Implicit-explicit Time Stepping with Spatial Discontinuous Finite Elements

W.H. Hundsdorfer, J. Jaffré

Implicit-Explicit Time Stepping
with Spatial Discontinuous Finite Elements

Willem Hundsdorfer

CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Jérôme Jaffré

INRIA-Rocquencourt, BP 105, 789153 Le Chesnay Cédex, France

Abstract

In this paper a combination of discontinuous, piecewise linear, finite elements with implicit-explicit time stepping is considered for convection-reaction equations. Combined with low order quadrature rules, this leads to convenient schemes. We shall consider the effect of such low order quadrature rules on accuracy and stability for one-dimensional problems.

1   Introduction

Discontinuous finite elements were originally introduced and analyzed for linear neutron transport equations, see Hill & Reed [21], Lesaint & Raviart [19]. The method was then extended to time-dependent nonlinear porous media equations with Euler time stepping by Chavent, Cohen, Jaffré and others, see the references in [6]. Further development and analysis for CFD problems was performed primarily by Cockburn, Shu and co-workers, see [7] for an extended bibliography. The main interest of discontinuous finite elements is to provide locally conservative schemes in a finite element framework, with the possibility to enforce monotonicity and TVD properties by limiting.

In this paper we consider convection-reaction equations in one spatial dimension (1D), and we investigate the combination of discontinuous piecewise linear finite elements in space with time stepping by a BDF2 type scheme which is explicit for convection and implicit for reactions, thus allowing stiffness in the reaction terms. The analysis presented here includes the effect of low order numerical quadrature in the scheme, a question that was raised already in [6, 15].

Consider the 1D convection-reaction equation

$$u_t + f(u)_x = g(u) \tag{1.1}$$

with $0 \leq x \leq L$, $0 \leq t \leq T$ and given initial and boundary data. In this paper we shall deal with the so-called method of lines approach where first a spatial discretization is performed,

giving an approximation $w(t) = u_h(t)$ in a finite dimensional space. Then a fully discretized solution $w_n = u_{h,n}$ at time level t_n is obtained by a suitable time stepping method.

We consider a partitioning of the interval $[0, L]$ into cells $\mathcal{C}_i = (x_{i-1/2}, x_{i+1/2})$ with midpoints x_i and mesh widths $h_i = x_{i+1/2} - x_{i-1/2}$. For discretization in space the discontinuous Galerkin scheme is considered with $w = u_h \in M^1$, the space of piecewise linear discontinuous functions on the given partitioning. We shall refer to this as the DG^1 scheme. The discretization is specified by the requirement that for all test functions $v \in M^1$ it holds that

$$\int_{\mathcal{C}_i} \partial_t w(x,t)\, v(x)\, dx - \int_{\mathcal{C}_i} f(w(x,t))\, \partial_x v(x)\, dx - f^*_{i-1/2}(t)\, v(x^+_{i-1/2}) +$$

$$+ f^*_{i+1/2}(t)\, v(x^-_{i+1/2}) \;=\; \int_{\mathcal{C}_i} g(w(x,t))\, v(x)\, dx. \tag{1.2}$$

The numerical fluxes $f^*_{i+1/2}(t)$ are constructed by monotone flux splitting using the left and right values $w(x^-_{i+1/2}, t)$ and $w(x^+_{i+1/2}, t)$. Examples of such flux splittings are the Godunov and Engquist-Osher splittings. For scalar equations with monotonically increasing f these choices all reduce to upstream fluxes $f^*_{i+1/2}(t) = f(w(x^-_{i+1/2}, t))$.

For a better understanding of this discretization, we consider the decomposition $M^1 = M^0 + \widetilde{M}^1$ into mean values and first moments,

$$w(x, t) = \overline{w}_i(t) + \widetilde{w}_i(t)\varphi_i(x) \qquad \text{on } \mathcal{C}_i, \tag{1.3}$$

with $\varphi_i(x) = 2h_i^{-1}(x - x_i)$ restricted to $\mathcal{C}_i$. Applying the above weak formulation with test functions $v = \mathbb{1}_{\mathcal{C}_i} \in M^0$ and $v = \varphi_i \in \widetilde{M}^1$, we obtain

$$h_i \frac{d\overline{w}_i(t)}{dt} = f^*_{i-1/2}(t) - f^*_{i+1/2}(t) + \int_{\mathcal{C}_i} g(w(x,t))\, dx, \tag{1.4a}$$

$$\int_{\mathcal{C}_i} \varphi_i(x)^2\, dx\, \frac{d\widetilde{w}_i(t)}{dt} = \frac{2}{h_i} \int_{\mathcal{C}_i} f(w(x,t))\, dx - f^*_{i-1/2}(t) - f^*_{i+1/2}(t) +$$

$$+ \int_{\mathcal{C}_i} g(w(x,t))\varphi_i(x)\, dx. \tag{1.4b}$$

A very convenient formula is obtained if midpoint quadrature (1-point Gauss) is applied to the integrals on the right hand side. To approximate $\int_{\mathcal{C}_i} \varphi_i(x)^2 dx$ in (1.4b) we can use Simpson's rule, which is exact here, giving weight $\frac{1}{3}h_i$, or the trapezoidal rule giving weight h_i. For the moment we leave this choice open by using weight factor $\frac{1}{3\kappa}h_i$ with parameter $\kappa > 0$. So the resulting formula reads

$$\frac{d\overline{w}_i(t)}{dt} = \frac{1}{h_i}\left(f^*_{i-1/2}(t) - f^*_{i+1/2}(t)\right) + g(\overline{w}_i(t)), \tag{1.5a}$$

$$\frac{d\widetilde{w}_i(t)}{dt} = \frac{-3\kappa}{h_i}\left(f^*_{i-1/2}(t) - 2f(\overline{w}_i(t)) + f^*_{i+1/2}(t)\right). \tag{1.5b}$$

Due to the midpoint quadrature the reaction term g is only evaluated in the mean values $\overline{w}_i$, which is attractive when using an implicit method in time for this term. It will be shown that the resulting spatial discretization is still second order accurate.

2

Trapezoidal quadrature for the integral on the left hand side of (1.4b) ($\kappa = \frac{1}{3}$) was considered in Chavent & Jaffré [6] based on the practical observation that this choice leads to better stability properties than Simpson's rule ($\kappa = 1$) with explicit Euler time stepping. Here we shall consider second order time stepping to achieve better accuracy and intrinsic stability, but it will turn out that this observation for explicit Euler still carries over to some extent.

To introduce the time integration formula, let $w(t) \in \mathbb{R}^{2m}$ be the vector containing all $\overline{w}_i(t)$ and $\widetilde{w}_i(t)$ values, with m being the number of grid points in space. The above semi-discrete system can be written as

$$\frac{d}{dt}w(t) = F(w(t)) + G(w(t)) \tag{1.6}$$

with $F, G : \mathbb{R}^{2m} \to \mathbb{R}^{2m}$ containing the convection and reaction terms, respectively. Let $t_n = n\tau$ with step size $\tau > 0$ and let w_n be the fully discrete approximation to $w(t_n)$. In this paper we shall primarily consider the following implicit-explicit BDF2 type scheme

$$w_n = \tfrac{4}{3}w_{n-1} - \tfrac{1}{3}w_{n-2} + \tfrac{2}{3}\tau\Big(F(2w_{n-1} - w_{n-2}) + G(w_n)\Big), \tag{1.7}$$

where explicit convection is combined with implicit reaction. The starting value $w_0 = w(0)$ is given, and $w_1 \approx w(t_1)$ will be computed with the implicit-explicit Euler type scheme

$$w_1 = w_0 + \tau F(w_0) + \tau G(w_1). \tag{1.8}$$

Note that with the spatial discretization (1.5) the implicit algebraic relations in (1.7),(1.8) will only involve the mean values $\overline{w}_i$. In this respect, the computational complexity of the scheme is comparable to standard finite volume methods.

The implicit-explicit scheme (1.7) was proposed by Crouzeix [10] and Varah [23]. It is second order accurate in time. The reaction terms G are calculated with the standard implicit BDF2 scheme which has excellent stability and damping properties, see for instance [13, 16], and therefore arbitrary stiffness of the reaction term is allowed. In case the reaction term is nonstiff, then G can of course also be treated explicitly. For the combined implicit-explicit scheme (1.7) stability is usually determined by the CFL restriction for the explicit convection part, see [10, 12]. With the present DG^1 spatial discretization this CFL restriction will be discussed by means of Fourier transformation (von Neumann analysis) for linear equations.

To avoid spatial oscillations and negative values near shocks we shall use a standard slope limiter introduced by van Leer [18]. Numerical tests will show that the resulting scheme then still has order 2 accuracy, approximately. For convection equations it will be shown that the combined space-time scheme (1.5),(1.7) is total variation diminishing (TVD) for Courant numbers up to 5/16. It should be noted that the above explicit scheme that is used in (1.7) for the convection part, falls outside the class of so-called TVD multi-step schemes introduced by Shu [22]. We use this explicit scheme because it combines in a natural way with the implicit BDF2 method, which is one of the best known implicit time step formulas. In the discussion on time stepping for convection a TVD multi-step scheme of Shu will be included for comparison.

The contents of this paper is as follows. In Section 2 a detailed error analysis is presented for the spatial discretization (1.5) applied to the linear convection equation $u_t + u_x = 0$ on a uniform grid. Linear stability results with nonuniform grids are discussed in Section 3. In Section 4 spatial consistency of (1.4) and (1.5) is analyzed for nonlinear equations. Stability and TVD properties of the time integration formula applied to convection equations are considered in Section 5. This final section also contains numerical tests and concluding remarks.

2   Linear convection on uniform grid

In this section a detailed accuracy analysis is presented for the linear convection equation

$$u_t + u_x = 0, \tag{2.1}$$

with $0 \le x \le L$ and $0 \le t \le T$, where the solution is assumed to be smooth. We consider here only the errors of the spatial discretization on a uniform grid. Nonuniform grids and time discretization will be discussed in later sections. In the following let $h = L/m$ and $x_i = (i - \frac{1}{2})h$ for $i = 1, 2, ..., m$. For the linear equation (2.1), the DG^1 scheme without limiting reads

$$\frac{d}{dt}\overline{w}_i(t) = \frac{1}{h}\Big((\overline{w}_{i-1}(t) + \widetilde{w}_{i-1}(t)) - (\overline{w}_i(t) + \widetilde{w}_i(t))\Big),$$

$$\frac{d}{dt}\widetilde{w}_i(t) = \frac{-3\kappa}{h}\Big((\overline{w}_{i-1}(t) + \widetilde{w}_{i-1}(t)) - 2\overline{w}_i(t) + (\overline{w}_i(t) + \widetilde{w}_i(t))\Big).$$

It is assumed that we have either a Dirichlet condition at the left boundary, $\overline{w}_0(t) + \widetilde{w}_0(t) = \gamma_0(t)$, or periodicity in space, $\overline{w}_0(t) + \widetilde{w}_0(t) = \overline{w}_m(t) + \widetilde{w}_m(t)$. It is important to note that there is no question here on how to approximate the integral $\int_{\mathcal{C}_i} f(w(x,t))dx$ in (1.4). Since f is linear any quadrature rule will give a contribution $h\overline{w}_i(t)$.

For brevity we shall often omit in the notation the explicit t-dependence of $\overline{w}_i(t)$ and $\widetilde{w}_i(t)$. The subindex i will always refer to the position in space; to indicate time levels, as in formula (1.7), the subindex n will be used. Separating the $\overline{w}_i$ and $\widetilde{w}_i$ terms we obtain the following form for the semi-discrete system,

$$\frac{d}{dt}\overline{w}_i = \frac{1}{h}(\overline{w}_{i-1} - \overline{w}_i) + \frac{1}{h}(\widetilde{w}_{i-1} - \widetilde{w}_i), \tag{2.2a}$$

$$\frac{d}{dt}\widetilde{w}_i = \frac{-3\kappa}{h}(\overline{w}_{i-1} - \overline{w}_i) - \frac{3\kappa}{h}(\widetilde{w}_{i-1} + \widetilde{w}_i). \tag{2.2b}$$

2.1   Von Neumann stability

In this subsection we briefly consider stability of the spatial discretization, assuming periodicity in space. Writing the semi-discrete system (2.2) as $w'(t) = Aw(t)$ in $\mathbb{R}^{2m}$, Fourier transformation gives $\widehat{w}'(t) = \widehat{A}\widehat{w}(t)$ in $\mathbb{C}^2$ with

$$\widehat{A} = \frac{1}{h}\begin{pmatrix} e^{i\phi_k} - 1 & e^{i\phi_k} - 1 \\ -3\kappa(e^{i\phi_k} - 1) & -3\kappa(e^{i\phi_k} + 1) \end{pmatrix},$$

where $\phi_k = 2\pi kh$ $(k = 1, 2, ..., m = 1/h)$, $i = \sqrt{-1}$. Stability of the semi-discrete system, in the sense of von Neumann, requires that the eigenvalues of the matrix $\widehat{A}$ have a nonpositive real part. The eigenvalues λ are given by

$$h\lambda_\pm = \tfrac{1}{2}(e^{i\phi_k} - 1) - \tfrac{3}{2}\kappa(e^{i\phi_k} + 1) \pm \sqrt{\tfrac{1}{4}((e^{i\phi_k} - 1) - 3\kappa(e^{i\phi_k} + 1))^2 + 6\kappa(e^{i\phi_k} - 1)},$$

see also [4], where we take the branch of the square root such that Re $\sqrt{z} \ge 0$ for $z \in \mathbb{C}$. These eigenvalues are plotted in the following Figure 2.1 with stars for λ_+ and circles for the λ_- eigenvalues.
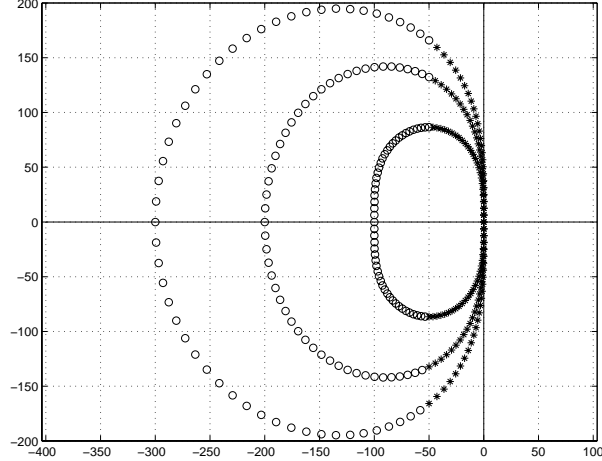
4

FIGURE 2.1. Eigenvalues with $h = 1/50$ for $\kappa = 1$ (outer), $\kappa = 2/3$ (middle) and $\kappa = 1/3$ (inner).

Note that for small k, that is for the low-frequency Fourier modes, we have

$$\widehat{A} = \frac{1}{h}\begin{pmatrix} 0 & 0 \\ 0 & -6\kappa \end{pmatrix} + \mathcal{O}(1)$$

with corresponding $\lambda_+ = \mathcal{O}(1)$ and $\lambda_- = -\frac{1}{h}6\kappa + \mathcal{O}(1)$. Therefore, the λ_- eigenvalues are for small k primarily associated with the equation for the $\widetilde{w}_i$, and hence a perturbation in this equation will be quickly damped. This indicates that we can allow much larger errors in the second equation in (2.2) than in the first one, and this will be important in the accuracy analysis.

It is obvious from Figure 2.1 that the eigenvalues are all in the left half plane and it can be shown that the scheme is stable in the L_2-norm for the linear equation on uniform grid with periodicity in space. This follows from the observation that the two eigenvectors of $\widehat{A}$ are orthogonal with respect to the inner product $(\xi, \zeta) = \bar{\xi}_1\zeta_1 + (3\kappa)^{-1}\bar{\xi}_2\zeta_2$ for $\xi, \zeta \in \mathbb{C}^2$.

A more general, direct proof of L_2-stability will be given in the next section. It will not rely on the eigenvalues. Knowledge of the eigenvalues will be used in a later section to obtain stability restrictions for explicit time integration schemes; for the moment we note already that according to Figure 2.1 the choice $\kappa = \frac{1}{3}$ will yield a more relaxed CFL restriction than with the larger values of κ.

In the remainder of this section it will simply be assumed that the semi-discrete system (2.2) is stable in some suitable norm $\|\cdot\|$.

2.2 Spatial errors

If we insert exact solution values $\overline{u}_i(t), \widetilde{u}_i(t)$, associated with the exact PDE solution $u(x,t)$, in the semi-discrete system (2.2), we get

$$\frac{d}{dt}\overline{u}_i = \frac{1}{h}(\overline{u}_{i-1} - \overline{u}_i) + \frac{1}{h}(\widetilde{u}_{i-1} - \widetilde{u}_i) + \overline{\delta}_i, \tag{2.3a}$$

5

$$\frac{d}{dt}\widetilde{u}_i = \frac{-3\kappa}{h}(\overline{u}_{i-1} - \overline{u}_i) - \frac{3\kappa}{h}(\widetilde{u}_{i-1} + \widetilde{u}_i) + \widetilde{\delta}_i, \tag{2.3b}$$

with *(spatial) truncation errors* $\overline{\delta}_i$ and $\widetilde{\delta}_i$. Expressions for these truncation errors are easily found by a Taylor series expansion.

To find a relation for the time evolution of the *spatial discretization errors*

$$\overline{\varepsilon}_i(t) = \overline{u}_i(t) - \overline{w}_i(t) \quad \text{and} \quad \widetilde{\varepsilon}_i(t) = \widetilde{u}_i(t) - \widetilde{w}_i(t),$$

we subtract (2.2) from (2.3) to obtain

$$\frac{d}{dt}\overline{\varepsilon}_i = \frac{1}{h}(\overline{\varepsilon}_{i-1} - \overline{\varepsilon}_i) + \frac{1}{h}(\widetilde{\varepsilon}_{i-1} - \widetilde{\varepsilon}_i) + \overline{\delta}_i, \tag{2.4a}$$

$$\frac{d}{dt}\widetilde{\varepsilon}_i = \frac{-3\kappa}{h}(\overline{\varepsilon}_{i-1} - \overline{\varepsilon}_i) - \frac{3\kappa}{h}(\widetilde{\varepsilon}_{i-1} + \widetilde{\varepsilon}_i) + \widetilde{\delta}_i. \tag{2.4b}$$

The standard convergence analysis now goes as follows. Writing this error evolution equation as

$$\varepsilon'(t) = A\varepsilon(t) + \delta(t) \tag{2.5}$$

with vectors $\varepsilon = (\overline{\varepsilon}_1, \cdots, \overline{\varepsilon}_m, \widetilde{\varepsilon}_1, \cdots, \widetilde{\varepsilon}_m)^T$, $\delta = (\overline{\delta}_1, \cdots, \overline{\delta}_m, \widetilde{\delta}_1, \cdots, \widetilde{\delta}_m)^T$ in $\mathbb{R}^{2m}$, we thus have

$$\varepsilon(t) = \exp(tA)\varepsilon(0) + \int_0^t \exp((t-s)A)\delta(s)\,ds.$$

Stability tells us that $\|\exp(tA)\| \le C$ for all $t \ge 0$. Hence we obtain the bound

$$\|\varepsilon(t)\| \le C\|\varepsilon(0)\| + C\,t\,\max_{0 \le s \le t}\|\delta(s)\|.$$

Consequently, if there is no initial error, $\varepsilon(0) = 0$, and if we know for the truncation errors that $\|\delta(t)\| \le Ch^p$ (consistency of order p), then we have $\|\varepsilon(t)\| \le Kh^p$ on any bounded time interval $[0, T]$ (convergence of order p).

2.3   Consistency

Since we are working in a finite element setting, the natural choice for the exact solution values is

$$\overline{u}_i(t) = \frac{1}{h}\int_{\mathcal{C}_i} u(x,t)dx, \qquad \widetilde{u}_i(t) = \frac{3}{h}\int_{\mathcal{C}_i} \varphi_i(x)u(x,t)dx,$$

so that $\overline{u}_i(t)\mathbb{1}_{\mathcal{C}_i}(x) + \widetilde{u}_i(t)\varphi_i(x)$ corresponds to the L_2-projection of $u(x,t)$ on M^1. Note that there is a scaling factor $3/h$ here because equation (1.4b) is divided by this factor to obtain (1.5b) when using exact quadrature. Further note that with this choice we have

$$\overline{u}_i(t) = u(x_i,t) + \tfrac{1}{24}h^2 u_{xx}(x_i,t) + \cdots, \tag{2.6a}$$

$$\widetilde{u}_i(t) = \tfrac{1}{2}hu_x(x_i,t) + \tfrac{1}{80}h^3 u_{xxx}(x_i,t) + \cdots, \tag{2.6b}$$

so up to the $\mathcal{O}(h^2)$ level we could also have taken the point values $u(x_i,t)$ and $\tfrac{1}{2}hu_x(x_i,t)$, which would be a natural choice in a finite difference setting. By a Taylor series expansion (calculations are given later on), we find from (2.3) that the truncation errors are

$$\overline{\delta}_i(t) = -\tfrac{1}{12}h^2 u_{xxx}(x_i,t) + \mathcal{O}(h^3), \tag{2.7a}$$

6

$$\widetilde{\delta}_i(t) = -\tfrac{1}{2}hu_{xx}(x_i,t) + \tfrac{1}{4}\kappa h^2 u_{xxx}(x_i,t) + \mathcal{O}(h^3). \qquad (2.7b)$$

This does not look very promising: it seems that the scheme is first order only, and that we might just as well take $\kappa = 0$, that is the first-order upwind Godunov finite volume scheme (if $\widetilde{w}_i(0) \equiv 0$). Fortunately, the scheme is not as bad as it appears here at first sight. It will be shown that we have in fact convergence with order 2 if $\kappa > 0$ and even with order 3 in the mean values if $\kappa = 1$.

2.4 Improved error estimation

The truncation errors are only an indication for the true, global error. To show that they are too pessimistic in the above, we shall use the following lemma.

Lemma 2.1. Consider the error relation (2.4) with $\kappa > 0$. Assume that the relation is stable and

$$\overline{\varepsilon}_i(0) = \mathcal{O}(h^q), \quad \widetilde{\varepsilon}_i(0) = \psi_0(x_i)h^{q-1} + \mathcal{O}(h^q),$$
$$\overline{\delta}_i(t) = \mathcal{O}(h^q), \quad \widetilde{\delta}_i(t) = \psi(x_i,t)h^{q-1} + \mathcal{O}(h^q)$$

with smooth functions ψ_0, ψ, and with remainder terms $\mathcal{O}(h^q)$ uniformly in x_i and t. Then

$$\overline{\varepsilon}_i(t) = \mathcal{O}(h^q), \quad \widetilde{\varepsilon}_i(t) = \exp\Big(\frac{-6\kappa t}{h}\Big)\psi_0(x_i)h^{q-1} + \mathcal{O}(h^q)$$

uniformly on bounded intervals $[0, T]$.

Proof. Let $\lambda = -6\kappa/h$ and introduce $\overline{\varepsilon}_i^*(t) = \overline{\varepsilon}_i(t)$,

$$\widetilde{\varepsilon}_i^*(t) = \widetilde{\varepsilon}_i(t) - \frac{1}{6\kappa}\psi(x_i,t)h^q - \exp(\lambda t)\,\psi_0(x_i)\,h^{q-1}.$$

Then with these modified errors we easily find that $\overline{\varepsilon}_i^*(0), \widetilde{\varepsilon}_i^*(0) = \mathcal{O}(h^q)$ and

$$\frac{d}{dt}\overline{\varepsilon}_i^* = \frac{1}{h}(\overline{\varepsilon}_{i-1}^* - \overline{\varepsilon}_i^*) + \frac{1}{h}(\widetilde{\varepsilon}_{i-1}^* - \widetilde{\varepsilon}_i^*) - \exp(\lambda t)\,\psi_0'(x_i)\,h^{q-1} + \mathcal{O}(h^q),$$

$$\frac{d}{dt}\widetilde{\varepsilon}_i^* = \frac{-3\kappa}{h}(\overline{\varepsilon}_{i-1}^* - \overline{\varepsilon}_i^*) - \frac{3\kappa}{h}(\widetilde{\varepsilon}_{i-1}^* + \widetilde{\varepsilon}_i^*) + \mathcal{O}(h^q).$$

Note that

$$\int_0^t \exp(\lambda s)\,\psi_0'(x_i)\,h^{q-1}\,ds = \mathcal{O}(h^q).$$

The standard convergence approach, with the variation of constants formula, applied to the above relation now gives $\overline{\varepsilon}_i^*(t), \widetilde{\varepsilon}_i^*(t) = \mathcal{O}(h^q)$, and consequently we have $\overline{\varepsilon}_i(t) = \mathcal{O}(h^q)$ and $\widetilde{\varepsilon}_i(t) - \exp(\lambda t)\psi_0(x_i)h^{q-1} = \mathcal{O}(h^q)$. $\qquad\square$

First consider $\psi_0 = 0$. Then the ε^*-errors in the above proof can be interpreted as the errors with respect to *modified* exact solution values: instead of $\overline{u}_i, \widetilde{u}_i$ one should consider $\overline{u}_i$ and $\widetilde{u}_i - \frac{1}{6\kappa}\psi(x_i,t)h^q$. Application to (2.6), (2.7) thus shows that we should regard

$$\overline{u}_i = u(x_i,t) + \mathcal{O}(h^2), \qquad \widetilde{u}_i = \tfrac{1}{2}hu_x(x_i,t) + \frac{1}{12\kappa}h^2 u_{xx}(x_i,t) + \mathcal{O}(h^3).$$

7

Then, with respect to these modified exact solution values we obtain the following truncation errors (calculations follow),

$$\overline{\delta}_i = -\frac{1}{12}\left(1 - \frac{1}{\kappa}\right)h^2 u_{xxx}(x_i, t) + \mathcal{O}(h^3), \quad \widetilde{\delta}_i = \mathcal{O}(h^2). \tag{2.8}$$

Thus we see that we do indeed have second order convergence. In case $\kappa = 1$ we can apply the same argument once more to show third-order convergence of the mean values $\overline{w}_i$ (super convergence).

Theorem 2.2. Consider the DG[1] spatial discretization (2.2) with a given norm and assume the discretization is stable. Then we have convergence with order 2 for any $\kappa > 0$. If $\kappa = 1$ we have convergence of order 3 for the mean values.

Proof. Consider

$$\overline{u}_i(t) = u(x_i, t) + \xi h^2 u_{xx}(x_i, t),$$

$$\widetilde{u}_i(t) = \tfrac{1}{2}h u_x(x_i, t) + \eta h^2 u_{xx}(x_i, t) + \zeta h^3 u_{xxx}(x_i, t),$$

with constants ξ, η, ζ to be specified later. Then, with all the following tems u, u_x, u_{xx} evaluated in (x_i, t),

$$\frac{1}{h}(\overline{u}_{i-1} - \overline{u}_i) = -u_x + \tfrac{1}{2}h u_{xx} - (\tfrac{1}{6} + \xi)h^2 u_{xxx} + \cdots,$$

$$\frac{1}{h}(\widetilde{u}_{i-1} - \widetilde{u}_i) = -\tfrac{1}{2}h u_{xx} + (\tfrac{1}{4} - \eta)h^2 u_{xxx} + \cdots,$$

$$\frac{1}{h}(\widetilde{u}_{i-1} + \widetilde{u}_i) = u_x + (2\eta - \tfrac{1}{2})h u_{xx} + (2\zeta - \eta + \tfrac{1}{4})h^2 u_{xxx} + \cdots.$$

Using the differential equation $u_t + u_x = 0$ to express the time derivatives of $\overline{u}_i$ and $\widetilde{u}_i$ in terms of spatial derivatives, we now easily obtain

$$\overline{\delta}_i = (\eta - \tfrac{1}{12})h^2 u_{xxx} + \mathcal{O}(h^3),$$

$$\widetilde{\delta}_i = (6\kappa\eta - \tfrac{1}{2})h u_{xx} + \left(\tfrac{1}{4}\kappa - (1 + 3\kappa)\eta + 3\kappa(2\zeta - \xi)\right)h^2 u_{xxx} + \mathcal{O}(h^3).$$

This gives the Taylor expansions (2.7), (2.8) for $\eta = 0$ and $\eta = \frac{1}{12\kappa}$, respectively. In accordance to Lemma 2.1, we thus have second order convergence whenever $\kappa > 0$.

If $\kappa = 1$, then the choice $\eta = \frac{1}{12}$ shows third order convergence towards the modified exact solution values. Note that ξ, ζ are still free; these will affect the numerical result only through the initial values. For the preparation of the initial data we can simply take (2.6), that is $\eta = 0$, since application of Lemma 2.1 with $\psi_0(x_i) = \frac{1}{12}u_{xx}(x_i, 0)$ shows that this will not affect the third order result for the mean values $\overline{w}_i$. $\square$

The above theorem gives a generalization of a result of Lesaint & Raviart [19] and Cockburn [7, Thm.3.2] who showed second order convergence for the global errors in case $\kappa = 1$. Some consistency results for $\kappa \neq 1$ and different interpretations of the exact solution values were given in a technical note of Chavent & Cockburn [4].

In Section 4 we shall discuss consistency for nonlinear equations and nonuniform grids. From the above it is seen that consistency should be regarded in a generalized sense: one should allow for modified exact solution values.

2.5   Numerical illustration

We consider the linear convection equation (2.1) with $T = \frac{1}{2}$, $L = 2$ and with smooth solution

$$u(x,t) = \sin^2(\pi(x - t)), \qquad 0 \le t \le \tfrac{1}{2}, \ 0 \le x \le 2.$$

Here we present some accuracy results in the L_2-norm for the spatial discretizations with $\kappa = \frac{1}{3}, \frac{2}{3}, 1$. Accuracy in the max-norm was also measured, but this gave comparable results.

Figure 2.2 contain plots of the discrete L_2-norms for the mean value errors $\|\overline{\varepsilon}\|_2$ and the global errors $\|\varepsilon\|_2$,

$$\|\overline{\varepsilon}\|_2 = \Big(\sum_{j=1}^{m} h_j |\overline{\varepsilon}_j|^2\Big)^{1/2}, \qquad \|\varepsilon\|_2 = \Big(\sum_{j=1}^{m} h_j |\overline{\varepsilon}_j|^2 + \frac{1}{3} h_j |\widetilde{\varepsilon}_j|^2\Big)^{1/2}.$$

The errors are plotted as function of $1/m$ where m is the number of grid points.

The third order convergence for the mean values with $\kappa = 1$ is clearly visible in Figure 2.2. When considering the global spatial error $\|\varepsilon\|_2$, the convergence rate is the same for the three values of κ, but with $\kappa = 1$ we have the smallest error constants.

Accuracy results for the DG[1] spatial discretization on nonuniform grids are presented in Figure 2.3. With the nonuniform grids the mesh widths h_j were chosen randomly: first random numbers $\omega_j \in (0,1)$ were generated and subsequently we used $h_j = L\omega_j / \sum_{k=1}^{m} \omega_k$ to obtain a discretization grid for $0 \le x \le L$. With these random grids 10 runs were performed for each value of m; in the figure the means of the errors of these runs are plotted together with the estimated standard deviations, indicated by error bars. It should be noted that the expectation of the maximal mesh width is now $2/m$.

The results on these random grids are more or less the same as on the uniform grids. Apparently we still have third order convergence for the mean values if $\kappa = 1$. Consistency on nonuniform meshes is discussed later but those results will not be detailed enough to capture this feature.

Finally we also present results with limiting in Figure 2.4. To obtain the results for the spatial discretizations we used the explicit time stepping (1.7) with very small time steps, so that temporal errors are negligible. A limiter can be applied after each time step as follows: given approximations $\overline{w}_i, \widetilde{w}_i$ at time level t_n the first moments $\widetilde{w}_i$ are adjusted according to

$$\widetilde{w}_i := \mathcal{M}(\widetilde{w}_i, \overline{w}_{i+1} - \overline{w}_i, \overline{w}_i - \overline{w}_{i-1}) \qquad (2.9)$$

with minmod function $\mathcal{M}$ defined by

$$\mathcal{M}(a,b,c) = \begin{cases} \text{sign}(a) \min(|a|, |b|, |c|) & \text{if } \text{sign}(a) = \text{sign}(b) = \text{sign}(c), \\ 0 & \text{otherwise.} \end{cases}$$
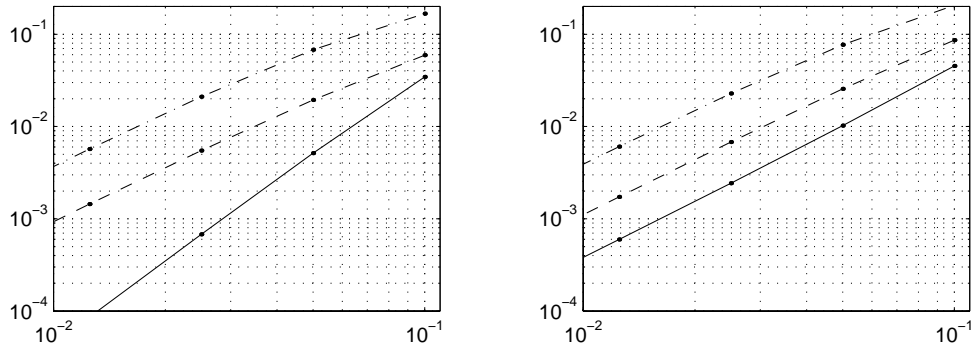
FIGURE 2.2. L_2-errors versus $1/m$ with mean-values (left) and global solution (right) for linear convection on uniform grids. Results for $\kappa = 1$ (solid), $\kappa = 2/3$ (dashed) and $\kappa = 1/3$ (dash-dot).
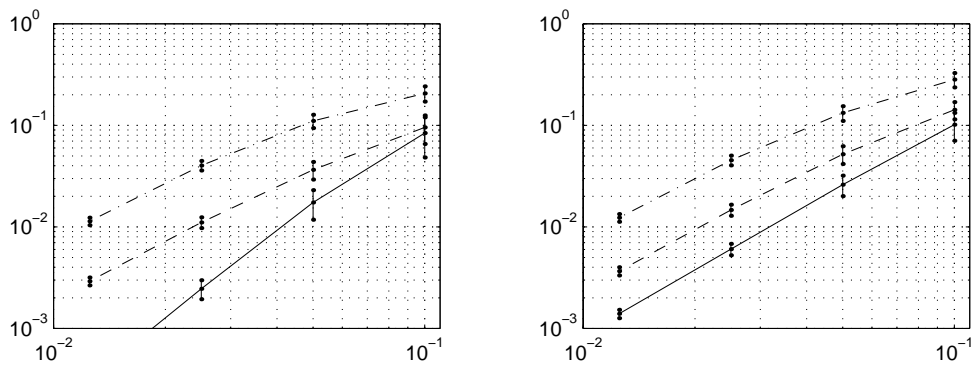

FIGURE 2.3. L_2-errors versus $1/m$ with mean-values (left) and global solution (right) for linear convection on random grids. Results for $\kappa = 1$ (solid), $\kappa = 2/3$ (dashed) and $\kappa = 1/3$ (dash-dot).
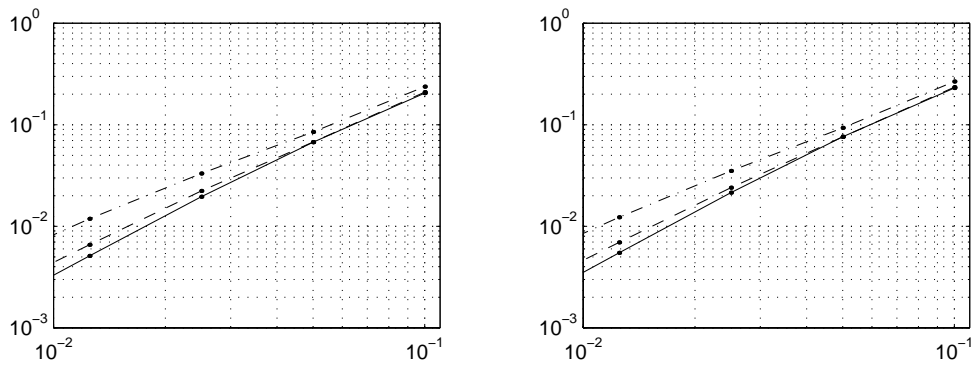

FIGURE 2.4. L_2-errors versus $1/m$ with mean-values (left) and global solution (right) for linear convection, uniform grids with limiting. Results for $\kappa = 1$ (solid), $\kappa = 2/3$ (dashed) and $\kappa = 1/3$ (dash-dot).

10

This is a standard limiter due to van Leer [18], see also [20] or [7] for other choices. For the smooth solution in this example such limiting is not necessary, in fact it will decrease the accuracy near extremal values. We take it into consideration since limiting is necessary for nonsmooth solutions, and the results here indicate the accuracy that can be expected in regions where the solution is smooth with extremal values. Note that in a region where the solution is smooth without extremal values the generic case is $\text{sign}(\widetilde{w}_i) = \text{sign}(\overline{w}_{i+1} - \overline{w}_i) = \text{sign}(\overline{w}_i - \overline{w}_{i-1})$ with $\widetilde{w}_i$ having the smallest modulus, so there the limiter will not be active.

The result of Figure 2.4 shows that with limiting there is very little difference between the three values of κ. Apparently the accuracy is determined here by the limiter, not by the underlying scheme. Results with limiting on nonuniform meshes were also obtained but these were very similar to the uniform case.

3 Stability for linear convection – nonuniform grids

In this section stability of the spatial discretization (1.5) is considered for the linear convection equation $u_t + u_x = 0$ on nonuniform grids. We consider the following norm on the space M^1,

$$\|w\| = \left(\sum_{j=1}^{m} h_j \overline{w}_j^2 + \frac{1}{3\kappa} h_j \widetilde{w}_j^2\right)^{1/2}. \tag{3.1}$$

If $\kappa = 1$ this is just the L_2-norm $\|\cdot\|_2$ on M^1. For other $\kappa > 0$ it is equivalent to this norm,

$$\sqrt{\kappa}\|w\| \le \|w\|_2 \le \|w\|.$$

Due to this norm equivalence, stability in the norm $\|\cdot\|$ also implies stability in the L_2-norm.

Denote $\langle y, z\rangle = \sum_{j=1}^{m} y_j z_j$, $|z| = \langle z, z\rangle^{1/2}$ and let E stand for the backward shift operator

$$(Ez)_i = z_{i-1} \qquad \text{for } z \in \mathbb{R}^m.$$

If we have periodicity in space, we use the convention $z_0 = z_m$. With a Dirichlet condition at the inflow boundary $x = 0$ we put $z_0 = 0$. This last choice corresponds with a homogeneous Dirichlet condition, which we consider here since stability deals with the difference between solutions, and we assume there is no perturbation on the given boundary data.

From (1.5), with $f(u) = u$, $g(u) = 0$, we see that

$$\frac{1}{2}\frac{d}{dt}\|w\|^2 = \sum_{j=1}^{m} h_j \overline{w}_j \frac{d\overline{w}_j}{dt} + \frac{1}{3\kappa} h_j \widetilde{w}_j \frac{d\widetilde{w}_j}{dt} =$$

$$= \langle \overline{w}, (E - I)\overline{w} + (E - I)\widetilde{w}\rangle - \langle \widetilde{w}, (E - I)\overline{w} + (E + I)\widetilde{w}\rangle =$$

$$= -|\overline{w}|^2 - |\widetilde{w}|^2 + \langle \overline{w} - \widetilde{w}, E(\overline{w} + \widetilde{w})\rangle.$$

Note that due to the choice of norm, the parameter κ no longer appears in this expression. Therefore we can follow the arguments of Cockburn [7, Lem.3.1] to prove stability. For completion these arguments are repeated here.

11

Consider

$$\vartheta(w) = \Big(\sum_{j=1}^{m'} [w]_{j-1/2}^2 \Big)^{1/2} \tag{3.2}$$

where $[w]_{j-1/2}$ denotes the jump at $x_{j-1/2}$,

$$[w]_{j-1/2} = -w(x_{j-1/2}^-) + w(x_{j-1/2}^+) = -(\overline{w}_{j-1} + \widetilde{w}_{j-1}) + (\overline{w}_j - \widetilde{w}_j),$$

and where

$$\begin{cases} m' = m \quad \text{and} \quad \overline{w}_0 = \overline{w}_m, \ \widetilde{w}_0 = \widetilde{w}_m \quad \text{for the periodicity condition,} \\ m' = m+1, \ \overline{w}_j = \widetilde{w}_j = 0 \ (j=0, m+1) \quad \text{for the Dirichlet condition.} \end{cases}$$

Consider first the case with periodicity condition. Then we have

$$\vartheta(w)^2 = |E(\overline{w} + \widetilde{w}) - (\overline{w} - \widetilde{w})|^2 =$$

$$= |E(\overline{w} + \widetilde{w})|^2 + |\overline{w} - \widetilde{w}|^2 - 2\langle \overline{w} - \widetilde{w}, E(\overline{w} + \widetilde{w})\rangle =$$

$$= 2|\overline{w}|^2 + 2|\widetilde{w}|^2 - 2\langle \overline{w} - \widetilde{w}, E(\overline{w} + \widetilde{w})\rangle.$$

Likewise for the Dirichlet condition we obtain

$$\vartheta(w)^2 = |E(\overline{w} + \widetilde{w}) - (\overline{w} - \widetilde{w})|^2 + |\overline{w}_m + \widetilde{w}_m|^2 =$$

$$= 2|\overline{w}|^2 + 2|\widetilde{w}|^2 - 2\langle \overline{w} - \widetilde{w}, E(\overline{w} + \widetilde{w})\rangle.$$

By combining the above formulas, the following result follows.

Theorem 3.1. For the spatial discretization (1.5) of the linear convection equation (2.1), it holds that

$$\frac{d}{dt}\|w(t)\|^2 = -\vartheta(w(t))^2.$$

$\square$

It should be emphasized that the above result is a straightforward modification of Lemma 3.1 in Cockburn [7]. It is a modification in the sense that we allow $\kappa \neq 1$ and Dirichlet conditions. The result of Cockburn is more general since it also applies to higher order DG schemes.

Results in other norms, for instance max-norm or L_1-norm, are lacking. However, numerical experiments do indicate that the scheme is also stable in these norms. As an illustration we plotted in the Figures 3.1, 3.2 the time evolution of $\|\exp(tA)\|$ with the L_2-norm and max-norm. Here A is the difference operator on $\mathbb{R}^{2m}$ such that (2.2) assumes the form $w'(t) = Aw(t)$ in the vector notation (1.6) (see also (2.4),(2.5)). A homogeneous Dirichlet boundary condition is imposed at $x = 0$. In Figure 3.1 the mesh width is uniform $h = 1/m$. For the plots in Figure 3.2 the mesh widths h_j were chosen randomly as in Section 2.5 to obtain a partitioning for $0 \leq x \leq L = 1$. Results are presented here for $\kappa = 1, \frac{2}{3}$ and $m = 50$. With $\kappa = \frac{1}{3}$ and other choices of m the results were similar.

The numerical results with the random grids are not very dependent on the actual grid choice. Results for two repeated experiments are also plotted in Figure 3.2, indicated with

dotted lines; these are not well visible since for a large part these results nearly coincide with the original experiment, illustrating the insensitivity with respect to the grid choice. Further we note that in Figure 3.1 with $\kappa = 2/3$ we do not have $\|\exp(tA)\|_2 \leq 1$; for small t we get a bound slightly larger than 1. The result for the scaled norm (3.1) implies that $\|\exp(tA)\|_2 \leq \sqrt{1/\kappa}$, which gives a bound 1.22 approximately, but it is obvious from the plots that this is an over-estimation. The fact that $\|\exp(tA)\|$ tends to 0 rapidly for $t > 1$ is a consistency property. Since we use a homogeneous inflow condition for $u_t + u_x = 0$, $0 \leq x \leq 1$, the exact solution operator for this equation equals 0 for $t > 1$.

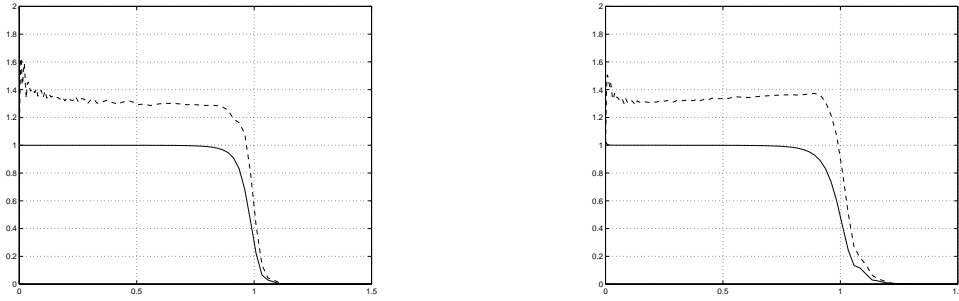

FIGURE 3.1. Stability bounds with $m = 50$ on uniform grid. Plots of $\|\exp(tA)\|$ as function of t in max-norm (dashed) and L_2-norm (solid). Left picture for $\kappa = 1$, right picture for $\kappa = 2/3$.
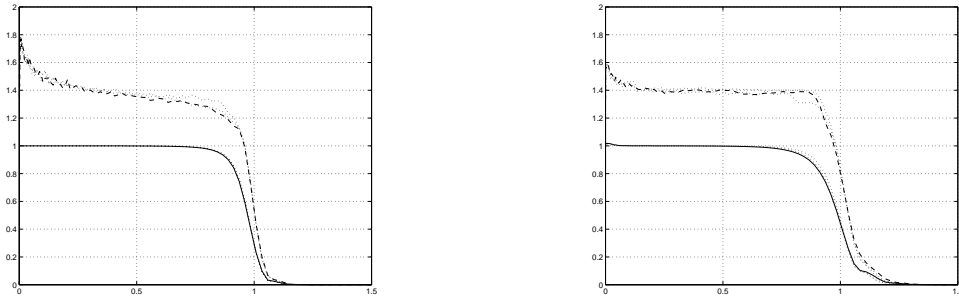

FIGURE 3.2. Example of stability bounds with $m = 50$ on random grids. Plots of $\|\exp(tA)\|$ as function of t in max-norm (dashed) and L_2-norm (solid). Dotted lines indicate repeated experiments. Left picture for $\kappa = 1$, right picture $\kappa = 2/3$.

Finally we note that the above stability results for linear convection are easily generalized to include a stable linear reaction term. For the equation

$$u_t + u_x = \gamma u, \quad \gamma < 0,$$

the semi discrete system will read $w'(t) = Aw(t) + Bw(t)$ with $Bw = \gamma[\overline{w}, 0]^T$ in scheme (1.5) or with $Bw = \gamma w = \gamma[\overline{w}, \widetilde{w}]^T$ using higher order quadrature to evaluate the g terms in (1.4).

In both cases it follows directly that

$$\frac{1}{2}\frac{d}{dt}\|w\|^2 \le -|\overline{w}|^2 - |\widetilde{w}|^2 + \langle \overline{w} - \widetilde{w}, E(\overline{w} + \widetilde{w})\rangle = -\vartheta(w)^2.$$

4 Consistency – nonlinear equations

In this section we shall consider the spatial truncation errors up to the $\mathcal{O}(h^2)$ level for non-linear equations (1.1) on nonuniform grids. For linear equations on nonuniform grids without limiting, convergence can then be proved as in Section 2 by using the above stability results. At the end of this section numerical results are presented that show the relevance for nonlinear equations with limiting. We shall consider both the scheme (1.5), with midpoint quadrature, and the scheme that uses higher order (Simpson) quadrature.

4.1 Midpoint quadrature

Consider scheme (1.5) for a scalar convection-reaction equation on a nonuniform grid with mesh widths $h_i = x_{i+1/2} - x_{i-1/2}$ and $x_i = \frac{1}{2}(x_{i-1/2} + x_{i+1/2})$. Let $\theta_i = h_{i-1}/h_i$ and assume that θ_i is bounded by some fixed constant.

For convenience it is assumed that $f'(u) \ge 0$ for any u, so that the numerical fluxes are given explicitly by

$$f^*_{i+1/2} = f(\overline{w}_i + \widetilde{w}_i).$$

The local truncation errors of scheme (1.5) are thus given by

$$\overline{\delta}_i = \frac{d}{dt}\overline{u}_i - \frac{1}{h_i}\Big(f(\overline{u}_{i-1} + \widetilde{u}_{i-1}) - f(\overline{u}_i + \widetilde{u}_i)\Big) - g(\overline{u}_i), \tag{4.1a}$$

$$\widetilde{\delta}_i = \frac{d}{dt}\widetilde{u}_i + \frac{3\kappa}{h_i}\Big(f(\overline{u}_{i-1} + \widetilde{u}_{i-1}) - 2f(\overline{u}_i) + f(\overline{u}_i + \widetilde{u}_i)\Big). \tag{4.1b}$$

Guided by the previous analysis for linear equations on uniform grids we consider

$$\overline{u}_i(t) = u(x_i,t) + h_i^2\phi(x_i,t), \qquad \widetilde{u}_i(t) = \tfrac{1}{2}h_i u_x(x_i,t) + h_i^2\psi(x_i,t), \tag{4.2}$$

with smooth functions ϕ,ψ to be determined later. Then

$$\overline{u}_i(t) + \widetilde{u}_i(t) = u(x_{i+1/2},t) - \tfrac{1}{8}h_i^2 u_{xx}(x_i,t) + h_i^2\Big(\phi(x_i,t) + \psi(x_i,t)\Big) + \mathcal{O}(h_i^3),$$

and hence we have

$$f(\overline{u}_i(t) + \widetilde{u}_i(t)) = f(u(x_{i+1/2},t)) +$$
$$+ h_i^2 f'(u(x_i,t))\Big(-\tfrac{1}{8}u_{xx}(x_i,t) + \phi(x_i,t) + \psi(x_i,t)\Big) + \mathcal{O}(h_i^3),$$

$$f(\overline{u}_{i-1}(t) + \widetilde{u}_{i-1}(t)) = f(u(x_{i-1/2},t)) +$$
$$+ \theta_i^2 h_i^2 f'(u(x_i,t))\Big(-\tfrac{1}{8}u_{xx}(x_i,t) + \phi(x_i,t) + \psi(x_i,t)\Big) + \mathcal{O}(h_i^3).$$

Further it is easily found that

$$f(\overline{u}_i(t)) = f(u(x_i,t)) + f'(u(x_i,t))h_i^2\phi(x_i,t) + \mathcal{O}(h_i^4),$$

14

$$g(\overline{u}_i(t)) = g(u(x_i, t)) + \mathcal{O}(h_i^2),$$

$$\frac{d}{dt}\overline{u}_i = -f(u(x_i, t))_x + g(u(x_i, t)) + \mathcal{O}(h_i^2),$$

$$\frac{d}{dt}\widetilde{u}_i = \tfrac{1}{2}h_i\Big(-f(u(x_i, t))_{xx} + g(u(x_i, t))_x\Big) + \mathcal{O}(h_i^2).$$

For the local truncation errors we thus obtain the following expressions, with all terms on the right hand side evaluated in the point (x_i, t),

$$\overline{\delta}_i = -(1 - \theta_i^2)h_i f'(u)\Big(\tfrac{1}{8}u_{xx} - \phi - \psi\Big) + \mathcal{O}(h_i^2), \tag{4.3a}$$

$$\widetilde{\delta}_i = (-\tfrac{1}{2} + \tfrac{3}{4}\kappa)h_i f(u)_{xx} + \tfrac{1}{2}h_i g(u)_x + 3\kappa(1 + \theta_i^2)h_i f'(u)(-\tfrac{1}{8}u_{xx} + \phi + \psi) -$$

$$- \ 6\kappa h_i f'(u)\phi + \mathcal{O}(h_i^2).$$

This last expression can be written as

$$\widetilde{\delta}_i = -\tfrac{1}{2}h_i f'(u)u_{xx} + (-\tfrac{1}{2} + \tfrac{3}{4}\kappa)h_i f''(u)u_x^2 + \tfrac{1}{2}h_i g'(u)u_x + 6\kappa h_i f'(u)\psi +$$

$$+ \ 3\kappa(1 - \theta_i^2)h_i f'(u)(\tfrac{1}{8}u_{xx} - \phi - \psi) + \mathcal{O}(h_i^2). \tag{4.3b}$$

If $f'(u) \neq 0$, or more generally if $f'(u)^{-1}f''(u)u_x^2$ and $f'(u)^{-1}g'(u)u_x$ are bounded, then we can choose ψ, ϕ such that

$$\psi = \frac{1}{12\kappa}u_{xx} - (\frac{1}{8} - \frac{1}{12\kappa})f'(u)^{-1}f''(u)u_x^2 - \frac{1}{12\kappa}f'(u)^{-1}g'(u)u_x,$$

$$\phi = \tfrac{1}{8}u_{xx} - \psi,$$

to obtain the second order estimate $\overline{\delta}_i, \widetilde{\delta}_i = \mathcal{O}(h_i^2)$.

The assumption on $f'(u)$ seems essential: if we consider the degenerate case $f \equiv 0$ then scheme (1.5) reduces to

$$\frac{d}{dt}\overline{w}_i = g(\overline{w}_i), \qquad \frac{d}{dt}\widetilde{w}_i = 0$$

and here we always have an order 1 truncation error for the second component $\widetilde{\delta}_i$. Note however that this error does not affect $\overline{\delta}_i$. In general, for nonlinear f, the choice $\varphi = \tfrac{1}{8}u_{xx} - \psi$ leads to $\overline{\delta}_i = \mathcal{O}(h_i^2)$, $\widetilde{\delta}_i = \mathcal{O}(h_i)$. Although Lemma 2.1 has been formulated only for linear problems, it does suggest that also here we will have second order convergence of the mean values.

4.2 Simpson quadrature

We compare the truncation errors of scheme (1.5) with those obtained for the scheme with Simpson quadrature. Assuming as before that $f'(u) \geq 0$, this quadrature rule leads to local errors

$$\overline{\delta}_i = \frac{d}{dt}\overline{u}_i - \frac{1}{h_i}\Big(f(\overline{u}_{i-1} + \widetilde{u}_{i-1}) - f(\overline{u}_i + \widetilde{u}_i)\Big) + \tfrac{1}{6}\Big(g(\overline{u}_i - \widetilde{u}_i) + 4g(\overline{u}_i) + g(\overline{u}_i + \widetilde{u}_i)\Big), \tag{4.4a}$$

$$\widetilde{\delta}_i = \frac{d}{dt}\widetilde{u}_i + \frac{\kappa}{h_i}\Big(3f(\overline{u}_{i-1} + \widetilde{u}_{i-1}) - f(\overline{u}_i - \widetilde{u}_i) - 4f(\overline{u}_i) + 2f(\overline{u}_i + \widetilde{u}_i)\Big) +$$
$$+ \frac{\kappa}{2}\Big(g(\overline{u}_i - \widetilde{u}_i) - g(\overline{u}_i + \widetilde{u}_i)\Big). \qquad (4.4b)$$

For the moment we still leave $\kappa > 0$ unspecified, although $\kappa = 1$ is a very natural choice here. Insertion of

$$g(\overline{u}_i - \widetilde{u}_i) + g(\overline{u}_i + \widetilde{u}_i) = 2g(\overline{u}_i) + g''(\overline{u}_i)\widetilde{u}_i^2 + \cdots,$$

into (4.4a) and of

$$f(\overline{u}_i - \widetilde{u}_i) + f(\overline{u}_i + \widetilde{u}_i) = 2f(\overline{u}_i) + f''(\overline{u}_i)\widetilde{u}_i^2 + \cdots,$$

$$g(\overline{u}_i - \widetilde{u}_i) - g(\overline{u}_i + \widetilde{u}_i) = -2g'(\overline{u}_i)\widetilde{u}_i + \cdots,$$

into (4.4b), shows by comparison with scheme (1.5) that we now have the truncation errors

$$\overline{\delta}_i = \overline{\delta}_i^* + \mathcal{O}(h_i^2),$$

$$\widetilde{\delta}_i = \widetilde{\delta}_i^* - \tfrac{1}{4}\kappa h_i f''(u)u_x^2 - \tfrac{1}{2}\kappa h_i g'(u)u_x + \mathcal{O}(h_i^3)$$

where $\overline{\delta}_i^*$ and $\widetilde{\delta}_i^*$ are the truncation errors of scheme (1.5). From the expressions (4.3) we thus find that

$$\overline{\delta}_i = -(1 - \theta_i^2)h_i f'(u)\Big(\tfrac{1}{8}u_{xx} - \phi - \psi\Big) + \mathcal{O}(h_i^2), \qquad (4.5a)$$

$$\widetilde{\delta}_i = -\tfrac{1}{2}h_i f'(u)u_{xx} + \tfrac{3}{4}(\kappa - 1)h_i f''(u)u_x^2 - \tfrac{1}{2}(\kappa - 1)h_i g'(u)u_x + 6\kappa h_i f'(u)\psi +$$
$$+ 3\kappa(1 - \theta_i^2)h_i f'(u)(\tfrac{1}{8}u_{xx} - \phi - \psi) + \mathcal{O}(h_i^2), \qquad (4.5b)$$

where as in (4.3) the terms on the right hand side are all evaluated in (x_i, t).

If $\kappa \neq 1$ the structure of the local errors is the same as with midpoint quadrature, but $\kappa = 1$ is a natural choice here since it corresponds to Simpson quadrature for all integrals in (1.4). With this choice $\kappa = 1$ we now obtain a uniform order 2 estimate for the (modified) truncation errors, also if $f'(u) = 0$. This is achieved with

$$\psi = \frac{1}{12}u_{xx} \quad \text{and} \quad \phi = \frac{1}{24}u_{xx}.$$

In conclusion, the use of higher order quadrature gives simpler truncation errors, but this will only lead to an improved consistency for the first moments in case $f'(u) = 0$. If $f'(u) \neq 0$ the structure of the truncations errors is the same as with scheme (1.5).

4.3 Numerical illustration

In the following figure the numerical solutions with midpoint and Simpson quadrature are plotted for Burgers equation

$$u_t = f(u)_x, \qquad f(u) = \tfrac{3}{4}u^2,$$

on $0 \leq t \leq T = \frac{1}{2}$, $0 \leq x \leq L = 2$, with initial step function $u_0(x) = 0$ for $0 < x \leq \frac{1}{2}$ and $u_0(x) = 1$ elsewhere, and with the inflow condition $u(0, t) = 1$. A uniform grid was used with $h = 1/25$. We note that the values of L and T are chosen as before and the scaling factor

$\frac{3}{4}$ in $f(u)$ serves only to keep the significant solution features within the spatial range. The results at the output time $T = \frac{1}{2}$ are given in Figure 4.1 for the slope limited DG^1 scheme with midpoint quadrature and with Simpson quadrature. The results here are with $\kappa = 1$ since Simpson quadrature makes little sense otherwise.

The results are very similar, Simpson quadrature only gives slightly better results near $f'(u) = 0$. This is in agreement with the above analysis. It should be noted however that these numerical results are with nonsmooth solutions and limiting, and therefore the previous analysis only gives an indication for the actual results here.
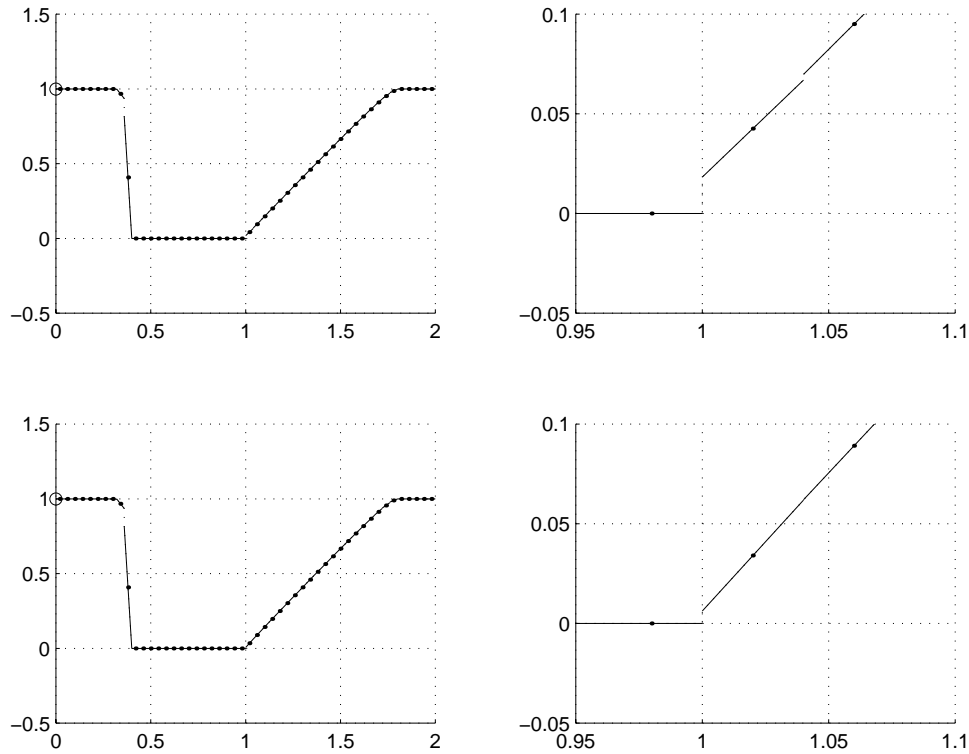

FIGURE 4.1. Schemes with $\kappa = 1$ for Burgers equation with midpoint quadrature (top) and Simpson quadrature (bottom). Right pictures contain a zoom of the left ones near $x = 1$.

5 Temporal discretization

In this section some results are presented for the BDF2 type scheme (1.7). First we only consider the explicit scheme for convection equations. After spatial discretization with the DG^1 scheme the resulting semi-discrete system reads

$$\frac{d}{dt}w(t) = F(w(t)) \tag{5.1}$$

and the time discretization scheme is

$$w_n = \tfrac{4}{3}w_{n-1} - \tfrac{1}{3}w_{n-2} + \tfrac{2}{3}\tau F(2w_{n-1} - w_{n-2}). \qquad (5.2)$$

Here $\tau > 0$ is the time step and w_n is the fully discrete discretization at time level $t_n = n\tau$.

For comparison we will also consider a 3-step TVD scheme of Shu [22]

$$w_n = \tfrac{3}{4}w_{n-1} + \tfrac{1}{4}w_{n-3} + \tfrac{3}{2}\tau F(w_{n-1}), \qquad (5.3)$$

and for theoretical purposes also the first-order explicit Euler scheme

$$w_n = w_{n-1} + \tau F(w_{n-1}). \qquad (5.4)$$

The methods (5.2),(5.3) are both second order accurate.

Below we shall refer to the *stability region* $\mathcal{S}$ of a time stepping method, which is the set of values $z = \tau\lambda \in \mathbb{C}$ such that the method is stable for the complex, scalar test equation $w'(t) = \lambda w(t)$ with step size τ. With the explicit Euler method we have $\mathcal{S} = \{z \in \mathbb{C} : |1 + z| \le 1\}$ The stability regions of the second order methods (5.2) and (5.3) are plotted in Figure 5.1. The boundaries of these regions were found from the root–locus curves, see [16].
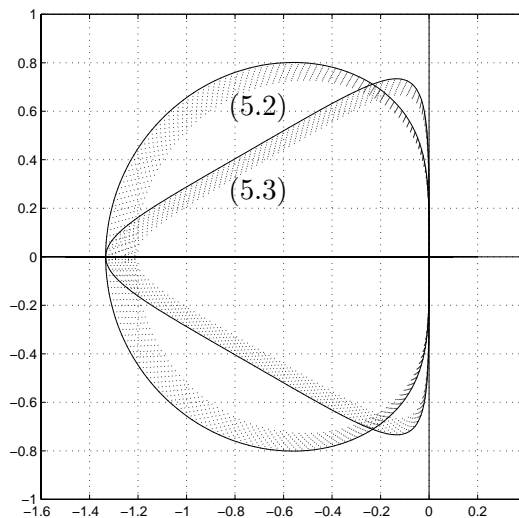

FIGURE 5.1. Stability regions of the methods (5.2) and (5.3).

5.1   Linear stability

Consider the spatial discretization (2.2) for the linear convection problem (2.1) on a uniform grid without limiting. For this problem the explicit Euler scheme is unconditionally unstable for any fixed ratio τ/h, it is only stable under a severe step size restriction of the form $\tau h^{-3/2} \le C$, see Chavent & Cockburn [5]. The reason for this unfavourable restriction is the fact that the stability region of the explicit Euler method $\{z \in \mathbb{C} : |1+z| \le 1\}$ cannot contain

all $\tau\lambda_\pm$ with eigenvalues $\lambda_\pm$ given in Section 2.1. In the vicinity of the origin the eigenvalues are too close to the imaginary axis.

With the second order methods (5.2) and (5.3) the stability regions, are better fitted to these eigenvalues, compare Figures 2.1 and 5.1. In Table 5.1 the stability restrictions are given in terms of maximal Courant number τ/h. These stability restrictions have been found experimentally from the requirement $\tau\lambda_\pm \in \mathcal{S}$. It is obvious that smaller values of κ allow larger time steps, see also Figure 2.1. Furthermore the stability restriction with method (5.2) is somewhat more relaxed than with (5.3).

	(5.2)	(5.3)
$\kappa = 1/3$	0.44	0.35
$\kappa = 2/3$	0.27	0.20
$\kappa = 1$	0.20	0.14

TABLE 5.1. Maximal Courant numbers for stability of the methods (5.2), (5.3).

It can be shown that the eigenvalue condition $\tau\lambda_\pm \in \mathcal{S}$ is sufficient here for having stability with periodicity in space, by the observation that eigenvectors of the Fourier transform $\widehat{A}$ are orthogonal with respect to a scaled inner product on $\mathbb{C}^2$, see Section 2.1.

Stability for the equation $u_t + u_x = \gamma u$ with the implicit-explicit scheme (1.7) and $\gamma < 0$ arbitrary, follows from [12], under the condition that the Courant number is such that the explicit scheme is stable for the case $\gamma = 0$.

5.2 TVD properties

Consider the scalar equation $u_t + f(u)_x = 0$ and spatial discretization with the DG^1 scheme and with the slope limiter (2.9). For $v \in M^1$ with $v(x) = \overline{v}_i \mathbb{1}_{\mathcal{C}_i}(x) + \widetilde{v}_i \varphi_i(x)$ on $\mathcal{C}_i$ as in (1.3), let

$$|v|_{\overline{TV}} = \sum_{j=1}^{m-1} |\overline{v}_j - \overline{v}_{j+1}|.$$

This corresponds to the total variation of the mean values $\overline{v}_j$. The time stepping method is called TVD (total variation diminishing) for the present spatial discretization if $|w_n|_{\overline{TV}}$ is nonincreasing in n.

Let $\nu = \tau h^{-1} \max |f'(u)|$ be the Courant number. Then the explicit Euler method is TVD for $\nu \leq \frac{1}{2}$, see Cockburn [7] for instance. For method (5.3) the TVD property holds under the condition $\nu \leq \frac{1}{4}$. This easily follows from the result for Euler's method, together with

$$|w_n|_{\overline{TV}} \leq \tfrac{3}{4}|w_{n-1} + 2\tau F(w_{n-1})|_{\overline{TV}} + \tfrac{1}{4}|w_{n-3}|_{\overline{TV}}.$$

In fact this method was constructed such that it can be written as a convex combination of Euler steps, see Shu [22]. We note this way of proving the TVD property is closely related to results on the preservation of positivity obtained by Bolley & Crouzeix [2].

The above explicit BDF2 type scheme cannot be written as such a convex combination of explicit Euler steps. Yet it can be shown that this scheme is TVD by taking starting

procedures into consideration. Let

$$v_n = 2w_n - w_{n-1}.$$

Then the recursion (5.2), $n \geq 2$, can also be written as

$$v_n = \theta v_{n-1} + \tfrac{4}{3}\tau F(v_{n-1}) + (\tfrac{5}{3} - 2\theta)w_{n-1} - (\tfrac{2}{3} - \theta)w_{n-2}, \tag{5.6}$$

where we shall take $\theta = \tfrac{2}{3}$ or $\tfrac{5}{6}$.

First consider the case were the first approximation w_1 is calculated with the Euler scheme,

$$w_1 = w_0 + \tau F(w_0). \tag{5.7}$$

Then $v_1 = w_0 + 2\tau F(w_0)$, and consequently we have

$$|w_1|_{\overline{TV}} \leq |w_0|_{\overline{TV}}, \quad |v_1|_{\overline{TV}} \leq |w_0|_{\overline{TV}}$$

whenever $\nu \leq \tfrac{1}{4}$. The proof of the TVD property can now be given by induction: assuming that $\max(|w_{n-1}|_{\overline{TV}}, |v_{n-1}|_{\overline{TV}}) \leq |w_0|_{\overline{TV}}$ and $\nu \leq \tfrac{1}{4}$, we obtain from (5.6) with $\theta = \tfrac{2}{3}$,

$$|v_n|_{\overline{TV}} \leq \tfrac{2}{3}|v_{n-1}|_{\overline{TV}} + \tfrac{1}{3}|w_{n-1}|_{\overline{TV}} \leq |w_0|_{\overline{TV}},$$

and also

$$|w_n|_{\overline{TV}} = |\tfrac{1}{2}v_n + \tfrac{1}{2}w_{n-1}|_{\overline{TV}} \leq |w_0|_{\overline{TV}}.$$

Hence, (5.2) with (5.7) is TVD provided that $\nu \leq \tfrac{1}{4}$.

Using formula (5.6) with $\theta = \tfrac{5}{6}$ this can be slightly improved, but then we need an other starting procedure. Consider the explicit trapezoidal rule

$$w_1^* = w_0 + \tau F(w_0), \qquad w_1 = w_0 + \tfrac{1}{2}\tau F(w_0) + \tfrac{1}{2}\tau F(w_1^*). \tag{5.8}$$

Note that we have $w_1 = \tfrac{1}{2}w_0 + \tfrac{1}{2}(w_1^* + \tau F(w_1^*))$. Hence $v_1 = w_1^* + \tau F(w_1^*)$, and using induction we thus see from (5.6) with $\theta = \tfrac{5}{6}$ that the combination of (5.2), (5.8) is TVD under the restriction $\nu \leq \tfrac{5}{16}$.

In conclusion, both with respect to stability and TVD the explicit BDF2 type method (5.2) has a slight advantage over the 3-step scheme (5.3). Moreover (5.2) combines in a natural way with the implicit BDF2 scheme as in (1.7) to treat stiff terms, whereas for (5.3) no implicit counterpart with comparable favourable stability properties is known. Stability results for the combined implicit-explicit scheme (1.7) can be found in [1, 10, 12].

Remark. The above TVD results are related to results on positivity preservation given in [17, 24]. Note that we also have TVD results for the explicit Euler method for fixed Courant number $\nu \leq \tfrac{1}{2}$ whereas the underlying non-limited scheme is unstable for fixed Courant numbers. This inherent instability manifests itself with limiting by strong compression, where smooth profiles are turned into staircase shaped profiles.

Further we note that apart from (5.6), the vectors $v_n = 2w_n - w_{n-1}$ also satisfy the recursion

$$v_n = \tfrac{4}{3}v_{n-1} - \tfrac{1}{3}v_{n-2} + \tfrac{4}{3}\tau F(v_{n-1}) - \tfrac{2}{3}\tau F(v_{n-2}),$$

for $n \geq 3$. This is a standard explicit linear multi-step formula, whereas (5.2) is a so-called one-leg formula, see [11, 16]. The above shows that for this linear multi-step formula the total variation of the means will stay bounded for $\nu \leq \tfrac{5}{16}$, provided special starting procedures are considered to calculate v_1 and v_2.

20

5.3   Numerical test – accuracy versus CPU

In this section a numerical test is considered to determine which values of κ are optimal in terms of accuracy versus CPU time for non-smooth solutions with limiting. We saw in Section 2 that $\kappa = 1$ gives the best spatial accuracy, but on the other hand with smaller values of κ a larger time step is allowed.

As test example we regard a simple adsorption-desorption model. Consider a given external flow field q together with unknown dissolved concentration u and adsorbed concentration v. The evolution of the concentrations is given by the system of convection-reaction equations

$$u_t + (qu)_x = k\Big(v - \psi(u)\Big),$$
$$v_t = -k\Big(v - \psi(u)\Big) \tag{5.9}$$

where $k > 0$ is the reaction rate and

$$\psi(u) = \frac{k_1 u}{1 + k_2 u}$$

describes the steady state ratio between u and v, see [14] for related examples. In the limit $k \to \infty$ we have $v = \psi(u)$, or

$$(u + \psi(u))_t + (qu)_x = 0,$$

which can be formulated as a nonlinear hyperbolic equation in standard form for $w = u + \psi(u)$,

$$w_t + (q\phi(w))_x = 0, \tag{5.10}$$

with ϕ implicitly defined by the relation

$$w = u + \psi(u) \quad \Longrightarrow \quad u = \phi(w).$$

We shall take the values $k = 1000$, $k_1 = k_2 = 100$, and we solve equation (5.9) as a stiff convection-reaction system using the method (1.5)–(1.7). The velocity q is spatially homogeneous and given by

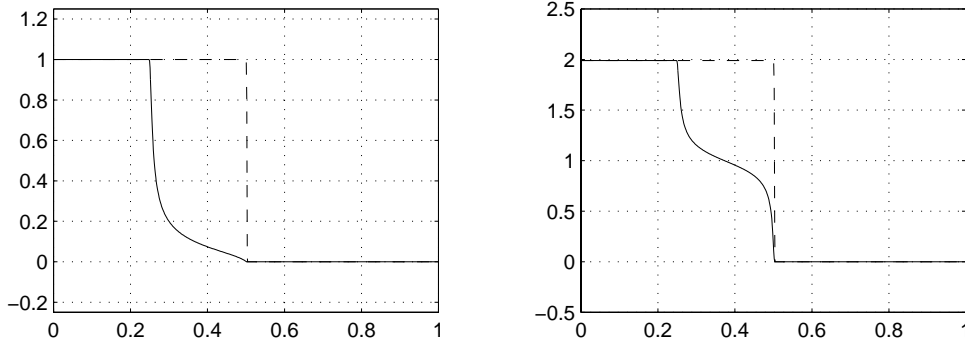$$q(t) = \begin{cases} 1 & \text{if } t \le 1, \\ -1 & \text{if } t > 1. \end{cases}$$

FIGURE 5.2. Adsorption-desorption test (5.9). Plots of dissolved concentration u (left) and total concentration $u + v$ (right) at time $t = 1$ (dashed) and $t = \frac{5}{4}$ (solid).

The initial condition is $u, v \equiv 0$, and we have given boundary conditions $u(0, t) = 1$ for $t \le 1$, $u(1, t) = 0$ for $t > 1$. The equation is considered on the time interval $0 \le t \le T = \frac{5}{4}$. An illustration of the solution is given in Figure 5.2, where the dissolved concentration u and total concentration $u + v$ is plotted as function of x at time $t = 1$ and $t = T = \frac{5}{4}$.

For this large value of k the solution is almost identical to (5.10). We consider (5.9) since it presents a relatively simple example of more general convection-reaction systems. In the numerical solution (1.5)–(1.7) the slope limiter (2.9) is applied after each time step. For the time step (1.7) a nonlinear algebraic system of the form

$$\overline{u} = U + \tfrac{2}{3}\tau k\left(\overline{v} - \psi(\overline{u})\right), \qquad \overline{v} = V - \tfrac{2}{3}\tau k\left(\overline{v} - \psi(\overline{u})\right)$$

has to be solved where U, V contain past information and the (explicit) convection update. We have $\overline{v} = V + U - \overline{u}$ so actually it is an algebraic relation for $\overline{u}$ only. This is solved with Newton iteration where care should be taken to end up on the correct branch $\overline{u} \ge 0$ of the quadratic algebraic relation. This can easily be achieved here by starting the Newton iteration with initial guess $\overline{u} = 0$ instead of $\overline{u} = U$, which might seem more natural. This argument indicates that spatial discretization with higher order quadrature, giving an algebraic system for $\overline{u}, \widetilde{u}$ would already for this simple test example be much more cumbersome.

In the numerical test we compared the three parameter values $\kappa = \frac{1}{3}, \frac{2}{3}, 1$. The step size τ was chosen as $1/N$ with N is a multiple of 4, so as to have $T = \frac{5}{4}$ as step point, and such that the Courant restriction of Table 5.1 is satisfied. So if we write $N_0 = 4\lceil (4Ch)^{-1}\rceil$ with C the entry for (5.2) in Table 5.1 and with $\lceil \cdot \rceil$ denoting upward integer rounding, then $\tau = 1/N_0$ corresponds to the maximal time step allowed for stability of explicit method for convection. With $\kappa = \frac{1}{3}$ this is larger than the step size needed to ensure the TVD property, and indeed slight oscillations were observed in the rarefaction wave that is formed for $t > 1$. Therefore also the choice $\tau = 1/(2N_0)$ is considered in the test. In Figure 5.3 the Euclidian mean-value errors in $u + v$ are plotted as function of the inverse of CPU time (in seconds for a MATLAB program on a SUN ULTRA 5 workstation) with number of points in space $m = 25, 50, 100, 200$; an accurate reference solution was calculated with $m = 800$ and small time steps.
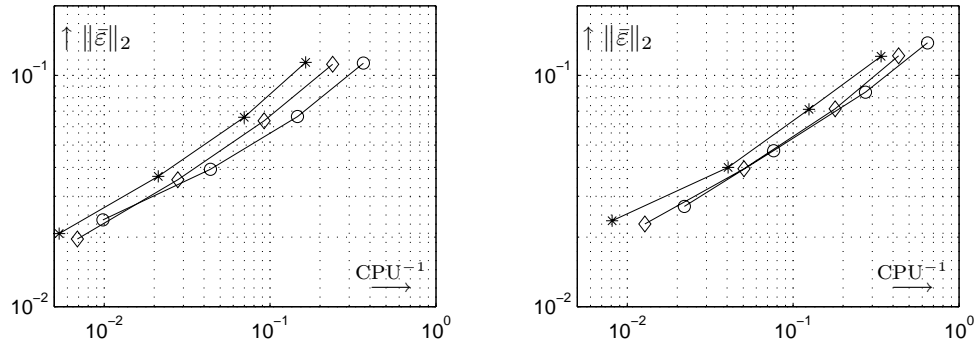

FIGURE 5.3. Mean errors $\|\bar{\varepsilon}\|_2$ of total concentration $u + v$ at output time $t = \frac{5}{4}$ as function of the inverse of CPU timings for $\kappa = 1$ (∗-marks), $\kappa = \frac{2}{3}$ (⋄-marks) and $\kappa = \frac{1}{3}$ (o-marks). Left plot with $N = 2N_0$, right plot with $N = N_0$.

From Figure 5.3 the following observations can be made. The value $\kappa = \frac{1}{3}$ gives larger spatial errors than $\kappa = \frac{2}{3}, 1$, but due to the fact that larger time steps are allowed this value comes out slightly better than the larger ones with respect to efficiency. The difference is not very large; in this test the value $\kappa = 1$ requires roughly twice the CPU time of $\kappa = \frac{2}{3}$ to achieve a certain accuracy level. Taking the step size close to the stability limit introduces a significant temporal error. Taking $\tau = 1/(2N_0)$ has almost the same efficiency. Finally we note that in this test the convergence rates are approximately $\mathcal{O}(h^1)$, due to the fact that the solution is not continuous for $t < 1$. Moreover, if h is decreased by a factor 2 the amount of work increases almost by a factor 4, since also twice as many time steps are needed.

5.4   Conclusions and remarks

Based on the tests in Section 2.5 and 5.3 the following conclusions can be drawn. Use of midpoint quadrature in (1.4) leads to a numerical scheme that is very suited for implicit-explicit time integration by a method such as (1.7). Higher order quadrature seems to offer only very small advantages in accuracy, see Section 4, but it leads to much more complicated (time consuming) implicit relations if stiff reactions are present. Of course, these conclusions for 1D problems cannot be directly generalized to multi-dimensional problems but the results here indicate that the use of low order quadrature is favourable in important situations.

If the solution is smooth and no limiting is applied then taking $\kappa = 1$ is optimal with respect to spatial accuracy. For non-smooth solutions with limiting the advantage in accuracy with $\kappa = 1$ almost vanishes and then taking smaller values of κ can have clear advantages since larger step sizes can be taken. Although the efficiency of the various κ values was comparable in the above test, for larger reactive systems the increase in step size allowed with $\kappa = \frac{1}{3}$ will yield a more pronounced gain in efficiency.

Diffusion has not been considered here. The standard approach would be by a mixed formulation as presented in Chavent & Jaffré [6]. Formally, with an inverse mass matrix, the resulting system can then still be written in the form (1.6) where G then also contains the discretized diffusion, and thus the method (1.7) can still be applied with implicit diffusion treatment as for the case of a stiff reaction term. More recently, different diffusion implementations have been considered in [3, 9] where the diffusion is treated completely within the discontinuous Galerkin approach.

References

[1] U.M. Ascher, S.J. Ruuth, B. Wetton, *Implicit-Explicit Methods for Time-Dependent PDE's.* SIAM J. Numer. Anal. 32 (1995), pp. 797-823.

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

[3] C.E. Baumann, J.T. Oden, *A discontinuous hp finite element method for the Euler and Navier-Stokes equations.* Int. J. Numer. Methds. Fluids; to appear.

[4] G. Chavent, B. Cockburn, *Convergence et stabilité des schémas LRG.* INRIA Report # 370, Rocquencourt, 1987.

[5] G. Chavent, B. Cockburn, *The local projection P^0-P^1-discontinuous Galerkin method for scalar conservation laws.* M^{2}AN 23 (1989), pp. 565-592.

[6] G. Chavent, J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation.* Studies in Mathematics and its Applications 17, North-Holland, Amsterdam, 1986.

[7] B. Cockburn, *An introduction to the discontinuous Galerkin method for convection dominated problems.* In: Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, A. Quarteroni (ed.), Lecture Notes in Mathematics 1697, Springer Verlag, 1998, pp. 151-268.

[8] B. Cockburn, C.-W. Shu, *The Runge-Kutta local projection P^1-discontinuous Galerkin finite element method for scalar conservation laws.* M^{2}AN 25 (1991), pp. 337-361.

[9] B. Cockburn, C.-W. Shu, *The local discontinuous Galerkin finite element method for convection-diffusion systems.* SIAM J. Numer. Anal.; to appear.

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

[11] G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems.* Numerical Analysis, Dundee 1975. Springer Lecture Notes in Mathematics 506, pp. 60-74.

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

[13] C.W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice-Hall, New Jersey, 1971.

[14] C.H. Giles, D. Smith, A. Huitson, *A general treatment and classification of the solute adsorption isotherm.* J. Coll. Int. Sci. 47 (1974), pp. 755-765.

[15] V. Gowda, J. Jaffré, *Discontinuous finite elements for nonlinear conservation laws.* INRIA Report # 1848, Rocquencourt, 1993.

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

[17] W. Hundsdorfer, *Partially implicit BDF2 blends for convection dominated flows.* CWI Report MAS-R9831, CWI, Amsterdam. To appear in SIAM J. Numer. Anal.

[18] 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.

[19] P. Lesaint, P.A. Raviart, *On a finite element method for solving the neutron transport equation.* In: Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor (ed.). Academic Press Inc, 1974, pp. 89-123.

[20] R.J. LeVeque, *Numerical methods for conservation laws.* Lecture Notes in Mathematics, ETH Zürich, Birkhäuser Verlag, Basel, 1992.

[21] W. Reed, T. Hill, *Triangular mesh methods for the neutron transport equation.* Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

[22] C.-W. Shu, *Total-variation-diminishing time discretizations.* SIAM J. Sci. Stat. Comp. 9 (1988), pp. 1073-1084.

[23] 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.

[24] J.G. Verwer, W. Hundsdorfer, J.G. Blom, *Numerical time integration for air pollution models.* CWI Report MAS-R9825, CWI, Amsterdam. To appear in Surveys on Mathematics for Industry, Vol. 10/4.