Rational methods for abstract semilinear problems without order reduction

Carlos Arranz-Simón[1], Begoña Cano[1], César Palencia[1]

[1]Applied Mathematics Department, IMUVA,University of Valladolid, P/ Belen, 7, Valladolid, 47011, Spain.

Contributing authors: carlos.arranz@uva.es; bcano@uva.es; cesar.palencia@uva.es;

Abstract

Rational methods are intended to time integrate linear homogeneous problems. However, their scope can be extended so as to cover linear nonhomogeneous problems. In this paper the integration of semilinear problems is considered. The resulting procedure requires the same computational cost than the one of a linked Runge–Kutta method, with the advantage that the order reduction phenomenon is avoided. Some numerical illustrations are included showing the predicted behaviour of the proposed methods.

Keywords: order reduction; rational methods; Runge–Kutta methods; partial differential equations; abstract evolution equations.

1 Introduction

It is well-known that many phenomena in science, engineering or social sciences can be described by semilinear initial value problems, which can be written in abstract form like

$$\begin{cases} u'(t) = Au(t) + f(t, u(t)), & t > 0, \\ u(0) = u_0, \end{cases} \tag{1}$$

where A is a space differential operator and $f(t, u)$ is a nonlinear term which may contain derivatives in space of at least one order less than those of A.

1

On the other hand, it is also well-known that integrating these problems in time with methods with stages, as the widely used Runge–Kutta methods, lead to order reduction, i.e. the order of convergence which is observed is less than that corresponding to the same method when integrating a non-stiff ODE [27]. Because of this, several techniques have been devised in the literature to avoid it.

The first one consists of designing methods satisfying certain stiff (or weak stage) order conditions and not only the non-stiff ones [8–10, 18, 19]. The disadvantage of this is that there is less freedom in the choice of coefficients, which may influence the computational efficiency of the suggested methods. Moreover, in the case of weak stage order conditions, the analysis has just been performed for linear problems and just seen numerically for nonlinear problems till order 3.

Another technique for linear problems was suggested in [14], 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 [13], where the non-natural hypothesis that $f(t, u)$ vanishes for nul u must be made.

A third procedure consists of correcting the boundary values for the stages [1–4, 24]. In the case of linear problems, if an analytic expression for the source term is known for which space and time derivatives can be calculated, order reduction can be completely avoided [2, 3]. In the case of nonlinear problems with time-dependent nonlinearity, numerical differentiation is required [24] if order ≥ 3 is pursued with Runge–Kutta methods. In the case of Rosenbrock methods, which are Runge–Kutta type integrators which make use of the Jacobian of the vector field to integrate nonlinear problems more efficiently, numerical differentiation is just required to get order ≥ 5 if f does not contain derivatives in space and to get order ≥ 4 if it does [4]. This technique is very cheap since just some additional calculations on the boundaries of the stages must be performed. The only disadvantage is that, in order to get high accuracy, numerical differentiation is required, which is well-known to be unstable for small grids [26].

This paper deals with the problem of avoiding order reduction when integrating the semilinear problem (1) by generalizing the technique already described in [7] for linear problems. The main idea is to consider an $A(\theta)$-acceptable rational method, which can be that associated to a Runge–Kutta method when integrating a homogeneous problem, and to write suitably problem (1) as a homogeneous one. This technique has the advantage that no numerical differentiation is required if f has no space derivatives and it is also very cheap since the same number of linear systems as that corresponding to the Runge–Kutta method when integrating a linear system must be solved. On the other hand, once some starting values are calculated, just one nonlinear system per step must be calculated if the method is implicit in the part corresponding to the source term. Moreover, there is no barrier on the order of accuracy which can be achieved. The final technique resembles that suggested in [17] for Garlerkin/Runge–Kutta discretizations for semilinear parabolic problems, but it is much more general in its deduction and application.

In contrast to [7], due to the nonlinearity, a more complex analysis must be performed distinguishing between different types of nonlinearities and considering sharp regularization estimates. On the other hand, it is not an aim of the paper to consider initial boundary value problems with time-dependent boundary conditions. For linear problems, that has already been studied in [6] and we numerically know that it also works properly for semilinear problems, but its thorough analysis deserves further research.

The paper is structured as follows: In Section 2, the derivation of the methods is described. Then, in Section 3, some technical lemmas are given, which allow to prove convergence of the method in Section 4. As the method requires in principle to calculate f at some previous values (as with multistep methods), the procedure to calculate the starting values is described and justified in Section 5. Finally, in Section 6, some numerical experiments are shown which corroborate in both hyperbolic and parabolic problems the avoidance of order reduction.

2 Derivation of the methods

The purpose of this paper is to extend the family of rational methods for linear problems of the form

$$\begin{cases} u'(t) = Au(t) + f(t), & t > 0, \\ u(0) = u_0, \end{cases} \tag{2}$$

which were introduced in [7], to semilinear problems of the form (1). In this section we aim to explain how to do this. For the convenience of the reader, we briefly review what these methods consist of and the main assumptions we consider.

Our analysis will be based on the formulation of (1) as an evolution equation in a Banach space $(X, \| \cdot \|)$. We assume standard properties for the operator A, as in [20, 25].

Throughout the article, 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 1.

- 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

$$\| (\lambda I - A)^{-n} \| \leq \frac{M}{(\operatorname{Re} \lambda - \omega)^n}, \tag{3}$$

for $n = 1, 2, \ldots$, on the plane $\{\lambda \in \mathbb{C} : \operatorname{Re} \lambda > \omega\}$ for $M \geq 1$, $\omega \in \mathbb{R}$.

Under this assumption, the operator A is the infinitesimal generator of a $\mathcal{C}_0$ semigroup $\left\{ \mathrm{e}^{tA} \right\}_{t \geq 0}$ that satisfies the growth estimate

$$\|\mathrm{e}^{tA}\| < M\mathrm{e}^{t\omega}. \tag{4}$$

- 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

$$\| (\lambda I - A)^{-1} \| \leq \frac{M}{|\lambda - \omega|} \tag{5}$$

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 $0 < \theta < \pi/2$.

Under this assumption, the operator A is the infinitesimal generator of an analytic semigroup $\{\mathrm{e}^{tA}\}_{t \geq 0}$. Fixed $\omega^* \geq \omega$, 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 (4), we now also have the estimate

$$\|t^\alpha \tilde{A}^\alpha \mathrm{e}^{tA}\| \leq M \mathrm{e}^{t\omega}. \tag{6}$$

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

Hypothesis 2.
- 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 globally Lipschitz continuous. Thus, there exists a real number L such that

$$\|f(t, \xi) - f(t, \eta)\| \leq L\|\xi - \eta\| \tag{7}$$

for all $t \in [0, T]$.
- 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 globally Lipschitz. Thus, there exists a real number L such that

$$\|f(t, \xi) - f(t, \eta)\| \leq L\|\xi - \eta\|_\alpha \tag{8}$$

for all $t \in [0, T]$.

We note that for the convergence proofs in the chapter, it is sufficient that (7) and (8) hold in a strip along the exact solution. Although, for simplicity, we assume that f is globally Lipschitz.

The methods designed in [7] start from a rational mapping

$$r(z) = r_\infty + \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} r_{\ell,j} (1 - w_\ell z)^{-j}, \tag{9}$$

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.

4

Let $r(z)$ be a rational mapping that approximates the exponential e^z with order $p \geq 1$, that is,

$$r(z) - e^z = O\left(z^{p+1}\right), \qquad r(z) - e^z \neq O\left(z^{p+2}\right), \quad \text{as } z \to 0. \tag{10}$$

Hypothesis 3.

- If $\alpha = 0$, we assume that r is A-acceptable, i.e., that $|r(z)| \leq 1$ when $\mathrm{Re}\, z \leq 0$.
- If $\alpha > 0$, we assume that $\gamma = |r_\infty| < 1$ and, if the semigroup has sectorial angle θ, we also assume that r is A(ϑ)-acceptable with $\vartheta > \theta$. This means that $|r(z)| \leq 1$ when z is in $\{z \in \mathbb{C} : |\arg(-z)| < \vartheta\}$.

Under these assumptions, there exists a number $\tau_0 > 0$ such that the operator

$$r(\tau A) = r_\infty I + \sum_{\ell=1}^{k} \sum_{j=1}^{m_l} r_{\ell,j} \left(I - \tau w_\ell A\right)^{-j} \tag{11}$$

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

Basic convergence and stability results for our analysis in relation to these operators can be found in [12]. In this paper, the following stability bound is stated

$$\|r^n (\tau A)\| \leq M\, C_s(n)\, e^{\omega^+ t_n}, \quad t_n = \tau\, n, \quad C_s = O\left(n^\sigma\right), \tag{12}$$

where $0 \leq \sigma \leq 1/2$ with $\omega^+ = \max\{0, \omega\}$. The stability bound is optimal (it is sharp for $A = d/dx$ in the maximum norm [11]) and may be improved depending on the nature of r and A. For our purposes, it is interesting to note that $C_s(n)$ is such that $\sigma = 0$, when one of the following is satisfied:

(a) $r(z) = 1/(1 - z)$, which corresponds to the implicit Euler method,
(b) for $r_{m,n}(z) = P_m/Q_n$ the Padé approximant of e^z with $\deg(P_m) = m$ and $\deg(Q_n) = n$, whenever $n = m - 1$. This is the case of Radau methods.
(c) X is a Hilbert space and A an ω-dissipative operator
(d) A generates an analytic semigroup in a Banach space X.

In [12] it is also proved a convergence estimate of the homogeneous problem $(f = 0)$

$$\|r^n (\tau A)\, u_0 - e^{t_n A} u_0\| \leq C_e\, M\, t_n\, \tau^p\, e^{\omega^+ \kappa t_n} \|A^{p+1}\, u_0\|, \qquad n \geq 1, \tag{13}$$

for $u_0 \in D\left(A^{p+1}\right)$, $\kappa = \kappa(r) \geq 1$ and $C_e = C_e(r) > 0$. Note that the previous bound shows the convergence of the rational method to the solution of the homogeneous problem with order p, even in cases where $C_s(n)$ is not $O(1)$.

The fact that rational methods preserve the order of convergence p when applied to homogeneous problems is the key starting point of our work. This is because this property is not preserved when we consider non-homogeneous problems, where the corresponding Runge–Kutta method exhibits a reduced order of convergence that has to do with the stage order of the method, rather than to p itself. This phenomenon is what is known as order reduction. The optimal orders that can be achieved have been studied in [5, 22, 23]. In [7], a brief outline of the causes of this reduction is presented.

5

The strategy proposed in [7] to design the methods is the introduction of an evolution problem in a suitable product space $Z = X \times Y$, so that the first component of this system has the original problem as its solution. This is achieved by considering the space of uniformly continuous, bounded functions $Y = \mathcal{C}_{ub}([0,\infty), X)$, the operator $L : v \in Y \mapsto v(0) \in X$ and the operator $B : v \in D(B) \subset Y \mapsto v' \in Y$, whose domain is $D(B) = \{v \in Y : v' \in Y\}$. The operator B is the infinitesimal generator of the translation semigroup, that is,

$$e^{tB} v(s) = v(t+s) \quad \text{for } v \in Y.$$

Thus, the evolution problem

$$\begin{cases} u'(t) = Au(t) + Lv(t), & t \geq 0, \\ v'(t) = \qquad\quad Bv(t), & t \geq 0, \\ u(0) = u_0, \\ v(0) = f, \end{cases} \tag{14}$$

is introduced. A direct calculation proves that the first component of the solution of this problem is the solution of (2). Once this is achieved, the idea is to apply the rational method to the homogeneous problem (14), which gives as a result the approximation

$$\begin{aligned}
\bar{u}_{n+1} &= r_\infty \bar{u}_n \\
&\quad + \sum_{\ell=1}^{k} \sum_{j=1}^{m_l} r_{\ell,j} (I - \tau w_\ell A)^{-j} \left(\bar{u}_n + \tau w_\ell \sum_{i=1}^{j} (I - \tau w_\ell A)^{i-1} L (I - \tau w_\ell B)^{-i} r^n(\tau B) f \right) \\
&= r(\tau A) \bar{u}_n + \tau E(\tau A) r^n(\tau B) f,
\end{aligned} \tag{15}$$

for a certain operator $E(\tau A) : Y \to X$. This scheme has the disadvantage that the resolvents of B are more difficult to compute than the action of the semigroup itself. This is solved by the following approximation result (Lemma 4.3 in [7])

$$\| L (I - \tau w_\ell B)^{-i} f(t_n + \cdot) - \boldsymbol{\gamma}_{\ell,i}^n \cdot f(t_n + \tau \boldsymbol{c}_n) \| \leq C \tau^p \| f^{(p)}(t_n + \cdot) \|_\infty, \tag{16}$$

for certain values $\boldsymbol{\gamma}_{\ell,i}^n \in \mathbb{R}^p$ whenever $\boldsymbol{c}_n \in \mathbb{R}^p$ with $c_i \neq c_j$ for $i \neq j$. This result allows to approximate one of these resolvents by a linear combination of p values of the source term. An advantage of these methods is that they leave freedom to choose the abscissae on which the source term f is evaluated. We can choose different nodes $\boldsymbol{c}_n \in \mathbb{R}^p$ in every step so that the scheme uses the values

$$f(t_n + \tau \boldsymbol{c}_n) = [f(t_n + \tau c_n^1), \ldots, f(t_n + \tau c_n^p)]^T \in X^p$$

to compute the approximation of u at time t_{n+1}. After choosing the nodes, the appropriate coefficients $\boldsymbol{\gamma}_{\ell,j}^n \in \mathbb{R}^p$ must be computed solving the $p \times p$ linear Vandermonde

systems

$$\begin{pmatrix} 1 & \cdots & 1 \\ c_n^1 & & c_n^p \\ \vdots & \ddots & \vdots \\ (c_n^1)^{p-1} & \cdots & (c_n^p)^{p-1} \end{pmatrix} \boldsymbol{\gamma}_{\ell,j}^n = \begin{pmatrix} 0! F_{\ell,j}^0 \\ 1! F_{\ell,j}^1 \\ \vdots \\ (p-1)! F_{\ell,j}^p \end{pmatrix},$$

where the right hand side is given by the Taylor expansion

$$(1 - w_\ell z)^{-j} = \sum_{k=0}^{\infty} F_{\ell,j}^k z^k. \tag{17}$$

In our implementations, only a few different $\boldsymbol{c}_n$ are used, so that they can be easily precomputed before starting the integration. After this, a step of the method is computed by

$$u_{n+1} = r_\infty u_n + \sum_{\ell=1}^{k} \sum_{j=1}^{m_l} r_{\ell,j} (I - \tau w_\ell A)^{-j} \left(u_n + \tau w_\ell \sum_{i=1}^{j} (I - \tau w_\ell A)^{i-1} \boldsymbol{\gamma}_{\ell,j}^n \cdot f(t_n + \tau \boldsymbol{c}_n) \right)$$

$$= r(\tau A) u_n + \tau E_n(\tau A) f(t_n + \tau \boldsymbol{c}_n), \tag{18}$$

for a certain operator $E_n(\tau A) : X^p \to X$, whose dependence on n is only due to the possibility of choosing different nodes at each step. Note that the formula in the first line is written in a more explicit way (that shows how the method can be implemented), while the second line is more compact and will be useful in the analysis of the method. The main result in [7] shows that this scheme approximates the solution of (2) with order p, except perhaps for a reduction of $1/2$ due to the stability constant $C_s(n)$. Since we are now interested in semilinear problems of the form (1), we 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 by $f(t_n, u(t_n)) \approx f(t_n, u_n)$. Although integer equispaced nodes lead to bad conditioned Vandermonde systems for high p (see [6, 7]), that is not a problem for our purpose of avoiding order reduction (at least for $p \le 6$, as it can be observed in the numerical experiments of Section 6).

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 (18),

$$u_{n+1} = r(\tau A) u_n + \tau E_n(\tau A) f(t_n + \tau \boldsymbol{c}_n, \boldsymbol{U}_n), \quad n \ge p - 1. \tag{19}$$

Starting values $u_0, u_1, \cdots, u_{p-1}$, must be provided. In Section 5 we explain how to compute the first values within this framework. In Section 6 we discuss in depth the consequences of choosing each of the node possibilities.

3 Preliminaries: discrete inequalities and regularisation

In this section we state some results which are required to prove the convergence of the scheme (19) in the following sections.

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 article. 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:*

$$\sum_{m=k}^{m^+(k)} t_{n-m}^{-\alpha} \leq p^{1+\alpha} \, t_{n-k}^{-\alpha}, \tag{20}$$

$$\tau \sum_{m=0}^{n-1} t_{n-m}^{-\alpha} \leq \frac{t_n^{1-\alpha}}{1-\alpha}, \tag{21}$$

$$\sum_{m=k}^{m^+(k)} \delta^{n-m-1} \leq p \, \delta^{1-p} \delta^{n-k-1}, \tag{22}$$

$$\sum_{m=1}^{n-1} \delta^{n-m-1} \leq \frac{1}{1-\delta}. \tag{23}$$

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

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

which proves (20), since the sum has at most p terms. For the second inequality, notice that

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

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

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

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

We now state the following lemma, which is a variant of the discrete Gronwall lemma.

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

$$(a) \ \xi_n^p = \sum_{k=n-p+1}^{n} \xi_k \quad or \quad (b) \ \xi_n^p = \sum_{k=n-p+2}^{n+1} \xi_k, \quad for \quad p-1 \leq n \leq N-1. \quad (24)$$

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

$$\max_{0 \leq k \leq p-1} \xi_k \leq K_0 \quad (25)$$

and that

$$\xi_{n+p-1} \leq K_0 + K_1 \sum_{m=0}^{n-1} \left(\tau \, t_{n-m}^{-\alpha} + \tau^{1-\alpha} \delta^{n-m-1} \right) \xi_{m+p-1}^p, \quad n \geq 0. \quad (26)$$

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

$$\xi_n \leq K K_0 \qquad for \quad n \geq p-1. \quad (27)$$

Proof. . We first assume case (a) in (24). Notice that

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

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

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 formulae (21), (23) 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 (20), (22),

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

9

Then, we combine the latter and (25) to get that, for $n \geq 0$,

$$\xi_{n+p-1} \leq K\,K_0 + KK_1 \sum_{k=1}^{n-1} \left(\tau\,t_{n-k}^{-\alpha} + \tau^{1-\alpha}\delta^{n-k-1}\right)\xi_{k+p-1}, \tag{28}$$

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 (28). The proof concludes applying Lemma 2.1. in [15]. $\quad\square$

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.

Notice that $r\,(\tau A) - r_\infty I$ is a linear combination of powers of resolvents of A. This implies, by (3), that for $0 < \tau < \tau_0$,

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

and in the analytic case, Lemma 2.2 in [15] also implies that

$$\|A\left(r\,(\tau A) - r_\infty I\right)x\| \leq \frac{K}{\tau}\|x\|, \quad x \in X,$$

for another constant $K > 0$ that may depend on r and T. Then, by interpolation (see e.g. [28]), we get that

$$\| \left(r\,(\tau A) - r_\infty I\right)x\|_\alpha \leq \frac{K}{\tau^\alpha}\|x\|, \quad x \in X. \tag{29}$$

This shows that the linear part of the numerical scheme regularises the solution after one step. For several steps, the equation (7) in [15], although stated in a more general framework of variable step size problems, can be realised in our context as a generalisation of the above to several steps,

$$\| \left(r^n\,(\tau A) - r_\infty^n I\right)x\|_\alpha \leq \frac{K}{t_n^\alpha}\|x\|, \quad x \in X. \tag{30}$$

Taking into account formulae (15) and (18), it can be seen that the operators $E(\tau A)$ and $E_n(\tau A)$ are given by a linear combination of resolvents of A. Therefore, using the

10

same argument, it can be directly proved that, for $0 \leq \beta \leq \alpha$,

$$\|E(\tau A)\,v\|_\beta \leq K\,\tau^{-\beta}\|v\|_Y, \quad \text{for } v \in Y, \tag{31}$$

$$\|E_n(\tau A)\,v_p\|_\beta \leq K\,\tau^{-\beta}\|v_p\|_{X^p}, \quad \text{for } v_p \in X^p, \tag{32}$$

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

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

Lemma 3.3. *Let $0 < \alpha < 1$. Under hypotheses 1 and 3, let $\xi_m \in X_\alpha$, $0 \leq m \leq n$ and $0 < \tau < \tau_0$. Then, there exists a positive constant K (that may be different in each case) such that the following estimates hold*

$$\left\| \tau \sum_{m=0}^{n-1} r^{n-m-1}\,(\tau A)\,\xi_m \right\|_\alpha \leq K\tau \sum_{m=0}^{n-2} \left(\frac{\|\xi_m\|}{t_{n-m-1}^\alpha} + \gamma^{n-m-1}\|\xi_m\|_\alpha \right) + \tau\|\xi_{n-1}\|_\alpha, \tag{33}$$

$$\left\| \tau \sum_{m=0}^{n-1} r^{n-m-1}\,(\tau A)\,\xi_m \right\|_\alpha \leq K \left(\max_{0 \leq m \leq n-2} \|\xi_m\| + \tau \max_{0 \leq m \leq n-1} \|\xi_m\|_\alpha \right). \tag{34}$$

Proof. . Taking into account the regularization estimate (30), the left hand side in (33) and (34) is bounded by

$$\left\| \tau \sum_{m=0}^{n-1} \left(r^{n-m-1}\,(\tau A) - r_\infty^{n-m-1} \right) \xi_m \right\|_\alpha + \left\| \tau \sum_{m=0}^{n-1} r_\infty^{n-m-1}\xi_m \right\|_\alpha$$

$$\leq \tau \sum_{m=0}^{n-2} \frac{K}{t_{n-m-1}^\alpha}\|\xi_m\| + \tau\|\xi_{n-1}\|_\alpha + \tau \sum_{m=0}^{n-2} \gamma^{n-m-1}\|\xi_m\|_\alpha,$$

which proves (33). To prove (34), notice that

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

$$\leq \frac{\tau^{1-\alpha}n^{1-\alpha}}{1-\alpha} \max_{0 \leq m \leq n-2} \|\xi_m\| \leq \frac{T^{1-\alpha}}{1-\alpha} \max_{0 \leq m \leq n-2} \|\xi_m\|,$$

and

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

$\square$

11

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

$$\left\| \tau \sum_{m=0}^{n-1} r^{n-m-1}\,(\tau A)\,\xi_m \right\| \leq \tau\,C_s(n) \sum_{m=0}^{n-1} \|\xi_m\|. \tag{35}$$

4 Convergence of the method

Before stating the main theorem, we have to prove an auxiliar result, which is an extension of the convergence theorem (Theorem 4.4 in [7]) to the framework of spaces X_α. Its proof is similar to the one of that theorem, now taking into account the regularization estimates.

For the rest of the section, assume hypotheses 1, 2 and 3, and let $h : [0, T] \to X$ be $h(t) = f(t, u(t))$. The linear problem

$$\begin{cases} v'(t) = Av(t) + h(t), & t \geq 0, \\ v(0) = u_0, \end{cases} \tag{36}$$

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

$$v_{n+1} = r\,(\tau A)\,v_n + \tau E_n\,(\tau A)\,h\,(t_n + \tau \boldsymbol{c}_n), \quad n \geq 1, \tag{37}$$

for some sequence $\{\boldsymbol{c}_n\}_{n=0}^{N-1}$.

Theorem 4.1. *Under hypotheses of Lemma 3.3, let $u : [0, \infty) \to X_\alpha$ be the solution of (36) to be approximated on the interval $[0, T]$ with constant step-size $0 < \tau = T/N < \tau_0$. Assume also that $u \in \mathcal{C}^{p+1}\left([0, T], X_\alpha\right)$, $h \in \mathcal{C}^{p+1}\left([0, T], X\right)$. If v_n is the numerical approximation to $u(t_n)$ given by (37),*

$$\|u(t_n) - v_n\|_\alpha \leq K\,\tau^p \left(\|u^{(p+1)}\|_{\alpha, \infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty \right), \qquad 0 \leq n \leq N. \tag{38}$$

Proof. . Using the notation of [7], it is straightforward to prove that S_G is a semigroup in $X_\alpha \times Y$ and the main result in [12] guarantees that, for the abstract scheme

$$\bar{v}_{n+1} = r\,(\tau A)\,\bar{v}_n + \tau E(\tau A)\,r^n\,(\tau B)\,h, \quad n \geq 1, \tag{39}$$

we get global error of order p,

$$\|u(t_n) - \bar{v}_n\|_\alpha \leq C\,\tau^p \left(\|u^{(p+1)}\|_{\alpha, \infty} + \|h^{(p+1)}\|_\infty \right). \tag{40}$$

Then subtracting (39) from (37),

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

12

$$+ \ \tau \left(E(\tau A) \left(h(t_n + \cdot) - r^n \left(\tau B \right) h \right) \right), \quad \text{for } n \geq 0,$$

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

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

$$(II) = \left\| \tau \sum_{m=0}^{n-1} r^{n-m-1} \left(\tau A \right) E(\tau A) \left(h(t_m + \cdot) - r^m \left(\tau B \right) h \right) \right\|_\alpha.$$

The proof is concluded taking into account (34), the regularisation estimates (32) and (31), and the approximation estimate (16). 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 4.2. *For $0 \leq \alpha < 1$, let $u : [0, T] \to X_\alpha$ be the solution of (1) to be approximated in the interval $[0, T]$ with constant step size $0 < \tau = T/N < \tau_0$. Let us assume hypotheses 1, 2 and 3 and also that $u \in \mathcal{C}^{p+1}\left([0, T], X_\alpha\right)$ and $h \in \mathcal{C}^{p+1}\left([0, T], X\right)$. If u_n is the numerical approximation to $u(t_n)$ given by (19), and $u_0, \cdots, u_{p-1} \in X_\alpha$ are starting values satisfying*

$$\|u(t_n) - u_n\|_\alpha \leq C_0 \, \tau^p, \qquad 0 \leq n \leq p - 1, \tag{41}$$

then,

$$\|u(t_n) - u_n\|_\alpha \leq K \, C_s(n) \, \tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty \right), \qquad 0 \leq n \leq N. \tag{42}$$

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 $0 \leq n \leq N$. Set

$$e_n^p = \begin{cases} \displaystyle\sum_{k=n-p+1}^{n} e_k, & \text{if } \boldsymbol{c}_n = [-p+1, \cdots, 0], \\ \displaystyle\sum_{k=n-p+2}^{n+1} e_k, & \text{if } \boldsymbol{c}_n = [-p+2, \cdots, 1], \end{cases} \tag{43}$$

for $p - 1 \leq n \leq N$. We recall (19) and (37) to get, for $0 \leq n \leq N$,

$$u_{n+p} - v_{n+p} = r\left(\tau A \right) \left(u_{n+p-1} - v_{n+p-1} \right) + \tau E_{n+p-1}(\tau A) \left(\mathbf{f}_{n+p-1} - \mathbf{h}_{n+p-1} \right). \tag{44}$$

By the discrete variation-of-constants formula,

$$u_{n+p-1} - v_{n+p-1} = r\left(\tau A \right)^n \left(u_{p-1} - v_{p-1} \right)$$

13

$$+ \tau \sum_{m=0}^{n-1} r\,(\tau A)^{n-m-1}\, E_{m+p-1}(\tau A)\,(\mathbf{f}_{m+p-1} - \mathbf{h}_{m+p-1})\,, \quad (45)$$

for $0 \le n \le N - p + 1$. We use (32) and the Lipschitz property of f to get

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

and

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

Then, we bound the sum in (45) by combining the previous estimates together with (33). On the other hand, the first term in (45) is bounded using (12), (38) and (41), giving rise to

$$\|u_{n+p-1} - v_{n+p-1}\|_\alpha \le M\,C\,C_s(n)\,\tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty \right)$$
$$+ K\,L\,C_s(n) \sum_{m=0}^{n-1} \left(\tau\,t_{n-m}^{-\alpha} + \tau^{1-\alpha}\,\gamma^{n-m-1} \right) e_{m+p-1}^p\,,$$

for some other constants C and K. Notice that, due to (35), 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 4.1 to get

$$e_{n+p-1} \le M\,C\,C_s(n)\,\tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty \right)$$
$$+ K\,L\,C_s(n) \sum_{m=0}^{n-1} \left(\tau\,t_{n-m}^{-\alpha} + \tau^{1-\alpha}\,\gamma^{n-m-1} \right) e_{m+p-1}^p\,,$$

and the proof concludes by using the version of discrete Gronwall lemma in Lemma 3.2. Notice that the hypothesis (41) is fully taken into account in this step. $\qquad \square$

5 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

$$u_{n+1} = r\,(\tau A)\,u_n + \tau E_n\,(\tau A)\,f\,(t_n + \tau \boldsymbol{c}_n, \mathbf{U}_n)\,, \quad 0 \le n \le p - 2, \qquad (46)$$

where in the first steps c_n is such that $t_n + \tau c_n = [0, \tau, \dots, (p-1)\tau]^T$ and $\mathbf{U}_n = [u_0, \dots, u_{p-1}]^T$ for $0 \le n \le p-2$. It is then necessary to show that the system (46) has a unique solution that approximates the values $u(t_n)$, $1 \le n \le p-1$, within the adequate order. To see this, we rewrite the system as a fixed point equation in X_α^{p-1}

$$U^* = \mathcal{N}(U^*), \tag{47}$$

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

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

where the $\tilde{u}_j$, $0 \le j \le p-1$, are defined recursively by $\tilde{u}_0 = u_0$ and

$$\tilde{u}_{j+1} = r(\tau A)\,\tilde{u}_j + \tau E_n(\tau A)\, f\left(t_j + \tau c_j, [\tilde{u}_0, \tilde{u}_1, \dots, \tilde{u}_j, u_{j+1}, \dots, u_{p-1}]^T\right),$$

for $0 < j < p-1$. To show that (47) 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

$$\tilde{u}_1 - \tilde{v}_1 = \tau E_n(\tau A)\left(f\left(t_0 + \tau c_0, [u_0, \mathbf{U}^T]^T\right) - f\left(t_0 + \tau c_0, [u_0, \mathbf{V}^T]^T\right)\right), \tag{48}$$

so using (32) and the Lipschitz property of f

$$\|\tilde{u}_1 - \tilde{v}_1\|_\alpha \le K L \tau^{1-\alpha} \|\mathbf{U} - \mathbf{V}\|_\alpha, \tag{49}$$

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 $\|\tilde{u}_k - \tilde{v}_k\|_\alpha \le k\,K^k\,L\tau^{1-\alpha}\|\mathbf{U} - \mathbf{V}\|_\alpha$ and proceed by induction for $1 \le k \le p-2$,

$$\|\tilde{u}_{k+1} - \tilde{v}_{k+1}\|_\alpha \le K\|\tilde{u}_k - \tilde{v}_k\|_\alpha + K L \tau^{1-\alpha} \max\left\{\max_{1 \le j \le k}\|\tilde{u}_j - \tilde{v}_j\|_\alpha, \|\mathbf{U} - \mathbf{V}\|_\alpha\right\}$$
$$\le (k+1)\,K^{k+1}\,L\,\tau^{1-\alpha}\|\mathbf{U} - \mathbf{V}\|_\alpha.$$

Therefore,
$$\|\mathcal{N}(\mathbf{U}) - \mathcal{N}(\mathbf{V})\|_\alpha \le p\,K^p\,L\,\tau^{1-\alpha}\|\mathbf{U} - \mathbf{V}\|_\alpha, \tag{50}$$

and the mapping is contractive for sufficiently small τ. In that case, the contractive mapping theorem guarantees that (47) has a unique solution $\mathbf{U} = [u_1 \dots, u_{p-1}]^T$. To conclude, we show that this fixed point approximates the solution with order p. We set $\mathbf{U}(t_n) = [u_0, u(t_1), \dots, u(t_{p-1})]^T$ for $0 \le n \le p-2$. Under the assumptions of Theorem 4.1, the scheme

$$\bar{u}_{n+1} = r(\tau A)\,\bar{u}_n + \tau E_n(\tau A) f(t_n + \tau c_n, \mathbf{U}(t_n)), \quad 0 \le n \le p-2, \tag{51}$$

15

is such that

$$\|\bar{u}_n - u(t_n)\|_\alpha \le C\,K\,\tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty \right), \quad 0 \le n \le p-1. \tag{52}$$

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

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

and by the discrete variation-of-constants formula and (32), we get that $\|u_n - \bar{u}_n\|_\alpha$ is bounded by

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

By the triangle inequality, the previous bound and (52)

$$\|u_n - u(t_n)\|_\alpha \le K\,\tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty \right)$$
$$+ pKL\,\tau^{1-\alpha} \sup_{1 \le k \le p-1} \|u(t_k) - u_k\|_\alpha. \tag{53}$$

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

$$\sup_{1 \le k \le p-1} \|u(t_k) - u_k\|_\alpha \le K\,\tau^p \left(\|u^{(p+1)}\|_{\alpha,\infty} + \|h^{(p)}\|_\infty + \|h^{(p+1)}\|_\infty \right). \tag{54}$$

In practice, to solve (47), we can compute the initial approximation 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.

6 Numerical illustrations

In this section we show numerical results which are obtained with the numerical scheme (19) and different choices of the nodes. We deal with simple PDEs which are integrated by the method of lines. For the spatial discretization, we use finite difference methods. We take $h > 0$ as discretization parameter and we are led to systems of ODEs

$$\begin{cases} u'_h(t) = A_h u_h(t) + f_h(t, u_h(t)), & t \ge 0, \\ u_h(0) = u_{0,h}. \end{cases} \tag{55}$$

We adjust f_h in such a way that the restriction of u to the discrete mesh is the exact solution of (55), so that we consider only due to the time integration. The implementation of the scheme (19) 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 such systems, which matrices typically have a sparse structure that can be exploited.

After the spatial discretization, scheme (19) is applied to the PDE in the discretized form (55). In Section 6 of [7], accurate details on the implementation issues of the linear version (18) 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 $\boldsymbol{c}_n = [-p+1, \ldots, 0]^T \in \mathbb{R}^p$, so $u_{n-p+1}, \ldots, u_n$ 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 $\boldsymbol{c}_n = [-p+2, \ldots, 1]$ and using $u_{n-p+2}, \ldots, \tilde{u}_{n+1}$ to compute u_{n+1}.
3. *Implicit mode.* We set a tolerance TOL and iterate the previous process until two successive iterants $\tilde{u}_{n+1}^{[k]}$ 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 numerical experiments below, the tolerance to calculate implicitly the starting values or for the iteration in the implicit mode has been $TOL = 10^{-14}$.

Example 1.

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

$$\begin{cases} u_t(t,x) = u_{xx}(t,x) + \lambda \left(\int_0^1 u(t,x) \, dx \right) u_x + f(t,x), & 0 \leq t \leq 1, \quad 0 \leq x \leq 1, \\ u(0,x) = u_0(x), & 0 \leq x \leq 1, \\ u(t,0) = 0, & 0 \leq t \leq 1, \\ u(t,1) = 0, & 0 \leq t \leq 1. \end{cases}$$

$$(56)$$

where $f : [0,1] \times X_\alpha \to \mathbb{C}$, $u_0 : [0,1] \to X_\alpha$, $\lambda > 0$. In order to fit the problem in our framework, we take $X = L^2[0,1]$, $A = d^2/dx^2$ with $D(A) = H^2[0,1] \cap H_0^1[0,1]$ and $\alpha = 1/2$, so that $\|\cdot\|_\alpha = \|\cdot\|_{H^1(0,1)}$. We adjust the data u_0 and f in such a way that $u(t,x) = x(1-x)\,e^t$, $0 \leq t, x \leq 1$, is the solution of the problem. 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 $x_j = jh, 1 < j < J$, in $(0,1)$, with $h = 1/(J+1)$. 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
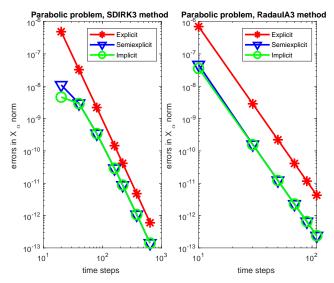
17

Fig. 1 This figure shows the error in the discrete norm $\| \cdot \|_{H^1(0,1)}$ for the parabolic problem (56) with $\lambda = 1$, two different time integrators and $J = 100$.

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.

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

step size	Explicit error	order	Semiexplicit error	order	Implicit 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

As time integrators, we use the scheme (19) with the rational functions of the Runge–Kutta methods SDIRK3 ($p = 4$) and the 3-stages RadauIA3 ($p = 5$) [16]. Notice that the application of these RK methods does not give its classical order of convergence p. According to the main result in [5], we should expect orders $p' = 3.25$ and 4.25 for SDIRK3 and RadauIA3, while Tables 1 and 2 show that the scheme (19) avoids the order reduction, as it is predicted by Theorem 4.2. Both tables show results taking $\lambda = 1$. Notice that both methods satisfy hypothesis 3 because they are A-stable and satisfy $|r_\infty| < 1$.

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 an order of magnitude. However, the implicit mode does not practically improve on the semiexplicit mode,

Table 2 Errors and orders of convergence of Example 1, $\lambda = 1$, Radau IA3

step size	Explicit		Semiexplicit		Implicit	
	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

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, Figure 1 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.
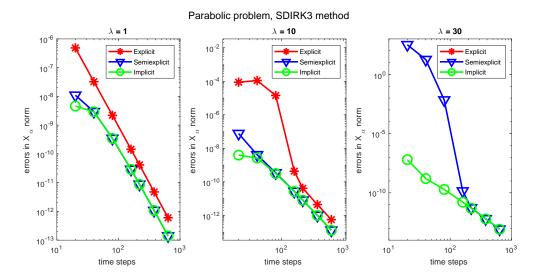

Fig. 2 This figure shows the error in the discrete norm $\| \cdot \|_{H^1(0,1)}$ for the parabolic problem with $\lambda = 1, 10, 30$ and the method SDIRK3.

Figures 2 and 3 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.

Fig. 3 This figure shows the error in the discrete norm $\| \cdot \|_{H^{3/2}(\Omega)}$ for the parabolic problem with $\lambda = 1, 10, 15$ and the method RadauIA3.

Example 2.

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

$$
\begin{cases}
u_t(t,x,y) = \Delta u(t,x,y) + u^2 + f(t,x,y), & 0 \le t \le 1, \quad (x,y) \in \Omega, \\
u(0,x,y) = u_0(x,y), & (x,y) \in \Omega, \\
u(t,x,y) = 0, & 0 \le t \le 1, \quad (x,y) \in \partial\Omega.
\end{cases}
\tag{57}
$$

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 $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$ and $\alpha = 3/4$, so that $\| \cdot \|_\alpha = \| \cdot \|_{H^{3/2}(\Omega)}$. 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 $x_j = jh, 1 < j < J$, $y_k = kh, 1 < k < J$, 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. In this way, since the exact solution is a polynomial of degree two in x and y, there are no spatial errors.

As time integrators, we use again the scheme (19) 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 and 4 show that the scheme (19) does, as it is predicted by Theorem 4.2.

Table 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

Table 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

Example 3.

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

$$\begin{cases} u_t(t,x) = -u_x(t,x) + u - u^3 + f(t,x), & 0 \le t \le 1, \quad 0 \le x \le 1, \\ u(0,x) = u_0(x), & 0 \le x \le 1, \\ u(t,0) = u(t,1), & 0 \le t \le 1, \end{cases} \tag{58}$$

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 $u(t,x) = x^3 \, e^t \, \sin(\pi x) + (1 - e^t)$, $0 \le t, x \le 1$, is the solution of the problem. We fix again number J of uniformly distributed nodes $x_j = jh$, $1 < j < J$, in $(0,1)$, with $h = 1/(J+1)$. The spatial derivatives are approximated by upwind finite differences.

Tables 5 and 6 show the results for the hyperbolic problem with the scheme (19) and the rational function of SDIRK3 and 3-stages RadauIA3, respectively. According to [5], these RK methods have orders $p = 3.5$ and $p = 4.5$, respectively, when applied to this problem. We find the results to be similar to the parabolic one.

Again, we obtain an improvement on the size of the error after the semiexplicit correction for a fixed stepsize and we observe that the order of the methods is in agreement with that being predicted by Theorem 4.2.
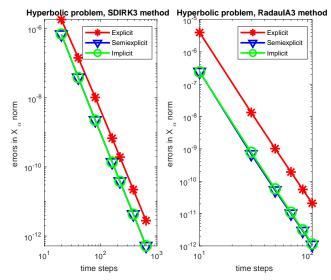
Funding

Fig. 4 This figure shows the error in the discrete norm associated to $\| \cdot \|_{H^1(0,1)}$ for the hyperbolic problem (58) with two different time integrators and $J = 100$.

Table 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

References

[1] S. Abarbanel, D. Gottlieb, and M.H. Carpenter. On the removal of boundary errors caused by Runge–Kutta integration of nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 17:777–782, 1996.

[2] I. Alonso-Mallo. Runge–Kutta methods without order reduction for linear initial boundary value problems. *Numerische Mathematik*, 91:577–603, 2002.

[3] I. Alonso-Mallo and B. Cano. Spectral/Rosenbrock discretizations without order reduction for linear parabolic problems. *Applied Numerical Mathematics*, 47:247–268, 2002.

Table 6 Errors and orders of convergence of Example 3 with RadauIA3, J=100.

step size	Explicit error	order	Semiexplicit error	order	Implicit 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

[4] I. Alonso-Mallo and B. Cano. Efficient time integration of nonlinear partial differential equations by means of Rosenbrock methods. *Mathematics*, 9:1970, 2021.

[5] I. Alonso-Mallo and C. Palencia. Optimal orders of convergence for Runge–Kutta methods and linear initial boundary value problems. *Applied Numerical Mathematics*, 44:1–19, 2003.

[6] C. Arranz-Simón, B. Cano, and C. Palencia. Rational methods for abstract linear initial boundary value problems without order reduction. Submitted for publication.

[7] C. Arranz-Simón and C. Palencia. Rational methods for abstract linear inhomogeneous problems without order reduction. *SIAM Journal on Numerical Analysis*, 63(1):422–436, 2025.

[8] A. Biswas, D.I. Ketcheson, S. Roberts, B. Seibold, and D. Shirokoff. Explicit Runge–Kutta methods that alleviate order reduction. *SIAM Journal on Numerical Analysis*, 63(4):1398–1426, 2025.

[9] A. Biswas, D.I. Ketcheson, B. Seibold, and D. Shirokoff. Design of DIRK schemes with high weak stage order. *Communications in Applied Mathematics and Computational Science*, 18(1):1–28, 2023.

[10] A. Biswas, D.I. Ketcheson, B. Seibold, and D. Shirokoff. Algebraic structure of the weak stage order conditions for Runge–Kutta methods. *SIAM Journal on Numerical Analysis*, 62(1):48–72, 2024.

[11] P. Brenner and V. Thomée. Stability and convergence rates in L_p for certain difference schemes. *Mathematica Scandinavica*, 26:5–23, 1970.

[12] P. Brenner and V. Thomée. On rational approximations of semigroups. *SIAM Journal on Numerical Analysis*, 16(4):683–694, 1979.

[13] M.P. Calvo, J. Frutos, and J. Novo. An efficient way to avoid the order reduction of linearly implicit Runge–Kutta methods for nonlinear IBVPs. In K. Antreich, R. Bulirsch, A. Gilg, and P. Rentrop, editors, *Mathematical Modelling, Simulation and Optimization of Integrated Circuits*, International Series of Numerical

Mathematics, vol. 146, pages 321–332. Birkhäuser Verlag, Basel, 2003.

[14] M.P. Calvo and C. Palencia. Avoiding the order reduction of Runge–Kutta methods for linear initial boundary value problems. *Mathematics of Computation*, 71(240):1529–1543, 2001.

[15] C. González and C. Palencia. Stability of Runge–Kutta methods for abstract time-dependent parabolic problems: the Hölder case. *Mathematics of Computation*, 68(225):73–89, 1999.

[16] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, 1991.

[17] S.L. Keeling. Galerkin/Runge–Kutta discretizations for semilinear parabolic equations. *SIAM Journal on Numerical Analysis*, 27(2):394–418, 1990.

[18] J. Lang and J.G. Verwer. ROS3P: An accurate third-order Rosenbrock solver designed for parabolic problems. *BIT Numerical Mathematics*, 41:731–738, 2001.

[19] C. Lubich and A. Ostermann. Linearly implicit time discretizations of nonlinear parabolic equations. *IMA Journal of Numerical Analysis*, 15:555–583, 1995.

[20] A. Lunardi. *Analytic Semigroups and Optimal Regularity in Parabolic Problems*. Springer, 2012.

[21] S. McKee. Generalised discrete Gronwall lemmas. *Zeitschrift für Angewandte Mathematik und Mechanik*, 62:429–434, 1982.

[22] A. Ostermann and M. Roche. Runge–Kutta methods for partial differential equations and fractional orders of convergence. *Mathematics of Computation*, 59(200):403–420, 1992.

[23] A. Ostermann and M. Roche. Rosenbrock methods for partial differential equations and fractional orders of convergence. *SIAM Journal on Numerical Analysis*, 30(4):1084–1098, 1993.

[24] D. Pathria. The correct formulation of intermediate boundary conditions for Runge–Kutta time integration of initial boundary value problems. *SIAM Journal on Scientific Computing*, 18:1255–1266, 1997.

[25] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Springer, 2012.

[26] J.M. Sanz-Serna. *Diez Lecciones de Cálculo Numérico*. Universidad de Valladolid, 1998.

[27] J.M. Sanz-Serna, J.G. Verwer, and W.H. Hundsdorfer. Convergence and order reduction of Runge–Kutta schemes applied to evolutionary problems in partial

differential equations. *Numerische Mathematik*, 50:405–418, 1986.

[28] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators.* North Holland, 1978.