thomas algorithm. The spatial derivatives u_x and u_{xx} are approximated by using central finite differences and the standard three-point finite difference scheme, respectively; while the integral has been approximated by using the composite Simpson's rule. In this way, since the exact solution is a polynomial of degree two in x, there are no spatial errors.

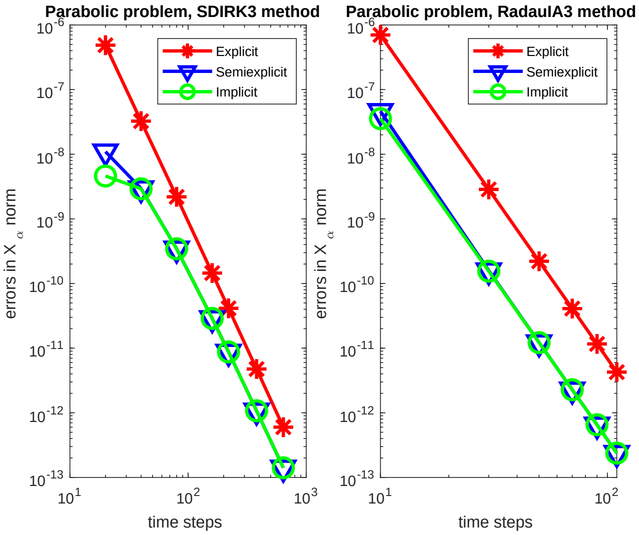


Figure 3.1. Errors in the discrete norm $\|\cdot\|_{H^1(0,1)}$ against number of time steps for the parabolic problem (3.46) with $\lambda = 1$, two different time integrators and $J = 100$.

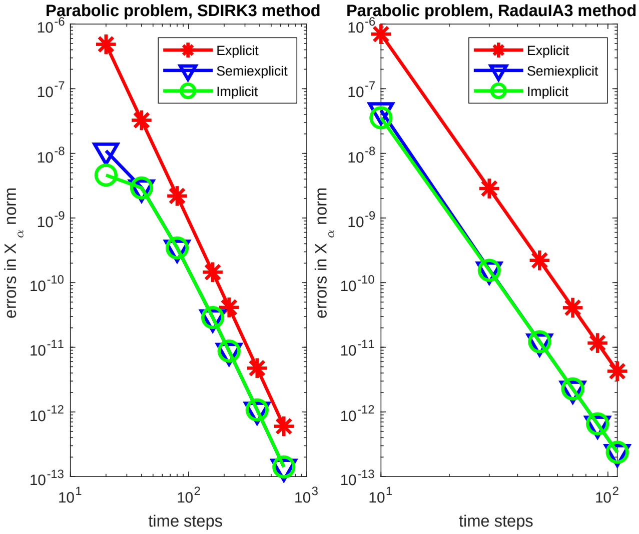


Figure 3.2. Errors in the discrete norm $\|\cdot\|_{H^1(0,1)}$ against number of time steps for the parabolic problem (3.46) with $\lambda = 1$, two different time integrators and $J = 100$.

Table 3.1: Errors and order of convergence in Example 1 with the rational SDIRK3 method.

	Explicit		Semiexplicit		Implicit	
step size	error	order	error	order	error	order
5.000e-02	4.864e-07	–	1.095e-08	–	4.606e-09	–
2.500e-02	3.250e-08	3.90	2.952e-09	1.89	2.919e-09	0.66
1.250e-02	2.192e-09	3.89	3.451e-10	3.10	3.445e-10	3.08
7.692e-03	3.281e-10	3.91	6.201e-11	3.54	6.196e-11	3.53
4.545e-03	4.137e-11	3.94	8.748e-12	3.72	8.745e-12	3.72
2.632e-03	4.757e-12	3.96	1.075e-12	3.84	1.075e-12	3.83
1.563e-03	5.991e-13	3.97	1.410e-13	3.90	1.409e-13	3.90

Table 3.2: Errors and orders of convergence of Example 1, $\lambda = 1$, Radau IA3

	Explicit		Semiexplicit		Implicit	
step size	error	order	error	order	error	order
1.000e-01	6.995e-07	–	4.621e-08	–	3.550e-08	–
3.333e-02	2.858e-09	5.01	1.587e-10	5.16	1.553e-10	4.94
2.000e-02	2.209e-10	5.01	1.222e-11	5.02	1.218e-11	4.98
1.429e-02	4.094e-11	5.01	2.274e-12	5.00	2.274e-12	4.99
1.111e-02	1.163e-11	5.01	6.528e-13	4.97	6.540e-13	4.96
9.091e-03	4.261e-12	5.00	2.378e-13	5.03	2.377e-13	5.04

As time integrators, we use the scheme (3.10) with the rational functions of the Runge–Kutta methods SDIRK3 ($p = 4$) and the 3-stages RadauIA3 ($p = 5$) [38]. Notice that the application of these RK methods does not give its classical order of convergence p, while Tables 3.1 and 3.2 show that the scheme (3.10) does, as it is predicted by Theorem 3.7. Both tables show results taking $\lambda = 1$.

In this case ($\lambda = 1$), it is interesting to note that in both cases the semiexplicit mode involves an improvement of the error by slightly more than one order of magnitude. However, the implicit mode does not practically improve on the semiexplicit mode, so its higher computational cost is not justified. Moreover, if we take the number of systems being solved in the integration as a magnitude for the computational cost, Figures 3.1 and 3.2 suggests that the explicit mode is more efficient than the semiexplicit one, since the computational cost of the latter is twice that of the first for the same number of steps.

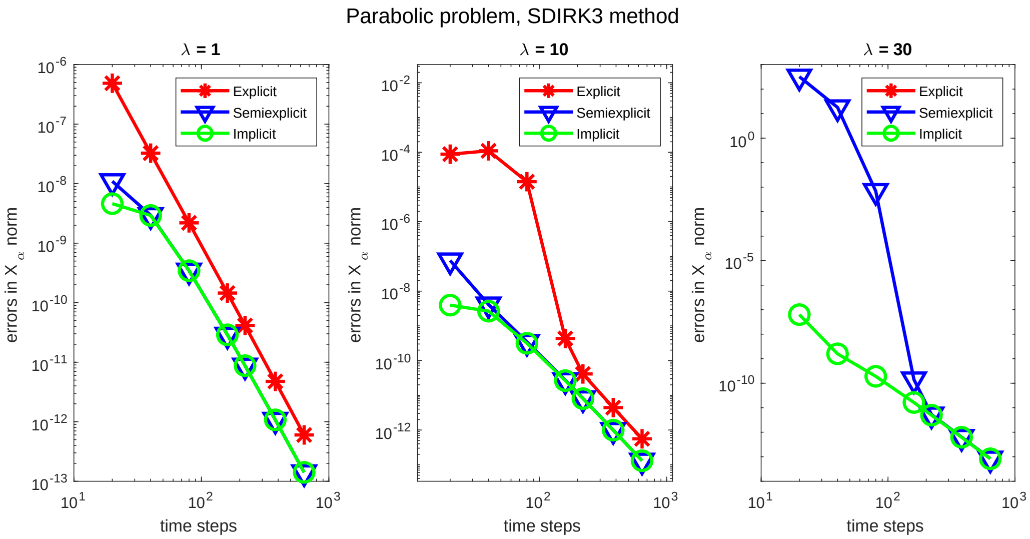


Figure 3.3. Errors in the discrete norm $\|\cdot\|_{H^1(0,1)}$ against number of time steps for the parabolic problem (3.46) with $\lambda = 1, 10, 30$ and the method SDIRK3.

Figures 3.1 and 3.2 show the error in the integration when the parameter λ is modified. We take $\lambda = 1, 10, 30$ for SDIRK3 and $\lambda = 1, 10, 15$ for RadauIA3. With both methods we observe that the explicit method does not perform out well when λ increases, while the implicit one does not almost vary. The semiexplicit has an intermediate behaviour. This is consistent with the well-known sensitivity of explicit methods to stiff problems, whereas implicit methods handle it better due to their greater stability.

Parabolic problem in 2D

We consider a semilinear parabolic problem in the domain $\Omega = (0, 1) \times (0, 1)$ with homogeneous boundary conditions,

$$\left\{ \begin{array}{rcll} u_t(t,x,y) & = & \Delta u(t,x,y) + u^2 + f(t,x,y), & 0 \le t \le 1, \quad (x,y) \in \Omega, \\ u(0,x,y) & = & u_0(x,y), & (x,y) \in \Omega, \\ u(t,x,y) & = & 0, & 0 \le t \le 1, \quad (x,y) \in \partial\Omega. \end{array} \right. \tag{3.47}$$

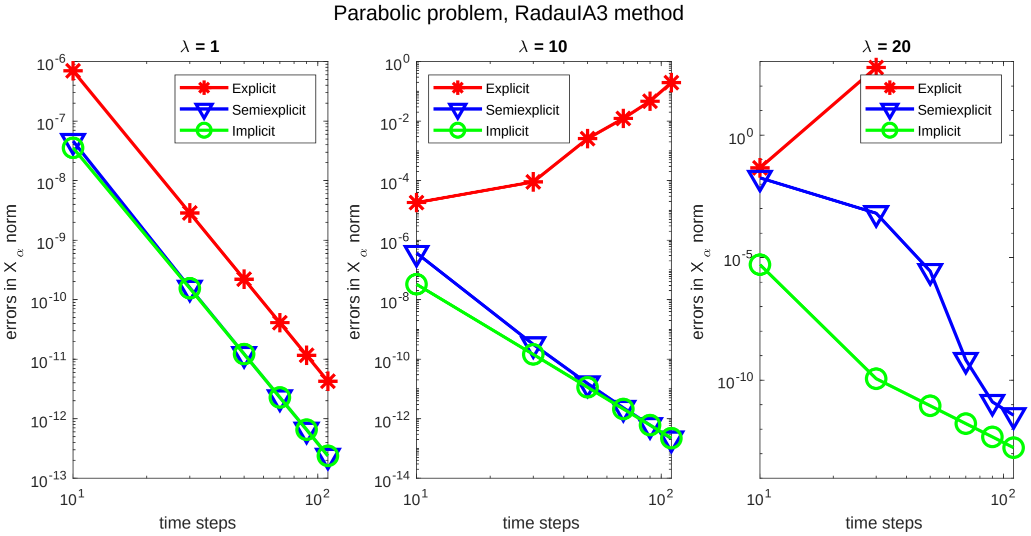


Figure 3.4. Errors in the discrete norm $\|\cdot\|_{H^1(0,1)}$ against number of time steps for the parabolic problem (3.46) with $\lambda = 1, 10, 15$ and the method RadauIA3.

where $f : [0,1] \times \Omega \to \mathbb{C}$, $u_0 : \Omega \to \mathbb{C}$. In order to fit the problem in our framework, we take $X = L^2(\Omega)$, $A = \Delta$ with $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$. We set $\alpha = 3/4$, so that according to (1.52), $X_\alpha = H^{3/2}(\Omega) \cap H_0^1(\Omega)$ and $\|\cdot\|_\alpha = \|\cdot\|_{H^{3/2}(0,1)}$. We adjust the data u_0 and f in such a way that $u(t,x,y) = x(1-x)y(1-y)\,\mathrm{e}^t$, $0 \le t, x \le 1$, is the solution of the problem.

We discretize the problem in space by means of finite differences. To this end, we fix a number J of uniformly distributed nodes $x_j = jh$, $y_k = kh, 1 \le j,k \le J$, in $(0,1)$, with $h = 1/(J+1)$. The spatial derivatives u_{xx} and u_{yy} are approximated by using the standard three-point finite difference scheme, respectively, leading to the matrix (2.53) under the lexicographical order. We solve the corresponding systems using the conjugate gradient method. In this way, since the exact solution is a polynomial of degree two in x and y, there are no spatial errors.

Table 3.3: Errors and orders of convergence of Example 2 with SDIRK3, $J = 50$.

	Explicit		Semiexplicit		Implicit	
step size	error	order	error	order	error	order
2.500e-02	1.402e-08	–	2.047e-10	–	1.791e-10	–
1.250e-02	9.234e-10	3.92	9.012e-11	1.18	8.930e-11	1.00
6.250e-03	6.197e-11	3.90	1.033e-11	3.12	1.031e-11	3.11
3.125e-03	4.095e-12	3.92	8.581e-13	3.59	8.571e-13	3.59
1.563e-03	2.641e-13	3.95	6.128e-14	3.81	6.128e-14	3.81

As time integrators, we use again the scheme (3.10) with the rational functions of the Runge–Kutta methods SDIRK3 ($p = 4$) and 3-stages RadauIA3 ($p = 5$). Notice that the application of these RK methods does not give its classical order of convergence p, while Tables 3.3 and 3.4 show that the scheme (3.10) does, as it is predicted by Theorem 3.7.

Table 3.4: Errors and orders of convergence of Example 2 with RadauIA3, $J = 50$.

	Explicit		Semiexplicit		Implicit	
step size	error	order	error	order	error	order
1.000e-01	5.600e-05	–	2.869e-06	–	3.188e-06	–
5.000e-02	2.322e-06	4.59	1.241e-07	4.53	1.320e-07	4.59
2.500e-02	8.220e-08	4.82	4.511e-09	4.78	4.665e-09	4.82
1.250e-02	2.721e-09	4.92	1.515e-10	4.90	1.542e-10	4.92
6.250e-03	8.746e-11	4.96	5.300e-12	4.84	5.340e-12	4.85

Hyperbolic problem

We consider a semilinear hyperbolic problem in the unit interval with periodic boundary conditions,

$$\left\{\begin{array}{lcll} u_t(t,x) & = & -u_x(t,x) + u - u^3 + f(t,x), & 0 \le t \le 1, \quad 0 \le x \le 1, \\ u(0,x) & = & u_0(x), & 0 \le x \le 1, \\ u(t,0) & = & u(t,1), & 0 \le t \le 1, \end{array}\right. \tag{3.48}$$

where $f : [0,1] \times [0,1] \to X$, $u_0 : [0,1] \to X$. In order to fit the problem in our framework, we take $X = H^1[0,1]$, $A = -d/dx$ with $D(A) = \{u \in H^1[0,1] : u(0) = u(1)\}$ and $\alpha = 0$, so that $\|\cdot\|_\alpha = \|\cdot\|_X = \|\cdot\|_{H^1(0,1)}$. We adjust the data u_0 and f in such a way that $u(t,x) = x^3\,\mathrm{e}^t \sin(\pi x) + (1-\mathrm{e}^t)$, $0 \le t, x \le 1$, is the solution of the problem. We fix again number J of uniformly distributed nodes $x_j = jh, 1 < j \le J$, in $(0,1)$, with $h = 1/J$, that leads to the matrix

$$A_h = \frac{1}{h}\begin{pmatrix} -1 & & & 1 \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}. \tag{3.49}$$

The corresponding systems are solved using a convenient adaptation of the Thomas algorithm, so only $O(N)$ operations and memory usage is required. The spatial derivatives are approximated by upwind finite differences.

Tables 3.5 and 3.6 show the results for the hyperbolic problem with the scheme (3.10) and the rational function of SDIRK3 and 3-stages RadauIA3, respectively. We find the results to be similar to the parabolic problem.

Again, we obtain an improvement on the size of the error after the semiexplicit correction for a fixed step size and we observe that the order of the methods is in agreement with that being predicted by Theorem 3.7.

Figure 3.5 presents the same information in a graphical form. In this problem, since the semiexplicit method is approximately twice as costly per step as the explicit one, the latter proves to be computationally more efficient.

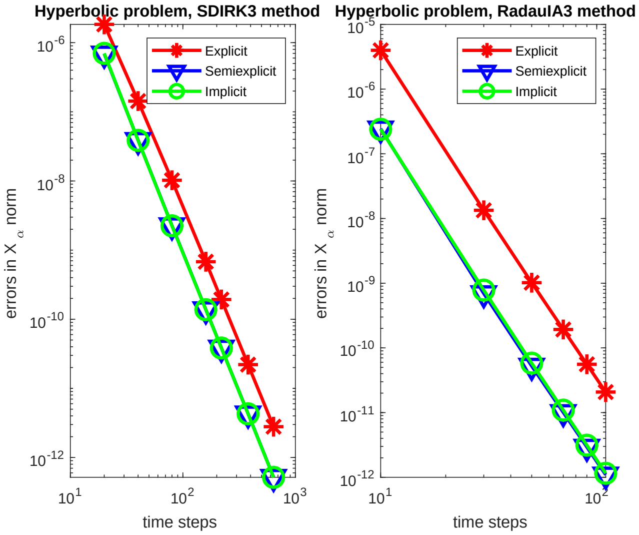


Figure 3.5. Error in the discrete norm associated to $\|\cdot\|_{H^1(0,1)}$ against number of time steps for the hyperbolic problem (3.48) with two different time integrators and $J = 100$.

Table 3.5: Errors and order of convergence in Example 3 with the rational SDIRK3 method, J=100.

	Explicit		Semiexplicit		Implicit	
step size	error	order	error	order	error	order
5.000e-02	1.839e-06	–	6.896e-07	–	7.013e-07	–
2.500e-02	1.430e-07	3.69	3.879e-08	4.15	3.867e-08	4.18
1.250e-02	1.024e-08	3.80	2.266e-09	4.09	2.251e-09	4.10
7.692e-03	1.537e-19	3.91	3.185e-10	4.04	3.168e-10	4.04
4.545e-03	1.929e-10	3.95	3.849e-11	4.02	3.837e-11	4.01
2.632e-03	2.206e-11	3.97	4.299e-12	4.01	4.286e-12	4.01
1.563e-03	2.791e-12	3.97	5.206e-13	4.01	5.200e-13	4.02

Table 3.6: Errors and orders of convergence of Example 3 with RadauIA3, J=100.

	Explicit		Semiexplicit		Implicit	
step size	error	order	error	order	error	order
1.000e-01	3.976e-06	–	2.440e-07	–	2.397e-07	–
3.333e-02	1.339e-08	5.18	6.939e-10	5.34	7.826e-10	5.21
2.000e-02	1.018e-09	5.04	5.255e-11	5.05	5.815e-11	5.09
1.429e-02	1.925e-10	4.95	1.009e-11	4.91	1.092e-11	4.97
1.111e-02	5.582e-11	4.93	2.942e-12	4.90	3.135e-12	4.97
9.091e-03	2.078e-11	4.92	1.093e-12	4.93	1.154e-12	4.98

Chapter 4

Rational methods for abstract, initial boundary value problems without order reduction

> A boundary condition is not merely a constraint; it is part of the problem's identity, shaping the solution space like the banks shape the river.
>
> Anonymous

4.1 Introduction

In this chapter, we consider the semigroup setting introduced in [8, 64], which we briefly describe, in order to study the time integration of abstract, linear initial boundary value problems of the form

$$\begin{cases} u'(t) = Au(t) + f(t), \quad t > 0, \\ u(0) = u_0, \\ \partial u(t) = g(t), \quad t > 0, \end{cases} \tag{4.1}$$

where X and Y are two complex Banach spaces, A and ∂ are two linear operators $A : D(A) \subset X \to X$ and $\partial : D(A) \subset X \to Y$, such that $[A, \partial]^T : D(A) \to X \times Y$ is closed (i.e., $[A, \partial]^T(w) = [Aw, \partial w]^T$), $u_0 \in X$, $f : [0, \infty) \to X$ and $g : [0, \infty) \to Y$. Set

$$D(A_0) = \operatorname{Ker} \partial = \{\, x \in D(A) \,:\, \partial x = 0 \,\}$$

and let $A_0 : D(A_0) \subset X \to X$ be the restriction of A to $D(A_0)$. In the context of the PDEs we are interested in, X is typically an L^p space, as in the previous chapters, and Y is an appropriate Sobolev space well suited to the boundary conditions under consideration. The operator ∂, which may or may not involve derivatives (for instance, for Neumann or Robin boundary conditions versus Dirichlet boundary conditions), defines the boundary conditions.

We assume the following hypothesis along the chapter:

Hypothesis H1. $A_0 \in \mathcal{G}(X, M, \omega)$ that is, A_0 is the infinitesimal generator of a $\mathcal{C}_0$ semigroup of bounded operators in X, with $M \geq 1$ and $\omega \in \mathbb{R}$.

Hypothesis H2. There exists a bounded, linear operator $E : Y \to X$, such that

$$Ev \in D(A) \text{ and } \partial Ev = v, \qquad v \in Y,$$

and $AE : Y \to X$ is also bounded. In the PDE's context, E is a linear extension operator that, given v in the Sobolev space Y, provides an element $w = Ev$ in the Sobolev space $D(A)$ such that $\partial w = v$. The existence of E is studied in the Extension Theory [51].

It turns out that [64], for $\text{Re}(\lambda) > \omega$ and $v \in Y$, the eigenvalue problem

$$\left\{ \begin{array}{rcl} Au &=& \lambda u, \\ \partial u &=& v, \end{array} \right. \tag{4.2}$$

admits a unique solution, denoted by $u = K(\lambda)v$, that belongs to $D(A)$. This gives rise to bounded, linear operators $K(\lambda) : Y \to D(A)$, $\text{Re}(\lambda) > \omega$, than can be expressed, independently of the extension operator E, as

$$K(\lambda) = [I - (\lambda - A_0)^{-1}(\lambda - A)]E, \quad \text{Re}(\lambda) > \omega. \tag{4.3}$$

From this representation we readily obtain

$$\|K(\lambda)\| \leq \|E\| + \frac{|\lambda| \, \|E\| + \|AE\|}{\text{Re}(\lambda) - \omega}, \quad \text{Re}(\lambda) > \omega.$$

Remark 4.1. Under Hypotheses H1 and H2, the operator $[A, \partial]^T : D(A) \subset X \to X \times Y$ is automatically closed. This can be deduced from the existence of the boundary correction operators $K(\lambda)$ for $\text{Re}(\lambda) > \omega$, defined in (4.3), which facilitate the decomposition $I = (\lambda - A_0)^{-1}(\lambda - A) + K(\lambda)\partial$ on $D(A)$. This identity, in turn, implies that the graph norm of $[A, \partial]^T$ is equivalent to the norm of X on $D(A)$, ensuring the fact that the operator is closed. Thus, stating Hypotheses H1 and H2 suffices.

The chapter is mainly focused on the IBVPs (4.1) with data $u_0 \in D(A)$, $f : [0, +\infty) \to X$ continuous, and $g : [0, +\infty) \to Y$ of class $\mathcal{C}^1$. In case this problem admits a genuine solution $u : [0, +\infty) \to D(A)$, we can justify that,

$$g'(t) = \partial w'(t) = \partial(Aw(t) + f(t)), \qquad t \geq 0. \tag{4.4}$$

In particular, for $t = 0$, the data must satisfy the so-called natural compatibility condition

$$g'(0) = \partial(Au_0 + f(0)), \tag{4.5}$$

which is a necessary requirement in order to obtain a $\mathcal{C}^1$ solution of (4.1). Higher regularity imposes additional natural compatibility conditions on the data. It has been proved [64] that condition (4.5) is not only necessary but also sufficient; that is, under this assumption, (4.1) admits a unique genuine solution u, and consequently, (4.4) remains valid for all $t \geq 0$. Moreover, the IBVP (4.1) is well posed [64] in the sense that the genuine solution u depends continuously on the initial and boundary data, namely $u_0 \in D(A)$, $f \in \mathcal{C}([0,+\infty),X)$, and $g \in \mathcal{C}^1([0,+\infty),Y)$. Here, the L^1 norm is considered for f, and the total variation norm for g. This well-posedness allows us to extend the notion of solution to generalized solutions of (4.1) within the framework of

$$X \times L^1_{\text{loc}}([0,+\infty),X) \times BV_{\text{loc}}([0,+\infty),Y).$$

It is clear that, for $\text{Re}(\lambda) > \omega$, the solution of (4.1) can be expressed as

$$u(t) = \tilde{u}(t) + K(\lambda)g(t), \qquad t \geq 0,$$

where, since $\partial\tilde{u}(t) = 0$, $t \geq 0$, and $AK(\lambda) = \lambda K(\lambda)$, the term $\tilde{u} : [0,\infty) \to X$ solves

$$\left\{ \begin{array}{lcl} \tilde{u}'(t) & = & A_0\tilde{u}(t) + f(t) + K(\lambda)(\lambda g(t) - g'(t)), \qquad t \geq 0, \\ \tilde{u}(0) & = & u_0 - K(\lambda)g(0). \end{array} \right.$$

so that, when $\omega < 0$, the choice $\lambda = 0$, results in the simpler IVP

$$\left\{ \begin{array}{lcl} \tilde{u}'(t) & = & A_0\tilde{u}(t) + f(t) - K(0)g'(t), \qquad t \geq 0, \\ \tilde{u}(0) & = & u_0 - K(0)g(0). \end{array} \right. \tag{4.6}$$

Alternatively, the IBVP (4.1) is reduced to an IVP by using an available extension operator E, instead of $K(0)$, but then we must use the source term $f + AEg - Eg'$.

Let us notice that we can easily reduce the problem to the situation $\omega < 0$. To this end, we just fix $\alpha > \omega$ and write the solution of u of the IBVP (4.1) in the form

$$\tilde{u}(t) = \mathrm{e}^{t\alpha}\tilde{u}_\alpha(t), \qquad t \geq 0,$$

where $\tilde{u}_\alpha : [0,\infty) \to X$ is the solution of the conjugate problem

$$\left\{ \begin{array}{lcl} \tilde{u}_\alpha'(t) & = & (A - \alpha I)\,\tilde{u}_\alpha(t) + \mathrm{e}^{-t\alpha}f(t), \quad t \geq t_0, \\ \tilde{u}_\alpha(t_0) & = & \tilde{u}_0, \\ \partial\tilde{u}_\alpha(t) & = & \mathrm{e}^{-t\alpha}g(t), \quad t \geq t_0. \end{array} \right.$$

The infinitesimal generator of the above problem is $A_0 - \alpha I$, whose spectral abscissa is $\omega - \alpha < 0$. This reduction simplifies the presentation and it is also interesting from the numerical point of view. On the one side, no restriction on the used step size in required and, on the other, the basic estimates [18] involving rational approximations are simpler when $\omega \leq 0$. However, the main reason why we introduce this simplification for the first time in this chapter is to work with the operators $K(0)$ and simplify the evolution equation in $\tilde{u}$. Thus, in the rest of the paper, we will make the simplifying assumption that $\omega < 0$. We will mainly focus on the time discretization of IVPs with the format (4.6).

In Section 4.2.3 we treat the full discretization of (4.1) and, to deal with the spatial consistency, we will introduce two Banach spaces $(W, \|.\|_W)$ and $(H, \|\cdot\|_H)$, continuously embedded in X and Y, such that

$$W \subset D(A) \quad \text{and} \quad K(0)H \subset W. \tag{4.7}$$

Under these conditions, the restriction of $K(0)$, resp. ∂, to H, resp. to W, are continuous from H to W, resp., from W to H. In the common applications, W and H are Sobolev spaces, with norms finer than those of X and Y. Their orders must account for the nature of ∂, that may contain derivatives or not, and the Extension Theory is the tool to provide the existence of an extension operator $E : H \to W$ such that $\partial E = I$. Once E is obtained, shows that $K(0)\partial$ leaves W invariant, as soon as $A_0^{-1}(AW) \subset W$, that is the usual situation.

It is important to remark that, for boundary data $g : [0, +\infty) \to Y$ of class $\mathcal{C}^1$, the IBVP (4.1) also makes sense in the framework of W. Actually, (4.1) admits a genuine solution $u : [0, +\infty) \to W$ if, and only if, the natural compatibility condition (4.5) is satisfied, in which case (4.4) remains valid. In fact, if we set $W_0 = W \cap \operatorname{Ker}(\partial)$, endowed with norm $\|\cdot\|_{W_0}$ induced by $\|\cdot\|_W$, (4.6) is a standard, non-homogeneous, linear problem in W_0. Besides, for $t \geq 0$, it is clear that

$$\|u(t)\|_{W_0} \leq \|w(t)\|_W + \|K(0)g(t)\| \leq \|w(t)\|_W + \|K(0)\partial\|_{W\to W}\|w(t)\|_W, \tag{4.8}$$

an important estimate when considering the spatial discretization. Notice also that this estimate implies that the restriction of $S_{A_0}(t)$, ≥ 0, to W_0, form a $\mathcal{C}_0$-semigroup on W_0, whose infinitesimal generator is the restriction of A_0 to W_0.

4.2 Derivation and analysis of the method

4.2.1 Motivation of the scheme and previous results

We repeat the process we followed to construct the scheme for linear problems in Chapter 2. Given a linear and bounded operator $K : Y \to X$, we embed the non-homogeneous IVP (4.6) into an enlarged, homogeneous one. To this end, we consider both the semigroup of translations on $\mathcal{C}_{ub}([0,\infty), X)$ and on $\mathcal{C}_{ub}([0,\infty), Y)$, that are denoted by $T_B(t)$ and $T_{B_Y}(t)$, $t \geq 0$, respectively. Accordingly, their generators are denoted by B and B_Y. A direct consequence of Proposition 2.1 is that $D(B^m) = \mathcal{C}^m_{ub}([0,\infty), X)$, $D(B^m_Y) = \mathcal{C}^m_{ub}([0,\infty), Y)$, for $m \geq 0$.

Let $L : D(B) \to \mathcal{C}_{ub}([0,\infty), X)$ and $L_Y : \mathcal{C}_{ub}([0,\infty), Y) \to Y$ be the delta operators

$$Lf = f(0),\ f \in \mathcal{C}_{ub}([0,\infty), X), \qquad L_Y g = g(0),\ g \in \mathcal{C}_{ub}([0,\infty), Y).$$

The product $\widehat{Z} = X \times \mathcal{C}_{ub}([0,\infty), X) \times \mathcal{C}^1_{ub}([0,\infty), Y)$, endowed with the norm

$$\|[u, f, g]^T\| = \|u\| + \|f\|_\infty + \max\{\|g\|_\infty, \|B_Y g\|_\infty\}, \qquad [u, f, g]^T \in \widehat{Z},$$

is a Banach space. On the domain $D(\widehat{G}) = D(A_0) \times D(B) \times D(B_Y^2) \subset \widehat{Z}$, let us define the operator $\widehat{G} : D(\widehat{G}) \subset \widehat{Z} \to \widehat{Z}$ by

$$\widehat{G} = \begin{pmatrix} A_0 & L & -KL_Y B_Y \\ 0 & B & 0 \\ 0 & 0 & B_Y \end{pmatrix},$$

and consider the linear, homogeneous IVP on $\widehat{Z}$

$$\left\{\begin{array}{rcll} U'(t) & = & \widehat{G}U(t) & t \geq 0, \\ U(0) & = & U_0 \in D(\widehat{G}). & \end{array}\right. \tag{4.9}$$

Writing $U(0) = [\tilde{u}_0, f, g]^T$, $U(t) = [\tilde{u}(t), \phi(t), \psi(t)]^T \in \widehat{Z}$, $t \geq 0$, the last two components trivially yield

$$\phi(t) = \mathrm{e}^{Bt}(t)f = f(t + \cdot), \qquad \psi(t) = \mathrm{e}^{B_Y t}g = g(t + \cdot),$$

while the first one fits into the equation

$$\tilde{u}'(t) = A_0\tilde{u}(t) + f(t) - Kg'(t), \qquad t \geq 0.$$

These remarks show that (4.9) admits a unique genuine solution that, using the variation-of-constant formula, can be represented as $U(t) = \mathrm{e}^{\widehat{G}t}U(0)$, $t \geq 0$, where $\mathrm{e}^{\widehat{G}t} : \widehat{Z} \to \widehat{Z}$, $t \geq 0$, is the linear operator

$$\mathrm{e}^{\widehat{G}t} = \begin{pmatrix} \mathrm{e}^{A_0 t} & \int_0^t \mathrm{e}^{A_0(t-s)} L \mathrm{e}^{Bs} \cdot \mathrm{d}s & -\int_0^t \mathrm{e}^{A_0(t-s)} K L_Y B_Y \, \mathrm{e}^{B_Y s} \cdot \mathrm{d}s \\ 0 & \mathrm{e}^{Bs} & 0 \\ 0 & 0 & \mathrm{e}^{B_Y s} \end{pmatrix}.$$

Clearly, $\left\{\mathrm{e}^{\widehat{G}t}\right\}_{t \geq 0}$ is a strongly continuous semigroup on $\widehat{Z}$ and

$$\|\mathrm{e}^{\widehat{G}t}\| \leq M(1 + 2t). \tag{4.10}$$

It will be useful to introduce the family of seminorms $|||\cdot|||_{m,t}$ in the product space $\mathcal{C}^m_{ub}([0,\infty), X) \times \mathcal{C}^m_{ub}([0,\infty), X) \times \mathcal{C}^{m+1}_{ub}([0,\infty), Y)$ given, for $[v, \phi, \psi]^T$ in such a product, by the expression

$$|||[v, \phi, \psi]^T|||_{m,t} = \|v\|_{m,t} + \|\phi\|_{m,t} + \|\psi\|_{m+1,t} \tag{4.11}$$

($\|\cdot\|_{m,t}$ is defined in (2.3)).

Let us stress that, for $m \geq 1$, $U \in D(\widehat{G}^m)$ if and only if the map $t \mapsto e^{t\widehat{G}}U_0$, $t \geq 0$, belongs to $C_{\mathrm{ub}}([0, +\infty), X)$. This is equivalent to have $u \in \mathcal{C}^m_{ub}([0,\infty), X)$, $f \in \mathcal{C}^m_{ub}([0,\infty), X)$ and $g \in \mathcal{C}^{m+1}_{ub}([0,\infty), Y)$. Therefore, under such smoothness conditions on u, f and g, the solution U of (4.9) takes values in $D(\widehat{G}^m)$ and, since $\widehat{G}^m U(s) = U^{(m)}(s)$, for $s \geq 0$, it turns out that

$$\|\widehat{G}^m U(s)\| = \|U^{(m)}(s)\| \leq |||U|||_{m,t}, \qquad 0 \leq s \leq t. \tag{4.12}$$

Notice that, for $m \geq 1$, $D(\widehat{G}^m) \subset D(A_0^m) \times \mathcal{C}^m_{ub}([0,\infty), X) \times \mathcal{C}^{m+1}_{ub}([0,\infty), Y)$, but the equality is true only for $m = 1$. Actually, to guarantee that $U \in \mathcal{C}^m_{ub}([0,\infty), X)$ we must imposes several compatibility conditions on the initial data.

After introducing the notation, we follow a similar approach to that of Chapter 2 to construct the methods; that is, we apply the rational method to the enlarged, homogeneous problem. Since we assume that $\omega < 0$, the operational calculus results on Chapter 1 guarantee that the operators

$$r(\tau\widehat{G}) = r_\infty I + \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell j}(I - \tau w_\ell \widehat{G})^{-j}, \tag{4.13}$$

are well defined and are uniformly bounded for $\tau > 0$.

Given an initial value $U_0 \in \widehat{Z}$ and $\tau > 0$, the recurrence

$$U_{n+1} = r(\tau \widehat{G}) U_n, \qquad n \geq 1, \tag{4.14}$$

defines the numerical approximation $U_n \in \widehat{Z}$ to $U(t_n)$, at $t_n = n\tau$, by means of the rational method based on $r(z)$. Let us explore the form of $r(\tau G)$.

Proposition 4.2. For $\mathrm{Re}(\lambda) > 0$, we claim that

$$(I - \lambda \widehat{G})^{-j} = \begin{pmatrix} (I-\lambda A_0)^{-j} & \lambda Q_{0,j}(\lambda) & -Q_{1,j}(\lambda) \\ 0 & (I-\lambda B)^{-j} & 0 \\ 0 & 0 & (I-\lambda B_Y)^{-j} \end{pmatrix},$$

for $j \geq 1$, where

$$\left\{ \begin{array}{lcl} Q_{0,j}(\lambda) & = & \sum_{i=1}^{j} (I-\lambda A_0)^{-j+i-1} L (I-\lambda B)^{-i}, \\ Q_{1,j}(\lambda) & = & \sum_{i=1}^{j} (I-\lambda A_0)^{-j+i-1} K L_Y \lambda B_Y (I-\lambda B_Y)^{-i}. \end{array} \right. \tag{4.15}$$

Proof. Notice that for $j = 1$, let us set

$$(I - \lambda \widehat{G})^{-1} = \begin{pmatrix} (I-\lambda A_0)^{-1} & \lambda Q_{0,1}(\lambda) & -Q_{1,1}(\lambda) \\ 0 & (I-\lambda B)^{-1} & 0 \\ 0 & 0 & (I-\lambda B_Y)^{-1} \end{pmatrix}, \qquad j \geq 1$$

for suitable operators

$$Q_{0,1}(\lambda) = (I-\lambda A_0)^{-1} L (I-\lambda B)^{-1}, \quad Q_{1,1}(\lambda) = (I-\lambda A_0)^{-1} K L_Y \lambda B (I-\lambda B)^{-1}.$$

By induction, using the variations-of-constant formula, it is straightforward to get the result for $j \geq 1$. □

This result, used in (4.14) with $\lambda = \tau w_\ell$, $1 \leq \ell \leq k$, $1 \leq j \leq m_\ell$, readily yields

$$r(\tau \widehat{G}) = \begin{pmatrix} r(\tau A_0) & E(\tau) & -F(\tau) \\ 0 & r(\tau B) & 0 \\ 0 & 0 & r(\tau B_Y) \end{pmatrix}, \tag{4.16}$$

where

$$E(\tau) = \tau \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell j} w_\ell \sum_{i=1}^{j} (I - \tau w_\ell A_0)^{-j+i-1} L (I - \tau w_\ell B)^{-i}, \tag{4.17}$$

$$F(\tau) = \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell j} w_\ell \sum_{i=1}^{j} (I - \tau w_\ell A_0)^{-j+i-1} K(0) L_Y \tau B_Y (I - \tau w_\ell B_Y)^{-i}. \tag{4.18}$$

In summary, for a given initial datum $U_0 = [\tilde{u}_0, f, g]^T \in \widehat{Z}$, the rational method (4.14) generates approximations $U_n = [\tilde{u}_n, f_n, g_n]^T$, $n \geq 0$, with $f_n = r(\tau B)^n f$, $g_n = r(\tau B_Y)^n g$ and the first components are provided by the recurrence

$$\tilde{u}_{n+1} = r(\tau A_0)\tilde{u}_n + E(\tau) r(\tau B)^n f - F(\tau) r(\tau B_Y)^n g, \qquad n \geq 0, \tag{4.19}$$

Thus, as in the previous chapter, one step requires solving $s := \sum_{\ell=1}^{k} m_\ell$ linear systems with the different operators $(I - \tau w_\ell A_0)$, $1 \le \ell \le k$, something that we can assume to be numerically affordable within some tolerance; however, as we commented, it also requires solving linear systems with the operators $(I - \tau w_\ell B)^{-1}$ and $(I - \tau w_\ell B_Y)^{-1}$, a difficulty we will avoid again by using Lemma 2.4.

4.2.2 Analysis of convergence

We are mainly interested in the consistency. In the rest of the section we will assume that $\tilde{u} \in \mathcal{C}_{ub}^{p+1}([0,\infty), X)$, $f \in \mathcal{C}_{ub}^{p+1}([0,\infty), X)$ and $g \in \mathcal{C}_{ub}^{p+2}([0,\infty), Y)$. Formula (1.38), applied with $n = 1$, shows that for $U_0 \in D(\widehat{G}^{p+1})$,

$$\|\mathrm{e}^{\widehat{G}\tau} U_0 - r(\tau\widehat{G})U_0\| \le C_{el} M \tau^{p+1} \|\widehat{G}^{p+1} U_0\|,$$

where the constant K_{el} depends exclusively on $r(z)$. This result, applied to $\mathrm{e}^{\widehat{G}t_n}U_0 \in D(\widehat{G}^{p+1})$, $n \ge 0$, and recalling (4.10), leads to

$$\|\mathrm{e}^{\widehat{G}t_{n+1}} U_0 - r(\tau\widehat{G})\mathrm{e}^{\widehat{G}t_n} U_0\| \le C_c M^2 \tau^{p+1}(1 + 2t_n)\|\widehat{G}^{p+1} U_0\|.$$

Recalling (4.11), we conclude that for smooth solutions, we have

$$\|\mathrm{e}^{\widehat{G}t_{n+1}} U_0 - r(\tau G)\mathrm{e}^{\widehat{G}t_n} U_0\| \le C_{el} M^2 \tau^{p+1}(1 + 2t_n)\, |||[u, f, g]^T|||_{p+1,t_n}. \tag{4.20}$$

Notice that the first component $\rho_n \in X$ of the local error $\mathrm{e}^{\widehat{G}t_{n+1}} U_0 - r(\tau\widehat{G})\mathrm{e}^{\widehat{G}t_n} U_0$, $n \ge 0$, satisfies

$$\tilde{u}(t_{n+1}) = r(\tau A_0)u(t_n) + E(\tau)\mathrm{e}^{t_n B} f - F(\tau)\mathrm{e}^{t_n B_Y} g + \rho_n, \tag{4.21}$$

and (4.20) implies that

$$\|\rho_n\| \le C_{el} M^2 \tau^{p+1}(1 + 2t_n)\, |||[u, f, g]^T|||_{p+1,t_n}, \qquad n \ge 0. \tag{4.22}$$

As we mentioned, the practical difficulty of the rational method (4.14) is evaluating the different expressions $L(I - \tau w_\ell B)^{-i}$ and $L_Y \tau B_Y (I - \tau w_\ell B_Y)^{-i}$, $1 \le \ell \le k$, $1 \le i \le m_\ell$, occurring in (4.17) and (4.18). We have already suggested a way to overcome it by using Lemma 2.4. The evaluations of $r(\tau B)$ and $r(\tau B_Y)$ will be substituted by that of the operators $\mathrm{e}^{\tau B}$ and $\mathrm{e}^{\tau B_Y}$, which are simply translations of the corresponding functions.

Given $\tau > 0$, we start by selecting two compact sets $\mathcal{K}_0 \subset \mathbb{R}_d^p$ and $\mathcal{K}_1 \subset \mathbb{R}_d^{p+1}$ and two arbitrary sequences $\boldsymbol{c}_n \in \mathcal{K}_0$, $\boldsymbol{d}_n \in \mathcal{K}_1$, $n \ge 0$, with the condition that $t_n + \tau\boldsymbol{c}_n$ and $t_n + \tau\boldsymbol{d}_n$ have non-negative components. Thus, in principle, we will have as many versions of the method we suggest as many possible selections of the compact sets and sequences. In Section 4.3 we discuss different choices of nodes.

Let us fix $1 \le \ell \le k$ and $1 \le i \le m_\ell$. For every $n \ge 0$, we solve the Vandermonde system (2.22) corresponding to the nodes $\boldsymbol{c}_n$, resp. $\boldsymbol{d}_n$, and the rational mappings $(1 - w_\ell z)^{-i}$, resp. $z(1 - w_\ell z)^{-i}$. This yields vectors $\boldsymbol{\gamma}_{n,\ell,i} \in \mathbb{C}^p$ and $\boldsymbol{\eta}_{n,\ell,i} \in \mathbb{C}^{p+1}$ such that, by Lemma 2.4, satisfy

$$\|L(I - \tau w_\ell B)^{-i} f(t_n + \cdot) - \boldsymbol{\gamma}_{n,\ell,i}^T \cdot f(t_n + \tau\boldsymbol{c}_n)\| \le \kappa\tau^{p+1}\|f^{(p+1)}\|_\infty$$

and

$$\|L_Y \tau B_Y(I-\tau w_\ell B_Y)^{-i} g(t_n+\cdot) - \boldsymbol{\eta}_{n,\ell,i}^T \cdot g(t_n+\tau \boldsymbol{c}_n)\| \leq \kappa \tau^{p+2} \|g^{(p+2)}\|_\infty,$$

for some $\kappa = \kappa(\mathcal{K}_0, \mathcal{K}_1) > 0$, uniformly in $\tau > 0$. These facts suggests to consider, instead of (4.17) and (4.18), the bounded operators $E_n(\tau) : \mathcal{C}_{ub}([0,\infty), X) \to X$ and $F_n(\tau) : \mathcal{C}_{ub}([0,\infty), Y) \to X$ defined by

$$E_n(\tau) f = \tau \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell j} w_\ell \sum_{i=1}^{j} (I-\tau w_\ell A_0)^{-j+i-1} \boldsymbol{\gamma}_{n,l,i}^T \cdot f(t_n+\tau \boldsymbol{c}_n), \tag{4.23}$$

$$F_n(\tau) g = \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell j} w_\ell \sum_{i=1}^{j} (I-\tau w_\ell A_0)^{-j+i-1} K \boldsymbol{\eta}_{n,\ell,i}^T \cdot g(t_n+\tau \boldsymbol{d}_n). \tag{4.24}$$

At this point it is clear that

$$\tau \|E(\tau) f - E_n(\tau) f\| \leq \kappa C_{aux} M \tau^{p+1} \|f^{(p)}\|_\infty, \quad f \in \mathcal{C}_{ub}^p([0,\infty), X), \tag{4.25}$$

and

$$\|F(\tau) g - F_n(\tau) g\| \leq \kappa C_{aux} M \|K\| \tau^{p+1} \|g^{(p+1)}\|_\infty, \quad g \in \mathcal{C}_{ub}^{p+1}([0,\infty), X), \tag{4.26}$$

where C_{aux} is a constant depending only on r.

Let $\tilde{u} \in \mathcal{C}([0,\infty), X)$ be the solution of (4.6) with data $u_0 \in X$, $f \in \mathcal{C}_{ub}([0,\infty), X)$, $g \in \mathcal{C}_{ub}^1([0,\infty), Y)$. We propose the recurrence

$$\bar{u}_{n+1} = r(\tau A_0) \bar{u}_n + E_n(\tau) f(t_n+\cdot) - F_n(\tau) g(t_n+\cdot), \qquad n \geq 0, \tag{4.27}$$

with $\bar{u}_0 = u_0$, as the numerical procedure to time integrate (4.6). This procedure avoids the practical difficulty of solving systems like (4.14), avoids the use of $r(\tau B)$ and $r(\tau B_Y)$, and it only requires evaluations of f and g.

The analysis of the convergence of (4.27) is rather simple, once all these concepts have been introduced. It is similar to that in Chapter 2, which is carried out by the classical approach of stability and consistency. Recall (1.37), where A is to be replaced by A_0 under the notation adopted in this chapter.

Lemma 4.3. Let $\delta_n \in X$, $n \geq 0$, be a sequence of perturbations and let $\bar{u}_n^* \in X$, $n \geq 0$, be the solution of the perturbed recurrence

$$\bar{u}_{n+1}^* = r(\tau A_0) \bar{u}_n^* + E_n(\tau) f(t_n+\cdot) - F_n(\tau) g(t_n+\cdot) + \delta_n \qquad n \geq 0,$$

with initial value $\bar{u}_0^* = u_0 + \delta_0$. Then

$$\|\bar{u}_n^* - \bar{u}_n\| \leq C_s(n) \sum_{j=0}^{n-1} \|\delta_j\|, \qquad n \geq 0.$$

Proof. Since the difference $\bar{u}_n^* - \bar{u}_n$, $n \geq 0$, satisfies

$$\bar{u}_{n+1}^* - \bar{u}_{n+1} = r(\tau A_0)(\bar{u}_n^* - \bar{u}_n) + \delta_n, \qquad n \geq 0,$$

the proof is trivial (note that $C_s(n) \geq 1$). □

The local error of method, at time level $n \geq 0$, is now the residual $\bar{\rho}_n \in X$ satisfying

$$\tilde{u}(t_{n+1}) = r(\tau A_0)\tilde{u}(t_n) + E_n(\tau)\mathrm{e}^{t_n B} f - F_n(\tau)\mathrm{e}^{t_n B_Y} g + \bar{\rho}_n. \tag{4.28}$$

The consistency is provided by the next lemma.

Lemma 4.4. Let us assume that $\tilde{u} \in \mathcal{C}^{p+1}([0,\infty), X)$, $f \in \mathcal{C}^{p+1}([0,\infty), X)$ and $g \in \mathcal{C}^{p+2}([0,\infty), Y)$. Then the corresponding local errors $\bar{\rho}_n$, $n \geq 0$, of the method (4.27) satisfy

$$\|\bar{\rho}_n\| \leq C_{c,n}\tau^{p+1}(1+2t_n)|||[u,f,g]^T|||_{p+1,t_n},$$

where $C_{c,n} = M^2 C_{el} + \kappa_n M C_{aux}(1+\|K\|)$.

Proof. We will also use the first component of the local error ρ_n, $n \geq 0$, given in (4.21), of the rational discretization (4.27). By subtracting (4.28) from (4.21), we obtain

$$\rho_n = \bar{\rho}_n - (E(\tau) - E_n(\tau))\mathrm{e}^{t_n B} f + (F(\tau) - F_n(\tau))\mathrm{e}^{t_n B_Y} g, \quad n \geq 0,$$

and the proof concludes by using (4.22), (4.25) and (4.26). □

The convergence is now a plain consequence of the lemmas. We will assume, by simplicity, $\bar{\kappa} = \sup_{n\geq 0} \kappa_n < +\infty$. This is obvious when the sequences $\{\mathbf{c}_n\}_{n=0}^{\infty}$ and $\{\mathbf{d}_n\}_{n=0}^{\infty}$ are eventually constant. Set $C_c = M^2 C_{el} + M C_{aux}(1+\|K\|)$. General choices of the auxiliary sequences can equally be considered just by using $\bar{\kappa}_0 = \kappa_0$ and $\bar{\kappa}_n = \max_{0\leq j\leq n} \kappa_n$, for $n \geq 1$.

Theorem 4.5. Assume that the solution $\tilde{u}$ of (4.6) belongs to $\mathcal{C}^{(p+1)}([0,\infty), X)$, that $f \in \mathcal{C}^{(p+1)}([0,\infty), X)$ and that $g \in \mathcal{C}^{(p+2)}([0,\infty), Y)$. Then the approximations $\bar{u}_n$, $n \geq 1$, generated by a given version of the method (4.27) satisfy

$$\|\tilde{u}(t_n) - \bar{u}_n\| \leq C_c \tau^p t_n (1+2t_n)|||[\tilde{u},f,g]^T|||_{p+1,t_n}.$$

Proof. Since $\tilde{u}(t_n)$, $n \geq 0$, fits into the recurrence

$$\tilde{u}(t_{n+1}) = r(\tau A_0) + E_n(t_n) f - F_n(\tau) g + \bar{\rho}_n,$$

by Lemma 4.3 we have

$$\|\tilde{u}(t_n) - \bar{u}_n\| \leq n\, C_s(n) \max_{0\leq j\leq n-1} \|\bar{\rho}_n\|, \qquad n \geq 1,$$

and the proof concludes by using Lemma 4.4. □

Let us point out that, by the well-known Lax–Richtmyer Theorem, and since the stability of (4.27) depends only on that of $r(\tau A_0)$, it turns out that the method (4.27) converges, without any order of convergence, for data in $X \times L^1_{loc}([0,+\infty), X) \times BV_{loc}([0,+\infty), Y)$.

Concerning the IBVP (4.1), we just propose

$$u_n = K(0)g(t_n) + \bar{u}_n, \qquad n \geq 0, \tag{4.29}$$

where $\bar{u}_n \in X$ is the approximation to the solution u of the IVP (4.6) with $K = K(0)$, provided by the method (4.27).

Since $\tilde{u} = u - K(0)g$, it turns out that $\tilde{u} \in \mathcal{C}^{p+1}_{ub}([0,\infty), X)$ and that

$$|||[\tilde{u}, f, g]^T|||_{p+1,t+n} \leq |||[u, f, g]^T|||_{p+1,t+n} + \|K(0)\| \|g\|_{p+1,t_n}, \tag{4.30}$$

and then Theorem 4.5 shows that, for $n \geq 0$,

$$\|u_n - u(t_n)\| \leq MC_c\tau^p t_n(1+2t_n)\left\{|||[u, f, g]^T|||_{p+1,t_n} \|K(0)\| \|g\|_{p+1,t_n}\right\},$$

that bounds the error in terms of the smoothness of the solution u and that of f and g.

Remark 4.6. A natural choice for $\boldsymbol{c}_n$, $n \geq 0$, is one fulfilling that the last $p-1$ components of $t_n + \tau_n \boldsymbol{c}_n$ are the first $p-1$ components of $t_{n+1} + \tau_{n+1}\boldsymbol{c}_{n+1}$ at least for $n \geq n_0$, for a certain $n_0 \leq p$ (recall (2.39 - 2.41)). The same argument applies for $\boldsymbol{d}_n$ and p of its $p+1$ components.

For instance, we can choose $\boldsymbol{c}_0 = [0, 1, \ldots, p-1]$, $\boldsymbol{c}_n = -n + \boldsymbol{c}_0$, for $1 \leq n \leq n_0$ and $\boldsymbol{c}_n = \boldsymbol{c}_{n_0}$, for $n > n_0$. For $\boldsymbol{d}_n$, we can proceed in the same way for $\boldsymbol{d}_n$, just by changing $p-1$ by p. In this way, instead of the p evaluations of f and the $p+1$ evaluations of Kg per step, needed for arbitrary choices, we pass to perform just one evaluation of f and g, for $n \geq n_0$. If n_0 is roughly p or $p+1$, the first evaluations are used for the first steps, and globally we need one evaluation of f and g. Moreover, recalling Remark 2.5, a good choice is to deal with sequences such that, for $n \geq n_0$, t_n is the mean point of both the nodes $t_n + \tau\boldsymbol{c}_n$ and $t_n + \tau\boldsymbol{d}_n$. This requires, to use half-integer numbers and different evaluation points for f and for Kg, and both are harmless features.

If we do not insist on using these natural choices, we will pay the price of performing more evaluations, but we are free to choose the auxiliary nodes and improve the value of κ.

Remark 4.7. The approach we suggest can be easily extended to deal with higher order derivatives in the source term, that is, to consider time discretizations of problems with the format

$$u'(t) = A_0 u(t) + \sum_{j=0}^{M} K_j f_j^{(j)}(t), \qquad t \geq 0,$$

where $K_j : Y_j \to X$ are linear, bounded operators, acting on different Banach spaces Y_j, $0 \leq j \leq m$, and the nonhomogeneous terms are mappings $f_j \in \mathcal{C}^j_{ub}([0,\infty), Y_j)$, $0 \leq j \leq m$. In particular, $m = 2$ would be of interest in the context of wave equations.

Remark 4.8. Given a sequence of step sizes τ_n, $n \geq 0$, fulfilling the requirements in Proposition 2.2.3, the suggested method (4.27) becomes

$$\bar{u}_{n+1} = r(\tau_n A_0)\bar{u}_n + E_n(\tau_n)f(t_n + \cdot) - F_n(\tau_n)g(t_n + \cdot), \qquad n \geq 0.$$

and in Theorem 4.5 we must change $t_n\tau^p$ by $\sum_{j=0}^{n-1} \tau_j^{p+1}$ in order to prove the same result.

4.2.3 Full discretization

In this section we consider the full discretization of (4.1). We will consider the extensively used method of lines, which consists of discretizing firstly in space and then in time. The framework introduced here does not rely on any particular discretization. We adopt a general spatial discretization setting, while the discussion focuses on standard finite difference, finite element, and spectral methods. We thus arrive at the following standard abstract framework.

Let $0 < h \leq h_0$ denote the parameter governing the spatial discretization. All the operators that we consider are going to be linear. Associated with each h-value, we introduce two Banach spaces X_h and Y_h of finite dimension and two operators $P_h : X \to X_h$ and $Q_h : Y \to Y_h$. Their norms, as well as the norms of associated operators are, by default, denoted by $\| \cdot \|_h$. The norms $\| \cdot \|_h$ reflect the ones of X_h and Y_h and it is well-understood that, for $w \in X$, $P_h w \in X_h$ contains enough information so as to provide an approximation of $w \in W$. The same idea is behind $Q_h v$, for $v \in Y$. For instance, in finite elements, P_h and Q_h are L^2 projectors (considered even for $X = L^q$, $q \neq 2$) and, in finite differences, they are sampling operators (or concentrated averages, for $q < +\infty$).

The connection to IBVP (4.1) is given by:

(a) An onto operator $\partial_h : X_h \to Y_h$. We set $X_{h,0} = \operatorname{Ker}(\partial_h)$.

(b) Two operators $A_h : X_h \to X_h$ and $A_{h,0} : X_{h,0} \to X_{h,0}$ such that $A_h = A_{h,0}$ on $X_{h,0}$. The semigroup generated by $A_{h,0}$ on $X_{h,0}$ is denoted by $S_{A_{h,0}}(t)$, $t \geq 0$.

(c) An operator $K_h : Y_h \to X_h$.

(d) An operator $P_{h,0} : X \to X_{h,0}$ such that

$$\|P_h - P_{h,0}\|_{W_0 \to X_h} \leq M_P \|\partial_h P_h\|_{W_0 \to X_h},$$

and another operator $K_{h,0} : Y_h \to X_{h,0}$. If $X_h \hookrightarrow X$ so that $P_{h,0}$ is also defined in X_h, a usual condition in finite elements, then one can substitute $K_{h,0}$ by $K_{h,0} Q_h g = P_{h,0}(K_h Q_h g)$ and it is not necessary to include a new operator.

As the notation suggests, A_h, $A_{h,0}$, $\partial_h P_h$, $K_h Q_h$ and $K_{h,0} Q_h$ try to fit with $P_h A$, $P_{h,0} A_0$, $Q_h \partial$, $P_h K$ and $P_{h,0} K$.

The idea is that $P_{h,0} w$ contains specific information to reconstruct $w \in \operatorname{Ker}\partial$. In the standard methods, for boundary conditions other than the Dirichlet ones, it turns out that $\partial w = 0$ does not imply that $\partial_h P_h w = 0$, for $w \in X_h$, this is why it is important to introduce $P_{h,0}$.

The consistency refers to a couple of Banach spaces $(W, \| \cdot \|_Z)$, with $W \subset D(A)$ and $(H, \| \cdot \|_H)$, with $H \subset Y$ that are continuously embedded in $D(A)$ and Y and satisfy (4.7). We set $W_0 = \operatorname{Ker}(\partial) \cap W$, endowed with the norm $\| \cdot \|_{W_0}$, that is the one induced by $\| \cdot \|_W$.

To reflect the fact that the norm of W and H are finer that those X and Y, we must consider new, adequate norms $\| \cdot \|_{h,W}$ on X_h and $\| \cdot \|_{h,H}$ on Y_h. The spaces $(W_h, \| \cdot \|_{h,H})$

and $(H_h, \|\cdot\|_{H_h})$ are denoted W_h and H_h, i.e., $W_h = X_h$ and $H_h = Y_h$, but endowed with different norms.

Set

$$L_h = \max\{\|P_h\|_h, \|P_{h,0}\|_h, \|Q_h\|_h, \|\partial_h\|_{W_h \to Y_h}, \|Q_h\|_{H \to H_h}\},$$

and

$$M_h = \sup_{t \geq 0} \|S_{A_{h,0}}(t)\|.$$

The ideal stability hypotheses are

(S1) $M_L := \sup_{0<h\leq h_0} L_h < +\infty$.

(S2) $M_s := \sup_{0<h\leq h_0} M_h < +\infty$

However, frequently L_h and, particularly, M_h exhibit a weak growth, for instance $M_h = O(|\ln h|)$, as $h \to 0+$ (this is why we have singled out M_h). By simplicity, we will assume (S.1) and (S.2), keeping in mind that the final convergence estimate is valid by using L_h and M_h.

Consistency is expressed in terms of an infinitesimal $\epsilon : (0, h_0] \to (0, +\infty)$ that measures the quality of the convergence. Typically $\epsilon(h) = C\, h^m$, for some order $m > 0$. Tracing the value of all the involved constants, by following their proofs, is an impossible task that leads to pessimistic and huge values for a general situation. In the absence of particular properties, we just try to catch the order and later, if possible, to use other techniques to get the leading constant for a particular problem. With this idea in mind, we will simplify the statements and proofs by expressing the different infinitesimals in the form $O(\epsilon(h))$, understanding that the hidden constants are bounded by a common value, uniformly for $0 < h \leq h_0$.

Consistency assumes that for $0 < h \leq h_0$:

(C1) $\|A_{h,0}P_{h,0} - P_{h,0}A_0\|_{W_0 \to X_h} \leq O(\epsilon(h))$.

(C2) $\|\partial_h P_h - Q_h \partial\|_{W \to Y_h} \leq O(\epsilon(h))$.

(C3) $\|P_h K(0) - K_h Q_h\|_{H \to X_h} \leq O(\epsilon(h))$.

(C4) $\|P_{h,0} K(0) - K_{h,0} Q_h\|_{H \to X_{h,0}} \leq O(\epsilon(h))$.

Is is important to note that, while (C1) is standard for IVPs, when dealing with IBVPs, it is natural to require that both the boundary condition and problem (4.2) can be discretized accurately for data in the consistency class, that is (C3), and that the solution depends continuously on the data, that is the reason of (C2). If $X_h \hookrightarrow X$ so that $K_{h,0}$ can be omitted and $P_{h,0}$ is just the projection from X to X_h, then (C4) is a plain consequence of (d) and (C3).

Let u be the solution of the IBVP (4.1). Henceforth, we suppose that the additional conditions $u \in \mathcal{C}^1([0, +\infty), W)$, $f \in \mathcal{C}_{ub}([0, +\infty), X)$ and $g \in \mathcal{C}^1_{ub}([0, +\infty), H)$ and set $u = \tilde{u} + K(0)g$, where $\tilde{u}$ solves (4.6). We have already remarked that $\tilde{u}$, due to the natural compatibility condition, not only takes values in W_0 but is also the solution of (4.6), that turns out to be an IVP in W_0.

Given $t \geq 0$, while for $K(0)g(t)$ the goal is to approximate $P_h K(0)g(t)$, that brings the information to reconstruct $K(0)g(t)$, for $u(t) \in W_0$ it is rather to approximate $P_{h,0}u(t)$, that codifies the information to reconstruct $u(t)$. With these goals in mind, we just propose $K_h Q_h g(t)$ as the spatial discretization of $K(0)g(t)$, and $\tilde{u}_h(t)$, where $\tilde{u}_h$ is the solution of the IVP

$$\left\{ \begin{array}{lcl} \tilde{u}_h'(t) & = & A_{h,0}\tilde{u}_h(t) + P_{h,0}f(t) - K_{h,0}Q_h g'(t)), \quad t > 0, \\ \tilde{u}_h(0) & = & P_{h,0}u_0 - K_{h,0}Q_h g(0), \\ \partial_h \tilde{u}_h(t) & = & 0, \quad t \geq 0, \end{array} \right. \tag{4.31}$$

as the h-disretization of u. The proposed spatial discretization of u is $u_h = K_h Q_h g + \tilde{u}_h$, due to the consistency hypothesis (C3).

We stress that (4.31) is an IVP in $X_{h,0}$ indeed (the boundary condition is thus redundant). In contrast, notice that the related problem with source term $P_h f(t) - P_h K(0) g'(t)$ results in an IVP in X_h, not in $X_{h,0}$, no matter that its solution takes values in $X_{h,0}$.

In consonance with (2.3), we set

$$\|w\|_{1,t,W} = \max_{0 \leq j \leq 1} \max_{0 \leq s \leq t} \|w^{(j)}(s)\|_W, \quad t \geq 0.$$

We are now in a position to prove the next result.

Theorem 4.9. Assume that the solution w of the IBVP (4.1) belongs to $\mathcal{C}^1_{ub}([0,+\infty), W)$, $f \in \mathcal{C}_{ub}([0,+\infty), X)$ and $g \in \mathcal{C}^1_{ub}([0,+\infty), Z)$. Then, the error of the full semidiscrete approximation $e_h = P_{h,0}\tilde{u} - \tilde{u}_h$ can be estimated by

$$\|e_h(t)\|_h \leq M_s t O(\epsilon(h)) \left(\|u\|_{0,t} + \|g\|_{1,t,H}\right), \qquad t \geq 0. \quad 0 < h \leq h_0. \tag{4.32}$$

Proof. The stability and consistency of (4.31) are carried out by a well know approach, based on the variation-of-constants formula. Plugging $P_{h,0}$ in (4.6) and taking the difference with (4.31) shows that the error $e_h = P_{h,0}\tilde{u} - \tilde{u}u_h$ fits into the IVP

$$\left\{ \begin{array}{lcl} e_h'(t) & = & A_{h,0}\, e_h(t) + \delta_h(t), \quad t > 0, \\ e_h(0) & = & K_{h,0}Q_h g(0) - P_{h,0}K(0)g(0), \\ \partial_h e_h(t) & = & 0, \quad t \geq 0, \end{array} \right.$$

where the truncation error is

$$\delta_h(t) = \left(P_{h,0}A_0\tilde{u}(t) - A_{h,0}P_{h,0}\tilde{u}(t)\right) - \left(P_{h,0}K(0)g'(t) - K_{h,0}Q_h g'(t)\right), \qquad t > 0.$$

Therefore, by (C1), (C4) and the fundamental estimate (4.8), for $t > 0$,

$$\|\delta_h(t)\|_h \leq O(\epsilon(h)) \left(\|u\|_{0,t} + \|g\|_{1,t,H}\right),$$

whereas

$$\|e_h(0)\|_h \leq O(\epsilon(h)) \|g(0)\|_H,$$

Now, by taking norms in the variation-of-constant formula,

$$e_h(t) = \mathrm{e}^{tA_{h,0}} e_h(0) + \int_0^t \mathrm{e}^{(t-s)A_{h,0}} \delta_h(s)\, \mathrm{d}s, \qquad t \geq 0,$$

we readily conclude that (4.32) is true. □

For $0 < h \leq h_0$ and $t_n = n\tau$, $n \geq 0$, it is now natural to propose, as the full approximation to $u(t_n)$, the sum

$$u_{h,n} = K_h Q_h g(t_n) + \bar{u}_{h,n},$$

where $\bar{u}_{h,n}$ is the time approximation, at time level n, to the solution u_h of the IVP

$$\left\{ \begin{array}{rcll} \tilde{u}_h'(t) & = & A_{h,0}\tilde{u}_h(t) + P_{h,0}f(t) - K_{h,0}Q_h g'(t), & t > 0, \\ \tilde{u}_h(0) & = & P_{h,0}u_0 - K_{h,0}Q_h g(0), & \end{array} \right. \tag{4.33}$$

provided by some version of a rational method (4.27), based on $r(z)$. Since

$$\begin{aligned} \|P_h u(t_n) - u_{h,n}\|_h \leq \|(P_h - P_{h,0})\tilde{u}(t_n)\|_h + \|P_{h,0}\tilde{u}(t_n) - \tilde{u}_h(t_n)\|_h + \|\tilde{u}_h(t_n) - \bar{u}_{h,n}\|_h \\ + \|P_{h,0}K(0)g(t_n) - K_{h,0}Q_h g(t_n)\|_h, \end{aligned}$$

and combining Theorem 4.5, (d), (C2) and (C4), we get the estimate for the error of the full discretization that we just state afterwards. Recalling Theorem 4.5, notice that the estimate for the time discretization of (4.31) uses the norm

$$|||[u_{h,0}, P_h f, Q_h g]^T|||_{p+1,t,X_h} = \|u_{h,0}\|_h + \max_{0\leq j\leq p+1} \|P_h f^{(j)}(s)\|_h + \max_{0\leq j\leq p+2} \|Q_h g^{(j)}(s)\|_h,$$

that, by (S1), is bounded by $M_L|||[u, f, g]_{p+1,t}|||$, for $t \geq 0$.

Theorem 4.10. Assume that the solution u of the IBVP (4.1) belongs to the space $\mathcal{C}^1([0, +\infty), W) \cap \mathcal{C}^{p+1}([0, +\infty), X)$, $f \in \mathcal{C}_{ub}^{p+1}([0, +\infty), X)$ and $g \in \mathcal{C}_{ub}^{p+2}([0, +\infty), H)$. Let $0 < h \leq h_0$, $\tau > 0$ and $n \geq 0$. Then, the error of the full discretization can be estimated by

$$\|P_h u(t_n) - u_{h,n}\|_h \leq ERT_n(\tau) + ERS_n(h),$$

where

$$ERT_n(\tau) = C_c \tau^p t_n(1 + 2t_n) \left(|||[u, f, g]^T|||_{p+1,t_n} + \|K(0)\| \|g\|_{p+1,t_n} \right),$$

and

$$ERS_n(h) = t_n M_s M_L O(\epsilon(h)) \|w\|_{1,t,W}.$$

Again by the Lax–Richtmyer theorem, it can be proved that the proposed full discretization of (4.1) converges for data $u_0 \in X$, $f \in \mathcal{C}_{ub}([0, +\infty), X)$ and $g : [0, +\infty) \to Z$ of finite total variation.

4.3 Numerical experiments

The aim of this section is to corroborate the results of the previous ones. For that, we consider the domain $\Omega = (0, 1) \times (0, 1) \subset \mathbb{R}^2$ and we integrate the following parabolic initial boundary value problem with Dirichlet boundary conditions

$$\begin{cases} u_t(t, x, y) = \Delta u(t, x, y) - \sin(x + y + t) \\ \qquad\qquad + 2\cos(x + y + t), \quad (x, y) \in \Omega,\ t \in (0, T), \\ u(0, x, y) = \cos(x + y), \quad (x, y) \in \Omega, \\ u(t, x, y) = \cos(x + y + t), \quad (x, y) \in \partial\Omega,\ t \in (0, T), \end{cases} \tag{4.34}$$

for some $T > 0$.This problem has the form of (4.1) and satisfies (4.5), so the function

$$u(t, x, y) = \cos(x + y + t), \quad (x, y) \in \Omega,\, t \in (0, T)$$

is a genuine solution of (4.1) which in fact has as high regularity as required for any of the theorems of the previous section. We notice that for $X = L^2(\Omega)$, $Y = H^{\frac{3}{2}}(\partial\Omega)$, $A = \Delta$, $D(A) = H^2(\Omega)$ and $A_0 = A|_{\ker(\partial)}$ is the infinitesimal generator of an analytic semigroup of negative type, such that $D(A_0) = H^2(\Omega) \cap H^1_0(\Omega)$ [66]. We also notice that f and g have as much regularity as required in Theorems 4.5 and 4.10.

For the spatial discretization, we consider $h = 1/J$ for a certain integer J and a uniform grid $(x_i, y_j) = (ih, jh)$ for $0 \leq i, j \leq J$. We denote by

$$\mathring{\Omega}_h = \{(x_i, y_j) = (ih, jh) : 0 < i, j < J\}, \quad \partial\Omega_h = \Omega \setminus \mathring{\Omega}_h$$

the set of the interior and boundary nodes of the grid. This naturally leads to $X_h = \mathbb{R}^{(J+1)^2}$, endowed with the discrete L^2 norm, and $Y_h = \mathbb{R}^{4J}$ stands for the boundary values of such a matrix, so that $\partial_h : X_h \to Y_h$ sends a matrix $u_h \in X_h$ into the vector containing its boundary values. Then, $X_{h,0} = \ker \partial_h$ is isomorphic to $\mathbb{R}^{(J-1)^2}$. In such a case, P_h will be the sampling on the nodes of the grid when applied to a C^1-function and, for any function in $L^2(\Omega)$, as $C^1(\Omega)$ is dense in $L^2(\Omega)$, it corresponds to the limit in the discrete L^2-norm. The same applies for Q_h and the projection on the nodes of the boundary of the grid, in such a way that $Q_h\partial u = \partial_h P_h u$ for every $u \in X$, so the consistency condition (C2) is automatically satisfied. For the sake of simplicity, we denote by $U_h = P_h u$ and $U_\partial = \partial_h P_h u$ for every $u \in X$, and $G_h = Q_h g$ for $g \in Y$.

In order to discretize the operator A we use the well-known 4th-order nine-point formula for the Laplacian [74]. It is well known that for any elliptic problem of the form

$$Au = f, \quad \partial w = g, \tag{4.35}$$

with $f \in X$ and $g \in Y$, a solution $u \in W = H^6(\Omega)$ is discretized in $\mathring{\Omega}$ by

$$\tilde{A}_{h,0} U_h + \tilde{C}_h G_h = M_h F_h + \tilde{D}_h F_\partial + O(h^4), \tag{4.36}$$

where $\tilde{A}_{h,0}$ and M_h are the $(J-1) \times (J-1)$ tridiagonal block-matrices

$$\tilde{A}_{h,0} = \frac{1}{h^2}\begin{pmatrix} -\frac{10}{3}I + \frac{2}{3}T & \frac{2}{3}I + \frac{1}{6}T & 0 & \dots & 0 \\ \frac{2}{3}I + \frac{1}{6}T & -\frac{10}{3}I + \frac{2}{3}T & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & I \\ 0 & \dots & 0 & \frac{2}{3}I + \frac{1}{6}T & -\frac{10}{3}I + \frac{2}{3}T \end{pmatrix},$$

$$M_h = \frac{1}{12}\begin{pmatrix} 8I + T & I & 0 & \dots & 0 \\ I & 8I + T & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & I \\ 0 & \dots & 0 & I & 8I + T \end{pmatrix},$$

with I the identity matrix and

$$T = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix},$$

and where $\tilde{C}_h$ and $\tilde{D}_h$ respectively correspond to the lacking coefficients in $\tilde{A}_{h,0}$ and M_h associated to the nodes on the boundary. To fit this setting in our abstract framework, we need to consider

$$A_{h,0} = M^{-1}\tilde{A}_{h,0}, \quad C_h = M_h^{-1}\tilde{C}_h, \quad D_h = M_h^{-1}\tilde{D}_h.$$

We emphasize that the multiplication by M_h^{-1} is only a theoretical device to embed this problem into our abstract framework. In practice, such multiplications are never performed: neither the explicit computation of M_h^{-1} nor the solution of systems with matrix M_h is required. On the one hand, by Gerschgorin theorem, the eigenvalues of M_h are in $(1/3, 1)$, so those of M_h^{-1} are in $(1, 3)$ and its discrete L^2-norm is bounded by 3. Since $h^2\tilde{A}_{h,0} = (4I + T) \otimes (4I + T)/6 - 6I$ and the eigenvalues of T are $2\cos(2\pi jh)$ for $j = 1, \dots, J-1$, the eigenvalues of $h^2\tilde{A}_{h,0}$ are

$$(4 + 2\cos(2\pi jh))^2/6 - 6 = -\frac{4}{3}\pi^2 j^2 h^2 + O(h^4) = O(h^2), \qquad \text{for } j = 1, \dots, J-1.$$

Since $\tilde{A}_{h,0}$ and M_h commute, they can be diagonalized in the same set of eigenvectors and so the eigenvalues of the symmetric matrix $A_{h,0}$ are negative. Thus, the corresponding exponentials are bounded and (S2) is true.

Now, we are in position to define the operators $P_{h,0}$, $K_{h,0}$, that are not only the projection on the interior points of the grid, but it also contains some information on the nodes of the boundary, as they come from the nine-point formula. We set

$$P_{h,0}u = \begin{cases} U_h + M_h^{-1}\tilde{D}_h U_\partial & \text{in } \mathring{\Omega}_h, \\ 0 & \text{in } \partial\Omega_h, \end{cases} \tag{4.37}$$

Then, the consistency condition (C1) comes from the consistency of the elliptic discretization (4.36). For $u \in W_0$ and f as in (4.35), notice that in $\mathring{\Omega}_h$

$$\begin{aligned} A_{h,0}P_{h,0}u &= M_h^{-1}\tilde{A}_{h,0}U_h, \\ P_{h,0}A_0u &= P_{h,0}f = F_h + M_h^{-1}\tilde{D}_h F_\partial, \end{aligned}$$

so (4.36) and the bound $\|M_h^{-1}\|_h \leq 3$ leads to condition (C1) in the form

$$\|A_{h,0}P_{h,0}u - P_{h,0}A_0u\| = O(h^4).$$

To conclude, we must explain how to define K_h and $K_{h,0}$ when using the nine-point formula. Since for $g \in Y$, $K(0)g$ is the solution of the problem

$$AK(0)g = 0, \quad \partial K(0)g = g,$$

the discretization (4.36) naturally leads to

$$P_h K(0) g = -\tilde{A}_{h,0}^{-1} \tilde{C}_h (Q_h g) + O(h^4),$$

and we define

$$K_h(Q_h g) = \begin{cases} -\tilde{A}_{h,0}^{-1} \tilde{C}_h G_h & \text{in } \mathring{\Omega}_h, \\ G_h & \text{in } \partial\Omega_h, \end{cases} \tag{4.38}$$

so that condition (C3) is automatically satisfied in $H = H^{11/2}(\Omega)$. We also define $K_{h,0}$ in an analogous way to (4.37), by

$$K_{h,0}(Q_h g) = \begin{cases} K_h G_h + M^{-1} D G_h & \text{in } \mathring{\Omega}_h, \\ 0 & \text{in } \partial\Omega_h. \end{cases} \tag{4.39}$$

The fulfilment of the consistency condition (C4) reduces, in view of the definitions of $P_{h,0}$ and $K_{h,0}$, almost directly to the fulfilment of condition (C3). As anticipated when introducing the operators in the abstract setting, the definition of $K_{h,0}$ is practically identical to that of $P_{h,0}$, and hence (C4) is an almost redundant condition. Nevertheless, it must be additionally assumed in the case of finite differences. Observe, for example, that it makes no sense to speak of $P_{h,0} G_h$, even though the operators are formally identical.

We have already checked that this spatial discretization fits the abstract framework presented in Section 4.2.3. Before presenting the numerical results, notice that, for instance, for $u \in W$, $\tau > 0$, in the grid $\mathring{\Omega}$ it is true that

$$(I - \tau A_{h,0})^{-1} P_{h,0} u = (M_h - \tau \tilde{A}_{h,0})^{-1} \left(M_h U_h + \tilde{D}_h U_\partial \right),$$

Thus, we avoid computing the inverses of M_h, working instead with the corresponding matrix-vector products. This is advantageous since M_h is a sparse matrix. The vector to which $(M_h - \tau \tilde{A}_{h,0})^{-1}$ is applied can be computed efficiently by arranging the values of $P_h u$ in a matrix that mirrors the structure of $\bar{\Omega}_h$, as it is simply a linear combination of the entries of $P_h u$ with weights $8/12$ and $1/12$.

Furthermore, the resulting system can be solved efficiently using the conjugate gradient method, since both M_h and $\tilde{A}_{h,0}$ are sparse, and the products with $(M_h - \tau \tilde{A}_{h,0})$ can be implemented straightforwardly.

We integrate the semidiscretized version of (4.34) with the 3-stages SDIRK and 3-stages Gauss methods, already described in Chapter 2.

Integration with SDIRK3

Problem (4.34) is firstly integrated in time by the 3-stages SDIRK Runge–Kutta method by using the standard method of lines: discretizing firstly in space with the nine-point formula for the Laplacian and then in time. More precisely, the SDIRK3 method is applied to

$$\begin{cases} U_h'(t) = A_{h,0} U_h(t) + C_h G_h(t) + F_h(t) + D_h(F_\partial(t) - \dot{G}_h(t)), \\ U_h(0) = P_h u_0. \end{cases}$$

We notice that the fact that $D_h \neq 0$ with this spatial discretization implies that we still require $\dot{g}$, which is something which we try to avoid. For other spatial discretizations for which $D_h \equiv 0$ that would not be necessary, but the accuracy in space would be smaller. We implement SDIRK so that no matrix product with $\tilde{A}_{h,0}$ is necessary (which would be unstable for small h). In this case, we write the complete formulas for the integration for the sake of clarity. We calculate

$$
\begin{aligned}
K_{1,n,h} &= -\frac{1}{\gamma}U_{n,h} + (M_h - \tau\gamma\tilde{A}_{h,0})^{-1}[\frac{1}{\gamma}M_hU_{n,h} + \tau(\tilde{C}_hG_h(t_n) + M_hF_h(t_n) \\
&\qquad\qquad\qquad\qquad\qquad\qquad +\tilde{D}_h(F_\partial(t_n) - \dot{G}_h(t_n)))], \\
K_{2,n,h} &= -\frac{1}{\gamma}(U_{n,h} + a_{21}K_{1,n,h}) \\
&\quad +(M_h - \tau\gamma\tilde{A}_{h,0})^{-1}[\frac{1}{\gamma}M_h(U_{n,h} + a_{21}K_{1,n,h}) + \tau(\tilde{C}_hG_h(t_n + c_2\tau) \\
&\qquad\qquad +M_hF_h(t_n + c_2\tau) + \tilde{D}_h(F_\partial(t_n + c_2\tau) - \dot{G}_h(t_n + c_2\tau)))] \\
K_{3,n,h} &= -\frac{1}{\gamma}(U_{n,h} + a_{31}K_{1,n,h} + a_{32}K_{2,n,h}) \\
&\quad +(M_h - \tau\gamma\tilde{A}_{h,0})^{-1}[\frac{1}{\gamma}M_h(U_{n,h} + a_{31}K_{1,n,h} + a_{32}K_{2,n,h}) + \tau(\tilde{C}_hG_h(t_n + c_3\tau) \\
&\qquad\qquad +M_hF_h(t_n + c_3\tau) + \tilde{D}_h(F_\partial(t_n + c_3\tau) - \dot{G}_h(t_n + c_3\tau)))] \\
U_{n+1,h} &= U_{n,h} + b_1K_{1,n,h} + b_2K_{2,n,h} + b_3K_{3,n,h},
\end{aligned}
$$

where $\gamma, c_2, c_3, a_{21},\ a_{31},\ a_{32},\ b_1, b_2, b_3$ are the corresponding coefficients of the Butcher tableau. We notice that three linear systems with sparse matrices $(M_h - \tau\gamma\tilde{A}_{h,0})$ must be solved per step, as well as three evaluations of f, the same for g and $\dot{g}$. We have integrated till time $T = 1/2$ with $N = 200$ so that the error in space is negligible and the linear systems have been solved with the iterative gradient conjugate method with $tol = 10^{-14}$. The results on the global error turn up in the second column of Table 4.1. We notice that the order of convergence 2.25 expected by [8] is obtained. The global order behaves as the local one, instead of one less, when $r_\infty \neq 1$ [38, 60] due to a summation-by-parts argument, which in parabolic problems explains this fact. The local order can be seen to also behave as $O(\tau^{2+1/4})$, as Table 4.2 shows.

Table 4.1: Global errors and order of convergence with SDIRK3 method and suggested rational SDIRK3 methods.

	Runge–Kutta		Rational Explicit		Rational Implicit	
step size	error	order	error	order	error	order
1.000e-01	2.379e-04	–	8.721e-07	–	1.797e-07	–
5.000e-02	5.083e-05	2.23	6.656e-08	3.71	8.792e-09	4.35
2.500e-02	1.083e-05	2.23	4.025e-09	3.98	4.874e-10	4.17
1.250e-02	2.254e-06	2.26	2.555e-10	4.04	3.386e-11	3.85
6.250e-03	4.672e-07	2.27	1.569e-11	4.03	2.667e-12	3.67

Table 4.2: Local errors at $t = \tau$ and local order of convergence with SDIRK3 method and suggested rational SDIRK3 methods.

	Runge–Kutta		Rational Explicit		Rational Implicit	
step size	error	order	error	order	error	order
1.000e-01	3.964e-04	–	1.308e-06	–	4.727e-08	–
5.000e-02	8.315e-05	2.25	5.195e-08	4.65	2.223e-09	4.41
2.500e-02	1.734e-05	2.26	2.024e-09	4.68	8.524e-11	4.70
1.667e-02	6.921e-06	2.27	2.981e-10	4.72	1.239e-11	4.76

On the other hand, we have integrated the same problem with the suggested technique in this paper. We have considered two possibilities for the nodes $\mathbf{c}_n$ and $\mathbf{d}_n$ in (4.27), which we will denote by explicit and implicit method because of the difference in $\mathbf{c}_n$ for $n \geq 3$:

$$\begin{array}{llll}
 & \mathbf{c}_0 = [0,1,2,3] & \mathbf{d}_0 = [0,1,2,3,4] & \\
 & \mathbf{c}_1 = [-1,0,1,2] & \mathbf{d}_1 = [-1,0,1,2,3] & \\
 & \mathbf{c}_2 = [-2,-1,0,1] & \mathbf{d}_2 = [-2,-1,0,1,2] & \\
n \geq 3 & \mathbf{c}_n = [-3,-2,-1,0] & \mathbf{d}_n = [-3,-2,-1,0,1] & \text{explicit method} \\
 & \mathbf{c}_n = [-2,-1,0,1] & \mathbf{d}_n = [-3,-2,-1,0,1] & \text{implicit method}
\end{array} \tag{4.40}$$

Notice that, as $p = 4$, $\mathbf{c}_n \in \mathbb{R}^4$ while $\mathbf{d}_n \in \mathbb{R}^5$. There are many other possibilities but we have chosen these among the ones which just imply at most one function evaluation of f and g per step. Recall the stability function of the SDIRK3 method in (2.45). In such a way, $k = 1$, $m_1 = 3$ and $w_1 = \gamma$ in (1.25). Then, Lemma 2.4 can be applied to

$$\begin{aligned}
H_{11}(z) &= \frac{1}{1-z\gamma} = 1 + \gamma z + \gamma^2 z^2 + \gamma^3 z^3 + O(z^4), \\
H_{12}(z) &= \frac{1}{(1-z\gamma)^2} = 1 + 2\gamma z + 3\gamma^2 z^2 + 4\gamma^3 z^3 + O(z^4), \\
H_{13}(z) &= \frac{1}{(1-z\gamma)^3} = 1 + 3\gamma z + 6\gamma^2 z^2 + 10\gamma^3 z^3 + O(z^4), \\
I_{11}(z) &= \frac{z}{1-z\gamma} = z + \gamma z^2 + \gamma^2 z^3 + \gamma^3 z^4 + O(z^5), \\
I_{12}(z) &= \frac{z}{(1-z\gamma)^2} = z + 2\gamma z^2 + 3\gamma^2 z^3 + 4\gamma^3 z^4 + O(z^5), \\
I_{13}(z) &= \frac{z}{(1-z\gamma)^3} = z + 3\gamma z^2 + 6\gamma^2 z^3 + 10\gamma^3 z^4 + O(z^5),
\end{aligned}$$

giving rise to the coefficients $\boldsymbol{\gamma}^n_{1,i} \in \mathbb{R}^4$ and $\boldsymbol{\eta}^n_{1,i} \in \mathbb{R}^5$ for $i = 1, 2, 3$, which turn up in (4.27). What is finally implemented with our spatial discretization in order to use sparse matrices is

$$\bar{u}_{h,0} = P_h u_0 + \tilde{A}^{-1}_{h,0} \tilde{C}_h g(t_0), \tag{4.41}$$

and then, recursively, trying to minimize calculations,

$$\begin{aligned}
K_1 &= (M_h - \tau\gamma\tilde{A}_{h,0})^{-1}[M_h X_{h,n} - \gamma{\boldsymbol{\eta}_{11}^n}^T(-M_h\tilde{A}_{h,0}^{-1}\tilde{C}_h + D_h)G_h(t_n + \tau\mathbf{d}_n) \\
&\qquad +\tau\gamma{\boldsymbol{\gamma}_{11}^n}^T(M_h P_h + \tilde{D}_h\partial)f(t_n + \tau\mathbf{c}_n)], \\
K_2 &= (M_h - \tau\gamma\tilde{A}_{h,0})^{-1}[M_h K_1 - \gamma{\boldsymbol{\eta}_{12}^n}^T(-M_h\tilde{A}_{h,0}^{-1}\tilde{C}_h + D_h)G_h(t_n + \tau\mathbf{d}_n) \\
&\qquad +\tau\gamma{\boldsymbol{\gamma}_{12}^n}^T(M_h P_h + \tilde{D}_h\partial)F_h(t_n + \tau\mathbf{c}_n)], \\
K_3 &= (M_h - \tau\gamma\tilde{A}_{h,0})^{-1}[M_h K_2 - \gamma{\boldsymbol{\eta}_{13}^n}^T(-M_h\tilde{A}_{h,0}^{-1}\tilde{C}_h + D_h)G_h(t_n + \tau\mathbf{d}_n) \\
&\qquad +\tau\gamma{\boldsymbol{\gamma}_{13}^n}^T(M_h P_h + \tilde{D}_h\partial)F_h(t_n + \tau\mathbf{c}_n)], \\
\bar{u}_{h,n+1} &= (1 - \frac{1}{\gamma})\bar{u}_{h,n} + r_{11}K_1 + r_{12}K_2 + r_{13}K_3,
\end{aligned}$$

and, for any t_n in which we want to approximate the solution of (4.1), we take

$$u_{h,n} = \begin{cases} \bar{u}_{h,n} - \tilde{A}_{h,0}^{-1}\tilde{C}_h g(t_n) & \text{in } \mathring{\Omega}_h, \\ G(t_n) & \text{in } \partial\Omega_h, \end{cases} \tag{4.42}$$

Then, at each step we just solve three linear systems with matrix $(M_h - \tau\gamma\tilde{A}_{h,0})$ and another one with matrix $\tilde{A}_{h,0}$ to calculate $\tilde{A}_{h,0}^{-1}\tilde{C}_h g(t_n)$, which can be reused whenever necessary, in the same way that happens with the evaluation of f and g at each step. Table 3.1 shows the global error for several step sizes for both the explicit and implicit rational implementation suggested in this paper. It can be observed that, in both cases, the order is very near 4, which is the classical order of the method, as Theorem 4.5 predicts. On the other hand, the local order corresponding to the error after the first step when using $\mathbf{c}_n$ and $\mathbf{d}_n$ in the last lines of (4.40) approaches 5 as τ diminishes, as Table 4.2 shows and as predicted by Lemma 4.4.

Comparing the errors for a fixed step size, we observe that even for the larger timestep $\tau = 0.1$, the errors are much smaller with the rational approach proposed in this paper than with the Runge–Kutta method applied through the standard method of lines. Since order reduction is also avoided, this difference becomes even more significant as τ decreases. Moreover, at least for this problem, the error for the same step size is considerably smaller with the implicit choice of $\mathbf{c}$ than with the explicit one.

We also provide a comparison in terms of CPU time among the three methods. In particular, Figure 4.2 shows the global error against CPU time. We can observe that, for a fixed step size, the CPU time required (at least for the shorter time step sizes) by the Runge–Kutta method is about three-quarters of that required by the rational method. This is explained by the fact that most of the computational cost comes from solving linear systems and, as already mentioned, the rational method requires solving four linear systems per step instead of three for Runge–Kutta.

When the step size is smaller, it appears that the conjugate-gradient method used to solve some of the linear systems requires less time, and therefore the fact that only one function evaluation of f per step is needed (instead of three) makes the rational methods more favourable. In any case, for this problem and a fixed CPU time, the difference in accuracy is about two orders of magnitude for larger time step sizes between the Runge–Kutta method and the explicit rational method, and about four orders of magnitude for

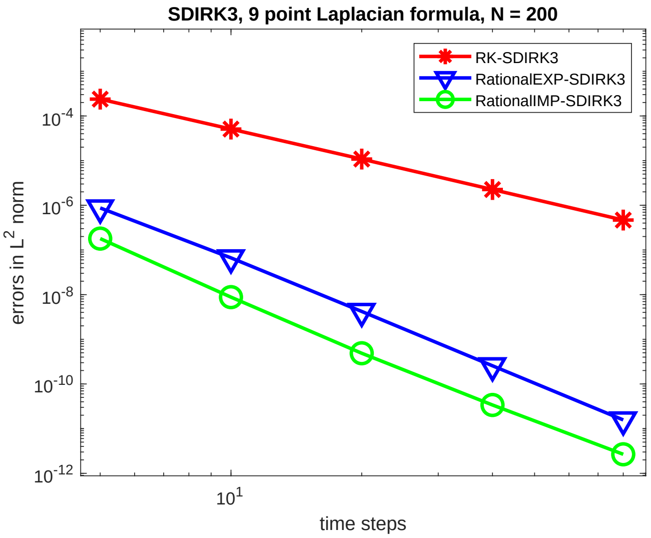


Figure 4.1. Error against time steps when integrating problem (4.34) with SDIRK3 method and suggested rational SDIRK3 methods.

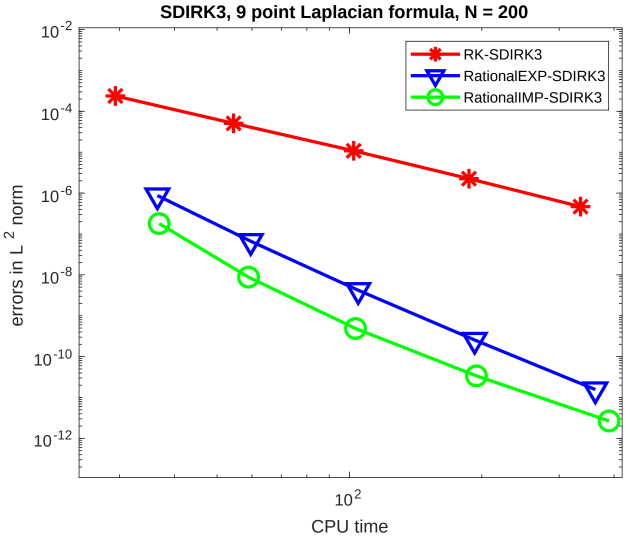


Figure 4.2. Error against CPU time when integrating problem (4.34) with SDIRK3 method and suggested rational SDIRK3 methods.

the smaller step sizes. Furthermore, if the implicit rational method is considered, the error is roughly eight times smaller than with the explicit one.

Integration with Gauss3

We also take as time integrator the 3-stages Gauss method. When using it to integrate (4.34) with u_0 in (4.34), order 3.25 for the global error would be expected in general according to [8]. However, in the same way as for SDIRK3 method, a summation by parts argument applies because the problem is parabolic and the global order is one order higher, i.e. 4.25, as it can be approximately observed in the second column of Table 4.3 when integrating till time $T = 1$, where the same values of N and *tol* as for SDIRK3 have been considered for the spatial discretization and the iterative solution of the linear systems. We recall the stability function of the Gauss3 method (2.44). Using this, the implementation of the Runge–Kutta method means solving, at each step, three linear systems with matrices $(M_h - \tau w_\ell \tilde{A}_{h,0})$ as well as the three evaluations of $f(t_n + c_j\tau)$, $g(t_n + c_j\tau)$ and $\dot{g}(t_n + c_j\tau)$ for $j = 1, 2, 3$.

Table 4.3: Errors and order of convergence with Gauss3 method and the suggested rational Gauss3 methods.

	Runge–Kutta		Rational Explicit		Rational Implicit	
step size	error	order	error	order	error	order
1.000e-01	6.117e-06	–	2.276e-08	–	1.712e-09	–
6.667e-02	1.105e-06	4.22	2.029e-09	5.96	1.291e-10	6.38
5.000e-02	3.131e-07	4.38	3.576e-10	6.03	2.100e-11	6.31
4.000e-02	1.225e-07	4.20	9.245e-11	6.06	5.408e-12	6.08
3.333e-02	5.508e-08	4.39	3.059e-11	6.07	1.819e-12	5.98

On the other hand, we have taken the suggested rational method (4.27) based on this Gauss method, by considering $\mathbf{c}_n$ and $\mathbf{d}_n$ in a similar way to (4.40), but now with $\mathbf{c}_n \in \mathbb{R}^6$ and $\mathbf{d}_n \in \mathbb{R}^7$. Moreover, the coefficients $\boldsymbol{\gamma}^n_{l,1}$ and $\boldsymbol{\eta}^n_{l,1}$ are found as Lemma 2.4 states for

$$\begin{aligned} H_{l1}(z) &= \frac{1}{1 - w_\ell z} = \sum_{j=0}^{5} w_\ell^j z^j + O(z^6), \\ I_{l,1}(z) &= \frac{z}{1 - w_\ell z} = \sum_{j=0}^{5} w_\ell^j z^{j+1} + O(z^7), \quad l = 1, 2, 3. \end{aligned}$$

The final formulas for the implementation of both explicit and implicit methods by using

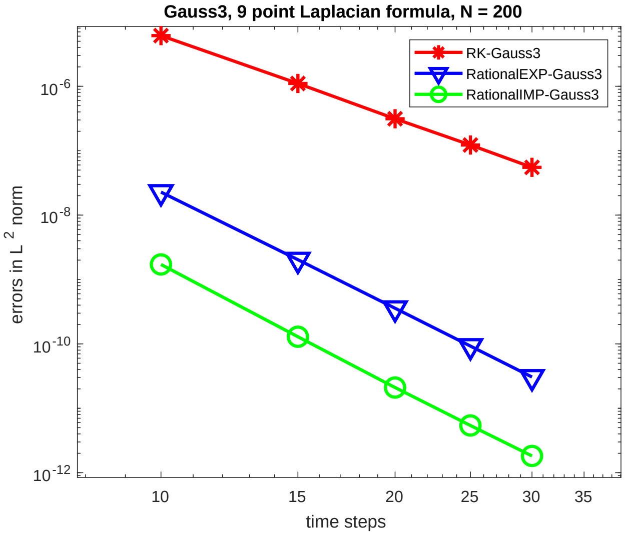


Figure 4.3. Error against time steps when integrating problem (4.34) with Gauss3 method and suggested rational Gauss3 methods.

the nine-point formula in space are then

$$\begin{aligned}
\bar{u}_{h,n+1} \;=\; & -\bar{u}_{h,n} \\
& +r_{11}(M_h-\tau w_1\tilde{A}_{h,0})^{-1}[M_hX_{h,n}-w_1{\boldsymbol{\eta}_{11}^n}^T(-M_h\tilde{A}_{h,0}^{-1}\tilde{C}_h+\tilde{D}_h)G_h(t_n+\tau\mathbf{d}_n) \\
& \qquad\qquad +\tau w_1{\boldsymbol{\gamma}_{11}^n}^T(M_hP_h+\tilde{D}_h\partial)F_h(t_n+\tau\mathbf{c}_n)] \\
& +r_{21}(M_h-\tau w_2\tilde{A}_{h,0})^{-1}[M_hX_{h,n}-w_2{\boldsymbol{\eta}_{21}^n}^T(-M_h\tilde{A}_{h,0}^{-1}\tilde{C}_h+\tilde{D}_h)G_h(t_n+\tau\mathbf{d}_n) \\
& \qquad\qquad +\tau w_2{\boldsymbol{\gamma}_{21}^n}^T(M_hP_h+\tilde{D}_h\partial)F_h(t_n+\tau\mathbf{c}_n)] \\
& +r_{31}(M_h-\tau w_3{\boldsymbol{\eta}_{31}^n}^T\tilde{A}_{h,0})^{-1}[M_hX_{h,n}-w_3(-M_h\tilde{A}_{h,0}^{-1}\tilde{C}_h+\tilde{D}_h)g(t_n+\tau\mathbf{d}_n) \\
& \qquad\qquad +\tau w_3{\boldsymbol{\gamma}_{31}^n}^T(M_hP_h+\tilde{D}_h\partial)F_h(t_n+\tau\mathbf{c}_n)],
\end{aligned}$$

once $\bar{u}_{h,0}$ is calculated through (4.41) and, whenever the approximation to u in (4.34) is required, (4.42) is considered. Then, in the same way as for the suggested rational SDIRK methods, three linear systems with matrix $(M_h - \tau w_\ell \tilde{A}_{h,0})$ must be solved at each step and, by keeping the calculation from one step to the other, just one more linear system must be solved with matrix $\tilde{A}_{h,0}$, as well as just one function evaluation of f and g.

The global errors which are obtained with the rational methods can be observed in Table 4.3, where it is clear that no order reduction is observed, i.e. the errors diminish like $O(\tau^6)$. Moreover, the size of the errors is much smaller than with the Runge–Kutta method, even for the biggest value of τ. As for the comparison in efficiency, Figure 4.4 shows that, for a fixed time step size, the suggested rational methods are more expensive than the Runge–Kutta in an approximate proportion of 4 to 3, which is natural because of the number of linear systems to be solved at each step. However, as the errors are so small with the rational methods, with a fixed CPU time, the size of the errors with the explicit rational method is still two orders of magnitude smaller than with the Runge–Kutta method and, with the implicit rational method, even three orders of magnitude smaller. Moreover, the comparison is more favourable for rational methods when the time step size is small.

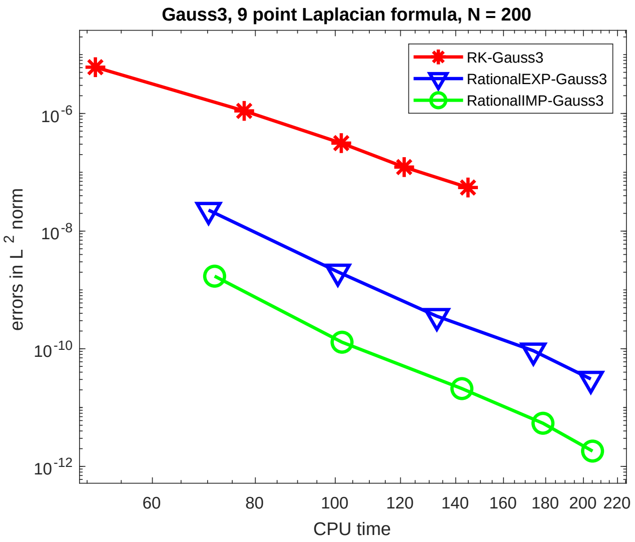


Figure 4.4. Error against CPU time when integrating problem (4.34) with Gauss3 method and suggested rational Gauss3 methods.

4.4 Semilinear problems with boundary conditions

After introducing the rational methods in Chapter 2, we extended them in two directions: to semilinear problems (3.1) in Chapter 3, and to initial boundary value problems (4.1) in Chapter 4. These two developments are somewhat independent, in the sense that the results in each case are built upon the framework established in Chapter 2.

The final section of this dissertation is devoted to synthesizing these results to address semilinear IBVPs of the form

$$\begin{cases} u'(t) = Au(t) + f(t, u(t)), \quad 0 < t < T, \\ u(0) = u_0, \\ \partial u(t) = g(t), \quad 0 < t < T, \end{cases} \tag{4.43}$$

where X and Y are two complex Banach spaces, A and ∂ are two linear operators $A : D(A) \subset X \to X$ and $\partial : D(A) \subset X \to Y$. For a certain $\alpha \geq 0$, $u_0 \in X_\alpha$, $f : [0, \infty) \times X_\alpha \to X$ and $g : [0, \infty) \to Y$. Set

$$D(A_0) = \operatorname{Ker} \partial = \{ x \in D(A) \, : \, \partial x = 0 \}$$

and let $A_0 : D(A_0) \subset X \to X$ be the restriction of A to $D(A_0)$. Since the convergence analysis of the method for these problems does not require fundamentally new techniques, and it is essentially a combination of the previous results, it is more convenient to summarize it in this section rather than dedicating an entire chapter to this problem.

We start recalling the hypotheses we need for this setting.

Hypothesis H1.

- If $\alpha = 0$, $A_0 \in \mathcal{G}(X, M, \omega)$, that is, A_0 is the infinitesimal generator of a $\mathcal{C}_0$ semigroup of bounded operators in X, with $M \geq 1$ and $\omega \in \mathbb{R}$. .

- If $\alpha > 0$, $A_0 \in \mathcal{G}(X, M, \omega, \theta)$ that is, A_0 is the infinitesimal generator of an analytic semigroup of bounded operators in X, with $M \geq 1$, $\omega \in \mathbb{R}$ and $0 < \theta < \pi/2$.

Hypothesis H2.

- If $\alpha = 0$ and $\left\{\mathrm{e}^{tA_0}\right\}_{t\geq 0}$ is a $\mathcal{C}_0$ semigroup, we assume that $f : [0,T] \times X \to X$ is locally Lipschitz continuous. Thus, there exists a real number L such that
$$\|f(t,\xi) - f(t,\eta)\| \leq L\|\xi - \eta\|$$
for all $t \in [0,T]$ and $\max\left(\|\xi\|, \|\eta\|\right) \leq R$.
- If $\alpha > 0$ and $\left\{\mathrm{e}^{tA_0}\right\}_{t\geq 0}$ is analytic,we assume that $f : [0,T] \times X_\alpha \to X$ is locally Lipschitz. Thus, there exists a real number L such that
$$\|f(t,\xi) - f(t,\eta)\| \leq L\|\xi - \eta\|_\alpha$$
for all $t \in [0,T]$ and $\max\left(\|\xi\|_\alpha, \|\eta\|_\alpha\right) \leq R$.

Hypothesis H3.

- If $\alpha = 0$, we assume that r is A-acceptable.
- If $\alpha > 0$ and $A \in \mathcal{G}\left(X, M, \omega, \theta\right)$, we assume that r is strongly A(ϑ)-acceptable with $\vartheta > \theta$.

Hypothesis H4. There exists a bounded, linear operator $E : Y \to X$, such that
$$Ev \in D(A) \text{ and } \partial Ev = v, \qquad v \in Y,$$
and $AE : Y \to X$ is also bounded.

Under these hypotheses, we can assume that there exists a function $u : [0,T] \to X_\alpha$ satisfying (4.43). For the reasons explained at the beginning of this chapter, we assume, without loss of generality, that $\omega < 0$ and we will work with the extension operator $K = K(0)$. As in the beginning of Chapter 4, the solution of (4.43) can be written as
$$u(t) = \tilde{u}(t) + K(0)g(t), \tag{4.44}$$
where, since $\partial\tilde{u}(t) = 0$, $t \geq 0$, and $AK(0) = 0$, the function $\tilde{u} : [0,T] \to X_\alpha$ solves
$$\left\{\begin{array}{lcll} \tilde{u}'(t) & = & A_0\tilde{u}(t) + f(t, \tilde{u}(t) + Kg(t)) - Kg'(t), & \quad 0 < t < T, \\ \tilde{u}(0) & = & \tilde{u}_0 := u_0 - Kg(0). & \end{array}\right. \tag{4.45}$$
Notice that $u_0 - Kg(0) \in X_\alpha$ whenever $u_0 \in X_\alpha$, since $Kg(0) \in D(A) \subset X_\alpha$. Moreover, the function $\tilde{f} : [0,T] \times X_\alpha \to X$ defined by
$$\tilde{f}(t,v) = f(t, v + Kg(t))$$
satisfies Hypothesis H2 whenever f does. We follow the same strategy as in Section 3.2.3: first, we analyze the integration of the linear version of the method in the norm of X_α, and then we apply Gronwall's lemma together with the Lipschitz property to establish the convergence of the nonlinear case. For this purpose, we require an adaptation of Lemma 3.3 to the setting with boundary values.

Lemma 4.11. Let $\alpha \in (0,1)$, $A \in \mathcal{G}(X, M, \omega, \theta)$ and $F(\tau)$, $F_n(\tau)$ be the operators defined in (4.18) and (4.24), respectively. For $0 \leq \beta \leq \alpha$, it is true that

$$\|F(\tau)\, w\|_\beta \leq K\, \tau^{-\beta} \|w\|_\infty, \quad \text{for } v \in \mathcal{C}_{ub}\left([0,\infty), Y\right), \tag{4.46}$$

$$\|F_n(\tau)\, \mathbf{w}\|_\beta \leq K\, \tau^{-\beta} \|\mathbf{w}\|_{X^p}, \quad \text{for } \mathbf{w} \in Y^p, \tag{4.47}$$

where $\| \cdot \|_{Y^p}$ corresponds to the maximum of the norm of each component in Y.

We define a function $h : [0,T] \to X$ by $h(t) = f(t, \tilde{u}(t) + Kg(t))$ and consider the integration of the problem

$$\left\{ \begin{array}{rcll} v'(t) & = & A_0 v(t) + h(t) - Kg'(t), & 0 < t < T, \\ v(0) & = & \tilde{u}_0, & \end{array} \right.$$

which has $\tilde{u}$ as solution, with a certain time step $\tau > 0$ and the scheme

$$v_{n+1} = r(\tau A_0) v_n + \tau E_n(\tau) h(t_n + \tau \mathbf{c}_n) - F_n(\tau) g(t_n + \tau \mathbf{c}_n), \quad n \geq 0 \tag{4.48}$$

We need to adapt the seminorms to take into account the space X_α. Set, for a function $v \in \mathcal{C}^m_{ub}([0,\infty), X_\alpha)$, $m \geq 0$ and $0 \leq t \leq \infty$,

$$\|h\|_{m,t,\alpha} = \max_{0 \leq j \leq m} \sup_{0 \leq s \leq t} \|h^{(j)}(s)\|_\alpha. \tag{4.49}$$

Then, in the product space $\mathcal{C}^m_{ub}([0,\infty), X_\alpha) \times \mathcal{C}^m_{ub}([0,\infty), X) \times \mathcal{C}^{m+1}_{ub}([0,\infty), Y)$ we consider $||| \cdot |||_{m,t,\alpha}$ given, for $[v, \phi, \psi]^T$ in such a product, by the expression

$$|||[v, \phi, \psi]^T|||_{m,t,\alpha} = \|v\|_{m,t,\alpha} + \|\phi\|_{m,t} + \|\psi\|_{m+1,t} \tag{4.50}$$

Theorem 4.12. Let $\tilde{u} : [0,T] \to X_\alpha$ be the solution of (4.45) to be approximated on the interval $[0,T]$. Assume also that $\tilde{u} \in \mathcal{C}^{p+1}\left([0,T], X_\alpha\right)$, $h \in \mathcal{C}^{p+1}\left([0,T], X\right)$ and that $g \in \mathcal{C}^{p+2}([0,\infty), Y)$. If v_n is the numerical approximation to $\tilde{u}(t_n)$ given by (4.48) with constant step size $0 < \tau = T/N < \tau_0$, there exists a constant K depending on M, T and α such that

$$\|\tilde{u}(t_n) - v_n\|_\alpha \leq K\, \tau^p |||[\tilde{u}, h, g]^T|||_{p+1, t_n, \alpha}, \qquad 0 \leq n \leq N. \tag{4.51}$$

Proof. The proof is similar to that in Theorem 3.6 but additionally taking into account the regularisation estimates (4.46), (4.47). □

We consider the nodes $\boldsymbol{c}_n = [-p+1, \ldots, 0] \in \mathbb{Z}^p$ or $\boldsymbol{c}_n = [-p+2, \ldots, 1] \in \mathbb{Z}^p$. It happens again that the first choice requires the use of the previous values $\boldsymbol{U}_n = [u_{n-p+1}, \ldots, u_n]$ to compute u_{n+1}, so it is explicit; whereas the second choice requires $\boldsymbol{U}_n = [u_{n-p+2}, \ldots, u_{n+1}]$, and an implicit scheme turns up. We are free to choose any nodes $\boldsymbol{d}_n \in \mathbb{R}^p$ for the boundary conditions. Then the proposed scheme to integrate (4.43) is

$$\begin{aligned} u_{n+1} = {} & K(0) g(t_{n+1}) + r\left(\tau A\right)\left(u_n - K(0) g(t)\right) + \tau E_n(\tau) f\left(t_n + \tau \boldsymbol{c}_n, \boldsymbol{U}_n\right) \\ & - F_n(\tau) K(0) g(t_n + \tau \boldsymbol{d}_n), \end{aligned} \tag{4.52}$$

for $n \geq p-1$. The motivation to define (4.52) may be clear taking intro account formulas (2.29), (3.10), (4.27), (4.44). Recall that starting values u_n, for $0 < n < p-1$ must be provided. This can be done using the scheme (4.52) following the theory in Section 3.2.4. We conclude by stating the final result that brings this doctoral thesis to its culmination.

Theorem 4.13. For $0 \leq \alpha < 1$, let $u : [0,T] \to X_\alpha$ be the solution of (4.43) to be approximated in the interval $[0,T]$. Let us assume hypotheses H1, H2, H3 and H4 and also that $u \in \mathcal{C}^{p+1}([0,T], X_\alpha)$, $h \in \mathcal{C}^{p+1}([0,T], X)$ and $g \in \mathcal{C}^{p+2}([0,T], X)$. If u_n is the numerical approximation to $u(t_n)$ given by (4.52) with constant step size $0 < \tau = T/N < \tau_0$, and $u_0, \cdots, u_{p-1} \in X_\alpha$ are starting values satisfying

$$\|u(t_n) - u_n\|_\alpha \leq C_0\, \tau^p, \qquad 0 \leq n \leq p-1, \tag{4.53}$$

then, there exists a constant $K > 0$ depending on T, α and the nodes $\{\mathbf{c}_n\}_{n=0}^{N-1}$, $\{\mathbf{d}_n\}_{n=0}^{N-1}$ such that

$$\|u(t_n) - u_n\|_\alpha \leq K\, C_s(n)\, \tau^p\, |||[u,h,g]^T|||_{p+1,t_n\alpha}, \qquad 0 \leq n \leq N. \tag{4.54}$$

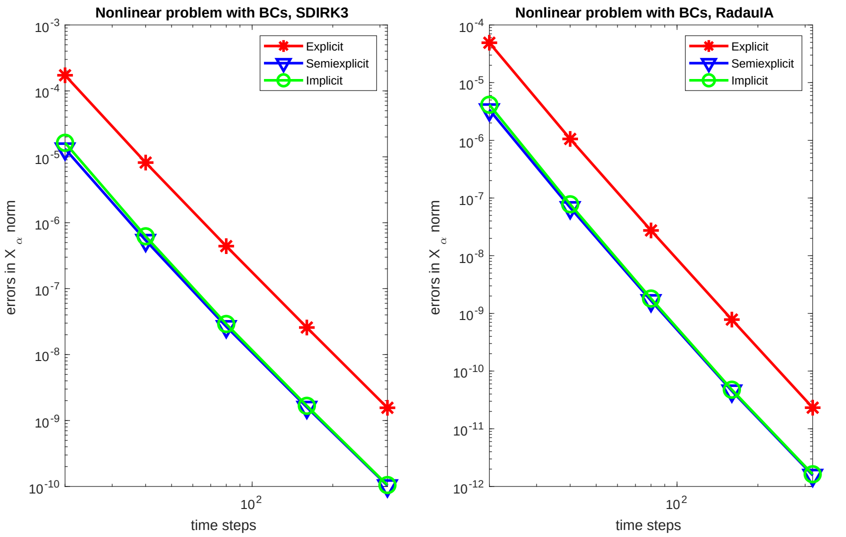


Figure 4.5. Error in the discrete norm associated to $\|\cdot\|_{H^1(0,1)}$ for the problem (4.55) with two different time integrators and $J = 100$.

We conclude the chapter with an example that illustrates the order of convergence predicted by Theorem 4.13. We use a version of the method (4.52) to integrate the parabolic PDE

$$\begin{cases} u_t = u_{xx} + u^2 + f(t), & 0 \leq x \leq 1,\ 0 \leq t \leq 0.5, \\ u(0,x) = 3x - x^2, & 0 \leq x \leq 1, \\ u(t,0) = 1 - \exp(-\pi^2 t),\ \ u(t,1) = 1 + \exp(-\pi^2 t), & 0 \leq t \leq 0.5, \end{cases} \tag{4.55}$$

where $f : [0,1] \to X$ is such that

$$u(x,t) = 1 + (2x-1)\exp(-\pi^2 t) + (x-x^2)\exp(t)$$

is the solution of the equation. The problem fits our framework setting $X = L^2[0,1]$, $A = d^2/dx^2$ with $D(A) = H^2[0,1]$. We take $\alpha = 1/2$, so that according to (1.52), $X_\alpha = H^1[0,1]$ and $\|\cdot\|_\alpha = \|\cdot\|_{H^1(0,1)}$. Notice that the source term satisfies the Lipschitz condition, since the Sobolev embedding (1.51) guarantees that $H^1[0,1] \hookrightarrow L^\infty[0,1]$. This is also true for every $\alpha > 1/4$. To deal with the boundary conditions, we take $Y = \mathbb{R}^2$, $\partial u = [u(0), u(1)]$, which makes sense for every $u \in D(A)$, and then $D(A_0) = D(A) \cap \ker \partial = H^2[0,1] \cap H_0^1[0,1]$.

We choose the version of (4.52) which is based on the stability rational functions of the previously described 3-stage SDIRK and 3-stage RadauIA methods. As nodes, we take $\boldsymbol{c}_n = [-p+1, \ldots, 0] \in \mathbb{Z}^p$ or $\boldsymbol{c}_n = [-p+2, \ldots, 1] \in \mathbb{Z}^p$ with the various possible implementations described in Section 3.3 and the equispaced nodes for the boundary conditions, like in (4.40).

For the spatial discretization, we discretize A with the second order central differences on the uniform grid $x_i = ih$, for $i = 1, \ldots, N$, with $h = 1/(N+1)$, leading to the matrix (2.51). The corresponding systems are solved using Thomas algorithm together with an LU-factorization version.

Table 4.4: Errors and orders of convergence with the rational SDIRK3 method (4.52) integrating problem (4.55).

	Explicit		Semiexplicit		Implicit	
step size	error	order	error	order	error	order
2.500e-02	1.725e-04	–	1.321e-05	–	1.631e-05	–
1.250e-02	8.176e-06	4.40	5.407e-07	4.61	6.229e-07	4.71
6.250e-03	4.417e-07	4.21	2.662e-08	4.34	2.913e-08	4.42
3.125e-03	2.575e-08	4.10	1.589e-09	4.07	1.671e-09	4.12
1.563e-03	1.560e-09	4.04	1.026e-10	3.95	1.050e-10	3.99

Table 4.5: Errors and orders of convergence with the rational Gauss3 method 4.52 integrating problem (4.55).

	Explicit		Semiexplicit		Implicit	
step size	error	order	error	order	error	order
2.500e-02	4.914e-05	–	3.388e-06	–	4.136e-06	–
1.250e-02	1.057e-06	5.54	6.824e-08	5.63	7.781e-08	5.73
6.250e-03	2.740e-08	5.27	1.668e-09	5.35	1.809e-09	5.43
3.125e-03	7.806e-10	5.13	4.537e-11	5.20	4.751e-11	5.25
1.563e-03	2.320e-11	5.07	1.563e-12	4.86	1.627e-12	4.87

The results are reported in Tables 4.4 and 4.5, which display good agreement with the predictions of Theorem 4.13. In both cases, it can be observed that the first iteration provided by the semi-explicit mode reduces the error by an additional order of convergence.

However, the subsequent corrections carried out by the implicit mode do not produce any further improvement in accuracy. Nevertheless, it is worth emphasizing that, as we have already seen in Section 3.3, the implicit mode may still be of practical interest in situations where the nonlinear term exhibits a stiff character.

Conclusions

The results presented in this thesis fulfil the objective set at the beginning, namely, the development of a series of rational methods that integrate abstract evolution problems while achieving the optimal order of convergence. These methods possess properties that make them particularly attractive for practical applications. On the one hand, they are efficient, requiring the solution of the same number of linear systems per step as a Runge–Kutta method when integrating an initial value problem. On the other hand, they exhibit favourable numerical stability properties, as they avoid numerical differentiation of the data, a common requirement in methods designed to avoid the order reduction.The numerical experiments demonstrate that, for a fixed step size, the reduction in error relative to Runge–Kutta methods is significant. Moreover, since the number of linear systems to be solved remains the same while the number of function evaluations can be substantially reduced, the proposed methods emerge as a highly efficient computational tool.

The core of the work is in Chapter 2, where the functional analysis framework we use is introduced, the main properties are proved, and the rational methods for nonhomogeneous linear problems are established. As this is the chapter in which they are introduced, special attention is paid to their practical implementation, as well as to the consequences of the possible choices of nodes at which the source term f is evaluated. These insights into the effects of the nodes guide the development of the subsequent chapters.

In Chapter 3 we extend the methods to deal with semilinear problems. Although the framework is rather more complicated, the ideas to extend the methods are simple. The computational aspect reveals interesting results with the possible implementations, showing advantages in different cases depending on the stiffness of the source term and the system matrix. This line of research remains open. For example, one could investigate ideas similar to Lagrangian methods to attempt evaluating the source term at intermediate points, or explore the possibility of taking multiple steps implicitly, as it is done with the starting values.

Finally, we extend the methods to integrate initial boundary value problems in Chapter 4. We also synthesize these results with those of semilinear problems to consider both features at the same time. The numerical experiments of this chapter are of special interest. On the one hand, they treat the kind of problems that appear in most realistic applications. On the other, we work with a more sophisticated spatial discretization with a mass matrix similar to those in finite elements. This feature shows that our methods are versatile and can be used in realistic computations that require complex geometries.

Moreover, we have developed a new framework to analyze the full discretization of the methods, that is, to take into account the error due to the spatial discretization.

The framework turns out to be very general and cover standard methods such as finite differences, finite elements and spectral methods.

There are some lines of research that remain open after this work. A first possibility is to pursue this line of theoretical research further, extending the methods to a wider class of problems. Firstly, one can consider linear nonautonomous equations of the type

$$u'(t) = A(t)u(t) + f(t), \quad 0 < t \leq T, \quad u(0) = u_0,$$

where, for every $t \in [0, T]$, the linear operator $A(t) : D(A(t)) \subset X \mapsto X$ is sectorial, and the function $t \mapsto A(t)$ has certain degree of smoothness. Secondly, fully nonlinear equations in a open set $\mathcal{O} \subset X$ of the form

$$u'(t) = F(t, u(t)), \quad 0 < t \leq T, \quad u(0) = u_0,$$

can be considered within this framework. For a detailed study of these problems, see the references [53, 66].

From another perspective, other common topics in numerical analysis where the phenomenon of order reduction arises could also be studied using this approach. To mention a few examples, alternating direction methods, the design of splitting methods, or problems involving convolution quadrature appear to be promising areas in which the proposed techniques could provide new insights and improvements.

Finally, the favourable features of the developed methods—such as their efficiency, stability, compatibility with spatial discretizations involving mass matrices, and the possibility of implementing them with variable step sizes—make them well suited for applications in more realistic computational settings. A relevant example is the use of IMEX (implicit-explicit) methods, which are designed to handle stiff components separately and fit naturally with the time integrators proposed in this thesis. This strategy is particularly effective for solving reaction-advection-diffusion equations and fluid mechanics problems, such as the Navier–Stokes equations.

Bibliography

[1] S. Abarbanel, D. Gottlieb, M. H. Carpenter, *On the removal of boundary errors caused by Runge–Kutta integration of nolinear partial differential equations*, SIAM J. Sci. Comput. 17 (1996), pp. 777–782.

[2] I. Alonso-Mallo, *Rational methods with optimal order of convergence for partial differential equations*, Appl. Numer. Math., 35 (2000), pp. 265–292.

[3] I. Alonso-Mallo, *Runge–Kutta methods without order reduction for linear initial boundary value problems*, Numer. Math., 91 (2002), pp. 577–603.

[4] I. Alonso-Mallo, B. Cano, *Spectral/Rosenbrock discretizations without order reduction for linear parabolic problems*, Appl. Num. Math. 47 (2002), pp. 247–268.

[5] I. Alonso-Mallo, B. Cano, *Avoiding order reduction of Runge–Kutta discretizations for linear time-dependent parabolic problems*, BIT Numer. Math., 44 (2004), pp. 1–20.

[6] I. Alonso-Mallo, B. Cano, *Efficient time integration of nonlinear partial differential equations by means of Rosenbrock methods*, Mathematics 9 (2021), 1970.

[7] I. Alonso-Mallo, C. Palencia, *On the convolution operators arising in the study of abstract initial boundary value problems*, Proc. Roy. Soc. Edinburgh Sect. A 126 (1996), pp. 515–539.

[8] I. Alonso-Mallo, C. Palencia, *Optimal orders of convergence for Runge–Kutta methods and linear, initial boundary value problems*, Appl. Numer. Math., 44 (2003), pp. 1–19.

[9] C. Arranz-Simón, A. Ostermann, *Exponential integrators for parabolic problems with non-homogeneous boundary conditions*, arXiv:2510.21381 (2025).

[10] C. Arranz-Simón, C. Palencia, *Rational methods for abstract, linear, inhomogeneous problems without order reduction*, SIAM J. Numer. Anal., 63.1 (2025), pp. 422–436.

[11] C. Arranz-Simón, B. Cano, C. Palencia, *Rational methods for abstract, semilinear problems without order reduction*, arXiv:2509.17984 (2025).

[12] C. Arranz-Simón, B. Cano, C. Palencia, *Rational methods for abstract, linear initial boundary value problems without order reduction*, arXiv:2510.15481 (2025).

[13] J. Bergh, J. Löfström, Interpolation spaces: an introduction. Vol. 223. Springer Science and Business Media, 2012.

[14] A. Biswas, D. I. Ketcheson, B. Seibold, D. Shirokoff, *Design of DIRK schemes with high weak stage order*, Communications in Applied Mathematics and Computational Science 18.1 (2023), pp. 1–28.

[15] A. Biswas, D. I. Ketcheson, B. Seibold, D. Shirokoff, *Algebraic structure of the weak stage order conditions for Runge–Kutta methods*, SIAM J. Numer. Anal. 62.1 (2024), pp. 48–72.

[16] A. Biswas, D. I. Ketcheson, S. Roberts, B. Seibold, D. Shirokoff, *Explicit Runge–Kutta methods that alleviate order reduction*, SIAM J. Numer. Anal. 63. 4 (2025), pp. 1398-1426.

[17] P. Brenner, V. Thomée, *Stability and convergence rates in L_p for certain difference schemes*, Math. Scand., 26 (1970), pp. 5–23.

[18] P. Brenner, V. Thomée, *On rational approximations of semigroups*, SIAM J. Numer. Anal., 16.4 (1979), pp. 683–694.

[19] H. Brezis, Functional analysis, Sobolev spaces and partial differential equations, New York: Springer, 2000.

[20] M. P. Calvo, J. Frutos, J. Novo, *An efficient way to avoid the order reduction of linearly implicit Runge–Kutta methods for nonlinear IBVP's*, Mathematical Modelling, Simulation and Optimization of Integrated Circuits, Antreich, K.; Bulirsch, R.; Gilg, A.; Rentrop, P., Eds.; International Series of Numerical Mathematics Vol. 146, Birkhauser Verlag, Basel Switzerland, 2003; pp. 321–332.

[21] M. P. Calvo, C. Palencia, *Avoiding the order reduction of Runge–Kutta methods for lineal initial boundary value problems*, Math. Comput., 71.240 (2001), pp. 1529–1543.

[22] M. P. Calvo, C. Palencia, *A class of explicit multistep exponential integrators for semilinear problems*, Numer. Math., 102.3 (2006), pp. 367–381.

[23] B. Cano, M. J. Moreta, *Exponential quadrature rules without order reduction for integrating linear initial boundary value problems*, SIAM J. Num. Anal. 56-3 (2018), pp. 1187–1209.

[24] B. Cano, M. J. Moreta, *Solving nonlinear initial boundary value problems with explicit Runge-Kutta exponential methods without order reduction*, ESAIM: M2AN 58 (2024), pp. 1053–1085.

[25] B. Cano, M. J. Moreta, *Efficient exponential Rosenbrock methods till order four*, J. Comput. Appl. Math. 453 (2025) pp. 116158.

[26] B. Cano, M. J. Moreta, *Simplified explicit exponential Runge-Kutta methods without order reduction*, to appear in J. Comput. Math. (2025).

[27] B. Cano, M. J. Moreta, *Exponential Rosenbrock methods without order reduction when integrating nonlinear initial boundary value problems*, to appear in ESAIM: M2AN (2025).

[28] M. H. Carpenter, D. Gottlieb, S. Abarbanel, W. S. Don, *The theoretical accuracy of Runge–Kutta time discretizations for the initial boundary value problem: A study of the boundary error*, SIAM J. Sci. Comput., 16 (1995), pp. 1241–1252.

[29] M. Crouzeix, S. Larsson, S. Piskarev, V. Thomée, *The stability of rational approximations of analytic semigroups*, BIT Numer. Math., 33 (1993), pp. 74–84.

[30] N. Dunford, J. T. Schwartz, Linear operators, part 1: general theory, John Wiley and Sons, 1988.

[31] K. J. Engel, R. Nagel, One-parameter semigroups for linear evolution equations, New York: Springer, 2000.

[32] K. J. Engel, R. Nagel, A Short Course on Operator Semigroups, Springer Science and Business Media, 2006.

[33] M. Fabrizio, C. Giorgi, V. Pata, *A new approach to equations with memory*, Arch. Ration. Mech. Anal., 198.1 (2010), pp. 189–232.

[34] W. Gautschi, *Norm estimates for inverses of Vandermonde matrices*, Numer. Math., 23.4 (1974), pp. 337–347.

[35] W. Gautschi, *Optimally conditioned Vandermonde matrices*, Numer. Math., 24 (1975), pp. 1–12.

[36] C. González, C. Palencia, *Stability of Runge–Kutta methods for abstract time-dependent parabolic problems: the Hölder case*, Math. Comput., 68.225 (1999), pp. 73–89.

[37] E. Hairer, G. Wanner, Solving Ordinary Differential Equations I: Nonstiff problems., Springer-Verlag, Berlin, 1993.

[38] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II: Stiff and Diferential-Algebraic problems, Springer-Verlag, Berlin, 1996.

[39] D. Henry, Geometric theory of semilinear parabolic equations, Vol. 840, Springer, 2006.

[40] R. Hersh, T. Kato, *High-accuracy stable difference schemes for well-posed initial-value problems*, SIAM J. Numer. Anal., 16.4 (1979), pp. 670–682.

[41] E. Hille, R. S. Phillips, Functional Analysis and Semi-groups, Vol. 31, American Mathematical Society, 1996.

[42] M. Hochbruck, A. Ostermann, *Exponential Runge–Kutta methods for parabolic equations*, Appl. Numer. Math., 53 (2005), pp. 323–339.

[43] M. Hochbruck, A. Ostermann, *Explicit exponential Runge–Kutta methods for semilinear parabolic problems* SIAM J. Numer. Anal., 43 (2005), pp. 1069–1090.

[44] M. Hochbruck, A. Ostermann, *Exponential integrators* Acta Numer., 19 (2010), pp. 209–286.

[45] S. L. Keeling, *Galerkin/Runge–Kutta Discretizations for Semilinear Parabolic Equations*, SIAM J. Numer. Anal. 27.2 (1990), pp. 394–418.

[46] C. A. Kennedy, M. H. Carpenter, *Diagonally implicit Runge–Kutta methods for ordinary differential equations. A review*, Tech. Report TM-2016-219173, National Aeronautics and Space Administration, Langley Research Center, 2016.

[47] C. A. Kennedy, M. H. Carpenter, *Diagonally implicit Runge–Kutta methods for stiff ODEs*, Appl Numer. Math., 146 (2019), pp. 221–244.

[48] J. Lang, J. G. Verwer, *ROS3P - An accurate third-order Rosenbrock solver designed for parabolic problems*, BIT Num. Math., 41 (2001), pp. 731–738.

[49] S. Larsson, V. Thomee, and L. B. Wahlbin, *Finite-element methods for a strongly damped wave equation*, IMA J. Numer. Anal., 11.1 (1991), pp. 115–142.

[50] J. Löfström, *Interpolation of boundary value problems of Neumann type on smooth domains*, J Lond. Math. Soc., 2.3 (1992), pp. 499–516.

[51] J. L. Lions, E. Magenes, *Non-homogeneous boundary value problems and applications: Vol. 1.*, Springer Science and Business Media, 2012.

[52] C. Lubich, A. Ostermann, *Linearly implicit time discretizations of nonlinear parabolic equations*, IMA J. Num. Anal. 15 (1995), pp. 555–583.

[53] A. Lunardi, Analytic semigroups and optimal regularity in parabolic problems, Springer Science and Business Media, 2012.

[54] A. Lunardi, Interpolation theory, Scuola Normale Superiore, Pisa, third edition, 2018.

[55] S. McKee, *Generalised discrete Gronwall lemmas*, Z. Angew. Math. Mech., 62 (1982), pp. 429–434.

[56] R. Metzler, J. Klafter, *The random walk's guide to anomalous diffusion: a fractional dynamics approach*, Phys. Rep., 339.1 (2000), pp. 1–77.

[57] R. Metzler, J. H. Jeon, A. G. Cherstvy, E. Barkai, *Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking*, Phys. Chem. Chem. Phys., 16.44 (2014), pp. 24128–24164.

[58] M. L. Minion, R. I. Saye, *Higher-order temporal integration for the incompressible Navier–Stokes equations in bounded domains.* J. Comput. Phys. 375 (2018), pp. 797-822.

[59] A. Ostermann, M. Roche, *Runge–Kutta methods for partial differential equations and fractional orders of convergence*, Math. Comput., 59.200 (1992), pp. 403–420.

[60] A. Ostermann, M. Roche, *Rosenbrock methods for partial differential equations and fractional orders of convergence*, SIAM J. Numer. Anal., 30.4 (1993), pp. 1084–1098.

[61] C. Palencia, *A stability result for sectorial operators in Banach spaces*, SIAM J. Numer. Anal., 30.5 (1993), pp. 1373–1384.

[62] C. Palencia, *On the stability of variable stepsize rational methods for holomorphic semigroups*, Math. Comput., 62.205 (1994), pp. 93–103.

[63] C. Palencia, *Stability of rational multistep approximations of holomorphic semigroups*, Math. Comput., 64.210 (1995), pp. 591–599.

[64] C. Palencia, I. Alonso-Mallo, *Abstract initial boundary value problems*, Proc. Roy. Soc. Edinburgh Sect. A 124 (1994), pp. 879–908.

[65] D. Pathria, *The correct formulation of intermediate boundary conditions for Runge–Kutta time integration of initial boundary value problems*, SIAM J. Sci. Comput. 18 (1997), pp. 1255–1266.

[66] A. Pazy, Semigroups of Linear Operators and Applications to Partial Differential Equations, Vol. 44, Springer Science and Business Media, 2012.

[67] L. Perko, Differential equations and dynamical systems, Vol. 7, Springer Science and Business Media, 2013.

[68] S. Roberts, A. Sandu, *Eliminating order reduction on linear, time-dependent ODEs with GARK methods*, arXiv preprint arXiv:2201.07940, (2022).

[69] W. Rudin, Functional analysis, International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York , 1991.

[70] J. M. Sanz-Serna, *Diez Lecciones de Cálculo Numérico*, Universidad de Valladolid, 1998.

[71] J.M. Sanz-Serna, J.G. Verwer, W.H. Hundsdorfer, *Convergence and order reduction of Runge–Kutta schemes applied to evolutionary problems in partial differential equations*, Numer. Math., 50 (1986), pp. 405–418.

[72] V. Sposini, A. V. Chechkin, F. Seno, G. Pagnini, R. Metzler, *Random diffusivity from stochastic equations: comparison of two models for Brownian yet non-Gaussian diffusion*, New J. Physics, 20(4) (2018), pp. 043044.

[73] J. Stiller, *A spectral deferred correction method for incompressible flow with variable viscosity.* J. Comput. Phys. 423 (2020), pp. 109840.

[74] J. C. Strikwerda, Finite difference schemes and partial differential equations, Pacific Grove California: Wadsworth and Brooks, 1989.

[75] H. Triebel, Interpolation Theory, Function Spaces, Differential Operators, North Holland, 1978.

[76] J. G. Verwer, *Convergence and order reduction of diagonally implicit Runge–Kutta schemes in the method of lines*, in Numerical Analysis, D. F. Griffiths and G. A. Watson eds., Pitman, Boston, MA, 1986, pp. 220—237.

[77] A. Yagi, Abstract Parabolic Evolution Equations and their Applications, Springer Science and Business Media, 2009.

[78] K. Yosida, Functional analysis, Springer Science and Business Media, 2012.