J. Fac. Sci. Univ. Tokyo
Sect. IA, Math.
35 (1988), 77-101.

Finite element analysis of a nondifferentiable nonlinear problem related to MHD equilibria

By Fumio KIKUCHI

Summary. We consider finite element approximation of a nondifferentiable nonlinear eigenvalue problem related to MHD (**m**agneto**h**ydro**d**ynamics) equilibria. In an abstract setting, we first present two Newton-like iteration schemes as generalizations of the usual Newton method and the modified Newton method, and consider their convergence properties with an implicit function theorem derived. Then we apply the results to a nonlinear eigenvalue problem described by a semilinear elliptic problem with a nondifferentiable nonlinear term. Finally, we introduce a simple finite element model to this problem, to which we show that the iteration schemes are applicable. Order estimates of the errors of the finite element solutions are also given under some assumptions. A few numerical results are illustrated to see the validity of the analysis.

Key words. Newton-like methods, finite element method, MHD equilibria, nondifferentiable nonlinear problem

Contents

1. Introduction

Let Ω be a bounded domain in $\boldsymbol{R}^n$ $(n=1, 2, 3)$ with a boundary $\partial\Omega$. The independent variable of $\boldsymbol{R}^n$ is denoted by $\xi=\{\xi_1, \cdots, \xi_n\}$. In the

equilibrium analysis of confined MHD (**magneto**hydro**dynamics**) plasmas or thin stretched membranes partially covered with water [12, 15], we have the following semilinear elliptic problem for a real parameter λ and a real function $u=u(\xi)$:

$$(1) \qquad -\Delta u=\lambda f(u) \quad \text{in } \Omega, \qquad u=0 \quad \text{on } \partial\Omega,$$

where Δ is the Laplace operator, and $f(u)=(u-1)^+$, that is,

$$(2) \qquad f(u)(\xi)=\max\{0,\ u(\xi)-1\} \quad \text{for} \quad \xi\in\Omega.$$

Clearly, $f(u)$ is nonlinear in u. The function $u(\xi)\equiv 0$ satisfies (1) for any λ but we will seek other non-trivial pairs $\{\lambda, u\}$, and hence the present problem may be considered a nonlinear eigenvalue problem.

Analysis of this type of problems is drawing much attention due to the importance in various fields of mathematical physics, see e.g. Berestycki-Fernandez Cara-Glowinski [2], Kikuchi-Nakazato-Ushijima [12], and Temam [15]. An essential feature of the above problem is the non-smoothness of the nonlinear term $f(u)$. Due to this term, the Newton method and its variants in their original forms are unavailable to solve (1), although they are standard computational techniques for nonlinear equations with Fréchet differentiable operators [13]. At the same time, such methods appear to become available if suitable modifications are introduced, since the present model problem is in general nondifferentiable but is in a sense "almost differentiable".

In this paper, we will perform numerical analysis of the above problem with a slightly more general nonlinear term. In its process, we will present iteration methods, which were already applied to finite element analysis of MHD equilibria by Kikuchi-Aizawa [10, 11] with satisfactory results but without mathematical justifications.

We first consider an abstract problem, to which we propose two iteration schemes with a theorem on their convergence under some assumptions (Section 2). The nonlinear operator appearing in the problem is not necessarily differentiable, and the iteration schemes are generalizations of the usual Newton method and the modified Newton method. Similar methods were early proposed by Keller [7] for more specialized problems and may be called *Newton-like methods*. Moreover, we derive an implicit function theorem as a slight extension of those by Girault-Raviart [5] and Rappaz [14], which can be used to establish the local existence of a path of solutions. Then the abstract

results are applied to the afore-mentioned semilinear elliptic problem: we show the local existence of a path of solutions and the feasibility of the proposed iteration schemes under certain assumptions on the nonlinear term and the exact solutions (Section 3). In this process, we generalize the estimates for the nonlinear term $f(u)$ given by Kikuchi [8, 9] and Rappaz [14]. See also Caloz [3] for related results. Finally, we show the corresponding results for a simple finite element analog of the problem (Section 4). Moreover, we obtain order estimates of the errors of finite element solutions when Ω is a convex polyhedral domain (Section 5). We also illustrate some simple numerical results to see the validity of the present theoretical analysis (Section 6). Especially, it is confirmed that the asymptotic behaviors of errors predicted by the theory is in good agreement with the numerical results.

2. Newton-like methods and an implicit function theorem

In this section, we will introduce Newton-like methods and an implicit function theorem for an abstract nondifferentiable problem.

The norm of a Banach space X will be denoted by $\|\cdot\|_X$. The closed ball of radius $\delta>0$ in X centered at $x\in X$ is designated by $B(x, \delta, X)$. For two Banach spaces X and Y, $L(X, Y)$ indicates the Banach space of all linear bounded operators from X into Y. If $T\in L(X, Y)$ is bijective, we denote its inverse by T^{-1}, which belongs to $L(Y, X)$ by the open mapping theorem [16].

Let X, Y and Z be three Banach spaces, and let us consider two mappings $F: X\times Y\to Z$ and $F_y: X\times Y\to L(Y, Z)$. The values of F and F_y at $\{x, y\}\in X\times Y$ are denoted by $F(x, y)$ and $F_y(x, y)$, respectively, while that of $F_y(x, y)$ at $y^*\in Y$ is described by $F_y(x, y)y^*$. Let $\{x_0, y_0\}$ be a point in $X\times Y$, and let us set $z_0=F(x_0, y_0)$. We make the following assumptions.

[H1] There exist positive constants A and δ such that the following conditions hold.

(i) $F_y(x_0, y_0)\in L(Y, Z)$ is bijective with

$$\|F_y(x_0, y_0)^{-1}\|_{L(Z,Y)}\leqq A. \tag{3}$$

(ii-1) (4) $\|F(x, y_1)-F(x, y_2)-F_y(x_0, y_0)(y_1-y_2)\|_Z\leqq\dfrac{1}{4A}\|y_1-y_2\|_Y$;

$$\forall x \in B\left(x_0, \frac{\delta}{8A^2}, X\right), \quad \forall y_1, y_2 \in B(y_0, \delta, Y).$$

(ii-2) (5) $\|F(x_1, y)-F(x_2, y)\|_Z \leqq A\|x_1-x_2\|_X$;

$$\forall x_1, x_2 \in B\left(x_0, \frac{\delta}{8A^2}, X\right), \quad \forall y \in B(y_0, \delta, Y).$$

(ii-3) (6) $\|F_y(x, y)-F_y(x_0, y_0)\|_{L(Y,Z)} \leqq \frac{1}{4A}$;

$$\forall x \in B\left(x_0, \frac{\delta}{8A^2}, X\right), \quad \forall y \in B(y_0, \delta, Y).$$

(iii) (7) $$\|z_0\|_Z \leqq \frac{\delta}{8A}.$$

Under these hypotheses, let us find $y \in B(y_0, \delta, Y)$, for each $x \in B(x_0, \delta/(8A^2), X)$, such that

(8) $$F(x, y)=0.$$

To this end, we should first establish the following lemma.

LEMMA 1. *Under* [H1], $F_y(x, y) \in L(Y, Z)$ *is bijective with*

(9) $$\|F_y(x, y)^{-1}\|_{L(Z,Y)} \leqq \frac{4A}{3}$$

for any $x \in B(x_0, \delta/(8A^2), X)$ *and any* $y \in B(y_0, \delta, Y)$. *Moreover,*

(10) $$\|F(x, y_1)-F(x, y_2)-F_y(x^*, y^*)(y_1-y_2)\|_Z \leqq \frac{1}{2A}\|y_1-y_2\|_Y;$$

$$\forall x, x^* \in B\left(x_0, \frac{\delta}{8A^2}, X\right), \quad \forall y_1, y_2, y^* \in B(y_0, \delta, Y).$$

PROOF. Let us fix $z \in Z$ and consider an equation $F_y(x, y)y^*=z$ for $y^* \in Y$. From (i) of [H1], this is equivalent to

(a) $$y^*=F_y(x_0, y_0)^{-1}[z-\{F_y(x, y)-F_y(x_0, y_0)\}y^*].$$

Due to (i) and (ii-3) of [H1], we find that

$$\|F_y(x_0, y_0)^{-1}\{F_y(x, y)-F_y(x_0, y_0)\}\|_{L(Y,Y)} \leqq A \cdot \frac{1}{4A}=\frac{1}{4},$$

and hence we can apply the principle of contraction mappings to solve

(a). Thus we can show the existence and uniqueness of y^* with the estimate $\|y^*\|_Y \leqq (4A/3)\|z\|_Z$, and the former part of the lemma is proven. On the other hand, (10) follows from (ii-1) and (ii-3) of [H1], and the proof is complete.

From Lemma 1, we can rewrite (8) as

$$y = y - F_y(x^*, y^*)^{-1}F(x, y) \tag{11}$$

for any $x^* \in B(x_0, \delta/(8A^2), X)$ and $y^* \in B(y_0, \delta, Y)$. Suggested by (11), let us consider the following two iteration schemes.

Scheme-1. Fix $x, x^* \in B(x_0, \delta/(8A^2), X)$ and $y^{(0)}, y^* \in B(y_0, \delta, Y)$, and then obtain a sequence $\{y^{(k)}\}_{k=1}^{\infty}$ in Y recursively by

$$y^{(k)} = y^{(k-1)} - F_y(x^*, y^*)^{-1}F(x, y^{(k-1)}) \qquad (k=1, 2, \cdots). \tag{12}$$

Scheme-2. Fix $x \in B(x_0, \delta/(8A^2), X)$ and $y^{(0)} \in B(y_0, \delta, Y)$, and then obtain a sequence $\{y^{(k)}\}_{k=1}^{\infty}$ in Y recursively by

$$y^{(k)} = y^{(k-1)} - F_y(x, y^{(k-1)})^{-1}F(x, y^{(k-1)}) \qquad (k=1, 2, \cdots). \tag{13}$$

We expect that the sequences above converge to a solution of (8). Note that Schemes-1 and -2 respectively coincide with the modified Newton method and the usual Newton one when F_y is the Fréchet derivative of F, see Liusternik-Sobolev [13]. So, we call the present schemes *Newton-like methods.*

For the convergence of these schemes, we have the following theorem. See also Keller [7] for related results.

THEOREM 1. *Under* [H1], *each of Schemes*-1 *and* -2 *gives a sequence contained in and converging in* $B(y_0, \delta, Y)$, *and the limit is a unique solution of* (8) *in* $B(y_0, \delta, Y)$ *for each* $x \in B(x_0, \delta/(8A^2), X)$.

PROOF. (i) *Scheme*-1. It suffices to show that the mapping $F^*: B(y_0, \delta, Y) \to Y$ defined by

$$F^*(y) = y - F_y(x^*, y^*)^{-1}F(x, y) \quad \text{for each} \quad y \in B(y_0, \delta, Y)$$

is contractive when x, x^*, and y^* are fixed as indicated in Scheme-1. Since $z_0 = F(x_0, y_0)$, we have

$$F^*(y) - y_0 = F_y(x^*, y^*)^{-1}\{F_y(x^*, y^*)(y - y_0) + F(x, y_0) - F(x, y) - F(x, y_0) + F(x_0, y_0) - z_0\}.$$

Then it follows from (5), (7), (9), and (10) that

$$\|F^*(y)-y_0\|_Y \leqq \frac{4A}{3}\left(\frac{1}{2A}\|y-y_0\|_Y + A\|x-x_0\|_X + \|z_0\|_Z\right)$$
$$\leqq \frac{4A}{3}\left(\frac{1}{2A}\cdot\delta + A\cdot\frac{\delta}{8A^2} + \frac{\delta}{8A}\right)=\delta,$$

and hence the image of F^* is contained in $B(y_0, \delta, Y)$. For any y_1, $y_2 \in B(y_0, \delta, Y)$, we find

$$F^*(y_1)-F^*(y_2)=F_y(x^*, y^*)^{-1}\{F_y(x^*, y^*)(y_1-y_2)-F(x, y_1)+F(x, y_2)\}.$$

Applying Lemma 1 to the above, we have

$$\|F^*(y_1)-F^*(y_2)\|_Y \leqq \frac{4A}{3}\cdot\frac{1}{2A}\|y_1-y_2\|_Y=\frac{2}{3}\|y_1-y_2\|_Y.$$

Thus we have proven that the mapping F^* is contractive.

(ii) *Scheme*-2. Define $F^*: B(y_0, \delta, Y)\to Y$ by

$$F^*(y)=y-F_y(x, y)^{-1}F(x, y) \qquad \text{for each } y \in B(y_0, \delta, Y),$$

where $x \in B(x_0, \delta/(8A^2), X)$ is fixed as a parameter. Then (13) is expressed by $y^{(k)}=F^*(y^{(k-1)})$. Like part (i), the image of F^* is shown to be contained in $B(y_0, \delta, Y)$, and hence the iteration process is well-defined. In this case, however, we cannot rely upon the principle of contraction mappings. But, in part (i), we have already shown the unique existence of the solution of (8) for each $x \in B(x_0, \delta/(8A^2), X)$. Thus, denoting the solution by $G(x)$, let us evaluate $F^*(y)-G(x)$ for $y \in B(y_0, \delta, Y)$. Since

$$F^*(y)-G(x)=F_y(x, y)^{-1}\{F_y(x, y)(y-G(x))-F(x, y)+F(x, G(x))\},$$

we have from Lemma 1 that

$$\|F^*(y)-G(x)\|_Y \leqq \frac{4A}{3}\cdot\frac{1}{2A}\|y-G(x)\|_Y=\frac{2}{3}\|y-G(x)\|_Y.$$

This estimation shows that the sequence based on Scheme-2 converges to the unique solution $G(x)$ of (8) in $B(y_0, \delta, Y)$, and the proof is complete.

REMARK 1. If it holds that, besides [H1],

$$\|F_y(x, y_1)-F_y(x, y_2)\|_{L(Y,Z)} \leqq \frac{1}{4A\delta}\|y_1-y_2\|_Y; \tag{14}$$

$$\forall x \in B\left(x_0, \frac{\delta}{8A^2}, X\right), \quad \forall y_1, y_2 \in B(y_0, \delta, Y),$$

then we can show that the iteration process based on Scheme-2 is also contractive.

We can also derive an implicit function theorem from [H1] as a generalization of those by Girault-Raviart [5] and Rappaz [14].

THEOREM 2. *Under* (i), (ii-1), (ii-2) *and* (iii) *of* [H1], *there exists a unique mapping* $G: B(x_0, \delta/(8A^2), X) \to B(y_0, \delta, Y)$ *such that* $F(x, G(x))=0$ *for each* $x \in B(x_0, \delta/(8A^2), X)$. *Moreover,*

$$\|G(x)-y_0\|_Y \leqq \frac{4A}{3}(A\|x-x_0\|_X + \|z_0\|_Z); \ \forall x \in B\left(x_0, \frac{\delta}{8A^2}, X\right), \tag{15}$$

$$\|G(x_1)-G(x_2)\|_Y \leqq \frac{4A^2}{3}\|x_1-x_2\|_X; \ \forall x_1, x_2 \in B\left(x_0, \frac{\delta}{8A^2}, X\right). \tag{16}$$

PROOF. We have already proven the existence and uniqueness of such a mapping in part (i) of the proof of Theorem 1. Note also that (ii-3) of [H1] is unnecessary if we take $\{x^*, y^*\}$ as $\{x_0, y_0\}$. Applying (i), (ii-1) and (ii-2) of [H1] to the identity

$$G(x)-y_0 = F_y(x_0, y_0)^{-1}\{F_y(x_0, y_0)(G(x)-y_0) - F(x, G(x)) + F(x, y_0) - F(x, y_0) + F(x_0, y_0) - z_0\},$$

we find that

$$\|G(x)-y_0\|_Y \leqq A\left(\frac{1}{4A}\|G(x)-y_0\|_Y + A\|x-x_0\|_X + \|z_0\|_Z\right),$$

from which (15) follows. Similarly, we can obtain (16) by the use of the identity

$$G(x_1)-G(x_2) = F_y(x_0, y_0)^{-1}\{F_y(x_0, y_0)(G(x_1)-G(x_2)) - F(x_1, G(x_1)) + F(x_1, G(x_2)) - F(x_1, G(x_2)) + F(x_2, G(x_2))\},$$

and the proof is complete.

3. Analysis of a nondifferentiable nonlinear eigenvalue problem

Let Ω be a bounded domain in $\boldsymbol{R}^n$ $(n=1, 2, 3)$ with a Lipschitz continuous boundary $\partial\Omega$, and let $f: \boldsymbol{R}^1 \to \boldsymbol{R}^1$ be a given continuous function.

As a generalization of (1), let us consider the problem of finding a real number λ and a real function $u=u(\xi)$ such that

$$-\Delta u=\lambda f(u) \quad \text{in } \Omega, \qquad u=0 \quad \text{on } \partial\Omega, \tag{17}$$

where $f(u)$ denotes the composite function of f and u. In this section, we will analyze this problem under some assumptions.

For Ω, we consider the usual real Sobolev spaces $H^k(\Omega)$ and $H_0^k(\Omega)$ $(k=0, 1, 2, \cdots)$. The norm of $H^k(\Omega)$ (and $H_0^k(\Omega)$) is denoted by $\|\cdot\|_k$. In particular, the inner product and the norm of $H^0(\Omega)=L_2(\Omega)$ are designated by $(\cdot, \cdot)$ and $\|\cdot\|$, respectively. The dual space of $H_0^1(\Omega)$ and its norm are denoted by $H^{-1}(\Omega)$ and $\|\cdot\|_{-1}$, respectively. We will also use $(\cdot, \cdot)$ as the duality pairing over $H^{-1}(\Omega)\times H_0^1(\Omega)$ with the understanding that $L_2(\Omega)\subset H^{-1}(\Omega)$. The norm of $L_p(\Omega)$ for $p\geqq 1$ is denoted by $\|\cdot\|_{L_p}$. We will often use abbreviated notations H_0^1, L_p, H^{-1} etc.

We employ the following hypotheses for $f=f(t)$.

[H2] Let $\{t_i\}_{i=1}^m$ ($m=$positive integer) be a sequence in $\boldsymbol{R}^1$ such that $t_1<t_2<\cdots<t_m$. Define open intervals $\{I_i\}_{i=0}^m$ by $I_0=]-\infty, t_1[$, $I_i=]t_i, t_{i+1}[$ for $1\leqq i\leqq m-1$ $(m\geqq 2)$, and $I_m=]t_m, \infty[$. We assume that f satisfies:

(i) $f=f(t)$ is continuously differentiable in $\boldsymbol{R}^1$ except at $t=t_i$ for $1\leqq i\leqq m$. We denote the derivative by $f'=f'(t)$.

(ii) For any i with $1\leqq i\leqq m$, both $f'(t_i-0)$ and $f'(t_i+0)$ exist and are finite.

(iii) One of the following three conditions holds according to the value of n.

(iii-1) $n=1$. For each $\varepsilon>0$, there exists a positive number $M_1(\varepsilon)$ such that for any i with $0\leqq i\leqq m$

$$|f'(t)-f'(t^*)|\leqq M_1(\varepsilon)|t-t^*|;\ \forall t, t^*\in B(0, \varepsilon, \boldsymbol{R}^1)\cap I_i. \tag{18}$$

(iii-2) $n=2$. There exist a positive constant M_2 and a nonnegative constant α such that for any i with $0\leqq i\leqq m$

$$|f'(t)-f'(t^*)|\leqq M_2\max\{1, |t|^\alpha, |t^*|^\alpha\}|t-t^*|;\ \forall t, t^*\in I_i. \tag{19}$$

(iii-3) $n=3$. There exists a positive constant M_3 such that for any i with $0\leqq i\leqq m$

$$|f'(t)-f'(t^*)|\leqq M_3\max\{1, |t|, |t^*|\}|t-t^*|;\ \forall t, t^*\in I_i. \tag{20}$$

REMARK 2. One of the main reasons why the conditions above differ with n is attributed to the Sobolev imbedding theorem to be used later.

From [H2], we have the following lemma, whose proof is elementary and is omitted here.

LEMMA 2. *Under* [H2], $f'=f'(t)$ *and* $f=f(t)$ *satisfy one of the following three conditions according to the value of* n.

(i) $n=1$. *For each* $\varepsilon>0$, *there exists a positive number* $M_1^*(\varepsilon)$ *such that*

$$|f'(t)|\leqq M_1^*(\varepsilon);\ \forall t\in B(0,\ \varepsilon,\ \boldsymbol{R}^1)\backslash\{t_i\}_{i=1}^m, \tag{21}$$

$$|f(t)-f(t^*)|\leqq M_1^*(\varepsilon)|t-t^*|;\ \forall t,\ t^*\in B(0,\ \varepsilon,\ \boldsymbol{R}^1), \tag{22}$$

$$|f(t)|\leqq M_1^*(\varepsilon);\ \forall t\in B(0,\ \varepsilon,\ \boldsymbol{R}^1). \tag{23}$$

(ii) $n=2$. *There exists a positive constant* M_2^* *such that*

$$|f'(t)|\leqq M_2^*\max\{1,\ |t|^{\alpha+1}\};\ \forall t\in\boldsymbol{R}^1\backslash\{t_i\}_{i=1}^m, \tag{24}$$

$$|f(t)-f(t^*)|\leqq M_2^*\max\{1,\ |t|^{\alpha+1},\ |t^*|^{\alpha+1}\}|t-t^*|;\ \forall t,\ t^*\in\boldsymbol{R}^1, \tag{25}$$

$$|f(t)|\leqq M_2^*\max\{1,\ |t|^{\alpha+2}\};\ \forall t\in\boldsymbol{R}^1, \tag{26}$$

where α *is the nonnegative constant in* (iii-2) *of* [H2].

(iii) $n=3$. *There exists a positive constant* M_3^* *such that*

$$|f'(t)|\leqq M_3^*\max\{1,\ t^2\};\ \forall t\in\boldsymbol{R}^1\backslash\{t_i\}_{i=1}^m, \tag{27}$$

$$|f(t)-f(t^*)|\leqq M_3^*\max\{1,\ t^2,\ (t^*)^2\}|t-t^*|;\ \forall t,\ t^*\in\boldsymbol{R}^1, \tag{28}$$

$$|f(t)|\leqq M_3^*\max\{1,\ |t|^3\};\ \forall t\in\boldsymbol{R}^1. \tag{29}$$

Let us introduce an extension ∂f of f' to $\boldsymbol{R}^1$ such that it has finite values at $t=t_i$ $(1\leqq i\leqq m)$. A typical example employed in [10, 11] is

$$\partial f(t)=\begin{cases} f'(t) & \text{if } t\neq t_i \text{ for any } i\ (1\leqq i\leqq m), \\ \dfrac{f'(t_i-0)+f'(t_i+0)}{2} & \text{if } t=t_i \text{ for some } i\ (1\leqq i\leqq m). \end{cases} \tag{30}$$

Hereafter, we will denote the composite function of ∂f and $u\in H^1$ by $\partial f(u)$. It is easy to see that $f(u)$ and $\partial f(u)$ are measurable functions

in Ω for any $u \in H^1$. Moreover, by the Sobolev imbedding theorem and Lemma 2, it holds that, for $n=1, 2, 3$ and any $u, v \in H^1$,

$$f(u) \in L_2, \quad \partial f(u) \in L_3, \quad \partial f(u)v \in L_2. \tag{31}$$

As a weak formulation of (17), let us consider the problem: find a pair $\{\lambda, u\} \in \boldsymbol{R}^1 \times H_0^1$ such that

$$\langle u, v\rangle = \lambda(f(u), v); \ \forall v \in H_0^1, \tag{32}$$

where

$$\langle u, v\rangle = \sum_{i=1}^{n} \int_\Omega \frac{\partial u}{\partial \xi_i} \frac{\partial v}{\partial \xi_i} d\xi. \tag{33}$$

We make the following assumptions on this problem.

[H3] There exists a pair $\{\lambda_0, u_0\} \in \boldsymbol{R}^1 \times H_0^1$ that satisfies (32) and the conditions below.

(i) The Lebesgue measure of the following set is zero:

$$\Omega_0 = \{\xi \in \Omega \mid u_0(\xi) = t_i \text{ for some } i \text{ with } 1 \leqq i \leqq m\}. \tag{34}$$

(ii) The problem of finding $w \in H_0^1$ such that

$$\langle w, v\rangle - \lambda_0(\partial f(u_0)w, v) = 0 \qquad (\forall v \in H_0^1) \tag{35}$$

has the trivial solution $w=0$ only.

REMARK 3. Due to (i), condition (ii) does not alter with the definition of ∂f.

Let us derive some lemmas from [H2] and [H3].

LEMMA 3. *For each $g \in H^{-1}$, there exists a unique function w in H_0^1 such that*

$$\langle w, v\rangle - \lambda_0(\partial f(u_0)w, v) = (g, w); \quad \forall v \in H_0^1. \tag{36}$$

Moreover, there exists a positive constant C independent of g such that

$$\|w\|_1 \leqq C\|g\|_{-1}. \tag{37}$$

PROOF. By the Sobolev imbedding theorem, H^1 is continuously imbedded to L_p for $1 \leqq p \leqq 6$ and $n \leqq 3$. Thus by the Hölder inequality and (31), we have for any $w, v \in H_0^1$ that

$$|(\partial f(u_0)w, v)| \leqq \|\partial f(u_0)\|_{L_3}\|w\|_{L_2}\|v\|_{L_6} \leqq C^*\|\partial f(u_0)\|_{L_3}\|w\|_{L_2}\|v\|_{H^1},$$

where C^* is a positive constant. That is, the multiplication of $w \in H_0^1$ by $\partial f(u_0)$ is a linear bounded operator from H_0^1 to H^{-1}. Moreover, this operator is compact. To see this fact, let us denote this operator by T and consider an arbitrary sequence $\{u_k\}_{k=1}^{\infty}$ in H_0^1 that converges weakly to u^* in H_0^1. Note that this sequence also converges strongly to u^* in L_2 due to the Rellich theorem of choice. From the above inequality, we have

$$\sup_{v \in H_0^1 \setminus \{0\}} \frac{|(T(u_k-u^*), v)|}{\|v\|_{H^1}} \leqq C^*\|\partial f(u_0)\|_{L_3}\|u_k-u^*\|_{L_2}.$$

Thus, Tu_k converges strongly to Tu^* in H^{-1} as $k \to \infty$, and hence T is a compact operator from H_0^1 to H^{-1}. Now we can use the Fredholm alternative with [H3]-(ii) to draw the conclusions. This completes the proof.

LEMMA 4. *For any $\varepsilon>0$, there exists a positive number $M(\varepsilon)$ such that*

$$\|f(u_1)-f(u_2)\|_{-1} \leqq M(\varepsilon)\|u_1-u_2\|; \quad \forall u_1, u_2 \in B(u_0, \varepsilon, H^1). \tag{38}$$

PROOF. For simplicity, we will prove (38) only for $n=3$: the proof is essentially the same for $n=1, 2$. From (28), we have for $u_1, u_2 \in H^1$ and $v \in H_0^1$ that

$$\begin{aligned}|(f(u_1)-f(u_2), v)| &\leqq \int_\Omega |f(u_1(\xi))-f(u_2(\xi))| \cdot |v(\xi)| d\xi \\ &\leqq M_3^* \int_\Omega \max\{1, |u_1(\xi)|^2, |u_2(\xi)|^2\}|u_1(\xi)-u_2(\xi)| \cdot |v(\xi)| d\xi.\end{aligned}$$

As in the proof of the preceding lemma, the above may be further evaluated as

$$|(f(u_1)-f(u_2), v)| \leqq M_3^*(1+\|u_1\|_{L_6}^2+\|u_2\|_{L_6}^2)\|u_1-u_2\|_{L_2}\|v\|_{L_6}.$$

The desired estimation immediately follows from the above since H^1 is continuously imbedded to L_6, and the proof is complete.

LEMMA 5. *For each $\kappa>0$, there exists a positive number $\gamma(\kappa)$ such that, for any $u_1, u_2 \in B(u_0, \gamma(\kappa), H^1)$,*

$$\|f(u_1)-f(u_2)-\partial f(u_0)(u_1-u_2)\|_{-1} \leqq \kappa\|u_1-u_2\|_1, \tag{39}$$

$$\|\{\partial f(u_1)-\partial f(u_2)\}v\|_{-1}\leqq\kappa\|v\|_1;\quad \forall v\in H^1. \tag{40}$$

REMARK 4. Due to (39), $f(u)$, as an operator from H^1 into H^{-1}, is Fréchet differentiable at $u=u_0$ and the value of the derivative there is equal to the multiplication by $\partial f(u_0)$. However, $f(u)$ is not necessarily Fréchet differentiable except there.

PROOF. This is an extension of Lemma 4.2 of Kikuchi [8] and Lemma 4 of Rappaz [14]. As in Lemma 4, we will give a proof only for $n=3$. Let u_1 and u_2 be arbitrary functions in H^1. For $r>0$ and $u\in H^1$, define

$$\Omega_r=\{\xi\in\Omega \mid \min_{1\leqq i\leqq m}|u_0(\xi)-t_i|>r\},$$
$$\omega_r(u)=\{\xi\in\Omega \mid |u(\xi)-u_0(\xi)|\leqq r\},$$

where $\{t_i\}_{i=1}^m$ is the sequence in $\boldsymbol{R}^1$ given in [H2].

For $\xi\in\Omega_r\cap\omega_r(u_1)\cap\omega_r(u_2)$, we find

$$|u_j(\xi)-t_i|\geqq|u_0(\xi)-t_i|-|u_j(\xi)-u_0(\xi)|>r-r=0\quad (j=1,\,2;\ 1\leqq i\leqq m).$$

Clearly, $u_0(\xi)\in I_i$ for some i with $0\leqq i\leqq m$. Then the above estimates show that $u_1(\xi)\in I_i$ and $u_2(\xi)\in I_i$. Thus, by the mean value theorem, there exists a number θ such that $0<\theta<1$ and

$$\begin{aligned}&f(u_1(\xi))-f(u_2(\xi))-f'(u_0(\xi))\{u_1(\xi)-u_2(\xi)\}\\&\quad=[f'(u_2(\xi)+\theta\{u_1(\xi)-u_2(\xi)\})-f'(u_0(\xi))]\{u_1(\xi)-u_2(\xi)\}.\end{aligned}$$

Applying (20) to this identity with the estimation $|u_2(\xi)+\theta\{u_1(\xi)-u_2(\xi)\}-u_0(\xi)|\leqq r$ taken into account, we have

$$\begin{aligned}(a)\qquad&|f(u_1(\xi))-f(u_2(\xi))-f'(u_0(\xi))\{u_1(\xi)-u_2(\xi)\}|\\&\quad\leqq M_3 r\max\{1,\,|u_0(\xi)|,\,|u_1(\xi)|,\,|u_2(\xi)|\}|u_1(\xi)-u_2(\xi)|.\end{aligned}$$

For $\xi\in\Omega^*\equiv\Omega\backslash\{\Omega_0\cup(\Omega_r\cap\omega_r(u_1)\cap\omega_r(u_2))\}$, we use (27) and (28) to obtain

$$\begin{aligned}(b)\qquad&|f(u_1(\xi))-f(u_2(\xi))-f'(u_0(\xi))\{u_1(\xi)-u_2(\xi)\}|\\&\quad\leqq|f(u_1(\xi))-f(u_2(\xi))|+|f'(u_0(\xi))|\cdot|u_1(\xi)-u_2(\xi)|\\&\quad\leqq M_3^*\max\{1,\,|u_0(\xi)|^2,\,|u_1(\xi)|^2,\,|u_2(\xi)|^2\}|u_1(\xi)-u_2(\xi)|.\end{aligned}$$

As in the proof of Lemma 3, it follows from (a), (b), [H3]-(i) and the Hölder inequality that, for any $v\in H_0^1$,

$$
\begin{aligned}
&|(f(u_1)-f(u_2)-\partial f(u_0)(u_1-u_2),\ v)| \\
&\quad\leqq\int_{\Omega\setminus\Omega_0}|f(u_1(\xi))-f(u_2(\xi))-f'(u_0(\xi))\{u_1(\xi)-u_2(\xi)\}|\cdot|v(\xi)|d\xi \\
&\quad\leqq M_3 r(1+\|u_0\|_{L_3}+\|u_1\|_{L_3}+\|u_2\|_{L_3})\|u_1-u_2\|_{L_3}\|v\|_{L_3} \\
&\qquad+M_3^*\Big(\int_{\Omega^*}d\xi\Big)^{1/3}(1+\|u_0\|_{L_6}^2+\|u_1\|_{L_6}^2+\|u_2\|_{L_6}^2)\|u_1-u_2\|_{L_6}\|v\|_{L_6}.
\end{aligned}
$$

Notice here that $\int_{\Omega^*}d\xi$ is equal to the measure of Ω^*, which can be made arbitrarily close to zero by choosing r, $\|u_1-u_0\|_1$ and $\|u_2-u_0\|_1$ appropriately small. Thus we can obtain (39).

Estimation (40) may be derived similarly. From (20), we find for $\xi\in\Omega_r\cap\omega_r(u_1)\cap\omega_r(u_2)$ that

$$
|f'(u_1(\xi))-f'(u_2(\xi))|\leqq M_3\max\{1,\ |u_1(\xi)|,\ |u_2(\xi)|\}|u_1(\xi)-u_2(\xi)|.
$$

On the other hand, for $\xi\in\Omega^*$, we find from (27) that

$$
|f'(u_1(\xi))-f'(u_2(\xi))|\leqq 2M_3^*\max\{1,\ |u_1(\xi)|^2,\ |u_2(\xi)|^2\}.
$$

From these two estimates, we have for any $v\in H^1$ and $w\in H_0^1$ that

$$
\begin{aligned}
&|(\{\partial f(u_1)-\partial f(u_2)\}v,\ w)| \\
&\quad\leqq M_3(1+\|u_1\|_{L_3}+\|u_2\|_{L_3})\|u_1-u_2\|_{L_6}\|v\|_{L_6}\|w\|_{L_3} \\
&\qquad+2M_3^*\Big(\int_{\Omega^*}d\xi\Big)^{1/3}(1+\|u_1\|_{L_6}^2+\|u_2\|_{L_6}^2)\|v\|_{L_6}\|w\|_{L_6}.
\end{aligned}
$$

Thus we obtain (40) by choosing r, $\|u_1-u_0\|_1$ and $\|u_2-u_0\|_1$ appropriately small, and the proof is complete.

Before applying the general theory of Section 2 to the present problem, let us consider the problem: given $g\in H^{-1}$, find $w\in H_0^1$ such that

$$
\langle w,\ v\rangle=(g,\ v);\quad \forall v\in H_0^1. \tag{41}
$$

It is well-known that the solution w exists uniquely in H_0^1 for each $g\in H^{-1}$ and satisfies

$$
\|w\|_1\leqq C\|g\|_{-1}, \tag{42}
$$

where C is a positive constant independent of g. Henceforth, we will use C, C^* etc. as the notations of generic positive constants. So, C above is not necessarily the same as that in (37).

From the preceding observations, the following definitions make sense:

$$X=\boldsymbol{R}^1, \quad Y=Z=H_0^1, \tag{43}$$

$$S\in L(H^{-1}, Y); \quad Sg=w\in Y \text{ of (41) for each } g\in H^{-1}, \tag{44}$$

$$F: X\times Y\to Y; \quad F(\lambda, u)=u-\lambda Sf(u) \text{ for each } \{\lambda, u\}\in X\times Y, \tag{45}$$

$$F_y: X\times Y\to L(Y, Y); \quad F_y(\lambda, u)v=v-\lambda S\partial f(u)v \text{ for each } \{\lambda, u, v\}\in X\times Y\times Y. \tag{46}$$

Clearly, we have

$$\langle F(\lambda, u), v\rangle=\langle u, v\rangle-\lambda(f(u), v); \quad \forall\{\lambda, u, v\}\in X\times Y\times Y, \tag{47}$$

$$\langle F_y(\lambda, u)w, v\rangle=\langle w, v\rangle-\lambda(\partial f(u)w, v); \quad \forall\{\lambda, u, v, w\}\in X\times Y\times Y\times Y. \tag{48}$$

Thus, (32) is equivalent to finding $\{\lambda, u\}\in X\times Y$ such that $F(\lambda, u)=0$.

With the preparations above, let us check [H1] by putting $\{x_0, y_0\}=\{\lambda_0, u_0\}$ and $z_0=F(\lambda_0, u_0)$ $(=0)$, where $\{\lambda_0, u_0\}$ is the same as in [H3]. First, Lemma 3 assures (i) of [H1] thanks to (48). Next, (ii) of [H1] follows from (31) and Lemmas 4 and 5. Finally, (iii) of [H1] holds since $z_0=0$. Although we do not give explicit expressions of A and δ in [H1], the existence of such quantities is clear. Thus we have the following theorem.

THEOREM 3. *Let us consider f, ∂f and $\{\lambda_0, u_0\}\in\boldsymbol{R}^1\times H_0^1$ which are introduced in this section and satisfy* [H2] *and* [H3]. *Then there exist two positive numbers μ_0 and δ_0 as well as a continuous mapping $u(\cdot): [\lambda_0-\mu_0, \lambda_0+\mu_0]\to B(u_0, \delta_0, H_0^1)$ such that*

(i) *$u(\lambda)$ for each $\lambda\in[\lambda_0-\mu_0, \lambda_0+\mu_0]$ satisfies $F(\lambda, u(\lambda))=0$ or* (32) *uniquely in $B(u_0, \delta_0, H_0^1)$. Moreover, $u(\lambda_0)=u_0$.*

(ii) *$u(\lambda)$ is Lipschitz continuous with respect to λ in the sense that there exists a positive constant C such that*

$$\|u(\lambda_1)-u(\lambda_2)\|_1\leqq C|\lambda_1-\lambda_2|; \quad \forall\lambda_1, \lambda_2\in[\lambda_0-\mu_0, \lambda_0+\mu_0]. \tag{49}$$

(iii) *When Schemes-1 and -2 introduced in Section 2 are applied to solving $F(\lambda, u)=0$, they give sequences converging to $u(\lambda)$ in H_0^1 for each $\lambda\in[\lambda_0-\mu_0, \lambda_0+\mu_0]$ provided that $\{x^*, y^*\}=\{\lambda^*, u^*\}$ of Scheme-1 is in $[\lambda_0-\mu_0, \lambda_0+\mu_0]\times B(u_0, \delta_0, H_0^1)$ and that $y^{(0)}=u^{(0)}$ of Schemes-1 and -2 is in $B(u_0, \delta_0, H_0^1)$.*

4. Finite element approximation

In this section, we consider a finite element approximation to (32). Hereafter, we restrict our analysis to the case when Ω is a bounded polyhedral domain in $\boldsymbol{R}^n$ ($n=1, 2, 3$). Then we consider a regular family of triangulations $\{T^h\}_{h>0}$ of Ω by n-simplices (i.e., finite elements), where $h>0$ is the maximum side length of all n-simplices in each triangulation T^h. The precise meaning of "regular" is given in Ciarlet [4]. Roughly speaking, the family $\{T^h\}_{h>0}$ is said to be regular if: (i) each T^h is composed of finite number of n-simplices without any gaps and overlaps so that any face of a simplex in T^h either lies on $\partial\Omega$ or coincides with a face of another simplex in T^h, (ii) the simplices in $\{T^h\}_{h>0}$ are not too flat, and (iii) we can find a sequence of triangulations in $\{T^h\}_{h>0}$ such that $h\to 0$. These conditions are essential to establish some basic properties of finite element approximations such as [H4] below.

Let us consider a finite dimensional subspace V^h of H_0^1 associated with each T^h. We assume that V^h is equipped with the norm of H_0^1 and employ the following hypotheses for the family $\{V^h\}_{h>0}$.

[H4] The family $\{V^h\}_{h>0}$ satisfies

$$\inf_{v_h\in V^h}\|v_h-v\|_1\longrightarrow 0 \quad \text{as } h\to 0 \text{ for each } v\in H_0^1, \tag{50}$$

$$\inf_{v_h\in V^h}\|v_h-v\|_1\leqq Ch\|v\|_2 \qquad \text{for each } v\in H^2\cap H_0^1, \tag{51}$$

where C is a positive constant independent of h and v.

A typical and the simplest example of a family of subspaces that satisfies [H4] is the piecewise linear finite element approximation associated with $\{T^h\}_{h>0}$, see Ciarlet [4].

Now a finite element analog of (32) is to find a pair $\{\lambda, u_h\}\in \boldsymbol{R}^1\times V^h$ such that

$$\langle u_h, v_h\rangle=\lambda(f(u_h), v_h); \quad \forall v_h\in V^h. \tag{52}$$

As an approximation of (41), let us consider the problem: given $g\in H^{-1}$, find $w_h\in V^h$ such that

$$\langle w_h, v_h\rangle=(g, v_h); \quad \forall v_h\in V^h. \tag{53}$$

As is well-known, the solution w_h exists uniquely in V^h for each $g\in H^{-1}$ and satisfies

$$\|w_h\|_1 \leqq C\|g\|_{-1}, \tag{54}$$

where C is a positive constant independent of g and h. Thus we can define the following approximation operator of S in (44).

$$S_h \in L(H^{-1}, V^h); \quad S_h g = w_h \in V^h \text{ of (53) for each } g \in H^{-1}. \tag{55}$$

From (50) and (54), we can show for each $g \in H^{-1}$ that

$$\|S_h g\|_1 \leqq C\|g\|_{-1}, \quad \lim_{h\to 0} \|S_h g - Sg\|_1 = 0. \tag{56}$$

As in Section 3, let us make the following definitions:

$$X = \boldsymbol{R}^1, \quad Y = Z = V^h, \tag{57}$$

$$F_h: X \times Y \to Y; \quad F_h(\lambda, u_h) = u_h - \lambda S_h f(u_h) \text{ for each } \{\lambda, u_h\} \in X \times Y, \tag{58}$$

$$F_{hy}: X \times Y \to L(Y, Y); \quad F_{hy}(\lambda, u_h)v_h = v_h - \lambda S_h \partial f(u_h)v_h \text{ for each } \{\lambda, u_h, v_h\} \in X \times Y \times Y, \tag{59}$$

where F_h and F_{hy} are respectively the approximations of F and F_y in Section 3. Then we find that

$$\langle F_h(\lambda, u_h), v_h\rangle = \langle u_h, v_h\rangle - \lambda(f(u_h), v_h); \quad \forall\{\lambda, u_h, v_h\} \in X \times Y \times Y, \tag{60}$$

$$\langle F_{hy}(\lambda, u_h)w_h, v_h\rangle = \langle w_h, v_h\rangle - \lambda(\partial f(u_h)w_h, v_h); \quad \forall\{\lambda, u_h, v_h, w_h\} \in X \times Y \times Y \times Y, \tag{61}$$

and hence (52) is equivalent to finding $\{\lambda, u_h\} \in X \times Y$ such that $F_h(\lambda, u_h) = 0$. Moreover, let us define $u_{h0} \in X = V^h$ by

$$u_{h0} = \lambda_0 S_h f(u_0), \tag{62}$$

which is well-defined thanks to (31), and will be used as an approximation of u_0 $(= \lambda_0 S f(u_0))$ in [H3]. By (56), we have

$$\lim_{h\to 0} \|u_{h0} - u_0\|_1 = 0. \tag{63}$$

Before applying the general theory of Section 2 to the present approximate problem, we also prepare a few lemmas.

LEMMA 6. *u_{h0} defined by (62) for $h>0$ satisfies, as $h \to 0$,*

$$\|F_h(\lambda_0, u_{h0})\|_1 \longrightarrow 0, \tag{64}$$

$$\sup_{w \in H_0^1 \setminus \{0\}} \frac{\|\{\partial f(u_0) - \partial f(u_{h0})\}w\|_{-1}}{\|w\|_1} \longrightarrow 0. \tag{65}$$

PROOF. From (60) and (62), $F_h(\lambda_0, u_{h0})$ satisfies for any $v_h \in V^h$

$$\langle F_h(\lambda_0, u_{h0}), v_h\rangle = \langle u_{h0}, v_h\rangle - \lambda_0(f(u_{h0}), v_h) = \lambda_0(f(u_0) - f(u_{h0}), v_h),$$

and hence

$$\|F_h(\lambda_0, u_{h0})\|_1 \leqq |\lambda_0| \cdot \|f(u_0) - f(u_{h0})\|_{-1},$$

since $F_h(\lambda_0, u_{h0})$ is in V^h. Applying Lemma 4 with (63) to this estimation, we have (64). On the other hand, (65) follows from Lemma 5 and (63), and the proof is complete.

LEMMA 7. *There exists a positive number h_0 such that the following holds:*
(i) *For $h \in]0, h_0]$ and each given $g \in H^{-1}$, there exists a unique function $w_h \in V^h$ such that*

$$\langle w_h, v_h\rangle - \lambda_0(\partial f(u_0)w_h, v_h) = (g, v_h); \quad \forall v_h \in V^h. \tag{66}$$

Moreover, the present w_h satisfies

$$\|w_h\|_1 \leqq C\|g\|_{-1}, \tag{67}$$

where C is a positive constant independent of g and h.
(ii) *For $h \in]0, h_0]$ and each given $g \in H^{-1}$, there exists a unique function $w_h^* \in V^h$ such that*

$$\langle w_h^*, v_h\rangle - \lambda_0(\partial f(u_{h0})w_h^*, v_h) = (g, v_h); \quad \forall v_h \in V^h, \tag{68}$$

or, equivalently, $F_{hy}(\lambda_0, u_{h0})w_h^* = S_h g$, where u_{h0} is defined in (62). *Moreover, the present w_h^* satisfies*

$$\|w_h^*\|_1 \leqq C^*\|g\|_{-1}, \tag{69}$$

where C^* is a positive constant similar to C above.

PROOF. These two problems (66) and (68) are discrete analogs of (36), and (i) follows directly from Theorem 1.5 in Ch. 3 of Aubin [1]. On the other hand, (ii) is obtained by regarding (68) as a perturbation problem of (66) and noting (65). This completes the proof.

With the preparations above, let us check [H1] by putting $\{x_0, y_0\} = \{\lambda_0, u_{h0}\}$ and $z_0 = F_h(\lambda_0, u_{h0})$, where λ_0 is the same as in [H3] and u_{h0} is

defined by (62) for u_0 in [H3]. Of course, F_h and F_{hy} respectively correspond to F and F_y in [H1], and we assume that $h>0$ is small enough. First, (ii) of Lemma 7 assures (i) of [H1]. Next, (ii) of [H1] follows from (65) and Lemmas 4 and 5. Finally, (iii) of [H1] holds thanks to (64). We do not give explicit expressions of A and δ in [H1], but the existence of such quantities is clear, and, moreover, they can be chosen to be independent of $h>0$ when h is sufficiently small. Thus we have the following theorem.

THEOREM 4. *Let us consider f, ∂f and $\{\lambda_0, u_0\} \in \boldsymbol{R}^1 \times H_0^1$ which are introduced in Section 3 and satisfy* [H2] *and* [H3]. *Moreover, assume that* [H4] *holds and consider u_{h0} in* (62). *Then there exist positive numbers h_1, μ_1 and δ_1 as well as a continuous mapping $u_h(\cdot): [\lambda_0-\mu_1, \lambda_0+\mu_1] \to B(u_{h0}, \delta_1, V^h)$ for $h \in]0, h_1]$ such that*

(i) *$u_h(\lambda)$ for each $\lambda \in [\lambda_0-\mu_1, \lambda_0+\mu_1]$ satisfies $F_h(\lambda, u_h(\lambda))=0$ or* (52) *uniquely in $B(u_{h0}, \delta_1, V^h)$. Moreover, $u_h(\lambda_0) \to u_0$ in H_0^1 as $h \to 0$.*

(ii) *$u_h(\lambda)$ is Lipschitz continuous with respect to λ in the sense that there exists a positive constant C (independent of $h \in]0, h_1]$) such that*

$$\|u_h(\lambda_1)-u_h(\lambda_2)\|_1 \leqq C|\lambda_1-\lambda_2|; \quad \forall \lambda_1, \lambda_2 \in [\lambda_0-\mu_1, \lambda_0+\mu_1]. \tag{70}$$

(iii) *When Schemes-1 and -2 introduced in Section 2 are applied to solving $F_h(\lambda, u_h)=0$, they give sequences converging to $u_h(\lambda)$ in H_0^1 for each $\lambda \in [\lambda_0-\mu_1, \lambda_0+\mu_1]$ provided that $\{x^*, y^*\}=\{\lambda^*, u_h^*\}$ of Scheme-1 is in $[\lambda_0-\mu_1, \lambda_0+\mu_1] \times B(u_{h0}, \delta_1, V^h)$ and that $y^{(0)}=u_h^{(0)}$ of Schemes-1 and -2 is in $B(u_{h0}, \delta_1, V^h)$.*

Here, δ_1 can be chosen such that $\delta_1 \leqq \delta_0$, where δ_0 is a positive number appearing in Theorem 3.

REMARK 5. We omit the proof since it is almost clear, but the fact that $\lim_{h\to 0} \|u_h(\lambda_0)-u_0\|_1=0$ follows from (15) and (63).

5. Error analysis of the finite element solutions

In this section, we will perform error analysis of the finite element solutions $\{u_h(\lambda)\}_{h>0}$ as $h \to 0$ under the same assumptions as in Sections 3 and 4. In particular, Ω is a bounded polyhedral domain in $\boldsymbol{R}^n$ ($n=1, 2, 3$). Moreover, we will obtain error estimates in terms of $h>0$ when Ω is convex.

If Ω is a bounded *convex* polyhedral domain in $\boldsymbol{R}^n$ ($n=1, 2, 3$), *Sg*

belongs to $H^2 \cap H_0^1$ for each $g \in L_2$ and

$$\|Sg\|_2 \leqq C\|g\|, \tag{71}$$

where S is the operator defined in (44) and $C>0$ is independent of $g \in L_2$, see Grisvard [6]. Then, by (51) and Nitsche's trick,

$$\|S_h g - Sg\|_q \leqq C^* h^{2-q}\|g\| \qquad (q=0, 1), \tag{72}$$

where S_h is the operator defined in (55) and $C^*>0$ is independent of $h>0$ and $g \in L_2$, see Ciarlet [4].

We obtain the following results for error estimates.

THEOREM 5. *Assume that* [H2], [H3] *and* [H4] *hold. Then there exist positive numbers* h_2, μ_2 *and* δ_2 *such that for* $h \in]0, h_2]$:
(i) $u(\lambda)$ *in Theorem* 3 *and* $u_h(\lambda)$ *in Theorem* 4 *both exist for each* $\lambda \in [\lambda_0 - \mu_2, \lambda_0 + \mu_2]$, (ii) $u(\lambda)$ *and* $u_h(\lambda)$ *are unique respectively in* $B(u_0, \delta_2, H_0^1)$ *and* $B(u_{h0}, \delta_2, V^h)$, *and* (iii) *they satisfy*

$$\|u_h(\lambda) - u(\lambda)\|_1 \leqq C \inf_{v_h \in V^h} \|v_h - u(\lambda)\|_1 \quad \textit{for} \quad |\lambda - \lambda_0| \leqq \mu_2, \tag{73}$$

where $C>0$ *is independent of* h *and* λ. *Moreover,*

$$\lim_{h \to 0} \max_{|\lambda - \lambda_0| \leqq \mu_2} \|u_h(\lambda) - u(\lambda)\|_1 = 0, \tag{74}$$

that is, $\|u_h(\cdot) - u(\cdot)\|_1$ *converges uniformly to* 0. *Furthermore, if* Ω *is a convex bounded polyhedral domain in* $\boldsymbol{R}^n$ $(n=1, 2, 3)$,

$$\|u_h(\lambda) - u(\lambda)\|_q \leqq C^* h^{2-q} \qquad (|\lambda - \lambda_0| \leqq \mu_2;\ q=0, 1), \tag{75}$$

where C^* *is a positive number similar to* C *above.*

REMARK 6. Estimation (75) is best possible with respect to the order of h, as far as the piecewise linear finite element method is concerned.

PROOF. From Theorems 3 and 4, it is obvious that there exist h_2, μ_2 and δ_2 such that (i) and (ii) hold. Hereafter, we will rechoose these positive numbers appropriately smaller whenever we need. We will prove the rest of the theorem in three steps.

1° Let us define $u_h^*(\lambda)$ for $|\lambda - \lambda_0| \leqq \mu_2$ by

$$u_h^*(\lambda) = \lambda S_h f(u(\lambda)),$$

which satisfies, by the relation $u(\lambda) = \lambda S f(u(\lambda))$ and the fundamental

property of the Ritz-Galerkin method [4],

(a) $$\|u_h^*(\lambda)-u(\lambda)\|_1 \leqq C_1 \inf_{v_h \in V^h} \|v_h-u(\lambda)\|_1,$$

where $C_1>0$ is independent of h and λ (we will use similar constants C_2, C_3 and C_4 later). Moreover, when Ω is convex, we have, by (72) and the fact that $f(u(\lambda))$ belongs to L_2 and $\|f(u(\lambda))\|$ is uniformly bounded for $|\lambda-\lambda_0| \leqq \mu_2$,

(b) $$\|u_h^*(\lambda)-u(\lambda)\|_q \leqq C_2 h^{2-q} \qquad (|\lambda-\lambda_0| \leqq \mu_2;\ q=0,\ 1).$$

2° Next, let us evaluate $u_h(\lambda)-u_h^*(\lambda)$. Hereafter, we will denote $u(\lambda)$, $u_h(\lambda)$ and $u_h^*(\lambda)$ simply by u, u_h and u_h^*, respectively. Since $u_h=\lambda S_h f(u_h)$ and $u_h^*=\lambda S_h f(u)$,

(c) $$\begin{aligned} u_h-u_h^*-\lambda_0 S_h \partial f(u_0)(u_h-u_h^*) &= \lambda_0 S_h\{f(u_h)-f(u_h^*)-\partial f(u_0)(u_h-u_h^*)\} \\ &\quad +(\lambda-\lambda_0)S_h(f(u_h)-f(u_h^*))+\lambda S_h(f(u_h^*)-f(u)). \end{aligned}$$

From Lemmas 4 and 5, we have

(d) $$\begin{aligned} &\|f(u_h)-f(u_h^*)-\partial f(u_0)(u_h-u_h^*)\|_{-1} \leqq \varepsilon \|u_h-u_h^*\|_1, \\ &\|f(u_h)-f(u_h^*)\|_{-1} \leqq C_3\|u_h-u_h^*\| \leqq C_3\|u_h-u_h^*\|_1, \\ &\|f(u_h^*)-f(u)\|_{-1} \leqq C_3\|u_h^*-u\|, \end{aligned}$$

where ε is a positive number which can be made arbitrarily close to 0 by rechoosing μ_2 and h_2. From (c), (d) and Lemma 7-(i), we find

$$\|u_h-u_h^*\|_1 \leqq C_4\{(\varepsilon+\mu_2)\|u_h-u_h^*\|_1+\|u_h^*-u\|\}.$$

Thus, by rechoosing μ_2 and h_2 so that $C_4(\varepsilon+\mu_2) \leqq 1/2$, we have

$$\|u_h-u_h^*\|_1 \leqq 2C_4\|u_h^*-u\|.$$

3° Now we can conclude that (73) holds due to (a) and the triangle inequality, and, if Ω is convex, (75) also holds due to (b). To show (74), notice from Theorem 3-(ii) that $u(\cdot)$ is uniformly continuous as an H_0^1-valued function defined on $[\lambda_0-\mu_2, \lambda_0+\mu_2]$. Then it follows from (50) that

$$\lim_{h\to 0} \max_{|\lambda-\lambda_0| \leqq \mu_2} \inf_{v_h \in V^h} \|v_h-u(\lambda)\|_1 = 0,$$

and hence (74) holds true. This completes the proof.

6. Numerical results

To see the validity of the proposed method, let us consider the one-dimensional case of (1) with $\Omega=]-1/2, 1/2[$:

$$-\frac{d^2u}{d\xi^2}=\lambda(u-1)^+ \quad (-1/2<\xi<1/2), \quad u(-1/2)=u(1/2)=0, \tag{76}$$

where ξ is the independent variable. This problem was already considered in [8], [9], and [12], but observations of errors are not fully given yet. Clearly, [H2] holds for the present $f(u)$. In this case, the problem is described by the two-point boundary value problem of a simple ordinary differential equation, and hence we can make full use of the elementary quadrature method. First, we can see that the above problem has a non-trivial solution if and only if $\lambda>\pi^2$. Moreover, for each $\lambda>\pi^2$, such a non-trivial solution is unique, which we denote here by $u_\lambda=u_\lambda(\xi)$ and is explicitly expressed by

$$u_\lambda(\xi)=\begin{cases}1+2\cos(\sqrt{\lambda}\,\xi)/(\sqrt{\lambda}-\pi) & (|\xi|\leqq\pi/(2\sqrt{\lambda})),\\ 1+(\pi-2\sqrt{\lambda}\,|\xi|)/(\sqrt{\lambda}-\pi) & (|\xi|>\pi/(2\sqrt{\lambda})).\end{cases} \tag{77}$$

From this expression, we can see that (i) of [H3] holds true. To check (ii) of [H3], we should first notice that $\partial f(u_\lambda)$ for the above u_λ with $\lambda>\pi^2$ is given by

$$\partial f(u_\lambda)(\xi)=1 \text{ for } |\xi|<\pi/(2\sqrt{\lambda}); \ =0 \text{ for } |\xi|>\pi/(2\sqrt{\lambda}), \tag{78}$$

where the values of $\partial f(u_\lambda)$ at $\xi=\pm\pi/(2\sqrt{\lambda})$ may be specified arbitrarily. Then, again by using the quadrature method, we can see that problem (35) for $\lambda_0=\lambda>\pi^2$ with $u_0=u_\lambda$ has only the trivial solution $w=0$.

To obtain finite element solutions, we partition Ω into N (= positive integer) subintervals (or finite elements) with a common length $h=1/N$. In the examples below, N is always an even positive integer. We employ the piecewise linear polynomials as approximation functions for u, and apply Scheme-2 to solve the discretized nonlinear equations based on the finite element method. The value of $\partial f(t)$ at $t=1$ is taken as 1/2 according to (30). In the present case, the integrations to derive discrete equations can be exactly evaluated: in particular, we need not use numerical integrations for the nonlinear terms $f(u)$ and $\partial f(u)$, see Kikuchi-Nakazato-Ushijima [12]. As a starting approximation $u_h^{(0)}$ of u_h, we use

$$u_h^{(0)}(\xi)=\beta(1-2|\xi|) \quad \text{for} \quad \xi\in\Omega, \tag{79}$$

where β is a parameter such that $\beta>1$. The employed stopping condition for iteration is

$$\|u_h^{(k)}-u_h^{(k-1)}\|/\|u_h^{(k)}\|\leqq 10^{-10}, \tag{80}$$

where $u_h^{(k)}$ is the k-th ($k\geqq 0$) approximation of the iteration process. For completeness, we describe Scheme-2 applied to the present finite element analysis: fix $u_h^{(0)}\in V^h$ and obtain $\{u_h^{(k)}\}_{k=1}^{\infty}$ in V^h recursively by

$$\begin{aligned}\langle u_h^{(k)}, v_h\rangle-\lambda(\partial f(u_h^{(k-1)})u_h^{(k)}, v_h)\\ =\lambda(f(u_h^{(k-1)}), v_h)-\lambda(\partial f(u_h^{(k-1)})u_h^{(k-1)}, v_h);\quad \forall v_h\in V^h.\end{aligned} \tag{81}$$

All the computations are performed in double precision arithmetic. Hereafter, we will denote the computed nontrivial finite element solutions by $u_{h\lambda}=u_{h\lambda}(\xi)$.

We obtain numerical solutions for various values of λ and N. The values of λ are taken as $\lambda=10$, 15, 20, 50, 100, while those of N are taken as 10, 20, 40, 80, 100, 200, 1000. The parameter β is usually taken as 2, but, for $\lambda=10$, it is taken as 10 so that $u_h^{(k)}$ does not converge to the trivial solution $u_h=0$. When λ is close to π^2, say $\lambda=10$, the linear operator appearing in the left-hand side of (36) becomes almost singular, and hence the present iteration scheme is not sufficiently stable. In such cases, the iteration schemes proposed by

Table 1-(1). Numerical results-1 ($\lambda=10$, 15)
(k=number of iterations)

λ	10			15		
β	10			2		
N	k	$u_{h\lambda}(0)$	$\|u_{h\lambda}-u_\lambda\|$	k	$u_{h\lambda}(0)$	$\|u_{h\lambda}-u_\lambda\|$
10	4	262.253	1.148E+2	5	3.82687	4.522E−2
20	4	115.941	1.274E+1	5	3.75725	1.115E−2
40	4	101.694	2.796E+0	4	3.74018	2.780E−3
80	4	98.6602	6.783E−1	4	3.73593	6.943E−4
100	4	98.3082	4.325E−1	4	3.73542	4.443E−4
200	4	97.8426	1.076E−1	4	3.73474	1.111E−4
1000	4	97.6946	4.299E−3	4	3.73453	4.443E−6
$u_\lambda(0)$		97.6884			3.73452	

Kikuchi [8, 9] Kikuchi-Nakazato-Ushijima [12], and Rappaz [14] may be expected to be stable and efficient.

In Table 1, we give numerical results for the numbers of iterations required for convergence, L_2-errors $\|u_{h\lambda}-u_\lambda\|$ of the finite element solutions, and the values of $u_{h\lambda}(0)$. Here, the number of iterations means the value of k where (80) is first satisfied, and L_2-errors are calculated by the elementwise use of the three-point Gauss quadrature formula. From (77), the exact value of $u_\lambda(0)$ is given by $u_\lambda(0)=1+2/(\sqrt{\lambda}-\pi)$

Table 1-(2). Numerical results-2 ($\lambda=20$, 50)
(k=number of iterations)

λ	20			50		
β	2			2		
N	k	$u_{h\lambda}(0)$	$\|u_{h\lambda}-u_\lambda\|$	k	$u_{h\lambda}(0)$	$\|u_{h\lambda}-u_\lambda\|$
10	5	2.54550	1.825E−2	4	1.52815	9.730E−3
20	4	2.51371	4.563E−3	4	1.51373	2.486E−3
40	4	2.50578	1.140E−3	3	1.51017	6.246E−4
80	4	2.50380	2.849E−4	3	1.50927	1.565E−4
100	4	2.50357	1.823E−4	3	1.50916	1.002E−4
200	4	2.50325	4.559E−5	3	1.50902	2.505E−5
1000	4	2.50315	1.823E−6	3	1.50898	1.002E−6
$u_\lambda(0)$	2.50315			1.50897		

Table 1-(3). Numerical results-3 ($\lambda=100$)
(k=number of iterations)

λ	100		
β	2		
N	k	$u_{h\lambda}(0)$	$\|u_{h\lambda}-u_\lambda\|$
10	4	1.30859	1.049E−2
20	4	1.29607	2.760E−3
40	4	1.29272	7.004E−4
80	3	1.29189	1.757E−4
100	3	1.29179	1.125E−4
200	3	1.29166	2.814E−5
1000	3	1.29161	1.126E−6
$u_\lambda(0)$	1.29161		

for $\lambda > \pi^2$. We can see that the convergence of the iteration process is fairly rapid, and the present method may be effectively used for practical problems on a much larger scale. In fact, this method is already applied to equilibrium analysis of MHD plasmas with bifurcation phenomena by Kikuchi-Aizawa [10, 11]. It is also clear that the L_2-errors of the finite element solutions are in quite good proportion to h^2 when h is sufficiently small. This observation is consistent with our theoretical result given in Theorem 5 that $\|u_{h\lambda} - u_\lambda\|$ is of the order of h^2, where Ω in this case is of course convex.

7. Concluding remarks

We have considered Newton-like methods and an implicit function theorem for nondifferentiable (but almost differentiable) nonlinear problems, and apply them to finite element approximations to a semilinear elliptic eigenvalue problem related to MHD equilibria and some other physical phenomena. It is shown that the Newton-like methods are in fact applicable to the discrete problems based on the finite element method under some assumptions on the continuous problem and the finite element spaces. We also give some simple numerical results.

In actual finite element computations, the integrations related to the nonlinear terms must be carried out numerically except when $f(u)$ and the finite element space are very simple. In [12], some observations are given for $f(u) = (u-1)^+$. Since analysis of the effects of numerical integrations is very important, the author is now attacking this problem in more general cases. It is also important to analyze problems with more general types of nonlinear terms.

Acknowledgements.

The author is very grateful to Professor T. Yamamoto of Ehime University for his valuable comments on the Newton method. Above all, he informed the author of the paper by Keller [7]. This work is partially supported by the Grant-in-Aids for Scientific Research from the Ministry of Education.

References

[1] Aubin, J.-P., Approximation of Elliptic Boundary-value Problems, Wiley-Intersciences, New York-London-Sydney-Toronto, 1972.

[2] Berestycki, H., Fernandez Cara, E. and R. Glowinski, A numerical study of some

questions in vortex rings theory, RAIRO Anal. Numér. **18** (1984), 7-85.
[3] Caloz, G., A free boundary problem related to axisymmetric MHD equilibria: existence and numerical approximation of solutions, preprint, Département de Mathématiques, École Polytechnique Fédérale de Lausanne, Suisse, 1984.
[4] Ciarlet, P. G., The Finite Element Method for Elliptic Problems, North-Holland Publishing Company, Amsterdam-New York-Oxford, 1978.
[5] Girault, V. and P.-A. Raviart, An analysis of upwind schemes for the Navier-Stokes equations, SIAM J. Numer. Anal. **19** (1982), 312-333.
[6] Grisvard, P., Elliptic Problems in Nonsmooth Domains, Pitman Publ. Ltd., Boston-London-Melbourne, 1985.
[7] Keller, H. B., Newton's method under mild differentiability conditions, J. Comput. System Sci. **4** (1970), 15-28.
[8] Kikuchi, F., Construction of a path of MHD equilibrium solutions by an iterative method, ISAS (Institute of Space and Aeronautical Science, University of Tokyo) Report **44** (1979), 97-111.
[9] Kikuchi, F., An iteration scheme for a nonlinear eigenvalue problem, Theoretical Appl. Mech. **29** (1981), 319-333.
[10] Kikuchi, F. and T. Aizawa, Finite element analysis of MHD equilibria, Theoretical Appl. Mech. **30** (1981), 513-527.
[11] Kikuchi, F. and T. Aizawa, Finite element analysis of ideal MHD plasmas in torus regions, In: Finite Elements in Fluids (R. H. Gallagher, J. T. Oden, O. C. Zienkiewicz, T. Kawai, M. Kawahara, ed.), Vol. 5, John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore, 1984, pp. 311-323.
[12] Kikuchi, F., Nakazato, K. and T. Ushijima, Finite element approximation of a nonlinear eigenvalue problem related to MHD equilibria, Japan J. Appl. Math. **1** (1984), 369-403.
[13] Liusternik, L. A. and V. J. Sobolev, Elements of Functional Analysis, Frederick Unger Publ. Co., New York, 1961.
[14] Rappaz, J., Approximation of a nondifferentiable nonlinear problem related to MHD equilibria, Numer. Math. **45** (1984), 117-133.
[15] Temam, R., A nonlinear eigenvalue problem: the shape of equilibrium of a confined plasma, Arch. Rational Mech. Anal. **60** (1975), 51-73.
[16] Yosida, K., Functional Analysis, 2nd ed., Springer Verlag, Berlin-Heidelberg-New York, 1968.

(Received February 28, 1987)

Department of Mathematics
College of Arts and Sciences
University of Tokyo
Komaba, Meguro-ku, Tokyo
153 Japan