ON THE LOCAL AND GLOBAL ERRORS OF SPLITTING APPROXIMATIONS OF REACTION-DIFFUSION EQUATIONS WITH HIGH SPATIAL GRADIENTS

STÉPHANE DESCOMBES, THIERRY DUMONT, VIOLAINE LOUVET, AND MARC MASSOT

This paper is dedicated to Michel Crouzeix.

ABSTRACT. In this paper we study the approximation by splitting techniques of the ordinary differential equation $\dot{U} + AU + BU = 0$, $U(0) = U_0$ with A and B two matrices. We assume that we have a stiff problem in the sense that A is ill-conditionned and U_0 is a vector which is the discretization of a function with a very high derivative. This situation may appear for example when we study the discretization of a partial differential equation. We prove some error estimates for two general matrices and in the stiff case, where the estimates are independent of U_0 and the commutator between A and B.

1. INTRODUCTION

Let $\mathcal{M}_n(\mathbb{R})$ be the vector space of matrices of size n, let A and B in $\mathcal{M}_n(\mathbb{R})$ and U_0 in $\mathbb{R}^n$. In this article we study the approximation of the following system of ordinary differential equations :

$$\begin{cases} \dfrac{\partial U}{\partial t} + AU + BU = 0 \quad t > 0, \\ U(0) = U_0, \end{cases} \tag{1.1}$$

by splitting techniques in the situation when A is ill-conditionned and the associated initial condition U_0 results in a very stiff system. The A and B matrices can be thought of as coming from the discretization of a linear partial differential equation where A corresponds to the discretization of the Laplacian operator and the initial condition, U_0, is a vector which is the discretization of a function with a very high derivative. The exact solution of (1.1) is given by $\exp(-t(A+B))U_0$ and is most of the time approximated when t is small enough by Lie or Strang's formula. The Lie formula is given by

$$L(t)U_0 = \exp(-tA)\exp(-tB)U_0, \tag{1.2}$$

which is an approximation of local order 2 in the sense that there exist R_1 independent of t and $R_2(t)$ belonging to $\mathcal{M}_n(\mathbb{R})$ such that for t small

$$e^{-t(A+B)}U_0 - e^{-tA}e^{-tB}U_0 = t^2 R_1 U_0 + t^3 R_2(t) U_0 \tag{1.3}$$

and then is an approximation of global order 1.

The Strang's formula ([12], [13]) is given by

$$S(t)U_0 = \exp(-tA/2)\exp(-tB)\exp(-tA/2)U_0.$$

It is an approximation of local order 3 in the sense that there exist R_3 independent of t and $R_4(t)$ belonging to $\mathcal{M}_n(\mathbb{R})$ such that for t small

$$e^{-t(A+B)}U_0 - e^{-tA/2}e^{-tB}e^{-tA/2}U_0 = t^3\, R_3U_0 + t^4R_4(t)U_0$$

and then is an approximation of global order 2. For these two formulæ R_1 and R_3 are well known and can be for example found in the book of Hairer, Lubich and Wanner [6]. To give the exact value of R_1 and R_3, we have to use the commutator between A and B, denoted by $\partial_A B$, defined by

$$\partial_A B = AB - BA.$$

This provides an elegant way of writting the iterated commutators between A and B, such as :

$$\partial_A^2 B = \partial_A(\partial_A B) = A^2B - 2ABA + B^2A.$$

With this notation we have

$$R_1 = -\frac{1}{2}\partial_A B, \qquad R_3 = -\frac{1}{4}\partial_A^2 B - \frac{1}{2}\partial_B^2 A.$$

Let us assume that A and B are positive definite matrices and come from the discretization of a continuous initial-value problem such as :

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} + V(x)u = 0 & x \in \mathbb{R}, t \in]0,T[, \\ u(x,0) = u_0(x) & x \in \mathbb{R}. \end{cases} \tag{1.4}$$

Let us also assume that the L^2 norm of the derivative of the smooth initial condition u_0 is very high. Let us assume that V is a bounded function from $\mathbb{R}$ to $\mathbb{R}$ of class C^∞ with all bounded derivatives. We consider the operator b corresponding to the multiplication by V and the operator $a = -\partial_x^2$ (minus the second partial derivative with respect to x in one dimension), their commutator is given by

$$\partial_a b = -(\partial_x^2 V) - 2(\partial_x V)\partial_x\;; \tag{1.5}$$

if we assume, at least formally, that we can generalized equation (1.3) to this example which is of infinite dimension, we will have :

$$e^{t(\partial_x^2 - V)}u_0 - e^{t\partial_x^2}e^{-tV}u_0 = -\frac{t^2}{2}(\partial_x^2 V)u_0 - t^2(\partial_x V)(\partial_x u_0) + O(t^3). \tag{1.6}$$

However, the L^2-norm of $\partial_x u_0$ is by assumption very high and the previous asymptotics is only interesting if the splitting time step t is very small, as usually in a numerical analysis study. Nevertheless, from a simulation point of view, the time step can not be chosen too small, and there is a limit associated to the computer precision. Consequently, it is especially relevant in this stiff configuration to obtain alternative error estimates which do not involve the derivative of the initial condition.

In this article we study the associated problem in finite dimension by assuming that $(\partial_A B)U_0$ is very high. The key issue is to obtain a alternative control of the local and global errors when we use an approximation based on Lie or Strang's formula, which involve bounded constants independent of $(\partial_A B)U_0$.

Let us introduce a natural assumption on the commutator of A and B in the discretized case for which the stiffness is similar to the one considered in the continuous case mentionned previously. In formula (1.5), we can see that the commutator can be decomposed in two parts, one which is bounded due to the assumptions on

V and one which behaves like "the square root of a". Since A and B are positive definite matrices, we can define $\sqrt{A}$, the square root of the matrix A. We make the following assumption denoted H1: there exists two constants $C_0 > 0$ and $C_1 > 0$ such that $\partial_A B$ is constituted of two parts $(\partial_A B)_0$ and $(\partial_A B)_1$ such that

$$\|(\partial_A B)_0\|_2 \leq C_0, \quad \|(\partial_A B)_1(\sqrt{A})^{-1}\|_2 \leq C_1.$$

We can notice that this assumption clearly corresponds to the continous case (1.5). We also point out that C_0 and C_1 must be generic constants which do not depend on the matrix A in order to be consistant with (1.5). Such an estimate has already been used for splitting methods for example in [10] or [4].

In this article, for a matrix A, we use the following notation

$$\|A\|_2 = \sqrt{\rho(AA^*)}$$

with $\rho(AA^*)$ the maximum of the absolute values of the eigenvalues of AA^* and we recall that for A and B two matrices

$$\|AB\|_2 = \|BA\|_2.$$

The key point in the article is a formula which provides an exact representation of the local errors and thus yields two complementary results (Corollaries 2.3 and 2.4) for $t > 0$:

$$\left\| e^{-t(A+B)} - L(t) \right\|_2 \leq \frac{t^2}{2} \|\partial_A B\|_2$$

and

$$\left\| e^{-t(A+B)} - L(t) \right\|_2 \leq \frac{C_0\, t^2}{2} + \frac{C_1\, t\sqrt{t}}{3\sqrt{2e}}.$$

The first one is more relevant for very small splitting time steps t. However, as t groves, the second one can be more accurate because of the bounds on the two constants C_0 and C_1 involved, whereas $\|\partial_A B\|_2$ can be very big.

Some similar results with the Strang's formula are obtained. We also present the study of a time discretization of $\exp(-tA)$ by an implicit Euler method and apply these ideas to the example of the discretization of a linear partial differential equation similar to (1.4) but on $[0, 1]$ with Dirichlet boundary conditions and in that case, all the constants are explicit. The paper is organized as follow : in section 1, we prove a formula which provides an exact form of the errors between $\exp(-t(A + B))$ and his Lie approximation and his Strang approximation. In section 2, we derive some bounds of the errors and specially some alternative bounds which are independant of $(\partial_A B)U_0$. In section 3, we focus on the discretization of the linear PDE (3.1) and prove that the considered assumptions are satisfied by the resulting matrices and initial conditions. In section 4, we present the study of the time discretization of $\exp(-tA)$ by an implicit Euler method and characterize its influence on the local error estimate. Section 5 allows us to conclude on the local order reduction and we give some results on global order. Finally, in section 6, we present some numerical results on a nonlinear reaction-diffusion equation which provides a nice numerical illustration of the order reduction related to the strong gradients in the intial solution.

2. Error formulae in the case of a finite dimensional setting

Let A and B belong to $\mathcal{M}_n(\mathbb{R})$, in this section, we compute an exact form of the errors between $\exp(-t(A+B))$ and Lie and Strang's approximations. We first recall the following result given in [4] :

Lemma 2.1. *The following identity holds :*

$$\partial_{e^{-tA}} B = -\int_0^t e^{-(t-s)A}\big(\partial_A B\big)e^{-sA}\,ds. \tag{2.1}$$

This lemma allows us to obtain two error formulae which are give in the following theorem

Theorem 2.2. *Let A and B belong to $\mathcal{M}_n(\mathbb{R})$, the following identities hold :*

$$L(t) = e^{-t(A+B)} + \int_0^t\int_0^s e^{-(t-s)(A+B)}e^{-(s-r)A}\big(\partial_A B\big)e^{-rA}e^{-sB}\,dr\,ds, \tag{2.2}$$

$$\begin{aligned} S(t) = {}& e^{t(A+B)} \\ &+\frac{1}{4}\int_0^t\int_0^s (s-r)e^{-(t-s)(A+B)}e^{-(s-r)A/2}\big(\partial_A^2 B\big)e^{-rA/2}e^{-sB}e^{-sA/2}\,dr\,ds \\ &+\frac{1}{2}\int_0^t\int_0^s (s-r)e^{-(t-s)(A+B)}e^{-sA/2}e^{-rB}\big(\partial_B^2 A\big)e^{-(s-r)B}e^{-sA/2}\,dr\,ds. \end{aligned} \tag{2.3}$$

Proof. Concerning (2.2), we have

$$\begin{aligned} \frac{d}{dt}L(t) &= -Ae^{-tA}e^{-tB} - e^{-tA}Be^{-tB} \\ &= -(A+B)e^{-tA}e^{-tB} - \left(\partial_{e^{-tA}}B\right)e^{-tB}. \end{aligned}$$

Using Duhamel's formula this yields

$$L(t) = e^{-t(A+B)} - \int_0^t e^{-(t-s)(A+B)}\partial_{e^{-sA}}Be^{-sB}\,ds$$

and formula (2.2) can be obtained by using the form of $\partial_{e^{sA}}B$ given in Lemma 2.1. Identity (2.3) has been already proved in [4]. □

A simple corollary of this theorem gives some bounds of the errors in operator norm, in the spirit of the results given in [1].

Corollary 2.3. Let A and B belong to $\mathcal{M}_n(\mathbb{R})$, for $t \geq 0$, we have the following error bounds

$$\left\|e^{-t(A+B)} - L(t)\right\|_2 \leq \frac{t^2}{2}e^{t(\|A\|_2+\|B\|_2)}\|\partial_A B\|_2$$

and

$$\left\|e^{-t(A+B)} - S(t)\right\|_2 \leq \frac{t^2}{24}e^{t(\|A\|_2+\|B\|_2)}\left(2\|\partial_A^2 B\|_2 + \|\partial_B^2 A\|_2\right).$$

If moreover we assume that A and B are positive definite matrices, we have

$$\left\|e^{-t(A+B)} - L(t)\right\|_2 \leq \frac{t^2}{2}\|\partial_A B\|_2$$

and

$$\left\|e^{-t(A+B)} - S(t)\right\|_2 \leq \frac{t^2}{24}\left(2\|\partial_A^2 B\|_2 + \|\partial_B^2 A\|_2\right).$$

Proof. The proof is obvious, we only have to take into account for example that for $t \geq 0$, $\| \exp(-t(A+B))\|_2 \leq \exp(t(\|A\|_2 + \|B\|_2))$ and when A and B are positive definite matrices that $\| \exp(-t(A+B))\|_2 \leq 1$. □

So far we only assume that A and B are positive definite matrices but as we have seen in the introduction we may add some assumptions on the commutator between A and B, like for example assumption H1. In the next corollary we assume that H1 is satisfied and study the influence of this assumption on the error bounds. Since we also study the Strang's formula which involves some iterated commutators, we need an assumption on these iterated commutators. We call H2 the following assumption : there exists four constants $C_i > 0$, $1 \leq i \leq 3$ such that

$$\|\partial_B^2 A\|_2 \leq C_0$$

and such that $\partial_A^2 B$ is constituted of three parts $(\partial_A^2 B)_0$, $(\partial_A^2 B)_1$ and $(\partial_A^2 B)_2$ such that

$$\|(\partial_A^2 B)_0\|_2 \leq C_1,\ \|(\partial_A^2 B)_1(\sqrt{A})^{-1}\|_2 \leq C_2,\ \|(\partial_A^2 B)_2(\sqrt{A})^{-1}A^{-1}\|_2 \leq C_3.$$

Assumption H2 can be a little bit surprising. If we take the example with $a = -\partial_x^2$ and $b = V$ given in the introduction then $\partial_a^2 b$ is a differential operator of order 2 and then behaves like a. It would be natural to assume that $(\partial_A^2 B)_2$ is dominated by A in the sense that $\|(\partial_A^2 B)_2 A^{-1}\|_2$ is bounded and not only dominated by $\sqrt{A}\,A$. But we will see in the next session that the boundary conditions may really influence the power of A which dominates $\partial_A^2 B$. With these assumptions, the following corollary gives much refined results than corollary 2.3 since the bounds do not depend on the commutator and iterated commutators of A and B.

Corollary 2.4. Let A and B belong to $\mathcal{M}_n(\mathbb{R})$ such that A and B are positive definite matrices satisfying H1, for $t \geq 0$, we have the following error bound :

$$\left\| e^{-t(A+B)} - L(t) \right\|_2 \leq \frac{C_0\, t^2}{2} + \frac{C_1\, t\sqrt{t}}{3\sqrt{2e}}. \tag{2.4}$$

If moreover A and B satisfy H2, we have with

$$\alpha = \int_0^1 \frac{(1-u)^{1/4}}{u^{3/4}} du, \tag{2.5}$$

the following estimate

$$\left\| e^{-t(A+B)} - S(t) \right\|_2 \leq \frac{(C_0 + 2C_1)\, t^3}{12} + \frac{C_2\, t^2\sqrt{t}}{15\sqrt{2e}} + \frac{C_3 \alpha\, t\sqrt{t}}{4}. \tag{2.6}$$

Proof. We have

$$\begin{aligned}
\|e^{-t(A+B)} - L(t)\|_2 &\leq \int_0^t \int_0^s \|\big(\partial_A B\big) e^{-rA} e^{-sB}\|_2 \, dr\, ds \\
&\leq \frac{C_0\, t^2}{2} + \int_0^t \int_0^s \|\big(\partial_A B\big)_1 e^{-rA} e^{-sB}\|_2 \, dr\, ds \\
&\leq \frac{C_0\, t^2}{2} \\
&\quad + \int_0^t \int_0^s \|(\partial_A B)_1 (\sqrt{A})^{-1}\|_2 \|\sqrt{A} e^{-rA}\|_2 \|e^{-sB}\|_2 \, dr\, ds \\
&\leq \frac{C_0\, t^2}{2} + C_1 \int_0^t \int_0^s \|\sqrt{A} e^{-rA}\|_2 \, dr\, ds,
\end{aligned}$$

we have for $r > 0$,

$$\|\sqrt{A} e^{-rA}\|_2 \leq \sup_{\lambda \geq 0} \Big(\sqrt{\lambda} e^{-r\lambda}\Big) = \frac{1}{\sqrt{r}} \sup_{\lambda \geq 0} \Big(\sqrt{r\lambda} e^{-r\lambda}\Big) = \frac{1}{\sqrt{2e\, r}}, \tag{2.7}$$

thus

$$\|e^{-t(A+B)} - L(t)\|_2 \leq \frac{C_0\, t^2}{2} + \frac{C_1\, t\sqrt{t}}{3\sqrt{2e}}.$$

(2.6) can be obtained with the same techniques using the relation

$$\|(\partial_A^2 B)_2 (\sqrt{A})^{-1} A^{-1}\|_2 = \|A^{-3/4} (\partial_A^2 B)_2 A^{-3/4}\|_2$$

and is left to the reader. □

It is worth insisting on the fact that estimates (2.4) and (2.6) are given in operator norm thus are independent of U_0 and the second provides a more relevant alternative bound on the error when $\|\partial_A B\|_2$ is very big and t is not small enough.

3. A priori errors on a simple linear problem with low or high gradients

In this section we consider a simple linear partial differential equation in one dimension similar to the one given in the introduction but on a bounded domain. Let $T > 0$ and V a positive function from $[0,1]$ to $\mathbb{R}$ of class C^∞, we consider the following problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} + V(x) u = 0 & x \in]0,1[,\ t \in]0,T[, \\ u(0,t) = u(1,t) = 0 & t \in]0,T[, \\ u(x,0) = u_0(x) & x \in]0,1[. \end{cases} \tag{3.1}$$

We want to solve this problem with finite differences method in space and splitting in time. We introduce n belonging to $\mathbb{N}^*$, $h = 1/(n+1)$ and a subdivision

$$(x_i = ih)_{i \in [0, n+1]}$$

of $[0,1]$. For the sake of simplicity, we denote $U_0 = (u_0(x_1), \cdots, u_0(x_n))^t$, $V_i = V(x_i)$ for $1 \leq i \leq n$ and more generally $V_i^{(j)} = V^{(j)}(x_i)$ the j-th derivative of V at the point x_i for $1 \leq i \leq n$ and $1 \leq j \leq 2$. We also introduce two positive definite matrices, the first is the well known matrix corresponding to the discretization of the Laplace operator, $A = (a_{ij})_{1\le i,j\le n}$ given by

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & & 0 \\ 0 & -1 & 2 & -1 & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & -1 & 2 & -1 \\ 0 & & \dots & & -1 & 2 \end{pmatrix} \tag{3.2}$$

and the second one comes from the discretization of the product Vu

$$B = \begin{pmatrix} V_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & V_2 & 0 & 0 & & 0 \\ 0 & 0 & V_3 & 0 & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & 0 & V_{n-1} & 0 \\ 0 & & \dots & & 0 & V_n \end{pmatrix}. \tag{3.3}$$

With these notations, we now have to find an a priori bound of

$$E_L(t)U_0 = e^{-t(A+B)}U_0 - e^{-tA}e^{-tB}U_0$$

and

$$E_S(t)U_0 = e^{-t(A+B)}U_0 - e^{-tA/2}e^{-tB}e^{-tA/2}U_0.$$

We will do this by using the results of the previous section and the first step is to compute the commutator between A and B. We want to obtain the discrete equivalent of formula (1.5) and it is the purpose of the next lemma. For the sequel, for $0 \le i \le 4$, we denote

$$\kappa_i = \max\{|V^{(i)}(x)|,\, x \in [0,1]\} \tag{3.4}$$

and we introduce the matrix D defined by

$$D = \frac{1}{h}\begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & & 0 \\ 0 & -1 & 1 & 0 & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & -1 & 1 & 0 \\ 0 & \dots & \dots & 0 & -1 & 1 \end{pmatrix}.$$

Lemma 3.1. *There exist two matrices B' and B'' and two diagonal matrices M_2 and M_4 satisfying*

$$\|M_2\|_2 \le \kappa_2$$

and

$$\|M_4\|_2 \le \frac{\kappa_4}{12},$$

such that

$$\partial_A B = -B'' - B'\left(D - D^*\right) + h^2 M_4 + hM_2\left(D - D^*\right).$$

Proof. For W in $\mathbb{R}^n$, we have for $1 \leq i \leq n$

$$\begin{aligned}(ABW)_i - (BAW)_i &= \sum_{k=1}^{n} (a_{ik}V_kW_k - a_{ik}V_iW_k) \\ &= \frac{1}{h^2}\left(V_iW_{i+1} - V_{i-1}W_{i-1} - V_{i+1}W_{i+1} + V_iW_{i+1}\right),\end{aligned}$$

with the convention that $W_0 = W_{n+1} = 0$. More precisely, we have

$$\begin{aligned}(ABW)_i - (BAW)_i =& \frac{1}{h^2}(-V_{i-1} + 2V_i - V_{i+1})W_i \\ &+ \frac{V_{i-1} - V_i}{h}\frac{W_i - W_{i-1}}{h} + \frac{V_{i+1} - V_i}{h}\frac{W_i - W_{i+1}}{h}.\end{aligned}$$

Let us recall the following notations : $V_i' = V'(x_i)$ and $V_i'' = V''(x_i)$ for $1 \leq i \leq n$. Denoting by $C(V^{(2)})$ (resp. $C(V^{(4)})$) a constant depending on the second derivative of V (resp. fourth derivative of V), a simple computation gives

$$\begin{aligned}((\partial_A B)W)_i &= -V_i^{''}W_i - V_i^{'}(DW)_i - V_i^{'}(D^*W)_i \\ &\quad + C(V^{(4)})h^2W_i + C(V^{(2)})h\frac{W_{i+1} - W_{i-1}}{h}.\end{aligned}$$

This formula gives the existence of two matrices B' and B'' and two diagonal matrices M_2 and M_4 satisfying

$$\|M_2\|_2 \leq \kappa_2$$

and

$$\|M_4\|_2 \leq \frac{\kappa_4}{12},$$

such that

$$(\partial_A B)W = -B''W - B'\left(D - D^*\right)W + h^2M_4W + hM_2\left(D - D^*\right)W. \tag{3.5}$$

This concludes the proof of lemma 3.1. □

We now prove that assumptions H1 and H2 are satisified with A and B defined in (3.2) and (3.3).

Theorem 3.2. *The commutator $\partial_A B$ can be shown to be constituted of two parts $(\partial_A B)_0$ and $(\partial_A B)_1$ such that*

$$\|(\partial_A B)_0\|_2 \leq \kappa_2 + \frac{\kappa_4 h^2}{12},\ \|(\partial_A B)_1(\sqrt{A})^{-1}\|_2 \leq 4(\kappa_1 + h\kappa_2).$$

Moreover

$$\|\partial_B^2 A\|_2 \leq 2\kappa_0(\kappa_1 + h\kappa_2)$$

and $\partial_A^2 B$ is constituted of three parts $(\partial_A^2 B)_0$, $(\partial_A^2 B)_1$ and $(\partial_A^2 B)_2$ such that

$$\|(\partial_A^2 B)_0\|_2 \leq\ \kappa_4 + O(h^2),\ \|(\partial_A^2 B)_1(\sqrt{A})^{-1}\|_2 \leq 2\kappa_3 + O(h)$$

and

$$\|(\partial_A^2 B)_2(\sqrt{A})^{-1}A^{-1}\|_2 \leq 4\kappa_1 + \kappa_2 + O(h).$$

Proof. Let

$$(\partial_A B)_0 = -B'' + h^2M_4$$

and

$$(\partial_A B)_1 = -B'\left(D - D^*\right) + hM_2\left(D - D^*\right).$$

It is clear that

$$\|(\partial_A B)_0\|_2 \leq \kappa_2 + \frac{\kappa_4 h^2}{12}.$$

We will now estimate the most difficult term that is to say

$$\left\|(-B^{'}(D-D^*) + hM_2(D-D^*))(\sqrt{A})^{-1}\right\|_2.$$

We first have

$$\left\|(-B^{'}(D-D^*) + hM_2(D-D^*))(\sqrt{A})^{-1}\right\|_2 \leq 4(\kappa_1 + h\kappa_2)\left\|D(\sqrt{A})^{-1}\right\|_2.$$

Since

$$\|D(\sqrt{A})^{-1}\|_2^2 = \rho(D(\sqrt{A})^{-1}(\sqrt{A})^{-1}D^*) = \rho(DD^*A^{-1}),$$

we only have to find a bound of $\rho(DD^*A^{-1})$. But $DD^* = A - R/h^2$ with R the matrix such that $R_{NN} = 1$ and $R_{ij} = 0$ for $(i,j) \neq (N,N)$. Thus

$$DD^*A^{-1} = Id - R\frac{A^{-1}}{h^2}, \tag{3.6}$$

this matrix is a lower triangular matrix with eigenvalues equal to 1 and $1/(N+1)$, we then deduce that

$$\rho(DD^*A^{-1}) = 1$$

and

$$\|(\partial_A B)_1(\sqrt{A})^{-1}\|_2 \leq 4(\kappa_1 + h\kappa_2).$$

The proofs concerning the iterated commutators $\partial_A^2 B$ and $\partial_B^2 A$ can be obtained with the same techniques and tedious computations. We only want to point out that we can not assume that $(\partial_A^2 B)_2$ is dominated by A since it involves $\partial_A(D-D^*)$ which is a diagonal matrix with only two terms different from zero and equal to $2/h^3$. □

Remark 3.3. To make a comparison with the continuous case, we can see that the matrix $D - D^*$ corresponds approximately to the derivation. In fact formula (3.6) shows that, because of boundary conditions, D is not exactly the square root of A.

Thank's to Lemma 3.1 and Theorem 3.2 we have seen that assumptions H1 and H2 are satisfied and the constants are explicit. We can now obtain the estimates on E_L and E_S.

Theorem 3.4. *Let $C_{L0} = \kappa_2 + \kappa_4 h^2/12$, $C_{L1} = 4(\kappa_1 + h\kappa_2)$, $C_{S0} = 2\kappa_0(\kappa_1 + h\kappa_2)$ and three terms which are constant up to a negligeable quantity, $C_{S1} = \kappa_4 + O(h^2)$, $C^{S2} = 2\kappa_3 + O(h)$ and $C_{S3} = 4\kappa_1 + \kappa_2 + O(h)$. For $t \geq 0$, the following estimates hold :*

$$\|E_L(t)U_0\|_2 \leq \left(C_{L0}\|U_0\|_2 + C_{L1}\|\sqrt{A}U_0\|_2\right)\frac{t^2}{2}, \tag{3.7}$$

$$\|E_L(t)U_0\|_2 \leq \left(\frac{C_{L0}\,t^2}{2} + \frac{C_{L1}\,t\sqrt{t}}{3\sqrt{2e}}\right)\|U_0\|_2 \tag{3.8}$$

and

$$\|E_S(t)U_0\|_2 \leq \left((2C_{S0} + C_{S1})\|U_0\|_2 + C_{S2}\|\sqrt{A}U_0\|_2 + C_{S3}\|A\sqrt{A}U_0\|_2\right)\frac{t^3}{24}, \tag{3.9}$$

(3.10)

$$\|E_S(t)U_0\|_2 \leq \frac{(C_{S0} + 2C_{S1})\,t^3}{12} + \frac{C_{S2}\,t^2\sqrt{t}}{15\sqrt{2e}} + \frac{C_{S3}\alpha\,t\sqrt{t}}{4},$$

α being defined in (2.5).

Proof. As in the proof of corollary 2.4, we have

$$\begin{aligned}\|E_L(t)U_0\|_2 &\leq \int_0^t \int_0^s \|\big(\partial_A B\big)e^{-rA}e^{-sB}U_0\|_2 \, dr\, ds \\ &\leq \frac{C_{L0}\, t^2}{2}\|U_0\|_2 + \int_0^t \int_0^s \|\big(\partial_A B\big)_1 e^{-rA}e^{-sB}U_0\|_2 \, dr\, ds \\ &\leq \frac{C_{L0}\, t^2}{2}\|U_0\|_2 \\ &\quad + C_{L1}\int_0^t \int_0^s \|e^{-rA}\sqrt{A}e^{-sB}U_0\|_2 \, dr\, ds.\end{aligned}$$

But we have

$$\begin{aligned}\int_0^t \int_0^s \|e^{-rA}\sqrt{A}e^{-sB}U_0\|_2 \, dr\, ds &\leq \int_0^t \int_0^s \|e^{-rA}\|_2\|\sqrt{A}e^{-sB}(\sqrt{A})^{-1}\sqrt{A}U_0\|_2 \, dr\, ds \\ &\leq \int_0^t \int_0^s \|\sqrt{A}e^{-sB}(\sqrt{A})^{-1}\|_2\|\sqrt{A}U_0\|_2 \, dr\, ds \\ &\leq \int_0^t \int_0^s \|e^{-sB}\|_2\|\sqrt{A}U_0\|_2 \, dr\, ds \\ &\leq \frac{t^2\, \|\sqrt{A}U_0\|_2}{2}.\end{aligned}$$

Thus

$$\|E_L(t)U_0\|_2 \leq t^2 \left(\frac{\kappa_2 + \kappa_4 h^2/12}{2}\|U_0\|_2 + 2(\kappa_1 + h\kappa_2)\ \|\sqrt{A}U_0\|_2 \right),$$

and this yields (3.7). Estimate (3.8) is just a consequence of (2.4), estimate (3.9) is left to the reader, since it is completely similar to the proof of (3.7). Finally (3.10) is a direct consequence of (2.6). □

4. On the study of a time discretization

Most of the time the exponential of A can not be computed exactly and we have to take into account a time discretization. In the sequel we come back to the general case of two matrices, A and B are now not coming from the discretization of equation (3.1). We will focus on the Lie formula and we will see how the approximation of the exponential of A will change the local error. Denoting by $\mathbf{1}$ the identity matrix, we appoximate $\exp(-tA)$ by $(\mathbf{1}+tA)^{-1}$ that we assume well defined. This approximation is the standard implicit Euler method and we denote the approximation by $r(tA)$. We then consider $r(tA)\exp(-tB)$ and we assume that $\exp(-tB)$ can be computed exactly as in the previous section.

Theorem 4.1. *Let A and B belong to $\mathcal{M}_n(\mathbb{R})$, we have :*

$$\begin{aligned}r(tA)\, e^{-tB} - e^{-t(A+B)} &= -\int_0^t s\, e^{-(t-s)(A+B)} r(sA)\ (\partial_A B)\ r(sA) e^{-sB}\, ds \\ &\quad + \int_0^t s\, e^{-(t-s)(A+B)} A^2 r(sA)^2\, e^{-sB}\, ds.\end{aligned} \tag{4.1}$$

Proof. We have

$$\begin{aligned}\frac{d}{dt}r(tA)\,e^{-tB} &= \left(\frac{d}{dt}r(tA)\right)e^{-tB} - r(tA)\,Be^{-tB}\\ &= -Ar(tA)^2\,e^{-tB} - \left(\partial_{r(tA)}B\right)e^{-tB} - Br(tA)\,e^{-tB}\\ &= -(A+B)r(tA)\,e^{-tB} - \left(\partial_{r(tA)}B\right)e^{-tB} + Ar(tA)(\mathbf{1}-r(tA))e^{-tB}.\end{aligned}$$

Define $R(t) = -\left(\partial_{r(tA)}B\right)e^{-tB} + Ar(tA)(\mathbf{1}-r(tA))e^{-tB}$. Using Duhamel's formula we have

$$r(tA)\,e^{-tB} = e^{-t(A+B)} + \int_0^t e^{-(t-s)(A+B)}R(s)\,ds.$$

Estimate (4.1) is now a consequence of the two formulae

$$\partial_{r(tA)}B = t\,r(tA)\,(\partial_A B)\,r(tA)$$

and

$$\mathbf{1} - r(tA) = t\,A\,r(tA).$$

□

Corollary 4.2. Let A and B belong to $\mathcal{M}_n(\mathbb{R})$ such that A and B are positive definite matrices, for $t \geq 0$, we have the following error bounds :

$$\|r(tA)\,e^{-tB} - e^{-t(A+B)}\|_2 \leq \frac{t^2}{3}\left(\|\partial_A B\|_2 + \|A^2\|_2\right). \tag{4.2}$$

If moreover we assume that assumption H1 is satisfied, we have

$$\|r(tA)\,e^{-tB} - e^{-t(A+B)}\|_2 \leq \frac{C_0 t^2}{2} + \frac{C_1 t\sqrt{t}}{3} + t\|A\|_2. \tag{4.3}$$

Proof. As in the proof of corollary 2.4, noticing that for $t \geq 0$ $\|r(tA)\|_2 \leq 1$, we have

$$\|r(tA)\,e^{-tB} - e^{-t(A+B)}\|_2 \leq \frac{t^2}{2}\|\partial_A B\|_2 + \frac{t^2}{2}\|A^2\|_2.$$

Concerning estimate (4.3) we have

$$\begin{aligned}\|r(tA)\,e^{-tB} - e^{-t(A+B)}\|_2 \leq C_0\int_0^t s ds + C_1\int_0^t \sqrt{s}\|\sqrt{sA}\,r(sA)\|_2\,ds\\ + \int_0^t \|A\|_2\|sA\,r(sA)\|_2\,ds.\end{aligned}$$

We have for $s > 0$,

$$\|\sqrt{sA}\,r(sA)\|_2 \leq \sup_{\lambda\geq 0}\left(\sqrt{\lambda}r(\lambda)\right) = \frac{1}{2}$$

and

$$\|sA\,r(sA)\|_2 \leq \sup_{\lambda\geq 0}(\lambda r(\lambda)) = 1,$$

thus

$$\|r(tA)\,e^{-tB} - e^{-t(A+B)}\|_2 \leq \frac{C_0 t^2}{2} + \frac{C_1 t\sqrt{t}}{3} + t\|A\|_2.$$

This concludes the proof of corollary 4.2. □

It is worth insisting on the fact that when the exponential is discretized, we can not obtain an error bound which is independent of $\|A\|_2$. This will play a crucial role in the next section. Let us define for $t \geq 0$,

$$E_{\text{dis}}(t)U_0 = e^{-t(A+B)}U_0 - r(tA)e^{-tB}U_0.$$

We come back to the example given in the previous section with A and B given in (3.2) and (3.3).

Theorem 4.3. *The following estimates hold :*

$$\|R_{\text{dis}}(t)U_0\|_2 \leq \frac{t^2}{2}\left(C_{L0}\,\|U_0\|_2 + C_{L1}\,\|\sqrt{A}U_0\|_2 + \|A^2U_0\|_2\right) \tag{4.4}$$

and

$$\|R_{\text{dis}}(t)U_0\|_2 \leq \left(\frac{C_{L0}\,t^2}{2} + \frac{C_{L1}\,t\sqrt{t}}{3\sqrt{2e}}\right)\|U_0\|_2 + t\|AU_0\|_2. \tag{4.5}$$

Proof. As in the proof of corollary 4.2, we have

$$\begin{aligned}
\|R_{\text{dis}}(t)U_0\|_2 &\leq \int_0^t s\|r(sA)\,(\partial_A B)\,r(sA)\,e^{-sB}U_0\|_2\,ds \\
&\quad + \int_0^t s\|A^2\,r(sA)^2\,e^{-sB}U_0\|_2\,ds \\
&\leq \int_0^t s\|\,(\partial_A B)\,r(sA)\,e^{-sB}U_0\|_2\,ds + \int_0^t s\|A^2e^{-sB}U_0\|_2\,ds \\
&\leq \frac{C_{L0}\,t^2}{2}\|U_0\|_2 + \int_0^t s\|\big(\partial_A B\big)_1 r(sA)\,e^{-sB}U_0\|_2\,dr\,ds \\
&\quad + \frac{t^2}{2}\|A^2U_0\|_2, \\
&\leq \frac{C_{L0}\,t^2}{2}\|U_0\|_2 + \frac{C_{L1}\,t^2}{2}\|\sqrt{A}U_0\|_2 + \frac{t^2}{2}\|A^2U_0\|_2,
\end{aligned}$$

and this yields (4.4). Estimate (4.5) is a consequence of (4.3). □

So far we have given some estimates to control the local errors and a natural question is the influence on the global behaviour, this is the subject of the next section.

5. Some conclusions on the local and global errors for equations with high spatial gradients

In this section we would like to focus on the behaviour of the errors when we have a stiff problem, for example when U_0 is a vector having components equal to some values of a function with a very high derivative. We have seen in the previous sections that an order reduction may appear in the local error and we address the following question : is there an influence on the behaviour of global errors ? We infer from Theorem 3.4 that for $t \geq 0$,

$$\|E_L(t)U_0\|_2 \leq \max\left(\frac{C_{L0}\|U_0\|_2 + C_{L1}\|\sqrt{A}U_0\|_2}{2}t^2, \left(\frac{C_{L0}\,t^2}{2} + \frac{C_{L1}\,t\sqrt{t}}{3\sqrt{2e}}\right)\|U_0\|_2\right)$$

and a similar estimate for $\|E_S(t)U_0\|_2$. This estimate provides a very good control of the local error, the first term being more relevant when t is small and the second when t is not small enough and $\|\sqrt{A}U_0\|_2$ is very high. More precisely there exists an explicit constant $\theta > 0$ depending on $\|\sqrt{A}U_0\|_2$ such that for $t \leq \theta$, $\|E_L(t)U_0\|_2$ behaves like t^2 and for $t \geq \theta$, $\|E_L(t)U_0\|_2$ behaves like $t\sqrt{t}$. This explains why an order reduction may appear if we have very high spatial gradients, since

$$\|DU_0\|_2 \leq \|D(\sqrt{A})^{-1}\sqrt{A}U_0\|_2 \leq \|D(\sqrt{A})^{-1}\|_2\|\sqrt{A}U_0\|_2.$$

We will now show that in a linear framework with two positive definite matrices, a local order reduction has no influence on the global order and only help us to provide some good a priori estimates. If we try to decompose the global error, we can write for example that the difference $(\exp(-tA)\exp(-tB))^n - \exp(-nt(A+B))$ can be split as follows :

$$\begin{aligned}&\left(e^{-tA}e^{-tB}\right)^n U_0 - e^{-nt(A+B)}U_0\\ &= \sum_{j=1}^{n-1}\left(e^{-tA}e^{-tB}\right)^{n-j-1} E_L(t)e^{-jt(A+B)}U_0 + (e^{-tA}e^{-tB})^{n-1}E_L(t)U_0.\end{aligned} \tag{5.1}$$

The worse term that we can have in this formula is $t^2\|\sqrt{A}e^{-jt(A+B)}U_0\|_2$ and when A is a positive definite matrice we may have

$$t^2\|\sqrt{A}e^{-jt(A+B)}U_0\|_2 \leq C\frac{t^2}{\sqrt{jt}}\|U_0\|_2,$$

this gives finally a global order equal to 1. Unfortunately such a computation will not give, for example, an order one for the approximation of e^{-tA} by an implicit Euler method since we have in that case the term $t\,A^2U_0$. It is clear that this method which works also in a nonlinear case and has been used in [2], fails and does not prove that there is no global order reduction. For this, we have to use the results of Ichinose and all [8], [9], who proved that when A and B are two positive definite matrices then the Lie formula even if we use an implicit Euler approximation is still of order 1 and the Strang's formula is of order 2. As we will see numerically in the next section for the Strang's formula, the situation is very different when we work in a nonlinear framework.

6. Numerical Simulations - KPP equation

Following the theoretical investigations in the linear case we have presented, we focus in this section on the numerical evidence of the order reduction associated to Strang splitting in a typical nonlinear framework of stiff travelling waves. Let us first recall the Kolmogorov-Petrovskii-Piskunov model. In their original paper dated in 1937 [11], these authors introduced a model describing the propagation of a virus and the first rigourous analysis of a stable travelling wave solution of a nonlinear reaction-diffusion equation [5, 14]. The equation is the following :

$$\partial_t\beta - D\,\Delta\beta = k\,\beta^2(1-\beta). \tag{6.1}$$

The non-dimensionalization process and the structure of the exact solution can be found in [5]. In the case of $D = 1$ and $k = 1$, the velocity of the self-similar travelling wave is $c = 1/\sqrt{2}$ and the maximal gradient value reaches $1/\sqrt{32}$. The structure of the wave can be observed in Figure 1 with a discretization of 5000 points of the interval $[-70, 70]$ and a time varying in $[0, 15]$ divided into eight time intervals.

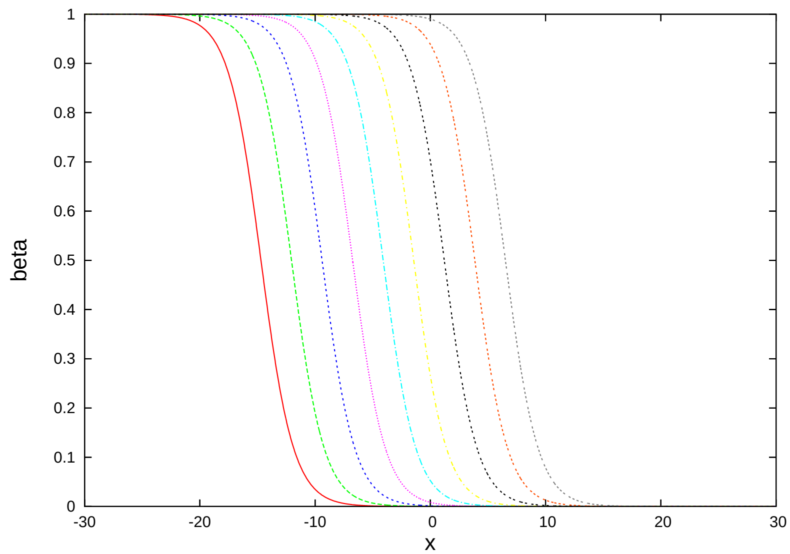

FIGURE 1. Standard KPP travelling wave, discretization with 5000 pts with a zoom on the $[-30, 30]$ region

The one point we want to insist on, is the fact that the velocity of the travelling wave is proportional to $(k\,D)^{1/2}$, whereas the maximal gradient is proportional to $(k/D)^{1/2}$. Thus switching to values $k = 10.0$ and $D = 0.1$, the velocity is preserved, but the maximal gradient is multiplied by a factor of 10 and introduces stiffness in the equation, as presented in Figure 2. Let us underline that, for the discretizations considered in the following, the wave, however "stiff", is always well resolved on the grid.

In order to perform a numerical illustration of the order reduction of the operator splitting technique, we need, on the one side the "quasi-exact" temporal integration of the semi-discretized problem 6.1, and on the other side the Strang splitting on the same equation, where each of the three substeps are also integrated by a "quasi-exact" solver in time, so that the only error coming from the splitting technique can be characterized. This "quasi-exact" resolution in time is conducted using the method of lines with a finite difference second order discretization in space; we use a BDF method [7] with a very fine both relative and absolute tolerance of 10^{-12}. We start from an already self-similar solution and present the ability of the splitted solution to reproduce the correct propagation speed and self-similar profile.

However, one could infer that there is no need to compute a "quasi-exact" solution since an analytical solution is easily available [5]. In fact, the spatial discretization needed in order to reach an error difference with respect to the exact solution, that does not interfere with the splitting error itself, is much too high for the parametric studies we envision (40000 points for an l^2 error below 10^{-7}, a treshold for limited interferences). Consequently, given a spatial discretization, the good way to proceed is to compare the "quasi-exact" solution of 6.1 and the splitted solution with "quasi-exact" temporal integration of each substeps. In such a way, we decouple the errors originating in the splitting itself, from the errors coming

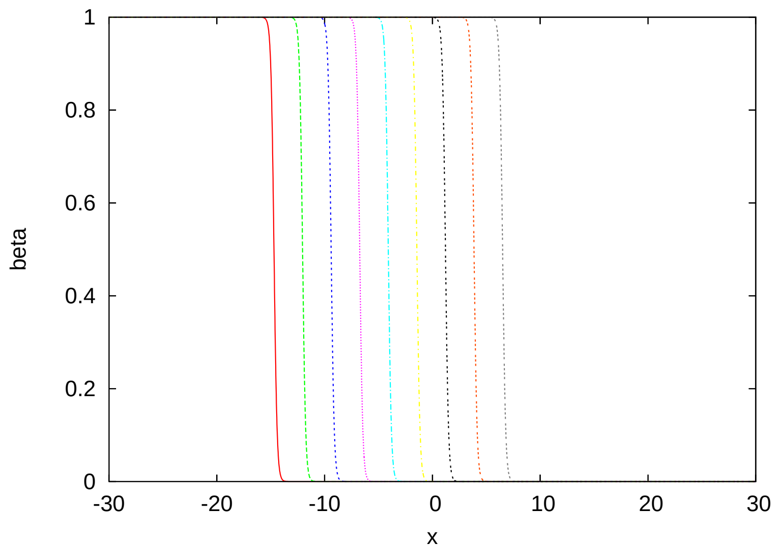

FIGURE 2. "Stiff" KPP travelling wave, discretization with 8000 pts

from the temporal integration of the substeps and their coupling with the splitting error, and it is coherent with the theoretical study conducted in the paper.

We firstly observe the usual second order of the Strang splitting applied to the standard KPP equation in Figure 3, with splitting time steps ranging from 15/1024 up to 15/16. We have plotted a straight line of slope 2 in order to show the very good agreement of the global order with the local order predicted by the theory.

The key point in the paper is the order reduction exhibited by the second case, where the velocity of the wave is the same, but the maximal gradient is ten times higher. Let us insist on the fact that this loss of accuracy is not related to the spatial disretization, as proved by the various points obtained with two different discretizations, 2000 and 8000, in Figure 4. There is a treshold, beyond which the second order predicted by the classical theory is not observed any more and we switch to an order that is closer to 1.5 as predicted by the theory. When the splitting time step is getting too big (i.e. beyond 1), the velocity of the travelling wave as well as the wave profile are not resolved any more. However, there is still a range of splitting time steps for which the order 3/2 can be nicely observed.

REFERENCES

1. Sergio Blanes and Fernando Casas, *On the convergence and optimization of the Baker-Campbell-Hausdorff formula*, Linear Algebra Appl. **378** (2004), 135–158. MR MR2031789 (2004k:17008)
2. Stéphane Descombes, *Convergence of a splitting method of high order for reaction-diffusion systems*, Math. Comp. **70** (2001), no. 236, 1481–1501 (electronic). MR MR1836914 (2002c:65152)
3. Stéphane Descombes and Marc Massot, *Operator splitting for nonlinear reaction-diffusion systems with an entropic structure: singular perturbation and order reduction*, Numer. Math. **97** (2004), no. 4, 667–698. MR MR2127928 (2006i:65131)
4. Stéphane Descombes and Michelle Schatzman, *Strang's formula for holomorphic semi-groups*, J. Math. Pures Appl. (9) **81** (2002), no. 1, 93–114. MR MR1994884

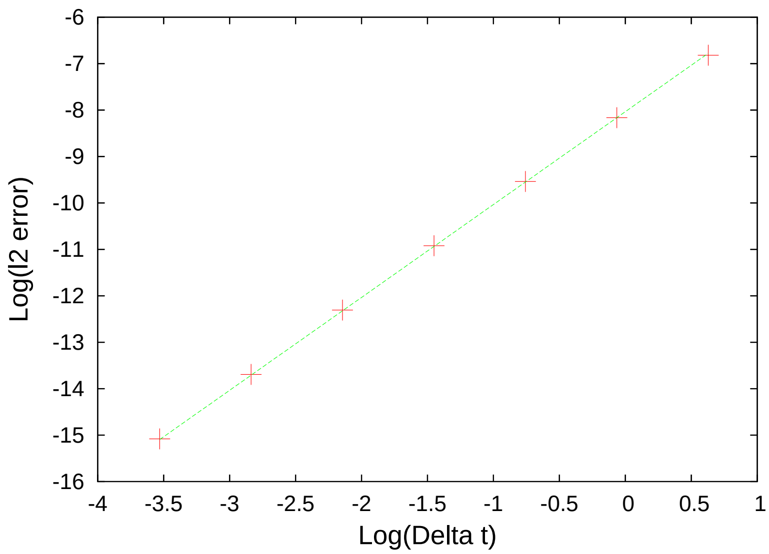

FIGURE 3. Standard KPP travelling wave, l^2 global error at time $t = 1.875$

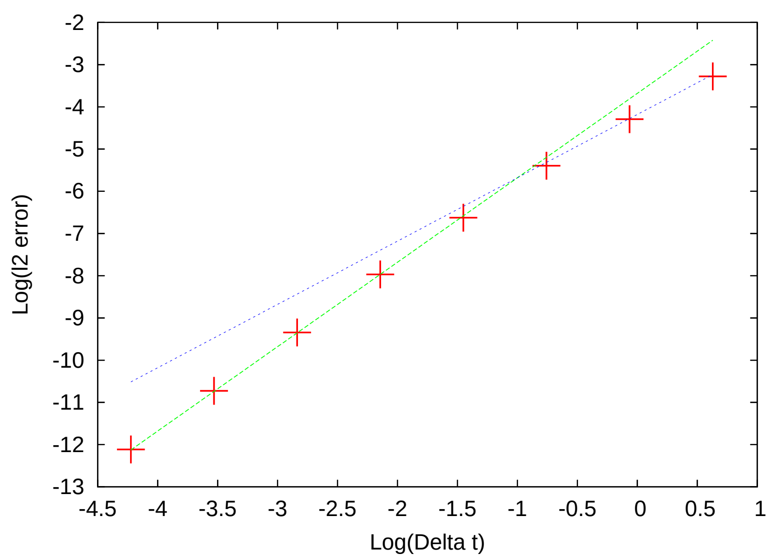

FIGURE 4. "Stiff" KPP travelling wave, l^2 global error at time $t = 1.875$ for a spatial discrtization with 2000 and 8000 points; the two lines indicate slopes of 2 and 1.5

5. P. Gray and S. K. Scott, *Chemical oscillations and instabilites*, Oxford University Press, 1994.
6. Ernst Hairer, Christian Lubich, and Gerhard Wanner, *Geometric numerical integration*, second ed., Springer Series in Computational Mathematics, vol. 31, Springer-Verlag, Berlin,

2006, Structure-preserving algorithms for ordinary differential equations. MR MR2221614 (2006m:65006)

7. A.C. Hindmarsh, *Odepack, a systematized collection of ode solvers, in scientific computing*, North-Holland, Amsterdam, 1983.
8. Takashi Ichinose and Hideo Tamura, *The norm convergence of the Trotter-Kato product formula with error bound*, Comm. Math. Phys. **217** (2001), no. 3, 489–502. MR MR1822104 (2002e:47048a)
9. Takashi Ichinose, Hideo Tamura, Hiroshi Tamura, and Valentin A. Zagrebnov, *Note on the paper: "The norm convergence of the Trotter-Kato product formula with error bound" by T. Ichinose and H. Tamura [Comm. Math. Phys.* **217** *(2001), no. 3, 489–502; MR1822104 (2002e:47048a)]*, Comm. Math. Phys. **221** (2001), no. 3, 499–510. MR MR1852050 (2002e:47048b)
10. Tobias Jahnke and Christian Lubich, *Error bounds for exponential operator splittings*, BIT **40** (2000), no. 4, 735–744. MR MR1799313 (2001k:65143)
11. A.N. Kolmogoroff, I.G. Petrovsky, and N.S. Piscounoff, *Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application a un problème biologique*, Bulletin de l'Université d'état Moscou, Série Internationale Section A Mathématiques et Mécanique **1** (1937), 1–25.
12. Gilbert Strang, *Accurate partial difference methods. I. Linear Cauchy problems*, Arch. Rational Mech. Anal. **12** (1963), 392–402. MR 26 #4489
13. ———, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal. **5** (1968), 506–517. MR MR0235754 (38 #4057)
14. Aizik I. Volpert, Vitaly A. Volpert, and Vladimir A. Volpert, *Traveling wave solutions of parabolic systems*, American Mathematical Society, Providence, RI, 1994, Translated from the Russian manuscript by James F. Heyda.

Unité de Mathématiques Pures et Appliquées - CNRS UMR 5669 - Ecole Normale Supérieure de Lyon, 46, Allée d'Italie, 69364 LYON Cedex 07 France

E-mail address: `sdescomb@umpa.ens-lyon.fr`

Institut Camille Jordan Universit Claude Bernard Lyon 1 43 boulevard du 11 novembre 1918 69622 Villeurbanne cedex France

E-mail address: `tdumont@math.univ-lyon1.fr, louvet@math.univ-lyon1.fr`

Laboratoire EM2C - UPR CNRS 288, Ecole Centrale Paris, Grande Voie des Vignes, 92295 Chatenay-Malabry Cedex France

E-mail address: `Marc.Massot@em2c.ecp.fr`