HAL open science

Asymptotic behavior of splitting schemes involving time-subcycling techniques

Guillaume Dujardin, Pauline Lafitte

To cite this version:

Guillaume Dujardin, Pauline Lafitte. Asymptotic behavior of splitting schemes involving time-subcycling techniques. 2012. hal-00751217v2

HAL Id: hal-00751217

https://hal.science/hal-00751217v2

Preprint submitted on 4 Oct 2013 (v2), last revised 6 Oct 2015 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Asymptotic behavior of splitting schemes involving time-subcycling techniques

Guillaume Dujardin[*] *Pauline Lafitte*[†]

Abstract

In order to integrate numerically a well-posed multiscale evolutionary problem such as a Cauchy problem for an ODE system or a PDE system, using time-subcycling techniques consists in splitting the vector field in a fast part and a slow part and taking advantage of this decomposition, for example by integrating the fast equation on a much smaller time step than the slow equation (instead of having to integrate the whole system with a very small time step to ensure stability for example). These techniques are designed to improve the computational efficiency and have been very widely used for designing schemes, that may have (at least) one component that has to be computed through an explicit scheme thus constrained by a limitation of the time step (CFL). In this paper, we 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. We develop our analysis on ODE and PDE toy-models and illustrate our results numerically on more complex systems.

1 Introduction

Time-subcycling is a way to speed up computations for a multiscale problem by splitting the underlying operator and treating the different steps of the resulting numerical scheme with adapted time-steps. Our aim is to determine how appropriately 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 may 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, 10, 4]. Related local time-stepping techniques have been developed extensively for multiscale problems arising in computational fluid and structural dynamics [17, 7, 8]. Simulating 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, and an ever larger number of steps 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 [9, 11, 12] and references therein).

The applications we have specifically in mind are related to the recent development of the "asymptotic-preserving" schemes in the sense of Jin [14, 15] 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 [10, 16, 5, 4]. However, if subcycling techniques have been used in several test-cases, up to our knowledge, the asymptotic error to the long-time solution has never been precisely analyzed. Note that the computation of the iterated numerical solutions of the fast equations required by subcycled schemes could be computed using for example multi-revolution composition methods (see for example [3, 6]), even if we will not use these techniques in this paper. We aim here at studying the convergence (error

[*]INRIA Lille Nord Europe, EPI SIMPAF & Laboratoire Paul Painlevé, Université Lille Nord de France, CNRS UMR 8525
[†]École Centrale Paris, Lab. MAS & INRIA Lille Nord Europe, EPI SIMPAF

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 splitting, 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 [13] and of the analysis led by Temam [18]. Both systems have exact and explicit solutions so one can do any computation and estimate involving the exact flows. The first one (analyzed in Section 2) is linear and reads[1]

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

where $c \geq 0$ and $N \in \mathbb{N}$, with N being large: it is the stiffness parameter in the problem. The second system (analyzed in Section 3) is nonlinear and reads[2]

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

For the numerical solutions of the linear system (1), we consider linear 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

$$M_f(\lambda_f) := \begin{pmatrix} \lambda_f & 1-\lambda_f \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad M_s(\lambda_s) := \begin{pmatrix} 1 & 0 \\ 1-\lambda_s & \lambda_s \end{pmatrix}. \tag{3}$$

The expressions of the parameters λ will depend on the choice of integrator (exact flow or θ-scheme) and the composition of the matrices will depend on the type of splitting one wants to use (Lie or Strang type). We introduce the concepts of asymptotic error and asymptotic order (see Definition 2.3), and prove properties about the asymptotic orders of the schemes (see Propositions 2.5 and 2.6) which are illustrated by several numerical experiments. We comment on the differences between schemes with and without subcycling. In Section 4, we perform the same kind of analysis in an infinite dimensional setting 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 a numerical point of view, we have to take into account the spatial discretization[3] and the long-time behavior of the schemes is to be analyzed also with respect to the spatial discretization parameter. 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 4.4) and we compare this rate to that of the exact solution (see Theorem 4.2). For inhomogeneous Dirichlet boundary conditions, we compare several linear 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-δt second order L^2-approximation of the exact asymptotic equilibrium state under a CFL-like condition (see Theorem 4.7).

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

[3] In fact, the stiffness parameter depends on the mesh.

2 Full analysis of the linear system

2.1 The exact solution

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

$$(-(N+1),\ P = -A/(N+1)) \text{ and } \left(0,\ Q = (1,1)^t\,(1,N)/(N+1)\right).$$

So the exact solution is, for all $t \in \mathbb{R}$,

$$W(t) := (u(t), v(t))^t = \left(\mathrm{e}^{-(N+1)ct}P + Q\right)(u^0, v^0)^t$$

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 $Q(u^0, v^0)^t$ when t tends to infinity. In the following, we fix $T > 0$ and we define

$$F_T = \mathrm{e}^{-(N+1)cT}P + Q, \tag{4}$$

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

2.2 General properties of linear splitting schemes

Let $G_{\delta t}$ be defined for $\delta t \in \mathcal{I}_N$ as the 2-by-2 matrix of a numerical flow which is a product of matrices of the form (3), where $\mathcal{I}_N$ is the intersection, that may depend on N, of the stability intervals of the related schemes (see examples in Section 2.3). In the following, for all $n \in \mathbb{N}$, we will denote by

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

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

Lemma 2.1: 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 $(1, 1)^t$ 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 α and β. 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,

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

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

Proof. Since all the matrices M_s and M_f have $(1, 1)^t$ 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 2.1, we can show that the exact and numerical propagators share an interesting property:

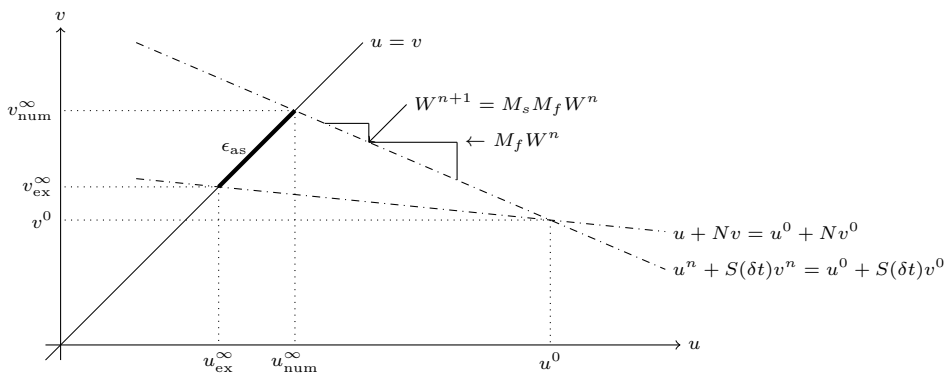

Fig. 1: Evolution of the exact and numerical solution in the phase space $\mathbb{R}_u \times \mathbb{R}_v$. We note $W^n = (u^n, v^n)^t$.

Proposition 2.2: For any fixed $\delta t > 0$, $F^n_{\delta t} = F_{n\delta t}$ projects the vector $(u^0, v^0)^t$ onto the line of equation $u = v$ when n tends to infinity and so does $G^n_{\delta t}$ 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 2.1, we get $\forall n \in \mathbb{N},\ G^n_{\delta t} = (\mu(\delta t))^n P(\delta t) + Q(\delta t)$, with $|\mu(\delta t)| < 1$ and the result follows. ■

Let us denote those limits (which depend on δt)

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

Let us 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 $\forall n \geq 0$, $u^n + S(\delta t)v^n = u^0 + S(\delta t)v^0$, the asymptotic error $(u^\infty_{\text{num}}, v^\infty_{\text{num}})^t - (u^\infty_{\text{ex}}, v^\infty_{\text{ex}})^t = (Q(\delta t) - Q)(u^0, v^0)^t$ can be measured as 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).

Therefore, we set the following

Definition 2.3: The relative asymptotic error is the scaled difference

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

and we say that the asymptotic order (A-order) is at least $p \in \mathbb{N}^\star$ if when δt tends to 0, we have

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

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

Our first result is the following

Theorem 2.4: 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 (3). If the local order of $G_{\delta t}$ is at least $p + 1$[4], 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

$$G_{\delta t} - F_{\delta t} = \begin{pmatrix} 1 - \alpha(\delta t) & \alpha(\delta t) \\ \beta(\delta t) & 1 - \beta(\delta t) \end{pmatrix} - \mathrm{e}^{-(N+1)c\delta t} P - Q = \mathcal{O}(\delta t^{p+1}).$$

This implies the following Taylor expansions for α and β:

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

[4] hence its global order is at least p

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 linear splitting schemes for the linear differential system (1), based on the composition of exact flows of the split vector fields and on θ-schemes discretizing the split equations. We focus on their asymptotic behavior. We know from Proposition 2.2 and Theorem 2.4 that for all initial datum $(u^0, v^0) \in \mathbb{R}^2$, 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 δt 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 δt ? Can we do better than the estimate on the relative A-order provided by Theorem 2.4 ?

2.3 Lie and Strang splitting schemes

Denoting by δt the numerical time step related to the "slow" equation, the time step associated to the "fast" equation is then $\delta t/N$. 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.

The (exact or numerical) integration of the fast (resp. slow) equation of (2) over a time step δt yields the flow

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

with $\lambda_s(\delta t) = \lambda_f(\delta t/N)$. In the following, we may use the superscript θ, which is either a parameter within $[0,1]$ of a θ-scheme or $\theta = \text{ex}$ for the exact solution. This means that

$$\begin{cases} \lambda_f^{\theta_f}(\delta t) = \dfrac{1 - Nc\theta_f\delta t}{1 + (1-\theta_f)Nc\delta t}; & \lambda_s^{\theta_s}(\delta t) = \dfrac{1 - c\theta_s\delta t}{1 + (1-\theta_s)c\delta t}; \\ \lambda_f^{\text{ex}}(\delta t) = e^{-Nc\delta t}; & \lambda_s^{\text{ex}}(\delta t) = e^{-c\delta t}. \end{cases}$$

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 associated schemes are A-stable). The stability interval $\mathcal{I}_N$ is the intersection of these domains in δt.

Remark 2: Recall that $\lambda_f^{\theta_f}(\delta t) = 1 - Nc\delta t + N^2c^2(1-\theta_f)\delta t^2 + \mathcal{O}(\delta t^3)$.

For any functions of δt λ_f, λ_s, we consider the following four schemes: given $W^n \in \mathbb{R}^2$, we set

- **Scheme #1**: (Lie type - slow time - subcycled) $W^{n+1} = G_1(\delta t)W^n$ where

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

- **Scheme #2**: (Lie type - fast time - no subcycling) $W^{n+1} = G_2(\delta t)W^n$ where

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

- **Scheme #3**: (Strang type - slow time - subcycled) $W^{n+1} = G_3(\delta t)W^n$ where

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

[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.

- **Scheme #4**: (Strang type - fast time - no subcycling) $W^{n+1} = G_4(\delta t)W^n$ where

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

Using the notations of Lemma 2.1, we obtain

Scheme #1	$\alpha_1(\delta t) = 1 - (\lambda_f(\delta t/N))^N$	$\beta_1(\delta t) = (1 - \lambda_s(\delta t))(\lambda_f(\delta t/N))^N$
Scheme #2	$\alpha_2(\delta t) = 1 - \lambda_f(\delta t/N)$	$\beta_2(\delta t) = (1 - \lambda_s(\delta t/N))\lambda_f(\delta t/N)$
Scheme #3	$\alpha_3(\delta t) = (1 - \lambda_f(\delta t/N)^N)[\lambda_s(\delta t/2)]^N$	$\beta_3(\delta t) = (1 - \lambda_s(\delta t/2))(1 + [\lambda_f(\delta t/N)^N]\lambda_s(\delta t/2))$
Scheme #4	$\alpha_4(\delta t) = (1 - \lambda_f(\delta t/N))\lambda_s(\delta t/2)$	$\beta_4(\delta t) = (1 - \lambda_s(\delta t/2))(1 + \lambda_f(\delta t/N)\lambda_s(\delta t/2))$

Asymptotic order The above computations enable us to prove the following

Proposition 2.5: A linear Lie-splitting method such as Scheme #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 2.4. Let us consider Scheme #1 and assume that we have the following Taylor expansion for $\lambda_s(\delta t)$ and $\lambda_f(\delta t/N)$:

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

We derive that

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

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

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

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 3: The crucial point lies in the fact that the linear combination of the derivatives 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. Let us examine the θ-schemes case. One infers from Remark 2 that $(A_f, A_s) = (1-\theta_f, 1-\theta_s)$. So, as soon as $N > 1$, one cannot build schemes of type #1 or #2 of A-order at least 2 with θ-schemes, unless, in Scheme #2, the slow scheme is fully explicit and the fast scheme is fully implicit. In this very particular case, the A-order is infinite because $\alpha_2 = N\beta_2$. 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 δt from the fast scheme part.

Proposition 2.6: A linear Strang-splitting method such as Scheme #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.

Proof. The fact that a Strang-splitting method involving two methods of order 2 is of A-order 2 comes from Theorem 2.4. Assume we have the same Taylor expansion as in the proof of Proposition 2.5. For Scheme #3, we have

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

and for Scheme #4

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

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 4: 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. In particular, in case the fast and slow schemes are θ-schemes, the above condition $1 - 2A_s + 4A_f - 2 = 0$ for Scheme #3 reads $4\theta_f - 2 + 1 - 2\theta_s = 0$ so that we have a fairly natural one-parameter family of couples of schemes of order 1, not depending on N, that lead to a subcycled scheme (#3) of A-order at least 2. In particular, one can choose to use an explicit Euler scheme for the slow equation ($\theta_f = 1$) and a semi-implicit scheme for the fast equation ($\theta_s = 1/4$) so that the subcycled Strang-splitting scheme #3 is at least of A-order 2. It is also possible to build a second A-order scheme #4 with θ-schemes provided one solves $2N(1 - 2\theta_s) + 2\theta_f - 1 = 0$. One sees in that case that the influence of the choice of θ_f weakens as N increases.

Remark 5: 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 δt^2 is then $4\theta_s - 2 + 1 - 2\theta_f = 0$ (resp. $2(1 - 2\theta_s) + N(2\theta_f - 1) = 0$) for Scheme $\widetilde{\#3}$ (resp. $\widetilde{\#4}$). One concludes easily that it is then possible to build a $\widetilde{\#3}$ scheme of A-order 2 with an explicit fast scheme ($\theta_f = 1$) and a semi-implicit slow scheme ($\theta_s = 3/4$). For $\widetilde{\#4}$ schemes, one notes that the choice of the fast scheme is now the most important.

Convergence rate Let us perform the same analysis on the convergence rate to equilibrium, *i.e.* the eigenvalues μ_i, $i \in \{1, \ldots, 4\}$. We get the following Taylor expansions of $\rho_i(\delta t) = \mu_i(\delta t) - \mathrm{e}^{-c(N+1)\delta t}$, that we summarize in the following table in the (A_f, A_s) form:

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

One notes at once that second order fast and slow schemes generate a second 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 be exactly 1. The only combination of θ-schemes leading to a second order approximated rate and of 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). In any case, the choice of the fast scheme plays a greater role than that of the slow scheme for the approximated rate. As predicted in Appendix A, schemes $\widetilde{\#3}$ and $\widetilde{\#4}$ have similar rates of convergence to that of schemes #3 and #4: the choice of the fast scheme is always predominant.

2.4 Conclusion

Let us remind the reader that the applications we have in mind involve a fast equation for which an implicit scheme is hard to solve, thus implying the use of an explicit scheme, inducing a stability constraint on the numerical time-step δt. 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 a scheme of type $\widetilde{\#3}$ with $\theta_f = 1$ (explicit), $\theta_s = 1/4$ (semi-implicit), which will be of second A-order, even though it is (locally) consistent of order 1 with (1) and has a rate of convergence which approximates the exact rate at order 1. It is the only scheme, among the four types described above and involving an explicit resolution of the fast equation, that achieves a second A-order.

3 Analysis of the nonlinear system

3.1 Analysis of the exact solutions

In this section, we investigate the long time behavior of the two-scale nonlinear system (2). Let us first write this system in the form

$$\begin{cases} u' &= -N(u-v)[c+(u-v)] \\ v' &= (u-v)[c+(u-v)]. \end{cases} \tag{8}$$

This way, we are able to derive the following

Proposition 3.1: 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 $X' = 0, \quad Y' = -(N+1)Y(c - Y)$. 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 follows from this observation and the fact that the system reads (8). ■

3.2 Splitting schemes with or without subcycling for the nonlinear problem (2)

Let us recall this result providing an estimate of the 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 3.2: 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 $\varphi_{e,\delta t}$ the exact flow of this equation. Let us denote by $\varphi_f(\delta t)$ (respectively) $\varphi_s(\delta t)$ the propagators at time δt of the two split equations:

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

Assume that $S_{f,\delta t}$ and $S_{s,\delta t}$ are numerical methods of respective orders p and q. Assume that a splitting method is defined for $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{C}$ 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 r. Then the order of the method $\Phi_{\delta t}$ is at least $\min(p,q,r)$, and so is the order of the method with subcycling

$$\Phi_{\delta t}^{sc} = \Pi_{i=1}^n (S_{s,b_i\delta t} \circ (S_{f,a_i\delta t/N})^N). \tag{9}$$

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

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

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

$$S^p_{f,\delta t/N} = \varphi^p_{f,\delta t/N} + \mathcal{O}(\delta t^{q+1}),$$

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

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

This implies that

$$\begin{aligned}\Phi^{sc}_{\delta t} = \Pi^n_{i=1}(S_{s,b_i\delta t} \circ (S_{f,a_i\delta t/N})^N) &= \Pi^n_{i=1}(\varphi_{s,b_i\delta t} + \mathcal{O}(\delta t^{p+1})) \circ (\varphi_{f,a_i\delta t} + \mathcal{O}(\delta t^{q+1})) \\ &= \Pi^n_{i=1}(\varphi_{s,b_i\delta t} \circ \varphi_{f,a_i\delta t}) + \mathcal{O}(\delta t^{\min(p,q)+1}) \\ &= \varphi_{e,\delta t} + \mathcal{O}(\delta t^{\min(p,q,r)+1}),\end{aligned}$$

since the splitting method is assumed to have order r when used with the exact flows. This proves the result for $\Phi^{sc}_{\delta t}$. The result for $\Phi_{\delta t}$ is even simpler. ■

In the following, we consider numerical splitting methods for the nonlinear problem (2) 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, and Scheme #4 is a FSF Strang-splitting method without subcycling.

Once again, we consider numerical flows for the integration of the split equations described by θ-schemes, *i.e.* for the fast equation, the first component of $\Phi^{\theta_f}_{f,\delta t}(u^n, v^n)$ 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 and, for the slow equation, the second component of $\Phi^{\theta_f}_{f,\delta t}(u^n, v^n)$ solves the equation in X

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

while its first component is its first argument.

3.3 Numerical examples of splitting methods for problem (2)

We run the four schemes with four different values of the couple (θ_f, θ_s). We sum up the results on the asymptotic order in Table 1 and provide numerical results in Figure 2. These results were obtained with final time $T = 2.0$, speed $c = 1$, factor $N = 50$, initial datum $(u^0, v^0) = (5, 1)$, so that, using the analysis carried out in the proof of Proposition 3.1, the exact solution at final time is within a distance smaller than 10^{-40} of its asymptotic limit $55/51 \times (1, 1)$.

By Theorem 3.2, we know that the Lie-splitting schemes (Scheme #1 and Scheme #2) are of classical order 1 for any possible choice of (θ_f, θ_s). The first two columns of Table 1 show that the asymptotic order is also 1 in these cases. Theorem 3.2 also implies that the Strang-splitting scheme #3 has at least order 1 with the choice $(\theta_f, \theta_s) = (1, 0)$ and the asymptotic orders collected at the end of the first line of Table 1 show that the 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 at the end of the second line of Table 1 show that the asymptotic order is also 2 in this case. The end of the 2 last lines is surely the most interesting part of this section: for $(\theta_f, \theta_s) = (1, 3/4)$ and $(\theta_f, \theta_s) = ((N-1)/(2N), 1/4)$, the classical order of the splitting method is, by Theorem 3.2 at least 1. In the first case $(\theta_f, \theta_s) = (1, 3/4))$, the numerical results show that the subcycled scheme #3 has A-order 2 while the Strang-splitting 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 4). 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 (linear) Scheme #4 has A-order 2.

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 (2) 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.

(θ_f, θ_s)	Scheme #1	Scheme #2	Scheme #3	Scheme #4
$(1.0, 0.0)$	0.9671	1.0000	1.2808	1.0499
$(0.5, 0.5)$	0.9677	1.0000	2.0071	1.9952
$(1.0, 0.75)$	0.9686	1.0021	1.9889	1.0548
$(\frac{N-1}{2N}, 0.25)$	0.9672	0.9991	1.5041	1.9932

Tab. 1: Asymptotic error for the 4 schemes for some values of (θ_f, θ_s).

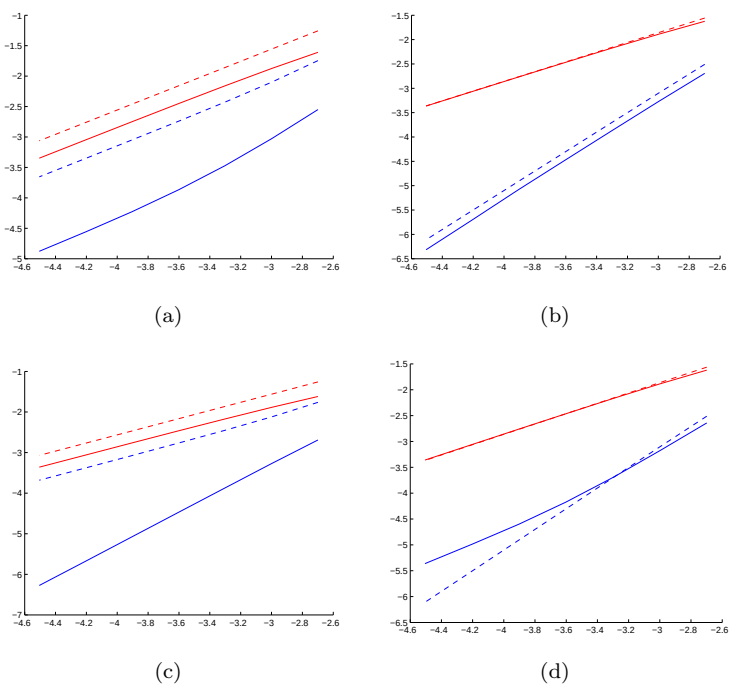

Fig. 2: Logarithm of the asymptotic error as a function of the time step: Scheme #1 (full red line), Scheme #2 (dotted red line), Scheme #3 (full blue line), Scheme #4 (dotted blue line). $(\theta_f, \theta_s) = (1.0, 0.0)$ (a), $(\theta_f, \theta_s) = (0.5, 0.5)$ (b), $(\theta_f, \theta_s) = (1.0, 0.75)$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

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

$$\left\{ \begin{array}{rcl} \partial_t u & = & \nu_1 \Delta u + c_1(v-u) \\ \partial_t v & = & \nu_2 \Delta v + c_2(u-v) \end{array} \right. \qquad t>0, x \in (0,L), \tag{10}$$

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. Of course, $\Delta = \partial_x^2$ is the Laplace operator and $L>0$ is given. We focus on the case where one of the equations in System (10) is "fast" and the other is "slow". Moreover, we assume the "speed" ratios allow us to actually do subcycling. This means that

$$\frac{\nu_1}{\nu_2} = \frac{c_1}{c_2} = N \in \mathbb{N}^\star, \tag{11}$$

and $N >> 1$[6]. 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 (10) is the "fast" one, so u is the "fast" unknown and the second one (on v) is the slow one. Let us recall that we have the following

Theorem 4.1: For all initial data $(u^0, v^0) \in L^2(0,L)^2$, System (10) 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

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

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 λ_k and μ_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 $+\infty$,

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

■

The following theorem deals with the asymptotic behavior of the solutions of System (10):

Theorem 4.2: For all solutions (u,v) of System (10) and all $t \geq 0$, we have

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

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

Proof. Let (u, v) be a smooth solution of (10). We compute

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

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 $M \times M$ matrix $A = toeplitz(-1, 2, -1, 0)$ where $M \in \mathbb{N}^\star$ and $\delta x = L/(M+1)$. We note for all $i \in \{0, \ldots, M+1\}$, $x_i = i.\delta x$ and $U = (u_1, \ldots, u_M)$ will be the solution of the discretized problem. Let us recall that the eigenvalues and associated eigenvectors of A are, for $1 \leq p \leq M$,

$$\left(\lambda_p = 4\sin^2\left(\frac{p\pi}{2(M+1)}\right), (\sin(1p\pi/(M+1)), \sin(2p\pi/(M+1)), \ldots, \sin(Mp\pi/(M+1)))\right). \tag{12}$$

In the following, we denote by

$$A = PDP^{-1} \tag{13}$$

the corresponding diagonalization of A.

The time discretization: numerical analysis of the rate of convergence for several time-splitting schemes 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 (10). We consider $(p, p', q, q') \in (\mathbb{N}^*)^4$ such that

$$\frac{q'}{q} = \frac{p'}{Np}. \tag{14}$$

The "fast" one is discretized on an interval of length $\delta t/(Np)$ and we denote by $\Phi_{fast,\delta t/(Np)}$ 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 $\Phi_{slow,\delta t/q}$ 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

$$\Psi_{Lie,\delta t} = \Phi_{slow,\delta t} \circ \Phi^N_{fast,\delta t/N}, \tag{15}$$

corresponding to $(p, p', q, q') = (1, N, 1, 1)$. As we did in Section 2 and in Section 3, we consider θ-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:

$$\Phi_{fast,\delta t/N}(u^n, v^n) = \left[\left(I - \theta_f\delta t\left(cI + \nu\frac{1}{(\delta x)^2}A\right)\right)\left(I + (1-\theta_f)\delta t\left(cI + \nu\frac{1}{(\delta x)^2}A\right)\right)^{-1} u^n + c\delta t v^n\,,\, v^n\right], \tag{16}$$

and

$$\Phi_{slow,\delta t}(u^n, v^n) = \left[u^n\,,\, \left(I - \theta_s\delta t\left(cI + \nu\frac{1}{(\delta x)^2}A\right)\right)\left(I + (1-\theta_s)\delta t\left(cI + \nu\frac{1}{(\delta x)^2}A\right)\right)^{-1} v^n + c\delta t u^n\right]. \tag{17}$$

This way, a stability condition reads

$$\delta t \leq \frac{1}{c + 4\nu/(\delta x)^2}. \tag{18}$$

Note also that the stability condition (18) on the scheme is actually independent on 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 δt 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 δt. The matrices of the linear mappings $\Phi_{slow,\delta t/q}$ and $\Phi_{fast,\delta t/(Np)}$ in the canonical basis of $\mathbb{R}^{2M}$ read respectively

$$M_s(\delta t/q) = \begin{pmatrix} I & 0 \\ c\frac{\delta t}{q} C_s^{-1} & B_s C_s^{-1} \end{pmatrix} \qquad \text{and} \qquad M_f(\delta t/(Np)) = \begin{pmatrix} B_f C_f^{-1} & c\frac{\delta t}{p} C_f^{-1} \\ 0 & I \end{pmatrix}. \tag{19}$$

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 $\Phi_{fast,\delta t/(Np)}^{p'}$ 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 (14), we define $\Psi_{\delta t,p,p',q,q'} = \Phi_{slow,\delta t/q}^{q'} \circ \Phi_{fast,\delta t/(Np)}^{p'}$ the matrix of which reads

$$\begin{pmatrix} (B_f C_f^{-1})^{p'} & c\frac{\delta t}{p} C_f^{-1} \Sigma_{f,p'} \\ c\frac{\delta t}{q} C_s^{-1} (B_f C_f^{-1})^{p'} \Sigma_{s,q'} & (B_s C_s^{-1})^{q'} + c^2 \frac{\delta t^2}{pq} C_s^{-1} C_f^{-1} \Sigma_{s,q'} \Sigma_{f,p'} \end{pmatrix}. \tag{20}$$

In particular, if $q = q' = p = 1$ and $p' = N$, $\Psi_{\delta t,p,p',q,q'} = \Psi_{Lie,\delta t}$ and, if $q = q' = 2$, $p = 1$ and $p' = N$, $\Psi_{\delta t,p,p',q,q'}$ and $\Psi_{Strang,\delta t} = \Phi_{slow,\delta t/2} \circ \Phi_{fast,\delta t/N}^N \circ \Phi_{slow,\delta t/2}$ are similar and thus share the same spectrum.

Denoting by $\mathcal{P}$ the matrix (see (13))

$$\mathcal{P} = \begin{pmatrix} P & 0 \\ 0 & P \end{pmatrix}, \tag{21}$$

we obtain that the matrix $\mathcal{D} := \mathcal{P}^{-1} \Psi_{\delta t,p,p',q,q'} \mathcal{P}$ is exactly the same as that of (20) 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 of size $2M \times 2M$, each of which is diagonal. We infer that all the eigenvalues of $\Psi_{\delta t,p,p',q,q'}$ are the roots of the M polynomial equations

$$\lambda^2 - \left((\phi_f^{-1}\psi_f)^{p'} + (\phi_s^{-1}\psi_s)^{q'} + c^2 \frac{\delta t^2}{pq} \phi_f^{-1} \phi_s^{-1} \widetilde{\Sigma}_{s,q'} \widetilde{\Sigma}_{f,p'} \right) \lambda + (\phi_f^{-1}\psi_f)^{p'} (\phi_s^{-1}\psi_s)^{q'} = 0, \tag{22}$$

where

$$\psi_{f,s}(\mu) = 1 - \theta_{f,s} \frac{\delta t}{p} \mu \quad \text{and} \quad \phi_{f,s}(\mu) = 1 + (1-\theta_{f,s}) \frac{\delta t}{p} \mu, \tag{23}$$

$$\widetilde{\Sigma}_{f,p'} = \sum_{k=0}^{p'-1} (\phi_f^{-1}\psi_f)^k \quad \text{and} \quad \widetilde{\Sigma}_{s,q'} = \sum_{k=0}^{q'-1} (\phi_s^{-1}\psi_s)^k, \tag{24}$$

and μ is an eigenvalue of $cI + \nu A/(\delta x)^2$. We extend these six real-valued functions of μ to the continuous interval $(c, c + 4\nu/(\delta x)^2)$. The functions $\mu \mapsto \phi_i^{-1}(\mu)$ and $\mu \mapsto \psi_i(\mu)$ are smooth, decreasing 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 decreasing on $(c, c+4\nu/(\delta x)^2)$. For example,

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

are positive decreasing functions on $(c, c+4\nu/(\delta x)^2)$. Note that the discriminant of the polynomial (22) is

$$\begin{aligned}
\mathcal{D}(\mu) :=& \Big(P(\mu)+Q(\mu)+\Sigma(\mu)\Big)^2 - 4Q(\mu)P(\mu) \\
=& \Big(Q(\mu)-P(\mu)+\Sigma(\mu)\Big)^2 + 4P(\mu)\Sigma(\mu) > 0 \qquad (25)\\
=& \Big(P(\mu)-Q(\mu)+\Sigma(\mu)\Big)^2 + 4Q(\mu)\Sigma(\mu) > 0, \qquad (26)
\end{aligned}$$

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

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

for $\mu \in (c, c+4\nu/(\delta x)^2)$. Note that, with the stability condition (18), we have for all μ, $0 < \lambda^-(\mu) < \lambda^+(\mu)$. Moreover, we have the following monotonicity property:

Lemma 4.3: The map $\mu \mapsto \lambda^+(\mu)$ is decreasing in $(c, c+4\nu/(\delta x)^2)$[7].

Proof. Note that, thanks to (25), $\sqrt{\mathcal{D}(\mu)} > Q(\mu) - P(\mu)$ if $Q(\mu) > P(\mu)$. Similarly, (26) leads to $\sqrt{\mathcal{D}(\mu)} > P(\mu) - Q(\mu)$ if $P(\mu) > Q(\mu)$ since P, Q, Σ are positive functions. So $\sqrt{\mathcal{D}} > |P-Q|$. Differentiating the function $\mu \mapsto \lambda^+(\mu)$ with respect to μ yields

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

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

Hence the biggest eigenvalue of $\Psi_{\delta t,p,p',q,q'}$ is $\lambda^+(\mu_1)$ with $\mu_1 := c + \nu\lambda_1/(\delta x)^2$ (see (12)). Of course, an asymptotic expansion of that biggest eigenvalue as $\delta t \to 0^+$ helps us controlling the exponential decay of the L^2 norm of the numerical solution provided by $\Psi_{\delta t,p,p',q,q'}$. This allows us to prove the following

Theorem 4.4: Let $c, \nu > 0$, $N \geq 2$ and $\Phi_{\delta t,p,p',q,q'}$ be defined as above. Assume $M \in \mathbb{N}^\star$ is given. There exists[8] $C, \gamma, h > 0$ such that for all $T > 0$, all $U^0, V^0 \in \mathbb{R}^M$, all $\delta t \in (0,h)$ and all $n \in \mathbb{N}$ with $n\delta t \leq T$, we have

$$\|\Psi^n_{Lie,\delta t}(U^0,V^0)\|_2 \leq C\mathrm{e}^{-\gamma n\delta t}\|(U^0,V^0)\|_2. \qquad (27)$$

Remark 6: 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 $M \to +\infty$), we have, at least asymptotically with respect to δx, a numerical decay rate of the appropriate order with respect to the parameters ν and L: we compare the exact decay rate $\nu\pi^2/L^2$ from Theorem 4.2 with the asymptotic numerical one $N\nu\pi^2/(L^2(N+1))$ (recall that N is large).

[7] Note that $\mathcal{D}$ is not a decreasing function of μ in general.
[8] The reason for the constant C is the lack of symmetry of the matrix $\mathcal{D}$.

Proof. Let $M \in \mathbb{N}^\star$ be fixed. 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

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

Following the same way, we obtain Taylor expansions for $P(\mu_1)$, $Q(\mu_1)$, $\Sigma(\mu_1)$ and then $\mathcal{D}(\mu_1)$ and eventually $\lambda^+(\mu_1)$ when δt tends to 0:

$$\lambda_1^+ = 1 - \frac{(N+1)\mu_1 - \sqrt{(N-1)^2\mu_1^2 + 4Nc^2}}{2}\delta t + \mathcal{O}(\delta t^2), \tag{28}$$

and therefore,

$$\frac{T}{\delta t}\ln(\lambda_1^+) = -\frac{T}{2}\Big((N+1)\mu_1 - \sqrt{(N-1)^2\mu_1^2 + 4Nc^2}\Big) + \mathcal{O}(\delta t).$$

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

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

and therefore

$$(N+1)\mu_1 - \sqrt{(N-1)^2\mu_1^2 + 4Nc^2} > 0.$$

Since λ_1^+ is the biggest eigenvalue of $\Psi_{Lie,\delta t}$, this proves the result. Note also that the constant γ can be taken arbitrary close to

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

Using the mean value theorem, for some $c_\theta \in (0, 4N(\mu_1^2 - c^2))$, the latter quantity is equal to

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

■

4.2 The non homogeneous Dirichlet problem

The continuous problem In this section we consider System (10) equipped with inhomogeneous Dirichlet boundary conditions, namely

$$u(t,0) = u_l, \qquad u(t,L) = u_r, \qquad v(t,0) = v_l, \qquad v(t,L) = v_r, \tag{29}$$

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 4.5: The PDE system (10) with non homogeneous Dirichlet boundary conditions has a unique stationary solution given by

$$\begin{cases} u_{\text{ex}}^\infty : \ x \mapsto \frac{u_l+v_l}{2} + \frac{(u_r+v_r-u_l-v_l)x}{2L} + \frac{(u_l-v_l)[\cosh(x/\alpha)-\cosh(L/\alpha)\sinh(x/\alpha)/\sinh(L/\alpha)]}{2} + \frac{(u_r-v_r)\sinh(x/\alpha)/\sinh(L/\alpha)}{2} \\ v_{\text{ex}}^\infty : \ x \mapsto \frac{u_l+v_l}{2} + \frac{(u_r+v_r-u_l-v_l)x}{2L} - \frac{(u_l-v_l)[\cosh(x/\alpha)-\cosh(L/\alpha)\sinh(x/\alpha)/\sinh(L/\alpha)]}{2} - \frac{(u_r-v_r)\sinh(x/\alpha)/\sinh(L/\alpha)}{2} \end{cases} \tag{30}$$

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 non homogeneous reaction-diffusion system (10)-(29) 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 (30) to it (see Theorem 4.1). Moreover, for all initial datum (u^0, v^0), the solution of the non homogeneous System (10) converges exponentially fast to the stationary solution (30).

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.

Space and time discretizations Using the same space discretization as above (see Section 4.1), we consider two θ-schemes for the time discretization in the spirit of what we did for the homogeneous problem (see (16)-(17)), with parameters θ_f and θ_s. Taking into account the non homogeneous Dirichlet boundary conditions yields a sequence $((U^n, V^n)^t)_{n\in\mathbb{N}}$ defined by an arithmetic-geometric recursion: given $W^0 = (U^0, V^0)^t \in \mathbb{R}^{2M}$, we have for all $n \geq 0$,

$$W^{n+1} = \mathcal{M}W^n + \mathcal{M}_u \begin{pmatrix} U_{l,r} \\ 0_M \end{pmatrix} + \mathcal{M}_v \begin{pmatrix} 0_M \\ V_{l,r} \end{pmatrix} =: \mathcal{M}W^n + \Upsilon \tag{31}$$

where $\mathcal{M}$ is defined as a product of matrices of the form (19), $U_{l,r} = (u_l, 0, \ldots, 0, u_r)^t$, $V_{l,r} = (v_l, 0, \ldots, 0, v_r)^t$ and $\mathcal{M}_u$ and $\mathcal{M}_v$ are $2M$-by-$2M$ matrices, depending on δt, δx and the choice of the splitting method between the two θ-schemes.

Let us list the numerical experiments we conducted:

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

$$\mathcal{M} = M_s M_f^N, \;\; \mathcal{M}_u = \nu \frac{\delta t}{\delta x^2} M_s \sum_{k=0}^{N-1} M_f^k \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \;\text{ and }\; \mathcal{M}_v = \nu \frac{\delta t}{\delta x^2} \begin{pmatrix} 0 & 0 \\ 0 & C_s^{-1} \end{pmatrix} \tag{32}$$

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

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

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

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

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

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

Equilibrium states of the splitting schemes We prove the existence of a unique equilibrium state for the splitting schemes above, comment on the rate of convergence of the schemes towards their equilibrium state and also analyze how close the equilibrium state of each scheme is to a projection on the numerical space grid of the equilibrium state (30) of the continuous reaction-diffusion system (10) with non homogeneous Dirichlet conditions (29) in an L^2 sense. Following (18), we denote by $\mathrm{CFL}(M)$ the positive real number

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

We endow $\mathbb{R}^M$ and $\mathbb{R}^M \times \mathbb{R}^M$ with the classical Euclidian norms denoted by $\|\cdot\|_2$ and the corresponding algebras of square matrices with the induced norms denoted by $|||\cdot|||_2$. To compute the asymptotic numerical solution of a given method of type (31), we need to solve the $2M$-by-$2M$ linear system

$$(I_{2M} - \mathcal{M})W = \Upsilon. \tag{33}$$

Proposition 4.6: Let $\delta t, \delta x > 0$ satisfying (18) be fixed. For a general Lie-splitting method of the form (15), the numerical asymptotic state of any of the splitting schemes defined by the recursion relation (31) above is the unique solution W^{∞}_{num} of the linear system (33).

Proof. Since $\delta t, \delta x$ satisfy (18), we know from Theorem 4.4 that the spectral radius of the matrix $\mathcal{M}$ of $\Psi_{Lie,\delta t}$ in the canonical basis of $\mathbb{R}^{2M}$ is less than 1. Hence, the matrix $I_{2M} - \mathcal{M}$ 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 4.4.

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

Theorem 4.7: Provided that $\delta t \in (0, \text{CFL}(M))$, the asymptotic state of Scheme #1 is a uniform-in-δt second order approximation of the exact asymptotic state given in Proposition 30[9]:

$$\begin{pmatrix} \Pi_{\delta x}\big(u^{\infty}_{\text{ex}}\big) \\ \Pi_{\delta x}\big(v^{\infty}_{\text{ex}}\big) \end{pmatrix} - W^{\infty}_{\text{num}}(\delta t) = \mathcal{O}(\delta x^2).$$

Proof. To analyze the asymptotic convergence of Scheme #1, we put the projections $\Pi_{\delta x}\big(u^{\infty}_{\text{ex}}\big)$ and $\Pi_{\delta x}\big(v^{\infty}_{\text{ex}}\big)$ of the exact solutions u^{∞}_{ex} and v^{∞}_{ex} defined in (30) in the numerical scheme. Using the identity

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

and the fact that the functions u^{∞}_{ex} and v^{∞}_{ex} are solutions of (10) with the non-homogeneous Dirichlet boundary conditions (29), we first compute

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

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

$$M_f^N \begin{pmatrix} \Pi_{\delta x}\big(u^{\infty}_{\text{ex}}\big) \\ \Pi_{\delta x}\big(v^{\infty}_{\text{ex}}\big) \end{pmatrix} = \begin{pmatrix} \Pi_{\delta x}\big(u^{\infty}_{\text{ex}}\big) \\ \Pi_{\delta x}\big(v^{\infty}_{\text{ex}}\big) \end{pmatrix} - \nu \frac{\delta t}{\delta x^2} \sum_{k=0}^{N-1} M_f^k \begin{pmatrix} C_f^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_{l,r} \\ 0 \end{pmatrix} + \mathcal{O}(\delta t (\delta x)^2), \tag{34}$$

where, once again, the constant in the $\mathcal{O}$ is independent of δt and δx provided that the CFL condition (18) 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 $\delta t \in (0, \text{CFL}(M))$ thanks to Lemma 4.8 that follows. ■

Lemma 4.8: There exists a positive constant $C > 0$ such that, for all $M \in \mathbb{N}^{\star}$ and all $\delta t \in (0, \text{CFL}(M))$, we have

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

[9] for $w \in C^0([0, L])$, $\Pi_{\delta x}(w) = (w(x_1), \ldots, w(x_M))^t$

Remark 7: 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

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

The CFL condition (18) 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 1. Hence, using the algebra property, it is sufficient to prove the result for the second matrix in the product above, which is *not* symmetric. Using the fact that, for any square matrix R with real coefficients, $|||R|||_2^2 = \rho(A^t A)$, where ρ denotes the spectral radius, we are left with the computation of the eigenvalues of the symmetric non-negative matrix

$$N_f = \begin{pmatrix} B_f^2 & c\delta t B_f \\ c\delta t B_f & (1+c^2\delta t^2) I_M \end{pmatrix}.$$

The eigenvalues of the matrix N_f are the $2M$ roots of the M polynomials

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

where $(\mu_p)_{1\le p\le M}$ denotes the list of the eigenvalues of B_f. The CFL condition (18) ensures that for all $p \in \{1,\dots,M\}$, $\mu_p \in [0,1]$. 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 δt which yileds the result with $C = \sqrt{2(2+c^2/(c+16\nu/L^2)^2)}$. ■

Multiplying (34) by M_s, we finally get

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

Comparing this relation with that defining the numerical equilibrium state (31) (with the right-hand side defined in (32)), we infer that

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

where the constant in the $\mathcal{O}$ is independent of δt and δx provided that the CFL condition (18) is fulfilled. Finally, let us state and prove the following Proposition about the inverse of the matrix $(I_{2M} - M_s M_f^N)$:

Proposition 4.9: There exists a positive constant $C > 0$ such that for all $M \in \mathbb{N}^\star$ and all $\delta t \in (0, \text{CFL}(M))$,

$$|||(I - M_s M_f^N)^{-1}|||_2 \le \frac{C}{\delta t}. \tag{35}$$

Proof. Let us fix $M \in \mathbb{N}^\star$ and $\delta t \in (0, \text{CFL}(M))$. Using the conjugation with the orthogonal matrix $\mathcal{P}$ (see (21)), we have that the $|||\cdot|||_2$-norm of $I_{2M} - M_s M_f^N$ is equal to that of the same matrix where A is replaced with D (see (13)). The latter matrix has a very particular structure: the four M-by-M matrices defining it are diagonal. Let us denote by $(a_i)_{1\le i\le M}$, $(b_i)_{1\le i\le M}$, $(c_i)_{1\le i\le M}$, and $(d_i)_{1\le i\le M}$ these entries such that

$$Z := \mathcal{P}^{-1}(I - M_s M_f^N)\mathcal{P} = \begin{pmatrix} a_1 & 0 & 0 & b_1 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & a_M & 0 & 0 & b_M \\ c_1 & 0 & 0 & d_1 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & c_M & 0 & 0 & d_M \end{pmatrix}.$$

The eigenvalues of Z lie in $(0,1)$ (see Theorem 4.4). Hence, Z is invertible and its inverse is given by

$$Z^{-1} = \mathcal{P}^{-1}(I - M_s M_f^N)^{-1}\mathcal{P} = \begin{pmatrix} \alpha_1 & 0 & 0 & \beta_1 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \alpha_M & 0 & 0 & \beta_M \\ \gamma_1 & 0 & 0 & \delta_1 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \gamma_M & 0 & 0 & \delta_M \end{pmatrix},$$

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

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

One can check easily that

$$|||Z^{-1}|||_2 = \max_{1 \le i \le M} |||m_i|||_2.$$

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

$$|||m_i|||_2^2 \le \frac{b_i^2 + c_i^2}{(a_i d_i - b_i c_i)^2} + \frac{a_i^2 + d_i^2}{(a_i d_i - b_i c_i)^2}, \tag{36}$$

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 (20), we have

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

and

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

For the first term in the upper bound (36), let us show that the numerator is $\mathcal{O}(\delta t^2)$ while the denominator is bounded from below by a positive constant times δt^4.
On the one hand, we have

$$|b_i|^2 \le c^2 N^2 \delta t^2 \qquad \text{and} \qquad |c_i|^2 \le c^2 \delta t^2. \tag{37}$$

On the other hand, for all $i \in \{1, \dots, M\}$, 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 (18) 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 (23), we have

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

so that

$$a_i d_i - b_i c_i = \delta t\frac{(1 - (\psi_f\phi_f^{-1})^N(\mu_i))}{\phi_s(\mu_i)}\frac{\mu_i^2 - c^2}{\mu_i}. \tag{38}$$

The CFL condition (18) 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

$$1-(\psi_f\phi_f^{-1})^N(\mu_i) \geq 1-(\psi_f\phi_f^{-1})(\mu_i) = \frac{\delta t\mu_i}{1+(1-\theta_f)\delta t\mu_i} \geq \frac{\delta t\mu_i}{2}. \tag{39}$$

This allows to bound from below

$$a_id_i - b_ic_i \geq \frac{\delta t^2}{4}\underbrace{(\mu_i+c)}_{\geq c}(\underbrace{\mu_i-c}_{=\nu\lambda_i/\delta x^2}) \geq c\nu\frac{\delta t^2}{4}\frac{\lambda_1}{\delta x^2}.$$

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

$$\frac{\lambda_1}{\delta x^2} = \frac{4}{\delta x^2}\sin^2\left(\frac{\pi}{2}\frac{1}{(M+1)}\right) \geq 4\frac{(M+1)^2}{L^2}\frac{4}{\pi^2}\frac{\pi^2}{4}\frac{1}{(M+1)^2} \geq \frac{4}{L^2}. \tag{40}$$

This proves

$$a_id_i - b_ic_i \geq \frac{c\nu}{L^2}\delta t^2. \tag{41}$$

Using (37) and (41), there exists a positive constant C such that

$$\forall M\in\mathbb{N}^\star,\quad \forall\delta t\in(0,\mathrm{CFL}(M)),\qquad \frac{b_i^2+c_i^2}{(a_id_i-b_ic_i)^2} \leq \frac{C}{\delta t^2}. \tag{42}$$

Let us now bound the second term in the right hand side of (36). Let us fix $M\in\mathbb{N}^\star$ and $i\in(0,\mathrm{CFL}(M))$ again. From (38), we have

$$\frac{1}{(a_id_i-c_ib_i)^2} = \frac{1}{\delta t^2}\frac{\phi_s^2(\mu_i)}{(1-(\psi_f\phi_f^{-1})^N(\mu_i))^2}\left(\frac{\mu_i}{\mu_i^2-c^2}\right)^2.$$

A similar direct calculation yields

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

We infer

$$\frac{a_i^2+d_i^2}{(a_id_i-c_ib_i)^2} = \frac{1}{\delta t^2}\phi_s^2(\mu_i)\left(\frac{\mu_i}{\mu_i^2-c^2}\right)^2\left[1+\frac{1}{\phi_s^2(\mu_i)}\left(\frac{\mu_i\delta t}{1-(\psi_f\phi_f^{-1})^N(\mu_i)} - \frac{c^2}{\mu_i}\delta t\right)^2\right]. \tag{43}$$

We can bound the terms in the product above as follows. The CFL condition (18) implies that $\phi_s^2(\mu_i)\leq 4$. Moreover, using (40), 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 (39), 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 (43), taking products and using Young's inequality, we infer that

$$\forall M \in \mathbb{N}^\star, \quad \forall \delta t \in (0, \mathrm{CFL}(M)), \qquad \frac{a_i^2 + d_i^2}{(a_i d_i - b_i c_i)^2} \leq \frac{11}{4} \frac{L^4}{\nu^2} \frac{1}{\delta t^2}. \tag{44}$$

The inequalities (42) and (44) together with (36) prove the result. ■

The numerical tests we conducted for several values of θ_f, θ_s and N showed that the matrix $I_{2M} - \mathcal{M}$ 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, $M = 20, 40, 80, 160$, $\delta x = L/(M+1)$:

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

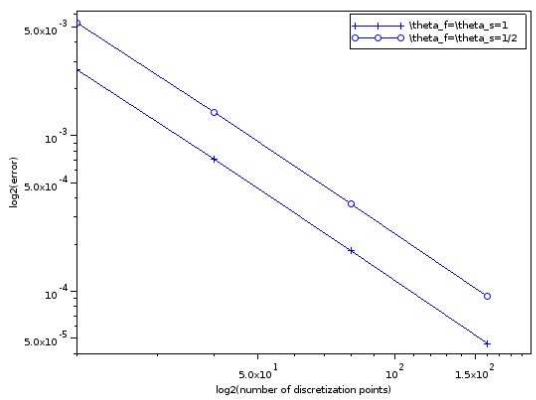

Fig. 3: L^∞-error of the asymptotic numerical and exact states for explicit/explicit and Crank-Nicolson/Crank-Nicolson schemes. The numerical order is 1.94.

From Figure 3, we conclude that the asymptotic state depends only on the spatial discretization through δx and does not depend on the time discretization δt or the values of θ. Moreover, the numerical order is close to 2 in δx.

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 δt (and δx 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 non homogeneous Dirichlet boundary conditions, we were able to prove that the asymptotic numerical solution obtained through a subcycled scheme is a uniform-in-δt second order approximation in δx 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 2.1 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

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

References

[1] Denise Aregba-Driollet, Maya Briani, and Roberto 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] Manuel Calvo, Laurent O. Jay, Juan I. Montijano, and Luis Ràndez. Approximate compositions of a near identity map by multi-revolution runge-kutta methods. *Numer. Math.*, 97:635–666, 2004.

[4] José Antonio Carrillo, Thierry Goudon, and Pauline Lafitte. Simulation of fluid and particles flows: asymptotic preserving schemes for bubbling and flowing regimes. *J. Comput. Phys.*, 227(16):7929–7951, 2008.

[5] José Antonio Carrillo, Thierry Goudon, Pauline Lafitte, and Francesco Vecil. Numerical schemes of diffusion asymptotics and moment closures for kinetic equations. *J. Sci. Comput.*, 36(1):113–149, 2008.

[6] Philippe Chartier, Joseba Makazaga, Ander Murua, and Gilles Vilmart. Multi-revolution composition methods for highly oscillatory differential equations. Preprint, 2013.

[7] William John Trevor Daniel. A study of the stability of subcycling algorithms in structural dynamics. *Computer Methods in Applied Mechanics and Engineering*, 156(14):1 – 13, 1998.

[8] W.J.T. Daniel. A partial velocity approach to subcycling structural dynamics. *Computer Methods in Applied Mechanics and Engineering*, 192:375 – 394, 2003.

[9] Julien Diaz and Marcus J. Grote. Energy conserving explicit local time stepping for second-order wave equations. *SIAM J. Sci. Comput.*, 31(3):1985–2014, 2009.

[10] Pauline Godillon-Lafitte and Thierry Goudon. A coupled model for radiative transfer: Doppler effects, equilibrium, and nonequilibrium diffusion asymptotics. *Multiscale Model. Simul.*, 4(4):1245–1279 (electronic), 2005.

[11] Marcus J. Grote and Teodora Mitkova. Explicit local time-stepping methods for Maxwell's equations. *J. Comput. Appl. Math.*, 234(12):3283–3302, 2010.

[12] Marcus J. Grote and Teodora Mitkova. High-order explicit local time-stepping methods for damped wave equations. *J. Comput. Appl. Math.*, 239:270–289, 2013.

[13] E. Hairer and G. Wanner. *Solving ordinary differential equations II: Stiff and differential-algebraic problems*, volume 2. Springer, 2004.

[14] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J. Sci. Comput.*, 21(2):441–454, 1999.

[15] 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.

[16] 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.

[17] Serge 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.

[18] Roger Temam. Multilevel methods for the simulation of turbulence. A simple model. *J. Comput. Phys.*, 127(2):309–315, 1996.