# Asymptotic behavior of splitting schemes involving time-subcycling techniques

Guillaume Dujardin, Pauline Lafitte

**BIT manuscript No.**
(will be inserted by the editor)

# Asymptotic behavior of splitting schemes involving time-subcycling techniques

**Guillaume Dujardin · Pauline Lafitte**

October 10, 2014

**Abstract** This paper deals with the numerical integration of well-posed multiscale systems of ODEs or evolutionary PDEs. In this context, time-subcycling techniques are widely used in engineering problems to improve computational efficiency. These methods rely on a decomposition of the vector field in a fast part and a slow part and take advantage of that decomposition. This way, one can integrate the fast equations with a much smaller time step than that of the slow equations, instead of having to integrate the whole system with a very small time step to ensure stability. Then, one can build a numerical integrator using a standard composition method, such as a Lie or a Strang formula for example. The goal of this paper is to study the long time behavior of such schemes, that are primarily designed to be convergent in short-time to the solution of the original problem. In particular, when the solutions of the systems converge in time to an asymptotic equilibrium state, the question of the accuracy of the numerical long-time limit of the schemes as well as that of the rate of convergence is addressed. The asymptotic error is then defined as the difference between the exact and numerical asymptotic states. Its analysis is developed on simple linear ODE and PDE toy-models and is illustrated with several numerical experiments on these toy-models as well as on more complex systems.

**Keywords** Lie and Strang splitting schemes · $\theta$-schemes · Long time asymptotics · Asymptotic error · Asymptotic order

**Mathematics Subject Classification (2000)** 65L20 · 65M12

## 1 Introduction

Time-subcycling is a way to speed up numerical computations for an evolutionary multiscale problem by splitting the underlying operator and treating its different parts with adapted time-steps to build up a numerical integrator. The analysis of these methods over finite time intervals is rather similar to that of composition methods over finite time intervals. In contrast, in this paper, our aim is to determine how well the subcycling techniques capture the right asymptotic state for continuous dynamical systems described by ODEs or PDEs, the solutions of which converge to a steady state as time goes to infinity. In order to save computational time, the subcycling techniques have been very widely used for schemes associated with multiscale systems, that have (at least) one component that has to be computed through an explicit scheme thus constrained by a limitation of the time step (CFL) [2,9,3]. Related local time-stepping techniques have been developed extensively for multiscale problems arising in computational fluid and structural dynamics [16,6,7]. The simulation of transport or diffusive phenomena in the presence of complex geometries requires local mesh refinement, that imposes the use of finite element or discontinuous Galerkin methods. An ever larger number of steps is needed if the chosen scheme is explicit, due to the CFL condition, or the inversion of large matrices if an implicit scheme is preferred in order to alleviate the time step restriction. The local convergence of these methods has been established in a variety of cases (see [8,10,11] and references therein).

Inria Lille Nord-Europe, Équipe MÉPHYSTO & Laboratoire Paul Painlevé, Université Lille Nord de France, CNRS UMR 8525

École Centrale Paris, Lab. MAS

Address(es) of author(s) should be given

The applications we have specifically in mind are related to the recent development of the "asymptotic-preserving" schemes in the sense of Jin [13,14] for kinetic equations. Splitting systems with respect to suitable timescales was indeed proved efficient for Boltzmann-type and Fokker-Planck equations by way of micro-macro decompositions [9,15,4,3]. However, if subcycling techniques have been used in several test-cases, up to our knowledge, the asymptotic error between the exact and numerical long-time solutions has never been precisely analyzed.

We aim here at studying the convergence (error estimates and rate of convergence) of subcycled schemes and comparing them to non-subcycled schemes in simple situations. In particular, we exhibit the remarkable and unexpected asymptotic behavior of some Strang splitting schemes, which approximate better the solution in long time than locally predicted, in the spirit of the asymptotic high-order schemes developed by Aregba-Driollet, Briani and Natalini [1].

We develop our analysis on several examples which write as autonomous Cauchy problems of order one in time with a fast and a slow component in the vector field. For every example, we introduce several schemes, with and without subcycling, we perform numerical experiments on the long-time behaviour of the proposed schemes, and we provide the reader with a mathematical analysis of the numerical results. This paper is organized as follows. Sections 2 and 3 are devoted to two different examples of differential systems with two different time scales, in the spirit of the analysis of the Dahlquist equation when studying the asymptotic stability of schemes for stiff ODEs [12] and of the analysis led by Temam [17]. Both systems have exact and explicit solutions so one can do any computation and estimate involving the exact flows. The first one is analyzed in Section 2, is linear and reads[1]

$$\begin{cases} u' = -Nc(u-v) \\ v' = c(u-v), \end{cases} \tag{1}$$

where $c > 0$ and $N \in \mathbb{N}$, with $N$ being large: it is the stiffness parameter in the problem. For the numerical solutions of the linear system (1), we consider splitting schemes between the fast (*i.e.* first) equation of the system and the slow (*i.e.* second) equation. Therefore the numerical schemes will always lead to a product of matrices of the form

$$M_f(\lambda_f) := \begin{pmatrix} \lambda_f & 1-\lambda_f \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad M_s(\lambda_s) := \begin{pmatrix} 1 & 0 \\ 1-\lambda_s & \lambda_s \end{pmatrix}. \tag{2}$$

The expressions of the parameters $\lambda$ will depend on the choice of integrator (exact flow or $\theta$-scheme) and the composition of the matrices will depend on the type of splitting one wants to use (Lie or Strang type for example). We introduce the concepts of asymptotic error and asymptotic order (see Definition 1), and prove properties about the asymptotic orders of the schemes (see Propositions 2, 3, and 4) which are illustrated by several numerical experiments. We comment on the differences between schemes with and without subcycling. In Section 3, the second system that is analyzed is nonlinear and reads[2]

$$\begin{cases} u' = -Nc(u-v) - N(u-v)^2 \\ v' = c(u-v) + (u-v)^2. \end{cases} \tag{3}$$

In Section 4, we perform the same kind of analysis for a 1D linear coupled reaction-diffusion system. For this problem, the boundary conditions play a crucial role in the existence of attractive equilibrium states. We focus on two cases of boundary conditions (homogeneous and inhomogeneous Dirichlet conditions). For homogeneous boundary conditions, we introduce a subcycled Lie-splitting scheme, we address the question of the rate of convergence towards the equilibrium state (see Theorem 5) and we compare this rate to that of the exact solution (see Theorem 4). For inhomogeneous Dirichlet boundary conditions, we compare several splitting schemes with and without subcycling and we address the question of the asymptotic error which depends on both the time and space discretization parameters. For the subcycled Lie-splitting scheme, we prove that the asymptotic equilibrium state of the scheme is a uniform-in-$\delta t$ second order $L^2$-approximation of the exact asymptotic equilibrium state under a CFL-like condition (see Theorem 6). We illustrate numerically the asymptotic behavior of the Strang and Csomós[3] schemes.

[1] From the dimensional point of view, $c$ is homogeneous to the inverse of a characteristic time.

[2] Any solution of system (1) or system (3) satisfies $u' + Nv' = 0$. Hence the corresponding trajectory is included in a straight line of slope $-1/N$ in the phase space $\mathbb{R} \times \mathbb{R}$.

[3] See [5]

## 2 Full analysis of the linear system (1)

### 2.1 The exact solutions of the linear system (1)

Let us compute the exact solution of (1). We consider the matrix

$$A = \begin{pmatrix} -N & N \\ 1 & -1 \end{pmatrix}.$$

It is diagonalizable and its eigenvalues and associated spectral projectors are

$$(-(N+1),\ P_{\rm ex} = -A/(N+1)) \text{ and } \left(0,\ Q_{\rm ex} = (1,1)^{\mathsf{t}}(1,N)/(N+1)\right).$$

So the exact solution of system (1) is, for all $t \in \mathbb{R}$,

$$W(t) := (u(t), v(t))^{\mathsf{t}} = \left(\mathrm{e}^{-(N+1)ct} P_{\rm ex} + Q_{\rm ex}\right)(u^0, v^0)^{\mathsf{t}},$$

for the initial values $u^0$ and $v^0$ at time $t = 0$. In particular, we note that all the solutions converge to the equilibrium state $Q_{\rm ex}(u^0, v^0)^{\mathsf{t}}$ when $t$ tends to infinity. In the following, we fix $T > 0$ and define

$$F(T) = \mathrm{e}^{-(N+1)cT} P_{\rm ex} + Q_{\rm ex}, \tag{4}$$

the matrix of the exact flow of the system (1), the eigenvalues of which are $\mathrm{e}^{-(N+1)cT}$ and 1.

### 2.2 General properties of splitting schemes for the linear system (1)

Let $G(\delta t)$ be defined for $\delta t \in \mathcal{I}_N$ as the 2-by-2 matrix of any linear numerical flow that is a product of matrices of the form (2), where $\mathcal{I}_N$ is the intersection, that may depend on $N$, of the stability intervals of the involved schemes (see examples in Section 2.3). In the following, for all $n \in \mathbb{N}$, we will denote by

$$W^n := (u^n, v^n)^{\mathsf{t}} = (G(\delta t))^n W^0$$

the numerical solution at time $n\delta t$ starting from the initial datum $W^0 = (u^0, v^0)^{\mathsf{t}}$.

**Lemma 1** *For all $\delta t \in \mathcal{I}_N$, the matrix $G(\delta t)$ is diagonalizable, with two distinct real eigenvalues. One of these eigenvalues is 1 and the other one lies in $(0,1)$. The vector $(1,1)^{\mathsf{t}}$ is an eigenvector of $G(\delta t)$ associated to the eigenvalue 1. Hence the matrix $G(\delta t)$ reads*

$$G(\delta t) = \begin{pmatrix} 1-\alpha(\delta t) & \alpha(\delta t) \\ \beta(\delta t) & 1-\beta(\delta t) \end{pmatrix}, \tag{5}$$

*for two real-valued functions $\alpha$ and $\beta$. Moreover, the spectral decomposition of the matrix $G(\delta t)$ reads*

$$G(\delta t) = \mu(\delta t) P(\delta t) + Q(\delta t), \tag{6}$$

*where $P(\delta t)$ is the matrix of the spectral projector of $G(\delta t)$ associated to the eigenvalue $\mu(\delta t) = 1 - \alpha(\delta t) - \beta(\delta t)$ and $Q(\delta t)$ is that associated to the eigenvalue 1. In particular,*

$$Q(\delta t) = (1,1)^{\mathsf{t}}(\beta(\delta t), \alpha(\delta t))/(\alpha(\delta t) + \beta(\delta t)). \tag{7}$$

*Remark 1* We will sometimes use in the following the notation $G[\alpha, \beta]$ in reference to (5).

*Remark 2* The functions $\alpha$ and $\beta$ are polynomials of functions of type $\lambda_f$ and $\lambda_s$ (see the form of the matrices $M_f$ and $M_s$ in (2); see also examples page 6).

*Proof* Since all the matrices $M_s$ and $M_f$ have $(1,1)^{\mathsf{t}}$ for eigenvector associated with 1, so does any (finite) product of such matrices and this explains the form of the matrix $G(\delta t)$ in (5). Moreover, since all the matrices $M_s$ and $M_f$ also have their other real eigenvalue in $(0,1)$, the determinant of a product of such matrices is in $(0,1)$. Hence for all $\delta t \in \mathcal{I}_N$, $G(\delta t)$ is diagonalizable with eigenvalues 1 and $\mu(\delta t) = \mathrm{Tr}(G(\delta t)) - 1 = \det(G(\delta t)) \in (0,1)$.

With Lemma 1, we can show that the exact and numerical propagators share an interesting property:

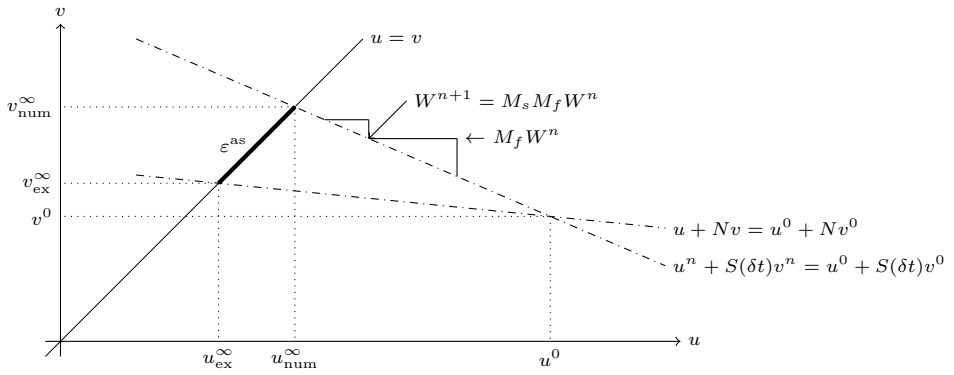


**Fig. 1** Evolution of the exact and numerical solution in the phase space $\mathbb{R}_u \times \mathbb{R}_v$. We note $W^n = (u^n, v^n)^{\mathsf{t}}$.

**Proposition 1** *For any fixed $\delta t > 0$, $(F(\delta t))^n = F(n\delta t)$ projects the vector $(u^0, v^0)^{\mathsf{t}}$ onto the line of equation $u = v$ when $n$ tends to infinity and so does $(G(\delta t))^n$ for all $\delta t \in \mathcal{I}_N$.*

*Proof* The projection property for $F(n\delta t)$ as $n \to +\infty$ relies on the decomposition (4). Using Lemma 1, we get $\forall n \in \mathbb{N},\ (G(\delta t))^n = (\mu(\delta t))^n P(\delta t) + Q(\delta t)$, with $|\mu(\delta t)| < 1$ and the result follows.

Let us denote the numerical and exact limits in time by

$$(u^\infty_{\text{num}}, v^\infty_{\text{num}})^{\mathsf{t}} = \lim_{n\to+\infty} (G(\delta t))^n (u^0, v^0)^{\mathsf{t}} \quad \text{and} \quad (u^\infty_{\text{ex}}, v^\infty_{\text{ex}})^{\mathsf{t}} = \lim_{n\to+\infty} (F(\delta t))^n (u^0, v^0)^{\mathsf{t}}.$$

Note that the numerical limit $(u^\infty_{\text{num}}, v^\infty_{\text{num}})^{\mathsf{t}}$ actually depends on $\delta t$. Therefore, we set the following definition:

**Definition 1** Let us call the asymptotic error $\varepsilon^{\text{as}} := (u^\infty_{\text{num}}, v^\infty_{\text{num}})^{\mathsf{t}} - (u^\infty_{\text{ex}}, v^\infty_{\text{ex}})^{\mathsf{t}}$. We say that the asymptotic order (A-order) is at least $p \in \mathbb{N}^\star$ if when $\delta t$ tends to 0, we have

$$\varepsilon^{\text{as}} = \mathcal{O}(\delta t^p).$$

Of course, as usual, the A-order is the supremum of the set of such $p$.

Note that, for the linear system (1), $\varepsilon^{\text{as}} = (Q(\delta t) - Q_{\text{ex}})(u^0, v^0)^{\mathsf{t}}$. We define $S(\delta t)$ as the ratio $\alpha(\delta t)/\beta(\delta t)$. Since $\forall t \in \mathbb{R}$, $u(t) + Nv(t) = u(0) + Nv(0)$, and $\forall n \geq 0$, $u^n + S(\delta t)v^n = u^0 + S(\delta t)v^0$, $\varepsilon^{\text{as}}$ can be measured in terms of the difference of the slopes of the two straight lines $u + Nv = u^0 + Nv^0$ and $u + S(\delta t)v = u^0 + S(\delta t)v^0$ (see Figure 1). More precisely,

$$\|\varepsilon^{\text{as}}\|_2 = \sqrt{2}\frac{N}{N+1}\frac{|u_0 - v_0|}{(S(\delta t)+1)}\frac{|S(\delta t) - N|}{N}.$$

Therefore

**Definition 2** The relative asymptotic error is defined as the scaled difference

$$\varepsilon^\infty := \frac{|S(\delta t) - N|}{N}.$$

For the linear system (1), the asymptotic order is studied in the following by means of the relative asymptotic error $\varepsilon^\infty$.

Our first result is the following link between the final-time classical order of a splitting method $G(\delta t)$ defined as above for the solution of system (1) and its A-order.

**Theorem 1** *Let $G(\delta t)$ be defined for $\delta t \in \mathcal{I}_N$, associated with a discretization of* (1) *and assume that it is a product of matrices of the form* (2)*. If the local order of $G(\delta t)$ is at least $p+1$, so of global order at least $p$, then its A-order is at least $p$.*

*Proof* Since the numerical flow $G(\delta t)$ has local order $p+1$, its difference with the exact flow $F(\delta t)$ reads

$$G(\delta t) - F(\delta t) = \begin{pmatrix} 1-\alpha(\delta t) & \alpha(\delta t) \\ \beta(\delta t) & 1-\beta(\delta t) \end{pmatrix} - \mathrm{e}^{-(N+1)c\delta t} P_{\mathrm{ex}} - Q_{\mathrm{ex}} = \mathcal{O}(\delta t^{p+1}).$$

This implies the following Taylor expansions for $\alpha$ and $\beta$:

$$\alpha(\delta t) = (1-\mathrm{e}^{-c(N+1)\delta t})(N/(N+1)) + \mathcal{O}(\delta t^{p+1}) \text{ and } \beta(\delta t) = (1-\mathrm{e}^{-c(N+1)\delta t})/(N+1) + \mathcal{O}(\delta t^{p+1}).$$

We infer that the slope of the equilibrium state is $S(\delta t) = \alpha(\delta t)/\beta(\delta t) = N + \mathcal{O}(\delta t^p)$.

Now, we define splitting schemes for the linear differential system (1), based on the composition of exact flows of the split vector fields and on $\theta$-schemes discretizing the split equations. We focus on their asymptotic behavior. We know from Proposition 1 and Theorem 1 that for all initial data $(u^0, v^0) \in \mathbb{R}^2$, the numerical solutions provided by such splitting schemes (assuming they are consistent with equation (1)) converge to an asymptotic state when the numerical time $n\delta t$ tends to infinity (and $\delta t$ is fixed). The typical questions of interest are the following: What is the size of this relative asymptotic error with respect to the numerical time step $\delta t$ ? Can we do better than the estimate on the A-order provided by Theorem 1 ?

### 2.3 Lie, Strang, and Csomós splitting schemes with and without subcycling for the linear system (1)

Denoting by $\delta t > 0$ the numerical time step related to the "slow" equation, the time step associated to the "fast" equation is then $\delta t/N$. The (exact or numerical) integration of the fast (resp. slow) equation of (1) over a time step $\delta t/N$ (resp $\delta t$) yields the flow

$$\Phi_{f,\delta t/N} \text{ (resp. } \Phi_{s,\delta t}) \qquad \text{with matrix} \qquad M_f(\lambda_f(\delta t/N)) \text{ (resp. } M_s(\lambda_s(\delta t))),$$

with $\lambda_s(\delta t), \lambda_f(\delta t/N) \in \mathbb{R}$. We write the Taylor expansions in $\delta t$ of $\lambda_s(\delta t)$ and $\lambda_f(\delta t/N)$ in the following way:

$$\lambda_f(\delta t/N) = 1 - c\delta t + c^2 A_f \delta t^2 + \mathcal{O}(\delta t^3) \quad \text{and} \quad \lambda_s(\delta t) = 1 - c\delta t + c^2 A_s \delta t^2 + \mathcal{O}(\delta t^3). \tag{8}$$

For any functions of $\delta t$, $\lambda_f$, $\lambda_s$, we consider the following six schemes: given $i \in \{1, \dots, 6\}$ and $W^n \in \mathbb{R}^2$, we set

$$W^{n+1} = G_i(\delta t) W^n.$$

- **Scheme #1**: (Lie type - slow time - subcycled)

$$G_1(\delta t) = M_s(\lambda_s(\delta t)) M_f(\lambda_f(\delta t/N))^N$$

- **Scheme #2**: (Lie type - fast time - no subcycling)

$$G_2(\delta t) = \big(M_s(\lambda_s(\delta t/N)) M_f(\lambda_f(\delta t/N))\big)^N$$

- **Scheme #3**: (Strang type - slow time - subcycled)

$$G_3(\delta t) = M_s(\lambda_s(\delta t/2))\ M_f(\lambda_f(\delta t/N))^N\ M_s(\lambda_s(\delta t/2))$$

- **Scheme #4**: (Strang type - fast time - no subcycling)

$$G_4(\delta t) = \big(M_s(\lambda_s(\delta t/(2N)))\ M_f(\lambda_f(\delta t/N))\ M_s(\lambda_s(\delta t/(2N)))\big)^N$$

- **Scheme #5**: (Csomós[4] type - with subcycling)

$$G_5(\delta t) = \frac{1}{2}\Big(M_s(\lambda_s(\delta t)) M_f(\lambda_f(\delta t/N))^N + M_f(\lambda_f(\delta t/N))^N M_s(\lambda_s(\delta t))\Big)$$

- **Scheme #6**: (Csomós type - without subcycling)

$$G_6(\delta t) = \frac{1}{2^N}\big(M_s(\lambda_s(\delta t/N)) M_f(\lambda_f(\delta t/N)) + M_f(\lambda_f(\delta t/N)) M_s(\lambda_s(\delta t/N))\big)^N$$

[4] See [5]

*Remark 3* Since in actual applications, the ratio $N$ between fast and slow scales in the system may not be know accurately (one may only know that it is, say, of order $10^3$), the advantage of using subcyling techniques (with a subcycling number of the same order as that of $N$) is that one can expect to achieve higher order without having to know that ratio exactly, at least on the very academic linear problem (1).

*Remark 4* When dealing with slow/fast Lie-splitting methods, one has to choose which equation will be integrated first: either the slow equation first, and then the fast one (which we denote by FS)[5], or the fast equation and then the slow one (which we denote by SF). Note that, in our very simple linear setting, the eigenvalues, eigenvectors, spectral projectors, etc of any FS splitting method can be deduced from those of a SF splitting formula in a way explained in Appendix A and the analysis extends straightforwardly. Therefore, we restrict ourselves to the study of SF Lie-splitting schemes. We also focus on FSF Stang-splitting schemes. For Csomós schemes, we take advantage of the symmetry and use both SF and FS schemes.

Using the notations of Lemma 1, we obtain the results presented in Table 1.

| Scheme # | function $\alpha$ | function $\beta$ |
|---|---|---|
| #1 | $\alpha_1(\delta t) = 1 - (\lambda_f(\delta t/N))^N$ | $\beta_1(\delta t) = (1 - \lambda_s(\delta t))(\lambda_f(\delta t/N))^N$ |
| #2 | $\alpha_2(\delta t) = 1 - \lambda_f(\delta t/N)$ | $\beta_2(\delta t) = (1 - \lambda_s(\delta t/N))\lambda_f(\delta t/N)$ |
| #3 | $\alpha_3(\delta t) = (1 - \lambda_f(\delta t/N)^N)[\lambda_s(\delta t/2)]^N$ | $\beta_3(\delta t) = (1 - \lambda_s(\delta t/2))(1 + [\lambda_f(\delta t/N)^N]\lambda_s(\delta t/2))$ |
| #4 | $\alpha_4(\delta t) = (1 - \lambda_f(\delta t/N))\lambda_s(\delta t/2)$ | $\beta_4(\delta t) = (1 - \lambda_s(\delta t/2))(1 + \lambda_f(\delta t/N)\lambda_s(\delta t/2))$ |
| #5 | $\alpha_5(\delta t) = (1 - \lambda_f(\delta t/N))^N)(1 + \lambda_s(\delta t))/2$ | $\beta_5(\delta t) = (1 - \lambda_s(\delta t))(1 + \lambda_f(\delta t/N))^N)/2$ |
| #6 | $\alpha_6(\delta t) = (1 - \lambda_f(\delta t/N)))(1 + \lambda_s(\delta t/N))/2^N$ | $\beta_6(\delta t) = (1 - \lambda_s(\delta t/N))(1 + \lambda_f(\delta t/N)))/2^N$ |

**Table 1** The functions $\alpha$ and $\beta$ for the schemes #1, #2, #3, #4, #5, and #6

*Asymptotic order* The above computations enable us to prove the following

**Proposition 2** *A Lie-splitting method such as Schemes #1 and #2 has an A-order of at least 1. Moreover, if it involves two schemes of order at least 2, then its A-order is at most 1. However, it is possible to build linear Lie-splitting methods of A-order at least 2 involving schemes of order 1.*

*Proof* The fact that Schemes #1 and #2 have order at least 1 follows from Theorem 1. Let us consider Scheme #1 and write, using the Taylor expansions (8),

$$S_1(\delta t) = \alpha_1(\delta t)/\beta_1(\delta t) = N + cN(A_s - A_f + (N+1)/2)\delta t + \mathcal{O}(\delta t^2). \tag{9}$$

When the two schemes are of order at least 2, we have $A_f = A_s = 1/2$, so that the A-order is exactly 1. To build a Lie-splitting scheme such that its A-order is at least 2, one just has to solve the equation $A_f - A_s = (N+1)/2$ for $A_f$ and $A_s$. A similar computation yields

$$S_2(\delta t) = \alpha_2(\delta t)/\beta_2(\delta t) = N + c((1 - A_f)N + A_s)\delta t + \mathcal{O}(\delta t^2). \tag{10}$$

Hence, the choice $(A_f, A_s) = (1, 0)$ leads to a Lie-splitting method of A-order at least 2 with two underlying methods of order 1.

*Remark 5* The crucial point lies in the fact that the linear combination of the coefficients $A_f$ and $A_s$ involves $N$ in both cases, so that the slow and fast schemes have to be specifically designed with the knowledge of $N$ if one wants to achieve the second A-order. The only case for which the A-order is at least 2 and $A_f$ and $A_s$ do not depend on $N$ is the exception above ($(A_f, A_s) = (1, 0)$ for Scheme #2).

**Proposition 3** *A Strang-splitting method such as Schemes #3 and #4 involving only schemes of order at least 2 has an A-order of at least 2. Moreover, it is possible to build a Strang-splitting scheme of A-order at least 2 involving two schemes of order only 1.*

[5] We chose this notation because of the usual convention on the composition of flows: the first to be applied is written on the right-hand side of the others.

*Proof* The fact that a Strang-splitting method involving two methods of order 2 is of A-order 2 comes from Theorem 1. Assume we have the same Taylor expansion as in the proof of Proposition 2. For Scheme #3, we have

$$S_3(\delta t) = \alpha_3(\delta t)/\beta_3(\delta t) = N + Nc(2A_s - 1 + 2 - 4A_f)\delta t/4 + \mathcal{O}(\delta t^2), \tag{11}$$

and for Scheme #4

$$S_4(\delta t) = \alpha_4(\delta t)/\beta_4(\delta t) = N + c(N(2A_f - 1) + 2 - 4A_s)\delta t/4 + \mathcal{O}(\delta t^2). \tag{12}$$

For example, one can choose $(A_f, A_s) = (1/4, 0)$ to have a Scheme #3 of A-order at least 2 involving two schemes of order 1.

*Remark 6* In contrast to what occurs in the Lie case, the dependence upon $N$ in the Strang subcycled scheme #3 is decoupled from the combination of $A_f$ and $A_s$.

*Remark 7* We can exchange the influence of the choices of $A_s$ and $A_f$ in the A-order by Strang-splitting with the order FSF, that is, by introducing

$$\widetilde{G_3}(\delta t) = M_f(\lambda_f(\delta t/(2N)))^N \; M_s(\lambda_s(\delta t)) \; M_f(\lambda_f(\delta t/(2N)))^N,$$
$$\widetilde{G_4}(\delta t) = (M_f(\lambda_f(\delta t/(2N))) \; M_s(\lambda_s(\delta t)) \; M_f(\lambda_f(\delta t/(2N))))^N,$$

thanks to the computations detailed in Appendix A. The coefficient in front of $\delta t^2$ is then $1 - 4A_s + 2A_f$ (resp. $2(2A_s - 1) + N(1 - 2A_f)$) for Scheme $\widetilde{\#3}$ (resp. $\widetilde{\#4}$).

**Proposition 4** *A Csomós-splitting method such as Schemes #5 and #6 has A-order at least 1. The numerical slopes of Schemes #5 and #6 read*

$$S_5(\delta t) = N + Nc\left(A_s - A_f\right)\delta t + \mathcal{O}(\delta t^2), \tag{13}$$

*and*

$$S_6(\delta t) = N + c\left(N - 1 - 2NA_f + 2A_s\right)\delta t/2 + \mathcal{O}(\delta t^2). \tag{14}$$

*Remark 8* We note that, using subcycling, one can build a one-parameter family of schemes of type #5 that has A-order 2 and whose coefficients $A_s$ and $A_s$ do not depend on $N$. Without subcycling, however, one has to face that $A_s$ and $A_f$ do depend on $N$ (except if $A_s = A_f = 1/2$) to achieve A-order 2 with a scheme of type #6.

*Convergence rate* Let us perform the same analysis on the convergence rate to equilibrium, *i.e.* the eigenvalues $\mu_i(\delta t)$, $i \in \{1, \ldots, 4\}$ of the matrices $G_i(\delta t)$ defined in Lemma 1. We get the Taylor expansions of

$$\rho_i(\delta t) = \mu_i(\delta t) - \mathrm{e}^{-c(N+1)\delta t},$$

that we summarize in Table 2. One notes at once that second order fast and slow schemes generate a second

| $i$ | $(A_f, A_s)$ |
|---|---|
| $\rho_1(\delta t)$ | $c^2(N(2A_f - 1) + 2A_s - 1)\delta t^2/2 + \mathcal{O}(\delta t^3)$ |
| $\rho_2(\delta t)$ | $c^2(N^2(2A_f - 1) + 2A_s - 1)\delta t^2/(2N) + \mathcal{O}(\delta t^3)$ |
| $\rho_3(\delta t)$ | $c^2(2N(2A_f - 1) + 2A_s - 1)\delta t^2/4 + \mathcal{O}(\delta t^3)$ |
| $\rho_4(\delta t)$ | $c^2(2N^2(2A_f - 1) + 2A_s - 1)\delta t^2/(4N) + \mathcal{O}(\delta t^3)$ |
| $\rho_5(\delta t)$ | $c^2(NA_f + A_s - (N + 1)/2)\delta t^2 + \mathcal{O}(\delta t^3)$ |
| $\rho_6(\delta t)$ | $c^2\left(N^2(2A_f - 1) + (2A_s - 1)\right)\delta t^2/(2N) + \mathcal{O}(\delta t^3)$ |

**Table 2** The functions $\rho$ for the schemes #1, #2, #3, #4, #5 and #6

order approximation of the convergence rate, as well as an A-order of 2 for Schemes #3 and #4. Besides, one can manage to construct a second order approximated rate choosing at least one of the fast and slow schemes to be of order 1, but the A-order will then be exactly 1.

*Application to $\theta$-schemes* In this paragraph, we consider two $\theta$-schemes for the numerical solutions of the fast and slow equations of system (1). We take $(\theta_f, \theta_s) \in [0,1]^2$ and we set

$$\lambda_f(\delta t) = \frac{1 - Nc\theta_f \delta t}{1 + (1-\theta_f)Nc\delta t} \qquad \text{and} \qquad \lambda_s(\delta t) = \frac{1 - c\theta_s \delta t}{1 + (1-\theta_s)c\delta t}.$$

In particular, we have

$$(A_f, A_s) = (1-\theta_f, 1-\theta_s). \tag{15}$$

Classically, in order to ensure that the associated schemes are A-stable in the classical sense (see [12]), in case $\theta_f \in (1/2, 1]$ (resp. $\theta_s \in (1/2, 1]$), we assume that $(2\theta_f - 1)cN\delta t/N < 2$ (resp. $(2\theta_s - 1)c\delta t < 2$) so that $\lambda_f^{\theta_f}(\delta t/N) \in (0,1)$ (resp. $\lambda_s^{\theta_s}(\delta t) \in (0,1)$). The stability interval $\mathcal{I}_N$ defined at the beginning of Section 2.2 is the intersection of the corresponding domains in $\delta t$. Our choice of different time steps for the slow and fast equations in order to use subcycling techniques implies that $\mathcal{I}_N$ is independent of $N$ in that case.

The results of the previous paragraphs provide us with the following propositions, when the underlying numerical integration methods are $\theta$-schemes. For Lie-splitting methods (Schemes #1 and #2):

**Proposition 5** *Assume $N > 1$. The only scheme of type #1 or #2 of A-order at least 2 involving two $\theta$-schemes is of type #2 with $\theta_s = 1$ (fully explicit) and $\theta_f = 0$ (fully implicit). In this very particular case, the A-order is infinite because $\alpha_2 = N\beta_2$.*

*Proof* Plugging relation (15) in the Taylor expansions (9) and (10), the result follows by cancelling the terms of order 1.

*Remark 9* Note that, if a fully implicit scheme is at hand for the fast equation, it seems unwise to use a subcycling technique anyway, since there is no stability constraint on $\delta t$ from the fast scheme part.

*Remark 10* In particular, in view of relation (15) and of the Taylor expansions in the first two lines of Table 2, no Lie-splitting scheme of type #1 or #2 has A-order 2 with an approximation of order 3 of the rate of convergence.

For Strang-splitting methods (Schemes #3 and #4) involving $\theta$-schemes:

**Proposition 6** *There exists a one-parameter family of schemes of type #3 with A-order 2, and another one of schemes of type #4 with A-order 2.*

*Proof* Plugging relation (15) in the Taylor expansions (11) and (12), the result follows by cancelling the terms of order 1.

*Remark 11* Note that, without subcycling (Scheme #4), the one-parameter family of schemes depends on $N$ (through the equation $2N(1-2\theta_s) + 2\theta_f - 1 = 0$). On the contrary, with subcycling (Scheme #3), the one-parameter family is independant of $N$ (since the link between $\theta_f$ and $\theta_s$ is $4\theta_f - 2 + 1 - 2\theta_s = 0$).

**Proposition 7** *Using $\theta$-schemes, it is then possible to build a scheme of type $\widetilde{\#3}$ (see Remark 7) of A-order 2 with an explicit fast scheme ($\theta_f = 1$) and a semi-implicit slow scheme ($\theta_s = 3/4$).*

*Proof* The coefficient in front of $\delta t^2$ in the asymptotic error expansion is then $4\theta_s - 2 + 1 - 2\theta_f = 0$.

*Remark 12* Once again, in addition to having more reasonable computational costs and relaxing stability constraints, using subcycling techniques allows to derive families of schemes involving explicit schemes and with reasonable high A-order (2, in this example with a Strang composition method).

**Proposition 8** *Assume $N > 1$. The only combination of $\theta$-schemes leading to a third order approximated rate of convergence and having A-order 2 consists in taking the Crank-Nicolson scheme for both the fast and slow schemes, using the Strang splitting (Schemes #3 and #4).*

*Proof* For schemes of type #3, plugging the relation (15) in the Taylor expansion (11) and cancelling the term of order 1 yields a link between $\theta_f$ and $\theta_s$ which does not match the condition of cancellation of the term of order 2 in $\rho_3(\delta t)$ (see Table 2) except when $(\theta_f, \theta_s) = (0.5, 0.5)$. The proof is very similar for schemes of type #4.

**Proposition 9** *There exists a one-parameter family of schemes involving $\theta$-schemes, one of type #5 and another one of type #6, with A-order 2.*

*Proof* Plugging (15) into the Taylor expansion (13) and (14) yields the result.

*Remark 13* Note once again that the one-parameter family is independent of $N$ for the subcycled Csomós scheme #5, while it depends on $N$ for the non-subcycled Csomós scheme #6. This is an extra advantage of subcycled schemes when $N$ is not known exactly (see Remark 3). Moreover, using a Csomós scheme with subcycling (type #5) allows to use a composition of two explicit schemes ($\theta_f = \theta_s = 1$) which has A-order 2 (see Fig. 2).

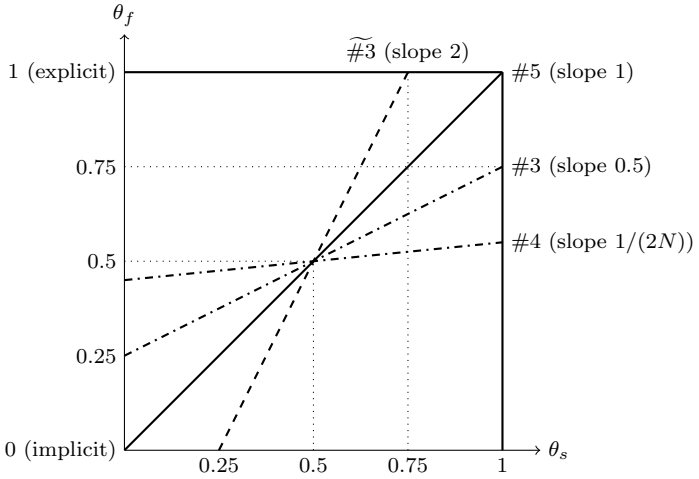


**Fig. 2** One-parameter families of splitting schemes of A-order 2

### 2.4 Conclusion

Let us remind the reader that the applications we have in mind are by far more complicated than the system (1). However, they share with the system (1) the property that they involve a fast equation for which an implicit scheme is costly or hard to solve, thus implying the use of an explicit scheme, inducing a stability constraint on the numerical time-step $\delta t$. In that case, the subcycling techniques are computationally less costly, thus relevant.

We proved in this section that, in view of the aforementioned goal, we can indeed build two schemes, one of type $\widetilde{\#3}$ (Strang with subcycling) with $\theta_f = 1$ (explicit) and $\theta_s = 3/4$ (semi-implicit), and one of type #5 (Csomós) with $\theta_s = \theta_f = 1$ (explicit/explicit) which are of A-order 2, even though they are (locally) consistent of order 1 with (1) and have a rate of convergence which approximates the exact rate at order 2. Moreover, the coefficients $\theta_f$ and $\theta_s$ of these schemes are independent of $N$.

We postpone the numerical illustration of these results to the study of a nonlinear system in the following section.

## 3 Analysis of the nonlinear system (3)

### 3.1 Analysis of the exact solutions of the system (3)

In this section, we investigate the long time behavior of the two-scale nonlinear system (3). Let us first write this system in the form

$$\begin{cases} u' = -N(u-v)[c+(u-v)] \\ v' = \quad (u-v)[c+(u-v)]. \end{cases} \tag{16}$$

This way, we are able to derive the following

**Proposition 10** *Let $(u^0, v^0) \in \mathbb{R}^2$ be given. The maximal solution starting at $(u^0, v^0)$ lies on the straight line of equation $u + Nv = u^0 + Nv^0$. It is defined for all non-negative time if $u^0 + c \geq v^0$ and it dies in finite positive time if $u^0 + c < v^0$. Moreover, if $u_0 + c = v_0$ then the solution is constant, and if $u^0 + c > v^0$ then the solution tends to the intersection of the two straight lines of equations $u + Nv = u^0 + Nv^0$ and $u = v$, i.e. to the point of coordinates $(u^0 + Nv^0)/(N+1) \times (1,1)$.*

*Proof* The linear change of variable $(X, Y) = (u + Nv, u - v)$ yields the equivalent differential system

$$\begin{cases} X' = 0, \\ Y' = -(N+1)Y(c+Y). \end{cases}$$

The second equation of this system has for maximal solution starting at $t = 0$ in $Y^0 \in \mathbb{R}$ the function $Y(t) = Y^0 \mathrm{e}^{-c(N+1)t}/(1 + (1 - \mathrm{e}^{-c(N+1)t})Y^0/c)$ defined as long as $-c < Y^0(1 - \mathrm{e}^{-c(N+1)t})$. The result on the existence time for the maximal solutions of (16) follows from this observation. Moreover, if $Y^0 > 0$, then $Y(t)$ tends to 0 when $t$ tends to $+\infty$. This proves the asymptotic behavior of the corresponding maximal solutions.

Hence, for the range of interest of initial values ($(u_0, v_0)$ such that $u_0 + c > v_0$), the qualitative behaviour is the same for the linear system (1) and for the nonlinear system (16): the solutions evolve on straight lines of equation $u + Nv = \text{cste}$ and converge to an equilibrium point located on the line of equation $u = v$. Therefore, we extend the Definition 1 of the asymptotic error $\varepsilon^{\text{as}}$ to this nonlinear case as well.

### 3.2 Splitting schemes with or without subcycling for the nonlinear problem (3)

Let us prove this somehow classical result providing an estimate of the local order of a splitting scheme (with or without subcycling) as a function of the order of the underlying schemes and the order of the splitting method.

**Theorem 2** *Let us consider a differential system of the form*

$$\begin{cases} u' = Nf(u,v) \\ v' = \ g(u,v), \end{cases}$$

*where $f$ and $g$ are smooth functions from $\mathbb{R}^2$ to $\mathbb{R}$. We denote by $\varphi_{\text{ex},\delta t}$ the exact flow at time $\delta t$ of this equation. Let us denote by $\varphi_f(\delta t)$ (respectively $\varphi_s(\delta t)$) the propagators at time $\delta t$ of the two split equations:*

$$\begin{cases} u' = Nf(u,v) \\ v' = \quad 0 \end{cases} \qquad \text{(resp.)} \qquad \begin{cases} u' = \quad 0 \\ v' = g(u,v). \end{cases}$$

*Assume that $S_{f,\delta t}$ and $S_{s,\delta t}$ are numerical methods of respective orders $n_f$ and $n_s$. Assume that a splitting method is defined for $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{R}$ by the formula*

$$\Phi_{\delta t} = \Pi_{i=1}^{n} (S_{s,b_i \delta t} \circ S_{f,a_i \delta t}),$$

*so that this method with the exact flows has order $n_{\text{ex}}$. Then the order of the method $\Phi_{\delta t}$ is at least $\min(n_f, n_s, n_{\text{ex}})$, and so is the order of the method with subcycling*

$$\Phi_{\delta t}^{sc} = \Pi_{i=1}^{n} (S_{s,b_i \delta t} \circ (S_{f,a_i \delta t/N})^N). \tag{17}$$

*Proof* Since the method $S_{s,\delta t}$ has order $n_s$, we may write, when $\delta t \to 0$,

$$S_{s,\delta t} = \varphi_{s,\delta t} + \mathcal{O}(\delta t^{n_s+1}) \quad \text{and} \quad S_{f,\delta t/N} = \varphi_{f,\delta t/N} + \mathcal{O}(\delta t^{n_f+1}).$$

The smoothness of the propagators implies that for all $j \in \mathbb{N}^\star$,

$$S_{f,\delta t/N}^j = \varphi_{f,\delta t/N}^j + \mathcal{O}(\delta t^{n_f+1}),$$

where the constant in the Landau symbol depends on $j$. In particular, for $j = N$, using the semi-group property of the exact flow, we have

$$S_{f,\delta t/N}^N = \varphi_{f,\delta t} + \mathcal{O}(\delta t^{n_f+1}).$$

This implies that

$$\begin{aligned} \Phi_{\delta t}^{sc} = \Pi_{i=1}^{n} (S_{s,b_i \delta t} \circ (S_{f,a_i \delta t/N})^N) &= \Pi_{i=1}^{n} (\varphi_{s,b_i \delta t} + \mathcal{O}(\delta t^{n_s+1})) \circ (\varphi_{f,a_i \delta t} + \mathcal{O}(\delta t^{n_f+1})) \\ &= \Pi_{i=1}^{n} (\varphi_{s,b_i \delta t} \circ \varphi_{f,a_i \delta t}) + \mathcal{O}(\delta t^{\min(n_f,n_s)+1}) \\ &= \varphi_{\text{ex},\delta t} + \mathcal{O}(\delta t^{\min(n_f,n_s,n_{\text{ex}})+1}), \end{aligned}$$

since the splitting method is assumed to have order $n_{\text{ex}}$ when used with the exact flows. This proves the result for $\Phi_{\delta t}^{sc}$. The proof for $\Phi_{\delta t}$ is even simpler.

In the following, we consider numerical splitting methods for the nonlinear problem (3) in the same way as for the linear problem (1) in Section 2.3:

- Scheme #1 is a SF Lie-splitting method with subcycling,
- Scheme #2 is a SF Lie-splitting method without subcycling,
- Scheme #3 is a FSF Strang-splitting method with subcycling,
- Scheme #4 is a FSF Strang-splitting method without subcycling,
- Scheme #5 is a Csomós-splitting method with subcycling, and
- Scheme #6 is a Csomós-splitting method without subcycling.

Once again, we use $\theta$-schemes to integrate the split equations numerically: we chose $(\theta_f, \theta_s) \in [0,1]^2$ and define $\Phi_{f,\delta t}$ and $\Phi_{s,\delta t}$ as follows. For the fast equation, the first component $u^{n+1}$ of $\Phi_{f,\delta t}(u^n, v^n)$ solves the equation in $X$

$$X - u^n = N\delta t\theta_f\big(c(v^n - u^n) - (u^n - v^n)^2\big) + N\delta t(1-\theta_f)\big(c(v^n - X) - (X - v^n)^2\big),$$

while its second one is its second argument $v^n$. For the slow equation, the second component $v^{n+1}$ of $\Phi_{s,\delta t}(u^{n+1}, v^n)$ solves the equation in $X$

$$X - v^n = \delta t\theta_s\big(c(u^{n+1} - v^n) + (u^{n+1} - v^n)^2\big) + \delta t(1-\theta_s)\big(c(u^{n+1} - X) + (u^{n+1} - X)^2\big),$$

while its first component is its first argument $u^{n+1}$.

### 3.3 Numerical examples of splitting methods for problem (3)

We run the six schemes with six different values of the couple $(\theta_f, \theta_s)$. We sum up the results on the asymptotic order in Table 3 and provide numerical results in Figure 3. These results were obtained with final time $T = 5.0$, speed $c = 1$, factor $N = 10$, initial datum $(u^0, v^0) = (5, 1)$, so that, using the analysis carried out in the proof of Proposition 10, the exact solution at final time is within a distance smaller than $10^{-20}$ of its asymptotic limit $15/11 \times (1,1)^{\mathsf{t}}$.

By Theorem 2, we know that the Lie-splitting schemes (Scheme #1 and Scheme #2) are of classical order 1 for any possible choice of $(\theta_f, \theta_s)$. The first two columns of Table 3 show that the asymptotic order is also 1 in these cases, except when $(\theta_f, \theta_s) = (0, 1)$. This is in accordance with the results obtained in Proposition 5 for the linear system (1) since in this case, the A-order of Schemes #1 and #2 is 1 except when $(\theta_f, \theta_s) = (0, 1)$ and the A-order is infinite (see Proposition 5). Theorem 2 also implies that the Strang-splitting scheme #3 is at least of classical order 1 with the choice $(\theta_f, \theta_s) = (1, 0)$ and the asymptotic orders collected in the middle of the third line of Table 3 show that the numerical asymptotic order is also 1 in this case. The same theorem also ensures that Scheme #3 has order 2 when applied with $(\theta_f, \theta_s) = (1/2, 1/2)$. The asymptotic orders displayed in the middle of the fourth line of Table 3 show that the asymptotic order is also 2 in this case. The last two lines are even more interesting: for $(\theta_f, \theta_s) = (1, 3/4)$ and $(\theta_f, \theta_s) = ((N-1)/(2N), 1/4)$, the classical order of the Strang splitting method is, by Theorem 2 at least 1. In the first case $(\theta_f, \theta_s) = (1, 3/4)$, the numerical results suggest that the subcycled Scheme #3 has A-order 2 while the non-subcycled Scheme #4 has A-order 1. We recall that, for these parameters, the Scheme #3 was of A-order 2 in the linear setting (see Remark 6). In the second case $(\theta_f, \theta_s) = ((N-1)/(2N), 1/4)$, the same phenomenon occurs: Scheme #3 has A-order 1 while Scheme #4 has A-order 2. We recall that these values of the parameters were chosen in the linear setting in such a way that the Scheme #4 has A-order 2. The Csomós scheme without subcycling (Scheme #6) applied to the nonlinear problem (16) is of numerical A-order 1 except when $\theta_s = \theta_f = 1/2$ and the numerical A-order is 3 (see Table 3). This is in good accordance with results for the linear case proved in Section 2.3 since, for the linear problem (1), we have

$$S_6(\delta t) = N + \frac{1}{2}\,c\,\big(1 - 2\theta_s - N(1-2\theta_f)\big)\,\delta t + \frac{1}{4}\,c^2\,\big((2\theta_f - 1)(1 - N + 2N\theta_f - 2\theta_s)\big)\,\delta t^2 + \mathcal{O}(\delta t^3),$$

and the terms of order 1 and 2 in the Taylor expansion of $S_6(\delta t)$ vanish for these values of $\theta_s$ and $\theta_f$. The Csomós splitting scheme with subcycling (Scheme #5) applied to the nonlinear problem (16) is indeed of numerical A-order 2 in general when $\theta_f = \theta_s$, and is of numerical A-order 2 in other cases. The two relatively high values on the last 2 lines of the corresponding row of Table 3 are due to the fact that $\delta t$ was not small enough to reach the actual rate. These results are in good accordance with the results proved for the linear problem (1) (see (13) and (15)).

Roughly speaking, a subcycled scheme (odd number) requires half as many numerical computations as the corresponding not-subcycled scheme (even number), since the computational ratio is of order $(N+1)/(2N) \sim 1/2$. Therefore, for a given precision $\varepsilon > 0$ to be achieved on the asymptotic state, the previous analysis suggests to use a subcycled method with high order. For example, for the integration of the nonlinear problem (16), provided $T > 0$ is chosen big enough, the subcycled Scheme #3, which has A-order 2 (and whose coefficients $\theta_f$ and $\theta_s$ do not depend on the value of $N$[6]), will require $\mathcal{O}((N+1) \times T/\varepsilon^{1/2})$ computations, while its not-subcycled analogue Scheme #4, which has A-order 1, will require $\mathcal{O}(2N \times T/\varepsilon)$ computations.

[6] See Remark 3

### 3.4 Conclusion

These examples suggest that, in this context, the $A$-order of a scheme applied to the linear problem is the same as the $A$-order of the scheme applied to the nonlinear problem. This can be explained by the fact that the two problems (1) and (3) have the same set of attractive equilibrium points (the straight line $u = v$), they project the initial datum $(u_0, v_0)$ (chosen in an appropriate subset of the phase plane $(u^0 + c > v^0)$) on the same equilibrium point $(u^0 + Nv^0)/(N + 1) \times (1, 1)$, and in the neighborhood of this equilibrium point, $(u - v)^2 << |u - v|$. In particular, these examples show that it is possible to build in the nonlinear setting, as well in the linear setting, splitting methods with asymptotic order greater than the classical order of the schemes used for solving the split-equations.

| $(\theta_f, \theta_s)$ | Scheme #1 | Scheme #2 | Scheme #3 | Scheme #4 | Scheme #5 | Scheme #6 |
|---|---|---|---|---|---|---|
| $(0.0, 1.0)$ | 0.8642 | – | 0.7700 | 0.9787 | 1.2941 | 1.0001 |
| $(1.0, 1.0)$ | 0.8693 | 1.0072 | 1.4769 | 1.0409 | <u>1.9647</u> | 1.0055 |
| $(1.0, 0.0)$ | 0.8404 | 1.0000 | 1.1860 | 1.0229 | 0.8734 | 1.0002 |
| $(0.5, 0.5)$ | 0.8534 | 1.0000 | <u>1.8313</u> | <u>1.9984</u> | <u>1.8888</u> | <u>2.4817</u> |
| $(1.0, 0.75)$ | 0.8617 | 1.0053 | <u>1.8674</u> | 1.0354 | 1.8373 | 1.0042 |
| $(\frac{N-1}{2N}, 0.25)$ | 0.8463 | 0.9975 | 1.3994 | <u>1.9926</u> | 1.8184 | 0.9955 |

**Table 3** Asymptotic error for the 6 schemes for some values of $(\theta_f, \theta_s)$. Figures are underlined when the method is of A-order at least 2.

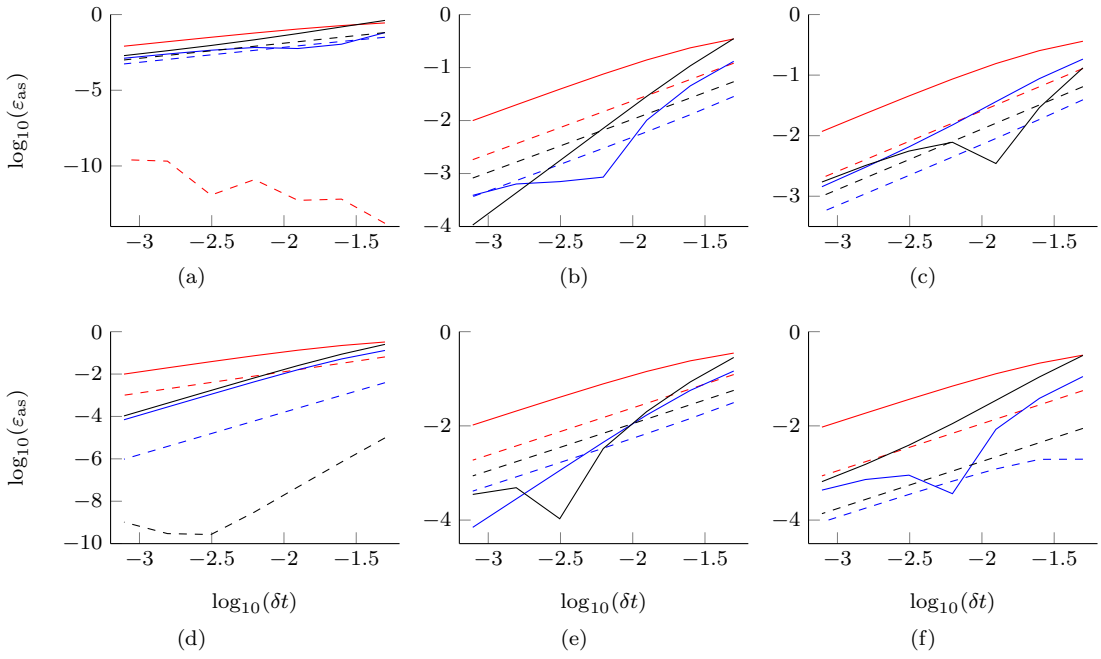


**Fig. 3** Logarithm of the asymptotic error as a function of the logarithm of the time step: Scheme #1 (solid red line), Scheme #2 (dotted red line), Scheme #3 (solid blue line), Scheme #4 (dotted blue line), Scheme #5 (solid black line), Scheme #6 (dotted black line). $(\theta_f, \theta_s) = (0.0, 1.0)$ (a), $(\theta_f, \theta_s) = (1.0, 1.0)$ (b), $(\theta_f, \theta_s) = (1.0, 0.0)$ (c), $(\theta_f, \theta_s) = (0.5, 0.5)$ (d), $(\theta_f, \theta_s) = (1.0, 0.75)$ (e) and $(\theta_f, \theta_s) = ((N - 1)/(2N), 0.25)$ (f).

## 4 A coupled reaction-diffusion system

### 4.1 The homogeneous Dirichlet problem

*The continuous problem* This section aims at studying the behavior of time-splitting schemes involving subcycling techniques for solving the following system of partial differential equations

$$\begin{cases} \partial_t u = \nu_1 \Delta u + c_1(v-u) \\ \partial_t v = \nu_2 \Delta v + c_2(u-v) \end{cases} \qquad t>0, x\in(0,L), \tag{18}$$

with homogeneous Dirichlet boundary conditions at $x=0$ and $x=L$, and given initial data $u^0$ and $v^0$ in an appropriate function space. Here, $\Delta = \partial_x^2$ is the Laplace operator and $L>0$ is given. Moreover, $\nu_1$ and $\nu_2$ are real positive diffusion parameters and $c_1$ and $c_2$ are real positive reaction speed parameters. We focus on the case where one of the equations in System (18) is "fast" and the other is "slow". Moreover, we assume the "speed" ratios allow us to actually do subcycling. This means that

$$\frac{\nu_1}{\nu_2} = \frac{c_1}{c_2} = N \in \mathbb{N}^\star, \tag{19}$$

and $N >> 1$[7]. Recall that one can expect to have similar results when only the order of magnitude of $N$ is known (See Remark 3 for the ODE system of Section 2), but we assume that $N$ is exactly known via relation (19) to keep the notations and the analysis simple. Consequently, in accordance with Section 2, we will use the notation $\nu = \nu_2$ and $c = c_2$. In that case, the first equation in (18) is the "fast" one, and the second one is the slow one since it reads

$$\begin{cases} \partial_t u = N\nu \Delta u + Nc(v-u) \\ \partial_t v = \nu \Delta v + c(u-v) \end{cases} \qquad t>0, x\in(0,L), \tag{20}$$

Therefore, $u$ is referred to as the fast unknown and $v$ as the slow one. Let us recall that we have the following

**Theorem 3** *For all initial data $(u^0,v^0)\in L^2(0,L)^2$, System (20) has a unique solution $t\mapsto (u(t),v(t))$ in $C^0([0,+\infty),L^2(0,L)^2)\cap C^\infty((0,+\infty)\times[0,L],\mathbb{R}^2)$, satisfying $(u,v)(0)=(u^0,v^0)$.*

*Proof* If one looks for solutions of the form

$$u(t,x) = \sum_{k=1}^{+\infty} \alpha_k(t)\sin\left(k\pi x/L\right) \quad \text{and} \quad v(t,x) = \sum_{k=1}^{+\infty} \beta_k(t)\sin\left(k\pi x/L\right),$$

then the coefficients satisfy the differential systems

$$\dot\alpha_k(t) = -N\left(c+\nu\frac{k^2\pi^2}{L^2}\right)\alpha_k(t) + Nc\beta_k(t), \qquad \dot\beta_k(t) = c\alpha_k(t) - \left(c+\nu\frac{k^2\pi^2}{L^2}\right)\beta_k(t),$$

and the eigenvalues $\lambda_k$ and $\mu_k$ of the matrices $M_k = \begin{pmatrix} -N\left(c+\nu\frac{k^2\pi^2}{L^2}\right) & +Nc \\ +c & -\left(c+\nu\frac{k^2\pi^2}{L^2}\right) \end{pmatrix}$ are both real, negative and satisfy, when $k$ tends to $+\infty$,

$$\lambda_k \sim -N\nu\frac{k^2\pi^2}{L^2} \quad \text{and} \quad \mu_k \sim -\nu\frac{k^2\pi^2}{L^2}.$$

The following theorem deals with the asymptotic behavior of the solutions of System (20):

**Theorem 4** *For all solutions $(u,v)$ of System (20) and all $t\geq 0$, we have*

$$\int_0^L (|u^2| + N|v|^2)(t)\mathrm{d}x \leq \Big(\int_0^L (|u|^2 + N|v|^2)(0)\mathrm{d}x\Big)\mathrm{e}^{-\frac{2\pi^2\nu}{L^2}t}. \tag{21}$$

[7] Yet, we are not interested in the limit $N\to+\infty$.

*Proof* Let $(u,v)$ be a smooth solution of (20). We compute

$$\begin{aligned}\left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\int_0^L(|u|^2+N|v|^2)\mathrm{d}x\right)(t) &= N\nu\int_0^L u(t)\Delta u(t)+N\nu\int_0^L v(t)\Delta v(t)+Nc\int_0^L(u(v-u)+v(u-v))(t)\\ &= -N\nu\int_0^L|\nabla u(t)|^2-\nu\int_0^L N|\nabla v(t)|^2-Nc\int_0^L|u(t)-v(t)|^2\\ &\leq -\frac{2\pi^2\nu}{L^2}\frac{1}{2}\int_0^L(|u(t)|^2+N|v(t)|^2)\mathrm{d}x,\end{aligned}$$

using that $N\geq 1$ and Poincaré's inequality.

The goal of the next paragraphs is to show how this exponential convergence to 0 in $L^2(0,L)$ is reproduced by splitting schemes with (or without) subcycling.

*The space discretization* In the following, we will use the classical finite-difference discretization of minus the Laplace operator, using the symmetric tridiagonal $J\times J$ matrix $A=toeplitz(-1,2,-1,0)$ where $J\in\mathbb{N}^\star$ and $\delta x=L/(J+1)$. We note for all $i\in\{0,\dots,J+1\}$, $x_i=i.\delta x$ and $U=(u_1,\dots,u_J)^{\mathsf{t}}$ will be the solution of the discretized problem. Let us recall that the eigenvalues and associated eigenvectors of $A$ are, for $1\leq j\leq J$,

$$\left(\lambda_j=4\sin^2\left(\frac{j\pi}{2(J+1)}\right),\left(\sin(1j\pi/(J+1)),\sin(2j\pi/(J+1)),\dots,\sin(Jj\pi/(J+1))\right)^{\mathsf{t}}\right). \tag{22}$$

In the following, we denote by

$$A=ZDZ^{-1}, \tag{23}$$

the corresponding diagonalization of $A$. We endow $\mathbb{R}^J$ with the classical Euclidian norm

$$\forall(u_1,\dots,u_J)^{\mathsf{t}}\in\mathbb{R}^J,\qquad \|(u_1,\dots,u_J)^{\mathsf{t}}\|_2:=\sqrt{\frac{1}{J+1}\sum_{i=1}^J|u_i|^2}=\sqrt{\frac{\delta x}{L}\sum_{i=1}^J|u_i|^2}.$$

We use a similar definition for the Euclidian norm on $\mathbb{R}^J\times\mathbb{R}^J$, which we also denote by $\|\cdot\|_2$. We use the induced norms on the corresponding algebras of square matrices which we denote by $|||\cdot|||_2$.

*The time discretization* Assume $\delta t>0$ is given. The methods we have in mind all share the same basic idea: we discretize in time separately the spatially-discretized versions of both equations of System (20). We consider $(p,p',q,q')\in(\mathbb{N}^*)^4$ such that

$$\frac{q'}{q}=\frac{p'}{Np}. \tag{24}$$

The "fast" one is discretized on an interval of length $\delta t/(Np)$ and we denote by $\Phi_{f,\delta t/(Np)}$ its numerical flow. We iterate this method $p'$ times. The "slow" one is discretized on an interval of length $\delta t/q$ and we denote by $\Phi_{s,\delta t/q}$ its numerical flow. We iterate this method $q'$ times. Then, we compute numerical flows using splitting methods and subcycling by considering numerical flows such as

$$\Psi_{Lie,\delta t}=\Phi_{s,\delta t}\circ\Phi_{f,\delta t/N}^N, \tag{25}$$

corresponding to $(p,p',q,q')=(1,N,1,1)$. As we did in Section 2 and in Section 3, we consider $\theta$-schemes for the solution of the slow and fast equations. We choose two parameters $(\theta_f,\theta_s)\in[0,1]^2$. The numerical integrators involved in the splitting scheme therefore read:

$$\Phi_{f,\delta t/N}(u^n,v^n)=\left[\left(I-\theta_f\delta t\left(cI+\nu\frac{1}{(\delta x)^2}A\right)\right)\left(I+(1-\theta_f)\delta t\left(cI+\nu\frac{1}{(\delta x)^2}A\right)\right)^{-1}u^n+c\delta t v^n\,,\,v^n\right], \tag{26}$$

and

$$\Phi_{s,\delta t}(u^n,v^n)=\left[u^n\,,\,\left(I-\theta_s\delta t\left(cI+\nu\frac{1}{(\delta x)^2}A\right)\right)\left(I+(1-\theta_s)\delta t\left(cI+\nu\frac{1}{(\delta x)^2}A\right)\right)^{-1}v^n+c\delta t u^n\right], \tag{27}$$

where $I$ stands for the identity matrix. This way, a stability condition reads

$$\delta t\leq\frac{1}{c+4\nu/(\delta x)^2}. \tag{28}$$

Note also that the stability condition (28) of the scheme is actually independent of $N$, and this is a very interesting feature of splitting schemes involving subcycling. Let us define for $i \in \{s, f\}$,

$$B_i(\delta t) := I - \theta_i \delta t \left( cI + \nu \frac{1}{(\delta x)^2} A \right) \qquad \text{and} \qquad C_i(\delta t) := I + (1-\theta_i)\delta t \left( cI + \nu \frac{1}{(\delta x)^2} A \right).$$

For the sake of simplicity, we omit the dependence in $\delta t$ of $C$ and $B$, thus noting $(B,C)_s = (B,C)_s(\delta t/q)$ and $(B,C)_f = (B,C)_f(\delta t/p)$. Since they are polynomials in $A$, the matrices $I$, $C_s$, $C_f$, $B_s$, $B_f$, $C_s^{-1}$, $C_f^{-1}$ and $A$ do commute for all values (distinct or not) of $\delta t$. The matrices of the linear mappings $\Phi_{s,\delta t/q}$ and $\Phi_{f,\delta t/(Np)}$ in the canonical basis of $\mathbb{R}^{2J}$ read respectively

$$M_s(\delta t/q) = \begin{pmatrix} I & 0 \\ c\frac{\delta t}{q} C_s^{-1} & B_s C_s^{-1} \end{pmatrix} \qquad \text{and} \qquad M_f(\delta t/(Np)) = \begin{pmatrix} B_f C_f^{-1} & c\frac{\delta t}{p} C_f^{-1} \\ 0 & I \end{pmatrix}. \tag{29}$$

Let us define $\Sigma_{i,m} = \sum_{k=0}^{m-1} (C_i^{-1} B_i)^k$ for $m \geq 1$ and $i \in \{s, f\}$. Therefore, the matrix of $\Phi^{p'}_{f,\delta t/(Np)}$ reads

$$M_f(\delta t/(Np))^{p'} = \begin{pmatrix} (B_f C_f^{-1})^{p'} & c\frac{\delta t}{p} C_f^{-1} \Sigma_{f,p'} \\ 0 & I \end{pmatrix}.$$

Recalling (24), we define $\Psi_{\delta t,p,p',q,q'} = \Phi^{q'}_{s,\delta t/q} \circ \Phi^{p'}_{f,\delta t/(Np)}$ the matrix of which reads

$$\begin{pmatrix} (B_f C_f^{-1})^{p'} & c\frac{\delta t}{p} C_f^{-1} \Sigma_{f,p'} \\ c\frac{\delta t}{q} C_s^{-1} (B_f C_f^{-1})^{p'} \Sigma_{s,q'} & (B_s C_s^{-1})^{q'} + c^2 \frac{\delta t^2}{pq} C_s^{-1} C_f^{-1} \Sigma_{s,q'} \Sigma_{f,p'} \end{pmatrix}. \tag{30}$$

In particular, if $q = q' = p = 1$ and $p' = N$, $\Psi_{\delta t,p,p',q,q'} = \Psi_{Lie,\delta t}$.

*Numerical analysis of the rate of convergence for the subcycled SF Lie-splitting scheme* The central result of this subsection is the following analysis of the rate of convergence to 0 of the numerical solutions of Problem (20) provided by the subcycled SF Lie method:

**Theorem 5** *Let $c, \nu > 0$, $N \geq 2$. Assume $J \in \mathbb{N}^\star$ is given. There exists*[8] *$C, \gamma, h > 0$ such that for all $T > 0$, all $U^0, V^0 \in \mathbb{R}^J$, all $\delta t \in (0, h)$ and all $n \in \mathbb{N}$ with $n\delta t \leq T$, we have*

$$\|\Psi^n_{Lie,\delta t}(U^0, V^0)\|_2 \leq C \mathrm{e}^{-\gamma n \delta t} \|(U^0, V^0)\|_2. \tag{31}$$

*Remark 14* Note that one can impose $\gamma \geq N\nu\lambda_1/((N+1)(\delta x)^2)$ in this case (provided $h$ is small enough). Since $N\nu\lambda_1/(\delta x)^2 \to N\nu\frac{\pi^2}{L^2}$ as $\delta x \to 0^+$ (or equivalently as $J \to +\infty$), we have, at least asymptotically with respect to $\delta x$, a numerical decay rate of the appropriate order with respect to the parameters $\nu$ and $L$: we compare the exact decay rate $\nu\pi^2/L^2$ from Theorem 4 (21) with the asymptotic numerical one $N\nu\pi^2/(L^2(N+1))$ (recall that $N$ is large).

*Proof* We perform a numerical analysis of the linear splitting method $\Psi^n_{Lie,\delta t}$. We determine its eigenvalues, show that they are real positive and control the biggest one to obtain the exponential decay stated in (31). Let $(p, p', q, q')$ be positive integers satisfying (24). Denoting by $\mathcal{Z}$ the matrix (see (23))

$$\mathcal{Z} = \begin{pmatrix} Z & 0 \\ 0 & Z \end{pmatrix}, \tag{32}$$

we obtain that the matrix $\mathcal{D} := \mathcal{Z}^{-1} \Psi_{\delta t,p,p',q,q'} \mathcal{Z}$ is exactly the same as that of (30) where $A$ is replaced with $D$ in the definition of the matrices $B_f, B_s, C_f$ and $C_s$. In particular, it consists in four square blocks, each of size $J \times J$, each of which is diagonal. We infer that all the eigenvalues of $\Psi_{\delta t,p,p',q,q'}$ are the roots of the $J$ polynomial equations

$$\tau^2 - \left( (\phi_f^{-1}\psi_f)^{p'} + (\phi_s^{-1}\psi_s)^{q'} + c^2 \frac{\delta t^2}{pq} \phi_f^{-1} \phi_s^{-1} \widetilde{\Sigma}_{s,q'} \widetilde{\Sigma}_{f,p'} \right) \tau + (\phi_f^{-1}\psi_f)^{p'} (\phi_s^{-1}\psi_s)^{q'} = 0, \tag{33}$$

[8] There is a constant $C \geq 1$ due to the lack of symmetry of the matrix $\mathcal{D}$.

where

$$\psi_{f,s}(\mu) = 1 - \theta_{f,s}\frac{\delta t}{p}\mu \quad \text{and} \quad \phi_{f,s}(\mu) = 1 + (1-\theta_{f,s})\frac{\delta t}{p}\mu, \tag{34}$$

$$\widetilde{\Sigma}_{f,p'} = \sum_{k=0}^{p'-1} (\phi_f^{-1}\psi_f)^k \quad \text{and} \quad \widetilde{\Sigma}_{s,q'} = \sum_{k=0}^{q'-1} (\phi_s^{-1}\psi_s)^k, \tag{35}$$

and $\mu$ is an eigenvalue of $cI + \nu A/(\delta x)^2$. We extend these six real-valued functions of $\mu$ to the continuous interval $(c, c + 4\nu/(\delta x)^2)$. For $i \in \{s, f\}$, the functions $\mu \mapsto \phi_i^{-1}(\mu)$ and $\mu \mapsto \psi_i(\mu)$ are smooth, non-increasing on $(c, c + 4\nu/(\delta x)^2)$ with values in $(0, 1]$. Hence, any finite product of such functions and any finite sum is smooth and non-increasing on $(c, c + 4\nu/(\delta x)^2)$. Indeed,

$$P : \mu \mapsto (\phi_f^{-1}(\mu)\psi_f(\mu))^{p'}, \quad Q : \mu \mapsto (\phi_s^{-1}(\mu)\psi_s(\mu))^{q'}, \quad \Sigma : \mu \mapsto c^2\frac{\delta t^2}{pq}\phi_f^{-1}(\mu)\phi_s^{-1}(\mu)\widetilde{\Sigma}_{s,q'}(\mu)\widetilde{\Sigma}_{f,p'}(\mu),$$

are positive non-increasing functions on $(c, c + 4\nu/(\delta x)^2)$. Note that the discriminant of the polynomial (33) is

$$\begin{aligned}\mathcal{D}(\mu) :=& \Big(P(\mu) + Q(\mu) + \Sigma(\mu)\Big)^2 - 4Q(\mu)P(\mu)\\ =& \Big(Q(\mu) - P(\mu) + \Sigma(\mu)\Big)^2 + 4P(\mu)\Sigma(\mu) > 0 \qquad (36)\\ =& \Big(P(\mu) - Q(\mu) + \Sigma(\mu)\Big)^2 + 4Q(\mu)\Sigma(\mu) > 0, \qquad (37)\end{aligned}$$

so that the eigenvalues of $\Psi_{\delta t,p,p',q,q'}$ are real and can be expressed using the functions

$$\tau^-(\mu) = \frac{P(\mu) + Q(\mu) + \Sigma(\mu) - \sqrt{\mathcal{D}(\mu)}}{2} \qquad \text{and} \qquad \tau^+(\mu) = \frac{P(\mu) + Q(\mu) + \Sigma(\mu) + \sqrt{\mathcal{D}(\mu)}}{2},$$

for $\mu \in (c, c + 4\nu/(\delta x)^2)$. Note that, with the stability condition (28), we have $0 < \tau^-(\mu) < \tau^+(\mu)$. Moreover, we have a monotonicity property for the function $\mu \mapsto \tau^+(\mu)$ on the interval $(c, c + 4\nu/(\delta x)^2)$ (see Lemma 2). Hence the biggest eigenvalue of $\Psi_{\delta t,p,p',q,q'}$ is $\tau^+(\mu_1)$ with $\mu_1 := c + \nu\lambda_1/(\delta x)^2$ (see (22)).

We compute an asymptotic expansion of that biggest eigenvalue when $\delta t \to 0^+$ to control the exponential decay of the $L^2$ norm of the numerical solution provided by $\Psi_{\delta t,p,p',q,q'}$. Let $J \in \mathbb{N}^\star$ be fixed. We number the eigenvalues of $cI + \nu A\delta x^2$ as follows:

$$\forall i \in \{1, \dots, J\}, \qquad \mu_i = c + \nu\frac{\lambda_i}{\delta x^2}. \tag{38}$$

Since $\phi_f^{-1}(\mu_1)\psi_f(\mu_1) = (1 - \theta_f\delta t\mu_1)/(1 + (1-\theta_f)\delta t\mu_1)$, we may write

$$\forall k \in \{0, \dots, p'\}, \qquad (\phi_f^{-1}(\mu_1)\psi_f(\mu_1))^k = 1 - k\mu_1\delta t + \mathcal{O}(\delta t^2),$$

We infer that

$$\sum_{k=0}^{p'-1} (\phi_f^{-1}(\mu_1)\psi_f(\mu_1))^k = p' - \mu_1\frac{p'(p'-1)}{2}\delta t + \mathcal{O}(\delta t^2).$$

We obtain Taylor expansions for $P(\mu_1)$, $Q(\mu_1)$, $\Sigma(\mu_1)$ and then $\mathcal{D}(\mu_1)$ similarly. Eventually, for the Lie-splitting SF method ($q = q' = p = 1$ and $p' = N$), we obtain the following Taylor expansion for $\tau^+(\mu_1)$ when $\delta t$ tends to 0:

$$\tau^+(\mu_1) = 1 - \gamma_0\delta t + \mathcal{O}(\delta t^2)$$

with

$$\gamma_0 := \frac{(N+1)\mu_1 - \sqrt{(N-1)^2\mu_1^2 + 4Nc^2}}{2}.$$

Therefore,

$$\frac{1}{\delta t}\ln(\tau^+(\mu_1)) = -\gamma_0 + \mathcal{O}(\delta t). \tag{39}$$

Note that, since $0 < c < \mu_1$, we have $0 < 4Nc^2 < 4N\mu_1^2$. Hence

$$(N+1)^2\mu_1^2 - (N-1)^2\mu_1^2 = 4N\mu_1^2 > 4Nc^2,$$

and $\gamma_0 > 0$. Since $\tau^+(\mu_1)$ is the biggest eigenvalue of $\Psi_{Lie,\delta t}$, this proves the result. Note also that

$$\gamma_0 = \frac{(N+1)\mu_1 - \sqrt{(N+1)^2\mu_1^2 - 4N(\mu_1^2 - c^2)}}{2}. \tag{40}$$

Using the mean value theorem, for some $c_\theta \in (0, 4N(\mu_1^2 - c^2))$, we conclude that

$$\gamma_0 = \frac{1}{2}\frac{1}{2}\frac{4N(\mu_1^2 - c^2)}{\sqrt{(N+1)^2\mu_1^2 - c_\theta}} > N\frac{\mu_1^2 - c^2}{(N+1)\mu_1} = \frac{N}{N+1}\underbrace{\frac{(\mu_1 + c)}{\mu_1}}_{\geq 1}\underbrace{(\mu_1 - c)}_{=\nu\lambda_1/\delta x^2} \geq \frac{N}{N+1}\nu\frac{\lambda_1}{(\delta x)^2}.$$

Putting together (39) and (40) allows for the expected choice of $\gamma$.

**Lemma 2** *The map $\mu \mapsto \tau^+(\mu)$ is non-increasing in $(c, c + 4\nu/(\delta x)^2)$*[9].

*Proof* We use the notations of Theorem 5. Note that, thanks to (36), $\sqrt{\mathcal{D}(\mu)} > Q(\mu) - P(\mu)$ if $Q(\mu) > P(\mu)$. Similarly, (37) leads to $\sqrt{\mathcal{D}(\mu)} > P(\mu) - Q(\mu)$ if $P(\mu) > Q(\mu)$ since $P$, $Q$, $\Sigma$ are positive functions. So $\sqrt{\mathcal{D}} > |P - Q|$. Differentiating the function $\mu \mapsto \tau^+(\mu)$ with respect to $\mu$ yields

$$\begin{aligned}
2\sqrt{\mathcal{D}}\frac{\mathrm{d}}{\mathrm{d}\mu}\tau^+ &= (\underbrace{P' + Q' + \Sigma'}_{<0})\sqrt{\mathcal{D}} + (P + Q + \underbrace{\Sigma}_{>0})(\underbrace{P' + Q' + \Sigma'}_{<0}) - 2(PQ)' \\
&< (P' + Q' + \Sigma')|Q - P| + (P + Q)(P' + Q' + \Sigma') - 2P'Q - 2PQ' \\
&< P'(|P - Q| + P - Q) + Q'(|Q - P| + Q - P) \\
&\leq 0.
\end{aligned}$$

This implies that the derivative of $\mu \mapsto \tau^+(\mu)$ is non-positive on $(c, c + 4\nu/(\delta x)^2)$ and proves the lemma.

### 4.2 The inhomogeneous Dirichlet problem

*The continuous problem* In this section we consider System (20) equipped with inhomogeneous Dirichlet boundary conditions, namely

$$u(t,0) = u_l, \qquad u(t,L) = u_r, \qquad v(t,0) = v_l, \qquad v(t,L) = v_r, \tag{41}$$

where $u_l, v_l, u_r$ and $v_r$ are four given real numbers. As in the homogeneous case above (see Section 4.1), there is a unique stationary solution to the boundary value problem:

**Proposition 11** *The PDE system* (20) *with non homogeneous Dirichlet boundary conditions has a unique stationary solution given by*

$$\begin{cases} u_{\mathrm{ex}}^\infty : \ x \mapsto \frac{u_l+v_l}{2} + \frac{(u_r+v_r-u_l-v_l)x}{2L} + \frac{(u_l-v_l)[\cosh(x/\alpha) - \cosh(L/\alpha)\sinh(x/\alpha)/\sinh(L/\alpha)]}{2} + \frac{(u_r-v_r)\sinh(x/\alpha)/\sinh(L/\alpha)}{2} \\ v_{\mathrm{ex}}^\infty : \ x \mapsto \frac{u_l+v_l}{2} + \frac{(u_r+v_r-u_l-v_l)x}{2L} - \frac{(u_l-v_l)[\cosh(x/\alpha) - \cosh(L/\alpha)\sinh(x/\alpha)/\sinh(L/\alpha)]}{2} - \frac{(u_r-v_r)\sinh(x/\alpha)/\sinh(L/\alpha)}{2} \end{cases} \tag{42}$$

*where* $\alpha = \sqrt{\nu/(2c)}$.

Therefore, using the linearity of the problems, for all $(u^0, v^0) \in L^2(0,L)^2$, the inhomogeneous reaction-diffusion system (20)-(41) has a unique solution in $C^0([0,+\infty), L^2(0,L)^2) \cap C^\infty((0,+\infty) \times [0,L], \mathbb{R}^2)$ satisfying $(u,v)(0) = (u^0, v^0)$, which is obtained from that of the homogeneous Dirichlet problem (with a modified initial datum) by adding the constant-in-time function (42) to it (see Theorem 3). Moreover, for all initial datum $(u^0, v^0)$, the solution of the inhomogeneous System (20) converges exponentially fast to the stationary solution (42).

The goal of the next paragraphs is to illustrate how well this convergence towards (a discretized version of) the stationary solution is achieved by numerical methods using subcycling techniques.

[9] Note that $\mathcal{D}$ is not a non-increasing function of $\mu$ in general.

*Space and time discretizations* Using the same space discretization as above (see Section 4.1), we consider two $\theta$-schemes for the time discretization in the spirit of what we did for the homogeneous problem (see (26)-(27)), with parameters $\theta_f$ and $\theta_s$. Taking into account the inhomogeneous Dirichlet boundary conditions yields a sequence $((U^n, V^n)^{\mathsf{t}})_{n\in\mathbb{N}}$ defined by an arithmetic-geometric recursion: given $W^0 = (U^0, V^0)^{\mathsf{t}} \in \mathbb{R}^{2J}$, we have for all $n \geq 0$,

$$W^{n+1} = \mathcal{M}W^n + \mathcal{M}_u \begin{pmatrix} U_{l,r} \\ 0_J \end{pmatrix} + \mathcal{M}_v \begin{pmatrix} 0_J \\ V_{l,r} \end{pmatrix} =: \mathcal{M}W^n + \Upsilon \tag{43}$$

where $\mathcal{M}$ is defined as a product of matrices of the form (29), $U_{l,r} = (u_l, 0, \ldots, 0, u_r)^{\mathsf{t}}$, $V_{l,r} = (v_l, 0, \ldots, 0, v_r)^{\mathsf{t}}$ and $\mathcal{M}_u$ and $\mathcal{M}_v$ are $2J$-by-$2J$ matrices, depending on $\delta t$, $\delta x$ and the choice of the splitting method between the two $\theta$-schemes.

Let us list the numerical experiments we conducted:

- Scheme #1 (Lie - SF - slow time - subcycled): $M_s := M_s(\delta t)$ and $M_f := M_f(\delta t/N)$

$$\mathcal{M} = M_s M_f^N, \;\; \mathcal{M}_u = \nu\frac{\delta t}{\delta x^2} M_s \sum_{k=0}^{N-1} M_f^k \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \;\text{ and }\; \mathcal{M}_v = \nu\frac{\delta t}{\delta x^2} \begin{pmatrix} 0 & 0 \\ 0 & C_s^{-1} \end{pmatrix} \tag{44}$$

- Scheme #2 (Lie - SF - fast time - no subcycling): $M_s := M_s(\delta t/N)$ and $M_f := M_f(\delta t/N)$

$$\mathcal{M} = M_s M_f, \;\; \mathcal{M}_u = \nu\frac{\delta t}{\delta x^2} M_s \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \;\text{ and }\; \mathcal{M}_v = \frac{\nu}{N}\frac{\delta t}{\delta x^2} \begin{pmatrix} 0 & 0 \\ 0 & C_s^{-1} \end{pmatrix}$$

- Scheme #3 (Strang - SFS - slow time - subcycled): $M_s := M_s(\delta t/2)$ and $M_f := M_f(\delta t/N)$

$$\mathcal{M} = M_s M_f^N M_s, \;\; \mathcal{M}_u = \nu\frac{\delta t}{\delta x^2} M_s \sum_{k=0}^{N-1} M_f^k \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \;\text{ and }\; \mathcal{M}_v = \nu\frac{\delta t}{2\delta x^2}(I_{2J} + M_s M_f^N) \begin{pmatrix} 0 & 0 \\ 0 & C_s^{-1} \end{pmatrix}$$

- Scheme #4 (Strang - SFS - fast time - no subcycling): $M_s := M_s(\delta t/(2N))$ and $M_f := M_f(\delta t/N)$

$$\mathcal{M} = M_s M_f M_s, \;\; \mathcal{M}_u = \nu\frac{\delta t}{\delta x^2} M_s \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \;\text{ and }\; \mathcal{M}_v = \nu\frac{\delta t}{2N\delta x^2}(I_{2J} + M_s M_f) \begin{pmatrix} 0 & 0 \\ 0 & C_s^{-1} \end{pmatrix}$$

- Scheme #5 (Csomós - slow time - subcycled): $M_s := M_s(\delta t)$ and $M_f := M_f(\delta t/N)$

$$\mathcal{M} = \frac{1}{2}(M_s M_f^N + M_f^N M_s),$$

$$\mathcal{M}_u = \nu\frac{\delta t}{\delta x^2}\frac{I_{2J} + M_s}{2} \sum_{k=0}^{N-1} M_f^k \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \;\text{ and }\; \mathcal{M}_v = \nu\frac{\delta t}{\delta x^2}(I_{2J} + M_f^N) \begin{pmatrix} 0 & 0 \\ 0 & C_s^{-1} \end{pmatrix}$$

- Scheme #6 (Csomós - fast time - no subcycling): $M_s := M_s(\delta t/N)$ and $M_f := M_f(\delta t/N)$

$$\mathcal{M} = \frac{1}{2}(M_s M_f + M_f M_s), \;\; \mathcal{M}_u = \nu\frac{\delta t}{\delta x^2}\frac{I_{2J} + M_s}{2} \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \;\text{ and }\; \mathcal{M}_v = \frac{\nu}{N}\frac{\delta t}{\delta x^2}\frac{I_{2J} + M_f}{2} \begin{pmatrix} 0 & 0 \\ 0 & C_s^{-1} \end{pmatrix}$$

*Equilibrium states of the splitting schemes* We prove the existence of a unique equilibrium state for the splitting Scheme #1 above, comment on the rate of convergence of the scheme towards its equilibrium state and also analyze how close the equilibrium state of the scheme is to a projection on the numerical space grid of the equilibrium state (42) of the continuous reaction-diffusion system (20) with inhomogeneous Dirichlet conditions (41) in an $L^2$ sense. Following (28), we denote by $\mathrm{CFL}(J)$ the positive real number

$$\mathrm{CFL}(J) = \frac{1}{c + 4\nu/\delta x^2} = \frac{1}{c + 4\nu(J+1)^2/L^2}.$$

To compute asymptotic numerical solutions of a given method of type (43), we need to solve the $2J$-by-$2J$ linear system

$$(I_{2J} - \mathcal{M})W = \Upsilon. \tag{45}$$

**Proposition 12** *Let $\delta t, \delta x > 0$ satisfying (28) be fixed. For a subcycled SF Lie-splitting method of the form (44), there exists a unique numerical asymptotic state defined as the unique solution $W^{\infty}_{\mathrm{num}}$ of the linear system (45).*

*Proof* Since $\delta t, \delta x$ satisfy (28), we know from Theorem 5 that the spectral radius of the matrix $\mathcal{M}$ of $\Psi_{Lie,\delta t}$ in the canonical basis of $\mathbb{R}^{2J}$ is less than 1. Hence, the matrix $I_{2J} - \mathcal{M}$ is invertible and the numerical asymptotic state is well-defined and unique.

Using the linearity of the problems, we infer that the numerical rate of convergence towards this asymptotic state is then given by Theorem 5.

Let us state and prove the central result of this section, *i.e.* the convergent asymptotic behavior of the subcycled SF Lie scheme (Scheme #1):

**Theorem 6** *Provided that $\delta t \in (0, \mathrm{CFL}(J))$, the asymptotic state of Scheme #1 is a* uniform-in-$\delta t$ *second order approximation of the exact asymptotic state given in Proposition 42:*

$$\begin{pmatrix} \Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big) \\ \Pi_{\delta x}\big(v_{\mathrm{ex}}^{\infty}\big) \end{pmatrix} - W_{\mathrm{num}}^{\infty}(\delta t) = \mathcal{O}(\delta x^2),$$

*where for $w \in C^0([0,L])$, $\Pi_{\delta x}(w) = (w(x_1), \ldots, w(x_J))^{\mathsf{t}}$.*

*Proof* To analyze the asymptotic convergence of Scheme #1, we put the projections $\Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big)$ and $\Pi_{\delta x}\big(v_{\mathrm{ex}}^{\infty}\big)$ of the exact solutions $u_{\mathrm{ex}}^{\infty}$ and $v_{\mathrm{ex}}^{\infty}$ defined in (42) in the numerical scheme. Using the identity

$$\frac{1}{\delta x^2} A \Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big) = -\Pi_{\delta x}\big(\Delta u_{\mathrm{ex}}^{\infty}\big) + U_{l,r} + \mathcal{O}(\delta x^2),$$

and the fact that $(u_{\mathrm{ex}}^{\infty}, v_{\mathrm{ex}}^{\infty})$ is an equilibrium state of problem (20) with the inhomogeneous Dirichlet boundary conditions (41), we first compute

$$M_f \begin{pmatrix} \Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big) \\ \Pi_{\delta x}\big(v_{\mathrm{ex}}^{\infty}\big) \end{pmatrix} = \begin{pmatrix} \Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big) \\ \Pi_{\delta x}\big(v_{\mathrm{ex}}^{\infty}\big) \end{pmatrix} - \nu \frac{\delta t}{\delta x^2} \begin{pmatrix} C_f^{-1} U_{l,r} \\ 0 \end{pmatrix} + \mathcal{O}(\delta t (\delta x)^2),$$

where the constant in the $\mathcal{O}$ is independent of $\delta t$ and $\delta x$ provided that the CFL condition is fulfilled. Iterating this computation, we obtain

$$M_f^N \begin{pmatrix} \Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big) \\ \Pi_{\delta x}\big(v_{\mathrm{ex}}^{\infty}\big) \end{pmatrix} = \begin{pmatrix} \Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big) \\ \Pi_{\delta x}\big(v_{\mathrm{ex}}^{\infty}\big) \end{pmatrix} - \nu \frac{\delta t}{\delta x^2} \sum_{k=0}^{N-1} M_f^k \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_{l,r} \\ 0 \end{pmatrix} + \mathcal{O}(\delta t (\delta x)^2), \tag{46}$$

where, once again, the constant in the $\mathcal{O}$ is independent of $\delta t$ and $\delta x$ provided that the CFL condition (28) is fulfilled. This is due to the fact that we have

$$M_f \mathcal{O}(\delta t(\delta x^2)) = \mathcal{O}(\delta t(\delta x^2)),$$

provided that $\delta t \in (0, \mathrm{CFL}(J))$ thanks to Lemma 3 (see Appendix), that gives uniform estimates of $|||M_{s,f}|||_2$. Multiplying (46) by $M_s$ and using again that $(u_{\mathrm{ex}}^{\infty}, v_{\mathrm{ex}}^{\infty})$ is an equilibrium state of problem (20) with the inhomogeneous Dirichlet boundary conditions (41), we finally get

$$(I_{2J} - M_s M_f^N) \begin{pmatrix} \Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big) \\ \Pi_{\delta x}\big(v_{\mathrm{ex}}^{\infty}\big) \end{pmatrix} = \nu \frac{\delta t}{\delta x^2} M_s \sum_{k=0}^{N-1} M_f^k \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_{l,r} \\ 0 \end{pmatrix} + \nu \frac{\delta t}{\delta x^2} \begin{pmatrix} 0 & 0 \\ 0 & C_s^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ V_{l,r} \end{pmatrix} + \mathcal{O}(\delta t (\delta x)^2).$$

Comparing this relation with that defining the numerical equilibrium state (43) (with the right-hand side defined in (44)), we infer that

$$(I_{2J} - M_s M_f^N) \left( \begin{pmatrix} \Pi_{\delta x}\big(u_{\mathrm{ex}}^{\infty}\big) \\ \Pi_{\delta x}\big(v_{\mathrm{ex}}^{\infty}\big) \end{pmatrix} - W_{\mathrm{num}}^{\infty} \right) = \delta t \mathcal{O}(\delta x^2), \tag{47}$$

where the constant in the $\mathcal{O}$ is independent of $\delta t$ and $\delta x$ provided that the CFL condition (28) is fulfilled. Finally, we can use the result of Proposition 13 (see Appendix) which states that there exists a constant $C > 0$ such that for all $\delta t$ and $\delta x$ satisfying the CFL condition, we have $|||(I - M_s M_f^N)^{-1}|||_2 \leq \frac{C}{\delta t}$. This estimate together with that written in (47) proves the result.

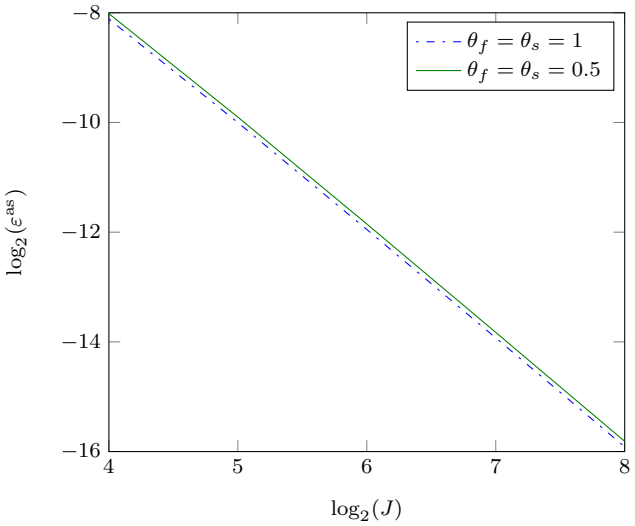


**Fig. 4** $L^\infty$-error of the asymptotic numerical and exact states for explicit/explicit and Crank-Nicolson/Crank-Nicolson schemes. The numerical order is 1.95.

*Numerical tests* The numerical tests we conducted for several values of $\theta_f$, $\theta_s$ and $N$ showed that the matrix $I_{2J} - \mathcal{M}$ is also invertible for Schemes #3 and #4. We show here the graph obtained with Scheme #1 for the following sets of parameters, $N = 10$ being fixed, $J = 20, 40, 80, 160$, $\delta x = L/(J+1)$:

- $(u_l, u_r, v_l, v_r) = (1, 2, -1, 4)$, $\delta t = \delta x^2/\nu_1/2$, $(\theta_f, \theta_s) = (1, 1)$ [explicit,explicit]
- $(u_l, u_r, v_l, v_r) = (2, 4, -1, 4)$, $J = 20, 40, 80, 160$, $\delta x = L/(J+1)$, $\delta t = 0.01$, $(\theta_f, \theta_s) = (1/2, 1/2)$ [Crank-Nicolson,Crank-Nicolson]

From Figure 4, we conclude that the asymptotic state depends only on the spatial discretization through $\delta x$ and does not depend on the time discretization $\delta t$ or the values of $\theta$. Moreover, the numerical order is close to 2 in $\delta x$.

## 5 Conclusion and perspectives

Speeding up computations through a subcycling procedure is widely used, but the asymptotic behavior of the numerical solution in large time is a concern. Indeed, there are two limits involved, as $\delta t$ (and $\delta x$ in the PDE case) tend to 0 and as the final time $T$ tends to $+\infty$. We proved for an illustrative case of ODE systems that the asymptotic error is at least of the same order of convergence as the local-in-time error, and can even be better since there exists a Strang combination of (local) first order schemes that leads to a second asymptotic order ! The analysis of the convergence rate of the subcycled scheme has been performed for ODE and PDE toy-models, showing that the Strang splitting associated with Crank-Nicolson schemes was the only way to get a second order approximation of the exact rate. Finally, in the case of a coupled reaction-diffusion system with inhomogeneous Dirichlet boundary conditions, we were able to prove that the asymptotic numerical solution obtained through a subcycled scheme is a uniform-in-$\delta t$ second order approximation in $\delta x$ of the exact asymptotic state. The aim is now to tackle the much more difficult case of a fully coupled hyperbolic-parabolic system, in particular as the limit of a system consisting of a kinetic equation in the diffusive regime and a transport equation.

## A FS to SF computations

Let us define the matrix

$$\Pi := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and let us denote by $G[\alpha, \beta]$ a matrix of the form (5). Let $A$ be a 2-by-2 matrix. Then $\Pi\, A$ exchanges the lines of $A$ and $A\, \Pi$ exchanges the columns. Thus, if $\lambda \in \mathbb{R}$,

$$\Pi\, M_s(\lambda)\, \Pi = M_f(\lambda),$$

and, if $\alpha, \beta \in (0, 1)$,

$$\Pi\, G[\alpha, \beta]\, \Pi = G[\beta, \alpha].$$

Since $\Pi^2 = I$, it means that $M_s(\lambda)$ and $M_f(\lambda)$ are similar, thus share the same spectrum. In Section 2, we computed the A-orders and rates of convergence of SF (fast, then slow) and FSF (fast, then slow, then fast) type schemes. We show here that the results we obtained can easily be applied to FS and SFS schemes.

*Lie-splitting schemes* Consider $\lambda_s, \lambda_f \in (0,1)$. According to Lemma 1 and Remark 1, we define $\alpha(\lambda_s, \lambda_f)$ and $\beta(\lambda_s, \lambda_f)$ as

$$M_s(\lambda_s)M_f(\lambda_f) = G[\alpha(\lambda_s, \lambda_f), \beta(\lambda_s, \lambda_f)].$$

Since

$$M_f(\lambda_f)M_s(\lambda_s) = \Pi\, M_s(\lambda_f)M_f(\lambda_s)\, \Pi,$$

we infer that

$$M_f(\lambda_f)M_s(\lambda_s) = \Pi\, G[\beta(\lambda_f, \lambda_s), \alpha(\lambda_f, \lambda_s)]\, \Pi$$

Consequently, we can deduce the convergence rate and the A-order of the FS methods at once from the results we obtained for the SF methods.

*Strang-splitting methods* In the same way, knowing $M_f(\lambda_f)M_s(\lambda_s)M_f(\lambda_f)$, one can deduce the convergence rate and the A-order of $M_f(\lambda_f)M_s(\lambda_s)M_f(\lambda_f)$ by noting that

$$M_s(\lambda_s)M_f(\lambda_f)M_s(\lambda_s) = \Pi\, M_f(\lambda_s)M_s(\lambda_f)M_f(\lambda_s)\, \Pi.$$

## B Helpful estimates for the proof of Theorem 6

The following lemma is helpful for the proof of Theorem 6.

**Lemma 3** *There exists a positive constant $C > 0$ such that, for all $J \in \mathbb{N}^\star$ and all $\delta t \in (0, \mathrm{CFL}(J))$, we have*

$$|||M_s|||_2 \leq C \qquad \text{and} \qquad |||M_f|||_2 \leq C.$$

*Remark 15* Note that the constant $C$ above is in fact greater than 1, even if the matrices have their spectrum in the interval $[0,1]$. This is due to the lack of symmetry in those matrices.

*Proof* Since the situation for $M_s$ and $M_f$ is very similar, we prove the inequality for $M_f$ only, and we start with the decomposition

$$M_f = \begin{pmatrix} C_f^{-1} & 0 \\ 0 & I_J \end{pmatrix} \times \begin{pmatrix} B_f & c\delta t I_J \\ 0 & I_J \end{pmatrix}.$$

Recall that for any square matrix $R$ with real coefficients, $|||R|||_2^2 = \rho(R^{\mathrm{t}}R)$, where $\rho$ denotes the spectral radius. The CFL condition (28) ensures that the spectrum of $C_f^{-1}$ lies in $(0,1]$. Since the first matrix in the product above is symmetric, we infer that its norm is $\sqrt{\rho(I_J)} = 1$. Hence, using the algebra property for $|||\cdot|||_2$, it is sufficient to prove the result for the second matrix in the product above, which is *not* symmetric. We are left with the computation of the eigenvalues of the symmetric non-negative matrix

$$\begin{pmatrix} B_f^2 & c\delta t B_f \\ c\delta t B_f & (1 + c^2\delta t^2) I_J \end{pmatrix},$$

the eigenvalues of which are the $2J$ roots of the $J$ polynomials

$$X^2 - (\mu_p^2 + (1 + c^2\delta t^2))X + \mu_p^2, \qquad 1 \leq p \leq J,$$

where $(\mu_p)_{1\leq p\leq J}$ denotes the list of the eigenvalues of $B_f$. The CFL condition (28) ensures that for all $p \in \{1, \dots, J\}$, $\mu_p \in [0,1]$. Hence, the greatest eigenvalue of the corresponding polynomial above is less than $2(1 + 1 + c^2\delta t^2)$. Moreover, the CFL condition also provides us with an estimate on $\delta t$ which yields the result with $C = \sqrt{2(2 + c^2/(c + 16\nu/L^2)^2)}$.

One can control the inverse of the matrix of System (45) by the following proposition to prove Theorem 6.

**Proposition 13** *There exists a positive constant $C > 0$ such that for all $J \in \mathbb{N}^\star$ and all $\delta t \in (0, \mathrm{CFL}(J))$,*

$$|||(I - M_s M_f^N)^{-1}|||_2 \leq \frac{C}{\delta t}. \tag{48}$$

*Proof* Let us fix $J \in \mathbb{N}^\star$ and $\delta t \in (0, \mathrm{CFL}(J))$. Using the conjugation with the orthogonal matrix $\mathcal{Z}$ (see (32)), we have that the $|||\cdot|||_2$-norm of $I_{2J} - M_s M_f^N$ is equal to that of the same matrix where $A$ is replaced with $D$ (see (23)). The latter matrix has a very particular structure: the four $J$-by-$J$ matrices defining it are diagonal. Let us denote by $(a_i)_{1\leq i\leq J}$, $(b_i)_{1\leq i\leq J}$, $(c_i)_{1\leq i\leq J}$, and $(d_i)_{1\leq i\leq J}$ these entries such that

$$\zeta := \mathcal{Z}^{-1}(I - M_s M_f^N)\mathcal{Z} = \begin{pmatrix} a_1 & 0 & 0 & b_1 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & a_J & 0 & 0 & b_J \\ c_1 & 0 & 0 & d_1 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & c_J & 0 & 0 & d_J \end{pmatrix}.$$

The eigenvalues of $\zeta$ lie in $(0,1)$ (see Theorem 5). Hence, $\zeta$ is invertible and its inverse is given by

$$\zeta^{-1} = \mathcal{Z}^{-1}(I - M_s M_f^N)^{-1}\mathcal{Z} = \begin{pmatrix} \alpha_1 & 0 & 0 & \beta_1 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \alpha_J & 0 & 0 & \beta_J \\ \gamma_1 & 0 & 0 & \delta_1 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \gamma_J & 0 & 0 & \delta_J \end{pmatrix},$$

where for all $i \in \{1, \dots, J\}$,

$$\begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix}^{-1} = \begin{pmatrix} \alpha_i & \beta_i \\ \gamma_i & \delta_i \end{pmatrix} =: m_i.$$

One can check easily that

$$|||\zeta^{-1}|||_2 = \max_{1 \le i \le J} |||m_i|||_2.$$

Moreover, we have

$$|||m_i|||_2^2 = \frac{a_i^2 + b_i^2 + c_i^2 + d_i^2 + \sqrt{(a_i^2 + b_i^2 + c_i^2 + d_i^2)^2 - 4(a_i d_i - b_i c_i)^2}}{2(a_i d_i - b_i c_i)^2} \le \frac{a_i^2 + b_i^2 + c_i^2 + d_i^2}{(a_i d_i - b_i c_i)^2}.$$

We split the upper bound above as follows

$$|||m_i|||_2^2 \le \frac{b_i^2 + c_i^2}{(a_i d_i - b_i c_i)^2} + \frac{a_i^2 + d_i^2}{(a_i d_i - b_i c_i)^2}, \tag{49}$$

and we prove an estimate of the form $\mathcal{O}(1/\delta t^2)$ for the two terms in the sum above. In view of (30), we have

$$a_i = 1 - P(\mu_i),\ b_i = -c\delta t(\phi_f^{-1}\tilde{\Sigma}_{f,N})(\mu_i),$$

and

$$c_i = -c\delta t(\phi_s^{-1}P)(\mu_i) \text{ and } d_i = 1 - Q(\mu_i) - c^2\delta t^2(\phi_s^{-1}\phi_f^{-1}\tilde{\Sigma}_{f,N})(\mu_i),$$

where the $\mu_i$ are defined by (38) as the ordered eigenvalues of $cI + \nu A/\delta x^2$. For the first term in the upper bound (49), let us show that the numerator is $\mathcal{O}(\delta t^2)$ while the denominator is bounded from below by a positive constant times $\delta t^4$. On the one hand, we have

$$|b_i|^2 \le c^2 N^2 \delta t^2 \qquad \text{and} \qquad |c_i|^2 \le c^2 \delta t^2. \tag{50}$$

On the other hand, for all $i \in \{1, \dots, J\}$, we have

$$\begin{aligned} a_i d_i - b_i c_i &= (1 - P(\mu_i))(1 - Q(\mu_i)) - c^2\delta t^2(\phi_s^{-1}\phi_f^{-1}\tilde{\Sigma}_{f,N})(\mu_i) \\ &= \left(1 - (\psi_f\phi_f^{-1})^N(\mu_i)\right)(1 - Q(\mu_i)) - c^2\delta t^2\left(\phi_s^{-1}\phi_f^{-1}\frac{1 - (\psi_f\phi_f^{-1})^N}{1 - \psi_f\phi_f^{-1}}\right)(\mu_i) \\ &= \left(\left(\frac{1 - (\psi_f\phi_f^{-1})^N}{\phi_s}\right)\left(\phi_s - \psi_s - \frac{c^2\delta t^2}{\phi_f - \psi_f}\right)\right)(\mu_i). \end{aligned}$$

The CFL condition (28) ensures that $\delta t\mu_i$, $\psi_s(\mu_i), \phi_s^{-1}(\mu_i), \psi_f(\mu_i)$, $\phi_f^{-1}(\mu_i)$ and $P(\mu_i)$ belong to $(0,1]$. In view of the definitions (34), we have

$$(\phi_s - \psi_s)(\mu_i) = \delta t\mu_i = (\phi_f - \psi_f)(\mu_i),$$

so that

$$a_i d_i - b_i c_i = \delta t\frac{(1 - (\psi_f\phi_f^{-1})^N(\mu_i))}{\phi_s(\mu_i)}\frac{\mu_i^2 - c^2}{\mu_i}. \tag{51}$$

The CFL condition (28) implies that $1/\phi_s(\mu_i) \ge 1/2$ and

$$0 < (\psi_f\phi_f^{-1})^N(\mu_i) \le (\psi_f\phi_f^{-1})(\mu_i) = \frac{1 - \theta_f\delta t\mu_i}{1 + (1 - \theta_f)\delta t\mu_i}.$$

Therefore, we have

$$1 - (\psi_f\phi_f^{-1})^N(\mu_i) \ge 1 - (\psi_f\phi_f^{-1})(\mu_i) = \frac{\delta t\mu_i}{1 + (1 - \theta_f)\delta t\mu_i} \ge \frac{\delta t\mu_i}{2}. \tag{52}$$

This allows to bound from below

$$a_i d_i - b_i c_i \ge \frac{\delta t^2}{4}(\underbrace{\mu_i + c}_{\ge c})(\underbrace{\mu_i - c}_{=\nu\lambda_i/\delta x^2}) \ge c\nu\frac{\delta t^2}{4}\frac{\lambda_1}{\delta x^2}.$$

Recall that for all $x \in (0, \pi/2)$, $\sin(x) \geq 2x/\pi$, so that

$$\frac{\lambda_1}{\delta x^2} = \frac{4}{\delta x^2} \sin^2\Big(\frac{\pi}{2}\frac{1}{(J+1)}\Big) \geq 4\frac{(J+1)^2}{L^2}\frac{4}{\pi^2}\frac{\pi^2}{4}\frac{1}{(J+1)^2} \geq \frac{4}{L^2}. \tag{53}$$

This proves

$$a_i d_i - b_i c_i \geq \frac{c\nu}{L^2}\delta t^2. \tag{54}$$

Using (50) and (54), there exists a positive constant $C$ such that

$$\forall J \in \mathbb{N}^\star, \quad \forall \delta t \in (0, \mathrm{CFL}(J)), \qquad \frac{b_i^2 + c_i^2}{(a_i d_i - b_i c_i)^2} \leq \frac{C}{\delta t^2}. \tag{55}$$

Let us now bound the second term in the right hand side of (49). Let us fix $J \in \mathbb{N}^\star$ and $i \in (0, \mathrm{CFL}(J))$ again. From (51), we have

$$\frac{1}{(a_i d_i - c_i b_i)^2} = \frac{1}{\delta t^2}\frac{\phi_s^2(\mu_i)}{(1-(\psi_f\phi_f^{-1})^N(\mu_i))^2}\Big(\frac{\mu_i}{\mu_i^2 - c^2}\Big)^2.$$

A similar direct calculation yields

$$\begin{aligned}
a_i^2 + d_i^2 &= \Big(1 - (\psi_f\phi_f^{-1})^N(\mu_i)\Big)^2 + \Big(\frac{\phi_s(\mu_i) - \psi_s(\mu_i)}{\phi_s(\mu_i)} - c^2\delta t^2 \frac{1}{\phi_s\phi_f(\mu_i)}\frac{1-(\psi_f\phi_f^{-1})^N(\mu_i)}{1-\psi_f\phi_f^{-1}(\mu_i)}\Big)^2 \\
&= \Big(1 - (\psi_f\phi_f^{-1})^N(\mu_i)\Big)^2\Big[1 + \frac{1}{\phi_s^2(\mu_i)}\Big(\frac{\mu_i\delta t}{1-(\psi_f\phi_f^{-1})^N(\mu_i)} - c^2\delta t^2\frac{1}{\phi_f(\mu_i) - \psi_f(\mu_i)}\Big)^2\Big] \\
&= \Big(1 - (\psi_f\phi_f^{-1})^N(\mu_i)\Big)^2\Big[1 + \frac{1}{\phi_s^2(\mu_i)}\Big(\frac{\mu_i\delta t}{1-(\psi_f\phi_f^{-1})^N(\mu_i)} - \frac{c^2}{\mu_i}\delta t\Big)^2\Big].
\end{aligned}$$

We infer

$$\frac{a_i^2 + d_i^2}{(a_i d_i - c_i b_i)^2} = \frac{1}{\delta t^2}\phi_s^2(\mu_i)\Big(\frac{\mu_i}{\mu_i^2 - c^2}\Big)^2\Big[1 + \frac{1}{\phi_s^2(\mu_i)}\Big(\frac{\mu_i\delta t}{1-(\psi_f\phi_f^{-1})^N(\mu_i)} - \frac{c^2}{\mu_i}\delta t\Big)^2\Big]. \tag{56}$$

We can bound the terms in the product above as follows. The CFL condition (28) implies that $\phi_s^2(\mu_i) \leq 4$. Moreover, using (53), we have

$$\frac{\mu_i}{\mu_i^2 - c^2} = \underbrace{\frac{\mu_i}{(\mu_i + c)}}_{\leq 1}\frac{1}{(\mu_i - c)} \leq \frac{\delta x^2}{\nu\lambda_i} \leq \frac{\delta x^2}{\nu\lambda_1} \leq \frac{L^2}{4\nu}.$$

Recall that $1/\phi_s(\mu_i)^2 \leq 1$. From (52), we obtain $\mu_i\delta t/(1-(\psi_f\phi_f^{-1})^N(\mu_i)) \leq 2$. For the last term in the product, we have

$$\frac{c^2}{\mu_i}\delta t = \underbrace{c\delta t}_{\leq 1}\underbrace{\frac{c}{c + \nu\lambda_i/\delta x^2}}_{\leq 1} \leq 1.$$

Using these inequalities in (56), taking products and using Young's inequality, we infer that

$$\forall J \in \mathbb{N}^\star, \quad \forall \delta t \in (0, \mathrm{CFL}(J)), \qquad \frac{a_i^2 + d_i^2}{(a_i d_i - b_i c_i)^2} \leq \frac{11}{4}\frac{L^4}{\nu^2}\frac{1}{\delta t^2}. \tag{57}$$

The inequalities (55) and (57) together with (49) prove the result.

## References


1. D. Aregba-Driollet, M. Briani, and R. Natalini. Asymptotic high-order schemes for $2\times2$ dissipative hyperbolic systems. *SIAM Journal on Numerical Analysis*, 46(2):869–894, 2008.
2. M. O. Bristeau, R. Glowinski, B. Mantel, J. Periaux, and G. S. Singh. On the use of subcycling for solving the compressible Navier-Stokes equations by operator-splitting and finite element methods. *Communications in Applied Numerical Methods*, 4(3):309–317, 1988.
3. J. A. Carrillo, T. Goudon, and P. Lafitte. Simulation of fluid and particles flows: asymptotic preserving schemes for bubbling and flowing regimes. *Journal of Computational Physics*, 227(16):7929–7951, 2008.
4. J. A. Carrillo, T. Goudon, P. Lafitte, and F. Vecil. Numerical schemes of diffusion asymptotics and moment closures for kinetic equations. *Journal of Scientific Computing*, 36(1):113–149, 2008.
5. P. Csomós, I. Faragó, and Á. Havasi. Weighted sequential splittings and their analysis. *Computers & Mathematics with Applications*, 50(7):1017–1031, 2005.
6. W.J.T. Daniel. A study of the stability of subcycling algorithms in structural dynamics. *Computer Methods in Applied Mechanics and Engineering*, 156(14):1 – 13, 1998.

7. W.J.T. Daniel. A partial velocity approach to subcycling structural dynamics. *Computer Methods in Applied Mechanics and Engineering*, 192:375 – 394, 2003.
8. J. Diaz and M. J. Grote. Energy conserving explicit local time stepping for second-order wave equations. *SIAM Journal on Scientific Computing*, 31(3):1985–2014, 2009.
9. P. Godillon-Lafitte and T. Goudon. A coupled model for radiative transfer: Doppler effects, equilibrium, and nonequilibrium diffusion asymptotics. *Multiscale Modeling & Simulation. A SIAM Interdisciplinary Journal*, 4(4):1245–1279 (electronic), 2005.
10. M. J. Grote and T. Mitkova. Explicit local time-stepping methods for Maxwell's equations. *Journal of Computational and Applied Mathematics*, 234(12):3283–3302, 2010.
11. M. J. Grote and T. Mitkova. High-order explicit local time-stepping methods for damped wave equations. *Journal of Computational and Applied Mathematics*, 239:270–289, 2013.
12. E. Hairer and G. Wanner. *Solving ordinary differential equations II: Stiff and differential-algebraic problems*, volume 2. Springer, 2004.
13. S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing*, 21(2):441–454, 1999.
14. S. Jin. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. *Lecture Notes for Summer School on Methods and Models of Kinetic Theory (M&MKT), Porto Ercole (Grosseto, Italy)*, 2010.
15. M. Lemou and L. Mieussens. A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit. *SIAM Journal on Scientific Computing*, 31(1):334–368, 2008.
16. S. Piperno. Explicit/implicit fluid/structure staggered procedures with a structural predictor and fluid subcycling for 2d inviscid aeroelastic simulations. *International Journal for Numerical Methods in Fluids*, 25(10):1207–1226, 1997.
17. R. Temam. Multilevel methods for the simulation of turbulence. A simple model. *Journal of Computational Physics*, 127(2):309–315, 1996.