An Error Analysis for Galerkin Approximations to an Equation of Mixed Elliptic-Parabolic Type

Todd Arbogast

CRPC-TR90092
October, 1990

Center for Research on Parallel Computation
Rice University
P.O. Box 1892
Houston, TX 77251-1892

AN ERROR ANALYSIS FOR GALERKIN APPROXIMATIONS TO AN EQUATION OF MIXED ELLIPTIC-PARABOLIC TYPE

TODD ARBOGAST

Abstract. Error estimates are derived for both continuous and discrete time Galerkin approximations to a highly nonlinear flow equation which is formally parabolic. It has a nonlinear, monotone accumulation term with a Hölder continuous derivative that is allowed to vanish, in which case the equation becomes elliptic in nature. This equation includes Richard's equation, which models unsaturated and saturated groundwater flow. For smooth solutions, upper bounds are obtained for $L^\infty(L^2)$ and $L^2(H^1)$-norms of the error as well as for a certain nonlinear form that gives a measure of the error's size. The rates of convergence are determined by the nature of the accumulation term's nonlinearity; in the strictly parabolic case where the accumulation term is C^2 and has its first derivative positively bounded from above and below, these rates are of optimal order. The rate of convergence of the nonlinear form is generally of optimal order.

1. Introduction. In this paper we present an analysis of the error in Galerkin approximations to an equation of the following form:

$$\begin{aligned}
&(1.1) && \partial_t\theta(x,u) - \nabla\cdot\left[a(x,\theta(x,u))\mathbf{k}(x)\nabla u + \mathbf{b}(x,t,\theta(x,u))\right] \\
& && \qquad + c(x,\theta(x,u))u \\
& && \qquad = f(x,t,\theta(x,u)) \qquad \text{in } \Omega\times J, \\
&(1.2a) && u = u_{\mathrm{D}}(x,t) \qquad \text{on } \Gamma_{\mathrm{D}}\times J, \\
&(1.2b) && -\left[a(x,\theta(x,u))\mathbf{k}(x)\nabla u + \mathbf{b}(x,t,\theta(x,u))\right]\cdot\nu - \lambda(x,t)u \\
& && \qquad = g(x,t) \qquad \text{on } \Gamma_{\mathrm{N}}\times J, \\
&(1.3) && u = u^0(x) \qquad \text{on } \Omega\times 0,
\end{aligned}$$

where $J=(0,T]$ and function evaluations are nonlinear in $\theta(u)$ only. The function $\theta(u)$ is assumed to be monotone nondecreasing and uniformly $C^{1,\gamma}$ in $u\in\mathbb{R}$; that is, $\partial_u\theta(u)$ is bounded from above and Hölder continuous, but $\partial_u\theta(u)$ is not necessarily positively bounded from below. As one consequence, the problem is parabolic in nature only in regions

This work was supported by the National Science Foundation under grant DMS–8905505; it was done while the author was visiting the University of Houston and Rice University, both of Houston, Texas.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

where the solution u gives rise to a nonconstant $\theta(u)$; otherwise, it is elliptic in nature, with some free boundary separating the two regions.

Equation (1.1)–(1.3) arises, for example, with $c(\theta) \equiv 0$ in the modeling of groundwater flow (and it is often called *Richard's* equation) [2], [7]. The saturated flow of incompressible groundwater is modeled when the moisture content $\theta(u)$ is at its maximum (saturated) value, and the flow is of elliptic type. The unsaturated flow is given by the parabolic equation when $\theta(u)$ is an increasing function of u; $\theta(u)$ represents the accumulation of water due to highly nonlinear capillary effects. The advantage of the form (1.1)–(1.3) is that the solution can be approximated without a direct approximation of the free boundary, which is poorly understood (see, e.g., [1], [6], [8], [10], and [11]). This fact has been exploited computationally by, for example, Celia *et al.* [3] and Knabner [12].

Wheeler and Dupont [17, Section 3.4] gave a continuous time error analysis of Galerkin approximations to a version of this equation when the problem is smooth and strictly parabolic, i.e., when $\partial_u\theta(u)$ is Lipschitz and positively bounded from below. Rachford [13] considered the problem in nonconservative form (i.e., with $\partial_t\theta(u)$ written as $\partial_u\theta(u)\partial_t u$). He gave a discrete time error analysis of a Crank-Nicolson type scheme, again when the problem is smooth and strictly parabolic. These earlier results do not in any direct way extend to the present problem.

With U representing the approximate continuous time solution, we will obtain upper bounds on two norms and a nonlinear form,

$$\|\theta(u) - \theta(U)\|_{L^\infty(J;L^2(\Omega))}, \quad \|u - U\|_{L^2(J;H^1(\Omega))},$$
$$\text{and} \quad \int_0^T \big(\theta(u) - \theta(U), u - U\big)\, dt,$$

when u is sufficiently smooth ($W^{k,p}$ is the standard Sobolev space of k-differentiable functions in L^p, and $H^k = W^{k,2}$). For the type of accumulation θ that we consider, note that for some $q, Q > 0$ and any v and w,

$$q|\theta(v) - \theta(w)| \le [(\theta(v) - \theta(w))(v - w)]^{1/2} \le Q|v - w|,$$

so our nonlinear form indeed measures in some sense how much w deviates from v. As we will see (Theorem 1, Section 5), the rate of convergence of U to u in either of the norms depends on the rate at which the graph of θ may become flat and on the Hölder constant of $\partial_u\theta$; that is, upon β and γ defined in (A16)–(A17) below. The nonlinear form estimate is better behaved, converging to zero at the optimal rate (up to a possible restriction on the spatial dimension). Somewhat better

results are obtained in the special cases of strictly parabolic flow and of monotone flow (Theorems 2-3, Section 5). The discrete time case has analogous results (Theorems 4-6, Section 6).

2. The Galerkin procedures. Before we present the general Galerkin approximation procedures considered in this paper, we define a weak form of the problem. Let $(\cdot\,,\,\cdot)$ denote the $L^2(\Omega)$ inner product, $\langle\cdot\,,\,\cdot\rangle$ denote the $L^2(\Gamma_{\mathrm{N}})$ inner product, and

$$\mathcal{V} = \{v \in H^1(\Omega) : v|_{\Gamma_{\mathrm{D}}} = 0\}.$$

Then $u \in \mathcal{V} + u_{\mathrm{D}}$ satisfies

$$\begin{aligned}(2.1)\qquad &\big(\partial_t\theta(u), v\big) + \big(a(\theta(u))\mathbf{k}\nabla u + \mathbf{b}(\theta(u)), \nabla v\big)\\ &\qquad + \langle\lambda u, v\rangle + \big(c(\theta(u))u, v\big)\\ &\qquad = \big(f(\theta(u)), v\big) - \langle g, v\rangle \qquad \text{for all } v \in \mathcal{V}.\end{aligned}$$

For $h > 0$, let $\{\mathcal{V}_h\}_h$ be a family of finite dimensional subspaces of $\mathcal{V}$. (For simplicity we do not consider nonconforming Galerkin methods in this paper.) Now the continuous time Galerkin procedure is the following: Find $U \in \mathcal{V}_h + u_{\mathrm{D}}$ such that

$$\begin{aligned}(2.2)\qquad &\big(\partial_t\theta(U), v\big) + \big(a(\theta(U))\mathbf{k}\nabla U + \mathbf{b}(\theta(U)), \nabla v\big)\\ &\qquad + \langle\lambda U, v\rangle + \big(c(\theta(U))U, v\big)\\ &\qquad = \big(f(\theta(U)), v\big) - \langle g, v\rangle \qquad \text{for all } v \in \mathcal{V}_h,\end{aligned}$$

where $U(x,0)$ is defined to be some reasonable approximation to $u^0(x)$ in $\mathcal{V}_h + u_{\mathrm{D}}(\cdot\,,0)$ (we tacitly assume that $u^0(x) = u_{\mathrm{D}}(x,0)$ for $x \in \Gamma_{\mathrm{D}}$).

For discrete time, let $0 = t_0 < t_1 < \cdots < t_N = T$ be a given partition of J, define $\Delta t^n = t_n - t_{n-1}$ and $\Delta t = \max_n \Delta t^n$, let v^n denote $v(t_n)$ for any function v of time, and denote the backward difference operator by

$$\partial v^n = \frac{v^n - v^{n-1}}{\Delta t^n}.$$

The discrete time Galerkin procedure is as follows: For $n = 1, 2, ..., N$, find $U^n \in \mathcal{V}_h + u_{\mathrm{D}}^n$ such that

$$\begin{aligned}(2.3)\qquad &\big(\partial\theta(U)^n, v\big) + \big(a(\theta(U^n))\mathbf{k}\nabla U^n + \mathbf{b}^n(\theta(U^n)), \nabla v\big)\\ &\qquad + \langle\lambda^n U^n, v\rangle + \big(c(\theta(U^n))U^n, v\big)\\ &\qquad = \big(f^n(\theta(U^n)), v\big) - \langle g^n, v\rangle \qquad \text{for all } v \in \mathcal{V}_h,\end{aligned}$$

and as above U^0 is some reasonable approximation to u^0 in $\mathcal{V}_h + u_{\mathrm{D}}^0$.

REMARK 1: The existence and uniqueness of our approximate solutions can be established in a manner analogous to that for the true solution [10], [1]. (Uniqueness can also be shown by the techniques of this paper.) The fully discrete scheme (2.3) is computable if, for example, $\mathcal{V}_h$ is some standard finite element space and an iterative procedure such as Newton's method is used to solve approximately the nonlinear equations.

We close this section by making the following assumptions explicit. Only (A16) appears to be novel. Throughout the paper let ϵ, q, and Q denote generic positive constants that are independent of x, t, h, n, and the Δt^n, where ϵ can be taken to be as small as needed.

(A1) $\Omega \subset \mathbb{R}^d$ is a connected, bounded domain with a sufficiently regular boundary Γ (e.g., $\Gamma \in C^{1,1}$). Also, Γ is partitioned into two subdomains, Γ_{D} and Γ_{N}, with Γ_{D} having positive measure.

(A2) $\theta(x,u)$ is monotone nondecreasing in $u \in \mathbb{R}$ for each fixed $x \in \Omega$, uniformly Lipschitz in both x and u, and uniformly bounded from above and below.

(A3) $\mathbf{k}(x)$ is uniformly Lipschitz in $x \in \Omega$, and $\mathbf{k}$ is a uniformly positive-definite symmetric tensor.

(A4) $a(x,\theta)$ is uniformly positive, uniformly bounded, and uniformly Lipschitz in both x and θ; moreover, $|\nabla_x \partial_\theta a|$ and $|\partial_\theta^2 a|$ are uniformly bounded.

(A5) The vector $\mathbf{b}(x,t,\theta)$ is continuous in t and uniformly Lipschitz in θ, and $\sup_\theta |\mathbf{b}(\theta)| \in C^0(\bar{J}; L^2(\Omega))$.

(A6) $c(x,\theta) \geq 0$ is uniformly bounded and uniformly Lipschitz in θ.

(A7) $f(x,t,\theta)$ is continuous in t and uniformly Lipschitz in θ, and is such that $\sup_\theta |f(\theta)| \in C^0(\bar{J}; L^2(\Omega))$.

(A8) $u_{\mathrm{D}} \in C^0(\bar{J}; W^{1,\infty}(\Omega)) \cap W^{1,1}(J; L^1(\Omega))$.

(A9) $g \in C^0(\bar{J}; H^{-1/2}(\Gamma_{\mathrm{N}}))$.

(A10) $\lambda \in W^{1,\infty}(J; W^{1,\infty}(\Gamma_{\mathrm{N}}))$ and $\lambda \geq 0$.

(A11) $u^0 \in L^2(\Omega)$.

(A12) For some $r \geq 2$,

$$\inf_{\chi \in \mathcal{V}_h} \|v - \chi\|_1 \leq Q \|v\|_k h^{k-1} \qquad \text{for } 1 \leq k \leq r,$$

where $\|\cdot\|_k$ denotes the $H^k(\Omega)$-norm.

(A13) The following inverse inequality holds for $v \in \mathcal{V}_h$:

$$\|v\|_{L^4(\Omega)} \leq Q \|v\|_0 h^{-d/4}.$$

(A14) $u - u_{\mathrm{D}} \in H^1(J; H^r(\Omega))$ and $u \in W^{1,\infty}(J; W^{1,\infty}(\Omega))$.

(A15) $\|u^0 - U^0\|_0 \le Qh^r$.

(A16) On its domain of definition, uniformly in $x \in \Omega$, $(\partial_u\theta) \circ \theta^{-1}$ is Hölder continuous of order β, $0 < \beta \le 1$; that is, for any $v, w \in \mathbb{R}$,

$$\left|\partial_u\theta(v) - \partial_u\theta(w)\right| \le Q|\theta(v) - \theta(w)|^\beta.$$

(A17) $\partial_u\theta$ is uniformly Hölder continuous of order γ, $0 < \gamma \le 1$: For any $v, w \in \mathbb{R}$,

$$\left|\partial_u\theta(v) - \partial_u\theta(w)\right| \le Q|v - w|^\gamma.$$

(A18) $q\Delta t^n \le \Delta t^{n+1} \le Q\Delta t^n$ for all n and $\partial_t^2\theta(u) \in L^2(J; L^2(\Omega))$.

REMARK 2: Assumptions (A12)–(A13) hold for standard finite element spaces defined over quasi-uniform meshes [4]. Moreover, (A1) and (A12) imply that

$$\inf_{\chi \in \mathcal{V}_h} \|v - \chi\|_0 \le Q\|v\|_k h^k \qquad \text{for } 0 \le k \le r.$$

Assumption (A15) can be satisfied easily if $u^0 \in H^r(\Omega)$; for example, define U^0 to be an elliptic (see (4.1) below) or L^2 projection of u^0.

REMARK 3: Clearly (A2) and (A16) imply (A17) for some $\gamma \ge \beta$, and if $\partial_u\theta$ is positively bounded from below, then $\beta = \gamma$. As an example, suppose that $\theta(u) \propto 1 - |\min\{0, u\}|^\alpha$ for some $\alpha \in (1, \infty)$. Then (A16) holds with $\beta = (\alpha - 1)/\alpha$ and (A17) holds with $\gamma = \min(\alpha - 1, 1)$. Both β and γ measure the smoothness of the graph of θ, but β also measures in some sense the manner in which the graph of θ becomes constant wherever it does so.

3. Some stability results. In this section, we consider the question of overall stability of the two Galerkin procedures. We begin with an observation, easily verified as an exercise in calculus.

PROPOSITION 1. *Assume* (A2). *Then for any v and w in* $\mathbb{R}$,

$$(3.1) \quad \left(2\sup_u|\partial_u\theta|\right)^{-1}\left(\theta(v) - \theta(w)\right)^2 \le \int_w^v \left(\theta(\mu) - \theta(w)\right) d\mu \le \left(\theta(v) - \theta(w)\right)(v - w).$$

LEMMA 1. *Assume* (A1)–(A11), (A15). *If U is defined by* (2.2), *then*

$$(3.2) \quad \|U\|_{L^2(J;H^1(\Omega))} \le Q\Big\{\Big\|\sup_\theta|f(\theta)|\Big\|_{L^2(J;L^2(\Omega))} + \|g\|_{L^2(J;H^{-1/2}(\Gamma_N))} + \|u_D\|_{L^2(J;H^1(\Omega))} + \|u_D\|_{W^{1,1}(J;L^1(\Omega))} + \Big\|\sup_\theta|\mathbf{b}(\theta)|\Big\|_{L^2(J;L^2(\Omega))} + \|U^0\|_{L^1(\Omega)}\Big\}.$$

PROOF: In (2.2), let $v = U - u_{\mathrm{D}} \in \mathcal{V}_h$. The use of Poincaré's inequality and some manipulation result in

$$(3.3)\quad \begin{aligned}&(\partial_t\theta(U), U - u_{\mathrm{D}}) + q\|U\|_1^2 \\ &\quad\le Q\{\|\sup_\theta |\mathbf{b}(\theta)|\|_0^2 + \|\sup_\theta |f(\theta)|\|_0^2 + \|g\|_{H^{-1/2}(\Gamma_{\mathrm{N}})}^2 + \|u_{\mathrm{D}}\|_1^2\},\end{aligned}$$

where q and Q depend upon the bounds for $a\mathbf{k}$, λ, and c. Because

$$(3.4)\quad \begin{aligned}&\partial_t\theta(U)(U - u_{\mathrm{D}}) \\ &\quad= \partial_t\Big\{\int_{U^0}^{U} (\theta(U) - \theta(\mu))\, d\mu + \theta(U)(U^0 - u^D)\Big\} + \theta(U)\partial_t u_{\mathrm{D}},\end{aligned}$$

integration in time followed by an application of Proposition 1 gives the lemma. ▌

LEMMA 2. *Assume* (A1)–(A11), (A15). *If* U^n *is defined by* (2.3), *then*

$$(3.5)\quad \begin{aligned}&\Big\{\sum_{n=1}^{N} \|U^n\|_1^2\, \Delta t^n\Big\}^{1/2} \\ &\quad\le Q\Big\{\Big(\sum_{n=1}^{N} [\|\sup_\theta |f^n(\theta)|\|_0^2 + \|g^n\|_{H^{-1/2}(\Gamma_{\mathrm{N}})}^2 \\ &\qquad\qquad + \|u_{\mathrm{D}}^n\|_1^2 + \|\sup_\theta |\mathbf{b}^n(\theta)|\|_0^2]\Delta t^n\Big)^{1/2} \\ &\qquad + \|u_{\mathrm{D}}\|_{W^{1,1}(J;L^1(\Omega))} + \|U^0\|_{L^1(\Omega)}\Big\}.\end{aligned}$$

PROOF: In (2.3), let $v = U^n - u_{\mathrm{D}}^n \in \mathcal{V}_h$ to see as in the previous argument that

$$(3.6)\quad \begin{aligned}&(\partial\theta(U)^n, U^n - u_{\mathrm{D}}^n) + q\|U^n\|_1^2 \\ &\quad\le Q\{\|\sup_\theta |\mathbf{b}^n(\theta)|\|_0^2 + \|\sup_\theta |f^n(\theta)|\|_0^2 \\ &\qquad + \|g^n\|_{H^{-1/2}(\Gamma_{\mathrm{N}})}^2 + \|u_{\mathrm{D}}^n\|_1^2\}.\end{aligned}$$

In discrete form we have that

$$(3.7)\quad \begin{aligned}&\partial\theta(U)^n(U^n - u_{\mathrm{D}}^n) \\ &\quad= \partial\Big(\int_{U^0}^{U} (\theta(U) - \theta(\mu))\, d\mu + \theta(U)(U^0 - u^D)\Big)^n \\ &\qquad + \theta(U^{n-1})\partial u_{\mathrm{D}}^n + \frac{1}{\Delta t^n}\int_{U^{n-1}}^{U^n} (\theta(\mu) - \theta(U^{n-1}))\, d\mu.\end{aligned}$$

The last term above is nonnegative, so the proof is completed as above after multiplying through be Δt^n and summing on n. ▌

4. An elliptic projection. It has proven convenient in the analysis of parabolic Galerkin methods to analyze the error in two parts, the first between the actual solution and some elliptic projection, and the second between this elliptic projection and the approximate solution [17]. This will be convenient for our mixed problem as well. Hence we define an elliptic projection $\tilde{u} \in \mathcal{V}_h + u_{\mathrm{D}}$ for $u \in \mathcal{V} + u_{\mathrm{D}}$ by

$$(4.1)\qquad \big(a(\theta(u))\mathbf{k}\nabla(u-\tilde{u}), \nabla v\big) + \langle \lambda\,(u-\tilde{u}), v\rangle \\ + \big(c(\theta(u))(u-\tilde{u}), v\big) = 0 \qquad \text{for all } v \in \mathcal{V}_h.$$

The following estimates hold.

LEMMA 3. *Assume* (A1)-(A12), (A14), (A17). *Then for almost every* $t \in J$,

$$(4.2)\qquad \|u-\tilde{u}\|_0 + \|u-\tilde{u}\|_1 h \le Q\|u-u_{\mathrm{D}}\|_r h^r,$$

$$(4.3)\qquad \|\partial_t(u-\tilde{u})\|_0 \le Q\{\|u-u_{\mathrm{D}}\|_r + \|\partial_t(u-u_{\mathrm{D}})\|_r\}h^{r+\gamma-1}.$$

PROOF: The elliptic equation (4.1) is linear for the given u, so it is well known how to obtain the first estimate [4]. Furthermore, we can obtain an estimate of the time derivative of $u-\tilde{u}$ by considering the differentiated equation

$$(4.4)\qquad \big(a(\theta(u))\mathbf{k}\nabla\partial_t(u-\tilde{u}), \nabla v\big) + \langle \lambda\partial_t(u-\tilde{u}), v\rangle \\ + \big(c(\theta(u))\partial_t(u-\tilde{u}), v\big) + \big(\partial_t[a(\theta(u))]\mathbf{k}\nabla(u-\tilde{u}), \nabla v\big) \\ + \langle \partial_t\lambda\,(u-\tilde{u}), v\rangle + \big(\partial_t[c(\theta(u))](u-\tilde{u}), v\big) = 0 \\ \text{for all } v \in \mathcal{V}_h.$$

Easily with (4.2) we have that

$$(4.5)\qquad \|\partial_t(u-\tilde{u})\|_1 \le Q\{\|u-u_{\mathrm{D}}\|_r + \|\partial_t(u-u_{\mathrm{D}})\|_r\}h^{r-1},$$

since $\big|\partial_t[a(\theta(u))]\big|$ and $\big|\partial_t[c(\theta(u))]\big|$ are uniformly bounded.

For the L^2 estimate of $\partial_t(u-\tilde{u})$, note that if $\lambda \equiv 0$ and $\gamma = 1$, this estimate appears in [5, Section 4]. The proof given there needs some modification in our situation. Solve the following dual problem for $\psi \in \mathcal{V}$:

$$(4.6\text{a})\qquad -\nabla\cdot\big[a(\theta(u))\mathbf{k}\nabla\psi\big] + c(\theta(u))\psi = \partial_t(u-\tilde{u}) \qquad \text{in } \Omega,$$

$$(4.6\text{b})\qquad -\big[a(\theta(u))\mathbf{k}\nabla\psi\big]\cdot\nu - \lambda\psi = 0 \qquad \text{on } \Gamma_{\mathrm{N}}.$$

As is well known, $\|\psi\|_2 \le Q\|\partial_t(u-\tilde{u})\|_0$. Now integration by parts for the first equality and (4.4) for the second shows that

$$
\begin{aligned}
(4.7)\quad & \big(\partial_t(u-\tilde{u}), \partial_t(u-\tilde{u})\big) \\
&= \big(a(\theta(u))\mathbf{k}\nabla\partial_t(u-\tilde{u}), \nabla\psi\big) \\
&\quad + \langle\lambda\partial_t(u-\tilde{u}), \psi\rangle + \big(c(\theta(u))\partial_t(u-\tilde{u}), \psi\big) \\
&= \big(a(\theta(u))\mathbf{k}\nabla\partial_t(u-\tilde{u}), \nabla(\psi - v)\big) \\
&\quad + \langle\lambda\partial_t(u-\tilde{u}), \psi - v\rangle + \big(c(\theta(u))\partial_t(u-\tilde{u}), \psi - v\big) \\
&\quad + \big(\partial_t[a(\theta(u))]\mathbf{k}\nabla(u-\tilde{u}), \nabla(\psi - v)\big) \\
&\quad + \langle\partial_t\lambda\,(u-\tilde{u}), \psi - v\rangle + \big(\partial_t[c(\theta(u))](u-\tilde{u}), \psi - v\big) \\
&\quad - \big(\partial_t[a(\theta(u))]\mathbf{k}\nabla(u-\tilde{u}), \nabla\psi\big) \\
&\quad - \langle\partial_t\lambda\,(u-\tilde{u}), \psi\rangle - \big(\partial_t[c(\theta(u))](u-\tilde{u}), \psi\big) \\
&\le Q\big\{\big[\|\partial_t(u-\tilde{u})\|_1 + \|u-\tilde{u}\|_1\big]\|\psi - v\|_1 + \|u-\tilde{u}\|_0\|\psi\|_0\big\} \\
&\quad - \big(\partial_t[a(\theta(u))]\mathbf{k}\nabla(u-\tilde{u}), \nabla\psi\big) - \langle\partial_t\lambda\,(u-\tilde{u}), \psi\rangle \\
&\qquad\qquad\qquad\qquad \text{for all } v \in \mathcal{V}_h.
\end{aligned}
$$

For the next to last term on the far right side of (4.7), using negative norms (which we take to be those of the dual spaces of the corresponding positively indexed spaces),

$$
\begin{aligned}
(4.8)\qquad & \big|\big(\partial_t[a(\theta(u))]\mathbf{k}\nabla(u-\tilde{u}), \nabla\psi\big)\big| \\
&\quad \le \|\nabla(u-\tilde{u})\|_{-\gamma}\|\partial_t[a(\theta(u))]\mathbf{k}\nabla\psi\|_\gamma \\
&\quad \le \|\nabla(u-\tilde{u})\|_{-\gamma}\|\nabla\psi\|_\gamma,
\end{aligned}
$$

since $\partial_t[a(\theta(u))] = \partial_\theta a(\theta(u))\,\partial_u\theta(u)\,\partial_t u \in C^{0,\gamma}(\overline{\Omega})$. Also, for $\delta \ge 0$, there is a continuous linear extension operator $E : H^\delta(\Omega) \to H^\delta(\mathbb{R}^d)$, so for any j,

$$
\begin{aligned}
(4.9)\qquad & \|\partial_{x_j}(u-\tilde{u})\|_{-\gamma} \le \|\partial_{x_j}E(u-\tilde{u})\|_{H^{-\gamma}(\mathbb{R}^d)} \\
&\quad \le Q\|E(u-\tilde{u})\|_{H^{1-\gamma}(\mathbb{R}^d)} \le Q\|u-\tilde{u}\|_{1-\gamma}.
\end{aligned}
$$

By interpolation on (4.2), we obtain that

$$
(4.10)\qquad \|u-\tilde{u}\|_{1-\gamma} \le Q\|u - u_{\mathrm{D}}\|_r h^{r+\gamma-1}.
$$

(A nice summary of the Sobolev space theory used above can be found in [9]).

For the last term on the far right side of (4.7), we note[1] that a duality argument implies

$$\|u-\tilde{u}\|_{H^{-1/2}(\Gamma_{\mathrm{N}})} \le Q\|u-u_{\mathrm{D}}\|_r h^r; \tag{4.11}$$

hence, we make the estimate

$$\begin{aligned} \left|\langle \partial_t \lambda(u-\tilde{u}), \psi\rangle\right| &\le Q\|u-\tilde{u}\|_{H^{-1/2}(\Gamma_{\mathrm{N}})}\|\partial_t\lambda\,\psi\|_1 \\ &\le Q\|u-\tilde{u}\|_{H^{-1/2}(\Gamma_{\mathrm{N}})}\|\psi\|_1. \end{aligned} \tag{4.12}$$

Since $\inf_{v\in\mathcal{V}_h}\|\psi-v\|_1 \le Q\|\psi\|_2 h$, combining (4.2), (4.5), (4.7)–(4.12), we obtain that

$$\begin{aligned} &\|\partial_t(u-\tilde{u})\|_0^2 \\ &\quad\le Q\big\{\big[\|\partial_t(u-\tilde{u})\|_1 + \|u-\tilde{u}\|_1\big]h \\ &\qquad + \|u-\tilde{u}\|_{1-\gamma} + \|u-\tilde{u}\|_{H^{-1/2}(\Gamma_{\mathrm{N}})}\big\}\|\psi\|_2 \\ &\quad\le Q\big\{\|u-u_{\mathrm{D}}\|_r + \|\partial_t(u-u_{\mathrm{D}})\|_r\big\}h^{r+\gamma-1}\|\partial_t(u-\tilde{u})\|_0, \end{aligned} \tag{4.13}$$

and the last estimate of the lemma follows. ▮

We need to make the following assumption, which holds for many $\mathcal{V}_h$:

(A19) $\|\tilde{u}\|_{L^\infty(J;W^{1,\infty}(\Omega))} \le Q$.

(See, e.g., [16], [14], and [15] and some of the references cited therein for this result on certain standard finite element spaces defined over quasi-uniform meshes.)

It remains only to estimate the deviation of U from $\tilde{u}$.

5. An analysis of the continuous time procedure. In this section, we present an analysis of the error $u-U$ when U is defined by (2.2). Combine (4.1) with (2.1) and (2.2) to obtain an equation for the error:

$$\begin{aligned} &\big(\partial_t[\theta(u)-\theta(U)], v\big) + \big(a(\theta(u))\mathbf{k}\nabla\tilde{u} - a(\theta(U))\mathbf{k}\nabla U, \nabla v\big) \\ &\qquad + \big(\mathbf{b}(\theta(u)) - \mathbf{b}(\theta(U)), \nabla v\big) + \langle\lambda(\tilde{u}-U), v\rangle \\ &\qquad + \big(c(\theta(u))\tilde{u} - c(\theta(U))U, v\big) \\ &\quad = \big(f(\theta(u)) - f(\theta(U)), v\big) \qquad \text{for all } v\in\mathcal{V}_h. \end{aligned} \tag{5.1}$$

[1] The author is indebted to Professor Mary F. Wheeler for this observation.

For Q_1 fixed below, our first choice of a test function in (5.1) will be

$$v(x,t,\bar{t}) = \int_t^{\bar{t}} (\tilde{u}(x,\tau) - U(x,\tau))\, e^{-Q_1\tau}\, d\tau \in \mathcal{V}_h. \tag{5.2}$$

We include the exponential expression for use in a Gronwall argument. It is more customary to make a direct estimate and then apply Gronwall's inequality, but this simple approach seems to fail.

By the inverse inequality (A13),

$$\begin{aligned} &\big(a(\theta(u))\mathbf{k}\nabla\tilde{u} - a(\theta(U))\mathbf{k}\nabla U, \nabla v\big) \\ &\quad = \big(a(\theta(u))\mathbf{k}\nabla(\tilde{u} - U), \nabla v\big) \\ &\qquad + \big([a(\theta(u)) - a(\theta(U))]\mathbf{k}\nabla\tilde{u}, \nabla v\big) \\ &\qquad - \big([a(\theta(u)) - a(\theta(U))]\mathbf{k}\nabla(\tilde{u} - U), \nabla v\big) \\ &\quad \geq \big(a(\theta(u))\mathbf{k}\nabla(\tilde{u} - U), \nabla v\big) \\ &\qquad - Q\|\theta(u) - \theta(U)\|_0 \|\nabla v\|_0 \\ &\qquad\qquad \times \big[\|\nabla\tilde{u}\|_{L^\infty(\Omega)} + \|\nabla(\tilde{u} - U)\|_0 h^{-d/2}\big]. \end{aligned} \tag{5.3}$$

Further, we have that

$$\begin{aligned} &a(\theta(u))\mathbf{k}\nabla(\tilde{u} - U)\cdot\nabla\int_t^{\bar{t}} (\tilde{u} - U)\, e^{-Q_1\tau}\, d\tau \\ &\quad = -\frac{1}{2}\partial_t\Big[a(\theta(u))\Big(\int_t^{\bar{t}} \mathbf{k}^{1/2}\nabla(\tilde{u} - U)\, e^{-Q_1\tau}\, d\tau\Big)^2 e^{Q_1 t}\Big] \\ &\qquad + \frac{1}{2}\big[Q_1 a(\theta(u)) + \partial_\theta a(\theta(u))\partial_t\theta(u)\big] \\ &\qquad\qquad \times \Big(\int_t^{\bar{t}} \mathbf{k}^{1/2}\nabla(\tilde{u} - U)\, e^{-Q_1\tau}\, d\tau\Big)^2 e^{Q_1 t}. \end{aligned} \tag{5.4}$$

Similar expressions hold for

$$\big(c(\theta(u))\tilde{u} - c(\theta(U))U, v\big), \quad c(\theta(u))\,(\tilde{u} - U)\int_t^{\bar{t}} (\tilde{u} - U)e^{-Q_1\tau}\, d\tau,$$

$$\text{and} \quad \lambda\,(\tilde{u} - U)\int_t^{\bar{t}} (\tilde{u} - U)e^{-Q_1\tau}\, d\tau.$$

We now easily obtain from (5.1) with Poincaré's inequality that

$$
\begin{aligned}
(5.5)\quad &\Big(\partial_t[\theta(u)-\theta(U)], \int_t^{\bar t} (\tilde u - U)\, e^{-Q_1 \tau}\, d\tau\Big) \\
&\quad - \frac{1}{2}\partial_t \int_\Omega a(\theta(u))\Big(\int_t^{\bar t} \mathbf{k}^{1/2}\nabla(\tilde u - U)\, e^{-Q_1\tau}\, d\tau\Big)^2 e^{Q_1 t}\, dx \\
&\quad + \frac{1}{2}Q_1 \int_\Omega a(\theta(u))\Big(\int_t^{\bar t} \mathbf{k}^{1/2}\nabla(\tilde u - U)\, e^{-Q_1\tau}\, d\tau\Big)^2 e^{Q_1 t}\, dx \\
&\quad - \frac{1}{2}\partial_t \int_{\Gamma_{\mathrm N}} \lambda\Big(\int_t^{\bar t} (\tilde u - U)\, e^{-Q_1\tau}\, d\tau\Big)^2 e^{Q_1 t}\, ds(x) \\
&\quad - \frac{1}{2}\partial_t \int_\Omega c(\theta(u))\Big(\int_t^{\bar t} (\tilde u - U)\, e^{-Q_1\tau}\, d\tau\Big)^2 e^{Q_1 t}\, dx \\
&\le Q\Big\| \int_t^{\bar t} \nabla(\tilde u - U)\, e^{-Q_1\tau}\, d\tau\Big\|_0^2 e^{Q_1 t} \\
&\quad + \epsilon\big[1 + \|\nabla(\tilde u - U)\|_0^2 h^{-d}\big]\, \|\theta(u)-\theta(U)\|_0^2 e^{-Q_1 t}.
\end{aligned}
$$

Integration by parts shows

$$
\begin{aligned}
(5.6)\quad &\int_0^{\bar t} \Big(\partial_t[\theta(u)-\theta(U)], \int_t^{\bar t} (\tilde u - U)\, e^{-Q_1\tau}\, d\tau\Big) dt \\
&= -\Big(\theta(u^0)-\theta(U^0), \int_0^{\bar t} (\tilde u - U)\, e^{-Q_1 t}\, dt\Big) \\
&\quad + \int_0^{\bar t} \big(\theta(u)-\theta(U), u - U\big)\, e^{-Q_1 t}\, dt \\
&\quad - \int_0^{\bar t} \big(\theta(u)-\theta(U), u - \tilde u\big)\, e^{-Q_1 t}\, dt \\
&\ge \int_0^{\bar t} \big(\theta(u)-\theta(U), u - U\big)\, e^{-Q_1 t}\, dt \\
&\quad - Q\Big\{\int_0^{\bar t} \|u-\tilde u\|_0^2 e^{-Q_1 t}\, dt + \|\theta(u^0)-\theta(U^0)\|_0^2\Big\} \\
&\quad - \epsilon\Big\{\int_0^{\bar t} \|\theta(u)-\theta(U)\|_0^2 e^{-Q_1 t}\, dt + \Big\|\int_0^{\bar t} (\tilde u - U) e^{-Q_1 t}\, dt\Big\|_0^2\Big\},
\end{aligned}
$$

so with Poincaré's inequality and a sufficiently large Q_1, (5.5) integrated

over $[0,\bar{t}]$ yields

$$(5.7)\qquad \begin{aligned} &\frac{1}{2}\int_0^{\bar{t}} \big(\theta(u)-\theta(U), u-U\big)\, e^{-Q_1 t}\, dt \\ &\quad + \frac{1}{4}\int_\Omega a(\theta(u^0))\Big(\int_0^{\bar{t}} \mathbf{k}^{1/2}\nabla(\tilde{u}-U)\, e^{-Q_1 t}\, dt\Big)^2 dx \\ &\quad + \frac{1}{2}\int_{\Gamma_N} \lambda^0\Big(\int_0^{\bar{t}} (\tilde{u}-U)\, e^{-Q_1 t}\, dt\Big)^2 ds(x) \\ &\quad + \frac{1}{2}\int_\Omega c(\theta(u^0))\Big(\int_0^{\bar{t}} (\tilde{u}-U)\, e^{-Q_1 t}\, dt\Big)^2 dx \\ &\le Q\Big\{\int_0^{\bar{t}} \|u-\tilde{u}\|_0^2 e^{-Q_1 t}\, dt + \|\theta(u^0)-\theta(U^0)\|_0^2\Big\} \\ &\quad + \epsilon h^{-d}\int_0^{\bar{t}} \|\nabla(\tilde{u}-U)\|_0^2 \|\theta(u)-\theta(U)\|_0^2 e^{-Q_1 t}\, dt \\ &\le Qh^{2r} + \epsilon h^{-d}\int_0^{\bar{t}} \|\nabla(\tilde{u}-U)\|_0^2 \|\theta(u)-\theta(U)\|_0^2 e^{-Q_1 t}\, dt, \end{aligned}$$

using (4.2) and (A15). We leave this partial estimate for the time being.

Now take in (5.1) as a second test function

$$(5.8)\qquad v = \tilde{u} - U \in \mathcal{V}_h.$$

Since

$$(5.9)\qquad \begin{aligned} &a(\theta(u))\nabla\tilde{u} - a(\theta(U))\nabla U \\ &\qquad = a(\theta(U))\nabla(\tilde{u}-U) + [a(\theta(u)) - a(\theta(U))]\nabla\tilde{u} \end{aligned}$$

and a similar expression holds for $c(\theta(u))\tilde{u} - c(\theta(U))U$, we obtain immediately from (5.1) with Poincaré's inequality that

$$(5.10)\qquad \begin{aligned} &\big(\partial_t[\theta(u)-\theta(U)], u-U\big) + \tfrac{1}{2}\big(a(\theta(U))\mathbf{k}\nabla(\tilde{u}-U), \nabla(\tilde{u}-U)\big) \\ &\qquad + \langle\lambda(\tilde{u}-U), \tilde{u}-U\rangle + \big(c(\theta(U))(\tilde{u}-U), \tilde{u}-U\big) \\ &\qquad \le Q\|\theta(u)-\theta(U)\|_0^2 + \big(\partial_t[\theta(u)-\theta(U)], u-\tilde{u}\big). \end{aligned}$$

Note that for any Q_2,

$$(5.11)\qquad \begin{aligned} &\partial_t[\theta(u)-\theta(U)](u-U)e^{-Q_2 t} \\ &\quad = \partial_t\Big[\int_U^u \big(\theta(\mu)-\theta(U)\big)\, d\mu\, e^{-Q_2 t}\Big] \\ &\quad + Q_2\int_U^u \big(\theta(\mu)-\theta(U)\big)\, d\mu\, e^{-Q_2 t} \\ &\quad + \big\{\partial_t\theta(u)(u-U) - (\theta(u)-\theta(U))\partial_t u\big\}e^{-Q_2 t}. \end{aligned}$$

(We remark that Wheeler and Dupont use a version of this type of identity; see [17, Section 3.4].) The first two terms on the right side are well behaved by Proposition 1; the last term is poorly behaved. By (A16), it can be bounded as follows. For some w between u and U,

(5.12)
$$\begin{aligned}&\big|\partial_t\theta(u)(u-U)-(\theta(u)-\theta(U))\partial_t u\big| = \big|(\partial_u\theta(u)-\partial_u\theta(w))(u-U)\partial_t u\big|\\ &\qquad\le Q|\theta(u)-\theta(w)|^{\beta}\,|u-U| \le Q|\theta(u)-\theta(U)|^{\beta}\,|u-U|\\ &\qquad\le Q\big[(\theta(u)-\theta(U))(u-U)\big]^{2\beta/(1+\beta)}+\epsilon\{|\tilde{u}-U|^2+|u-\tilde{u}|^2\};\end{aligned}$$

the last step is by the well known inequality

$$|ab| \le \frac{1}{\epsilon^{p/p'}p}|a|^p + \frac{\epsilon}{p'}|b|^{p'} \quad \text{for any } 1<p<\infty,\ \frac{1}{p}+\frac{1}{p'}=1,$$

which implies that

$$|a|^{\beta}|b| = |ab|^{\beta}|b|^{1-\beta} \le Q|ab|^{2\beta/(1+\beta)} + \epsilon b^2.$$

We now multiply (5.10) through by $e^{-Q_2 t}$, combine the result with (5.11)-(5.12), and integrate in time from 0 to t. The last term can be integrated by parts. Then Proposition 1, Poincaré's inequality, and a large enough Q_2 yield the estimate

(5.13)
$$\begin{aligned}&q\|\theta(u)-\theta(U)\|_0^2 e^{-Q_2 t}\\ &\quad+\frac{1}{4}\int_0^t (a(\theta(U))\mathbf{k}\nabla(\tilde{u}-U),\nabla(\tilde{u}-U))e^{-Q_2\tau}\,d\tau\\ &\quad+\int_0^t \langle\lambda(\tilde{u}-U),\tilde{u}-U\rangle e^{-Q_2\tau}\,d\tau\\ &\quad+\int_0^t (c(\theta(U))(\tilde{u}-U),\tilde{u}-U)e^{-Q_2\tau}\,d\tau\\ &\le Q\bigg\{\int_0^t \|\partial_t(u-\tilde{u})\|_0^2\, e^{-Q_2\tau}\,d\tau + \|u-\tilde{u}\|^2_{L^\infty(J;L^2(\Omega))}\\ &\quad+(\theta(u^0)-\theta(U^0),u^0-U^0)\\ &\quad+\int_0^t (\theta(u)-\theta(U),u-U)^{2\beta/(1+\beta)}e^{-Q_2\tau}\,d\tau\bigg\}\\ &\le Q\bigg\{h^{2(r+\gamma-1)}+\bigg(\int_0^t(\theta(u)-\theta(U),u-U)\\ &\qquad\qquad\times e^{-Q_2(1+\beta)\tau/2\beta}\,d\tau\bigg)^{2\beta/(1+\beta)}\bigg\},\end{aligned}$$

using Lemma 3 and (A15).

We can now combine (5.7) and (5.13) to obtain an estimate. A continuation argument is required, so for some fixed Q_0 independent of h, let $T' = T'_h \leq T$ be the largest value of time for which (where $J' \equiv [0, T']$)

$$\|\nabla(\tilde{u} - U)\|_{L^2(J';L^2(\Omega))} \leq Q_0 h^{d\beta/(3\beta-1)}. \tag{5.14}$$

Because of Lemma 1, $T' > 0$ (but perhaps not uniformly so in h). Fix $t = \bar{t}$ where $\|\theta(u) - \theta(U)\|_0$ attains its essential maximum on J'. Note that in general for $1/2 < \delta \leq 1$,

$$|abc|^\delta = |b|^{1-\delta}\big(|b|^{2\delta-1}|ac|^\delta\big) \leq |b| + \big(|a|^{\delta/(2\delta-1)}|b|\big)^{(2\delta-1)/\delta}|c|,$$

and that $r + \gamma - 1 \geq r + \beta - 1 \geq 2r\beta/(1+\beta)$. Then from (5.13) and (5.7), with $\bar{J} \equiv [0, \bar{t}]$,

$$\begin{aligned}
&\|\theta(u) - \theta(U)\|^2_{L^\infty(J';L^2(\Omega))} + \|\nabla(\tilde{u} - U)\|^2_{L^2(\bar{J};L^2(\Omega))} \\
&\quad \leq Q\bigg\{h^{4r\beta/(1+\beta)} + \Big(\epsilon h^{-d}\|\nabla(\tilde{u} - U)\|^2_{L^2(\bar{J};L^2(\Omega))} \\
&\qquad\qquad \times \|\theta(u) - \theta(U)\|^2_{L^\infty(J';L^2(\Omega))}\Big)^{2\beta/(1+\beta)}\bigg\} \\
&\quad \leq Q\bigg\{h^{4r\beta/(1+\beta)} + \epsilon^{2\beta/(1+\beta)}\Big[\|\nabla(\tilde{u} - U)\|^2_{L^2(\bar{J};L^2(\Omega))} \\
&\qquad + \Big(h^{-2d\beta/(3\beta-1)}\|\nabla(\tilde{u} - U)\|^2_{L^2(\bar{J};L^2(\Omega))}\Big)^{(3\beta-1)/2\beta} \\
&\qquad\qquad \times \|\theta(u) - \theta(U)\|^2_{L^\infty(J';L^2(\Omega))}\Big]\bigg\},
\end{aligned} \tag{5.15}$$

provided that $3\beta > 1$. We hide two terms for ϵ small enough (with (5.14)). Repeating the above for $t = T'$ yields that

$$\begin{aligned}
\|\theta(u) - \theta(U)\|_{L^\infty(J';L^2(\Omega))} + \|\nabla(\tilde{u} - U)\|_{L^2(J';L^2(\Omega))}& \\
\leq Q h^{2r\beta/(1+\beta)}&.
\end{aligned} \tag{5.16}$$

To complete the continuation argument, suppose that $T' < T$ and assume that $r > d(1+\beta)/2(3\beta-1) > 0$. Then we have shown that

$$\|\nabla(\tilde{u} - U)\|_{L^2(J';L^2(\Omega))} \leq Q h^{2r\beta/(1+\beta)} \leq \tfrac{1}{2} Q_0 h^{d\beta/(3\beta-1)} \tag{5.17}$$

for all $h < h_0$ for some $h_0 > 0$, since Q is independent of h and T'. Hence the maximality of T' requires that $T' = T$ and the argument is completed.

By our analysis (see especially (4.2)–(4.3), (5.7), (5.13), and (5.16)), with a few modifications, we have the following results.

THEOREM 1. *Assume* (A1)–(A12), (A14)–(A17), (A19). *Let U be defined by* (2.2). *Define the two cases:*

(1) (A13) *holds and* $r > d(1+\beta)/2(3\beta - 1) > 0$;

(2) *a and c are independent of θ.*

Then in either case, with $\delta = 2r\beta/(1+\beta)$, $\bar{\delta} = r$,

$$\|\theta(u) - \theta(U)\|_{L^\infty(J;L^2(\Omega))} + \|\tilde{u} - U\|_{L^2(J;H^1(\Omega))} \le Qh^\delta, \tag{5.18}$$

$$\|u - U\|_{L^2(J;H^1(\Omega))} \le Q\{h^\delta + h^{r-1}\}, \tag{5.19}$$

$$\left\{ \int_0^T (\theta(u) - \theta(U), u - U)\, dt \right\}^{1/2} \le Qh^{\bar{\delta}}. \tag{5.20}$$

PROOF: Case (1) is exactly the full argument given above. Note that (5.20) is of optimal order since $4r\beta/(1+\beta) - d/2 \ge r$.

Case (2). In the case that a and c are independent of θ, our argument simplifies substantially because the continuation and inverse inequality arguments are not needed. To see this, note that only the first term on the far right side of (5.3) remains; consequently, the same is true of (5.7), which is now a complete error estimate. Finally, the estimate (5.13) is completed directly with (5.7). ▮

REMARK 4: In Case (1), if $\theta(u) \propto 1 - |\min\{0, u\}|^\alpha$ (recall that then $\beta = (\alpha - 1)/\alpha$), then $3\beta > 1$ requires that $\alpha > 3/2$; furthermore, $r/d > (2\alpha - 1)/(4\alpha - 6)$.

REMARK 5: Case (2) (a and c independent of θ) arises, for example, in Richard's equation ($c \equiv 0$) after the Kirchoff transformation $u \mapsto \bar{u} = \int_0^u a(\theta(\xi))\, d\xi$ and $\theta(u) \mapsto \bar{\theta}(\bar{u})$ [7].

Our results can be restricted to the strictly parabolic case.

THEOREM 2. *Assume* (A1)–(A12), (A14)–(A15), (A17), (A19), *with $\partial_u \theta(u)$ uniformly positively bounded from below. Let U be defined by* (2.2). *Define the two cases:*

(1) (A13) *holds and* $r > d/2\gamma$;

(2) *a and c are independent of θ.*

Then in either case, (5.18)–(5.20) *hold with* $\delta = r(1+\gamma)/2$, $\bar{\delta} = r$.

PROOF: If the problem is strictly parabolic, then

$$q|\theta(u) - \theta(U)| \le |u - U| \le Q|\theta(u) - \theta(U)| \tag{5.21}$$

and $\beta = \gamma$.

Case (1). Because of (5.21), (5.12) can be stopped at the second inequality with β replaced by γ. Replace $2\beta/(1+\beta)$ by $(1+\gamma)/2$ in

(5.13) and (5.15)–(5.17), and replace $\beta/(3\beta-1)$ by $(1+\gamma)/4\gamma$ in (5.14)–(5.15) and (5.17). Finally, $r+\gamma-1 \geq r(1+\gamma)/2$ and $r(1+\gamma)-d/2 \geq r$.

Case (2). This is a combination of Case (1) and Case (2) of Theorem 1. ∎

For monotone flows we have the following specialized result.

THEOREM 3. *Assume* (A1)–(A12), (A14)–(A15), (A17), (A19), *with either* $\partial_t u \leq 0$ *and* $\partial_u \theta$ *monotone nonincreasing in* u, *or* $\partial_t u \geq 0$ *and* $\partial_u \theta$ *monotone nondecreasing in* u. *Let* U *be defined by* (2.2). *Then* (5.18)–(5.20) *hold with* $\delta = r+\gamma-1$, $\bar{\delta} = \delta$. *Moreover, if* (A13) *holds,* (5.20) *holds with* $\bar{\delta} = \min\{r, 2\delta - d/2\}$.

PROOF: If $\partial_t u \leq 0$ and $\partial_u \theta$ decreases, or if $\partial_t u \geq 0$ and $\partial_u \theta$ increases, then (5.12) can be replaced simply by

$$\begin{aligned} (5.12') \qquad & \partial_t \theta(u)(u-U) - (\theta(u)-\theta(U))\partial_t u \\ & \quad = \big(\partial_u \theta(u) - \partial_u \theta(w)\big)(u-U)\partial_t u \geq 0 \end{aligned}$$

(for some w between u and U). As a consequence, the last term on the far right side of (5.13) is absent, so (5.13) is a complete error estimate and no continuation argument is needed.

We have (5.20) because either we can avoid the inverse inequality argument of (5.3) by noting that

$$(5.22) \quad \big|([a(\theta(u)) - a(\theta(U))]\mathbf{k}\nabla(\tilde{u}-U), \nabla v)\big| \leq Q\|\nabla(\tilde{u}-U)\|_0 \|\nabla v\|_0,$$

or with (A13) we can follow the argument as it is given. ∎

REMARK 6: In each of the preceding results, the norms of the error are bounded by h to the optimal power r or $r-1$ in the case where uniformly in x, $\partial_u \theta(u)$ is positively bounded from below and uniformly Lipschitz as a function of u (then $\beta = \gamma = 1$). Of course in this case, the argument can be further simplified; in fact, only the second test function is needed since the expression after the second inequality in (5.12) is controlled by a large enough Q_2 (see [17, Section 3.4]).

6. An analysis of the discrete time procedure. In this section, we present an analysis of the error $u^n - U^n$ when U^n is defined by (2.3). Our proof follows closely that given in the last section.

Combine (4.1) with (2.1) and (2.3) to obtain an equation for the error:

$$\begin{aligned}
(6.1)\qquad &\big(\partial[\theta(u)-\theta(U)]^n, v\big) + \big(a(\theta(u^n))\mathbf{k}\nabla\tilde{u}^n - a(\theta(U^n))\mathbf{k}\nabla U^n, \nabla v\big)\\
&\quad + \big(\mathbf{b}^n(\theta(u^n)) - \mathbf{b}^n(\theta(U^n)), \nabla v\big) + \langle \lambda^n(\tilde{u}^n - U^n), v\rangle\\
&\quad + \big(c(\theta(u^n))\tilde{u}^n - c(\theta(U^n))U^n, v\big)\\
&= \big(f^n(\theta(u^n)) - f^n(\theta(U^n)), v\big)\\
&\quad + \big(\partial\theta(u)^n - \partial_t\theta(u^n), v\big) \qquad \text{for all } v \in \mathcal{V}_h.
\end{aligned}$$

With the discrete exponential

$$\eta_1^{\pm n} = \Big(\prod_{k=1}^{n}(1 + Q_1\Delta t^k)\Big)^{\pm 1} \qquad \text{for } n \geq 0,$$

set in (6.1)

$$(6.2)\qquad v(x,n,\bar{n}) = \sum_{k=n}^{\bar{n}}(\tilde{u}^k - U^k)\eta_1^{-k}\,\Delta t^k \in \mathcal{V}_h.$$

Again (5.3) holds (at time t_n), as well as

(6.3)

$$\begin{aligned}
&a(\theta(u^n))\mathbf{k}\nabla(\tilde{u}^n - U^n)\cdot\nabla\sum_{k=n}^{\bar{n}}(\tilde{u}^k - U^k)\eta_1^{-k}\,\Delta t^k\\
&= -\frac{1}{2\Delta t^n}\Bigg[a(\theta(u^{n+1}))\Big(\sum_{k=n+1}^{\bar{n}}\mathbf{k}^{1/2}\nabla(\tilde{u}^k - U^k)\eta_1^{-k}\,\Delta t^k\Big)^2\eta_1^n\\
&\qquad - a(\theta(u^n))\Big(\sum_{k=n}^{\bar{n}}\mathbf{k}^{1/2}\nabla(\tilde{u}^k - U^k)\eta_1^{-k}\,\Delta t^k\Big)^2\eta_1^{n-1}\Bigg]\\
&\quad + \frac{1}{2}Q_1 a(\theta(u^n))\Big(\sum_{k=n}^{\bar{n}}\mathbf{k}^{1/2}\nabla(\tilde{u}^k - U^k)\eta_1^{-k}\,\Delta t^k\Big)^2\eta_1^{n-1}\\
&\quad + \frac{a(\theta(u^{n+1})) - a(\theta(u^n))}{2\Delta t^n}\Big(\sum_{k=n+1}^{\bar{n}}\mathbf{k}^{1/2}\nabla(\tilde{u}^k - U^k)\eta_1^{-k}\,\Delta t^k\Big)^2\eta_1^n\\
&\quad + \frac{1}{2}a(\theta(u^n))|\mathbf{k}^{1/2}\nabla(\tilde{u}^n - U^n)|^2\eta_1^{-n}\,\Delta t^n
\end{aligned}$$

and similar expressions for terms with c and λ. Now from (6.1) we can

obtain with Poincaré's inequality that

(6.4)

$$
\begin{aligned}
&\Big(\partial[\theta(u)-\theta(U)]^n, \sum_{k=n}^{\bar n}(\tilde u^k - U^k)\eta_1^{-k}\,\Delta t^k\Big) \\
&\quad - \frac{1}{2}\partial\Big[\int_\Omega a(\theta(u^{(\cdot)+1}))\Big(\sum_{k=(\cdot)+1}^{\bar n} \mathbf{k}^{1/2}\nabla(\tilde u^k - U^k)\eta_1^{-k}\,\Delta t^k\Big)^2 \eta_1^{(\cdot)}\,dx\Big]^n \\
&\quad + \frac{1}{2}Q_1\int_\Omega a(\theta(u^n))\Big(\sum_{k=n}^{\bar n}\mathbf{k}^{1/2}\nabla(\tilde u^k - U^k)\eta_1^{-k}\,\Delta t^k\Big)^2\eta_1^{n-1}\,dx \\
&\quad - \frac{1}{2}\partial\Big[\int_{\Gamma_{\mathrm N}} \lambda^{(\cdot)+1}\Big(\sum_{k=(\cdot)+1}^{\bar n}(\tilde u^k - U^k)\,\eta_1^{-k}\,\Delta t^k\Big)^2\eta_1^{(\cdot)}\,ds(x)\Big]^n \\
&\quad - \frac{1}{2}\partial\Big[\int_\Omega c(\theta(u^{(\cdot)+1}))\Big(\sum_{k=(\cdot)+1}^{\bar n}(\tilde u^k - U^k)\,\eta_1^{-k}\,\Delta t^k\Big)^2\eta_1^{(\cdot)}\,dx\Big]^n \\
&\le Q\Big\{\sum_{j=n}^{n+1}\Big\|\sum_{k=j}^{\bar n}\mathbf{k}^{1/2}\nabla(\tilde u^k - U^k)\,\eta_1^{-k}\,\Delta t^k\Big\|_0^2\eta_1^{j-1} \\
&\qquad + \|\partial_t^2\theta(u)\|^2_{L^2([t_{n-1},t_n];L^2(\Omega))}\Delta t^n\eta_1^{-n+1}\Big\} \\
&\quad + \epsilon\big[1 + \|\nabla(\tilde u^n - U^n)\|_0^2 h^{-d}\big]\,\|\theta(u^n) - \theta(U^n)\|_0^2\eta_1^{-n+1},
\end{aligned}
$$

wherein we use that

$$
\|\partial\theta(u)^n - \partial_t\theta(u^n)\|_0 \le Q\|\partial_t^2\theta(u)\|_{L^2([t_{n-1},t_n];L^2(\Omega))}(\Delta t^n)^{1/2}. \tag{6.5}
$$

Now note that summation by parts yields

$$
\begin{aligned}
\sum_{n=1}^{\bar n}&\Big(\partial[\theta(u)-\theta(U)]^n, \sum_{k=n}^{\bar n}(\tilde u^k - U^k)\,\eta_1^{-k}\,\Delta t^k\Big)\Delta t^n \\
&= -\Big(\theta(u^0) - \theta(U^0), \sum_{k=1}^{\bar n}(\tilde u^k - U^k)\,\eta_1^{-k}\,\Delta t^k\Big) \\
&\quad + \sum_{n=1}^{\bar n}\big(\theta(u^n) - \theta(U^n), u^n - U^n\big)\,\eta_1^{-n}\,\Delta t^n \\
&\quad - \sum_{n=1}^{\bar n}\big(\theta(u^n) - \theta(U^n), u^n - \tilde u^n\big)\,\eta_1^{-n}\,\Delta t^n;
\end{aligned} \tag{6.6}
$$

hence, with Poincaré's inequality, a sufficiently large Q_1, a sufficiently small $\Delta t \equiv \max_k \Delta t^k$, and (A18), (6.4) multiplied by Δt^n and summed on n from 1 to $\bar{n}$ gives

$$
\begin{aligned}
(6.7)\quad & \frac{1}{2}\sum_{n=1}^{\bar{n}} \left(\theta(u^n) - \theta(U^n), u^n - U^n\right) \eta_1^{-n}\,\Delta t^n \\
& \quad + \frac{1}{4}\int_{\Omega} a(\theta(u^1))\Big(\sum_{k=1}^{\bar{n}} \mathbf{k}^{1/2}\nabla(\tilde{u}^k - U^k)\eta_1^{-k}\,\Delta t^k\Big)^2\,dx \\
& \leq Q\Big\{(\Delta t)^2 + \sum_{n=1}^{\bar{n}} \|u^n - \tilde{u}^n\|_0^2\eta_1^{-n}\,\Delta t^n + \|\theta(u^0) - \theta(U^0)\|_0^2\Big\} \\
& \quad + \epsilon h^{-d}\sum_{n=1}^{\bar{n}} \|\nabla(\tilde{u}^n - U^n)\|_0^2\|\theta(u^n) - \theta(U^n)\|_0^2\eta_1^{-n+1}\,\Delta t^n \\
& \leq Q\{h^{2r} + (\Delta t)^2\} \\
& \quad + \epsilon h^{-d}\sum_{n=1}^{\bar{n}} \|\nabla(\tilde{u}^n - U^n)\|_0^2\|\theta(u^n) - \theta(U^n)\|_0^2\eta_1^{-n+1}\,\Delta t^n.
\end{aligned}
$$

We turn to the second estimate. In (6.1) let

$$
(6.8)\qquad v = \tilde{u}^n - U^n \in \mathcal{V}_h.
$$

Again noting (5.9), we easily obtain with Poincaré's inequality that

$$
\begin{aligned}
&(6.9)\\
&\left(\partial[\theta(u) - \theta(U)]^n, u^n - U^n\right) + \tfrac{1}{2}\left(a(\theta(U^n))\mathbf{k}\nabla(\tilde{u}^n - U^n), \nabla(\tilde{u}^n - U^n)\right) \\
&\qquad + \langle\lambda(\tilde{u}^n - U^n), \tilde{u}^n - U^n\rangle + \left(c(\theta(U^n))(\tilde{u}^n - U^n), \tilde{u}^n - U^n\right) \\
&\qquad \leq Q\{\|\theta(u^n) - \theta(U^n)\|_0^2 + \|\partial_t^2\theta(u)\|^2_{L^2([t_{n-1},t_n];L^2(\Omega))}\Delta t^n\} \\
&\qquad + \left(\partial[\theta(u) - \theta(U)]^n, u^n - \tilde{u}^n\right).
\end{aligned}
$$

Analogous to (5.11), we note that with $\eta_2^{\pm n} = \left(\prod_{k=1}^n(1 + Q_2\Delta t^k)\right)^{\pm 1}$,

$$
\begin{aligned}
(6.10)\qquad & \partial[\theta(u) - \theta(U)]^n(u^n - U^n)\eta_2^{-n+1} \\
& = \partial\Big[\int_U^u (\theta(\mu) - \theta(U))\,d\mu\,\eta_2^{-(\cdot)}\Big]^n \\
& \qquad + Q_2\int_{U^n}^{u^n} (\theta(\mu) - \theta(U^n))\,d\mu\,\eta_2^{-n} - E^n\eta_2^{-n+1},
\end{aligned}
$$

where in the third term, $E^n = E_1^n + E_2^n$ can be written as

$$\text{(6.11a)}\quad E_1^n = -\partial\theta(u)^n(u^n - U^n) + (\theta(u^n) - \theta(U^n))\partial u^n,$$

$$\text{(6.11b)}\quad E_2^n = (\theta(U^n) - \theta(U^{n-1}))\partial u^n - \frac{1}{\Delta t^n}\int_{u^{n-1}}^{u^n} (\theta(u^n) - \theta(\mu))\, d\mu$$
$$- \frac{1}{\Delta t^n}\int_{U^{n-1}}^{U^n} (\theta(\mu) - \theta(U^{n-1}))\, d\mu.$$

Analogous to the previous section, multiply equation (6.9) through by ${\eta_2}^{-n+1}\Delta t^n$, combine the result with (6.10), and sum on n. After summing the last term in (6.9) by parts, Proposition 1, a large enough Q_2, and a small enough Δt yield the estimate

(6.12)
$$\begin{aligned}
&q\|\theta(u^n) - \theta(U^n)\|_0^2{\eta_2}^{-n} \\
&\qquad + \frac{1}{2}\sum_{k=1}^{n} (a(\theta(U^k))\mathbf{k}\nabla(\tilde{u}^k - U^k), \nabla(\tilde{u}^k - U^k)){\eta_2}^{-k+1}\,\Delta t^k \\
&\quad \le Q\Big\{(\Delta t)^2 + \sum_{k=1}^{n} \|\partial(u - \tilde{u})^k\|_0^2{\eta_2}^{-k+1}\,\Delta t^k + \|u - \tilde{u}\|_{L^\infty(J;L^2(\Omega))}^2 \\
&\qquad + (\theta(u^0) - \theta(U^0), u^0 - U^0)\Big\} + \sum_{k=1}^{n} E^k{\eta_2}^{-k+1}\,\Delta t^k \\
&\quad \le Q\{h^{2(r+\gamma-1)} + (\Delta t)^2\} + \sum_{k=1}^{n} E^k{\eta_2}^{-k+1}\,\Delta t^k,
\end{aligned}$$

wherein we use (A18) to account for the time lag in a summation by parts term, (A15), and Lemma 3.

We now obtain bounds for E^n defined by (6.11). For some w between u^n and u^{n-1} and some W between u^n and U^n, by (A16)–(A17) and analogous to (5.12),

(6.13)
$$\begin{aligned}
E_1^n &= -\big[\partial_u\theta(w) - \partial_u\theta(u^n) + \partial_u\theta(u^n) - \partial_u\theta(W)\big](u^n - U^n)\partial u^n \\
&\le Q\big[|w - u^n|^\gamma + |\theta(u^n) - \theta(W)|^\beta\big]|u^n - U^n| \\
&\le Q\big[(\Delta t^n)^\gamma + |\theta(u^n) - \theta(U^n)|^\beta\big]|u^n - U^n| \\
&\le Q\big\{(\Delta t^n)^{2\gamma} + \big[(\theta(u^n) - \theta(U^n))(u^n - U^n)\big]^\delta\big\} + \epsilon|u^n - U^n|^2,
\end{aligned}$$

where δ can be either $2\beta/(1+\beta)$ or β. For the three terms in E_2^n, note that the last two are negative, so they can be used to control the other

term. By using the Mean Value Theorem twice, once for functions and once for integrals, note that for some w between u^n and u^{n-1},

$$\begin{aligned}(6.14)\qquad \int_{u^{n-1}}^{u^n} (\theta(u^n) - \theta(\mu))\, d\mu &= \int_{u^{n-1}}^{u^n} \partial_u\theta(w(\mu))(u^n - \mu)\, d\mu \\ &= \tfrac{1}{2}\partial_u\theta(w)(u^n - u^{n-1})^2.\end{aligned}$$

With a similar expression for the other integral term in E_2^n, we have for some W_1 and W_2 between U^n and U^{n-1} that

(6.15)
$$\begin{aligned}E_2^n &= \big[\partial_u\theta(W_1)\partial U^n\, \partial u^n - \tfrac{1}{2}\partial_u\theta(w)(\partial u^n)^2 - \tfrac{1}{2}\partial_u\theta(W_2)(\partial U^n)^2\big]\Delta t^n \\ &= \big[(\partial_u\theta(W_1) - \partial_u\theta(W_2))\partial U^n \partial u^n + \tfrac{1}{2}(\partial_u\theta(W_2) - \partial_u\theta(w))(\partial u^n)^2 \\ &\qquad - \tfrac{1}{2}\partial_u\theta(W_2)(\partial u^n - \partial U^n)^2\big]\Delta t^n \\ &\le Q\big\{|\theta(W_1) - \theta(W_2)|^\beta |U^n - U^{n-1}| + |W_2 - w|^\gamma \Delta t^n\big\}.\end{aligned}$$

This last expression is easily estimated:

(6.16)
$$\begin{aligned}&|\theta(W_1) - \theta(W_2)|^\beta |U^n - U^{n-1}| + |W_2 - w|^\gamma \Delta t^n \\ &\quad \le Q\big\{\big[|\theta(u^n) - \theta(U^n)|^\beta + |\theta(u^{n-1}) - \theta(U^{n-1})|^\beta + (\Delta t^n)^\beta\big] \\ &\qquad\quad \times \big[|u^n - U^n| + |u^{n-1} - U^{n-1}| + \Delta t^n\big] \\ &\qquad + \big[|u^n - U^n|^\gamma + |u^{n-1} - U^{n-1}|^\gamma + (\Delta t^n)^\gamma\big]\Delta t^n\big\} \\ &\quad \le Q\big\{|\theta(u^n) - \theta(U^n)|^{2\beta} + |\theta(u^{n-1}) - \theta(U^{n-1})|^{2\beta} + (\Delta t^n)^{2\beta}\big\} \\ &\qquad + \epsilon\big\{|u^n - U^n|^2 + |u^{n-1} - U^{n-1}|^2\big\},\end{aligned}$$

since $1/(2-\gamma) \ge \gamma \ge \beta$. Combining (6.13)–(6.16),

$$\begin{aligned}(6.17)\qquad E^n &= E_1^n + E_2^n \\ &\le Q\big\{(\Delta t^n)^{2\beta} + \big[(\theta(u^n) - \theta(U^n))(u^n - U^n)\big]^\beta \\ &\qquad + \big[(\theta(u^{n-1}) - \theta(U^{n-1}))(u^{n-1} - U^{n-1})\big]^\beta\big\} \\ &\quad + \epsilon\big\{|\tilde{u}^n - U^n|^2 + |u^n - \tilde{u}^n|^2 \\ &\qquad + |\tilde{u}^{n-1} - U^{n-1}|^2 + |u^{n-1} - \tilde{u}^{n-1}|^2\big\}.\end{aligned}$$

Combining (6.17) with (6.12), we obtain with Poincaré's inequality and (A18) that

(6.18)
$$\begin{aligned}
&\|\theta(u^n)-\theta(U^n)\|_0^2\eta_2^{-n}+\sum_{k=1}^{n}\|\nabla(\tilde{u}^k-U^k)\|_0^2\eta_2^{-k+1}\,\Delta t^k\\
&\quad\le Q\Big\{h^{2(r+\gamma-1)}+(\Delta t)^{2\beta}\\
&\qquad+\sum_{k=1}^{n}\big(\theta(u^k)-\theta(U^k),u^k-U^k\big)^\beta\eta_2^{-k}\,\Delta t^k\\
&\qquad+\big(\theta(u^0)-\theta(U^0),u^0-U^0\big)^\beta\Delta t^1+\|u^0-U^0\|_0^2\Delta t^1\Big\}\\
&\quad\le Q\Big\{h^{2(r+\gamma-1)}+(\Delta t)^{2\beta}+h^{2r\beta}\Delta t\\
&\qquad+\|\theta(u^n)-\theta(U^n)\|_0^\beta\|\nabla(u^n-U^n)\|_0^\beta\eta_2^{-n}\,\Delta t^n\\
&\qquad+\Big(\sum_{k=1}^{n-1}(\theta(u^k)-\theta(U^k),u^k-U^k)\eta_2^{-k/\beta}\,\Delta t^k\Big)^\beta\Big\}.
\end{aligned}$$

Finally, we require an induction argument, so for $N'\geq 1$ suppose that

$$\sum_{n=1}^{N'-1}\|\nabla(\tilde{u}^n-U^n)\|_0^2\,\Delta t^n\le Q_0h^{d\beta/(2\beta-1)}, \tag{6.19}$$

which is vacuous for $N'=1$. Fix $n=\bar{n}$ where $\|\theta(u^n)-\theta(U^n)\|_0$ attains its maximum on integers between 1 and N'. Note that in general

$$|ab|^\beta\Delta t\le\epsilon\{a^2+b^2\Delta t\}+Q(\Delta t)^{(2-\beta)/2(1-\beta)},$$

so (6.18) with (6.7) gives (if $2\beta>1$ and since $r+\gamma-1\geq r\beta$ and

$(2-\beta)/2(1-\beta) \geq 2\beta)$

(6.20)

$$\begin{aligned}
&\|\theta(u^{\bar{n}}) - \theta(U^{\bar{n}})\|_0^2 + \sum_{k=1}^{\bar{n}} \|\nabla(\tilde{u}^k - U^k)\|_0^2 \, \Delta t^k \\
&\quad \leq Q\Biggl\{ h^{2r\beta} + (\Delta t)^{2\beta} + \|\theta(u^{\bar{n}}) - \theta(U^{\bar{n}})\|_0^{\beta} \|\nabla(u^{\bar{n}} - U^{\bar{n}})\|_0^{\beta} \, \Delta t^{\bar{n}} \\
&\qquad + \Bigl(\epsilon h^{-d} \sum_{k=1}^{\bar{n}-1} \|\nabla(\tilde{u}^k - U^k)\|_0^2 \, \Delta t^k \, \|\theta(u^{\bar{n}}) - \theta(U^{\bar{n}})\|_0^2 \Bigr)^{\beta} \Biggr\} \\
&\quad \leq Q\Biggl\{ h^{2r\beta} + (\Delta t)^{2\beta} + \epsilon[\|\theta(u^{\bar{n}}) - \theta(U^{\bar{n}})\|_0^2 + \|\nabla(u^{\bar{n}} - U^{\bar{n}})\|_0^2 \, \Delta t^{\bar{n}}] \\
&\qquad + \epsilon^{\beta} \Biggl[\sum_{k=1}^{\bar{n}-1} \|\nabla(\tilde{u}^k - U^k)\|_0^2 \, \Delta t^k \\
&\qquad\quad + \Bigl(h^{-d\beta/(2\beta-1)} \sum_{k=1}^{\bar{n}-1} \|\nabla(\tilde{u}^k - U^k)\|_0^2 \, \Delta t^k \Bigr)^{(2\beta-1)/\beta} \\
&\qquad\qquad \times \|\theta(u^{\bar{n}}) - \theta(U^{\bar{n}})\|_0^2 \Biggr] \Biggr\}.
\end{aligned}$$

With the induction hypothesis (6.19) we hide four terms, and then a repetition of the above for $n = N'$ yields that

$$\begin{aligned}
(6.21) \quad &\max_{0 \leq n \leq N'} \|\theta(u^n) - \theta(U^n)\|_0^2 + \sum_{n=1}^{N'} \|\nabla(\tilde{u}^n - U^n)\|_0^2 \, \Delta t^n \\
&\qquad \leq Q\{h^{2r\beta} + (\Delta t)^{2\beta}\}.
\end{aligned}$$

To complete the induction argument, make the assumptions that $r > d/2(2\beta - 1) > 0$ and, as h and Δt tend to zero, $h^{-d/2(2\beta-1)} \Delta t = o(1)$. Then

$$(6.22) \quad \sum_{n=1}^{N'} \|\nabla(\tilde{u}^n - U^n)\|_0^2 \, \Delta t^n \leq Q\{h^{2r\beta} + (\Delta t)^{2\beta}\} \leq Q_0 h^{d\beta/(2\beta-1)}$$

for all $h < h_0$ for some fixed $h_0 > 0$, and thus the induction can be continued.

From (4.2)–(4.3), (6.7), (6.12), (6.18), and (6.20) we have the following results.

THEOREM 4. *Assume* (A1)–(A12), (A14)–(A19). *Let* U^n *be defined by* (2.3). *Define the four cases:*

(1) (A13) *holds*, $r > d/2(2\beta-1) > 0$, $h^{-d/2(2\beta-1)}\Delta t = o(1)$ *as* h *and* Δt *tend to zero, and* $\delta_1 = r\beta$, $\delta_2 = \beta$, $\bar{\delta}_1 = r$, $\bar{\delta}_2 = 1$;

(2) a *and* c *are independent of* θ *and* $\delta_1 = r\beta$, $\delta_2 = \beta$, $\bar{\delta}_1 = r$, $\bar{\delta}_2 = 1$;

(3) (A13) *holds*, $r > d(1+\beta)/2(3\beta-1) > 0$, $h^{-d\beta/(3\beta-1)}(\Delta t)^{\delta_2} = o(1)$ *as* h *and* Δt *tend to zero, and* $\delta_1 = 2r\beta/(1+\beta)$, $\delta_2 = \min\{1/2, \gamma\}$, $\bar{\delta}_1 = r$, $\bar{\delta}_2 = \delta_2(1+\beta)/2\beta$;

(4) a *and* c *are independent of* θ *and* $\delta_1 = 2r\beta/(1+\beta)$, $\delta_2 = \min\{1/2, \gamma, 2\beta/(1+\beta)\}$, $\bar{\delta}_1 = r$, $\bar{\delta}_2 = 1$.

Then in each case, for Δt *(and* h *in Cases* (1) *and* (3)*) sufficiently small,*

$$\max_{0\leq n\leq N} \|\theta(u^n) - \theta(U^n)\|_0 + \left\{\sum_{n=1}^{N} \|\nabla(\tilde{u}^n - U^n)\|_0^2 \,\Delta t^n\right\}^{1/2} \leq Q\{h^{\delta_1} + \Delta t^{\delta_2}\}, \tag{6.23}$$

$$\left\{\sum_{n=1}^{N} \|\nabla(u^n - U^n)\|_0^2 \,\Delta t^n\right\}^{1/2} \leq Q\{h^{\delta_1} + h^{r-1} + \Delta t^{\delta_2}\}, \tag{6.24}$$

$$\left\{\sum_{n=1}^{N} \left(\theta(u^n) - \theta(U^n), u^n - U^n\right) \Delta t^n\right\}^{1/2} \leq Q\{h^{\bar{\delta}_1} + (\Delta t)^{\bar{\delta}_2}\}. \tag{6.25}$$

PROOF: Case (1) is the given argument. The last estimate is optimal because of the assumptions relating r, d, β, γ, h, and Δt.

Case (2). The last term on the right side of (6.7) is missing, and this estimate directly completes (6.18) if we do not extract the nth term from the last sum.

For the last two cases, note that by Proposition 1, we can replace (6.15) with

$$E_2^n \leq Q|\theta(U^n) - \theta(U^{n-1})| - q(\Delta t^n)^{-1}|\theta(U^n) - \theta(U^{n-1})|^2 \leq Q\Delta t^n, \tag{6.15$'$}$$

and then use $\delta = 2\beta/(1+\beta)$ in (6.13). Continuing the argument, we obtain Cases (3)–(4) in a manner similar to Cases (1)–(2) (see also the proof of Theorem 1). ▮

In the strictly parabolic case we have the following results.

THEOREM 5. *Assume* (A1)–(A12), (A14)–(A19), *with* $\partial_u\theta(u)$ *uniformly positively bounded from below. Let* U^n *be defined by* (2.3). *Define the two cases:*

(1) (A13) *holds*, $r > d/2\gamma$, *and* $h^{-d/2\gamma}\Delta t = o(1)$ *as* h *and* Δt *tend to zero;*

(2) *a and c are independent of θ.*

Then in either case, for Δt (and h in Case (1)) sufficiently small, (6.23)–(6.25) *hold with $\delta_1 = r(1+\gamma)/2$, $\delta_2 = (1+\gamma)/2$, $\bar{\delta}_1 = r$, $\bar{\delta}_2 = 1$.*

PROOF: This theorem is analogous to Theorem 2. We recall (5.21) and that $\beta = \gamma$. Noting that

$$|a|^{\gamma}|b| \leq |a|^{1+\gamma} + |b|^{1+\gamma}$$

and simplifying (6.13)–(6.16), (6.17) can be replaced by

$$(6.17') \qquad E^n \leq Q\{(\Delta t^n)^{1+\gamma} + |u^n - U^n|^{1+\gamma} + |u^{n-1} - U^{n-1}|^{1+\gamma}\}.$$

This allows us to obtain the results as in the proof of Theorem 2. ∎

REMARK 7: Cases (1)–(2) of Theorem 4 (which do not make use of (6.15′)) and Theorem 5 give optimal bounds for the error in the case where uniformly in x, $\partial_u\theta(u)$ is positively bounded from below and uniformly Lipschitz as a function of u (recall then $\beta = \gamma = 1$). Of course in this case, the argument can be further simplified.

In the case of monotone flows, we have the following.

THEOREM 6. *Assume* (A1)–(A12), (A14)–(A19), *with either $\partial_t u \leq 0$ and $\partial_u\theta$ monotone nonincreasing in u, or $\partial_t u \geq 0$ and $\partial_u\theta$ monotone nondecreasing in u. Let U^n be defined by* (2.3)*. Then, for Δt sufficiently small,* (6.23)–(6.25) *hold with $\delta_1 = r + \gamma - 1$, $\delta_2 = \min\{1/2, \gamma\}$, $\bar{\delta}_1 = \delta_1$, $\bar{\delta}_2 = \delta_2$. Moreover, if* (A13) *holds, with now $\bar{\delta}_1 = \min\{r, 2\delta_1 - d/2\}$,*

$$(6.26) \quad \left\{\sum_{n=1}^{N} \left(\theta(u^n) - \theta(U^n), u^n - U^n\right) \Delta t^n\right\}^{1/2} \leq Q\{h^{\bar{\delta}_1} + \Delta t + \Delta t^{2\delta_2} h^{-d/2}\}.$$

PROOF: If $\partial_t u \leq 0$ and $\partial_u\theta$ decreases, or if $\partial_t u \geq 0$ and $\partial_u\theta$ increases, then in (6.13), $-\left[\partial_u\theta(u^n) - \partial_u\theta(W)\right](u^n - U^n)\partial u^n \leq 0$, so $E_1^n \leq Q(\Delta t^n)^{2\gamma} + \epsilon|u^n - U^n|^2$. With (6.15′), (6.12) is easily completed.

We have (6.26) because either the inverse inequality argument does not arise if we use (5.22), or (A13) allows us to follow the given argument. ∎

REFERENCES

1. H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z. **183** (1983), 311–341.
2. J. Bear, "Dynamics of Fluids in Porous Media," Elsevier, New York, 1972.

3. M. A. Celia, E. T. Bouloutas, and R. L. Zarba, *A general mass-conservative numerical solution for the unsaturated flow equation*, Wat. Resour. Res. **26** (1990), 1483–1496.
4. P. Ciarlet, "The Finite Element Method for Elliptic Problems," North-Holland, Amsterdam, 1978.
5. T. Dupont, G. Fairweather, and J. P. Johnson, *Three-level Galerkin methods for parăbolic equations*, SIAM J. Numer. Anal. **11** (1974), 392–410.
6. C. J. van Duyn and J. M. de Graaf, *Limiting profiles in contaminant transport through porous media*, SIAM J. Math. Anal. **18** (1987), 728–743.
7. R. A. Freeze and J. A. Cherry, "Groundwater," Prentice-Hall, Englewood Cliffs, New Jersey, 1979.
8. B. H. Gilding, *Improved theory for a nonlinear degenerate parabolic equation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **16** (1989), 165–224.
9. P. Grisvard, "Elliptic problems in nonsmooth domains," Pitman, Boston, 1985.
10. U. Hornung, *A parabolic-elliptic variational inequality*, Manuscripta Math. **39** (1982), 155–172.
11. J. Hulshof and N. Wolanski, *Monotone flows in n-dimensional partially saturated porous media: Lipschitz-continuity of the interface*, Arch. Rational Mech. Anal. **102** (1988), 287–305.
12. P. Knabner, *Finite-element simulation of saturated-unsaturated flow through porous media*, preprint.
13. H. H. Rachford, Jr., *Two-level discrete-time Galerkin approximations for second order nonlinear parabolic partial differential equations*, SIAM J. Numer. Anal. **10** (1973), 1010–1026.
14. R. Rannacher and R. Scott, *Some optimal error estimates for piecewise linear finite element approximations*, Math. Comp. **38** (1982), 437–445.
15. A. H. Schatz and L. B. Wahlbin, *On the quasi-optimality in L_∞ of the $\overset{o}{H}{}^1$ projection into finite element spaces*, Math. Comp. **38** (1982), 1–22.
16. R. Scott, *Optimal L^∞ estimates for the finite element method on irregular meshes*, Math. Comp. **30** (1976), 681–697.
17. M. F. Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal. **10** (1973), 723–759.

Keywords. error estimates, elliptic-parabolic, nonlinear, Galerkin, finite element, Richard's equation
1980 *Mathematics subject classifications*: 65M15, 65M60, 76S05

Department of Mathematics, Purdue University, West Lafayette, Indiana 47907