Journal of Mathematical Analysis and Applications **234,** 514–528 (1999)
Article ID jmaa.1998.6364, available online at http://www.idealibrary.com on IDEAL®

Reliable Solution for a 1D Quasilinear Elliptic Equation with Uncertain Coefficients

Jan Chleboun*

Submitted by Hans-Görg Roos

Received June 27, 1997

The solution of a quasilinear elliptic state equation depends on the coefficient function belonging to an admissible set. The solution is evaluated by a cost functional the value of which is to be maximized over the admissible set, i.e., the reliable (safe) solution is searched for. Due to the nature of the equation, the Kirchhoff transformation can be applied to obtain both the existence of the true state solution and a cost sensitivity formula. In many cases, the latter makes it possible to determine the reliable solution immediately. The problem is approximated by means of the finite element method, and some convergence results are proven. Numerical examples illustrate the theory which can be directly generalized to spatial problems.

Key Words: quasilinear elliptic equation; reliable solution; uncertain data; Kirchhoff transformation.

1. INTRODUCTION: MAXIMIZATION PROBLEM

The notion *reliable solution* was introduced in [2] to label "the worst" case among a set of possible solutions, where possibility is induced by uncertain input data, and the degree of badness is measured by a cost functional. The highest local temperature of a heated body, the conductivity coefficients of which are not known exactly, can serve as an industrial example. Another problem is treated in [4].

In this pilot study, we apply the theory presented in articles [2, 3, 5] to a class of particular problems analyzed theoretically and, finally, solved

* E-mail: chleb@math.cas.cz

0022-247X/99 $30.00

numerically. In detail, we deal with the following equation for an unknown function $u \in C^1(\overline{\Omega}) \cap C^2(\Omega)$, $\Omega = (0, 1)$,

$$-(a(u)u')' = f \quad \text{in } \Omega, \tag{1.1}$$

$$u = \overline{u} \quad \text{on } \Gamma_1, \tag{1.2}$$

$$a(u)u' = g \quad \text{on } \Gamma_2, \tag{1.3}$$

where $\Gamma_1 = \{0\}$ or $\Gamma_1 = \{0, 1\}$, $\Gamma_2 = \{0, 1\} \setminus \Gamma_1$, $a \in C^{(0),1}(\mathbb{R})$ (Lipschitz continuous functions on $\mathbb{R}$), $\overline{u}$ is a constant, and f is a right-hand side function. The prime stands for $\mathrm{d}/\mathrm{d}x$.

Let the problem (1.1)–(1.3) be reformulated into the weak form: Find $u \in H^1(\Omega)$ such that

$$u - \overline{u} \in V = \{v \in H^1(\Omega): v|_{\Gamma_1} = 0\}, \tag{1.4}$$

$$(a(u)u', v')_{0,\Omega} = \langle f, v\rangle_\Omega + \langle g, v\rangle_{\Gamma_2} \qquad \forall v \in V, \tag{1.5}$$

where $H^1(\Omega)$ is the Sobolev space (with the norm $\|\cdot\|_{1,\Omega}$) of functions continuous on $\overline{\Omega}$ and with a square integrable generalized derivative on Ω, $f \in V^*$ (dual space to V). The symbols $(\cdot,\cdot)_{0,\Omega}$ and $\langle\cdot,\cdot\rangle$ stand for the inner product in the space $L^2(\Omega)$ and the duality pairing, respectively. In detail, the rightmost term is equal to either $g(1)v(1)$ or zero, with the latter holding if $\Gamma_2 = \emptyset$. Let us be reminded that $H^1(\Omega)$ is continuously embedded into the space $C(\overline{\Omega})$ of continuous functions on $\overline{\Omega}$ provided with the common norm $\|\cdot\|_{0,\infty,\Omega}$. The embedding is even compact.

The uncertainty mentioned in the first paragraph concerns the function a which belongs to the set

$$\mathcal{U}_{\mathrm{ad}} = \{a \in \mathcal{U}^0_{\mathrm{ad}}(C_L): a_{\min}(t) \le a(t) \le a_{\max}(t) \qquad \forall t \in \mathbb{R}\}$$

defined with the aid of

$$\begin{aligned}\mathcal{U}^0_{\mathrm{ad}}(C_L) = \{a \in C^{(0),1}(\mathbb{R}): |\mathrm{d}a/\mathrm{d}t| \le C_L \text{ a.e.},\\ a(t) = a(T_\mathrm{l}) \quad \text{for } t < T_\mathrm{l},\\ a(t) = a(T_\mathrm{r}) \quad \text{for } t > T_\mathrm{r}\},\end{aligned}$$

$$a_{\min}, a_{\max} \in \widetilde{\mathcal{U}}^0_{\mathrm{ad}}(C_L) = \{a \in \mathcal{U}^0_{\mathrm{ad}}(C_L): 0 < \tilde{a}_{\min} \le a(t) \le \tilde{a}_{\max} \quad \forall t \in \mathbb{R}\},$$

and given constants $C_L > 0$, $\tilde{a}_{\min}$, $\tilde{a}_{\max}$, T_l, T_r, $T_\mathrm{l} < T_\mathrm{r}$.

Next, we introduce intervals $G_j \subset \Omega$, $j = 1, \ldots, J$, and functionals

$$\begin{aligned}\Phi_j(v) &= (\operatorname{meas} G_j)^{-1} \int_{G_j} v \,\mathrm{d}x,\\ \Phi(v) &= \max_{1\le j\le J} \Phi_j(v), \qquad v \in L^2(\Omega).\end{aligned} \tag{1.6}$$

Denoting by $u(a)$ the solution of (1.4)–(1.5) and defining $\Psi_j(a) = \Phi_j(u(a))$, $\Psi(a) = \Phi(u(a))$, we set the Maximization Problem: Find

$$a^0 = \arg\max_{a \in \mathcal{U}_{\mathrm{ad}}} \Psi(a). \tag{1.7}$$

Remark 1.1. To give the Maximization Problem a physical meaning, we can interpret $u(a)$ as a temperature distribution in Ω. Then Φ evaluates the maximum of its value averaged over chosen intervals G_j, and we search for the coefficient a^0 inducing the maximum of Φ. Since $H^1(\Omega)$ is embedded into $C(\overline{\Omega})$ we could consider pointwise values of $u(a)$ as well. ∎

The problem (1.1)–(1.3) is covered by the article [5] but, in contrast to a more general class of equations studied there, it allows the use of the Kirchhoff transformation. This tool will be found helpful in solving (1.7).

2. KIRCHHOFF TRANSFORMATION

Let us suppose u is the solution of the problem (1.4)–(1.5), the variable a is omitted for a while. We can define a function $z\colon \overline{\Omega} \to \mathbb{R}$ by the equality (Kirchhoff transformation)

$$z(x) = \int_{\bar{u}}^{u(x)} a(t)\,\mathrm{d}t, \qquad x \in \overline{\Omega}. \tag{2.1}$$

The transformation is invertible because a is a positive function. Observing that

$$z' = a(u)u', \tag{2.2}$$

we can transform the problem (1.4) and (1.5) into the following *linear* elliptic equation: Find $z \in V$ such that

$$(z', v')_{0,\Omega} = \langle f, v\rangle_{\Omega} + \langle g, v\rangle_{\Gamma_2} \qquad \forall v \in V. \tag{2.3}$$

Problem (2.3) has a unique solution z which does not depend on $a \in \mathcal{U}_{\mathrm{ad}}$. The function z yields the solution u of (1.4)–(1.5):

THEOREM 2.1. *Suppose $a \in \widetilde{\mathcal{U}}^0_{\mathrm{ad}}(C_L)$ and $z \in V$ solves* (2.3). *Let the function u be defined by the equality* (2.1). *Then $u'(x)$ is defined for all $x \in \overline{\Omega}$, where $z'(x)$ exists, and u is the unique solution of the equations* (1.4)–(1.5) *with the coefficient a.*

Proof. The function u is uniquely determined by (2.1) because $a(t) > 0$ for any $t \in \mathbb{R}$. The continuity of u is the consequence of the inequality

$$\widetilde{a}_{\min}|u(x_1) - u(x_2)| \le \left| \int_{u(x_2)}^{u(x_1)} a(t)\, \mathrm{d}t \right| = |z(x_1) - z(x_2)|, \qquad x_1, x_2 \in \overline{\Omega}.$$

To prove differentiability, we introduce

$$b_1(x,t) = \begin{cases} \min\limits_{s \in [u(x), u(x+t)]} a(s) & \text{if } u(x) \le u(x+t), \\ \max\limits_{s \in [u(x+t), u(x)]} a(s) & \text{if } u(x) > u(x+t), \end{cases}$$

$$b_2(x,t) = \begin{cases} \max\limits_{s \in [u(x), u(x+t)]} a(s) & \text{if } u(x) \le u(x+t), \\ \min\limits_{s \in [u(x+t), u(x)]} a(s) & \text{if } u(x) > u(x+t), \end{cases}$$

$$q(x,t) = (u(x+t) - u(x))/t, \qquad x, x+t \in \overline{\Omega}.$$

The equality (2.1), the continuity of a and u, and the inequalities

$$q(x,t) b_1(x,t) \le \frac{1}{t} \int_{u(x)}^{u(x+t)} a(s)\, \mathrm{d}s \le q(x,t) b_2(x,t)$$

imply

$$\limsup_{t \to 0} q(x,t) \le \frac{z'(x)}{a(u(x))} \le \liminf_{t \to 0} q(x,t),$$

if $z'(x)$ exists at $x \in \overline{\Omega}$. It means $u'(x) = z'(x)/a(u(x))$ for a.a. $x \in \overline{\Omega}$ as $z'(x)$ is defined for a.a. $x \in \overline{\Omega}$. Moreover, $u' \in L^2(\Omega)$ and u solves (1.4)–(1.5) due to (2.2).

By virtue of (2.1) and the positiveness of a, the solution u is unique. This also follows from [5, Theorem 3.2]. ∎

LEMMA 2.1. *Let $u(a)$ and z be the solution of the problem* (1.4) *and* (1.5) *and* (2.3), *respectively. Then for all $x \in \overline{\Omega}$* $\operatorname{sign}(u(a)(x) - \overline{u}) = \operatorname{sign}(z(x))$ *regardless of $a \in \widetilde{\mathcal{U}}^0_{\mathrm{ad}}(C_L)$ and $C_L > 0$. Moreover,* $\operatorname{sign}(u'(a)(x)) = \operatorname{sign}(z'(x))$ *if the derivatives exist at $x \in \overline{\Omega}$.*

Proof. Since a is positive, the assertions follow from (2.1) and (2.2). ∎

Remark 2.1. Lemma 2.1 and similar results for higher derivatives which are available under smoothness assumptions help to graph $u(a)$ on the basis of solving the easy problem (2.3). ∎

LEMMA 2.2. *Let $\widetilde{\varepsilon} > 0$ be a parameter such that $C_0 \equiv \widetilde{a}_{\min} - \widetilde{\varepsilon} > 0$, and let $a_0 \in \widetilde{\mathcal{U}}^0_{\mathrm{ad}}(C_L)$, $a_\varepsilon \in \mathcal{U}^0_{\mathrm{ad}}(\widehat{C})$, $\widehat{C} > 0$, be two functions, $\|a_\varepsilon\|_{0,\infty,\mathbb{R}} \leq \widetilde{\varepsilon}$. If $u(a_0)$ and $u(a)$ are the solutions of the problem* (1.4)–(1.5) *with a_0 and $a = a_0 + a_\varepsilon$, respectively, then*

$$\|u(a_0) - u(a)\|_{1,\Omega} \leq C\|a_\varepsilon\|_{0,\infty,\mathbb{R}},$$

where the constant $C > 0$ does not depend on a_ε, a_0, and $\widehat{C}$.

Proof. Let us set $u \equiv u(a)$, $u_0 \equiv u(a_0)$. The Kirchhoff transform applied to both u_0 and u results in the unique solution z of the problem (2.3). For any $x \in \overline{\Omega}$, we have

$$0 = \int_{\overline{u}}^{u(x)} a(t)\,\mathrm{d}t - \int_{\overline{u}}^{u_0(x)} a_0(t)\,\mathrm{d}t = \int_{u_0(x)}^{u(x)} a_0(t)\,\mathrm{d}t + \int_{\overline{u}}^{u(x)} a_\varepsilon(t)\,\mathrm{d}t.$$

Thus

$$\begin{aligned}\widetilde{a}_{\min}|u(x) - u_0(x)| &\leq \left|\int_{u_0(x)}^{u(x)} a_0(t)\,\mathrm{d}t\right| = \left|\int_{\overline{u}}^{u(x)} \frac{a(t)}{a(t)} a_\varepsilon(t)\,\mathrm{d}t\right| \\ &\leq \frac{\varepsilon}{C_0}\left|\int_{\overline{u}}^{u(x)} a(t)\,\mathrm{d}t\right| \leq \frac{\varepsilon}{C_0}\|z\|_{0,\infty,\Omega}, \qquad (2.4)\end{aligned}$$

where $\varepsilon = \|a_\varepsilon\|_{0,\infty,\mathbb{R}}$.

We focus on u', u_0' now. To this end, we define the function $b(t_1, t_2) = a_0(t_1) - a_0(t_2)$, $t_1, t_2 \in \mathbb{R}$, complying with the obvious inequality

$$|b(t_1, t_2)| \leq C_L|t_1 - t_2|. \qquad (2.5)$$

We infer from (1.5) that for any $v \in V$

$$\begin{aligned}0 &= (a_0(u)u' - a_0(u_0)u_0', v')_{0,\Omega} + (a_\varepsilon(u)u', v')_{0,\Omega} \\ &= \big((a_0(u_0) + b(u, u_0))u' - a_0(u_0)u_0', v'\big)_{0,\Omega} \\ &\quad + (a_\varepsilon(u)u', v')_{0,\Omega} \\ &= (a_0(u_0)(u' - u_0'), v')_{0,\Omega} + (b(u, u_0)u', v')_{0,\Omega} \\ &\quad + (a_\varepsilon(u)u', v')_{0,\Omega}. \qquad (2.6)\end{aligned}$$

On the basis of (2.6) where $v = u_0 - u$ is considered, we get

$$\begin{aligned}\widetilde{a}_{\min}\|u_0' - u'\|^2_{0,\Omega} &\leq (a_0(u_0)v', v')_{0,\Omega} \\ &= (b(u, u_0)u', v')_{0,\Omega} + (a_\varepsilon(u)u', v')_{0,\Omega} \\ &\leq (C_L\|v\|_{0,\infty,\Omega} + \varepsilon)\|u'\|_{0,\Omega}\|v'\|_{0,\Omega} \\ &\leq C_1\varepsilon\|u'\|_{0,\Omega}\|u_0' - u'\|_{0,\Omega}. \qquad (2.7)\end{aligned}$$

To derive (2.7), the definition of $\widetilde{\mathcal{U}}^0_{\text{ad}}(C_L)$ and the estimates (2.5) and (2.4) were employed.

The assertion follows from (2.4) and (2.7) provided $\|u'\|_{0,\Omega}$ can be bounded independently of a_0, a_ε, and $\widehat{C}$ which is true by virtue of (2.2). ∎

THEOREM 2.2. *Problem* (1.7) *has at least one solution.*

Proof. The solution $u(a)$ depends continuously on $a \in \mathcal{U}_{\text{ad}}$ (Lemma 2.2), the functionals Φ and Ψ are continuous and, by the Arzelà–Ascoli theorem [8], the set $\mathcal{U}_{\text{ad}}$ is compact in $C(\mathbb{R})$ (see [2, Lemma 1.2], for example). ∎

3. APPROXIMATION

In this section, we follow ideas of [2, 5] to address convergence and uniqueness questions. This is why we give only sketches of proofs in some instances. As opposed to Sections 2 and 4, the invariability of $\bar{u}$ is not important in this section.

Let the interval Ω be uniformly subdivided into N subintervals e_i of the length $h = 1/N$. To approximate the space V, we introduce its subspace

$$V_h = \{v_h \in V\colon v_h|_{e_i} \in P_1(e_i),\ i = 1, \dots, N\},$$

where $P_1(e_i)$ denotes linear polynomials on e_i.

Instead of solving (1.4)–(1.5) we search for a Galerkin approximation u_h such that

$$u_h - \bar{u} \in V_h, \tag{3.1}$$

$$\big(a(u_h)u_h', v_h'\big)_{0,\Omega} = \langle f, v_h\rangle_\Omega + \langle g, v_h\rangle_{\Gamma_2} \quad \forall v_h \in V_h. \tag{3.2}$$

The following theorem is, in fact, a combination of [2, Appendix] and [5, Theorem 2.6]. However, we avoid assumptions bound with the parameter h (see [5, Theorem 2.6(ii)]).

Let us define $\widehat{a}_{\min} = \min_{t\in\mathbb{R}} a_{\min}(t)$. Let us recall the Friedrichs inequality (see [7], for example) $\|v\|_{0,\Omega} \le C_F\|v'\|_{0,\Omega}$ and the embedding inequality $\|v\|_{0,\infty,\Omega} \le C_0\|v\|_{1,\Omega}$ valid for all $v \in V$; $C_F, C_0 > 0$.

It can be shown (see the proof of Theorem 3.1) that the Galerkin solution is among functions the first seminorm of which is less or equal to a positive constant C_B.

THEOREM 3.1. *Let* $a \in \widetilde{\mathcal{U}}^0_{\text{ad}}(C_L)$ *be arbitrary. Then a Galerkin approximation* u_h *to the problem* (3.1)–(3.2) *exists. The function* u_h *is unique if at least one of the following conditions takes place*:

(i) $\widehat{a}_{\min}^{-1} C_0 C_B C_L \sqrt{1 + C_F^2} < 1$;

(ii) $\|u_h'\|_{0,\infty,\Omega} < \widehat{a}_{\min}/(C_F C_L)$.

In the case (i), *the Galerkin solution* u_h *can be calculated via the Katchanov* (*secant moduli*) *method*: *Let* $y_0 \in V_h$ *be arbitrary. If* $y_k \in V_h$ *is known, let* y_{k+1} *be defined by the relation*

$$(a(\overline{u}+y_k)(\overline{u}+y_{k+1})', v')_{0,\Omega} = \langle f, v\rangle_\Omega + \langle g, v\rangle_{\Gamma_2} \qquad \forall v \in V_h.$$

Then $\|u_h - (\overline{u}+y_k)\|_{1,\Omega} \to 0$ *as* $k \to \infty$.

Proof. To prove the existence of the solution u_h we follow ideas of the proof of [5, Theorem 2.6(i)].

We define a mapping S: $V_h \to V_h$ by the relation

$$\big(a(\overline{u}+y)(\overline{u}+Sy)', v'\big)_{0,\Omega} = \langle f, v\rangle_\Omega + \langle g, v\rangle_{\Gamma_2} \qquad \forall v \in V_h.$$

By virtue of the Lax–Milgram lemma (see, e.g., [7]), S is uniquely defined. Taking $v = Sy$, we can easily show that a constant C_B independent of a exists such that $C_B \geq \|(Sy)'\|_{0,\Omega}$ $\forall y \in V_h$.

Let us consider arbitrary $y, z \in V_h$ and denote $v = Sy - Sz \in V_h$. We get

$$\begin{aligned}
\widehat{a}_{\min}\|v'\|^2_{0,\Omega} &\leq (a(\overline{u}+y)v', v')_{0,\Omega}\\
&= 1\big|(a(\overline{u}+z)(\overline{u}+Sz)', v')_{0,\Omega} - (a(\overline{u}+y)(\overline{u}+Sz)', v')_{0,\Omega}\big|\\
&= \big|\big((a(\overline{u}+z) - a(\overline{u}+y))(\overline{u}+Sz)', v'\big)_{0,\Omega}\big|\\
&\leq C_L\|z-y\|_{0,\infty,\Omega}\|(\overline{u}+Sz)'\|_{0,\Omega}\|v'\|_{0,\Omega}\\
&\leq C_L C_0(\|\overline{u}'\|_{0,\Omega} + C_B)\|v'\|_{0,\Omega}\|z-y\|_{1,\Omega}.
\end{aligned}$$

Cancelling $\|v'\|_{0,\Omega}$ on both sides of the inequalities and using the Friedrichs inequality and $\overline{u}' = 0$ ($\overline{u}$ is a constant), we infer

$$\|Sy - Sz\|_{1,\Omega} = \|v\|_{1,\Omega} \leq \widehat{a}^{-1}_{\min} C_0 C_L \sqrt{1 + C_F^2}\, C_B\|z-y\|_{1,\Omega}.$$

The mapping S is Lipschitz continuous and allows application of the Brower theorem [1] which gives the existence of $y = Sy$, i.e., of $u_h = \overline{u} + y$. Under the assumption (i), the mapping S is contractive, the fixed point is unique and can be gained via the Katchanov method (details in [5]).

The condition (ii) is calculated from bounds put on $\|u_h\|_{0,\infty,\Omega}$ and $\|u_h\|_{1,\infty,\Omega}$ to prove the uniqueness of u_h in [2, Appendix]. ∎

To approximate the admissible set $\mathcal{U}_{\mathrm{ad}}$, we introduce equally spaced points T_i, $i = 1, \ldots, M$, into the interval $[T_{\mathrm{l}}, T_{\mathrm{r}}]$, $T_1 \equiv T_{\mathrm{l}}$, $T_M \equiv T_{\mathrm{r}}$, and then define the set

$$\begin{aligned}
\mathcal{U}^M_{\mathrm{ad}} = \big\{a \in \mathcal{U}^0_{\mathrm{ad}}(C_L) :\ & a_{\min}(T_i) \leq a(T_i) \leq a_{\max}(T_i),\ i = 1, \ldots, M,\\
& a|_{[T_i, T_{i+1}]} \in P_1([T_i, T_{i+1}]), i = 1, \ldots, M-1\big\},
\end{aligned}$$

which can be identified with the set of vectors

$$\widehat{\mathcal{U}}_{\mathrm{ad}}^{M} = \left\{ \alpha \in \mathbb{R}^M \colon \exists a \in \mathcal{U}_{\mathrm{ad}}^{M} \quad \alpha = (\alpha_1, \ldots, \alpha_M) = (a(T_1), \ldots, a(T_M)) \right\}. \tag{3.3}$$

Though $\mathcal{U}_{\mathrm{ad}}^{M} \subset \widetilde{\mathcal{U}}_{\mathrm{ad}}^{0}(C_L)$, in general $\mathcal{U}_{\mathrm{ad}}^{M} \not\subset \mathcal{U}_{\mathrm{ad}}$, however.

Let us *suppose* that for any $a \in \mathcal{U}_{\mathrm{ad}}^{M}$ there is *a unique solution of the problem* (3.1)–(3.2). This assumption is backed by some industrial measurements showing that C_L can often be a small number (see [5, Sect. 4]). If we define $\Psi^h(a) = \Phi(u_h(a))$ then the Approximate Maximization Problem reads: Find

$$a_h^{0M} = \arg\max_{a \in \mathcal{U}_{\mathrm{ad}}^{M}} \Psi^h(a), \tag{3.4}$$

where $u_h(a)$ solves the problem (3.1)–(3.2) with the coefficient $a \in \mathcal{U}_{\mathrm{ad}}^{M}$.

THEOREM 3.2. *Problem* (3.4) *has at least one solution*.

Proof. Though we cannot utilize the Kirchhoff transform as in the proof of Lemma 2.2, the continuity of Ψ^h can be proved by other means, cf. [2, Theorem 2.1, Theorem 1.1, Proposition 1.2]. Thus (3.4) and (3.3) lead to the maximization of a continuous function over a compact set. ∎

Remark 3.1. If the uniqueness of the Galerkin solution u_h is not supposed, problem (3.4) can be properly modified, see [2, Lemma 2.2], and, under an additional assumption (see [2, Theorem 2.1]), the rest of this section remains valid. ∎

Two discretization parameters, i.e., h and M, are used in the definition of the Approximate Maximization Problem (3.4). Let us bind them together supposing M depends on h in such a way that the sequence $\{M(h)\}$, $h \to 0+$, is nondecreasing.

LEMMA 3.1. *Assume that a sequence* $\{a_{M(h)}\}$, $h \to 0+$, $a_{M(h)} \in \mathcal{U}_{\mathrm{ad}}^{M(h)}$, *converges in* $C(\mathbb{R})$ *to a function* a. *Let* $\{u_{M(h)}\}$ *be the corresponding sequence of the solutions of the problem* (3.1)–(3.2). *Then* $a \in \mathcal{U}_{\mathrm{ad}}^{\widehat{M}}$ *for some* $\widehat{M}$ (*if* $M(h)$ *is bounded*) *or* $a \in \mathcal{U}_{\mathrm{ad}}$. *Moreover,*

$$\|u_{M(h)} - u\|_{1,\Omega} \to 0, \tag{3.5}$$

where u, $u - \overline{u} \in V$, *solves* (1.4)–(1.5) *with the coefficient* a.

Proof. We can follow basic ideas presented in the proof of [5, Theorem 2.9] and thus reduce some parts of our proof to a sketch. Since the cited proof deals with a convergence of u_h for a fixed we will pay more attention to those parts where substantial modifications have to be done to treat the fact that we deal with the sequence $\{a_{M(h)}\}$.

If $\{M(h)\}$ is bounded then a number $\widehat{M}$ exists such that $a \in \mathcal{U}_{\text{ad}}^{\widehat{M}}$, else $a \in \mathcal{U}_{\text{ad}}$ as proven in [2, Lemma 3.2].

As in [5], the sequence $\{u_{M(h)}\}$ is bounded in $H^1(\Omega)$ as we can infer from (3.1)–(3.2) and the properties of $\mathcal{U}_{\text{ad}}^{M(h)}$. Thus a subsequence, denoted again by $\{a_{M(h)}\}$, and a function $w \in H^1(\Omega)$ exist such that

$$u_{M(h)} \rightharpoonup w \text{ (weakly)} \quad \text{in } H^1(\Omega). \tag{3.6}$$

Moreover, $w \in \overline{u} + V$ because $\overline{u} + V$ is convex and closed in $H^1(\Omega)$ and $u_{M(h)} \in \overline{u} + V$.

The next step is to prove $w \equiv u$. To this end we choose an arbitrary $v \in V \cap C^\infty(\overline{\Omega})$ ($C^\infty(\overline{\Omega})$ stands for smooth functions on $\overline{\Omega}$) and consider a sequence $\{v_h\}$, $v_h \in V_h$, $v_h \to v$ in $H^1(\Omega)$ as $h \to 0+$.

Let us estimate the following value with the aid of (3.2):

$$\begin{aligned}
&\left|(a(w)w', v')_{0,\Omega} - \langle f, v\rangle_\Omega - \langle g, v\rangle_{\Gamma_2}\right| \\
&\quad \le \left|(a(w)w', v')_{0,\Omega} - (a_{M(h)}(u_{M(h)})u'_{M(h)}, v'_h)_{0,\Omega}\right| \\
&\qquad + \left|\langle f, v_h - v\rangle_\Omega\right| + \left|\langle g, v_h - v\rangle_{\Gamma_2}\right| = I_1(h) + I_2(h) + I_3(h).
\end{aligned}$$

Further,

$$\begin{aligned}
I_1(h) &\le \left|(a(w)w', v')_{0,\Omega} - (a(w)u'_{M(h)}, v')_{0,\Omega}\right| \\
&\quad + \left|(a(w)u'_{M(h)}, v')_{0,\Omega} - (a(u_{M(h)})u'_{M(h)}, v')_{0,\Omega}\right| \\
&\quad + \left|(a(u_{M(h)})u'_{M(h)}, v')_{0,\Omega} - (a_{M(h)}(u_{M(h)})u'_{M(h)}, v'_h)_{0,\Omega}\right| \\
&= I_{11}(h) + I_{12}(h) + I_{13}(h).
\end{aligned}$$

By (3.6), $\lim_{h\to 0+} I_2(h) = 0 = \lim_{h\to 0+} I_3(h) = \lim_{h\to 0+} I_{11}(h)$.

We also have $\lim_{h\to 0+} I_{12}(h) = 0$ due to the boundedness of $\{u_{M(h)}\}$, (3.6), and the compact embedding $H^1(\Omega) \subset\subset C(\overline{\Omega})$. Finally, $\lim_{h\to 0+} I_{13}(h) = 0$ as $\{v_h\}$ converges strongly. We get (1.5) (with w substituted for u) for any $v \in V \cap C^\infty(\overline{\Omega})$. The density argument leads to the equality for all $v \in V$ which implies $w \equiv u$.

According to Theorem 2.1, u is the unique solution and, as a consequence, not only a subsequence but the whole sequence $\{u_{M(h)}\}$ converges weakly to u.

To show the strong convergence, we introduce a sequence of functions $w_h \in V_h$ such that

$$\lim_{h\to 0+} \|w_h - u + \overline{u}\|_{1,\Omega} = 0 \tag{3.7}$$

and define functions $v_h = u_{M(h)} - \overline{u} - w_h$, $v_h \in V_h$, $h \to 0+$. If v_h tends to zero then (3.5) holds by virtue of (3.7) and the triangle inequality.

Let us estimate $\|v_h\|_{1,\Omega}$ using (3.2) and (1.5) with v_h substituted for v:

$$\begin{aligned}
\widetilde{a}_{\min}\|v_h\|^2_{1,\Omega} &\le (a_{M(h)}(u_{M(h)})u'_{M(h)}, v'_h)_{0,\Omega} \\
&\quad - (a_{M(h)}(u_{M(h)})(\overline{u}+w_h)', v'_h)_{0,\Omega} \\
&= (a(u)u', v'_h)_{0,\Omega} - (a_{M(h)}(u_{M(h)})(\overline{u}+w_h)', v'_h)_{0,\Omega} \\
&\le \left|\big((a(u)-a_{M(h)}(u))u', v'_h\big)_{0,\Omega}\right| \\
&\quad + \left|\big(a_{M(h)}(u)(u-(\overline{u}+w_h))', v'_h\big)_{0,\Omega}\right| \\
&\quad + \Big|\big(a_{M(h)}(u)(\overline{u}+w_h)', v'_h\big)_{0,\Omega} \\
&\qquad - \big(a_{M(h)}(u_{M(h)})(\overline{u}+w_h)', v'_h\big)_{0,\Omega}\Big| \\
&\le \Big[\|a-a_{M(h)}\|_{0,\infty,\mathbb{R}}\|u\|_{1,\Omega} + \widetilde{a}_{\max}\|w_h-u+\overline{u}\|_{1,\Omega} \\
&\qquad + C_L\|u-u_{M(h)}\|_{0,\infty,\mathbb{R}}\|\overline{u}+w_h\|_{1,\Omega}\Big]\|v_h\|_{1,\Omega}.
\end{aligned}$$

If $h \to 0+$ then the right-hand side tends to zero as a consequence of (3.7), (3.6) (where $w \equiv u$), and the compact embedding $H^1(\Omega) \subset\subset C(\overline{\Omega})$. ∎

The final theorem of this section takes pattern from [2, Theorem 3.1] but has simpler assumptions.

THEOREM 3.3. *Let* $\{a^0_{M(h)}\}$, $h \to 0+$, *be a sequence of solutions of the Approximate Maximization Problem* (3.4), *and let* $M(h) \to \infty$. *Then there exists a subsequence* $\{a^0_{M(\widehat{h})}\} \subset \{a^0_{M(h)}\}$ *such that*

$$\begin{aligned}
a^0_{M(\widehat{h})} &\to a^0 && \text{in } C(\mathbb{R}), \\
u_{\widehat{h}}(a^0_{M(\widehat{h})}) &\to u(a^0) && \text{in } H^1(\Omega), \\
\Psi^{\widehat{h}}(a^0_{M(\widehat{h})}) &\to \Psi(a^0),
\end{aligned}$$

as $\widehat{h} \to 0+$, *where* a^0 *is a solution of the Maximization Problem* (1.7).

Proof. We can follow the proof of [2, Theorem 3.1] taking into account Lemma 3.1 instead of [2, Proposition 3.2]. ∎

4. SENSITIVITY ANALYSIS

Let us choose $a_0 \in \widetilde{\mathcal{U}}^0_{\text{ad}}(C_L)$ and $a \in \mathcal{U}^0_{\text{ad}}(\widehat{C})$, where $\widehat{C}$ is an arbitrary positive constant. If $\tau_0 > 0$ is sufficiently small then the function $a_\tau = a_0 + \tau a \ge C > 0$ on $\mathbb{R}$ for any $\tau \in (-\tau_0, \tau_0)$, and a unique state solution $u_\tau \equiv u(a_\tau)$ of (1.4)–(1.5) exists. We examine the Gâteaux derivative of the state solution $u(a_0)$ in a certain norm and in the direction determined by a. We write u_0 instead of $u(a_0)$ in what follows.

THEOREM 4.1. *There is a unique function* $\dot{u}_0 \in H^1(\Omega)$ *such that*

$$\lim_{\tau\to 0}\left\| \frac{u_\tau - u_0}{\tau} - \dot{u}_0 \right\|_{0,\infty,\Omega} = 0.$$

The function $\dot{u}_0$ *reads*

$$\dot{u}_0(x) = \frac{-1}{a_0(u_0(x))}\int_{\overline{u}}^{u_0(x)} a(t)\,\mathrm{d}t, \qquad x \in \overline{\Omega}. \tag{4.1}$$

Proof. Since the Kirchhoff transform (2.1) applied to u_τ, $\tau \in (-\tau_0, \tau_0)$ results in a unique function z, we have for any $x \in \overline{\Omega}$

$$\begin{aligned} 0 &= \int_{\overline{u}}^{u_\tau(x)} a_\tau(t)\,\mathrm{d}t - \int_{\overline{u}}^{u_0(x)} a_0(t)\,\mathrm{d}t \\ &= \int_{u_0(x)}^{u_\tau(x)} a_0(t)\,\mathrm{d}t + \tau\int_{\overline{u}}^{u_\tau(x)} a(t)\,\mathrm{d}t. \end{aligned} \tag{4.2}$$

By this equality, Lemma 2.2 and the embedding $H^1(\Omega) \subset\subset C(\overline{\Omega})$,

$$\lim_{\tau\to 0}\frac{1}{\tau}\int_{u_0(x)}^{u_\tau(x)} a_0(t)\,\mathrm{d}t = \lim_{\tau\to 0} -\int_{\overline{u}}^{u_\tau(x)} a(t)\,\mathrm{d}t = -\int_{\overline{u}}^{u_0(x)} a(t)\,\mathrm{d}t. \tag{4.3}$$

We observe that $a_0(t) \le a_0(t_1) + C_L(t - t_1)$ and $a_0(t) \ge a_0(t_1) - C_L(t - t_1)$ if $t \ge t_1$. The first integral at the right-hand side of (4.2) can be estimated from below and above by the inequalities

$$\begin{aligned} \int_{u_0(x)}^{u_\tau(x)} a_0(t)\,\mathrm{d}t &\le \int_{u_0(x)}^{u_\tau(x)} [a_0(u_0(x)) + C_L(t - u_0(x))]\,\mathrm{d}t \\ &= (u_\tau(x) - u_0(x))[a_0(u_0(x)) + C_L(u_\tau(x) - u_0(x))/2] \\ &= (u_\tau(x) - u_0(x))c_1(x,\tau), \\ \int_{u_0(x)}^{u_\tau(x)} a_0(t)\,\mathrm{d}t &\ge (u_\tau(x) - u_0(x))[a_0(u_0(x)) - C_L(u_\tau(x) - u_0(x))/2] \\ &= (u_\tau(x) - u_0(x))c_2(x,\tau) \end{aligned}$$

which are valid for any $x \in \overline{\Omega}$ and irrespective of $\operatorname{sign}(u_\tau(x) - u_0(x))$. If τ_0 is sufficiently small then $0 < c_1(x,\tau)$, $0 < c_2(x,\tau)$ for any $x \in \overline{\Omega}$ and $\tau \in (-\tau_0, \tau_0)$. We can suppose, without loss of generality, that $\tau > 0$. Combining the inequalities and dividing by τ, c_1, and c_2, we arrive at

$$\frac{1}{\tau c_1(x,\tau)}\int_{u_0(x)}^{u_\tau(x)} a_0(t)\,\mathrm{d}t \le \frac{u_\tau(x) - u_0(x)}{\tau} \le \frac{1}{\tau c_2(x,\tau)}\int_{u_0(x)}^{u_\tau(x)} a_0(t)\,\mathrm{d}t.$$

By Lemma 2.2 and the continuous embedding $H^1(\Omega)$ into $C(\overline{\Omega})$, the values $c_1(x,\tau)$, $c_2(x,\tau)$ tend to $a_0(u_0)$ uniformly on $\overline{\Omega}$ if $\tau \to 0+$.

Taking this and (4.3) into account, we can define $\dot{u}_0$ and express it by the following simple formula:

$$\dot{u}_0(x) \equiv \lim_{\tau\to 0} \frac{u_\tau(x) - u_0(x)}{\tau} = \frac{-1}{a_0(u_0(x))} \int_{\bar{u}}^{u_0(x)} a(t)\,\mathrm{d}t. \tag{4.4}$$

Since $\tau \to 0+$ and $\tau \to 0-$ lead to the identical result, we can write $\tau \to 0$ in (4.4). Due to Lemma 2.2 and the equality (4.3), the limit (4.4) is uniform on $\overline{\Omega}$. To see that $\dot{u}_0 \in H^1(\Omega)$ it is sufficient to differentiate the formula derived above. ∎

As a direct consequence of Theorem 4.1, we get

$$\begin{aligned}\dot{\Psi}_j(a_0, a) &\equiv \left.\frac{\mathrm{d}\Phi_j(u_\tau)}{\mathrm{d}\tau}\right|_{\tau=0} \\ &= -(\operatorname{meas} G_j)^{-1} \int_{G_j} \left(\frac{1}{a_0(u_0(x))} \int_{\bar{u}}^{u_0(x)} a(t)\,\mathrm{d}t \right) \mathrm{d}x. \end{aligned}\tag{4.5}$$

Since

$$\max_{a\in\mathcal{U}_{\mathrm{ad}}} \left\{ \max_{1\le j\le J} \Psi_j(a) \right\} = \max_{1\le j\le J} \left\{ \max_{a\in\mathcal{U}_{\mathrm{ad}}} \Psi_j(a) \right\},$$

the Maximization Problem (1.7) can be subdivided into J particular maximization problems defined as the search for the maximum of Ψ_j over $\mathcal{U}_{\mathrm{ad}}$, $j = 1, \ldots, J$. That is why we focus only on a functional Ψ_j in what follows.

In practice, we deal with a vector α and the set $\widehat{\mathcal{U}}_{\mathrm{ad}}^M$ (see (3.3)) rather than with a function a and the set $\mathcal{U}_{\mathrm{ad}}^M$, respectively.

Let us consider a vector $\alpha^i \in \mathbb{R}^M$ the components of which are equal to 0 except for the ith one which equals 1. There is a unique piecewise linear function $a_i \in \mathcal{U}_{\mathrm{ad}}^0((M-1)/(T_{\mathrm{r}} - T_{\mathrm{l}}))$ defined by the vector α^i of the nodal values at points T_k, $k = 1, \ldots, M$.

Any vector $\alpha = (\alpha_1, \ldots, \alpha_M) \in \widehat{\mathcal{U}}_{\mathrm{ad}}^M$ corresponds to a unique function $a_\alpha \in \mathcal{U}_{\mathrm{ad}}^M$ which implies a unique state solution $u(a_\alpha)$ and, as supposed, a unique Galerkin approximation $u_h(a_\alpha)$.

Setting $\widehat{\Psi}_j(\alpha) = \Phi_j(u(a_\alpha))$ and taking into account Lemma 2.2 and the derivative (4.5), we have

$$\widehat{\Psi}_{j,i}(\alpha) \equiv \frac{\partial\widehat{\Psi}_j}{\partial\alpha_i}(\alpha) = \dot{\Psi}_j(a_\alpha, a_i) \approx \Theta_{j,i}^h(\alpha), \qquad i = 1, \ldots, M. \tag{4.6}$$

The definition of the functional $\Theta_{j,i}^h(\alpha)$ coincides with the right-hand side of the equality (4.5), where the functions a_α, $u_h(a_\alpha)$, and a_i are substituted for a_0, u_0, and a, respectively. According to Lemma 3.1, where a fixed $M(h)$ independent of h is considered, $u_h(a_\alpha) \to u(a_\alpha)$ in $H^1(\Omega)$ so that $\Theta_{j,i}^h(\alpha) \to \dot{\Psi}_j(a_\alpha, a_i)$ as $h \to 0+$.

For simplicity reasons, let us suppose that there exists $i_0 \in \{1, \ldots, M\}$ such that $T_{i_0} = \overline{u}$.

LEMMA 4.1. *Let* T_i, $i \in \{1, \ldots, M\} \setminus \{i_0\}$, *and* $\alpha \in \widehat{\mathcal{U}}_{\mathrm{ad}}^M$ *be arbitrary. If* $T_i > \overline{u}$ *then* $\widehat{\Psi}_{j,i}(\alpha) \le 0$, *if* $T_i < \overline{u}$ *then* $\widehat{\Psi}_{j,i}(\alpha) \ge 0$, $1 \le j \le J$.

Proof. Since $a_k(T_i) = \delta_{ik}$ (the Kronecker symbol) and $a_\alpha > 0$, $a_i \ge 0$, the assertion is a direct consequence of (4.5) and (4.6). ∎

Lemma 4.1 plays a crucial role in solving the problem (1.7) formulated in terms of $\widehat{\Psi}_j$, α, and $\widehat{\mathcal{U}}_{\mathrm{ad}}^M$ now. To maximize $\widehat{\Psi}_j(\alpha)$, α_i tends to $a_{\min}(T_i)$ for $i > i_0$, and α_i tends to $a_{\max}(T_i)$ if $i < i_0$.

Remark 4.1. In applications, we can expect that the set G_j comprises a point where a (local) extremum of the function z or, equivalently, u_α is achieved (see Lemma 2.1). Let us suppose $z|_{G_j}$ is a nonnegative function. As in Lemma 4.1, we can infer from (4.6) and (4.5) that the ith component of the vector $\alpha_j^0 = \arg\max_{\alpha \in \widehat{\mathcal{U}}_{\mathrm{ad}}^M} \widehat{\Psi}_j(\alpha)$ is equal to $a_{\min}(T_i)$, $i = i_0, \ldots, M$. The remaining components can be arbitrary within the admissible set. They do not influence the state solution on G_j (cf. Lemma 2.1 and (4.5)).

A similar conclusion with $a_{\max}(T_i)$ and $i = 1, \ldots, i_0$ can be drawn for $z|_{G_j} \le 0$.

If z changes its sign on G_j, then the above tendencies are combined with the slope constraint parameter C_L. Roughly speaking, on a semiaxis, the "extremal" function a_j^0 sticks to $a_{\min}$ or $a_{\max}$, and switches to the other value on the neighborhood of T_{i_0}, where its maximum increase or decrease rate is bounded by C_L. ∎

Remark 4.2. If a *spatial* form of the problem (1.1)–(1.3) is considered such that the Kirchhoff transformation can be applied, i.e., if a is a *scalar* function, then, under some smoothness assumptions, (4.4) holds as well as (4.5). This would be a starting point to some generalizations of Lemma 4.1 and Remark 4.1 to problems, where $\Omega \subset \mathbb{R}^n$, $n = 2, 3$ for example. ∎

5. NUMERICAL EXAMPLES

A MATLAB program was coded to solve the Maximization Problem (3.4) with $J = 1$. The functionals $\Theta_{1,i}^h$ were evaluated by the trapezoidal rule applied to nodal values exactly calculated from the Galerkin approximation u_h. The stiffness matrices in the Katchanov algorithm (see Theorem 3.1) were also computed exactly. The Katchanov iterations ran until the $C(\overline{\Omega})$-norm of the difference of two successive iterative solutions was decreasing. The sequential quadratic programming (SQP) routine `E04UCF`, see [6], was chosen to solve (3.4). In examples, the symbol $\widehat{\Psi}_1^h(\alpha)$ stands for $\Phi_1(u_h(a_\alpha))$.

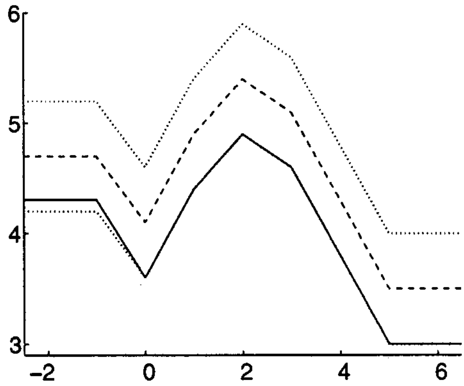

FIG. 1. Example 5.1: $a_{\min}$, $a_{\max}$, a_{ini}, $a_{\alpha_1^0}$.

Example 5.1. We chose $f = 100\chi(G_1)$ (χ is the characteristic function), $G_1 = (0.44, 0.66)$, $\Gamma_1 = \{0, 1\}$, $\overline{u} = 0$, $M = 7$, $T_l = -1$, $T_r = 5$, $C_L = 0.8$, and $h = 0.02$ (see Fig. 1 for $a_{\min}$, $a_{\max}$ (dotted lines)). It is $u_h(a_\alpha)|_{G_1} > 0$ for any admissible α. The initial value $\widehat{\Psi}_1^h(\alpha_{\text{ini}}) = 4.358$ corresponds to $a_{\text{ini}} = (a_{\min} + a_{\max})/2$ (dashed line in Fig. 1). After three SQP minimization steps and five functional evaluations, α_1^0 was found, $\widehat{\Psi}_1^h(\alpha_1^0) = 5.029$, (see Fig. 1 for $a_{\alpha_1^0}$ (solid line)).

Example 5.2. We set $f = 14000\sin(16x)(1 + (x - 0.4)^2)$, $\Gamma_1 = \{0, 1\}$, $\overline{u} = 0$, $G_1 = (0.281, 0.463)$, $a_{\min} = 4.0$, $a_{\max} = 10.1$, $T_l = -5$, $T_r = 5$, and $C_L = 1$. The function $u_h(a_\alpha)$ changes its sign on G_1. Optimization runs with various parameters M and h were performed. Always, $\alpha_{\text{ini}} = (7.05, \ldots, 7.05)$. The values $p(M) = \widehat{\Psi}_1^h(\alpha_M^0)$, $h = 0.02$, are graphed in Fig. 2. The graph is almost identical for $h = 0.01$, $h = 0.005$, or $h = 0.0025$

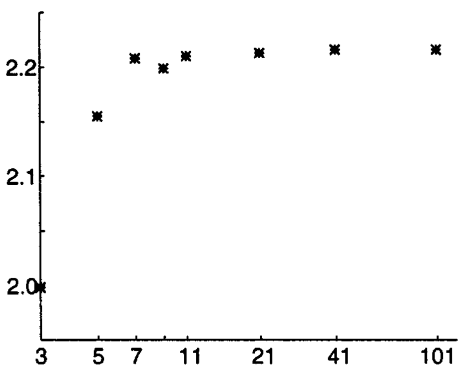

FIG. 2. Example 5.2: values $p(M)$

because the difference between the corresponding values of $\widehat{\Psi}_1^h(\alpha_M^0)$ is less than 1%. To give the reader an idea of the vector α_M^0, let us list its components for $M = 21$: $\alpha_1 = \alpha_2 = 10.1$, $\alpha_i = \alpha_{i-1} - 0.5$, $i = 3, \ldots, 18$, $\alpha_{19} = \alpha_{20} = \alpha_{21} = 4$.

ACKNOWLEDGMENTS

The author wishes to thank his colleague Dr. I. Hlaváček and also anonymous reviewers for valuable comments and suggestions. This work was supported by Grant 201/97/0217 of the Grant Agency of the Czech Republic. The support is gratefully acknowledged.

REFERENCES

1. L. E. J. Brower, Über Abbildung von Mannigfaltigkeiten, *Math. Ann.* **71** (1912), 97–115.
2. I. Hlaváček, Reliable solution of a quasilinear nonpotential elliptic problem of a nonmonotone type with respect to the uncertainty in coefficients, *J. Math. Anal. Appl.* **212** (1997), 452–466.
3. I. Hlaváček, "Reliable solutions of elliptic boundary value problems with respect to uncertain data," *in* Proceedings of WCNA 1996, *Nonlinear Anal.* **30** (1997), 3879–3890.
4. I. Hlaváček, Reliable solutions of problems in the deformation theory of plasticity with respect to uncertain material function, *Appl. Math.* **41** (1996), 447–466.
5. I. Hlaváček, M. Křížek, and J. Malý, On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type, *J. Math. Anal. Appl.* **184** (1994), 168–189.
6. "NAG Foundation Toolbox User's Guide," The Numerical Algorithms Group Ltd. and The MathWorks, Inc, Natick, 1996.
7. J. Nečas, "Les Méthodes Directes en Théorie des Equations Elliptiques," Academia, Prague, 1967.
8. A. E. Taylor, "Introduction to Functional Analysis," Wiley, New York, 1958.