Department of

APPLIED MATHEMATICS

Adaptive Characteristic Operator Splitting Techniques for Convection-Dominated Diffusion Problems in One and Two Space Dimensions

by

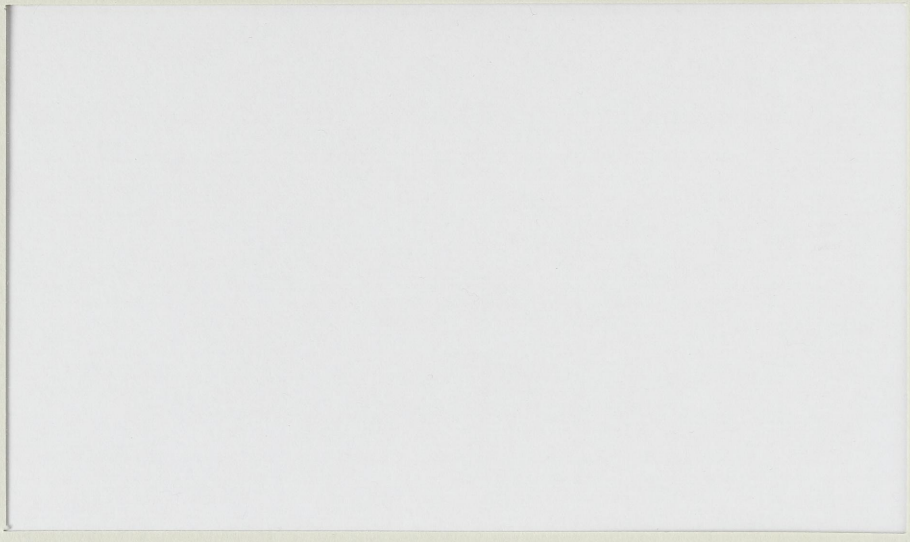
Helge K. Dahle

Report No. 85

December 1988

UNIVERSITY OF BERGEN

Bergen, Norway



Department of Mathematics
University of Bergen
5000 Bergen
Norway

ISSN 0084-778x

Adaptive Characteristic Operator Splitting Techniques for Convection-Dominated Diffusion Problems in One and Two Space Dimensions

by

Helge K. Dahle

Report No. 85

December 1988

ACKNOWLEDGMENTS

This research was supported in parts by the Norwegian Research Council of Science and Humanities (NAVF) and VISTA, a research cooperation between the Norwegian Academy of Science and Letters and Den norske stats oljeselskap a.s (Statoil)

Contents

1 Introduction

The dynamics of two-phase immiscible flow in a porous media strongly depends on the balance between capillary forces (diffusion) and convection. In commonly occuring cases the capillary forces are of importance only in small regions where the saturation gradients are large, i.e in areas where physical shocks are located, and the physics is essentially dominated by convection. This is reflected in an almost hyperbolic nature of the equations modeling such problems.

Equations of this kind are very difficult to treat by standard numerical methods. Normally one has to use time steps and a grid size determined by local behaviour in the shock region to accomplish a stable and accurate method. However, such methods are inefficient or worthless in terms of computer time. Usually one therefore determines the time steps and grid size from global behaviour of the problem and then introduces enough numerical diffusion to gain a stable solution. This is of course on the cost of accuracy, and even if mass balance is retained, the numerical solution might be smoothed far from the physical solution.

The goal of this report is to describe and analyze a numerical method developed to simulate the transport of a sharp saturation-front in one and two space dimensions, using a large time step and a spatial grid adapted to the local behaviour in the shock region. Thus, effects due to boundary conditions, inhomogeneities and nonestablished shocks are as far as possible neglected in the formulation of the problem.

In section 2 we give the governing equations for two-phase immiscible flow and specify the nonlinear fractional flow function and diffusion coefficient.

Section 3 develops the numerical schemes to be used in the one-dimensional case, see [1,2]; In section 3.1 we split the fractional flow function into a term describing the unique physical velocity for an established shock and another term which balances the diffusion at the shock.

An appropriate time step and length scale for the transport process is introduced, and we discretize the convective part of the equation in time by integrating backwards along the approximate characteristics, known as the modified method of characteristics, [3,4,5].

The sharp variation at the interface between oil and water requires a much finer grid than what is necessary to describe the transport phenomena. In section 3.2 we introduce local grid-refinement of the shock region.

Section 3.3 discusses the numerical modelling and discretization of the diffusion phenomena with an asymmetric transport term. We work out a Petrov-Galerkin formulation for this problem. Appropriate upstream weights of the test functions are obtained from a symmetrization technique introduced by Barrett and Morton [6], giving optimal approximation properties. Such test functions have been analyzed and used several places [7,8,9], and produce a stable numerical scheme around the shock.

Section 4 deals with error-estimates for the one dimensional case. We first develop an estimate in the supremum-norm, adequate for the purely hyperbolic part of the problem. Secondly, an error-estimate in the H^1-norm is developed for the inner problem. This estimate reduces to the one given by Douglas and Russel [4] in the symmetric

case. We may also refer to [10,11,12,13], for estimates on related problems.

Section 5 extends the numerical methods to two space dimensions [1]. We discuss composite trial and test functions and introduce a composite grid operator.

Recently efficient preconditioners have been developed for elliptic problems, based on domain decomposition techniques, [14,15]. Following these ideas, we construct a preconditioner for the composite operator in section 5.4.

In section 6, the practical implementation of the numerical code is documented, and some aspects of the operator-splitting technique are discussed in more detail.

Computational results and conclusions are presented in sections 7 and 8.

This report is my thesis for the partial requirement of the Dr. Scient. degree in applied mathematics. I wish to thank Professor Magne S. Espedal for all help and support during the work on this thesis. Without his optimism and encouragements, I would never have managed to finish it. I will also thank Tor Barkve and Øystein Pettersen for many helpful discussions and for making the reservoir group a pleasant place to work. In particular I have benefited from the computational work on the two dimensional part accomplished by Ove Sævareid. I want to express gratitude to members at the Institute of Scientific Computation, University of Wyoming, leaded by Professor Richard. E. Ewing, for many helpful suggestions and for reading through parts of the manuscript. I also want to thank members of the staff at the Institute of Mathematics, University of Bergen, for social stimulating inputs.

Finally, I will express my thanks to Gunvor for doing more than her part of the house work during the last months (at least).

2 Problem Definition

2.1 General Equations

We are studying two-phase immiscible flow in a two-dimensional square region Ω, representing a homogeneous oil field of constant thickness. We shall assume that the flow is incompressible, and we will neglect gravity forces. A suitable formulation of the dynamical equations for the total Darcy velocity $\mathbf{v}$, the total fluid pressure p, and the water saturation $u \in [0,1]$ is derived by Chavent [16], and will be used in the following nondimensional form:

$$\nabla \cdot \mathbf{v} = g_1(\mathbf{x},t), \quad \mathbf{x} \in \Omega, \quad t \in J = [0,T], \tag{1}$$

$$\mathbf{v} = -\mathbf{A}(u) \cdot \nabla p, \tag{2}$$

$$\mathbf{v} \cdot \mathbf{n} = g_2(\mathbf{x},t), \quad \mathbf{x} \in \partial\Omega, \quad t \in J, \tag{3}$$

$$\frac{\partial u}{\partial t} + \nabla \cdot (f(u)\mathbf{v}) - \epsilon \nabla \cdot (\mathbf{D}(u) \cdot \nabla u) = 0, \quad \epsilon \ll 1, \tag{4}$$

$$u(\mathbf{x},0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{5}$$

$$(\epsilon \mathbf{D} \cdot \nabla u - f(u)\mathbf{v}) \cdot \mathbf{n} = g_3(\mathbf{x},t), \quad \mathbf{x} \in \partial\Omega, \quad t \in J, \tag{6}$$

where

$$\mathbf{A}(u) = \mathbf{K}(\lambda_w + \lambda_o), \tag{7}$$

$$f(u) = \frac{\lambda_w}{\lambda_w + \lambda_o}, \tag{8}$$

$$\mathbf{D}(u) = \mathbf{K}\frac{\lambda_w \lambda_o}{\lambda_w + \lambda_o}\frac{dp_c}{du} \tag{9}$$

and g_i, $i = 1,2,3$, are zero away from the wells. $\mathbf{K}$ is the absolute permeability, λ_i, $i = w,o$, denotes the water and oil mobilities respectively, and p_c is the capillary pressure.

The objective of this report is to discuss a numerical scheme that efficiently resolves the dynamic saturation front defined by the transition zone between oil and water. We may note that the pressure equation in smooth cases is only weakly coupled to the saturation equation, giving an almost steady state velocity field with a $1/r$ dependency in the well regions where r is the distance from the well. Further, since $\mathbf{v}$ is the total fluid velocity it does not possess any shock behaviour in the transition zone between oil and water.

We have chosen to use an IMPES method in the numerical formulation of the equations, which means that we IMplicitly advance the Pressure in time and then Explicitly update the Saturation values in each time step. The pressure equation is solved to optimal order by a a mixed finite element code [17]. Thus, we may assume that

the velocity field $\mathbf{v}$, appearing in the saturation equations (4)-(6), is known explicitly for any given time t.

A typical shape of the fractional flow function $f(u)$ is depicted in Figure 1 (a). To obtain the S-shape shown in this figure, we may choose the water and oil mobilities to be given respectively by:

$$\lambda_w = u^p \text{ and } \lambda_o = (1-u)^p, \quad p = 2, 3, \ldots \tag{10}$$

Hence, for computational purposes we may use an analytical form of the fractional flow function defined by:

$$f(u) = \frac{u^p}{u^p + (1-u)^p}, \quad p = 2, 3, \ldots \tag{11}$$

(Note that the mobility ratio has been set to one.)

Further, for simplicity, we shall replace the absolute permeability tensor $\mathbf{K}$ by the identity matrix. Consequently, the components of the diffusion tensor $\mathbf{D}$ are given by:

$$D(u) = \frac{\lambda_w \lambda_o}{\lambda_w + \lambda_o} \frac{dp_c}{du}, \tag{12}$$

where

$$\mathbf{D}(u) = D(u)\mathbf{ii} + D(u)\mathbf{jj}.$$

Generally, the diffusion coefficient has the properties:

$$D(u) \geq 0, \quad D(0) = D(1) = 0. \tag{13}$$

A typical shape of the components of the diffusion tensor is depicted in Figure 1 (b). For simplicity we may use a computational form of the coefficients that qualitatively resembles Figure 1 (b), given by:

$$D(u) = 4u(1-u). \tag{14}$$

We once more emphasize that the purpose of this work is to simulate the transport of a sharp saturation front in an adequate manner. Therefor we have chosen to represent the initial profile $u_0(\mathbf{x})$ by an established shock, located somewhat away from the injection well, thus avoiding the singularity at the well. We shall also assume the monotonicity conditions:

$$\frac{\partial u_0}{\partial x} \leq 0, \quad \frac{\partial u_0}{\partial y} \leq 0. \tag{15}$$

Since the handling of general boundary conditions is outside the scope of this report, we may assume the wells to be given by:

$$u(\mathbf{x}_0) = 1 \tag{16}$$

and

$$\nabla u(\mathbf{x}_1) \cdot \mathbf{n} = 0, \tag{17}$$

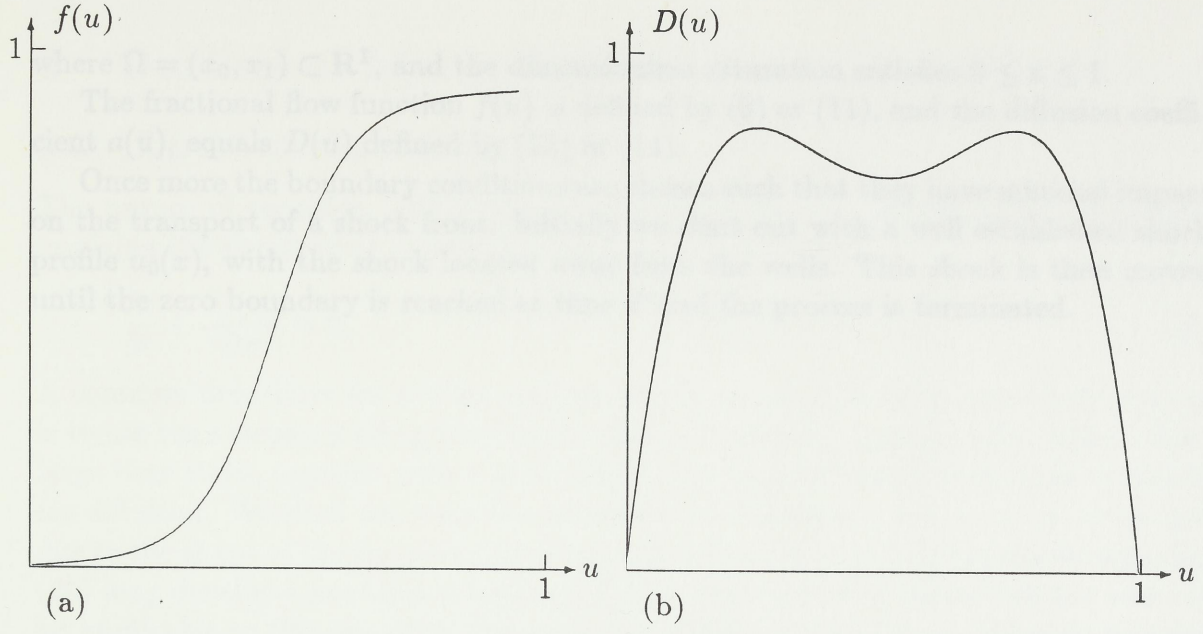


Figure 1: (a) The fractional flow function f as function of u. (b) Typical form of the diffusion coefficient D as function of u.

where $\mathbf{x}_0$ and $\mathbf{x}_1$ denotes the injection and production well respectively. Since $\mathbf{D}(1) = \mathbf{0}$, it follows that $g_2(\mathbf{x}, t)$ and $g_3(\mathbf{x}, t)$ have to satisfy the compability conditions:

$$g_3(\mathbf{x}_0, t) = -g_2(\mathbf{x}_0, t)$$

and

$$g_3(\mathbf{x}_1, t) = -f(u)g_2(\mathbf{x}_1, t).$$

These boundary conditions are essentially no flow conditions. Since condition (17) does not define an appropriate outflow condition, we have to terminate the process when a nonzero saturation value is transported into the production well at time $t = T$.

2.2 One-dimensional Equations

For the one-dimensional, incompressible case, the total fluid velocity is constant and the pressure equation decouples from the saturation equation. Hence, the general equations (1)-(6) reduces to the simpler problem for the water saturation $u(x, t)$:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) - \epsilon \frac{\partial}{\partial x}(a(u)\frac{\partial u}{\partial x}) = 0, \quad (x, t) \in \Omega \times [0, T], \quad \epsilon \ll 1, \tag{18}$$

and

$$\begin{aligned} &u(x, 0) = u_0(x), \quad x \in \Omega \\ &u(x_0, t) = 1, \quad \left.\frac{\partial u(x, t)}{\partial x}\right|_{x=x_1} = 0. \end{aligned} \tag{19}$$

λ(u)
1
0
1
(a)
D(u)
1
0
1
(b)

where $\Omega = (x_0, x_1) \subset \mathbf{R^1}$, and the dimensionless saturation satisfies $0 \leq u \leq 1$.

The fractional flow function $f(u)$ is defined by (8) or (11), and the diffusion coefficient $a(u)$, equals $D(u)$ defined by (12) or (14).

Once more the boundary conditions are chosen such that they have minimal impact on the transport of a shock front. Initially we start out with a well established shock profile $u_0(x)$, with the shock located away from the wells. This shock is then moved until the zero boundary is reached at time T and the process is terminated.

3 Outline of Methods, One-dimensional Case

3.1 Operator-Splitting

The nature of equation (18) is almost hyperbolic because of the small ϵ-parameter implying the dominating convective part:

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0. \tag{20}$$

A common procedure for solving such problems, which reflects its hyperbolic nature, is to use time stepping along the characteristics defined by equation (20), which allows large time steps, together with a finite element or finite difference technique to correct for diffusion. We shall use a similar procedure in this report, but notice at once that due to the shape of the fractional flow function defined by (11), the hyperbolic equation (20) may develop a nonunique solution. Hence, the method of characteristics may not be applicable to this equation. We shall resolve this problem by an operator-splitting technique introduced by Espedal and Ewing [1].

For an established shock we split the fractional flow function into two parts defined by:

$$f(u) = \bar{f}(u) + b(u)u, \tag{21}$$

where

$$\bar{f}(u) = \begin{cases} \dfrac{f(u_{BL})}{u_{BL}} \cdot u, & 0 \le u \le u_{BL}, \\ f(u), & u_{BL} < u \le 1, \end{cases} \tag{22}$$

and

$$b(u) = 0, \quad u_{BL} < u \le 1.$$

The Buckley-Leverett shock saturation u_{BL} is defined by equation (20) together with an appropriate entropy condition, and is given by the concave envelope of $f(u)$ as shown in section 7.1. For a growing shock, a dynamical definition of $\bar{f}(u)$ may be necessary, we shall discuss this in more detail in section 6.3.

We define the characteristic direction $\tau(u)$ in terms of the nonlinear operator:

$$\frac{\partial}{\partial \tau(u)} = \frac{1}{\psi(u)} \left(\frac{\partial}{\partial t} + \bar{f}'(u) \frac{\partial}{\partial x} \right), \tag{23}$$

where

$$\psi(u) = \sqrt{1 + \bar{f}'(u)^2}.$$

From the definition of $\bar{f}$ it follows that the characteristic direction is uniquely determined by (23), since a fully developed shock consists of a rarefaction wave and a contact discontinuity as given by $\bar{f}$. Further, we note that the equation:

$$\psi \frac{\partial u}{\partial \tau} = 0,$$

transports a shock with the same physical velocity as equation (20) together with an appropriate entropy condition.

Using the above definitions, equation (18) can be written in the equivalent form:

$$\psi(u)\frac{\partial u}{\partial \tau(u)} + \frac{\partial}{\partial x}(b(u)u) - \epsilon\frac{\partial}{\partial x}(a(u)\frac{\partial u}{\partial x}) = 0. \tag{24}$$

The characteristic derivative is now discretized in terms of the characteristic direction defined by (23); Let $\Delta t = t^n - t^{n-1}$, $\Delta t > 0$, and let $\bar{x}$ and $\bar{u}^{n-1}$ be the solution of the nonlinear equations:

$$\begin{aligned} \bar{x} &= x - \bar{f}'(\bar{u}^{n-1})\Delta t, \\ \bar{u}^{n-1} &= u(\bar{x}, t^{n-1}). \end{aligned} \tag{25}$$

Hence:

$$\psi\frac{\partial u^n}{\partial \tau} \approx \frac{u(x,t^n) - u(\bar{x},t^{n-1})}{\Delta t}. \tag{26}$$

We note that the characteristic curves defined by (23) are straight lines in the (x,t)-plane. Thus, if equation (25) is solved exactly, the only change in the solution along the characteristics is due to diffusion. For later reference, we shall refer to $\bar{u}(x,t)$ as the characteristic solution of equation (18).

By substitution of the approximate characteristic derivative (26) into (24) we get the following elliptic equation to solve in each time step

$$\frac{u^n - \bar{u}^{n-1}}{\Delta t} + \frac{d}{dx}(b(u^n)u^n) - \epsilon\frac{d}{dx}(a(u^n)\frac{du^n}{dx}) = 0. \tag{27}$$

3.2 Substructuring

There are two obvious spatial scales associated with our problem. Except for a thin shock layer, the gradients are small, and the following relations are certainly valid:

$$\psi\frac{\partial u}{\partial \tau} = O(\epsilon/h_o), \quad h_o = O(1). \tag{28}$$

In a neighborhood of an established shock the gradients are large and the reduced convective term $b(u)$, balances the diffusion term in equation (27). A small space scale is appropriate in this region and the following inequalities are valid (see [1]):

$$\left|\frac{\partial u}{\partial \tau}\right| \le \frac{1}{h_i}\max_{0\le u\le 1}\left|b(u)u - \epsilon a(u)\frac{\partial u}{\partial x}\right|, \quad \left|\frac{\partial u}{\partial x}\right| \le \frac{u}{h_i}, \quad h_i \ll 1. \tag{29}$$

The appearance of these two space scales motivate the use of a substructuring method to solve the elliptic equation (27).

A composite grid is constructed in the following manner; A uniform coarse grid is defined which is independent of time and adequate for the slow variation outside a shock layer. The position of a shock is then located on the coarse grid in each time step, and the elements containing the shock front are refined. We shall return to the

question of how to identify shock regions later; one should, however expect the shock front to be contained within at most two connected coarse grid elements for small ϵ.

With the composite grid defined in each time step, a solution procedure for equation (27) is developed as follows; On the coarse grid we neglect the diffusive terms, and the coarse grid solution at time t^n is given by the characteristic solution at each node defined by (26) and (28) to order ϵ. On the refined grid, we shall solve equation (27) by a Petrov-Galerkin method using the coarse grid values as boundary terms.

Let Ω_i be the refined elements on Ω and let h be a parameter defining the fine grid size, such that the number of fine grid nodes equals $1/h$. The boundary value problem we shall solve on Ω_i in each time step is then given by:

$$u + \Delta t \frac{d}{dx}(b(u)u) - \epsilon \Delta t \frac{d}{dx}(a(u)\frac{du}{dx}) = \bar{u}, \quad x \in \Omega_i \tag{30}$$

and

$$u = \bar{u}, \quad x \in \partial\Omega_i. \tag{31}$$

If Ω_i consists of several coarse grid elements, we will compute the solution separately on each element, although Ω_i will be treated as a single element in the following.

The parameter h introduces a coordinate stretching into equation (30). A consistent singular perturbation expansion of this equation around the shock-layer show that the shock-width $h_i = O(\epsilon)$, and further that the convective term balances the diffusion term to order ϵ :

$$\frac{\partial}{\partial x}(b(u)u - \epsilon a(u)\frac{\partial u}{\partial x}) = O(\epsilon). \tag{32}$$

For a thorough discussion of Burgers' equation, see [18,19]. Here we conclude that we have to choose $h = O(\epsilon)$, to completely resolve the shock.

If the shock is well resolved it follows from (29) that

$$u = \bar{u} + O(\delta \Delta t), \tag{33}$$

where $\delta = \epsilon$. If we are not able to resolve the shock front properly, i.e. in the limit $\epsilon \to 0$, the convective term dominates the diffusion term. In such cases standard numerical methods give rise to unstable solutions. To guarantee a stable solution, enough artificial diffusion has to be added to balance the convective term, such that equation (32) is retained to correct order. In the next section we show how this may be done by solving (30) with a Petrov-Galerkin method rather than a Galerkin method.

The importance of balancing the convective term is further recognized in the treatment of the nonlinear coefficients $a(u)$ and $b(u)$. If balance between diffusion and convection is achieved, as from the Petrov-Galerkin method outlined in the next section, we may assume (33) to be valid since (32) is satisfied, even if the shock is not well resolved. Then, by a Taylor expansion of the coefficients around the characteristic solution, we get:

$$\begin{aligned} b(u) &= b(\bar{u}) + O(\delta \Delta t) \\ a(u) &= a(\bar{u}) + O(\delta \Delta t), \end{aligned} \tag{34}$$

where $\delta = \max(\epsilon, h)$ depending on whether the shock is resolved or not. By (28) this expansion is uniformly valid for all $x \in \Omega$. We use this expansion to rewrite equation (30) in a linear form consistent with the discretization of the characteristic derivative to leading order:

$$u + \Delta t \frac{d}{dx}(\bar{b}(x)u) - \epsilon \Delta t \frac{d}{dx}(\bar{a}(x)\frac{du}{dx}) = \bar{u}, \tag{35}$$

where $\bar{a} = a(\bar{u})$ and $\bar{b} = b(\bar{u})$. Similarly, for completeness, we replace $a(u^n)$ and $b(u^n)$ with respectively $\bar{a}^{n-1}(x)$ and $\bar{b}^{n-1}(x)$ to linearize equation (27).

From (29) it follows that to leading order, the dynamics of the shock layer is governed by the diffusion-convection problem:

$$\frac{d}{dx}(\bar{b}(x)u) - \epsilon \frac{d}{dx}(\bar{a}(x)\frac{du}{dx}) = 0. \tag{36}$$

If the shock is well resolved the correction terms to this equation are of order ϵ, if the shock is not resolved the correction is of order h. Although we are going to solve the complete equation (35), we shall use this information when we determine an appropriate test space for the Petrov-Galerkin method and precisely construct the test functions in terms of problem (36). We also note that the effect of solving the diffusion convection problem is to get the correct shock-width, or if the shock is not well resolved, to keep the numerical diffusion within the resolution given by h. This may be further connected to the concepts of "TVD-schemes" and "flux-limiters", see [20,21,22,23].

To summarize, the solution at time t^n consists of a characteristic part on the coarse grid and a diffusion-convection part on the refined grid giving the correct shock-width, rather than the solution of (27) by a Petrov-Galerkin method on the composite grid.

We might judge the solution constructed in this manner as a simple preconditioner for an iterative method solving equation (27) on the composite grid. Numerical experiments performed show, however, that the solution obtained is well-behaved after the first-step of such an iteration, and it does not seem necessary to elaborate the solution any further. Obviously this is true because the characteristic solution is close to the real solution for small ϵ, and because the problem is in one space dimension. We note that without any iterations the fine grid solution is passed to the coarse grid only via the characteristic solution.

3.3 Petrov-Galerkin Method

As in the previous section, we let Ω_i denote the shock region, which for simplicity is normalized to be the interval $[0, 1]$. In the following we shall develop a Petrov-Galerkin discretization for the diffusion correction problem defined on Ω_i.

We first assume that $0 \leq \bar{u} \leq u_b$ where $u_{BL} \leq u_b < 1$, and that $\frac{d\bar{u}}{dx} \leq 0$ in the shock layer. It follows that (35) is singular for $\bar{u} = 0$, i.e. ahead of the shock front, since $a(0) = 0$, whereas the other zero of $a(u)$ is excluded by the assumptions.

Let H^m be the usual Sobolev space of functions with L_2-integrable derivatives of order $\leq m$. We shall define the following subsets of H^1:

$$V = \{v \in H^1(\Omega_i) \mid v = \bar{u} \text{ on } \partial\Omega_i\},$$
$$H_0^1 = \{v \in H^1(\Omega_i) \mid v = 0 \text{ on } \partial\Omega_i\}.$$

The weak formulation of equation (35) is then: Find $u \in V$ such that

$$A(u,v) = (\bar{u},v) \quad \forall v \in H_0^1, \tag{37}$$

where

$$A(u,v) = (u,v) + \Delta t\, B(u,v)$$

and

$$B(u,v) = ((\bar{b}u)',v) + (\epsilon\bar{a}u',v'). \tag{38}$$

With $\bar{a}(x) \in C^0(\overline{\Omega}_i)$ and $\bar{b}(x) \in H^1(\Omega_i)$ the bilinear forms $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ are continuous on $H_0^1 \times H_0^1$. Unfortunately they might not be coercive, because

$$((\bar{b}v)',v) = \frac{1}{2}(\bar{b}'v,v) \leq 0,$$

since

$$\frac{d\bar{u}}{dx} \leq 0, \quad \frac{d\bar{b}}{d\bar{u}} \geq 0 \text{ and } \bar{b}' = \frac{d\bar{b}}{d\bar{u}}\frac{d\bar{u}}{dx}.$$

However, by (32), it is reasonable to assume that the convective term $b(u)$ is dominated by the other terms, such that:

$$\left|\Delta t(\bar{b}'v,v)\right| < 2\{(v,v) + \epsilon\Delta t(\bar{a}v',v')\} \tag{39}$$

which implies that $A(\cdot,\cdot)$ is coercive. Existence and uniqueness of u satisfying (37) is then guaranteed by the Lax-Milgram theorem.

We shall use a conforming finite element technique to solve equation (37). Let $\{x_i; i = 0,1,\ldots,N\}$ be a uniform discretization of Ω_i such that $0 = x_0 < x_1 < \cdots < x_N = 1$ and $x_i - x_{i-1} = h$. We define the trial space S^h and the test space T^h to be discrete subspaces of H^1 of dimension $N+1$, spanned by θ_i and ψ_i, the trial and test functions respectively. Further we define the discrete subsets:

$$S_0^h = S^h \cap H_0^1, \quad T_0^h = T^h \cap H_0^1, \text{ and } S_V^h = S^h \cap V.$$

It may be noted that we always can expand functions satisfying the boundary conditions in terms of basis functions in S^h in the one-dimensional case, implying that $S^h \cap V$ is not the empty space. In higher dimensions this is not generally true.

The Petrov-Galerkin finite element formulation of equation (37) is then: Find $U \in S_V^h$ such that

$$A(U,\psi_i) = (\bar{u},\psi_i) \quad \forall \psi_i \in T_0^h. \tag{40}$$

We choose our basis functions to be the usual châpeau basis defined by:

$$\theta_i = \begin{cases} 0 & x \le x_{i-1}, \\ (x - x_i)/(x_i - x_{i-1}) & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/(x_{i+1} - x_i) & x_i \le x \le x_{i+1}, \\ 0 & x_{i+1} \le x. \end{cases} \tag{41}$$

A symmetrization technique introduced by Barrett and Morton [6], will be used to find close to optimal test functions for problem (40).

As already noted, the usual Galerkin method, where the test and basis spaces are equal, fails to work if the mesh Péclet number $\beta = \bar{b}h/\epsilon\bar{a}$, is large in absolute value ($\beta < 0$ since $\bar{b} < 0$). As shown by Morton [24], the mesh Péclet number enters the error estimate in the energy norm, giving a poor bound for large negative β. Computations show that the solution exhibits wild oscillations when this occurs.

The appearance of these oscillations may easily be demonstrated for the simple diffusion convection problem (see [24]): Find $U \in S_V^h$ such that

$$B(U, \psi_i) = 0 \quad \forall \psi_i \in T_0^h, \tag{42}$$

where $B(\cdot, \cdot)$ is defined by (38), $\bar{a}$ and $\bar{b}$ being constant, and $T_0^h = S_0^h$ is spanned by (41). The Galerkin system for this problem, reduces to the difference equations:

$$\delta^2 U_i - \beta \Delta_0 U_i = 0, \quad i = 0, 1, \ldots, 1/h = N,$$

where

$$U_i = U(x_i), \quad \delta^2 U_i = U_{i+1} - 2U_i + U_{i-1}, \quad \text{and} \quad \Delta_0 U_i = \frac{1}{2}(U_{i+1} - U_{i-1}).$$

The solution to these equations is easily obtained and without loss of generality we choose boundary values $U_0 = 1$ and $U_N = 0$ which gives the solution:

$$U_i = \frac{\mu_0^{N-i} - 1}{\mu_0^N - 1}, \quad \mu_0 = \frac{2 + \beta}{2 - \beta}.$$

If μ_0 is negative, i.e $\beta < -2$, this solution exhibits oscillations, which is of a purely numerical nature since the solution to the associated continuous problem is monotone. Hence, for large and negative Péclet numbers $\bar{b}/\epsilon\bar{a}$, we have to choose h correspondingly small to obtain monotonicity.

In our framework, one should expect β to be small since the grid refinement technique is supposed to resolve the front. However, as already noted, this may not be the case in the limit $\epsilon \to 0$. When ϵ is very small we only want to resolve the front to some $h_{\min}$, without really resolving it. Hence, we need test functions which gives a stable solution even when a front is not well resolved, i.e. when the mesh Péclet number is large in absolute value.

Another motivation for choosing a test space different from the trial space, is the singularity at the bottom of the shock front. We will later show that one effect of choosing a test space different from the trial space is to produce Galerkin equations which appear to be well behaved even at the singularity.

In the previous section we noticed that the dominating and troublesome part of our problem was the diffusion convection equation defined by equation (36), or in the weak

formulation, the part defined by the bilinear form $B(\cdot,\cdot)$. We shall therefor choose our test space in terms of this problem.

Following Barrett and Morton [6], we define a symmetric form

$$B^m(u,v) = (\epsilon a_0 u', v') \quad \forall u, v \in H_0^1, \tag{43}$$

where the coefficient $a_0(x) > 0$ should be of the same order as the physical diffusion given by $\epsilon\bar{a}$, or the numerical diffusion introduced from not properly resolving the shock-front. An obvious choice which satisfies these prescriptions is given by:

$$a_0 = a - bh/\epsilon. \tag{44}$$

We note that $B^m(\cdot,\cdot)$ is a continuous and coercive bilinear form on $H_0^1 \times H_0^1$ and therefor define an inner product on H_0^1. Thus, from the Riesz representation theorem there exists a unique continuous representation $R^m : H_0^1 \to H_0^1$, such that given $u \in H_0^1$:

$$B(u,v) = B^m(u, R^m v) \quad \forall v \in H_0^1,$$

since $B(u,\cdot)$ is a continuous linear functional on H_0^1. If we choose our test-space T_0^h to be spanned by the test functions ψ_i^*, such that:

$$span\{R^m\psi_i^*\} = span\{\theta_i\} = S_0^h, \tag{45}$$

we obtain the optimal approximation property:

$$\|u - U_m\|_{B^m}^2 = \inf_{V \in S_v^h} \|u - V\|_{B^m}^2, \tag{46}$$

for the problem defining the shape of a shock: Find $U^m \in V$ such that

$$B(U^m, \psi) = 0, \ \forall \psi \in T_0^h,$$

where $\|\cdot\|_{B^m}$ is the energy norm defined by $B^m(\cdot,\cdot)$. Motivated by this we choose the test-functions to be given by the relation:

$$B(u, \psi_i^*) = B^m(u, \theta_i), \quad \forall u \in H_0^1.$$

This problem reduces to the solution of the set of first order differential equations:

$$\epsilon\bar{a}\psi_i^{*\prime} + \int_0^x \bar{b}\psi_i^{*\prime} dt = \epsilon a_0 \theta_i' + C \quad i = 1, \ldots, N-1. \tag{47}$$

It follows that the representation R^m, which relates our bilinear forms, is defined by the expression:

$$R^m \ : \ \psi \to R^m\psi = \int_0^x \frac{1}{\epsilon a_0}\{\epsilon\bar{a}\psi' + \int_0^t \bar{b}\psi' ds - C\}dt, \tag{48}$$

where C has to be chosen such that $R^m\psi \in H_0^1$.

Although a general form of R^m is known, it is difficult to find the inverse operator[1] R^{m-1} which solves equation (47) and explicitly defines the optimal test functions. As a first approach we therefor confine ourselves to the case of constant coefficients, since this simplification will enable us to construct explicit test functions which are usable for practical purposes.

In this case we easily obtain the solution of (47) to be:

[1]Even though the notation R^{m-1} for the inverse representer may be confusing, it will be used here and in all the subsequent sections.

$$\psi_i^* = \frac{a_0}{\bar{a}} \int_0^x e^{-(\bar{b}/\epsilon\bar{a})(x-t)} \theta_i'(t)dt - \frac{a_0}{\bar{a}} \left(\frac{1 - e^{-(\bar{b}/\epsilon\bar{a})x}}{1 - e^{-(\bar{b}/\epsilon\bar{a})}} \right) \int_0^1 e^{-(\bar{b}/\epsilon\bar{a})(1-t)} \theta_i'(t)dt. \quad (49)$$

These test functions are nonlocal, and we need a procedure to obtain test functions ψ_i with local support that span the same test space as the global test functions. In general this is not possible, however, from solving the local problems ($\beta = \bar{b}h/\epsilon\bar{a}$):

$$\psi_i'' + \beta\psi_i' = 0, \quad i = 1, \ldots, N-1, \quad (50)$$

where ψ_i by definition has local support on (x_{i-1}, x_{i+1}), such that

$$\psi_i(x_{i-1}) = \psi_i(x_{i+1}) = 0 \text{ and } \psi_i(x_i) = 1,$$

we get local test functions [6]:

$$\psi_i = \begin{cases} 0 & x < x_{i-1} \\ (1 - e^{-\beta(x-x_{i-1})/h})/(1 - e^{-\beta}) & x_{i-1} \le x \le x_i \\ (e^{-\beta(x-x_i)/h} - e^{-\beta})/(1 - e^{-\beta}) & x_i \le x \le x_{i+1} \\ 0 & x > x_{i+1} \end{cases} \quad (51)$$

Operating on ψ_i with R^m we obtain

$$R^m\psi_i = \theta_i^* = \begin{cases} -\bar{\beta}x & x < x_{i-1} \\ \bar{\beta}\theta_i(x)/(1 - e^{-\beta}) - \bar{\beta}x & x_{i-1} \le x \le x_i \\ \bar{\beta}e^{-\beta}\theta_i(x)/(1 - e^{-\beta}) - \bar{\beta}(1-x) & x_i \le x \le x_{i+1} \\ \bar{\beta}(1-x) & x > x_{i+1} \end{cases}$$

where $\bar{\beta} = a_0\beta/\bar{a}$. Since $\theta_i^*(x)$ is a linear function with knots x_{i-1}, x_i and x_{i+1} we conclude that the optimal approximation property associated with the symmetric bilinear form, $span\{R^m\psi_i\} = S_0^h$, is satisfied.

Similarly, it obviously follows that this relation is true if ψ_i is constructed from the local problem (50) without assuming constant coefficients. We use this observation to obtain approximate local test functions for the original problem. Assume that the coefficients are slowly varying such that they can be replaced with averages on each element of the refined grid, consistent with the expansions already performed, then the solution of equation (50) is seen to be:

$$\psi_i = \begin{cases} 0 & x < x_{i-1} \\ (1 - e^{-\beta_{i-1}(x-x_{i-1})/h})/(1 - e^{-\beta_{i-1}}) & x_{i-1} \le x \le x_i \\ (e^{-\beta_i(x-x_i)/h} - e^{-\beta_i})/(1 - e^{-\beta_i}) & x_i \le x \le x_{i+1} \\ 0 & x_i < x \end{cases} \quad (52)$$

where β_i denote the average on element $[x_i, x_{i+1}]$.

At this point we may notice that ψ_i as defined by (51) or (52), and θ^*, as a consequence of the choice of $a_0(x)$, is bounded in the limit $|\beta| \to \infty$. The boundedness of ψ_i implies that the method is well defined even at the singularity of the diffusion coefficient $a(u)$.

The exponentials appearing in the test functions above may be difficult to handle in numerical computations. Several upwind schemes have been proposed producing

simpler test functions, which resembles the optimal test functions given by (51) and (52). Here, we shall derive simpler test functions introduced by Heinrich et al. [7].

We assume that the test functions approximating the optimal test functions (51), have the form:

$$\tilde{\psi}_i = \begin{cases} 0 & x < x_{i-1} \\ \theta_i + c_L\sigma_i & x_{i-1} \leq x \leq x_i \\ \theta_i + c_R\sigma_i & x_i \leq x \leq x_{i+1} \\ 0 & x_i < x \end{cases} \tag{53}$$

where c_L and c_R is to be determined in terms of the mesh Péclet number and σ_i is the second order polynomial given by:

$$\sigma_i = \begin{cases} (x - x_{i-1})(x - x_i)/h^2 & x_{i-1} \leq x \leq x_i \\ -(x - x_i)(x - x_{i+1})/h^2 & x_i \leq x \leq x_{i+1} \end{cases} \tag{54}$$

The procedure we intend to use to determine the coefficients is described in general as follows; We choose polynomial test functions $\tilde{\psi}_i(x; \alpha_1, \ldots, \alpha_k)$ where the α's are to be determined, and integrate analytically the elements of the stiffness matrixes given by:

$$a_{ij} = B(\theta_j, \psi_i) \text{ and } \tilde{a}_{ij} = B(\theta_j, \tilde{\psi}_i),$$

where ψ_i and $\tilde{\psi}_i$ are respectively the optimal and approximate optimal test functions. The simplest way of determining the α's is to compare the entries of the stiffness matrixes element for element. However, in general this will lead to over- or underdetermined systems of equations.

In case of an underdetermined system, more equations can be supplied by comparing the right hand sides of the Galerkin equations, using polynomial source functions. If the system is overdetermined we have to remove equations, for instance we may only compare the diagonals of the matrixes.

We note that this procedure depends upon our ability to integrate the Galerkin equations analytically. With the optimal test functions given by (51), this is easily accomplished and the problem of determining c_L and c_R reduces to solving the linear system:

$$\begin{pmatrix} \frac{1}{6} & 0 \\ -\frac{1}{6} & -\frac{1}{6} \\ 0 & \frac{1}{6} \end{pmatrix} \begin{pmatrix} c_L \\ c_R \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(\frac{2}{\beta} - \coth(\frac{\beta}{2}) \\ \coth(\frac{\beta}{2}) - \frac{2}{\beta} \\ \frac{1}{2}(\frac{2}{\beta} - \coth(\frac{\beta}{2}) \end{pmatrix}$$

This system of equations has a unique solution given by:

$$c \stackrel{\text{def}}{=} c_L = c_R = 3(\frac{2}{\beta} - \coth(\frac{\beta}{2})). \tag{55}$$

Thus, the test functions defined by (53) and (55) give the same stiffness matrix as the optimal test functions given by (51), the only difference is how the source function is sampled. On the other hand, we may easily verify that the area bounded by the optimal test functions is conserved by the approximate test functions, implying that constant source functions gives identical Galerkin equations for both sets of test functions. In analogy with what we did with the optimal test functions (51), we shall replace the Péclet number entering (55) with averages on each element. Hence, the test functions to be used in practical computations are given by:

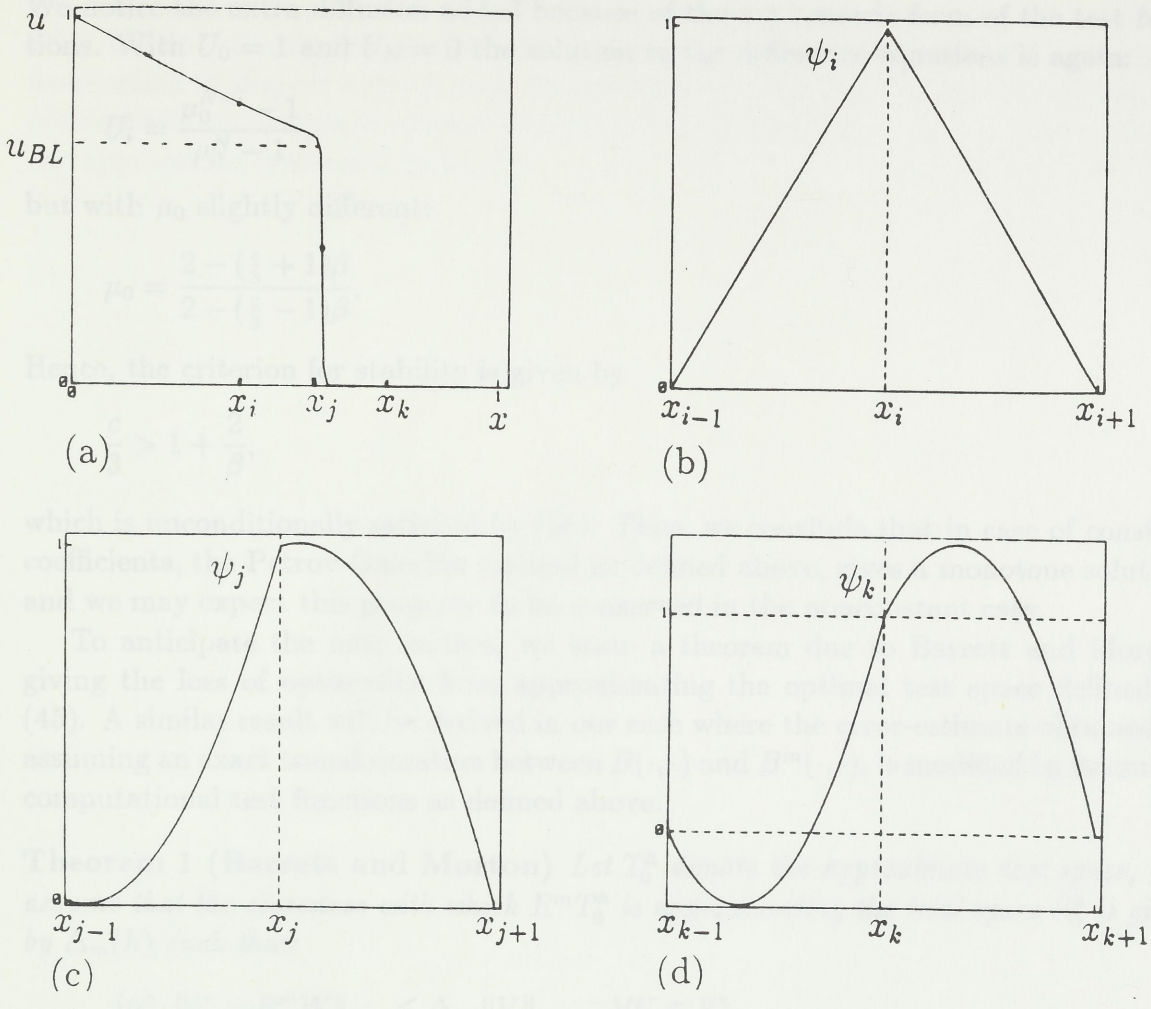


Figure 2: Typical test functions in one space dimension as function of the saturation u. (a) One-dimensional saturation profile. (b) $u(x_i) > u_{BL} \Rightarrow c^i = 0$. (b) $u(x_j) = 1/2\, u_{BL} \Rightarrow c^j = 1$. (c) $u(x_k) = 0 \Rightarrow c^k = 3$.

$$\tilde{\psi}_i = \begin{cases} 0 & x < x_{i-1} \\ \theta_i + c_{i-1}\sigma_i & x_{i-1} \leq x \leq x_i \\ \theta_i + c_i\sigma_i & x_i \leq x \leq x_{i+1} \\ 0 & x_i < x \end{cases} \tag{56}$$

where

$$c_i = 3(\frac{2}{\beta_i} - \coth(\frac{\beta_i}{2})).$$

The typical variation of these test functions through the shock region is depicted in Figure 2.

Returning to the motivating example defined by equation (42), we may now write out the Galerkin system with the asymmetric test functions given by (53) and (55). In this case the problem reduces to the difference equations

$$-(1 - \beta\frac{c}{6})\, \delta^2 U_i + \beta\Delta_0 U_i = 0, \qquad i = 1, \ldots, N-1.$$

(a)
(b)
(c)
(d)

We notice the extra diffusion added because of the asymmetric form of the test functions. With $U_0 = 1$ and $U_N = 0$ the solution to the difference equations is again:

$$U_i = \frac{\mu_0^{N-i} - 1}{\mu_0^N - 1},$$

but with μ_0 slightly different:

$$\mu_0 = \frac{2 - (\frac{c}{3} + 1)\beta}{2 - (\frac{c}{3} - 1)\beta}.$$

Hence, the criterion for stability is given by

$$\frac{c}{3} > 1 + \frac{2}{\beta},$$

which is unconditionally satisfied by (55). Thus, we conclude that in case of constant coefficients, the Petrov-Galerkin method as defined above, gives a monotone solution, and we may expect this property to be conserved in the nonconstant case.

To anticipate the next section, we state a theorem due to Barrett and Morton, giving the loss of optimality from approximating the optimal test space defined by (45). A similar result will be derived in our case where the error-estimate obtained by assuming an exact transformation between $B(\cdot,\cdot)$ and $B^m(\cdot,\cdot)$, is modified by assuming computational test functions as defined above.

Theorem 1 (Barrett and Morton) *Let T_0^h denote the approximate test space, and assume that the closeness with which $R^m T_0^h$ is approximating the trial space S_0^h is given by $\Delta_m(h)$ such that:*

$$\inf_{W \in T_0^h} \|V - R^m W\|_{B^m} \leq \Delta_m \|V\|_{B^m} \quad \forall V \in S_0^h.$$

Then, if $\Delta_m \in [0,1)$, there exist a unique solution to the problem:
Find $U \in S_V^h$ such that:

$$B(U, \psi_i) = (f, \psi_i) \quad \forall \psi_i \in T_0^h$$

Furthermore, the error in the solution satisfy

$$\|u - U\|_{B^m} \leq \left(1 - \Delta_m^2\right)^{-1/2} \leq \inf_{V \in S_V^h} \|u - V\|_{B^m}$$

For a proof of this theorem see [6].

In the simple case of constant coefficients, the parameter Δ_m has been calculated by Scotney [25] for several choices of upwind schemes. These calculations show that for the test functions introduced by Heinrich et al., $(1 - \Delta_m^2)^{-1/2}$ is bounded by approximately 1.3 for large mesh Péclet numbers. Thus, we may conclude that there is little loss of optimality by using the approximate test functions defined above. It is also reasonable to believe that this is true in the non-constant case.

We may finally remark that optimality should have been defined in terms of the complete operator $A(\cdot,\cdot)$, and an associated symmetric form A^m. Such test functions

are discussed by Demkowicz and Oden for a symmetric operator, i.e. $b(u) \equiv 0$, see [26,27]. However, symmetrization with respect to both the L_2-term and the asymmetric term seems to suggest a much more complicated problem for determining appropriate optimal and approximate optimal test functions, and it is not obvious how to choose the appropriate symmetric form A^m.

4 Error-estimate

In the subsequent sections error-estimates for the numerical solution of (18) and (19) based on the operator-splitting technique will be derived.

First we derive an error-estimate for the characteristic solution in the maximum norm, which demonstrates that diffusion correction is only necessary in the regions of Ω where the gradients are large. Secondly, an error bound in the H^1-norm is obtained for the complete convection-diffusion procedure, in the shock region.

Consequently, in each time-step, we divide Ω into two parts such that

$$\Omega = \Omega_o^n \cup \Omega_i^n, \quad \Omega_o^n \cap \Omega_i^n = \emptyset$$

We let Ω_o denote the outer region, where the solution is completely determined by the hyperbolic behaviour, and let Ω_i denote the shock layer.

Since the shock front is stable on the time scale we are dealing with, i.e. few of the characteristics diffuse out of the shock layer before the shock reaches the out-end, we may assume that the area of $\Omega_o(t)$ and $\Omega_i(t)$ is constant in time:

$$\begin{array}{l} meas(\Omega_0^{n-1}) = meas(\Omega_o^n) \\ meas(\Omega_i^{n-1}) = meas(\Omega_i^n) \end{array} \quad n = 1, \ldots, T/\Delta t \tag{57}$$

We shall assume that Ω_i is transported with the shock velocity between successive time steps. This implies that the shock region can be transformed onto a domain in phase-space with time independent boundaries, by a simple change of coordinates.

We notice that this transport necessarily conflicts with a fixed coarse grid, since coarse grid nodes eventually become internal nodes in the shock layer. In the analysis which follows we will assume a procedure which adjusts coarse grid nodes such that the shock layer Ω_i can be taken to be one coarse grid block, with the front placed approximately in the middle of the block.

In the computations performed, no procedure of this kind has been used, and the coarse grid defined is uniform and fixed. The appearance of internal coarse grid nodes does not seem to create much distortion on the numerical solution either. However, for practical reasons such as robustness, a procedure that removes internal nodes in the shock layer and adjusts the surrounding coarse grid nodes should be considered built into the code.

4.1 Some Definitions and Notation

The usual L_2-norm on Ω will be written:

$$\|v\|_{0,\Omega} = \left(\int_\Omega v^2 dx \right)^{1/2}, \tag{58}$$

and we define the maximum norm to be:

$$\|v\|_{\infty,\Omega} = \operatorname{ess\,sup}_{x \in \Omega} |u(x)|. \tag{59}$$

We shall norm H_0^1 with the seminorm:

$$\|v\|_{1,\Omega} = \left(\int_\Omega v_x^2 dx \right)^{1/2}. \tag{60}$$

We note that if $c \leq \epsilon a_0(x) \leq C$ on Ω, where c and C is positive constants, the H^1-norm is equivalent to the energy-norm defined by $B^m(\cdot,\cdot)$, giving:

$$c \|\cdot\|_{1,\Omega}^2 \leq \|\cdot\|_{B^m}^2 \leq C \|\cdot\|_{1,\Omega}^2 \tag{61}$$

Further we will need to norm the solution space defined on $[0,T] \times \Omega$. Let H be a normed linear space on Ω, then, if $w(x,t)$ is defined on $[0,T] \times \Omega$, we say that $w \in L_p([0,T],H)$, if $w(\cdot,t)$ is in H for all $t \in [0,T]$ and $\|w\|_H \in L_p([0,T])$. This space is normed with:

$$\|w\|_{L_p([0,T],H)} = \|v(t)\|_{L_p([0,T])}, \tag{62}$$

where

$$v(t) = \|w\|_{H(\Omega)}(t).$$

In the L_2- case this norm reduces to the simple integral expression:

$$\|w\|_{L_2([0,T],L_2(\Omega))}^2 = \|w\|_{L_2(\Omega \times [0,T])}^2 = \int_0^T \int_\Omega w^2 dx dt. \tag{63}$$

We shall denote the exact analytic solution by $u(x,t)$, $(x,t) \in \Omega \times [0,T]$ and the numerical solution by $u_h(x,t)$. By definition we have:

$$u_h^n(x) = u_h(x,t^n), \quad \bar{u}_h^{n-1}(x) = u_h^{n-1}(\bar{x}), \tag{64}$$

where $\bar{x}$ is defined by (25). Finally we define the differences:

$$\eta = u - w_h, \quad \xi = u_h - w_h, \quad \zeta = u - u_h. \tag{65}$$

where the elliptic projection w_h is to be defined in section 4.3.3, and we note that K denotes a positive generic constant.

4.2 Error-estimate for the Characteristic Solution

We shall develop an error-estimate in the maximum norm for the characteristic solution defined by equation (25). The estimate obtained will then show that the characteristic solution gives a good approximation to the solution of equation (18) on Ω_o.

Let u_h^n be the discrete solution to the characteristic problem (25) in each time step. For simplicity we assume that u_h^n is piecewise linear on a uniform grid $\{x_i\}_{i=0}^N$, with mesh size h_o, although the analysis is completely valid also on adaptive grids.

We further assume that the characteristic equation may be solved exactly in each time step and let v^n be the solution obtained by integrating backwards along the characteristics from $t = t^n$ to $t = t^{n-1}$ such that:

$$v^n(x) = \bar{u}_h^{n-1}. \tag{66}$$

Hence, by definition $u_h^n(x)$ is the linear interpolant of $v(x)$ at the nodes x_i.

If $u_h^{n-1}(x)$ is monotone and decreasing and $\bar{f}$ represents a fully established shock, $v(x)$ is uniquely determined in each time step, since one part of the solution represents a rarefaction and the second part is transported along parallel characteristics. If a more general $\bar{f}$ is used, e.g. representing a growing shock, the following condition is needed:

$$\max \left| \frac{du_h^{n-1}}{dx} \frac{d^2\bar{f}}{du^2} \right| \Delta t < 1, \quad n = 1, 2, \ldots, T/\Delta t, \tag{67}$$

where additional conditions are required at the points where $\bar{f}$ is not sufficiently differentiable. We shall discuss such conditions later. Here, however, we shall assume that (67) represents sufficient conditions to assure uniqueness on the characteristic solution in each time step.

By the triangle inequality the error is bounded by:

$$\|u^n - u_h^n\|_{\infty,\Omega} \leq \|u^n - v^n\|_{\infty,\Omega} + \|v^n - u_h^n\|_{\infty,\Omega}. \tag{68}$$

We shall show that the last term on the right hand side of this inequality, i.e. the interpolation error, is bounded by:

$$\|v^n - u_h^n\|_{\infty,\Omega} \leq K h_o \Delta t. \tag{69}$$

Let $\tilde{x}_i$ denote the shifted nodes defined by:

$$\tilde{x}_i = x_i + \Delta t \bar{f}'(u_h^{n-1}(x_i)), \tag{70}$$

which is equivalent with forward stepping along the characteristics from a known point. Since u_h^{n-1} by definition is piecewise linear with nodes x_i, $i = 0, 1, \ldots, N$, the derivatives of $v^n(x)$ on each of the shifted intervals $(\tilde{x}_i, \tilde{x}_{i+1})$ are given by:

$$\frac{d^j v}{dx^j} = \frac{d\bar{u}_h^{n-1}}{d\bar{x}} \frac{d^j \bar{x}}{dx^j}, \quad j = 1, 2, \ldots$$

We note that $\bar{f}''(u)$ is discontinuous for $u = u_{BL}$. In the following we shall assume that $\bar{f}$ is sufficiently differentiable to assure that $\bar{x}(x)$ is two times continuously differentiable on Ω, except for the point corresponding to $u = u_{BL}$. This point can easily be treated in the analysis that follows and we may also allow for a finite number of such points. However, to simplify the analysis somewhat we shall assume that $\bar{x}(x) \in C^2(\Omega)$, implying that $v(x) \in C^2(\tilde{x}_i, \tilde{x}_{i+1})$, $i = 0, 1, \ldots, N-1$.

Without loss of generality we will show (69) for the case $x_i < x \leq \tilde{x}_i \leq x_{i+1}$, see Figure 3. For convenience we drop sub- and superscript on $v^n(x)$ and $u_h^n(x)$ and let:

$$\begin{aligned} \delta_{i+1/2} &\stackrel{\text{def}}{=} \frac{\bar{f}'_{i+1} - \bar{f}'_i}{2}, \\ \bar{f}'_{i+1/2} &\stackrel{\text{def}}{=} \frac{\bar{f}'_{i+1} + \bar{f}'_i}{2}, \end{aligned} \tag{71}$$

where $\bar{f}'_i = \bar{f}'(u_h^{n-1}(x_i))$. We note that the following relations are satisfied:

$$\bar{f}'_{i+1} = \bar{f}'_{i+1/2} + \delta_{i+1/2}, \tag{72}$$

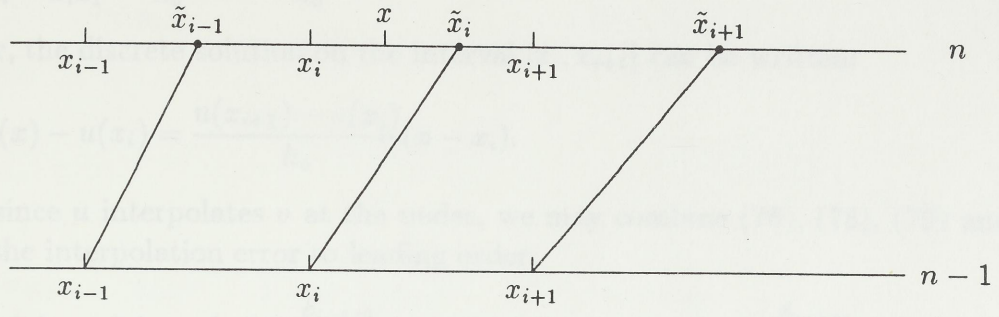


Figure 3: Shift of nodal values along characteristics between successive time steps.

$$\bar{f}_i' = \bar{f}_{i+1/2}' - \delta_{i+1/2}, \tag{73}$$

$$\tilde{x}_i - \tilde{x}_{i-1} = h_o + 2\delta_{i-1/2}\Delta t \tag{74}$$

and

$$2|\delta_{i+1/2}| < \delta h_o, \tag{75}$$

where $\delta = \max |\frac{du_h^{n-1}}{dx}\frac{d^2\bar{f}}{du^2}|$. Using the smoothness of $v(x)$ between shifted nodes, leads to the expansion:

$$v(x) - v(x_i) = \frac{v(\tilde{x}_i) - v(\tilde{x}_{i-1})}{\tilde{x}_i - \tilde{x}_{i-1}}(x - x_i) + \frac{d^2v}{dx^2}(\xi)(\tilde{x}_i + \tilde{x}_{i-1} - 2\xi)(x - x_i), \tag{76}$$

where $x_i \leq \xi \leq x$ and higher order terms are neglected. Using the smoothness of v once more, respectively on the intervals $(\tilde{x}_{i-1}, x_i)$ and $(\tilde{x}_i, x_{i+1})$, leads further to the expansions:

$$\begin{aligned} v(\tilde{x}_{i-1}) &= v(x_i) + \frac{dv}{dx}(x_i)(\tilde{x}_{i-1} - x_i), \\ v(\tilde{x}_i) &= v(x_{i+1}) + \frac{dv}{dx}(x_{i+1})(\tilde{x}_i - x_{i+1}). \end{aligned} \tag{77}$$

If $\tilde{x}_i = x_{i+1}$, the last expansion reduces to $v(\tilde{x}_i) = v(x_{i+1})$ and the analysis is somewhat simplified. Since

$$\tilde{x}_{i-1} - x_i = \Delta t \bar{f}_i' - h_o,$$

$$\tilde{x}_i - x_{i+1} = \Delta t \bar{f}_{i+1}' - h_o,$$

it follows from (72), (73) and (77) that:

$$\begin{aligned} v(\tilde{x}_i) - v(\tilde{x}_{i-1}) = \; & v(x_{i+1}) - v(x_i) + (\frac{dv}{dx}(x_{i+1}) - \frac{dv}{dx}(x_i))(\bar{f}_{i+1/2}'\Delta t - h_o) \\ & + (\frac{dv}{dx}(x_{i+1}) + \frac{dv}{dx}(x_i))\delta_{i+1/2}\Delta t. \end{aligned} \tag{78}$$

We observe that inequality (67) can equivalently be written $\delta\Delta t < 1$ which gives:

$$\frac{1}{\tilde{x}_i - \tilde{x}_{i-1}} \approx \frac{1}{h_o}(1 - 2\frac{\delta_{i-1/2}}{h_o}\Delta t). \tag{79}$$

Further, the discrete solution on the interval $[x_i, x_{i+1}]$ can be written:

$$u(x) - u(x_i) = \frac{u(x_{i+1}) - u(x_i)}{h_o}(x - x_i). \tag{80}$$

Then, since u interpolates v at the nodes, we may combine (76), (78), (79) and (80) to get the interpolation error to leading order:

$$\begin{aligned} v(x) - u(x) = \; & [-2K_1\frac{\delta_{i-1/2}}{h_o}\Delta t + K_2(\bar{f}'_{i+1/2}\Delta t - h_o) + K_3\frac{\delta_{i+1/2}}{h_o}\Delta t \\ & + K_4(\tilde{x}_i + \tilde{x}_{i-1} - 2\xi)](x - x_i) + O(\Delta t^2 h_o, \Delta t\, h_o^2), \end{aligned} \tag{81}$$

where

$$\begin{aligned} K_1 &= \frac{v(x_{i+1}) - v(x_i)}{h_o} = \frac{du_h^n}{dx}, \\ K_2 &= \frac{\frac{dv}{dx}(x_{i+1}) - \frac{dv}{dx}(x_i)}{h_o} \approx \frac{d^2v}{dx^2}(\xi), \\ K_3 &= \frac{dv}{dx}(x_{i+1}) + \frac{dv}{dx}(x_i), \\ K_4 &= \frac{d^2v}{dx^2}(\xi). \end{aligned}$$

The desired result is now obtained from equation (81), inequality (75), and the bounds:

$$\begin{aligned} & |x - x_i| < \min\{\bar{f}'_i\Delta t, h_o\}, \\ & |\bar{f}'_{i+1/2}\Delta t - h_o| \le h_o, \\ & |\tilde{x}_i + \tilde{x}_{i-1} - 2\xi| \le h_o. \end{aligned} \tag{82}$$

For completeness we note that the case $x = x_i$ is trivially satisfied, further, the cases $\tilde{x}_i \le x \le x_{i+1}$ and $\tilde{x}_i > x_{i+1}$ may be treated similarly without changing the outcome. We also note that the coefficients K_1, K_2, K_3, and K_4 are large on the inner region where the derivatives by assumption are large.

Finally we may remark that δ/h_o, K_2, and K_3 can be interpreted as second order derivative terms, hence the interpolation error can be viewed as numerical diffusion. We observe that if the characteristics are parallel lines, i.e. $\delta \equiv 0$, and the $\tilde{x}_i$'s coincide with the nodes, then the numerical diffusion is exactly zero. This is only possible if the fractional flow function is linear.

The first term on the right hand side of (68), is bounded as follows:

$$\|u^n - v^n\|_{\infty,\Omega} \le \left\|u^n - \bar{u}^{n-1}\right\|_{\infty,\Omega} + \left\|\bar{u}^{n-1} - \bar{u}_h^{n-1}\right\|_{\infty,\Omega}. \tag{83}$$

Using the coordinate transformation defined by the characteristic equation (25) and the fact that $u(x_0, t) = u_h(x_0, t)$ is known at the boundary, the second term on the right hand side of this inequality reduces to:

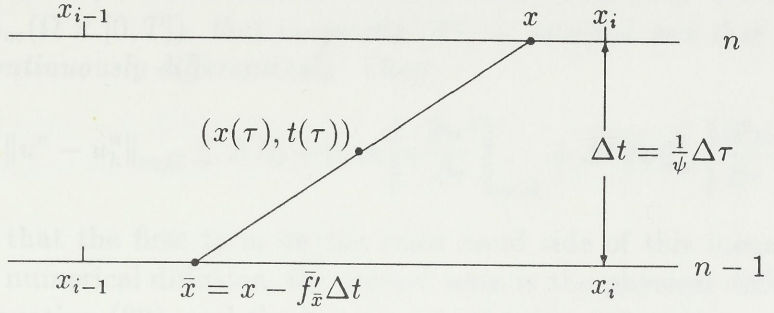


Figure 4: Approximate characteristics between successive time steps.

$$\left\|\bar{u}^{n-1} - \bar{u}_h^{n-1}\right\|_{\infty,\Omega} \leq \left\|u^{n-1} - u_h^{n-1}\right\|_{\infty,\Omega}. \tag{84}$$

To obtain a bound for $\|u^n - \bar{u}^{n-1}\|_{\infty,\Omega}$ we shall use an integral representation along the characteristic curves, see [4]. Let $\mathbf{r}(\tau) = [x(\tau), t(\tau)]$ denote the characteristic curve through the points (x, t^n) and $(\bar{x}, t^{n-1})$ as shown by Figure 4, and integrate along this curve to obtain:

$$u^n - \bar{u}^{n-1} = \Delta t\, \psi \frac{\partial u^n}{\partial \tau} - \int_{(\bar{x},t^{n-1})}^{(x,t^n)} \sqrt{(x(\tau)-\bar{x})^2 + (t(\tau)-t^{n-1})^2}\, \frac{\partial^2 u}{\partial \tau^2} d\tau. \tag{85}$$

The line integral in this formula may equivalently be expressed as:

$$\int_{(\bar{x},t^{n-1})}^{(x,t^n)} \sqrt{(x(\tau)-\bar{x})^2 + (t(\tau)-t^{n-1})^2}\, \frac{\partial^2 u}{\partial \tau^2} d\tau = \int_{\tau^{n-1}}^{\tau^n} (\tau - \tau^{n-1}) \frac{\partial^2 u}{\partial \tau^2} d\tau.$$

We have implicitly assumed that if the characteristic curve crosses the boundary, then $\bar{x} = x_0$ and $\bar{u}^{n-1}$, t^{n-1} and τ^{n-1} are consistently adjusted with the boundary values.

Since the characteristics are straight lines between successive time steps as shown in Figure 4, we get:

$$\left| \int_{\tau^{n-1}}^{\tau^n} (\tau - \tau^{n-1}) \frac{\partial^2 u}{\partial \tau^2} d\tau \right| \leq \Delta t^2 \|\psi\|_\infty^2 \left\| \frac{\partial^2 u}{\partial \tau^2} \right\|_{\infty,\Omega\times[t^{n-1},t^n]}, \tag{86}$$

where $\|\psi\|_\infty = \|\psi\|_{\infty,\Omega\times[0,T]}$. Hence, from (85) and (86) we obtain the estimate:

$$\left\|u^n - \bar{u}^{n-1}\right\|_{\infty,\Omega} \leq \Delta t \left\| \psi \frac{\partial u^n}{\partial \tau} \right\|_{\infty,\Omega} + \Delta t^2 \|\psi\|_{\infty,\Omega\times[0,T]}^2 \left\| \frac{\partial^2 u}{\partial \tau} \right\|_{\infty,\Omega\times[t^{n-1},t^n]}^2. \tag{87}$$

We finally combine (68), (69), (83), (84) and (87) to get the inequality:

$$\begin{aligned} \|u^n - u_h^n\|_{\infty,\Omega} \leq\ & \left\|u^{n-1} - u_h^{n-1}\right\|_{\infty,\Omega} + Kh_0\Delta t + \\ & \Delta t \left\| \psi \frac{\partial u^n}{\partial \tau} \right\|_{\infty,\Omega} + \Delta t^2 \|\psi\|_\infty^2 \left\| \frac{\partial^2 u}{\partial \tau} \right\|_{\infty,\Omega\times[t^{n-1},t^n]}. \end{aligned}$$

We may use this inequality recursively together with the inequality $n \leq T/\Delta t$ and the initial condition $u^0(x) = u_h^0(x)$, to obtain an upper error bound associated with the purely characteristic solution. The result is summarized in the following theorem:

Theorem 2 *Let u be the solution to the parabolic problem (18), and let u_h^n be the linear interpolant of the characteristic solution defined by (25). We assume that u, $\frac{\partial u}{\partial \tau}$ and $\frac{\partial^2 u}{\partial \tau^2} \in L_\infty(\Omega \times [0,T])$, that inequality (67) is satisfied and that $\bar{x}(x)$ is piecewise two times continuously differentiable. Then:*

$$\max_{0 \le t^n \le T} \|u^n - u_h^n\|_{\infty,\Omega} \le K h_o + \max \left\| \psi \frac{\partial u^n}{\partial \tau} \right\|_{\infty,\Omega} + \Delta t \, \|\psi\|_\infty^2 \left\| \frac{\partial^2 u}{\partial \tau} \right\|_{\infty,\Omega \times [0,T]} \tag{88}$$

We note that the first term on the right hand side of this inequality may be interpreted as numerical diffusion, the second term is the physical diffusion, which is of order ϵ by equation (29), and the last term is the time truncation error. If ϵ is small, the right hand side is small away from regions with large gradients, implying that the characteristic solution is close to the continuous solution in the outer region. In the inner region, some of the terms may become large even if h_o is small, which justifies a diffusion correction step in the solution procedure.

We finally remark that this estimate should be extendible to higher spatial dimensions.

4.3 H^1-estimate for the Shock Region

4.3.1 Formulation of the Inner Problem

Before we proceed with a formulation of the problem defining the inner solution, we notice that the boundary values associated with this problem are not exact. However, since the characteristic solution given on the coarse grid can be assumed to be fairly accurate, we shall carry out the error analysis under the assumption of exact boundary values.

The complete inner equations are given by:

$$\frac{\partial u}{\partial t} + \bar{f}'(u)\frac{\partial u}{\partial x} + \frac{\partial}{\partial x}(b(u)u) - \epsilon \frac{\partial}{\partial x}(a(u)\frac{\partial u}{\partial x}) = g(x,t), \quad x \in \Omega_i(t), \tag{89}$$

and

$$\begin{aligned} u(x,0) &= u_0(x), \qquad x \in \Omega_i(t) \\ u(x,t) &= 0, \qquad (x,t) \in \partial\Omega_i(t) \times [0,T], \end{aligned} \tag{90}$$

where $u_0(x) = 0$ on $\partial\Omega_i(0)$ and the nonzero boundary is taken care of by the right hand side of the equation.

Since the inner region is transported with the shock velocity between successive time steps, a natural shift of coordinates are given by:

$$\begin{aligned} x^* &= x - v_{BL}t, \\ t^* &= t. \end{aligned} \tag{91}$$

The inner region is then transformed from the parallelogram $\Omega_i(t) \times [0,T]$ in (x,t)-space, to the rectangular region $\Omega_i(0) \times [0,T]$ in (x^*,t)-space. Further, the derivatives transform as:

$$\begin{aligned} \frac{\partial}{\partial t} &= \frac{\partial}{\partial t^*} - v_{BL}\frac{\partial}{\partial x^*}, \\ \frac{\partial}{\partial x} &= \frac{\partial}{\partial x^*}. \end{aligned} \tag{92}$$

To avoid technical difficulties we shall confine ourselves to the somewhat simpler problem obtained by replacing the nonlinear coefficients with:

$$\begin{aligned} b(u) &\leftarrow b(u_0(x - v_{BL}t)) = b(x^*), \\ a(u) &\leftarrow a(u_0(x - v_{BL}t)) = a(x^*). \end{aligned} \tag{93}$$

Thus, rather than to update the coefficients in each time step by the characteristic solution, we follow the approximate characteristics given by the shock velocity, backwards to the initial profile. By doing this we will lose a bound on the time step caused by the linearization defined by (34), the basic ideas involved in developing an error-estimate should however still remain.

Substitution of (91),(92) and (93) into (89) and (90) then give the inner problem:

$$\frac{\partial u}{\partial t} + (\bar{f}'(u) - v_{BL})\frac{\partial u}{\partial x} + \frac{\partial}{\partial x}(b(x)u) - \epsilon\frac{\partial}{\partial x}(a(x)\frac{\partial u}{\partial x}) = g(x,t),\ x \in \Omega_i, \tag{94}$$

and

$$\begin{aligned} u(x,0) &= u_0(x), \quad && x \in \Omega_i \\ u(x,t) &= 0, && (x,t) \in \partial\Omega_i \times [0,T], \end{aligned} \tag{95}$$

where $u_0(x) = 0$ on $\partial\Omega_i$. For convenience we have dropped asterisks on the independent variables.

We note that by definition, $a(u_0(x^*))$ is zero on the part of the inner region Ω_i where $u_0 = 0$. Contrary to this, we may argue that $a(u)$ is strictly positive in the boundary layer defining the shock region; Ahead of the shock, where $u \equiv 0$, the solution is completely determined by convection, since diffusion processes are zero when $u = 0$. Hence, the right boundary of $\Omega_i(t)$, should be defined in terms of a curve dividing $\Omega \times [0,T]$ into a zero and a nonzero part. This curve is of course almost parallel to the characteristics defined by the shock velocity, but it may not coincide with any characteristic. The importance of this observation, however, is that the diffusion coefficient should be nonzero inside the shock region.

The rest of this section is organized as follows: First we obtain the variational formulation of the inner problem and the associated discrete equations. The symmetrization technique given by Barrett and Morton [6] is used and we state some properties needed on the Riesz-representation. The error-estimate is then developed following similar lines as Douglas and Russell [4]. We end this section with a discussion of approximate symmetrization leading to a similar result as given in Theorem 1.

4.3.2 Discrete Equations

We shall require the following standard regularity assumptions on the solution $u(x,t)$ of (94):

$$\begin{aligned} &\text{(a)}\ \ u \in L_\infty(0,T;H^q(\Omega_i)),\\ &\text{(b)}\ \ \frac{\partial u}{\partial t} \in L_2(0,T;H^{q-1}(\Omega_i)),\\ &\text{(c)}\ \ \frac{\partial^2 u}{\partial t^2} \in L_2(0,T;L_2(\Omega_i)), \end{aligned} \tag{96}$$

where $q \geq 2$.

Proceeding as before, we define the characteristic direction $\tau(u)$ in terms of the operator:

$$\frac{\partial}{\partial \tau(u)} = \frac{1}{\psi(u)}\left(\frac{\partial}{\partial t} + (\bar{f}'(u) - v_{BL})\frac{\partial}{\partial x}\right),$$

where

$$\psi(u) = \sqrt{1 + (\bar{f}'(u) - v_{BL})^2}.$$

Hence, the characteristics between successive time steps are given by:

$$\begin{aligned} &\bar{x} = x - \Delta t\left(\frac{d\bar{f}}{du}(\bar{u}^{n-1}) - v_{BL}\right),\\ &\bar{u}^{n-1} = u^{n-1}(\bar{x}). \end{aligned} \tag{97}$$

We note that inequality (67) is needed to assure uniqueness on the characteristic solution. The characteristic derivative will be replaced by the expression:

$$\psi\frac{\partial u}{\partial \tau} \approx \frac{u(x,t^n) - u(\bar{x},t^{n-1})}{\Delta t}. \tag{98}$$

Since the characteristic equation (97) is homogeneous in space, the error in this approximation is due to the continuous change in saturation along the characteristics caused by diffusion, as was shown in the previous section.

By using the characteristic derivative we may write equation (94) and (95) in the equivalent form:

$$\begin{aligned} &(\psi\frac{\partial u}{\partial \tau}, v) + B(u,v) = (g,v), \ \ \forall v \in H_0^1, \ \ t \in (0,T],\\ &B(u - u_0, v) = 0, \ \ \forall v \in H_0^1, \ \ t = 0, \end{aligned} \tag{99}$$

where

$$B(u,v) = ((bu)', v) + \epsilon(au', v').$$

We let S^h define our discrete trial space, $S^h \subset H^1(\Omega_i)$, spanned by $\{\theta_i\}$, and we shall assume that:

$$\forall w \in H^s : \inf_{X \in S^h}\left\{\|w - X\|_{0,\Omega_i} + h\|w - X\|_{1,\Omega_i}\right\} \leq Kh^s\|w\|_{s,\Omega_i}.$$

where $s \leq q$.

Since $B(u,v)$ is a linear continuous form in both its arguments, we may also assume existence and uniqueness of the Riesz-representation R^m given by:

$$B(u,v) = B^m(u, R^m v), \tag{100}$$

where $B^m(u,v) = \epsilon(a_0 u', v')$, $a_0 = a - bh/\epsilon$, defines an inner product on $H_0^1(\Omega_i)$.

Using the symmetrization technique introduced by Barrett and Morton [6], we choose our test space T_0^h, spanned by $\{\psi_i\}$ such that:

$$span\{R^m \psi_i\} = span\{\theta_i\}.$$

We note that for computational purposes, any linearly independent set of functions spanning T_0^h gives an equivalent linear system to solve, whereas for theoretical use, we shall need that the Riesz-mapping is *onto*, i.e.:

$$\forall \theta \in S_0^h, \exists \psi \in T_0^h : \psi = R^{m-1}\theta. \tag{101}$$

By (100) and (101), the discrete Galerkin equations may then be written: Find $u_h^n \in S_0^h$, $n = 1, 2, \ldots N$, such that

$$\begin{aligned} &(\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t}, R^{m-1}\theta) + B^m(u_h, \theta) = (g, R^{m-1}\theta), \quad \forall \theta \in S_0^h, \\ &B^m(u_h^0 - u_0, \theta) = 0, \quad \forall \theta \in S_0^h. \end{aligned} \tag{102}$$

We shall also require that the inverse representation satisfy:

$$\begin{aligned} &\text{I.} \quad \forall \theta \in S_0^h : (\theta, R^{m-1}\theta) \geq 0, \\ &\text{II.} \quad \forall \theta \in S_0^h : (R^{m-1}\theta, R^{m-1}\theta) \leq M_\epsilon(\theta, R^{m-1}\theta), \end{aligned} \tag{103}$$

where M_ϵ is a positive constant that may depend on ϵ, h and the coefficients $a(x)$ and $b(x)$.

We do not intend to prove I and II in the general case, which seems to be quite difficult. However, it is reasonable to assume that I and II is satisfied for small β, since $R^{m-1}\theta \to \theta$ when $\beta \to 0$. In the case of constant coefficients, we have worked out the following result in the opposite limit:

Lemma 1 *Let $\theta = \sum_{i=1}^{N-1}$ and let ψ_i be the optimal test functions associated with θ_i and defined by (47). Then, for sufficiently small ϵ, $(\theta, R^{m-1}\theta)$ is strictly positive and we get the following estimate on M_ϵ in (103):*

$$M_\epsilon \sim \frac{|b|h}{\epsilon a} = |\beta|. \tag{104}$$

A proof of this lemma is given in the appendix. We expect a similar result to be valid in general cases, with weak conditions on the coefficients $a(x)$ and $b(x)$. Further, the somewhat artificial condition II, seems intuitively to be superfluous in the error estimate, although it is required in the analysis leading to the estimate. The fact that $R^{m-1}\theta$ and the energy norm $\|\cdot\|_{B^m}$ is well defined in the limit $\epsilon \to 0$, (a_0 is strictly positive), therefor suggests that the error analysis is possible to carry out even if $a(x)$ is singular and condition II not satisfied. We shall discuss this point in more detail after the estimate is obtained.

Existence and uniqueness of the solution of the discrete system (102), now follows from property I and the fact that B^m defines an inner product on Ω_i.

4.3.3 Error-estimate on Inner Region

We define the elliptic projection w_h of the solution u by the equation:

$$B^m(u - w_h, \theta) = 0, \quad \forall \theta \in S_0^h, \quad 0 \leq t \leq T. \tag{105}$$

Standard results from elliptic theory then show that for $p = 2$ or ∞ and $1 \leq s \leq q$:

$$\|\eta\|_{L_p(0,T;H^1(\Omega_i))} \leq K h^{s-1} \|u\|_{L_p(0,T;H^s(\Omega_i))}, \tag{106}$$

where η is defined by equation (65). Since $B^m(\cdot,\cdot)$ is linear and independent of time, $\frac{\partial \eta}{\partial t}$ is also a solution of (105). Combining this with the regularity condition (96)(b) and the estimate above, gives for $q \geq 2$ and $1 \leq s \leq q$:

$$\left\| \frac{\partial \eta}{\partial t} \right\|_{L_2(0,T;L_2(\Omega_i))} \leq K h^{s-1} \left\| \frac{\partial u}{\partial t} \right\|_{L_2(0,T;H^{s-1}(\Omega_i))}. \tag{107}$$

Writing (99) in a similar form as (102), using the Riesz-representation, and subtracting these equations gives:

$$(\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t} - \psi \frac{\partial u^n}{\partial \tau}, R^{m-1}\theta) + B^m(u_h^n - u^n, \theta) = 0, \quad \forall \theta \in S_0^h.$$

From the definition of the elliptic projection w_h and the definitions (65), this equation may further be written:

$$\begin{aligned} &(\frac{\xi^n - \bar{\xi}^{n-1}}{\Delta t}, R^{m-1}\theta) + B^m(\xi^n, \theta) = \\ &(\psi \frac{\partial u^n}{\partial \tau} - \frac{u^n - \bar{u}^{n-1}}{\Delta t}, R^{m-1}\theta) + (\frac{\eta^n - \bar{\eta}^{n-1}}{\Delta t}, R^{m-1}\theta) \end{aligned}, \quad \forall \theta \in S_0^h. \tag{108}$$

To continue we will need the inequality:

$$(u, v) \leq \frac{1}{2\delta} \|u\|_{0,\Omega_i}^2 + \frac{\delta}{2} \|v\|_{0,\Omega_i}^2, \tag{109}$$

which is derived from the Cauchy-Schwartz inequality and the inequality:

$$ab \leq \frac{1}{2\delta} a^2 + \frac{\delta}{2} b^2.$$

It follows from (109) that the first term on the right hand side of (108) is bounded by:

$$(\psi \frac{\partial u^n}{\partial t} - \frac{u^n - \bar{u}^{n-1}}{\Delta t}, R^{m-1}\theta) \leq \frac{1}{2\delta} \left\| \psi \frac{\partial u^n}{\partial t} - \frac{u^n - \bar{u}^{n-1}}{\Delta t} \right\|_{0,\Omega_i}^2 + \frac{\delta}{2} (R^{m-1}\theta, R^{m-1}\theta).$$

We will estimate the error in the approximation of the characteristic derivative from the integral representation given by (85). Since the characteristics are straight lines between successive time steps as shown in Figure 4, we may use the Cauchy-Schwartz inequality to get:

$$\left(\int_{\tau^{n-1}}^{\tau^n} (\tau - \tau^{n-1}) \frac{\partial^2 u}{\partial \tau^2} d\tau \right)^2 \leq \frac{1}{3} \psi^4 \Delta t^3 \int_{t^{n-1}}^{t^n} \left(\frac{\partial^2 u}{\partial \tau^2} \right)^2 dt. \tag{110}$$

By combining (85) and (86) we obtain the estimate:

$$\left\| \psi \frac{\partial u^n}{\partial t} - \frac{u^n - \bar{u}^{n-1}}{\Delta t} \right\|_{0,\Omega}^2 \leq$$
$$\Delta t^2 \left\| \psi \right\|_\infty^4 \int_{\Omega_o^n} \int_{t^{n-1}}^{t^n} \left| \frac{\partial^2 u}{\partial \tau} \left(\frac{t^n - t}{\Delta t} \bar{x} + \frac{t - t^{n-1}}{\Delta t} x, t \right) \right|^2 dt dx.$$

The last integral has to be measured in a standard norm. We therefor define the transformation:

$$S : (x,t) \mapsto (z,t) = \left(\frac{t^n - t}{\Delta t} \bar{x} + \frac{t - t^{n-1}}{\Delta t} x, t \right) = (\theta(t)\bar{x} + (1 - \theta(t)x, t). \tag{111}$$

Since the inner region is transported with the shock velocity, characteristics may diffuse out of the left boundary. Hence, S maps $\Omega \times [t^{n-1}, t^n]$ onto a smaller or equal region, say $\bar{W} \subseteq \Omega \times [t^{n-1}, t^n]$. The Jacobian of S is given by:

$$DS = \begin{pmatrix} 1 - \theta \Delta t \left(\frac{d\bar{f}}{du} \right)'(x) & \frac{d\bar{f}}{du}(x) \\ 0 & 1 \end{pmatrix},$$

and the determinant of the Jacobian matrix is further given by:

$$\det DS = 1 + O(\Delta t),$$

It follows that the map S is invertible for sufficiently small time steps Δt and that the integral on the right hand side of (87) is bounded by:

$$\begin{aligned} &\int_\Omega \int_{t^{n-1}}^{t^n} \left| \frac{\partial^2 u}{\partial \tau^2}(x(t),t) \right|^2 dt dx \leq \int_{\bar{W}} \left| \frac{\partial^2 u}{\partial \tau^2}(z,t) \right|^2 dt dz + O(\Delta t) \leq \\ &\int_\Omega \int_{t^{n-1}}^{t^n} \left| \frac{\partial^2 u}{\partial \tau^2}(z,t) \right|^2 dt dz + O(\Delta t) = \left\| \frac{\partial^2 u}{\partial \tau^2} \right\|_{L^2(\Omega \times [t^{n-1}, t^n])}^2 + O(\Delta t). \end{aligned} \tag{112}$$

Thus, the error due to the time discretization is bounded by:

$$\left\| \psi \frac{\partial u^n}{\partial t} - \frac{u^n - \bar{u}^{n-1}}{\Delta t} \right\|_{0,\Omega}^2 \leq K \Delta t \left\| \frac{\partial^2 u}{\partial \tau^2} \right\|_{L_2(\Omega \times [t^{n-1}, t^n])}^2.$$

The first term on the right hand side of (108) may therefor be estimated as follows:

$$\frac{1}{2\delta} K \Delta t \left\| \frac{\partial^2 u}{\partial \tau^2} \right\|_{L_2(\Omega \times [t^{n-1}, t^n])}^2 + \frac{\delta}{2} (R^{m-1}\theta, R^{m-1}\theta). \tag{113}$$

Using inequality (109) once more implies that the last term on the right hand side of (108) is bounded by:

$$\left(\frac{\eta^n - \bar{\eta}^{n-1}}{\Delta t}, R^{m-1}\theta \right) \leq \frac{1}{2\delta} \left\| \frac{\eta^n - \bar{\eta}^{n-1}}{\Delta t} \right\|_{0,\Omega_i}^2 + \frac{\delta}{2} (R^{m-1}\theta, R^{m-1}\theta).$$

By the triangle inequality we get:

$$\frac{1}{2}\left\|\frac{\eta^n-\bar{\eta}^{n-1}}{\Delta t}\right\|^2_{0,\Omega_i} \le \left\|\frac{\eta^n-\eta^{n-1}}{\Delta t}\right\|^2_{0,\Omega_i} + \left\|\frac{\eta^{n-1}-\bar{\eta}^{n-1}}{\Delta t}\right\|^2_{0,\Omega_i}.$$

From the integral representation:

$$\eta^n-\eta^{n-1}=\int_{t^{n-1}}^{t^n}\frac{\partial\eta}{\partial t}dt,$$

the Cauchy-Shwartz inequality and (63) we obtain:

$$\left\|\frac{\eta^n-\eta^{n-1}}{\Delta t}\right\|^2_{0,\Omega_i} = \frac{1}{\Delta t^2}\int_{\Omega_i}\left(\int_{t^{n-1}}^{t^n}\frac{\partial\eta}{\partial t}dt\right)^2 dx \le$$
$$\frac{1}{\Delta t}\int_{t^{n-1}}^{t^n}\int_{\Omega_i}\left(\frac{\partial\eta}{\partial t}\right)^2 dxdt = \frac{1}{\Delta t}\left\|\frac{\partial\eta}{\partial t}\right\|_{L_2([t^{n-1},t^n],L_2(\Omega_i))}.$$

Further, Taylor expansion of $\eta^{n-1}-\bar{\eta}^{n-1}$ gives:

$$\left\|\frac{\eta^{n-1}-\bar{\eta}^{n-1}}{\Delta t}\right\|^2_{0,\Omega_i} \approx \left\|\frac{1}{\Delta t}\frac{\partial\eta^{n-1}}{\partial x}(x-\bar{x})\right\|^2_{0,\Omega_i} \le K\left\|\frac{\partial\eta^{n-1}}{\partial x}\right\|^2_{0,\Omega_i} = K\left\|\eta^{n-1}\right\|^2_{1,\Omega_i},$$

where the inequality follows from $|x-\bar{x}| \le K\Delta t$ by (97). Thus, the last term on the right hand side is bounded by:

$$\frac{1}{\delta}\left(\frac{K}{\Delta t}\left\|\frac{\partial\eta}{\partial t}\right\|_{L_2([t^{n-1},t^n],L_2(\Omega_i))} + K\left\|\eta^{n-1}\right\|^2_{1,\Omega_i}\right) + \frac{\delta}{2}(R^{m-1}\theta, R^{m-1}\theta). \tag{114}$$

We note that this bound is nonoptimal in the sense that a small "derivative" along the characteristics have been represented with derivatives along the coordinate-axes where the gradients by assumption are large.

To derive a bound on the solution of (102), we choose the test function:

$$\theta = \frac{\xi^n-\hat{\xi}^{n-1}}{\Delta t} \tag{115}$$

where $\hat{\xi}^{n-1} \in S^h$ is the interpolation of $\bar{\xi}^{n-1}$. We note that $\bar{\xi}^{n-1}$ in general is not an element of the trial space S^h.

A simple calculation show that:

$$B^m(\xi^n, \frac{\xi^n-\hat{\xi}^{n-1}}{\Delta t}) \le \frac{1}{2\Delta t}\left[B^m(\xi^n,\xi^n) - B^m(\hat{\xi}^{n-1},\hat{\xi}^{n-1})\right].$$

To obtain the desired estimate, we shall replace $\hat{\xi}^{n-1}$ with $\bar{\xi}^{n-1}$. This introduces an interpolation error which may be difficult to give an exact expression for in general cases. However, if the characteristics are parallel lines, which is mainly the case in the shock region, and the functions in S^h are piecewise linear, lemma 2 in the appendix shows that:

$$\left\|\hat{\xi}\right\|^2_1 = \left\|\bar{\xi}\right\|^2_1 - k(1-k)h^2\sum_i h\left(\frac{\bar{\xi}'_{i+1}-\bar{\xi}'_i}{h}\right)^2, \tag{116}$$

where $0 \leq k \leq 1$, $k = 0$ or 1 if the shifted nodes coincide with the regular grid. The last term of equation (116) may be interpreted as a numerical diffusion term. In the general case we shall assume that:

$$B^m(\hat{\xi}^{n-1}, \hat{\xi}^{n-1}) \leq (1 + Kk(1-k)h)B^m(\bar{\xi}^{n-1}, \bar{\xi}^{n-1}) + k(1-k)h^2D^2,$$

where D is an upper bound on the numerical diffusion caused by the interpolation. We may also assume that higher order interpolation reduces the effect of numerical diffusion. From the change of coordinates (97), we obtain:

$$B^m(\bar{\xi}^{n-1}, \bar{\xi}^{n-1}) \leq (1 + K\Delta t)B^m(\xi^{n-1}, \xi^{n-1}), \tag{117}$$

where we have allowed for divergent characteristics.

The first term on the left hand side of (108) can be written:

$$(\frac{\xi^n - \bar{\xi}^{n-1}}{\Delta t}, R^{m-1}\theta) = (\theta, R^{m-1}\theta) + (\frac{\hat{\xi}^{n-1} - \bar{\xi}^{n-1}}{\Delta t}, R^{m-1}\theta).$$

However, the interpolation error introduced on the left hand side of this equation can be neglected compared with the interpolation error in the previous term.

Thus, the left hand side of (108) is bounded below by the expression:

$$\begin{aligned}&(\theta, R^{m-1}\theta) + \frac{1}{2\Delta t}\left[B^m(\xi^n, \xi^n) - B^m(\xi^{n-1}, \xi^{n-1})\right] - \\ &-K_1 B^m(\xi^{n-1}, \xi^{n-1}) - k(1-k)\frac{h^2}{\Delta t}D^2,\end{aligned} \tag{118}$$

where $K_1 = O(1 + h/\Delta t)$. We typically have that $\Delta t = O(h_o)$, hence:

$$\frac{h}{\Delta t} \ll 1. \tag{119}$$

Combining inequality (113), (114) and (118) with (108), multiplying with Δt and choosing:

$$\delta = \frac{(\theta, R^{m-1}\theta)}{(R^{m-1}\theta, R^{m-1}\theta)}, \tag{120}$$

gives the recursion relation:

$$\begin{aligned}&B^m(\xi^n, \xi^n) - B^m(\xi^{n-1}, \xi^{n-1}) \leq K_1\Delta t B^(\xi^{n-1}, \xi^{n-1}) + \\ &+K\Delta t^2 \left\| \frac{\partial^2 u}{\partial \tau^2} \right\|^2_{L_2([t^{n-1}, t^n]; L_2(\Omega))} + K \left\| \frac{\partial \eta}{\partial t} \right\|^2_{L_2([t^{n-1}, t^n]; L_2(\Omega_i))} + \\ &+K\Delta t \|\eta^{n-1}\|^2_{1,\Omega_i} + k(1-k)h^2D^2.\end{aligned} \tag{121}$$

We note that the sharpness of this estimate depends on the number M_ϵ.

To continue we need the following discrete version of Gronwall's lemma:

Lemma 2 (Discrete Gronwall) *Assume that* $\|\xi^0\| = 0$ *and that for all* $n \leq N$, $\|\xi^n\|$ *satisfies the inequality*

$$\|\xi^n\| \leq B \sum_{i=1}^{n} \left\|\xi^{i-1}\right\| + A$$

where A and B are positive constants. Then it follows:

$$\max_{1\leq n\leq N} \|\xi^n\| \leq A \exp(N \cdot B)$$

For a proof of this lemma, see the appendix.

From (102) and (105) it follows that $\|\xi^0\|_{B^m} = 0$, thus, summing the recursion formula (121) in time, taking the square root and using the Gronwall lemma implies:

$$\begin{aligned} \max_{1\leq n\leq N} \|\xi^n\|_{B^m} \leq \; & K\Delta t \left\|\frac{\partial^2 u}{\partial \tau^2}\right\|_{L_2(0,T;L_2)} + K \left\|\frac{\partial \eta}{\partial t}\right\|_{L_2(0,T;L_2)} + \\ & + K \|\eta\|_{L_\infty(0,T;H^1)} + \sqrt{k(1-k)} \frac{h}{\Delta t^{1/2}} D, \end{aligned} \tag{122}$$

where we have used that

$$\Delta t \sum_{i=1}^{n} \left\|\eta^{n-1}\right\|_{1,\Omega_i}^2 \leq \|\eta\|_{L_\infty(0,T;H^1(\Omega_i))}^2 .$$

We note that the growth rate in the Gronwall lemma is small by the definition of K_1 and equation (119). Finally we get from (61) and (65):

$$\max_{1\leq n\leq N} \|\xi^n\|_{B^m} \geq \sqrt{c} K \left(\max_{1\leq n\leq N} \|u^n - u_h^n\|_{1,\Omega_i} - \|\eta\|_{L_\infty(0,T;H^1(\Omega_i))} \right),$$

hence, (106), (107) and (122) imply the following theorem:

Theorem 3 *Let $u(x,t)$ be the solution to the parabolic problem (99) and let u_h^n be the numerical approximation defined by equation (102). Assuming that $u(x,t)$ satisfy the regularity conditions (96), that the time step satisfy inequality (67) and that the inverse Riesz-representation satisfy the conditions (103) I and II, then the error in the approximate solution is bounded by:*

$$\begin{aligned} \max_{1\leq n\leq N} \|u^n - u_h^n\|_{B^m} \leq \; & K \left[\|u\|_{L_\infty(0,T;H^q(\Omega_i))} + \left\|\frac{\partial u}{\partial t}\right\|_{L_2(0,T;H^{q-1}(\Omega_i))} \right] h^{q-1} + \\ & + K\Delta t \left\|\frac{\partial^2 u}{\partial \tau^2}\right\|_{L_2(0,T;L_2(\Omega_i))} + \sqrt{k(1-k)} \frac{h}{\Delta t^{1/2}} D. \end{aligned} \tag{123}$$

We note that $q = 2$ if S^h is a linear interpolation space. We may further remark that if we had chosen

$$\theta = \frac{\xi^n - \xi^{n-1}}{\Delta t},$$

as a test function, we would obtain a similar expression as (118) without the numerical diffusion term, but with a $1/\epsilon$ factor in front of the negative B^m-term since the negative B^m-term this time would appear from the L_2-term. This would again lead to an inconvenient $\exp(1/\epsilon)$ factor in the discrete Gronwall lemma. The reason for this would of course once more be that a small "derivative" along the characteristics is

replaced by derivatives along the coordinate-axes where the gradients by assumption are large.

We also emphasize that the sharpness of the estimate in its present form depends on the number M_ϵ through the constants K. This number enters the estimate in a nonoptimal manner, when the L_2-term $(\theta, R^{m-1}\theta)$ on the left hand side of equation (108), is balanced against $(R^{m-1}\theta, R^{m-1}\theta)$ terms on the right hand side. We note that when inequality (109) is used to bound terms like $(\theta, R^{m-1}\theta)$ on the right hand side of equation (108), the original order of these terms is lost since $(R^{m-1}\theta, R^{m-1}\theta)$ terms are introduced. However, we have not been able to get a more accurate bound on the right hand side terms, such that this problem can be avoided, and is therefor dependent upon condition (103) II to obtain the estimate. We suspect that this condition is not necessary, and that the error estimate may not depend on the number M_ϵ in general.

We finally note that no arguments used to obtain the error estimate is limited to the one-dimensional case. By simple arguments, this estimate may be generalized to higher dimensions, although coupling to the pressure equations will give a more complex analysis in the general case, see [10,11].

4.3.4 Approximate Symmetrization

We end this section with a discussion of approximate test functions related to the theorem by Barrett and Morton stated previously.

We let T_0^h denote the approximate test space, and assume that the closeness with which $R^m T_0^h$ is approximating the trial space S_0^h is given by $\Delta_m(h)$ such that:

$$\inf_{\psi \in T_0^h} \|\theta - R^m \psi\|_{B^m} \leq \Delta_m \|\theta\|_{B^m} \quad \forall \theta \in S_0^h. \tag{124}$$

We shall further assume that an infimum, denoted ψ_θ, exists for all $\theta \in S_0^h$, and that these test function satisfy similar conditions as required by the inverse representation:

I. $\forall \theta \in S_0^h : (\theta, \psi_\theta) \geq 0$

II. $\forall \theta \in S_0^h : (\psi_\theta, \psi_\theta) \leq \tilde{M}_\epsilon (\theta, \psi_\theta)$,

where $\tilde{M}_\epsilon$ is a positive constants similar to M_ϵ.

Using the approximate optimal test functions, the Petrov-Galerkin equations are written: Find $u_h^n \in S_0^h$, $n = 1, 2, \ldots N$, such that

$$\begin{aligned} &(\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t}, \psi) + B(u_h, \psi) = (g, \psi), \quad \forall \psi \in T_0^1, \\ &B(u_h^0 - u_0, \psi) = 0, \quad \forall \psi \in T_0^1. \end{aligned} \tag{125}$$

Since equation (125) is linear and finite-dimensional, it suffices to show uniqueness to prove existence of a solution of the discrete equations. We shall therefor show that if $\theta_0 \in S_0^h$ is such that

$$(\theta_0, \psi) + B(\theta_0, \psi) = 0, \quad \forall \psi \in T_0^1, \tag{126}$$

then $\theta_0 \equiv 0$. To show this we note that (126) is satisfied for $\psi = \psi_{\theta_0}$. Choosing this as test function leads to

$$
\begin{aligned}
0 &= (\theta_0, \psi_{\theta_0}) + B(\theta_0, \psi_{\theta_0}) = \\
&= (\theta_0, \psi_{\theta_0}) + B^m(\theta_0, R^m \psi_{\theta_0}) \geq B^m(\theta_0, \theta_0) - B^m(\theta_0, \theta_0 - R^m \psi_{\theta_0}).
\end{aligned}
$$

From the Cauchy-Schwartz inequality and the assumption (124) on ψ_{θ_0} we obtain:

$$
B^m(\theta_0, \theta_0 - R^m \psi_{\theta_0}) \leq \|\theta_0\|_{B^m} \|\theta_0 - R^m \psi_{\theta_0}\|_{B^m} \leq \Delta_m \|\theta_0\|^2_{B^m}.
$$

Combining these results implies:

$$
(\theta_0, \psi_{\theta_0}) + (1 - \Delta_m) \|\theta_0\|^2_{B^m} \leq 0.
$$

and we deduce uniqueness and existence of a solution of (125) if $\Delta_m \in [0, 1)$ and condition I is satisfied.

An error-estimate is derived following identical lines as in the previous section. As in equation (108) we may write:

$$
\begin{aligned}
&(\frac{\xi^n - \bar{\xi}^{n-1}}{\Delta t}, \psi_\theta) + B^m(\xi^n, \theta) = (\frac{\eta^n - \bar{\eta}^{n-1}}{\Delta t}, \psi_\theta) + \\
&+ (\psi \frac{\partial u^n}{\partial t} - \frac{u^n - \bar{u}^{n-1}}{\Delta t}, \psi_\theta) + B^m(\xi^n, \theta - \psi_\theta)
\end{aligned}
\quad , \forall \theta \in S_0^h, \tag{127}
$$

where we have used that:

$$
B(\xi^n, \psi_\theta) = B^m(\xi^n, \theta) - B^m(\xi^n, \theta - R^m \psi_\theta).
$$

This relation can be treated exactly as before except for the last term on the right hand side. Using (124), this term is bounded by:

$$
B^m(\xi^n, \theta - R^m \psi_\theta) \leq \frac{1}{2\delta} \Delta_m^2 \|\xi^n\|^2_{B^m} + \frac{\delta}{2} \|\theta\|^2_{B^m}.
$$

Then, by choosing

$$
\theta = \frac{\xi^n - \hat{\xi}^{n-1}}{\Delta t},
$$

and

$$
\delta \approx \min\{\tilde{M}_\epsilon^{-1}, (\theta, \psi_\theta) / \|\theta\|^2_{B^m}\},
$$

as before, adding in time and using the discrete Gronwall lemma implies that:

$$
\max_{1 \leq n \leq N} \|u^n - u_h^n\|_{B^m} \leq (1 - K\Delta_m^2)^{-1/2} \{R.S.\}. \tag{128}
$$

where $\{R.S.\}$ denotes the right hand side of inequality (123). We note that the constant in front of Δ_m is proportional to the constant from the Gronwall lemma multiplied by δ^{-1}.

5 Solution Procedure, Two-dimensional Case

In the following section we shall extend the one-dimensional solution procedure to higher space dimensions. For simplicity we restricts ourselves to the two-dimensional case, partly because the computations performed are two-dimensional, yet we emphasize that the procedure described may as well be used in three space dimensions.

The methods to be outlined are mostly trivial extensions of the one-dimensional case, however, the exposition given is aimed to be fairly complete, and some arguments given previously will therefor be repeated.

5.1 Modified Method of Characteristics

The dominating part of the saturation equation (4), is the hyperbolic equation:

$$\frac{\partial u}{\partial t} + \nabla \cdot f(u)\mathbf{v} = 0. \tag{129}$$

Since the flow is incompressible, the divergence term may be written:

$$\nabla \cdot f(u)\mathbf{v} = f'(u)\mathbf{v}\nabla u.$$

It follows that the characteristic curves associated with equation (129) is given by:

$$\begin{aligned} \frac{du}{d\tau} &= u_t + f'(u)\mathbf{v} \cdot \nabla u = 0, \\ \frac{dt}{d\tau} &= 1, \\ \frac{d\mathbf{x}}{d\tau} &= f'(u)\mathbf{v}. \end{aligned} \tag{130}$$

We shall use the characteristic curves defined above, to reflect the hyperbolic nature of the saturation equation in the time discretization of this equation.

For simplicity we divide $[0, T]$ into N equal parts Δt, such that $T = N \cdot \Delta t$. Further, we shall assume that the total fluid velocity $\mathbf{v}(\mathbf{x})$, may be locally approximated in space by a constant vector. Hence, the solution of (130) between succesive time steps is given by the characteristic problem:

$$\begin{aligned} \bar{\mathbf{x}} &= \mathbf{x} - \Delta t \cdot f'(\bar{u}^{n-1})\mathbf{v}, \\ \bar{u}^{n-1} &= u(\bar{\mathbf{x}}, t^{n-1}). \end{aligned} \tag{131}$$

We shall refer to $\bar{u}^{n-1}$ as the characteristic solution of problem (4)-(6).

The usual procedure of time stepping backwards along the characteristics, is given by the following discretization of the convective derivative:

$$\frac{du}{d\tau} \approx \frac{u^n - \bar{u}^{n-1}}{\Delta t}, \tag{132}$$

where $\bar{u}^{n-1}$ is the characteristic solution defined above.

This scheme has been thoroughly analyzed several places [1,3,10,11] and it has been demonstrated that a time truncation error of order $\epsilon\Delta t$ is obtained. Obviously, in the limit $\epsilon \to 0$, the characteristic solution coincides with the exact solution of problem (4).

However, with the given initial data and a fractional flow function defined by (11), equation (129) develops a nonunique solution. This may be seen directly from (131), since a sufficient condition for uniqueness of this equation is given by:

$$\max \left| f''(u^{n-1}) \nabla u^{n-1} \cdot \mathbf{v} \right| \cdot \Delta t < 1. \tag{133}$$

If no diffusion is added to the hyperbolic part of the saturation equation, $|\nabla u^{n-1}|$ becomes infinitely large, and we have to choose Δt arbitrarily small to obtain a unique solution. Even in the presence of a small diffusion, which gives uniqueness to the complete saturation equation (4), inequality (133) gives a serious constraint on the time step, since $|\nabla u^{n-1}| = O(1/\epsilon)$ in the shock region for a well resolved shock. Thus, the given time discretization is not applicable in the presence of a shock solution.

To resolve this problem without losing the nice properties of the method of characteristics, we shall use the operator-splitting technique introduced by Espedal and Ewing [1]. From classical shock theory we know that in the limit $\epsilon \to 0$, and for a monotonically decreasing initial profile with maximum saturation $u = 1$ and minimum saturation $u = 0$, equation (129) develops a shock with top saturation $u = u_{BL}$ and bottom saturation $u = 0$. Here u_{BL} may be determined geometrically from Figure 21 (a) as the concave envelope of the fractional flow function, or from the equation:

$$\frac{f(u_{BL})}{u_{BL}} = f'(u_{BL}).$$

Further, the physical speed of this shock is given by:

$$\mathbf{v}_{BL} = v_{BL}\mathbf{v}(\mathbf{x}),$$

where

$$v_{BL} = f'(u_{BL}).$$

By definition, a shock solution with top saturation $u = u_{BL}$ and bottom saturation $u = 0$ is a fully established shock, and we assume that the initial profile $u_0(\mathbf{x})$, represents a fully established shock.

To avoid the constraint given by inequality (133), we will divide the fractional flow function into two parts, see [1], such that:

$$f(u) = \bar{f}(u) + b(u)u, \tag{134}$$

where

$$\bar{f}(u) = \begin{cases} f(u) & u_{BL}u \leq 1 \\ v_{BL} \cdot u & 0 \leq u \leq u_{BL}, \end{cases} \tag{135}$$

and

$$b(u) = \begin{cases} 0 & u_{BL}u \leq 1 \\ \dfrac{f(u)}{u} - v_{BL} & 0 \leq u \leq u_{BL}. \end{cases} \tag{136}$$

The convective derivative may then be rewritten in terms of $\bar{f}(u)$ rather than $f(u)$:

$$\frac{d}{d\tau} = \frac{\partial}{\partial t} + \bar{f}'(u)\mathbf{v} \cdot \nabla.$$

The associated characteristic problem is given by:

$$\begin{aligned} \bar{\mathbf{x}} &= \mathbf{x} - \Delta t \cdot \bar{f}'(\bar{u}^{n-1})\mathbf{v}, \\ \bar{u}^{n-1} &= u(\bar{\mathbf{x}}, t^{n-1}). \end{aligned} \tag{137}$$

By these definitions, we shall once more define the time discretization in terms of equation (132).

We notice that the characteristic solution defined by (137) coincides with the solution of equation (4) in the limit $\epsilon \to 0$ for an initial profile representing a fully established shock. Furthermore, the condition given by inequality (133), is not longer necessary since the characteristic solution consists of a rarefaction wave and a part transported along parallel characteristics.

We may object that this construction does not handle the time period when a shock is gradually building up, or more complicated cases when several local shocks are involved. Further, we have treated the velocity field as a constant vector locally in space in the analysis above, which of course is a reasonable assumption when the velocity field is slowly varying in time. However, for long time steps it may be only a crude approximation close to wells and corners where the velocity field changes rapidly.

5.2 Substructuring

By the definitions given in the previous section, we can write equation (4) in the following form:

$$\frac{\partial u}{\partial \tau} + \nabla \cdot (\mathbf{b}(u, \mathbf{x})u - \epsilon \mathbf{D}(u) \cdot \nabla u) = 0, \tag{138}$$

where

$$\mathbf{b}(u, \mathbf{x}) = b(u)\mathbf{v}(\mathbf{x}).$$

Further, by using the time discretization along the approximate characteristics, defined by (132) and (137), we may approximate (138) in each time step by the elliptic equation:

$$u^n + \nabla \cdot (\mathbf{b}(\tilde{u}^n, \mathbf{x})u^n - \epsilon \mathbf{D}(\tilde{u}^n) \cdot \nabla u^n) = \bar{u}^{n-1}, \tag{139}$$

where the approximation of the nonlinear term $\tilde{u}^n$, is yet to be decided. As will be shown later, $\tilde{u}^n = \bar{u}^{n-1}$ represents a good initial approximation.

Because of the small ϵ-term, we associate two space scales with this problem. Except for a thin shock layer, the gradients are small, and the solution is almost completely determined by the characteristic solution. In this outer region we have:

$$\frac{\partial u}{\partial \tau} = O(\epsilon)/h_o^2, \tag{140}$$

where h_o is the typical length scale of the slow spatial variation in the outer region.

An inner shock region is determined by the presence of large gradients. This region is characterized by balance between the transport term $\mathbf{b}(u, \mathbf{x})$ and diffusive forces. We may state this balance formally in terms of the inequality, (see [1]):

$$\left|\frac{\partial u}{\partial \tau}\right| \leq \frac{1}{h_i} \max |\mathbf{b}(u, \mathbf{x}) - \epsilon \mathbf{D}(u) \cdot \nabla u|, \tag{141}$$

where h_i is the typical length scale of the shock region.

By assumption we have:

$$\frac{h_i}{h_o} \ll 1, \tag{142}$$

which motivates the use of a substructuring method to solve equation (139). A composite grid is defined as follows; An outer uniform grid is defined with a mesh size compatible with h_o. We let this coarse grid be independent of time and it is assumed to resolve the slow spatial variation in the outer regions.

We shall further cover the coarse grid blocks containing the shock region with a uniform fine grid with mesh size compatible with h_i. The location of the fine grid will of course change from time step to time step. Later, we discuss how to locate the shock regions; here we may note that most of the information needed comes explicitly from the characteristic solution.

A typical composite grid is depicted in Figure 5. We denote the outer region by Ω_o, the inner region Ω_i and the interface between Ω_o and Ω_i by $\partial\Omega_b$. Hence, Ω is decomposed into three none-overlapping parts such that:

$$\Omega = \Omega_o \cup \Omega_i \cup \partial\Omega_b$$

and

$$\Omega_o \cap \Omega_i \cap \partial\Omega_b = \emptyset.$$

In the following we let h_o and h_i denote the mesh size of the grids covering Ω_o and Ω_i respectively.

We notice that since $h_o = O(1)$, it follows from (140) that the outer solution of equation (139) to lowest order is given by:

$$u^n = \bar{u}^{n-1} + O(\delta \Delta t), \tag{143}$$

where $\delta = \epsilon$.

To obtain a lowest order approximation in the inner region, we observe that h_i may be regarded as a stretching parameter introduced into equation (138). A consistent perturbation expansion of the shock region shows that the shock width is of order ϵ, and moreover, that the transport term, $\mathbf{b}(u, \mathbf{x})$, balances diffusion to the same order:

$$\nabla \cdot (\mathbf{b}(u, \mathbf{x}) - \epsilon \mathbf{D} \cdot \nabla u) = O(\epsilon). \tag{144}$$

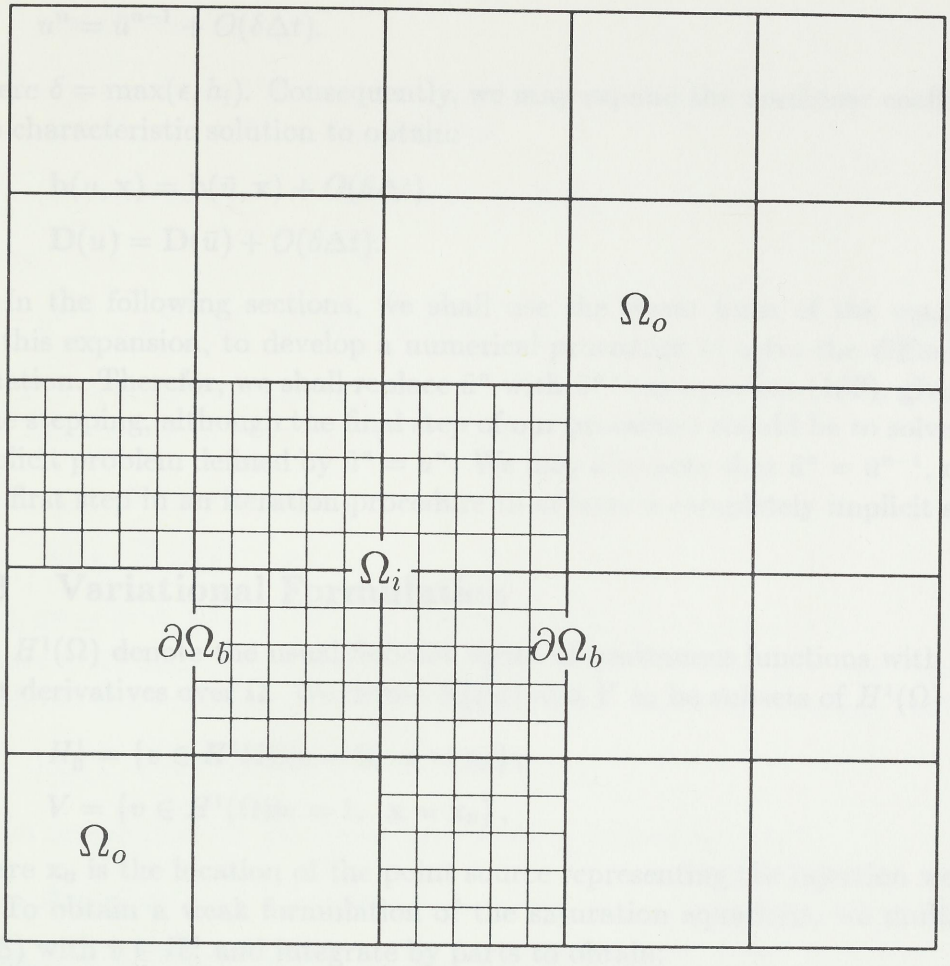


Figure 5: Example of composite grid, splitting Ω into three parts; An outer region Ω_o, an inner region Ω_i and the interface $\partial\Omega_b$, between the outer and inner region.

Thus, to completely resolve the shock, h_i has to be chosen of order ϵ, and expansion (143) is then valid also in the inner region.

If we are not able to resolve the shock region properly, which is the case when $\epsilon \to 0$, the transport term dominates the diffusion term and standard discretization schemes produces unstable numerical solutions. To avoid unstable solutions, we have to add artificial diffusion such that balance between the convection and diffusion term is retained on the given mesh discretization. A numerical scheme that adds the appropriate amount of artificial diffusion will be used in the next section.

Here, we emphasize that the solution of equation (138) in the shock region, is determined to leading order by:

$$\nabla \cdot (\mathbf{b}(u, \mathbf{x})u - \epsilon \mathbf{D}(u) \cdot \nabla u) = 0, \tag{145}$$

Thus, our numerical scheme should be constructed to first approximation in terms of the operator given by this equation.

Furthermore, the effect of balancing the transport term and the diffusion term is to minimize numerical diffusion without losing stability properties. If this is achieved, the solution in the shock region is given to lowest order by

Ω_o
Ω_i
$\partial\Omega_b$
Ω_o

$$u^n = \bar{u}^{n-1} + O(\delta \Delta t). \tag{146}$$

where $\delta = \max(\epsilon, h_i)$. Consequently, we may expand the nonlinear coefficients around the characteristic solution to obtain:

$$\begin{aligned} \mathbf{b}(u, \mathbf{x}) &= \mathbf{b}(\bar{u}, \mathbf{x}) + O(\delta \Delta t), \\ \mathbf{D}(u) &= \mathbf{D}(\bar{u}) + O(\delta \Delta t). \end{aligned} \tag{147}$$

In the following sections, we shall use the linear form of the equations defined by this expansion, to develop a numerical procedure to solve the diffusion correction equation. Therefor, we shall replace $\tilde{u}^n$ with $\bar{u}^{n-1}$ in equation (139), giving an explicit time stepping, although the final step of our procedure should be to solve a completely implicit problem defined by $\tilde{u}^n = u^n$. We may also note that $\tilde{u}^n = \bar{u}^{n-1}$, may represent the first step in an iteration procedure to achieve a completely implicit scheme.

5.3 Variational Formulation

Let $H^1(\Omega)$ denote the usual Sobolev space of continuous functions with L_2-integrable first derivatives over Ω. We define $H_0^1(\Omega)$ and V to be subsets of $H^1(\Omega)$ given by:

$$\begin{aligned} H_0^1 &= \{v \in H^1(\Omega) | v = 0, \ \ \mathbf{x} = \mathbf{x}_0\}, \\ V &= \{v \in H^1(\Omega) | v = 1, \ \ \mathbf{x} = \mathbf{x}_0\}, \end{aligned}$$

where $\mathbf{x}_0$ is the location of the point source representing the injection well.

To obtain a weak formulation of the saturation equations, we multiply equation (138) with $v \in H_0^1$ and integrate by parts to obtain:

$$(\frac{\partial u}{\partial \tau}, v) + B(u, v) = \int_{\partial\Omega} \epsilon \mathbf{D}(u) \cdot \nabla u \cdot \mathbf{n} d\gamma,$$

where

$$B(u, v) = (\nabla \cdot (\mathbf{b}(u, \mathbf{x})u), v) + \epsilon(\mathbf{D}(u) \cdot \nabla u, \nabla v) \tag{148}$$

and

$$(u, v) \stackrel{\text{def}}{=} \int_\Omega uv d\omega.$$

The right hand side of this integral equation is determined by using the boundary conditions (3), (6), (16) and (17). Away from the wells, $g_2(\mathbf{x}, t)$ and $g_3(\mathbf{x}, t)$ is zero and the line integral obviously vanishes except for the wells. The well regions are approximated by point sinks and sources represented by Dirac delta functions. However, the line integral also vanishes at these points since $\nabla u(\mathbf{x}_1, t) \cdot \mathbf{n} = 0$ by (17) and $v(\mathbf{x}_0) = 0$ by definition of H_0^1.

Thus, equation (138) can equivalently be written:

$$(\frac{\partial u}{\partial \tau}, v) + B(u, v) = 0 \ \ \forall v \in H_0^1, \ \ t \in [0, T] \tag{149}$$

and

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}).$$

Similarly, we obtain the weak formulation of equation (139):
Find $u^n \in V$, $n = 1, 2, \ldots, N$, such that

$$(u^n, v) + \Delta t B(u^n, v) = (\bar{u}^{n-1}, v) \ \ \forall v \in H_0^1, \tag{150}$$

where

$$\bar{u}^{n-1} = u^{n-1}(\bar{\mathbf{x}}), \ \ u^0(\mathbf{x}) = u_0(\mathbf{x}),$$

and $\bar{\mathbf{x}}$ is defined by equation (137).

Consistent with equation (147) we define:

$$\begin{aligned} \mathbf{b}(\mathbf{x}) &\stackrel{\text{def}}{=} \mathbf{b}(\bar{u}^{n-1}, \mathbf{x}), \\ \mathbf{D}(\mathbf{x}) &\stackrel{\text{def}}{=} \mathbf{D}(\bar{u}^{n-1}, \mathbf{x}). \end{aligned} \tag{151}$$

With $\tilde{u}^n = \bar{u}^{n-1}$, $\mathbf{b}(\mathbf{x}) \in [H^1(\Omega), H^1(\Omega)]$, and the components of $\mathbf{D}(\mathbf{x})$ in $C^0(\bar{\Omega})$, $B(u, v)$ defines a bilinear continuous form on $H_0^1 \times H_0^1$. Unfortunatly, the sign of $B(v, v)$ is indefinite due to the asymmetric transport term. However, relation (144) implies that:

$$|\Delta t(\nabla \cdot (\mathbf{b}v), v)| < \{(v, v) + \epsilon \Delta t(\mathbf{D}v, v)\},$$

which means that the complete bilinear form defined by (150):

$$A(u, v) = (u, v) + \Delta t B(u, v), \tag{152}$$

is coercive on $H^1(\Omega)$. Hence, we may deduce existence and uniqueness of u^n, $n = 1, 2, \ldots, N$, satisfying the weak formulation (150), from the Lax-Milgram theorem.

A Petrov-Galerkin formulation will be used to obtain a discrete approximation of the weak formulation (150) in each time step. Let $\{\mathbf{x}_{ij}\}$ be the nodes generating the rectangular mesh covering Ω as shown in Figure 5. We define a trial space S^h and a test space T^h to be discrete subspaces of H^1 spanned by $\theta_{ij}(\mathbf{x})$ and $\psi_{ij}(\mathbf{x})$, the trial and test functions respectively. Further, we shall need the subsets:

$$S_0^h = S^h \cap H_0^1, \ \ T_0^h = T^h \cap H_0^1$$

and

$$S_V^h = S^h \cap V.$$

The Petrov-Galerkin finite element formulation of equation (150) is then:
Find $u_h^n \in S_V^h$, $n = 1, \ldots, N$, such that

$$A(u_h^n, \psi) = (\bar{u}_h^{n-1}, \psi) \ \ \forall \psi \in T_0^h \tag{153}$$

and

$$\bar{u}_h^{n-1} = u_h^{n-1}(\bar{\mathbf{x}}), \ \ u_h^0(\mathbf{x}) = u_0(\mathbf{x}). \tag{154}$$

It is well known that using $T^h = S^h$ (as in the usual Galerkin formulation) is a bad choice of test space when $\epsilon \to 0$ and the transport term $\mathbf{b}(\mathbf{x})$ dominates the diffusion term. This appears as unphysical oscillations in the numerical solution in the presence of a shock. It may also be demonstrated that this problem is caused by the dominating, asymmetric transport term in the discretized bilinear form $B(u,v)$, i.e., the leading part of $A(u,v)$ in the shock region.

To handle such problems, Barrett and Morten [6], have developed a symmetrization technique to obtain optimal approximation properties in selected norms. From a practical point of view, this technique may be regarded as a method to add the appropriate amount of artificial diffusion or upstream weighting, to stabilize the numerical solution.

The symmetrization technique used in one space dimension is in principle easily extendible to higher dimensions, but the extension may be technically involved and produces test functions that is difficult to use in practical computations. A procedure that resolves this problem has been developed by Demkowicz and Oden, who introduce the concept of "numerical optimal" test functions [26,27].

Here, however, we shall use a straight forward extension of the test functions already developed; We choose bilinear elements spanned by the trial functions $\theta_{ij}(\mathbf{x}) = \theta_i(x)\theta_j(y)$ where $\theta_i(\cdot)$ is defined by equation (41). Consistent with this we choose the test functions

$$\psi_{ij}(\mathbf{x}) = [\theta_i(x) + c_1^I \sigma_i(x)] \cdot [\theta_j(y) + c_2^I \sigma_j(y)], \tag{155}$$

where $\sigma_i(\cdot)$ is given by equation (54) and c_k^I, $k = 1,2$ is determined from the components of $\mathbf{b} = [b_1, b_2]$ and $\mathbf{D} = D_{11}\mathbf{ii} + D_{22}\mathbf{jj}$ such that:

$$c_k^I = 3\left(\frac{2}{\beta_k^I} - \coth\left(\frac{\beta_k^I a}{2}\right)\right), \quad \beta_k^I = \frac{h b_k^I}{\epsilon D_{kk}^I}, \quad k = 1,2, \tag{156}$$

where $()^I$ denotes some sort of average over element I.

We observe that these test functions may be severely skewed in the shock region. The skewedness depends of course on the direction of the flow field through the components of $\mathbf{b} = b(u)\mathbf{v}(\mathbf{x})$. Away from the shock, $\mathbf{b}(\mathbf{x}) = 0$, and the trial and test functions coincide. We emphasize that the test functions (155), are constructed to stabilize the solution around sharp shocks, where the solution is mainly determined by $B(u,v)$. Away from the shock region, where the asymmetric transport term is zero, it may be necessary to construct optimal test functions with respect to the complete symmetric bilinear form $A(u,v)$, given by equation (152). Demkowicz and Oden [26,27], have constructed optimal test functions for such problems. Although we will not pursue the problem further here, we note that their test functions may be convenient to use away from the shock region,

With the test functions (155) defined, we have to construct a composite discrete operator from the bilinear form $A(u_h, \psi)$ on $\Omega = \Omega_o \cup \Omega_i \cup \partial\Omega_b$. We decompose the discrete solution into three parts, $u_h = (u_h^o, u_h^b, u_h^i)^T$, such that u_h^o represents the outer solution defined on Ω_o, u_h^b the solution on the interface $\partial\Omega_b$ and u_h^i represents the solution defined on the inner region Ω_i.

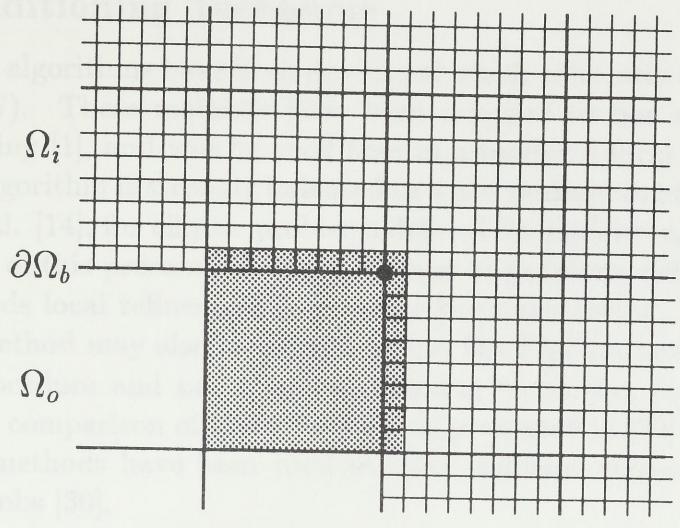


Figure 6: Support of basis and test functions associated with coarse grid nodes on $\partial\Omega_b$.

Our problem now is to patch the inner and outer solutions together at the interface $\partial\Omega_b$. To perform this, we first determine the nodal values at the fine grid nodes on $\partial\Omega_b$ by linear interpolation between successive coarse grid nodes on $\partial\Omega_b$. Thus, these nodes are regarded as "slave" nodes. Second, we decompose the trial functions into three sets $\{\theta^o_{ij}(\mathbf{x})\}$, $\{\theta^b_{ij}(\mathbf{x})\}$ and $\{\theta^i_{ij}(\mathbf{x})\}$, associated with the outer region, interface, and inner region respectively. The outer and inner trial functions are defined straight forward as products of the one-dimensional trial functions with support on adjacent mesh cells. The coarse grid trial functions with vertices on $\partial\Omega_b$, defining the set $\{\theta^b_{ij}(\mathbf{x})\}$, are similarly defined, but with a support confined to at most one fine grid cell into Ω_i, as shown in Figure 6.

In analogy with this, we split the test functions into three sets such that $\{\psi^o_{ij}(\mathbf{x})\}$ are test functions defined on Ω_o, $\{\psi^b_{ij}(\mathbf{x})\}$ are coarse grid test functions with vertices on $\partial\Omega_b$ and $\{\psi^i_{ij}(\mathbf{x})\}$ are defined on Ω_i.

Hence, we may write the composite discrete operator in matrix form:

$$\begin{pmatrix} A_{oo} & A_{ob} & 0 \\ A_{bo} & A_{bb} & A_{bi} \\ 0 & A_{ib} & A_{ii} \end{pmatrix} \begin{pmatrix} u^o_h \\ u^b_h \\ u^i_h \end{pmatrix} = \mathbf{d}^T, \tag{157}$$

where $\mathbf{d}^T$ is the right hand side of equation (150).

We observe that by definition of θ^b_{ij} and ψ^b_{ij}, the coupling between the inner and outer solution is restricted to a narrow strip along $\partial\Omega_b$. This is of course reasonable since the test functions involve the local parameter c^I_k and are defined relative the local behaviour in the shock region. Further, this imply that the inner coarse grid nodes are indistinguishable from the fine grid nodes. However, in the solution procedure these nodes may be of special importance.

Ω_f
$\partial\Omega_\delta$
Ω_c

5.4 Preconditioning Technique

Recently, several algorithms have been developed which efficiently solve linear systems of the form (157). These methods have been adapted to our specific problem by Espedal and Ewing [1], and will be used here in a modified form to solve (157). We note that their algorithm is strongly indebted to a preconditioning technique developed by Bramble et. al. [14], for elliptic problems defined on conforming grids.

An extension of this preconditioning technique to problems defining localized phenomena that needs local refinement is found in Bramble, Ewing, Pasciak and Schatz [15]. This last method may also be viewed within the FAC framework, although FAC is a solution procedure and not a preconditioning technique, (see McCormick and Thomas [28]). A comparison of these methods is presented in [29]. We may also note that multi-grid methods have been used to solve diffusion convection problems, see Schmidt and Jacobs [30].

Two strategies may be employed to solve equation (157). If the characteristic solution by inequality (140) is assumed to determine the outer solution within acceptable accuracy, then problem (157) reduces to determine A_{ii}^{-1}. In this case we may use the preconditioner defined by Bramble et. al. [14]. If we have to perform diffusion correction on the complete domain Ω, then we should use a scheme similar to the one developed by Bramble, Ewing, Pasciak and Schatz [15]. In both cases the main difficulty is to compute the inner solution, i.e. to determine the inverse of A_{ii}.

We will construct a solution within a conjugate gradient iteration scheme. A preconditioner $B(\cdot,\cdot)$, will be constructed in terms of the bilinear form $A(\cdot,\cdot)$, based on the results from [1,14,15]. (The preconditioner $B(\cdot,\cdot)$, should not be confused with the bilinear form defined by (148)).

For simplicity we shall follow the first strategy and let the outer solution u_o^h, be completely determined by the characteristic solution. It is then sufficient to solve the restriction of boundary value problem (153), to the inner region Ω_i: Find $u_i^h \in S_V^h(\Omega_i)$ such that

$$A(u_i^h, \psi) = (\bar{u}, \psi), \quad \forall \psi \in T_0^h(\Omega_i), \tag{158}$$

where

$$S_V^h(\Omega_i) = \{u^h \in S^h(\Omega_i) | u^h = \bar{u}, \quad \mathbf{x} \in \partial\Omega_b\}.$$

Since the bilinear form $A(u,v)$ is not symmetric and $S_0^h(\Omega_i) \neq T_0^h(\Omega_i)$, the construction of a preconditioner does not follow directly from the algorithm given by Bramble et. al. [14]. However, Espedal and Ewing [1] have shown that if $T_0^h(\Omega_i)$ is an optimal test space with respect to a coercive, symmetric and bilinear form $A^*(\cdot,\cdot)$, then the techniques given in [14] are applicable to our problem. Here we shall assume that the test space T_0^h is spanned by optimal or approximate optimal test functions with respect to a given $A^*(\cdot,\cdot)$, and proceed by constructing the preconditioner directly in terms of $S^h(\Omega_i)$, $T^h(\Omega_i)$ and $A(\cdot,\cdot)$.

A subdomain decomposition of Ω_i is obviously defined by the internal coarse grid nodes on Ω_i, which divide Ω_i into K separate square blocks, say (see Figure 7):

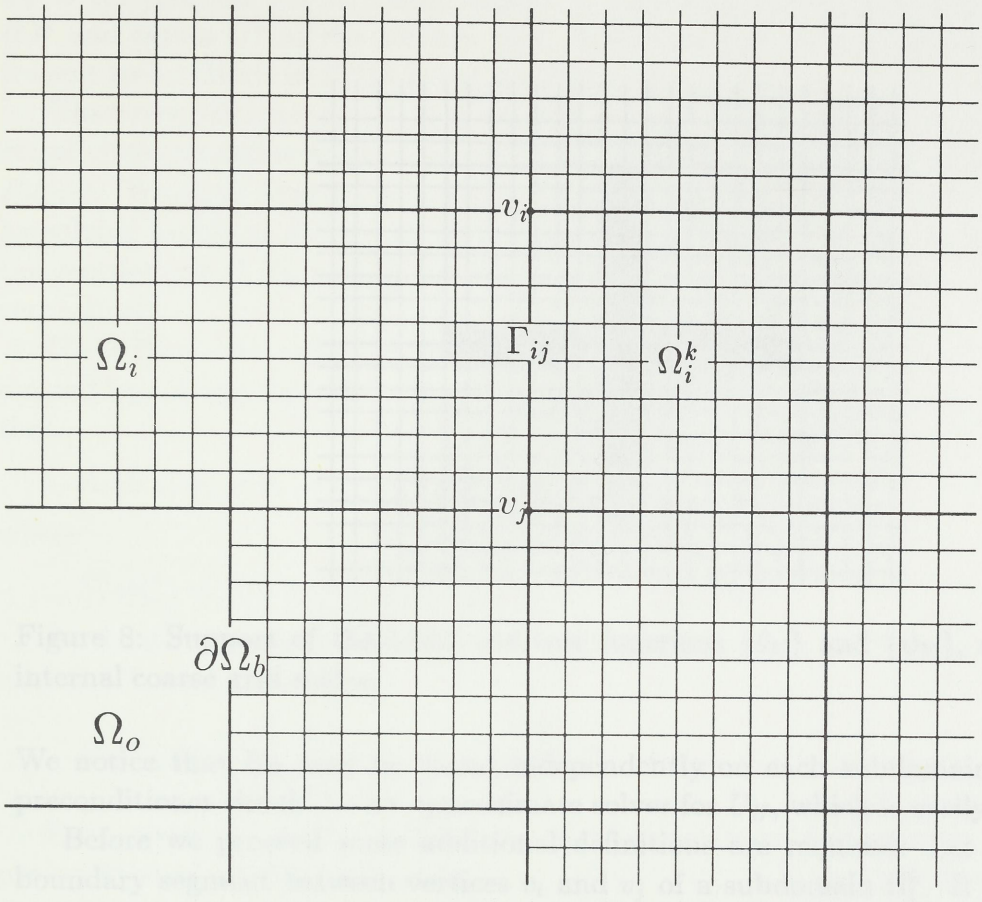


Figure 7: Example of subdomain decomposition of the inner region Ω_i; Ω_i^k denotes subdomain k, Γ_{ij} denotes an internal boundary segment and v_i is an internal coarse grid node.

$$\Omega_i = \bigcup_{k=1}^{K} \Omega_i^k.$$

We shall denote the internal coarse grid nodes or the vertices of each subdomain Ω_i^k, by the set $\{v_j\}$.

Define $\tilde{A}_k(u,v)$ to be the restriction of $A(u,v)$ to Ω_k and decompose functions in $S^h(\Omega_i)$ into two parts:

$$U^h = U_H + U_P, \tag{159}$$

where $U_P \in S_0^h(\Omega_i^1) \oplus \cdots \oplus S_0^h(\Omega_i^K)$ and satisfies

$$\tilde{A}_k(U_P, \psi) = \tilde{A}_k(U^h, \psi), \quad \forall \psi \in T_0^h(\Omega_i^k), \quad k = 1, \ldots, K. \tag{160}$$

It follows that U_H coincide with U^h on $\partial\Omega_i^k$ and therefor satisfies homogeneous Dirichlet problems on each subdomain:

$$\tilde{A}_k(U_H, \psi) = 0, \quad \forall \psi \in T_0^h(\Omega_i^k), \quad k = 1, \ldots, K. \tag{161}$$

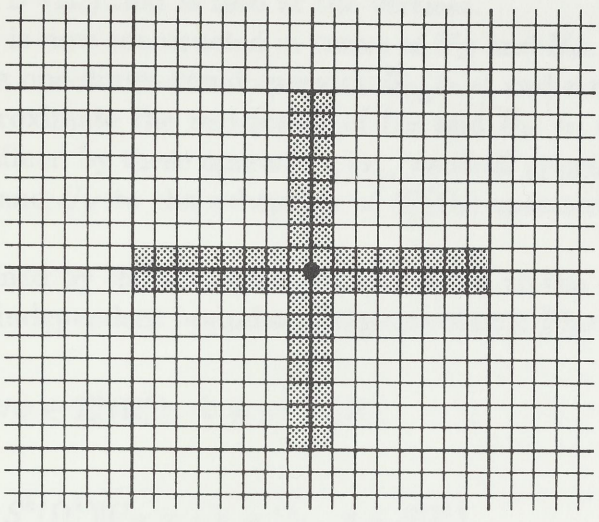


Figure 8: Support of the basis and test functions $\{\theta_V\}$ and $\{\psi_V\}$, associated with internal coarse grid nodes.

We notice that U_P may be found independently on each subdomain Ω_i^k, thus, the preconditioner should be an approximate solver for U_H, which is easily computed.

Before we proceed some additional definitions are required: Let Γ_{ij} denote the boundary segment between vertices v_i and v_j of a subdomain Ω_i^k. It is then natural to decompose basis functions spanning $S_0^h(\Omega_i)$ into three sets, $\{\theta_P\}$, $\{\theta_E\}$, and $\{\theta_V\}$ respectively, such that $\theta_P \in S_0^h(\Omega_i^k)$, $\theta_E \in S_0^h(\Gamma_{ij})$, and $\{\theta_V\}$ are basis functions with nodes at the vertices of each subdomain. $S_0^h(\Gamma_{ij})$ is a subspace of functions in $S_0^h(\Omega_i)$ with nodes on Γ_{ij}, i.e., functions which vanishes on the interior and at the vertices of each Ω_i^k.

We let θ_P and θ_E be straight forward extensions of the one-dimensional basis functions (41), on Ω_i, whereas θ_V is a "coarse grid" basis function with support localized to the internal boundary segments Γ_{ij} and with nodes at the vertices $\{v_j\}$. It follows that θ_V vanishes at the interior of each Ω_i^k, as shown on Figure 8.

In analogy with this we define the test functions $\{\psi_P\}$, $\{\psi_E\}$, and $\{\psi_V\}$ where ψ_P and ψ_E is straight forward defined by equation (155). However, it is not obvious how to choose ψ_V associated with θ_V, one possible implementation is given in section 6.2. We note that

$$S_0^h(\Omega_i) = span\{\theta_P\} \bigcup span\{\theta_E\} \bigcup span\{\theta_V\},$$

whereas in general

$$T_0^h(\Omega_i) \neq \left[\tilde{T}_0^h(\Omega_i) = span\{\psi_P\} \bigcup span\{\psi_E\} \bigcup span\{\psi_V\}\right].$$

To determine U_H we decompose U_H further into two parts such that

$$U_H = U_E + U_V. \tag{162}$$

U_V is the solution of (161) with respect to functions in $S^h(\Omega_i)$ that is linear on each $\partial\Omega_i^k$ and equals U^h at the vertices $\{v_j\}$. It follows that U_E should satisfy (161) with respect to functions in $S^h(\Omega_i)$ that is zero at the vertices.

The preconditioner is now constructed in terms of U_E and U_V. Following Bramble et. al. [14], we define a one-dimensional operator $B_{\Gamma_{ij}}(\cdot,\cdot)$ and a coarse grid operator $B_c(\cdot,\cdot)$. We shall approximate the restriction of U_E and U_V to $\partial\Omega_i^k$, $k = 1,\ldots,K$, by solving problems defined by these operators, i.e., we shall approximately determine the restrictions of U_E and U_V to the subspaces of $S_0^h(\Omega_i)$ spanned by $\{\theta_E\}$ and $\{\theta_V\}$ respectively.

U_H is then determined by the extension of U_E and U_V to the interior of each Ω_i^k, defined by solving the independent boundary value problems: Find $U_H \in S_V^h(\Omega_i^k)$ such that

$$\tilde{A}_k(U_H,\psi) = 0, \quad \forall\psi \in T_0^h(\Omega_i^k), \quad k = 1,\ldots,K, \tag{163}$$

where

$$S_V^h(\Omega_i^k) = \{U_H \in S^h(\Omega_i^k) | U_H = U_E + U_V, \quad \mathbf{x} \in \partial\Omega_i^k\}.$$

By the definitions given above, we shall formally write the preconditioner associated with $A(U^h,\psi)$:

$$B(U^h,\psi) = \sum_{k=1}^{K} \tilde{A}_k(U_P,\psi_P) + \sum_{\Gamma_{ij}} B_{\Gamma_{ij}}(U_E,\psi_E) + B_c(U_V,\psi_V). \tag{164}$$

For a given g, e.g. the residual obtained in each step of the iteration scheme, the preconditioned solution defined by

$$B(U^h,\psi) = (g,\psi), \quad \forall\psi \in \tilde{T}_0^h(\Omega_i),$$

is determined by the following algorithm, (see [14,15]):

1. Find a particular solution $U_P \in S_0^h(\Omega_i^k)$, on each subdomain Ω_i^k by solving:

$$\tilde{A}_k(U_P,\psi_P) = (g,\psi_P), \quad \forall\psi_P \in T_0^h(\Omega_i^k). \tag{165}$$

2. Find the restriction $U_E \in S_0^h(\Gamma_{ij})$ such that

$$B_{\Gamma_{ij}}(U_E,\psi_E) = (g,\psi_E) - \tilde{A}_k(U_P,\psi_E), \quad \forall\psi_E \in T_0^h(\Gamma_{ij}). \tag{166}$$

3. Determine U_V on the vertices $\{v_j\}$ by finding the restriction $U_V \in span\{\theta_V\}$ that satisfies

$$B_c(U_V,\psi_V) = (g,\psi_V) - \tilde{A}_k(U_P,\psi_V), \quad \forall\psi_V. \tag{167}$$

4. Extend U_E and U_V to the interior of each subdomain by solving equation (163), i.e., determine U_H.

5. The preconditioned solution is then given by:

$$U^h = U_H + U_P$$

We note that the solution of U_P on each subdomain may be computed in parallel. Similarly, step 2 and step 4 may be highly parallelized.

6 Implementation of the Numerical Methods

In this section we will give an overview of the practical implementation of the present one and two dimensional numerical codes written in FORTRAN 77. The codes have been implemented on an HP-9000/318 work station and an Alliant FX/8 parallel machine.

6.1 One-dimensional Code

6.1.1 Overview, Data-structure

The one-dimensional code consists of three main parts following naturally from the methods outlined previously, i.e. a set of routines solving the characteristic part, a set of routines to identify the shock regions, and finally a set of routines to perform the diffusion correction.

The structure of the code is shown on the flow charts Figure 9-11, and may shortly be described as follows; Input parameters are read from the file `data.dat` and the initial profile is computed by linear interpolation between the points given in the input file `iprofile.dat`. The mesh parameters are initialized by the routine `setpar` and an initial time step is chosen by the routine `setdt` as follows; A time step (*cfldt*), equivalent to travelling one coarse grid block with the largest physical velocity, i.e. the shock velocity, is computed. The user then gives a fraction of *cfldt* to determine the time step used in the code by the formula:

$$dt = frac \cdot cfldt. \tag{168}$$

This time step may of course be changed during a simulation by giving a new fraction of *cfldt*. We may also note that *frac* may easily be given such that the numerical diffusion due to interpolation is negligible in the shock region, i.e. $k \approx 0$ or $k \approx 1$ in the estimate (123).

The main part of the code is the time loop governed by the routine `mocdif`. This routine builds a solution in each time step, beginning with the solution on the coarse grid. The coarse grid solution is taken to be the characteristic solution determined by the routines `cgrid` and the solver of the characteristic equation `mthmoc`. We note that the only input required by the characteristic solver is the coordinate of the node which is to be updated and the saturation profile from the previous time step.

The coarse grid solution is passed to the routine `fpos`, and coarse grid blocks which have to be refined are identified. The refinement is performed by the routine `fgrid` which determines the characteristic solution on the refined nodes by calling the characteristic solver `mthmoc`. Using the coarse grid nodes as boundary values, the final solution in the current time step is obtained by adding diffusion to the fine grid characteristic solution. The elliptic solver performing the diffusion correction is governed by the routine `diesol`.

Each time step is concluded by calling the routine `count` which stores the computed saturation profile and updates the time loop. This routine also checks for an output time and computes the mass balance.

One of the important features of the method is the simplicity of the adaptive space discretization. We have chosen a fixed, uniform grid to represent the slow variation away from the front region. This grid is refined in each time step, by adding uniform subgrids on selected coarse grid blocks, which should resolve the sharp variations around the shock front.

A natural way to organize the variables, which allows one to easily generate the subgrids, is a "nested" data structure. The storage required are three long arrays denoted $x(0{:}nn_{\max})$, $s(0{:}nn_{\max})$, and $next(0{:}nn_{\max})$ containing respectively the coordinate values of the nodes, the solution at the nodes and a pointer to the index of the next node on the grid, ordered by increasing coordinate values.

Thus, assume that the present number of nodes equal nn, and that the indices i_1 and i_2 represent adjacent nodes such that:

$$x(i_1) < x(i_2) \quad \text{and}$$
$$next(i_1) = i_2.$$

Then, if a new node has to be added between i_1 and i_2, the following algorithm is used to update the data structure:

$i_3 = nn;$
$x(i_3) =$ 'coordinate of the new node';
$next(i_3) = next(i_1);$
$next(i_1) = i_3;$
$nn = nn + 1;$

Note that the index of the new node is nn since the arrays are counted from 0 to $nn-1$, nn being the number of nodes before i_3 is added.

In addition to the arrays described above, an array containing indices of the left nodes of refined coarse grid blocks, named *index*, are required by the code to easily identify these blocks.

Since the information computed will be needed in the next time step we have used two additional arrays, *ss* and *sx*, to store s and x respectively. These arrays are updated at the end of each time iteration by the routine **save**, and ordered with increasing coordinate values. Thus, the storage required consists basically of five equally long arrays.

We note that the static memory allocation used by FORTRAN 77 is very inconvenient for dynamical problems solved by adaptive methods, since we do not know in advance the length of the arrays required in each time step. A new FORTRAN version with dynamical memory allocation will therefor allow for a much more efficient code for such problems.

6.1.2 Characteristic Solver

The solution of the nonlinear characteristic equation at time $t = t^n$:

$$\begin{aligned} \bar{x} &= x - \Delta t \cdot \bar{f}'(\bar{u}^{n-1}), \\ \bar{u}^{n-1} &= u^{n-1}(\bar{x}), \end{aligned} \tag{169}$$

is determined by the routine `mthmoc(x,cs)`, where `cs` is the saturation value computed at the node `x` by the routine. The computational procedure is based on two basic assumptions:

(i). The solution is monotone and decreasing in space for all $t \in [0, T]$.

(ii). The derivative of the modified fractional flow function is divided into two parts:

$$\bar{f}'(u) = f'(u), \quad u_{BL} < u \leq 1,$$
$$\bar{f}'(u) = v_{BL}, \quad 0 \leq u \leq u_{BL},$$

such that the upper part $u_{BL} < u \leq 1$, represents a rarefaction wave, and the lower part $0 \leq u \leq u_{BL}$, represents a pure transport of the saturation profile along parallel characteristics..

Since the characteristic solution by **(i)** and **(ii)** consists of a rarefaction part and a part transported along parallel characteristics, the solution to (169) is uniquely defined in each time step.

We have used the assumptions given above to split the solution procedure into three separate cases depending on whether the characteristic curve is entirely behind the front region, crosses the front region or is entirely ahead of the front region. The front region in this context refers to the position of u_{BL} in the previous time step.

If the node to be updated is given by the coordinate x_i, the solution procedure in the different cases is given as follows;

Case I: If $u^{n-1}(x_i) \geq u_{BL}$, the characteristic curve through x_i is entirely behind the shock front and is a part of the rarefaction wave. The characteristic equation is solved by functional iteration using the following algorithm:

$\bar{u}_0^{n-1} = u^{n-1}(x_i)$;
for $j = 1, 2, \ldots$:
  $\bar{x}_i^j = x_i - \Delta t \cdot \bar{f}'(\bar{u}_{j-1}^{n-1})$;
  $\bar{u}_j^{n-1} = u^{n-1}(\bar{x}_i^j)$;
  if $\left|\bar{u}_j^{n-1} - \bar{u}_{j-1}^{n-1}\right| < \mathrm{TOL}$ then :
    $\bar{u}^{n-1}(x_i) = \bar{u}_j^{n-1}$;
    return;
  endif;
continue;

We note that this recursion shortly can be written:

$$\bar{u}_j^{n-1} = u^{n-1}(\bar{x}(\bar{u}_{j-1}^{n-1})), \quad j = 1, 2, \ldots \tag{170}$$

Thus if

$$\bar{u}^{n-1}(u_a) \stackrel{\text{def}}{=} u^{n-1}(\bar{x}(u_a)) \text{ and } \bar{x}(u_a) \stackrel{\text{def}}{=} x - \Delta t \cdot \bar{f}'(u_a),$$

then the general convergence condition for the recursion may be stated by the Lipschitz condition:

$$\forall u_a, u_b \in L, \exists k < 1 : |\bar{u}^{n-1}(u_a) - \bar{u}^{n-1}(u_b)| \leq k|u_a - u_b|, \tag{171}$$

where $L = [u_{BL}, 1]$.

If we assume that u^{n-1} and $f'(u)$ are continuously differentiable functions of x and u respectively, it follows from the mean value theorem that

$$\exists \tilde{u} \in [u_b, u_a] : \bar{u}^{n-1}(u_a) - \bar{u}^{n-1}(u_b) = -\Delta t \frac{du^{n-1}}{dx} \frac{d^2 f}{du^2}\bigg|_{u=\tilde{u}} (u_a - u_b),$$

implying that

$$k = \max_{u \in [u_{BL}, 1]} \Delta t \left| \frac{du^{n-1}}{dx} \frac{d^2 f}{du} \right|.$$

Hence, we obtain the convergence criterion:

$$\max_{u \in [u_{BL}, 1]} \left| \frac{du^{n-1}}{dx} \frac{d^2 f}{du^2} \right| < 1/\Delta t. \tag{172}$$

Although the characteristics are uniquely defined in the rarefaction part of the solution, this criterion may be compared with the sufficient condition for uniqueness of the characteristic solution given by inequality (67).

Case II: We define x_{BL} by the relation:

$$x_{BL} = x_i - \Delta t \cdot v_{BL}.$$

If $u^{n-1}(x_i) < u_{BL}$ and $u^{n-1}(x_{BL}) > u_{BL}$, the characteristic curve crosses the front region. In this case numerical experience has shown that a robust method is necessary to obtain a close approximation to the solution of the characteristic equation.

Let the solution of the characteristic equation $\bar{x}_i$, be given by the zero of the function $F(\bar{x})$, defined by:

$$F(\bar{x}) = x_i - \Delta t \cdot \bar{f}'(u^{n-1}(\bar{x})) - \bar{x}. \tag{173}$$

Since $f'(u) < v_{BL}$ when $u > u_{BL}$ by assumption **(ii)**, $F(\bar{x})$ changes sign in the interval $x_{BL} \leq \bar{x} \leq x_i$:

$$F(\bar{x}) = \begin{cases} -\Delta t \cdot v_{BL} < 0 \text{ if } \bar{x} = x_i \\ \Delta t \cdot (v_{BL} - f'(u^{n-1}(x_{BL})) > 0 \text{ if } \bar{x} = x_{BL}, \end{cases}$$

thus, as we expect, the zero of $F(\bar{x})$ is located to this interval. Exploiting the sign change, a robust method to determine the zero of $F(\bar{x})$ is the method of bisection, leading to the following algorithm:

$a_0 = x_{BL}$;
$b_0 = x_i$;

for $i = 1, 2, \ldots$:

$m_i = \frac{1}{2}(a_{i-1} + b_{i-1})$;

$F(m_i) = x_i - \Delta t \cdot \bar{f}'(u^{n-1}(m_i)) - m_i$;

$$(a_i, b_i) = \begin{cases} (m_i, b_{i-1}) \text{ if } F(m_i) < 0 \\ (a_{i-1}, m_i) \text{ if } F(m_i) > 0 \end{cases} ;$$

$lu = u^{n-1}(a_i)$;

$ru = u^{n-1}(b_i)$;

if $|ru - lu| < \text{TOL}$ then :

$\bar{u}^{n-1}(x_i) = \frac{1}{2}(lu + ru)$;

return;

endif;

continue;

The method of bisection always converges as long as $F(\bar{x})$ is continuous and changes sign on the initial interval. However, it is well known that the rate of convergence of this method is very slow. The usual way of increasing the convergence rate, i.e. by letting the method of bisection produce a good initial approximation for a more sophisticated method, has not yet been implemented.

Case III: If $u^{n-1}(x_{BL}) \leq u_{BL}$, the characteristic curve through x_i is entierly ahead of the front region. Since $\bar{f}'(u)$ is constant in this region the solution is simply given by:

$$\bar{u}^{n-1}(x_i) = u^{n-1}(x_{BL}),$$

which is exactly the transport of the profile with the shock velocity ahead of the front region.

To perform the computations in the three cases described above, `mthmoc` has been supplied with the two external routines `dfflow` and `find`. These routines computes respectively the derivative of the fractional flow function and u^{n-1} for a given x-value. In the present code, all the external function routines are given by explicit analytic expressions, which is also the case when computing the derivative of the modified fractional flow function.

To compute u^{n-1} for an arbitrary x-value, we use linear interpolation in the arrays *sx* and *ss*. Since these arrays are ordered by increasing coordinate value, we may efficiently identify the nodes to interpolate between by use of the bisection method; Starting with the entire domain as initial interval we recursively half each interval bounded by a pair of nodes and containing the x-value until we find the smallest interval containing x. This interval eventually defines the adjacent nodes to interpolate between to give the desired saturation value $u^{n-1}(x)$.

6.1.3 Adaptive Grid

The space discretization of the equation is made adaptive by locating front regions with sharp variations and then adding denser subgrids to these regions, to hopefully resolve the sharp variations.

The main problem involved in a refinement procedure is to locate the shock regions, since it is a simple task to add new nodes once the location of the nodes are known, as shown before. In the present code the shock regions are identified by the routine `fpos`.

Two approaches to this problem have been implemented, although only one is used currently. The first approach, is to use local *a priori* error estimates constructed from the associated elliptic operator [1,26,27]. Such estimates enables one to compute local error estimates on each coarse grid block from the characteristic solution. If the error exceeds a given tolerance the block has to be refined, else the characteristic coarse grid solution is assumed to approximate the actual solution.

Our experience with this approach is that the shock region is identified, but the computations necessary to obtain the error estimates are time consuming and quite involved, and the robustness of the procedure may be discussed.

Lately, local *a posteriori* estimates have been developed that have been shown to be very useful for computational purposes. Such error estimates are developed several places (see [31,32,33]) and may be adapted to our problem.

Due to the hyperbolic nature of the given problem, however, it seems wasteful to drive the local refinement from elliptic error estimators (see [34]). Therefor, in its present form, the numerical code uses the characteristic solution and the following simple assumption to identify the shock region:

(iii). One established shock exists, located to a few coarse grid blocks, with top saturation equal to u_{BL}, and bottom saturation equal to zero.

Assume that the coarse grid blocks are numbered from 1 to n and the nodes are numbered from 0 to n, further, let u_i equal the characteristic solution computed at node 'i'. By assumption **(iii)**, we are then lead to the following algorithm which defines the shock region:

for $i = 1, 2, \ldots, n$:
    if $u_{i-1} \geq u_{BL}$ and $u_i < u_{BL}$ then :
        refine block 'i';
        end loop;
    endif;
continue;

for $j = i, \ldots, n-1$:
    if $u_j \leq \mathrm{TOL}$ then :
        return;
    else :
        refine block 'j';
    endif;
continue;

if $u_n > \mathrm{TOL}$ then :
    the front has reached the out end,
    terminate the execution;
endif;

To summarize this quasi FORTRAN code; The first loop identifies the block containing the shock saturation and refines this block, the subsequent loop searches further for the block containing the bottom saturation and refines this block and the blocks between, which defines the shock width relative the coarse grid. By the definition of the bottom saturation, TOL should be set to zero, however, for computational purposes, TOL equals a small number greater than zero. The last 'if' statement is necessary if we wish to terminate the computations before the front has reached the out end boundary.

We note that the algorithm only needs the knowledge of the top and bottom saturation of a shock. Since these numbers should be defined by the splitting of the fractional flow function, the algorithm may easily be extended to more general cases. A more general strategy would be to use the gradient of the coarse grid characteristic solution as a criterion to determine the shock region, since we expect large gradients to coincide, at least to some extent, with the location of the shock region.

6.1.4 Diffusion Correction

A standard algorithm for constructing discrete finite element equations, is given by Axelsson and Barker [35], and has been used to implement the given Petrov-Galerkin method. In what follows we give a short description of the routines found in the flow chart Figure 11 and we refer to the standard algorithm given in [35], for details not specific for our implementation.

The main routine `d1esol` contains the call sequence to the routines needed beginning with the initialization routine `d1init`. This routine computes averages of the nonlinear coefficients which is used when the element equations are computed. The handling of the nonlinear coefficients is based on the following two assumptions:

(iv). The nonlinearities may be removed by substitution of the characteristic solution.

(v). The grid refinement allows us to treat the coefficients as constants on each element.

The first assumption follows from the expansion given by equation (34), which of course implies a constraint on the time step. However, if the diffusive time scale is slow compared to the characteristic time scale, i.e. a convection dominated process, this is not a serious constraint. Further, assumption **(iv)** can be replaced by an iteration procedure, solving a fully implicit set of equations, and using the characteristic solution as initial approximation. However, the numerical experiments performed show that very good results are obtained without iteration, as long as the characteristic solution is not to far from the real solution.

The second assumption, **(v)**, is made to simplify the integration of the element equations and involves an averaging procedure over each element to give the weight of the coefficients. There are several ways to perform this averaging process. Here, we have chosen to use the mean value of the characteristic solution as a point to define the average values of the coefficients. These values are stored in the arrays *a(1:nel)* and *b(1:nel)*, containing the diffusion and convection coefficients respectively, ordered from one to the number of elements *nel*.

A technical note should be given on the diffusion coefficients; Ahead of the front small oscillations may occur and give negative saturation values. If these values allow the nonlinear diffusion coefficient to become negative, the oscillations may grow unstable. We have therefor chosen to replace the diffusion coefficient with its absolute value to avoid such instabilities.

The entries of the linear system are computed by the routines `d1pstm` and `d1pstv`, which compute the stiffness matrix and the right hand side of the system respectively. These routines update the entries of the system, element for element, using the local stiffness matrices and right hand sides, computed successively by the routines `d1elsm` and `d1elsv`. The tridiagonal structure of the stiffness matrix allows us to store the entries in three arrays named *smkc*, *smka*, and *smkg* containing the subdiagonal, diagonal and superdiagonal of the matrix respectively, while the right hand side of the system is stored in the array *smvb*.

Because of the simplicity of the trial and test spaces and by assumption **(v)**, we have chosen to use analytic expressions for the entries of the linear system. The approximate optimal test functions associated with the linear trial space involves the computations of the parameters:

$$c_i = 3(\frac{2}{\beta_i} - \coth(\frac{\beta_i}{2})), \quad i = 1, 2, \ldots, nel, \tag{174}$$

where $\beta_i = b_i h / a_i$ is the local mesh Péclet number, and a_i and b_i is averages of the diffusion and convection coefficients over each element, defined above. These parameters is computed by the external function routine `cfunc`.

To avoid computing exponentials for large arguments, we use the asymptotic expansions:

$$c \to \begin{cases} 3 & \text{when } \beta \to -\infty, \\ -3 & \text{when } \beta \to \infty. \end{cases}$$

Similarly we avoid the removable singularity of c_i at $\beta = 0$ by using the limit of c_i when $\beta \to 0$ for small β. Thus, the computational form of c_i is defined to be:

$$c_i = \begin{cases} 3 & \beta \leq -3, \\ c_i(\beta) & -3 < \beta \leq -0.001, \\ 0 & -0.001 \leq \beta \leq 0.001, \\ c_i(\beta) & 0.001 \leq \beta < 3, \\ 3 & -3 \leq 3 \leq \beta, \end{cases}$$

where $c_i(\beta)$ is the analytic expression given by (174).

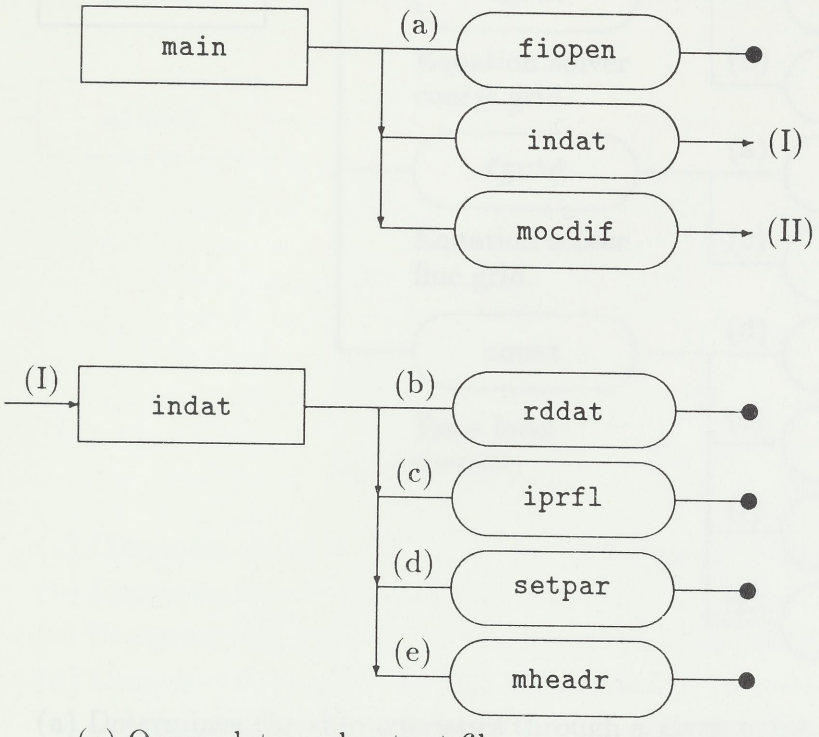
Typical test functions have been drawn in Figure 2 for negative mesh Péclet numbers. As noted before, these test functions gives the appropriate upstream weighting for our problem, although we should point out that the test functions are downstream relative the characteristic flow described by the problem. However, the upstream weighting is determined relative to the convective term $b(u)$, which represents a flow in the opposite direction of the characteristic flow.

A last assumption should be stated, which implicitly has been taken for granted:

(vi). The nonlinear mixed boundary conditions associated with the physical boundaries of the model problem considered, does not affect the characteristic solution. The inner boundaries defining a shock region may be taken to be the characteristic coarse grid solution at the surrounding coarse grid nodes.

Consequently, we are computing the transport of an established shock front, located somewhat away from the outer boundaries of the region defined by the problem. Thus, we do not have to consider the complicated mixed and nonlinear boundary conditions associated with the outer boundaries. The inner Dirichlet boundary conditions defined by the characteristic solution are eliminated by the routine `d1dbdr`.

The elimination of the complete system is performed by the LINPACK routine `sgtsl`, which solves linear systems involving general (not symmetric) tridiagonal matrices. The calling routine, initializing the LINPACK routine, is named `d1lsol`. For further details on the LINPACK Users' Guide [36].

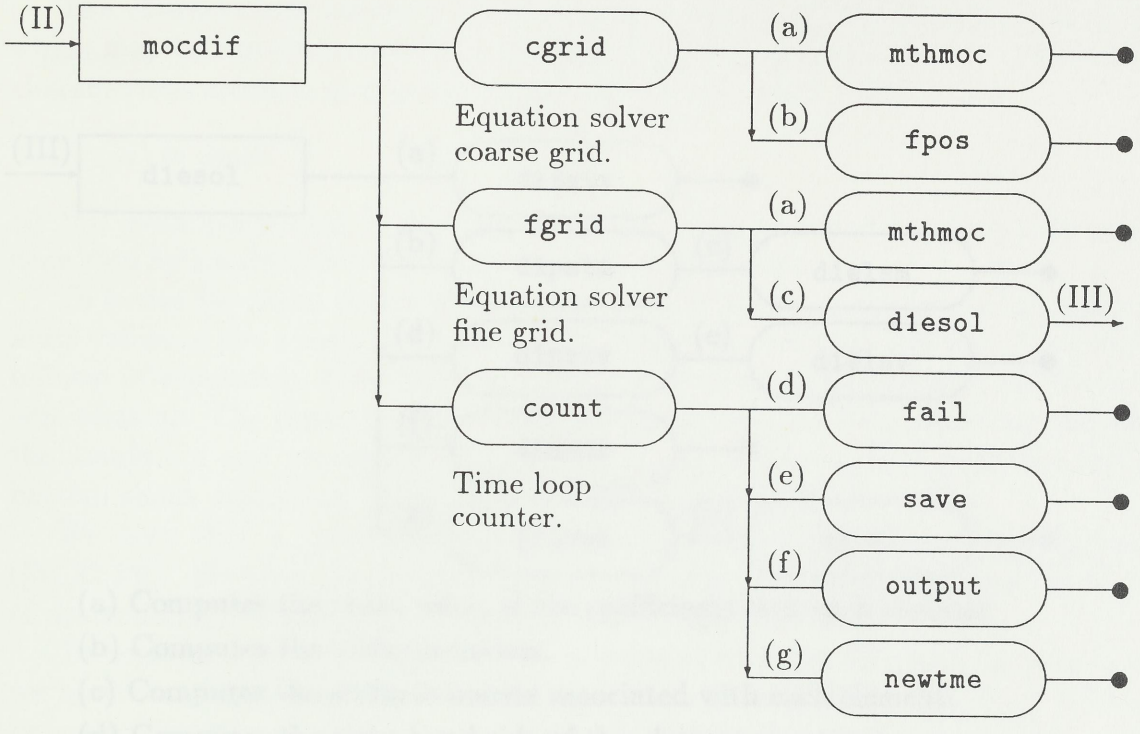


(a) Opens data and output files.
(b) Reads input parameters.
(c) Reads the initial profile.
(d) Initializes parameters/variables and determines an initial time step.
(e) Makes header information for the output files.

Figure 9: Main routines and input routines.

(a) Opens data and output files.
(b) Reads input parameters.
(c) Reads the initial profile.
(d) Initializes parameters/variables and determines an initial time step.
(e) Makes header information for the output files.

Figure 9: Main routines and input routines.

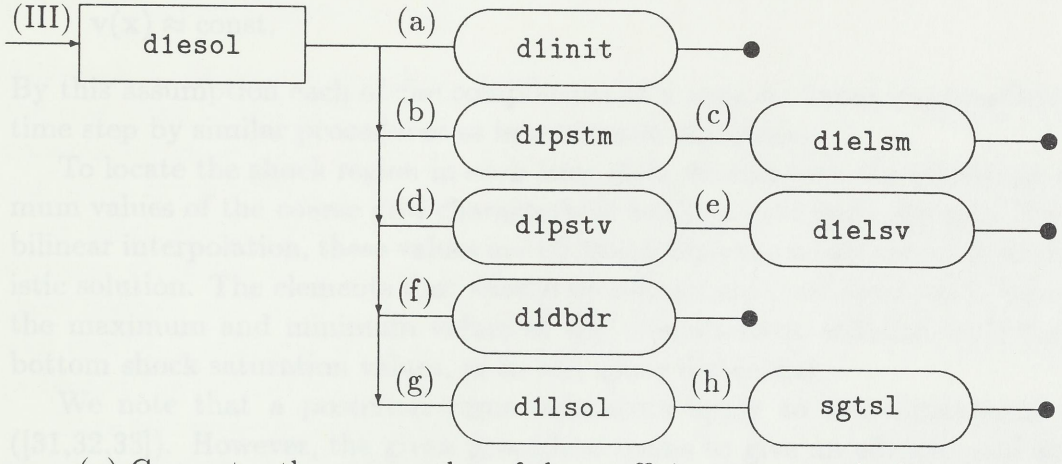


(a) Determines the characteristics through a given point.
(b) Determines the front position and and the elements to be refined.
(c) Determines diffusion correction on the refined elements.
(d) Checks if the equation solvers have converged.
(e) Saves the saturation profile for the next time step.
(f) Writes the solution to output files for a given output time.
(g) Changes next output time and the time step (if necessary).

Figure 10: Equation solvers.

(a) Determines the characteristics through a given point.
(b) Determines the front position and and the elements to be refined.
(c) Determines diffusion correction on the refined elements.
(d) Checks if the equation solvers have converged.
(e) Saves the saturation profile for the next time step.
(f) Writes the solution to output files for a given output time.
(g) Changes next output time and the time step (if necessary).

Figure 10: Equation solvers.



(a) Computes the mean value of the coefficients over each element.
(b) Computes the stiffness matrix.
(c) Computes the stiffness matrix associated with each element.
(d) Computes the right hand side of the element equations.
(e) Computes the local right hand side associated with each element.
(f) Adjusts for boundary conditions.
(g) Determines the solution of the linear system.
(h) Linear solver- Linpack routine.

Figure 11: Elliptic solver.

(a) Computes the mean value of the coefficients over each element.
(b) Computes the stiffness matrix.
(c) Computes the stiffness matrix associated with each element.
(d) Computes the right hand side of the element equations.
(e) Computes the local right hand side associated with each element.
(f) Adjusts for boundary conditions.
(g) Determines the solution of the linear system.
(h) Linear solver- Linpack routine.

Figure 11: Elliptic solver.

6.2 Two-dimensional Code

Implementation of the two-dimensional code is in large parts straightforward extensions of the one-dimensional code. Since much work still remains before a satisfactory code is obtained, the coding will only be briefly discussed here.

The pressure equation is solved by a mixed finite element code, developed at the Institute for Scientific Computation, University of Wyoming. This code is based on principles given in [17].

The characteristic solver is based on the assumption of a smooth velocity field which may be approximated by a constant locally in space. Hence, to compute the characteristic solution at a given point $\mathbf{x}$ we shall use that

$$\mathbf{v}(\mathbf{x}) \approx \text{const.}$$

By this assumption each of the components of $\bar{\mathbf{x}}$ may be found independently in each time step by similar procedures as in one space dimension.

To locate the shock region in each time step, we compute the maximum and minimum values of the coarse grid characteristic solution over each element. Since we use bilinear interpolation, these values are obtained from the nodal values of the characteristic solution. The elements that should be refined are then determined by comparing the maximum and minimum values of the characteristic solution with the top and bottom shock saturation values, as in one space dimension.

We note that *a posteriori* error estimators apply to two-dimensional problems ([31,32,33]). However, the given procedure seems to give an efficient and simple way of determining the shock region without costly computations.

To construct the element equations associated with the diffusion correction step, we have once more used the standard algorithms given in [35]. The main difference between the one and two-dimensional codes is therefor the construction of the preconditioner for the two-dimensional problem.

Although, the preconditioner is mainly determined by the algorithm given in section 5.4, step two and three of this algorithm gives room for different choices.

In the present code, step two is computed in terms of $B_{\Gamma_{ij}}$, given as follows:

$$B_{\Gamma_{ij}}(U_E, \psi_E) = (U_E, \psi_{Ek}) + ((b_k U_E)', \psi_{Ek}) + \epsilon(D_{kk} U_E', \psi_{Ek}')$$

where ψ_{Ek} is the one-dimensional restriction of ψ_E to the boundary segment Γ_{ij}, $k = 1$ or 2 depending on whether Γ_{ij} is parallel to the x- or y-axis and prime denotes derivation along Γ_{ij}. The right hand side of equation (166), is straightforward obtained in terms of the two-dimensional test functions ψ_E, with support on a strip of width $2h_i$ along Γ_{ij}.

Step three of the algorithm is simply computed by extending the coarse grid operator into Ω_i. Hence, the left hand side of equation (167) is computed as if no refinement of Ω_i is performed. We note that this differs from the definition of θ_V and ψ_V given in section 5.4. In a next version of the code, a more sophisticated coarse grid operator will be used to determine U_V, based on the definitions of θ_V and ψ_V given in section 5.4.

The preconditioned iteration implemented so far converges to a correct result, but the convergency rate is still fairly slow. However, it seems to be only minor modifications left before a good preconditioner is implemented.

6.3 Modifications of the Characteristic Solver

The characteristic solver used in the present codes limits the class of problems that is possible to solve. We shall give extensions of the characteristic solver in two directions, which may be implemented later. The extensions will be discussed in a two-dimensional framework, but apply as well to one or higher space dimensions.

6.3.1 Curved Characteristics

Although a constant $\mathbf{v}$ is a good local approximation in most of Ω, this is not at all the case close to the wells and the corners. We know that the velocity field has a $1/r$ dependency close to the wells, where r is the distance from the well. However, this does not cause much trouble since $u(t, \mathbf{x})$ by assumption is slowly varying in this region.

A more severe problem is that the velocity field is strongly curved close to the corners, as shown on Figure 21 (a). Using the constant velocity approximation in these regions, obviously gives a poor approximation to the characteristics, and may as well results in characteristics out of Ω. To handle this problem, we have so far simply projected the characteristics out of Ω onto the boundary $\partial\Omega$.

Since the velocity field in general may not be expected to behave as smoothly as shown on Figure 21 (a), these problems suggest a more closely approximation of the characteristics.

Consider the hyperbolic problem:

$$u_t + g(\mathbf{x}, u)\, \mathbf{v}(\mathbf{x}) \cdot \nabla u = 0.$$

where the $\mathbf{x}$-dependency of g allows for a heterogeneous reservoir. The characteristics between successive time steps associated with this problem are given by the integral expression:

$$\bar{\mathbf{x}} = \mathbf{x} - \int_{t^{n-1}}^{t^n} g(\mathbf{x}(\tau), u^{n-1}(\bar{\mathbf{x}}))\, \mathbf{v}(\mathbf{x}(\tau)) d\tau,$$
$$\bar{\mathbf{x}} = \mathbf{x}(t^{n-1}), \quad \mathbf{x} = \mathbf{x}(t^n).$$

To improve on the straight line approximation of the characteristics, we divide $[t^{n-1}, t^n]$ into m equal parts such that $\Delta t = t^n - t^{n-1} = m \cdot \delta t$, then $\bar{\mathbf{x}}$ may be approximated by the scheme:

$$\begin{aligned}
&\bar{\mathbf{x}}_0 = \mathbf{x};\\
&\text{for } i = 1, 2, \ldots, m:\\
&\qquad \bar{\mathbf{x}}_i = \bar{\mathbf{x}}_{i-1} - g(\bar{\mathbf{x}}_{i-1}, \bar{u}^{n-1})\delta t;\\
&\text{continue;}\\
&\bar{\mathbf{x}} = \mathbf{x}_k;
\end{aligned} \tag{175}$$

In this scheme we have treated $\bar{u}^{n-1}$ as a known value. To obtain a complete solution to the characteristic problem, the scheme has to be combined with the algorithms given in section 6.1.2.

6.3.2 Growing Shock Solutions

Another objection that may be raised against the operator-splitting in its present form, is its limitation to the transport of a well established shock. This problem relates to the definition of the modified fractional flow function. An appropriate dynamical definition of the modified fractional flow function should give the physical transport and further give a correct balance between diffusion and transport in the shock region, as stated by inequality (141).

An algorithm which directly determines the form of the modified fractional flow function when a shock is building up, has been presented by Espedal and Ewing [1]. However, this algorithm assumes that the shock is building up from a zero saturation value. Here we wants to give a more general approach to the problem.

For simplicity we assume an S-shaped fractional flow function as shown on Figure 1 and we shall assume that the shock is building up on a time scale which is fastest along the main diagonal between the wells. Our intention is to use inequality (133) along the diagonal to decide whether the fractional flow function has to be modified or not. If ∇u^{n-1} is large in a given region the time step can be severely limited by (133). Hence, if ∇u^{n-1} is larger than a prescribed value for $u_b^n \leq u \leq u_t^n$, we replace $f(u)$ with $\bar{f}^n(u)$, defined as follows:

$$\bar{f}^n(u) = \begin{cases} f(u) & u_t^n \leq u \leq 1, \\ \dfrac{f(u_t^n) - f(u_b^n)}{u_t^n - u_b^n} \cdot u & u_b^n \leq u \leq u_t^n, \\ f(u) & 0 \leq u \leq u_b^n. \end{cases}$$

We note that this expression implicitly defines $\mathbf{b}^n(\mathbf{x}, u)$.

If $u_t^n = u_{BL}$ and $u_b^n = 0$, we may continue as already described. If the shock is not fully established, inequality (133) gives a sufficient condition for uniqueness when $u \in [0, u_b^n)$ and $u \in (u_t^n, 1]$, say Δt_1. However, since $u \in [u_b^n, u_t^n]$ is moved with the shock velocity given by:

$$v^n = \frac{f(u_t^n) - f(u_b^n)}{u_t^n - u_b^n},$$

nonuniqueness may develop on each side of the shock.

Let Δu be the possible change in the shock saturations in each time step. Further, let Δx_t be the positive distance between the saturation values $u_t^n + \Delta u$ and u_t^n, and let Δx_b similarly be the positive distance between u_b^n and $u_b^n - \Delta u$. We compute the time steps:

$$\Delta t_2 = \frac{\Delta x_t}{f'(u_t^n + \Delta u) - v^n}$$

and

$$\Delta t_3 = \frac{\Delta x_t}{v^n - f'(u_b^n - \Delta u)}.$$

To determine the characteristic solution $\bar{u}^{n-1}$, we shall use the time step given by:

$$\Delta t = \min\{\Delta t_1, \Delta t_2, \Delta t_3\}.$$

Then, for $u_t^n + \Delta u \leq \bar{u}^{n-1} \leq 1$ and $0 \leq \bar{u}^{n-1} \leq u_b^n - \Delta u$, the solution is uniquely determined along the characteristics defined by $f(u)$. For $u_b^n \leq \bar{u}^{n-1} \leq u_t^n$, the solution is transported along parallel characteristics given by the shock velocity v^n. The gaps in the solution, i.e. $u_b^n - \Delta u \leq \bar{u}^{n-1} \leq u_b^n$ and $u_t^n \leq \bar{u}^{n-1} \leq u_t^n + \Delta u$, can then be uniquely given by linear interpolation. We finally update the shock saturations and check for a fully established shock.

7 Numerical Experiments

We shall present numerical results from three different diffusion-convection problems. First we consider a problem resembling a core-plug simulation in one space dimension, next we compute the solution to Burgers equation, then we conclude with a two-dimensional extension of the first problem.

Typical values are chosen for the parameter ϵ scaling the diffusion term, for the mesh spacing h_o and h_i associated with the outer and inner regions respectively, and for the time step Δt. We note that the solution obtained on a composite grid is only corrected for diffusion on the inner region Ω_i.

In addition to the adaptive grid option, (`op1`), the one-dimensional code also allows for a uniform grid solution, (`op2`), with mesh spacing given by h.

In all the computations presented the mass balance is good, with an error within a few percent. We may also note that when adaptive grids are used in one space dimension, the inner region covers at most two coarse grid blocks.

7.1 Core-plug Simulation

We consider the one-dimensional example defined by equations (11) and (14), see also [2], hence:

$$f(u) = \frac{u^p}{u^p + (1-u)^p}, \quad p = 2, 3, \ldots \tag{176}$$

$$a(u) = 4u(1-u). \tag{177}$$

We note that $a(u)$ satisfies the requirements (13).

An established shock is given by the concave envelope of f, shown in Figure 12, which again defines the top shock saturation u_{BL} and the shock velocity $v_{BL} = f(u_{BL})/v_{BL}$. Thus, $\bar{f}$ and $b(u)$ are given by:

$$\bar{f}(u) = \begin{cases} f(u), & u_{BL} < u \leq 1 \\ v_{BL} \cdot u, & 0 \leq u \leq u_{BL} \end{cases}, \tag{178}$$

$$b(u) = \begin{cases} 0, & u_{BL} < u \leq 1 \\ f(u)/u - v_{BL}, & 0 \leq u \leq u_{BL}. \end{cases} \tag{179}$$

In Figures 14-20, we have computed the numerical solution of the problem stated above for $p = 3$, and for an initial profile given by the established shock:

$$u_0(x) = \begin{cases} 2(u_{BL} - 1)x + 1 & 0 \leq x \leq 1/2 \\ 0 & 1/2 < x \leq 1. \end{cases}$$

Figures 14 (a)-(d) demonstrates the effect of different diffusion levels. The first computations, Figure 14 (a), show the pure characteristic solution and the solutions obtained by setting $\epsilon = 0$ and $\epsilon = 10^{-5}$. As we expected, the solutions are completely identical. We note the stability of the diffusion correction step for very small ϵ, and

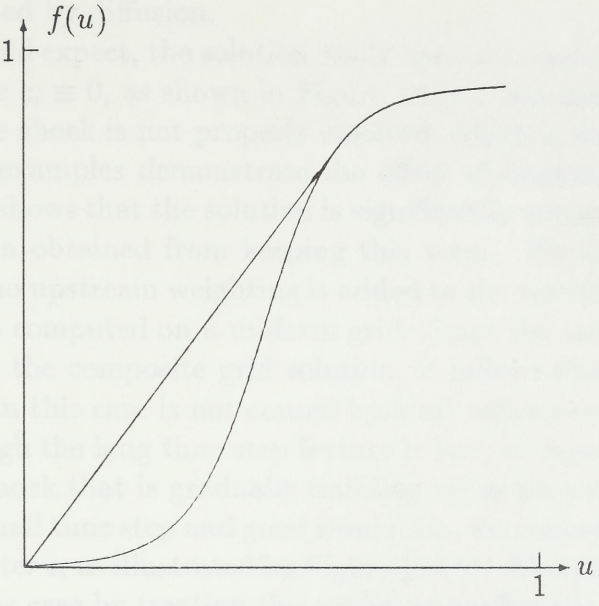


Figure 12: The concave envelope of the fractional flow function $f(u) = u^3/[u^3+(1-u)^3]$.

also that a very long time step is used, in fact, the successive profiles plotted refer to successive time steps.

Some more structure appears in Figure 14 (b). However, we also recognize the numerical diffusion introduced by not resolving the shock properly.

Figure 14 (c) shows the results obtained for a fairly large ϵ. We clearly observe the effect of the diffusion on the upper part of the shock, whereas the lower part is almost unaffected due to the singularity of the diffusion coefficient. By solving the same problem on a uniform grid with $h = h_i$, we obtain identical saturation profiles, but at a much higher computational cost. In the table below, we have compared the computations needed in each time step on the composite grid with the computations needed on the uniform grid. The shock front is assumed to be located within one coarse grid block.

	Number of nodes updated by method of characteristics	*Linear system inverted to obtain diffusion correction*
op1	50	40×40
op2	400	400×400

As further illustrated by Figure 14, the solution computed on a uniform grid with $h = h_o$ gives a very poor result.

Figure 14 (d) shows a computation of a process where the diffusion starts to dominate. Even in this case a fairly accurate result is obtained on an adaptive grid, compared with the uniform grid solution.

The profiles shown in Figures 14 (b) and (d) may be compared to the examples depicted in Figure 15. In this figure we have replaced the diffusion coefficient (177), with $a(u) \equiv 1$. The effect of changing the functional form of the diffusion coefficient, becomes evident from Figure 15 (b), where we clearly see that the lower part of the

shock is affected by diffusion.

As we should expect, the solution easily becomes unstable if the upstream weighting is removed, i.e $c_i \equiv 0$, as shown in Figure 16 (a). Similarly, the solution may become unstable if the shock is not properly resolved, which is illustrated by Figure 16 (b).

The next examples demonstrate the effect of omitting the convective term $b(u)$. Figure 17 (a) shows that the solution is significantly smeared when $b(u) \equiv 0$, compared to the solution obtained from keeping this term. We also observe that instabilities appear since no upstream weighting is added to the test functions. Figure 17 (b) is the same example computed on a uniform grid. Since the uniform grid solution is almost identical with the composite grid solution, it follows that the form of the composite grid solution in this case is not caused by local refinement.

Even though the long time step feature is lost, as expected, the given splitting also applies to a shock that is gradually building up as shown by Figure 18 (a). However, even with a small time step and good resolution, the numerical solution is very sensitive to the parameter ϵ, as illustrated by Figure 18 (b). We could probably obtain an better solution in this case by treating the nonlinear coefficients in an implicit manner.

7.2 Burgers Equation

We next compute the solution to Burgers equation, defined by the initial value problem:

$$\begin{aligned} &u_t + uu_x = \epsilon u_{xx}, \quad x \in \mathbf{R}, \quad t \in [0, T], \\ &u(x, 0) = u_0(x), \quad x \in \mathbf{R}. \end{aligned} \tag{180}$$

The build up of a shock for this problem has been computed by Russel [5], using the method of characteristics combined with a straight forward Galerkin element method to perform the diffusion correction. The characteristic operator used in [5] is defined by :

$$\frac{\partial}{\partial \tau} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x}.$$

Using inequality (67), we get the following constraint on the time step, (see [5]):

$$\max \left| u_x^{n-1} \right| \Delta t < 1 \tag{181}$$

Thus, when a shock like solution has developed, the time step is strongly inhibited by this condition.

Returning to our procedure, we note that the fractional flow function associated with Burgers equation may be written:

$$f(u) = \frac{1}{2} u^2,$$

The concave envelope of this function, defining an established shock, is shown in Figure 13. We get $u_{BL} = 1$ and $v_{BL} = 1/2$, consequently:

$$\begin{aligned} \bar{f}(u) &= \tfrac{1}{2} u \\ b(u) &= \tfrac{1}{2}(u - 1) \end{aligned}, \quad 0 \leq u \leq 1. \tag{182}$$

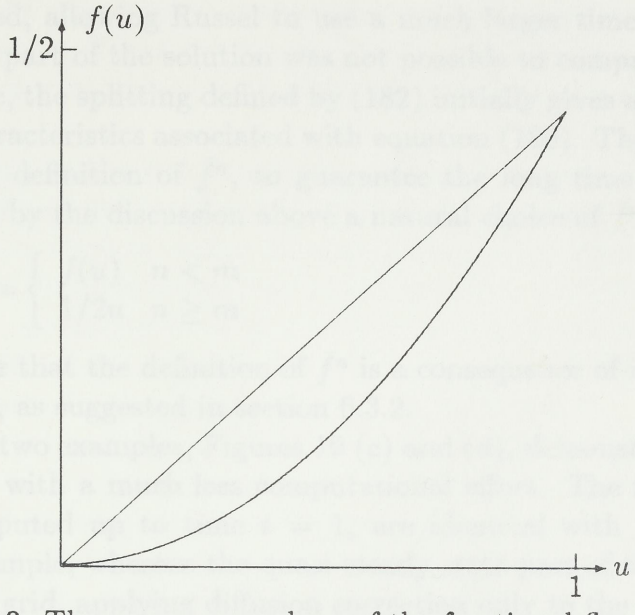


Figure 13: The concave envelope of the fractional flow function $f(u) = 1/2u^2$.

The characteristic curves defined by $\bar{f}$, is straight lines in the x,t-plane with slope $\frac{1}{2}$, and no constraints on the time step, similar to (181), is required by the characteristic solver.

To compare our results with the results given in [5], we have chosen an initial profile to be given by the continuous function:

$$u_0(x) = \begin{cases} 1 & x \leq 0 \\ 1 - 2x & 0 \leq x \leq \frac{1}{2} \\ 0 & x \geq \frac{1}{2}. \end{cases}$$

Due to the limitations on the time step given by (181), the computations presented in [5] are only carried out to time $t = 1$ when a shock like solution develops and strongly reduces the time step. For $t > 1$ an asymptotic solution of the problem may be found in [5,18], and is given by:

$$u(x,t) = \frac{1}{2} - \frac{1}{2}\tanh\left(\frac{1}{4\epsilon}\left[x - \left(\frac{1}{4} + \frac{1}{2}t\right)\right]\right). \tag{183}$$

This quasi-steady state solution is compared with the numerical experiments performed for $t > 1$.

We first compute the solution for a fairly large time step Δt. As shown by Figure 19 (a), the numerical solution immediately sets up unphysical oscillations, which clearly demonstrates that the linearization of the convective term $b(u)$ given by (34), introduces a severe constraint on the time step when the solution is rapidly changing along the approximate characteristics. Thus, to compute the first time period correctly, i.e. $t < 1$, we reduced the time step with a factor of five, whereas the steady state period $t \geq 1$, was computed without changing the time step. The result depicted in Figure 19 (b), show a monotone and physical solution for all t.

If we compare these results with the results obtained by Russel [5], we observe that inequality (181) does not restrict the time step seriously before a shock like solution

f(u)
1/2
u

has developed, allowing Russel to use a much larger time step initially, whereas the steady state part of the solution was not possible to compute due to (181).

Of course, the splitting defined by (182) initially gives a poor approximation of the physical characteristics associated with equation (180). The example therefor suggests a dynamical definition of $\bar{f}^n$, to guarantee the long time step feature for all t; Let $t^m = 1$, then by the discussion above a natural choice of $\bar{f}^n$ is:

$$\bar{f}^n(u) = \begin{cases} f(u) & n < m \\ 1/2u & n \geq m \end{cases}$$

We easily see that the definition of $\bar{f}^n$ is a consequence of inequality (133) and knowledge of u^{n-1}, as suggested in section 6.3.2.

The last two examples, Figures 19 (c) and (d), demonstrate that a good result can be obtained with a much less computational effort. The first four profiles in Figure 19 (c), computed up to time $t = 1$, are identical with the ones computed in the previous example, whereas the quasi-steady state part of the solution is computed on a composite grid, applying diffusion correction only to the inner region.

By a similar procedure, we have computed the solution for a smaller ϵ. The result which is presented in Figure 19 (d), reveals that a smaller mesh size and time step is needed to obtain a stable solution this time.

7.3 Two-dimensional Example

We consider a two-dimensional extension of the example computed in section 7.1, given as follows; Let λ_w, λ_o, $f(u)$ and $D(u)$ be defined by equations (10), (11) and (14) ($p = 2$), and let the permeability tensor $\mathbf{K}$ be equal to the identity matrix such that:

$$\mathbf{D}(u) = \begin{pmatrix} D(u) & 0 \\ 0 & D(u) \end{pmatrix}.$$

We further assume that the initial profile is given by:

$$u_0(\mathbf{x}) = \begin{cases} \frac{1}{R}(u_{BL} - 1)r + 1 & 0 \leq r \leq R \\ 0 & R < r \end{cases},$$

where r is the radial distance from the injection well and R is a distance locating the initial shock somewhat away from the well.

The splitting of f given by (178) and (179) directly apply to this problem, and we assume that the velocity field is constant locally to obtain the approximate characteristics.

In the numerical examples given by Figures 20-21 (a)-(d), the distances between successive saturation contours correspond to a change of 0.1 in the saturation. The time step is typically chosen such that the front is moved approximately one coarse grid block in each time step. As in one space dimension, the validity of using long time steps is clearly demonstrated since the shape of the front is very stable.

The computations presented show the effect of the variation of the small parameter ϵ, which scales the effect of capillary forces. It also demonstrates the effect of different local refinements of the coarse grid.

Some of the runs presented are compared with equivalent computations on a globally refined grid. There are no noticeable differences in the results. However, the computational work is increased, typically by a factor of three, to obtain the uniform grid solution.

In addition to the computational gain from using a composite grid that locally resolves the shock region, the code runs nicely in parallel on the Alliant FX/8, but so far we have not tried to exploit the parallelism in an optimal manner.

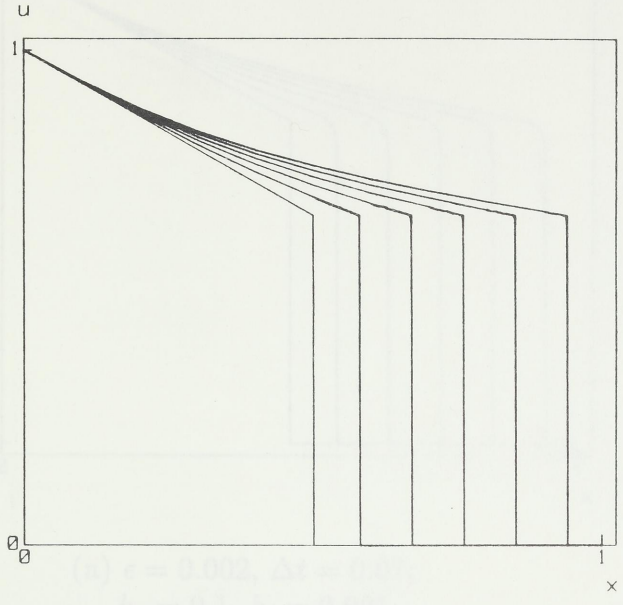
Figure 20 (a) shows a well resolved front for $\epsilon = 5 \times 10^{-3}$ with 20×20 nodes on the refined grids. A refinement of 10×10 gives a small numerical diffusion as in Figure 20 (c). Figures 20 (a)-(c) show the time evolution of the front. The last run in this series shows the start of a channeling into the production well.

Figure 21 (a) and (b) show that the reduction of the refinement of the coarse grid from 10×10 to 5×5 nodes still gives a well resolved front for $\epsilon = 10^{-2}$. We have also plotted a typical velocity field in Figure 21 (a).

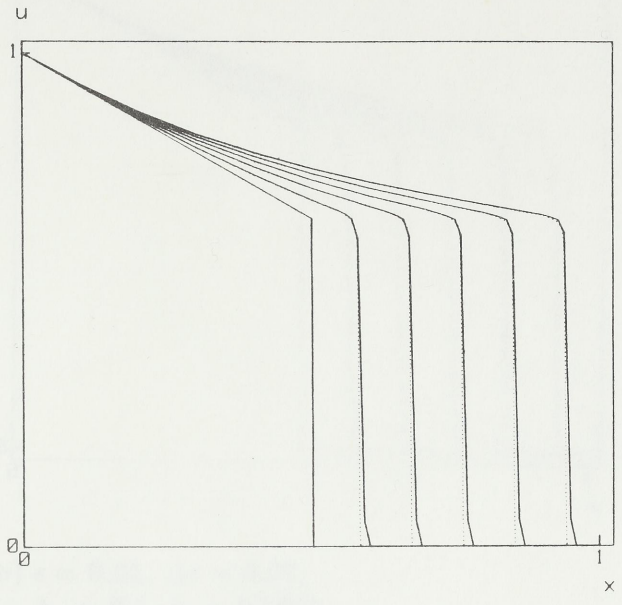
The ability of the code to handle very sharp fronts, is demonstrated by Figure 21 (c). To obtain this result we have chosen $\epsilon = 10^{-3}$.

Figure 21 (d) gives the saturation along the diagonal for different ϵ. It clearly show the effect of the diffusion term for different diffusion levels.

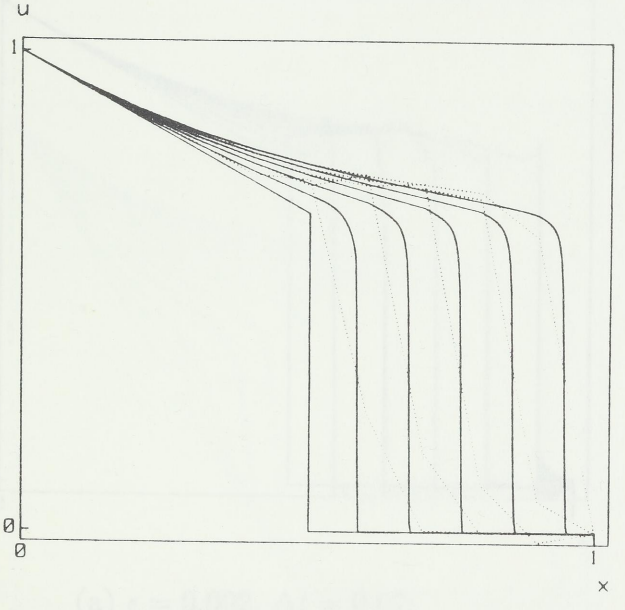
The computational results clearly demonstrates the feasibility of the method. So far, we have not optimized important parts of the code, e.g. the preconditioned conjugate gradient iteration. However, fairly large problems run very fast on an HP-9000/318 work station.



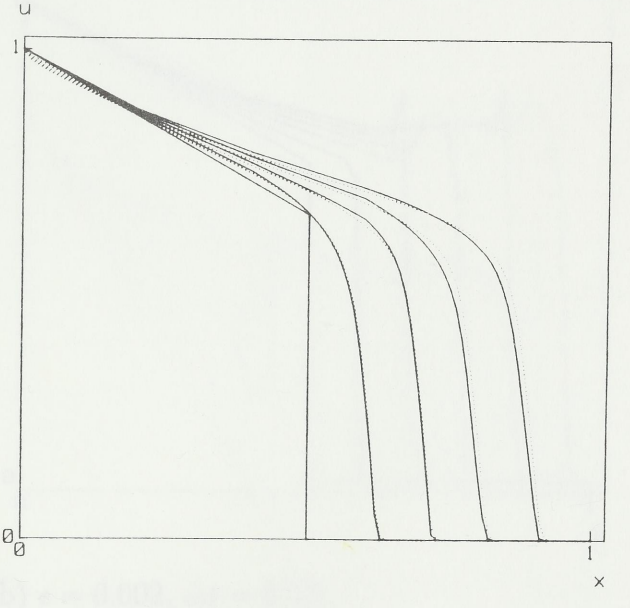
(a) $\epsilon = 0$, $\epsilon = 10^{-5}$ or
pure characteristic solution;
$\Delta t = 0.07$, $h_o = 0.1$, $h_i = 0.001$.



(b) $\epsilon = 0.002$, $\Delta t = 0.07$;
$h_o = 0.1$, $h_i = 0.001$(solid lines);
$h_o = 0.1$, $h_i = 0.01$ (dotted lines).

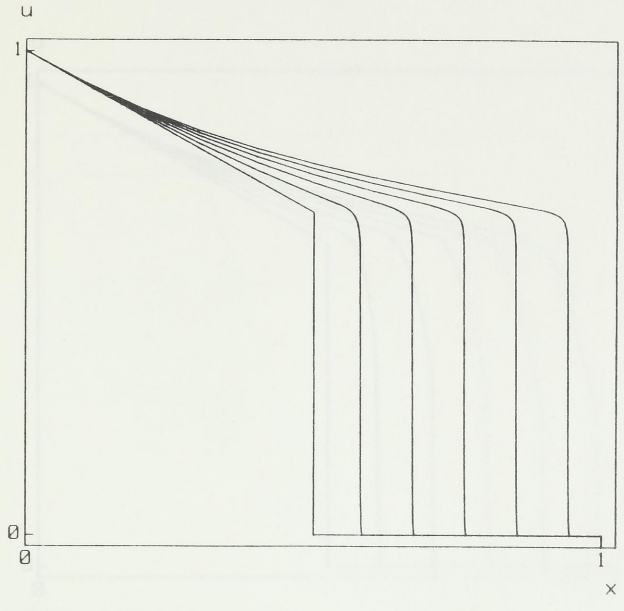


(c) $\epsilon = 0.01$, $\Delta t = 0.07$;
$h_o = 0.1$, $h_i = 0.0025$ (solid lines);
$h = 0.1$ (dotted lines).

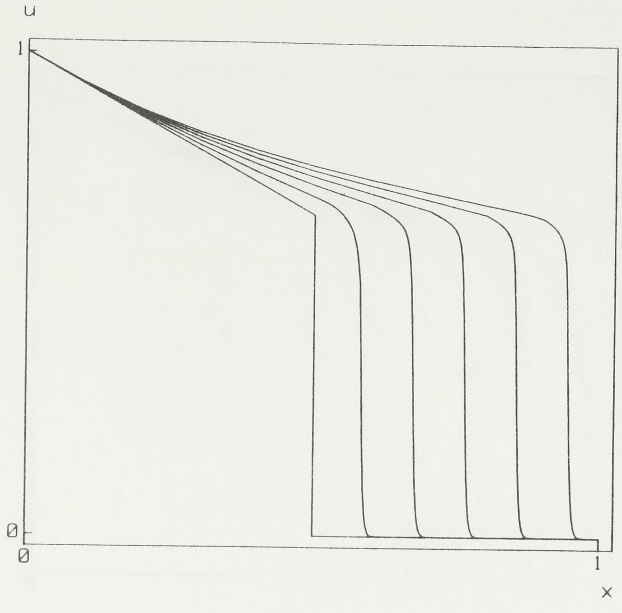


(d) $\epsilon = 0.1$, $\Delta t = 0.07$;
$h_o = 0.2$, $h_i = 0.01$ (solid lines);
$h = 0.01$ (dotted lines).

Figure 14: Core-plug simulations, effect of different diffusion levels.

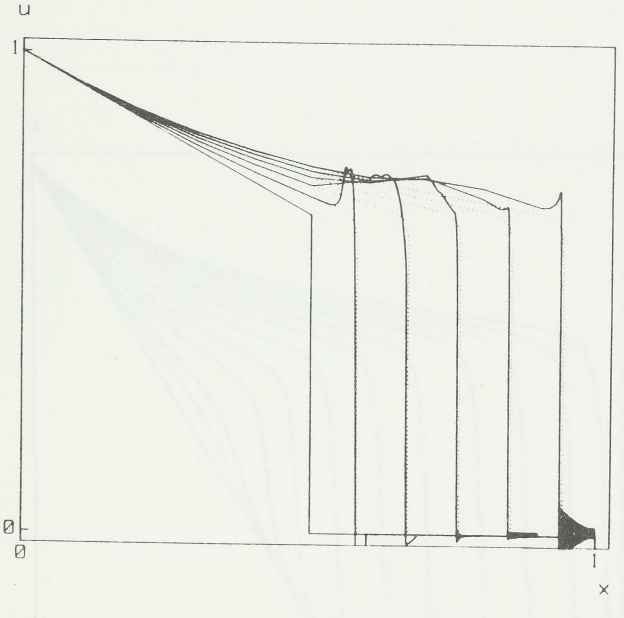


(a) $\epsilon = 0.002$, $\Delta t = 0.07$;
$h_o = 0.1$, $h_i = 0.001$.

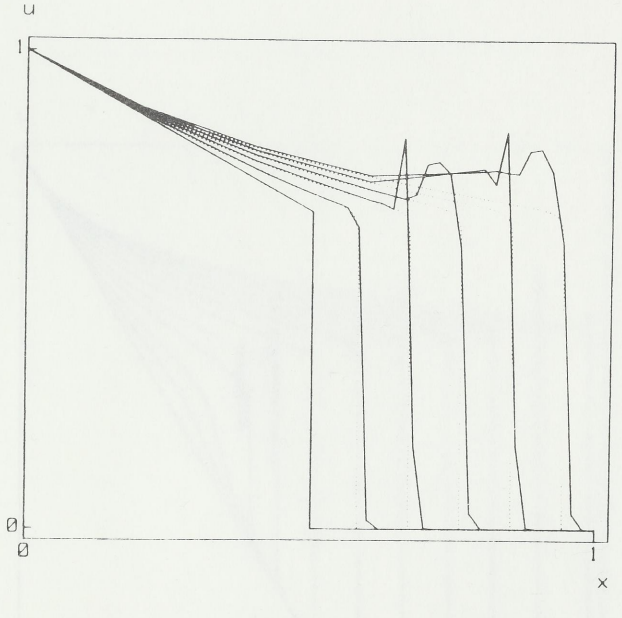


(b) $\epsilon = 0.01$, $\Delta t = 0.07$;
$h_o = 0.1$, $h_i = 0.0025$.

Figure 15: Core-plug simulations, $a(u) \equiv 1$.



(a) $\epsilon = 0.002$, $\Delta t = 0.07$;
$h_o = 0.1$, $h_i = 0.001$ (solid lines);
$h_o = 0.1$, $h_i = 0.001$ (dotted lines).



(b) $\epsilon = 0.002$, $\Delta t = 0.07$;
$h_o = 0.2$, $h_i = 0.02$ (solid lines);
$h_o = 0.1$, $h_i = 0.001$ (dotted lines).

Figure 16: Unstable solutions; (a) No upstream weight added; (b) The shock is not properly resolved.

(a) $\varepsilon = 0.002$, $\Delta t = 0.07$;
$h_0 = 0.1$, $h_l = 0.001$.

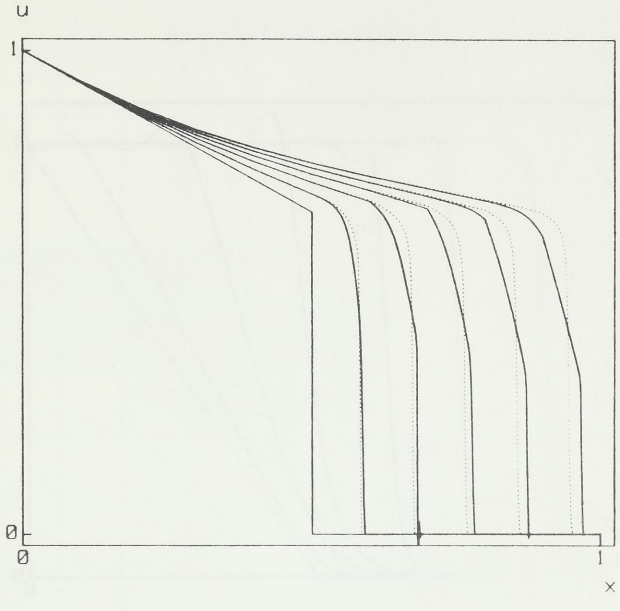
(b) $\varepsilon = 0.01$, $\Delta t = 0.07$;
$h_0 = 0.1$, $h_l = 0.0025$.

Figure 15: Core-plug simulations, $a(u) \equiv 1$.

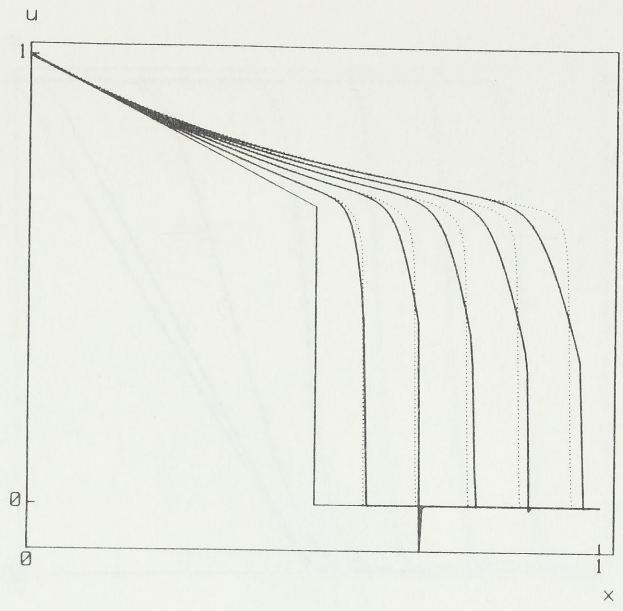
(a) $\varepsilon = 0.002$, $\Delta t = 0.07$;
$h_0 = 0.1$, $h_l = 0.001$ (solid lines);
$h_0 = 0.1$, $h_l = 0.001$ (dotted lines).

(b) $\varepsilon = 0.002$, $\Delta t = 0.07$;
$h_0 = 0.2$, $h_l = 0.02$ (solid lines);
$h_0 = 0.1$, $h_l = 0.001$ (dotted lines).

Figure 16: Unstable solutions; (a) No upstream weight added; (b) The shock is not properly resolved.

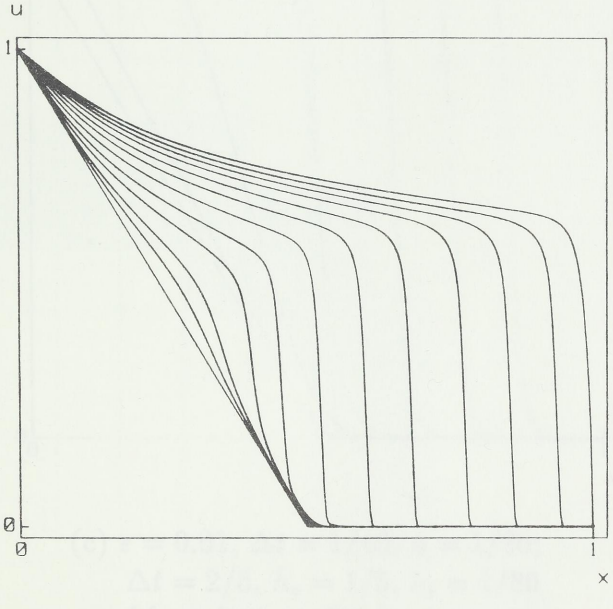


(a) $\epsilon = 0.002$, $\Delta t = 0.07$;
$h_o = 0.1$, $h_i = 0.0025$.

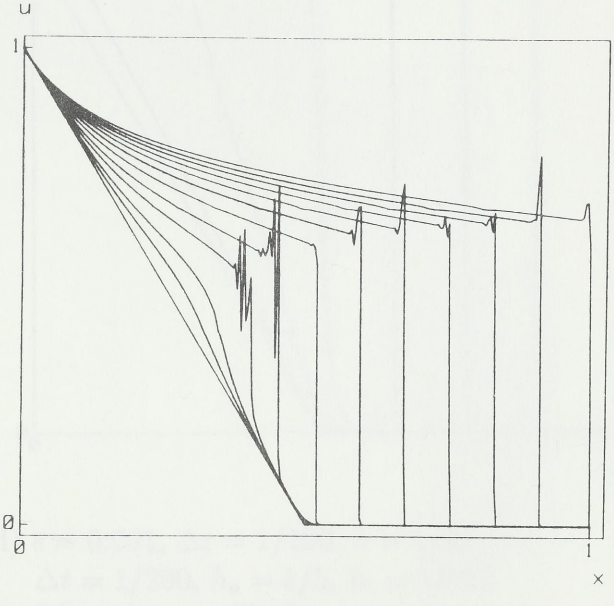


(b) $\epsilon = 0.002$, $\Delta t = 0.07$;
$h = 0.0025$.

Figure 17: Core-plug simulations, effect of omitting convective term, $b(u) \equiv 0$.



(a) $\epsilon = 0.01$, $\Delta t = 0.005$, $h = 0.004$.



(b) $\epsilon = 0.001$, $\Delta t = 0.003$, $h = 0.002$.

Figure 18: Growing shock solutions.

(a) $\epsilon = 0.002$, $\Delta t = 0.07$; $h_0 = 0.1$, $h_1 = 0.0025$.

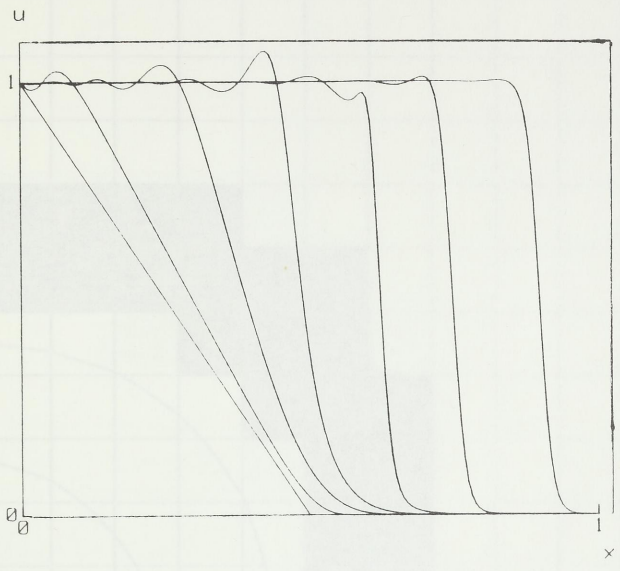
(b) $\epsilon = 0.002$, $\Delta t = 0.07$; $A = 0.0025$.

Figure 17: Core-plug simulations, effect of omitting convective term, $h(u) \equiv 0$.

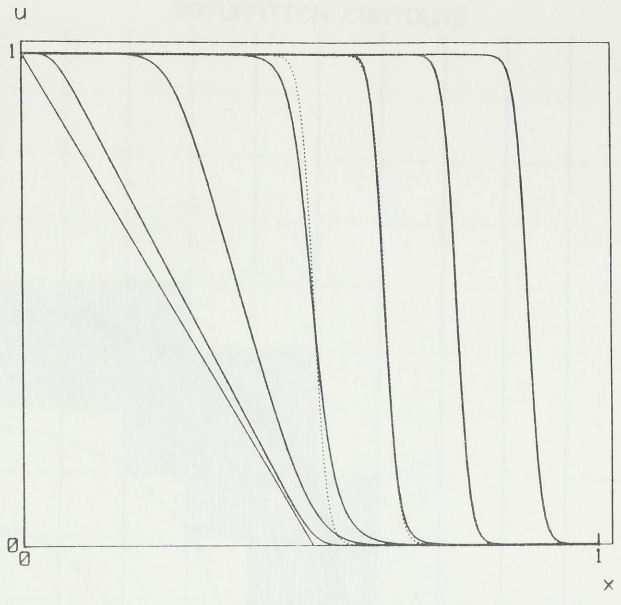
(a) $\epsilon = 0.01$, $\Delta t = 0.005$, $\lambda = 0.004$.

(b) $\epsilon = 0.001$, $\Delta t = 0.003$, $\lambda = 0.002$.

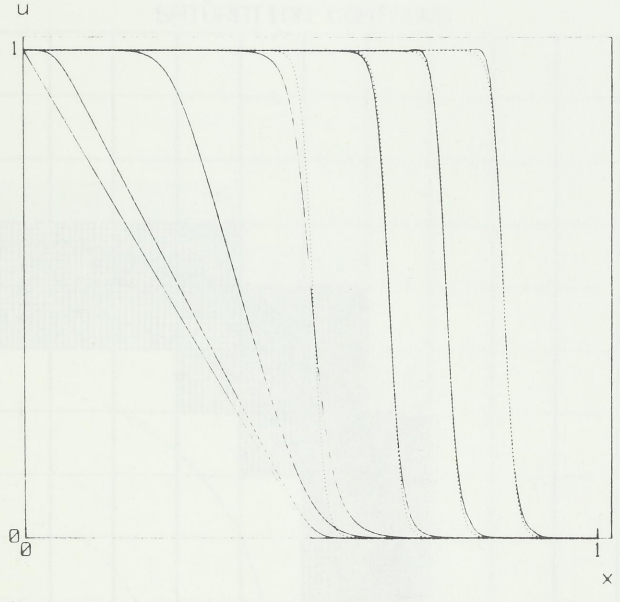
Figure 18: Growing shock solutions.



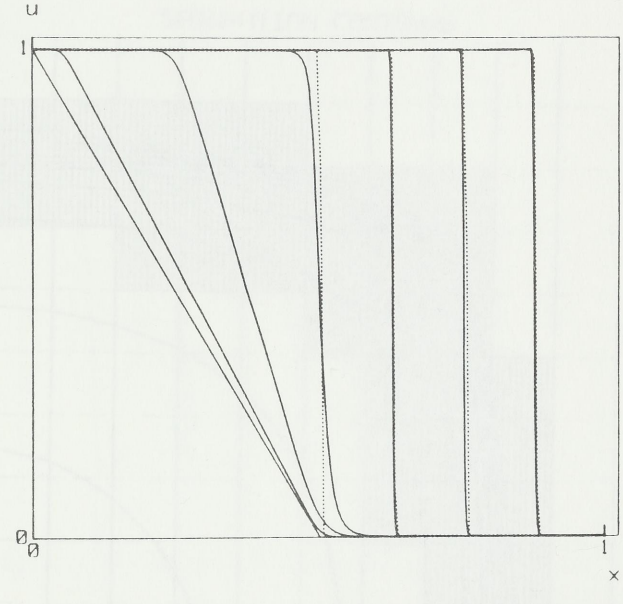
(a) $\epsilon = 0.01$, $\Delta t = 1/10$, $h = 1/80$.



(b) $\epsilon = 0.01$, $\Delta t = 1/40$, $h = 1/80$; $\Delta t = 1/10$ (three last profiles).



(c) $\epsilon = 0.01$, $\Delta t = 1/40$, $h = 1/80$; $\Delta t = 2/5$, $h_o = 1/5$, $h_i = 1/80$ (three last profiles).



(d) $\epsilon = 0.001$, $\Delta t = 1/100$, $h = 1/80$; $\Delta t = 1/200$, $h_o = 1/5$, $h_i = 1/250$ (three last profiles).

Figure 19: Computed solutions to Burgers equation. the dotted profiles represent the asymptotic solution.

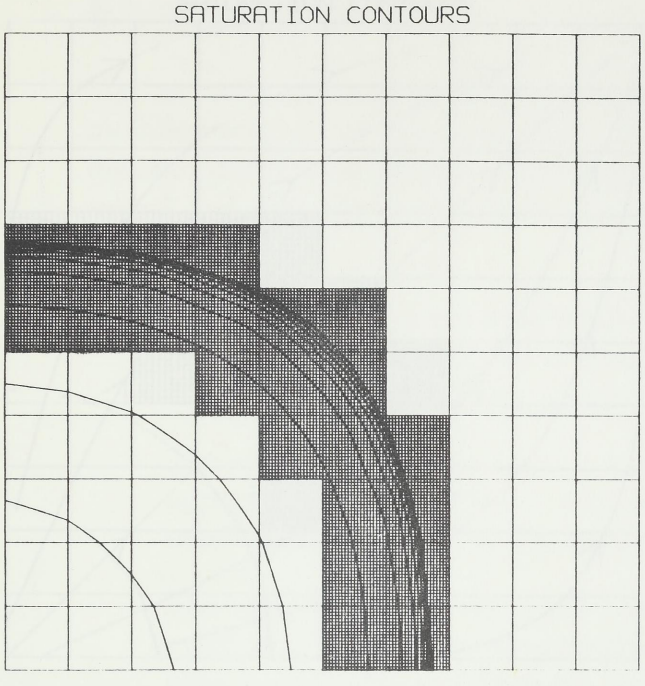
(a) $\epsilon = 0.01$, $\Delta t = 1/10$, $h = 1/80$.

(b) $\epsilon = 0.01$, $\Delta t = 1/40$, $h = 1/80$; $\Delta t = 1/10$ (three last profiles).

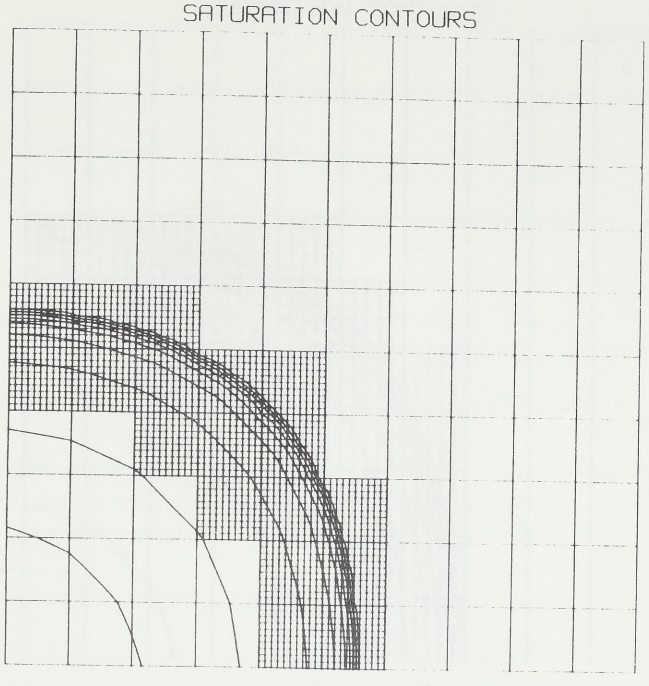
(c) $\epsilon = 0.01$, $\Delta t = 1/40$, $h = 1/80$; $\Delta t = 2/5$, $h_c = 1/5$, $h_f = 1/80$ (three last profiles).

(d) $\epsilon = 0.001$, $\Delta t = 1/100$, $h = 1/80$; $\Delta t = 1/200$, $h_c = 1/5$, $h_f = 1/250$ (three last profiles).

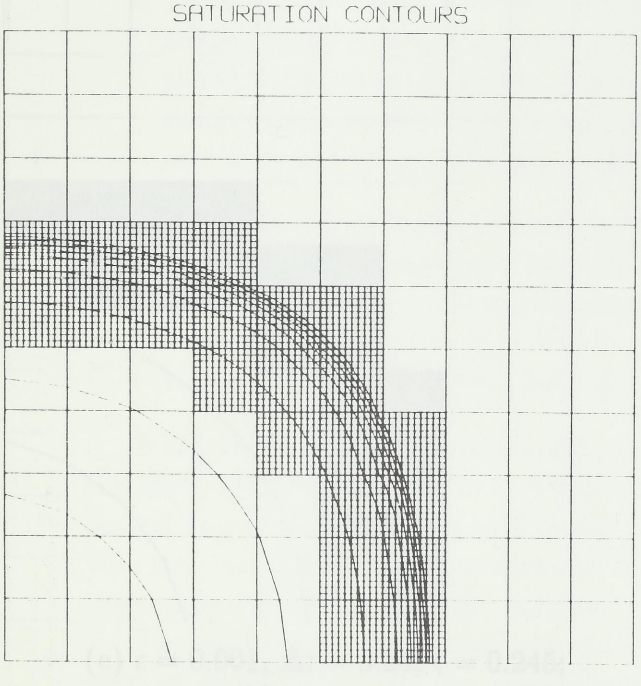
Figure 19: Computed solutions to Burgers equation. the dotted profiles represent the asymptotic solution.



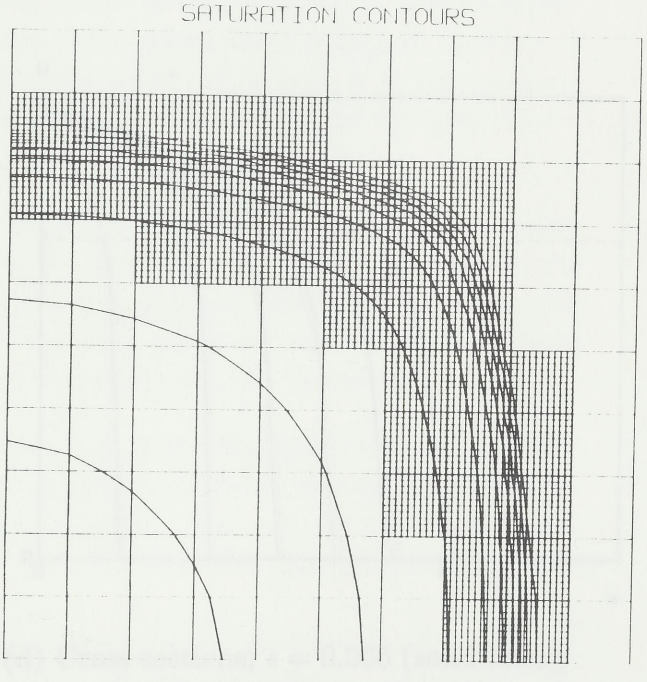
(a) $\epsilon = 0.005$, $\Delta t = 0.04$, $t = 0.245$; $h_o = 1/10$, $h_i = 1/200$.



(b) $\epsilon = 0.005$, $\Delta t = 0.04$, $t = 0.165$; $h_o = 1/10$, $h_i = 1/100$.

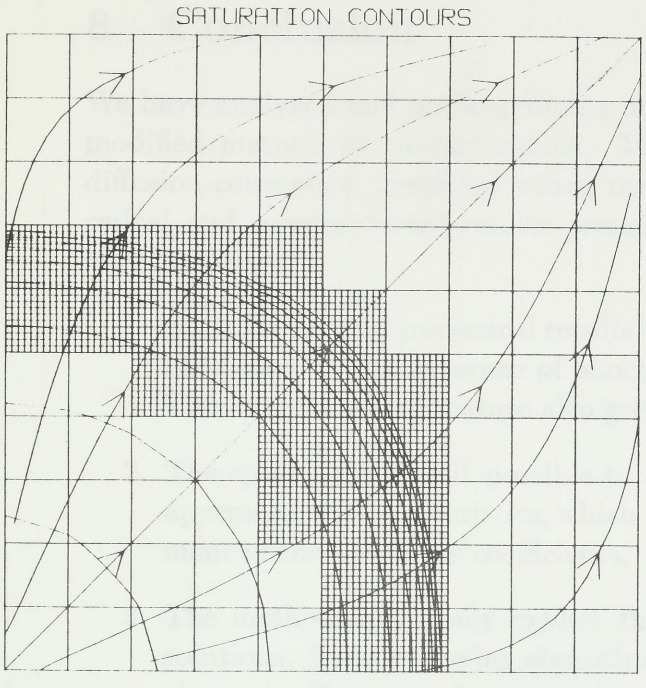


(c) $\epsilon = 0.005$, $\Delta t = 0.04$, $t = 0.245$; $h_o = 1/10$, $h_i = 1/100$.

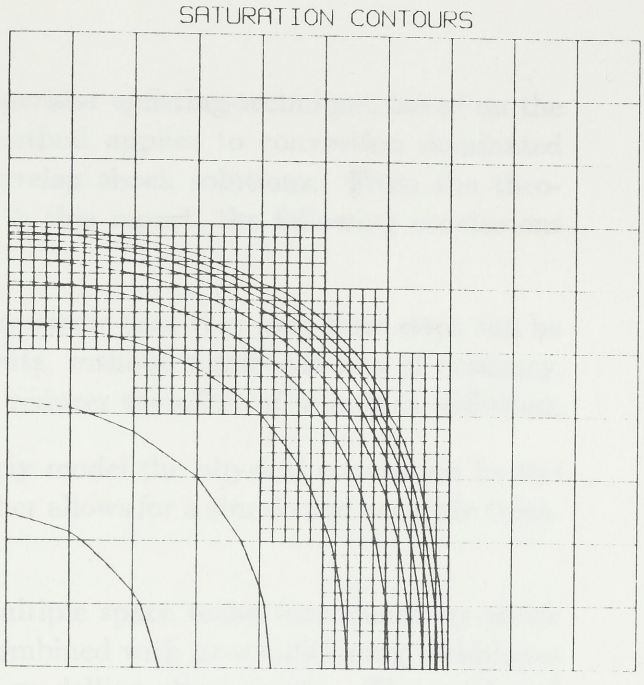


(d) $\epsilon = 0.005$, $\Delta t = 0.04$, $t = 0.405$; $h_o = 1/10$, $h_i = 1/100$.

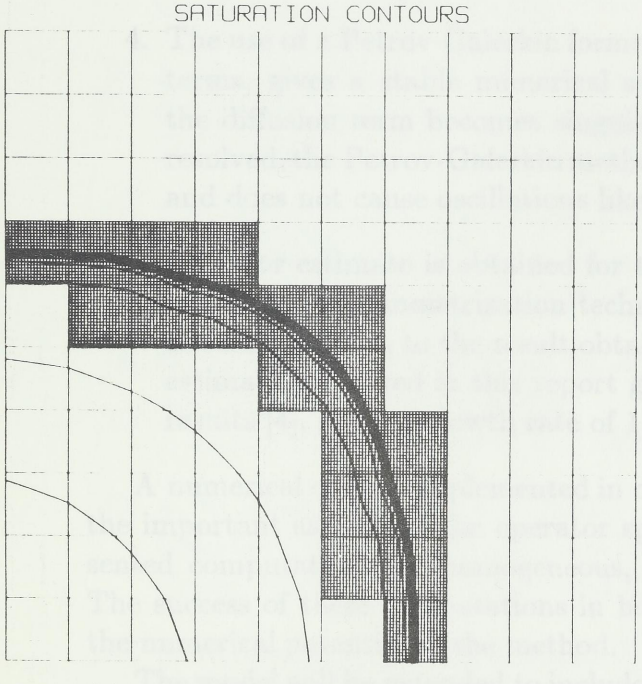
Figure 20: Saturation contours from two-dimensional computations.



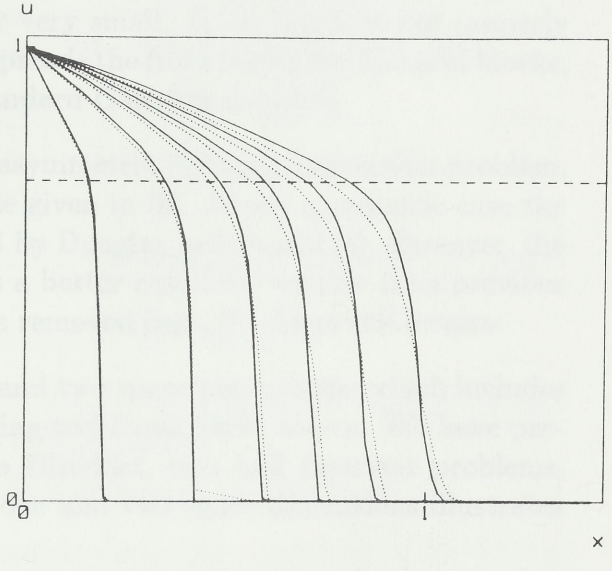
(a) $\epsilon = 0.01$, $\Delta t = 0.04$, $t = 0.245$; $h_o = 1/10$, $h_i = 1/100$.



(b) $\epsilon = 0.01$, $\Delta t = 0.04$, $t = 0.245$; $h_o = 1/10$, $h_i = 1/50$.



(c) $\epsilon = 0.001$, $\Delta t = 0.04$, $t = 0.245$; $h_o = 1/10$, $h_i = 1/200$.



(d) Cross sections; $\epsilon = 0.005$ (solid lines); $\epsilon = 0.01$ (dotted lines); $\Delta t = 0.04$; $h_o = 1/10$, $h_i = 1/100$.

Figure 21: Saturation contours from two-dimensional computations.

8 Conclusion

We have analyzed and implemented a new operator splitting-technique, based on the modified method of characteristics. The method applies to convection dominated diffusion-convection problems which may develop shock solutions. From the theoretical and computational results presented in this report, the following conclusions may be drawn:

1. Both theory and numerical results demonstrate that very long time steps can be used even in the presence of shock fronts, without significant loss of accuracy. The long time step feature also greatly reduces the effect of numerical diffusion.

2. The splitting makes it possible to closely model the physical convection by the approximate characteristics, which further allows for a simple and accurate treatment of the nonlinear coefficients.

3. The method may easily exploit the multiple space scales introduced by shock solutions. Substructuring algorithms combined with preconditioning techniques gives an efficient and accurate way of modelling shock fronts. The combined substructuring and preconditioning techniques are further well suited for parallel machine architectures.

4. The use of a Petrov-Galerkin formulation of the diffusion problem with transport terms, gives a stable numerical scheme, even when the coefficient in front of the diffusion term becomes singular or very small. If the shock is not properly resolved, the Petrov-Galerkin method spreads the front over a few fine grid blocks, and does not cause oscillations like standard Galerkin methods.

5. An error estimate is obtained for the asymmetric diffusion-convection problem, based on the symmetrization technique given in [6]. In the symmetric case the estimate reduces to the result obtained by Douglas and Russel [4]. However, the estimate presented in this report gives a better result for small ϵ than previous results [4], since a growth rate of $1/\epsilon$ is removed from the Gronwall lemma.

A numerical code is implemented in one and two space dimensions, which includes the important aspects of the operator splitting-technique listed above. We have presented computations for homogeneous, zero Dirichlet, two well reservoir problems. The success of these computations in both one and two space dimensions illustrates the numerical potential of the method.

The model will be extended to include heterogeneities and growing shocks. Another area of further study is the parallelism inherent in the algorithms. Especially the preconditioner needs to be further developed. Also, as noted earlier, it may be possible to make the error-estimate sharper.

References

[1] M. S. Espedal and R. E. Ewing. Characteristic Petrov-Galerkin subdomain methods for two-phase immiscible flow.
Comp. Meth. in Appl. Mech. and Eng. 64 (1987) 113-135 North Holland

[2] H. K. Dahle, M. S. Espedal and R. E. Ewing. Characteristic Petrov-Galerkin subdomain methods for convection diffusion problems.
Springer, the IMA Volumes in Mathematics and its Applications, Vol.11

[3] T. F. Russel. The time stepping along characteristics with incomplete iteration for Galerkin approximation of miscible displacement in porous media.
SIAM J. Numer. Anal. 22 (1986) 970-1013

[4] J. Douglas, Jr and T. F. Russel. Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures.
SIAM J. Numer. Anal. 19 (1982) 871-885

[5] T. F. Russel. Galerkin time stepping along characteristics for Burger's equation.
R. Stepleman et al. (eds.), Scientific Computing (IMACS/North-Holland Publishing Company)(1983) 183-192

[6] J. W. Barrett and K. W. Morton. Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems.
Comp. Meth. in Appl. Mech. and Eng. 45 (1984) 97-122 North Holland

[7] J. C. Heinrich, P. S. Huyakorn, A. R. Mitchell and O. C. Zienkiewicz. An upwind finite element scheme for two-dimensional convective transport equations.
Inter. J. Numer. Engrg. 11 (1977) 131-143

[8] M. Ng-Stynes, E. O'Riordan, M. Stynes. Numerical methods for time-dependent convection diffusion equations.
J. Comp. and Appl. Math. 21 (1988) 289-310 North-Holland

[9] A. C. Galeao, E. G. Do Carmo. A Consistent Approximate Upwind Petrov-Galerkin Method for Convection-Dominated Problems.
Comp. Meth. in Appl. Mech. and Eng. 68 (1988) 83-95 North Holland

[10] R. E. Ewing, T. F. Russel and M. F. Wheeler. Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics.
Comp. Meth. in Appl. Mech. and Eng. 47 (1984) 73-92 North Holland

[11] C. N. Dawson, T. F. Russel, M. F. Wheeler. Some Improved Error Estimates for the Modified Method of Characteristics. To appear.

[12] J. W. Jerome. Convection-Dominated Nonlinear Systems: Analysis of the Douglas-Russel Transport-Diffusion Algorithm Based on Approximate Characteristics and Invariant Regions.
SIAM J. Numer. Anal. 25 (1988) 815-836

[13] R. G. Duran. On the Approximation of Miscible Displacement in Porous Media by a Method of Characteristics Combined with a Mixed Method.
SIAM J. Numer. Anal. 25 (1988) 989-1001

[14] J. H. Bramble, J. E. Pasciak and A. H. Schatz. The construction of preconditioners for elliptic problems by substructuring.
J. Math. Comp. 47 (1986) 103-134

[15] J. H. Bramble, R. E. Ewing, J. E. Pasciak and A. H. Schatz. A preconditioning technique for the efficient solution of problems with local grid refinement.
Comp. Meth. in Appl. Mech. and Eng. 67 (1988) 149-159 North Holland

[16] G. Chavent, G. Cohen and J. Jaffra. Discontinuous upwinding and mixed finite elements for two-phase flows in reservoir simulation.
Comp. Meth. in Appl. Mech. and Eng. 47 (1984) 93-118 North Holland

[17] R. E. Ewing. The mathematics of reservoir simulation.
Frontiers in Applied Mathematics 1 (SIAM, Philadelphia, PA, 1983).

[18] J. D. Cole. On a quasi-linear parabolic equation occurring in aerodynamics.
Quart. of Appl. Math. 9 (1951), 225-236

[19] J. Kevorkian and J. D. Cole. Perturbation methods in applied mathematics.
Applied Mathematical Sciences, 34, Springer-Verlag, 1981.

[20] S. Osher and S. Chakravarthy. High resolution schemes and the entropy condition.
SIAM J. Numer. Anal. 21 (1984) 955-984

[21] H. C. Yee, R. F. Warming and A. Harten. Implicit Total Variation Diminishing (TVD) Schemes for Steady-State Calculations.
J. Comp. Phys. 57, 327-360 (1985)

[22] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. SIAM J. Numer. Anal. 21 (1984) 995-1011

[23] B. van Leer. Towards the ultimate conservation scheme: IV. A new approach to numerical convection. J. Comp. Phys. 23, 276-299 (1977)

[24] K. W. Morton. Finite element methods for non-self-adjoint problems. Lecture Notes in Mathematics 965 (Springer-Verlag, 1982) 113-148 Mathematics of Computation 47 (1986)103-134

[25] B. Scotney. Numerical experiments and error analysis for Petrov-Galerkin methods. Numer. Anal. Rept. 11/82, Univ. of Reading, 1982

[26] L. Demkowicz and J. T. Oden. An adaptive characteristic Petrov-Galerkin finite element method for convection-dominated linear and non-linear parabolic problems in one space variable.
TICOM Rept. 85-3, The Univ. of Texas at Austin, April 1985.

[27] L. Demkowicz and J. T. Oden. An adaptive characteristic Petrov-Galerkin finite element method for convection-dominated linear and non-linear parabolic problems in one space variable.
Comp. Meth. in Appl. Mech. and Eng. 55 (1986) 63-87 North Holland

[28] S. McCormick and J. W. Thomas. The Fast Adaptive Composite Grid (FAC) Method for Elliptic Equations.
Math. of Comp., Vol. 46, 174 (1986), pp 439-456

[29] R. E. Ewing and R. D. Lazarov. Local Refinement Techniques in the Finite Element and Finite Difference Methods.
Proc. Conf. on Numer. Meth. and Appl., Sofia, Bulgaria, August 22-27, 1988 (to appear).

[30] G. H. Schmidt, F. J. Jacobs. Adaptive Local Grid Refinement an Multi-grid in Numerical Reservoir Simulation.
J. Comp. Phys. 77, 140-165 (1988)

[31] I. Babuska and W. C. Rheinboldt. Reliable error estimation and mesh adaptation for the finite element method.
Comp. Meth. in Nonlin. Mech., (J. T. Oden, Ed.), North Holland, NY, pp. 67-108 (1980)

[32] I. Babuska and W. C. Rheinboldt. A survey of *a posteriori* error estimators and adaptive approach in the finite element method.
Proceedings of China-France Symposium on Finite Element Methods (F. Kang and J. L. Lions, Eds), Gordon and Breach, NY, 1983, pp 1-56.

[33] R. E. Ewing. Adaptive local grid refinement.
Proceedings of the SEG/SIAM/SPE Conference on Mathematical and Computational Methods in Seismic Exploration and Reservoir Modeling, Houston, Jan. 1985, (W. E. Fitzgibbon, Ed.), SIAM Publications Philadelphia, PA, 1986, pp. 235-247

[34] R. E. Ewing, M. S. Espedal, J. A. Puckett and R. J. Schmidt. Simulation techniques for multiphase and multicomponent flows.
Commun. in Appl. Num. Meth., Vol. 4, 335-342 (1988)

[35] O. Axelsson and V. A. Barker. Finite Element Solution of Boundary Value Problems, Theory and Computations.
Computer Science and Applied Mathematics, Academic Press, 1984

[36] J. J. Dongarra et al. LINPACK Users' Guide. SIAM 1979

Appendix

In the following appendix, we will give proofs for the lemmas used in the main text. We first obtain a result on the inverse Riesz-representation:

Lemma 1 *Let $\theta = \sum_{i=0}^{N} c_i\theta_i$, $c_0 = c_N = 0$, and let ψ_i be the optimal test functions associated with θ_i and defined by (47). Then, for sufficiently small ϵ, $(\theta, R^{m-1}\theta)$ is strictly positive and $(R^{m-1}\theta, R^{m-1}\theta)/(\theta, R^{m-1}\theta) \leq M_\epsilon$ where M_ϵ is estimated by:*

$$M_\epsilon \sim \frac{|b|h}{\epsilon a} = |\beta|. \tag{184}$$

Proof: Let $\delta \stackrel{\text{def}}{=} e^{b/\epsilon a}$, then, since $b < 0$, it follows that $\delta \ll 1$ since $|b|/\epsilon a > 1$. We integrate (47) to obtain:

$$\frac{a}{a_0}\beta(\delta-1)R^{m-1}\theta_i = \begin{cases} (1-\delta^x)(2\delta^{x_i-x} - \delta^{x_{i-1}-x} - \delta^{x_{i+1}-x}); & x \leq x_{i-1}, \\ -1+\delta+(1-\delta^x)(2\delta^{x_i-x} - \delta^{x_{i+1}-x}) & \\ \qquad -(\delta-\delta^x)\delta^{x_{i-1}-x} \quad ; & x_{i-1} \leq x \leq x_i, \\ 1-\delta+(\delta-\delta^x)(2\delta^{x_i-x} - \delta^{x_{i-1}-x}) & \\ \qquad -(1-\delta^x)\delta^{x_{i+1}-x} \quad ; & x_i \leq x \leq x_{i+1}, \\ (\delta-\delta^x)(2\delta^{x_i-x} - \delta^{x_{i-1}-x} - \delta^{x_{i+1}-x}); & x_{i+1} \leq x. \end{cases} \tag{185}$$

By definition of a_0, β and δ it follows that:

$$\lim_{\epsilon\to 0} \frac{a}{a_0}\beta(\delta-1) = 1,$$

consequently, $R^{m-1}\theta_i$ may to leading order be written:

$$R^{m-1}\theta_i \sim \psi_i^0 \stackrel{\text{def}}{=} \begin{cases} 0 & x \leq x_{i-1}, \\ -1 & x_{i-1} \leq x \leq x_i, \\ 1 & x_i \leq x \leq x_{i+1}, \\ 0 & x_{i+1} \leq x. \end{cases} \tag{186}$$

By linearity of R^{m-1} we have:

$$R^{m-1}\theta = \sum_{i=0}^{N} c_i\psi_i^0,$$

hence, to leading order:

$$(R^{m-1}\theta, R^{m-1}\theta) \sim h\sum_{i=1}^{N}(c_i - c_{i-1})^2. \tag{187}$$

To obtain an estimate on $(\theta, R^{m-1}\theta)$, we observe that:

$$(\sum_{i=0}^{N} c_i\psi_i^0, \sum_{i=0}^{N} c_i\theta_i) \equiv 0,$$

thus, higher order terms are required to get an positive estimate. Using the exact form (185) of $R^{m-1}\theta_i$, leads to:

$$(R^{m-1}\theta,\theta) = \frac{\epsilon a}{|b|}\sum_{i=1}^{N}(c_i - c_{i-1})^2 + \left(\frac{\epsilon a}{h|b|}\right)^2 h\sum_{i=1}^{N}(4c_ic_{i-1} - \frac{3}{2}(c_i^2 + c_{i-1}^2)) + O(\delta^h). \quad (188)$$

We observe that the first sum is strictly positive except for $c_i = c_{i-1}$, in which case the second term is positive. Hence, for sufficiently small $\epsilon > 0$ the given expression is strictly positive which proves the first part of the lemma. The second part obviously follows from (187) and (188) which conclude the proof. ■

We shall next prove a lemma which gives the interpolation error from linear interpolation, introduced when the solution is transported along parallel characteristics.

Lemma 2 *Let $u(x)$ be piecewise linear with nodes $\{x_i\}_{i=1}^{N+1}$, and $v(x)$ be piecewise linear with nodes $\{\tilde{x}_i\}_{i=0}^{N+1}$. Assume that $u(x)$ interpolate $v(x)$ on $\{x_i\}_{i=1}^{N+1}$, i.e. $u_i \equiv v_i$, and that:*

$$\begin{aligned} &\tilde{x}_i - x_i = kh, \quad x_{i+1} - \tilde{x}_i = (1-k)h, \quad i = 1, 2 \ldots, N, \\ &\tilde{x}_0 = x_1, \quad \tilde{x}_{N+1} = x_{N+1}, \end{aligned} \quad (189)$$

where $0 \le k \le 1$. Then:

$$\int_{x_1}^{x_{N+1}} u'^2 dx = \int_{x_1}^{x_{N+1}} v'^2 dx - k(1-k)h^2 \sum_{i=1}^{N} h\left(\frac{v'^2_{i+1} - v'^2_i}{h}\right)^2, \quad (190)$$

where

$$\begin{aligned} &v'_i = \frac{v(\tilde{x}_i) - v(\tilde{x}_{i-1})}{h}, \quad i = 2, 3 \ldots, N, \\ &v'_1 = \frac{v(\tilde{x}_1) - v_1}{kh}, \quad v'_{N+1} = \frac{v_{N+1} - v(\tilde{x}_N)}{(1-k)h}. \end{aligned} \quad (191)$$

Proof: By definitions (189) and (191), the difference between u and v is given by, $i = 1, 2, \ldots, N+1$:

$$u(x) - v(x) = \begin{cases} (1-k)(v'_{i+1} - v'_i)(x - x_i), & x_i \le x \le \tilde{x}_i, \\ -k(v'_{i+1} - v'_{i+1})(x - x_{i+1}), & \tilde{x}_i \le x \le x_{i+1}. \end{cases}$$

Hence, the left hand side of equation (190) can be written:

$$\begin{aligned} &\int_{x_1}^{x_{N+1}} u'^2 dx = \\ &= h\textstyle\sum_{i=1}^{N}\{(v'_i + (1-k)(v'_{i+1} - v'_i))^2 k + \\ &+ (v'_{i+1} - k(v'_{i+1} - v'_i))^2(1-k)\} = h\textstyle\sum_{i=1}^{N}(kv'_i + (1-k)v'_{i+1})^2 = \\ &= h\textstyle\sum_{i=1}^{N}\left\{k^2 v'^2_i + 2k(1-k)v'_i v'_{i+1} + (1-k)^2 v'^2_{i+1}\right\} = \\ &= h\textstyle\sum_{i=1}^{N}\left\{kv'^2_i + (1-k)v'^2_{i+1} - k(1-k)(v'_{i+1} - v'_i)^2\right\} \end{aligned}$$

We may then rearrange terms and use the identity:

$$\int_{x_1}^{x_{N+1}} v'^2 dx = khv'^2_1 + h\sum_{i=1}^{N} v'^2_i + (1-k)hv'^2_{N+1},$$

to obtain the desired result. ■

For completeness, we shall also give a proof of the discrete Gronwall lemma which is a consequence of the following elementary lemma:

Lemma 3 *Let A and B be positive constants and let $\{y_n\}_{n=0}^N$ be a sequence of numbers that satisfy:*

$$y_0 = 0 \tag{192}$$

and

$$|y_n| - |y_{n-1}| \leq B|y_{n-1}| + A, \quad n \leq N. \tag{193}$$

Then:

$$|y_n| \leq \frac{A}{B}\left(\exp(nB) - 1\right), \quad n \leq N.$$

Proof: Let η_n satisfy the difference equations:

$$\eta_n = (1+B)\eta_{n-1} + A, \quad n = 1, 2\ldots,$$

$$\eta_0 = 0.$$

The solution to these equations is:

$$\eta_n = \frac{A}{B}(1+B)^n - \frac{A}{B}.$$

Since $\exp(nx) \geq (1+x)^n$ for all $x \geq 0$, we get the bound:

$$\eta_n \leq \frac{A}{B}\left(\exp(nB) - 1\right). \tag{194}$$

We shall show that for all $n \leq N$:

$$|y_n| \leq \eta_n. \tag{195}$$

The inequality is obviously satisfied if $n = 0$ because of (192). If (195) is true for $n = k-1$, then:

$$\eta_k = (1+B)\eta_{k-1} + A \geq (1+B)|y_{k-1}| + A,$$

hence, by (193):

$$\eta_k \geq |y_k|.$$

Therefor by induction, inequality (195) is true for all $n \leq N$. The lemma now follows from inequalities (194) and (195). ■

Lemma 4 (Discrete Gronwall) *Let A and B be positive constants and let $\{e_n\}_{n=0}^N$ be a sequence of numbers that satisfy:*

$$e_0 = 0 \tag{196}$$

and

$$|e_n| \leq B \sum_{i=0}^{n-1} |e_i| + A, \quad n \leq N. \tag{197}$$

Then:

$$|e_{n+1}| \leq A \exp(nB), \quad n \leq N-1.$$

Proof: Define y_n to be the sum:

$$y_n = B \sum_{i=0}^{n} |e_i|, \tag{198}$$

hence:

$$y_n - y_{n-1} = B|e_n|.$$

By (196) and (197), the sequence $\{y_n\}_{n=0}^N$ satisfy the conditions given in the previous lemma:

$$y_0 = 0$$

and

$$|y_n| - |y_{n-1}| \leq B|y_{n-1}| + AB, \quad n \leq N.$$

Hence, we may deduce from the previous lemma that:

$$|y_n| \leq A(\exp(nB) - 1), \quad n \leq N$$

Combining this inequality with the inequalities (197) and (198) proves the discrete form of the Gronwall lemma. ■