Superconvergence of the Strang splitting when using the Crank-Nicolson scheme for parabolic PDEs with Dirichlet and oblique boundary conditions

Guillaume Bertoli, Christophe Besse, Gilles Vilmart

Superconvergence of the Strang splitting when using the Crank-Nicolson scheme for parabolic PDEs with Dirichlet and oblique boundary conditions

Guillaume Bertoli[1], Christophe Besse[2], and Gilles Vilmart[1]

April 20, 2021

Abstract

We show that the Strang splitting method applied to a diffusion-reaction equation with inhomogeneous general oblique boundary conditions is of order two when the diffusion equation is solved with the Crank-Nicolson method, while order reduction occurs in general if using other Runge-Kutta schemes or even the exact flow itself for the diffusion part. We prove these results when the source term only depends on the space variable, an assumption which makes the splitting scheme equivalent to the Crank-Nicolson method itself applied to the whole problem. Numerical experiments suggest that the second order convergence persists with general nonlinearities.

Key words. Strang splitting, Crank-Nicolson, diffusion-reaction equation, nonhomogeneous boundary conditions, order reduction

AMS subject classifications. 65M12, 65L04

1 Introduction

We consider a parabolic semilinear differential problem, in a smooth bounded domain Ω in $\mathbb{R}^d$ in dimension $d \geq 1$, for $t \in [0, T]$, of the form

$$\begin{aligned} \partial_t u(x,t) &= Du(x,t) + f(x, u(x,t)) \quad \text{in } \Omega \times (0,T], \\ Bu(x,t) &= b(x) \quad \text{on } \partial\Omega \times (0,T], \\ u(x,0) &= u_0(x) \quad \text{in } \Omega, \end{aligned} \tag{1.1}$$

where, for $1 < p < \infty$, $D : W^{2,p}(\Omega) \to L^p(\Omega)$ is a linear diffusion operator and $f : L^p(\Omega) \to L^p(\Omega)$ is a possibly nonlinear source term. The operator B represents boundary conditions of type Dirichlet, Neumann or Robin. When time-discretizing a problem of this form, it can be advantageous to use a splitting method in order to divide the problem (1.1) into two parts: the source equation

$$\partial_t u(x,t) = f(x, u(x,t)) \quad \text{in } \Omega \times (0,T], \tag{1.2}$$

[1]Université de Genève, Section de mathématiques, UNI DUFOUR, 24, rue du Général Dufour, Case postale 64, 1211 Genève 4, Switzerland, Guillaume.Bertoli@unige.ch, gilles.vilmart@unige.ch

[2]Institut de Mathématiques de Toulouse, U.M.R CNRS 5219, Université de Toulouse, CNRS, UPS IMT, 118 Route de Narbonne, 31062 Toulouse Cedex 9, France, christophe.besse@math.univ-toulouse.fr

and the diffusion equation

$$\partial_t u(x,t) = Du(x,t) \quad \text{in } \Omega \times (0,T], \qquad Bu(x,t) = b(x) \quad \text{on } \partial\Omega \times (0,T]. \tag{1.3}$$

We use respectively the notations ϕ_t^f and ϕ_t^D to denote the exact flows of the subproblems (1.2) and (1.3). The advantage of this subdivision is that equations (1.2) and (1.3) can often be solved more efficiently than the main problem (1.1). The classical splitting method used to approximate the main problem (1.1) is the Strang splitting. Starting from an arbitrary initial datum u_n, one step of the Strang splitting method, with a time step $\tau > 0$, applied to equation (1.1) is given by

$$u_{n+1} = \phi_{\frac{\tau}{2}}^f \circ \phi_\tau^D \circ \phi_{\frac{\tau}{2}}^f (u_n). \tag{1.4}$$

Interchanging the roles of the diffusion part and nonlinear part, it is also possible to define one step of the Strang splitting method as

$$u_{n+1} = \phi_{\frac{\tau}{2}}^D \circ \phi_\tau^f \circ \phi_{\frac{\tau}{2}}^D (u_n). \tag{1.5}$$

In both cases, we start the procedure with the initial condition u_0 of the parabolic problem (1.1). Both methods (1.4) and (1.5) are formally of order of accuracy two. However, a reduction of order occurs in general, particularly for inhomogeneous boundary conditions, has observed in [10] and [11]. A suitable correction of the splitting algorithm has been proposed in [5], to avoid order reduction phenomenon. In [2], an alternate correction was proposed that depends only on the flow $\phi_{\frac{\tau}{2}}^f$ and facilitates the calculation of the correction. In this paper, we will however not use the techniques developed in [5], [2]. We prove that when the Crank-Nicolson scheme is used to solve the diffusion equation (1.3) in the splitting (1.4), there is no reduction of order away from a neighbourhood of $t = 0$. This superconvergence property appears specific to the Crank-Nicolson scheme and when another Runge-Kutta method or even the exact flow itself is used to approximate the diffusion subproblem, the order reduction of the splitting (1.4) is not avoided. We denote by $\phi_t^{D,CN}$ the numerical flow of the Crank-Nicolson method for the diffusion problem (1.3). We obtain the following splitting method, where ϕ_τ^D has been replaced by $\phi_t^{D,CN}$ in the splitting (1.4),

$$u_{n+1} = \phi_{\frac{\tau}{2}}^f \circ \phi_\tau^{D,CN} \circ \phi_{\frac{\tau}{2}}^f (u_n). \tag{1.6}$$

We prove that the splitting (1.6) has no order reduction when the nonlinearity $f = f(x)$ only depends on the space variable. More precisely, we prove in this case the following exact representation of the error at time $t_n = n\tau$,

$$u_n - u(t_n) = \left(r(\tau A)^n - \mathrm{e}^{n\tau A}\right) A^{-1}(Du_0 + f), \tag{1.7}$$

where A is the restriction of the operator D to $\mathcal{D}(A) = \{u \in W^{2,p}(\Omega)\ ;\ Bu = 0 \text{ on } \partial\Omega\}$, the set of functions satisfying the homogeneous boundary condition $Bu(x) = 0$ on the boundary $\partial\Omega$, and where $r(z) = (1+\frac{z}{2})/(1-\frac{z}{2})$ is the stability function of the Crank-Nicolson scheme.

As seen in [13, Theorem 4.2 and Theorem 4.4], for the simplified case where f and b are both zeros in (1.1), that is for $\partial_t u(x,t) = Au(x,t)$, second order convergence results for A-stable methods (see [8, Chapter IV.3]) usually require $u_0 \in \mathcal{D}(A^2)$. In contrast, L-stable methods are second order convergent outside a neighbourhood of the origin even if $u_0 \in L^p(\Omega)$. The Crank-Nicolson scheme, although it is not L-stable but only A-stable,

is second order convergent outside a neighbourhood of the origin for $u_0 \in \mathcal{D}(A)$ (see [9, Theorem 2.1]). Maximal parabolic regularity of A-stable Runge-Kutta methods is studied in [12]. To the best of our knowledge, there exists no result in the literature which proves already that the Crank-Nicolson scheme is second order convergent outside a neighbourhood of the origin when applied to a nonlinear parabolic problem with inhomogeneous boundary conditions and an initial condition $u_0 \in \{u \in W^{2,p}(\Omega)\ ;\ Bu = b \text{ on } \partial\Omega\}$.

This is a direct consequence of the convergence result of the splitting (1.6) since for $f = f(x)$, the splitting with Crank-Nicolson (1.6) is equal to the Crank-Nicolson scheme (3.1) applied to the whole problem (1.1).

We also provide numerical experiments in the case where f depends on the solution u, in which case the splitting (1.6) is different from the Crank-Nicolson scheme, and observe that the splitting method (1.6) remains second order convergent in this case. Note also that the superconvergence property does not hold for the other splitting methods given by $u_{n+1} = \phi^{D,CN}_{\frac{\tau}{2}} \circ \phi^{f}_{\tau} \circ \phi^{D,CN}_{\frac{\tau}{2}}(u_n)$ nor does this splitting preserves stationary states for $f = f(x)$.

This paper is organized as follows. In Section 2, we describe an appropriate analytical framework for the analysis, where D is chosen to be a second order elliptic operator and B a first order differential operator corresponding to Dirichlet or oblique boundary conditions. In Section 3, we describe precisely the algorithm corresponding to the splitting (1.6). In Section 4, we restrict ourselves to the case where f does not depend on the solution u. We first provide an exact representation of the local error (Proposition 4.10). We then prove formula (1.7) for the global error (Theorem 4.3) and additionally conclude that the stationary states are preserved by the splitting method (1.6) (Remark 4.6). In Section 5, we provide numerical experiments to illustrate the properties of the splitting method (1.6) compared to several natural splitting methods in dimensions one and two for constant and nonlinear terms.

2 Analytical framework

We follow closely the framework and the notations given in the book [14, Chapter 3]. Let Ω be an open bounded subset of $\mathbb{R}^d$ with C^2 boundary $\partial\Omega$ and dimension $d \geq 1$. For $1 < p < \infty$, let $D : W^{2,p}(\Omega) \to L^p(\Omega)$ be a second order differential operator,

$$D = \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}\right) + \sum_{i=1}^{d} b_i(x)\frac{\partial}{\partial x_i},$$

where a_{ij} and b_i are real continuous functions on $\overline{\Omega}$. We assume that the matrix $[a_{ij}(x)]_{ij}$ is symmetric and uniformly positive definite on $\overline{\Omega}$, i.e for all $x \in \overline{\Omega}$ and for all $\xi \in \mathbb{R}^d$, $\xi^T[a_{ij}(x)]_{ij}\xi \geq c\xi^T\xi$, where $c > 0$ is independent of x. The source term $f : L^p(\Omega) \to L^p(\Omega)$ is assumed continuously differentiable and we assume that the initial conditions u_0 belongs to $W^{2,p}(\Omega)$.

The linear operator $B : W^{2,p}(\Omega) \to W^{1,p}(\Omega)$ is either defined for all $u \in W^{2,p}(\Omega)$ as $Bu = u$, which corresponds to Dirichlet boundary conditions or it is a first order differential operator defined for all $u \in W^{2,p}(\Omega)$ as

$$Bu(x) = \sum_{i=1}^{d} \beta_i(x)\frac{\partial u(x)}{\partial x_i} + \alpha(x)u(x),$$

where β_i and α are uniformly continuous and differentiable on $\overline{\Omega}$, which corresponds to Robin boundary conditions. We assume that α is not zero everywhere on $\partial\Omega$. The degenerate case, where α is the zero function, corresponding to Neumann boundary conditions, is discussed in Remark 2.1 below. If B is a first order operator, we assume that the uniform nontangentiality condition is satisfied for all $x \in \partial\Omega$,

$$\left| \sum_{i=1}^{d} \beta_i(x) \vec{n}_i(x) \right| \geq c,$$

where $c > 0$ is independent of x and where $\vec{n}(x)$ is the outwardly normal unit vector. On the boundary $\partial\Omega$, $Bu|_{\partial\Omega}$ is the trace of $Bu \in W^{1,p}(\Omega)$ on $\partial\Omega$ and is therefore an element of $L^p(\partial\Omega)$. To avoid heavy notations, we simply write $Bu = b$, on $\partial\Omega$. We assume that b is a twice continuously differentiable function on the boundary $\partial\Omega$. To avoid stiffness of the solution or boundary layers at the initial time $t = 0$, we assume in addition that the initial condition $u_0 \in W^{2,p}(\Omega)$ satisfies the boundary conditions $Bu_0(x) = b(x)$ on $\partial\Omega$.

The space $\{u \in W^{2,p}(\Omega) \ ; \ Bu = b \text{ on } \partial\Omega\}$ is difficult to handle since it is not a linear subspace of $L^p(\Omega)$ if b is not the zero function on $\partial\Omega$. Therefore, we provide a reformulation of the problem (1.1) with homogeneous boundary conditions. We choose a function $z \in W^{2,p}(\Omega)$ which satisfies the boundary conditions $Bz = b$ on $\partial\Omega$. Such a function always exists with the assumptions that we made on B, b and $\partial\Omega$. We define the function $\tilde{u} = u - z$, which satisfies the following differential problem with homogeneous boundary conditions,

$$\begin{aligned} \partial_t \tilde{u}(x,t) &= D\tilde{u}(x,t) + f(x, \tilde{u}(x,t) + z(x)) + Dz(x) \ \text{ in } \Omega \times (0,T], \\ B\tilde{u}(x,t) &= 0 \ \text{ on } \partial\Omega \times (0,T], \\ \tilde{u}(x,0) &= u_0(x) - z(x) \quad \text{in } \Omega. \end{aligned} \tag{2.1}$$

We define the operator $(A, \mathcal{D}(A))$ as the restriction of the operator D to the domain $\mathcal{D}(A) = \{u \in W^{2,p} \ ; \ Bu = 0 \text{ on } \partial\Omega\}$, i.e. $Au = Du$ for all $u \in \mathcal{D}(A)$. The operator A therefore includes the homogeneous boundary conditions in its domain. Under the above assumptions, the operator A is a closed densely defined linear operator satisfying the two following properties (see [14, Theorem 3.1.13] and [18, page 92]):

1. The resolvent set of A, $\rho(A) = \{\lambda \in \mathbb{C} \ ; \ \lambda I - A \text{ is an isomorphism}\}$, contains the closure of the set $\Sigma_\theta = \{z \in \mathbb{C} \ ; \ z \neq 0, \ |\arg(z)| < \pi - \theta\}$, where $\theta \in (0, \frac{\pi}{2})$ is fixed,

$$\rho(A) \supset \overline{\Sigma}_\theta. \tag{2.2}$$

2. For all $\lambda \in \Sigma_\theta$, the resolvent of A, $R(\lambda, A) = (\lambda I - A)^{-1}$, satisfies the following bound for the operator norm,

$$\|R(\lambda, A)\| \leq \frac{M}{|\lambda|}, \tag{2.3}$$

where $M \geq 1$.

Note that, since $0 \in \rho(A)$ by (2.2), the operator A is invertible and A^{-1} is bounded. The operator A is therefore the infinitesimal generator of an analytic uniformly bounded semigroup denoted e^{tA}, given by

$$\mathrm{e}^{tA} = \frac{1}{2\pi i} \int_\Gamma \mathrm{e}^{zt} R(z, A) \mathrm{dz},$$

where Γ is the boundary of Σ_θ with imaginary part increasing along Γ (see [17, Theorems 2.5.2 and 1.7.7]).

Since the homogeneous boundary conditions are included in the domain of A, we have the following reformulation of the problem (2.1),

$$\partial_t \tilde{u}(t) = A\tilde{u}(t) + f(\tilde{u}(t)+z) + Dz, \quad \text{for} \quad t \in (0,T], \qquad \tilde{u}(0) = u_0 - z, \tag{2.4}$$

where we omit the variable x in the notations, i.e. $\tilde{u}(t)$ denotes $\tilde{u}(x,t)$ and similarly for z, u_0, and f. This equation has a solution $\tilde{u} \in C^1([0,T], L^p(\Omega)) \cap C([0,T], \mathcal{D}(A))$ if T is sufficiently small (see [14, Proposition 7.1.10]), given by Duhamel formula,

$$\tilde{u}(t) = \mathrm{e}^{tA}(u_0 - z) + \int_0^t \mathrm{e}^{(t-s)A}(f(\tilde{u}(s)+z) + Dz)\mathrm{ds}.$$

For $\tau > 0$, denoting $t_n = n\tau$, since $\tilde{u}(t_n) \in \mathcal{D}(A)$, we have

$$u(t_n+\tau) = z + \mathrm{e}^{\tau A}(u(t_n) - z) + \int_0^\tau \mathrm{e}^{(\tau-s)A}(f(u(t_n+s)) + Dz)\mathrm{ds}. \tag{2.5}$$

Note that the above reformulation corresponds to the usual lifting methodology to handle inhomogeneous boundary conditions. A more general Banach space framework, that includes e.g. a bi-Laplacian diffusion problem, will be discussed in Remark 4.5.

Remark 2.1. *Alternatively, on can consider pure Neumann boundary conditions, which corresponds to*

$$B = \sum_{i=1}^d \beta_i(x) \frac{\partial}{\partial x_i}.$$

In this particular case, we consider the operator

$$D = \sum_{i,j=1}^d \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}\right) + \sum_{i=1}^d b_i(x)\frac{\partial}{\partial x_i} + c(x),$$

where $c(x) < c_{max} < 0$ with $c(x)$ uniformly continuous. If all the additional assumptions on D and B are satisfied, then the operator A, defined as above, satisfied the assumptions (2.2) *and* (2.3) *as required. Alternatively, if c is the null function, we can also consider the subspace of functions with zero average on Ω.*

Remark 2.2. *Although we choose to present Theorem 4.3 in $L^p(\Omega)$ to simplify the presentation, it remains true for general complex separable Banach spaces and suitable analytic semigroups. In this case, the hypotheses on A and f are described in Section 4, Remark 4.5.*

3 The splitting method based on the Crank-Nicolson scheme

We describe precisely the algorithm for the splitting (1.6) with a time step $\tau > 0$. The same time step τ is used for the Strang splitting (1.4) and for the Crank-Nicolson method used to approximate (1.3). One step of the Crank-Nicolson scheme with a time step τ and an initial condition u_0 is given by the solution u_1 of the following equation,

$$\frac{u_1(x) - u_0(x)}{\tau} = D\frac{u_1(x) + u_0(x)}{2} \quad \text{in } \Omega, \quad B\frac{u_0(x) + u_1(x)}{2} = b(x) \quad \text{on } \partial\Omega. \tag{3.1}$$

We denote by $u_1 = \phi_\tau^{D,CN}(u_0)$ the solution of the problem (3.1). Numerically, as explained in [1] (see also Remark 5.1), it is advantageous to save a linear system resolution and to define $v_1 = \frac{u_1+u_0}{2}$ and to first find the solution v_1 of

$$2\frac{v_1(x) - u_0(x)}{\tau} = Dv_1(x) \quad \text{in } \Omega, \qquad Bv_1(x) = b(x) \ \text{ on } \ \partial\Omega. \tag{3.2}$$

One step of the Crank-Nicolson method is then given by

$$u_1 = 2v_1 - u_0. \tag{3.3}$$

One step of the splitting method (1.6) is therefore given by the following algorithm.

Algorithm 3.1 (Main algorithm for the splitting (1.6))**.**

1. *Given $u_n(x)$, compute the solution $w(x, \frac{\tau}{2})$ of $\partial_t w(x,t) = f(x, w(x,t))$ in Ω, $w(x,0) = u_n(x)$.*
2. *Compute the solution $v(x)$ of $(I - \frac{\tau}{2}D)v(x) = w(\frac{\tau}{2})$ in Ω with $Bv(x) = b(x)$ on $\partial\Omega$. Compute $\hat{v}(x) = 2v(x) - w(x, \frac{\tau}{2})$.*
3. *Compute the solution $\hat{w}(x, \frac{\tau}{2})$ of $\partial_t \hat{w}(x,t) = f(x, \hat{w}(x,t))$ in Ω, $\hat{w}(x,0) = \hat{v}(x)$. Define $u_{n+1}(x) = \hat{w}(x, \frac{\tau}{2})$.*

Note that one step of the Algorithm 3.1 is given by the solution u_{n+1} of the following problem,

$$\begin{aligned} &\frac{\phi^f_{-\frac{\tau}{2}}(u_{n+1}) - \phi^f_{\frac{\tau}{2}}(u_n)}{\tau} = D\frac{\phi^f_{-\frac{\tau}{2}}(u_{n+1}) + \phi^f_{\frac{\tau}{2}}(u_n)}{2} \quad \text{in } \Omega, \\ &B\frac{\phi^f_{-\frac{\tau}{2}}(u_{n+1}) + \phi^f_{\frac{\tau}{2}}(u_n)}{2} = b \quad \text{on } \partial\Omega, \end{aligned} \tag{3.4}$$

which can be solved using a linear solver for computing $\hat{v} = \phi^f_{-\tau/2}(u_{n+1})$, then $u_{n+1} = \phi_{\tau/2}(\hat{v})$, combined for instance with a finite element discretization for the spatial discretization.

Remark 3.2. *Similarly to [5], the auxiliary function z is only used as a tool to introduce homogeneous boundary conditions in the analysis. It is never used throughout the algorithm 3.1 as seen in* (3.4)*.*

4 Convergence analysis for a solution independent source term

In this section, we give an estimate of the error of the splitting (1.6) when the source term $f = f(x)$ only depends on the space variable x. We assume $f \in L^p(\Omega)$, where again $1 < p < \infty$. We can write the parabolic problem (1.1) as

$$\begin{aligned} \partial_t u(x,t) &= Du(x,t) + f(x) \ \text{ in } \Omega \times (0,\infty), \\ Bu(x,t) &= b(x) \ \text{ on } \partial\Omega \times (0,\infty), \\ u(x,0) &= u_0(x) \quad \text{in } \Omega, \end{aligned} \tag{4.1}$$

Since we are interested in the semi-discretization in time, for brevity of notations, we write, from now on, $u(t)$ and f instead of $u(x,t)$ and $f(x)$. We denote by $r(y)$ the stability function of the Crank-Nicolson method given by,

$$r(y) = \frac{1+\frac{y}{2}}{1-\frac{y}{2}}. \tag{4.2}$$

We recall that, for any Runge-Kutta method applied with a time step $\tau > 0$ to the Dahlquist scalar test equation (see [8, Definition IV.2.1])

$$\frac{dx}{dt}(t) = \lambda x(t), \quad x(0) = x_0,$$

with $\lambda \in \mathbb{C}$, one obtains the induction $x_{n+1} = R(h\lambda)x_n$, where $R : \mathbb{C} \to \mathbb{C}$ is a rational approximation of the exponential that we call the stability function of the Runge-Kutta method. For a fixed degree of the numerator and denominator, the rational approximations that have the highest order of approximation are called the Padé approximations of the exponential (cf. [8, Chapter IV.3]) and efficient Runge-Kutta methods are typically constructed to have a stability function equal to a Padé approximation. The approximation $r(z)$ that corresponds to the Crank-Nicolson stability function is the (1,1)-Padé approximation (4.2). An A-stable method is by definition a Runge-Kutta method whose stability function verifies $|R(y)| \leq 1$ for all $y \in \mathbb{C}^- = \{z \in \mathbb{C} \ ; \ \Re(z) \leq 0\}$. We recall that the only Padé approximations that verify this property are the (j,k)-Padé approximations with $k \leq j \leq k+2$ (see [8, Theorem 4.12]).

Remark 4.1. *Consider the following ordinary differential equation*

$$\frac{dx}{dt}(t) = \lambda x(t) + b, \quad x(0) = x_0,$$

with $\lambda, b \in \mathbb{C}$. A Runge-Kutta method with stability function $R(y)$ yields

$$x_{n+1} = R(h\lambda)x_n + \frac{R(h\lambda)-1}{h\lambda}b. \tag{4.3}$$

The Strang splitting $x_{n+1} = \phi^b_{\frac{h}{2}} \circ \phi^{\lambda,RK}_h \circ \phi^b_{\frac{h}{2}}(x_n)$, where $\phi^b_{\frac{h}{2}}(x) = x + \frac{h}{2}b$ and $\phi^{\lambda,RK}_h(x) = R(h\lambda)x$ yields

$$x_{n+1} = R(h\lambda)(x_n + \frac{h}{2}b) + \frac{h}{2}b. \tag{4.4}$$

Equations (4.3) *and* (4.4) *coincide if and only if*

$$R(y) = r(y) = \frac{1+\frac{y}{2}}{1-\frac{y}{2}},$$

which is the stability function (4.2) *of the Crank-Nicolson scheme. Therefore, the only Runge-Kutta methods for which equations* (4.3) *and* (4.4) *coincide are the ones with a stability function equal to $r(y)$.*

Remark 4.2. *One step of the Crank-Nicolson scheme applied to the parabolic problem* (1.1) *is given by the solution u_{n+1} of the following equation,*

$$\frac{u_{n+1}-u_n}{\tau} = D\frac{u_{n+1}+u_n}{2} + \frac{f(u_{n+1})+f(u_n)}{2} \quad \text{in } \Omega,$$

$$B\frac{u_{n+1}+u_n}{2} = b \quad \text{on } \partial\Omega. \tag{4.5}$$

We observe that, for $f = f(x)$, formula (4.5) *is equal to formula* (3.4) *which describes one step of the splitting* (1.6)*. Therefore, we deduce that for $f = f(x)$, the Strang splitting with Crank-Nicolson* (1.6) *is equivalent to the Crank-Nicolson scheme itself applied to the whole problem* (4.1)*.*

4.1 Main results

The main result of this paper is the following theorem, which states that the splitting (1.6) yields a method of order 2 of accuracy away form a neighbourhood of the origin $t = 0$ on unbounded intervals ($t > 0$). In contrast, in a neighbourhood of zero, the order of accuracy reduces to one.

Theorem 4.3. *Let $e_n = u_n - u(t_n)$ be the global error of the splitting method* (1.6) *applied to the parabolic problem* (4.1)*, where $u_0 \in W^{2,p}(\Omega)$ and satisfies $Bu_0 = b$ on $\partial\Omega$. Then, for all $n \geq 0$, the global error e_n is given by*

$$e_n = \left(r(\tau A)^n - \mathrm{e}^{n\tau A}\right) A^{-1}(Du_0 + f) \tag{4.6}$$

and it satisfies the bound

$$\|e_n\|_{L^p(\Omega)} \leq \frac{C\tau^2}{t_n},$$

where C is a constant independent of τ, n, and $t_n = n\tau$.

Remark 4.4. *The estimate of Theorem 4.3 could be used to derive a fully discrete estimate of the form*

$$\|u_n^h - u(t_n)\| \leq C(\tau^2/t_n + h^p)$$

where u_n^h denotes a standard finite element discretization of order p on a spatial mesh with size h. The idea of the proof is to rely on the triangle inequality

$$\|u_n^h - u(t_n)\| \leq C\|u_n^h - u^h(t_n)\| + \|u^h(t_n) - u(t_n)\|$$

where $u^h(t_n)$ denotes the semi-discretization in space at time t_n. Then, observe that the estimate of Theorem 4.1 also holds for $u_n^h - u^h(t_n)$ uniformly with respect to the spatial mesh size h, using that the space discretization of the diffusion operator A is a self-adjoint operator that satisfies assumptions analogous to (2.2) *and* (2.3)*, see e.g. [3].*

Remark 4.5. *One can extend to more abstract problems the framework described in Section 2 and consider for example, a problem where D is the bi-Laplacian with appropriate boundary conditions. This formulation also include Galerkin approximation of parabolic problems (see [18]). More generally, for a complex separable Banach space X, we need $A : \mathcal{D}(A) \to X$, to be a closed densely defined linear operator satisfying the conditions* (2.2) *and* (2.3) *or equivalently we require A to be the infinitesimal generator of a uniformly bounded analytic semigroup ([17, Theorems 2.5.2]). We require $f : X \to X$ to be continuously differentiable. The operator $D : \mathcal{D}(D) \to X$ is assumed to be an extension of A, i.e. $\mathcal{D}(A) \subset \mathcal{D}(D)$. We require $z \in \mathcal{D}(D)$ and $u_0 - z \in \mathcal{D}(A)$. Then, the problem* (2.4)

has a unique solution $\tilde{u} \in C^1([0,T],X) \cap C([0,T],\mathcal{D}(A))$ *if* T *is sufficiently small (see [14, Proposition 7.1.10]). The main problem* (1.1)*, for* $u = \tilde{u} + z$ *becomes*

$$\partial_t u(t) = Du(t) + f(u(t)), \quad \text{for } t \in (0,T], \quad u(t) - z \in \mathcal{D}(A) \quad \text{for } t \in (0,T], \quad u(0) = u_0.$$

For this problem, the splitting (1.6)*, is defined as above for* (1.1)*, where the boundary conditions of the problem* (1.1) *have been replaced by* $u(t) - z \in \mathcal{D}(A)$*. Under all those conditions, for* $f = f(x)$*, the convergence analysis (Theorem 4.3) remains true, i.e. the splitting method* (1.6) *has no reduction of order away from the origin and the global error is given by formula* (4.6)*.*

Remark 4.6. *Since the splitting method* (1.6) *applied to the parabolic problem* (4.1) *is equivalent to the Crank-Nicolson scheme, it must preserve exactly the stationary states. Precisely, for* $Du_0 + f = 0$*, the error satisfies* $e_n = 0$*. For general nonlinearities that depend on the solution* $f = f(u)$*, the stationary states are not preserved. This is not surprising since in [15], it is proved that a method that is not a Butcher-series method, like the splitting methods, and which is invariant by an affine change of variable (precisely affine-equivalent), cannot preserve all stationary states.*

4.2 Preliminaries

We give some basic properties of the rational approximation $r(y)$ in (4.2). The following properties are obtained with straightforward computations.

Lemma 4.7. *We have the following formulas,*

$$r(y) + 1 = \frac{2}{1 - \frac{y}{2}}, \qquad r(y) - 1 = \frac{y}{1 - \frac{y}{2}}, \tag{4.7}$$

and

$$(r(y) - \mathrm{e}^y) = \frac{y}{2}(r(y) + 1) - (\mathrm{e}^y - 1). \tag{4.8}$$

The rational approximation $r(y)$ satisfies the following result, in the context of homogeneous parabolic problems, which is proved in the Appendix A (see [9, Theorem 2.1] for a proof in a more general case).

Theorem 4.8. *For* $u_0 \in \mathcal{D}(A)$*, where* A *satisfies the assumptions of Section 2 and* $r(y)$ *is defined in* (4.2)*, we have the following error estimate,*

$$\|(r(\tau A)^n - \mathrm{e}^{\tau n A})u_0\|_{L^p(\Omega)} \leq \frac{C\tau^2}{t_n}\|Au_0\|_{L^p(\Omega)},$$

where C *is a constant independent of* u_0*,* τ*,* n *and* $t_n = n\tau$*.*

Note that this corresponds to an estimate of the splitting (1.6) for the specific case where the problem is homogeneous, i.e. $f = 0$, and with homogeneous boundary conditions. In what follows, we give an exact representation of the numerical solution of the splitting (1.6) in term of the operator A. We recover homogeneous boundary conditions with an appropriate change of variable. Let $z \in W^{2,p}(\Omega)$ be the same function chosen in (2.5) and satisfying $Bz = b$ on $\partial\Omega$. Defining $\tilde{v}_1 = v_1 - z$, we rewrite equation (3.2) as follows,

$$2\frac{\tilde{v}_1 - u_0}{\tau} + \frac{2}{\tau}z = A\tilde{v}_1 + Dz,$$

where we recall that the homogeneous boundary conditions are included in the domain of A. This gives the following expression for $\tilde{v}_1$,

$$\tilde{v}_1 = \left(I - \frac{\tau}{2}A\right)^{-1} u_0 + \frac{\tau}{2}\left(I - \frac{\tau}{2}A\right)^{-1} Dz - \left(I - \frac{\tau}{2}A\right)^{-1} z.$$

We denote $u_1 = \phi_\tau^{D,CN}(u_0)$, the solution of the problem (3.1). Therefore, from equation (3.3), we have the following formula,

$$\begin{aligned} u_1 = \phi_\tau^{D,CN}(u_0) &= 2\tilde{v}_1 - u_0 + 2z \\ &= \left(2\left(I - \frac{\tau}{2}A\right)^{-1} - I\right)(u_0 - z) + \tau\left(I - \frac{\tau}{2}A\right)^{-1} Dz + z. \end{aligned}$$

Using the equalities (4.7), we obtain

$$\phi_\tau^{D,CN}(u_0) = z + r(\tau A)(u_0 - z) + (r(\tau A) - I)A^{-1}Dz. \tag{4.9}$$

One step of the splitting (1.6), is denoted by $\mathcal{S}_\tau$, i.e.

$$u_{n+1} = \mathcal{S}_\tau(u_n) := \phi_{\frac{\tau}{2}}^f \circ \phi_\tau^{D,CN} \circ \phi_{\frac{\tau}{2}}^f(u_n).$$

Using the following representation of the exact flow ϕ_t^f for $f = f(x)$,

$$\phi_t^f(u_0) = u_0 + tf, \tag{4.10}$$

and formula (4.9) for $\phi_\tau^{D,CN}$, we obtain the following expression for $\mathcal{S}_\tau(u_n)$:

$$\mathcal{S}_\tau(u_n) = z + r(\tau A)(u_n - z) + (r(\tau A) - I)A^{-1}Dz + \frac{\tau}{2}(r(\tau A) + I)f. \tag{4.11}$$

Remark 4.9. *Take any A-stable Runge-Kutta method and denote by $R(z)$ its stability function. Then if this Runge-Kutta method is used to solve the diffusion equation* (1.3)*, instead of the Crank-Nicolson method, one obtains formula* (4.11) *for the numerical solution with $r(\tau A)$ replaced with $R(\tau A)$. Indeed, let u be the solution of the diffusion equation* (1.3)*. Let $z \in W^{2,p}(\Omega)$ be a function satisfying $Bz = b$ on $\partial\Omega$. Then, we have $\partial_t u = A(u - z + A^{-1}Dz)$ for $t > 0$. Defining $y = u - z + A^{-1}Dz$, we obtain the following equivalent problem,*

$$\partial_t y = Ay \quad \text{for } t > 0, \quad y(0) = u_0 - z + A^{-1}Dz.$$

Applying one step of the Runge-Kutta method with initial condition $y(0)$ and with a time step τ gives,

$$y_1 = R(\tau A)(u_0 - z + A^{-1}Dz).$$

Hence, we obtain the same formula (4.9) *with $r(\tau A)$ replaced with $R(\tau A)$ for the numerical solution of the diffusion* (1.3)*,*

$$u_1 = R(\tau A)(u_0 - z + A^{-1}Dz) + z - A^{-1}Dz = z + R(\tau A)(u_0 - z) + (R(\tau A) - 1)A^{-1}Dz.$$

4.3 Local error

We start with the following proposition that gives an exact representation of the local error.

Proposition 4.10. *The local error* $\delta_{n+1} = \mathcal{S}_\tau(u(t_n)) - u(t_{n+1})$ *of the splitting* (1.6) *satisfies the following identity,*

$$\delta_{n+1} = (r(\tau A) - \mathrm{e}^{\tau A})A^{-1}\mathrm{e}^{t_n A}(Du_0 + f).$$

Proof. We have the following representation of δ_{n+1}:

$$\begin{aligned}\delta_{n+1} &= (r(\tau A) - \mathrm{e}^{\tau A})(u(t_n) - z) + (r(\tau A) - \mathrm{e}^{\tau A})A^{-1}Dz \\ &\quad + \frac{\tau}{2}(r(\tau A) + I)f - A^{-1}(\mathrm{e}^{\tau A} - I)f. \end{aligned} \tag{4.12}$$

From formula (4.8), we obtain,

$$\delta_{n+1} = (r(\tau A) - \mathrm{e}^{\tau A})(u(t_n) - z + A^{-1}Dz + A^{-1}f). \tag{4.13}$$

From equation (2.4), we know that $\tilde{u}(t_n)$ is the solution of the following problem,

$$\partial_t \tilde{u}(t) = A\tilde{u}(t) + f + Dz \quad \text{for } t \in (0, T], \qquad \tilde{u}(0) = u_0 - z.$$

Hence, δ_{n+1} satisfies

$$\delta_{n+1} = (r(\tau A) - \mathrm{e}^{\tau A})A^{-1}\partial_t \tilde{u}(t_n).$$

Using the variation of constant formula, we obtain,

$$\begin{aligned}\partial_t \tilde{u}(t_n) &= A\tilde{u}(t_n) + f + Dz \\ &= A\mathrm{e}^{t_n A}\tilde{u}_0 + A\int_0^{t_n} \mathrm{e}^{(t_n - s)A}(f + Dz)\mathrm{ds} + f + Dz \\ &= A\mathrm{e}^{t_n A}\tilde{u}_0 + (\mathrm{e}^{t_n A} - I)(f + Dz) + f + Dz \\ &= \mathrm{e}^{t_n A}(A\tilde{u}_0 + f + Dz) \\ &= \mathrm{e}^{t_n A}(Du_0 + f)\end{aligned}$$

where we use $A\tilde{u}_0 + Dz = D(\tilde{u}_0 + z) = Du_0$. This concludes the proof. □

Remark 4.11. *Assume that another Runge-Kutta method, with stability function* $R(y)$*, is used to solve the diffusion equation* (1.3) *involved in the splitting method* (1.6)*. Then, from Remark 4.9, we observe that the local error* δ_{n+1} *still satisfies formula* (4.12) *with* $r(\tau A)$ *replaced with* $R(\tau A)$*. However, the representation* (4.13) *is specific to Runge-Kutta methods having* $r(y)$ *as a stability function. Indeed, to find this representation of the local error, we used formula* (4.8) *which is a property satisfied only by the (1,1)-Padé approximation* (4.2)*.*

4.4 Global error

We are now in position to prove Theorem 4.3 for the global error $e_n = u_n - u(t_n)$ of the splitting (1.6). We observe that

$$e_{n+1} = \mathcal{S}_\tau(u_n) - \mathcal{S}_\tau u(t_n) + \mathcal{S}_\tau u(t_n) - u(t_{n+1}) = \mathcal{S}_\tau(u_n) - \mathcal{S}_\tau u(t_n) + \delta_{n+1}.$$

Proof of Theorem 4.3. From formula (4.11) for $\mathcal{S}_\tau(u_n)$, we obtain

$$\mathcal{S}_\tau(u_n) - \mathcal{S}_\tau u(t_n) = r(\tau A)(u_n - u(t_n)).$$

Therefore, we deduce

$$e_{n+1} = r(\tau A)e_n + \delta_{n+1}.$$

Hence, since $\mathrm{e}_0 = 0$, and using Proposition 4.10 for the local error, we obtain

$$\begin{aligned} e_n &= \sum_{k=0}^{n-1} r(\tau A)^{n-k-1}\delta_{k+1} \\ &= \left((r(\tau A) - \mathrm{e}^{\tau A})\sum_{k=0}^{n-1} r(\tau A)^{n-k-1}\mathrm{e}^{k\tau A}\right)A^{-1}(Du_0 + f) \\ &= \left(r(\tau A)^n - \mathrm{e}^{n\tau A}\right)A^{-1}(Du_0 + f). \end{aligned}$$

Therefore, using that A satisfies the hypotheses of Theorem 4.8, and since $A^{-1}(Du_0 + f) \in \mathcal{D}(A)$, we obtain the following estimate,

$$\|e_n\|_{L^p(\Omega)} \leq \frac{C}{t_n}\tau^2\|Du_0 + f\|_{L^p(\Omega)},$$

which concludes the proof of Theorem 4.3. □

Remark 4.12. *Formula* (4.6) *can also be obtained directly with an appropriate change of variable. This alternative proof makes clear why the global error e_n can be expressed as a difference between the rational function $r(\tau A)^n$ and the semigroup $\mathrm{e}^{n\tau A}$. Indeed by the affine change of variable $\hat{u}(x,t) = u(x,t) - \hat{z}(x)$, where $\hat{z}(x) = z(x) - A^{-1}Dz(x) - A^{-1}f(x)$, we obtain $\partial_t\hat{u}(x,t) = A\tilde{u}(x,t)$. Therefore, we can rewrite equation* (4.1) *as follows*

$$\begin{aligned} \partial_t\hat{u}(x,t) &= A\hat{u}(x,t) \quad \text{in } \Omega\times(0,\infty), \\ B\hat{u}(x,t) &= 0 \quad \text{on } \partial\Omega\times(0,\infty), \\ \hat{u}(x,0) &= u_0(x) - \hat{z}(x) \quad \text{in } \Omega, \end{aligned}$$

whose solution is given by $\hat{u}(x,t) = \mathrm{e}^{tA}A^{-1}(Du_0(x) + f(x))$. Hence, we have the following formula for the exact solution,

$$u(x,t) = \mathrm{e}^{tA}A^{-1}(Du_0(x) + f(x)) + \hat{z}(x). \tag{4.14}$$

By Remark 4.2, we know that the numerical solution u_n of the splitting (1.6) *is given by n iterations of the Crank-Nicolson method applied to the whole problem* (4.1)*. Since Runge-Kutta methods are affine invariant, we obtain the following formula for $u_n(x)$,*

$$u_n(x) = r(\tau A)^n A^{-1}(Du_0(x) + f(x)) + \hat{z}(x). \tag{4.15}$$

Subtracting (4.15) *and* (4.14) *at time t_n, we therefore obtain formula* (4.6) *for the global error. Note that for any A-stable Runge-Kutta method with stability function $R(y)$, applied to the whole problem* (4.1)*, the numerical solution u_n and the error e_n are given by formulas* (4.15) *and* (4.6) *with $r(\tau A)$ replaced with $R(\tau A)$.*

5 Numerical experiments

We first describe the parameters and the notations that we use for the numerical experiments that follow.

For the one dimensional problems, we choose $N = 1000$ uniform grid points to discretize the domain $\Omega = (0, 1)$, i.e. the mesh is of size $h = \frac{1}{N} = 10^{-3}$. We denote by $U_{n,l}$ and $U_l(t_n)$, the approximations of $u_n(x_l)$ and $u(t_n, x_l)$, where $x_l = lh$, i.e. U_n and $U(t_n)$ are vectors in $\mathbb{R}^N$. For the two dimensional problems, we discretize the domain $\Omega = (0, 1)^2$ with a uniform mesh of size $h = 10^{-2}$. We denote by $U_{n,l,m}$ and $U_{l,m}(t_n)$ the approximations of $u_n(x_l, y_m)$ and $u(t_n, x_l, y_m)$, where $y_m = mh$. The operators ∂_{xx} and $\partial_{xx} + \partial_{yy}$ are approximated with the standard second order finite difference approximation and ghost points are used for the normal derivatives in the boundary conditions. The final time is $T = 0.1$. We apply all considered splitting methods with the time steps $\tau = 0.02 \cdot 2^{-k}$ for $k = 0, \ldots, 6$.

In the splitting algorithms, the source term equation (1.2) is solved exactly. In the splitting (1.4) and (1.5), when we say that the diffusion equation is solved exactly, we mean that it is solved with a Krylov based algorithm developed in [16], with a tolerance close to machine precision. The reference solutions are computed with the Crank-Nicolson method (for $d = 1$) and the classical four stage Runge-Kutta method (for $d = 2$) with a small time step $\tau = 0.02 \cdot 2^{-10} \approx 2 \cdot 10^{-5}$.

The splitting (1.4) using the exact flow of the diffusion part is denoted *StrangEXP* and the splitting (1.6) is denoted *StrangCN*. We also consider the splitting methods *StrangGauss*, *StrangRadau* and *StrangLobatto*, which are constructed similarly to the splitting method *StrangCN*, but where the diffusion problem (1.3) is approximated with the two stage Gauss method (order 4), the two stage Radau 1a method (order 3) and the two stage Lobatto 3c method (order 2) (see the Runge-Kutta Butcher tableau and stability functions in Appendix B). Similarly, the splitting (1.5) is denoted by *StrangEXP2*. We denote by *StrangCN2*, *StrangGauss2*, *StrangRadau2* and *StrangLobatto2*, the splitting methods corresponding to (1.5), where the diffusion equation (1.3) is approximated with one of the methods described above.

The error of a splitting method at time $t_k = k\tau$ is defined as $u_k - u(k\tau)$, where $u(k\tau)$ is given by the reference solution at time $k\tau$. In the numerical experiments we always estimate the error with the trapezoidal approximation of the $L^2(\Omega)$ norm at final time $T = n\tau$ (except in Figure 1b). In dimension one, the estimate of the $L^2(0, 1)$ error E_k at time t_k is given by

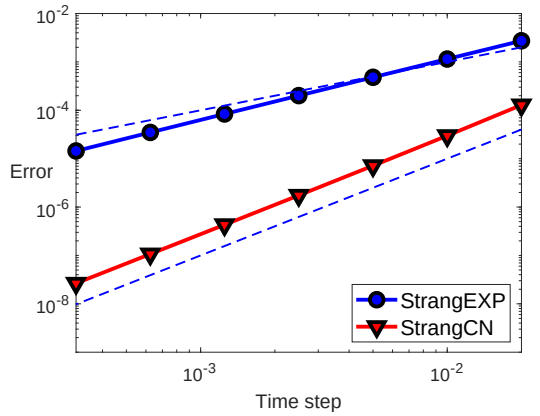
$$E_k^2 := h \sum_{l=1}^{N-1} \frac{|U_{k,l} - U_l(t_k)|^2 + |U_{k,l+1} - U_{l+1}(t_k)|^2}{2}.$$

In dimension two, the $L^2((0, 1)^2)$ error E_k is approximated similarly,

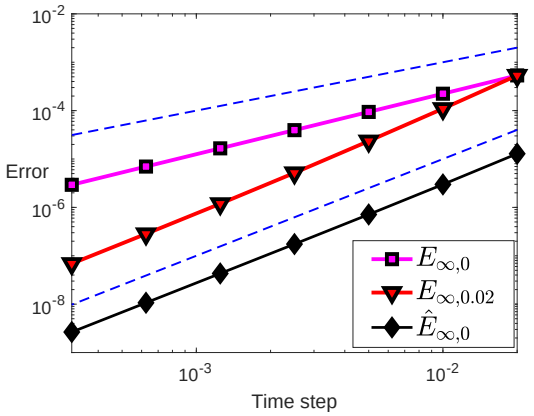
$$E_k^2 := h^2 \sum_{l,m=1}^{N-1} \frac{|U_{k,l,m} - U_{l,m}(t_k)|^2 + |U_{k,l+1,m} - U_{l+1,m}(t_k)|^2}{4} + \frac{|U_{k,l,m+1} - U_{l,m+1}(t_k)|^2 + |U_{k,l+1,m+1} - U_{l+1,m+1}(t_k)|^2}{4}.$$

In Figure 1b, we consider additional norms to estimate the error. We consider the approximation of the $L^\infty([0, T], L^2(0, 1))$ norm of the error,

$$E_{\infty,0} := \max_{k=0,\ldots,n} E_k. \tag{5.1}$$



(a) StrangCN *has no order reduction.*



(b) StrangCN *is not of order two near* $t = 0$.

Figure 1: *Solving the diffusion part* (1.3) *with the Crank-Nicolson scheme allows to remove the reduction of order of the Strang splitting method at final time* $T = 0.1$*, when applied to the 1d problem* $\partial_t u = \partial_{xx} u + 1$ *with inhomogeneous Dirichlet boundary conditions. However, in a neighbourhood of* $t = 0$*, the reduction of order is not avoided. Reference slopes one and two are given in dashed lines.*

Similarly, we consider the approximation of the $L^\infty([0.02, T], L^2(\Omega))$ norm of the error,

$$E_{\infty,0.02} := \max_{k=\frac{0.02}{\tau},\dots,n} E_k. \tag{5.2}$$

Another estimate of the error is provided where we compute an approximation of the $L^\infty([0, T], L^2(0, 1))$ norm of time multiplied by the error, precisely

$$\hat{E}_{\infty,0} := \max_{k=0,\dots,n} \|t_k E_k\|_2. \tag{5.3}$$

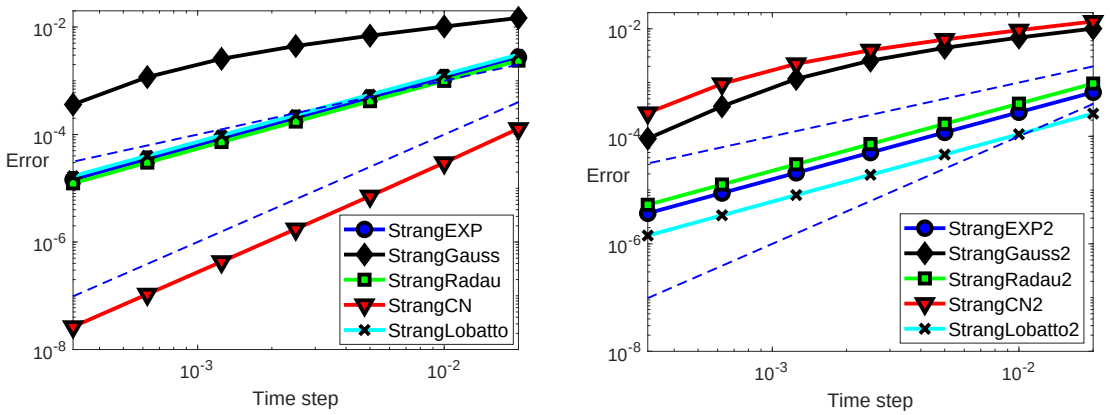
Remark 5.1. *In Figure 4, we implement the Crank-Nicolson method using the standard implementation of Runge-Kutta methods to avoid rounding error. To solve* $\frac{dx}{dt}(t) = Ax(t)$ *with* $x(0) = x_0$*, we start to resolve the linear system* $k_1 = Ax_0$ *and* $k_2 = Ax_0 + \frac{\tau}{2}A(k_1 + k_2)$*. Then, we write* $x_1 = x_0 + \frac{\tau}{2}k_1 + \frac{\tau}{2}k_2$*, and similarly for the two stage Gauss method. For even higher accuracy, one should use in addition a compensated summation algorithm (see [7, Algorithm VIII.5.1]).*

5.1 Solution independent source term

In the first series of experiments, we consider the following parabolic problem on $\Omega = (0, 1)$ with $t \in [0, T]$, with Dirichlet boundary conditions, which satisfies the hypotheses of Theorem 4.3,

$$\partial_t u(x, t) = \partial_{xx} u(x, t) + 1 \quad \text{in } (0, 1) \times (0, T], \qquad u(0, t) = u(1, t) = 1, \qquad u(x, 0) = 1. \tag{5.4}$$

In Figure 1a, we compare the splitting *StrangEXP* and the splitting *StrangCN* for the problem (5.4). This simple example illustrates the superiority of the splitting (1.6) compared to the splitting (1.4). Indeed, the splitting method (1.6) avoids order reduction



(a) $u_{n+1} = \phi^f_{\frac{\tau}{2}} \circ \phi^D_\tau \circ \phi^f_{\frac{\tau}{2}}(u_n)$.

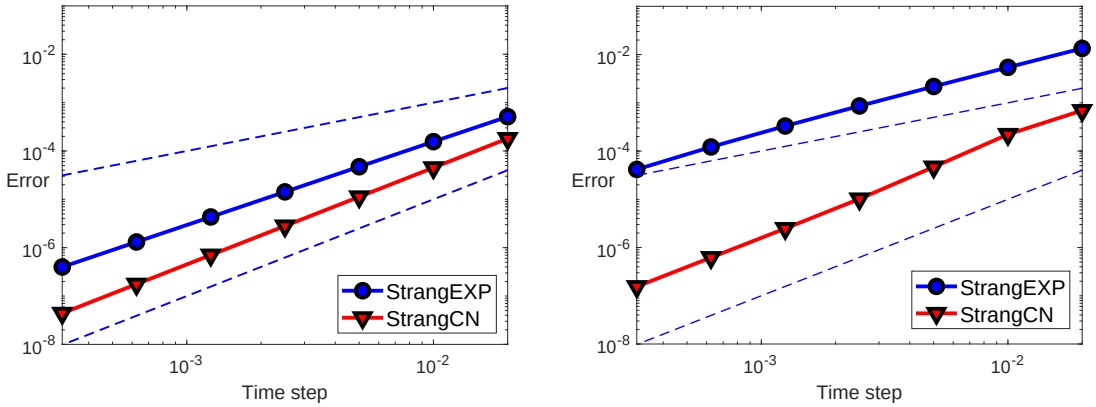
(b) $u_{n+1} = \phi^D_{\frac{\tau}{2}} \circ \phi^f_\tau \circ \phi^D_{\frac{\tau}{2}}(u_n)$.

Figure 2: *The reduction of order of the Strang splitting* (1.4) *method is not avoided when the following 2 stage Runge Kutta methods are used to solve the diffusion* (1.3)*: Gauss (order 4), Radau 1a (order 3) or Lobatto 3c (order 2). The 1d problem considered is* $\partial_t u = \partial_{xx} u + 1$ *with inhomogeneous Dirichlet boundary conditions. In addition, for the Strang splitting* (1.5)*, solving the diffusion* (1.3) *with the Crank-Nicolson method does not permit to remove the reduction of order. Reference slopes one and two are given in dashed lines.*

contrary to the Strang splitting (1.4) and it allows a high gain of accuracy for no additional computational cost. Note that in the non-generic case where the source term satisfies the condition $Bf(x) = 0$ on the boundary, there is no order reduction since no perturbation of the boundary occurs when solving the source equation (1.2) in the splitting (1.4). For example, for the problem $\partial_t u(x,t) = \partial_{xx} u(x,t) + \mathrm{e}^{-x}$ with $u(0,t) - \partial_n u(0,t) = 1$, $u(1,t) + \partial_n u(1,t) = 1$, and $u(x,0) = 1$, the splitting (1.4) is of order two and its convergence curve is superposed to the one of the splitting (1.6). The convergence curves are not drawn for conciseness.

In Figure 1b, we estimate the error of the splitting *StrangCN*, applied to the problem (5.4) with the norm $L^\infty([0,T], L^2(0,1))$ and the norm $L^\infty([0,T], L^2(0.02,1))$, i.e. we compare $E_{\infty,0}$ and $E_{\infty,0.02}$, given by (5.1) and (5.2). We observe that, $E_{\infty,0}$ does not decrease quadratically with respect to the time step τ. This is expected since the bound of the error of the splitting (1.6) given in Theorem 4.3 has order reduction down to one in a neighbourhood of $t = 0$. In comparison, with $E_{\infty,0.02}$, which avoids a neighbourhood of $t = 0$, we recover the second order convergence. We also observe that $\hat{E}_{\infty,0}$, given by formula (5.3), decays quadratically with respect to τ. This suggests that the error estimate $\mathcal{O}(\frac{\tau^2}{t_n})$ of Theorem 4.3 is optimal in a neighbourhood of $t = 0$.

In Figure 2a, we approximate the diffusion part (1.3) of the splitting (1.4) with a variety of Runge-Kutta methods. We use the 2 stage Gauss method (order 4), the 2 stage Radau 1a method (order 3) and the 2 stage Lobatto 3c method (order 2) (see Appendix B for the Butcher tableau of these methods). We also compute the error when the diffusion is solved exactly and when the Crank-Nicolson method is used, corresponding to the splitting method (1.4) and (1.6). Except for Crank-Nicolson, none of the classical Runge-Kutta methods that we tested allows to remove the order reduction. We observe that the 2 stage



(a) $f(u) = u$ *with Robin boundary conditions.* (b) $f(u) = u^2$ *with mixed boundary conditions.*

Figure 3: *The Strang splitting method* (1.4) *has no order order reduction when the diffusion equation* (1.3) *is solved using the Crank-Nicolson scheme. On the left picture, the equation is* $\partial_t u = \Delta u + u$ *with Robin boundary conditions. On the right picture, the equation is* $\partial_t u = \Delta u + u^2$ *with Neumann boundary conditions on the left and bottom boundaries and Dirichlet boundary conditions on the top and right boundaries. Reference slopes one and two are given in dashed lines.*

Gauss method is the method for which the error is the largest, when in comparison the Crank-Nicolson method (equivalently the 1 stage Gauss method) is by far the method for which the error is the smallest for all considered time steps τ.

In Figure 2b we apply the Strang splitting method (1.5) instead of the Strang splitting (1.4) to the problem (5.4). The same experiment is then performed where we approximate the diffusion equation (1.3) with different Runge-Kutta methods. We see that, for the Strang splitting (1.5), the Crank-Nicolson method does not allow to remove the order reduction. Surprisingly, it turns out that it is, amongst the methods tested, the scheme for which the error is the largest.

5.2 Solution dependent source term

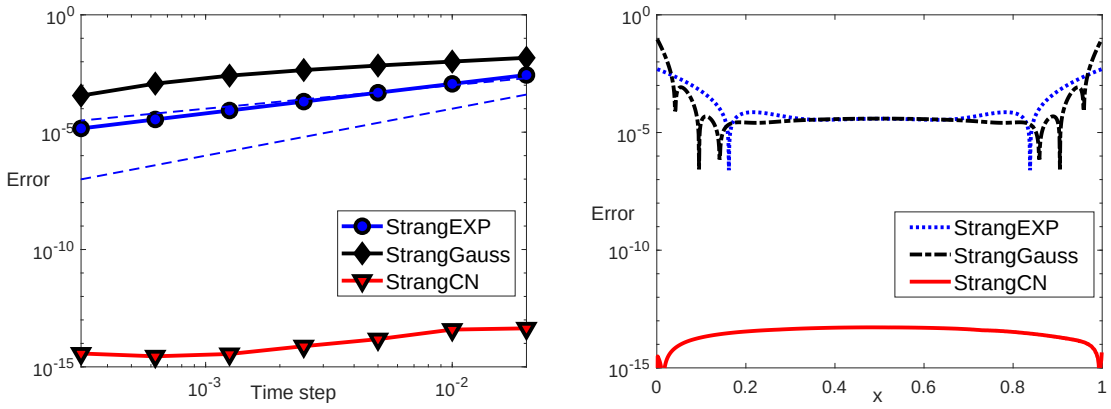
In Figure 3a, we consider the following two dimensional problem with Robin boundary conditions, for $(x, y) \in \Omega = (0, 1)^2$, $t \in [0, T]$,

$$\begin{aligned} \partial_t u(x, y, t) &= \partial_{xx} u(x, y, t) + \partial_{yy} u(x, y, t) + u(x, y, t), \\ u(0, y, t) + \partial_n u(0, y, t) = y^2, &\quad u(1, y, t) + \partial_n u(1, y, t) = y^2 + 2, \\ u(x, 0, t) + \partial_n u(x, 0, t) = x^2, &\quad u(x, 1, t) + \partial_n u(x, 1, t) = x^2 + 2, \\ u(x, y, 0) &= x^2 + y^2. \end{aligned} \tag{5.5}$$

As already observed, we see that the splitting *StrangEXP* suffers from order reduction, when in comparison the splitting (1.6) is of order two.

In Figure 3b, we consider the following problem, for $(x, y) \in \Omega = (0, 1)^2$, $t \in [0, T]$,

$$\begin{aligned} \partial_t u(x, y, t) &= \partial_{xx} u(x, y, t) + \partial_{yy} u(x, y, t) + u^2(x, y, t), \\ \partial_n u(0, y, t) = \frac{1}{2}, &\quad u(1, y, t) = \frac{\mathrm{e}^1 + \mathrm{e}^y}{2}, \end{aligned}$$



(a) *Error for a stationary problem* (b) $|u_n(x) - u(t_n, x)|$ *for* $\tau = 10^{-2}$

Figure 4: *The numerical solution given by the splitting* (1.6) *has error close to machine precision for the stationary problem* (5.7). *It contrast, for the splitting* StrangEXP (1.4), *the error is closed to* 10^{-5}. *When the two stage Gauss method is used to solve the diffusion equation* (1.3) *instead of the Crank-Nicolson method, the error deteriorates. The problem considered is* $\partial_t u = \partial_{xx} u - 1$ *with* $u_0 = x^2/2$. *Reference slopes one and two are drawn in dashed line in the left picture.*

$$\begin{aligned} \partial_n u(x,0,t) = \frac{1}{2}, \quad u(x,1,t) = \frac{\mathrm{e}^x + \mathrm{e}^1}{2}, \\ u(x,y,0) = \frac{\mathrm{e}^x + \mathrm{e}^y}{2}. \end{aligned} \tag{5.6}$$

For problem (5.6), the splitting (1.6) is significantly more accurate than the splitting (1.4). In particular, using the Crank-Nicolson method to solve the diffusion part allows to increase the precision by a factor close to 1000 for the smallest time step $\tau = 3.125 \cdot 10^{-4}$.

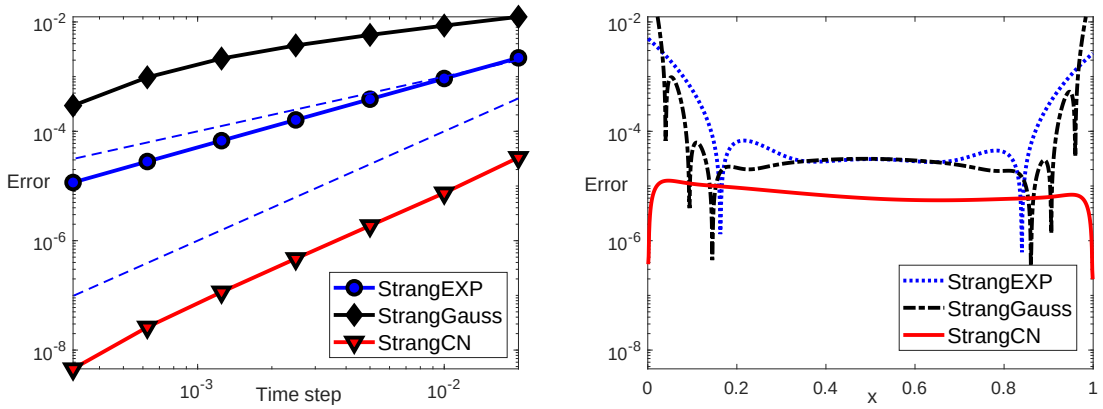
5.3 Stationary problems

In Figure 4, we consider the following stationary problem on $\Omega = (0,1)$, with $t \in [0,T]$,

$$\partial_t u(x,t) = \partial_{xx} u(x,t) - 1 \quad \text{in } (0,1), \qquad u(0,t) = 0,\ u(1,t) = \frac{1}{2}, \qquad u(x,0) = \frac{x^2}{2}. \tag{5.7}$$

Since the problem is stationary ($u(x,t) = u(x,0)$), we take $u(x,t) = \frac{x^2}{2}$ as the reference solution. In Figure 4, we observe that the error of the splitting (1.6) is close to the machine precision (10^{-15}) and it does not decay for smaller time steps. In comparison, the error of the splitting (1.5) is around 10^{-5} and it diminishes almost linearly when the time steps become smaller, i.e. the splitting (1.5) suffers from order reduction down to one even for stationary problems. We observe the same phenomenon of order reduction for the two stage Gauss method with an error worse than the splitting (1.4). For the other Runge-Kutta methods considered previously, the error curves are nearly identical to the one of the splitting (1.4) and they are not drawn for better readability.

In Figure 5, we consider a stationary problem where the source term f depends on the



(a) *Error for a stationary problem with* $f(u) = u$ (b) $|u_n(x) - u(t_n, x)|$ *for* $\tau = 10^{-2}$

Figure 5: *For a stationary problem with a source term that depends on the solution* u*, the splitting* StrangCN *is not exact, in contrast with a problem where the source term* f *only depends on the space variable (see Figure 4). The parabolic problem considered is* $\partial_t u = \partial_{xx} u + u$*,* $u_0 = \cos(x)$*, with inhomogeneous Dirichlet boundary conditions. Reference slopes one and two are given in dashed lines on the left picture.*

solution u, for $x \in [0, 1]$ with $t \in [0, T]$,

$$\partial_t u(x,t) = \partial_{xx} u(x,t) - u(x,t) \quad \text{in } (0,1), \quad u(0,t) = 1,\ u(1,t) = \cos(1), \quad u(x,0) = \cos(x). \tag{5.8}$$

The reference solution is $u(x, t) = \cos(x)$. We observe that the splitting *StrangCN* is not exact in comparison with the previous stationary problem (5.7), where f only depends on the space variable. However, even for this problem where $f = f(u)$ that does not satisfy the hypotheses of Theorem 4.3, we observe that the splitting *StrangCN* is second order convergent compared to the splittings *StrangEXP* and *StrangGauss*, which have a reduced order of convergence between one and two. As seen in Figure 5b, the loss of accuracy is mostly due to the large error made near the boundary of the domain. For better readability, we did not draw the convergence curves for *StrangLobatto* and *StrangRadau* since they are superposed to the convergence curve of *StrangEXP*.

We also observed numerically that the convergence analysis presented in this paper does not persist for dispersive problems, e.g. replacing $\partial_t u(x, t)$ by $i\partial_t u(x, t)$ formally in (5.4) to obtain a Schrödinger type problem (the convergence curves are not drawn for conciseness).

Acknowledgements. This work was partially supported by the Swiss National Science Foundation, grants No. 200020_184614 and No. 200020_178752. GB and GV would like to acknowledge great hospitality when visiting Institut de Mathématiques de Toulouse thanks to grant ANR project NABUCO, ANR-17-CE40-0025.

A The Crank-Nicolson rational approximation of an analytic semigroup

The aim of this Appendix is to prove Theorem 4.8, which is a direct consequence of [9, Theorem 2.1]. We present here a new direct and self contained proof of Theorem 4.8. Similarly to [9], the proof that we present holds for a general separable Banach space and not only for the special case $L^p(\Omega)$. This is useful in our context, if we consider a more abstract problem as described in Remark 4.5.

Let X be a separable complex Banach space with norm $\|\cdot\|$. Let A be a closed densely defined linear operator such that all eigenvalues are in a sector in the left half plane, or equivalently for $\alpha \in (0, \frac{\pi}{2})$ the resolvent set $\rho(A)$ contains the set $\overline{\Sigma}_\alpha$:

$$\rho(A) \supset \overline{\Sigma}_\alpha \tag{A.1}$$

where $\Sigma_\alpha = \{z \in \mathbb{C} \ ; \ z \neq 0, \ |\arg(z)| < \pi - \alpha\}$. Assume that for all $z \in \Sigma_\alpha$, the resolvent of A, $R(z, A) = (zI - A)^{-1}$, satisfies the following bound for the operator norm,

$$\|R(z, A)\| \leq \frac{M}{|z|}, \tag{A.2}$$

where $M \geq 1$. Under those assumptions, the operator A is the infinitesimal generator of an analytic semigroup given by integral formula (see [6, Definition I.3.4]),

$$\mathrm{e}^{tA} = \frac{1}{2\pi i} \int_\Gamma \mathrm{e}^{zt} R(z, A) \mathrm{dz}, \tag{A.3}$$

with $t \geq 0$ and where Γ is the boundary of Σ_α with imaginary part increasing along Γ. As before $r(z)$ denotes the stability function of the Crank-Nicolson scheme,

$$r(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}.$$

The purpose of this Appendix is to prove the following theorem :

Theorem A.1. *Let $u_0 \in \mathcal{D}(A)$ and let $n \geq 3$, then*

$$\|(r(\tau A)^n - \mathrm{e}^{\tau n A}) u_0\| \leq \frac{C\tau^2}{t_n} \|A u_0\|,$$

with C is a positive constant independent of u_0, τ, n, and $t_n = n\tau$.

In the following lemma, we give some estimates for of $r(z)$. The proof is inspired from [3].

Lemma A.2. *1. For all $z \in \mathbb{C}^-$ with $|z| \leq 1$, we have*

$$|r(z) - \mathrm{e}^z| \leq \frac{5}{12} |z|^3. \tag{A.4}$$

2. For all $z \in \mathbb{C}^-$,

$$|r(z)| \leq \max\left(e^{\frac{4}{5}\Re(z)}, e^{\frac{4}{5\Re(z)}}\right) \tag{A.5}$$

3. For all $y \in \mathbb{R}^+$ and for all integer $k \geq 1$, there exists C_k independent of y such that

$$\int_1^\infty \mathrm{e}^{-\frac{y}{x}} \frac{y^k}{x^{k+1}} \mathrm{dx} \leq C_k. \tag{A.6}$$

4. For all $y \in \mathbb{R}^+$ and for all integer $k \geq 1$, there exists C_k independent of y such that

$$\int_0^1 \mathrm{e}^{-yx} y^{k+1} x^k \mathrm{dx} \leq C_k. \tag{A.7}$$

Proof. 1. Since $z \neq -\frac{1}{2}$, we have

$$r(z) - \mathrm{e}^z = \frac{z^3}{4} \frac{1}{1 - \frac{z}{2}} - z^3 \int_0^1 \mathrm{e}^{(1-s)z} \frac{s^2}{2} \mathrm{ds}.$$

Therefore, for $z \in \mathbb{C}^-$ and $|z| \leq 1$, we obtain

$$|r(z) - \mathrm{e}^z| \leq |z^3| \left(\frac{1}{4} + \int_0^1 \frac{s^2}{2} \mathrm{ds} \right) = \frac{5}{12} |z^3|.$$

2. We first show that

$$r(-x) \leq \mathrm{e}^{-x} \tag{A.8}$$

for $x \in \mathbb{R}^+$. Since e^{-x} and $r(-x)$ are both equal to 1 for $x = 0$, it suffices to show that $\frac{\mathrm{d}}{\mathrm{dx}} r(-x) \leq \frac{\mathrm{d}}{\mathrm{dx}} \mathrm{e}^{-x}$ for $x \geq 0$. A calulation yields $\frac{\mathrm{d}}{\mathrm{dx}} r(-x) = \frac{-1}{(1+\frac{x}{2})^2} \leq -\mathrm{e}^{-x} = \frac{\mathrm{d}}{\mathrm{dx}} \mathrm{e}^{-x}$. For $\alpha \in (0, \frac{\pi}{2})$ and $0 \leq \rho \leq 1$, we have

$$\left| r(-\rho \mathrm{e}^{\pm i\alpha}) \right|^2 = \frac{1 + \frac{\rho^2}{4} - \rho \cos(\alpha)}{1 + \frac{\rho^2}{4} + \rho \cos(\alpha)} \leq \left(\frac{1 - \frac{2\rho}{5} \cos(\alpha)}{1 + \frac{2\rho}{5} \cos(\alpha)} \right)^2 = r(-\frac{4\rho}{5} \cos(\alpha))^2.$$

Indeed, we observe that

$$\frac{1 + \frac{\rho^2}{4} - \rho \cos(\alpha)}{1 + \frac{\rho^2}{4} + \rho \cos(\alpha)} \leq \left(\frac{1 - \frac{4\rho}{5} \cos(\alpha)}{1 + \frac{4\rho}{5} \cos(\alpha)} \right)^2 \Leftrightarrow \frac{\frac{8}{5} + \frac{2\rho^2}{5}}{2 + \frac{8\rho^2}{25} \cos(\alpha)^2} \cos(\alpha) \leq \cos(\alpha)$$

and the last inequality is true since for $0 \leq \rho \leq 1$ and $0 \leq \cos(\alpha) \leq 1$, we deduce

$$\frac{\frac{8}{5} + \frac{2\rho^2}{5}}{2 + \frac{8\rho^2}{25} \cos(\alpha)^2} \leq \frac{\frac{8}{5} + \frac{2}{5}}{2} \leq 1.$$

Using (A.8), we obtain

$$\left| r(-\rho \mathrm{e}^{\pm i\alpha}) \right| \leq \mathrm{e}^{-\frac{4\rho}{5} \cos(\alpha)}. \tag{A.9}$$

If $z = -\rho \mathrm{e}^{\pm i\alpha}$, with $\alpha \in (0, \frac{\pi}{2})$ and $\rho \geq 1$, we start to observe that

$$|r(z)| = \left| \frac{1 + \frac{4}{2z}}{1 - \frac{4}{2z}} \right| = \left| r\left(\frac{4}{z} \right) \right|.$$

Since $\arg(\frac{4}{z}) = -\arg(z)$, we write $\frac{4}{z} = \eta e^{\mp i\alpha}$ with $\eta = \frac{4}{\rho}$. Since $\rho \geq 1$, we deduce $0 < \eta \leq 4$. We observe that

$$\frac{1 + \frac{\eta^2}{4} - \eta\cos(\alpha)}{1 + \frac{\eta^2}{4} + \eta\cos(\alpha)} \leq \left(\frac{1 - \frac{\eta}{10}\cos(\alpha)}{1 + \frac{\eta}{10}\cos(\alpha)}\right)^2 = r(-\frac{\eta}{5}\cos(\alpha))^2.$$

Indeed, we have

$$\frac{1 + \frac{\eta^2}{4} - \eta\cos(\alpha)}{1 + \frac{\eta^2}{4} + \eta\cos(\alpha)} \leq \left(\frac{1 - \frac{\eta}{10}\cos(\alpha)}{1 + \frac{\eta}{10}\cos(\alpha)}\right)^2 \Leftrightarrow \frac{\frac{2}{5} + \frac{\eta^2}{10}}{2 + \frac{\eta^2}{50}\cos(\alpha)^2}\cos(\alpha) \leq \cos(\alpha)$$

and the last inequality is true since for $0 < \eta \leq 4$ and $0 \leq \cos(\alpha) \leq 1$, we deduce

$$\frac{\frac{2}{5} + \frac{\eta^2}{10}}{2 + \frac{\eta^2}{50}\cos(\alpha)^2} \leq \frac{\frac{2}{5} + \frac{16}{10}}{2} = \frac{20}{20} = 1.$$

Using (A.8), we obtain

$$\left|r(-\eta e^{\mp i\alpha})\right| \leq e^{-\frac{\eta}{5}\cos(\alpha)}.$$

This gives us for $\rho = \frac{4}{\eta}$,

$$\left|r(-\rho e^{\pm i\alpha})\right| \leq e^{-\frac{4}{5\rho}\cos(\alpha)}.$$

Together with (A.9), this concludes the proof.

3. The proof is made by induction. For $k = 1$, we obtain,

$$\int_1^\infty e^{-\frac{y}{x}}\frac{y}{x^2}dx = e^{-\frac{y}{x}}\Big|_1^\infty = 1 - e^{-y} \leq 1.$$

Assuming the result is true for k, we obtain

$$\int_1^\infty e^{-\frac{y}{x}}\frac{y^k}{x^{k+1}}dx = e^{-\frac{y}{x}}\frac{y^{k-1}}{x^{k-1}}\Big|_1^\infty + (k-1)\int_1^\infty e^{-\frac{y}{x}}\frac{y^{k-1}}{x^k}dx = -e^{-y}y^{k-1} + C \leq C,$$

where we used $e^{-y}y^{k-1} \leq C$, with C independent of y.

4. With the change of variable $\hat{x} = \frac{1}{x}$, we observe it is a direct consequence of the previous inequality (A.6).

□

We introduce the following notation for the stability function of the implicit Euler method,

$$r_0(z) = \frac{1}{1 - z}.$$

The rational approximation r_0 satisfy the following inequalities.

Lemma A.3. *Let $z \in \mathbb{C}^- = \{z \in \mathbb{C}\ ;\ \Re(z) \leq 0\}$ with $|z| \leq 1$, then*

$$\left|r_0\left(\frac{z}{2}\right) - e^{\frac{z}{2}}\right| \leq \frac{3}{8}|z|^2 \quad \textit{and} \quad \left|r_0\left(\frac{z}{2}\right)^2 - e^z\right| \leq \frac{6}{8}|z|^2.$$

Proof. Since $z \neq -\frac{1}{2}$, we have

$$r_0\left(\frac{z}{2}\right) - \mathrm{e}^{\frac{z}{2}} = \frac{z^2}{4(1-\frac{z}{2})} - \frac{z^2}{4}\int_0^1 \mathrm{e}^{(1-s)z} s\mathrm{ds}.$$

Therefore, for $z \in \mathbb{C}^-$ and $|z| \leq 1$, we obtain

$$\left|r_0\left(\frac{z}{2}\right) - \mathrm{e}^{\frac{z}{2}}\right| \leq \frac{|z^2|}{4}\left(1 + \int_0^1 s\mathrm{ds}\right) = \frac{3}{8}|z^2|.$$

For the second inequality we observe that

$$\left|r_0\left(\frac{z}{2}\right)^2 - \mathrm{e}^z\right| = \left|r_0\left(\frac{z}{2}\right)\left(r_0\left(\frac{z}{2}\right) - \mathrm{e}^{\frac{z}{2}}\right) + \left(r_0\left(\frac{z}{2}\right) - \mathrm{e}^{\frac{z}{2}}\right)\mathrm{e}^{\frac{z}{2}}\right| \leq \frac{6}{8}|z^2|,$$

which concludes the proof. □

We define F_n and G_n as follows,

$$F_n(z) = r(z)^{n-2} r_0\left(\frac{z}{2}\right)^2 - \mathrm{e}^{\tau(n-1)z} \quad \text{and} \quad G_n(z) = r(z)^{n-1} r_0\left(\frac{z}{2}\right) - \mathrm{e}^{\tau(n-\frac{1}{2})z}.$$

The term $r(z)^{n-2} r_0\left(\frac{z}{2}\right)^j$, $j = 1, 2$ correspond to the stability function of the composition of the Crank-Nicolson scheme with respectively one or two half steps of the Euler implicit scheme. This provides higher regularity for the solution since $r_0\left(\frac{\tau A}{2}\right)^j : \mathcal{D}(A^k) \to \mathcal{D}(A^{k+j})$ for k any integer. For more general results on composition of rational approximation, see [9].

Lemma A.4. *Let $n \geq 2$. Then $F_n(A)$ and $G_n(A)$ satisfy the following integral formula*

$$F_n(A) = \frac{1}{2\pi i}\int_\Gamma F_n(z) R(z, A)\mathrm{dz}, \quad \textit{and} \quad G_n(A) = \frac{1}{2\pi i}\int_\Gamma G_n(z) R(z, A)\mathrm{dz},$$

where $\Gamma = \{z \in \mathbb{C}\ ;\ |\arg(z)| = \pi - \alpha\}$.

Proof. The rational function $r(z)^{n-2} r_0\left(\frac{z}{2}\right)^2$ is holomorphic in a neighbourhood of the spectrum of A and it vanishes at infinity . Therefore (see [4, Theorem VII.9.4]), we have

$$r(A)^{n-2} r_0\left(\frac{A}{2}\right)^2 = \frac{1}{2\pi i}\int_\Gamma r(z)^{n-2} r_0\left(\frac{z}{2}\right)^2 R(z, A)\mathrm{dz}.$$

We conclude the proof using (A.3). The proof for $G_n(A)$ is similar and thus omitted. □

The proof that follows is inspired from [18, Theorem 9.3].

Proposition A.5. *For $n \geq 3$ and $u_0 \in X$, we have*

$$\left\| r(\tau A)^{n-2} r_0\left(\frac{\tau A}{2}\right)^2 u_0 - \mathrm{e}^{\tau(n-1)A} u_0 \right\| \leq C\frac{\tau^2}{t_n^2}\|u_0\| \tag{A.10}$$

and

$$\left\| r(\tau A)^{n-1} r_0\left(\frac{\tau A}{2}\right) u_0 - \mathrm{e}^{\tau(n-\frac{1}{2})A} u_0 \right\| \leq C\frac{\tau}{t_n}\|u_0\|, \tag{A.11}$$

where C is a constant independent of u_0, τ, and n.

Proof. Since $\frac{1}{n} = \frac{\tau}{t_n}$, we need to show that

$$\|F_n(A)\| \leq \frac{CM}{n^2}.$$

Let $z = -\rho \mathrm{e}^{\pm i\alpha}$ with $\rho \geq 1$. We have from inequality (A.5) that

$$|r_1(-\rho \mathrm{e}^{\pm i\alpha})| \leq \mathrm{e}^{-\frac{c}{\rho}},$$

where $0 < c < 1$ denotes a constant. Additionally, we have

$$\left| r_0\left(-\frac{\rho \mathrm{e}^{\pm i\alpha}}{2}\right)\right| = \frac{1}{\sqrt{1 + \frac{\rho^2}{4} + \rho\cos(\alpha)}} \leq \frac{2}{\rho}.$$

We recall that for all $x \in \mathbb{R}^+$, we have

$$\mathrm{e}^{-x} \leq \frac{C}{x^p}, \tag{A.12}$$

where C is independent of x. Therefore, since $|\mathrm{e}^{-(n-1)\rho \mathrm{e}^{i\alpha}}| \leq \mathrm{e}^{-(n-1)\rho\cos(\alpha)} \leq \frac{C}{(n-1)^2\rho^2}$, we obtain for $n \geq 3$,

$$|F_n(-\rho \mathrm{e}^{\pm i\alpha})| \leq \mathrm{e}^{-\frac{c(n-2)}{\rho}} \frac{4}{\rho^2} + \frac{C}{(n-1)^2\rho^2} \leq \left(\mathrm{e}^{-\frac{c(n-2)}{\rho}} \frac{(n-2)^2}{\rho^2} + \frac{1}{\rho^2}\right) \frac{C}{n^2}.$$

We use the inequality (A.6) with $y = c(n-2)$ and obtain

$$\int_1^\infty \mathrm{e}^{-\frac{c(n-2)}{\rho}} \frac{(n-2)^2}{\rho^3} \mathrm{d}\rho = \frac{1}{c^2} \int_1^\infty \mathrm{e}^{-\frac{c(n-2)}{\rho}} \frac{c^2(n-2)^2}{\rho^3} \mathrm{d}\rho \leq C.$$

This allows us to bound the integral,

$$\int_1^\infty |F_n(-\rho \mathrm{e}^{\pm i\alpha})| \|R(-\rho \mathrm{e}^{\pm i\alpha}, A)\| \mathrm{d}\rho \leq \frac{C}{n^2} \int_1^\infty \left(\mathrm{e}^{-\frac{c(n-2)}{\rho}} \frac{(n-2)^2}{\rho^2} + \frac{1}{\rho^2}\right) \frac{M}{\rho} \mathrm{d}\rho \leq \frac{CM}{n^2},$$

using $\int_1^\infty \frac{1}{\rho^3} \mathrm{d}\rho = \frac{1}{2}$ and (A.2).

For $\rho \leq 1$, we write

$$F_n(z) = r_0\left(\frac{z}{2}\right)^2 \left(r(z)^{n-2} - \mathrm{e}^{(n-2)z}\right) + \left(r_0\left(\frac{z}{2}\right)^2 - \mathrm{e}^z\right) \mathrm{e}^{(n-2)z}.$$

We observe that

$$r(z)^{n-2} - \mathrm{e}^{(n-2)z} = (r(z) - \mathrm{e}^z) \sum_{k=0}^{n-3} r(z)^{n-k-3} \mathrm{e}^{kz}.$$

Since, by estimate (A.4), $|r(-\rho \mathrm{e}^{\pm i\alpha}) - \mathrm{e}^{-\rho \mathrm{e}^{\pm i\alpha}}| \leq C\rho^3$ and since, by the inequality (A.5), $|r(-\rho \mathrm{e}^{\pm i\alpha})| \leq \mathrm{e}^{-\rho c}$, we obtain

$$\left| r\left(-\rho \mathrm{e}^{\pm i\alpha}\right)^{n-2} - \mathrm{e}^{-(n-2)\rho \mathrm{e}^{\pm i\alpha}} \right| \leq C\rho^3(n-2)\mathrm{e}^{-\rho(n-3)c} \leq C\rho^3(n-3)^3\mathrm{e}^{-\rho(n-3)c} \frac{C}{n^2}.$$

Using inequality (A.7) with $y = (n-3)c$, we deduce

$$\int_0^1 \rho^2 (n-3)^3 \mathrm{e}^{-\rho(n-3)c} \mathrm{d}\rho = \frac{1}{c^3} \int_0^1 \rho^2 (n-3)^3 c^3 \mathrm{e}^{-\rho(n-3)c} \mathrm{d}\rho \leq C.$$

From Lemma A.3, we also obtain,

$$\left| \left(r_0 \left(\frac{\rho}{2} \mathrm{e}^{\pm i\alpha} \right)^2 - \mathrm{e}^{-\rho \mathrm{e}^{\pm i\alpha}} \right) \mathrm{e}^{-(n-2)\rho \mathrm{e}^{\pm i\alpha}} \right| \leq C\rho^2 \mathrm{e}^{-(n-2)\rho\cos(\alpha)} \leq \frac{C}{n^2} \rho^2 (n-2)^2 \mathrm{e}^{-(n-2)\rho\cos(\alpha)}.$$

Using inequality (A.7) with $y = (n-2)\cos(\alpha)$, we obtain,

$$\int_0^1 \rho (n-2)^2 \mathrm{e}^{-(n-2)\rho\cos(\alpha)} \mathrm{d}\rho = \frac{1}{\cos(\alpha)^2} \int_0^1 \rho (n-2)^2 \cos(\alpha)^2 \mathrm{e}^{-(n-2)\rho\cos(\alpha)} \mathrm{d}\rho \leq C.$$

Therefore, since $\left| r_0(\frac{\rho}{2} \mathrm{e}^{\pm i\alpha}) \right| \leq 1$,

$$\begin{aligned}
&\int_0^1 |F_n(-\rho \mathrm{e}^{\pm i\alpha})| \|R(-\rho \mathrm{e}^{\pm i\alpha}, A)\| \mathrm{d}\rho \\
&\leq \int_0^1 \rho^2 (n-3)^3 \mathrm{e}^{-\rho(n-3)c} + \rho (n-2)^2 \mathrm{e}^{-(n-2)\rho\cos(\alpha)} \mathrm{d}\rho \frac{CM}{n^2} \leq \frac{CM}{n^2}
\end{aligned}$$

This concludes the proof for the first inequality (A.10). The second inequality (A.11) is obtained similarly and thus omitted. □

With the help of Proposition A.5, we can now prove Theorem A.1.

Proof of Theorem A.1. We write

$$\begin{aligned}
&(r(\tau A)^n - \mathrm{e}^{\tau n A}) u_0 \\
&= (r(\tau A)^{n-1} r_0 \left(\frac{\tau}{2} A \right) - \mathrm{e}^{\tau(n-\frac{1}{2})A})(1 + \frac{\tau}{2} A) u_0 + (1 + \frac{\tau}{2} A - \mathrm{e}^{\frac{\tau}{2} A}) \mathrm{e}^{\tau(n-\frac{1}{2})} u_0 \\
&= (r(\tau A)^{n-2} r_0 \left(\frac{\tau}{2} A \right)^2 - \mathrm{e}^{\tau(n-1)A})(1 + \frac{\tau}{2} A) u_0 + (1 + \frac{\tau}{2} A - \mathrm{e}^{\frac{\tau}{2} A}) \mathrm{e}^{\tau(n-1)} u_0 \\
&+ \frac{\tau}{2} (r(\tau A)^{n-1} r_0 \left(\frac{\tau}{2} A \right) - \mathrm{e}^{\tau(n-\frac{1}{2})A}) A u_0 + (1 + \frac{\tau}{2} A - \mathrm{e}^{\frac{\tau}{2} A}) \mathrm{e}^{\tau(n-\frac{1}{2})} u_0.
\end{aligned}$$

From Proposition A.5, we have

$$\|(r(\tau A)^{n-2} r_0 \left(\frac{\tau}{2} A \right)^2 - \mathrm{e}^{\tau(n-1)A})(1 + \frac{\tau}{2} A) u_0\| \leq C \frac{\tau^2}{t_n^2} (\|u_0\| + \tau \|A u_0\|)$$

and

$$\|\frac{\tau}{2} (r(\tau A)^{n-1} r_0 \left(\frac{\tau}{2} A \right) - \mathrm{e}^{\tau(n-\frac{1}{2})A}) A u_0\| \leq C \frac{\tau^2}{t_n} \|A u_0\|.$$

We observe that

$$(1 + \frac{\tau}{2} A - \mathrm{e}^{\frac{\tau}{2} A}) = -\frac{\tau^2}{4} A^2 \varphi_2(\frac{\tau}{2} A),$$

where $\varphi_2(\tau A)$ is the bounded operator given by

$$\varphi_2(z) = \int_0^1 \mathrm{e}^{(1-s)z} s \mathrm{ds}.$$

Using the smoothing property of e^{tA}, we obtain

$$\|(1+\frac{\tau}{2}A-\mathrm{e}^{\frac{\tau}{2}A})\mathrm{e}^{\tau(n-1)A}u_0\| \leq \frac{\tau^2}{4}\|\varphi_2(\frac{\tau}{2}A)\|\|A\mathrm{e}^{\tau(n-1)A}\|\|Au_0\| \leq \tau^2\frac{C}{t_{n-1}}\|Au_0\|$$

and

$$\|(1+\frac{\tau}{2}A-\mathrm{e}^{\frac{\tau}{2}A})\mathrm{e}^{\tau(n-\frac{1}{2})}u_0\| \leq \frac{\tau^2}{4}\|\varphi_2(\frac{\tau}{2}A)\|\|A\mathrm{e}^{\tau(n-\frac{1}{2})}\|\|Au_0\| \leq \tau^2\frac{C}{t_{n-\frac{1}{2}}}\|Au_0\|.$$

This concludes the proof of Theorem A.1. □

B Runge Kutta methods Butcher tableau

We give the Butcher tableau and the stability function of the A-stable implicit Runge-Kutta methods considered in the numerical experiments (see also [8, Chapter IV.5]).

The two stage Gauss method (order 4):

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad R(y)=\frac{1+\frac{y}{2}+\frac{y^2}{12}}{1-\frac{y}{2}+\frac{y^2}{12}}$$

The two stage Radau 1a method (order 3):

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \qquad R(y)=\frac{1+\frac{y}{3}}{1-\frac{2y}{3}+\frac{y^2}{6}}$$

The two stage Lobatto 3c method (order 2):

$$\begin{array}{c|cc} 0 & \frac{1}{2} & -\frac{1}{2} \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad R(y)=\frac{1}{1-y+\frac{y^2}{2}}.$$

References

[1] X. Antoine, C. Besse, and P. Klein. Absorbing boundary conditions for the one-dimensional Schrödinger equation with an exterior repulsive potential. *J. Comput. Phys.*, 228(2):312–335, 2009.

[2] G. Bertoli and G. Vilmart. Strang splitting method for semilinear parabolic problems with inhomogeneous boundary conditions: a correction based on the flow of the nonlinearity. *SIAM J. Sci. Comput.*, 42(3):A1913–A1934, 2020.

[3] M. Crouzeix. Approximation of parabolic equations. Lecture notes available at http://perso.univ-rennes1.fr/michel.crouzeix/, 2005.

[4] N. Dunford and J. T. Schwartz. *Linear operators. Part I.* Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988.

[5] L. Einkemmer and A. Ostermann. Overcoming order reduction in diffusion-reaction splitting. Part 2: Oblique boundary conditions. *SIAM J. Sci. Comput.*, 38(6):A3741–A3757, 2016.

[6] K.-J. Engel and R. Nagel. *One-parameter semigroups for linear evolution equations*, volume 194 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2000.

[7] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2010. Structure-preserving algorithms for ordinary differential equations, Reprint of the second (2006) edition.

[8] E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010. Stiff and differential-algebraic problems.

[9] A. Hansbo. Nonsmooth data error estimates for damped single step methods for parabolic equations in Banach space. *Calcolo*, 36(2):75–101, 1999.

[10] W. Hundsdorfer and J. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*, volume 33 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2003.

[11] W. Hundsdorfer and J. G. Verwer. A note on splitting errors for advection-reaction equations. *Appl. Numer. Math.*, 18(1-3):191–199, 1995. Seventh Conference on the Numerical Treatment of Differential Equations (Halle, 1994).

[12] B. Kovács, B. Li, and C. Lubich. A-stable time discretizations preserve maximal parabolic regularity. *SIAM J. Numer. Anal.*, 54(6):3600–3624, 2016.

[13] S. Larsson, V. Thomée, and L. B. Wahlbin. Finite-element methods for a strongly damped wave equation. *IMA J. Numer. Anal.*, 11(1):115–142, 1991.

[14] A. Lunardi. *Analytic semigroups and optimal regularity in parabolic problems*. Modern Birkhäuser Classics. Birkhäuser/Springer Basel AG, Basel, 2013.

[15] R. I. McLachlan, K. Modin, H. Munthe-Kaas, and O. Verdier. B-series methods are exactly the affine equivariant methods. *Numer. Math.*, 133(3):599–622, 2016.

[16] J. Niesen and W. M. Wright. Algorithm 919: a Krylov subspace algorithm for evaluating the ϕ-functions appearing in exponential integrators. *ACM Trans. Math. Software*, 38(3):Art. 22, 19, 2012.

[17] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*, volume 44 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1983.

[18] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.