CONVERGENCE OF IMPLICIT MONOTONE SCHEMES WITH APPLICATIONS IN MULTIPHASE FLOW IN POROUS MEDIA

FELIX KWOK AND HAMDI TCHELEPI

Abstract. Phase-based upstreaming, which is a commonly used spatial discretization for multiphase flow in reservoir simulation, naturally gives rise to implicit monotone schemes when implicit time-stepping is used. The nonlinear Gauss-Seidel and Jacobi algorithms are shown to converge to a unique bounded solution when applied to the resulting system of equations. Thus, for 1D problems, we obtain an alternate, constructive proof that such schemes are well-defined and converge to the entropy solution of the conservation law for arbitrary CFL numbers. The accuracy of phase-based upstream solutions is studied for various spatial and temporal grids, revealing the importance of unconditional stability when non-uniform grids and/or variable porosity is involved.

Key words. conservation laws, multiphase flow, reservoir simulation, upstream weighting, implicit monotone schemes

AMS subject classifications. 35L65, 65H10, 65M06, 65M12

1. Introduction. In the simulation of multiphase flow in petroleum reservoirs, the most commonly used model consists of a system of n conservation laws, where n is the number of immiscible fluid phases. Each conservation law is defined for all $\mathbf{x} \in \Omega \subset \mathbb{R}^k$ $(1 \leq k \leq 3)$ and has the form

$$\frac{\partial(\phi \rho_j S_j)}{\partial t} + \nabla \cdot (\rho_j \mathbf{v}_j) = \rho_j q_j, \qquad j = 1, \ldots, n, \tag{1.1}$$

where $\phi = \phi(\mathbf{x})$ is the porosity of the medium (with $0 < \phi_{min} \leq \phi \leq 1$), $K = K(\mathbf{x}) \geq K_{min} > 0$ is the absolute permeability; for each phase j, $S_j = S_j(\mathbf{x}, t)$ is the saturation (i.e., the volume fraction occupied by phase j in the neighborhood of $\mathbf{x}$), $\rho_j > 0$ is the density, $q_j = q_j(\mathbf{x})$ is the source or sink term, and $\mathbf{v}_j$ is the phase velocity, which is given by generalized Darcy's law:

$$\mathbf{v}_j = -K\lambda_j \left[\nabla p_j - \rho_j \mathbf{g}\right]. \tag{1.2}$$

Here $\lambda_j = \lambda_j(S_1, \ldots, S_n)$ is the phase mobility, p_j is the phase pressure, and $\mathbf{g} \in \mathbb{R}^k$ is a constant vector representing the gravitational acceleration. In addition, we have the following algebraic relations:

$$\text{Saturation constraint:} \qquad \sum S_j = 1, \tag{1.3}$$

$$\text{Capillary pressure constraint:} \quad p_j - p_{j+1} = P_{cj}(S_1, \ldots, S_n), \quad j = 1, \ldots, n-1. \tag{1.4}$$

In this paper we restrict our attention to incompressible flow with zero capillarity, i.e., we assume that the densities ρ_j are constant with respect to time and space, and that $p_1 = \cdots = p_n \equiv p$. The resulting system of PDEs exhibits a mixed hyperbolic-parabolic character, which becomes apparent when we consider the various limiting cases. If we multiply (1.1) by $1/\rho_j$ and sum over $j = 1, \ldots, N$, we obtain $\sum_j \nabla \cdot \mathbf{v}_j = \sum_j q_j$, where the $\partial/\partial t$ terms cancel because of the saturation constraint. We have

$$-\nabla \cdot \left[K\lambda_T \nabla p - K\mathbf{g}\textstyle\sum_j \lambda_j \rho_j\right] = \sum_j q_j, \tag{1.5}$$

where $\lambda_T = \sum_j \lambda_j$ is the total mobility, which is assumed to be positive and uniformly bounded away from zero. Thus, for a given saturation distribution, the pressure field

satisfies an elliptic PDE. On the other hand, if we define the total velocity $\mathbf{v}_T$ by $\mathbf{v}_T = \sum_j \mathbf{v}_j$, we can rewrite $\mathbf{v}_j$ as

$$\mathbf{v}_j = \frac{\lambda_j}{\lambda_T}\left[\mathbf{v}_T - K\mathbf{g}\textstyle\sum_\ell \lambda_\ell(\rho_\ell - \rho_j)\right]. \tag{1.6}$$

This means when $\mathbf{v}_T$ is constant over time (which is the case for flow in a 1D porous medium), the phase velocities $\mathbf{v}_j$ become functions of $\mathbf{x}$ and $S_1, \ldots, S_n$ only, so when we substitute (1.6) into (1.1), we get

$$\phi\frac{\partial S_j}{\partial t} + \nabla \cdot \mathbf{v}_j(\mathbf{x}, S_1, \ldots, S_n) = 0, \qquad j = 1, \ldots, n-1. \tag{1.7}$$

As a result, saturation behaves like the solution to a system of first-order hyperbolic conservation laws, so one should expect discontinuous saturation profiles. In higher dimensions, $\mathbf{v}_T$ generally varies over time, and a strong coupling exists between pressure and saturation because of the saturation dependence of λ_j and λ_T in (1.5), as well as the dependence of $\mathbf{v}_T$ on the pressure field. However, one can still observe the same kind of discontinuities in the saturation profile in such cases.

The vastly different behavior between saturation and pressure means these variables need to be treated differently in a numerical method. The shock-forming nature of (1.7) requires the use of a finite-volume method, whereas the elliptic nature of (1.5) means pressure variables must be treated implicitly to avoid a time step restriction of the form $\Delta t = \mathcal{O}(\Delta x^2)$. One can treat saturation variables either explicitly, for which a CFL condition of the form $\Delta t = \mathcal{O}(\Delta x)$ applies, or implicitly, where no such time-step limit exists. In problems of practical interest, the porosity ϕ and permeability K are highly oscillatory, non-smooth functions of $\mathbf{x}$, and $K(\mathbf{x})$ can vary by several orders of magnitude over the domain Ω. Thus, local CFL numbers can also exhibit large spatial variations. This means the time-step limit of an explicit method, which is determined by the maximum local CFL number, is often overly restrictive compared to the *average* CFL number. For this reason, the discretization of choice for most reservoir simulators is the *fully-implicit method* (FIM), which uses finite volume in space and backward Euler in time. The numerical flux function $F_{j,il}$, which approximates the continuous flux $\int_{\partial V_{il}} \rho_j \mathbf{v}_j \cdot \mathbf{n}_{il}\, ds$, is given by

$$F_{j,il} = |\partial V_{il}|\rho_j K_{il}\lambda_j(S^*)\left(-\frac{(p_l - p_i)(\mathbf{x}_l - \mathbf{x}_i)}{|\mathbf{x}_l - \mathbf{x}_i|^2} - \rho_j\mathbf{g}\right)\cdot\mathbf{n}_{il}, \tag{1.8}$$

where $|\partial V_{il}|$ is the area of the interface between cells i and l, $\mathbf{x}_i$ and $\mathbf{x}_l$ are the locations of the centers of cells i and l, and $\mathbf{n}_{il}$ is the unit normal to the cell interface, pointing from cell i to cell l. The mobility λ_j is evaluated at the *upstream* saturation:

$$S = \begin{cases} S(x_i) & \text{if } \left(-\dfrac{(p_l - p_i)(\mathbf{x}_l - \mathbf{x}_i)}{|\mathbf{x}_l - \mathbf{x}_i|^2} + \rho_j\mathbf{g}\right)\cdot\mathbf{n}_{il} \geq 0, \\ S(x_l) & \text{otherwise.} \end{cases} \tag{1.9}$$

As we will see in section 4, the resulting numerical flux functions are different from those used in classical CFD (e.g. the Godunov and Engquist-Osher schemes). Despite being only first-order accurate, phase-based upstreaming is the preferred upwind method in reservoir simulation because it is physically intuitive, and because it is generally easier to verify a consistency condition such as (1.9) than to locate potential sonic points, which vary over the domain and are strong functions of permeability

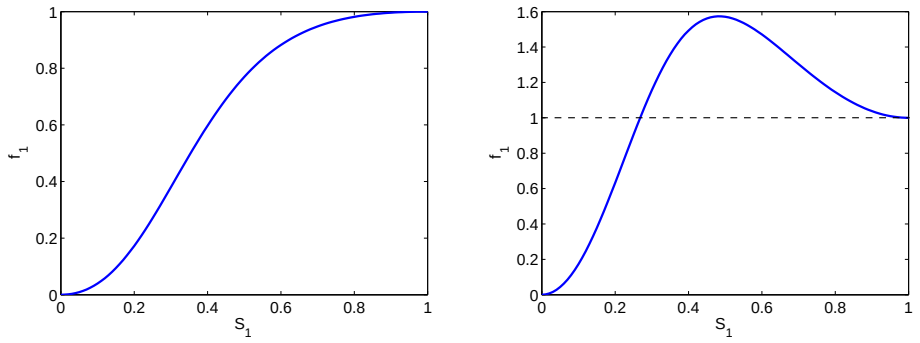


FIG. 1.1. *Flux functions for 1D incompressible two-phase flow: (a) Cocurrent flow (no buoyancy effects), (b) Countercurrent flow due to gravity.*

and total velocity. This is especially true for the fully-implicit method because the total velocity at time t^{n+1} is usually unknown.

Notice that in (1.9) it is possible for each phase to have a different upstream direction when the densities ρ_j are different, i.e., when buoyancy forces are significant; this is known as *countercurrent flow* in reservoir engineering literature. In one-dimensional porous media, countercurrent flow manifests itself through the presence of sonic points in the flux function v_j; thus, the flux function for a countercurrent flow problem would typically look like the one shown in Figure 1.1(b), whereas without countercurrent flow it would look more like Figure 1.1(a). A detailed treatment of phase-based upstreaming is given in [3], in which the authors show that, when explicit time-stepping is used on a two-phase flow problem, phase-based upstreaming leads to a monotone difference scheme, as long as the appropriate CFL condition is satisfied. This in turn implies that the solution of the explicit schemes converge to the entropy solution of the two-phase equations

$$\frac{\partial S_1}{\partial t} + \frac{\partial f_1}{\partial x} = 0, \tag{1.10}$$

$$f_1(x, S_1) = v_1(x, S_1) = \frac{\lambda_1(S_1)}{\lambda_1(S_1) + \lambda_2(S_1)} \left[v_T + K(x) g \lambda_2(S_1)(\rho_1 - \rho_2)\right] \tag{1.11}$$

as $\Delta t, \Delta x \to 0$ while satisfying the CFL condition. The goal of this paper is to extend this result for the fully-implicit case. This leads us to study the more general problem of implicit monotone schemes, of which the multiphase flow problem is a special case.

The use of implicit time stepping leads to a (typically large) system of nonlinear algebraic equations that must be solved for each time step. Moreover, the residual functions are generally non-differentiable due to upstreaming criteria of the form (1.9); thus, the existence of a unique solution to these systems of equations is not immediately obvious. For implicit monotone schemes for 1D scalar conservation laws, Lucier [9] showed that a unique solution to the discrete problem exists whenever the initial data is bounded and has bounded total variation. The proof of existence, which relies heavily on Crandall-Liggett theory [5], proceeds along the following lines (see [6, Ch. 3] for more details). First, one shows that the residual function R for the numerical scheme defines an m-accretive operator in the L^1 norm. Then by the

Crandall-Ligett theorem, the ODE

$$\frac{du}{dt} = -Ru, \quad u(0) = x \tag{1.12}$$

has a unique solution for $t \in [0, \infty)$ for any initial point x. Let $u(t; x)$ denote the solution of (1.12) with starting point x. Then one shows that the *Poincaré operator* P_ω, which maps the point x to the point $u(\omega, x)$, is strictly contractive. Then by Banach's fixed point theorem, P_ω has a unique fixed point x_0. One then proceeds to prove that $u(t; x_0) = x_0$ for all $0 \le t \le \omega$; thus, $du/dt = 0$, which implies $Rx_0 = 0$.

While this argument does prove the existence and uniqueness of a solution to the discretized problem, the proof does not suggest a practical algorithm for finding the solution. In section 3, we present an alternate constructive proof of existence by showing that the classical Gauss-Seidel and Jacobi iterations converge for this class of problems. In fact, we show that the iterative methods converge whenever the initial data for the discrete problem is bounded, so the implicit scheme is well-defined even when the initial data does not have bounded variation in $\mathbb{R}$. The well-definedness of the numerical scheme, together with the total variation diminishing (TVD) property and the existence of a discrete entropy inequality, imply that the numerical scheme converges to the entropy solution as the mesh is refined (i.e., as $\Delta x \to 0$). This result holds for any mesh ratio $\lambda = \Delta t/\Delta x$ (i.e., for any Courant number).

The remainder of this paper is organized as follows. Section 2 states the necessary assumptions of our framework. Section 3 contains the main result of this paper, which asserts that the nonlinear Gauss-Seidel and Jacobi processes converge when applied to monotone implicit schemes. This leads to a constructive proof of well-definedness of the numerical scheme, from which convergence of the numerical scheme under grid refinement follows [14]. In section 4, we derive the numerical flux functions for an implicit monotone scheme that is equivalent to the coupled problem (1.1) in the one-dimensional case, which would allow us to establish the well-definedness and convergence of the coupled problem. We also discuss the applicability of the above analysis for multidimensional problems. Finally, we discuss the convergence behavior and accuracy issues for implicit schemes with phase-based upstreaming in section 5.

2. Implicit monotone schemes. We consider the following fully-implicit, finite volume discretization

$$\phi_i(u_i^{n+1} - u_i^n) + \lambda(F_{i+1/2}^{n+1} - F_{i-1/2}^{n+1}) = 0, \qquad \lambda = \Delta t/\Delta x, \;\; i \in \mathbb{Z}, \tag{2.1}$$

where $F_{i+1/2}$ denotes the numerical flux across the interface between cells i and $i+1$. The above scheme approximates the 1D nonlinear conservation law

$$\phi(x)u_t + f(x,u)_x = 0, \qquad (x,t) \in \mathbb{R} \times \mathbb{R}^+, \tag{2.2}$$

$$u(x,0) = u^0(x), \tag{2.3}$$

which generalizes the problem (1.10), (1.11) to the variable porosity and permeability case. For simplicity, we assume a three-point scheme $F_{i+1/2}^{n+1} = F_{i+1/2}(u_i^{n+1}, u_{i+1}^{n+1})$; thus, the implicit stencil at cell i contains the value at cell i at time t^n, as well as the values at cells $i-1$, i and $i+1$ at the *future* time t^{n+1}. We also ignore the effects of boundary conditions by limiting ourselves to Cauchy problems for the moment; in section 4 we explain how certain boundary conditions can be incorporated. Assume that f and F are both locally Lipschitz continous (but not necessarily differentiable),

and that the numerical flux function $F_{i+1/2}$ is *consistent* with f in the sense that $F_{i+1/2}(u,u) = f(x_{i+1/2},u)$. Given we are interested in handling flux functions of the type shown in Figure 1.1(b), we cannot assume that the flux function $f(x,u)$ is monotonic in u, so sonic points may be present. However, the numerical flux function is assumed to satisfy the following condition, which also appears in [14] and is central to all our arguments in this paper:

ASSUMPTION 1 (Monotonic fluxes). *For all $i \in \mathbb{Z}$, the numerical flux function $F_{i+1/2}$ is non-decreasing in the first argument and non-increasing in the second argument, i.e. for any w, we have $F_{i+1/2}(u,w) \leq F_{i+1/2}(v,w)$ and $F_{i+1/2}(w,u) \geq F_{i+1/2}(w,v)$ whenever $u \leq v$.*

For two-phase flow problems in one dimension, the numerical flux functions obtained by phase-based upstreaming have the form

$$F_{i+1/2}(S_{1,i}, S_{1,i+1}) = \frac{\lambda_1(S_{1,i})}{\lambda_1(S_{1,i}) + \lambda_2(S_1^*)}[v_T + K\lambda_2 g(\rho_1 - \rho_2)], \tag{2.4}$$

where $g(\rho_1 - \rho_2) \geq 0$ and

$$S_1^* = \begin{cases} S_{1,i}, & \text{if } v_T - K\lambda_1(S_i)g(\rho_1-\rho_2) \geq 0, \\ S_{1,i+1} & \text{otherwise.} \end{cases} \tag{2.5}$$

The detailed derivation is deferred to section 4. The following theorem, which summarizes several results by Brenier and Jaffré [3] in the case of two-phase flow, guarantees that $F_{i+1/2}(S_i, S_{i+1})$ has the required properties.

THEOREM 2.1. *Assume that the mobility of phase j is increasing with S_j and decreasing with the saturation of the other phase for $j = 1, 2$ (e.g., water and oil). Then the numerical fluxes obtained from phase-based upstreaming defined by* (2.4), (2.5) *are consistent, Lipchitz continuous and satisfy Assumption 1.*

The hypothesis on phase mobilities is physically realistic [2]. Assumption 1 ensures that (2.1) is an implicit monotone scheme, in the sense that the resulting residual function defines an m-accretive operator in $\ell^1(\mathbb{Z})$ (see [7] for a proof). We do not use m-accretivity directly in this work. Instead, we show that Assumption 1 also implies that the residual function is an M-function in the sense of Rheinboldt [11]; this is generally not equivalent to m-accretivity [8], but nontheless allows us to prove existence and uniqueness of solutions. Once we establish that the implicit scheme is well-defined, we can apply the following theorem of Sanders [14, Theorem II] to conclude that the solutions of (2.1) converges to the unique entropy solution of (2.2) when f does not depend explicitly on x, i.e., when $f = f(u)$. Note that Assumption 1 implies that (2.1) is an E-scheme [10], so it is at most first-order accurate.

THEOREM 2.2 (Sanders). *Let $\mathbb{R} = \cup \mathcal{I}_i^n$ with $\mathcal{I}_i^n = [x_i, x_{i+1}) \times [t^n, t^{n+1})$ and $\delta = \sup_{i,n} |x_{i+1} - x_i| + |t^{n+1} - t^n|$. Suppose the conservation law* (2.2), (2.3) *with $f = f(u)$ is discretized using* (2.1), *where $F_{i+1/2}(u_i, u_{i+1}) \equiv F(u_i, u_{i+1})$ for all i. Assume that F is locally Lipschitz continuous, consistent and monotonic. Define the step function v_δ such that $v_\delta(x,t) = u_i^n$ when $(x,t) \in \mathcal{I}_i^n$, and suppose the initial data* (2.3) *is discretized via the averaging operator T_δ,*

$$T_\delta(u_0)(x) = \frac{1}{|x_{i+1} - x_i|} \int_{x_i}^{x_{i+1}} u_0(\xi) d\xi$$

when $x \in [x_i, x_{i+1})$. Then $v_\delta(x,t)$ converges in $L^\infty(L^1(\mathbb{R}); [0,T])$ to the unique entropy solution of (2.2), (2.3) *as δ tends to zero.*

3. Existence and uniqueness of solutions for the discretized problem. In this section, we show that the nonlinear Jacobi and Gauss-Seidel processes, when applied to the infinite system of nonlinear equations (2.1), converge to a unique bounded solution. Thus, we obtain an alternate constructive proof of the well-definedness of implicit monotone schemes. In addition, we show that Jacobi and Gauss-Seidel converge for any starting point that is bounded by the initial data, leading to a practical algorithm for computing the solution. Finally, we show how to extend the analysis to deal with problems with spatially-varying coefficients, as well as problems for which the flux function $f(u)$ is only defined over a finite interval rather than all of $\mathbb{R}$. An extension to multidimensional problems is considered in section 4.2. (*Note:* Throughout this section, x and t are generic variables and do not denote space or time.)

3.1. Nonlinear Jacobi and Gauss-Seidel processes. Suppose we want to solve a nonlinear system of algebraic equations $R(x) = 0$ for $x \in \ell^\infty(\mathbb{N})$, where $R = (r_1, r_2, \ldots)^T : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$. Then we can consider the *nonlinear Gauss-Seidel process*:

$$\begin{aligned} &\text{Solve} \quad r_i(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^*, x_{i+1}^k, x_{i+2}^k, \ldots) = 0 \text{ for } x_i^*, \\ &\text{Set} \qquad x_i^{k+1} = x_i^*, \quad i = 1, 2, \ldots, \quad k = 0, 1, 2, \ldots, \end{aligned} \tag{3.1}$$

as well as the *nonlinear Jacobi process*:

$$\begin{aligned} &\text{Solve} \quad r_i(x_1^k, \ldots, x_{i-1}^k, x_i^*, x_{i+1}^k, x_{i+2}^k, \ldots) = 0 \text{ for } x_i^*, \\ &\text{Set} \qquad x_i^{k+1} = x_i^*, \quad i = 1, 2, \ldots, \quad k = 0, 1, 2, \ldots. \end{aligned} \tag{3.2}$$

The only difference between (3.1)/(3.2) and classical Gauss-Seidel/Jacobi processes is that each Gauss-Seidel/Jacobi "sweep" now involves infinitely many variables and equations. To ensure that the iterations (3.1)/(3.2) make sense, we need the following assumptions on the residual function R:

ASSUMPTION 2 (Preservation of bounded sets). *$R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ is a mapping between bounded sequences for which there exists an increasing function $\zeta : [0, \infty) \to [0, \infty)$ such that $\|x\|_\infty \le B$ implies $\|R(x)\| \le \zeta(B)$.*

ASSUMPTION 3 (Finite number of dependencies). *For each i, the residual function $r_i(x_1, x_2, \ldots)$ is non-constant with respect to at most a finite number of x_j.*

In other words, the residual functions must come from a compact stencil and must preserve boundedness. Assumption 3 ensures that the iterative processes are well-defined, since it guarantees that for any given $i, k \in \mathbb{N}$, the value of x_i^{k+1} can be calculated using a finite number of univariate solves. It also guarantees that whenever R is continuous and (3.1)/(3.2) converges, the limit point x^* must satisfies $R(x^*) = 0$. Assumption 2 then allows us to extend Rheinboldt's analysis [11] to the infinite-dimensional case and prove that (3.1)/(3.2) converges to a unique bounded solution when R is an M-function.

Remark. Even though the nonlinear iteration (3.1)/(3.2) are well-defined in theory, a practical implementation would require that we either solve a finite-dimensional problem by supplying appropriate boundary conditions, or that we use delayed evaluation [1] to compute only those values of x_i^k that are needed.

3.2. M-function theory. M-functions are essentially generalizations of M-matrices in linear algebra. In the linear setting, it is well known [13] that the Gauss-Seidel method applied to $Ax = b$ converges for any right-hand side b and starting point x_0 if A is a non-singular M-matrix. M-functions have similar properties with

respect to the nonlinear Gauss-Seidel process, which is the subject of investigation in [11]. Here we provide extensions to the relevant definitions and theorems in [11] that would allow us to prove the existence and uniqueness of bounded solutions to (2.1).

For the remainder of the section, the natural partial ordering on $\ell^\infty(\mathbb{N})$ is written as $x \le y$, i.e., $x \le y$ iff $x_i \le y_i$ for all $i \in \mathbb{N}$. We denote by e^i the unit basis vectors with the i-th component one and all others zero. The following definitions are essentially identical to those in [11], except the domain of definition has been changed from $\mathbb{R}^n$ to $\ell^\infty(\mathbb{N})$ to handle vectors of infinite length:

DEFINITION 3.1. *Let $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$.*

1. *R is* isotone *(or* antitone*) if, for all $x, y \in \ell^\infty(\mathbb{N})$, $x \le y$ implies $R(x) \le R(y)$ (or $R(x) \ge R(y)$). It is* strictly isotone *(or antitone) if $x < y$ implies $R(x) < R(y)$ (or $R(x) > R(y)$).*
2. *R is* inverse isotone *if, for all $x, y \in \ell^\infty(\mathbb{N})$, $R(x) \le R(y)$ implies $x \le y$.*
3. *R is (strictly)* diagonally isotone *if, for all $x \in \ell^\infty(\mathbb{N})$, the functions*

$$\rho_{ii} : \mathbb{R} \to \mathbb{R}, \quad \rho_{ii}(t) = r_i(x + te^i), \quad i = 1, 2, \ldots \tag{3.3}$$

   are (strictly) isotone.
4. *R is* off-diagonally antitone *if, for any $x \in \ell^\infty(\mathbb{N})$ the functions*

$$\rho_{ij} : \mathbb{R} \to \mathbb{R}, \quad \rho_{ij}(t) = r_i(x + te^j), \quad i \ne j, \quad i, j = 1, 2, \ldots \tag{3.4}$$

   are antitone.
5. *R is an M-*function *if R is inverse isotone and off-diagonally antitone.*

One characterization of M-functions is given by the following theorem:

THEOREM 3.2. *Let $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ be off-diagonally antitone and satisfy Assumption 2 and 3. Then R is an M-function if, for each $B > 0$, there exists a positive sequence $\{w_i^B\}$ such that:*

1. *$\sum_{i=1}^\infty w_i^B < \infty$,*
2. *for any $\|x\|_\infty < B$, the function $Q(t) = (q_1(t), q_2(t), \ldots)$ defined by*

$$q_i(t) = \sum_{j=1}^{\infty} w_j^B r_j(x + te^i)$$

is strictly isotone over the interval $t \in (t_{min}, t_{max})$, where $t_{min} = -B - \inf_i x_i$ and $t_{max} = B - \sup_i x_i$.

Proof. The proof is based on [11, Theorem 5.1], suitably modified to handle the infinite-dimensional case. Suppose $R(x) \le R(y)$ for some $x, y \in \ell^\infty(\mathbb{N})$. Define the sets

$$N^- = \{i \in \mathbb{N} \mid y_i < x_i\}; \quad N^+ = \{i \in \mathbb{N} \mid y_i \ge x_i\}.$$

Suppose N^- is non-empty. For each $i \in N^-$, let $\gamma_i = (x_i - y_i)e^i$. We consider two cases:

1. If $|N^-| < \infty$, let $i_1 < i_2 < \cdots < i_m$ be the elements of N^-, and define

$$z^0 = y, \quad z^1 = y + \gamma_{i_1}, \quad \ldots, \quad z^m = y + \gamma_{i_1} + \cdots + \gamma_{i_m},$$

   and let $z^k = z^m = z$ for all $k > m$.
2. If $|N^-| = \infty$, let $i_1 < i_2 < \cdots$ be the elements of N^-, and define

$$z^0 = y, \quad z^1 = y + \gamma_{i_1}, \quad \ldots, \quad z^k = y + \gamma_{i_1} + \cdots + \gamma_{i_k}, \quad \ldots$$

   and let $z = \{z_i\}$ be such that $z_i = \max\{x_i, y_i\}$.

Define $R^k := R(z^k)$ and $R^\infty = R(z)$. In either case, we have the following properties:

1. $\|z^k\|_\infty < B$ and $\|z\|_\infty < B$, where $B = \max\{\|x\|_\infty, \|y\|_\infty\}$. Hence, by Assumption 2, $\|R^k\|_\infty \le \zeta(B)$ for all k (similarly for R^∞).
2. For each i, $z_i^k = z_i$ for large enough k, so by Assumption 3, $R_j^k \to R_j^\infty$ *pointwise* for each j.

Thus, the sequence $R^k = (R_j^k)_{j=1}^\infty$ is dominated by $G = (\zeta(B), \zeta(B), \ldots)$ for each $k \in \mathbb{N}$. Moreover $\sum_{j=1}^\infty w_j^B G_j < \infty$, so by the dominated convergence theorem [12],

$$\sum_{j=1}^\infty w_j^B R_j^k \to \sum_{j=1}^\infty w_j^B R_j^\infty \quad \text{as } k \to \infty.$$

By the strict isotonicity of Q, we have

$$\sum_{j=1}^\infty w_j^B R_j^0 \le \sum_{j=1}^\infty w_j^B R_j^1 \le \cdots$$

with at least one strict inequality (since N^- is non-empty). Thus, we must have

$$\sum_{j=1}^\infty w_j^B r_j(y) = \sum_{j=1}^\infty w_j^B R_j^0 < \sum_{j=1}^\infty w_j^B R_j^\infty = \sum_{j=1}^\infty w_j^B r_j(z). \tag{3.5}$$

Now split the last sum into two parts

$$\sum_{j=1}^\infty w_j^B r_j(z) = \sum_{j \in N^-} w_j^B r_j(z) + \sum_{j \in N^+} w_j^B r_j(z), \tag{3.6}$$

where the summation over N^+ may be empty. Then by off-diagonal antitonicity of R (and invoking the dominated convergence theorem whenever necessary), we can show similarly that

$$\sum_{j \in N^-} w_j^B r_j(z) \le \sum_{j \in N^-} w_j^B r_j(x), \qquad \sum_{j \in N^+} w_j^B r_j(z) \le \sum_{j \in N^+} w_j^B r_j(y), \tag{3.7}$$

using the fact that $z - x$ and $z - y$ vanish on N^- and N^+ respectively. Combining equations (3.5)–(3.7) gives

$$\sum_{j=1}^\infty w_j^B r_j(y) < \sum_{j \in N^-} w_j^B r_j(x) + \sum_{j \in N^+} w_j^B r_j(y), \tag{3.8}$$

which implies $\sum_{j \in N^-} w_j^B r_j(y) < \sum_{j \in N^-} w_j^B r_j(x)$. Thus, we must have $r_j(y) < r_j(x)$ for some $j \in N^-$, which contradicts the hypothesis $R(x) \le R(y)$. Hence N^- must be empty, so $x \le y$. ∎

COROLLARY 3.3. *Let R satisfy the hypotheses of Theorem 3.2. Let $z \in \ell^\infty(\mathbb{N})$. Then there is at most one bounded solution to the equation $R(x) = z$.*

Remark. In the context of discretized PDEs one normally assumes tacitly that the solution of interest must be bounded; this can be regarded as a boundary condition "at infinity". However, since such boundary conditions are not explicitly stated in the definition of M-functions, one must be careful to exclude any parasitic unbounded solutions that may arise. In fact, the solution is not necessarily unique if we allow

unbounded solutions. Consider the linear function $R = (r_1, r_2, \ldots)$ defined by $r_i(x) = x_i - \alpha x_{i+1}$ for $|\alpha| < 1$. Then for any $\|x\|_\infty < \infty$, we have $\|R(x)\|_\infty \le (1+\alpha)\|x\|_\infty$, so that Assumption 2 is satisfied. Assumption 3 (finitely many dependencies) is also satisfied because each r_i is only non-constant with respect to two components of x. Finally, if we let $w_j^B = \beta^j$ for any $|\alpha| < \beta < 1$, then $\sum_j \beta^j < \infty$ and

$$q_i(t) = \sum_{j=1}^{\infty} \beta^j \left[x_j + t\delta_{ij} - \alpha(x_{j+1} + t\delta_{i,j+1})\right] = (\beta-\alpha)\beta^{i-1}t + \beta x_1 + (\beta-\alpha)\sum_{j=2}^{\infty} \beta^{j-1} x_j,$$

so $q_i(t)$ is well-defined and is strictly increasing with respect to t whenever $\|x\|_\infty < \infty$. So the hypotheses of Theorem 3.2 are satisfied, and hence $x = 0$ is the only bounded solution of $R(x) = 0$. However, unbounded solutions of the form $y = \{K\alpha^{-i}\}$, $K \neq 0$ also satisfy $R(y) = 0$, so the theorem does not preclude these possibilities.

3.3. Convergence of nonlinear Jacobi and Gauss-Seidel. It turns out the hypotheses of Theorem 3.2 are enough to ensure convergence of nonlinear Jacobi and Gauss-Seidel for certain starting points, which will be described below. The following result is essentially Theorem 3.1 in [11], with modified hypotheses to accommodate ℓ^∞-bounded vectors with infinitely many components. The proof in [11] goes through verbatim, but is reproduced here because similar arguments also appear in the proof of Theorem 3.7. Note that by Assumption 3, each r_i depends on only finitely many arguments, so the standard arguments on limits, continuity and antitonicity hold without additional complications when they are used on individual components of R.

THEOREM 3.4 (Rheinboldt). *Let $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ satisfy the hypotheses of Theorem 3.2. Suppose for some $z \in \ell^\infty(\mathbb{N})$ there exist $x^0, y^0 \in \ell^\infty(\mathbb{N})$ such that*

$$x^0 \le y^0, \qquad R(x^0) \le z \le R(y^0).$$

Then the nonlinear Gauss-Seidel and Jacobi iterates $\{y^k\}$ and $\{x^k\}$, given by (3.1) *and* (3.2)*, and starting from y^0 and x^0, respectively, are uniquely defined and satisfy*

$$x^0 \le x^k \le x^{k+1} \le y^{k+1} \le y^k \le y^0, \qquad R(x^k) \le z \le R(y^k) \tag{3.9}$$

for $k = 0, 1, \ldots$. In addition, the pointwise limits

$$\lim_{k\to\infty} x^k = \lim_{k\to\infty} y^k = x^* \tag{3.10}$$

exist, and $R(x^*) = z$.

First we need the following lemma (which is part of Theorem 2.10 in [11]):

LEMMA 3.5. *Let $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ be an M-function. Then R is strictly diagonally isotone.*

Proof. Suppose that for some $x \in \ell^\infty(\mathbb{N})$, $s, t \in \mathbb{R}$, $s > t$ and index i we have $r_i(x + se^i) \le r_i(x + te^i)$. Off-diagonal antitonicity then implies that

$$r_j(x + se^i) \le r_j(x + te^i), \qquad j \neq i,$$

or, altogether, that $R(x + se^i) \le R(x + te^i)$. By inverse isotonicity, this leads to the contradiction $s \le t$, which shows that R must be strictly diagonally isotone. ∎

Proof. (Theorem 3.4) We only present the proof for convergence of Gauss-Seidel; the proof for Jacobi is similar. We proceed by induction and suppose that for some $k \ge 0$ and $i \ge 1$

$$x^0 \le x^k \le y^k \le y^0, \qquad R(x^k) \le z \le R(y^k), \tag{3.11a}$$

$$x_j^k \le x_j^{k+1} \le y_j^{k+1} \le y_j^k, \qquad j = 1, \ldots, i-1, \tag{3.11b}$$

where for $i = 1$ the relation (3.11b) is vacuous. Clearly, (3.11) is valid for $k = 0$ and $i = 1$. Define the functions

$$\alpha(s) = r_i(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, s, x_{i+1}^k, x_{i+2}^k, \ldots),$$
$$\beta(s) = r_i(y_1^{k+1}, \ldots, y_{i-1}^{k+1}, s, y_{i+1}^k, x_{i+2}^k, \ldots)$$

for $s \in [x_i^0, y_i^0]$. Then (3.11) and the off-diagonal antitonicity of R yield

$$\beta(s) \leq \alpha(s), \qquad s \in [x_i^0, y_i^0], \tag{3.12}$$
$$\beta(x_i^k) \leq \alpha(x_i^k) \leq r_i(x^k) \leq z_i \leq r_i(y^k) \leq \beta(y_i^k) \leq \alpha(y_i^k). \tag{3.13}$$

Since R is strictly diagonally isotone, α and β are both continuous and strictly increasing, so (3.13) implies the existence of unique $\hat{y}_i^k$ and $\hat{x}_i^k$ for which

$$\beta(\hat{y}_i^k) = z_i = \alpha(\hat{x}_i^k), \qquad x_i^k \leq \hat{x}_i^k \leq \hat{y}_i^k \leq y_i^k,$$

where the relation $\hat{x}_i^k \leq \hat{y}_i^k$ is a consequence of (3.12). But $x_i^{k+1} = \hat{x}_i^k$ and $y_i^{k+1} = \hat{y}_i^k$ by definition, so we have proved (3.11b) for $j = 1, \ldots, i$. Hence by induction (3.11b) holds for all $i \in \mathbb{N}$, and hence $x^k \leq x^{k+1} \leq y^{k+1} \leq y^k$. From this it follows again from off-diagonal antitonicity that

$$r_i(y^{k+1}) \geq r_i(y_1^{k+1}, \ldots, y_i^{k+1}, y_{i+1}^k, y_{i+2}^k \ldots) = z_i$$

and similarly that

$$r_i(x^{k+1}) \leq r_i(x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, x_{i+2}^k \ldots) = z_i.$$

This completes the induction on k and hence the proof of (3.9). Applying the monotone convergence theorem for sequences, we conclude that the *pointwise* limits

$$\lim_{k\to\infty} x_j^k = x_j^* \leq y_j^* = \lim_{k\to\infty} y_j^k$$

exist for each j, which allows us to define $x^* = \{x_j^*\}$ and $y^* = \{y_j^*\}$. Since each r_i is continuous and depends on only finitely many arguments, the definition of the Gauss-Seidel process then implies $r_i(x^*) = r_i(y^*) = z_i$ for each i, and hence $R(x^*) = R(y^*) = z$. Since both x^* and y^* are bounded, Corollary 3.3 implies that they are equal, completing the proof. □

3.4. Well-definedness of implicit monotone schemes. Using the theory in the last two sections, we can now prove that implicit monotone schemes (i.e, implicit schemes whose flux functions satisfy Assumption 1) are well-defined for bounded initial conditions. What we need to show is that the residual functions satisfy the hypotheses of Theorem 3.4. In the interest of clarity, in this section we only show convergence of the iterative schemes for problems whose coefficients do not vary in space (i.e., corresponding to the conservation law $u_t + f(u)_x = 0$, discretized on a uniform spatial grid). In the next section, we state the additional assumptions on ϕ_i and $F_{i+1/2}$ that are required for the spatially-varying case.

THEOREM 3.6. *Consider the numerical scheme* (2.1) *with*

$$F_{i+1/2}^{n+1} = F(u_i^{n+1}, u_{i+1}^{n+1}),$$

where $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is locally Lipschitz continuous and satisfies Assumption 1, i.e., non-decreasing in the first argument and non-increasing in the second. Assume that

the initial condition $\{u_i^0\}_{i=-\infty}^{\infty}$ *is bounded. Then* (2.1) *has a unique bounded solution* $\{u_i^{n+1}\}$ *for* $n = 0, 1, 2, \ldots$. *Moreover, this bounded solution satisfies the estimate*

$$\inf_{j\in\mathbb{Z}} u_j^n \le u_i^{n+1} \le \sup_{j\in\mathbb{Z}} u_j^n, \qquad \forall i \in \mathbb{Z}. \tag{3.14}$$

Proof. First, we need to define an ordering for the Gauss-Seidel sweeps, i.e., to permute the equations and variables so that the spatial indices go from 1 to ∞ rather than from $-\infty$ to ∞. After that, it suffices to check that all the hypotheses of Theorem 3.4 are satisfied for this ordering.

1. For $j = 1, 2, \ldots$, define $\sigma(j) = (-1)^j \lfloor j/2 \rfloor$, i.e. σ maps $\{1, 2, 3, 4, 5, \ldots\}$ to $\{0, 1, -1, 2, -2, \ldots\}$. Let τ be the inverse map, such that $\tau(\sigma(j)) = j$. Define $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ to be the reordered (and rescaled) set of residual equations, i.e.,

$$r_j(v) = \frac{v_j - u^n_{\sigma(j)}}{\lambda} + F(v_j, v_{\tau(\sigma(j)+1)}) - F(v_{\tau(\sigma(j)-1)}, v_j), \tag{3.15}$$

where $v_j = u^{n+1}_{\sigma(j)}$.

2. Since F is locally Lipschitz continuous, it is Lipschitz continuous over any compact set, so for any $B > 0$ there exists K_B (which can be chosen to be increasing with B) such that for any $(x, y) \in [-B, B] \times [-B, B]$,

$$|F(x, y) - F(0, 0)| \le K_B(|x| + |y|) \le 2K_B \cdot B.$$

Thus, for any $\|v\|_\infty \le B$, we have $|r_j(v)| \le \zeta(B)$ for all j, where

$$\zeta(B) = \Big(\frac{1}{\lambda} + 4K_B\Big)B + \frac{1}{\lambda}\|u^n\|_\infty.$$

Hence Assumption 2 is satisfied. Moreover, since each r_j depends only on v_j, $v_{\tau(\sigma(j)-1)}$ and $v_{\tau(\sigma(j)+1)}$, Assumption 3 (finite number of dependencies) is also satisfied.

3. By Assumption 1 (monotonic fluxes), F is clearly off-diagonally antitone. To satisfy the remaining hypotheses of Theorem 3.2, let $\{w_j^B\}$ take the form $w_j^B = \beta^{|\sigma(j)|}$ for some $0 < \beta < 1$, so that $\sum_{j=1}^\infty w_j^B < \infty$. An easy calculation shows that

$$q_i(t) := \sum_{j=1}^\infty w_j^B r_j(v + te^i) = \tilde{q}_i(t) + \sum_{j=1}^\infty w_j^B r_j(v),$$

where

$$\begin{aligned}\tilde{q}_i(t) = w_i^B t/\lambda &+ (w_i^B - w^B_{\tau(\sigma(i)+1)})\big[F(v_i + t, v_{\tau(\sigma(i)+1)}) - F(v_i, v_{\tau(\sigma(i)+1)})\big] \\ &+ (w^B_{\tau(\sigma(i)-1)} - w_i^B)\big[F(v_{\tau(\sigma(i)-1)}, v_i + t) - F(v_{\tau(\sigma(i)-1)}, v_i)\big].\end{aligned}$$

By the definition of w_i^B, we see that

$$|w_i^B - w^B_{\tau(\sigma(i)\pm1)}| \le \beta^{|\sigma(i)|-1}(1 - \beta),$$

which, when combined with the local Lipschitz continuity of F, gives

$$\beta^{|\sigma(i)|-1}\big[\beta t/\lambda - 2(1-\beta)K_B|t|\big] \le \tilde{q}_i(t) \le \beta^{|\sigma(i)|-1}\big[\beta t/\lambda + 2(1-\beta)K_B|t|\big].$$

Hence, $\tilde{q}_i(t)$ is strictly isotone whenever $\beta/\lambda > 2(1-\beta)K_B$, so picking

$$\frac{2\lambda K_B}{1 + 2\lambda K_B} < \beta < 1 \tag{3.16}$$

ensures isotonicity for $\tilde{q}_i(t)$ (and hence $q_i(t)$) for all i, as required in Theorem 3.2. (Note that the choice of β depends on B.)

4. We need to choose starting points x^0 and y^0 that satisfy the requirements of Theorem 3.4. Let x^0 and y^0 both be constant sequences with

$$x_i^0 = \inf_{j\in\mathbb{Z}} u_j^n, \qquad y_i^0 = \sup_{j\in\mathbb{Z}} u_j, \qquad \forall i \in \mathbb{N}.$$

Then clearly $x^0 \le y^0$, and for all $i \in \mathbb{N}$,

$$r_i(x^0) = \frac{1}{\lambda}\left(x_i^0 - u_{\sigma(i)}^n\right) = \frac{1}{\lambda}\left(\inf_{j\in\mathbb{Z}} u_j^n - u_{\sigma(i)}^n\right) \le 0,$$
$$r_i(y^0) = \frac{1}{\lambda}\left(y_i^0 - u_{\sigma(i)}^n\right) = \frac{1}{\lambda}\left(\sup_{j\in\mathbb{Z}} u_j^n - u_{\sigma(i)}^n\right) \ge 0,$$

so $R(x^0) \le 0 \le R(y^0)$. Thus, by Theorem 3.4, the nonlinear Gauss-Seidel iterates $\{y^k\}$ and $\{x^k\}$ both converge (pointwise) to the unique solution x^* with $R(x^*) = 0$; hence, a unique solution to (2.1) exists, i.e., $u_i^{n+1} = x^*_{\tau(i)}$. Moreover, we know that $x^0 \le x^* \le y^0$, which immediately implies (3.14). ∎

Remark. The initial condition $\{u_i^0\}_{i=-\infty}^{\infty}$ is not assumed to be in ℓ^1 nor in BV, so this result is somewhat more general than results that use Crandall-Liggett theory.

Remark. The definition of an M-function is invariant under symmetric permutations, i.e., $R(x)$ is an M-function if and only if $\sigma R(\sigma x)$ is also an M-function for any permutation $\sigma : \mathbb{N} \to \mathbb{N}$. Thus, the Gauss-Seidel process must converge regardless of the way the ordering is chosen in step 1 of the proof. However, the *rate* of convergence is sensitive to the ordering [8].

In fact, one can show that the nonlinear Jacobi and Gauss-Seidel processes converge for any starting point $\{z_i^{(0)}\}$ that is bounded by the initial data $\{u_i^n\}$. (In the sequel, superscripts in brackets indicate iterates within the Gauss-Seidel process, and superscripts without brackets indicate the time level in the numerical scheme.)

THEOREM 3.7. *Assume the hypotheses of Theorem 3.6. Suppose the initial guess* $\{z_i^{(0)}\}$ *satisfies*

$$\inf_{j\in\mathbb{Z}} u_j^n \le z_i^{(0)} \le \sup_{j\in\mathbb{Z}} u_j^n \tag{3.17}$$

for all $i \in \mathbb{Z}$. *Then the nonlinear Jacobi and Gauss-Seidel processes* (3.1) *and* (3.2) *are well-defined and converge to the unique bounded solution of* (2.1).

Proof. Again we only show convergence for the Gauss-Seidel process, since the proof for Jacobi is similar. Denote $\underline{u} = \inf_{j\in\mathbb{Z}} u_j^n$ and $\overline{u} = \sup_{j\in\mathbb{Z}} u_j^n$. First, we show that the Gauss-Seidel iterates are well-defined and that $\underline{u} \le z_j^{(k)} \le \overline{u}$ for all j, k. At each step we need to solve

$$r_j(z_j^*) = \frac{1}{\lambda}\left(z_j^* - u_j^n\right) + F(z_j^*, z_{j+1}) - F(z_{j-1}, z_j^*) = 0, \tag{3.18}$$

where $z_{j\pm1} = z_{j\pm1}^{(k)}$ or $z_{j\pm1}^{(k+1)}$ depending on the ordering of the Gauss-Seidel sweep, which by induction must lie between $\underline{u}$ and $\overline{u}$. But

$$\begin{aligned} r_j(\underline{u}) &= \frac{1}{\lambda}\left(\underline{u} - u_j^n\right) + F(\underline{u}, z_{j+1}) - F(z_{j-1}, \underline{u}) \\ &\le 0 + F(\underline{u}, \underline{u}) - F(\underline{u}, \underline{u}) = 0, \end{aligned}$$

where the inequality follows from Assumption 1. Similarly one obtains $r_j(\overline{u}) \geq 0$, so by continuity of F (and hence r_j) there must exist a solution z_j^* to (3.18), which by Lemma 3.5 must be unique. Hence, by induction, the Gauss-Seidel iterates are well-defined and are bounded above and below by $\overline{u}$ and $\underline{u}$ respectively.

Now consider the Gauss-Seidel iterates $\{x_j^{(k)}\}$ and $\{y_j^{(k)}\}$ with initial guess $x_j^{(0)} = \underline{u}$ and $y_j^{(0)} = \overline{u}$ for all j. By Theorem 3.6 these iterates converge pointwise to the same solution $\{x_j^*\}$. We show inductively that $x^{(k)} \leq z^{(k)} \leq y^{(k)}$ for all k, which would imply that $z_j^{(k)} \to x_j^*$ pointwise. Using the same reordering as in Theorem 3.6, assume that for some $k \geq 0$ and $i \geq 1$ we have

$$y^{(k)} \geq z^{(k)} \geq x^{(k)}, \qquad y_j^{(k+1)} \geq z_j^{(k+1)} \geq x_j^{(k+1)}, \qquad j = 1, \ldots, i-1,$$

which is valid for $k = 0$ and $i = 1$. Then by the same boundedness and antitonicity arguments as in Theorem 3.4, we have

$$\begin{aligned} r_i(y_1^{(k+1)}, \ldots, y_{i-1}^{(k+1)}, y_i^{(k+1)}, y_{i+1}^{(k)}, \ldots) = 0 &= r_i(z_1^{(k+1)}, \ldots, z_{i-1}^{(k+1)}, z_i^{(k+1)}, z_{i+1}^{(k)}, \ldots) \\ &\geq r_i(y_1^{(k+1)}, \ldots, y_{i-1}^{(k+1)}, z_i^{(k+1)}, y_{i+1}^{(k)}, \ldots), \end{aligned}$$

which, together with the strict diagonal isotonicity of r_i, implies that $y_i^{(k+1)} \geq z_i^{(k+1)}$. Similarly it follows that $z_i^{(k+1)} \leq x_i^{(k+1)}$. This completes the induction, and hence $z_j^{(k)} \to x_j^*$ pointwise. □

In particular, the nonlinear Gauss-Seidel and Jacobi processes converge if we use $\{u_j^n\}$ (i.e. the solution from the previous time step) as an initial guess. For small to moderate time-step sizes, one generally expects the solutions between consecutive time steps to be close to each other, so using $\{u_j^n\}$ often results in much faster convergence than either $\underline{u}$ or $\overline{u}$ as an initial guess.

3.5. Extensions. In this section we show how to extend the results of Theorems 3.6 and 3.7 to deal with conservation laws with non-uniform spatial grids and/or spatially-varying flux functions, as well as flux functions that are only defined over a closed interval $I \subset \mathbb{R}$.

3.5.1. Non-uniform grids and spatially-varying flux functions. Consider again the fully-implicit discretization (2.1):

$$\phi_i(u_i^{n+1} - u_i^n) + \lambda(F_{i+1/2}^{n+1} - F_{i-1/2}^{n+1}) = 0, \qquad \lambda = \Delta t/\Delta x, \;\; i \in \mathbb{Z},$$

where ϕ_i and $F_{i+1/2}$ can vary over the spatial index i. We assume that $0 < \phi_i \leq 1$. Note that the non-uniform grid case is automatically included: for any non-uniform discretization of the form

$$\frac{\tilde{\phi}_i(u_i^{n+1} - u_i^n)}{\Delta t} + \frac{F_{i+1/2}^{n+1} - F_{i-1/2}^{n+1}}{\Delta x_i} = 0, \tag{3.19}$$

we can multiply (3.19) by $\Delta t \Delta x_i / \Delta x_{max}$ to recover the form of (2.1) with

$$\phi_i = \tilde{\phi}_i \Delta x_i / \Delta x_{max}, \qquad \lambda = \Delta t / \Delta x_{max}.$$

To ensure convergence of the Jacobi and Gauss-Seidel processes, we need the following assumptions:

1. The family of flux functions $\{F_{i+1/2}\}_{i=-\infty}^{\infty}$ is equicontinuous (cf. [12]) with the same local Lipschitz constant K_B;
2. $\{\phi_i\}$ is uniformly bounded away from zero, i.e. there exists $\phi_{min} > 0$ such that $\phi_i \geq \phi_{min}$ for all $i \in \mathbb{Z}$.

While the equicontinuity condition may appear severe, it is usually satisfied in practice because the spatially-varying coeffients (e.g. $K(x)$ in (1.11)) tend to be uniformly bounded, ensuring equicontinuity in the flux functions. With the above assumptions, we can mimic Theorem 3.6 exactly by replacing λ with λ/ϕ_i. Then the proof goes through verbatim, except for (3.16), which must be modified to

$$\frac{2\lambda K_B}{\phi_{min} + 2\lambda K_B} < \beta < 1. \tag{3.20}$$

3.5.2. Bounded admissible solutions. Formally, Theorem 3.6 requires the discrete flux function $F(u_i, u_{i+1})$ to be defined on $\mathbb{R} \times \mathbb{R}$. In practice one may want to solve problems for which the flux function f is defined only on an interval $[u_{min}, u_{max}]$ rather than on all of $\mathbb{R}$, because states outside these bounds are unphysical. For instance, in the two-phase flow problem, we must have $S_i \in [0, 1]$ for all i, and the flux function $f(S)$ in (1.11) is not even defined outside this range. Fortunately, the estimate (3.14) ensures that as long as the initial conditions are within physical bounds, so will the solution for subsequent time steps $n > 0$, as well as any intervening Jacobi or Gauss-Seidel iterate. Thus, to apply Theorem 3.6 to these problems, one can *formally extend* the domain of the flux function f to $\mathbb{R}$ by defining, for instance,

$$\tilde{f}(u) = \begin{cases} f(u_{min}), & u < u_{min}, \\ f(u), & u_{min} \leq u \leq u_{max}, \\ f(u_{max}), & u > u_{max}, \end{cases}$$

and similarly for the discrete flux $F(u, v)$. Since all Gauss-Seidel iterates $\{y^k\}$ and $\{x^k\}$ satisfy the bound $x^0 \leq x^k \leq y^k \leq y^0$, the exact manner in which the extension is defined is unimportant as long as Assumption 1 (monotonicity) is valid.

4. Applications to Porous Media Flow.

4.1. 1D Buckley-Leverett problem with gravity. Consider a 1D incompressible two-phase flow problem with a constant-rate injection boundary condition on the left, a pressure boundary condition on the right, and no internal sources or sinks:

$$\phi(x)\frac{\partial S_j}{\partial t} + \frac{\partial v_j}{\partial x} = 0, \quad (x,t) \in (x_L, x_R) \times \mathbb{R}^+, \tag{4.1}$$

$$S_1(x, 0) = S^0(x), \quad x \in (x_L, x_R), \tag{4.2}$$

$$v_j(x_L, t) = v_{j,L}, \quad p(x_R, t) = x_R, \quad t \in \mathbb{R}^+ \tag{4.3}$$

for $j = 1, 2$, where $v_j = -K(x)\lambda_j(S_1)\Big(\dfrac{dp}{dx} - \rho_j \mathbf{g} \cdot \mathbf{i}\Big)$, $p \equiv p_1 = p_2$ (i.e., zero capillary pressure), and $\mathbf{i}$ is the unit vector along the x-direction. We assume that the injection velocities $v_{1,L}$ and $v_{2,L}$ are non-negative, and that the total velocity $v_{T,L} := v_{1,L} + v_{2,L}$ is strictly positive. (These assumptions cover the most interesting cases, such as oil recovery by water flooding.) This formulation, which contains pressure variables, is known as the *parabolic form* of the problem, since it represents the incompressible limit of a parabolic problem. We can also derive the *hyperbolic* or "fractional flow"

form of the problem by eliminating the pressure variables as follows. The discretized PDEs can be written as

$$\frac{\phi_i(S_{1,i} - S_{1,i}^{\text{old}})}{\Delta t} + \frac{F_{1,i+1/2} - F_{1,i-1/2}}{\Delta x} = 0, \tag{4.4a}$$

$$\frac{\phi_i(S_{1,i}^{\text{old}} - S_{1,i})}{\Delta t} + \frac{F_{2,i+1/2} - F_{2,i-1/2}}{\Delta x} = 0, \tag{4.4b}$$

where

$$F_{j,i+1/2} = K_{i+1/2}\lambda_{j,i+1/2}\left(\frac{p_i - p_{i+1}}{\Delta x} + \rho_j g\right) \tag{4.5}$$

for $j = 1, 2$, $i = 1, \ldots, N$, with $g = \mathbf{g} \cdot \mathbf{i}$. The numerical boundary conditions become

$$F_{j,1/2} = v_{j,L} \quad (j = 1, 2), \qquad p_{N+1} = 2p_R - p_N. \tag{4.6}$$

Assume without loss of generality that $g(\rho_1 - \rho_2) \geq 0$. To eliminate the pressure variables p_i, first note that summing equations (4.4a) and (4.4b) and rearranging gives

$$F_{1,i+1/2} + F_{2,i+1/2} = F_{1,i-1/2} + F_{2,i-1/2} = v_1 + v_2 =: v_T.$$

In other words, the total flux v_T is constant across any interface, which must then be equal to $v_{T,L}$. We can express the pressure gradient $(p_i - p_{i+1})/\Delta x$ in terms of v_T by summing (4.5) through $j = 1, 2$:

$$v_T = K_{i+1/2}\left[\lambda_{T,i+1/2}\frac{p_i - p_{i+1}}{\Delta x} + (\lambda_{1,i+1/2}\,\rho_1 g + \lambda_{2,i+1/2}\,\rho_2 g)\right],$$

where $\lambda_{T,i+1/2} = \lambda_{1,i+1/2} + \lambda_{2,i+1/2}$. Thus,

$$\frac{p_i - p_{i+1}}{\Delta x} = \frac{v_T - K_{i+1/2}(\lambda_{1,i+1/2}\,\rho_1 g + \lambda_{2,i+1/2}\,\rho_2 g)}{K_{i+1/2}(\lambda_{1,i+1/2} + \lambda_{2,i+1/2})}. \tag{4.7}$$

Substituting into (4.5) for $j = 1$ gives

$$\begin{aligned} F_{1,i+1/2} &= \frac{\lambda_{1,i+1/2}}{\lambda_{T,i+1/2}}\left[v_T + K_{i+1/2}\lambda_{2,i+1/2}(\rho_1 - \rho_2)g\right] \\ &= F_{1,i+1/2}(S_{1,i}, S_{1,i+1}), \end{aligned} \tag{4.8}$$

This, together with (4.4a):

$$\phi_i(S_{1,i} - S_{1,i}^{\text{old}}) + \frac{\Delta t}{\Delta x}\left(F_{1,i+1/2} - F_{1,i-1/2}\right) = 0, \tag{4.9}$$

leads to a numerical scheme that is identical to (2.1) except for the boundary conditions. Clearly, the treatment of boundary conditions will significantly affect the stability and accuracy of the numerical scheme. However, in order to understand the behavior of the numerical scheme at interior points, we will simply replace the initial-boundary value problem (4.1)–(4.3) with an initial value problem on an infinite domain with appropriate initial conditions. In particular, we replace the injection boundary condition with

$$S_1^0(x) = S_{1,L}, \qquad x < x_L, \tag{4.10}$$

where $S_{1,L}$ satisfies $f_1(x_L, S_{1,L}) = v_{1,L}/v_T$, with $f_1(x, S_1)$ given by (1.10). We also replace the pressure boundary condition with

$$S_1^0(x) = S_1^0(x_R), \qquad x > x_R. \tag{4.11}$$

The modified continuous problem will yield a solution identical to (4.1)–(4.3) for $0 < t < T_{BT}$, where T_{BT} is the breakthrough time (i.e., the time at which the shock front arrives at the pressure boundary). Note that f_1 is one-to-one over the interval $I = \{S : 0 \le f_1(S) < 1\}$ (see Figure 1.1), and $v_{1,L} \le v_T$ by assumption; thus, (4.10) is well-defined unless $v_{2,L} = 0$. (If $v_{2,L} = 0$, we define $S_1^0(x) = \inf f_1^{-1}(v_T)$, where f_1^{-1} denotes the inverse image.) For the remainder of this section, we will drop the phase subscript and denote $S_{1,i}$ and $F_{1,i+1/2}$ by S_i and $F_{i+1/2}$ respectively.

Phase-based upstreaming. Recall from section 1 (cf. Equation (1.9)) that the mobilities $\lambda_{j,i+1/2}$ are evaluated using the upstream saturations with respect to the flow direction of phase j:

$$\lambda_{j,i+1/2} = \begin{cases} \lambda_j(S_i) & \text{if } \frac{1}{\Delta x}(p_i - p_{i+1}) + \rho_j g \ge 0, \\ \lambda_j(S_{i+1}) & \text{otherwise.} \end{cases} \tag{4.12}$$

In light of (4.7), we can rewrite the upstream conditions as

$$\lambda_{j,i+1/2} = \begin{cases} \lambda_j(S_i) & \text{if } v_T + K_{i+1/2}\lambda_{j',i+1/2}(\rho_j - \rho_{j'})g \ge 0, \\ \lambda_j(S_{i+1}) & \text{otherwise,} \end{cases} \tag{4.13}$$

where the subscript $j' := 3 - j$ denotes the phase other than phase j. Even though pressure dependence has been eliminated, Equation (4.13) still does not explicitly define the upstream direction for λ_j, since the latter is defined in terms of the (yet undetermined) mobility of the other phase $\lambda_{j',i+1/2}$. For explicit numerical schemes, Brenier and Jaffré has shown in [3] how to explicitly determine the upstream direction for each phase for a given saturation profile $\{S_i^n\}$. In the special case of two-phase flow, they define the following quantities:

$$\theta_{1,i+1/2} = v_T + K_{i+1/2}g(\rho_1 - \rho_2)\lambda_2(S_{i+1}^n),$$
$$\theta_{2,i+1/2} = v_T - K_{i+1/2}g(\rho_1 - \rho_2)\lambda_1(S_i^n).$$

These quantities correspond precisely to the condition in (4.13), but the condition is evaluated at S_{i+1}^n for θ_1 and S_i^n for θ_2. Clearly $\theta_{1,i+1/2} > 0$, since $g(\rho_1 - \rho_2) \ge 0$. The correct upstream directions are then given by

$$\lambda_{1,i+1/2}^n = \lambda_1(S_i^n) \qquad \lambda_{2,i+1/2}^n = \lambda_2(S_i^n), \qquad \text{if } 0 \le \theta_{2,i+1/2} \le \theta_{1,i+1/2},$$
$$\lambda_{1,i+1/2}^n = \lambda_1(S_i^n) \qquad \lambda_{2,i+1/2}^n = \lambda_2(S_{i+1}^n), \qquad \text{if } \theta_{2,i+1/2} \le 0 \le \theta_{1,i+1/2}.$$

Thus, the numerical fluxes are well defined if one uses an explicit method with respect to saturation. However, this analysis does not address implicit-saturation schemes (such as FIM), which require the upstream directions to be consistent with saturation values *at the end of the time step*, i.e. with the saturation profile $\{S_i^{n+1}\}$. Because of this consistency requirement, it is not clear a priori that a solution to the parabolic form of the problem (4.4) even exists. Our approach to proving existence is to rely on the hyperbolic form of the problem (4.8)–(4.11). From the above derivation, it is evident that if $\{(S_i, p_i)\}_{i=1}^N$ is any solution to the parabolic form (4.4)–(4.6), then

$\{S_i\}_{i=1}^N$ must be a solution to the hyperbolic problem. Thus, the key idea is to first find the correct saturation profile $\{S_i\}$ using (4.8)–(4.11) with a numerical flux that automatically ensures consistency with the upstream directions; once the $\{S_i\}$ are known, we can easily solve for the pressure part because the pressure equation is linear. We distinguish two cases:

1. If $K_{i+1/2}g(\rho_1-\rho_2)\lambda_{1,\max} \le v_T$, then $\theta_{2,i+1/2} \ge 0$ always, so we revert to a single-point upstream scheme $F_{i+1/2} = F_{i+1/2}(S_i)$;

2. If $K_{i+1/2}g(\rho_1-\rho_2)\lambda_{1,\max} > v_T$, then by the monotonicity of $\lambda_1(S)$, there exists a unique $0 < S_c < 1$ such that $K_{i+1/2}g(\rho_1-\rho_2)\lambda_1(S_c) = v_T$. Then the numerical flux, which is to be evaluated at time t^{n+1}, is defined as

$$F_{i+1/2}(S_i, S_{i+1}) = \begin{cases} \dfrac{\lambda_1(S_i)\left[v_T + K_{i+1/2}g(\rho_1-\rho_2)\lambda_2(S_i)\right]}{\lambda_1(S_i)+\lambda_2(S_i)} & \text{if } 0 \le S_i \le S_c, \\ \dfrac{\lambda_1(S_i)\left[v_T + K_{i+1/2}g(\rho_1-\rho_2)\lambda_2(S_{i+1})\right]}{\lambda_1(S_i)+\lambda_2(S_{i+1})} & \text{if } S_c < S_i \le 1. \end{cases} \tag{4.14}$$

Figure 4.1 shows a plot of $F(S_i, S_{i+1})$ in the latter case, which corresponds to the continuous flux function in Figure 1.1(b). Note that we are able to recover $f(S)$ from the numerical flux $F(S_i, S_{i+1})$ because consistency implies $F(S,S) = f(S)$. Even though $f(S)$ itself is non-monotonic, the plot clearly shows that $F(S_i, S_{i+1})$ is an increasing function of S_i and a decreasing function of S_{i+1}, as predicted by Theorem 2.1. Also, the numerical flux is independent of the downstream saturation S_{i+1} inside the cocurrent region ($0 \le S_i \le S_c \approx 0.27$), but becomes a function of both variables when $S_i > S_c$. Finally, $F(S_i, S_{i+1})$ is Lipschitz continuous, but non-differentiable along the line $S_i = S_c$ due to the upstream condition (4.14). Since the numerical flux satisfies the monotonicity assumption, the analysis in the last two sections shows that the hyperbolic problem with implicit time-stepping has a unique solution $\{S_i^{n+1}\}$, which must also be the correct saturation profile for the parabolic problem. To solve for pressure, we use (4.7) and (4.6):

$$\frac{p_i - p_{i+1}}{\Delta x} = \frac{v_T - K_{i+1/2}(\lambda_{1,i+1/2}\, g\rho_1 + \lambda_{2,i+1/2}\, g\rho_2)}{K_{i+1/2}(\lambda_{1,i+1/2} + \lambda_{2,i+1/2})}, \qquad i = 1, \ldots, N, \tag{4.15}$$

$$p_{N+1} = 2p_R - p_N. \tag{4.16}$$

Since $\{S_i^{n+1}\}$ is now known, the right-hand side of (4.7) is also completely determined. Thus, the vector p of pressures satisfies $Ap = b$, where A is an $N \times N$ upper triangular matrix with a non-zero diagonal. So A is non-singular, which means there is a unique pressure profile $\{p_i^{n+1}\}$ that satisfies (4.7) and (4.6). It is easy to see that this pressure profile is consistent with the upstream condition (4.12): because of (4.7), this upstream condition is equivalent to (4.13), and the conditions therein are precisely the ones we use to define the numerical flux function (4.14) for the hyperbolic problem. Hence, we have shown that the parabolic form (4.4)–(4.6) has a unique solution, given by the above $\{(S_i^{n+1}, p_i^{n+1})\}$.

4.2. Multidimensional considerations. In multiple dimensions, one can no longer eliminate pressure variables as shown above, because the total velocity $\mathbf{v}_T$ is generally a function of space and time. Thus, the system of PDEs (1.1), (1.2) does not reduce to a purely hyperbolic problem, which means we cannot directly apply our existence and uniqueness results to the fully-implicit method in this case. Nonetheless, our analysis does apply to a related numerical scheme known as the *sequential-implicit*

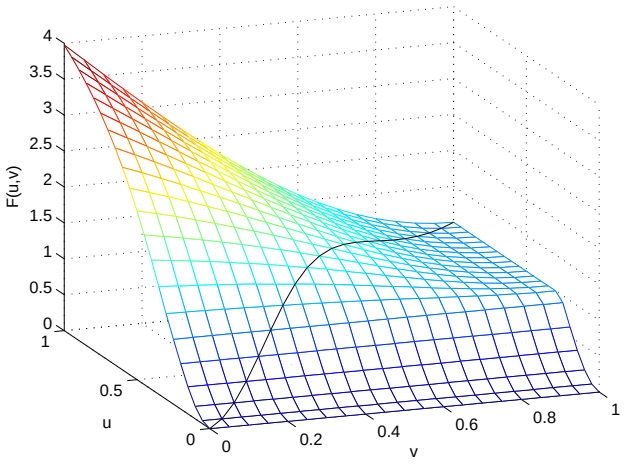


FIG. 4.1. *The numerical flux function $F(u,v)$ corresponding to the fractional flow in Figure 1.1(b). The black curve along the diagonal indicates the value of $F(u,u) = f(u)$.*

method (SEQ). At each time step in SEQ, we first solve the discrete version of the (linear) elliptic equation (1.5), in which the saturation-dependent coefficients are taken at time t^n. In other words, we solve for p^{n+1} via:

$$-\nabla \cdot \left[K\lambda_T(S^n)\nabla p^{n+1} - K\mathbf{g}\textstyle\sum_j \rho_j\lambda_j(S^n) \right] = \textstyle\sum_j q_j. \tag{4.17}$$

Next, we compute the total velocity

$$\mathbf{v}_T^* = \textstyle\sum_j \mathbf{v}_j^* = -\sum_j K\lambda_j(S^n)(\nabla p^{n+1} - \rho_j \mathbf{g}). \tag{4.18}$$

Finally, we compute the saturations S_j^{n+1} $(j = 1, \ldots, n-1)$ by solving the discrete version of (1.6) and (1.7) with *implicit time-stepping*:

$$\phi \frac{\partial S_j}{\partial t} + \nabla \cdot \mathbf{v}_j(x, S_1, \ldots, S_n) = 0, \tag{4.19}$$

$$\mathbf{v}_j = \frac{\lambda_j}{\lambda_T}\left(\mathbf{v}_T^* - K\mathbf{g}\textstyle\sum_\ell \lambda_\ell(\rho_\ell - \rho_j)\right). \tag{4.20}$$

Essentially, the SEQ method decouples the system into an elliptic and a hyperbolic subproblem. For two-phase flow problems, one can readily extend the convergence results of Theorems 3.6 and 3.7 to the hyperbolic subproblem above, as long as the spatial grid satisfies certain shape and connectivity requirements. Discretizing (4.19) yields the multidimensional analog of (2.1):

$$\phi_i(S_i^{n+1} - S_i^n) + \sum_{l \in \mathrm{adj}(i)} \lambda_{il} F_{il}(S_i^{n+1}, S_l^{n+1}) = 0. \tag{4.21}$$

Here, F_{il} is the flux (or velocity) from cell i to cell l, and $\lambda_{il} = \Delta t|\partial V_{il}|/|V_i|$, where $|\partial V_{il}|$ is the area of the surface separating cell i and l, $|V_i|$ is the volume of cell i and Δt is the time step. For a conservative scheme we must have $F_{il}(S_i, S_l) = -F_{li}(S_l, S_i)$,

and for monotonicity we require that F_{il} be non-decreasing with respect to the first argument and non-increasing with respect to the second. This requirement is satisfied for two-phase flow problems, since we can reproduce the derivation in section 4.1 to obtain the flux function

$$F_{il} = \frac{\lambda_{1,il}}{\lambda_{T,il}} \left[v_{il} + K_{il} g_{il} (\rho_1 - \rho_2) \lambda_{2,il} \right]$$

and the upstream condition

$$\lambda_{j,il} = \begin{cases} \lambda_j(S_i) & \text{if } v_{il} + K_{il} g_{il} (\rho_j - \rho_{j'}) \lambda_{j',il} \geq 0, \\ \lambda_j(S_l) & \text{otherwise,} \end{cases}$$

for $j = 1, 2$, where $v_{il} = \mathbf{v}_T^* \cdot \mathbf{n}_{il}$ and $g_{il} = \mathbf{g} \cdot \mathbf{n}_{il}$. In order to mimic the proof of Theorem 3.6, we need the following assumptions on the grid:

1. The number of cells (control volumes) adjacent to cell i is bounded for all i;
2. The ratio $|\partial V_{il}|/|V_i|$ is bounded for all pairs of adjacent cells (i, l);
3. The quantity $\phi_i |V_i|$ is uniformly bounded away from zero for all i;
4. For any cell i, the total number of cells reachable from i in k steps is $O(k^p)$ for some fixed $p > 0$ (i.e. grows at most polynomially in k).

The above assumptions are easily satisfied by regular Cartesian grids, and also by most unstructured grids of practical interest. We also need the following assumption on the numerical flux F_{il} (which is analogous to the assumption in section 3.5.1):

5. F_{il} is equicontinuous with the same Lipschitz constant for all pairs of adjacent cells (i, l).

These assumptions ensure that the residual functions are all bounded and have the same Lipschitz constant over the set $\{u \in \ell^\infty(N) \,|\, \|u\|_\infty < B\}$. The polynomial growth assumption (4) allows us to assign the weights $\{w_i^B\}$ to each cell i in the following manner: pick any node i_0 and let $w_i^B = \beta^{d(i_0,i)}$, where $d(i, j)$ is the shortest distance between node i and j in the graph-theoretic sense. Since the number of cells within k steps of i_0 grows polynomially in k, the series $\sum_i w_i^B$ converges for any $0 < \beta < 1$, so β can be chosen the same way as in step 3 of Theorem 3.6 and the same argument will hold. Hence, we can conclude that the hyperbolic subproblem in the SEQ method is well-defined for any time-step size Δt, and the nonlinear Gauss-Seidel and Jacobi processes are guaranteed to converge for these problems.

5. Accuracy of Phase-based Upstreamed Solutions. In this section, we investigate the accuracy of the numerical solution obtained from implicit phase-based upstreaming when we vary the spatial and temporal grid. Our 1D test case consists of a countercurrent flow problem inside the domain $\Omega = [0, 1]$, and our 2D example is a cocurrent flow problem in a heterogeneous reservoir. For the 1D problem, water is injected at the boundary $x = 0$ and a pressure boundary condition is maintained at $x = 1$. The hyperbolic form of the problem is described by (1.10), (1.11). The flux function $f_1(S)$, which is independent of x, is shown in Figure 1(b), with a sonic point at $S = 0.49$; countercurrent flow occurs whenever $S \geq S_c \approx 0.27$. The initial saturation profile is a step function with

$$S^0(x) = \begin{cases} 1, & 0 \leq x < 0.2, \\ 0, & 0.2 < x \leq 1. \end{cases}$$

The numerical solution is compared with the analytical solution at time $t = 0.15$. Because of the sonic point, the solution contains two shocks connected by a rarefaction;

one shock moves to the right with a velocity of 3.9, and the other travels to the left with a velocity of -1.2. When considering the accuracy of a numerical solution, two error measures are shown:

- The L^1-*error*, which is the difference between the numerical and the analytical solution in the L^1-norm;
- The *front dispersion*, which is the distance between analytical shock front and the leftmost point for which the numerical solution becomes zero.

We also measure how difficult the nonlinear problem is by showing, for each test case, the average number of nonlinear Gauss-Seidel iterations required to converge each time step. We remark that this measure is only useful for problems with countercurrent flow; in the cocurrent case, the flux function $F_{i+1/2}$ is a function of S_i only, which means Gauss-Seidel will always converge in one iteration if we solve the single-cell equations in the order $1, 2, \ldots, N$ (i.e., from upstream to downstream). In the countercurrent case, convergence is generally linear, and the rate of convergence is a function of the time step [8].

5.1. Refinement under fixed mesh ratio. Here we refine the grid at a fixed mesh ratio $\Delta x/\Delta t$ to maintain a fixed CFL number of 4.10, which is above the CFL limit for explicit schemes. Figure 5.1 shows the plots for $N = 50, 100, 200, 400$, and Table 5.1 shows the L^1-error and front dispersion data. The plots show that the numerical solution converges to the analytical solution even though the CFL number is greater than 1, which confirms our analysis. Moreover, both the L^1-error and the front dispersion are decreasing to zero a bit worse than linearly, with a ratio of about 0.61 and about 0.58 respectively for every refinement by a factor of two. Also note the poor resolution near the left boundary for $N = 50, 100$, where instead of approaching $S = 1$, the solution is closer to $S_c \approx 0.27$ at the left boundary. For these coarser grids, the numerical solution cannot decide whether the left-moving wave has reached the boundary, which is maintained at $S(0, t) = S_c$ (see Equation (4.10)). For higher resolutions ($N = 200, 400$), the artifact has disappeared and the numerical solution reproduces the back end of the saturation profile quite accurately. The average number of Gauss-Seidel iterations required for convergence are all similar, so refining the grid at a fixed mesh ratio does not increase the difficulty of the problem for the nonlinear solver.

5.2. Spatial refinement for fixed time steps. Here, we refine the spatial grid only while fixing the time-step size. Figure 5.2 and Table 5.2 show the results for $N = 25, 50, 100, 200, 400$ and a time-step size of $\Delta t = 0.0075$, i.e., we use 20 time steps to integrate up to $t = 0.15$. We see that even though the $N = 25$ case has a CFL number close to 1, the grid is clearly too coarse, and the shock fronts are very poorly resolved. The accuracy increases substantially when the spatial grid is refined to $N = 50, 100$, even though the CFL number becomes progressively larger; thus, the CFL number by itself is not a good measure of solution quality. However, the improvement due to spatial grid refinement becomes negligible for $N > 100$, since errors due to time discretization is now the dominant source of error. In addition, the average number of iterations required to attain convergence increases with each refinement: as we refine the grid, we are solving increasingly difficult problems, even though the improvement in solution accuracy will stagnate beyond a certain point. Thus, even though the fully-implicit method can tolerate arbitrarily large CFL numbers, one should not expect the solution accuracy to improve indefinitely simply by using a finer spatial grid, without making a corresponding reduction in time-step size.

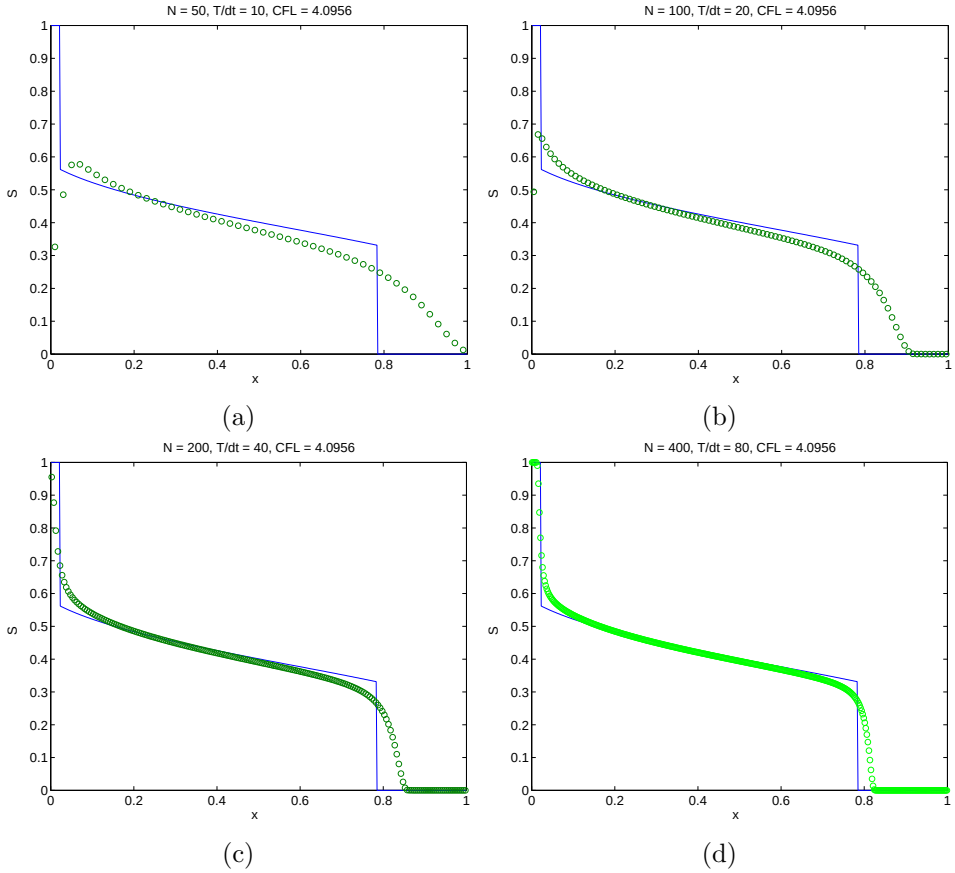


FIG. 5.1. *Numerical solution at different resolutions, CFL = 4.10,* $t = 0.15$.

5.3. Non-uniform grids. The real advantage of the fully-implicit method over an explicit scheme lies in its efficiency when applied to a heterogeneous problem, where the porosity $\phi(x)$ and permeability $K(x)$ can vary by orders of magnitude over the domain. In these problems, the CFL condition is determined by the minimum porosity in the domain, which can be much smaller than the average porosity. To illustrate this point, we show an example in which the spatial grid is non-uniform (which, from section 3.5.1, is equivalent to the spatially-varying porosity case). The non-uniform grid contains 50 gridblocks, with $\Delta x_{\max}/\Delta x_{\min} = 96$. Figure 5.3 and Table 5.3 compare the numerical solutions obtained from this grid to the uniform-grid solutions. We see that the solutions are qualitatively (from the plots) and quantitatively (from the L^1-error and front dispersion) not very different, even though the CFL number is 50 times larger in the non-uniform case. Thus, an explicit integrator would have to take unacceptably small time steps, whereas an implicit method allows time steps that are much more reasonable. In addition, the average number of iterations required for convergence is roughly the same for both cases, so the equations resulting from a non-uniform grid are not harder to solve, despite the large CFL numbers.

5.4. 2D example. The goal of this example is to compare the unconditionally stable SEQ method with the implicit-pressure-explicit-saturation (IMPES) method,

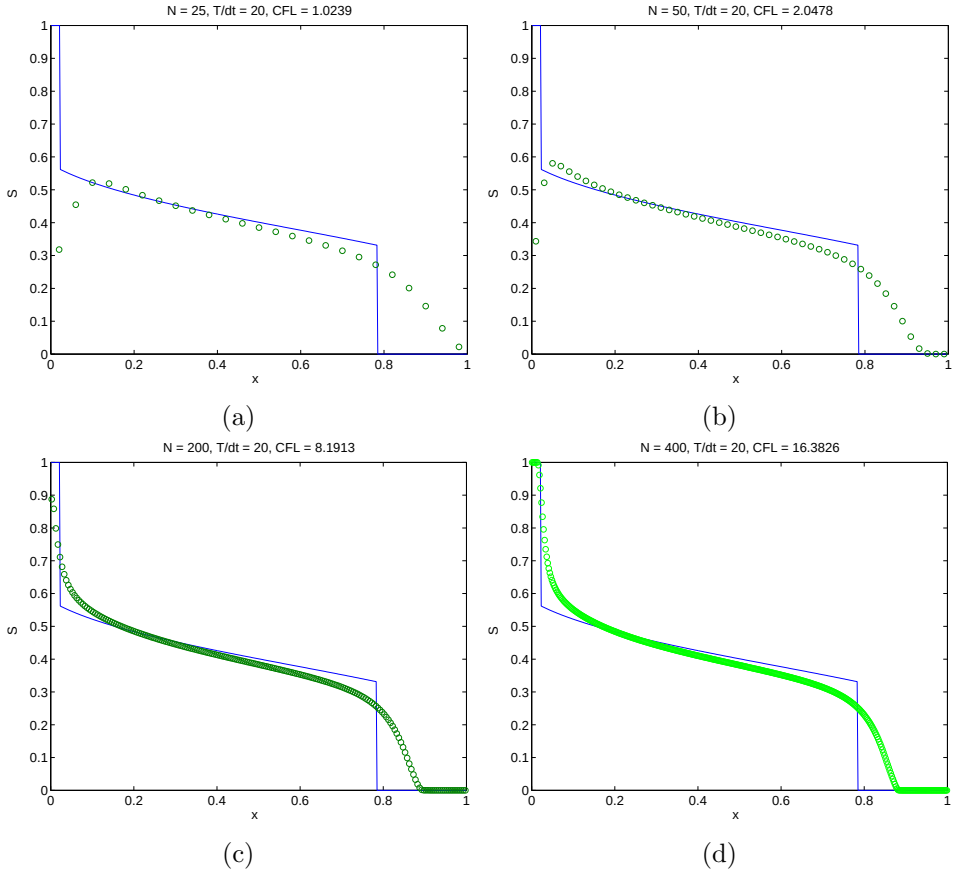


FIG. 5.2. *Numerical solution for different spatial grids, 20 time steps,* $t = 0.15$. *The* $N = 100$ *case is identical to Figure 5.1(b) and hence omitted.*

which carries a time-step restriction $\Delta t = \mathcal{O}(\Delta x)$. We use a 110×30 grid to discretize a 2D horizontal reservoir (i.e., no gravity). The porosity ϕ is constant throughout the reservoir, whereas the permeability field $K(\mathbf{x})$ is taken from the SPE 10 test set [4] and ranges from a minimum of 0.0052 to a maximum of 1219 (see Figure 5.4(b)). The reservoir is initially saturated with oil. Starting from $t = 0$, water is injected at a constant rate into cell $(1, 1)$, and fluid is produced at constant pressure from cell $(110, 30)$. We simulate the reservoir until $t = 30$ (0.182 pore volumes injected), which is roughly when breakthrough occurs at the outlet boundary. For IMPES, we take the largest time step allowed by the CFL criterion, whereas for SEQ we use two time-stepping strategies:

- Small Δt: $t = 1, 2, 3, 4$; after $t = 4$, $\Delta t = 2$ until $t = 30$;
- Large Δt: $t = 1, 3, 6, 10$; after $t = 10$, $\Delta t = 5$ until $t = 30$.

Figures 5.4(a), (c) and (e) show the saturation profiles obtained from IMPES and SEQ for the two time-step sizes, and Table 5.4 shows the error and running times for each case. Note that all three solutions are very similar; the largest discrepancies occur near flood fronts, where the SEQ solutions are noticeably more diffuse than the IMPES solution. However, the IMPES profile is sharp only because the method is forced to take tiny time steps to satisfy a very severe CFL condition. As a result,

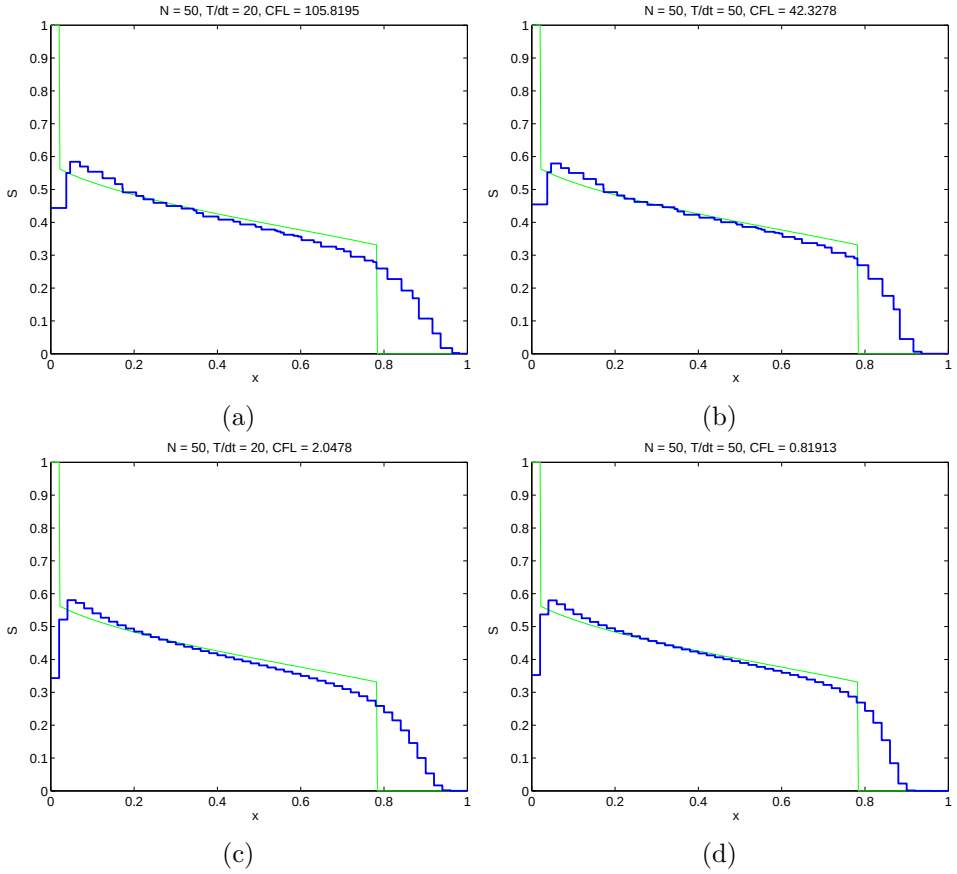


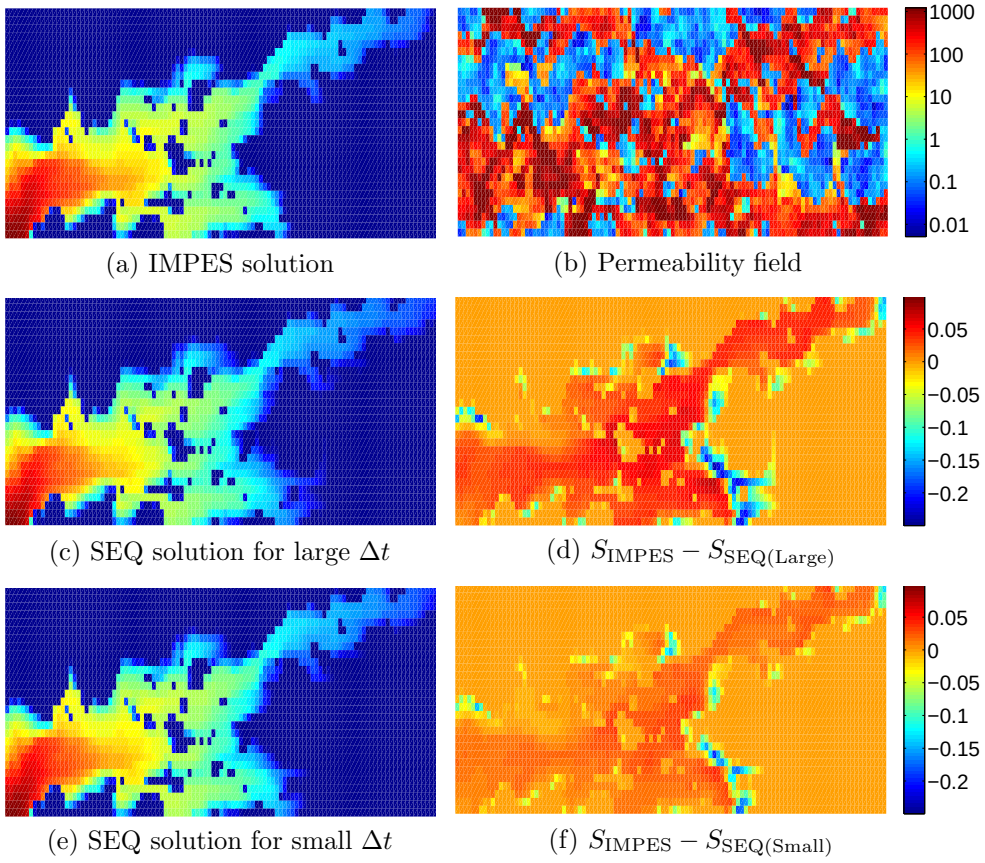
FIG. 5.3. *Numerical solutions obtained from a non-uniform grid ((a) and (b)) and their uniform-grid counterparts ((c) and (d)),* $t = 0.15$.

TABLE 5.1
Accuracy of numerical solutions for a fixed CFL number.

N	$t/\Delta t$	CFL	L^1-error	Front dispersion	Average # iterations
50	10	4.10	0.0665	> 0.215	4.9
100	20	4.10	0.0444	0.116	4.4
200	40	4.10	0.0273	0.066	4.2
400	80	4.10	0.0168	0.039	4.1

TABLE 5.2
Accuracy of numerical solutions for a fixed time step size.

N	$t/\Delta t$	CFL	L^1-error	Front dispersion	Average # iterations
25	20	1.02	0.0673	> 0.215	2.6
50	20	2.05	0.0529	0.156	3.3
100	20	4.10	0.0444	0.116	4.4
200	20	8.20	0.0378	0.101	6.4
400	20	16.40	0.0366	0.094	9.2



(a) IMPES solution (b) Permeability field

(c) SEQ solution for large Δt (d) $S_{\text{IMPES}} - S_{\text{SEQ(Large)}}$

(e) SEQ solution for small Δt (f) $S_{\text{IMPES}} - S_{\text{SEQ(Small)}}$

FIG. 5.4. *Permeability of the 2D reservoir, as well as saturation profiles that produced by the different numerical methods.*

TABLE 5.3
Accuracy of numerical solutions for a non-uniform grid.

	N	$t_D/\Delta t$	CFL	L^1-error	Front disp.	Avg. # iters
Non-uniform	50	20	105.80	0.0566	0.180	3.2
	50	50	42.30	0.0475	0.132	2.2
Uniform	50	20	2.05	0.0529	0.156	3.3
	50	50	0.82	0.0435	0.116	2.1

TABLE 5.4
Performance of different numerical methods on the 2D example.

	IMPES	SEQ (large Δt)	SEQ (small Δt)
No. of time steps	1350	8	17
Maximum Δt	0.0222	5	2
Maximum CFL	1	225	90
Breakthrough time (days)	29.1	24.6	27.7
$\|S_{\text{IMPES}} - S_{\text{SEQ}}\|_{L^1(\Omega)}$	–	0.0188	0.0091
$\|S_{\text{IMPES}} - S_{\text{SEQ}}\|_{L^\infty(\Omega)}$	–	0.2348	0.2004
Solution time (sec)	66017	1016	2130

IMPES runs 65 times slower than SEQ for the large Δt case, and 31 times slower for the small Δt case. Table 5.4 indicates that the maximum difference between the IMPES and SEQ solutions remains nearly constant when the time step is reduced. This is not surprising, because the numerical solution cannot converge uniformly when discontinuities are present in the solution. However, we do observe a decrease in the L^1-error, as well as a later breakthrough time, when Δt is reduced. This is consistent with our 1D results, where refining the grid reduces the L^1 and front dispersion errors, but the L^∞-error does not go to zero because of smearing across the shock front. This example shows that, for problems of practical interest, an implicit method can produce solutions of comparable quality at much lower computational costs than an explicit method.

6. Conclusion. We have shown that, for any residual function that arises from the implicit monotone discretization of a scalar hyperbolic conservation law, the nonlinear Gauss-Seidel and Jacobi processes converge to the unique bounded solution whenever the initial guess is bounded. This provides an alternate, constructive proof of the well-definedness of monotone implicit schemes, for which a solution algorithm is easily implementable. Convergence to the entropy solution for arbitrary CFL numbers follows immediately from the properties of the flux functions. These results are applicable to the fully-coupled two-phase flow problem in one-dimension, and to the hyperbolic subproblem in the sequential-implicit method in higher dimensions. Finally, we studied the accuracy of phase-based upstream solutions under different grid refinement schemes, and the importance of unconditional stability became evident when a non-uniform grid and/or a variable porosity field is used.

REFERENCES

[1] Hal Abelson, Jerry Sussman, and Julie Sussman, *Structure and Interpretation of Computer Programs*, MIT Press, 1984.

[2] Khalid Aziz and Antonin Settari, *Petroleum Reservoir Simulation*, Applied Science Publishers, New York, 1979.

[3] Yann Brenier and Jérôme Jaffré, *Upstream differencing for multiphase flow in reservoir simulation*, SIAM J. Numer. Anal., 28 (1991), pp. 685–696.

[4] M. A. Christie and M. J. Blunt, *Tenth SPE comparative solution project: A comparison of upscaling techniques*, SPE Reservoir Eval. Eng., 4 (2001), pp. 308–317.

[5] M. G. Crandall and T. M. Liggett, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.

[6] Klaus Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, 1985.

[7] Steinar Evje and Kenneth Hvistendahl Karlsen, *Degenerate convection-diffusion equations and implicit monotone difference schemes*, in Hyperbolic problems: Theory, Numerics, Applications, M. Fey and R. Jeltsch, eds., vol. 129, Birkhäuser Verlag, 1999, pp. 285–294.

[8] Felix Kwok, *Scalable Linear and Nonlinear Algorithms for Multiphase Flow in Porous Media*, PhD thesis, Stanford University, Stanford, CA, Dec. 2007.

[9] Bradley J. Lucier, *On nonlocal monotone difference schemes for scalar conservation laws*, Math. Comp., 47 (1986), pp. 19–36.

[10] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal., 21 (1984), pp. 217–235.

[11] Werner C. Rheinboldt, *On M-functions and their application to nonlinear Gauss-Seidel iterations and to network flows*, J. Math. Anal. Appl., 32 (1970), pp. 274–307.

[12] H. L. Royden, *Real Analysis*, Prentice-Hall, 1988.

[13] Yousef Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, 2nd ed., 2003.

[14] R. Sanders, *On convergence of monotone finite difference schemes with variable spatial differencing*, Math. Comp., 40 (1983), pp. 91–106.