

An explicit-implicit splitting method for a convection-diffusion problem

V. Thomée* and A. S. Vasudeva Murthy†

April 16, 2018

Dedicated to Amiya Pani on the occasion of his 60th birthday

Abstract

We analyze a second order accurate finite difference method for a spatially periodic convection-diffusion problem. The method is a time stepping method based on the Strang splitting of the spatially semidiscrete solution, in which the diffusion part uses the Crank-Nicolson method and the convection part the explicit forward Euler approximation on a shorter time interval. When the diffusion coefficient is small, the forward Euler method may be used also for the diffusion term.

1 Introduction

In this paper we shall consider the numerical solution of the convection-diffusion problem in the cube $\Omega = (0, 2\pi)^d$,

$$(1.1) \quad \frac{\partial U}{\partial t} = \operatorname{div}(a\nabla U) + b \cdot \nabla U, \quad \text{in } \Omega, \quad \text{for } t > 0, \quad \text{with } U(0) = V,$$

with periodic boundary conditions, where the positive definite $d \times d$ matrix $a(x) = (a_{ij}(x))$ and $b = b(x) = (b_1, \dots, b_d)$ are periodic and smooth.

*Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Gothenburg, Sweden; E-mail: thomee@chalmers.se

†TIFR Centre for Appl. Mathematics, Yelahanka New Town, Bangalore 560 065, India. E-mail: vasu@math.tifrbng.res.in

The equation (1.1) is a special case of the initial-value problem for the operator equation

$$(1.2) \quad \frac{dU}{dt} = -\mathcal{A}U + \mathcal{B}U, \quad \text{for } t \geq 0, \quad \text{with } U(0) = V,$$

where \mathcal{A} and \mathcal{B} represent different physical processes, in our case $\mathcal{A}U = -\text{div}(a\nabla U)$, $\mathcal{B}U = b \cdot \nabla U$. The solution of (1.2) may be formally expressed as

$$U(t) = \mathcal{E}(t)V = e^{-t(\mathcal{A}-\mathcal{B})}V, \quad \text{for } t \geq 0.$$

In some sense, in (1.1) \mathcal{A} and \mathcal{B} may be thought of as representing slow and fast physical processes. To discretize such an equation in time a common approach is to split $\mathcal{A}-\mathcal{B}$ into \mathcal{A} and $-\mathcal{B}$. With k a time step one introduces $t_n = nk$ and then, on each time interval (t_{n-1}, t_n) one may use the second order symmetric Strang splitting [8], [7],

$$(1.3) \quad \mathcal{E}(k) = e^{-k(\mathcal{A}-\mathcal{B})} \approx e^{\frac{1}{2}k\mathcal{B}} e^{-k\mathcal{A}} e^{\frac{1}{2}k\mathcal{B}},$$

which thus locally involves solutions of $U_t = -\mathcal{A}U$ and $U_t = \mathcal{B}U$, see e.g. Hundsdorfer and Verwer [5] and references therein.

The three exponentials on the right in (1.3) are then approximated by rational functions of \mathcal{A} and \mathcal{B} , respectively, such as the Crank-Nicolson method $r_0(k\mathcal{A})$, with $r_0(\lambda) = (1 - \frac{1}{2}\lambda)/(1 + \frac{1}{2}\lambda)$, for the middle factor. The choice of approximation involving \mathcal{B} is not so obvious. It has been suggested in the context of numerical weather prediction, e.g. in Baldauf [1], that time steps of different length could be used for the the two processes, with shorter time intervals for the fast process and longer for the slow one. Also Gassmann and Herzog [3] discuss the difficulties associated with splitting in such situations. In the case of reaction-diffusion equation, see Estep et al [2] and references therein. Our aim in this work is to discuss this problem in a somewhat rigorous fashion for our simple model problem.

We note that if \mathcal{A} and \mathcal{B} commute, which for (1.1) holds when a and b are independent of x , then $e^{-k(\mathcal{A}-\mathcal{B})} = e^{-k\mathcal{A}}e^{k\mathcal{B}} = e^{\frac{1}{2}k\mathcal{B}}e^{-k\mathcal{A}}e^{\frac{1}{2}k\mathcal{B}}$ so that the error in (1.3) is zero. When \mathcal{A} and \mathcal{B} do not commute, then formally, by Taylor expansion, $e^{-k(\mathcal{A}-\mathcal{B})} - e^{-k\mathcal{A}}e^{k\mathcal{B}} = O(k^2)$ and $e^{-k(\mathcal{A}-\mathcal{B})} - e^{\frac{1}{2}k\mathcal{B}}e^{-k\mathcal{A}}e^{\frac{1}{2}k\mathcal{B}} = O(k^3)$. Error estimates for the time splitting, depending on the regularity of the initial values may be found in Jahnke and Lubich [6], Hansen and Ostermann [4] and references therein.

In the method that we study in this paper, we begin by discretizing (1.1) in the spatial variables. We let $h = 2\pi/M$, where M is a positive integer, and define a corresponding uniform mesh

$$(1.4) \quad \Omega_h = \{x = x_j = jh; j = (j_1, \dots, j_d), 1 \leq j_l \leq M, l = 1, \dots, d\}.$$

For M -periodic vectors u with elements u_j , corresponding to the mesh-point x_j , and with $u_{j+Me_l} = u_j$, we consider the simple second order finite difference approximation of (1.1),

$$(1.5) \quad \frac{du}{dt} = \sum_{i,j=1}^d \bar{\partial}_j(\acute{a}_{ij}\partial_i u) + \sum_{j=1}^d \frac{1}{2}b_j(\partial_j + \bar{\partial}_j)u. \quad \text{in } \Omega_h, t > 0, \text{ with } u(0) = v.$$

Here ∂_j and $\bar{\partial}_j$ are forward and backward finite difference quotients in the direction of x_j , $\acute{a}_{ij}(x_l) = a_{ij}(x_l + \frac{1}{2}he_i)$, and v the restriction of V to Ω_h . The problem (1.5) may be written as a system of ODEs in time,

$$(1.6) \quad \frac{du}{dt} = -Au + Bu, \quad \text{for } t > 0, \quad \text{with } u(0) = v,$$

where the $M^d \times M^d$ matrices A and B correspond to the differential operators \mathcal{A} and \mathcal{B} . It is then to (1.6) that we will apply the splitting approach.

In the 1D case we may take, with $h = 2\pi/M$ the mesh-width and $x_l = lh$,

$$A = \frac{1}{h^2} \begin{bmatrix} d(x_1) & -a(x_{1.5}) & 0 & \dots & -a(x_{0.5}) \\ -a(x_{1.5}) & d(x_2) & -a(x_{2.5}) & \dots & 0 \\ 0 & -a(x_{2.5}) & d(x_3) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & -a(x_{M-0.5}) \\ -a(x_{0.5}) & 0 & 0 & \dots - a(x_{M-0.5}) & d(x_M) \end{bmatrix},$$

where $d(x_l) = a(x_l + 0.5h) + a(x_l - 0.5h)$ (recall $a(x_{M+0.5}) = a(x_{0.5})$), and

$$B = \frac{1}{2h} \begin{bmatrix} 0 & b(x_1) & \dots & -b(x_1) \\ -b(x_2) & 0 & b(x_2) \dots & 0 \\ \vdots & \vdots & \ddots & b(x_{M-1}) \\ b(x_M) & 0 & \dots - b(x_M) & 0 \end{bmatrix}.$$

The solution of (1.6), the spatially semidiscrete solution, is

$$u(t) = E(t)v = e^{-t(A-B)}v,$$

and we shall see that the error in this approximation is $O(h^2)$, under the appropriate regularity assumptions. For the time discretization we shall work with basic time intervals of length $k = 1/N$, where $N = 2p$ is an even positive integer, then apply the Strang splitting (1.3) on each of the interval, so that

$$(1.7) \quad E(k) = e^{-k(A-B)} \approx e^{\frac{1}{2}kB} e^{-kA} e^{\frac{1}{2}kB},$$

and finally approximate the three exponential factors.

We would like the time discretization error to match that of the discretization in space. For a second order time discretization method, this will require $k = O(h)$. For $k \leq \gamma h^2$, with γ appropriate, the problem may be solved by explicit approximations but since we prefer N to be relatively small, we will consider methods with h and k of the same order.

For the approximation of e^{-kA} in (1.7) we shall use the Crank-Nicolson method. Then, in order to approximate $e^{(1/2)kB}$ we would like to use the forward Euler method on a time interval of length $k_1 < k$. Assuming for the moment that b is constant and thus B skew-symmetric, we note that

$$\|I + k_1 B\| = (1 + k_1^2 \|B\|^2)^{1/2} \leq 1 + C k_1^2 h^{-2}.$$

Here and below, $\|\cdot\|$ denotes the standard matrix norm induced by the ℓ^2 vector inner product. Stability therefore holds if $k_1^2 h^{-2} \leq C k_1$, or if $k_1 \leq C h^2$. Since k should be of the same order as h , this makes it natural to choose $k_1 = k^2$. We thus subdivide the time intervals of length k into N subintervals of length $k^2 = k/N$ and apply an explicit forward Euler approximation on each of these. As we shall see, this approximation matches the second order of the Crank-Nicolson method.

Thus the diffusion part of the equation is approximated on intervals of length k and the convection part on intervals of length k^2 . We consider thus the time discrete solution at time $t_n = nk$,

$$(1.8) \quad u^n = E_k^n v, \quad \text{where } E_k = \mathcal{B}_k r_0(kA) \mathcal{B}_k, \quad \text{with } \mathcal{B}_k = (I + k^2 B)^p.$$

In fact, in the successive time stepping, only the matrix $\tilde{\mathcal{B}}_k = \mathcal{B}_k^2 = (I + k^2 B)^N$ is used, except in the first and last steps. The method proposed thus replaces the solution at each time step of a nonsymmetric problem, by the solution of a symmetric problem, plus applications of an explicit method, but successively repeated $N/2 = p$ times before and after the diffusion approximation, thus

covering the interval of length $Nk^2 = k$. We re-emphasize that the splitting is applied to the spatially semidiscrete problem and not to the continuous problem.

The analysis sketched above is carried out in Section 2 below. The analysis will use discrete Sobolev norms. After this, in Section 3, we discuss the case when the equation (1.1) contains a small diffusion coefficient ε . In this case we are able to show that if $\varepsilon \leq \gamma h$ with γ sufficiently small, then the approximation of $e^{-k\varepsilon A}$ can be done by the forward Euler method, and, with the convection part as before, we have a purely explicit second order approximation method. We close the paper by presenting some numerical illustrations in Section 4.

2 Basic Error Analysis

For the periodic problem in $\Omega = (0, 2\pi)^d$ and with Ω_h as in (1.4), we introduce the discrete inner product and norm

$$(2.1) \quad (v, w)_h = h^d \sum_{x_j \in \Omega_h} v_j w_j \quad \text{and} \quad \|v\|_h = (v, v)_h^{1/2}.$$

Further, we set $\partial_j u(x) = (u(x + he_j) - u(x))/h$ and $\partial^\alpha u = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d} u$ for $\alpha = (\alpha_1, \dots, \alpha_d)$, and define the discrete Sobolev norm, with $|\alpha| = \alpha_1 + \dots + \alpha_d$,

$$(2.2) \quad \|u\|_{h,s} = \left(\sum_{|\alpha| \leq s} \|\partial^\alpha u\|_h^2 \right)^{\frac{1}{2}}, \quad \text{for } s \geq 0.$$

We shall also use $\bar{\partial}_j u(x) = (u(x) - u(x - he_j))/h$. We define

$$\|w\|_{\mathbb{C}^s(R)} = \max_{|\alpha| \leq s} \sup_{x \in R} |D^\alpha w(x)|, \quad D^\alpha = \left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \dots \left(\frac{\partial}{\partial x_d} \right)^{\alpha_d}.$$

We shall write \mathbb{C}^s for $\mathbb{C}^s(\Omega)$, and \mathbb{C} for \mathbb{C}^0 . We note that defining U_h to be the restriction to the mesh Ω_h of a smooth function U , i.e. by $(U_h)_j = U(x_j)$, we have $\|U_h\|_{h,s} \leq C \|U\|_{\mathbb{C}^s}$.

Consider now the matrices A and B in our convection-diffusion problem (1.2). They satisfy, with C independent of h ,

$$(2.3) \quad \|Au\|_h \leq C \|u\|_{h,2} \quad \text{and} \quad \|Bu\|_h \leq C \|u\|_{h,1}.$$

Setting $Q(x) = \{y : |y_s - x_s| \leq h, s = 1, \dots, d\}$, we have, since the terms in AU_h are symmetric difference quotients of U at the meshpoints x_j , and the terms in $(\mathcal{A}U)_h$ are the corresponding derivatives of U at x_j ,

$$(2.4) \quad |(AU_h)(x_j) - (\mathcal{A}U)(x_j)| \leq Ch^2 \|u\|_{\mathbb{C}^4(Q(x))},$$

and

$$(2.5) \quad |(BU_h)(x_j) - (\mathcal{B}U)(x_j)| \leq Ch^2 \|u\|_{\mathbb{C}^3(Q(x))},$$

expressing, in particular, that (1.5) is a second order approximation of (1.1).

We note that, for $|\alpha| = s$,

$$(2.6) \quad \|\partial^\alpha Au - A\partial^\alpha u\|_h \leq C\|u\|_{h,s+1} \quad \text{and} \quad \|\partial^\alpha Bu - B\partial^\alpha u\|_h \leq C\|u\|_{h,s},$$

and we may conclude from (2.3) that

$$(2.7) \quad \|Au\|_{h,s} \leq C\|u\|_{h,s+2} \quad \text{and} \quad \|Bu\|_{h,s} \leq C\|u\|_{h,s+1}.$$

The matrix A is positive semidefinite, with

$$(2.8) \quad \|v\|_{h,1}^2 \leq C((Av, v)_h + \|v\|_h^2).$$

Further, it follows easily from the definition of A ,

$$(2.9) \quad \|Av\|_h \leq \alpha h^{-2} \|v\|_h, \quad \text{where } \alpha = 4d\lambda_{\max}(a).$$

For $B = (b_{il})$ we have

$$b_{il} = \begin{cases} \pm b_j(x_i), & \text{if } l = i \pm e_j, j = 1, \dots, d, \\ 0, & \text{for other } l, \end{cases}$$

and that then $b_{i+e_j, i} = -b_j(x_{i+e_j})$. It follows from this that

$$B = B_0 + B_1, \quad \text{with } B_0 = \frac{1}{2}(B - B^T), \quad B_1 = \frac{1}{2}(B + B^T).$$

Here B_0 is skew-symmetric and

$$(2.10) \quad \|B_0\| \leq h^{-1}\beta_0, \quad \beta_0 = \sum_{j=1}^d \|b_j\|_{\mathbb{C}}, \quad \|B_1\| \leq \beta_1 = \sum_{j=1}^d \left\| \frac{\partial b_j}{\partial x_j} \right\|_{\mathbb{C}}.$$

Note also that $(B_0 v, v) = 0$ for all v .

We begin with the straightforward standard analysis of the spatially semidiscrete problem (1.6), which we include for completeness. We first show the stability of the solution operator of (1.6) in discrete Sobolev norms.

Lemma 2.1. *Let $E(t) = e^{-t(A-B)}$. Then, for any $s \geq 0$, with c_s independent of h ,*

$$\|E(t)v\|_{h,s} \leq e^{c_s t} \|v\|_{h,s}, \quad \text{for } t \geq 0.$$

Proof. Let $s \geq 0$ and let $|\alpha| = s$. From (1.6) we find, for $u(t) = E(t)v$,

$$(2.11) \quad (\partial^\alpha u_t, \partial^\alpha u)_h + (\partial^\alpha Au, \partial^\alpha u)_h = (\partial^\alpha Bu, \partial^\alpha u)_h.$$

Here, by (2.6),

$$(2.12) \quad |(\partial^\alpha Au, \partial^\alpha u)_h - (A\partial^\alpha u, \partial^\alpha u)_h| \leq C\|u\|_{h,s+1}\|u\|_{h,s}.$$

Further, since $B = B_0 + B_1$ with B_0 skew-symmetric,

$$|(\partial^\alpha Bu, \partial^\alpha u)_h| \leq |(B\partial^\alpha u, \partial^\alpha u)_h| + C\|u\|_{h,s} \leq C\|u\|_{h,s}^2.$$

Therefore, by (2.11),

$$\frac{1}{2} \frac{d}{dt} \|\partial^\alpha u\|_h^2 + (A\partial^\alpha u, \partial^\alpha u)_h + \|\partial^\alpha u\|_h^2 \leq C\|u\|_{h,s+1}\|u\|_{h,s}.$$

Hence, using (2.8) and summing over $|\alpha| \leq s$ we find, with $c > 0$,

$$\frac{d}{dt} \|u\|_{h,s}^2 + c\|u\|_{h,s+1}^2 \leq C\|u\|_{h,s+1}\|u\|_{h,s} \leq c\|u\|_{h,s+1}^2 + C\|u\|_{h,s}^2,$$

or

$$(2.13) \quad \frac{d}{dt} \|u\|_{h,s}^2 \leq C\|u\|_{h,s}^2,$$

from which the lemma follows. □

Note that the special case of e^{-tA} is included for $B = 0$.

As a consequence, we have the following second order error estimate.

Theorem 2.1. *We have, for the solutions of (1.6) and (1.1), with $v = V_h$, and $C = C_T$,*

$$(2.14) \quad \|u(t) - U_h(t)\|_h \leq Ch^2 \|V\|_{\mathbb{C}^4}, \quad \text{for } nk \leq T < \infty.$$

Proof. Setting $\omega = u - U_h$, we find

$$\begin{aligned} \frac{d\omega}{dt} &= -A\omega + B\omega + \rho, \quad \text{in } \Omega_h, \quad \text{for } t > 0, \quad \text{with } \omega(0) = 0, \\ \text{where } \rho &= ((\mathcal{A}U)_h - AU_h) - ((\mathcal{B}U_h) - BU_h). \end{aligned}$$

Here, by (2.4) and (2.5),

$$\|\rho(t)\|_h \leq \|((\mathcal{A}U)_h - AU_h)(t)\|_h + \|((\mathcal{B}U_h) - BU_h)(t)\|_h \leq Ch^2 \|U(t)\|_{\mathbb{C}^4},$$

and hence

$$\omega(t) = \int_0^t E(t-s)\rho(s) ds, \quad \text{where } \|\rho(s)\|_h \leq Ch^2 \|U(s)\|_{\mathbb{C}^4}.$$

Hence, by Lemma 2.1, since $\|U(s)\|_{\mathbb{C}^4} \leq C\|V\|_{\mathbb{C}^4}$,

$$\|\omega(t)\|_h \leq C \int_0^t \|\rho(s)\|_h ds \leq Ch^2 \int_0^t \|U(s)\|_{\mathbb{C}^4} ds \leq Ch^2 \|V\|_{\mathbb{C}^4}. \quad \square$$

Turning to the analysis of the time discretization we first show the stability of e^{tB} .

Lemma 2.2. *For any $s \geq 0$, we have, with c_s independent of h ,*

$$(2.15) \quad \|e^{tB}v\|_{h,s} \leq e^{c_s t} \|v\|_{h,s}, \quad \text{for } t \geq 0.$$

Here we may choose $c_0 = \beta_1$, as in 2.10.

Proof. Let $s \geq 0$ and let $|\alpha| = s$. Then for the solution of (1.6) with $A = 0$,

$$(\partial^\alpha u_t, \partial^\alpha u)_h = (\partial^\alpha Bu, \partial^\alpha u)_h = (B\partial^\alpha u, \partial^\alpha u)_h + Q, \quad \text{where } Q \leq C_s \|u\|_{h,s}^2.$$

Since $(Bu, u)_h = (B_1u, u)_h \leq \beta_1 \|u\|_h^2$, we conclude that (2.13) holds, which shows (2.15). For $s = 0$ we have $Q = 0$ and hence (2.13) holds with $C = 2\beta_1$ which implies (2.15), with $c_0 = \beta_1$. \square

We now show the following error estimate for the Strang splitting.

Lemma 2.3. *We have, with C independent of h and k ,*

$$\|e^{-k(A-B)}v - e^{\frac{1}{2}kB}e^{-kA}e^{\frac{1}{2}kB}v\|_h \leq Ck^3 \|v\|_{h,6}.$$

Proof. Setting $F(k) = e^{-k(A-B)} - e^{\frac{1}{2}kB}e^{-kA}e^{\frac{1}{2}kB}$ and noting that $F(0) = F'(0) = F''(0) = 0$, we may use Taylor's formula to obtain

$$\|F(k)\| = \|(F(k) - F(0) - kF'(0) - \frac{1}{2}k^2F''(0))v\|_h \leq \frac{1}{6}k^3 \sup_{s \leq k} \|F'''(s)v\|_h.$$

Here, for $s \leq k$, using (2.3) and Lemmas 2.1 and 2.2,

$$\begin{aligned} \|F'''(s)v\|_h &\leq \|(A-B)^3 e^{-k(A-B)}v\|_h + C \sum_{i_1+i_2+i_3=3} \|B^{i_1} e^{\frac{1}{2}sB} A^{i_2} e^{-sA} B^{i_3} e^{\frac{1}{2}sB} v\|_h \\ &\leq C(\|v\|_{h,6} + \sum_{i_1+i_2+i_3=3} \|v\|_{h,i_1+2i_2+i_3}) \leq C\|v\|_{h,6}, \end{aligned}$$

which shows the lemma. \square

We now turn to the time stepping operator E_k defined in (1.8) and begin with the following stability result.

Lemma 2.4. *Let $k \leq \gamma h$. Then, with β_0, β_1 as in (2.10),*

$$(2.16) \quad \|\mathcal{B}_k\| = \|(I + k^2 B)^p\| \leq e^{\frac{1}{2}\mu k}, \quad \text{where } \mu = \frac{1}{2}(\gamma\beta_0)^2 + \beta_1.$$

Further, $E_k = \mathcal{B}_k r_0(kA) \mathcal{B}_k$ is stable, and

$$(2.17) \quad \|E_k^n\| \leq e^{\mu T}, \quad \text{for } nk \leq T.$$

Proof. Since B_0 is skew symmetric, we have

$$\|I + k^2 B_0\|^2 = 1 + k^4 \|B_0\|^2 \leq 1 + k^4 h^{-2} \beta_0^2 \leq 1 + (\gamma\beta_0)^2 k^2 \leq e^{(\gamma\beta_0)^2 k^2}.$$

Hence

$$(2.18) \quad \|I + k^2 B\| \leq \|I + k^2 B_0\| + k^2 \|B_1\| \leq e^{\frac{1}{2}(\gamma\beta_0)^2 k^2} + \beta_1 k^2 \leq e^{\mu k^2},$$

which shows (2.16) since $2pk = 1$. Since $\|r_0(kA)\| \leq 1$, (2.17) follows. \square

We start the analysis of the time discretization error with the following.

Lemma 2.5. *Let M be a square matrix, and assume $\|e^{sM}\| \leq C$ for $s \leq t$. Then we have, for $s \leq t$,*

$$(2.19) \quad \|(e^{tM} - (I + tM))v\|_h \leq Ct^2 \|M^2 v\|_h,$$

and

$$(2.20) \quad \|(e^{tM} - (I + tM + \frac{1}{2}t^2M^2))v\|_h \leq Ct^3\|M^3v\|_h.$$

If also $\|(I + \frac{1}{2}sM)^{-1}\| \leq C$ for $s \leq t$, then

$$(2.21) \quad \|(e^{-tM} - r_0(tM))v\|_h \leq Ct^3(\|M^3v\|_h + \|v\|_h).$$

Proof. By Taylor expansion we have

$$e^{tM} = I + tM + \int_0^t (t-s)e^{Ms}M^2ds,$$

and hence

$$\|(e^{tM} - (I + tM))v\|_h \leq C \int_0^t (t-s)\|M^2v\|_h ds = \frac{1}{2}Ct^2\|M^2v\|_h.$$

The estimate (2.20) follows analogously. For (2.21) we use (2.20) together with

$$r_0(tM) = I - tM + \frac{1}{2}t^2M^2 + \frac{1}{2} \int_0^t (t-s)^2 r_0'''(sM) ds,$$

where $r_0'''(\lambda) = -\frac{3}{2}(1 + \frac{1}{2}\lambda)^{-4}$, to complete the proof. \square

We now show the following error estimate for \mathcal{B}_k .

Lemma 2.6. *Let $k \leq \gamma h$. Then we have, with $C = C_\gamma$,*

$$\|(e^{\frac{1}{2}k^2B} - \mathcal{B}_k)v\|_h \leq Ck^3\|v\|_{h,2}.$$

Proof. Since $pk^2 = \frac{1}{2}k$ we may write

$$e^{\frac{1}{2}k^2B}v - (1 + k^2B)^p v = \sum_{j=0}^{p-1} e^{(p-j-1)k^2B} (e^{k^2B} - (I + k^2B))(I + k^2B)^j v.$$

By Lemma 2.5, we have

$$(2.22) \quad \|(e^{k^2B} - (1 + k^2B))v\|_h \leq Ck^4\|B^2v\|_h.$$

By Lemma 2.2, $\|e^{(p-j-1)k^2B}\| \leq e^{\frac{1}{2}k\beta_1}$ for $j \leq p-1$. Using also (2.18) we find

$$\begin{aligned} \|e^{\frac{1}{2}k^2B}v - (1 + k^2B)^p v\|_h &\leq Ck^4 \sum_{j=0}^{p-1} \|B^2(I + k^2B)^j v\|_h \\ &\leq Ck^4 \sum_{j=0}^{p-1} e^{\mu j k^2} \|B^2 v\|_h \leq Ck^4 p e^{\text{frac}12\mu k} \|B^2 v\|_h \leq Ck^3 \|v\|_{h,2} \end{aligned}$$

which completes the proof. \square

We now show the following error estimate for E_k .

Lemma 2.7. $k \leq \gamma h$. Then we have, with $C = C_\gamma$.

$$\|E(k)v - E_k v\|_h \leq Ck^3 \|v\|_{h,6},$$

Proof. In view of Lemma 2.3, it remains to show

$$\|e^{\frac{1}{2}k^2B} e^{-kA} e^{\frac{1}{2}k^2B} v - E_k v\|_h \leq Ck^3 \|v\|_{h,6}.$$

We have

$$\begin{aligned} e^{\frac{1}{2}k^2B} e^{-kA} e^{\frac{1}{2}k^2B} - \mathcal{B}_k r_0(kA) \mathcal{B}_k &= (e^{\frac{1}{2}k^2B} - \mathcal{B}_k) e^{-kA} e^{\frac{1}{2}k^2B} \\ &+ \mathcal{B}_k (e^{-kA} - r_0(kA)) e^{\frac{1}{2}k^2B} + \mathcal{B}_k r_0(kA) (e^{\frac{1}{2}k^2B} - \mathcal{B}_k) = J_1 + J_2 + J_3. \end{aligned}$$

Here by the above lemmas,

$$\begin{aligned} \|J_1 v\|_h &\leq Ck^3 \|e^{-kA} e^{\frac{1}{2}k^2B} v\|_{h,2} \leq Ck^3 \|v\|_{h,2}, \quad \|J_3 v\|_h \leq Ck^3 \|v\|_{h,2}, \\ \|J_2 v\|_h &\leq Ck^3 (\|A^3 e^{\frac{1}{2}k^2B} v\|_h + \|e^{\frac{1}{2}k^2B} v\|_h) \leq Ck^3 \|e^{\frac{1}{2}k^2B} v\|_{h,6} \leq Ck^3 \|v\|_{h,6} \end{aligned}$$

which completes the proof \square

We can now prove the following error estimate:

Theorem 2.2. Let $k \leq \gamma h$. Then we have for the solutions of (1.8) and (1.6), with $C = C_{\gamma,T}$,

$$\|u^n - u(nk)\|_h \leq Ck^2 \|v\|_{h,6}, \quad \text{for } nk \leq T.$$

Proof. We write

$$(2.23) \quad u^n - u(nk) = E_k^n v - E(nk)v = \sum_{j=0}^{n-1} E_k^{n-1-j} (E_k - E(k)) E(jk).$$

Using Lemmas 2.4, 2.7 and 2.1 we obtain, for $nk \leq T$,

$$\begin{aligned} \|E_k^n v - E(nk)v\|_h &\leq C \sum_{j=0}^{n-1} \|(E_k - E(k))E(jk)v\|_h \\ &\leq Ck^3 \sum_{j=0}^{n-1} \|E(jk)v\|_{h,6} \leq Cnk^3 \|v\|_{h,6} \leq Ck^2 \|v\|_{h,6}. \quad \square \end{aligned}$$

Since $\|V_h\|_{h,6} \leq C\|V\|_{\mathbb{C}^6}$, we immediately obtain from Theorems 2.1 and 2.2 the following total error estimate.

Theorem 2.3. *Let $v = V_h$ and $k \leq \gamma h$. Then we have for the solutions of (1.8) and (1.1), with $C = C_{\gamma,T}$,*

$$\|u^n - U_h(nk)\|_h \leq Ch^2 \|V\|_{\mathbb{C}^6}, \quad \text{for } nk \leq T.$$

3 The Case of a Small Diffusion Coefficient

In this section, we consider the variant of the problem (1.1) with a small diffusion coefficient $\varepsilon > 0$,

$$(3.1) \quad \frac{\partial U}{\partial t} = \varepsilon \operatorname{div}(a \nabla U) + b \cdot \nabla U, \quad \text{in } \Omega, \quad \text{for } t > 0, \quad \text{with } U(0) = V.$$

The corresponding semidiscrete system (1.6) may then be written

$$(3.2) \quad \frac{du}{dt} = -\varepsilon Au + Bu, \quad \text{for } t > 0, \quad \text{with } u(0) = v$$

where A and B are as before. We shall see that (3.2) is stable, and satisfies a $O(h^2)$ error estimate, independently of ε . Further, for ε and k small, or more precisely, if $\max(\varepsilon, k)/\leq \gamma h$ and $k\varepsilon \leq 2h^2/\alpha$, we will be able to show a $O(k^2)$ estimate for the time discretization error, even when we use the less

accurate forward Euler method for the A part of the time stepping operator, and with weaker regularity requirements than earlier. Also, we do not need to use the symmetric Strang splitting, and consider now, with $r_1(\lambda) = 1 - \lambda$,

$$U^n = \tilde{E}_k^n v, \quad \text{where } \tilde{E}_k = r_1(k\varepsilon A)\tilde{\mathcal{B}}_k, \quad \text{where } \tilde{\mathcal{B}}_k = \mathcal{B}_k^2 = (I + k^2 B)^m.$$

We note the inverse inequality $h\|u\|_{h,s} \leq C\|u\|_{h,s-1}$, and hence

$$(3.3) \quad \varepsilon\|Au\|_{h,s} \leq C\|u\|_{h,s+1}, \quad \text{for } \varepsilon \leq \gamma h, \quad \text{if } \gamma > 0.$$

As in Section 2 we first attend to the spatially semidiscrete problem.

Lemma 3.1. *Let $\tilde{E}(t) = e^{-t(\varepsilon A - B)}$. Then, for any $s \geq 0$, we have, with c_s independent of $\varepsilon > 0$ and $h > 0$,*

$$\|\tilde{E}(t)v\|_{h,s} \leq e^{c_s t} \|v\|_{h,s}, \quad \text{for } t \geq 0.$$

Proof. Following the steps in the proof of Lemma 2.1 we have, for $u(t) = \tilde{E}(t)v$ and $|\alpha| = s$,

$$(3.4) \quad (\partial^\alpha u_t, \partial^\alpha u)_h + \varepsilon(\partial^\alpha Au, \partial^\alpha u)_h = (\partial^\alpha Bu, \partial^\alpha u)_h.$$

Here, as in the proof of Lemma 2.1, $|(\partial^\alpha Bu, \partial^\alpha u)_h| \leq C\|u\|_{h,s}^2$. Hence, by (3.4), and using also (2.12),

$$\frac{1}{2} \frac{d}{dt} \|\partial^\alpha u\|_h^2 + \varepsilon(A\partial^\alpha u, \partial^\alpha u)_h + \varepsilon\|\partial^\alpha u\|_h^2 \leq C\varepsilon\|u\|_{h,s+1}\|u\|_{h,s} + C\|u\|_{h,s}^2.$$

and thus, by (2.8), with $c > 0$,

$$\frac{d}{dt} \|u\|_{h,s}^2 + \varepsilon c \|u\|_{h,s+1}^2 \leq \varepsilon c \|u\|_{h,s+1}^2 + C\|u\|_{h,s}^2.$$

This implies (2.13), with C independent of ε and h , and thus completes the proof. \square

In the same way as in Section 2 the stability shows the following error estimate.

Theorem 3.1. *We have, for the solutions of (3.2) and (1.1), with $C = C_T$ independent of ε ,*

$$(3.5) \quad \|u(t) - U_h(t)\|_h \leq Ch^2 \|V\|_{\mathbb{C}^4}, \quad \text{for } nk \leq T.$$

In the analysis of the time discretization we begin with the analogue of Lemma 2.3.

Lemma 3.2. *We have, with $C = C_\gamma$ independent of h and k ,*

$$\|E(k)v - e^{-k\varepsilon A} e^{kB}v\|_h \leq C(\varepsilon k^2 + k^3)\|v\|_{h,3}, \quad \text{for } \varepsilon \leq \gamma h.$$

Proof. With $F(k) = e^{-k(\varepsilon A - B)} - e^{-k\varepsilon A} e^{kB}$, we have $F(0) = F'(0) = 0$ and hence, by Taylor's formula,

$$\|F(k)v\|_h = \|(F(k) - F(0) - kF'(0))v\|_h \leq \frac{1}{2}k^2 \sup_{s \leq k} \|F''(s)v\|_h.$$

Here

$$\begin{aligned} F''(s) &= e^{-s(\varepsilon A - B)}(\varepsilon A - B)^2 - e^{-s\varepsilon A}(\varepsilon^2 A^2 - 2\varepsilon AB + B^2)e^{sB} \\ &= e^{-s(\varepsilon A - B)}(\varepsilon^2 A^2 - \varepsilon AB - \varepsilon BA) - e^{-s\varepsilon A}(\varepsilon^2 A^2 - 2\varepsilon AB)e^{sB} \\ &\quad + (e^{-s(\varepsilon A - B)} - e^{-s\varepsilon A}e^{sB})B^2 = G_1(s) + G_2(s) + G_3(s). \end{aligned}$$

Using (3.3), (2.7) and the boundedness of the exponentials, we find

$$\|G_1(s)v + G_2(s)v\|_h \leq C\varepsilon\|v\|_{h,3}, \quad \text{for } s \leq k.$$

Further, $G_3(0) = 0$ and hence

$$\|G_3(s)v\|_h \leq s \sup_{\sigma \leq s} \|G_3'(\sigma)v\|_h \leq Ck\|v\|_{h,3}, \quad \text{for } s \leq k.$$

Together these estimates complete the proof of the lemma. \square

We now turn to the time stepping operator \tilde{E}_k and begin with its stability.

Lemma 3.3. *If $k\varepsilon \leq 2h^2/\alpha$, then*

$$(3.6) \quad \|r_1(k\varepsilon A)\| = \|I - k\varepsilon A\| \leq 1.$$

If also $k \leq \gamma h$, then \tilde{E}_k is stable, or, with μ as in Lemma 2.4,

$$\|\tilde{E}_k^n\| \leq e^{\mu T}, \quad \text{for } nk \leq T.$$

Proof. We note that, since A is positive semidefinite,

$$\|I - k\varepsilon A\| \leq 1 \quad \text{if } k\varepsilon\|A\| \leq 2,$$

and thus (3.6) holds by (2.9). Hence, by Lemma 2.4,

$$\|\tilde{E}_k\| \leq \|r_1(kA)\| \|\mathcal{B}_k\|^2 \leq e^{\mu k}. \quad \square$$

We now turn to the error analysis and show the following.

Lemma 3.4. *If $\max(\varepsilon, k) \leq \gamma h$ and $k\varepsilon \leq 2h^2/\alpha$ we have*

$$\|E(k)v - \tilde{E}_k v\|_h \leq C(\varepsilon k^2 + k^3)\|v\|_{h,3}.$$

Proof. In view of Lemma 3.2 it remains to show

$$\|e^{-k\varepsilon A} e^{kB} v - \tilde{E}_k v\|_h \leq C(\varepsilon k^2 + k^3)\|v\|_{h,3}$$

We first note that by Lemma 2.5,

$$(3.7) \quad \|(e^{-k\varepsilon A} - r_1(k\varepsilon A))v\|_h \leq C\varepsilon^2 k^2 \|A^2 v\|_h \leq C\varepsilon k^2 \|v\|_{h,3}.$$

We write

$$e^{-k\varepsilon A} e^{kB} - \tilde{E}_k = (e^{-k\varepsilon A} - r_1(k\varepsilon A))e^{kB} + r_1(k\varepsilon A)(e^{kB} - \tilde{B}_k) = J_1 + J_2.$$

Here by (3.7) and Lemma 2.5, and by (3.6) and Lemma 2.4,

$$\|J_1 v\|_h \leq C\varepsilon k^2 \|e^{kB} v\|_{h,3} \leq C\varepsilon k^2 \|v\|_{h,3} \quad \text{and} \quad \|J_2 v\|_h \leq Ck^3 \|v\|_{h,2},$$

which completes the proof \square

The following is the resulting error estimate.

Theorem 3.2. *If $\max(\varepsilon, k) \leq \gamma h$ and $k\varepsilon \leq 2h^2/\alpha$ we have, with $C = C_{\gamma, T}$ independent of h, k and ε .*

$$\|u^n - u(nk)\|_h \leq C(\varepsilon k + k^2)\|v\|_{h,3}, \quad \text{for } nk \leq T.$$

Proof. Using again (2.23) we find

$$\begin{aligned} \|\tilde{E}_k^n v - E(nk)v\|_h &\leq C \sum_{j=0}^{n-1} \|(\tilde{E}_k - E(k))E(jk)v\|_h \\ &\leq C(\varepsilon k^2 + k^3) \sum_{j=0}^{n-1} \|E(jk)v\|_{h,3} \leq CT(\varepsilon k + k^2)\|v\|_{h,3}. \end{aligned} \quad \square$$

As in Section 2, our error estimates in Theorems 3.1 and 3.2 together show a total error estimate.

Theorem 3.3. *With $v = V_h$ and for $\max(\varepsilon, k) \leq \gamma h$ and $k\varepsilon \leq 2h^2/\alpha$ we have, with $C = C_{\gamma, T}$ independent of h, k and ε .*

$$\|u^n - U_h(nk)\|_h \leq Ch^2 \|V\|_{\mathbb{C}^4}, \quad \text{for } nk \leq T.$$

4 Numerical Illustrations

In this section we present some numerical computations to illustrate our error estimates. We restrict ourselves to the one-dimensional version of (1.1),

$$(4.1) \quad U_t = (a U_x)_x + b U_x, \quad \text{for } x \in (0, 2\pi), \quad t > 0, \quad \text{with } U(x, 0) = \sin x.$$

As before we shall choose $h = 2\pi/M$, $k = 1/N$, where M and N are positive integers, and study the effect of doubling these integers.

We begin with the simple case $a = b = 1$, in which case the exact solution is $U(x, t) = e^{-t} \sin(x + t)$. In Table 1 we compile the errors in the numerical solution at $t = 1$, first in the spatial discretization, then when our time stepping method is applied to the semidiscrete solution, and finally the total error. We use $N = 4, 8, 16, 32, 64$, and $M = 5N$, so that $k/h = 5/(2\pi) = 0.7958$. The successive ratios of the total errors are given in the last column and confirm the second order convergence estimates resulting from Theorems 2.1, 2.2 and 2.3.

Table 1					
$h = 2\pi/M, k = 1/N$					
M	N	$\ (u - U_h)(\cdot, 1)\ _h$	$\ u^N - u(\cdot, 1)\ _h$	$\ u^N - U_h(\cdot, 1)\ _h$	Ratio
20	4	0.01670	0.01199	0.02494	
40	8	0.00423	0.00300	0.00621	4.01
80	16	0.00106	0.00075	0.00155	4.01
160	32	0.00027	0.00019	0.00039	3.97
320	64	0.00007	0.00005	0.00010	3.90

We recall that in the case that a and b are constant, the matrices A and B involved in our method commute, and consequently the splitting error given in Lemma 2.2 vanishes. In order to also consider a situation when this does not happen, we let $a(x) = 1 + \frac{1}{2} \cos x$ and $b(x) = 1 + \frac{1}{2} \sin x$. To indicate that the matrices A and B do not commute in this case we consider the corresponding continuous operators

$$\mathcal{A}U = -\left((1 + \frac{1}{2} \cos x) U_x\right)_x, \quad \mathcal{B}U = (1 + \frac{1}{2} \sin x) U_x.$$

and find, after some effort,

$$\begin{aligned} (\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A})U &= -(a(bU_x)_x)_x + b(aU_x)_{xx} \\ &= \left(1 - \frac{1}{4}(2 - \sin x)/(2 + \cos x)\right) U_x - \left(\frac{1}{2} + \frac{1}{2} \sin x - \frac{1}{4} \sin^2 x + \cos x\right) U_{xx}. \end{aligned}$$

Thus \mathcal{A} and \mathcal{B} do not commute, and therefore neither could A and B . The exact solution U is taken to be the semidiscrete solution with $M = 2560, N = 512$. The errors are presented in Table 2. Again we see that the errors are of second order, which agrees with the error bounds in Theorems 2.1, 2.2 and 2.3

Table 2					
$h = 2\pi/M, k = 1/N$					
M	N	$\ (u - U_h)(\cdot, 1)\ _h$	$\ u^N - u(\cdot, 1)\ _h$	$\ u^N - U_h(\cdot, 1)\ _h$	Ratio
20	4	0.02641	0.01419	0.03323	
40	8	0.00651	0.00356	0.00817	4.07
80	16	0.00163	0.00089	0.00203	4.02
160	32	0.00041	0.00022	0.00051	3.98
320	64	0.00010	0.00005	0.00013	3.92

We finally consider a numerical example for Section 3, for which we use (4.1) with $a = \varepsilon = 0.01, b = 1$. Here $U(x, t) = e^{-\varepsilon t} \sin(x + t)$, and $u^n = \tilde{E}_k^n v$ with $\tilde{E}_k = r_1(kA)\tilde{\mathcal{B}}_k$. Note that the condition $\varepsilon k \leq 2h^2/\alpha$, with $\alpha = 4$, now reduces to $\varepsilon \leq (\pi/5)h < 0.8h$, or $\varepsilon \leq (\pi^2/800) = 0.0123$ for $N = 64$, which is satisfied for our choice of ε . The results are given in Table 3 and agree with the error bounds of Section 3.

Table 3					
$a = 0.01, b = 1, h = 2\pi/M, k = 1/N$					
M	N	$\ (u - U_h)(\cdot, 1)\ _h$	$\ u^N - u(\cdot, 1)\ _h$	$\ u^N - U_h(\cdot, 1)\ _h$	Ratio
20	4	0.02872	0.05381	0.06237	
40	8	0.00721	0.01365	0.01555	4.01
80	16	0.00180	0.00342	0.00388	4.00
160	32	0.00045	0.00085	0.00097	4.00
320	64	0.00011	0.00021	0.00024	4.04

References

- [1] M. Baldauf, *Linear stability analysis of Runge-Kutta-Based partial time-splitting schemes for the Euler equations*. Monthly Weather Review **138** (2010), 4475-4496.

- [2] D. Estep, V. Ginting, D. Ropp, J.N. Shadid and S. Tavener, *An a posteriori-a priori analysis of multiscale operator splitting*. SIAM J. Numer. Anal. **46** (2008), 1116-1146.
- [3] A. Gassmann and H.-J. Herzog, *A consistent time-split numerical scheme Applied to the nonhydrostatic compressible equations*. Monthly Weather Review. **135** (2007), 20-36.
- [4] E. Hansen and A. Ostermann, *Exponential splitting for unbounded operators*. Math. Comp. **78** (2009), 1485-1496.
- [5] W. Hundsdorfer and J. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Series in Computational Mathematics, Springer-Verlag Berlin Heidelberg 2003.
- [6] T. Jahnke and C. Lubich, *Error bounds for exponential operator splittings*. BIT **40** (2000), 735-744.
- [7] S. MacNamara and G. Strang, Gilbert, *Operator splitting*. Splitting methods in communication, imaging, science, and engineering, 95-114, Sci. Comput., Springer, Cham, 2016.
- [8] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal. **5** (1968), 506/517.