Electronic Transactions on Numerical Analysis.
Volume 18, pp. 174-187, 2004.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

# IMPLICIT FOR LOCAL EFFECTS AND EXPLICIT FOR NONLOCAL EFFECTS IS UNCONDITIONALLY STABLE*

MIHAI ANITESCU†, FARANAK PAHLEVANI‡, AND WILLIAM J. LAYTON §

**Abstract.** A combination of implicit and explicit timestepping is analyzed for a system of ODEs motivated by ones arising from spatial discretizations of evolutionary partial differential equations. Loosely speaking, the method we consider is implicit in local and stabilizing terms in the underlying PDE and explicit in nonlocal and unstabilizing terms. Unconditional stability and convergence of the numerical scheme are proved by the energy method and by algebraic techniques. This stability result is surprising because usually when different methods are combined, the stability properties of the least stable method plays a determining role in the combination.

**Key words.** unconditional stability, implicit-explicit methods, multiscale integration.

**AMS subject classifications.** 76D05, 35Q30, 90C31.

## 1. Introduction.

This work considers timestepping methods for systems of ordinary differential equations of the form

$$u'(t) + Au(t) + B(u)u(t) - Cu(t) = f(t), \tag{1.1}$$

in which $A$, $B(u)$, and $C$ are $n \times n$ matrices, $u(t)$ and $f(t)$ are $n$-vectors, and

$$A = A^T \succ 0,\ B(u) = -B(u)^T,\ C = C^T \succeq 0 \text{ and } A - C \succeq 0. \tag{1.2}$$

Here $\succ$ and $\succeq$ denote, respectively, the positive definite and the positive semidefinite ordering. The key properties motivating our work are that $A$ is sparse and that although $C$ is not sparse, the action of $C$ on a vector is inexpensive to calculate. This structure is motivated by multiscale discretizations of turbulence, but can also arise from closed-loop control problems and ensemble calculations. Given this structure of (1.1), the simplest scheme that is computationally feasible is explicit in the global, unstable part of (1.1), that is, $Cu$. Accordingly, we consider

$$\frac{u_{n+1} - u_n}{k} + Au_{n+1} + B(u_n)u_{n+1} - Cu_n = f_{n+1}, \tag{1.3}$$

where $k$ is the time step and $u_n$ is the approximation to $u(t = nk)$ produced by the above numerical scheme. This method is in the class of implicit-explicit methods for time-dependent partial differential equations [2, 1]. Usually when methods are combined, the stability properties of the explicit method play a determining role in the overall method. In Theorems 2.4 and 2.6, we prove the surprising result that (1.3) is unconditionally stable. This result is outside the realm of root condition stability analysis for uncoupled scalar problems.

In Section 2, unconditional stability and convergence of (1.3) are proved. We give two stability proofs. The first is algebraic. Since the constants depend on the dimension of the system, we also give an energy proof of stability (with uniform constants), that is potentially extensible to discretized PDEs. Section 3 presents numerical tests illustrating the theory. First, we briefly summarize some motivating problems leading to (1.1).

*Received October 1, 2003. Accepted for publication October 14, 2004. Recommended by D. Bertaccini.

†Argonne National Laboratory, Mathematics and Computer Science Division, 9700 South Cass Avenue, Argonne, IL 60439, U.S.A. E-mail: anitescu@mcs.anl.gov.

‡ University of Pittsburgh, Department of Mathematics, Pittsburgh, PA 15260, U.S.A. E-mail: fap4@pitt.edu.

§ University of Pittsburgh, Department of Mathematics, Pittsburgh, PA 15260, U.S.A. E-mail: wjl@pitt.edu.

The basic model of the turbulent dispersion is that it is dissipative in the mean (see [12], [16], [9]). A more accurate formulation is that its dissipative effects are focused on the smallest resolved scales (see [7]). This physical idea has led to algorithms for numerical stabilization of transport-dominated phenomena based on eddy diffusivity acting only on the smallest resolved scales (e.g., [10], [8], [15], [11], [5], [6], [7], [13], [14]). The natural realization of this idea for spatial discretizations of convection diffusion equations is, diffusive stabilization on all scales and then antidiffusing on the large scales. This leads to the system of ODEs

$$\dot{u}_{ij}(t) + b \cdot \nabla^h u_{ij} - (\epsilon_0(h) + \epsilon)\Delta^h u_{ij} + \epsilon_0(h) P_H(\Delta^h P_H(u_{ij})) = f_{ij}, \tag{1.4}$$

where standard notation is used: $\Delta^h$ is the discrete Laplacian, $\epsilon_0(h)$ is the artificial viscosity parameters and $P_H$ denotes a projection onto a coarser mesh( see Section 3 for details). The system (1.4) fits exactly the form (1.1), (1.2), where $C$ is provided as the matrix arising from $\epsilon_0(h)$ term. We shall also test one algorithm as a perturbation of the method (1.4), in which the projection is replaced by a nearest averaging $\overline{\Delta^h \overline{u_{ij}}}$. In both cases, the projection or averaging operator accounts for the nonlocal character (i.e., the large bandwidth) of $C$. On the other hand, averaging and projection are both embarrassingly parallel operators, whose action on a given vector, is cheap to perform.

REMARK 1.1. *(1) A second main application is discretization of turbulent flow problems which, although nonlinear and constrained, have a similar structure to the above (simple) linear convection diffusion problem.*
*(2) A known method of stabilizing the timestepping and the associated linear system (but not the spatial discretization ) corresponds to (1.1) without the averaging:*

$$\frac{u_{n+1} - u_n}{k} + b \cdot \nabla^h u_{n+1} - (\epsilon_0(h) + \epsilon)\Delta^h u_{n+1} + \epsilon_0(h)\Delta^h u_n = f_{n+1}. \tag{1.5}$$

*Each time step requires the inversion of the matrix corresponding to the operator $-(\epsilon_0(h) + \epsilon)\Delta^h + b \cdot \nabla^h + k^{-1}I$, which, for $\epsilon_0$ suitably chosen, is an $M$-matrix. Our analysis applies to this method as well.*

**2. Stability Analysis of the Numerical Scheme.** For our analysis, we assume that $B(u)$ is in $C^1(\Re^n)$ and $f(t)$ is in $C^1([0,\infty])$. For any $T > 0$, we denote by ,

$$F_T = \max_{t \in [0,T]} \|f(t)\|_2 .$$

LEMMA 2.1. *The system of ODEs (1.1), under the condition (1.2), with initial condition $u(0) = u_0$, has a unique solution on $[0, T]$, for any $T > 0$.*

**Proof** Since (1.1) can be written as $\dot{u} = \psi(t, u)$ with $\psi$ being of class $C^0$ in $t$ and $C^1$ in $u$, local existence and uniqueness follows from the classical theory of ODEs [4, Theorem V.8].

We now show that, the solution does not experience blow-up and can be extended everywhere. We multiply through (1.1) by $(u(t)^T)$ and we use (1.2) to obtain that,

$$u(t)^T u'(t) \le -u(t)^T (A - C)u(t) + u(t)^T f(t) \le u(t)^T f(t).$$

Using Cauchy-Schwarz, we obtain that,

$$\frac{d}{dt}\|u(t)\|_2^2 \le \|u(t)\|_2^2 + F_T^2.$$

In turn, this implies that,

$$\|u(t)\|_2^2 \le \|u(0)\|_2^2 \, e^t + F_T^2 \left(e^t - 1\right)$$

for any $t$ ,in an interval containing $0$, where $u(t)$ is defined. Since $u(t)$ does not experience blow-up in finite time, it can be extended uniquely over all of $[0,T]$. $\Box$

Note that, from (1.1) and from our assumption that $f(t)$ is of class $C^1([0,\infty])$, we get that $u(t)$ is of class $C^2([0,\infty])$. The fact that $u''(t)$ is continuous will be used in determining a bound for the truncation error.

Consider the system of ODEs (1.1), under the condition (1.2) and discretized by (1.3).

First, we note that each step of (1.3) requires the inversion of $I + kA + kB_n$.

LEMMA 2.2. *Under (1.2), the $n \times n$ matrix $I + kA + kB_n$ has a positive definite symmetric part and is, therefore, invertible.*

**Proof:** Let $x$ be any nonzero vector in $\Re^n$. Then ,

$$\begin{aligned} x^T(I + kA + kB_n)x &= x^T x + kx^T Ax + kx^T B_n x \\ &= \|x\|_2{}^2 + kx^T Ax > 0. \quad \Box \end{aligned}$$

Since $A$, $B_n = B(u_n)$ and $C$ do not commute, the stability of the numerical scheme cannot be analyzed by reduction to eigenvalues. Therefore, we formulate an energy norm that is not increased at each time step, that is, $\|u_{n+1}\|_E \le \|u_n\|_E$.

DEFINITION 2.3. *The energy norm of (1.3), $\|.\|_E$, is given by*

$$\|u\|_E{}^2 = u^T u + ku^T Cu, \tag{2.1}$$

*for $u \in \Re^n$, and its associated inner product is $< u, v >_E = (N_k v)^T (N_k u)$, with $N_k = (I + kC)^{\frac{1}{2}}$, for $u, v \in \Re^n$.*

It can be seen immediately that, the energy norm and the 2-norm satisfy the following inequality:

$$\sqrt{1 + k\lambda_{min}(C)}\, \|u\|_2 \le \|u\|_E \le \sqrt{1 + k\lambda_{max}(C)}\, \|u\|_2 \,,$$

where $\lambda_{min}(C)$ and $\lambda_{max}(C)$ are, respectively, the smallest and the largest eigenvalue of $C$. From this inequality and the positive semidefiniteness of $C$, we get that the induced matrix norms satisfy,

$$\|A\|_E \le \|A\|_2 \sqrt{1 + k\lambda_{max}(C)}.$$

THEOREM 2.4. *Let $u_n$ satisfy (1.3) with $f(.) \equiv 0$, under the condition (1.2) on the coefficients. Then,*

$$\|u_{n+1}\|_E \le \|u_n\|_E \,.$$

**Proof:** Multiplying with $u_{n+1}^T$ through the equation in (1.3), we obtain

$$u_{n+1}^T \frac{u_{n+1} - u_n}{k} + u_{n+1}^T A u_{n+1} + u_{n+1}^T B_n u_{n+1} = u_{n+1}^T C u_n$$

Since $B_n$ is skew symmetric, $u_{n+1}^T B_n u_{n+1} = 0$. Therefore

$$u_{n+1}^T \frac{u_{n+1} - u_n}{k} + u_{n+1}^T A u_{n+1} = u_{n+1}^T C u_n. \tag{2.2}$$

This is equivalent to

$$u_{n+1}^T u_{n+1} + k u_{n+1}^T A u_{n+1} = k u_{n+1}^T C u_n + u_{n+1}^T u_n. \tag{2.3}$$

Since $A \succeq C$, we have that

$$u_{n+1}^T u_{n+1} + k u_{n+1}^T C u_{n+1} \leq u_{n+1}^T u_n + k u_{n+1}^T C u_n. \tag{2.4}$$

Define $w = (u_{n+1}, k^{1/2} C^{1/2} u_{n+1})^T, v = (u_n, k^{1/2} C^{1/2} u_n)^T$. Then (2.4) can be written as $w^T w \leq w^T v$. Applying the Cauchy-Schwarz inequality, we get $\|w\|_2 \leq \|v\|_2$. Hence,

$$u_{n+1}^T u_{n+1} + k u_{n+1}^T C u_{n+1} \leq u_n^T u_n + k u_n^T C u_n \tag{2.5}$$

or

$$\|u_{n+1}\|_E \leq \|u_n\|_E . \square$$

The conclusion of the preceding theorem is that when (1.1) is homogeneous, that is, $f \equiv 0$, we obtain that $\|u_n\|_E \leq \|u_0\|_E$, $\forall n$, independent of $T$. This means that our method is, indeed, unconditionally stable.

Consider (1.3) with $f \equiv 0$, rewritten as

$$(I + kA + kB_n) u_{n+1} = (I + kC) u_n, B_n = B(u_n). \tag{2.6}$$

Equation (2.6) yields

$$u_{n+1} = (I + kA + kB_n)^{-1} (I + kC) u_n,$$

which, in turn, implies that

$$(I + kC)^{\frac{1}{2}} u_{n+1} = (I + kC)^{\frac{1}{2}} (I + kA + kB_n)^{-1} (I + kC)^{\frac{1}{2}} (I + kC)^{\frac{1}{2}} u_n.$$

Therefore, from the definition of $\|\cdot\|_E$, a sufficient condition to prove the unconditional stability result is, to prove that,

$$\left\| (I + kC)^{\frac{1}{2}} (I + kA + kB_n)^{-1} (I + kC)^{\frac{1}{2}} \right\|_2 \leq 1, \quad \forall n. \tag{2.7}$$

Using the assumption (1.2), this can be done by using the following Lemma.

LEMMA 2.5. *Let $D_1 = {D_1}^T \succ 0$ and $D_2 = {D_2}^T \succ 0$ be $n \times n$ matrices such that $D_1 - D_2 \succ 0$. Let $D_4 = D_2^{\frac{1}{2}}$. If $D_3$ is an $n \times n$ skew-symmetric matrix, then*

$$\| D_4 (D_1 + D_3)^{-1} D_4 \|_2 \leq 1. \tag{2.8}$$

**Proof:** Let $F = D_4 (D_1 + D_3)^{-1} D_4$. It is straightforward that, $F^{-1} = D_4^{-1} (D_1 + D_3) D_4^{-1}$. For any nonzero vector $x$ in $\Re^n$,

$$\begin{aligned} x^T F^{-1} x &= x^T D_4^{-1} (D_1 + D_3) D_4^{-1} x \\ &= x^T D_4^{-1} D_1 D_4^{-1} x + x^T D_4^{-1} D_3 D_4^{-1} x \end{aligned}$$

Using the fact that $D_1 - D_2$ is nonnegative and that $D_3$ is skew symmetric, we obtain

$$x^T F^{-1} x \geq x^T D_4^{-1} D_2 D_4^{-1} x = x^T x, \qquad \text{for any } 0 \neq x \in \Re^n.$$

**ETNA**
**Kent State University**
**etna@mcs.kent.edu**



This implies that,

$$\|x\|_2{}^2 \leq x^T F^{-1} x \leq \|x\|_2 \cdot \left\|F^{-1}x\right\|_2, \qquad \text{for any } \ 0 \neq x \in \Re^n,$$

that is,

$$\|x\|_2 \leq \left\|F^{-1}x\right\|_2, \qquad \text{for any } \ 0 \neq x \in \Re^n. \tag{2.9}$$

Obviously (2.9) is equivalent to

$$\|Fy\|_2 \leq \|y\|_2, \qquad \text{for any } \ 0 \neq y \in \Re^n. \tag{2.10}$$

Since the last equation holds for any nonzero vector $y$, then $\|F\|_2 \leq 1$.
□

As a consequence, the inequality (2.7) follows for any $k \geq 0$, by setting $D_1 = I + kA$, $D_2 = I + kC$, $D_3 = kB_n$. Using equation (2.6), this observation provides a new proof to Theorem 2.4.

The additional benefit of Lemma 2.5 is that, we can use it to establish the stability of the inhomogeneous problem over an arbitrary but finite time interval $[0, T]$.

Consider (1.3) with $f \not\equiv 0$. After some simple calculations, we get that, $u_n$ satisfies

$$u_{n+1} = (I + kA + kB_n)^{-1}(I + kC)u_n + k(I + kA + kB_n)^{-1} f_{n+1}. \tag{2.11}$$

We denote the range of the step index $n$, by $[0, N]$, where $kN = T$. To simplify the notation, we do not explicitly indicate that $N$ depends on $k$ and $T$.

THEOREM 2.6. *Let (1.2) hold. Then the solution of (2.11) satisfies the following bound:*

$$\begin{aligned}\|u_{n+1}\|_E &\leq \|u_0\|_E + \frac{k}{1 + k\lambda_{min}(C)} \Sigma_{p=0}^n \|f_{p+1}\|_E \\ &\leq \|u_0\|_E + \frac{T}{(1 + k\lambda_{min}(C))} \max_{t\in[0,T]} \|f(t)\|_E, \ \forall 0 \leq n \leq N-1.\end{aligned}$$

**Proof:** Let $N_k = (I + kC)^{\frac{1}{2}}$, $M_k = (I + kA + kB_n)^{-1}$. Then, from Lemma 2.5, with $M_k = D_1 + D_3$ and $N_k = D_4$, it follows that, $\|N_k M_k N_k\| \leq 1$. Consequently,

$$\begin{aligned}\|n_{n+1}\|_E^2 &= (N_k u_{n+1})^T (N_k u_{n+1}) = (N_k u_{n+1})^T (N_k M_k N_k) \left((N_k u_n) + k(N_k^{-1} f_{n+1})\right) \\ &\leq \|u_{n+1}\|_E \|N_k M_k N_k\| \left(\|u_n\|_E + k \left\|N_k^{-1} f_{n+1}\right\|\right) \\ &\leq \|u_{n+1}\|_E \left(\|u_n\|_E + k \left\|N_k^{-1} f_{n+1}\right\|\right).\end{aligned}$$

Moreover, by considering that,

$$\left\|N_k^{-1} f_{n+1}\right\| = \left\|N_k^{-2} N_k f_{n+1}\right\| \leq \frac{1}{1 + k\lambda_{min}(C)} \|f_{n+1}\|_E,$$

this implies,

$$\|u_{n+1}\|_E - \|u_n\|_E \leq \frac{k}{(1 + k\lambda_{min}(C))} \|f_{n+1}\|_E, \ 0 \leq n \leq N-1.$$

Summing from 0 to $n$ gives,

$$\|u_{n+1}\|_E - \|u_0\|_E \le \frac{k}{(1+k\lambda_{min}(C))}\Sigma_{p=0}^{n}\|f_{p+1}\|_E\,,\ \forall 0 \le n \le N-1,$$

that is,

$$\|u_{n+1}\|_E \le \|u_0\|_E + \frac{k}{(1+k\lambda_{min}(C))}\Sigma_{p=0}^{n}\|f_{p+1}\|_E\,,\ 0 \le n \le N-1,$$

which is the claimed first result. The second result follows immediately. □

**3. Convergence analysis of the numerical scheme.** The local truncation error, of the method (1.3), is clearly $O(k)$. In the error estimate (which follows) we need a precise statement of this fact, which we now derive. To simplify our notation, we use $\hat{u}_n$ to denote $u(t_n)$, where $u(\cdot)$ is the exact solution of (1.1).

At key points of our proofs we will use the following statement

(3.1)
$$v : [a,b] \to \Re^n,\ v \in \mathcal{C}[a,b] \ \Rightarrow\ \exists\zeta \in (a,b) \text{ such that } \left\|\int_{[a,b]} v(t)dt\right\| \le (b-a)\,\|v(\zeta)\|\,.$$

The statement follows for any $a, b \in \Re$ and any vector norm $\|\cdot\|$, by using the triangle inequality, the continuity of $v(\cdot)$ and the intermediate value theorem.

According to the definition of local truncation error [3],

(3.2)
$$\begin{aligned}\tau_{n+1} &= \frac{\hat{u}_{n+1}-\hat{u}_n}{k} + A\hat{u}_{n+1} + B(\hat{u}_n)\hat{u}_{n+1} - C\hat{u}_n \\ &\quad -[\hat{u}'_{n+1} + A\hat{u}_{n+1} + B(\hat{u}_{n+1})\hat{u}_{n+1} - C\hat{u}_{n+1}] \\ &= \frac{\hat{u}_{n+1}-\hat{u}_n}{k} - \hat{u}'_{n+1} - (B(\hat{u}_{n+1}) - B(\hat{u}_n))\hat{u}_{n+1} + C(\hat{u}_{n+1}-\hat{u}_n).\end{aligned}$$

Using the second-order integral form of the Taylor expansion around $t_{n+1}$, we obtain,

$$\hat{u}_{n+1} - \hat{u}_n - k\hat{u}'_{n+1} = -\int_{t_{n+1}}^{t_n} u''(t)(t-t_{n+1})dt,$$

which can be rewritten as,

$$\frac{\hat{u}_{n+1}-\hat{u}_n}{k} - \hat{u}'_{n+1} = -\frac{1}{k}\int_{t_{n+1}}^{t_n} u''(t)(t-t_{n+1})dt = -\frac{1}{k}\int_{t_n}^{t_{n+1}} u''(t)(t_{n+1}-t)dt.$$

Using the first-order integral form of the Taylor expansion around $t_n$, we obtain

$$(B(\hat{u}_{n+1}) - B(\hat{u}_n))\,\hat{u}_{n+1} - C(\hat{u}_{n+1}-\hat{u}_n) = \int_{t_n}^{t_{n+1}}\left(\tfrac{d}{dt}B(u(t))\hat{u}_{n+1} - Cu'(t)\right)dt.$$

Using the expression we have derived for the local truncation error $\tau_{n+1}$, and the preceding equations derived from Taylor's theorem, we obtain that

$$\begin{aligned}\tau_{n+1} &= -\frac{1}{k}\int_{t_n}^{t_{n+1}} u''(t)(t_{n+1}-t)dt - \int_{t_n}^{t_{n+1}}\left(\frac{d}{dt}B(u(t))\hat{u}_{n+1} - Cu'(t)\right)dt \\ &= \int_{t_n}^{t_{n+1}}\left(-\frac{t_{n+1}-t}{k}u''(t) - \frac{d}{dt}B(u(t))\hat{u}_{n+1} + Cu'(t)\right)dt.\end{aligned}$$

Using (3.1), we obtain that there exists $\xi_n \in (t_n, t_{n+1})$ such that ,

$$\|\tau_{n+1}\|_2 \leq \left\| -u''(\xi_n)(t_{n+1}-\xi_n) - k\, \frac{d}{dt}B(u(t))\Big|_{t=\xi_n} \hat{u}_{n+1} + kCu'(\xi_n) \right\|_2 . \tag{3.3}$$

Hence, using the fact that $0 \leq (t_{n+1}-\xi_n) \leq k$, we obtain that,

$$\begin{aligned} \|\tau_{n+1}\|_2 \quad &\leq \quad k \max_{t_n \leq s \leq t_{n+1}} \Big( \quad \|u''(s)\|_2 \\ &+ \quad \left\| \tfrac{d}{dt}B(u(t))\big|_{t=s} \right\|_2 \max_{t_n \leq \theta \leq t_{n+1}} \|u(\theta)\|_2 + \|Cu'(s)\|_2 \Big) . \end{aligned}$$

This proves the following lemma.

LEMMA 3.1. *Let $n \geq 0$. The method*

$$\frac{u_{n+1}-u_n}{k} + Au_{n+1} + B_n u_{n+1} - Cu_n = f_{n+1}, \tag{3.4}$$

*where $A = A^T \succ 0$ and $C = C^T \succeq 0$ are $n \times n$ symmetric matrices, $B_n$ an $n \times n$ skew-symmetric matrix, and $f_{n+1} = f((n+1)k)$, is consistent. That is, the local truncation error is $O(k)$.*

We now bound the total error. We consider first the energy norm of truncation error.

LEMMA 3.2. *Let $\tau_{n+1}$ be the local truncation error of method (3.4). Then*

$$\|\tau_{n+1}\|_E \leq k \max_{0\leq t\leq T} \left( \|u''(t)\|_E + \|Cu'(t)\|_E + \left\| \frac{d}{dt}B(u(t)) \right\|_E \max_{0\leq s\leq T} \|u(s)\|_E \right) . \tag{3.5}$$

**Proof:** By definition of the energy norm and using the expression we have obtained for $\tau_{n+1}$ from Taylor's Theorem, for which we apply (3.1), we get

$$\|\tau_{n+1}\|_E \leq \left\| -u''(\xi_n)(t_{n+1}-\xi_n) - k\, \frac{d}{dt}B(u(t))\Big|_{t=\xi_n} \hat{u}_{n+1} + kCu'(\xi_n) \right\|_E$$

for some $\xi_n \in (t_n, t_{n+1})$. The conclusion follows after applying the inequality $0 \leq t_{n+1} - \xi_n \leq k$, the triangle inequality, and the properties of the $\max$ function. Note that $\left\| \frac{d}{dt}B(u(t)) \right\|_E$ is the induced $\|\cdot\|_E$ of the corresponding matrix. □

We now give a convergence result for the solution of (1.3). First, we need to compute a certain estimate. We have, after using the chain rule when we compute $\frac{d}{d\theta}$, that

$$[B(\hat{u}_n) - B(u_n)]\, \hat{u}_{n+1} = \int_0^1 \frac{d}{d\theta}[B(\hat{u}_n\theta + u_n(1-\theta))]\hat{u}_{n+1} d\theta = W_n e_n$$

where $e_n = \hat{u}_n - u_n$ and

$$W_n = \int_0^1 d\theta \left( \; \nabla_u \left[ B(u)\hat{u}_{n+1} \right] \big|_{u=\hat{u}_n\theta + (1-\theta)u_n} \right) . \tag{3.6}$$

LEMMA 3.3. *Let u(.) be the solution of (1.1) and $u_n$ be the approximation to $u(nk)$, obtained from the numerical scheme (1.3). Then there exists $\Gamma$ such that, $\forall t \in [0,T]$ we have that ,*

$$\|W_n\|_2 \leq \Gamma, \text{ and } \|W_n\|_E \leq \Gamma_E = \Gamma\sqrt{1 + k\lambda_{max}(C)}, \; \forall 0 \leq n \leq N.$$

**ETNA**
Kent State University
etna@mcs.kent.edu



**Proof:** From Theorem 2.6, we have that,

$$\begin{aligned} \|u_n\|_2 &\leq \|u_n\|_E \leq \|u_0\|_E + T \max_{t\in[0,T]} \|f(t)\|_E \\ &\leq \sqrt{1+k\lambda_{max}(C)} \left(\|u_0\|_2 + T \max_{t\in[0,T]} \|f(t)\|_2\right), \ \forall 0 \leq n \leq N. \end{aligned}$$

We define

$$\Lambda_E = \sqrt{1+T\lambda_{max}(C)} \left( \|u_0\|_2 + T \max_{t\in[0,T]} \|f(t)\|_2 \right).$$

From Lemma 2.1 we have that $u(t)$ is continuous and bounded on $[0,T]$, and we define $\Lambda_u = \max_{t\in[0,T]} \|u(t)\|_2$. Since $B(\cdot)$ is of class $C^1$, we can define

$$\Gamma = \max_{\theta\in[0,1],\, \|u_1\|_2\leq\Lambda_E,\, \|u_2\|_2\leq\Lambda_u, \|v_2\|_2\leq\Lambda_u} \|(\nabla_u [B(\theta v_2 + (1-\theta)u_1)u_2])\|_2 .$$

Using the triangle inequality in $\|W_n\|_2$ following the definition (3.6) of $W_n$ we immediately obtain that

$$\|W_n\|_2 \leq \Gamma, \ \forall 0 \leq n \leq N.$$

The second part of the conclusion follows from the inequality between $\|\cdot\|_E$ and $\|\cdot\|_2$. $\square$

THEOREM 3.4. *Consider solving the inhomogeneous problem on the interval [0,T]*

$$u' + Au + B(u)u - Cu = f$$

*using method (1.3), i.e.,*

$$\frac{u_{n+1}-u_n}{k} + Au_{n+1} + B_n u_{n+1} - Cu_n = f_{n+1},$$

*where $B_n = B(u_n)$ and $f_{n+1} = f((n+1)k)$. Let $e_n = u(t_n) - u_n$ denote the global error. Assume that $e_0 = 0$. Then the method is convergent and*

$$\begin{aligned} \|e_{n+1}\|_E &\leq \frac{\left(1+\frac{k\Gamma_E}{1+k\lambda_{min}(C)}\right)^{n+1} - 1}{\frac{k\Gamma_E}{1+k\lambda_{min}(C)}} \frac{k^2 U}{1+k\lambda_{min}(C)} \\ &\leq \frac{e^{\frac{T\Gamma_E}{1+k\lambda_{min}(C)}} - 1}{\frac{k\Gamma_E}{1+k\lambda_{min}(C)}} \frac{k^2 U}{1+k\lambda_{min}(C)}, \ \forall 0 \leq n \leq N-1, \end{aligned}$$

*when $\Gamma_E \neq 0$, and*

$$\|e_{n+1}\|_E \leq (n+1)\frac{k^2 U}{1+k\lambda_{min}(C)} \leq T\frac{kU}{1+k\lambda_{min}(C)}, \ \forall 0 \leq n \leq N-1,$$

*when $\Gamma_E = 0$, where*

$$U = \max_{0\leq t\leq T} \left( \|u''(t)\|_E + \|Cu'(t)\|_E + \left\| \frac{d}{dt} B(u(t)) \right\|_E \max_{0\leq s\leq T} \|u(s)\|_E \right).$$

**Proof:** Following the definition of the truncation error $\tau_{n+1}$ and using the equation (3.6), we obtain that the error, $e_n = u(t_n) - u_n$, satisfies,

$$\frac{e_{n+1} - e_n}{k} + Ae_{n+1} + B_n e_{n+1} - Ce_n = \tau_{n+1} - W_n e_n.$$

After algebraic calculations, we find that,

$$e_{n+1} = (I + kA + kB_n)^{-1}(I + kC)e_n + k(I + kA + kB_n)^{-1}(\tau_{n+1} - W_n e_n).$$

We use the energy inner product to obtain,

$$\begin{aligned} &< e_{n+1}, e_{n+1} >_E = \\ &< (I + kA + kB_n)^{-1}(I + kC)e_n + k(I + kA + kB_n)^{-1}(\tau_{n+1} - W_n e_n), e_{n+1} >_E . \end{aligned}$$

Applying the definition of energy norm (2.1) and the substitutions $M_k = (I + kA + kB_n)^{-1}(I + kC)$, and $N_k = (I + kC)^{\frac{1}{2}}$, we find that ,

$$\begin{aligned} &(N_k e_{n+1})^T (N_k e_{n+1}) = \\ &(N_k e_{n+1})^T N_k M_k e_n + k(N_k e_{n+1})^T N_k (I + kA + kB_n)^{-1}(\tau_{n+1} - W_n e_n). \end{aligned}$$

Using the Cauchy-Schwarz inequality, we obtain that,

$$\begin{aligned} \|N_k e_{n+1}\|_2^2 \le{}& \|N_k e_{n+1}\|_2 . \left\|N_k M_k N_k^{-1}\right\|_2 . \|N_k e_n\|_2 \\ &+k \|N_k e_{n+1}\|_2 . \left\|N_k M_k N_k^{-1}\right\|_2 . \left\|N_k^{-1} (\tau_{n+1} - W_n e_n)\right\|_2 . \end{aligned}$$

Thus

$$\|N_k e_{n+1}\|_2 \le \left\|N_k M_k N_k^{-1}\right\|_2 . \|N_k e_n\|_2 + k \left\|N_k M_k N_k^{-1}\right\|_2 . \left\|N_k^{-1} (\tau_{n+1} - W_n e_n)\right\|_2 .$$

Using Lemma 2.5 with $D_2 = N_k^2$ and $D_1 + D_3 = M_k N_k^{-2}$, we obtain that $\left\|N_k M_k N_k^{-1}\right\|_2 \le 1$. Hence

$$\|e_{n+1}\|_E \le \|e_n\|_E + k \left\|N_k^{-2}\right\|_2 \|(\tau_{n+1} - W_n e_n)\|_E .$$

Equivalently, we obtain that,

$$\|e_{n+1}\|_E \le \|e_n\|_E + k \left\|(I + kC)^{-1}\right\|_2 (\|\tau_{n+1}\|_E + \|W_n\|_E \|e_n\|_E) .$$

Notice that, $(I + kC)^{-1}$ is a symmetric positive definite matrix and

$$\left\|(I + kC)^{-1}\right\|_2 = \frac{1}{1 + k\lambda_{min}(C)}.$$

On the other hand, by Lemma 3.3, there is a constant $\Gamma_E$ ,such that ,$\|W_n\|_E \le \Gamma_E$. Therefore,

$$\|e_{n+1}\|_E \le \left(1 + \frac{k\Gamma_E}{1 + k\lambda_{min}(C)}\right) \|e_n\|_E + \frac{k}{1 + k\lambda_{min}(C)} \|\tau_{n+1}\|_E . \tag{3.7}$$

This is a recursion formula of the following form:

$$r_{n+1} \le ar_n + b\theta_n,$$

**ETNA**
**Kent State University**
**etna@mcs.kent.edu**



which, when $a \neq 0$ has an upper bound of the type

$$r_{n+1} \leq a^{n+1} r_0 + \frac{a^{n+1}-1}{a} b \max_n \|\theta_n\|_E .$$

Using this fact, we obtain that, when $\Gamma_E \neq 0$, the following bound for the error holds , whenever $0 \leq n \leq N-1$.

$$\|e_{n+1}\|_E \leq \left(1 + \frac{k\Gamma_E}{1+k\lambda_{min}(C)}\right)^{n+1} \|e_0\|_E$$
$$+ \frac{\left(1+\frac{k\Gamma_E}{1+k\lambda_{min}(C)}\right)^{n+1} - 1}{\frac{k\Gamma_E}{1+k\lambda_{min}(C)}} \cdot \frac{k}{1+k\lambda_{min}(C)} \max_n \|\tau_{n+1}\|_E$$

Replacing $\|\tau_{n+1}\|_E$ by its bound (3.5) obtained in Lemma 3.2, and considering that $e_0 = 0$, we have, when $\Gamma_E \neq 0$ and $0 \leq n \leq N-1$, that

$$\|e_{n+1}\|_E \leq \frac{\left(1+\frac{k\Gamma_E}{1+k\lambda_{min}(C)}\right)^{n+1} - 1}{\frac{k\Gamma_E}{1+k\lambda_{min}(C)}} \cdot \frac{k^2 U}{1+k\lambda_{min}(C)}$$

with $U = \max_{0 \leq t \leq T} \left( \|u''(t)\|_E + \|Cu'(t)\|_E + \left\|\frac{d}{dt}B(u(t))\right\|_E \max_{0 \leq s \leq T} \|u(s)\|_E \right)$.

The second inequality for $\Gamma \neq 0$ follows from the inequality $(1+x)^n \leq e^{xn}$, for $x > 0$ and $n$ positive integer.

When $\Gamma_E = 0$, we immediately get from (3.7) and from Lemma 3.2 that,

$$\|e_{n+1}\|_E \leq (n+1) \frac{k^2 U}{1+k\lambda_{min}(C)}, \ \forall 0 \leq n \leq N-1,$$

which, together with $kN = T$ prove the inequalities for $\Gamma_E = 0$.

The convergence follows from the fact that, $\|\cdot\|_E$ converges to $\|\cdot\|_2$ as $k \to 0$, which implies that, $\|e_n\|_2 \to 0$ as $k \to 0$. $\square$

The case $\Gamma_E = 0$ occurs, for example, when $B(u)$ is constant (which we simulate numerically in the next section). For that case, the error increases only linearly with the size of the interval, assuming that the derivatives, up to order 2 of the solution $u(t)$ are uniformly bounded.

**4. Numerical Results.** Let $\Omega = [0,1] \times [0,1]$. For the equation

$$\begin{aligned} u_t + b \cdot \nabla u - \epsilon \Delta u &= f, \text{ over } \Omega, \\ u &= \phi(x) \text{ on } \delta\Omega, \\ u(x,0) &= u_0(x) \text{ in } \Omega, \end{aligned} \tag{4.1}$$

we use the method described in this work, with uniform mesh and central differences. A choice must be made for the antidiffusion operator: averaging or projection. We have selected averaging. Since it is just outside the theory, we will thereby test the robustness of the algorithm. Antidiffusion is completed by averaging, where $\bar{u}(p)$:=weighted average of nearest neighbors. This corresponds to filtering with $\delta = 2h$. The method becomes in our case,

$$\dot{u}_{ij}(t) + b \cdot \nabla^h u_{ij} - (\epsilon + \epsilon_0)\Delta^h u_{ij} + \epsilon_0 \overline{\Delta \overline{u_{ij}}}^q = f_{ij},$$

**ETNA**
Kent State University
etna@mcs.kent.edu



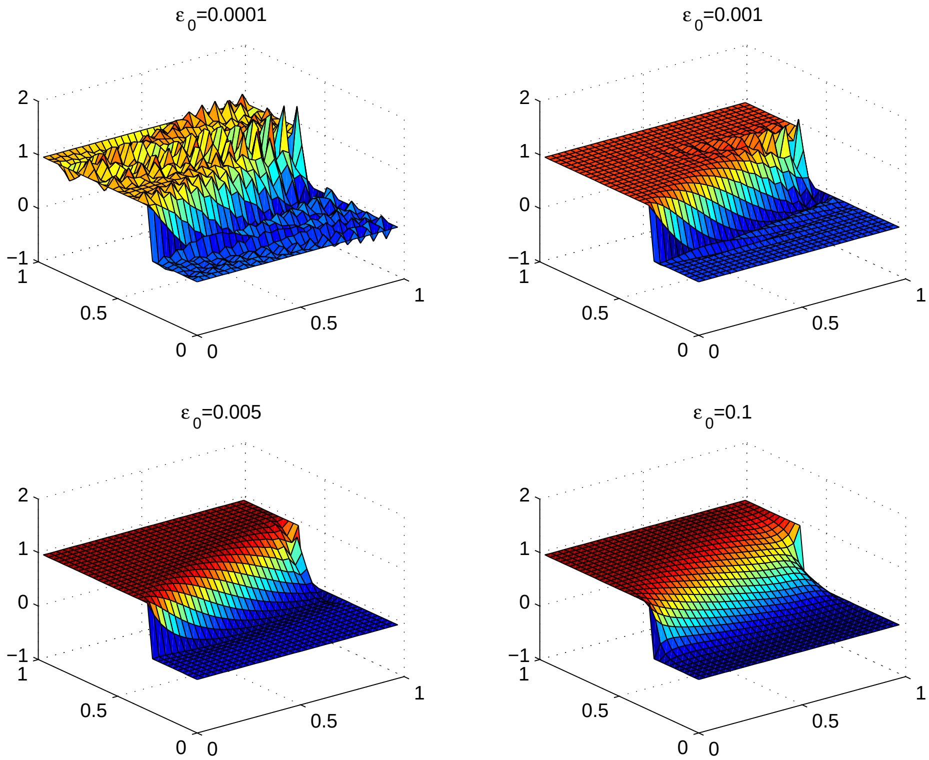


FIG. 4.1. *Spatial stability of the steady-state solution for various choices of the artificial viscosity parameter* $\epsilon_0$.

where $q$ denotes how many times the average operation is taken and $i, j$ are the indices of interior nodes. In our experiments ,we chose $q = 2$ and $\epsilon = 10^{-4}$. We take $b = (\cos(\theta), \sin(\theta))$, where $\theta = 17°$.

For the boundary and initial conditions, we take the line at angle $\theta$ through the center of the domain. On the north side of the line we take $\phi = 1$ on the boundary; on the south side we take $\phi = 0$ on the boundary. We take $f = 0$ and 0 as initial conditions. Note that ,this does not mean that the resulting ODE is homogeneous, since the boundary conditions will alter the right hand side of the ODE at inner points near the boundary of the domain.

We performed the following experiments, all on a $32 \times 32$ mesh.

1. We ran the simulation for $1,000$ steps with a time step of 10, with the artificial viscosity parameter $\epsilon_0$ having successively the values $10^{-1}$, $5 \times 10^{-3}$, $10^{-3}$, $10^{-4}$. We have presented no analysis for the spatial dependence of the solution with respect to $\epsilon_0$, but we have included this experiment for validation, since our choice of parameters should result roughly in the steady-state approximation for this mesh, which has been studied before in the literature.
   The results are depicted in Figure 4.1. We see that, when the artificial viscosity parameter $\epsilon_0$ is very small, a complete loss of coherence of the spatial structure results, whereas too large a parameter ($\epsilon_0 = 0.1$) alters the steady-state solution significantly. This effect is consistent with the typical behavior of centered methods for the skew step problem [17].
2. For $\epsilon_0 = 10^{-4}$, we ran the simulation for 100 steps with a time step of 1 and for $1,000$ steps with a time step of $0.1$. The energy norm comparison of these

**ETNA**
Kent State University
etna@mcs.kent.edu



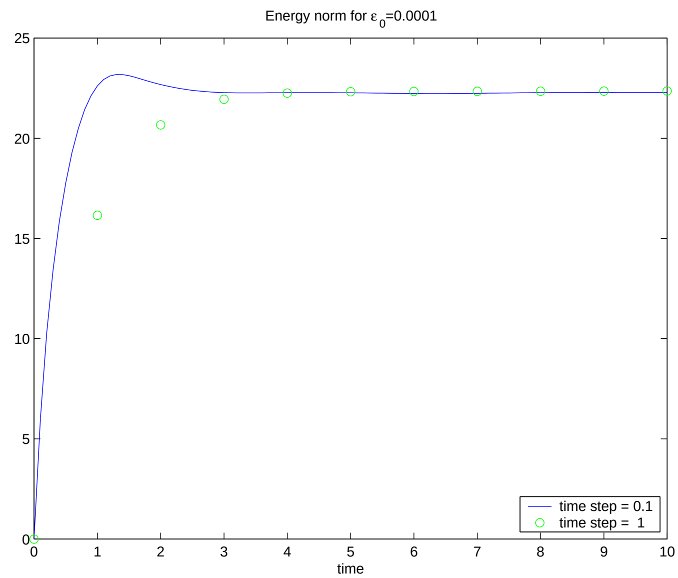


FIG. 4.2. *Stability of the numerical method demonstrated by the behavior of the energy norm*

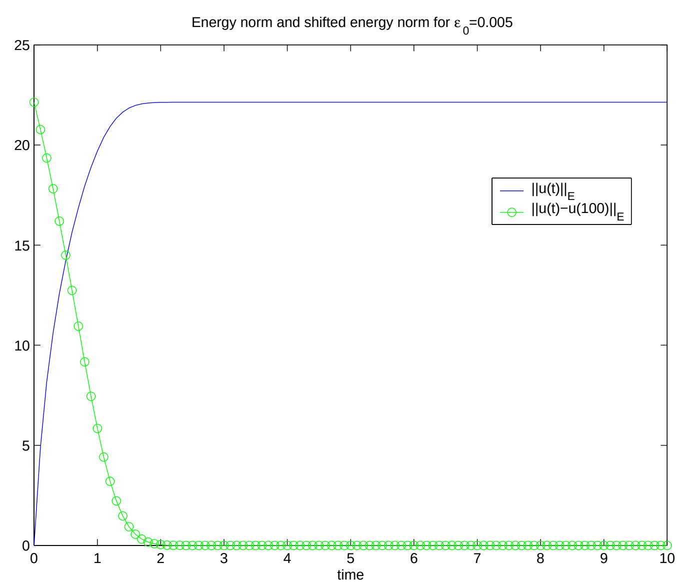


FIG. 4.3. *Numerical validation of Theorem 2.4*

computations is presented in Figure 4.2. We see that, even for the very large step, the energy norm stays bounded, consistent with our absolute stability claim.
We also present in Figure 4.3, a comparison between the energy norms of the distance between the successive iterates of the two cases and their outcome at time $100$. From Figure 4.2, we infer that $u(100)$ is a reasonable approximation to the steady-state solution. Since the equation (4.1) is linear, we have that $u_n - u(100)$ is the result of the numerical scheme applied to the homogeneous equation associated to

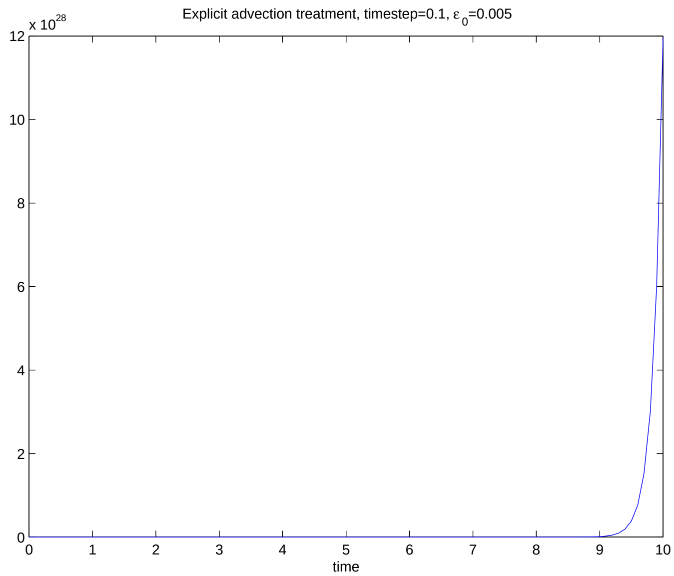


FIG. 4.4. *Exponential growth of the solution of the scheme that includes the advection term explicitly*

(4.1). From Theorem 2.4 we have that $\|u_n - u(100)\|_E$ must be a decreasing sequence, which is exactly what we observe from Figure 4.3. Note that $\|u_n\|_E$ is not a decreasing sequence, as can be seen in Figure 4.3. Moreover, the sequence $\|u_n\|_E$ may not even be monotonic, as seen in Figure 4.2, for $k = 0.1$.

3. We compare the results of our scheme with the similar scheme that takes into account explicitly the term that contains the skew-symmetric matrix $B(u_n)$. For the latter scheme we obtain the recursion

$$\frac{u_{n+1} - u_n}{k} + Au_{n+1} + B(u_n)u_n - Cu_n = f_{n+1}.$$

We apply this scheme to our example on a $32 \times 32$ mesh for $1000$ time steps of length $k = 1$. We see the rapid exponential growth that is typical for computations with the time step outside the region of stability.

This demonstrates that our scheme has significantly better stability properties than the alternative, which would result in linear systems of comparable sparsity. The numerical scheme, based on a backward Euler approach that considers all terms implicitly, though absolutely stable, will result in less sparse linear systems since the matrix $C$ contains an averaging operator that substantially reduces sparsity and is not considered here for comparison.

**Acknowledgements.** The work of Mihai Anitescu was supported by the Mathematical, Informatical and Computational Sciences Program of the Office of Science of the United States Department of Energy, through the Contract W-31-109-ENG-38. The work of the authors was supported by the National Science Foundation through awards DMS-0112239 (MA and WJL). and DMS-0207627 (FP and WJL). The authors are grateful to the anonymous referees whose comments have improved the paper.

REFERENCES

[1] U. ASCHER, S. RUUTH, AND R. SPITERI, *Implicit-explicit runge-kutta methods for time-dependent partial differential equations*, Applied Numerical Mathematics, 25 (1997), pp. 151–167.

[2] U. ASCHER, S. RUUTH, AND B. WETTON, *Implicit-explicit methods for time-dependent pdes*, SIAM J. Numerical Analysis, 32 (1995), pp. 797–823.

[3] K. E. ATKINSON, *An introduction to numerical analysis*, Wiley, 1989.

[4] G. BIRKHOFF AND G.-C. ROTA, *Ordinary Differential Equations*, Ginn and Company, Boston, 1962.

[5] J. L. GUERMOND, *Stabilization of Galerkin approximations of transport equations by subgrid modeling*, M2AN, 33 (1999), pp. 1293–1316.

[6] ———, *Stabilization par viscosite de sous-maille pour l'approximation de Galerkin des operateurs lineaires monotones*, C.R.A.S., 328 (1999), pp. 617–622.

[7] T. J. HUGHES, L. MAZZEI, AND K. E. JASEN, *Large eddy simulation and the variational multiscale method*, Comput.Visual Sci., 3 (2000), pp. 47–59.

[8] T. J. R. HUGHES, L. MAZZEI, AND K. E. JANSEN, *Large eddy simulation and the variational multiscale method*, Comput. Visual Sci., 3 (2000), pp. 47–59.

[9] T. ILIESCU AND W. LAYTON, *Approximating the largger eddies in fluid motion III: the Boussinesq model for turbulent fluctuations*, Analele Stiintifice ale Universitatii Al.l.Cuza, tomul XLIV (1998), pp. 245–261.

[10] S. KAYA, *Numerical analysis of a subgrid scale eddy viscosity method for higher reynolds number flow problem*, University of Pittsburgh,Technical report, (2002).

[11] S. KAYA AND W. LAYTON, *Subgrid-scale eddy viscosity methods are variational multiscale methods*, University of Pittsburgh,Technical report, (2002).

[12] H. KESTEN AND G. PAPANICOLAOU, *A limit theorem for stochastic acceleration*, Comm. Math. Phys., 78 (1980), pp. 19–63.

[13] W. LAYTON, *Approximating the larger eddies in fluid motion V: Kinetic energy balance of scale similarity models*, Math. and Computer Modeling, 31 (2000), pp. 1–7.

[14] ———, *A connection between subgrid scale eddy viscosity and mixed methods*, Appl. Math. and Computing, 133 (2002), pp. 147–157.

[15] Y. MADAY AND E. TADMOR, *Analysis of the spectral vanishing viscosity method for periodic conservation laws*, SIAM Journal on Numerical Analysis, 26 (1989), pp. 854–870.

[16] B. MOHAMMADI AND O. PIRONNEAU, *Analysis of the K-$\epsilon$ Turbulence Model*, Wiley, 1993.

[17] H. G. ROOS, M. STYNES, AND L. TOBISKA, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer, Berlin, 1996.