A MAXIMUM PRINCIPLE BASED NUMERICAL APPROACH TO POROUS MEDIUM EQUATIONS

IULIU SORIN POP[†,§] AND WEN-AN YONG[§]

1. Introduction

In this work we propose a maximum principle based numerical approach to some kind of degenerate parabolic equations typically describing the gas flow through porous media. The numerical difficulties in treating such problems arise from the nonlinear degeneracy of the equations.

A simple model with the above feature is

$$\begin{aligned} \partial_t u - \Delta\beta(u) &= 0, &&\text{in } \ Q_T \equiv (0,T)\times\Omega, \\ u(x,0) = u_0(x) &\geq 0, &&\text{in } \ \Omega, \\ u &= 0, &&\text{on } \ \partial\Omega. \end{aligned} \tag{1.1}$$

Here $\Omega \subset \mathbb{R}^d (d \geq 1)$ is a bounded domain with a Lipschitz continuous boundary, $T < \infty$ and $\beta : \mathbb{R} \to \mathbb{R}$ is a strictly increasing smooth function. By degeneracy we mean $\beta'(u) = 0$ for some u. Throughout this paper we assume, without loss of generality, that $\beta'(0) = 0$.

A possible treatement of such problems consists in the modification of β or β' so that the resulting diffusion coefficient does not vanish, resulting a regular parabolic problem. This suggests various numerical schemes to approximate the above degenerate problems (see [7], [8], [11], [14] and references therein).

On the other hand, it is well known that if positive solutions are sought, the difficulties due to the degeneracy can be overcome by perturbing the data such that the corresponding solutions do not take the value 0—the degenerate point (see [13]). The aim of this paper is to seek the possibility to use this technique numerically. More precisely, our approach is to solve the equations with the locally perturbed (initial and boundary) data, instead of the prescribed ones. The resulting problem has a solution taking values away from the degenerate point. This is guaranteed by the maximum principle obeyed by such equations. Thus the initial degenerate problem is reduced to a regular parabolic one, which can be solved numerically in different ways.

Because $\beta(u)$ is more regular then u itself, it is more convenient to consider the equation in the variable $\theta = \beta(u)$ and dicretize it correspondingly, as shown in the papers cited above. For the sake of simplicity we consider one-step discretization methods in time. Let n be an integer and $\tau = T/n$. The Euler implicit scheme reads

$$\begin{aligned} \beta^{-1}(\theta^{k+1}) - \beta^{-1}(\theta^k) =& \tau\Delta\theta^{k+1}, \\ \theta^{k+1}|_{\partial\Omega} \equiv& \beta(\varepsilon) \end{aligned} \tag{1.2}$$

1991 *Mathematics Subject Classification.* Primary 65M20; Secondary 35K65.

Key words and phrases. Numerical methods, degenerate parabolic equations, maximum principle, error estimates.

for $k = \overline{0, n-1}$ with $\theta^0 = \beta(u_0) + \beta(\varepsilon)v_0$, where ε is an artificial positive small number and $v_0 \in H^1(\Omega)$ an artificial function. Here θ^k approximates $\theta(t_k) = \beta(u(t_k))$ $(t_k = k\tau)$.

The choice of v_0 depends on the given initial data $u_0(x)$, so that

$$v_0|_{\{u_0 \geq \varepsilon\}} \equiv 0, \; v_0 \leq 1 \text{ and } \beta(u_0) + \beta(\varepsilon)v_0 \geq \beta(\varepsilon). \tag{1.3}$$

Here the inequalities hold almost everywhere (a.e.). A typical example is

$$v_0 := \left[1 - \frac{\beta(u_0)}{\beta(\varepsilon)}\right]_+, \tag{1.4}$$

where $[x]_+ = x$ if $x \geq 0$, otherwise $[x]_+ = 0$.

The above scheme is nonlinear. A linearized version is given below

$$\begin{aligned} \sigma_k(\theta^{k+1} - \theta^k) =& \tau\Delta\theta^{k+1}, \\ \theta^{k+1}|_{\partial\Omega} \equiv& \beta(\varepsilon), \\ \sigma_{k+1} =& (\beta^{-1})'(\theta^{k+1}) \end{aligned} \tag{1.5}$$

for $k = \overline{0, n-1}$, where $\theta^0 = \beta(u_0) + \beta(\varepsilon v_0$ and $\sigma_0 = (\beta^{-1})'(\theta^0)$. In order to make this scheme possible some assumptions on β are necessary. These are given below.

It is worth here to mention two aspects concerning the above schemes. First, the implicit method is similar to the one proposed in [7], [8] with a particular choice of the parameters. But the regularization there is done through a cut-off approach which, due to the shift of the data, is not necessary in (1.2). The same remark holds also in the linearized case, where a similarity to the scheme proposed in [14] holds up to the regularization method. In this last case β is perturbed globally, which again is not our situation.

For the spatial discretization the lumped mass version of the finite element method (see [17]) is considerred. This approach preserves the maximum principle and was adopted in many papers dealing with such types of problems (see, e.g, [5], [12], [15]).

2. Assumptions and formulation of the problem

About the problem in (1.1), we make the following assumptions.

(A1) β is continuously differentiable, $\beta(0) = 0$, $\beta'(u) \geq 0$ and β' may vanish at a single point at most (which we assume to be also 0).

(A2) $u_0 \geq 0$, $u_0 \in L^\infty(\Omega)$ and $\beta(u_0) \in H_0^1(\Omega)$.

In the analysis of the linearized version the Lipschitz continuity of β' is requested supplementary.

The assumption in (A1) implies that β has an inverse β^{-1}, which is continuously differentiable. Then there are two positive constants such that

$$0 < C_1(\varepsilon) \leq (\beta^{-1})'(\theta) \leq C_2(\varepsilon) < \infty, \qquad \forall\theta \in [\beta(\varepsilon), \beta(M)]. \tag{2.1}$$

Here M is a large constant satisfying $\theta^0 \leq \beta(M)$ a.e.. Moreover, $C_1(\varepsilon)$ often depends on ε in a regular way. To see this, we refer to the typical case where $\beta(u) = u^m (m \geq 1)$. So we will use C_1 to denote $C_1(\varepsilon)$.

Here and below we use the standard notation. In particular, we let $(\cdot, \cdot)$ stand for the inner product on $L^2(\Omega)$ or the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, $\|\cdot\|$ for the norm in $L^2(\Omega)$, $< u, v >$ for $(\nabla u, \nabla v)$, $\|\cdot\|_1$ and $|\cdot|_1$ for the respective norm and seminorm in $H^1(\Omega)$. In addition, we often write u or $u(t)$ instead of $u(t,x)$ and use C to denote a generic positive constant independent of τ and ε.

For the spatial discretization we consider a family $\{S_h\}_h$ of *regular* ([3]) decompositions of $\Omega \subset \mathbb{R}^d$ into closed d-simplices; h stands for the mesh-size. We assume, for simplicity, that $\overline{\Omega} = \cup_{T\in S_h} T$. Therefore the effect of numerical integration is neglected (a complete analysis in this sense can be found, e.g. in [12]). Let $V_h = \{\varphi \in C(\overline{\Omega}) : \varphi|_T$ is linear for all $T \in S_h, \varphi|_{\partial\Omega} = 0\}$ be the finite element space we are working with. The corresponding discrete inner product is defined by

$$(\chi, \varphi)_h = \sum_{T\in S_h} \int_T \Pi_h(\chi\varphi)dx = \sum_{T\in S_h} \frac{meas(T)}{d+1} \sum_{k=0}^{d} \chi(A_k^T)\varphi(A_k^T), \tag{2.2}$$

for any piecewise linear continuous functions χ, φ. Π_h stands for the local linear interpolation operator and $A_k^T, k = \overline{0,d}$ are the vertices of the element T. On V_h, this inner product is equivalent to $(\cdot,\cdot)$.

In our approach the maximum principle plays a crucial role. Hence the *acuteness* of S_h is assumed. This preserves the maximum principle in the fully discrete case ([4]).

A usual definition of the solution for (1.1) is

Definition. By a solution u of the problem in (1.1) we mean that

$$u \in L^\infty(0,T;L^2(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)), \quad \beta(u) \in L^2(0,T;H_0^1(\Omega)), \quad u(0) = u_0$$

and for all $\varphi \in H_0^1(\Omega)$ the equation

$$(\partial_t u(t), \varphi) + (\nabla\beta(u(t)), \nabla\varphi) = 0 \tag{2.3}$$

holds for a.e. $t \in (0,T)$.

Under the above assumptions, it is well known that this problem has a unique solution (see [9], [1] and the references therein).

In the rest of this paper we show the effectiveness of the schemes in (1.2), (1.5). The main theoretical results are

Theorem. Under the assumptions (A1) and (A2), if u is the solution of (1.1), then

$$\|\beta(u) - \theta_\Delta\|^2_{L^2(Q_T)} \le C\left[\tau + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2 + \beta^2(\varepsilon)\right],$$

where $\theta_\Delta(t) = \theta^k$ for $t \in (t_{k-1}, t_k]$, $k = \overline{1,n}$ and θ^k solves for each k the equation in (1.2).

Theorem. If, additionally, β' is Lipschitz continuous and θ^k is the solutions of (1.5), $k = \overline{1,n}$, then

$$\|\beta(u) - \theta_\Delta\|^2_{L^2(Q_T)} \le C\left[\tau + \tau C_2(\varepsilon)^2 + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2 + \beta^2(\varepsilon)\right].$$

Remark 2.1. If β satisfies $\varepsilon \cdot \inf\{\beta'(x) : x \in [\varepsilon, M]\} \ge C\beta(\varepsilon)$, which is true if $\beta(u) = u^m$ with $m \ge 1$, then the last two terms in the above error estimates become superfluous.

Remark 2.2. As in (1.3), the initial data are perturbed only locally. This is essential for the method to be efficient. However, the above results remains valid also for a global perturbation (for example, $v_0 \equiv 1$).

3. Linearization of the implicit scheme

In this section we study the nonlinear problem in (1.2) and its fully discrete counterpart. For each k a Dirichlet problem of a nonlinear elliptic equation is obtained, which will be written in a certain variational form. To this end, we define

$$V = \left\{\varphi \in H_0^1(\Omega) : 0 \le \varphi \le \beta(M) - \beta(\varepsilon), \quad a.e.\right\}$$

and for $\psi \in V$ and $\varphi \in H_0^1(\Omega)$,

$$\begin{aligned} a(\psi, \varphi) = & \quad (\beta^{-1}(\psi + \beta(\varepsilon)) - \varepsilon, \varphi) + \tau \langle \psi, \varphi \rangle, \\ f_k(\varphi) = & \quad \left(\beta^{-1}(\theta^k) - \varepsilon, \varphi\right), \end{aligned}$$

where θ^k will be shown to be in $\beta(\varepsilon) + V$. We note that V is convex and closed due to the Riesz theorem. Having defined these functionals, the variational formulation of the problem in (1.2) reads

$$a(\theta^{k+1} - \beta(\varepsilon), \varphi) = f_k(\varphi), \qquad \text{for all } \varphi \in H_0^1(\Omega). \tag{3.1}$$

Recall from (A2) that $\theta^0 \in \beta(\varepsilon) + V$. Furthermore, assume that $\theta^k \in \beta(\varepsilon) + V$ $(k \ge 0)$ is given. $\theta^{k+1} \in \beta(\varepsilon) + V$ can be obtained by an iteration scheme, which is based on the variational formulation in (3.1) instead of the numerical scheme in (1.2). To do so, let

$$K \ge C_2(\varepsilon) \tag{3.2}$$

be a constant and define, for $\psi, \varphi \in H_0^1(\Omega)$ and $\underline{\psi} \in V$,

$$\begin{aligned} a_K(\psi, \varphi) = & \quad K(\psi, \varphi) + \tau \langle \psi, \varphi \rangle, \\ f_K(\underline{\psi}; \varphi) = & \quad K(\underline{\psi}, \varphi) + (\beta^{-1}(\theta^k) - \beta^{-1}(\underline{\psi} + \beta(\varepsilon)), \varphi). \end{aligned}$$

The iteration scheme is induced through an operator on V. For this purpose, we apply the Lax-Milgram lemma to get a unique $T\psi \in H_0^1(\Omega)$ satisfying

$$a_K(T\psi, \varphi) = f_K(\psi; \varphi) \qquad \forall \varphi \in H_0^1(\Omega) \tag{3.3}$$

for each $\psi \in V$. This is a linear problem and therefore can be solved easily. In this way, we have defined an operator T from V to $H_0^1(\Omega)$. In fact, we will show below that T is a contraction mapping on the closed set V with an appropriate norm. Thus, the iteration scheme is well defined in the following way

$$\psi^{i+1} = T\psi^i \tag{3.4}$$

for $i \ge 0$ and $\psi^0 \in V$. A convenient choice of ψ^0 is $\psi^0 = \theta^k - \beta(\varepsilon)$. θ^{k+1} will be taken to be

$$\theta^{k+1} = \beta(\varepsilon) + \lim_{i \to \infty} \psi^i.$$

The existence of $\lim_{i\to\infty} \psi^i$ and θ^{k+1} in $\beta(\varepsilon) + V$ can be immediately seen by applying the fixed point theorem to T. This statement is supported by the following lemma

Lemma 3.1. *Assume (A1) and $\theta^k \in \beta(\varepsilon) + V$. Then $TV \subset V$ and there is a norm on $H_0^1(\Omega)$ equivalent to the usual one, such that T is contractive on the closed set V.*

Note that we do not assume the smallness of the time step τ.

Proof. We give here only a sketch of the proof, details can be found in [18]. The first part relies on the maximum principle, which is fulfilled by the problem in (3.3). For the second part note that $\|\varphi\|_K \equiv \sqrt{a_K(\varphi,\varphi)}$ is a norm on $H_0^1(\Omega)$. With this new norm, T is a contraction mapping on the closed subset V, as follows, for any $\psi_1, \psi_2 \in V$, from (3.3)

$$\begin{aligned} |a_K(T\psi_1 - T\psi_2, \varphi)| = &\ |K(\psi_1 - \psi_2, \varphi) - (\beta^{-1}(\psi_1 + \beta(\varepsilon)) - \beta^{-1}(\psi_2 + \beta(\varepsilon)), \varphi)| \\ = &\ |((K - (\beta^{-1})'(\theta))(\psi_1 - \psi_2), \varphi)| \\ \leq &\ \left(1 - \frac{C_1}{K}\right) \|\psi_1 - \psi_2\|_K \|\varphi\|_K. \end{aligned}$$

By taking $\varphi = T\psi_1 - T\psi_2$ the proof is completed. □

Remark 3.1. As resulting from the above proof, the convergence rate of the iteration in (3.4) is of order $1 - O(1/C_2(\varepsilon))$, which is close to 1 when ε is small. With the full discretization, an alternative iteration scheme has a better convergence rate, but depending on τ (see (3.5)).

Remark 3.2. The above results show that the problem in (3.1) has a unique solution in V. But the strict monotonicity of a can be easily proven, so the uniqueness holds in the whole $H_0^1(\Omega)$.

Remark 3.3. Lemma 3.1 implies also a maximum principle for the solution of the nonlinear problem in (3.1).

Remark 3.4. Only at $u = 0$, the original equation in (1.1) may be degenerate, that is, $\beta'(0) = 0$. The above maximum principle guarantees that if the (initial and boundary) data are away from 0, then the solution stays always away from 0. This is just the underlying idea of our approach to treat the degeneracy.

Remark 3.5. The idea of our iteration is based on the operator T, which has been used in [16], pp. 96. But we show that T is a contraction mapping, at least for the present problems, in the setting defined in the proof of Lemma 3.1. This fact makes the proof simpler, while the monotonicity of T can lead to an alternative one similar to that in [16].

As mentioned in Remark 3.1, the iteration scheme discussed up to now has a reduced convergence rate. The following iteration scheme suites better to practical purposes

$$\begin{aligned} &\bar\theta^i \in \beta(\varepsilon) + H_0^1(\Omega), \\ &(\sigma(\bar\theta^{i-1}, \theta^k)(\bar\theta^i - \theta^k), \varphi) + \tau < \bar\theta^i, \varphi >= 0, \\ &\sigma(\bar\theta^i, \theta^k) = \int_0^1 \left(\beta^{-1}\right)'(s\bar\theta^i + (1-s)\theta^k)ds \end{aligned} \tag{3.5}$$

for all $\varphi \in H_0^1(\Omega)$ and $i \geq 1$, where $\bar\theta^0 = \theta^k, \sigma(\bar\theta^0, \theta^k) = \left(\beta^{-1}\right)'(\theta^k)$.

For each $i > 0$, the discrete formulation of the problem in (3.5) reads

$$\begin{aligned} &\bar\theta_h^i \in \beta(\varepsilon) + V_h, \\ &(\sigma(\bar\theta_h^{i-1}, \theta_h^k)(\bar\theta_h^i - \theta_h^k), \varphi)_h + \tau < \bar\theta_h^i, \varphi >= 0 \end{aligned} \tag{3.6}$$

for all $\varphi \in V_h$, where $\bar\theta_h^0 = \theta_h^k, \sigma(\bar\theta_h^0, \theta_h^k) = \left(\beta^{-1}\right)'(\theta_h^k)$ and $\theta_h^k \in \beta(\varepsilon) + V_h$.

The resulting linear sistem is

$$(M_h^{i-1} + \tau A_h)\bar\vartheta^i = M_h^{i-1}\vartheta^k, \tag{3.7}$$

where $\bar{\vartheta}^i, \vartheta^k$ are vectors containing the nodal values of $\bar{\theta}^i_h, \theta^k_h$. A_h represents the stiffness matrix, while M^i_h is a diagonal matrix obtained from the lumped mass one by multiplication with the corresponding nodal values of $\sigma(\bar{\theta}^i_h, \theta^k_h)$. Because of the assumptions we have made on the decomposition S_h, $M^i_h + \tau A_h$ is a M-matrix. Therefore, if we have $\theta^k_h \in \beta(\varepsilon) + V_h$, the same holds for the entire array $\{\bar{\theta}^i_h, i \geq 0\}$ and the equations in (3.6) make sense.

Under the aditional assumptions on β', for small time steps τ, subtracting the equations in (3.7) for i and $i-1$, the convergence of the iteration can be shown by proving that there exists a positive constant $C(h,\varepsilon)$ such that $\|\bar{\vartheta}^i - \bar{\vartheta}^{i-1}\|_\infty < \tau C(h,\varepsilon)\|\bar{\vartheta}^{i-1} - \bar{\vartheta}^{i-2}\|_\infty$ for all $i > 1$. The details are skipped here. By virtue of this convergence we can take $\theta^{k+1}_h = \lim_{i\to\infty} \bar{\theta}^i_h$. It is easy to see that θ^{k+1}_h solves the equation

$$(\beta^{-1}(\theta^{k+1}_h) - \beta^{-1}(\theta^k_h), \varphi)_h + \tau < \theta^{k+1}_h, \varphi >= 0 \qquad \text{for all } \varphi \in V_h, \tag{3.8}$$

which is equivalent to the one in (3.6). In the numerical experiments we have adopted mainly this last iteration.

4. Error estimates

In this section we will show the convergence of our approach by proving error estimates for the schemes proposed in (1.2), (1.5) written in a variational form. For the nonlinear scheme this was given in (3.1), while in the linearized case it reads

$$(\sigma_k(\theta^{k+1} - \theta_k), \varphi) + \tau < \theta^{k+1}, \varphi >= 0 \tag{4.1}$$

for all $\varphi \in H^1_0(\Omega)$ and $k \geq 1$, where $\sigma_k = (\beta^{-1})'(\theta^k)$ and $\theta^{k+1} \in V$ (with V defined in the previuos section). The maximum principle holds for the equations in (4.1) too, thus for any $k \geq 0$ we have $\theta_k \in V$ and therefore the linearized scheme makes sense.

The following theorem establishes the stability of the approximation.

Theorem 4.1. *Assume (A1) and (A2). Then, for $p < n$ we have*

$$\tau\left(|\theta^{p+1}|_1^2 + \sum_{k=0}^{p} |\theta^{k+1} - \theta^k|_1^2\right) + 2C_1 \sum_{k=0}^{p} \left\|\theta^{k+1} - \theta^k\right\|^2 \leq \tau|\theta^0|_1^2, \tag{4.2}$$

if θ^k solves the problem in (3.1), while in the linearized case

$$\tau\left(|\theta^{p+1}|_1^2 + \sum_{k=0}^{p} |\theta^{k+1} - \theta^k|_1^2\right) + 2 \sum_{k=0}^{p} \left\|(\theta^{k+1} - \theta^k)\sqrt{\sigma_k}\right\|^2 \leq \tau|\theta^0|_1^2. \tag{4.3}$$

Proof. By taking $\phi = \theta^{k+1} - \theta^k \in H^1_0(\Omega)$ in (3.1), respectively (4.1), and summing up for $k = \overline{0,p}$ we get the desired inequalities. □

Besides we need the following lemma.

Lemma 4.2. *For $u \geq 0$, define $u^\varepsilon = \beta^{-1}(\beta(u) + \beta(\varepsilon))$. If β satisfies (A1), then*

$$0 < u^\varepsilon - u \leq \varepsilon + C_2(\varepsilon)\beta(\varepsilon), \tag{4.4}$$

where $C_2(\varepsilon)$ is given in (2.1).

Now we turn to the theorems stated in the beginning.

Theorem 4.3. *Under the assumptions (A1) and (A2), if u is the weak solution of the problem in (1.1) and θ^k solves for each $k>0$ the equation in (3.1), then*

$$\sum_{k=1}^n \int_{t_{k-1}}^{t_k} \|\beta(u(t))-\theta^k\|^2 dt \leq C\left\{\tau+\varepsilon^2+\left(C_2(\varepsilon)\beta(\varepsilon)\right)^2+\beta^2(\varepsilon)\right\}.$$

Proof. We follow the approach in [11], which was also used in [14], [10] and [6]. Integrating the equation in (2.3) over (t_{j-1},t_j), subtracting (3.1) from the result and summing up for $j=\overline{1,k}$ yields

$$(u(t_k)-\beta^{-1}(\theta^k),\varphi)-(u(t_0)-\beta^{-1}(\theta^0),\varphi)+\sum_{j=1}^k\left\langle\int_{t_{j-1}}^{t_j}\left[\beta(u(t))-\theta^j\right]dt,\varphi\right\rangle=0.$$

Taking $\varphi=\int_{t_{k-1}}^{t_k}\beta(u)dt+\tau(\beta(\varepsilon)-\theta^k)$ and summing up this relation for $k=\overline{1,n}$ we get

$$\sum_{k=1}^n\left(u(t_k)-\beta^{-1}(\theta^k),\int_{t_{k-1}}^{t_k}[\beta(u^\varepsilon)-\theta^k]dt\right)-\sum_{k=1}^n\left(u(t_0)-\beta^{-1}(\theta^0),\int_{t_{k-1}}^{t_k}[\beta(u^\varepsilon)-\theta^k]dt\right)$$

$$+\sum_{k=1}^n\sum_{j=1}^k\left\langle\int_{t_{j-1}}^{t_j}[\beta(u)-\theta^j]dt,\int_{t_{k-1}}^{t_k}[\beta(u)-\theta^k+\beta(\varepsilon)]dt\right\rangle=0. \tag{4.5}$$

Because of the positivity of the last term in (4.5), which can be easily shown following the papers cited above, we only need to concentrate on the first two ones, denoted by T_1 and T_2. T_1 can be decomposed as follows.

$$\begin{aligned} T_1 = &\sum_{k=1}^n\int_{t_{k-1}}^{t_k}\left(u(t_k)-u(t),\beta(u^\varepsilon(t))-\theta^k\right)dt\\ &+\sum_{k=1}^n\int_{t_{k-1}}^{t_k}\left(u(t)-u^\varepsilon(t),\beta(u^\varepsilon(t))-\theta^k\right)dt\\ &+\sum_{k=1}^n\int_{t_{k-1}}^{t_k}\left(u^\varepsilon(t)-\beta^{-1}(\theta^k),\beta(u^\varepsilon(t))-\theta^k\right)dt\\ \equiv\ &T_{11}+T_{12}+T_{13}.\end{aligned}$$

As in [11], we get

$$\begin{aligned}|T_{11}|=&\left|\sum_{k=1}^n\int_{t_{k-1}}^{t_k}\left(u(t_k)-u(t),\beta(u^\varepsilon(t))-\theta^k\right)dt\right|\\ \leq&\tau\left\|\partial_t u\right\|_{L^2(0,T;H^{-1}(\Omega))}\left\|\beta(u)-\theta_\Delta+\beta(\varepsilon)\right\|_{L^2(0,T;H_0^1(\Omega))}\leq C\tau.\end{aligned}\tag{4.6}$$

Employing Lemma 4.2, the following holds for any $\delta>0$

$$|T_{12}|\leq\sum_{k=1}^n\int_{t_{k-1}}^{t_k}\|u^\varepsilon-u\|_\infty\|\beta(u^\varepsilon)-\theta^k\|_{L^1}\leq(\varepsilon+C_2(\varepsilon)\beta(\varepsilon))\sum_{k=1}^n\int_{t_{k-1}}^{t_k}\|\beta(u^\varepsilon)-\theta^k\|_{L^1}$$

$$\leq\frac{\varepsilon+C_2(\varepsilon)\beta(\varepsilon)}{\delta}\sum_{k=1}^n\int_{t_{k-1}}^{t_k}\|\beta(u^\varepsilon)-\theta^k\|^2+\delta C\left(\varepsilon+C_2(\varepsilon)\beta(\varepsilon)\right). \tag{4.7}$$

Now, because of (A1), the estimate of T_{13} is a simple matter.

$$\begin{aligned} T_{13} =& \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \int_\Omega (u^\varepsilon - \beta^{-1}(\theta^k))(\beta(u^\varepsilon) - \theta^k) dx dt \\ \geq& C_1 \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u^\varepsilon(t)) - \theta^k\|^2 dt. \end{aligned} \tag{4.8}$$

As for T_{12} we arrive at

$$|T_2| \leq \frac{\varepsilon + C_2(\varepsilon)\beta(\varepsilon)}{\delta} \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u^\varepsilon) - \theta^k\|^2 + \delta C \left(\varepsilon + C_2(\varepsilon)\beta(\varepsilon)\right). \tag{4.9}$$

The inequalities in (4.6) - (4.9) in (4.5) yield

$$\begin{aligned} 0 \geq & \; -C\tau - 2\frac{\varepsilon + C_2(\varepsilon)\beta(\varepsilon)}{\delta} \textstyle\sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u^\varepsilon) - \theta^k\|^2 \\ & -2\delta C \left(\varepsilon + C_2(\varepsilon)\beta(\varepsilon)\right) + C_1 \textstyle\sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u^\varepsilon(t)) - \theta^k\|^2 dt. \end{aligned}$$

By a proper choice of δ we get

$$\sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u^\varepsilon) - \theta^k\|^2 \leq C \left\{\tau + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2\right\}. \tag{4.10}$$

Finally, notice that

$$\|\beta(u^\varepsilon) - \theta^k\| \geq \left| \|\beta(u) - \theta^k\| - C\beta(\varepsilon) \right|,$$

therefore

$$2\|\beta(u^\varepsilon) - \theta^k\|^2 \geq \|\beta(u) - \theta^k\|^2 - C\beta^2(\varepsilon).$$

This, together with (4.10), gives the desired result. □

Remark 4.1. The above error estimate show that the scheme in (1.2) behaves at least as good as the schemes in [11], [7] and [14]. In fact, the scheme in [11] is proved to have a convergence order $\tau^{1/2}$, while the scheme in [14] is of order $\tau^{1/6}$. The order $\tau^{1/4}$ is demonstrated in [6] for the method proposed in [7].

The next result concernes the linearized scheme in (4.1).

Theorem 4.4. *Under the conditions in Theorem 4.1, assuming additionally the Lipschitz continuity of β', if θ^k solves for each $k > 0$ the equation in (4.1), then*

$$\textstyle\sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u(t)) - \theta^k\|^2 dt \leq C \left\{\tau + \varepsilon^2 + \tau C_2(\varepsilon)^2 + (C_2(\varepsilon)\beta(\varepsilon))^2 + \beta^2(\varepsilon)\right\}.$$

Proof. The proof is similar to the one above. Following the same steps we arrive to

$$\begin{aligned} & \sum_{k=1}^{n} \sum_{i=1}^{k} \left(u(t_i) - u(t_{i-1}) - \sigma_{i-1}(\theta^i - \theta^{i-1}), \int_{t_{k-1}}^{t_k} [\beta(u^\varepsilon) - \theta^k] dt \right) \\ & + \sum_{k=1}^{n} \sum_{j=1}^{k} \left\langle \int_{t_{j-1}}^{t_j} [\beta(u) - \theta^j] dt, \int_{t_{k-1}}^{t_k} [\beta(u) - \theta^k + \beta(\varepsilon)] dt \right\rangle = 0. \end{aligned} \tag{4.11}$$

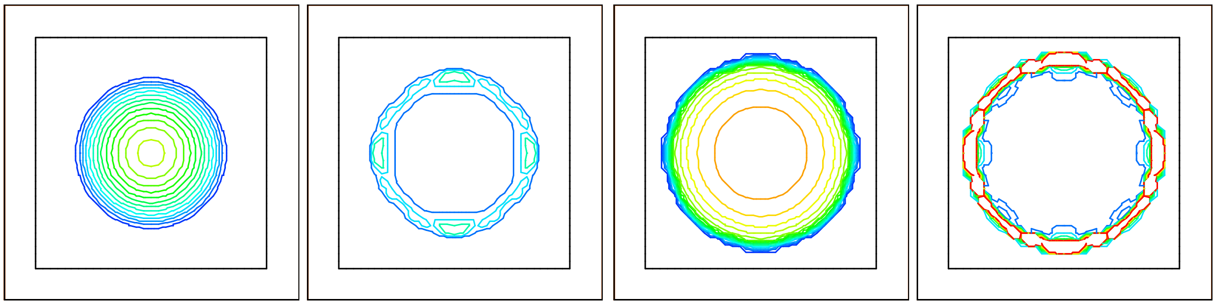

FIGURE 1. Solutions and absolute errors, t=1.0s, m=2 and m=6

Subtracting and adding in the first sums $\beta^{-1}(\theta^i) - \beta^{-1}(\theta^{i-1}) = \sigma(\theta^i, \theta^{i-1})(\theta^i - \theta^{i-1})$ (where $\sigma(\theta^i, \theta^{i-1})$ is defined in (3.5)) we arrive to a relation similar to the one in (4.5) up to the additional term

$$T \equiv \sum_{k=1}^{n} \sum_{i=1}^{k} \left((\sigma(\theta^i, \theta^{i-1}) - \sigma_{i-1})(\theta^i - \theta^{i-1}), \int_{t_{k-1}}^{t_k} [\beta(u^\varepsilon) - \theta^k] dt \right).$$

The Lipschitz continuity of β' yields

$$|\sigma(\theta^i, \theta^{i-1}) - \sigma_{i-1}| \le L_{\beta'} C_2(\varepsilon)^2 \sigma_{i-1} |\theta^i - \theta^{i-1}|.$$

This, together with the stability result in Theorem 4.1 leads to

$$|T| \le C(\tau C_2(\varepsilon))^2$$

and the rest of the proof follows as in the previous theorem. □

Remark 4.2. The linearized case has a reduced convergence order, depending on the properties of β' and the choice of ε. For example, if $\beta(u) = u^m$ we get an order of only $\tau^{1/2m}$. But practical computations show that the above estimates are quite pesimistic.

Finally we consider the convergence of the fully discretized scheme in (3.8). The proof of the following result is based on the ideeas in [5] and won't be given here.

Theorem 4.5. *Under the assumptions (A1) and (A2), if u is the weak solution of the problem in (1.1) and θ_h^k solves for each $k > 0$ the equation in (3.8) with $h = O(\tau)$, then*

$$\sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \|\beta(u(t)) - \theta_h^k\|^2 dt \le C \left\{ \tau + \varepsilon^2 + (C_2(\varepsilon)\beta(\varepsilon))^2 + \beta^2(\varepsilon) \right\}.$$

Remark 4.3. The restriction on the mesh-size h appears only theoretically. Different approaches in proving the convergence of the fully discrete scheme does not assume any relation between the discretization parameters, but the estimates depend aditionally on $C_2(\varepsilon)$ resulting a lower order of convergence compared with the semidiscrete case. This was avoided in the previous result.

The maximum principle based algorithm was tested on the porous media equation in two dimensions. With appropriately chosen initial data, an exact solution was given by G. Barenblatt. The cases $m = 2$ and $m = 6$ are considerred. The package *UG* [2], available at http://dom.ica3.uni-stuttgart.de/ is used in the computations. In both cases the solution u and the absolute error are diplayed using contour levels. No iterations are performed. The time step is $\tau = 0.01$, while for the spatial discretization the coarsest level contains a single interior node and is refined four times. The errors appear essentialy near the free boundary and range until 0.03 in the first case, respectively 0.2 in the second one.

Acknowledgements. We would like to express our thanks to Prof. W. Jäger for his support and suggestions and to Prof. J. Kačur for his remarks. We are indebted to Dr. N. Neuß for many useful discussions and for the kind assistance with *UG*. This work was supported by the Deutsche Forschungsgemeinschaft through SFB 359.

References

[1] H. W. Alt & S. Luckhaus, Quasilinear Elliptic-Parabolic Differential Equations, *Math. Z.*, 183(1983), 311 - 341.
[2] P. Bastian & K. Birken & K. Johannsen & S. Lang & N. Neuss & H. Rentz-Reichert & C. Wieners, UG - A Flexible Software Toolbox for Solving Partial Differential Equations, to appear in *Computing and Visualization in Science*.
[3] P. G. Ciarlet, The Finite Element Method for Elliptic Problems, *North-Holland, Amsterdam*, 1978.
[4] P. G. Ciarlet & P. A. Raviart, Maximum Principle and Uniform Convergence for the Finite Element Method, *Comput. Methods Appl. Mech. Engrg.*, 2(1973), 17 - 31.
[5] C. M. Elliott, Error Analysis of the Enthalpy Method for the Stefan Problem, *IMA J. Numer. Anal.*, 7(1987), 61 - 71.
[6] A. Handlovičová, Error Estimates of a Linear Approximation Scheme for Nonlinear Diffusion Problems, *Acta Math. Univ. Comenianae*, 61(1992), 27 - 39.
[7] W. Jäger & J. Kačur, Solution of Porous Medium Type Systems by Linear Approximation Schemes, *Numer. Math.*, 60(1991), 407 - 427.
[8] W. Jäger & J. Kačur, Solution of Doubly Nonlinear and Degenerate Parabolic Problems by Relaxation Schemes, *M^2AN (Math. Model. Numer. Anal.)*, 29(1995), 605 - 627.
[9] J. W. Jerome, Approximation of Nonlinear Evolution Systems, *Academic Press, New York*, 1983.
[10] J. Kačur & A. Handlovičová & M. Kačurová, Solution of Nonlinear Diffusion Problems by Linear Approximation Schemes, *SIAM J. Numer. Anal.*, 30(1993), 1703 - 1722.
[11] E. Magenes & R. H. Nochetto & C. Verdi, Energy Error Estimates for a Linear Scheme to Approximate Nonlinear Parabolic Problems, *M^2AN (Math. Model. Numer. Anal.)*, 21(1987), 655 - 678.
[12] R. H. Nochetto & C. Verdi, Approximation of Degenerate Parabolic Problems Using numerical Integration, *SIAM J. Numer. Anal.*, 25(1988), 784 - 814.
[13] O. A. Oleinik & A. S. Kalashnikov & Y.-L. Zhou, The Cauchy Problem and Boundary Value Problems for Equations of the Type of Unsteady Filtration, *Izv. Akad. Nauk SSSR Ser. Mat.*, 22(1958), 667 - 704.
[14] M. Slodička, Solution of Nonlinear Parabolic Problems by Linearization, *Preprint M3-92* (1992), Comenius University Bratislava, Faculty of Mathematics and Physics.
[15] M. Slodička, On a Numerical Approach to Nonlinear Degenerate Parabolic Problems, *Preprint M6-92* (1992), Comenius University Bratislava, Faculty of Mathematics and Physics.
[16] J. Smoller, Shock Waves and Reaction-Diffusion Equations, *Springer-Verlag, New York*, 1983.
[17] V. Thomée, Galerkin Finite Element Methods for Parabolic Problems, Lecture Notes in Math. 1054, *Springer-Verlag, Berlin*, 1984.
[18] W. A. Yong & I. S. Pop, A Numerical Approach to Porous Medium Equations, *Preprint 96-50 (SFB 359)* (1996), IWR, University of Heidelberg.

§Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany

†Faculty of Mathematics, "Babeş-Bolyai" University, M. Kogălniceanu 1, 3400 Cluj-Napoca, Romania

E-mail address: §`sorin@oasis.iwr.uni-heidelberg.de`, §`yong@esprit.iwr.uni-heidelberg.de`, †`ispop@cs.ubbcluj.ro`