A Finite Volume Element Method for a Nonlinear Elliptic Problem

P. Chatzipantelidis, V. Ginting and R. D. Lazarov

Department of Mathematics, Texas A&M University, College Station, TX, 77843

Dedicated to Owe Axelsson on the occasion of his 70th birthday

SUMMARY

We consider a finite volume discretization of second order nonlinear elliptic boundary value problems on polygonal domains. For sufficiently small data, we show existence and uniqueness of the finite volume solution using a fixed point iteration method. We derive error estimates in $H^1$, $L_2$ and $L_\infty$-norm. In addition a Newton’s method is analyzed for the approximation of the finite volume solution and numerical experiments are presented. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: finite volume element method, nonlinear elliptic equation, error estimates, fixed point iterations, Newton’s method

1. INTRODUCTION

We analyze a finite volume element method for the discretization of second order nonlinear elliptic partial differential equations on a polygonal domain $\Omega \subset \mathbb{R}^2$. Namely, for a given function $f$ we seek $u$ such that

$$L(u)u \equiv -\nabla \cdot (A(u) \nabla u) = f \quad \text{in } \Omega, \quad \text{and} \quad u = 0, \quad \text{on } \partial \Omega,$$

with $A: \mathbb{R} \to \mathbb{R}$ sufficiently smooth such that there exist constants $\beta_i$, $i = 1, 2, 3$, satisfying

$$0 < \beta_1 \leq A(x) \leq \beta_2, \quad |A'(x)| \leq \beta_3, \quad \text{for } x \in \mathbb{R}.$$

(1.2)

Finite volume approximations rely on the local conservation property expressed by the differential equation. Namely, integrating (1.1) over any region $V \subset \Omega$ and using Green’s formula, we obtain

$$-\int_{\partial V} (A(u) \nabla u) \cdot n \, ds = \int_V f \, dx,$$

(1.3)
where \( n \) denotes the unit exterior normal to \( \partial V \).

There are various approaches in deriving finite volume approximations of nonlinear elliptic equations. One, often called finite volume element method, uses a finite element partition of \( \Omega \), where the solution space consists of continuous piecewise linear functions, a collection of vertex centered control volumes and a test space of piecewise constant functions over the control volumes, cf., e.g., [5, 20, 19]. A second approach, usually called finite volume difference method, uses cell-centered grids and approximates the derivatives in the balance equation by \( \{T_z\}_{\partial h < 1} \) is a family of quasi-uniform triangulations of \( \Omega \), \( h \) denotes the maximum diameter of the triangles of \( T_h \). The discrete finite volume problem will satisfy a relation similar to (1.3) for \( V \) in a finite collection of subregions of \( \Omega \) called control volumes, the number of which will be equal to the dimension of the finite element space \( X_h \). These control volumes are constructed in the following way. Let \( z_K \) be the barycenter of \( K \in T_h \). We connect \( z_K \) with line segments to the midpoints of the edges of \( K \), thus partitioning \( K \) into three quadrilaterals \( K_z \), \( z \in Z_h(K) \), where \( Z_h(K) \) are the vertexes of \( K \). Then with each vertex \( z \in Z_h = \bigcup_{K \in T_h} Z_h(K) \) we associate a control volume \( V_z \), which consists of the union of the subregions \( K_z \), sharing the vertex \( z \) (see Figure 1). We denote the set of interior vertexes of \( Z_h \) by \( Z^0_h \).

The finite volume method is then to find \( u_h \in X_h \) such that

\[
- \int_{\partial V_z} (A(u_h) \nabla u_h) \cdot n \, ds = \int_{V_z} f \, dx, \quad \forall z \in Z^0_h, \tag{1.4}
\]
The Galerkin finite element method for (1.1) is: Find $u_h \in X_h$ such that

$$a(u_h; u_h, \chi) = (f, \chi), \quad \forall \chi \in X_h,$$

with $a(\cdot; \cdot, \cdot)$ the form defined by

$$a(v; w, \phi) = \int_{\Omega} A(v) \nabla w \cdot \nabla \phi \, dx.$$ 

It is known that the solution $u_h$ of (1.5) satisfies

$$|u_h - u| + h|\nabla(u_h - u)| \leq C(u, f)h^2$$

$$|u_h - u|_{L^\infty} \leq C_p \inf_{\chi \in X_h} \|\nabla(u - \chi)\|_{W_{1,p}^1}, \text{ with } p > 2.$$

Numerical methods for this type and more general problems has been considered by many authors, cf., e.g., [4, 13, 18, 21].

Here for sufficiently small data we shall derive similar results for the finite volume method. Li, in [20], considers a variation of the finite volume method under investigation here. The method differs in the construction of the control volumes. Instead of the barycenter $z_K$, the circumcenter is selected. For this finite volume method similar results with the finite element method, for the $H^1$-norm error estimate, are valid.

In Section 3, we establish existence of the finite volume solution $u_h$ of (2.3), using a fixed point iteration method. In particular, in Theorem 3.1 we show that the iterations remain inside a fixed ball with a radius that depends only on $f$. Then in Theorem 3.2 we show that for a sufficiently small data, $f$, the fixed point iteration operator is Lipschitz continuous with Lipschitz constant less that 1.

In Section 4 we derive optimal order $H^1$, $L_2$ and almost optimal $L_\infty$-norm error estimates. Note that for the $L_2$ estimation we assume that $A'$ is also Lipschitz continuous, $A' \in L_1(\mathbb{R})$ and $f \in H^1$.

Also in Section 5 we analyze a Newton’s method for the approximation of the finite volume solution $u_h$. We consider an inexact Newton iteration, a variant of the Newton iteration for nonlinear systems of equations, where the Jacobian of the system is solved approximately, cf., e.g., [2, 3, 11]. A similar approach for the finite element method is analyzed by Douglas and Dupont in [13]. As it is expected, one has to start the Newton iteration with an initial approximation $u_h^0$ sufficiently close to $u_h$. Also, following [13], we show that the Newton iterations converge to $u_h$ with order 2. Finally in Section 6 numerical results are presented.
2. PRELIMINARIES–THE FINITE VOLUME METHOD

There has been a tendency of analyzing finite volume element method using the existing results from its finite element counterpart, cf., e.g., [7, 8, 9, 10]. The investigations recorded in all these references were concentrated on elliptic and/or parabolic problems with coefficients independent of the solution, i.e., the function $A$ is only spatially varied. The finite volume element method is viewed as a perturbation of standard Galerkin finite element method with the help of an interpolation operator $I_h : C(\Omega) \rightarrow Y_h$, defined by

$$ I_h v = \sum_{z \in Z_h^0} v(z) \Psi_z, \quad (2.1) $$

where

$$ Y_h = \{ \eta \in L_2(\Omega) : \eta|_{V_z} = \text{constant}, \forall z \in Z_h^0; \eta|_{\partial \Omega} = 0, \forall z \in \partial \Omega \}, $$

and $\Psi_z$ is characteristic function of $V_z$. We note that $I_h : X_h \rightarrow Y_h$ is a bijection and bounded with respect to the $L_2$–norm, i.e., there exist $c_1, c_2 > 0$, such that

$$ c_1 \| \chi \| \leq \| I_h \chi \| \leq c_2 \| \chi \|, \quad \forall \chi \in X_h. \quad (2.2) $$

The finite volume problem (1.4) can be rewritten in a variational form. For an arbitrary $\eta \in Y_h$, we multiply the integral relation in (1.4) by $\eta(z)$ and sum over all $z \in Z_h^0$ to obtain the Petrov–Galerkin formulation, to find $u_h \in X_h$ such that

$$ a_h(u_h; u_h, \eta) = (f, \eta), \quad \forall \eta \in Y_h, \quad (2.3) $$

where the form $a_h(\cdot; \cdot, \cdot) : X_h \times X_h \times Y_h \rightarrow \mathbb{R}$ is defined by

$$ a_h(w; v, \eta) = -\sum_{z \in Z_h^0} \eta(z) \int_{\partial V_z} (A(w)\nabla v) \cdot n \, ds, \quad v, w \in X_h, \ \eta \in Y_h. \quad (2.4) $$

Obviously, $a_h(w; v, \eta)$ may be defined by (2.4) also for $v, w \in W_p^1(\Omega) \cap H_0^1(\Omega)$, $p > 2$, and using Green’s formula we easily see that

$$ a_h(w; v, \eta) = (L(w)v, \eta), \quad \text{for} \ v, w \in W_p^1(\Omega) \cap H_0^1(\Omega), \ \eta \in Y_h. \quad (2.5) $$

The bilinear form $a_h(w; \cdot, \cdot)$, with $w \in L_\infty$, of (2.4) may equivalently be written as

$$ a_h(w; v, \eta) = \sum_K \{ (L(w)v, \eta)_K + (A(w)\nabla v \cdot n, \eta)_{\partial K} \}, \quad \forall v \in X_h, \ \eta \in Y_h. \quad (2.6) $$

Indeed, by integration by parts, we obtain, for $z \in Z_h^0$ and $K \in T_h$,

$$ \int_{K_z} L(w)v \ dx = -\int_{\partial K_z \cap \partial K} (A(w)\nabla v) \cdot n \, ds - \int_{\partial K_z \cap \partial V_z} (A(w)\nabla v) \cdot n \, ds, \quad (2.7) $$

and (2.6) hence follows by multiplication by $\eta(z)$ and by summation first over the triangles that have $z$ as a vertex and then over the vertexes $z \in Z_h^0$. Also, we can easily see that $I_h$ has the following properties, cf., e.g., [7],

$$ \int_K I_h \chi \ dx = \int_K \chi \ dx, \ \forall \chi \in X_h, \ \text{for any} \ K \in T_h, \quad (2.8) $$
\[
\int_{e} I_h \chi \, ds = \int_{e} \chi \, ds, \quad \forall \chi \in X_h, \quad \text{for any side } e \text{ of } K \in T_h, \tag{2.9}
\]
\[
\|I_h \chi\|_{L^{\infty}(e)} \leq \|\chi\|_{L^{\infty}(e)}, \quad \forall \chi \in X_h, \quad \text{for any side } e \text{ of } K \in T_h, \tag{2.10}
\]
\[
\|\chi - I_h \chi\|_{L^p(K)} \leq h\|\nabla \chi\|_{L^1(K)}, \quad \forall \chi \in X_h, \quad 1 \leq p < \infty. \tag{2.11}
\]

In addition in \cite[Lemma 6.1, Remark 6.1, Lemma 5.1]{7} the following lemma was derived.

**Lemma 2.1.** Let \(e\) be a side of a triangle \(K \in T_h\). Then for \(v \in W^1_0(K)\) there exists a constant \(C_1 > 0\) independent of \(h\) such that
\[
|\int_{e} v(\chi - I_h \chi) \, ds| \leq C_1 h \|\nabla v\|_{L_p(K)} \|\nabla \chi\|_{L'_p(K)}, \quad \forall \chi \in X_h, \quad \text{with } \frac{1}{p} + \frac{1}{p'} = 1. \tag{2.12}
\]

Also, for \(f \in W^i_p, i = 0, 1\) and \(\chi \in X_h,\)
\[
|\varepsilon_h(f, \chi)| \leq Ch^i+\frac{1}{2} \|f\|_{W^i_p} \|\chi\|_{W^i_p}, \quad f \in W^i_p, \quad i, j = 0, 1, \quad \text{with } \frac{1}{p} + \frac{1}{p'} = 1, \tag{2.13}
\]
where \(\varepsilon_h : L_2 \times X_h \to \mathbb{R}\) is defined by
\[
\varepsilon_h(f, \chi) = (f, \chi - I_h \chi). \tag{2.14}
\]

**Lemma 2.2.** Let \(v \in W^2_q, 4/3 < q \leq 2\). The following identities hold.
\[
\sum_{K} \int_{\partial K} A(\tilde{w}) \nabla v \cdot n \, ds = 0, \tag{2.15}
\]
where \(\tilde{w}\) could be an element of \(X_h\) or the point value at the midpoint of the edge \(e\) of triangle \(K\), or an element of \(X_h\).

*Proof.* Note, that for \(v \in W^2_q\), the trace \(\nabla v \cdot n\) on \(\partial K\) exists for \(q > 4/3\). The left identity is obvious by rewriting the sum as integrals of jump terms over the interior edges of \(T_h\). These jumps obviously vanish due to the continuity of \(A(\tilde{w}) \nabla v \cdot n\) (in the trace sense). A similar argument gives the second identity. \(\square\)

Our analysis will be based on the corresponding one for linear problems, cf., e.g., \cite{7, 8}. There the error estimations are derived by bounding the error between the bilinear forms of the finite element, \(a\), and the finite volume methods, \(a_h\). This is shown to be \(O(h)\) uniformly in \(X_h\). Then for sufficiently small \(h\) the finite volume bilinear form \(a_h\) is coercive in \(X_h\), which leads to the existence and uniqueness of the finite volume approximation.

In the nonlinear case a similar estimation for the error functional \(\varepsilon_a\),
\[
\varepsilon_a(w; v_h, \chi) = a(w; v_h, \chi) - a_h(w; v_h, I_h \chi), \quad \forall v_h, \chi \in X_h, \quad w \in L_{\infty}, \tag{2.16}
\]
shows that this error is not \(O(h)\) uniformly in \(X_h\), cf. Lemma 2.3. This is due to the fact that the bound of \(\varepsilon_a(w; v_h, \chi)\), will depend on \(\|w_h\|_{L_{\infty}}\). Inverse inequalities of the form, cf., e.g., \cite{6},
\[
\|\nabla \chi\|_{L^s} \leq Ch^{2/2-s/2t} \|\nabla \chi\|_{L^t}, \quad \forall \chi \in X_h, \quad \text{with } 1 \leq t \leq s \leq \infty, \tag{2.17}
\]
which are true in a quasi-uniform mesh, give \(\varepsilon_a = O(h^{1-2/t})\), uniformly in a ball of \(X_h\) with respect to \(W^1_0\)-norm, for \(t > 2\).

In the sequel we derive estimations for \(\varepsilon_a\).
Lemma 2.3. There exists a constant $C_2 > 0$, independent of $h$, such that

$$|\varepsilon_a(w_h; v_h, \chi)| \leq C_2 \beta h \|\nabla w_h \cdot \nabla v_h\|_{L^p} \|\nabla \chi\|_{L^{p'}},$$  \quad \forall w_h, v_h, \chi \in X_h, \quad \frac{1}{p} + \frac{1}{p'} = 1. \quad (2.18)$$

Proof. In view of Green’s formula and (2.6), we may write $\varepsilon_a$ in the following form:

$$\varepsilon_a(w_h; v_h, \chi) = \sum_K \left\{ (L(w_h)v_h, \chi - I_h \chi)_{K} + (A(w_h)\nabla v_h \cdot n, \chi - I_h \chi)_{\partial K} \right\}$$

$$= \sum_K \{I_K + II_K\}. \quad (2.19)$$

Applying Hölder’s inequality to $I_K$, and using the fact that $w_h$ and $v_h$ are linear in $K$, and using (1.2) and (2.11), we have

$$|I_K| \leq \beta h \|\nabla w_h \cdot \nabla v_h\|_{L^p(K)} \|\chi - I_h \chi\|_{L^{p'}(K)} \leq \beta h \|\nabla w_h \cdot \nabla v_h\|_{L^p(K)} \|\nabla \chi\|_{L^{p'}(K)}. \quad (2.20)$$

For the $II_K$, we break the integration over the boundary of each triangle $K$, into the sum of integrations over its sides, and thus may use (2.12), and follow the same steps as in estimating $I_K$. Hence,

$$|II_K| \leq C_1 h |A(w_h)\nabla v_h|_{W^{1,2}_q(K)} \|\nabla \chi\|_{L^{p'}(K)} \leq C_1 \beta h \|\nabla w_h \cdot \nabla v_h\|_{L^p(K)} \|\nabla \chi\|_{L^{p'}(K)}. \quad (2.21)$$

Finally, (2.20) and (2.21) establish the desired estimate for $C_2 = C_1 + 1$. $\Box$

The following lemma will be used in Section 4 to estimate the error in the $L_2$-norm. For this estimation we will need to assume that $A'$ is Lipschitz continuous with constant $L$, i.e.

$$|A'(x) - A'(y)| \leq L|x - y|, \quad \forall x, y \in \mathbb{R}. \quad (2.22)$$

Lemma 2.4. Assume that $A'$ is Lipschitz continuous and $v \in W^2_q \cap H^1_0$, for $4/3 < q \leq 2$. Then there exists a constant $C > 0$ independent of $h$ such that for $w_h, v_h, \chi \in X_h$,

$$|\varepsilon_a(w_h; v_h, \chi)| \leq C h \left\{ |\nabla w_h|_{L^\infty} \left( |\nabla w_h \cdot \nabla v_h| + \|v\|_{W^2_q} \right) + h \|\nabla w_h \cdot \nabla (v_h - v)|_{L^q} \right\} |\nabla \chi|_{L^4}, \quad (2.23)$$

with $1/q + 1/q' = 1$.

Proof. Let $w_K$ and $w_e$ denote the average value of a function $w$ over triangle $K$ and the edge $e$, respectively. Since $v \in W^2_q$, Lemma 2.2 gives the identity

$$(A(w_h) - A(w_h,e))\nabla v \cdot n, \chi - I_h \chi)_{\partial K} = 0, \quad \forall \chi \in X_h.$$  

Employing this identity, the fact that $v_h$ is linear in $K$, Green’s formula, and (2.8) we get

$$\varepsilon_a(w_h; v_h, \chi) = \sum_K \left\{ (A'(w_h) - A'(w_h,K))\nabla w_h \cdot \nabla v_h, \chi - I_h \chi\right\}_K$$

$$+ \sum_K \left\{ (A(w_h) - A(w_h,e))\nabla (v_h - v) \cdot n, \chi - I_h \chi\right\}_{\partial K} = \sum_K \{I_K + II_K\}.$$  

Using now Hölder’s inequality, the fact that $w_h$ is linear in $K$, and (2.11), we can bound $I_K$,

$$|I_K| \leq C \int_K |w_h - w_h,K| |\nabla w_h \cdot \nabla v_h| |\chi - I_h \chi| \, dx$$

$$\leq Ch^2 \|\nabla w_h\|_{L^\infty} \|\nabla w_h \cdot \nabla v_h\|_{L^2(K)} \|\nabla \chi\|_{L^2(K)}. \quad (2.24)$$
For the estimation of $II_K$, we apply (2.12) and we get,
\[
|II_K| \leq Ch \left|(A(w_h) - A(w_{h,c}))\nabla (v_h - v)\right|_{W^1_2(K)} \|\nabla \chi\|_{L_p^\infty(K)}. \tag{2.25}
\]

Further, a simple calculation gives
\[
|(A(w_h) - A(w_{h,c}))\nabla (v_h - v)\|_{W^1_2(K)} \leq C(\|\nabla w_h \cdot \nabla (v_h - v)\|_{L_p(K)} + h}\|\nabla w_h\|_{L^\infty}\|v\|_{W^1_2(K)}).
\]

Summing now over all triangles, the relation above, (2.24), (2.25) and using the fact that $q' > 2$, we obtain (2.23). □

Next we will derive a “Lipschitz”-type estimation for $\varepsilon_a$.

**Lemma 2.5.** Let $v \in H^1 \cap L_\infty$, $w \in W^1_p$ with $p > 2$ and $A'$ be Lipschitz continuous with constant $L$, cf. (2.22). There exists $C_2 > 0$ such that
\[
|\varepsilon_a(v; \phi_h, \chi) - \varepsilon_a(w; \phi_h, \chi)| \leq C_2 h\|\nabla \phi_h\|_{L_\infty} (\beta_3 + L\|\nabla w\|_{L_p}) \|\nabla (v - w)\| \|\nabla \chi\|, \quad \forall \phi_h, \chi \in X_h,
\tag{2.26}
\]
where $\beta_3$ is the upper bound of $A'$, cf., (1.2).

**Proof.** We can easily see that
\[
\varepsilon_a(v; \phi_h, \chi) - \varepsilon_a(w; \phi_h, \chi) = \sum_K \left\{ \int_K \text{div}((A(v) - A(w))\nabla \phi_h)(\chi - I_h \chi) \, dx \\
+ \int_{\partial K} (A(v) - A(w))\nabla \phi_h \cdot n(\chi - I_h \chi) \, ds \right\}.
\]

Also, since $\phi_h$ is linear in $K$, $\text{div} (\nabla \phi_h) = 0$, therefore,
\[
\text{div}((A(v) - A(w))\nabla \phi_h) = \{ A'(v)\nabla (v - w) + (A'(v) - A'(w))\nabla w \} \cdot \nabla \phi_h, \quad \text{in } K.
\]

Then, this, (2.11), (2.12), the Hölder inequality
\[
\|vw\|_{L_t} \leq \|v\|_{L_s} \|w\|_{L_t}, \quad \text{with } t > s, \quad \frac{s}{t} + \frac{s}{l} = 1, \tag{2.27}
\]
for $s = 2$ and $t = p$ and the Sobolev inequality, cf. e.g., [6, 4.x.11],
\[
\|v\|_{L_s} \leq \|\nabla v\|, \quad \forall s < \infty, \tag{2.28}
\]
give for $C_2 = C_1 + 1$
\[
|\varepsilon_a(v; \phi_h, \chi) - \varepsilon_a(w; \phi_h, \chi)| \leq C_2 h(\beta_3\|\nabla (v - w)\| + L\|v - w\|\|\nabla w\|)\|\nabla \chi\| \|\nabla \phi_h\|_{L_\infty} \\
\leq C_2 h(\beta_3\|\nabla (v - w)\| + L\|v - w\|_{L_p}\|\nabla w\|_{L_p})\|\nabla \chi\| \|\nabla \phi_h\|_{L_\infty} \\
\leq C_2 h(\beta_3 + L\|\nabla w\|_{L_p})\|\nabla (v - w)\| \|\nabla \chi\| \|\nabla \phi_h\|_{L_\infty}. \quad \Box
\]

3. EXISTENCE OF FVE APPROXIMATIONS FOR SMALL DATA

In this section using a fixed point iteration we will show that a finite volume solution $u_h$ of (2.3) exists and is in the ball
\[
B_M = \{ \chi \in X_h : \|\nabla \chi\|_{L_p} \leq M \}, \quad \text{with } p > 2,
\]

Copyright © 2004 John Wiley & Sons, Ltd. Numer. Linear Algebra Appl. 2004; 00:1–26
Prepared using nlaauth.cls
where $M = M(f) > 0$, cf. Theorem 3.1. Further, if $M$ is sufficiently small, i.e., an appropriate norm of $f$ is small, the finite volume solution $u_h$ is unique, cf., Corollary 3.3.

For a fixed $f \in L^2$, we consider the iteration map $T_h : X_h \to X_h$ given by

$$a_h(v_h; T_h v_h, \eta) = (f, \eta), \quad \forall \eta \in Y_h. \quad \text{(3.1)}$$

In view of the Sobolev imbedding, $\|v\|_{L^\infty} \leq C\|v\|_{W^1_p}$, for $p > 2$, we shall employ the following inf-sup condition, cf., e.g., [6, Chapter 7]: There exist constants $\alpha = \alpha(A, \Omega) > 0$, $h_\alpha > 0$ and $\epsilon = \epsilon(A, \Omega) > 0$ such that for all $0 < h \leq h_\alpha$ and $v_h \in X_h$ and $w \in L^\infty$,

$$\|\nabla v_h\|_{L^p} \leq \alpha \sup_{0 \neq \chi \in X_h} \frac{a(w; v_h, \chi)}{\|\nabla \chi\|_{L^p}}, \quad \text{(3.2)}$$

with $2 \leq p \leq 2 + \epsilon$ and $\frac{1}{p} + \frac{1}{p'} = 1$.

In view of the identity $a(w_h; v_h, \chi) = a_h(w_h; v_h, I_h \chi) + \epsilon a(w_h; v_h, \chi)$ and the error estimate in Lemma 2.3 and (2.17),

$$|\epsilon a(w_h; v_h, \chi)| \leq Ch\|\nabla w_h \cdot \nabla v_h\|_{L^p}\|\nabla \chi\|_{L_p^\prime} \leq C h^{1-2/p}\|\nabla w_h\|_{L^p}\|\nabla v_h\|_{L^p}\|\nabla \chi\|_{L_p^\prime},$$

there exists $h_M > 0$ such that for all $0 < h \leq h_M \leq h_\alpha$

$$\|\nabla v_h\|_{L_p} \leq \alpha \sup_{0 \neq \chi \in X_h} \frac{a(w_h; v_h, I_h \chi)}{\|\nabla \chi\|_{L_p^\prime}}, \quad \forall v_h \in X_h, w_h \in B_M, \quad 2 < p < 2 + \epsilon. \quad \text{(3.3)}$$

Therefore, for $h < h_M$ and $v_h \in B_M$, $T_h v_h$ is well defined. Note that (3.3) holds also for $p = 2$ and $w_h \in B_M = \{\chi \in X_h : \|\nabla \chi\|_{L^2} \leq M\}$, with $\bar{p} > 2$.

In the following two theorems we will show that in a sufficiently small ball $B_M$ and data $f$, there exists a unique solution $u_h \in X_h$ of (2.3).

**Theorem 3.1.** There exists $h_M > 0$, such that for all $0 < h < h_M$, if $\|f\| \leq M \alpha^{-1}$ then $T_h$ maps $B_M$ into itself for $2 < p < 2 + \epsilon$.

**Proof.** Let $v_h \in B_M$ then in view of (3.3) we have

$$\|\nabla T_h v_h\|_{L_p} \leq \alpha \sup_{0 \neq \chi \in X_h} \frac{a_h(w_h; T_h v_h, I_h \chi)}{\|\nabla \chi\|_{L_p^\prime}} \leq \alpha \sup_{0 \neq \chi \in X_h} \frac{(f, I_h \chi)}{\|\nabla \chi\|_{L_p^\prime}}, \quad \text{(3.4)}$$

Then, using (2.2) and the Sobolev inequality $\|v\| \leq \|v\|_{W^2_p}$, for $p > 1$, cf. [6, 4.x.11], we get

$$\|\nabla T_h v_h\|_{L_p} \leq \alpha \|f\|, \quad \text{(3.5)}$$

which gives the desired result. \(\square\)

Next, we will show that the iteration map $T_h$ is Lipschitz continuous. For $M$ sufficiently small, $T_h$ is a contraction in $B_M$ in $H^1$-norm, which gives the uniqueness of the solution $u_h$ of (2.3) and the convergence of the fixed point iteration, $v_{h}^{n+1} = T_h v_{h}^{n} \to u_h$, as $n \to \infty$.

**Theorem 3.2.** Let $A'$ be Lipschitz continuous with constant $L$, cf. (2.22). Then there exists a constant $C_L = C_L(A, \Omega) > 0$ and $h'_M > 0$, such that for $\|f\| \leq M \alpha^{-1}$, $M < C_L^{-1}$ and all $0 < h \leq h'_M$, $T_h$ is a contraction, with constant $\ell = C_L M < 1$,

$$\|\nabla (T_h v - T_h w)\| \leq \ell \|\nabla (v - w)\|, \quad \forall v, w \in B_M. \quad \text{(3.6)}$$
Proof. Let \( v, w \in B_M \). Then, in view of the definition of \( T_h \), (3.1) we have
\[
a_h(v; T_h v, \eta) - a_h(w; T_h w, \eta) = 0, \quad \forall \eta \in Y_h.
\]
Therefore, we can easily see that for \( \eta = I_h \chi, \chi \in X_h \),
\[
a_h(v; T_h v - T_h w, I_h \chi) = a_h(w; T_h w, I_h \chi) - a_h(v; T_h w, I_h \chi)
= \varepsilon_a(v; T_h w, \chi) - \varepsilon_a(w; T_h w, \chi) + ((A(w) - A(v))\nabla T_h w, \nabla \chi). \tag{3.7}
\]
Using now the fact that for sufficiently small \( h, T_h w \in B_M \), cf., Theorem 3.1, the Hölder inequality (2.27) with \( s = 2 \) and \( t = p \) and the Sobolev (2.28), the last term of the right–hand side of (3.7) can be bounded for any \( \chi \in X_h \),
\[
|((A(w) - A(v))\nabla T_h w, \nabla \chi)| \leq \beta_3\|w - v\|_{L_p} \|\nabla \chi\| \leq \beta_3\|w - v\|_{L_p} \|\nabla \chi\| \leq \beta_3 M\|\nabla(v - w)\| \|\nabla \chi\|. \tag{3.8}
\]
Also, in view of Lemma 2.5 the remaining two terms in the right–handside of (3.7), give
\[
|\varepsilon_a(v; T_h w, \chi) - \varepsilon_a(w; T_h w, \chi)| \leq C_2 h^{1-2/p} M(\beta_3 + LM)\|\nabla(v - w)\| \|\nabla \chi\|. \tag{3.9}
\]
Since, \( a_h(v_k, w_k) \) is coercive for \( v_k \in B_M \) and \( h \) sufficiently small, choosing \( \chi = T_h v - T_h w \) in the above relation and in (3.7) and (3.8) gives that there exists a constant \( C_L = C(L, A, \Omega) > 0 \) such that
\[
\|\nabla(T_h v - T_h w)\| \leq C_L \|\nabla(v - w)\|.
\]
Therefore, for \( M < C_L^{-1} \), \( T_h \) is a contraction with constant \( 0 < \ell = C_L M < 1 \). \( \square \)

Finally, Theorems 3.1 and 3.2 give the following corollary,

**Corollary 3.3.** Assume that \( \alpha' \) is Lipschitz continuous with a constant \( L \). Then there exist constants \( C_L = C_L(A, \Omega) > 0 \) and \( \alpha > 0 \) such that if \( \|f\| \leq \alpha^{-1} C_L^{-1} \), with \( 2 < p < 2 + \epsilon \) then for \( h \) sufficiently small the problem (2.3), i.e., find \( u_h \in X_h \) such that
\[
a_h(u_h; u_h, I_h \chi) = (f, I_h \chi), \quad \forall \chi \in X_h,
\]
has a unique solution, with \( \epsilon \) given in (3.2).

### 4. ERROR ESTIMATES

In this section we shall derive \( W_1^s \), with \( 2 \leq s < p \), \( L_2 \)– and \( L_\infty \)-norm error estimates for the error \( u_h - u \) for \( f \in L_2 \). We shall assume that the nonlinear problem (1.1) has a unique solution \( u \in W_2^2 \cap H_0^1 \), with \( 4/3 < q \leq 2 \). In Section 3 we show that a finite volume solution \( u_h \) of (2.3) exists and is unique.

First, we will derive an a priori error estimate in \( \|\nabla \cdot \|_{L_p}, 2 \leq s < p \), norm. For \( s = 2 \) we get the usual \( H^1 \)–norm error bound. But for \( s > 2 \) this estimate combined with a standard Sobolev imbedding gives an \( L_\infty \)-norm error estimate, cf. Theorem 4.2.

**Theorem 4.1.** Let \( u_k \) and \( u \) be the solutions of (2.3) and (1.1), respectively, with \( f \in L_2 \). Then, if \( \gamma = \alpha \beta_3 M < 1 \) there exists a constant \( C = C(u, f) \), independent of \( h \), such that for \( 0 < h \leq h_M \)
\[
\|\nabla(u_h - u)\|_{L_p} \leq C(u, f)h^{1+2/s-2/q}, \quad \text{with} \quad 2 \leq s < p < 2 + \epsilon, \quad \frac{4}{3} < q \leq 2, \tag{4.1}
\]
where \( \alpha \) is the constant appeared in (3.2).
Proof. Using the triangle inequality we get
\[ \|\nabla(u_h - u)\|_{L^s} \leq \|\nabla(u - \chi)\|_{L^s} + \|\nabla(u_h - \chi)\|_{L^s}, \quad \forall \chi \in X_h. \quad (4.2) \]
In view of the approximation property of \( X_h \),
\[ \inf_{\chi \in X_h} \|\nabla(v - \chi)\|_{L^s} \leq C h^{1+2s/2} \|v\|_{W^2}, \quad \text{with} \ 4/3 < q \leq 2 \leq s, \quad (4.3) \]
the first term of the right-hand side of (4.2) is bounded as desired. Also, we can easily see that

\[ a(u; u_h - \chi, \psi) = a(u; u_h - u, \psi) + a(u; u - \chi, \psi) \leq a(u; u_h - u, \psi) + \beta_2 \|\nabla(u - \chi)\|_{L^s} \|\nabla\psi\|_{L^s}, \]

with \( 1/s + 1/s' = 1 \). Hence, in view of (3.2), we may write for \( 2 \leq s < p \),
\[ a(u; u_h - u, \psi) = a(u; u_h, \psi) - (f, \psi) \]
\[ = \{a(u; u_h, \psi) - a(u_h; u_h, \psi)\} + \{\varepsilon(a(u_h; u_h, \psi) - \varepsilon_h(f, \psi)\} = I + II. \quad (4.5) \]
Using then the fact that \( u_h \in B_M \), the Hölder inequality (2.27), with \( t = p \), and the Sobolev inequality (2.28), we have for any \( \chi, \psi \in X_h \),
\[ |I| = |a(u; u_h, \psi) - a(u_h; u_h, \psi)| \leq \beta_3 \|u_h - u\|_{L^p} \|\nabla u_h\|_{L^q} \|\nabla\psi\|_{L^{s'}} \]
\[ \leq \beta_3 M \|\nabla(u_h - \chi)\|_{L^s} + \|\nabla(u - \chi)\|_{L^s} \|\nabla\psi\|_{L^{s'}}. \quad (4.6) \]
The remaining term \( II \) can be bounded using Lemma 2.3 and (2.13), the inverse inequality (2.17) and the Hölder inequality (2.27), with \( t = 2q/(2 - q) \) and \( \ell = st/(t - s) \),
\[ |\varepsilon_h(f, \psi)| \leq C \|h\| \|\nabla\psi\| \leq C h^{1-2/s'} \|f\| \|\nabla\psi\|_{L^{s'}} = C h^{2/s} \|f\| \|\nabla\psi\|_{L^{s'}}, \quad (4.7) \]
and
\[ \|\varepsilon(a(u_h; u_h, \psi)) \leq C \|\nabla u_h \cdot \nabla(u_h - u)\|_{L^2} + \|\nabla u_h \cdot \nabla u\|_{L^2} \|\nabla\psi\|_{L^{s'}} \]
\[ \leq C (h^{1-2/p} M \|\nabla(u_h - u)\|_{L^2} + \|u_h\|_{L^q} \|\nabla u\|_{L^q} \|\nabla\psi\|_{L^{s'}} \]
\[ \leq C (h^{1-2/p} M \|\nabla(u_h - u)\|_{L^2} + h^{1+2/t-2/p} M \|\nabla u\|_{L^q}) \|\nabla\psi\|_{L^{s'}}. \quad (4.8) \]
Further, in view of the Sobolev imbedding, cf., e.g., [1]
\[ \|v\|_{L^r} \leq C \|v\|_{W_{1,r}}, \quad \forall v \in W^1_r, \ r \leq 2, \ \text{and} \ 2 \leq r/(2 - r), \quad (4.9) \]
and
\[ 1 + \frac{2}{t} - \frac{2}{p} = 1 - \frac{2}{t} + \frac{2}{s} - \frac{2}{p} = 2 - \frac{2}{q} + \frac{2}{s} - \frac{2}{p} > 1 + \frac{2}{s} - \frac{2}{q}, \]
Corollary 4.2. Let \( u_h \) and \( u \) be the solutions of (2.3) and (1.1), respectively, with \( f \in L_2 \). Then, if \( \gamma = \alpha \beta_3 M < 1 \) there exists a constant \( C_s = C_s(u, f) \), independent of \( h \), such that for \( 0 < h \leq h_M \)
\[
\| u - u_h \|_{L_\infty} \leq C_s(u, f) h^{1+2/s-2/q}, \quad \text{with} \ 2 < s < p < 2 + \epsilon.
\]

**Proof.** In view of the Sobolev imbedding \( \|v\|_{L_\infty} \leq C_s \| \nabla v \|_{L_s} \), \( s > 2 \) and Theorem 4.1 we can easily see that (4.12) holds.

Note that the constant \( C_s \) in Corollary 4.2 blows-up as \( s \to 2 \). Later, in Theorem 4.5, we will show an almost optimal order \( L_\infty \) error estimate. Next, we will show that the finite volume solution \( u_h \) is also bounded in \( \| \nabla \cdot \|_{L_q}, 2/q + 2/\bar{q} = 1 \). This will be used later in the \( L_2 \)-norm error estimation.

**Theorem 4.3.** Let \( u_h \) and \( u \) be the solutions of (2.3) and (1.1), respectively, with \( u \in W^2_q \cap H^1_0 \), \( 4/3 < q \leq 2 \). Then \( u_h \in W^1_q \), uniformly for all \( 0 < h \leq h_M \), i.e.,
\[
\| \nabla u_h \|_{L_q} \leq C(u, f), \quad \text{with} \ 2/q + 2/\bar{q} = 1.
\]

**Proof.** We rewrite \( u_h \) by adding and subtracting \( R_h u \) and \( \Pi_h u \), where \( R_h : H^1_0 \to X_h \) is the elliptic projection operator defined by
\[
a(u; R_h u, \chi) = a(u; u, \chi), \quad \forall \chi \in X_h,
\]
and \( \Pi_h : C(\Omega) \to X_h \) the standard nodal interpolant. Thus
\[
\| \nabla u_h \|_{L_q} \leq \| \nabla (u_h - R_h u) \|_{L_q} + \| \nabla R_h u \|_{L_q}
\leq \| \nabla (u_h - R_h u) \|_{L_q} + \| \nabla (R_h u - \Pi_h u) \|_{L_q} + \| \nabla \Pi_h u \|_{L_q}.
\]

In view of the approximation property, (4.3), \( \Pi_h \) satisfies
\[
| \nabla (\Pi_h v - v) |_{L_q} \leq C h^{1+2/s-2/q} \| v \|_{W^2_q}, \quad \text{for} \ 4/3 < q \leq 2 \leq s.
\]

Then, the last term in (4.14) can easily be estimated in view of (4.9) and (4.15), we have
\[
\| \nabla \Pi_h u \|_{L_q} \leq C \| u \|_{W^2_q},
\]

Also, we can easily see that the identity
\[
a(u; R_h u - u, R_h u - u) = a(u; R_h u - u, \Pi_h u - u),
\]
gives
\[ \| \nabla (R_h u - u) \| \leq C \| \nabla (\Pi_h u - u) \|. \]

Thus, using the inverse inequality (2.17), (4.15) and the fact that \( 2 - 2/q = 1 - 2/\bar{q} \), we can bound the second term in (4.14) by
\[ \| \nabla (R_h u - \Pi_h u) \|_{L_q} \leq C h^{2/\bar{q} - 1} \| \nabla (R_h u - \Pi_h u) \| \]
\[ \leq C h^{2/\bar{q} - 1} (\| \nabla (R_h u - u) \| + \| \nabla (\Pi_h u - u) \|) \leq C \| u \|_{W_q^2} \] (4.17)

Finally, the first term, in (4.14) can be estimated similarly. From Theorem 4.1, (2.17) and the fact that \( 2 - 2/q = 1 - 2/\bar{q} \) we have
\[ \| \nabla (u_h - R_h u) \|_{L_q} \leq C h^{2/\bar{q} - 1} \| \nabla (u_h - R_h u) \| \]
\[ \leq C h^{2/\bar{q} - 1} (\| \nabla (u_h - u) \| + \| \nabla (R_h u - u) \| + \| \nabla (\Pi_h u - u) \|) \leq C \| u, f \|. \] (4.18)

Combining now this with (4.14), (4.16) and (4.17), proves the theorem. \( \square \)

For the proof of the \( L_2 \)-norm error estimate we will employ a similar duality argument as the one used in [13]. Let us consider the following auxiliary problem. Let \( \varphi \in H_0^1 \) be such that
\[ a(u; \varphi, v) + (A'(u) \nabla u \nabla \varphi, v) = (u - u_h, v), \quad \forall v \in H_0^1. \] (4.19)

If \( A(u) \) is Lipschitz continuous and \( A'(u) \nabla u \in L_\infty \), then the solution \( \varphi \) of (4.19) satisfies the following elliptic regularity estimate,
\[ \| \varphi \|_{W_{q}^2} \leq C \| u_h - u \|, \quad \text{with } 4/3 < q_0 \leq 2, \] (4.20)
where \( q_0 \) depends on the biggest interior angle of \( \Omega \) and the coefficients \( A(u), A'(u) \nabla u \). If \( \Omega \) is convex then \( q_0 = 2 \), and if it is nonconvex then \( q_0 < 2 \).

**Theorem 4.4.** Let \( u_h \) and \( u \) be the solutions of (2.3) and (1.1), respectively, with \( u \in W_q^2 \cap H_0^1 \cap W_\infty^2 \), \( 4/3 < q \leq 2 \). Then, if \( u \) and \( A' \) are Lipschitz continuous, \( A'' \in L_1(\mathbb{R}) \), \( f \in H^1 \) and \( \gamma = \beta_1^{-1}/\beta_3 M < 1 \) there exists a constant \( C \), independent of \( h \), such that for sufficiently small \( h \),
\[ \| u_h - u \| \leq C \| u, f \| h^{4 - 2/\alpha - 2/q}. \] (4.21)

**Proof.** Before we begin the proof we note the following Taylor expansions
\[ A(u_h) - A(u) = (u_h - u) \int_0^1 A'(u - t(u_h - u)) \, dt = (u_h - u) A', \]
\[ A(u_h) - A(u) - A'(u)(u_h - u) = (u_h - u)^2 \int_0^1 A''(u - t(u_h - u))(1 - t) \, dt \]
\[ \equiv (u_h - u)^2 A''. \] (4.22)

Then, in view of (4.19), we have
\[ \| u - u_h \|^2 = a(u; u - u_h, \varphi) + (A'(u)(u - u_h) \nabla u, \nabla \varphi) \]
\[ = a(u; u, \varphi) - a(u_h; u, \varphi) + ((A(u_h) - A(u)) \nabla u_h, \nabla \varphi) \]
\[ - ((A(u_h) - A(u)) \nabla u, \nabla \varphi) + ((A(u_h) - A(u)) \nabla u, \nabla \varphi) - (A'(u)(u_h - u) \nabla u, \nabla \varphi) \]
\[ = a(u; u, \varphi) - a(u_h; u, \varphi) + ((A(u_h) - A(u)) \nabla (u_h - u), \nabla \varphi) \]
\[ + ((A(u_h) - A(u) - A'(u)(u_h - u)) \nabla u, \nabla \varphi). \]
Further, using (2.3) and (4.22), the relation above gives for any $\chi \in X_h$,
\[
\|u - u_h\|^2 = a(u; u, \varphi - \chi) - a(u_h; u_h, \varphi - \chi) + \varepsilon_h(f, \chi) - \varepsilon_a(u_h; u_h, \chi)
\]
\[
+ ((u_h - u)A\nabla(u_h - u) + (u_h - u)^2 A'' \nabla u, \nabla \varphi)
\]
\[
= \{a(u_h; u - u, \varphi - \chi) + ((u_h - u)A\nabla u, \nabla (\varphi - \chi)) + \varepsilon_h(f, \chi)\}
\]
\[
- \varepsilon_a(u_h; u_h, \chi) + \{(u_h - u)A\nabla (u_h - u) + ((u_h - u)^2 A'' \nabla u, \nabla \varphi)\}
\]
\[
= I_1 + I_2 + I_3.
\]
Choosing now $\chi = \Pi_h \varphi$ in (4.23) and using (2.13) and Lemma 2.4 we get
\[
|I_1| \leq C(||\nabla (u_h - u)|| + ||\nabla u||_{L_\infty} ||u_h - u||) ||\nabla (\varphi - \Pi_h \varphi)|| + Ch^2 ||f||_{H^1} ||\nabla \Pi_h \varphi||,
\]
\[
|I_2| \leq C \{h^2 ||\nabla u_h||_{L_\infty} ||\nabla u||^2 + ||u||_{W^{2}}^2 + h||\nabla u_h \cdot \nabla (u_h - u)||_{L_q} ||\nabla \Pi_h \varphi||_{L_{q'}}.\]
Since $2 < q = 2q/(2 - q)$, (4.16), the approximation property (4.15) and the fact that $2 \geq 3 - 2/q_0$, now give
\[
|I_1| \leq C h^{2-2/q_0} (||\nabla (u_h - u)|| + ||\nabla u||_{L_\infty} ||u_h - u|| + h||f||_{H^1}) ||\varphi||_{W^{2}_q}.\]
Using then Theorem 4.1 and (4.20), we obtain
\[
|I_1| \leq C(u) h^{2-2/q_0} (||\nabla (u_h - u)|| + h||f||_{H^1} + ||u||_{L_\infty}) ||u_h - u||
\leq C(u, f) h^{2/q_0} ||u_h - u|| + C(u, f) h^{2-2/q_0} ||u_h - u||^2.
\]
Also, using the fact that $q, q_0 > 4/3$ we get $q' \leq 2q_0/(2 - q_0)$, thus in view of (4.9) and (4.15),
\[
||\nabla \Pi_h \varphi||_{L_{q'}} \leq C||\varphi||_{W^{2}_q}.
\]
Then this, the inverse inequality (2.17), the Hölder inequality (2.27), with $s = 2, t = q$ and $s = q, t = 2$, and the fact that $2q/(q - 2) \leq q$, for $q > 4/3$, give
\[
|I_2| \leq C \{h^{2-2/q} ||\nabla u_h||_{L_q} (||\nabla u_h||_{L_q} ||\nabla u_h||_{L_2/(q-2)} + ||u||_{W^2_q})
\]
\[
+ h||\nabla u_h||_{L_q} ||\nabla (u_h - u)||\} ||\nabla \Pi_h \varphi||_{L_{q'}}
\leq C \||\nabla u_h||_{L_q} \{h^{2-2/q} (||\nabla u_h||_{L_2/q} + ||u||_{W^2_q}) + h||\nabla (u_h - u)||\} ||\varphi||_{W^{2}_q}.
\]
Using, next Theorems 4.1 and 4.3 and (4.20), we obtain
\[
|I_2| \leq C(u, f) (h^{2-2/q} + h||\nabla (u_h - u)||) ||u_h - u|| \leq C(u, f) h^{3-2/q} ||u_h - u||.\]
Next, we turn to the estimation of the term $I_3$ in (4.23). For this we use the Hölder inequality (2.27) with $t = q_0$; hence
\[
|I_3| \leq C ||\nabla (u_h - u)|| ||(u - u_h)\nabla \varphi|| \leq C ||\nabla (u_h - u)|| ||u_h - u||_{L_{q_0}} ||\nabla \varphi||_{L_{q_0}}.
\]
Then the interpolation inequality, cf., e.g., [15, Appendix B],
\[
||v||_{L_{q_0}} \leq ||v||^{1/2} ||v||^{1/2}_{L_s}, \text{ with } s = 2q_0/(4 - q_0),
\]
and the Sobolev inequality (2.28) give
\[
||u_h - u||_{L_{q_0}} \leq C ||\nabla (u_h - u)||^{1/2} ||u_h - u||^{1/2}.
\]
Therefore, using this and Theorem 4.1 in (4.28) give
\[ |I_3| \leq C \| \nabla (u - u_h) \|^{3/2} \| u - u_h \|^{1/2} \| \varphi \|_{W_0^2} \leq (C \| \nabla (u - u_h) \|^3 + \frac{1}{2} \| u - u_h \|) \| u - u_h \| \]
\[ \leq C(u, f) h^{3(2-2/q')} \| u - u_h \| + \frac{1}{2} \| u - u_h \|^2. \]

We can easily see that $3(2 - 2/q) > 4 - 2/q - 2/q_0$. Therefore, combining the relation above with (4.23), (4.26) and (4.27), we get
\[ \| u - u_h \|^2 \leq |I_1| + |I_2| + |I_3| \]
\[ \leq C(u, f) h^{4-2/q-2/q_0} \| u_h - u \| + C(u, f) h^{2-2/q_0} \| u_h - u \|^2 + C(u, f) h^{3-2/q_0} \| u - u_h \| \]
\[ + C(u, f) h^{3(2-2/q')} \| u - u_h \| + \frac{1}{2} \| u - u_h \|^2, \]

which for sufficiently small $h$ gives the desired estimate. \qed

**Theorem 4.5.** Let $u_h$ and $u$ be the solutions of (2.3) and (1.1), respectively. Then, if $\Omega$ is convex, $\gamma = C_{\Omega}^{-1/2} \beta_3 \| u \|_{W_0^2} < 1$, with $C_{\Omega} > 0$ a constant depending only on $\Omega$, $u \in W_0^2$ and $f \in L_\infty$, then there exists a constant $C$ independent of $h$, such that for sufficiently small $h$,
\[ \| u - u_h \|_{L_\infty} \leq C(u, f) h^{2} \log \left( \frac{1}{h} \right). \tag{4.29} \]

**Proof.** Using again a triangle inequality we get
\[ \| u_h - u \|_{L_\infty} \leq \| w_h - u \|_{L_\infty} + \| u_h - w_h \|_{L_\infty}, \]
where $w_h$ is the Galerkin finite element approximation of $u$, i.e.,
\[ a(w_h, w_h, \chi) = (f, \chi), \quad \forall \chi \in X_h. \tag{4.30} \]

In the case of the linear problem $-\text{div}(A(x)\nabla w) = f$, we have for $A \in W_\infty^2$, cf., eg., [6]
\[ \| w_h - w \|_{L_\infty} \leq C h^2 \log \left( \frac{1}{h} \right) \| w \|_{W_\infty^2}, \]
where $w_h$ is the finite element approximation of $w$. Since $f \in L_\infty$ and $u \in W_\infty^2$, then $A(u) \in W_\infty^2$. Therefore,
\[ \| R_h u - u \|_{L_\infty} \leq C(u) h^2 \log \left( \frac{1}{h} \right). \tag{4.31} \]

The estimation of $\| u_h - R_h u \|_{L_\infty}$ was derived in [21], where it shown that
\[ \| w_h - R_h u \|_{L_\infty} \leq \gamma \| w_h - u \|_{L_\infty}, \tag{4.32} \]
with $\gamma = C_{\Omega}^{-1/2} \beta_3 \| u \|_{W_\infty^2}$. Thus (4.31) and (4.32) give
\[ (1 - \gamma) \| w_h - u \|_{L_\infty} \leq C(u) h^2 \log \left( \frac{1}{h} \right), \tag{4.33} \]

We turn now to the estimation of $\| u_h - u_h \|_{L_\infty}$. Let $x_0 \in K_0 \subset T_h$ such that $\| u_h - u_h \|_{L_\infty} = \| w_h - u_h \|_{L_\infty} = \| (u_h - u_h)(x_0) \|$ and $\delta_{x_0} = \delta \in C^\infty_0(\Omega)$ a regularized Dirac $\delta$–function satisfying
\[ \langle \delta, \chi \rangle = \chi(x_0), \quad \forall \chi \in X_h. \]
For such a function $\delta$, cf., e.g., [6], we have

$$\supp \delta \subset B = \{ x \in \Omega : |x - x_0| \leq h/2 \};$$

$$\int_\Omega \delta = 1, \quad 0 \leq \delta \leq Ch^{-2}, \quad \|\delta\|_{L_p} \leq Ch^{2(1-p)/p}, \quad 1 < p < \infty.$$  

Also let us consider the corresponding regularized Green’s function $G \in H^1_0$, defined by

$$a(u_h; G, v) = (\delta, v), \quad \forall v \in H^1_0.$$  

(4.34)

Then, we have

$$\|u_h - u_h\|_{L_\infty} = (\delta, u_h - u_h) = a(u_h; G, u_h - u_h) = a(u_h; G_h, u_h - u_h)$$

$$= (f, G_h) - a(u_h; G_h)$$

$$= \varepsilon_h(f, G_h) - a(u_h; G_h) + \{ a(u_h; G_h) - a(u_h; u_h, G_h) \},$$

where $G_h \in X_h$ is the finite element approximation of $G$, i.e.,

$$a(u_h; G, \chi) = a(u_h; G_h, \chi), \quad \forall \chi \in X_h.$$  

Since $u \in W^2_{\infty}$, we have $u \in H^2$. Thus, in view of Theorem 4.3, $\|\nabla u_h\|_{L_\infty} \leq C$. Further, using Lemma 2.4 and (2.13), (1.2) and Theorem 4.4, we obtain

$$\|u_h - u_h\|_{L_\infty} \leq C \{ h^2(\|f\|_{H^1} + \|\nabla u_h\|_{L_2} \|\nabla u_h\| + \|\nabla u_h\|_{L_\infty} \|u\|_{H^2})$$

$$+ h \|\nabla u_h\|_{L_\infty} \|\nabla(u_h - u)\| + \|(u_h - u_h)\| \|\nabla u_h\| \} \|\nabla G_h\|$$

$$\leq Ch^2(\|f\|_{H^1} + \|u\|_{H^2} + \|u_h - u\| \|\nabla G_h\|).$$

(4.36)

The last term can be estimated by, cf., e.g., [13],

$$\|u_h - u\| \leq C(u, f)h^2.$$  

(4.37)

In addition in view of [22, Lemma 3.1] we get

$$\|G_h\|_{H^1} \leq C \|\nabla G\|_{L_2} \leq C \frac{1}{(s - 1)^{1/2}} \|\delta\|_{L_2},$$

(4.38)

with $s \Downarrow 1$. Choosing now $s = 1 + (\log(1/h))^{-1}$ we have

$$\|G_h\|_{H^1} \leq C(\log(1/h))^{1/2}. $$

(4.39)

Combining now (4.35)–(4.39), we obtain

$$\|u_h - u_h\|_{L_\infty} \leq C(u, f)h^2 \log(1/h)^{1/2}.$$  

(4.40)

From this and (4.33) for $\gamma < 1$ we get the desired estimation (4.29).
In this section we shall analyze Newton’s method for the computation of the finite volume solution $u_h$ of (2.3). We consider an inexact Newton iteration, a variant of the Newton iteration for nonlinear systems of equations, where the Jacobian of the system is solved approximately, cf., e.g., [2, 3, 11]. Our analysis is based on a similar approach for the finite element method, studied by Douglas and Dupont in [13].

Also here, we will assume that (1.1) has a unique solution $u \in H^2 \cap H_0^1$. For $\phi \in H^1$ we define the bilinear form $N(\phi; \cdot, \cdot)$ on $H_0^1 \times H_0^1$ by

$$N(\phi; v, w) = a(\phi; v, w) + d(\phi; v, w), \quad (5.1)$$

where $d$ is given by

$$d(\phi; v, w) = (A'(\phi)v \nabla \phi, \nabla w). \quad (5.2)$$

Further, let $N_h$ be the corresponding finite volume form to $N$, defined for $\phi \in H^2 \cap H_0^1$ on $(H^2 \cap H_0^1) + X_h \times (H^2 \cap H_0^1) + X_h$ by

$$N_h(\phi; v, w) = a_h(\phi; v, w) + d_h(\phi; v, w), \quad (5.3)$$

where $d_h$ is given by

$$d_h(\phi; v, w) = - \sum_K \int_K \text{div}(A'(\phi)v \nabla \phi)I_hw \, dx + \int_{\partial K} (A'(\phi)v \nabla \phi) \cdot nI_hw \, ds. \quad (5.4)$$

For $u_h^0 \in X_h$, the Newton approximations to the solution $u_h$ forms a sequence $\{u_h^k\}_{k=0}^\infty$ in $X_h$ satisfying

$$N_h(u_h^k; u_h^{k+1} - u_h^k, \chi) = (f, I_h \chi) - a_h(u_h^k; u_h^k, I_h \chi), \quad \forall \chi \in X_h. \quad (5.5)$$

We will show that $u_h^k \to u_h$ in $H^1$-norm as $k \to \infty$, with order two, provided that $u_h^0$ is sufficiently close to $u_h$. For this we will assume that $u_h$ converges to $u$ sufficiently fast,

$$\|u - u_h\|_{L^\infty} + \sigma_h \|u - u_h\|_{H^1} \to 0, \quad \text{as } h \to 0, \quad (5.6)$$

where

$$\sigma_h \equiv \sup\{\|\chi\|_{L^\infty}/\|\chi\|_{H^1} : 0 \neq \chi \in X_h\}. \quad (5.7)$$

Since $T_h$ is a quasi-uniform mesh, there exists a constant $C$, independent of $h$ such that

$$|\sigma_h| \leq C \log(\frac{1}{h}). \quad (5.8)$$

Further, let $C_3$ be another constant, independent of $h$, satisfying

$$\|u_h\|_{W^{1,\infty}} \leq C_3. \quad (5.9)$$

Note that this assumption holds, for $u \in H^2$, cf. Section 3. In addition we assume that $A''$ is bounded and is Lipschitz continuous, i.e.,

$$|A''(x)| \leq \beta_k, \quad |A''(x) - A''(y)| \leq L_2 |x - y|, \quad \forall x, y \in \mathbb{R}. \quad (5.10)$$

Next, we will show various auxiliary results that helps in the proof of Theorem 5.1. We start by stating the following lemma of Douglas and Dupont, [13].
Lemma 5.1. Given $\tau > 0$, there exists positive constants $\delta$, $h_0$ and $C_4$ such that the following holds. If $0 < h < h_0$, if $\phi \in W_h^1 \cap L_\infty$ satisfies
\[
\|\phi\|_{W_h^1} \leq \tau \quad \text{and} \quad \sigma_h\|\phi - u\|_{H^1} \leq \delta,
\]
and if $G$ is a linear functional on $H^1_0$ with
\[
\|G\| = \sup_{0 \neq \chi \in X_h} \frac{|G(\chi)|}{\|\chi\|_{H^1}},
\]
then there exists a unique $v \in X_h$ satisfying the equations
\[
N(\phi; v, \chi) = G(\chi), \quad w \in X_h.
\]
Furthermore, $v$ satisfies the bound
\[
\|v\|_{H^1} \leq C_4\|G\|.
\]

Lemma 5.2. For $\phi \in X_h$ the error functional $\varepsilon_N$ satisfies
\[
|\varepsilon_N(\phi; \psi, \chi)| \leq Ch\|\nabla \phi\|_{L_\infty}(1 + \sigma_h\|\phi\|_{H^1})\|\psi\|_{H^1}\|\chi\|_{H^1}, \quad \forall \chi, \psi \in X_h.
\]

Proof. From the definition of $\varepsilon_N$ we can easily see that, $\varepsilon_N = \varepsilon_a + (d - d_h)$. Therefore in view of Lemma 2.3, it suffices to bound $d - d_h$. Following the proof of Lemma 2.3 we have,
\[
d(\phi; \psi, \chi) - d_h(\phi; \psi, \chi)
= \sum_K \{\text{div}((A'(\phi)\psi)\nabla \phi), \chi - I_h\chi) + (A'(\phi)\psi)\cdot n, \chi - I_h\chi)\}
\]
\[
= \sum_K \{I_K + II_K\}. \tag{5.13}
\]

Applying Hölder’s inequality to $I_K$, and using the fact that $\phi$ is linear in $K$, (1.2), (5.10) and (2.11), we have
\[
|I_K| \leq (\beta_3\|\nabla \phi \cdot \nabla \psi\|_{L_2(K)} + \beta_4\|\nabla \phi^2 \psi\|_{L_2(K)})\|\chi - I_h\chi\|_{L_2(K)}
\leq Ch(\beta_3\|\nabla \phi \cdot \nabla \psi\|_{L_2(K)} + \beta_4\|\nabla \phi^2 \psi\|_{L_2(K)})\|\nabla \chi\|_{L_2(K)}.
\]

For the $II_K$, we break the integration over the boundary of each triangle $K$, into the sum of integrations over its sides, and thus may use (2.12), and follow the same steps as in estimating $I_K$. Hence,
\[
|II_K| \leq Ch\|\nabla (A'(\phi)\psi)\nabla \phi\|_{H^1(K)}\|\nabla \chi\|_{L_2(K)}
\leq Ch(\beta_3\|\nabla \phi \cdot \nabla \psi\|_{L_2(K)} + \beta_4\|\nabla \phi^2 \psi\|_{L_2(K)})\|\nabla \chi\|_{L_2(K)}.
\]

Then combining this with Lemma 2.3 and (5.14), we get
\[
|\varepsilon_N(\phi; \psi, \chi)| \leq Ch(\|\nabla \phi\|_{L_\infty}\|\nabla \psi\| + \|\nabla \phi\|_{L_\infty}\|\psi\|_{L_\infty}\|\nabla \phi\|)\|\chi\|_{H^1}.
\]

Finally, in view of the definition of $\sigma_h$ we get the desired estimate. \hfill \Box

Next, we derive a “Lipschitz”–type estimation for $\varepsilon_N$. 

Copyright © 2004 John Wiley & Sons, Ltd. Numer. Linear Algebra Appl. 2004; 00:1–26
Prepared using nlaauth.cls
Lemma 5.3. Let \( v, w, \phi, \chi \in X_h \) then

\[
|\varepsilon_N(v; \phi, \chi) - \varepsilon_N(w; \phi, \chi)| \leq C h \left\{ |\nabla (v - w) \cdot \nabla \phi| + \|\nabla w\|_{L_\infty} |(v - w) \nabla \phi| + \|\nabla w\|_{L_\infty}^2 |(v - w) \nabla \phi| \right\} |\nabla \chi|.
\] (5.15)

Proof. Similarly as in the proof of the previous lemma, we can easily see that \( \varepsilon_N = \varepsilon_a + (d - d_h) \). Thus in view of Lemma 2.5, it suffices to estimate \( d(v; \phi, \chi) - d_h(w; \phi, \chi) \).

Using a similar decomposition as in (5.13) and then applying (2.11) and (2.12) we get

\[
|d(v; \phi, \chi) - d_h(w; \phi, \chi)| \leq C h \left\{ |\nabla (A'(v) \nabla v - A'(w) \nabla w) \phi| + |(A'(v) \nabla v - A'(w) \nabla w) \phi|_{H^1} \right\} |\nabla \chi|.
\] (5.16)

Next, since \( \phi \in X_h \), we have

\[
\begin{align*}
& \quad \text{div}(\nabla (A'(v) \nabla v - A'(w) \nabla w) \phi) \\
&= (A''(v)|\nabla v|^2 - A''(w)|\nabla w|^2) \phi + (A'(v) \nabla v - A'(w) \nabla w) \cdot \nabla \phi \\
&= (A''(v)|\nabla v|^2 - |\nabla w|^2) \phi + (A''(v) - A''(w))|\nabla w|^2 \phi \\
&\quad + (A'(v) \nabla v - A'(w) \nabla w) \cdot \nabla \phi + (A'(v) - A'(w)) \nabla w \cdot \nabla \phi.
\end{align*}
\] (5.17)

Therefore, (5.16) gives

\[
|d(v; \phi, \chi) - d_h(w; \phi, \chi)| \leq C h (|\nabla (v - w) \cdot \nabla \phi| + |\nabla w|_{L_\infty} |(v - w) \nabla \phi|) |\nabla \chi| \\
+ C h (|\nabla w|_{L_\infty}^2 + |\nabla w|_{L_\infty}^2 |(v - w) \nabla \phi|) |\nabla \chi|.
\] (5.18)

Finally, this estimation and Lemma 2.5 give the desired (5.15). \( \square \)

Next, we show an error bound that we will employ in the proof of Theorem 5.1.

Lemma 5.4. For \( v_h, w_h, \chi \in X_h \), we have

\[
|\varepsilon_N(v_h; w_h - v_h, \chi) + \varepsilon_a(v_h; v_h, \chi) - \varepsilon_a(w_h; w_h, \chi)| \\
\leq C h \left( \sigma_h(\|\nabla v_h\|_{L_\infty}^2 + \|\nabla (w_h + v_h)\|_{L_\infty}) + h^{-1} \right) \|w_h - v_h\|_{H^1} \|\chi\|_{H^1}.
\] (5.19)

Proof. In view of the definition of \( \varepsilon_N \) and \( \varepsilon_a \) we have

\[
\begin{align*}
\varepsilon_N(v_h; w_h - v_h, \chi) + \varepsilon_a(v_h; v_h, \chi) - \varepsilon_a(w_h; w_h, \chi) \\
= \sum_K \int_K \text{div} \left( A(v_h) \nabla (w_h - v_h) + A'(v_h)(w_h - v_h) \nabla v_h + A(v_h) \nabla v_h \\
- A(w_h) \nabla w_h \right) \chi - I_h \chi \right) dx \\
+ \sum_K \int_{\partial K} \left( A(v_h) \nabla (w_h - v_h) + A'(v_h)(w_h - v_h) \nabla v_h + A(v_h) \nabla v_h \\
- A(w_h) \nabla w_h \right) \cdot n(\chi - I_h \chi) ds.
\end{align*}
\]

Then, since \( v_h, w_h \) are linear in \( K \in T_h \), we get

\[
\begin{align*}
\text{div} \left( A(v_h) \nabla (w_h - v_h) + A'(v_h)(w_h - v_h) \nabla v_h + A(v_h) \nabla v_h - A(w_h) \nabla w_h \right) \\
= 2 A'(v_h) \nabla v_h \cdot (w_h - v_h) + A''(v_h)(w_h - v_h) |\nabla v_h|^2 + A'(v_h) |\nabla v_h|^2 - A'(w_h) |\nabla w_h|^2 \\
= A''(v_h)(w_h - v_h) |\nabla v_h|^2 + A'(v_h) |\nabla v_h|^2 - A'(w_h) |\nabla w_h|^2 \\
&\quad + A'(w_h) |\nabla w_h|^2 - A'(w_h) |\nabla w_h|^2 + 2 A'(v_h) \nabla v_h \cdot (w_h - v_h).
\end{align*}
\]
Theorem 5.1. There exists positive constants $h_0$, $\delta$ and $C_5$ such that if $0 < h \leq h_0$ and $\sigma_h ||u_h^k - u_h||_{H^1} \leq \delta$ then for $k \geq 0$, $v_k = ||u_h^k - u_h||_{H^1}$ is a decreasing sequence satisfying

$$v_{k+1} \leq C_5 \sigma_h v_k^2.$$ \hspace{1cm} (5.20)

Proof. The proof is based on a similar result of Douglas and Dupont, [13], for the finite element method. First we show that for $h_0$ and $\delta$ are sufficiently small, and $\sigma_h ||u_h^k - u_h||_{H^1} = \sigma_h v_k \leq \delta$, with $0 < h \leq h_0$, there exists a unique $u_h^{k+1}$, given by (5.5). It suffices to show that if

$$N_h(u_h^k; v, \chi) = 0, \quad \forall \chi \in X_h,$$

then $v \equiv 0$, or else $||v||_{H^1} \leq 0$. For this we will employ Lemma 5.1 and demonstrate that $C_4 ||G|| < ||v||_{H^1}$, for an appropriately defined functional $G$. We can easily see that

$$N(u_h; v, \chi) = G(\chi),$$

where $G$ is given by

$$G(\chi) = N(u_h; v, \chi) - N_h(u_h^k; v, \chi) = \{N(u_h; v, \chi) - N(u_h^k; v, \chi)\} + \varepsilon_N(u_h^k; v, \chi) = I + II,$$

Following the proof in [13] we have that that

$$|I| \leq C h \sigma_h ||u_h - u_h^k||_{H^1} ||v||_{H^1} ||\chi||_{H^1} = C \sigma_h v_k ||v||_{H^1} ||\chi||_{H^1}.$$ \hspace{1cm} (5.21)

For the estimation of $II$ we use the inverse inequality, (2.17), (5.9), Lemma 5.2 and the fact that induction hypothesis and (5.6) give

$$||u_h^k||_{H^1} \leq v_k + ||u_h||_{H^1} \leq \sigma_h^{-1} \delta + ||u_h||_{H^1} \leq C,$$ \hspace{1cm} (5.22)
to get
$$|II| \leq C(\nu_k + \sigma_h ||u_h^k||_{H^1}) + h(1 + \sigma_h ||u_h^k||_{H^1}) ||u_h||_{W^1_2}) ||v||_{H^1} ||\chi||_{H^1} \leq C\sigma_h \nu_k ||v||_{H^1} ||\chi||_{H^1} + Ch \sigma_h ||v||_{H^1} ||\chi||_{H^1}. \quad (5.23)$$

Hence, since $\sigma_h \leq C \log(1/h)$, (5.21) and (5.23) give for $\delta$ and $h$ sufficiently small, $||v||_{H^1} \leq C_0 \sigma_h (\nu_k + h \log(1/h)) ||v||_{H^1} < ||v||_{H^1}$; thus $v = 0$.

In order to show (5.20) we will employ again Lemma 5.1 for a different functional $G$. This time let
$$N(u_h; u_h^{k+1} - u_h, \chi) = G(\chi), \quad \forall \chi \in X_h,$$

where $G$ is defined by
$$G(\chi) = N(u_h; u_h^k - u_h, \chi) + N(u_h; u_h^{k+1} - u_h^k, \chi) + N(u_h; u_h^{k+1} - u_h^k, \chi) - N(u_h^k; u_h^k, \chi)$$
$$+ \{N(u_h^k; u_h^k - u_h, \chi) + a(u_h^k; u_h, \chi) - a(u_h^k; u_h^k, \chi)\}$$
$$+ \{\varepsilon_N(u_h^k; u_h^{k+1} - u_h^k, \chi) - \varepsilon_a(u_h^k; u_h, \chi) + \varepsilon_a(u_h^k; u_h^k, \chi)\}$$
$$+ \{N(u_h^k; u_h^{k+1} - u_h^k, \chi) - N(u_h^k; u_h^k - u_h^k, \chi)\} = I + II + III. \quad (5.24)$$

We will show that
$$|||G||| \leq C\sigma_h \nu_k (\nu_k + \nu_{k+1}) + Ch \sigma_h \nu_{k+1}. \quad (5.25)$$

Then Lemma 5.1, and $\sigma_h \nu_k \leq \delta$, give
$$\nu_{k+1} \leq C_4 |||G||| \leq C\sigma_h \nu_k (\nu_k + \nu_{k+1}) + Ch \sigma_h \nu_{k+1}$$
$$\leq C\sigma_h \nu_k^2 + C(\delta + h \log(1/h)) \nu_{k+1}. \quad (5.26)$$

Finally for sufficiently small $\delta$ and $h$, the desired estimate, (5.20), follows easily.

Let us turn now to the estimation of $|||G|||$, for $G$ given by (5.24). The terms $I$ and $III$ are similar to the ones that appear in the analysis of the finite element method in [13], thus using the same arguments we get
$$|I + III| \leq C\sigma_h \nu_k (\nu_k + \nu_{k+1}) ||\chi||_{H^1}. \quad (5.27)$$

Then, we can easily see that $II$ can be rewritten in the following way,
$$II = \varepsilon_N(u_h^k; u_h^{k+1} - u_h^k, \chi) - \varepsilon_N(u_h^k; u_h^{k+1} - u_h^k, \chi)$$
$$+ \varepsilon_N(u_h^k; u_h^{k+1} - u_h^k, \chi) - \varepsilon_a(u_h; u_h, \chi) + \varepsilon_a(u_h^k; u_h^k, \chi)$$
$$= \{\varepsilon_N(u_h^k; u_h^{k+1} - u_h^k, \chi) - \varepsilon_N(u_h^k; u_h^{k+1} - u_h^k, \chi)\}$$
$$+ \{\varepsilon_N(u_h^k; u_h^{k+1} - u_h^k, \chi) - \varepsilon_a(u_h; u_h, \chi) - \varepsilon_a(u_h^k; u_h^k, \chi)\} = II_1 + II_2 + II_3. \quad (5.28)$$

Using Lemma 5.3, (5.9), inverse inequality, (2.17), (5.6) and (5.22), we can bound $II_1$ in the
Further, using Lemma 5.2, (5.9) and (5.6), we can easily bound $I_1$,
\[
|I_1| \leq Ch \left\{ \left( \| \nabla (u_h^k - u_h) \|_{L_\infty} + \| \nabla u_h \|_{L_\infty} \| u_h^k - u_h \|_{L_\infty} \right) \| \nabla (u_h^{k+1} - u_h^k) \| \\
+ \left( \| \nabla (u_h^k + u_h) \|_{L_\infty} \| \nabla (u_h^k - u_h) \| \\
+ \| \nabla u_h \|_{L_\infty}^2 \| u_h^k - u_h \|_2 \| u_h^{k+1} - u_h^k \|_{L_\infty} \right) \| \chi \|_{H^1} \right\}. \tag{5.29}
\]

Finally using, Lemma 5.4 and the fact that $\| \nabla u_h \|_{L_\infty}$ can be estimated by
\[
C(1 + (\| u_h^k + u_h \|_{H^1} + \nu) \sigma_k (\nu_k + \nu_{k+1})) \| \chi \|_{H^1} \leq C \sigma \nu_k (\nu_k + \nu_{k+1}) \| \chi \|_{H^1}.
\]

Further, using Lemma 5.2, (5.9) and (5.6), we can easily bound $I_2$,
\[
|I_2| \leq Ch \left\{ \left( \| \nabla u_h \|_{L_\infty} + \sigma_k \| \nabla u_h \|_{L_\infty} \| u_h^k \|_{H^1} \right) \| u_h^{k+1} - u_h \|_{H^1} \| \chi \|_{H^1} \right\}
\leq Ch(1 + \sigma_k) \nu_{k+1} \| \chi \|_{H^1} \tag{5.30}.
\]

Finally using, Lemma 5.4 and the fact that $\| \nabla u_h \|_{L_\infty} \leq C_3$ and $h \| \nabla u_h^k \|_{L_\infty} \leq C_4 \| u_h^k \|_{H^1} \leq C$, $I_3$ can be estimated by
\[
|I_3| \leq Ch \left\{ h \| \nabla u_h \|_{L_\infty}^2 + h \sigma_k \| \nabla (u_h^k + u_h) \|_{L_\infty} + 1 \right\} \| u_h^k - u_h \|_{H^1} \| \chi \|_{H^1}
\leq C \sigma_k (1 + \nu_k) \| \chi \|_{H^1}. \tag{5.31}
\]

Therefore combining (5.27) and (5.29)–(5.31), we get the desired (5.25). \qed

6. NUMERICAL IMPLEMENTATIONS

In this section we present procedures for implementing the finite volume method for the nonlinear problem. A series of numerical examples is given to further assess the theories that were previously deduced. Following the previous mathematical works, we implement two iterative schemes to solve the nonlinear finite volume problems, namely the fixed point iteration and the Newton iteration. As will be clear in the following subsection, these two schemes are built in the finite dimensional setting, i.e., using the finite element space $X_h$. We denote $\{ \phi_i \}_{i=1}^d$ to be the standard piecewise linear basis functions of $X_h$. Then the finite volume element solution may be written as
\[
u_h = \sum_{i=1}^d \alpha_i \phi_i \quad \text{for some} \quad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)^T
\]

6.1. Fixed Point Iteration vs Newton Iteration

To describe the schemes, we begin with several notations, noting that some of them have already been mentioned. Let $Z_h$ be the collection of vertices $z_i$ that belong to all triangles $K \in T_h$ and $Z_h^0 = \{ z_i \in Z_h : z_i \notin \Gamma_D \}$. Let $I = \{ i : z_i \in Z_h^0 \}$, $I_K = \{ m : z_m \text{ is a vertex of } K \}$, $T_{h,i} = \{ K \in T_h : i \in I_K \}$, and $I_i = \{ m \in I : z_m \text{ is a vertex of } K \in T_{h,i} \}$. Let $V_i$ be the control volume surrounding the vertex $z_i$.

Now we may write this finite volume problem as to find $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)^T$ such that
\[
F(\alpha) = 0, \tag{6.1}
\]
where $F : \mathbb{R}^d \to \mathbb{R}^d$ is a nonlinear operator with

$$F_i(\alpha) = -\int_{\partial V_i} A(u_h) \nabla u_h \cdot n \, ds - \int_{V_i} f \, dx \quad \forall i \in I. \quad (6.2)$$

The fixed point iteration is derived from the linearization of (6.1) on the coefficient $A(u)$ in (6.2). Thus, given an initial iterate $\alpha^0$ (i.e., equivalently $u^0_h = \sum_{i=0}^d \alpha_i^0 \phi_i$), for $k = 0, 1, 2, \ldots$ until convergence solve the linear system $M(\alpha^k) \alpha^{k+1} = q$, where $M(\alpha^k)$ is the resulting stiffness matrix evaluated at $u^k_h = \sum_{i=0}^d \alpha_i^k \phi_i$, whose entries are

$$M_{ij}^k = -\int_{\partial V_i} A(u_h^k) \nabla \phi_j \cdot n \, ds.$$

On the other hand, the classical Newton iteration relies on the first order Taylor expansion of $F(\alpha)$. It results in solving a linear system $F(\alpha^k) \alpha^{k+1} = q$ (6.1), where $F$ is a variation of Newton iteration for nonlinear system of equations in that the system Jacobian is only solved approximately, cf. e.g., [2, 3, 11]. To be specific, given an initial iterate $\alpha^0$, for $k = 0, 1, 2, \ldots$ until convergence do the following:

(a) Solve $F'(\alpha^k) \delta^k = -F(\alpha^k)$ until $\|F(\alpha^k) + F'(\alpha^k)\delta^k\| \leq \beta_k \|F(\alpha^k)\|);

(b) Update $\alpha^{k+1} = \alpha^k + \delta^k$.

In this algorithm $F'(\alpha^k)$ is the Jacobian matrix evaluated at iteration $k$. For iterative technique solving a linear system such as the Krylov method we only need the action of the Jacobian to a vector. It has been common practice to use the following finite difference approximation for such an action:

$$F'(\alpha^k) v \approx \frac{F(\alpha^k + \sigma v) - F(\alpha^k)}{\sigma}, \quad (6.3)$$

where $\sigma$ is a small number computed as follows:

$$\sigma = \text{sign}(\alpha^k \cdot v) \sqrt{\varepsilon \max_{v \neq 0} (|\alpha^k \cdot v|, \|v\|_1)},$$

with $\varepsilon$ being the machine unit round-off number. We note that when $\beta_k = 0$ then we have recovered the classical Newton iteration. One common used relation is

$$\beta_k = 0.001 \left( \frac{\|F(\alpha^k)\|}{\|F(\alpha^{k-1})\|} \right)^2,$$

with $\beta_0 = 0.001$. Choosing $\beta_k$ this way we avoid oversolving the Jacobian system when $\alpha^k$ is still considerably far from the exact solution.

Instead of using (6.3), we will present below an explicit construction of the Jacobian matrix. We note that we may decompose $F_i(\alpha)$ as follows:

$$F_i(\alpha) = \sum_{K \in T_h,i} F_{i,K}(\alpha), \quad \text{where} \quad F_{i,K}(\alpha) = -\int_{K \cap \partial V_i} A(u_h) \nabla u_h \cdot n \, ds - \int_{K \cap V_i} f \, dx.$$

From the above description it is apparent that $F_i(\alpha)$ is not fully dependent on all $\alpha_1, \alpha_2, \ldots, \alpha_d$. Consequently, $\frac{\partial F_i(\alpha)}{\partial \alpha_j} = 0$ for $j \notin I_i$. Next we want to find an explicit form of $\frac{\partial F_i(\alpha)}{\partial \alpha_j}$ for $j \in I_i$. 

Copyright © 2004 John Wiley & Sons, Ltd. Numer. Linear Algebra Appl. 2004; 00:1–26

Prepared using nlaauth.cls
Now suppose the edge \( e^j_{i2} \) is shared by triangles \( K_i, K_r \in T_{h,i} \). Then
\[
\frac{\partial F_i}{\partial \alpha_j} = - \sum_{e \in \partial_r} \int_{K_{i} \cap \partial e_i} (A'(u_h) \phi_j \nabla u_h \cdot n + A(u_h) \nabla \phi_j \cdot n) \, ds.
\]
Furthermore,
\[
\frac{\partial F_i}{\partial \alpha_i} = - \sum_{K \in T_{h,i}} \int_{K \cap \partial V_i} (A'(u_h) \phi_i \nabla u_h \cdot n + A(u_h) \nabla \phi_i \cdot n) \, ds.
\]
From this derivation it is obvious that the Jacobian matrix is not symmetric but sparse. Computation of this Jacobian matrix is similar to computing the stiffness matrix resulting from standard finite volume element, in that each entry is formed by accumulation of element by element contribution. Once we have the matrix stored in memory, then its action to a vector is straightforward. Since it is a sparse matrix, devoting some amount of memory for entries storage is not very expensive.

6.2. Numerical Examples

In this subsection we present several numerical experiments to verify the theoretical investigations. We solve a set of Dirichlet boundary value problems in \( \Omega = [0, 1] \times [0, 1] \). We compare the fixed point iteration and the Newton iteration. In both schemes, the iteration is stopped once \( \|u_h^k - u_h^{k-1}\|_{L_\infty} < 10^{-10} \). In all examples below, the initial iteration is taken to be \( \alpha = (0, 0, \ldots, 0)^T \).

The first example is solving \(-\nabla \cdot (k(u) \nabla u) = f \) in \( \Omega \) where the function \( f \) is chosen such that the known solution is \( u(x, y) = (x - x^2)(y - y^2) \). The nonlinearity comes from the coefficient with \( k(u) = \frac{1}{(1+u)^2} \). The results are listed in Table I. First column represents the mesh size. The domain is discretized into \( N \) numbers of rectangle in each direction. Each of these rectangle is divided into two triangles. Second and third columns correspond to the number of iterations performed until the stopping criteria is reached for fixed point iteration (FP) and Newton iteration (NW), respectively. The table shows that a superconvergence is observed in \( H^1 \)-norm due to the smoothness of the solution. Number of iterations in both schemes do not depend on the the mesh size. The numerical results for the second example are presented in Table II. Here

<table>
<thead>
<tr>
<th>( h )</th>
<th># iter</th>
<th>( H^1 )-seminorm</th>
<th>( L_2 )-norm</th>
<th>( L_\infty )-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 1/16 )</td>
<td>7</td>
<td>17.1931</td>
<td>3.73555</td>
<td>7.51200</td>
</tr>
<tr>
<td>( 1/32 )</td>
<td>7</td>
<td>4.31635</td>
<td>0.94094</td>
<td>1.88100</td>
</tr>
<tr>
<td>( 1/64 )</td>
<td>7</td>
<td>1.08075</td>
<td>0.23568</td>
<td>0.47000</td>
</tr>
<tr>
<td>( 1/128 )</td>
<td>7</td>
<td>0.27778</td>
<td>0.05894</td>
<td>0.01180</td>
</tr>
</tbody>
</table>

Table I. Error of FVEM for nonlinear elliptic BVP, with \( u = (x - x^2)(y - y^2) \) and \( k(u) = 1/(1+u)^2 \)

the exact solution is chosen to be \( u = 40(x - x^2)(y - y^2) \) and \( k(u) = 0.125(-u^2 + 4u^2 - 7u + 8) \) if \( u < 1 \) and \( k(u) = 1/(1+u) \) if \( u \geq 1 \). Again a superconvergence is observed for this example.

Furthermore, number of iterations needed are slightly higher than the previous example, which
may be due to larger source term \( f \). In this case the Newton iteration is shown to converge faster than the fixed point iteration. Next we consider a problem with known solution \( u(x, y) = x^{1.6} \)

Table II. Error of FVEM for nonlinear elliptic BVP, with \( u = 40(x - x^2)(y - y^2) \) and \( k(u) = 0.125(-u^3 + 4u^2 - 7u + 8) \) if \( u < 1 \) and \( k(u) = 1/(1 + u) \) if \( u \geq 1 \)

<table>
<thead>
<tr>
<th>N</th>
<th># iter</th>
<th>( H^1 )-seminorm</th>
<th>( L_2 )-norm</th>
<th>( L_\infty )-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Error ( \times 10^{-2} ) Rate</td>
<td>Error ( \times 10^{-2} ) Rate</td>
<td>Error ( \times 10^{-2} ) Rate</td>
</tr>
<tr>
<td>1/16</td>
<td>16</td>
<td>33.65484            -</td>
<td>7.33022        -</td>
<td>13.3000        -</td>
</tr>
<tr>
<td>1/32</td>
<td>15</td>
<td>9.10047             1.89</td>
<td>1.98347        1.89</td>
<td>3.57150        1.90</td>
</tr>
<tr>
<td>1/64</td>
<td>15</td>
<td>2.32645             1.97</td>
<td>0.50708        1.97</td>
<td>0.91120        1.97</td>
</tr>
<tr>
<td>1/128</td>
<td>15</td>
<td>0.58451             1.99</td>
<td>0.12740        1.99</td>
<td>0.22880        1.99</td>
</tr>
</tbody>
</table>

with \( k(u) = 1 + u \). Obviously, this solution is an element of \( H^2(\Omega) \) but not in \( H^3(\Omega) \). Also the resulting source term \( f \) only belongs to \( L^2(\Omega) \). The results are presented in Table III. These experiments show that the \( H^1 \)-norm of the error decreases at first order. The \( L_2 \)-norm of the error decreases slower than second order. Again, this case shows that the Newton iteration is relatively faster than the fixed point iteration.

Table III. Error of FVEM for nonlinear elliptic BVP with \( u(x, y) = x^{1.6} \) and \( k(u) = 1 + u \)

<table>
<thead>
<tr>
<th>N</th>
<th># iter</th>
<th>( H^1 )-seminorm</th>
<th>( L_2 )-norm</th>
<th>( L_\infty )-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Error ( \times 10^{-4} ) Rate</td>
<td>Error ( \times 10^{-4} ) Rate</td>
<td>Error ( \times 10^{-4} ) Rate</td>
</tr>
<tr>
<td>1/16</td>
<td>11</td>
<td>34.1671             -</td>
<td>3.71216        -</td>
<td>8.97536        -</td>
</tr>
<tr>
<td>1/32</td>
<td>11</td>
<td>17.5558             0.96</td>
<td>1.44873        1.36</td>
<td>3.53674        1.34</td>
</tr>
<tr>
<td>1/64</td>
<td>11</td>
<td>8.68644             1.02</td>
<td>0.53714        1.43</td>
<td>1.33414        1.40</td>
</tr>
<tr>
<td>1/128</td>
<td>11</td>
<td>4.20084             1.05</td>
<td>0.19272        1.48</td>
<td>0.48582        1.46</td>
</tr>
</tbody>
</table>

Tables IV and V illustrate Theorem 5.1. In this theorem, it has been shown that there exists a sequence of solutions in the Newton iteration such that their errors with respect to the finite volume solution \( u_h \) are a decreasing sequence. Using the notation in that theorem, \( \nu_k = \| u_k^h - u_h \|_{H^1} \) is a decreasing sequence satisfying

\[ \nu_{k+1} \leq C_5 \sigma_h \nu_k^2, \quad k = 0, 1, 2, \cdots. \]

We would like to examine the numerical behavior of this sequence for a fixed mesh size \( h \). It is obvious that given \( \nu_0 \) we have

\[ \nu_k \leq (C_5 \sigma_h)^{2^k-1} \nu_0^{2^k}, \quad k = 1, 2, \cdots, \]

which after dividing by \( \nu_0^{2^k} \) and taking logarithm on both sides give

\[ |\log(\nu_k/\nu_0^{2^k})| \leq C_5 \sigma_h (2^k - 1), \quad k = 1, 2, \cdots. \]

Hence we should expect that the sequence \( \nu_k \) would decrease exponentially as \( k \to \infty \).
Table IV. Results for case 2

<table>
<thead>
<tr>
<th>k</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>$</td>
<td>\log(v_k/v_0^o)</td>
<td>$</td>
</tr>
<tr>
<td>1</td>
<td>1.13</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>2</td>
<td>3.40</td>
<td>3.40</td>
<td>3.40</td>
</tr>
<tr>
<td>3</td>
<td>7.97</td>
<td>7.96</td>
<td>7.96</td>
</tr>
<tr>
<td>4</td>
<td>16.8</td>
<td>16.8</td>
<td>16.6</td>
</tr>
</tbody>
</table>

The Tables IV and V show the decreasing behavior of the sequence resulting from the Newton iteration for last two model problems described above. In each table, $k$ represents the iteration level, $h$ is the mesh size, and $m$ is the value of row $k$ divided by the value of row $k - 1$.

For case 2 presented in Table IV, in which the problem has a piecewise continuous coefficient and larger source term, we see that the decreasing behavior of the sequence is approximately exponential, and it is independent of the mesh size. Similar trends are also evident for case 3 shown in Table V.

Table V. Results for case 3

<table>
<thead>
<tr>
<th>k</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>$</td>
<td>\log(v_k/v_0^o)</td>
<td>$</td>
</tr>
<tr>
<td>1</td>
<td>1.17</td>
<td>1.32</td>
<td>1.45</td>
</tr>
<tr>
<td>2</td>
<td>3.57</td>
<td>3.86</td>
<td>4.19</td>
</tr>
<tr>
<td>3</td>
<td>8.04</td>
<td>8.72</td>
<td>9.26</td>
</tr>
<tr>
<td>4</td>
<td>16.8</td>
<td>18.2</td>
<td>19.6</td>
</tr>
<tr>
<td>5</td>
<td>32.7</td>
<td>36.9</td>
<td>40.1</td>
</tr>
</tbody>
</table>

REFERENCES