

ERROR ESTIMATES OF AN EFFICIENT LINEARIZATION SCHEME FOR A NONLINEAR ELLIPTIC PROBLEM WITH A NONLOCAL BOUNDARY CONDITION*

MARIAN SLODIČKA¹

Abstract. We consider a nonlinear second order elliptic boundary value problem (BVP) in a bounded domain $\Omega \subset \mathbb{R}^N$ with a nonlocal boundary condition. A Dirichlet BC containing an unknown additive constant, accompanied with a nonlocal (integral) Neumann side condition is prescribed at some boundary part Γ_n . The rest of the boundary is equipped with Dirichlet or nonlinear Robin type BC. The solution is found *via* linearization. We design a robust and efficient approximation scheme. Error estimates for the linearization algorithm are derived in $L_2(\Omega)$, $H^1(\Omega)$ and $L_\infty(\Omega)$ spaces.

Mathematics Subject Classification. 65N15, 35J60.

Received: July 6, 2000. Revised: November 13, 2000.

1. INTRODUCTION

Nonlinear convection-diffusion problems are known to have a lot of important applications. The time discretization of such a problem leads to the solution of a sequence of nonlinear elliptic BVPs, solving of which already belongs to the classical subjects of the numerical analysis. Many special techniques have been developed to solve these problems. Some of them lead to the solution of a nonlinear system of algebraic equations others rely on linearization. This, of course, can be done in various ways. Some algorithms use the Newton type linearization. This usually needs a good starting data which means that the time step should be relatively small. An other group is based on so called relaxation schemes. Here, the nonlinear elliptic PDE is replaced by a linear one along with an algebraic equation for corrections. Both of them are solved iteratively. One of the most general and effective algorithms is the so called Jäger-Kačur scheme (see [10, 11]). This has been presented in various modifications which can be found in the literature. After the time discretization, the convergence of iterations at each time point has been proved only in the 1D case (*cf.* [11]). In the higher dimensional situation one needs a uniform bound for iterations in the $C^{0,\alpha}(\overline{\Omega})$ space, which remains an open question. Moreover, Kačur has employed in [11] a regularization of a nonlinear function using the time step τ , which is independent of the iteration parameter k . Then, the author has proved the following estimate

$$|u - u_k| \leq C\tau^r |u - u_{k-1}|, \quad r > 0,$$

Keywords and phrases. Nonlinear elliptic BVP, error estimates, nonstandard boundary condition, linearization.

* This work was supported by the VEO-project No. 011 VO 697 of the Flemish Government.

¹ Department of Mathematical Analysis, Faculty of Engineering, Ghent University, Galglaan 2, B-9000 Ghent, Belgium.
e-mail: ms@cage.rug.ac.be

which represents a contraction only in the case when the time step τ is sufficiently small. This is intuitively clear, because the regularization error must depend on the regularization parameter. We will design such a scheme, where the iteration process will converge without any dependence on additional parameters coupled with the discretization.

Another group of linearization methods uses the monotonicity of iterations. Examples of such schemes can be found in [6, 14, 17] or p. 155 of [12]. The linearization of a nonlinear problem relies on the ordering properties of solutions. One defines recursive sequences starting from a *sub-* and a *super-solution*, respectively. Then there in fact will exist a solution lying between sub- and super-solution. The disadvantage of this approach is the fact that one has to start far away from the real solution and the information from the previous time step cannot be used as the starting point for the approximation scheme. Otherwise it is not possible to prove the monotonicity of iterations.

The need of a reliable, efficient and robust iteration scheme for the solution of nonlinear elliptic BVPs, which can start from arbitrary initial data, is evident. We propose such an algorithm in this paper. We consider a nonlinear second order elliptic PDE, where the nonlinearity can appear as a source term or at the Robin type BC. In both situations we assume that the functions are monotonically increasing and globally Lipschitz continuous. We distinguish between a regular ($0 < \delta \leq \beta'(\cdot) \leq L$) and a degenerate ($0 \leq \beta'(\cdot) \leq L$) case. Here, the function β represents the nonlinearity appearing in the equation or in the BC. The degeneracy ($\beta'(s) = 0$) is allowed at a single point, only. The case when β' vanishes in an interval is not covered in this paper, due to the proof-technique which is used here, and it will be studied later.

In the regular case, our algorithm coincides with the scheme from Evans [6], Slodička-Van Keer [17] or Slodička [14], but the main difference is that we prove the convergence without using the ordering property of approximations. Thus, we can start from arbitrary data and the iteration scheme will converge to the exact solution. In the degenerate case, we first apply a local regularization to the nonlinear function β , and then we use a similar linearization to the regular instance. Here, of course, the argument for convergence is more delicate. In both situations (regular and degenerate) we establish the error estimates for the linearization procedure in the $L_2(\Omega)$, $H^1(\Omega)$ and $L_\infty(\Omega)$ spaces.

Nonstandard boundary value problems become of growing interest, as they sometimes model the physical situation more adequately. Various mathematical models containing nonlocal BCs can be found in the literature, *e.g.*, in p. 520 of [7] in the so called *plasma problem*; in the computation of the electromagnetic losses in a lamination of an electric machine – see [18]; in Navier-Stokes equations *cf.* [13]; or in the Stokes problem, *cf.* [1, 3, 5, 15].

We demonstrate our approximation scheme on a second order nonlinear convection-diffusion equation with a divergence structure accompanied with a nonstandard BC on some part Γ_n of the boundary. Here, the Dirichlet BC contains an unknown additive constant, and the total flux through Γ_n is given. Such a type of BC can be found in some applications (*cf.* §4.1 of [8] or [18]). To enhance the comprehension we give a simple example. Consider the movement of air through a porous medium. Boundary conditions reflect the behavior of the solution u (which represents the air pressure) or the flux \mathbf{q} at the boundary. Let Γ_n be the boundary of an air pumping well. The nonlocal BC has the form

$$u = c \in \mathbb{R} \text{ (unknown) on } \Gamma_n, \quad \int_{\Gamma_n} \mathbf{q} \cdot \boldsymbol{\nu} = s \text{ (given)}.$$

This means, that the air pressure along Γ_n is supposed to be constant but unknown, whereas the discharge (total flux through Γ_n) is prescribed. Let us note, that the normal component of the flux cannot be measured point-wise. The possible inhomogeneities in the vicinity of the well disable the use of a constant Neumann BC at the well screen.

Such a type of BC in a nonlinear elliptic BVP has been studied in [16]. The authors have dealt with classical solutions and their proof-technique relies on the comparison principle – see Theorem 9.2 in [9]. Later, Slodička and Van Keer [17] enriched the problem setting from [16] by a linear Robin type BC. Moreover, the authors developed a suitable variational framework, thus the assumptions on the data have been weakened. The

nonlinear Robin type BC has been studied in Slodička [14]. The basic idea of the variational approach relies on a special choice of the test function space, elements of which have constant trace on Γ_n . A consequence of such a selection is the presence of an additive term containing the total discharge s in the variational formulation.

We focus the attention to the explanation of a new efficient linear approximation scheme and we present a new proof-technique, which allows the use of arbitrary starting data for iterations. At the end of this paper we present some numerical examples to demonstrate the robustness and the efficiency of the algorithm.

Let us note that all the presented results remain valid also in the case without a nonlocal BC. Then, of course, there exist practical applications leading to a nonlinear elliptic equation of the type we study in our paper, *e.g.*, models in molecular interactions and subsonic flows with the reaction function $g(u) = u^p$, $p \geq 1$ (see p. 134 in [12]). Other interesting application comes from the free boundary problems of the type

$$\frac{dg(u)}{dt} - \Delta u = f$$

with $g(u) = u|u|^p$, $p \geq 0$. This, after the time discretization leads to a nonlinear elliptic problem, which can be solved by the proposed numerical scheme.

Throughout the remainder of the paper C denotes a generic positive constant which can depend on the domain, on the data functions or on the constants δ and L , but it is independent of the iteration parameter k .

2. PROBLEM FORMULATION AND ASSUMPTIONS

Let $\Omega \subset \mathbb{R}^N$, $N \geq 2$, be a bounded domain with a Lipschitz continuous boundary Γ , which is divided into three parts Γ_D , Γ_N and Γ_n , corresponding to Dirichlet, Neumann and nonlocal part, respectively. Throughout the whole paper we assume that

$$|\Gamma_D| > 0, \quad \overline{\Gamma_n} \cap \overline{\Gamma_D} = \emptyset, \quad |\Gamma_n| > 0. \tag{1}$$

The last inequality means that we are dealing with a nonlocal BC on Γ_n . In this paper we study the following nonlinear elliptic BVP in divergence form

$$\left\{ \begin{array}{ll} \nabla \cdot (-\mathbf{A}_{\text{dif}} \nabla u - \mathbf{a}_{\text{con}} u) + g(u) = f & \text{in } \Omega \\ u = g_{\text{Dir}} & \text{on } \Gamma_D \\ (-\mathbf{A}_{\text{dif}} \nabla u - \mathbf{a}_{\text{con}} u) \cdot \boldsymbol{\nu} - g_{\text{Rob}}(u) = g_{\text{Neu}} & \text{on } \Gamma_N \\ u = g_n + \text{const} & \text{on } \Gamma_n \\ G(u) \equiv \int_{\Gamma_n} (-\mathbf{A}_{\text{dif}} \nabla u - \mathbf{a}_{\text{con}} u) \cdot \boldsymbol{\nu} = s \in \mathbb{R}. \end{array} \right. \tag{2}$$

The matrix \mathbf{A}_{dif} fulfills the inequality

$$C_0 |w|_{1,\Omega}^2 \leq (\mathbf{A}_{\text{dif}} \nabla w, \nabla w)_\Omega \leq C |w|_{1,\Omega}^2, \quad \forall w \in H^1(\Omega) \tag{3}$$

for some positive constants C_0 and C . Here, $(w, z)_M$ stands for the usual L_2 -inner product of any real or vector-valued functions w, z on a set M , *i.e.*, $(w, z)_M = \int_M wz$. The fact that $|\Gamma_D| > 0$ implies that the seminorm $|\cdot|_{1,\Omega}$ represents an equivalent norm in $H^1(\Omega)$ to the usual norm $\|\cdot\|_{1,\Omega}$.

The convection term \mathbf{a}_{con} obeys the following relations

$$\left\{ \begin{array}{ll} |\mathbf{a}_{\text{con}}| \leq C & \text{a.e. in } \Omega \\ \mathbf{a}_{\text{con}} \cdot \boldsymbol{\nu} \geq 0 & \text{a.e. on } \Gamma_N \cup \Gamma_n \\ \nabla \cdot \mathbf{a}_{\text{con}} = 0 & \text{a.e. in } \Omega. \end{array} \right. \tag{4}$$

The condition $\nabla \cdot \mathbf{a}_{\text{con}} = 0$ a.e. in Ω physically means that the convection is caused by an independent steady state process without spatially distributed sources.

The nonlinear functions g and g_{Rob} are supposed to be globally Lipschitz continuous and monotonically increasing

$$0 \leq \beta' \leq L, \quad \text{a.e. in } \mathbb{R}, \quad \beta = g, g_{\text{Rob}},$$

where β' is the a.e. classical derivative of β . Later, we will adopt also some new assumptions on g and g_{Rob} depending on the regular or the degenerate case.

The boundary data $f, g_{\text{Neu}}, g_{\text{Dir}}$ and g_n fulfill

$$f \in L_2(\Omega), \quad g_{\text{Neu}} \in L_2(\Gamma_N) \tag{5}$$

and there exists a function $\tilde{g} \in H^1(\Omega)$ such that

$$\tilde{g} = \begin{cases} g_n & \text{on } \Gamma_n \\ g_{\text{Dir}} & \text{on } \Gamma_D. \end{cases} \tag{6}$$

When dealing with such a general setting as (2), one cannot expect that the solution will be classical. The lack of regularity can be caused by properties of the data entering (2) even in the case when $\Gamma_n = \emptyset$. Therefore, we stay in a variational framework. First, we introduce the following subspace V of $H^1(\Omega)$

$$V = \{\varphi \in H^1(\Omega); \varphi = 0 \text{ on } \Gamma_D, \varphi = \text{const on } \Gamma_n\}, \tag{7}$$

which is clearly a Hilbert space, too. The V can be equipped with the induced norm from $H^1(\Omega)$. Now, we define the bilinear form $a : H^1(\Omega) \times H^1(\Omega) \rightarrow \mathbb{R}$ as

$$a(u, \varphi) = (\mathbf{A}_{\text{dif}} \nabla u + \mathbf{a}_{\text{con}} u, \nabla \varphi)_\Omega \quad \forall u, \varphi \in H^1(\Omega)$$

and the linear functional $F : V \rightarrow \mathbb{R}$

$$\langle F, \varphi \rangle = (f, \varphi)_\Omega - (g_{\text{Neu}}, \varphi)_{\Gamma_N} - s\varphi|_{\Gamma_n} \quad \forall \varphi \in V. \tag{8}$$

The continuity of F follows from (5) and the following obvious inequality

$$|s\varphi|_{\Gamma_n}| = \frac{|s|}{|\Gamma_n|} \int_{\Gamma_n} |\varphi| \leq C \|\varphi\|_{0, \Gamma_n} \leq C \|\varphi\|_{1, \Omega}. \tag{9}$$

The appropriate variational formulation of (2) reads as: Find $u \in H^1(\Omega)$ such that $u - \tilde{g} \in V$ and

$$a(u, \varphi) + (g(u), \varphi)_\Omega + (g_{\text{Rob}}(u), \varphi)_{\Gamma_N} = \langle F, \varphi \rangle \quad \forall \varphi \in V. \tag{10}$$

This variational formulation can be found in [14], where the well-posedness of the problem has been proved under some additional assumptions on g and g_{Rob} . More precisely, the graph of g and g_{Rob} should vary within two parallel increasing lines. We will not use this growth assumption in the proofs.

Next theorem states the existence and uniqueness of a weak solution to the nonlinear elliptic BVP (10).

Theorem 2.1 (existence and uniqueness of u). *Let the assumptions (1) and (3–6) be satisfied. Then there exists a unique solution $u \in H^1(\Omega)$ to the BVP (10).*

Proof. Let w be any function from V . The relation (4) together with the Friedrichs inequality and Green's theorem imply

$$\left\{ \begin{aligned} C|w|_{1,\Omega}^2 \geq (\mathbf{a}_{\text{con}}w, \nabla w)_\Omega &= \frac{1}{2}(\mathbf{a}_{\text{con}}, \nabla w^2)_\Omega \\ &= -\frac{1}{2}(\nabla \cdot \mathbf{a}_{\text{con}}, w^2)_\Omega + \frac{1}{2}(\mathbf{a}_{\text{con}}\boldsymbol{\nu}, w^2)_\Gamma \\ &= \frac{1}{2}(\mathbf{a}_{\text{con}}\boldsymbol{\nu}, w^2)_{\Gamma_N \cup \Gamma_n} \\ &\geq 0. \end{aligned} \right. \tag{11}$$

Hence, in view of (3) we have

$$C|w|_{1,\Omega}^2 \geq a(w, w) \geq C_0|w|_{1,\Omega}^2 \quad \forall w \in V. \tag{12}$$

Let $A : V \rightarrow V^*$ be the operator defined by

$$A(u)\varphi := a(u, \varphi) + (g(u), \varphi)_\Omega + (g_{\text{Rob}}(u), \varphi)_{\Gamma_N} = \langle F, \varphi \rangle.$$

It is a standard matter to check that A is a maximal monotone and coercive operator on V . Hence, accounting for Brézis (Corollaire 2.4, page 31 in [4]), and $F \in V^*$ one has that the variational equation $Au = F$, equivalent indeed to (10) for all $\varphi \in V$, admits exactly one solution in $H^1(\Omega)$.

Let us note that $u - \tilde{g}$ is constant, which follows from the special choice of the test function space V . At this stage (for the existence proof) it is not necessary to stress the way how this unknown boundary data may be recovered. This will be explained later, namely in Section 5. To do this, we have to explain the linearization process first, and then we show a simply but tricky way of the recovery of the unknown constant (see (41)). \square

Our next task is to construct the solution of (10). We will define a linear approximation scheme, the solution of which will approach the exact one in some functional spaces. We distinguish between regular and degenerate cases.

3. REGULAR CASE

Throughout this section we assume that both functions g and g_{Rob} do not degenerate, *i.e.*, there exists a positive constant δ such that

$$0 < \delta \leq \beta' \leq L, \quad \text{a.e. in } \mathbb{R}, \quad \beta = g, g_{\text{Rob}}. \tag{13}$$

Let u_0 be an arbitrary function satisfying

$$u_0 \in L_2(\Omega) \cap L_2(\Gamma_N). \tag{14}$$

We introduce a sequence $\{u_k\}_{k=0}^\infty$ which is defined recursively. More precisely, u_k for $k = 1, 2, \dots$ is a weak solution to the following linear elliptic BVP: Find $u_k \in H^1(\Omega)$ such that $u_k - \tilde{g} \in V$ and

$$\begin{aligned} a(u_k, \varphi) + (Lu_k, \varphi)_\Omega + (Lu_k, \varphi)_{\Gamma_N} &= \langle F, \varphi \rangle + (Lu_{k-1}, \varphi)_\Omega - (g(u_{k-1}), \varphi)_\Omega \\ &\quad + (Lu_{k-1}, \varphi)_{\Gamma_N} - (g_{\text{Rob}}(u_{k-1}), \varphi)_{\Gamma_N} \end{aligned} \tag{15}$$

holds for any $\varphi \in V$.

Our first concern is to prove the well-posedness of the BVP (15).

Lemma 3.1 (existence and uniqueness of u_k). *Let the assumptions (1, 3–6, 13, 14) be satisfied. Then the sequence $\{u_k\}_{k=1}^\infty \subset H^1(\Omega)$ is well defined.*

Proof. According to the relation (12) we see that the left-hand side of (15) is a V -elliptic continuous bilinear form.

Let $k = 1$. Then, according to (5, 9, 13, 14) the right-hand side of (15) is a bounded linear functional on V . Thus, the Lax-Milgram lemma guarantees the existence and uniqueness of the weak solution $u_1 \in H^1(\Omega)$ to the BVP (15) for $k = 1$. Since u_{k-1} belongs to $H^1(\Omega)$, the right-hand side of (15) is a bounded linear functional on V . Hence, there exists a unique $u_k \in H^1(\Omega)$ satisfying (15).

The unknown boundary constant for each u_k can be recovered using the continuous dependence of a solution on the BCs and taking into account the principle of superposition. More details can be found in Section 5. \square

Our next interest is to derive the error estimates for the linearization scheme (15). First, we introduce the following notation

$$h(s) = g(s) - Ls, \quad h_{\text{Rob}}(s) = g_{\text{Rob}}(s) - Ls, \quad s \in \mathbb{R}. \tag{16}$$

When we subtract (10) from (15), we get the variational formulation for the error of the linearization scheme

$$a(u_k - u, \varphi) + L(u_k - u, \varphi)_\Omega + L(u_k - u, \varphi)_{\Gamma_N} = (h(u) - h(u_{k-1}), \varphi)_\Omega + (h_{\text{Rob}}(u) - h_{\text{Rob}}(u_{k-1}), \varphi)_{\Gamma_N}, \tag{17}$$

which holds for any $\varphi \in V$.

Now, we are in a position to prove the error estimates of the approximations in the $H^1(\Omega)$ space.

Theorem 3.2 ($H^1(\Omega)$ -error estimate). *Let the assumptions of Lemma 3.1 be satisfied. Then there exists a positive constant $C = C(C_0, \delta, L)$ such that*

$$\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 + |u_k - u|_{1,\Omega}^2 \leq C \left(1 - \frac{2\delta}{L + \delta}\right)^k \left[\|u_0 - u\|_{0,\Omega}^2 + \|u_0 - u\|_{0,\Gamma_N}^2\right]$$

holds for all $k = 1, 2, \dots$

Proof. We set $\varphi = u_k - u \in V$ in (17) and get

$$a(u_k - u, u_k - u) + L\|u_k - u\|_{0,\Omega}^2 + L\|u_k - u\|_{0,\Gamma_N}^2 = (h(u) - h(u_{k-1}), u_k - u)_\Omega + (h_{\text{Rob}}(u) - h_{\text{Rob}}(u_{k-1}), u_k - u)_{\Gamma_N}. \tag{18}$$

Using the relation (13), one can easily deduce the following estimate for the functions h and h_{Rob}

$$\begin{aligned} \delta - L &\leq h'(s) = g'(s) - L \leq 0 && \text{a.e. in } \mathbb{R} \\ \delta - L &\leq h'_{\text{Rob}}(s) = g'_{\text{Rob}}(s) - L \leq 0 && \text{a.e. in } \mathbb{R}. \end{aligned}$$

Therefore, $|h'(s)| \leq L - \delta$ and $|h'_{\text{Rob}}(s)| \leq L - \delta$ are valid a.e. in \mathbb{R} . The terms on the right-hand side of (18) can be estimated using the Cauchy and Young inequalities as follows

$$\begin{aligned} |(h(u) - h(u_{k-1}), u_k - u)_\Omega| &\leq \|h(u) - h(u_{k-1})\|_{0,\Omega} \|u_k - u\|_{0,\Omega} \\ &\leq (L - \delta) \|u - u_{k-1}\|_{0,\Omega} \|u_k - u\|_{0,\Omega} \\ &\leq \frac{L - \delta}{2} \|u - u_{k-1}\|_{0,\Omega}^2 + \frac{L - \delta}{2} \|u_k - u\|_{0,\Omega}^2, \end{aligned}$$

and analogously we deduce

$$|(h_{\text{Rob}}(u) - h_{\text{Rob}}(u_{k-1}), u_k - u)_{\Gamma_N}| \leq \frac{L - \delta}{2} \|u - u_{k-1}\|_{0,\Gamma_N}^2 + \frac{L - \delta}{2} \|u_k - u\|_{0,\Gamma_N}^2.$$

In virtue of the V -ellipticity of the bilinear form a (see (12)), we estimate the left-hand side of (18) from below by

$$L \|u_k - u\|_{0,\Omega}^2 + L \|u_k - u\|_{0,\Gamma_N}^2 + C_0 |u_k - u|_{1,\Omega}^2.$$

Summarizing the foregoing results we arrive at

$$\frac{L + \delta}{2} \left[\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 \right] + C_0 |u_k - u|_{1,\Omega}^2 \leq \frac{L - \delta}{2} \left[\|u - u_{k-1}\|_{0,\Omega}^2 + \|u - u_{k-1}\|_{0,\Gamma_N}^2 \right],$$

which after a simple calculation gives

$$\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 + \frac{2C_0}{L + \delta} |u_k - u|_{1,\Omega}^2 \leq \left(1 - \frac{2\delta}{L + \delta} \right) \left[\|u - u_{k-1}\|_{0,\Omega}^2 + \|u - u_{k-1}\|_{0,\Gamma_N}^2 \right]. \quad (19)$$

Omitting the third term on the left, we obtain the recursion formula

$$\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 \leq \left(1 - \frac{2\delta}{L + \delta} \right) \left[\|u - u_{k-1}\|_{0,\Omega}^2 + \|u - u_{k-1}\|_{0,\Gamma_N}^2 \right], \quad (20)$$

which after k iterations gives

$$\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 \leq \left(1 - \frac{2\delta}{L + \delta} \right)^k \left[\|u_0 - u\|_{0,\Omega}^2 + \|u_0 - u\|_{0,\Gamma_N}^2 \right].$$

The rest of the proof follows from the last inequality and (19). \square

The original problem setting (2) has one degree of freedom, namely the unknown real constant at the nonlocal part of the boundary Γ_n , which we will call α . The approximated value of this constant can be recovered from u_k . Let us denote by α_k the approximation of α which is given by u_k , i.e., $\alpha_k = u_k - g_n|_{\Gamma_n} \in \mathbb{R}$. The next theorem gives the corresponding error estimate for α_k .

Theorem 3.3 ($|\alpha - \alpha_k|$ -estimate). *Let the assumptions of Lemma 3.1 be fulfilled. Then there exists a positive constant C such that*

$$|\alpha - \alpha_k|^2 \leq C \left(1 - \frac{2\delta}{L + \delta} \right)^k \left[\|u_0 - u\|_{0,\Omega}^2 + \|u_0 - u\|_{0,\Gamma_N}^2 \right]$$

holds for all $k = 1, 2, \dots$

Proof. Applying the trace theorem we can write

$$|\alpha - \alpha_k|^2 = |u - u_k|_{\Gamma_n}^2 = \frac{1}{|\Gamma_n|} \int_{\Gamma_n} |u - u_k|^2 \leq C \|u - u_k\|_{0,\Gamma}^2 \leq C \|u - u_k\|_{1,\Omega}^2. \quad (21)$$

The rest of the proof follows from Theorem 3.2. \square

The technique of establishing the error estimate in the L_∞ -norm is based on the so called cut-off functions. Similar approach can be found in the proof of some weak maximum principles.

Theorem 3.4 ($L_\infty(\Omega)$ -error estimate). *Let the assumptions of Lemma 3.1 be satisfied. Moreover we suppose $u \in L_\infty(\Omega) \cap L_\infty(\Gamma_N)$. Then, for $k = 1, 2, \dots$ we have*

$$\max \left\{ \|u_k - u\|_{L_\infty(\Omega)}, \|u_k - u\|_{L_\infty(\Gamma_N)} \right\} \leq \left(1 - \frac{\delta}{L} \right)^k \max \left\{ \|u_0 - u\|_{L_\infty(\Omega)}, \|u_0 - u\|_{L_\infty(\Gamma_N)} \right\}.$$

Proof. Let us fix the iteration parameter k for a moment. Now, we introduce the real constants A and B as

$$\begin{aligned} A &= L^{-1} \|h(u_{k-1}) - h(u)\|_{L^\infty(\Omega)} \\ B &= L^{-1} \|h_{\text{Rob}}(u_{k-1}) - h_{\text{Rob}}(u)\|_{L^\infty(\Gamma_N)}. \end{aligned}$$

Further, we denote by Ω^- and Γ_N^- the following sets

$$\begin{aligned} \Omega^- &= \{\mathbf{x} \in \Omega; \quad u_k(\mathbf{x}) - u(\mathbf{x}) + \max\{A, B\} < 0\} \\ \Gamma_N^- &= \{\mathbf{x} \in \Gamma_N; \quad u_k(\mathbf{x}) - u(\mathbf{x}) + \max\{A, B\} < 0\}. \end{aligned}$$

Let us suppose that at least one of the sets Ω^- or Γ_N^- has a positive measure, *i.e.*,

$$|\Omega^-| + |\Gamma_N^-| > 0.$$

We use the same notation here for the N - and for the $(N - 1)$ -dimensional measure! Further, we denote by f^- the usual cut-off function (see, *e.g.*, p. 32 in [9]) which is defined by

$$f^-(s) = \min\{f(s), 0\}.$$

Now, we start with the relation (17) for the error of the linearization scheme and choose $\varphi = (u_k - u + \max\{A, B\})^- \in V$. We get

$$\begin{aligned} a(u_k - u, (u_k - u + \max\{A, B\})^-) &+ L \left(u_k - u, (u_k - u + \max\{A, B\})^- \right)_\Omega \\ &+ L \left(u_k - u, (u_k - u + \max\{A, B\})^- \right)_{\Gamma_N} \\ &= \left(h(u) - h(u_{k-1}), (u_k - u + \max\{A, B\})^- \right)_\Omega \\ &\quad + \left(h_{\text{Rob}}(u) - h_{\text{Rob}}(u_{k-1}), (u_k - u + \max\{A, B\})^- \right)_{\Gamma_N}. \end{aligned}$$

This can be rewritten in the following form

$$\left\{ \begin{aligned} 0 &= \left(\mathbf{A}_{\text{dif}} \nabla(u_k - u), \nabla(u_k - u + \max\{A, B\})^- \right)_\Omega \\ &\quad + \left(\mathbf{a}_{\text{con}}(u_k - u), \nabla(u_k - u + \max\{A, B\})^- \right)_\Omega \\ &\quad + L \left(u_k - u - \frac{h(u) - h(u_{k-1})}{L}, (u_k - u + \max\{A, B\})^- \right)_\Omega \\ &\quad + L \left(u_k - u - \frac{h_{\text{Rob}}(u) - h_{\text{Rob}}(u_{k-1})}{L}, (u_k - u + \max\{A, B\})^- \right)_{\Gamma_N} \\ &= M_1 + M_2 + M_3 + M_4. \end{aligned} \right. \tag{22}$$

The term M_1 is nonnegative in virtue of the V -ellipticity of the matrix \mathbf{A}_{dif} (see (3)), more precisely

$$\begin{aligned} 0 &\leq \left(\mathbf{A}_{\text{dif}} \nabla(u_k - u + \max\{A, B\})^-, \nabla(u_k - u + \max\{A, B\})^- \right)_\Omega \\ &= \left(\mathbf{A}_{\text{dif}} \nabla(u_k - u + \max\{A, B\}), \nabla(u_k - u + \max\{A, B\})^- \right)_\Omega \\ &= \left(\mathbf{A}_{\text{dif}} \nabla(u_k - u), \nabla(u_k - u + \max\{A, B\})^- \right)_\Omega \\ &= M_1. \end{aligned}$$

The second term M_2 will also be nonnegative. To check this, we use the Green theorem and the relations (4) and (11). We successively obtain

$$\begin{aligned}
 M_2 &= \left(\mathbf{a}_{\text{con}}(u_k - u), \nabla (u_k - u + \max\{A, B\})^- \right)_{\Omega} \\
 &= \left(\mathbf{a}_{\text{con}}(u_k - u + \max\{A, B\}), \nabla (u_k - u + \max\{A, B\})^- \right)_{\Omega} \\
 &\quad - \max\{A, B\} \left(\mathbf{a}_{\text{con}}, \nabla (u_k - u + \max\{A, B\})^- \right)_{\Omega} \quad (\pm \max\{A, B\}) \\
 &= \underbrace{\left(\mathbf{a}_{\text{con}}(u_k - u + \max\{A, B\})^-, \nabla (u_k - u + \max\{A, B\})^- \right)_{\Omega}}_{\geq 0} \\
 &\quad + \max\{A, B\} \left(\underbrace{\nabla \cdot \mathbf{a}_{\text{con}}}_{=0}, (u_k - u + \max\{A, B\})^- \right)_{\Omega} \\
 &\quad - \max\{A, B\} \left(\mathbf{a}_{\text{con}} \cdot \boldsymbol{\nu}, (u_k - u + \max\{A, B\})^- \right)_{\Gamma} \quad (\text{using Green's thm.}) \\
 &\geq - \max\{A, B\} \left(\mathbf{a}_{\text{con}} \cdot \boldsymbol{\nu}, (u_k - u + \max\{A, B\})^- \right)_{\Gamma} \quad (\text{using (11) and (4)}) \\
 &= - \max\{A, B\} \left(\underbrace{\mathbf{a}_{\text{con}} \cdot \boldsymbol{\nu}}_{\geq 0}, (u_k - u + \max\{A, B\})^- \right)_{\Gamma_N \cup \Gamma_n} \quad (\text{using (4)}) \\
 &\geq 0.
 \end{aligned}$$

The following inequality holds a.e. in the set Ω^-

$$u_k - u - \frac{h(u) - h(u_{k-1})}{L} \leq u_k - u + A \leq u_k - u + \max\{A, B\} < 0.$$

Thus

$$M_3 = L \left(u_k - u - \frac{h(u) - h(u_{k-1})}{L}, (u_k - u + \max\{A, B\})^- \right)_{\Omega} > 0.$$

There is a similar inequality valid a.e. in Γ_N^- , i.e.,

$$u_k - u - \frac{h_{\text{Rob}}(u) - h_{\text{Rob}}(u_{k-1})}{L} \leq u_k - u + B \leq u_k - u + \max\{A, B\} < 0.$$

Hence, for the fourth term M_4 we get

$$M_4 = L \left(u_k - u - \frac{h_{\text{Rob}}(u) - h_{\text{Rob}}(u_{k-1})}{L}, (u_k - u + \max\{A, B\})^- \right)_{\Gamma_N} > 0.$$

Summarizing the estimates for M_1, \dots, M_4 we arrive at

$$M_1 + M_2 + M_3 + M_4 > 0,$$

which contradicts the relation (22). So, the assumption that $|\Omega^-| + |\Gamma_N^-| > 0$ fails to hold. Thus, we have just proved

$$\begin{aligned}
 u_k - u &\geq -\max\{A, B\} && \text{a.e. in } \Omega \\
 u_k - u &\geq -\max\{A, B\} && \text{a.e. on } \Gamma_N.
 \end{aligned} \tag{23}$$

Our next concern is to prove the following inequalities

$$\begin{aligned} u_k - u &\leq \max\{A, B\} && \text{a.e. in } \Omega \\ u_k - u &\leq \max\{A, B\} && \text{a.e. on } \Gamma_N. \end{aligned} \tag{24}$$

To do this, we first define the sets Ω^+ and Γ_N^+ by

$$\begin{aligned} \Omega^+ &= \{ \mathbf{x} \in \Omega; \quad u_k(\mathbf{x}) - u(\mathbf{x}) - \max\{A, B\} > 0 \} \\ \Gamma_N^+ &= \{ \mathbf{x} \in \Gamma_N; \quad u_k(\mathbf{x}) - u(\mathbf{x}) - \max\{A, B\} > 0 \}. \end{aligned}$$

We start again with the relation (17) for the error of the linearization scheme and choose $\varphi = (u_k - u - \max\{A, B\})^+ = \max\{u_k - u - \max\{A, B\}, 0\} \in V$. Then, we follow exactly the same strategy as we did in the proof of (23). In this way we prove (24). We omit the details.

The relations (23) and (24) imply

$$\max \left\{ \|u_k - u\|_{L_\infty(\Omega)}, \|u_k - u\|_{L_\infty(\Gamma_N)} \right\} \leq \max\{A, B\}. \tag{25}$$

The fact that $|h'(s)| \leq L - \delta$ and $|h'_{\text{Rob}}(s)| \leq L - \delta$ a.e. in \mathbb{R} allows us to write

$$\begin{aligned} \max\{A, B\} &= L^{-1} \max \left\{ \|h(u_{k-1}) - h(u)\|_{L_\infty(\Omega)}, \|h_{\text{Rob}}(u_{k-1}) - h_{\text{Rob}}(u)\|_{L_\infty(\Gamma_N)} \right\} \\ &\leq \left(1 - \frac{\delta}{L}\right) \max \left\{ \|u_{k-1} - u\|_{L_\infty(\Omega)}, \|u_{k-1} - u\|_{L_\infty(\Gamma_N)} \right\}. \end{aligned}$$

So, we have obtained the following recursion formula

$$\max \left\{ \|u_k - u\|_{L_\infty(\Omega)}, \|u_k - u\|_{L_\infty(\Gamma_N)} \right\} \leq \left(1 - \frac{\delta}{L}\right) \max \left\{ \|u_{k-1} - u\|_{L_\infty(\Omega)}, \|u_{k-1} - u\|_{L_\infty(\Gamma_N)} \right\}.$$

Now, we let the iteration parameter k free, and we iterate the recursion formula k -times to conclude the proof. □

4. DEGENERATE CASE

In this section we allow both functions g and g_{Rob} to degenerate, *i.e.*, their derivatives can approach zero. We do it in a single point, only. First, we introduce the following class \mathcal{Q}_b of all real valued functions β associated with any point $b \in \mathbb{R}$ and satisfying the next relations for some given positive constants δ, δ_0 and L

$$\left\{ \begin{array}{ll} 0 \leq \beta'(s) \leq L & \text{a.e. in } \mathbb{R} \\ \beta'(s_+) \beta'(s_-) = 0 \implies s = b & \\ 0 < \delta \leq \beta'(s) \leq L & \text{a.e. in } |s - b| > \delta_0 \\ & \beta' \text{ is nondecreasing in } (0, \delta_0) \\ & \beta' \text{ is nonincreasing in } (-\delta_0, 0). \end{array} \right. \tag{26}$$

We have chosen such criteria in the definition of \mathcal{Q}_b to cover the most interesting situations depicted in Figure 1.

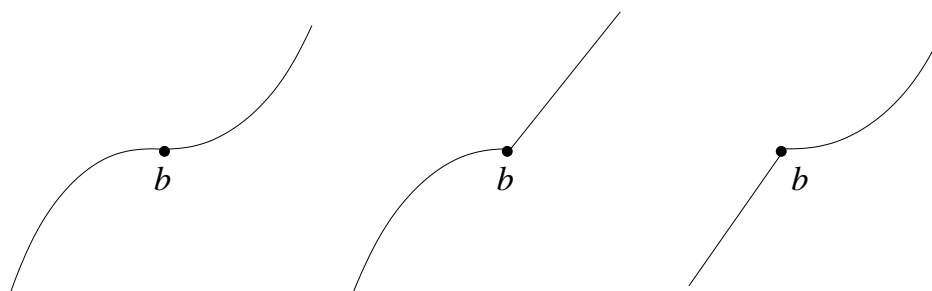


FIGURE 1. Examples of nonlinear functions from \mathcal{Q}_b .

Let us give a practical example of a function from \mathcal{Q}_0 . In the problems of molecular interactions and subsonic flows a simple model for the reaction function g is given by

$$g(u) = \sigma u^p,$$

where σ, p are positive constants with $p \geq 1$. This model also describes the temperature in radiating bodies (or gases) and in nuclear reactors with positive temperature feedback. The fact that the solution has to be nonnegative, allows the modification of f for negative arguments. Further, the solution has to be bounded from the physical point of view, which also allows the linearization of g for large arguments.

Functions g and g_{Rob} can of course belong to different classes, but without loss of generality we will assume that $g, g_{\text{Rob}} \in \mathcal{Q}_0$. In the case when some of the functions g, g_{Rob} belongs to the \mathcal{Q}_b for $b \neq 0$, we have to apply its “regularization” (28) at the point b instead of 0. This means that we modify the function locally at the point of the degeneration.

In view of the fact that the function β (stands for g or g_{Rob}) can degenerate, we regularize it first, and then we define a linearized approximation scheme. We assume that there exists a sequence $\{\beta_k\}_{k=1}^\infty$ and positive real numbers ε, δ such that

$$\left\{ \begin{array}{ll} 0 < \varepsilon, \delta & \\ \frac{\delta}{k} \leq \beta'_k \leq L & \text{a.e. in } \mathbb{R} \\ |\beta(s) - \beta_k(s)| \leq Ck^{-1-\varepsilon} & \forall k \in \mathbb{N}, \forall s \in \mathbb{R} \\ \beta = g, g_{\text{Rob}}. & \end{array} \right. \quad (27)$$

Please note, that the real numbers δ from (26) and (27) are the same. In the case that they should be different, we redefine both by their minimum.

We give a simple example of the regularization to enhance the readability. Let $\alpha > 1$ be a given real number. We define the function β as $\beta(s) = s|s|^{\alpha-1}$ in a neighborhood of 0. Outside this neighborhood β is Lipschitz continuous with some minimal growth condition. The regularization β_k of β can be given as (cf. Fig. 2)

$$\beta_k(s) = \begin{cases} \max \left\{ \beta(s), \frac{\delta s}{k} \right\} & s > 0 \\ \min \left\{ \beta(s), \frac{\delta s}{k} \right\} & s \leq 0. \end{cases} \quad (28)$$

Clearly $\frac{\delta}{k} \leq \beta'_k \leq L$ and one can check that

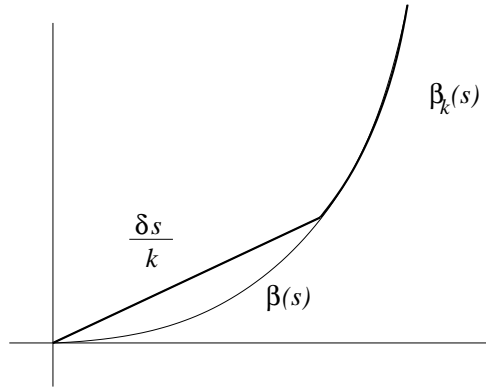


FIGURE 2. Local regularization of β .

$$|\beta(s) - \beta_k(s)| \leq (\alpha - 1)\alpha^{\frac{\alpha}{1-\alpha}} \left(\frac{\delta}{k}\right)^{1+\frac{1}{\alpha-1}} \leq C \left(\frac{1}{k}\right)^{1+\frac{1}{\alpha-1}}.$$

Now, we introduce a linearized scheme, the solution of which should approach the weak solution of (10). First, we replace the nonlinearity $\beta = g, g_{\text{Rob}}$ by its regularization β_k , and then we apply a similar scheme to the regular case (15). Hence, the linear scheme reads as: Find $u_k \in H^1(\Omega)$ such that $u_k - \tilde{g} \in V$ and

$$a(u_k, \varphi) + (Lu_k, \varphi)_\Omega + (Lu_k, \varphi)_{\Gamma_N} = \langle F, \varphi \rangle + (Lu_{k-1}, \varphi)_\Omega - (g_k(u_{k-1}), \varphi)_\Omega + (Lu_{k-1}, \varphi)_{\Gamma_N} - (g_{\text{Rob},k}(u_{k-1}), \varphi)_{\Gamma_N} \tag{29}$$

holds for any $\varphi \in V$.

The well-posedness of the linear elliptic BVP (29) can be proved exactly in the same way as in Lemma 3.1, therefore we omit the details.

Similarly as in (16), we define

$$h_k(s) = g_k(s) - Ls, \quad h_{\text{Rob},k}(s) = g_{\text{Rob},k}(s) - Ls, \quad s \in \mathbb{R}. \tag{30}$$

We subtract (10) from (29) and get the variational formulation for the error of the linearization scheme

$$a(u_k - u, \varphi) + L(u_k - u, \varphi)_\Omega + L(u_k - u, \varphi)_{\Gamma_N} = (g(u) - g_k(u), \varphi)_\Omega + (h_k(u) - h_k(u_{k-1}), \varphi)_\Omega + (g_{\text{Rob}}(u) - g_{\text{Rob},k}(u), \varphi)_{\Gamma_N} + (h_{\text{Rob},k}(u) - h_{\text{Rob},k}(u_{k-1}), \varphi)_{\Gamma_N}, \tag{31}$$

which holds for any $\varphi \in V$.

The following lemma plays a crucial role in the derivation of the error estimates for the approximations u_k .

Lemma 4.1 (algebraic). *Let ε, δ and L be positive real numbers satisfying $0 < \varepsilon < \frac{\delta}{L}$. Assume that $\{y_k\}_{k=0}^\infty$ is a sequence of nonnegative real numbers obeying the following recursion formula*

$$y_k \leq \frac{C}{k^{1+\varepsilon}} + \left(1 - \frac{\delta}{kL}\right) y_{k-1}, \quad k = 1, 2, \dots$$

Then there exists a positive constant $C = C(y_0, \varepsilon, \delta, L)$ such that

$$y_k \leq Ck^{-\varepsilon} \quad k = 1, 2, \dots$$

Proof. Suppose we have a recursion formula of the type

$$y_k \leq a_k + b_k y_{k-1} \quad k = 1, 2, \dots$$

One can prove by induction that

$$y_k \leq a_k + \sum_{j=1}^{k-1} a_j \prod_{i=j+1}^k b_i + y_0 \prod_{i=1}^k b_i \tag{32}$$

holds for all $k \in \mathbb{N}$. The details are left to the reader.

In our case we have $a_k = \frac{C}{k^{1+\varepsilon}}$ and $b_k = 1 - \frac{\delta}{kL}$. Now, we estimate all terms on the right-hand side of (32). We start with an obvious inequality for real numbers

$$1 + x \leq e^x, \quad \forall x \in \mathbb{R},$$

which immediately gives

$$\prod_{i=1}^m (1 + x_i) \leq \exp\left(\sum_{i=1}^m x_i\right), \quad \forall x_i \in \mathbb{R}, x_i \geq -1.$$

Therefore

$$\left\{ \begin{aligned} y_0 \prod_{i=1}^k \left(1 - \frac{\delta}{iL}\right) &\leq y_0 \exp\left(-\frac{\delta}{L} \sum_{i=1}^k \frac{1}{i}\right) \\ &\leq y_0 \exp\left(-\frac{\delta}{L} \int_1^{k+1} \frac{dx}{x}\right) \\ &\leq y_0 \exp\left(-\frac{\delta}{L} \ln k\right) \\ &= y_0 k^{-\frac{\delta}{L}}. \end{aligned} \right. \tag{33}$$

Similarly we estimate also the next term

$$\left\{ \begin{aligned} \sum_{j=1}^{k-1} \frac{C}{j^{1+\varepsilon}} \prod_{i=j+1}^k \left(1 - \frac{\delta}{iL}\right) &\leq C \sum_{j=1}^{k-1} \frac{1}{j^{1+\varepsilon}} \exp\left(-\frac{\delta}{L} \sum_{i=j+1}^k \frac{1}{i}\right) \\ &\leq C \sum_{j=1}^{k-1} \frac{1}{j^{1+\varepsilon}} \exp\left(-\frac{\delta}{L} [\ln(k+1) - \ln(j+1)]\right) \\ &= C \sum_{j=1}^{k-1} \frac{1}{j^{1+\varepsilon}} \left(\frac{j+1}{k+1}\right)^{\frac{\delta}{L}} \\ &\leq C \left(\frac{1}{k}\right)^{\frac{\delta}{L}} \sum_{j=1}^{k-1} j^{\frac{\delta}{L}-1-\varepsilon} \\ &\leq C \left(\frac{1}{k}\right)^{\frac{\delta}{L}} \int_0^k x^{\frac{\delta}{L}-1-\varepsilon} dx \\ &\leq C k^{-\varepsilon}. \end{aligned} \right. \tag{34}$$

Summarizing the relations (32–34) and taking into account the assumption $0 < \varepsilon < \frac{\delta}{L}$, we conclude the proof. \square

Now, we are in a position to derive the error bounds for the linearized scheme (29) in the $H^1(\Omega)$ norm.

Theorem 4.2 ($H^1(\Omega)$ - error). *Let $g, g_{\text{Rob}} \in \mathcal{Q}_0$ along with $0 < \varepsilon < \frac{\delta}{L}$. Moreover, we assume (1, 3–6) and (14). Then there exists a positive constant C such that*

$$\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 + |u_k - u|_{1,\Omega}^2 \leq Ck^{-2\varepsilon}$$

is valid for all $k \in \mathbb{N}$.

Proof. We choose $\varphi = u_k - u \in V$ in (31) and get

$$\begin{aligned} a(u_k - u, u_k - u) + L(u_k - u, u_k - u)_\Omega + L(u_k - u, u_k - u)_{\Gamma_N} \\ = (g(u) - g_k(u), u_k - u)_\Omega + (h_k(u) - h_k(u_{k-1}), u_k - u)_\Omega \\ + (g_{\text{Rob}}(u) - g_{\text{Rob},k}(u), u_k - u)_{\Gamma_N} + (h_{\text{Rob},k}(u) - h_{\text{Rob},k}(u_{k-1}), u_k - u)_{\Gamma_N}. \end{aligned} \quad (35)$$

The functions h_k and $h_{\text{Rob},k}$ belong to the class \mathcal{Q}_0 , thus one can deduce

$$\begin{aligned} \frac{\delta}{k} - L &\leq h'_k(s) = g'_k(s) - L \leq 0 && \text{a.e. in } \mathbb{R} \\ \frac{\delta}{k} - L &\leq h'_{\text{Rob},k}(s) = g'_{\text{Rob},k}(s) - L \leq 0 && \text{a.e. in } \mathbb{R}, \end{aligned}$$

and the relations $|h'_k(s)| \leq L - \frac{\delta}{k}$ and $|h'_{\text{Rob},k}(s)| \leq L - \frac{\delta}{k}$ are valid a.e. in \mathbb{R} .

We estimate the terms on the right-hand side of (35) using the Cauchy and Young inequalities as follows

$$\begin{aligned} |(g(u) - g_k(u), u_k - u)_\Omega| &\leq \|g(u) - g_k(u)\|_{0,\Omega} \|u_k - u\|_{0,\Omega} \\ &\leq \frac{C}{k^{1+\varepsilon}} \|u_k - u\|_{0,\Omega} \\ &= \frac{C\sqrt{k}}{k^{1+\varepsilon}\sqrt{\delta}} \frac{\sqrt{\delta}}{\sqrt{k}} \|u_k - u\|_{0,\Omega} \\ &\leq \frac{C}{k^{1+2\varepsilon}} + \frac{\delta}{2k} \|u_k - u\|_{0,\Omega}^2 \end{aligned}$$

and analogously we deduce

$$\begin{aligned} |(h_k(u) - h_k(u_{k-1}), u_k - u)_\Omega| &\leq \|h_k(u) - h_k(u_{k-1})\|_{0,\Omega} \|u_k - u\|_{0,\Omega} \\ &\leq \left(L - \frac{\delta}{k}\right) \|u - u_{k-1}\|_{0,\Omega} \|u_k - u\|_{0,\Omega} \\ &\leq \left(\frac{L}{2} - \frac{\delta}{2k}\right) \|u - u_{k-1}\|_{0,\Omega}^2 + \left(\frac{L}{2} - \frac{\delta}{2k}\right) \|u_k - u\|_{0,\Omega}^2. \end{aligned}$$

The terms containing the function g_{Rob} can be estimated in the same manner and we obtain

$$|(g_{\text{Rob}}(u) - g_{\text{Rob},k}(u), u_k - u)_\Omega| \leq \frac{C}{k^{1+2\varepsilon}} + \frac{\delta}{2k} \|u_k - u\|_{0,\Gamma_N}^2,$$

and

$$\begin{aligned} |(h_{\text{Rob},k}(u) - h_{\text{Rob},k}(u_{k-1}), u_k - u)_{\Gamma_N}| &\leq \left(\frac{L}{2} - \frac{\delta}{2k}\right) \|u - u_{k-1}\|_{0,\Gamma_N}^2 \\ &\quad + \left(\frac{L}{2} - \frac{\delta}{2k}\right) \|u_k - u\|_{0,\Gamma_N}^2. \end{aligned}$$

We use (12) and estimate the left-hand side of (35) from below by

$$L \|u_k - u\|_{0,\Omega}^2 + L \|u_k - u\|_{0,\Gamma_N}^2 + C_0 |u_k - u|_{1,\Omega}^2.$$

Summarizing the foregoing results we arrive at

$$\frac{L}{2} \left[\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 \right] + C_0 \|u_k - u\|_{1,\Omega}^2 \leq \frac{C}{k^{1+2\varepsilon}} + \left(\frac{L}{2} - \frac{\delta}{2k} \right) \left[\|u - u_{k-1}\|_{0,\Omega}^2 + \|u - u_{k-1}\|_{0,\Gamma_N}^2 \right],$$

which after a simple calculation gives

$$\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 + \frac{2C_0}{L} \|u_k - u\|_{1,\Omega}^2 \leq \frac{C}{k^{1+2\varepsilon}} + \left(1 - \frac{\delta}{kL} \right) \left[\|u - u_{k-1}\|_{0,\Omega}^2 + \|u - u_{k-1}\|_{0,\Gamma_N}^2 \right]. \tag{36}$$

We omit the third term on the left for a moment and obtain the recursion formula

$$\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 \leq \frac{C}{k^{1+2\varepsilon}} + \left(1 - \frac{\delta}{kL} \right) \left[\|u - u_{k-1}\|_{0,\Omega}^2 + \|u - u_{k-1}\|_{0,\Gamma_N}^2 \right].$$

An application of Lemma 4.1 gives

$$\|u_k - u\|_{0,\Omega}^2 + \|u_k - u\|_{0,\Gamma_N}^2 \leq Ck^{-2\varepsilon}.$$

The rest of the proof comes from the last estimate and (36). □

Once we have derived the error estimate in the $H^1(\Omega)$ space, we can also establish the error bound for $\alpha_k = u_k - g_n|_{\Gamma_n}$. The task is a simple consequence of the relation (21) and the foregoing Theorem 4.2, thus we omit the proof.

Theorem 4.3 (α_k -error estimate). *Let the assumptions of Theorem 4.2 be fulfilled. Then there exists a positive constant C such that*

$$|\alpha - \alpha_k| \leq C k^{-\varepsilon}$$

holds for all $k \in \mathbb{N}$.

Our next concern is to estimate the approximation error in the space $L_\infty(\Omega)$.

Theorem 4.4 ($L_\infty(\Omega)$ -error estimate). *Let the assumptions of Theorem 4.2 be satisfied. In addition we suppose $u \in L_\infty(\Omega) \cap L_\infty(\Gamma_N)$. Then there exists a positive constant C such that*

$$\max \left\{ \|u_k - u\|_{L_\infty(\Omega)}, \|u_k - u\|_{L_\infty(\Gamma_N)} \right\} \leq C k^{-\varepsilon}$$

holds for all $k \in \mathbb{N}$.

Proof. We enjoy the fact that we have already proved Theorem 3.4, and therefore we will try to keep the proof as short as possible. We follow the same idea as in Theorem 3.4, thus we point out the differences, only.

First, the definition of the constants A and B changes to

$$\begin{aligned} A &= L^{-1} \|g(u) - g_k(u) + h_k(u) - h_k(u_{k-1})\|_{L_\infty(\Omega)} \\ B &= L^{-1} \|g_{\text{Rob}}(u) - g_{\text{Rob},k}(u) + h_{\text{Rob},k}(u) - h_{\text{Rob},k}(u_{k-1})\|_{L_\infty(\Gamma_N)}. \end{aligned}$$

Instead of (22) we have

$$\begin{aligned}
 0 &= \left(\mathbf{A}_{\text{dif}} \nabla(u_k - u), \nabla(u_k - u + \max\{A, B\})^- \right)_{\Omega} \\
 &+ \left(\mathbf{a}_{\text{con}}(u_k - u), \nabla(u_k - u + \max\{A, B\})^- \right)_{\Omega} \\
 &+ L \left(u_k - u - \frac{g(u) - g_k(u) + h_k(u) - h_k(u_{k-1})}{L}, (u_k - u + \max\{A, B\})^- \right)_{\Omega} \\
 &+ L \left(u_k - u - \frac{g_{\text{Rob}}(u) - g_{\text{Rob},k}(u) + h_{\text{Rob},k}(u) - h_{\text{Rob},k}(u_{k-1})}{L}, (u_k - u + \max\{A, B\})^- \right)_{\Gamma_N} \\
 &= M_1 + M_2 + M_3 + M_4.
 \end{aligned}$$

Further we keep the line of the proof till we get the relation (25).

Since the function g belongs to the class \mathcal{Q}_0 , we estimate

$$\begin{aligned}
 A &= L^{-1} \|g(u) - g_k(u) + h_k(u) - h_k(u_{k-1})\|_{L_{\infty}(\Omega)} \\
 &\leq L^{-1} \left(\|g(u) - g_k(u)\|_{L_{\infty}(\Omega)} + \|h_k(u) - h_k(u_{k-1})\|_{L_{\infty}(\Omega)} \right) \\
 &\leq \frac{C}{k^{1+\varepsilon}} + \left(1 - \frac{\delta}{kL} \right) \|u - u_{k-1}\|_{L_{\infty}(\Omega)} \\
 &\leq \frac{C}{k^{1+\varepsilon}} + \left(1 - \frac{\delta}{kL} \right) \max \left\{ \|u_{k-1} - u\|_{L_{\infty}(\Omega)}, \|u_{k-1} - u\|_{L_{\infty}(\Gamma_N)} \right\}.
 \end{aligned}$$

Analogously we derive

$$\begin{aligned}
 B &= L^{-1} \|g_{\text{Rob}}(u) - g_{\text{Rob},k}(u) + h_{\text{Rob},k}(u) - h_{\text{Rob},k}(u_{k-1})\|_{L_{\infty}(\Gamma_N)} \\
 &\leq L^{-1} \left(\|g_{\text{Rob}}(u) - g_{\text{Rob},k}(u)\|_{L_{\infty}(\Gamma_N)} + \|h_{\text{Rob},k}(u) - h_{\text{Rob},k}(u_{k-1})\|_{L_{\infty}(\Gamma_N)} \right) \\
 &\leq \frac{C}{k^{1+\varepsilon}} + \left(1 - \frac{\delta}{kL} \right) \|u - u_{k-1}\|_{L_{\infty}(\Gamma_N)} \\
 &\leq \frac{C}{k^{1+\varepsilon}} + \left(1 - \frac{\delta}{kL} \right) \max \left\{ \|u_{k-1} - u\|_{L_{\infty}(\Omega)}, \|u_{k-1} - u\|_{L_{\infty}(\Gamma_N)} \right\}.
 \end{aligned}$$

Now, the relation (25) together with the just proved bounds of the constants A and B yield the following recursion formula valid for $k \in \mathbb{N}$

$$\max \left\{ \|u_k - u\|_{L_{\infty}(\Omega)}, \|u_k - u\|_{L_{\infty}(\Gamma_N)} \right\} \leq \frac{C}{k^{1+\varepsilon}} + \left(1 - \frac{\delta}{kL} \right) \max \left\{ \|u_{k-1} - u\|_{L_{\infty}(\Omega)}, \|u_{k-1} - u\|_{L_{\infty}(\Gamma_N)} \right\}.$$

The rest of the proof can be obtained by a simple application of Lemma 4.1. □

5. NUMERICAL EXPERIMENT

Let $\Omega = [0, 1] \times [0, 1]$ be the unit square in \mathbb{R}^2 , the boundary of which is split into three parts: Γ_D (right), Γ_N (top and bottom) and Γ_n (left part of Γ), see Figure 3.

For simplicity we consider the same nonlinear function in the domain and on Γ_n , *i.e.*, $g \equiv g_{\text{Rob}}$, which is defined as

$$g(s) = \begin{cases} 2(s - \sqrt{2}) + 2 & \text{for } s > \sqrt{2} \\ s^2 & \text{for } s \in [0, \sqrt{2}] \\ s & \text{for } s < 0. \end{cases}$$

This is clearly continuous, but there are jumps of the first derivative of g at the points $s = 0, \sqrt{2}$. We will later choose such an exact solution, the range of which will contain the interval $[0, \sqrt{2}]$.

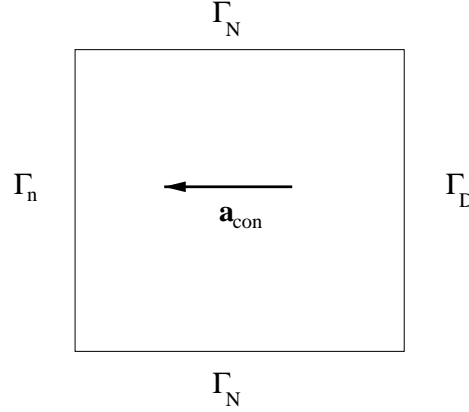


FIGURE 3. Domain Ω with the convection \mathbf{a}_{con} .

The convection term \mathbf{a}_{con} is given as $\mathbf{a}_{\text{con}} = (-1, 0)$, and clearly fulfills the assumption (4). We consider the following nonlinear elliptic BVP: Find $(u, \alpha) \in (H^1(\Omega), \mathbb{R})$ such that

$$\left\{ \begin{array}{ll} \nabla \cdot (-\nabla u - \mathbf{a}_{\text{con}}u) + g(u) = f & \text{in } \Omega \\ u = g_{\text{Dir}} & \text{on } \Gamma_D \\ (-\nabla u - \mathbf{a}_{\text{con}}u) \cdot \boldsymbol{\nu} - g(u) = g_{\text{Neu}} & \text{on } \Gamma_N \\ u(x, y) = -1 - y^2 + \alpha & \text{on } \Gamma_n \\ \int_{\Gamma_n} (-\nabla u - \mathbf{a}_{\text{con}}u) \cdot \boldsymbol{\nu} = -\frac{4}{3}, & \end{array} \right. \quad (37)$$

where the data functions f, g_{Dir} and g_{Neu} are defined in such a way that the exact solution of this BVP is

$$\begin{aligned} u(x, y) &= x^3 - y^2 + x \\ \alpha &= 1. \end{aligned}$$

Thus, we have to solve a linear BVP with a nonlocal BC at Γ_n at each single iteration. The choice of the space V , of all admissible test functions with constant traces on Γ_n , is not standard. Therefore, an application of any usual FE package for the numerical computation of such a problem is not straightforward. Here, we have applied the ideas from [16] and [17] to avoid this difficulty. We explain briefly the main idea.

Suppose, we have to solve the following general linear elliptic BVP at a given iteration k : Find $(u, \alpha) \in (H^1(\Omega), \mathbb{R})$ such that

$$\left\{ \begin{array}{ll} \nabla \cdot (-\tilde{\mathbf{A}}_{\text{dif}}\nabla u - \tilde{\mathbf{a}}_{\text{con}}u) + \tilde{a}_{\text{sou}}u = \tilde{f} & \text{in } \Omega \\ u = \tilde{g}_{\text{Dir}} & \text{on } \Gamma_D \\ (-\tilde{\mathbf{A}}_{\text{dif}}\nabla u - \tilde{\mathbf{a}}_{\text{con}}u) \cdot \boldsymbol{\nu} - \tilde{g}_{\text{Rob}}u = \tilde{g}_{\text{Neu}} & \text{on } \Gamma_N \\ u = \tilde{g}_n + \alpha & \text{on } \Gamma_n \\ G(u) = \int_{\Gamma_n} (-\tilde{\mathbf{A}}_{\text{dif}}\nabla u - \tilde{\mathbf{a}}_{\text{con}}u) \cdot \boldsymbol{\nu} = \tilde{s}. & \end{array} \right. \quad (38)$$

The relation, *e.g.* to the (29), is the following: $\tilde{\mathbf{A}}_{\text{dif}} := \mathbf{A}_{\text{dif}}, \tilde{\mathbf{a}}_{\text{con}} := \mathbf{a}_{\text{con}}, \tilde{a}_{\text{sou}} := L, \tilde{f} := f + Lu_{k-1} - g_k(u_{k-1}), \tilde{g}_{\text{Dir}} := g_{\text{Dir}}, \tilde{g}_{\text{Rob}} := L, \tilde{g}_{\text{Neu}} := g_{\text{Neu}} + Lu_{k-1} - g_{\text{Rob},k}(u_{k-1}), \tilde{g}_n := g_n$ and $\tilde{s} := s$.

We will find the solution in three steps. First, we solve the BVP

$$\left\{ \begin{array}{ll} \nabla \cdot \left(-\tilde{\mathbf{A}}_{\text{dif}} \nabla v - \tilde{\mathbf{a}}_{\text{con}} v \right) + \tilde{a}_{\text{sou}} v = \tilde{f} & \text{in } \Omega \\ v = \tilde{g}_{\text{Dir}} & \text{on } \Gamma_D \\ \left(-\tilde{\mathbf{A}}_{\text{dif}} \nabla v - \tilde{\mathbf{a}}_{\text{con}} v \right) \cdot \boldsymbol{\nu} - \tilde{g}_{\text{Rob}} v = \tilde{g}_{\text{Neu}} & \text{on } \Gamma_N \\ v = \tilde{g}_n & \text{on } \Gamma_n, \end{array} \right. \quad (39)$$

which is similar to (38) apart from the unknown constant α at Γ_n , which has been omitted here. Second BVP to be solved is

$$\left\{ \begin{array}{ll} \nabla \cdot \left(-\tilde{\mathbf{A}}_{\text{dif}} \nabla z - \tilde{\mathbf{a}}_{\text{con}} z \right) + \tilde{a}_{\text{sou}} z = 0 & \text{in } \Omega \\ z = 0 & \text{on } \Gamma_D \\ \left(-\tilde{\mathbf{A}}_{\text{dif}} \nabla z - \tilde{\mathbf{a}}_{\text{con}} z \right) \cdot \boldsymbol{\nu} - \tilde{g}_{\text{Rob}} z = 0 & \text{on } \Gamma_N \\ z = 1 & \text{on } \Gamma_n. \end{array} \right. \quad (40)$$

In the third step, we choose the appropriate value of the real parameter α , for which the total flux through Γ_n fulfills the relation

$$G(u_\alpha) \equiv G(v + \alpha z) = G(v) + \alpha G(z) = s.$$

This implies

$$\alpha = \frac{s - G(v)}{G(z)}. \quad (41)$$

Here, $G(z) \neq 0$, which follows from the following consideration.

Each problem of the type (38) admits at most one solution, which comes from the ellipticity of the corresponding differential operator. Suppose for a moment that $G(z) = 0$. Then, the function z can be seen as a solution of the following problem

$$\left\{ \begin{array}{ll} \nabla \cdot \left(-\tilde{\mathbf{A}}_{\text{dif}} \nabla z - \tilde{\mathbf{a}}_{\text{con}} z \right) + \tilde{a}_{\text{sou}} z = 0 & \text{in } \Omega \\ z = 0 & \text{on } \Gamma_D \\ \left(-\tilde{\mathbf{A}}_{\text{dif}} \nabla z - \tilde{\mathbf{a}}_{\text{con}} z \right) \cdot \boldsymbol{\nu} - \tilde{g}_{\text{Rob}} z = 0 & \text{on } \Gamma_N \\ z = \text{const} & \text{on } \Gamma_n \\ G(z) = \int_{\Gamma_n} \left(-\tilde{\mathbf{A}}_{\text{dif}} \nabla z - \tilde{\mathbf{a}}_{\text{con}} z \right) \cdot \boldsymbol{\nu} = 0. \end{array} \right. \quad (42)$$

Problem (42) has of course a trivial solution which is identically equal to 0. Thus, (42) must have at least two different solutions, which is not possible. Hence, our assumption $G(z) = 0$ was wrong.

Applying the principle of superposition, one can easily see that the function u_α , for α given by (41), solves the BVP (38).

We have used the mixed nonconforming finite element formulation for numerical solution of each linear elliptic BVP. This is equivalent to the mixed-hybrid method (see [2]). We explain briefly the main idea of this approximation.

Consider a regular triangulation \mathcal{T}_h (h stands for the mesh diameter) of the domain Ω . On each element $\mathcal{T} \in \mathcal{T}_h$ we define three linear basis functions associated with edges of \mathcal{T} , *i.e.*, a basis function has the value 1 at the midpoint of one edge and vanishes at the midpoints of the other edges of one triangle. Further we define a bubble function on \mathcal{T} , which is a polynomial function of third order vanishing on the boundary $\partial\mathcal{T}$, such that its integral average value on \mathcal{T} is 1. In this way we have enriched the standard linear nonconforming space by

bubbles, and we solve a linear elliptic problem in this space replacing the velocity field \mathbf{q} by its projection on the Raviart-Thomas space RT_0 . For more details see [2].

The linearization process has already been described in previous sections. We begin with the starting data u_0 satisfying (14). We present two computations. In the first one, we start relatively close to the solution u , more exactly we have taken

$$u_0(\mathbf{x}) = u(\mathbf{x})(1 + 0.2 \operatorname{ran}(\mathbf{x})),$$

where ran is a random function whose range is uniformly distributed over $(-1, 1)$. This situation should model the computation of a fixed time step of an evolution process after the time discretization. Therefore, one can assume that the initial data for iterations, which is normally given from the last time step computation, is sufficiently close to the solution. We have chosen a small modification (up to 20% error) of u as u_0 . In the second case, we start from u_0 , which is far away from the solution u , *i.e.*,

$$u_0(\mathbf{x}) = 100 \operatorname{ran}(\mathbf{x}).$$

Let us note that the random function ran has been evaluated once per a given triangle or an edge.

If u_0 would have been chosen as a sub- or a super-solution to the BVP (37), then the sequence $\{u_k\}_{k=0}^{\infty}$ would be monotone – see, *e.g.*, [14]. We note that the iterations in our computations cannot be monotone, because of the random choice of u_0 .

We have used a fixed uniform mesh consisting of 5 000 triangles, which corresponds to $\Delta x = \Delta y = 0.02$, and we have computed 25 iterations in both cases. Then we have evaluated various errors of u_k and plotted them *versus* iterations $k = 1, \dots, 25$. In order to get a better feeling about the rate of convergence, we have depicted *logarithms* of errors instead of errors on the y -axes. The results can be seen in Figures 4 to 8. Here, the left column represents the case for a good starting point u_0 , while on the right, there are the pictures corresponding to very badly chosen u_0 .

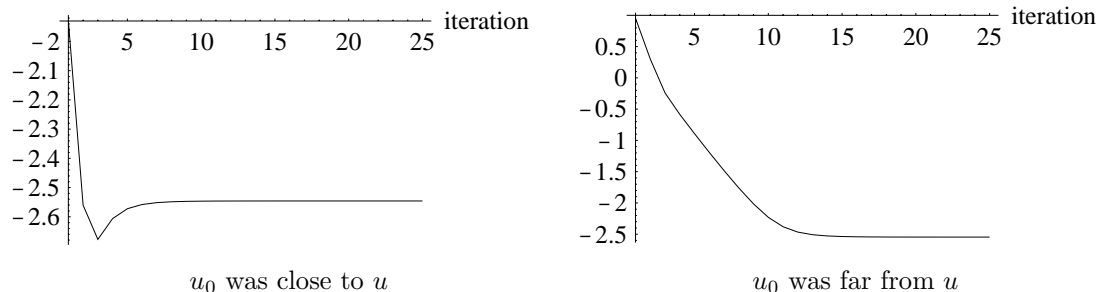


FIGURE 4. Logarithms of $L_2(\Omega)$ -errors for u_k *versus* iterations.

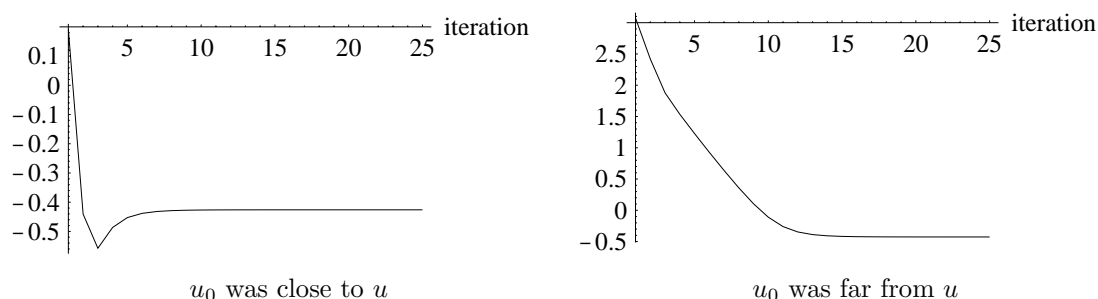


FIGURE 5. Logarithms of relative $L_2(\Omega)$ -errors for u_k *versus* iterations.

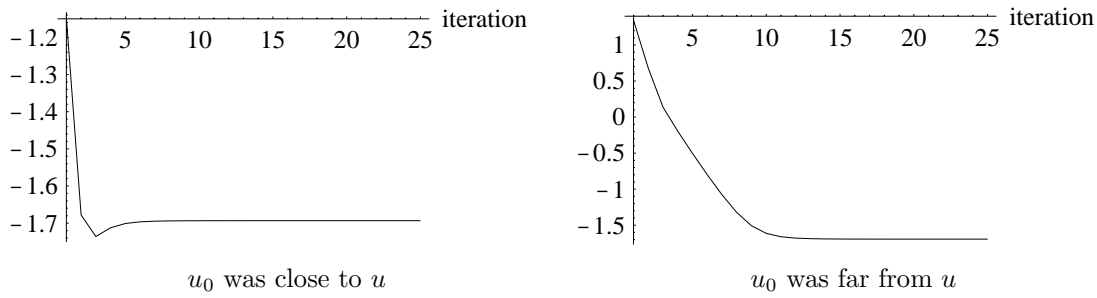


FIGURE 6. Logarithms of $L_2(\Omega)$ -errors for $\mathbf{q}_k = -\nabla u_k - \mathbf{a}_{\text{con}} u_k$ versus iterations.

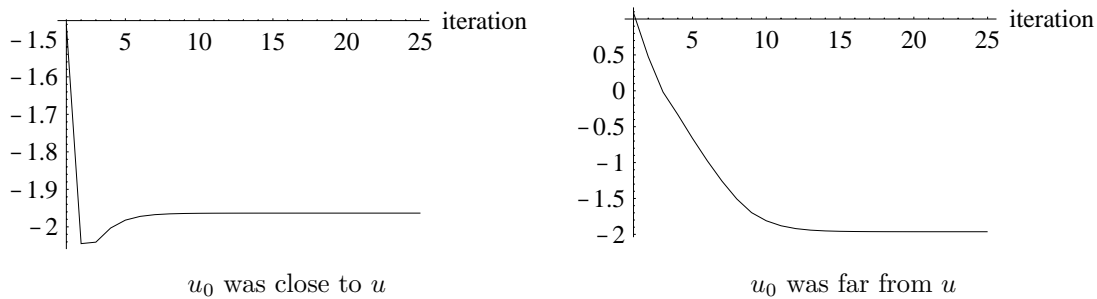


FIGURE 7. Logarithms of $L_\infty(\Omega)$ -errors for u_k versus iterations.

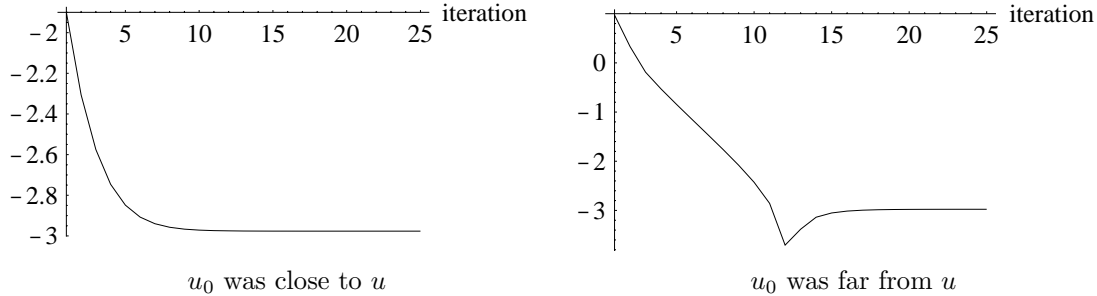


FIGURE 8. Logarithms of α_k -errors versus iterations.

The behavior of all graphs is similar. First, we observe a rapidly decreasing part of the curve, later the graph changes its behavior and becomes more or less constant. This can be easily explained. The resulting error consists of two parts: the linearization and the discretization error. At the beginning of the iteration process, the linearization error is superior to the discretization one. Later the opposite becomes true. Therefore, it makes no sense to iterate when the discretization error becomes dominant. Of course, the discretization error can be diminished by taking smaller mesh diameter h . The monotonically decreasing behavior of error can also be violated at the point where the discretization error becomes dominant to the linearization error. Recall, that the relation (20) shows the monotonicity of the error, but the discretization error has been neglected there.

Summarizing our results, we can see that the proposed linearization scheme is robust and it works even in the case when the iterations start far away from the real solution. In particular, we have needed 12 iterations in our example to get the best possible error for the given discretization. In the instance of a good starting point u_0 , it is enough to do 3 – 5 iterations to achieve the discretization error, which makes the scheme very efficient.

The robustness of the scheme allows the use of large time steps in the computation of evolution problems. The reason is, that the convergence of approximations at each time point of any time partitioning is independent of the time step.

REFERENCES

- [1] D. Andreucci and R. Gianni, Global existence and blow up in a parabolic problem with nonlocal dynamical boundary conditions. *Adv. Differ. Equ.* **1** (1996) 729–752.
- [2] D.N. Arnold and F. Brezzi, Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér.* **19** (1985) 7–32.
- [3] J.H. Bramble and P. Lee, On variational formulations for the Stokes equations with nonstandard boundary conditions. *RAIRO Modél. Math. Anal. Numér.* **28** (1994) 903–919.
- [4] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Math. Stud. **5**, Notas de matemática **50**, North-Holland Publishing Comp., Amsterdam, London; American Elsevier Publishing Comp. Inc., New York (1973).
- [5] H. De Schepper and M. Slodička, Recovery of the boundary data for a linear 2nd order elliptic problem with a nonlocal boundary condition. *ANZIAM J.* **42E** (2000) C488–C505. ISSN 1442-4436 (formerly known as *J. Austral. Math. Soc., Ser. B*).
- [6] L.C. Evans, Partial differential equations, *Graduate Studies in Mathematics* **19**, American Mathematical Society (1998).
- [7] A. Friedman, *Variational principles and free-boundary problems*. Wiley, New York (1982).
- [8] H. Gerke, U. Hornung, Y. Kelanemer, M. Slodička and S. Schumacher, Optimal Control of Soil Venting: Mathematical Modeling and Applications, *ISNM* **127**, Birkhäuser, Basel (1999).
- [9] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*. Springer, Berlin, Heidelberg (1983).
- [10] W. Jäger and J. Kačur, Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes. *RAIRO Modél. Math. Anal. Numér.* **29** (1995) 605–627.
- [11] J. Kačur, Solution to strongly nonlinear parabolic problems by a linear approximation scheme. *IMA J. Numer. Anal.* **19** (1999) 119–145.
- [12] C.V. Pao, *Nonlinear parabolic and elliptic equations*. Plenum Press, New York (1992).
- [13] R. Rannacher and S. Turek, Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations. *Internat. J. Numer. Methods Fluids* **22** (1996) 325–352.
- [14] M. Slodička, A monotone linear approximation of a nonlinear elliptic problem with a non-standard boundary condition, in *Algoritmy 2000*, A. Handlovičová, M. Komorníková, K. Mikula and D. Ševčovič, Eds., Bratislava (2000) 47–57.
- [15] M. Slodička and H. De Schepper, On an inverse problem of pressure recovery arising from soil venting facilities. *Appl. Math. Comput.* (to appear).
- [16] M. Slodička and H. De Schepper, A nonlinear boundary value problem containing nonstandard boundary conditions. *Appl. Math. Comput.* (to appear).
- [17] M. Slodička and R. Van Keer, A nonlinear elliptic equation with a nonlocal boundary condition solved by linearization. *Internat. J. Appl. Math.* **6** (2001) 1–22.
- [18] R. Van Keer, L. Dupré and J. Melkebeek, Computational methods for the evaluation of the electromagnetic losses in electrical machinery. *Arch. Comput. Methods Engrg.* **5** (1999) 385–443.