SIAM J. NUMER. ANAL.
Vol. 38, No. 6, pp. 1763–1783

PARTIALLY IMPLICIT BDF2 BLENDS FOR CONVECTION DOMINATED FLOWS*

WILLEM HUNDSDORFER†

Abstract. In this paper we consider various blends of implicit and explicit time integration schemes, based on the well-known BDF2 method, applied to convection-diffusion problems with dominating convection. A fully implicit treatment of convection terms is often not very efficient. We shall deal with second order schemes that are implicit in the convection terms only locally in space, without introducing the internal inconsistencies that are common with many time-splitting methods. Along with implementation aspects of the implicit relations, we shall discuss accuracy of the schemes, positivity and monotonicity properties.

Key words. numerical analysis, initial-boundary value problems, BDF methods, implicit-explicit methods, splitting methods

AMS subject classifications. 65M06, 65M12, 65M20

PII. S0036142999364741

1. Introduction.

When adopting the method of lines approach, space discretization of multidimensional, time-dependent partial differential equations (PDEs) results in large systems of ordinary differential equations (ODEs) which are to be integrated in time by an appropriate time stepping scheme. Frequently in such applications one is confronted with problems having both stiff and nonstiff parts. Diffusion, for example, leads to stiff terms that need implicit treatment. Convection terms can usually be taken explicitly, but if we have locally large convective velocities an explicit treatment is unfavorable due to the CFL restrictions on stability, whereas a fully implicit approach leads to systems of algebraic equations that are rather difficult to solve numerically. Here we shall deal with partial implicit treatment of convective terms in such a way that the resulting scheme is fully implicit only in those spatial regions where the solution is smooth and the convective velocities are large.

The focus in this paper is on convection dominated equations. First, consider the convection equation without any diffusion,

$$u_t + \nabla \cdot (q(x,t) f(u)) = 0, \qquad x \in \Omega,\ t \geq 0, \tag{1.1}$$

on a spatial domain $\Omega \subset \mathbb{R}^d$ with appropriate initial and boundary conditions. Here $q(x,t) \in \mathbb{R}^d$ is a given velocity and f is a scalar flux function. Discretization of the spatial derivatives leads to a large system of ODEs, the so-called *semidiscrete system*,

$$w'(t) = F(t, w(t)), \qquad t \geq 0, \tag{1.2}$$

where F contains the discretized convective terms, and an initial value $w_0 = w(0)$ is given. We consider numerical time integration schemes with step size $\tau > 0$, yielding approximations $w_n \approx w(t_n)$ at the time levels $t_n = n\tau$. For spatial discretization we shall deal with limited second order finite volume or finite difference formulas. The dimension of the semidiscrete system is proportional to the number of grid points, and components $w_i(t_n)$ of $w(t_n)$ refer to approximations at the grid point x_i or to an

*Received by the editors November 29, 1999; accepted for publication (in revised form) August 21, 2000; published electronically January 5, 2001.
http://www.siam.org/journals/sinum/38-6/36474.html

†CWI, P.O. Box 94079, 1090 GB, Amsterdam, The Netherlands (Willem.Hundsdorfer@cwi.nl).

average value on a cell Ω_i around x_i. With multidimensional problems i will denote a multi-index.

One of the most popular implicit methods for solving (1.2) is the second order BDF2 method

$$\frac{3}{2}w_n - 2w_{n-1} + \frac{1}{2}w_{n-2} = \tau F(t_n, w_n) \tag{1.3}$$

with $n \geq 2$; see [9]. Along with w_0, this two-step method needs w_1 as starting value. It can be computed by a one-step method, for instance, implicit Euler. The popularity of this BDF2 method is due to its stability and damping properties; see [10], for instance. These are crucial properties for efficient solution of diffusion equations.

Convection equations, on the other hand, are often treated more efficiently by an explicit method. Here we consider the related second order scheme

$$\frac{3}{2}w_n - 2w_{n-1} + \frac{1}{2}w_{n-2} = \tau F(t_n, \overline{w}_n), \quad \text{where} \quad \overline{w}_n = 2w_{n-1} - w_{n-2}, \tag{1.4}$$

to which we shall refer as the explicit BDF2 method. Note that $\overline{w}_n = 2w_{n-1} - w_{n-2}$ is just an explicit prediction by linear extrapolation. As with any standard explicit method, we now have a CFL condition for stability. Therefore, if we deal with large velocities or fine spatial grids, very small time steps have to be taken.

As we shall see, the fully implicit method also gives us difficulties when applied to large Courant numbers. This is due to slow convergence of the Newton iterations for the implicit relations but also due to loss of monotonicity. In this paper we therefore consider a partially implicit convection treatment, where only those parts in the domain with little spatial variation in the solution are treated implicitly. The resulting formula is

$$\frac{3}{2}w_n - 2w_{n-1} + \frac{1}{2}w_{n-2} = \tau F(t_n, \Theta w_n + (I - \Theta)\overline{w}_n), \tag{1.5}$$

where Θ is a diagonal matrix with entries $\theta_i = 0$ if the convection term is taken explicitly at the grid point x_i, and $\theta_i \in (0, 1]$ otherwise. The actual choice for the θ_i is discussed in section 4.

With convection-diffusion problems,

$$u_t + \nabla \cdot (q(x,t)f(u)) = \nabla \cdot (D(x,t,u) \cdot \nabla u), \tag{1.6}$$

the resulting semidiscrete system will be of the form

$$w'(t) = F(t, w(t)) + G(t, w(t)), \qquad t \geq 0, \tag{1.7}$$

where F contains the convective terms and G denotes discretized diffusion. The above formula (1.5) for the convection part can be well combined with implicit treatment of the diffusion term by considering

$$\frac{3}{2}w_n - 2w_{n-1} + \frac{1}{2}w_{n-2} = \tau F(t_n, \Theta w_n + (I - \Theta)\overline{w}_n) + \tau G(t_n, w_n), \tag{1.8}$$

so that we obtain a formula that benefits from the damping properties of the fully implicit BDF2 scheme for the diffusion part.

If $\Theta = O$ this is an implicit-explicit method of the type that was introduced by Crouzeix [5] and Varah [21]. Stability results can be found in [1, 5, 8, 21], for example, and a practical application in the field of air pollution was discussed in [22]. In general, the stability of this method is completely determined by the CFL restriction for the explicit convection part.

Note that all of the above methods are different from the usual time-splitting techniques, where different subproblems, such as $v'(t) = F(t, v(t))$ and $v'(t) = G(t, v(t))$, are solved subsequently on small time intervals. This leads to intermediate results which have little physical meaning, since they are not consistent with the total equation. Boundary conditions or interface conditions are usually lacking for these intermediate results. With the above BDF2-type methods we use only fully consistent approximations w_n and no intermediate results.

Further we note that if $\Theta = \theta I$ the above formula (1.5) is a two-step extension of the more familiar θ-method

$$w_n - w_{n-1} = \tau F(t_{n+\theta}, \theta w_n + (1-\theta) w_{n-1}) \tag{1.9}$$

with the explicit Euler method, $\theta = 0$, and the implicit Euler method, $\theta = 1$, as boundary cases. We shall not consider these methods here since both the implicit and explicit Euler methods are not well suited for convection problems. The implicit Euler method is much too diffusive, whereas the explicit Euler method is unstable for the spatial convection discretizations considered in this paper. Actually, in a method of lines setting, the explicit Euler method is unstable for all well-known spatial convection discretizations except for the (diffusive) first order upwind discretization.

A related method has been formulated by Blunt and Rubin [4] for one-dimensional (1D) problems, where the implicit Euler scheme was combined with an explicit, direct space-time scheme (Lax–Wendroff-type) with limiting. However, for multidimensional problems this combined scheme needs dimensional splitting since the formulation of such a direct space-time scheme for multidimensional problems is different than with the implicit Euler scheme; see also [13]. Moreover, due to the use of implicit Euler, the order is 1 at most.

In this paper we shall consider the second order BDF2 blends (1.5) mainly for purely convective problems. If diffusion is added as in (1.8), the method becomes implicit over the whole spatial domain, but in those regions where the entries θ_i are zero the implicit relations have a nice symmetric, diagonally dominant structure, so that standard linear solvers, such as conjugate gradients, will be very efficient.

Spatial discretization of the convective terms will be done by limiting in order to avoid oscillations and negative solution values. In section 2 we discuss by means of 1D examples implementation issues and qualitative behavior. As we shall see, the standard implicit BDF2 method (1.3) becomes rather expensive, and, more importantly, the results are also rather disappointing with respect to qualitative behavior and accuracy. This is due to the poor monotonicity properties of the standard implicit BDF2 method.

In section 3 we consider formula (1.5) with $\Theta = \theta I$, with the aim of selecting values of θ with better monotonicity properties than $\theta = 1$. To obtain theoretical results we shall concentrate on positivity for linear systems. The results in this section can be regarded as an extension of the positivity theory of Bolley and Crouzeix [2].

In section 4 we consider implementations of (1.5) with variable entries θ_i. The actual choices will be motivated by the preceding results. We shall discuss the accuracy of the schemes with variable entries in some detail in section 5, since the standard local truncation error no longer gives proper information about the accuracy of these schemes. This is similar to the situation for stiff ODEs as considered in Hundsdorfer and Steininger [12]. Numerical results will be presented in section 6 for a test example from reservoir simulation, where we have locally large convective velocities q near injection and production wells and moderate or small velocities elsewhere in

the spatial region. It will be seen that the locally implicit schemes can be much more efficient than the fully implicit counterparts such as (1.3), whereas this locally implicit approach allows step sizes much larger than with explicit schemes such as (1.4).

2. One-dimensional examples. In this paper we shall deal with convection-diffusion discretizations for 1D or two-dimensional (2D) problems. For ease of presentation we first consider the 1D convection problem

$$u_t + (q(x,t)f(u))_x = 0, \tag{2.1}$$

on $\Omega = [0,1]$, with monotonically increasing flux function f. Further, it is assumed that an initial profile $u(x,0)$ and appropriate boundary conditions are given. In this section we shall discuss the advantages and disadvantages of the implicit BDF2 method (1.3) compared to its explicit counterpart (1.4).

2.1. The spatial discretizations. For the spatial derivative in (2.1) we consider discretizations in flux form on a uniform mesh,

$$w_i' = \frac{1}{h}\Big(q_{i-\frac{1}{2}}\, f(w_{i-\frac{1}{2}}) - q_{i+\frac{1}{2}}\, f(w_{i+\frac{1}{2}})\Big), \tag{2.2}$$

with grid points $x_i = ih$ and $q_{i\pm 1/2} = q(x_i \pm \frac{1}{2}h, t)$. Here $w_i = w_i(t)$ stands for a semidiscrete approximation to the average value of $u(x,t)$ over the cell $\Omega_i = [x_i - \frac{1}{2}h, x_i + \frac{1}{2}h]$. The choice for the cell boundary values $w_{i\pm 1/2}$ determines the actual discretization.

It is well known that the first order upwind approximation $w_{i+1/2} = w_i$, for $q > 0$, gives very inaccurate and diffusive results. On the other hand, higher order linear discretizations, such as second order central $w_{i+1/2} = \frac{1}{2}(w_i + w_{i+1})$ or second order upwind $w_{i+1/2} = \frac{1}{2}(-w_{i-1} + 3w_i)$, give results that are very oscillatory. For that reason, discretizations with limiters have become increasingly popular.

In the following, let

$$\vartheta_i = \frac{w_i - w_{i-1}}{w_{i+1} - w_i}.$$

In (2.2) we shall deal with limited approximations for the cell boundary values of the form

$$w_{i+\frac{1}{2}} = \begin{cases} w_i + \frac{1}{2}\psi(\vartheta_i)(w_{i+1} - w_i) & \text{if } q_{i+\frac{1}{2}} \geq 0, \\ w_{i+1} + \frac{1}{2}\psi(1/\vartheta_{i+1})(w_i - w_{i+1}) & \text{if } q_{i+\frac{1}{2}} < 0, \end{cases} \tag{2.3}$$

where ψ is the limiter function. For this limiter function two choices are considered,

$$\psi(\vartheta) = \frac{\vartheta + |\vartheta|}{1 + |\vartheta|}, \tag{2.4}$$

$$\psi(\vartheta) = \max\left(0, \min\left(2, \frac{2}{3} + \frac{1}{3}\vartheta, 2\vartheta\right)\right). \tag{2.5}$$

The first limiter is due to van Leer [16], the second to Koren [14]. The limiters provide a suitable balance between the monotone first order upwind flux and higher order fluxes. Formal statements on accuracy are difficult, due to the built-in switches,

but simple numerical tests for smooth solutions show that the spatial discretizations are approximately second order in the L_2-norm. With both limiters we have $w(t) \geq 0$ whenever $w(0) \geq 0$, together with monotonicity properties such as the total variation diminishing (TVD) property; see, for instance, [15, 17] for more details.

For points adjacent to the boundaries, some of the w_j values that are needed in (2.3) might be missing, and for those, constant extrapolation is used, which means that we switch locally to first order upwind. The above discretizations extend easily to more dimensions on Cartesian meshes.

We observed that the explicit BDF2 method (1.4) is stable with these spatial discretizations up to Courant number 1/2, approximately. This is an experimental bound; precise results can be obtained for the corresponding linear nonlimited discretizations; see [8, 22].

2.2. Implementation. For test purposes we consider the linear 1D convection problem, (2.1), with

$$f(u) = u, \qquad q \equiv 1. \tag{2.6}$$

Note that even for this linear problem the resulting semidiscrete system will be nonlinear, due to the limiter. Therefore, with implicit time integration, some form of Newton iteration is required, which in turn needs an approximation to the Jacobian matrix $A \approx \frac{\partial}{\partial w} F(t, w)$. Within the Newton iteration for (1.3) the matrix $\frac{3}{2}I - \tau A$ is used. The first choice to be considered is the first order upwind approximation

$$A = A_1 \equiv \frac{q}{h} \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}$$

in stencil notation. The resulting iteration scheme is related to the defect correction approach used in [6, 18], for instance. Other choices for the Jacobian approximation can be obtained by realizing that the above flux formulas are nonlinear counterparts of formulas obtained by linearizing around $\vartheta = 1$ (replacement of $\psi(\vartheta)$ in (2.3) by $\psi(1) + \psi'(1)(\vartheta - 1)$). For the van Leer limiter (2.4) this leads to

$$A = A_2 \equiv \frac{q}{4h} \begin{bmatrix} -1 & 4 & -1 & -2 & 0 \end{bmatrix}$$

corresponding to the linear Fromm scheme. For the Koren limiter (2.5) we get

$$A = A_3 \equiv \frac{q}{6h} \begin{bmatrix} -1 & 6 & -3 & -2 & 0 \end{bmatrix},$$

which corresponds to the well-known linear third order upwind-biased scheme. Finally, we also consider the choice $A = 0$, which gives standard functional iteration.

In Table 2.1 the average number of Newton iterations per step are listed for the implicit BDF2 method (1.3) with these various choices and several Courant numbers $\nu = \tau/h$. As starting procedure to calculate w_1, the implicit Euler method was taken. The solutions were calculated on the spatial interval [0,1] with periodicity. The results are given here for an initial block-profile

$$u(x, 0) = \begin{cases} 0 & \text{for } 0 < x < \frac{1}{2}, \\ 1 & \text{otherwise,} \end{cases}$$

and for a smooth initial profile

$$u(x, 0) = \sin^2(\pi x).$$

In this test, the mesh width has been chosen as $h = 1/100$ and output time is $T = \frac{1}{4}$. The convergence criterion for the Newton iteration is that the max-norm of the residual should be less than 10^{-6}. This is rather strict, but accurate solution of the implicit relations is necessary to maintain the monotonicity of the limiting procedure. The maximum number of Newton iterations per step is set to 100. If convergence is still not reached, then the calculations are aborted and $**$ is used for the corresponding entry in Table 2.1. Actually, with $A = 0, \nu = 1$ this means genuine divergence, with the other cases in the table extremely slow convergence.

TABLE 2.1

Linear convection test (2.1),(2.6) *with implicit BDF2 method. The entries are the average number of Newton iterations per step with block-profile and* $\sin^2$*-profile, respectively.*

Limiter	A	$\nu = 1$	$\nu = 1/2$	$\nu = 1/4$
(2.4)	A_1	10.8 – 8.0	8.6 – 6.6	6.9 – 4.5
(2.4)	A_2	13.9 – 11.0	9.3 – 6.5	6.8 – 4.3
(2.4)	0	$**$ – $**$	23.7 – $**$	7.9 – 5.7
(2.5)	A_1	14.7 – 11.0	13.5 – 7.5	8.4 – 4.9
(2.5)	A_3	$**$ – $**$	24.6 – 12.2	9.5 – 5.3
(2.5)	0	$**$ – $**$	$**$ – $**$	9.1 – 6.8

The first observation from Table 2.1 is that the choices $A = A_2$ and $A = A_3$ do not perform well. Especially with (2.5) and $A = A_3$ we get a convergence behavior that is hardly better than with functional iteration. The only choice that does perform reasonably here is $A = A_1$. Moreover, we see that the algebraic relations with limiter (2.4) are easier to solve than with (2.5). It should be noted that the latter gives slightly better results with respect to accuracy, with somewhat less numerical diffusion, but the differences are small. Even with explicit methods the limiter (2.5) is more expensive than (2.4), due to the max-min calculations.

Therefore we consider in the following only the limiter (2.4) with first order upwind approximation for the Jacobian. This implementation seems quite robust. For example, in the above test, if only one time step is performed, $\tau = T$, $\nu = 25$, the Newton process still converges (with 16 iterations for both profiles). Moreover, with first order upwind approximations for the Jacobian the resulting linear system is diagonally dominant, which is of importance in more space dimensions in connection with iterative linear solvers.

However, even with this choice a rather large number of Newton iterations is needed per step. Note that in the above test, the explicit version of the BDF2 method could be used up to Courant number $\nu = 1/2$, and with this explicit method the CPU time per step is much smaller than with the implicit scheme. Therefore, we can conclude that accurately solving the implicit relations with limiting is expensive in terms of CPU time. Some gain could be achieved by setting the tolerance in the convergence criterion to less strict values, but it was observed that even with small Courant numbers negative values arise that are of the same order of magnitude as this tolerance. Numerical tests in 1D with Burgers and Buckley–Leverett equations gave results comparable to those in Table 2.1.

With multidimensional problems we shall adopt the same implementation as above. The Jacobian required in the Newton iteration is approximated by the Jacobian that corresponds to first order upwind spatial discretization.

2.3. Qualitative behavior. The advantage of an implicit time stepping method is the possibility of taking large step sizes without introducing instabilities. However, in several numerical tests we observed that the quality of the implicit solutions are rather poor with large, or even moderately large, Courant numbers if the solution has steep gradients. As an example, consider the 1D Buckley–Leverett equation given by (2.1) with

$$f(u) = \frac{3u^2}{3u^2 + (1-u)^2}, \qquad q \equiv 1, \tag{2.7}$$

and initial block-profile

$$u(x,0) = \begin{cases} 0 & \text{for } 0 < x < \frac{1}{2}, \\ 1 & \text{otherwise.} \end{cases}$$

At the inflow the boundary condition is $u(0,t) = 1$. For the mesh width we take $h = 1/100$ and the endpoint in time is $T = \frac{1}{4}$. In the following figures the numerical solutions are plotted with solid lines. Dashed lines are used to indicate the reference solution that uses the same mesh width h but computed with a very small time step; this corresponds to the exact solution of the semidiscrete system. In Figure 2.1 the implicit (1.3) and explicit (1.4) numerical solutions are plotted as function of x with 100 time steps, $\tau = 1/400$. There is little difference between the two solutions and they are close to the reference solution.

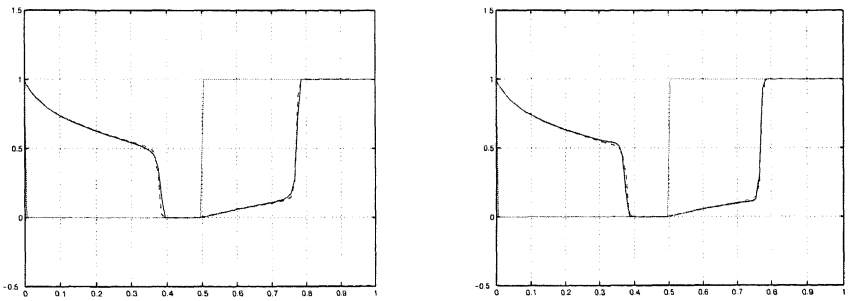

FIG. 2.1. *Numerical solutions at* $T = \frac{1}{4}$ *with Buckley–Leverett equation,* $h = \frac{1}{100}$, $\tau = \frac{1}{400}$. *Left picture: explicit method* (1.4), *right picture: implicit BDF2 method* (1.3).

If the number of time steps is decreased to 50, $\tau = 1/200$, we see from Figure 2.2 that now the explicit solution becomes unstable, but at the same time the implicit solution becomes very inaccurate. Both the shock speed and the shock height are no longer correct.

With linear convection $f(u) = u$, the same phenomenon was observed: if the solution has steep gradients, then the implicit method gives poor results whenever the step sizes are significantly larger than those that can be taken with the explicit method. As we shall see in the following section, this disappointing qualitative behavior of the implicit BDF2 method is due to loss of monotonicity for large step sizes. Although this can be somewhat improved with variants of the implicit BDF2 method (see next section), tests with other implicit schemes of Runge–Kutta or linear multistep-type consistently showed a similar behavior. This means that implicit methods can be

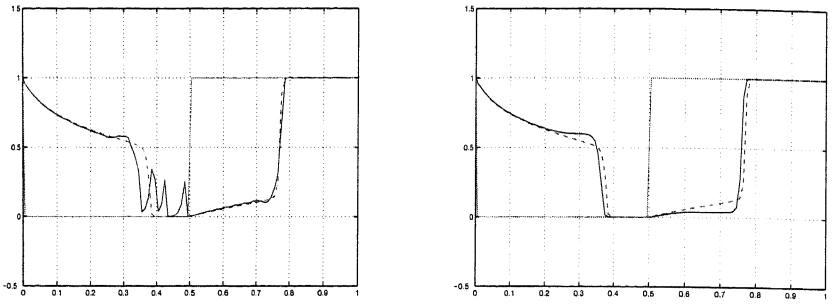

FIG. 2.2. *Numerical solutions at* $T = \frac{1}{4}$ *with Buckley–Leverett equation,* $h = \frac{1}{100}$, $\tau = \frac{1}{200}$. *Left picture: explicit method* (1.4), *right picture: implicit BDF2 method* (1.3).

used well only with large Courant numbers if the solution has little temporal or spatial variation. In case this is valid, an implicit treatment will be more efficient than an explicit one.

In the following sections we shall consider combinations of the implicit and explicit BDF2 methods with the aim of combining the favorable aspects of these two methods.

3. The θ-BDF2 methods. As a first step to combine the implicit and explicit methods we consider the following class of methods, with parameter $\theta \in [0, 1]$:

$$\tfrac{3}{2}w_n - 2w_{n-1} + \tfrac{1}{2}w_{n-2} = \tau F(t_n, \theta w_n + (1-\theta)\overline{w}_n), \tag{3.1}$$

where as before $\overline{w}_n = 2w_{n-1} - w_{n-2}$. Clearly, for $\theta = 0$ and $\theta = 1$ we reobtain the methods (1.3), (1.4), respectively. As we shall see later on, the above methods have order 2 for any choice of θ. Moreover, the methods are A-stable for $\theta \geq \frac{3}{4}$ and consequently we then have unconditional stability for convection-diffusion problems. In fact, if $\theta = \frac{3}{4}$ the stability region consists precisely of the left half complex plane. With this value of θ the method has no inherent damping. For diffusion problems the fully implicit BDF2 method with $\theta = 1$ is therefore preferred. For convection, on the other hand, damping is not necessarily a favorable property and we shall see that $\theta = \frac{3}{4}$ has better monotonicity properties, and consequently it gives a better qualitative behavior for convection problems.

3.1. Positivity properties. We shall consider monotonicity and positivity properties of the θ-BDF2 method (3.1) for linear equations

$$w'(t) = Aw(t) + g(t). \tag{3.2}$$

In the following we shall write $v \geq 0$ for a vector v if all its components are nonnegative. It will be assumed that the matrix $A = (a_{ij}) \in \mathbb{R}^{m\times m}$ has no real positive eigenvalues and

$$a_{ij} \geq 0 \quad (\text{for } i \neq j), \qquad a_{ii} \geq -\alpha \quad (\text{for all } i), \tag{3.3}$$

with $\alpha > 0$. The class of matrices satisfying this condition is denoted by $\mathcal{M}_\alpha$. By a continuity argument (on $\tau > 0$) it can be shown that for any $A \in \mathcal{M}_\alpha$

$$(I - \tau A)^{-1} \geq 0 \quad \text{for all } \tau > 0. \tag{3.4}$$

Further we consider $g(t) \geq 0$ for all $t \geq 0$ in (3.2). Under these assumptions it holds that

$$w(t) \geq 0 \quad \text{whenever } t \geq 0 \text{ and } w(0) \geq 0, \tag{3.5}$$

irrespective of the value of $\alpha \in \mathbb{R}$; see [2]. We note that for linear systems $w'(t) = Aw(t)$ with the property that $Ae = 0$ for $e = (1, 1, \ldots, 1)^T$, it easily follows that the solution will also satisfy a maximum principle

$$\min_i w_i(0) \leq w_j(t) \leq \max_i w_i(0).$$

A rational function φ is said to be *absolutely monotonic* on the interval $[-\gamma, 0]$ if φ and all its derivatives are nonnegative on this interval. It was shown by Bolley and Crouzeix [2] that

$$\varphi(\tau A) \geq 0 \quad \text{for all } A \in \mathcal{M}_\alpha \quad \text{iff} \quad \varphi \text{ is absolutely monotonic on } [-\tau\alpha, 0].$$

This result gives necessary and sufficient conditions for having

$$w_n \geq 0,\ n = 1, 2, \ldots \quad \text{whenever } w_0 \geq 0$$

with one-step time discretizations, such as Runge–Kutta methods. The condition of absolute monotonicity is already necessary for $A = h^{-1}(E - I) \in \mathbb{R}^{m\times m}$ with backward shift operator $E \in \mathbb{R}^{m\times m}$, $\alpha = m = h^{-1}$, provided that the dimension m is sufficiently large. Note that this is simply the semidiscrete system obtained from $u_t + u_x = 0$ with first order upwind discretization in space and homogeneous Dirichlet condition at the inflow boundary. In particular, for the one-step θ-method (1.9), we get the condition on the step size

$$\tau\alpha \leq \frac{1}{1-\theta}.$$

Therefore, with the implicit Euler method there is no step size restriction for positivity. With all other well-known methods we do get a restriction on the allowable step sizes, since unconditional positivity implies that the order of the method is at most 1; see [2].

Application of method (3.1) to the linear system (3.2) gives the recursion

$$w_n = \psi_1(\tau A)w_{n-1} + \psi_2(\tau A)w_{n-2} + \varphi(\tau A)\tau g(t_n) \tag{3.6}$$

with rational functions

$$\psi_1(z) = \frac{4(1 + (1-\theta)z)}{3 - 2\theta z},\ \psi_2(z) = \frac{-(1 + 2(1-\theta)z)}{3 - 2\theta z},\ \varphi(z) = \frac{2}{3 - 2\theta z}. \tag{3.7}$$

Positivity results with arbitrary nonnegative starting values w_0, w_1 were derived by Bolley and Crouzeix [2] for a class of linear multistep methods (see also Spijker [19] and Shu [20] for related results). These results, however, require that $\psi_1(\tau A), \psi_2(\tau A)$, $\varphi(\tau A) \geq 0$, and therefore they are not applicable to the BDF schemes. Due to the fact that $\psi_2(0) = -\frac{1}{3}$ one never has $w_2 \geq 0$ for arbitrary starting values $w_0, w_1 \geq 0$.

We shall derive positivity results for the θ-BDF2 methods (3.1) under the assumption that w_1 is obtained by a suitable starting procedure from w_0, for instance, by Euler's method. The derivation of these results is partly based on discussions with M. van Loon (1996, private communications). Results of this type for general multistep methods seem unknown.

3.2. The threshold function. The positivity results will be obtained by considering the above recursion (3.6) with suitable linear combinations $w_n - \epsilon w_{n-1}$. In this subsection some technical results will be derived. The final result is given in Theorem 3.1. In the following we denote

$$C(z) = \begin{pmatrix} \psi_1(z) & \psi_2(z) \\ 1 & 0 \end{pmatrix}, \qquad V = \begin{pmatrix} 1 & -\epsilon \\ 0 & 1 \end{pmatrix}.$$

Then

$$VC(z)V^{-1} = \begin{pmatrix} \varphi_1(z) & \varphi_2(z) \\ 1 & \epsilon \end{pmatrix}$$

with

$$\varphi_1(z) = \psi_1(z) - \epsilon, \quad \varphi_2(z) = \epsilon\psi_1(z) + \psi_2(z) - \epsilon^2.$$

We shall determine $\epsilon > 0$ such that the entries of $VC(z)V^{-1}$ are absolutely monotonic on the interval $[-\gamma, 0]$ with γ as large as possible. Since the φ_j are fractional linear (i.e., rational with linear denominator and numerator), it follows that this is equivalent to $\varphi_j'(0) \geq 0$ and $\varphi_j(z) \geq 0$ for $z \in [-\gamma, 0]$, $j = 1, 2$.

It is straightforward to verify that $\varphi_j(0) \geq 0$ and $\varphi_j'(0) \geq 0$ for $j = 1, 2$ iff

$$\epsilon_0 \leq \epsilon \leq 1 \qquad \text{with} \quad \epsilon_0 = \max\left(\frac{1}{3}, \frac{3-2\theta}{6-2\theta}\right). \tag{3.8}$$

Further we want $\varphi_j(z) \geq 0$. As we consider $z \leq 0$, this is seen to be equivalent with

$$|z| \leq r(\epsilon), \qquad q(\epsilon)|z| \leq p(\epsilon), \tag{3.9}$$

where

$$r(\epsilon) = \big(2\theta\epsilon + 4(1-\theta)\big)^{-1}(4 - 3\epsilon),$$

$$p(\epsilon) = (1-\epsilon)(3\epsilon - 1), \quad q(\epsilon) = 2\theta\epsilon^2 + 4(1-\theta)\epsilon - 2(1-\theta).$$

The optimal choice for ϵ will depend on the location of the largest zero λ_2 of $q(\epsilon)$. We have

$$q(\epsilon) = 2\theta(\epsilon - \lambda_1)(\epsilon - \lambda_2), \qquad \lambda_{1,2} = -\frac{1-\theta}{\theta} \pm \frac{1}{\theta}\sqrt{1-\theta}.$$

Note that $r(\epsilon)$ is monotonically decreasing in ϵ, and to satisfy $|z| \leq r(\epsilon)$ for $z \in [-\gamma, 0]$ with γ as large as possible, we should take $\epsilon \in [\epsilon_0, 1]$ as small as possible, but of course within the second constraint of (3.9).

First, assume that $\lambda_2 \geq \frac{1}{3}$, that is, $\theta \leq \frac{3}{4}$. Then $q(\epsilon) \leq 0$ for $\epsilon \in [\frac{1}{3}, \lambda_2]$, and thus the second constraint in (3.9) will be automatically satisfied for these ϵ. Therefore we can choose $\epsilon = \epsilon_0$, yielding the restriction $\gamma \leq r(\epsilon_0)$. Thus the optimal γ is given by

$$\gamma(\theta) = \frac{15 - 2\theta}{24 - 26\theta + 4\theta^2}, \qquad \theta \leq \tfrac{3}{4}. \tag{3.10}$$

For the second case $\lambda_2 < \frac{1}{3}$, that is, $\theta \geq \frac{3}{4}$, we get the condition

$$\gamma \leq \max_{\frac{1}{3} \leq \epsilon \leq 1} \min\left(r(\epsilon), \frac{p(\epsilon)}{q(\epsilon)}\right).$$

By some tedious calculations it can be shown that the second constraint is now the dominating one and that the above condition is least restrictive with $\epsilon = [(3-2\theta) + \sqrt{(4\theta-3)}]/(6-2\theta)$. This leads to the optimal γ given by

$$\gamma(\theta) = \frac{3+2\theta-3\sqrt{4\theta-3}}{2(6-5\theta)+2\theta\sqrt{4\theta-3}}, \qquad \theta > \tfrac{3}{4}. \tag{3.11}$$

The threshold function $\gamma(\theta)$ from (3.10), (3.11) is plotted in Figure 3.1. In the next subsections the relevance of this function is discussed.

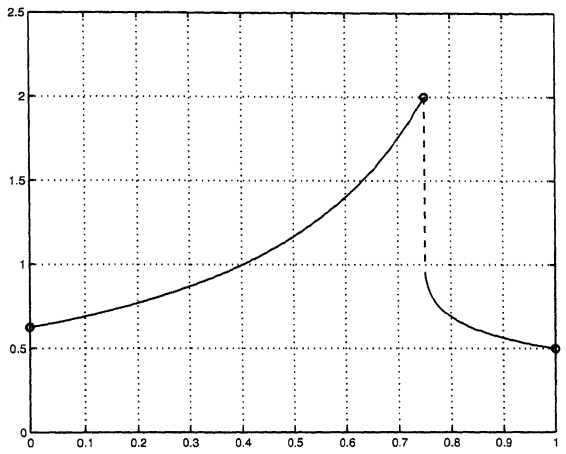

FIG. 3.1. *Positivity threshold function $\gamma(\theta)$ versus $\theta \in [0,1]$ according to* (3.10), (3.11).

3.3. Results for linear systems. From the calculations in section 3.2 it is easy to obtain positivity results for linear systems. In the following, $\gamma(\theta)$ refers to the threshold function given by (3.10), (3.11) and $\epsilon = \epsilon(\theta)$ stands for the optimal value such that $VC(z)V^{-1} \geq 0$ for all $z \in [-\gamma(\theta), 0]$.

THEOREM 3.1. *Consider the linear semidiscrete system* (3.2) *with $A \in \mathcal{M}_\alpha$ and $g(t) \geq 0$. Then $w_n \geq 0$ whenever $\tau\alpha \leq \gamma(\theta)$, $w_0 \geq 0$, and $w_1 - \epsilon w_0 \geq 0$.*

Proof. Denote

$$W_n = \begin{pmatrix} w_n \\ w_{n-1} \end{pmatrix}, \quad C(\tau A) = \begin{pmatrix} \psi_1(\tau A) & \psi_2(\tau A) \\ I & O \end{pmatrix}, \quad G_n = \begin{pmatrix} \varphi(\tau A) g(t_n) \\ 0 \end{pmatrix}.$$

Recursion (3.6) can be written as

$$W_n = C(\tau A) W_{n-1} + \tau G_n.$$

We consider

$$U_n = VW_n \qquad \text{with} \quad V = \begin{pmatrix} I & -\epsilon I \\ O & I \end{pmatrix}.$$

Then

$$U_n = VC(\tau A)V^{-1}U_{n-1} + \tau G_n.$$

From the results in section 3.2 it follows that the entries of the block matrix $VC(\tau A)V^{-1}$ are nonnegative provided that $\tau\alpha \leq \gamma(\theta)$. Further we have $G_n \geq 0$ and $U_0 \geq 0$. Therefore $U_n \geq 0$ for all n, and consequently the same holds for the W_n. □

Whether the condition $w_1 - \epsilon w_0 \geq 0$ is satisfied will of course depend on the starting procedure used to calculate w_1. It will hold if w_1 is calculated from one implicit Euler step. However, if $\theta = 0$ it is more natural to use an explicit Euler step. Since $\epsilon = \frac{1}{2}$ if $\theta = 0$, we then get

$$w_1 - \epsilon w_0 = w_1 - \tfrac{1}{2}w_0 = \tfrac{1}{2}w_0 + \tau A w_0 + \tau g(0),$$

and this is guaranteed to be nonnegative only if $\tau\alpha \leq \frac{1}{2}$. This condition is slightly more restrictive than with the threshold value $\gamma(0) = \frac{5}{8}$ for the explicit BDF2 method itself. This extra time step restriction due to the explicit Euler start can be easily avoided by calculating w_1 by another starting procedure; for example,

$$w_1^* = w_0 + \tau F(t_0, w_0), \quad w_2^* = w_1^* + \tau F(t_1, w_1^*), \quad w_1 = \tfrac{1}{2}(w_0 + w_2^*),$$

in which case it is seen that $w_1 - \frac{1}{2}w_0 = \frac{1}{2}w_2^* \geq 0$ whenever $\tau\alpha \leq 1$.

3.4. Test with the van Leer limiter. The above theoretical results give sufficient conditions for nonnegative solutions with linear problems. To test the relevance with the nonlinear semidiscrete systems obtained with limited spatial discretization (2.2)–(2.4), we consider once more the 1D test equation $u_t + u_x = 0$, $0 \leq t \leq \frac{1}{4}$ with a block-function as initial profile and $h = \frac{1}{100}$. In Table 3.1 we have listed the minimal number of steps $\mathcal{N}(r)$ needed to obtain numerical solutions with minimum larger than -10^{-r} with $r = 3, 4$. As before, the convergence criterion in the Newton iteration was that the max-norm of the residual should be less than 10^{-6} (same results with smaller tolerances), and the starting value w_1 was computed with the implicit Euler method.

TABLE 3.1
Linear convection test (2.1), (2.6) *with* θ*-BDF2 methods. Number of steps required for (almost) nonnegative solutions. The Courant numbers are* $\tau/h = 25/\mathcal{N}$.

θ	0	.7	.74	.75	.76	.8	1
$\mathcal{N}(4)$	40	21	21	24	31	46	75
$\mathcal{N}(3)$	39	21	21	23	26	38	63

For the larger values of θ the number of steps needed to achieve minimal values larger than -10^{-4} and -10^{-3} are relatively far apart; we do not have an explanation for this. We see from Table 3.1 that the theoretical results obtained for the linear class of problems do have a relevance for the van Leer limiter. In particular, if θ is close to 0.75, we can take significantly larger steps than with θ equal to 0 or 1. On the other hand, in this test the largest step sizes could be taken with values of θ slightly less than 0.75, in contrast to Figure 3.1. Also, the allowable step size with $\theta = 0$ seems somewhat larger than one would expect on the basis of Figure 3.1 in comparison with θ equal to 0.75 or 1.

It should be noted that the semidiscrete system obtained here with limiting can be written in the quasi-linear form

$$w_i' = \frac{1}{h}\, q\, a_i(w)(w_{i-1} - w_i) \quad \text{with} \quad 0 \le a_i(w) \le 2;$$

see [11]. The results for the linear systems therefore suggest positivity if the Courant numbers $\nu = q\tau/h$ are not larger than $\frac{1}{2}\gamma(\theta)$. In the above experiment this condition indeed seems sufficient, but it also seems a bit too strict, probably due to the fact that the limiter switches locally to first order upwind discretization for which the condition $\nu \le \gamma(\theta)$ is sufficient (and necessary). Similar to [11] for explicit Runge–Kutta methods, we can conclude that the linear theory does give reasonable qualitative predictions for more difficult, nonlinear situations, but these predictions should not be taken too literally.

As noted before, the θ-BDF2 methods are unconditionally stable for convection-diffusion problems iff $\theta \ge \frac{3}{4}$. Based on the linear theory and practical experience, we do prefer the implicit method with $\theta = \frac{3}{4}$ over the standard fully implicit BDF2 method with $\theta = 1$ for convection. For instance, with the 1D Buckley–Leverett test problem (2.7) the choice $\theta = \frac{3}{4}$ still gives accurate results with $\tau = 1/200$, $h = 1/100$ for which the standard BDF2 method produces qualitatively poor results; see Figure 2.2. Note, however, that basically we still have the same problems as with $\theta = 1$, namely, the high cost of solving the implicit relations and the fact that large Courant numbers lead to loss of monotonicity. Therefore we would like to apply this method with $\theta = \frac{3}{4}$ only if the temporal or spatial variation in the solution is not too large. This will be achieved by considering different values for θ in different parts of the spatial domain.

4. The Θ-BDF2 scheme. To combine implicit and explicit formulas we shall allow θ to vary over the spatial grid. Let in the following $\Theta = \mathrm{diag}(\theta_i)$, where θ_i will correspond with grid point x_i. We consider once more (1.5) but now with specification of θ_i,

$$\left.\begin{aligned} \tfrac{3}{2}w_n - 2w_{n-1} + \tfrac{1}{2}w_{n-2} &= \tau F\big(t_n, \Theta w_n + (I-\Theta)(2w_{n-1} - w_{n-2})\big), \\ \theta_i &= \begin{cases} 0 & \text{if } \nu_i \le \nu^*, \\ \theta^* & \text{otherwise,} \end{cases} \end{aligned}\right\} \tag{4.1}$$

with ν_i denoting the local Courant number at grid point x_i. We choose $\theta^* = \frac{3}{4}$ since this appeared the best choice to aim for with respect to stability and positivity, and $\nu^* = \frac{1}{2}$ since the explicit scheme appears to be stable and positive for $\nu_i \le \frac{1}{2}$.

Note that for 1D problems (2.1) the local Courant number is given by $\nu_i = \tau|q(x_i)f'(w_i)|/h_i$, where h_i is the length of the cell Ω_i around x_i. For multidimensional problems on Cartesian grids, ν_i is taken as the sum of the 1D contributions. When implemented with variable time steps the matrix Θ will also become variable in time even for linear convection with constant velocities. In section 6 we shall consider a simple variable step size selection procedure that essentially limits the max-norm of the displacement $w_n - w_{n-1}$. As a consequence, the scheme will be implicit only in those spatial regions where the velocities are large, but the solution is smooth.

With the above choice for Θ we apply the explicit scheme as much as possible within the stability constraint, and we switch to $\theta = \frac{3}{4}$ elsewhere. With this choice there are abrupt changes in the values of the θ_i over the grid. The effect of this on the accuracy is discussed next.

First we take a look at the truncation error of (4.1). Let $\overline{w}(t_n) = 2w(t_{n-1}) - w(t_{n-2})$. Insertion of the exact solution of (1.2) into the scheme gives

$$\frac{3}{2}w(t_n) - 2w(t_{n-1}) + \frac{1}{2}w(t_{n-2}) = \tau F\big(t_n, \Theta w(t_n) + (I-\Theta)\overline{w}(t_n)\big) + \tau r_n \tag{4.2}$$

with truncation error r_n. By a Taylor expansion we obtain

$$\frac{3}{2}w(t_n) - 2w(t_{n-1}) + \frac{1}{2}w(t_{n-2}) = \tau w'(t_n) - \frac{1}{3}\tau^3 w'''(t_n) + \mathcal{O}(\tau^4),$$

$$\Theta w(t_n) + (I-\Theta)\overline{w}(t_n) = w(t_n) - (I-\Theta)\tau^2 w''(t_n) + \mathcal{O}(\tau^3),$$

and hence

$$r_n = -\frac{1}{3}\tau^2 w'''(t_n) + \tau^2 A_n (I-\Theta) w''(t_n) + \mathcal{O}(\tau^3) \tag{4.3}$$

with Jacobian matrix $A_n = \frac{\partial}{\partial w} F(t_n, w(t_n))$. If a diffusion term is added as in (1.8) this formula for the truncation error is still valid.

The truncation error is often a good measure of the accuracy. Indeed, if we are dealing with a *fixed* ODE system, then the truncation error is $\mathcal{O}(\tau^2)$, reflecting the second order accuracy of the formula. However, in our situation where the ODE system is a semidiscrete PDE, the function F and its derivatives will contain negative powers of the mesh width h. In particular, the term $\tau^2 A_n(I-\Theta)w''(t_n)$ in (4.3) will be only a genuine $\mathcal{O}(\tau^2)$ term if Θ is sufficiently smooth in space. With the choice (4.1) this does not hold. Yet, as we shall see, the accuracy is not affected by this. Instead of looking only at the truncation error, a more refined error analysis is needed. This will be presented in the next section for linear systems.

We note that in (4.1) the linear combination with Θ is taken "within" the function F to ensure mass conservation. The related method

$$\begin{aligned} \tfrac{3}{2}w_n - 2w_{n-1} + \tfrac{1}{2}w_{n-2} &= \tau\Theta F(t_n, w_n) \\ &+ 2(I-\Theta)F(t_{n-1}, w_{n-1}) - \tau(I-\Theta)F(t_{n-2}, w_{n-2}) \end{aligned} \tag{4.4}$$

has smaller truncation errors in general. By Taylor expansion it is easily seen that the truncation error of (4.4) is equal to

$$\frac{1}{\tau}\big(\tfrac{3}{2}w(t_n) - 2w(t_{n-1}) + \tfrac{1}{2}w(t_{n-2})\big) - \Theta w'(t_n) - 2(I-\Theta)w'(t_{n-1})$$

$$+ (I-\Theta)w'(t_{n-2}) = \tau^2\big(\tfrac{2}{3}I - \Theta\big)w'''(t_n) + \mathcal{O}(\tau^3).$$

Therefore, as far as local accuracy is concerned, the form (4.4) is better than (4.1) in general. This is similar to genuine multistep formulas versus the so-called one-leg formulations; see [10]. However, the form (4.4) is not mass conserving.

Suppose that the discrete mass is given by $\mu^T w(t) = \sum \mu_i w_i(t)$ with components μ_i denoting the length of grid cell Ω_i, or area or volume in more dimensions; then mass conservation of the semidiscrete system (1.2) means that $\mu^T w(t)$ should remain constant in time for all starting values $w(0)$. This is equivalent to the condition

$$\mu^T F(t, w) = 0 \qquad \text{for all } t, w.$$

Now, suppose that $\mu^T w_0 = \mu^T w_1$. Then with (4.1) it easily follows by induction that we will have

$$\mu^T w_n = \mu^T w_0 \quad \text{for all } n.$$

With formula (4.4), however, this will hold only if $\Theta = \theta I$, that is, Θ constant over the space. Therefore, even though (4.4) has smaller truncation errors in general, we shall continue with the form (4.1).

5. Global accuracy results. In this section an error analysis for the Θ-BDF2 scheme (4.1) will be presented for linear systems

$$w'(t) = Aw(t) + g(t), \tag{5.1}$$

where the matrix A is assumed to be a finite difference approximation to a convective operator. Stability results with a Θ that varies over the space according to (4.1) are not available. The variation in Θ over space has as a consequence that the standard von Neumann analysis, based on Fourier decompositions, is no longer applicable. In the numerical tests the scheme (4.1) never encountered stability problems. In the following it will therefore simply be assumed that the scheme is stable in a given norm $\|\cdot\|$ for the above linear system, and we will consider global accuracy of the scheme under this assumption.

Let $\varepsilon_n = w(t_n) - w_n$ be the global discretization error. From (1.5) and (4.2) we obtain the error recursion

$$\varepsilon_n - \tfrac{4}{3}\varepsilon_{n-1} + \tfrac{1}{3}\varepsilon_{n-2} = \tfrac{2}{3}Z(\Theta\varepsilon_n + (I-\Theta)(2\varepsilon_{n-1} - \varepsilon_{n-2})) + \tfrac{2}{3}\tau r_n, \tag{5.2}$$

where $Z = \tau A$ and r_n is the local truncation error. This can be written in the more transparent form

$$\varepsilon_n = \Psi_1\varepsilon_{n-1} + \Psi_2\varepsilon_{n-2} + \delta_n \tag{5.3}$$

with matrices

$$\Psi_1 = \tfrac{4}{3}\big(I - \tfrac{2}{3}Z\Theta\big)^{-1}\big(I + Z(I-\Theta)\big), \quad \Psi_2 = -\tfrac{1}{3}\big(I - \tfrac{2}{3}Z\Theta\big)^{-1}\big(I + 2Z(I-\Theta)\big)$$

determining the propagation of previous errors, and with δ_n the local discretization error introduced in the step from t_{n-1} to t_n,

$$\delta_n = \big(I - \tfrac{2}{3}Z\Theta\big)^{-1}\tfrac{2}{3}\tau r_n.$$

For the linear system (5.1) this local discretization error equals

$$\delta_n = (I - \tfrac{2}{3}Z\Theta)^{-1}(-\tfrac{2}{9}\tau^3 w'''(t_n) + \tfrac{2}{3}\tau^2 Z(I-\Theta)w''(t_n)) + \mathcal{O}(\tau^3). \tag{5.4}$$

Here the last term contains only genuine $\mathcal{O}(\tau^3)$ terms; there are no hidden negative powers of h in the constant.

Our tacit stability assumption can now be specified: we assume that from the error recursion (5.3) it can be concluded that

$$\|\varepsilon_n\| \le C\left(\|\varepsilon_0\| + \|\varepsilon_1\| + \sum_{j=2}^{n}\|\delta_j\|\right), \tag{5.5}$$

with $C > 0$ a moderate stability constant, independent of the mesh width h. In particular, this assumption implies that $\|\Psi_1\|$ and $\|\Psi_2\|$ are bounded, from which it easily follows that terms like $\|(I - \frac{2}{3}Z\Theta)^{-1}\|$ and $\|(I - \frac{2}{3}Z\Theta)^{-1}Z\|$ are also bounded (by moderate constants, independent of h).

It thus follows from (5.4) that $\|\delta_n\| = \mathcal{O}(\tau^2)$. Note that this deviates from the estimate that would be obtained in the standard ODE case with a fixed, bounded matrix A. In that case $\|Z\| = \mathcal{O}(\tau)$ and consequently $\|\delta_n\| = \mathcal{O}(\tau^3)$.

Since we are dealing with semidiscrete systems arising from PDEs, where A will contain negative powers of h, the local error δ_n is merely $\mathcal{O}(\tau^2)$ in general. Thus one might expect the global errors to be first order only. However, similar to [12] and [10, sect.V.7] for stiff ODEs, it will be shown here that due to cancellation and damping effects we still have global convergence with order 2.

To demonstrate this second order convergence, define

$$\varepsilon_n^* = \varepsilon_n + \tau^2(I - \Theta)w''(t_n), \tag{5.6}$$

which will turn out to behave more regular than ε_n. By observing that

$$I - \Psi_1 - \Psi_2 = -\tfrac{2}{3}\big(I - \tfrac{2}{3}Z\Theta\big)^{-1}Z,$$

it follows that these transformed errors ε_n^* satisfy the recursion

$$\varepsilon_n^* = \Psi_1\varepsilon_{n-1}^* + \Psi_2\varepsilon_{n-2}^* + \delta_n^*$$

with transformed local error

$$\delta_n^* = \delta_n - \tau^2(I - \Theta)w''(t_n) + \Psi_1\tau^2(I - \Theta)w''(t_{n-1})$$

$$+ \Psi_2\tau^2(I - \Theta)w''(t_{n-2}) = -\big(I - \tfrac{2}{3}Z\Theta\big)^{-1}\tfrac{2}{9}\tau^3 w'''(t_n) + \mathcal{O}(\tau^3).$$

This transformed local error is genuinely of order 3, independent of the mesh width h. The stability argument applied to the recursion of the transformed errors now yields in a standard way order 2 convergence for the ε_n^*. Hence it follows that we also have for our original errors $\|\varepsilon_n\| = \mathcal{O}(\tau^2)$, uniformly for $t_n \leq T$, independent of the mesh width h.

Although this is not a complete convergence proof, since we had to assume that the scheme is stable, it does show that the choice for Θ in (4.1), with abrupt changes in θ_i over the grid, will not lead to an order reduction.

Remark. The above analysis carries over to linear systems

$$w'(t) = Aw(t) + Bw(t) + g(t),$$

where B is a diffusion term that is treated fully implicitly as in (1.8). The transformed errors should then be defined as

$$\varepsilon_n^* = \varepsilon_n + \tau^2 X(I - \Theta)w''(t_n)$$

with $X = (A + B)^{-1}A$. We then obtain second order convergence provided that $X = \mathcal{O}(1)$ uniformly in h.

6. Numerical results. In this section numerical results are presented for a 2D test convection problem arising from the quarter of five spots problem in reservoir simulations; see [7, 18], for example. On a square region $\Omega = [0,1]^2$ we have a source term σ at the point $x = (0,0)$, with volumetric rate $\sigma = \frac{1}{4}\pi$, and a sink term $-\sigma$ at $x = (1,1)$, corresponding to an injection and production well, respectively. It is assumed here that the permeability K and viscosity μ in the actual reservoir problem are constant, say, $K/\mu = 1$. The velocity q and pressure p are then given by

$$q = -\nabla p, \qquad \Delta p + s = 0, \tag{6.1}$$

with $s = s(x)$ describing the sources and sinks and with homogeneous Neumann boundary conditions for p. This determines p up to an additive constant. The resulting convection problem is

$$u_t + \nabla.(qf(u)) = s^+ + s^- u, \tag{6.2}$$

where $s^+ = \max(s, 0)$ and $s^- = \min(s, 0)$. The initial condition is $u \equiv 0$. For the flux function f we shall consider both the linear flux function (2.6) and the Buckley–Leverett flux function (2.7). These are simplified model situations for miscible and immiscible reservoir flows. Illustrations for the behavior of the solutions on $\Omega = [0, 1]^2$ are given in Figure 6.1.

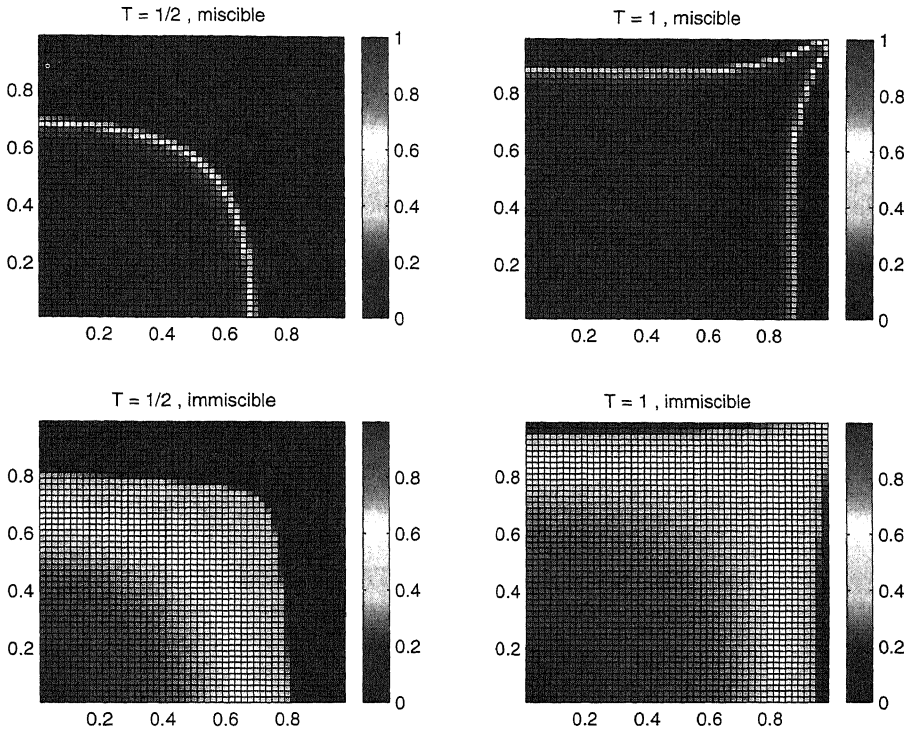

FIG. 6.1. *Numerical solutions at* $T = \frac{1}{2}$ *and* $T = 1$ *on* 50×50 *grids for the miscible model with linear convection (top pictures) and the immiscible model with Buckley-Leverett fluxes (bottom pictures).*

In the numerical tests, the pressure equation was solved using standard second order finite differences on a uniform $m \times m$ grid, mesh width $h = 1/m$, resulting in a first order approximation of the velocities at the cell edges. The injection well was modelled as a source term σ/h^2 in the lower left grid block. Likewise, for the production well we get a sink term $-\sigma w_{m,m}/h^2$ at the upper right grid block. For real reservoir simulations the pressure equations are usually solved in a more sophisticated manner; see, for instance, the contribution of Russell and Wheeler in [7]. With the above test problem the pressure could even be calculated analytically, but numerical solution directly leads to approximations for the velocities that are divergence-free in a discrete fashion. The convection terms in (6.2) are discretized on the same uniform grid with the van Leer limiter as described in section 2; see also Molenaar [18].

The velocities are large only at the corners where the wells are located, approximately $\frac{1}{2}\nabla \log r$ with distance r to the well near $(0,0)$ and $(1,1)$, respectively. Due to the injection at $x = (0,0)$ a front has formed at $t = 0$, which is roughly halfway to the production well at time $t = \frac{1}{2}$; see Figure 6.1. Therefore, in the vicinity of the sharp front we could then use the explicit BDF2 method. Near the wells the solution is smooth, so that there an implicit method could easily be applied. A combination of this is provided by the blended scheme (4.1).

The time integrations in the numerical tests were started with a small initial time step $\tau_0 = \frac{1}{100}h^2$ and subsequently a simple variable step size selection was used,

$$t_{n+1} = t_n + \tau_n, \quad \tau_n = \omega\tau_{n-1}, \quad \omega = \min(2, \mathrm{TOL}\ \|u_n\|_\infty / \|u_n - u_{n-1}\|_\infty). \tag{6.3}$$

The variable step size form of the Θ-BDF2 method was taken as

$$\begin{aligned} &(1+2\omega)w_{n+1} - (1+\omega)^2 w_n + \omega^2 w_{n-1} \\ &= (1+\omega)\tau_n F\big(\Theta_n w_{n+1} + (I-\Theta_n)((1+\omega)w_n - \omega w_{n-1})\big), \end{aligned} \tag{6.4}$$

where the coefficients are similar to the standard implicit BDF2 method; see [9], for example. The initial step is taken with the Euler method, implicit if $\theta^* > 0$ and explicit if $\theta^* = 0$. We note that the step size selection used here is the same as in [18]. Results with a more sophisticated selection, based on an estimate of higher derivatives, gave comparable results. Since the focus here is on the methods and not on step size selections, only the results for the above implementation are presented.

The implicit relations were solved with a modified Newton iteration, using first order upwind discretizations for Jacobian approximations, as described in section 2. In the Newton iteration the initial guess for w_{n+1} in (6.4) was taken as

$$\Theta_n \overline{w}_{n+1} + (I - \Theta_n)\left(\frac{(1+\omega)^2}{1+2\omega} w_n - \frac{\omega^2}{1+2\omega} w_{n-1} + \frac{1+\omega}{1+2\omega}\tau_n F(t_{n+1}, \overline{w}_{n+1})\right).$$

To solve the arising linear systems the Bi-CGSTAB method [23] was used without preconditioning. Note that due to the first order upwind approximation in the Newton iteration the linear system is diagonally dominant. This choice for the linear solver was guided by experiments in [3], where several linear solvers were compared for more general porous media equations. Both the Newton iteration and the Bi-CGSTAB iteration were stopped as soon as the norm of the residue was below 10^{-6}. The norm used here is the maximum norm, as in the step size selection, instead of the more common weighted L_2-norm as in [3], since we also want to resolve the steep solutions gradients accurately.

In Tables 6.1 and 6.2 the statistics are presented for output time $T = \frac{1}{2}$ with the implicit, explicit, and blended scheme (4.1). Along with a CPU timing in seconds on a SUN SPARC4 workstation, also given are the average number of Newton iterations per step (N-it) and the average number of Bi-CGSTAB iterations per Newton iteration (L-it). In the step size selection we used TOL = 0.1 for the implicit and partially implicit scheme, and TOL = 0.01 for the explicit scheme. With the explicit scheme this smaller value of TOL was needed to avoid oscillations (mild instabilities) near the inflow well. With this choice, the accuracy of the various schemes was very similar; the spatial discretization errors are the dominating ones.

Since the errors of the three methods were similar in the experiments, the CPU time is a measure of efficiency here. Obviously this is most favorable with the blended

TABLE 6.1
Statistics for 2D linear convection at $T = \frac{1}{2}$ on 50×50 and 100×100 grid.

	θ^*	ν^*	TOL	Grid	Steps	CPU (s)	N-it	L-it
Implicit	1	0	.1	50×50	218	217	3.34	2.52
Blended	.75	.5	.1	50×50	226	44	0.25	1.14
Explicit	0	0	.01	50×50	2142	131	-	-
Implicit	1	0	.1	100×100	340	2205	3.92	4.19
Blended	.75	.5	.1	100×100	364	413	0.51	2.37
Explicit	0	0	.01	100×100	4016	963	-	-

TABLE 6.2
Statistics for 2D Buckley-Leverett at $T = \frac{1}{2}$ on 50×50 and 100×100 grid.

	θ^*	ν^*	TOL	Grid	Steps	CPU (s)	N-it	L-it
Implicit	1	0	.1	50×50	292	288	3.57	1.55
Blended	.75	.5	.1	50×50	280	65	0.21	1.00
Explicit	0	0	.01	50×50	2985	227	-	-
Implicit	1	0	.1	100×100	531	2318	3.90	1.60
Blended	.75	.5	.1	100×100	498	445	0.24	0.99
Explicit	0	0	.01	100×100	5515	1603	-	-

method. It should be noted, however, that the explicit scheme also performs quite well. With the step size selection described above, the maximal Courant numbers are much larger than unity without introducing instabilities. There are still some small oscillations with the explicit method near the inflow corner, but on the scale of Figure 6.1 these are not visible. Apparently, relatively large Courant numbers can be taken here with the explicit scheme since the velocities are large only near the wells and possible instabilities are transported to the interior domain where they are damped.

However, the step sizes that can be taken with the implicit and blended scheme are much larger, but the fully implicit scheme is not efficient due to the amount of work that has to be performed in solving the algebraic relations. The blended scheme is initially fully explicit, since the step sizes selected according to (6.3) are small if the sharp front is in a region with large velocities. After awhile this scheme becomes implicit near the wells, but then the implicit relations are easy to solve since the solution does not vary much anymore near the wells.

It should be noted that the performance of the explicit scheme will decrease if a local grid refinement is used near the wells. This is often done in practice to capture small-scale geological features. In such a situation a more pronounced advantage of the blended scheme can be expected. This has not been tested, since for the present model problem such a grid refinement would be very artificial.

Numerical tests with small diffusion terms added to the convection equation, implemented as in (1.8), did give very similar results. Finally it should be noted that our implementation of the blended scheme in the above experiments was not very sophisticated. For example, the whole function F was calculated in each Newton iteration step, whereas this is not necessary inside the region where $\Theta = 0$ (more precisely, at those grid points where $\theta_i = 0$ for the grid point itself, its neighbors and

adjacent points). For ease of programming it was decided to use the same subroutines as for the fully implicit scheme.

In view of these experiments, we conclude that the blended scheme works very well for problems of the above type, where there are locally large velocities. If the size of the velocities is more or less uniform, and the solution is not very smooth, an explicit treatment of the convective terms will be more efficient in general. Fully implicit methods seem to be efficient only if the solution is sufficiently smooth in space, but with convection dominated flows steep gradients in the solution are the generic case.

REFERENCES

[1] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 797–823.

[2] C. Bolley and M. Crouzeix, *Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques*, RAIRO Anal. Numer., 12 (1978), pp. 237–245.

[3] J. G. Blom, J. G. Verwer, and R. A. Trompert, *A comparison between direct and iterative methods to solve the linear systems arising from a time-dependent 2D groundwater flow model*, Comp. Fluid Dyn., 1 (1993), pp. 95–113.

[4] M. Blunt and B. Rubin, *Implicit flux limiting schemes for petroleum reservoir simulation*, J. Comput. Phys., 102 (1992), pp. 194–210.

[5] M. Crouzeix, *Une méthode multipas implicite-explicite pour l' approximation des équations d' évolution paraboliques*, Numer. Math., 35 (1980), pp. 257–276.

[6] J.-A. Désidéri and P. W. Hemker, *Convergence analysis of the defect-correction iteration for hyperbolic problems*, SIAM J. Sci. Comput., 16 (1995), pp. 88–118.

[7] R. E. Ewing, ed., *The Mathematics of Reservoir Simulation*, Frontiers Appl. Math. 1, SIAM, Philadelphia, 1984.

[8] J. Frank, W. Hundsdorfer, and J. G. Verwer, *On the stability of implicit-explicit linear multistep methods*, Appl. Numer. Math., 25 (1997), pp. 193–205.

[9] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I—*Nonstiff Problems*, Springer-Verlag, Berlin, 1987.

[10] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II—*Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.

[11] W. Hundsdorfer, B. Koren, M. van Loon, and J. G. Verwer, *A positive finite-difference advection scheme*, J. Comput. Phys., 117 (1995), pp. 35–46.

[12] W. Hundsdorfer and B. I. Steininger, *Convergence of linear multistep and one-leg methods for stiff nonlinear initial value problems*, BIT, 31 (1991), pp. 124–143.

[13] W. Hundsdorfer and R. Trompert, *Method of lines and direct discretization: A comparison for linear advection*, Appl. Numer. Math., 13 (1994), pp. 469–490.

[14] B. Koren, *A robust upwind discretization for advection, diffusion and source terms*, in Numerical Methods for Advection-Diffusion Problems, C. B. Vreugdenhil and B. Koren, eds., Notes Numer. Fluid Mech. 45, Vieweg, Braunschweig, 1993.

[15] D. Kröner, *Numerical Schemes for Conservation Laws*, Wiley and Teubner, Chichester, UK, Stuttgart, 1997.

[16] B. van Leer, *Towards the ultimate conservative difference scheme* II. *Monotonicity and conservation combined in a second order scheme*, J. Comput. Phys., 14 (1974), pp. 361–370.

[17] R. J. LeVeque, *Numerical methods for conservation laws*, Lectures Math. ETH Zürich, Birkhäuser-Verlag, Basel, 1992.

[18] J. Molenaar, *Multigrid methods for high-order accurate fully implicit simulations of flow in porous media*, in Proceedings of the Second ECCOMAS Conference on Numerical Methods in Engineering, J.-A. Désidéri, P. Le Tallec, E. Onate, J. Périaux, and E. Stein, eds., John Wiley, 1996.

[19] M. N. Spijker, *Contractivity in the numerical solution of initial boundary value problems*, Numer. Math., 42 (1983), pp. 271–290.

[20] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 1073–1084.

[21] J. M. VARAH, *Stability restrictions on second order, three level finite difference schemes for parabolic equations*, SIAM J. Numer. Anal., 17 (1980), pp. 300–309.

[22] J. G. VERWER, J. G. BLOM, AND W. HUNDSDORFER, *An implicit-explicit approach for atmospheric transport-chemistry problems*, Appl. Numer. Math., 20 (1996), pp. 191–209.

[23] H. A. VAN DER VORST, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.