64 (2019) APPLICATIONS OF MATHEMATICS No. 2, 129–167

ON THE EFFECT OF NUMERICAL INTEGRATION IN THE FINITE ELEMENT SOLUTION OF AN ELLIPTIC PROBLEM WITH A NONLINEAR NEWTON BOUNDARY CONDITION

Ondřej Bartoš, Miloslav Feistauer, Filip Roskovec, Praha

Received July 11, 2018. Published online March 20, 2019.

Abstract. This paper is concerned with the analysis of the finite element method for the numerical solution of an elliptic boundary value problem with a nonlinear Newton boundary condition in a two-dimensional polygonal domain. The weak solution loses regularity in a neighbourhood of boundary singularities, which may be at corners or at roots of the weak solution on edges. The main attention is paid to the study of error estimates. It turns out that the order of convergence is not dampened by the nonlinearity if the weak solution is nonzero on a large part of the boundary. If the weak solution is zero on the whole boundary, the nonlinearity only slows down the convergence of the function values but not the convergence of the gradient. The same analysis is carried out for approximate solutions obtained by numerical integration. The theoretical results are verified by numerical experiments.

Keywords: elliptic equation; nonlinear Newton boundary condition; weak solution; finite element discretization; numerical integration; error estimation; effect of numerical integration

MSC 2010: 65N30, 65N15, 65D30

Introduction

There are many numerical techniques for solving partial differential equations. The effectivity of the respective methods is often closely related to the properties of the equations in question. We are concerned with the study of the finite element method (FEM) for the solution of an elliptic equation with a nonlinear Newton boundary condition in a bounded two-dimensional polygonal domain with numerical integration. Such boundary value problems have applications in science and engineering, see [13], [2]. We suppose that the nonlinear term has a general "polynomial"

This research was supported by the grant 17-01747S of the Czech Science Foundation.

growth. This can be found in the modelling of electrolysis of aluminium with the aid of the stream function. The nonlinear boundary condition describes turbulent flow in a boundary layer ([21]). Similar nonlinearity appears in a radiation heat transfer problem ([20], [19]) or in nonlinear elasticity ([15], [14]). A parabolic equation with a nonlinear Newton boundary condition is solved with the use of finite elements in [5] and [24], but the growth of the nonlinearity is only linear.

Paper [7] deals with the problem arising in the investigation of the electrolytical production of aluminium. The problem is discretized by piecewise linear conforming triangular elements and the effect of the numerical integration applied to this problem is investigated in [8]. Using monotone operator theory in [12] and assuming regularity of the weak solution, paper [9] gives error estimates. Paper [10] investigates this problem using discontinuous Galerkin method and piecewise polynomial functions, but does not consider the effect of numerical integration.

In this paper we study an elliptic boundary value problem with nonlinear Newton boundary condition in a polygonal domain. The goal is to analyse both FEM used on conforming shape regular meshes with piecewise polynomial functions and the effect of numerical integration while considering the actual regularity of the weak solution. In Section 1 the boundary value problem is introduced, the weak solution is defined and some auxiliary results are introduced. In Section 2 the finite element approximation of the weak solution is introduced and some properties of the discrete problem are proved. It turns out that the order of convergence depends on whether the exact weak solution is zero on the boundary or not. Section 3 is devoted to the discretization with numerical integration and some important estimates are proved. Section 4 is concerned with abstract error estimates under the aplication of numerical integration. Section 5 is devoted to the analysis of the boundedness of interpolated functions. These results are used in Section 6 which is devoted to error estimation in terms of the size of the triangulation. Finally, Section 7 supports theoretical results by numerical experiments.

1. Formulation of the continuous problem

We denote the set of real numbers by $\mathbb{R}$, the set of positive integers by $\mathbb{N}$, and the set of non-negative integers by $\mathbb{N}_0$. Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain with Lipschitz continuous boundary $\partial\Omega$. We consider a boundary value problem with nonlinear Newton boundary condition: find $u\colon \overline{\Omega} \to \mathbb{R}$ such that

$$-\Delta u = f \quad \text{in } \Omega, \tag{1.1}$$

$$\frac{\partial u}{\partial n} + \kappa |u|^{\alpha} u = \varphi \quad \text{on } \partial\Omega \tag{1.2}$$

with given functions $f\colon \Omega \to \mathbb{R}$, $\varphi\colon \partial\Omega \to \mathbb{R}$ and constants $\kappa > 0$, $\alpha \geqslant 0$. By a classical solution of (1.1) with boundary conditions (1.2) we understand a function $u \in C^2(\overline{\Omega})$ satisfying (1.1) pointwise at every point in Ω and satisfying (1.2) at every point on $\partial\Omega$ such that the outer normal unit vector n is defined.

In what follows we use standard notation of function spaces: $C^k(\overline{\Omega})$, $C^{k,\lambda}(\overline{\Omega})$, $C^\infty(\overline{\Omega})$, $L^p(\Omega)$, $L^p(\partial\Omega)$, $W^{k,p}(\Omega)$, $W^{k,p}(\partial\Omega)$. We denote the following norms in the spaces $L^p(\Omega), L^p(\partial\Omega)$ with $p \geqslant 1$ by

$$\|f\|_{0,p,\Omega} = \left(\int_\Omega |f|^p \,\mathrm{d}x\right)^{1/p}, \quad \|f\|_{0,p,\partial\Omega} = \left(\int_{\partial\Omega} |f|^p \,\mathrm{d}S\right)^{1/p}.$$

For $p \in [1,\infty)$ and $k \in \mathbb{N}$ we consider Sobolev spaces $W^{k,p}(\Omega)$, $W^{k,p}(\partial\Omega)$ with seminorms

$$|f|_{k,p,\Omega} = \left(\sum_{|\beta|=k} \int_\Omega |D^\beta f|^p \,\mathrm{d}x\right)^{1/p}, \quad |f|_{k,p,\partial\Omega} = \left(\sum_{|\beta|=k} \int_{\partial\Omega} |D^\beta f|^p \,\mathrm{d}S\right)^{1/p},$$

where $\beta = (\beta_1, \beta_2)$ is a multi-index with $|\beta| = \beta_1 + \beta_2$, and the norms

$$\|f\|_{k,p,\Omega} = \left(\sum_{|\beta|\leqslant k} \int_\Omega |D^\beta f|^p \,\mathrm{d}x\right)^{1/p}, \quad \|f\|_{k,p,\partial\Omega} = \left(\sum_{|\beta|\leqslant k} \int_{\partial\Omega} |D^\beta f|^p \,\mathrm{d}S\right)^{1/p}.$$

For $k \in \mathbb{N}$ and $p > 1$ we denote by $W^{k-1/p,p}(\partial\Omega)$ the space of traces from $W^{k,p}(\Omega)$ with the norm

$$\|f|_{\partial\Omega}\|_{k-1/p,p,\partial\Omega} = \inf\{\|g\|_{k,p,\Omega};\ g \in W^{k,p}(\Omega), g|_{\partial\Omega} = f|_{\partial\Omega}\}.$$

We also denote $W^{k,2}(\Omega) = H^k(\Omega)$ and $W^{0,p}(\Omega) = L^p(\Omega)$. The following continuous embeddings known as Sobolev embeddings hold for domains $\Omega \subset \mathbb{R}^n$ (in our case $n = 2$) with Lipschitz continuous boundaries (see Section 5.6 in [6]):

$$\begin{aligned} &W^{1,p}(\Omega) \hookrightarrow L^{np/(n-p)}(\Omega), && p \in [1,n), \\ &W^{1,n}(\Omega) \hookrightarrow L^q(\Omega), && q \in [1,\infty), \\ &W^{1,p}(\Omega) \hookrightarrow C^{0,1-n/p}(\overline{\Omega}), && p \in (n,\infty), \\ &W^{n,1}(\Omega) \hookrightarrow C(\overline{\Omega}). \end{aligned} \tag{1.3}$$

The following continuous trace embeddings also hold for domains with Lipschitz continuous boundaries (see Section 5.5 in [6] or Theorems 1.4.4.1 and 1.5.1.1 in [16]):

$$\begin{aligned} &W^{1,p}(\Omega) \hookrightarrow L^{(n-1)p/(n-p)}(\partial\Omega), && p \in [1,n), \\ &W^{1,n}(\Omega) \hookrightarrow L^q(\partial\Omega), && q \in [1,\infty), \\ &W^{1,p}(\Omega) \hookrightarrow C^{0,1-n/p}(\partial\Omega), && p \in (n,\infty), \\ &W^{n,1}(\Omega) \hookrightarrow C(\partial\Omega). \end{aligned} \tag{1.4}$$

If $G \subset \partial\Omega$, then by $|G|$ we denote the one-dimensional measure defined on $\partial\Omega$ of the set G. By 5.8.1 in [6] the following result holds.

Theorem 1.1 (Poincaré inequality)**.** *Let Ω be a domain with a Lipschitz continuous boundary. Let $u \in W^{1,p}(\Omega)$. Let $G \subset \partial\Omega$ with $|G| > 0$. Then there exists a constant $c_P > 0$ dependent on Ω, G and p such that*

$$\|u\|_{1,p,\Omega} \leqslant c_P(|u|_{1,p,\Omega} + \|u\|_{0,p,G}). \tag{1.5}$$

Now we introduce the concept of a weak solution. Let

$$f \in L^2(\Omega), \quad \varphi \in L^2(\partial\Omega). \tag{1.6}$$

We introduce the following forms for u, $v \in H^1(\Omega)$:

$$\begin{gathered} b(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad d(u,v) = \kappa \int_{\partial\Omega} |u|^\alpha uv \,\mathrm{d}S, \\ L^\Omega(v) = \int_\Omega fv \,\mathrm{d}x, \quad L^{\partial\Omega}(v) = \int_{\partial\Omega} \varphi v \,\mathrm{d}S, \\ L(v) = L^\Omega(v) + L^{\partial\Omega}(v), \quad a(u,v) = b(u,v) + d(u,v). \end{gathered} \tag{1.7}$$

Definition 1.1. We say that a function $u\colon \Omega \to \mathbb{R}$ is the weak solution of problem (1.1)–(1.2) if

$$\begin{gathered} u \in H^1(\Omega), \\ a(u,v) = L(v) \quad \forall\, v \in H^1(\Omega). \end{gathered} \tag{1.8}$$

The existence and uniqueness of the weak solution is a consequence of properties of the form a. Let us note that

$$a(u, u-v) - a(v, u-v) = \int_\Omega |\nabla u - \nabla v|^2 \,\mathrm{d}x + \kappa \int_{\partial\Omega} (|u|^\alpha u - |v|^\alpha v)(u-v) \,\mathrm{d}S. \tag{1.9}$$

In [9] it was proved that

$$(|\eta|^\alpha \eta - |\xi|^\alpha \xi)(\eta - \xi) \geqslant 2^{-\alpha} |\eta - \xi|^{\alpha+2}, \quad \eta, \xi \in \mathbb{R},\ \alpha \geqslant 0, \tag{1.10}$$

from which the following lemma follows.

Lemma 1.1. *Let $u, v \in H^1(\Omega)$. Then*

$$a(u, u-v) - a(v, u-v) \geqslant |u-v|^2_{1,2,\Omega} + \kappa 2^{-\alpha} \|u-v\|^{\alpha+2}_{0,\alpha+2,\partial\Omega}. \tag{1.11}$$

In [9] and [8], most of the following theorem was also proved.

Lemma 1.2. *The following assertions hold.*

(a) *L is a continuous linear functional on $H^1(\Omega)$.*

(b) *The functional $a(u,\cdot)$ from $H^1(\Omega)$ into $\mathbb{R}$ is continuous and linear for every $u \in H^1(\Omega)$.*

(c) *a is uniformly monotone:*

$$a(u,u-v) - a(v,u-v) \geqslant \varrho(\|u-v\|_{1,2,\Omega}) \quad \forall\, u,v \in H^1(\Omega), \tag{1.12}$$

where

$$\varrho(t) = \begin{cases} C_0\kappa 2^{-\alpha} t^{\alpha+2} & \text{for } 0 \leqslant t \leqslant 2\kappa^{-1/\alpha}, \\ C_0 t^2 & \text{for } t \geqslant 2\kappa^{-1/\alpha}. \end{cases} \tag{1.13}$$

For $\alpha = 0$ we set $\kappa^{-1/\alpha} = 0$.

(d) *The functional $a(\cdot,v)$ from $H^1(\Omega)$ into $\mathbb{R}$ is continuous for every $v \in H^1(\Omega)$ in the following sense: There exists a positive constant $C_1 > 0$ independent of v such that*

$$|a(u,v) - a(w,v)| \leqslant C_1(1 + \|u\|^\alpha_{1,2,\Omega} + \|w\|^\alpha_{1,2,\Omega})\|u-w\|_{1,2,\Omega}\|v\|_{1,2,\Omega} \tag{1.14}$$

for all $u, w \in H^1(\Omega)$.

(e) *The form $a(u,u)$ is coercive in the following sense: There exists a positive constant $C_2 > 0$ such that*

$$a(u,u) \geqslant C_2\|u\|^2_{1,2,\Omega} \tag{1.15}$$

holds for all $u \in H^1(\Omega)$ such that $\|u\|_{1,2,\Omega} \geqslant 1$.

P r o o f. Assertions (a), (b), (c), (e) were proved in [8] and [9]. It remains to prove the part (d). We have

$$|a(u,v) - a(w,v)| \leqslant \left|\int_\Omega \nabla(u-w)\cdot\nabla v\,\mathrm{d}S\right| + \left|\kappa\int_{\partial\Omega}(|u|^\alpha u - |w|^\alpha w)v\,\mathrm{d}x\right|.$$

The Cauchy inequality applied to the first term yields

$$\left|\int_\Omega \nabla(u-w)\cdot\nabla v\,\mathrm{d}S\right| \leqslant |u-w|_{1,2,\Omega}|v|_{1,2,\Omega}.$$

The second term can be estimated using the relation

$$|u|^\alpha u - |w|^\alpha w = \int_w^u \frac{\mathrm{d}}{\mathrm{d}t}(|t|^\alpha t)\,\mathrm{d}t = (\alpha+1)\int_w^u |t|^\alpha\,\mathrm{d}t.$$

The function $|t|^\alpha$ of $t \in \mathbb{R}$ is monotone in $(-\infty, 0)$ and in $(0, \infty)$ and its global minimum is reached for $t = 0$. Hence,

$$|t|^\alpha \leqslant (|u|^\alpha + |w|^\alpha), \quad t \in [u, w].$$

Take any $p_1, p_2, p_3 > 1$ such that $1/p_1 + 1/p_2 + 1/p_3 = 1$. Then these relations and the Hölder inequality imply that

$$\begin{aligned}\left|\kappa \int_{\partial\Omega} (|u|^\alpha u - |w|^\alpha w)v\,\mathrm{d}S\right| &\leqslant \kappa(\alpha+1)\int_{\partial\Omega}(|u|^\alpha + |w|^\alpha)|u-w||v|\,\mathrm{d}S \\ &\leqslant \kappa(\alpha+1)(\|u\|^\alpha_{0,\alpha p_1,\partial\Omega} + \|w\|^\alpha_{0,\alpha p_1,\partial\Omega})\|u-w\|_{0,p_2,\partial\Omega}\|v\|_{0,p_3,\partial\Omega}.\end{aligned}$$

The trace embedding (1.4) completes the proof of (1.14). □

It follows from the monotone operator theory [12] and properties in Lemma 1.2 that problem (1.8) has exactly one solution.

In the error estimates, the regularity of the weak solution will play an important role. In [10] the following results are proved.

Theorem 1.2. *Let $u \in H^1(\Omega)$ be a weak solution of* (1.8) *in a polygonal domain Ω. By ω_0 we denote the largest inner angle in Ω. Let $f \in L^q(\Omega)$, $\varphi \in W^{1-1/q,q}(\partial\Omega)$, where*

$$\begin{aligned} &q = 1 + \frac{\pi}{2\omega_0 - \pi} - \varepsilon < 2 \quad \text{for } \omega_0 > \pi, \\ &q = 1 + \frac{\pi}{2\omega_0 - \pi} - \varepsilon > 2 \quad \text{for } \frac{\pi}{2} < \omega_0 < \pi, \\ &q \geqslant 1 \text{ is arbitrary} \qquad\quad \text{for } \omega_0 \leqslant \frac{\pi}{2}, \end{aligned} \tag{1.16}$$

and $\varepsilon > 0$ is arbitrarily small. Then $u \in W^{2,q}(\Omega)$.

Since all inner angles ω in Ω are less than 2π, we shall consider

$$q > \frac{4}{3}. \tag{1.17}$$

Now we introduce some auxiliar results.

Lemma 1.3. *Let us assume that $u \in W^{k,q}(\Omega)$, $k \in \mathbb{N}$, $q \geqslant 1$, and $\beta = (\beta_1, \beta_2)$ is a multi-index with $\beta_1, \beta_2 \in \mathbb{N}_0$ such that $|\beta| = \beta_1 + \beta_2 \leqslant k$. Then*

$$D^\beta(|u|^\alpha u) = \frac{\partial^{|\beta|}(|u|^\alpha u)}{\partial x_1^{\beta_1} \partial x_2^{\beta_2}}$$

can be expressed as a finite sum of terms of a form

$$c|u|^{\alpha+1-J} \operatorname{sgn} u^{J+1} \prod_{j=1}^{J} D^{\gamma_j} u, \tag{1.18}$$

where $J \in \mathbb{N}_0$ and γ_j, $j = 1, \ldots, J$ are multi-indices such that $\sum_{j=1}^{J} \gamma_j = \beta$. Here, the constant c is dependent on α, β and multi-indices γ_j. If $\alpha \in \mathbb{N}_0$, then $D^\beta(|u|^\alpha u)$ only contains terms with non-negative exponent of $|u|$, i.e. $c = 0$ if $\alpha + 1 - J$ is a negative integer.

P r o o f. Let k, q be given. We will proceed using induction on $|\beta|$. If $|\beta| = 0$, then the only possible term has $J = 0$, $c = 1$ and $\prod_{j=1}^{0} D^{\gamma_j} u = 1$. If $|\beta| = 1$, then $c = \alpha + 1$, $J = 1$, and either $\gamma_1 = (1, 0)$ or $\gamma_1 = (0, 1)$.

Suppose that the lemma holds for all multi-indices with length smaller than $|\beta|$. In particular, we have

$$D^\beta(|u|^\alpha u) = \frac{\partial(D^{\beta'}(|u|^\alpha u))}{\partial x_i}$$

for some $i \in \{1, 2\}$ and β' such that $|\beta| = |\beta'| + 1$. Then we only need to apply $\partial/\partial x_i$ to the terms $c|u|^{\alpha+1-J'} \operatorname{sgn} u^{J'+1} \prod_{j=1}^{J'} D^{\gamma'_j} u$ which have $\sum_{j=1}^{J'} \gamma'_j = \beta'$. If the partial derivative $\partial/\partial x_i$ is applied to any factor in $\prod_{j=1}^{J'} D^{\gamma'_j} u$, then the resulting term does have the desired form with $J = J'$, one of the multi-indices γ'_j increased, and $\sum_{j=1}^{J} \gamma_j = \beta$. If $\partial/\partial x_i$ is applied to $|u|^{\alpha+1-J'}$, then the resulting term has $J = J' + 1$, $\sum_{j=1}^{J'} \gamma'_j + \gamma_{J'+1} = \beta$, where $\gamma_{J'+1}$ is either $(1, 0)$ or $(0, 1)$ depending on x_i, and therefore also has the desired form.

Suppose that $\alpha \in \mathbb{N}_0$. Then the exponent $\alpha + 1 - J$ in $|u|^{\alpha+1-J}$ is integer for any J. The only possibility to obtain a negative exponent in the induction step would be to apply $\partial/\partial x_i$ to $|u|^{\alpha+1-J'}$ for J' such that $\alpha + 1 - J' \in [0, 1)$, i.e. $\alpha + 1 - J' = 0$. But then $\partial|u|^0/\partial x_i = 0$ and the constant c would in fact be zero. □

Lemma 1.4. *Let $u \in W^{k,q}(\Omega)$, where $k \geqslant 2$ is an integer, $q > 1$ and let $\alpha + 1 \geqslant k$ or $\alpha \in \mathbb{N}_0$. Then $|u|^\alpha u|_{\partial\Omega} \in W^{k-1/q,q}(\partial\Omega)$ and the estimate*

$$\||u|^\alpha u\|_{k-1/q,q,\partial\Omega} \leqslant c\|u\|^{\alpha+1}_{k,q,\Omega} \tag{1.19}$$

with a constant $c > 0$ dependent on Ω, k, q, α, holds.

P r o o f. We will prove that $|u|^\alpha u \in W^{k,q}(\Omega)$. Consider any multi-index $\beta = (\beta_1, \beta_2)$ such that $|\beta| = \beta_1 + \beta_2 \leqslant k$. Our goal is to show that

$$D^\beta(|u|^\alpha u) = \frac{\partial^{|\beta|}(|u|^\alpha u)}{\partial x_1^{\beta_1} \partial x_2^{\beta_2}} \in L^q(\Omega).$$

The expression $D^\beta(|u|^\alpha u)$ is a sum of several terms of the form (1.18) given in Lemma 1.3. Due to the triangle inequality in Lebesgue spaces, we only need to show that all of these terms belong to the space $L^q(\Omega)$ and are estimated by the right-hand side of (1.19). The assumption $\alpha + 1 \geqslant k$ or $\alpha \in \mathbb{N}_0$ guarantees that the exponents $\alpha + 1 - J$ in (1.18) are non-negative for all terms. Since $u \in W^{k,q}(\Omega) \hookrightarrow C^{k-2}(\overline{\Omega})$, we can trivially estimate the terms which only have derivatives of orders up to $k-2$:

$$\left\| |u|^{\alpha+1-J} \prod_{j=1}^{J} D^{\gamma_j} u \right\|_{0,q,\Omega} \leqslant c\|u\|^{\alpha+1}_{k,q,\Omega}.$$

Consider the term $c|u|^\alpha D^\beta u$. Since $u \in W^{k,q}(\Omega) \hookrightarrow C(\overline{\Omega})$ and $D^\beta u \in L^q(\Omega)$, we have

$$\||u|^\alpha D^\beta u\|^q_{0,q,\Omega} = \int_\Omega |u|^{\alpha q} |D^\beta u|^q \,\mathrm{d}x \leqslant \|u\|^{\alpha q}_{C(\overline{\Omega})} \int_\Omega |D^\beta u|^q \,\mathrm{d}x \leqslant c\|u\|^{\alpha q+q}_{k,q,\Omega}.$$

The only remaining terms are $c|u|^{\alpha-1} \prod_{j=1}^{2} D^{\gamma_j} u$, where γ_1 has length 1 and γ_2 has length $k-1$. If $k \geqslant 3$, then we again estimate

$$\left\| |u|^{\alpha-1} \prod_{j=1}^{2} D^{\gamma_j} u \right\|^q_{0,q,\Omega} \leqslant \|u\|^{(\alpha-1)q}_{C(\overline{\Omega})} \||\nabla u|\|^q_{C(\overline{\Omega})} \int_\Omega |D^{\gamma_2} u|^q \,\mathrm{d}x \leqslant c\|u\|^{\alpha q+q}_{k,q,\Omega}.$$

If $k = 2$, then γ_2 has length 1, and we use embedding (1.3) to get

$$\left\| |u|^{\alpha-1} \prod_{j=1}^{2} D^{\gamma_j} u \right\|^q_{0,q,\Omega} \leqslant \|u\|^{(\alpha-1)q}_{C(\overline{\Omega})} \||\nabla u|\|^{2q}_{0,2q,\Omega} \leqslant c\|u\|^{\alpha q+q}_{k,q,\Omega},$$

where the last inequality was obtained because

▷ $W^{1,q}(\Omega) \hookrightarrow C(\overline{\Omega})$ for $q > 2$,
▷ $H^1(\Omega) \hookrightarrow L^4(\Omega)$ for $q = 2$,
▷ $W^{1,q}(\Omega) \hookrightarrow L^{2q}(\Omega)$ for $q \in [1, 2)$ as $2q/(2-q) \geqslant 2q$.

Combining these inequalities, we conclude that $|u|^\alpha u \in W^{k,q}(\Omega)$ and $\||u|^\alpha u\|_{k,q,\Omega} \leqslant c\|u\|^{\alpha+1}_{k,q,\Omega}$. The trace $|u|^\alpha u|_{\partial\Omega}$ therefore satisfies (1.19). □

Functions in $W^{2,q}(\Omega)$ are continuous. Therefore, it is possible to distinguish on which parts of the boundary $\partial\Omega$ is the weak solution u nonzero.

Lemma 1.5. *Let $u \in W^{k,q}(\Omega)$, where $k \in \mathbb{N}$, $k \geqslant 2$, $q > 1$ and Ω is a polygonal domain. Let $\alpha + 1 < k$. Let G be a closed subset of $\partial\Omega$. If $|G| > 0$ and $|u| > \varepsilon > 0$ on G, then $|u|^\alpha u|_G \in W^{k-1/q,q}(G)$.*

P r o o f. Function u is continuous in Ω. Therefore, we can find an open neighbourhood of G in Ω denoted by Ω_G such that $|u| > \varepsilon > 0$ in Ω_G. We can proceed similarly to the proof of Lemma 1.4. This time we cannot guarantee that the exponents $\alpha + 1 - J$ are non-negative. If $\alpha + 1 - J < 0$, then we cannot use the estimate $|u|_{\Omega_G}|^{\alpha+1-J} \leqslant \|u\|^{\alpha+1-J}_{C(\overline{\Omega})}$. However, it can be replaced by the inequality $|u|_{\Omega_G}|^{\alpha+1-J} \leqslant \varepsilon^{\alpha+1-J}$. The lowest possible negative exponent is $\alpha + 1 - k$ and the same arguments as in the proof of Lemma 1.4 lead to the estimate

$$\||u|^\alpha u\|_{k,q,\Omega_G} \leqslant c(\|u\|^{\alpha+1}_{k,q,\Omega_G} + \varepsilon^{\alpha+1-k}\|u\|^k_{k,q,\Omega_G}),$$

where c depends also on Ω_G and possibly on both G and u. □

2. Discretization

We assume that the domain $\Omega \subset \mathbb{R}^2$ is polygonal. We construct its triangulation $\mathcal{T}_h$ consisting of a finite number of closed triangles T. We will consider only conforming triangulations satisfying the following conditions:

$$\overline{\Omega} = \bigcup_{T\in\mathcal{T}_h} T, \text{ if } T_1, T_2 \in \mathcal{T}_h,\ T_1 \neq T_2, \text{then } T_1 \cap T_2 = \emptyset, \text{ or } T_1 \cap T_2 \text{ is either a common vertex or a common side of } T_1 \text{ and } T_2. \tag{2.1}$$

We say that $T \in \mathcal{T}_h$ is a boundary triangle if T has a side $S \subset \partial\Omega$ and we denote the set of all sides $S \subset \partial\Omega$ by s_h. Then $\bigcup_{S\in s_h} S = \partial\Omega$. For simplicity, we assume that each boundary triangle has only one boundary edge S and thus can be referred to as T_S. If a triangle is not a boundary triangle, we call it an inner triangle.

By h_T and ϱ_T we denote the length of the maximal side of T and the radius of the maximal circle inscribed into T, respectively. We further set

$$h = \max_{T \in \mathcal{T}_h} h_T. \tag{2.2}$$

Let us consider a shape-regular system of triangulations $\{\mathcal{T}_h\}_{h \in (0,h_0)}$, $0 < h_0$, of the domain Ω: there exists $\sigma > 0$ such that

$$\frac{h_T}{\varrho_T} \leqslant \sigma \quad \forall\, T \in \mathcal{T}_h \quad \forall\, h \in (0, h_0). \tag{2.3}$$

Let $r \in \mathbb{N}$ and $T \in \mathcal{T}_h$. We denote the space of all polynomials in x_1, x_2 on T of degree $\leqslant r$ by

$$P_r(T) = \left\{ p_T \colon T \to \mathbb{R};\ p_T(x_1, x_2) = \sum_{\substack{i,j \in \mathbb{N}_0 \\ i+j \leqslant r}} a_{i,j} x_1^i x_2^j,\ a_{i,j} \in \mathbb{R} \right\}. \tag{2.4}$$

An approximate solution will be sought in the space

$$H_h^r = \{ v_h \in C(\overline{\Omega});\ v_h|_T \in P_r(T),\ T \in \mathcal{T}_h \}. \tag{2.5}$$

Now, we can define the Galerkin approximation U_h of the solution u.

Definition 2.1. We say that $U_h \in H_r^h$ is the Galerkin approximation of the weak solution $u \in H^1(\Omega)$ given by (1.8) if

$$a(U_h, v_h) = L(v_h) \quad \forall\, v_h \in H_h^r. \tag{2.6}$$

Since $H_r^h \subset H^1(\Omega)$, it follows that the form a has all the properties in Lemma 1.2 and the existence and uniqueness of an approximate solution follows from the monotone operator theory in [12].

We can further improve the monotonicity of the form a by assuming that one of the functions in question is not close to zero on a part of $\partial\Omega$. More precisely, we suppose that

$$\begin{aligned} &G \subset \partial\Omega, \quad |G| > 0, \\ &|u| > \varepsilon > 0 \quad \text{on } G. \end{aligned} \tag{2.7}$$

Theorem 2.1. *Let $u \in H^1(\Omega)$ and let conditions* (2.7) *hold. Then there exists a constant $C_3 = C_3(\Omega, G, \varepsilon) > 0$ such that*

$$a(u, u - v) - a(v, u - v) \geqslant C_3 \|u - v\|_{1,2,\Omega}^2 \quad \forall\, v \in H^1(\Omega). \tag{2.8}$$

P r o o f. Since $|u|^\alpha - |v|^\alpha$ and $u^2 - v^2$ have the same sign, we see that $(|u|^\alpha - |v|^\alpha)(u^2 - v^2) \geqslant 0$ or equivalently $|u|^\alpha u^2 + |v|^\alpha v^2 \geqslant |u|^\alpha v^2 + |v|^\alpha u^2$. Thus, we can write

$$
\begin{aligned}
2(|u|^\alpha u - |v|^\alpha v)(u - v) &= |u|^\alpha(2u^2 - 2uv) + |v|^\alpha(2v^2 - 2uv) \\
&\geqslant |u|^\alpha(u^2 - 2uv + v^2) + |v|^\alpha(v^2 - 2uv + u^2) \\
&= (|u|^\alpha + |v|^\alpha)(u - v)^2.
\end{aligned} \tag{2.9}
$$

From this and equation (1.9) it directly follows that

$$
a(u, u - v) - a(v, u - v) \geqslant |u - v|^2_{1,2,\Omega} + \frac{1}{2}\kappa\varepsilon^\alpha \|u - v\|^2_{0,2,G}. \tag{2.10}
$$

The existence of the constant C_3 from the statement of this theorem follows from the Poincaré inequality (1.5). □

Under conditions (2.7), we can redefine the function ϱ from (1.12), (1.13) as

$$
\varrho(t) = C_3 t^2, \quad t \in [0, \infty) \tag{2.11}
$$

with a constant $C_3 > 0$.

Theorem 2.2. *Let $u \in H^1(\Omega)$ be a weak solution of* (1.8) *and let $U_h \in H_h^r$ be a Galerkin approximation defined by* (2.6). *Then there exists a constant $c > 0$ independent of h such that*

$$
\varrho_1(\|u - U_h\|_{1,2,\Omega}) \leqslant c \inf_{v_h \in H_h^r} \|u - v_h\|_{1,2,\Omega}, \tag{2.12}
$$

where

$$
\varrho_1(t) = \varrho(t)/t. \tag{2.13}
$$

(*We can remind that, in general, $\varrho(t)$ is defined by* (1.13) *or by* (2.11) *under* (2.7).)

P r o o f. By Lemma 1.2 it is possible to show that the approximate solution satisfies

$$
\varrho(\|U_h\|_{1,2,\Omega}) \leqslant a(U_h, U_h) = L(U_h) \leqslant c(\|f\|_{0,2,\Omega} + \|\varphi\|_{0,2,\partial\Omega})\|U_h\|_{1,2,\Omega}, \tag{2.14}
$$

where we have used the trace embedding in the last inequality. This shows that $\varrho_1(\|U_h\|_{1,2,\Omega})$ is bounded independently of h and U_h is uniformly bounded. Another consequence of formulas (2.6) and (1.8) is the relation

$$
a(u, u - U_h) - a(U_h, u - U_h) = a(u, u - v_h) - a(U_h, u - v_h) \quad \forall\, v_h \in H_h^r. \tag{2.15}
$$

Then, by (1.12), (2.15) and (1.14) for arbitrary $v_h \in H_h^r$, we have

$$
\begin{aligned}
(2.16)\quad \varrho(\|u-U_h\|_{1,2,\Omega}) &\leqslant a(u,u-U_h)-a(U_h,u-U_h)\\
&= |a(u,u-U_h)-a(U_h,u-U_h)|\\
&= |a(u,u-v_h)-a(U_h,u-v_h)|\\
&\leqslant C_1(1+\|u\|_{1,2,\Omega}^{\alpha}+\|U_h\|_{1,2,\Omega}^{\alpha})\|u-U_h\|_{1,2,\Omega}\|u-v_h\|_{1,2,\Omega},
\end{aligned}
$$

which yields (2.12) with a constant $c>0$ dependent on $\|u\|_{1,2,\Omega}^{\alpha}$, but independent of h. □

In what follows, we use Theorem 3.1.5 from [3]:

Theorem 2.3. *Let $r,m \in \mathbb{N}_0$, $p,q \geqslant 1$. Let the piecewise Lagrange interpolation π_h preserve polynomials of degree at most r. Let the triangulation $\mathcal{T}_h$ be shape-regular according to* (2.3). *Let the following embeddings hold:*

$$
(2.17)\qquad W^{r+1,q}(T) \hookrightarrow C(T), \quad W^{r+1,q}(T) \hookrightarrow W^{m,p}(T).
$$

Then there exists a constant $C_4>0$ such that for all $T \in \mathcal{T}_h$ and $h \in (0,h_0)$ we have

$$
(2.18)\qquad |u-\pi_h u|_{m,p,T} \leqslant C_4 |u|_{r+1,q,T} h_T^{r+1-m+2/p-2/q} \quad \forall\, u \in W^{r+1,q}(T).
$$

Let $k,r \in \mathbb{N}$, $q \geqslant 1$. In what follows we assume that $u \in W^{k+1,q}(\Omega)$ is the weak solution of (1.8) and $U_h \in H_r^h$ is the Galerkin approximation defined in (2.6). Let us set $\nu = \min(r,k)$. We get the following result.

Theorem 2.4. *Let the piecewise Lagrange interpolation π_h preserve polynomials of degree $\leqslant r$. Let the triangulations $\mathcal{T}_h$, $h \in (0,h_0)$, be shape-regular according to* (2.3). *Then there exists a constant $C_4>0$ such that*

$$
\begin{aligned}
(2.19)\qquad \|u-\pi_h u\|_{1,2,T} &\leqslant C_4 h_T^{\nu+1-2/q}|u|_{\nu+1,q,T}\\
&\qquad \forall\, u \in W^{k+1,q}(T) \quad \forall\, T \in \mathcal{T}_h \quad \forall\, h \in (0,h_0).
\end{aligned}
$$

Lemma 2.1. *Let $\beta \geqslant 1$, $n \in \mathbb{N}$, $x_i \geqslant 0$, $w_i > 0$, $i=1,\dots,n$. Then the following inequalities hold:*

$$
(2.20)\qquad \sum_{i=1}^{n} x_i^{\beta} \leqslant \left(\sum_{i=1}^{n} x_i\right)^{\beta},
$$

$$
(2.21)\qquad \left(\sum_i w_i x_i \Big/ \sum_i w_i\right)^{\beta} \leqslant \sum_i w_i x_i^{\beta} \Big/ \sum_i w_i.
$$

P r o o f. Both inequalities are a consequence of Jensen's inequality, cf. [25]. □

Theorem 2.5. *We have*

$$\|u - U_h\|_{1,2,\Omega} \leqslant \begin{cases} \varrho_1^{-1}(c|u|_{\nu+1,q,\Omega} h^{\nu+1-2/q}), & q \in [1,2), \\ \varrho_1^{-1}(c|u|_{\nu+1,q,\Omega} h^{\nu}), & q \in [2,\infty). \end{cases} \tag{2.22}$$

Here ϱ_1^{-1} denotes the inverse to ϱ_1 from inequality (2.12).

P r o o f. Using Theorem 2.2 for $v_h = \pi_h u$ and Theorem 2.3, we obtain

$$\begin{aligned} \varrho_1(\|u - U_h\|_{1,2,\Omega}) &\leqslant c\|u - \pi_h u\|_{1,2,\Omega} = c\Big(\sum_{T\in\mathcal{T}_h} \|u - \pi_h u\|^2_{1,2,T}\Big)^{1/2} \\ &\leqslant c\Big(\sum_{T\in\mathcal{T}_h} |u|^2_{\nu+1,q,T} h_T^{2\nu+2-4/q}\Big)^{1/2}. \end{aligned} \tag{2.23}$$

For $q < 2$ we use (2.20) with $\beta = \frac{2}{q}$, $x_i = |u|^q_{\nu+1,q,T} h_T^{q\nu+q-2}$ and get

$$\begin{aligned} \Big(\sum_{T\in\mathcal{T}_h} |u|^2_{\nu+1,q,T} h_T^{2\nu+2-4/q}\Big)^{1/2} &\leqslant \Big(\sum_{T\in\mathcal{T}_h} |u|^q_{\nu+1,q,T} h_T^{q\nu+q-2}\Big)^{1/q} \\ &\leqslant |u|_{\nu+1,q,\Omega} h^{\nu+1-2/q}. \end{aligned} \tag{2.24}$$

Inequality (2.21) can be rewritten as

$$\Big(\sum_i w_i x_i\Big)^{\beta} \leqslant \Big(\sum_i w_i x_i^{\beta}\Big)\Big(\sum_i w_i\Big)^{\beta-1}. \tag{2.25}$$

For $q \geqslant 2$ we use this inequality with $\beta = q/2$, $w_i = h_T^2$, $x_i = |u|^2_{\nu+1,q,T} h_T^{2\nu-4/q}$ and get

$$\Big(\sum_{T\in\mathcal{T}_h} |u|^2_{\nu+1,q,T} h_T^{2\nu+2-4/q}\Big)^{1/2} \leqslant \Big(\sum_{T\in\mathcal{T}_h} h_T^2 |u|^q_{\nu+1,q,T} h_T^{q\nu-2}\Big)^{1/q} \Big(\sum_{T\in\mathcal{T}_h} h_T^2\Big)^{1/2-1/q}. \tag{2.26}$$

Due to the shape regularity of the triangulations $\mathcal{T}_h$, there exists a constant $\widetilde{C}_R$ independent of h such that $\sum\limits_{T\in\mathcal{T}_h} h_T^2 \leqslant \widetilde{C}_R|\Omega|$. Then we get (2.22). □

Further, we show that if the exact solution is zero on the whole boundary, we can improve the estimate for the $H^1(\Omega)$ seminorm.

Theorem 2.6. *Let the weak solution $u \in W^{k+1,q}(\Omega)$ given by* (1.8) *be zero on $\partial\Omega$. Then*

$$|u - U_h|_{1,2,\Omega} \leqslant \begin{cases} c|u|_{\nu+1,q,\Omega} h^{\nu+1-2/q}, & q \in [1,2), \\ c|u|_{\nu+1,q,\Omega} h^{\nu}, & q \in [2,\infty). \end{cases} \tag{2.27}$$

P r o o f. Neglecting the second term on the right-hand side of (1.11) gives us

$$|u - U_h|^2_{1,2,\Omega} \leqslant a(u, u - U_h) - a(U_h, u - U_h). \tag{2.28}$$

The Galerkin orthogonality (2.15) for a piecewise Lagrange interpolation yields

$$a(u, u - U_h) - a(U_h, u - U_h) = a(u, u - \pi_h u) - a(U_h, u - \pi_h u). \tag{2.29}$$

The fact that $\pi_h u$ is also zero on $\partial\Omega$ and the Hölder inequality gives us

$$\begin{aligned} a(u, u - \pi_h u) - a(U_h, u - \pi_h u) &= \int_\Omega \nabla(u - U_h) \cdot \nabla(u - \pi_h u)\,\mathrm{d}x \\ &\leqslant |u - U_h|_{1,2,\Omega}\, |u - \pi_h u|_{1,2,\Omega}. \end{aligned} \tag{2.30}$$

Dividing this by $|u - U_h|_{1,2,\Omega}$ leads to an estimate

$$|u - U_h|_{1,2,\Omega} \leqslant |u - \pi_h u|_{1,2,\Omega}. \tag{2.31}$$

Using Theorem 2.3 for $H^1(T)$ seminorm instead of a norm and the same arguments as in the proof of Theorem 2.5 gives us the sought estimate. □

3. Discrete problem with numerical integration

In practical computation, integrals in the definition of the forms are evaluated by numerical integration. In this section, we are concerned with the analysis of the effect of numerical integration.

Let us consider the reference triangle $\widehat{T}$ with vertices $(0,0)$, $(1,0)$, $(0,1)$. We approximate an integral of a continuous function $\widehat{\psi}$ over $\widehat{T}$ using values at M different points x_μ and M weights ω_μ, $\mu = 1, \ldots, M$. Considering that the area of $\widehat{T}$ is $1/2$, we then have

$$\int_{\widehat{T}} \widehat{\psi}\,\mathrm{d}x \approx \frac{1}{2} \sum_{\mu=1}^{M} \omega_\mu \widehat{\psi}(x_\mu). \tag{3.1}$$

Any triangle T can be obtained using an affine function F_T such that $F_T(\widehat{T}) = T$. Then the nodes become $x_{T,\mu} = F(x_\mu)$, $\mu = 1, \dots, M$ and we obtain a quadrature formula for a function ψ defined on T

$$\int_T \psi \,\mathrm{d}x \approx |T| \sum_{\mu=1}^{M} \omega_\mu \psi(x_{T,\mu}), \quad T \in \mathcal{T}_h. \tag{3.2}$$

Analogically, we introduce numerical integration over edges S on $\partial\Omega$. As a reference element, we use the interval $[0,1]$ with m nodes x_μ and weights β_μ, $\mu = 1, \dots, m$. The quadrature formula on reference interval is

$$\int_0^1 \widehat{\vartheta} \,\mathrm{d}S \approx \sum_{\mu=1}^{m} \beta_\mu \widehat{\vartheta}(x_\mu), \tag{3.3}$$

and the quadrature formula on edges is

$$\int_S \vartheta \,\mathrm{d}S \approx |S| \sum_{\mu=1}^{m} \beta_\mu \vartheta(x_{S,\mu}), \quad S \in s_h. \tag{3.4}$$

The errors of integration are

$$\begin{aligned}
E_T(\psi) &= \int_T \psi \,\mathrm{d}x - |T| \sum_{\mu=1}^{M} \omega_\mu \psi(x_{T,\mu}), \\
E_S(\vartheta) &= \int_S \vartheta \,\mathrm{d}S - |S| \sum_{\mu=1}^{m} \beta_\mu \vartheta(x_{S,\mu}), \\
E_\Omega(\psi) &= \int_\Omega \psi \,\mathrm{d}x - \sum_{T\in\mathcal{T}_h} |T| \sum_{\mu=1}^{M} \omega_\mu \psi(x_{T,\mu}) = \sum_{T\in\mathcal{T}_h} E_T(\psi), \\
E_{\partial\Omega}(\vartheta) &= \int_{\partial\Omega} \vartheta \,\mathrm{d}S - \sum_{S\in s_h} |S| \sum_{\mu=1}^{m} \beta_\mu \vartheta(x_{S,\mu}) = \sum_{S\in s_h} E_S(\vartheta).
\end{aligned} \tag{3.5}$$

The approximations of forms are defined as

$$\begin{aligned}
d_d(u,v) &= \kappa \sum_{S\in s_h} |S| \sum_{\mu=1}^{m} \beta_\mu (|u|^\alpha uv)(x_{S,\mu}), \\
L_d^{\partial\Omega}(v) &= \sum_{S\in s_h} |S| \sum_{\mu=1}^{m} \beta_\mu (\varphi v)(x_{S,\mu}), \\
L_d^{\Omega}(v) &= \sum_{T\in\mathcal{T}_h} |T| \sum_{\mu=1}^{M} \omega_\mu (fv)(x_{T,\mu}).
\end{aligned} \tag{3.6}$$

We assume that the form b will be evaluated exactly as its arguments will be polynomials of degree $\leqslant 2r-2$. Furthermore, we again define forms

$$a_d(u,v) = b(u,v) + d_d(u,v), \quad L_d(v) = L_d^{\Omega}(v) + L_d^{\partial\Omega}(v). \tag{3.7}$$

Definition 3.1. Let E_T be the error of the numerical quadrature on a triangle $T \in \mathcal{T}_h$. We say that a quadrature on triangles is exact for polynomials of degree $\leqslant R$ if $E_T(v_h) = 0$ for any $v_h \in P_R(T)$, $T \in \mathcal{T}_h$.

Let E_S be the error of numerical quadrature on an edge $S \in s_h$. We say that a quadrature on edges is exact for polynomials of degree $\leqslant R$ if $E_S(v_h) = 0$ for any $v_h \in P_R(S)$, $S \in s_h$.

We will use error estimates from Theorems 7.36 and 7.37 in [4].

Theorem 3.1. *Let $S \in s_h$. Let the quadrature formula on edges be exact for polynomials of degree $\leqslant r+s_1-1$. Let $q, q' \in [1,\infty]$ be such that $1/q+1/q' = 1$ (we set $1/\infty = 0$). Then there exists a constant $c > 0$ such that for any $\varphi \in W^{s_1,q}(S)$, $v_h \in P_r(S)$, we have:*

$$|E_S(\varphi v_h)| \leqslant c|S|^{s_1}|\varphi|_{s_1,q,S}\|v_h\|_{0,q',S}. \tag{3.8}$$

Let $T \in \mathcal{T}_h$, where $\mathcal{T}_h$ are shape regular triangulations, let the quadrature formula on triangles be exact for polynomials of degree $\leqslant r+s_2-1$ and let $q, q' \in [1,\infty]$ be such that $1/q + 1/q' = 1$. Then there exists a constant $c > 0$ such that for any $f \in W^{s_2,q}(T)$, $v_h \in P_r(T)$, we have:

$$|E_T(fv_h)| \leqslant ch_T^{s_2}|f|_{s_2,q,T}\|v_h\|_{0,q',T}. \tag{3.9}$$

On the basis of these estimates we prove the following theorem.

Theorem 3.2. *Let the quadrature formula on edges be exact for polynomials of degree $\leqslant r+s_1-1$ on each $S \in s_h$ and let $q \in (1,\infty)$. Then there exists a constant $c > 0$ such that for any $\varphi \in W^{s_1,q}(\partial\Omega)$, $v_h \in H_h^r$, we have*

$$|E_{\partial\Omega}(\varphi v_h)| \leqslant ch^{s_1}|\varphi|_{s_1,q,\partial\Omega}\|v_h\|_{1,2,\Omega}. \tag{3.10}$$

Let the quadrature formula on triangles be exact for polynomials of degree $\leqslant r+s_2-1$ on each $T \in \mathcal{T}_h$, where $\mathcal{T}_h$ are shape regular and let $q \in (1,\infty)$. Then there exists a constant $c > 0$ such that for any $f \in W^{s_2,q}(\Omega)$, $v_h \in H_h^r$, we have:

$$|E_\Omega(fv_h)| \leqslant ch^{s_2}|f|_{s_2,q,\Omega}\|v_h\|_{1,2,\Omega}. \tag{3.11}$$

P r o o f. By (3.8), we have

$$|E_{\partial\Omega}(\varphi v_h)| = \sum_{S\in s_h} |E_S(\varphi v_h)| \leqslant ch^{s_1} \sum_{S\in s_h} |\varphi|_{s_1,q,S}\|v_h\|_{0,q',S}.$$

Applying the discrete Hölder inequality with parameters q and q' such that $1/q + 1/q' = 1$, we find that

$$\sum_{S\in s_h} |\varphi|_{s_1,q,S}\|v_h\|_{0,q',S} \leqslant |\varphi|_{s_1,q,\partial\Omega}\|v_h\|_{0,q',\partial\Omega}.$$

Finally, applying trace embedding $H^1(\Omega) \hookrightarrow L^{q'}(\partial\Omega)$ on v_h, we obtain the first error estimate (3.10). Analogically, we also obtain

$$|E_\Omega(fv_h)| \leqslant ch^{s_2} \sum_{T\in\mathcal{T}_h} |f|_{s_2,q,T}\|v_h\|_{0,q',T} \leqslant ch^{s_2}|f|_{s_2,q,\Omega}\|v_h\|_{0,q',\Omega},$$

and we complete the proof of (3.11) by the embedding $H^1(\Omega) \hookrightarrow L^{q'}(\Omega)$. □

4. Approximate solution

Definition 4.1. We call $u_h \in H_h^r$ an approximate solution of problem (1.1)–(1.2) if

$$a_d(u_h, v_h) = L_d(v_h) \quad \forall\, v_h \in H_h^r. \tag{4.1}$$

In order to obtain error estimates of the approximate solution we need an analogy to monotonicity results for the new form a_d.

Theorem 4.1. *Let the quadrature* (3.4) *have at least* $r+1$ *nodes and only positive weights, i.e.*

$$m \geqslant r+1, \quad \beta_\mu > 0, \quad \mu = 1,\dots,m. \tag{4.2}$$

Then there exists a constant $c > 0$ *independent of* h *such that the following inequality holds for every* $u_h, v_h \in H_h^r$*:*

$$a_d(u_h, u_h - v_h) - a_d(v_h, u_h - v_h) \geqslant |u_h - v_h|^2_{1,2,\Omega} + c\|u_h - v_h\|^{\alpha+2}_{0,\alpha+2,\partial\Omega}. \tag{4.3}$$

Let $s_{h1} \subset s_h$ *be a set of some boundary segments and* $G_h = \bigcup\limits_{S\in s_{h1}} S$*. If*

$$|v_h| > \varepsilon > 0 \quad \text{on } G_h \tag{4.4}$$

for some $\varepsilon > 0$*, then the following inequality holds for* $c > 0$ *independent of* h*:*

$$a_d(u_h, u_h - v_h) - a_d(v_h, u_h - v_h) \geqslant |u_h - v_h|^2_{1,2,\Omega} + c\|u_h - v_h\|^2_{0,2,G_h}. \tag{4.5}$$

P r o o f. The proof can be carried out in a similar way as in [9], Lemma 4.31. □

Lemma 4.1. *Let the assumptions of Theorem* 4.1 *hold. Then*

$$a_d(u_h, u_h - v_h) - a_d(v_h, u_h - v_h) \geqslant \widetilde{\varrho}(\|u_h - v_h\|_{1,2,\Omega}), \tag{4.6}$$

where

$$\widetilde{\varrho}(t) = \begin{cases} C_{0d} t^{\alpha+2} & \text{for } 0 \leqslant t \leqslant 1, \\ C_{0d} t^2 & \text{for } t \geqslant 1. \end{cases} \tag{4.7}$$

If condition (4.4) *holds, then we can redefine* $\widetilde{\varrho}$ *as*

$$\widetilde{\varrho}(t) = C_{1d} t^2. \tag{4.8}$$

P r o o f. The Hölder inequality for

$$\frac{1}{2} = \frac{1}{\alpha+2} + \Big(\frac{1}{2} - \frac{1}{\alpha+2}\Big)$$

applied to the right-hand side of (4.3) gives us

$$\begin{aligned} \|u_h - v_h\|_{0,2,\partial\Omega} &\leqslant \|u_h - v_h\|_{0,\alpha+2,\partial\Omega} \|1\|_{0,(1/2-1/(\alpha+2))^{-1},\partial\Omega} \\ &= \|u_h - v_h\|_{0,\alpha+2,\partial\Omega} |\partial\Omega|^{1/2-1/(\alpha+2)}. \end{aligned}$$

Poincaré inequality (1.5) then yields (4.6) with $\widetilde{\varrho}$ defined in (4.7). Analogously (4.6) with $\widetilde{\varrho}$ defined in (4.8) follows from (4.5). □

Uniform monotonicity of the form a_d on the finite dimensional space H_h^r guarantees the existence and the uniqueness of the approximate solution u_h given by (4.1).

Let us set

$$R(t) = \widetilde{\varrho}(t)/t, \tag{4.9}$$

and let R^{-1} be the inverse of R. Hence,

$$R^{-1}(t) = \begin{cases} \Big(\dfrac{t}{C_{0d}}\Big)^{1/(\alpha+1)} & \text{for } 0 \leqslant t \leqslant C_{0d}, \\ \dfrac{t}{C_{0d}} & \text{for } t \geqslant C_{0d}, \end{cases} \tag{4.10}$$

which can be replaced under condition (4.4) by

$$R^{-1}(t) = \frac{t}{C_{1d}}. \tag{4.11}$$

Then we have the following abstract error estimate.

Theorem 4.2. *Let u_h be the approximate solution of problem* (4.1) *and let $u \in H^1(\Omega)$ be the weak solution defined by* (1.8). *If $v_h \in H_h^r$, then*

$$
\begin{aligned}
(4.12)\qquad \|u-u_h\|_{1,2,\Omega} &\leqslant \|u-v_h\|_{1,2,\Omega} \\
&\quad + R^{-1}\Big(C_1\|u-v_h\|_{1,2,\Omega}(1+\|u\|_{1,2,\Omega}^{\alpha}+\|v_h\|_{1,2,\Omega}^{\alpha}) \\
&\quad + \sup_{0\neq w_h\in H_h^r}\frac{|a(v_h,w_h)-a_d(v_h,w_h)|}{\|w_h\|_{1,2,\Omega}} + \sup_{0\neq w_h\in H_h^r}\frac{|L(w_h)-L_d(w_h)|}{\|w_h\|_{1,2,\Omega}}\Big).
\end{aligned}
$$

P r o o f. By (4.6),

$$\widetilde{\varrho}(\|u_h-v_h\|_{1,2,\Omega}) \leqslant a_d(u_h,u_h-v_h)-a_d(v_h,u_h-v_h).$$

Using the relations

$$a_d(u_h,u_h-v_h)=L_d(u_h-v_h), \quad L(u_h-v_h)=a(u,u_h-v_h),$$

and adding and subtracting the same terms, we get

$$
\begin{aligned}
a_d(u_h,u_h-v_h)-a_d(v_h,u_h-v_h) &= [L_d(u_h-v_h)-L(u_h-v_h)] \\
&\quad + [a(u,u_h-v_h)-a(v_h,u_h-v_h)] + [a(v_h,u_h-v_h)-a_d(v_h,u_h-v_h)].
\end{aligned}
$$

The first bracket on the right-hand side can be estimated directly using the inequality from the definition of a norm of a linear operator:

$$|L_d(u_h-v_h)-L(u_h-v_h)| \leqslant \sup_{0\neq w_h\in H_h^r}\frac{|L(w_h)-L_d(w_h)|}{\|w_h\|_{1,2,\Omega}}\|u_h-v_h\|_{1,2,\Omega}.$$

The second bracket can be estimated using the continuity (1.14) of the form a:

$$
\begin{aligned}
|a(u,u_h-v_h)-a(v_h,u_h-v_h)| &\leqslant C_1(1+\|u\|_{1,2,\Omega}^{\alpha}+\|v_h\|_{1,2,\Omega}^{\alpha}) \\
&\quad \times \|u-v_h\|_{1,2,\Omega}\|u_h-v_h\|_{1,2,\Omega}.
\end{aligned}
$$

The third bracket can be estimated similarly to the first bracket:

$$|a(v_h,u_h-v_h)-a_d(v_h,u_h-v_h)| \leqslant \sup_{0\neq w_h\in H_h^r}\frac{|a(v_h,w_h)-a_d(v_h,w_h)|}{\|w_h\|_{1,2,\Omega}}\|u_h-v_h\|_{1,2,\Omega}.$$

Combining these estimates with the definition of R in (4.9) gives us

$$
\begin{aligned}
(4.13) \qquad R(\|u_h - v_h\|_{1,2,\Omega}) \leqslant & \sup_{0 \neq w_h \in H_h^r} \frac{|L(w_h) - L_d(w_h)|}{\|w_h\|_{1,2,\Omega}} \\
& + C_1 \|u - v_h\|_{1,2,\Omega} (1 + \|u\|_{1,2,\Omega}^{\alpha} + \|v_h\|_{1,2,\Omega}^{\alpha}) \\
& + \sup_{0 \neq w_h \in H_h^r} \frac{|a(v_h, w_h) - a_d(v_h, w_h)|}{\|w_h\|_{1,2,\Omega}}.
\end{aligned}
$$

Using the triangle inequality $\|u - u_h\|_{1,2,\Omega} \leqslant \|u - v_h\|_{1,2,\Omega} + \|u_h - v_h\|_{1,2,\Omega}$, we arrive at (4.12). □

Recall that

$$L(w_h) - L_d(w_h) = E_\Omega(f w_h) + E_{\partial\Omega}(\varphi w_h)$$

represents the error of integration of terms derived from the right-hand sides of (1.1) and (1.2). This error can be estimated using (3.11) and (3.10) from Theorem 3.2:

$$
(4.14) \qquad |L(w_h) - L_d(w_h)| \leqslant c(h^{s_2} |f|_{s_2,q,\Omega} + h^{s_1} |\varphi|_{s_1,q,\partial\Omega}) \|w_h\|_{1,2,\Omega}.
$$

The term

$$a(v_h, w_h) - a_d(v_h, w_h) = E_{\partial\Omega}(|v_h|^{\alpha} v_h w_h)$$

is the error of integration of the nonlinear term on the boundary $\partial\Omega$. It cannot be estimated directly using (3.10), because the continuous piecewise polynomial function v_h may have jumps in its derivatives at vertices of boundary triangles and some derivatives of $|v_h|^{\alpha} v_h$ may become nonintegrable near the roots of v_h in the case of noninteger parameter α. Using (3.8) and repeating arguments from the proof of Theorem 3.2 on separate parts of the boundary will lead to an estimate similar to (3.10). But first, we shall need to prove boundedness of $|v_h|^{\alpha} v_h$ on the boundary $\partial\Omega$ in a norm of some Sobolev space.

For the purpose of error estimation we need to choose $v_h \in H_h^r$ in such a way that

$$\|u - v_h\|_{1,2,\Omega} (1 + \|u\|_{1,2,\Omega}^{\alpha} + \|v_h\|_{1,2,\Omega}^{\alpha}) \to 0 \quad \text{for } h \to 0.$$

Therefore, we will set $v_h = \pi_h u$, where π_h is the continuous piecewise Lagrange interpolation operator.

5. Boundedness of interpolated functions

In this section, we are concerned with estimating derivatives of functions on the boundary of Ω. As we are aiming to use (3.8), it is sufficient to consider one segment at a time. Let $S \in s_h$ be a side of a boundary triangle T_S of a triangulation $\mathcal{T}_h$, let F be an affine mapping of $[0, |S|]$ onto S (this guarantees that $\|F'\| = 1$). Set $\widetilde{v_h} = v_h \circ F$ for a continuous piecewise polynomial function $v_h \in H_h^r$. Function $\widetilde{v_h}$ is therefore a polynomial of degree r defined on an interval $[0, |S|]$.

Let us begin by expressing the actual terms which appear after using chain rule on derivatives of $|\widetilde{v_h}|^\alpha \widetilde{v_h}$. We can proceed similarly as in the proof of Lemma 1.3 and get the following result.

Lemma 5.1. *Let $\widetilde{v_h}$ be a polynomial of degree $\leqslant r$ on the interval $[0, |S|]$. Let $\alpha \geqslant 0$ and $\beta \in \mathbb{N}$. Then*

$$(|\widetilde{v_h}|^\alpha \widetilde{v_h})^{(\beta)}$$

can be expressed as a finite sum of terms of the form

$$c|\widetilde{v_h}|^{\alpha+1-J} \operatorname{sgn} u^{J+1} \prod_{j=1}^{J} \widetilde{v_h}^{(\gamma_j)}, \tag{5.1}$$

where $J \in \mathbb{N}_0$ and $\gamma_j \in \mathbb{N}$, $j = 1, \ldots, J$ are such that $\sum_{j=1}^{J} \gamma_j = \beta$. Here, the constant c is dependent on α, β and γ_j. If $\alpha + 1 - J$ is a negative integer, then $c = 0$.

To estimate integrals of terms of expression (5.1), the most straightforward way is to estimate most of its factors in L^∞-norm and take them out of the integral. If we assume that $u \in W^{r+1,q}(\Omega)$ with $r \in \mathbb{N}$, $q \geqslant 1$, then the embeddings

$$\begin{aligned} W^{r+1,q}(\Omega) &\hookrightarrow C^r(\overline{\Omega}), \quad q > 2, \\ H^{r+1}(\Omega) &\hookrightarrow C^{r-1,\lambda}(\overline{\Omega}), \quad \lambda \in [0, 1), \\ W^{r+1,q}(\Omega) &\hookrightarrow C^{r-1,2-2/q}(\overline{\Omega}), \quad q \in [1, 2), \end{aligned} \tag{5.2}$$

follow from (1.3). Then we can approach estimating of lower derivatives by considerations applying to continuous functions rather than using properties of Sobolev spaces.

Lemma 5.2. *Let $u \in C^r(T_S)$, let $\pi_h u$ be its Lagrange interpolation of degree r using $r + 1$ nodes on the sides of T_S. Let F be an affine mapping of $I = [0, |S|]$*

onto S. Let $\widetilde{u} = u \circ F$ and $\widetilde{\pi_h u} = (\pi_h u) \circ F$. Then for any $i \in \{0, \dots, r\}$ we have the estimate

$$|\widetilde{\pi_h u}|_{i,\infty,I} \leqslant \sum_{j=i}^{r} |S|^{j-i} |\widetilde{u}|_{j,\infty,I}. \tag{5.3}$$

P r o o f. We use the induction on i from r downward to 0.

Let $i = r$. Our goal is to show that $|\widetilde{\pi_h u}|_{r,\infty,I} \leqslant |\widetilde{u}|_{r,\infty,I}$. Function $\widetilde{\pi_h u}$ is a polynomial of degree r and its rth derivative is a constant. Therefore, it is sufficient to prove that $\widetilde{\pi_h u}^{(r)} = \widetilde{u}^{(r)}(t)$ for some $t \in I$ or that a continuous function $(\widetilde{\pi_h u} - \widetilde{u})^{(r)}$ has a root. The Lagrange interpolation is exact at all nodes and thus $\widetilde{\pi_h u} - \widetilde{u}$ has at least $r+1$ roots in I. It follows from Rolle's theorem that $(\widetilde{\pi_h u} - \widetilde{u})'$ has r roots in I and repeating this argument r times gives us a root of $(\widetilde{\pi_h u} - \widetilde{u})^{(r)}$ in I.

Let inequality (5.3) hold for $i+1$. Take arbitrary $t, t_0 \in I$. Considering that $|t_0 - t| \leqslant |I| = |S|$, we have

$$|\widetilde{\pi_h u}^{(i)}(t)| = |\widetilde{\pi_h u}^{(i)}(t_0) + \int_{t_0}^{t} \widetilde{\pi_h u}^{(i+1)}(\tau)\,\mathrm{d}\tau| \leqslant |\widetilde{\pi_h u}^{(i)}(t_0)| + |S||\widetilde{\pi_h u}|_{i+1,\infty,I}.$$

Using (5.3) for $i+1$ and the definition of $L^\infty(I)$-norm, we have

$$\begin{aligned} |\widetilde{\pi_h u}|_{i,\infty,I} &\leqslant |\widetilde{\pi_h u}^{(i)}(t_0)| + |S| \sum_{j=i+1}^{r} |S|^{j-(i+1)} |\widetilde{u}|_{j,\infty,I} \\ &= |\widetilde{\pi_h u}^{(i)}(t_0)| + \sum_{j=i+1}^{r} |S|^{j-i} |\widetilde{u}|_{j,\infty,I}. \end{aligned}$$

To complete the induction step, it suffices to find some $t_0, t_1 \in I$ such that $\widetilde{\pi_h u}^{(i)}(t_0) = \widetilde{u}^{(i)}(t_1)$. Take $i+1$ of the $r+1$ nodes of interpolation. Construct a polynomial v of degree at most i such that $\widetilde{u}$, $\widetilde{\pi_h u}$ and v are equal at these nodes. Functions $\widetilde{u} - v$ and $\widetilde{\pi_h u} - v$ have $i+1$ roots in I and they both belong to a space $C^i(I)$. By Rolle's theorem, there are t_0 and t_1 such that $(\widetilde{\pi_h u} - v)^{(i)}(t_0) = (\widetilde{u} - v)^{(i)}(t_1) = 0$. This together with the fact that $v^{(i)}$ is a constant completes the proof. □

The case when $u \in C^{r-1,2-2/q}(T_S)$ for $q \in (1,2)$ is almost identical.

Lemma 5.3. *Let $u \in C^{r-1,\lambda}(T_S)$, where $\lambda \in (0,1)$, let $\pi_h u$ be its Lagrange interpolation of order r using $r+1$ nodes at the sides of T_S. Let F be an affine mapping of $I = [0, |S|]$ onto S. Let $\widetilde{u} = u \circ F$ and $\widetilde{\pi_h u} = (\pi_h u) \circ F$. Then there exists a constant $c > 0$ such that for any $i \in \{0, \dots, r\}$ we have the estimate*

$$|\widetilde{\pi_h u}|_{i,\infty,I} \leqslant c|S|^{r-1+\lambda-i} |\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)} + \sum_{j=i}^{r-1} |S|^{j-1+\lambda-i} |\widetilde{u}|_{j,\infty,I}. \tag{5.4}$$

P r o o f. Again, we use induction on i from r downward to 0. Let $i = r$. Our goal is to show that

$$|\widetilde{\pi_h u}|_{r,\infty,I} \leqslant c|S|^{-1+\lambda}|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)}.$$

Let v be a Taylor polynomial of the function $\widetilde{u}$ of degree $r-1$ at point 0, i.e., v is a polynomial of degree $\leqslant r-1$ and it has the same derivatives of orders up to $r-1$ at the point 0 as $\widetilde{u}$. Function $\widetilde{u} - v$ has the same $(r-1)$th derivative as $\widetilde{u}$ up to a constant and also has the same seminorm (Hölder constant) $|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)}$. Its interpolation $\widetilde{\pi_h u} - v$ has the rth derivative unchanged. We only need to show that

$$|\widetilde{\pi_h u} - v|_{r,\infty,I} \leqslant c|S|^{-1+\lambda}|(\widetilde{u} - v)^{(r-1)}|_{C^{0,\lambda}(I)},$$

where $(\widetilde{u} - v)$ satisfies $(\widetilde{u} - v)^{(j)}(0) = 0$ for all $j = 0, \dots, r-1$.

It follows from $(\widetilde{u} - v)^{(r-1)}(0) = 0$ and the definition of the Hölder continuity that

$$|\widetilde{u} - v|_{r-1,\infty,I} \leqslant |S|^{\lambda}|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)}.$$

Since $(\widetilde{u}-v)^{(r-2)}(0)=0$ (if $r \geqslant 2$), it follows that $|\widetilde{u} - v|_{r-2,\infty,I} \leqslant |S|^{1+\lambda}|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)}$. Repeating this argument yields $|\widetilde{u} - v|_{0,\infty,I} \leqslant |S|^{r-1+\lambda}|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)}$. Consider an affine transformation of $\widetilde{u} - v$ and $\widetilde{\pi_h u} - v$ from $I = [0, |S|]$ onto $[0, 1]$. Denote the resulting functions by $\widehat{u - v}$ and $\widehat{\pi_h u - v}$. The function $\widehat{u - v}$ is also bounded in $L^\infty(I)$-norm by $|S|^{r-1+\lambda}|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)}$. The interpolation $\widehat{\pi_h u - v}$ of $\widehat{u - v}$ is therefore bounded by

$$|\widehat{\pi_h u - v}|_{0,\infty,[0,1]} \leqslant c|S|^{r-1+\lambda}|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)},$$

where $c > 0$ is a constant dependent only on the choice of nodes of interpolation on the reference interval $[0, 1]$. The space of polynomials of degree $\leqslant r$ on $[0, 1]$ is a finite-dimensional space. Every seminorm on a finite-dimensional space can be estimated from above by any norm. Taking a seminorm $|\cdot|_{r,\infty,[0,1]}$ and a norm $\|\cdot\|_{0,\infty,[0,1]}$ thus yields

$$|\widehat{\pi_h u - v}|_{r,\infty,[0,1]} \leqslant c|S|^{r-1+\lambda}|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)},$$

where $c > 0$ is again some constant dependent only on π. Since affine transformation from I onto $[0, 1]$ multiplies the rth derivative by $|S|^r$, we have

$$|S|^r|\widetilde{\pi_h u} - v|_{r,\infty,I} = |\widehat{\pi_h u - v}|_{r,\infty,[0,1]} \leqslant c|S|^{r-1+\lambda}|\widetilde{u}^{(r-1)}|_{C^{0,\lambda}(I)}.$$

This is (5.4) for $i = r$.

Since functions in the space $C^{r-1,\lambda}(I)$ are also in $C^j(I)$ for $j = 0, \dots, r-1$, the whole induction step in the proof of the previous lemma works here too and we again have

$$|\widetilde{\pi_h u}|_{i,\infty,I} \leqslant |\widetilde{u}|_{i,\infty,I} + |S||\widetilde{\pi_h u}|_{i+1,\infty,I}. \tag{5.5}$$

Combining (5.5) and inequality (5.4) for $i+1$ gives (5.4) for i. □

To estimate the interpolation in a norm of Sobolev spaces we use a special case of Theorem 3.1.5 in [3] (one-dimensional variant of Theorem 2.3).

Corollary 5.1. *Let the piecewise Lagrange interpolation π_h preserve polynomials of degree $\leqslant r$. Let the restriction of the interpolated function $\pi_h u$ on any side of a triangle be given only by the values of u on that side (that is, let it have $r+1$ nodes on every side of a triangle). Let $k \in \mathbb{N}$, $k \geqslant r$, $q \geqslant 1$. Then there exists a constant $C(\pi) > 0$ such that*

$$|\widetilde{u} - \widetilde{\pi_h u}|_{k+1,q,I} \leqslant C|\widetilde{u}|_{k+1,q,I} \quad \forall\, \widetilde{u} \in W^{k+1,q}(I),$$

and it follows from the triangle inequality that we also have

$$|\widetilde{\pi_h u}|_{k+1,q,I} \leqslant (C+1)|\widetilde{u}|_{k+1,q,I} \quad \forall\, \widetilde{u} \in W^{k+1,q}(I). \tag{5.6}$$

When we use polynomials of degree r and consider only numerical quadrature for boundary nonlinear terms satisfying (4.2), we expect the order of convergence in H^1-norm to be r. But we need in addition to the regularity of the exact weak solution u an upper bound for the rth derivative of $(|\widetilde{\pi_h u}|^\alpha \widetilde{\pi_h u})$. It is necessary to have some upper estimate for the terms of the form

$$c|\widetilde{\pi_h u}|^{\alpha+1-J} \operatorname{sgn} \widetilde{\pi_h u}^{J+1} \prod_{j=1}^{J} \widetilde{\pi_h u}^{(\gamma_j)}, \quad \sum_{j=1}^{J} \gamma_j = i \leqslant r. \tag{5.7}$$

If α is an integer, then all exponents $\alpha + 1 - J$ in powers of $|\widetilde{\pi_h u}|$ are non-negative (those that are negative are in terms multiplied by $c = 0$) and we only need an upper estimate of $|\widetilde{\pi_h u}|$. The lowest possible exponent is $\alpha + 1 - r$ and therefore in the case of $\alpha \geqslant r - 1$, we also only need an upper estimate.

Lemma 5.4. *Let $u \in W^{r+1,q}(\Omega)$, where $r \in \mathbb{N}_0$, $q > 1$. Let T_S be a boundary triangle of the triangulation $\mathcal{T}_h$, and $I = [0, |S|]$. Let π_h be a continuous piecewise Lagrange interpolation of order r that uses $r+1$ nodes on the sides of triangles. Let*

F be the affine transformation of I onto S and let $\widetilde{\pi_h u} = (\pi_h u|_S) \circ F$. Let $i \in \mathbb{N}_0$, $i \leqslant r$, and let $\alpha \geqslant 0$ be the constant from (1.2). *Let either $\alpha \in \mathbb{N}_0$ or $\alpha \geqslant i-1$. Then $|\widetilde{\pi_h u}|^\alpha \widetilde{\pi_h u} \in W^{i,q}(I)$ and there exists a constant $c = c\,(\alpha, r, i, q) > 0$ such that*

$$||\widetilde{\pi_h u}|^\alpha \widetilde{\pi_h u}|_{i,q,I} \leqslant c\|u\|^\alpha_{r+1,q,\Omega} \|\widetilde{u}\|_{i,q,I}. \tag{5.8}$$

P r o o f. Due to the triangle inequality in Lebesgue spaces, we only need to estimate the terms of the form given in (5.7) (with $\sum_{j=1}^{J} \gamma_j = i$) by the right-hand side of (5.8). The assumption $\alpha \geqslant i-1$ or $\alpha \in \mathbb{N}_0$ guarantees that the exponents $\alpha + 1 - J$ in (5.7) are non-negative for all terms that need to be estimated. Since we have an embedding (5.2), all derivatives of orders up to $r-1$ can be estimated in L^∞-norm by $\|\widetilde{u}\|_{k+1,q,\Omega}$ due to (5.4) for $q \in (1,2]$ and all derivatives of orders up to r due to (5.3) for $q \in (2,\infty)$.

Let us take a term of the form (5.7):

$$c|\widetilde{\pi_h u}|^{\alpha+1-J} \operatorname{sgn} \widetilde{\pi_h u}^{J+1} \prod_{j=1}^{J} \widetilde{\pi_h u}^{(\gamma_j)}, \quad \sum_{j=1}^{J} \gamma_j = i.$$

Write the seminorm as an integral

$$\left| |\widetilde{\pi_h u}|^{\alpha+1-J} \prod_{j=1}^{J} \widetilde{\pi_h u}^{(\gamma_j)} \right|_{0,q,I} = \left(\int_I |\widetilde{\pi_h u}|^{(\alpha+1-J)q} \prod_{j=1}^{J} |\widetilde{\pi_h u}^{(\gamma_j)}|^q \,\mathrm{d}S \right)^{1/q}.$$

All terms that are continuous can be simply taken out of the integral and give us some upper bound for the seminorm. Without loss of generality assume that γ_J is the largest order of derivative. Suppose for the moment that all other factors are continuous and can be estimated in the following way:

$$\|\widetilde{\pi_h u}^{(\gamma_j)}\|_{0,\infty,I} \leqslant c\|\widetilde{u}\|_{C^r(I)} \leqslant c\|u\|_{C^r(\partial\Omega)} \leqslant c\|u\|_{r+1,q,\Omega} \tag{5.9}$$

replacing the C^r-norm by the $C^{r-1,\lambda}$-norm if $q \in (1,2]$. Then we have an estimate

$$\left(\int_I |\widetilde{\pi_h u}|^{(\alpha+1-J)q} \prod_{j=1}^{J} |\widetilde{\pi_h u}^{(\gamma_j)}|^q \,\mathrm{d}S \right)^{1/q} \leqslant c\|u\|^{(\alpha+1-J)+(J-1)}_{r+1,q,\Omega} \left(\int_I |\widetilde{\pi_h u}^{(\gamma_J)}|^q \,\mathrm{d}S \right)^{1/q}.$$

Using (5.6) gives an estimate of the last remaining part

$$\left(\int_I |\widetilde{\pi_h u}^{(\gamma_J)}|^q \,\mathrm{d}S \right)^{1/q} = |\widetilde{\pi_h u}|_{\gamma_J,q,I} \leqslant c|\widetilde{u}|_{\gamma_J,q,I} \leqslant c\|\widetilde{u}\|_{i,q,I}.$$

Combining these estimates yields

$$\left| |\widetilde{\pi_h u}|^{\alpha+1-J} \prod_{j=1}^{J} \widetilde{\pi_h u}^{(\gamma_j)} \right|_{0,q,I} \leqslant c\|u\|^{\alpha}_{r+1,q,\Omega} \|\widetilde{u}\|_{i,q,I}.$$

The assumption that all factors in $|\widetilde{\pi_h u}|^{\alpha+1-J} \prod_{j=1}^{J} \widetilde{\pi_h u}^{(\gamma_j)}$ besides $\widetilde{\pi_h u}^{(\gamma_J)}$ can be estimated by (5.9) follows from (5.3) and (5.4) if

- $\gamma_J \leqslant r-1$,
- $\gamma_J = r$ and $q > 2$,
- $\gamma_J = i$ (in this case $J = 1$ and there are no other factors with derivatives).

Since we have $\gamma_J \leqslant i \leqslant r$, one of these cases always holds and we have in fact completed the proof. □

If neither $\alpha \in \mathbb{N}_0$ nor $\alpha \geqslant r-1$ and we are still trying to use estimate (3.8) of order r, we need to obtain some positive lower bounds on $\widetilde{\pi_h u}$. These estimates can be derived with some aid from the Lebesgue constants if we include an assumption that $\max\limits_I |\widetilde{u}|$ and $\min\limits_I |\widetilde{u}|$ are relatively close, see Chapter 3 in [22].

Let us consider a fixed Lagrange interpolation π_h of order r preserving polynomials of degree $\leqslant r$ on the boundary. More precisely, the nodes of interpolation on the reference triangle $\widehat{T}$ are in one fixed position for all triangles $T \in \mathcal{T}_h$ and there are $r+1$ nodes of interpolation on every side of this triangle. Take an arbitrary function $\widetilde{u} \in C(I)$ such that $\|\widetilde{u}\|_{0,\infty,I} \leqslant 1$. Then there exists a constant Λ_π such that $\|\widetilde{\pi_h u}\|_{C(I)} \leqslant \Lambda_\pi$ for all such $\widetilde{u}$. It can be defined as

$$\Lambda_\pi = \max_{\substack{\widetilde{u} \in C(I) \\ \|\widetilde{u}\|_{0,\infty,I} \leqslant 1}} \|\widetilde{\pi_h u}\|_{0,\infty,I}.$$

Taking into account that $\widetilde{\pi_h u}$ is given by a finite $(r+1)$-amount of values of $\widetilde{u}$ and the interpolation operator π_h is linear, the maximum in the definition of Λ_π can be found by taking functions which have either 1 or -1 at each node (that is 2^{r+1} combinations). If we further consider that rescaling a function from one interval onto another with a linear substitution will not change the function's extremes, we see that this constant Λ_π is shared for all segments in s_h for all triangulations $\{T_h\}, h > 0$.

If we now take a function $\widetilde{u} \in C(I)$ which is bounded by $a+b$ from above and by $a-b$ from below for some $a \in \mathbb{R}$ and $b > 0$, it follows that the interpolated function $\widetilde{\pi_h u} \in P_r(I)$ is bounded by $a + \Lambda_\pi b$ from above and by $a - \Lambda_\pi b$ from below. Suppose that the values of $\widetilde{u}$ are in $[C_L, 1]$ for some constant $C_L \in (0,1)$. Then we have $a = \frac{1}{2}(1+C_L)$ and $b = \frac{1}{2}(1-C_L)$, and $\widetilde{\pi_h u}$ is estimated from below by

$$\frac{1}{2}(C_L+1) - \frac{\Lambda_\pi}{2}(1-C_L) = \frac{1}{2}(C_L(\Lambda_\pi+1) - (\Lambda_\pi - 1)),$$

which is zero for the choice $C_L = (\Lambda_\pi - 1)/(\Lambda_\pi + 1)$. Then, from the conditions

$$C_L = \frac{\Lambda_\pi - 1}{\Lambda_\pi + 1}, \quad C_l \in (C_L, 1), \quad \frac{\min_I |\widetilde{u}|}{\max_I |\widetilde{u}|} \geqslant C_l \tag{5.10}$$

follows the lower bound estimate

$$\min_I |\widetilde{\pi_h u}| \geqslant \frac{1}{2}(C_l(\Lambda_\pi + 1) - (\Lambda_\pi - 1)) \max_I |\widetilde{u}|.$$

Therefore, we have an estimate in $L^\infty(I)$-norm for a negative power of the interpolated function

$$\| |\widetilde{\pi_h u}|^{-\gamma} \|_{0,\infty,I} \leqslant \Big(\frac{2}{C_l(\Lambda_\pi + 1) - (\Lambda_\pi - 1)} \Big)^\gamma \|\widetilde{u}\|^{-\gamma}_{0,\infty,I}, \quad \gamma > 0. \tag{5.11}$$

If the triangulation is refined by dividing some triangles into smaller ones, the maximum of $|u|$ on any new segment is bounded from above by the old maximum and the new minimum is bounded from below by the old minimum. Thus, the new segment also satisfies conditions (5.10) and the constant C_l might even be increased.

Choosing linearly transformed Chebyshev nodes for the interpolation π_h gives an estimate for the Lebesgue constant

$$\Lambda_\pi = \frac{2}{\pi}\Big(\log(r+1) + \gamma + \log\frac{8}{\pi} \Big) + O(r^{-2}), \tag{5.12}$$

where r is the degree of interpolation and $\gamma = 0.577215$ is the Euler-Mascheroni constant, see [11], [18]. Using the optimal Lebesgue constants for $r \leqslant 4$ (formulas (3.3) and (7.4) in [23]) gives us some possible values for the constant C_L in (5.10). Constants Λ_π and C_L are contained in Table 1 for $r = 1, 2, 3, 4$.

r	1	2	3	4
Λ_π	1.000000	1.250000	1.422919	1.559490
C_L	0.000000	0.111111	0.174549	0.218594

Table 1. Values of Lebesgue constant for polynomials of degrees up to 4 and corresponding constants C_L for an optimal choice of nodes.

Lemma 5.5. *Let $u \in W^{r+1,q}(\Omega)$, let $S \in s_h$ be a boundary segment such that $u|_S$ is non-zero and does not change sign, and furthermore, let $\min_S |u| / \max_S |u| \geqslant C_l$. Suppose that $C_l > C_L$, where C_L is defined above. Let $\widetilde{u}$ be the affine transformation of $u|_S$ onto $I = [0, |S|]$ as defined above. Then there exists a constant $c = c(\alpha, r, q) > 0$ such that*

$$||\widetilde{\pi_h u}|^\alpha \widetilde{\pi_h u}|_{r,q,I} \leqslant c \|u\|^\alpha_{r+1,q,\Omega} \|\widetilde{u}\|_{r,q,I}. \tag{5.13}$$

P r o o f. We can proceed similarly as we did in the proof of Lemma 5.4. The only new concern is that now the exponents $\alpha+1-J$ in the terms of the form

$$c|\widetilde{\pi_h u}|^{\alpha+1-J}\operatorname{sgn}\widetilde{\pi_h u}^{J+1}\prod_{j=1}^{J}\widetilde{\pi_h u}^{(\gamma_j)},\quad \sum_{j=1}^{J}\gamma_j=r$$

can be negative. Whereas we previously used an estimate (5.9) for non-negative $\alpha+1-J$, we now use

(5.14)
$$\||\widetilde{\pi_h u}|^{\alpha+1-J}\|_{0,\infty,I}\leqslant\Big(\frac{2}{C_l(\Lambda_{\pi_h}+1)-(\Lambda_{\pi_h}-1)}\Big)^{J-\alpha-1}\|\widetilde{u}\|_{0,\infty,I}^{\alpha+1-J}\leqslant c\|u\|_{r+1,q,\Omega}^{\alpha+1-J}$$

for negative $\alpha+1-J<0$. This estimate leads to the inequality

$$\Big||\widetilde{\pi_h u}|^{\alpha+1-J}\prod_{j=1}^{J}\widetilde{\pi_h u}^{(\gamma_j)}\Big|_{0,q,I}\leqslant c\|u\|_{r+1,q,\Omega}^{(\alpha+1-J)+(J-1)}\|\widetilde{u}\|_{r,q,I}.$$

It follows that inequality (5.13) holds. □

6. Error estimation

The purpose of this section is to estimate the error of quadrature on the boundary $\partial\Omega$ denoted by $E_{\partial\Omega}(|\pi_h u|^\alpha(\pi_h u)w_h)$. We can divide the boundary segments $S\in s_h$ into three disjoint sets $s_h=s_{h0}\cup s_{h1}\cup s_{h2}$.

- ▷ s_{h0} contains segments S with $u|_S=0$. Then also $\pi_h u|_S=0$ and the quadrature is exact there, i.e. $E_S(|\pi_h u|^\alpha(\pi_h u)w_h)=0$.
- ▷ If $\alpha+1\geqslant r$ or $\alpha\in\mathbb{N}_0$, then s_{h1} contains all segments not in s_{h0}. If $\alpha\notin\mathbb{N}_0$ and $\alpha+1<r$, then s_{h1} contains all segments not in s_{h0} satisfying $\min_S|u|/\max_S|u|\geqslant C_l$, where C_l is given by (5.10). Then combining (5.13) (or (5.8)) and (3.8) gives us an error estimate of order r.
- ▷ s_{h2} contains the remaining segments, i.e. for $\alpha\notin\mathbb{N}_0$ and $\alpha+1<r$, s_{h2} contains segments satisfying $\min\limits_S|u|/\max\limits_S|u|<C_l$ and u is not identically zero on S. Let us set $h_2=\max\{|S|;\ S\in s_{h2}\}$ (or $h_2=0$ if there are no segments in s_{h2}). Combining (5.8) and (3.8) gives us an error estimate of order

(6.1)
$$r_2=\lfloor\alpha\rfloor+1.$$

Theorem 6.1. *Let the weak solution u given in* (1.8) *belong to $W^{r+1,q}(\Omega)$ and let the right-hand side functions belong to spaces $f\in W^{r,q}(\Omega)$ and $\varphi\in W^{r,q}(\partial\Omega)$.*

Let $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ be a shape regular system of triangulations of Ω according to (2.3). *Let its boundary segments s_h be divided according to the cases above for a piecewise continuous Lagrange interpolation π_h of order r with $r+1$ nodes on the sides of triangles. Let the approximate solution be given by* (4.1). *Let the quadrature formulas on edges and on triangles be exact for polynomials of degree $\leqslant 2r-1$ and let the quadrature formula on edges satisfy* (4.2). *Then there exist constants $c_1 = c_1(u, r, q, \Omega) > 0$, $c_2 = c_2(u, r, q, \Omega, \alpha) > 0$, $c_3 = c_3(u, r, q, \Omega, \alpha) > 0$, $c_4 = c_4(f, \varphi, r, \Omega, \pi) > 0$ such that*

$$\|u - u_h\|_{1,2,\Omega} \leqslant c_1 h^{r+1-2/q} + R^{-1}(c_2 h^{r+1-2/q} + c_3(h^r + h_2^{r_2}) + c_4 h^r)) \tag{6.2}$$

if $q \in (1, 2)$ and

$$\|u - u_h\|_{1,2,\Omega} \leqslant c_1 h^r + R^{-1}(c_2 h^r + c_3(h^r + h_2^{r_2}) + c_4 h^r) \tag{6.3}$$

if $q \geqslant 2$, where R^{-1} is defined in (4.10)–(4.11).

P r o o f. It follows from Theorem 4.2 that the error $\|u - u_h\|_{1,2,\Omega}$ is bounded from above by

$$\|u - \pi_h u\|_{1,2,\Omega} + R^{-1}\Big(c\|u - \pi_h u\|_{1,2,\Omega}(1 + \|u\|_{1,2,\Omega}^{\alpha} + \|\pi_h u\|_{1,2,\Omega}^{\alpha}) + \sup_{0\neq w_h\in H_h^r} \frac{|a(\pi_h u, w_h) - a_d(\pi_h u, w_h)|}{\|w_h\|_{1,2,\Omega}} + \sup_{0\neq w_h\in H_h^r} \frac{|L(w_h) - L_d(w_h)|}{\|w_h\|_{1,2,\Omega}}\Big).$$

Estimation of $\|u - \pi_h u\|_{1,2,\Omega}$ by

$$\|u - \pi_h u\|_{1,2,\Omega} \leqslant c|u|_{r+1,q,\Omega} h^{r+1-2/q}$$

for $q \in (1, 2)$ and by

$$\|u - \pi_h u\|_{1,2,\Omega} \leqslant c|u|_{r+1,q,\Omega} h^r$$

for $q \geqslant 2$ was done in the proof of Theorem 2.5. Inequality

$$\|\pi_h u\|_{1,2,\Omega} \leqslant c\|u\|_{1,2,\Omega}$$

follows from (5.6) if we take into account that $\|\cdot\|_{W^{k,q}(I)} = \|\cdot\|_{W^{k,q}(S)}$.

Since the quadrature formulas are exact for polynomials of degree $\leqslant 2r - 1$ and

$$L(w_h) - L_d(w_h) = E_\Omega(f w_h) + E_{\partial\Omega}(\varphi w_h),$$

it follows from Theorem 3.2 that

$$\sup_{0\neq w_h\in H_h^r} \frac{|L(w_h) - L_d(w_h)|}{\|w_h\|_{1,2,\Omega}} \leqslant c h^r(|f|_{r,q,\Omega} + |\varphi|_{r,q,\partial\Omega}).$$

Finally, we have

$$|a(\pi_h u, w_h) - a_d(\pi_h u, w_h)| = \sum_{S \in s_h} E_S(|\pi_h u|^\alpha (\pi_h u) w_h).$$

Errors on the segments $s_h = s_{h0} \cup s_{h1} \cup s_{h2}$ are estimated separately using Theorem 3.1. Since $u = \pi_h u = 0$ on segments $S \in s_{h0}$, we have

$$\sum_{S \in s_{h0}} E_S(|\pi_h u|^\alpha (\pi_h u) w_h) = 0.$$

On segments $S \in s_{h1}$, we can use the estimate

$$|E_S(|\pi_h u|^\alpha (\pi_h u) w_h)| \leqslant c|S|^r ||\pi_h u|^\alpha \pi_h u|_{r,q,S} \|w_h\|_{0,q',S}$$

for $1/q + 1/q' = 1$, and then either (5.13) or (5.8), implying

$$||\pi_h u|^\alpha \pi_h u|_{r,q,S} \leqslant c\|u\|^\alpha_{r+1,q,\Omega} \|u\|_{r,q,S},$$

which yields

$$|E_S(|\pi_h u|^\alpha (\pi_h u) w_h)| \leqslant c(u) h^r \|u\|_{r,q,S} \|w_h\|_{0,q',S}.$$

Summing over all $S \in s_{h1}$, using the discrete Hölder inequality and trace embedding, we conclude that

$$\begin{aligned} \left| \sum_{S \in s_{h1}} E_S(|\pi_h u|^\alpha (\pi_h u) w_h) \right| &\leqslant \sum_{S \in s_{h1}} c(u) h^r \|u\|_{r,q,S} \|w_h\|_{0,q',S} \\ &\leqslant c(u) h^r \|u\|_{r,q,\bigcup s_{h1}} \|w_h\|_{0,q',\bigcup s_{h1}} \\ &\leqslant c(u) h^r \|u\|_{r+1,q,\Omega} \|w_h\|_{1,2,\Omega}. \end{aligned}$$

On segments $S \in s_{h2}$ we can similarly use the estimate

$$|E_S(|\pi_h u|^\alpha \pi_h u w_h)| \leqslant c|S|^{r_2} ||\pi_h u|^\alpha \pi_h u|_{r_2,q,S} \|w_h\|_{0,q',S}.$$

By (5.8) we have

$$||\pi_h u|^\alpha \pi_h u|_{r_2,q,S} \leqslant c\|u\|^\alpha_{r+1,q,\Omega} \|u\|_{r_2,q,S},$$

which leads to

$$\left| \sum_{S \in s_{h2}} E_S(|\pi_h u|^\alpha \pi_h u w_h) \right| \leqslant c(u) h_2^{r_2} \|u\|_{r+1,q,\Omega} \|w_h\|_{1,2,\Omega}.$$

Combining these estimates yields inequalities (6.2) and (6.3). □

Note that the function R^{-1} defined in (4.10)–(4.11) is linear if the exact solution u is sufficiently distant from zero on a large part of the boundary $\partial\Omega$. Our theoretical estimates for the order of convergence in the H^1-norm are divided by $\alpha+1$ only if the exact solution is zero on most of the boundary. Similarly to the Galerkin approximation, we can improve the estimate for the rate of convergence in H^1-seminorm by omitting the denominator $\alpha+1$ if the exact solution u is zero on the whole boundary $\partial\Omega$. In this case, we also need to assume that the right-hand side integrals are evaluated exactly, that is

$$\int_\Omega f v_h \,\mathrm{d}x, \quad \int_{\partial\Omega} \varphi v_h \,\mathrm{d}S$$

can be evaluated exactly for the given functions f, φ from (1.1)–(1.2), and $v_h \in H_h^r$, whereas

$$\int_{\partial\Omega} |v_h|^\alpha v_h w_h \,\mathrm{d}S, \quad v_h, w_h \in H_h^r$$

is evaluated using numerical quadrature. The argument is similar to Theorem 2.6.

Theorem 6.2. *Let the weak solution $u \in W^{r+1,q}(\Omega)$ given in* (1.8) *be zero on $\partial\Omega$. Let an approximate solution $u_h \in H_h^r$ be given by*

$$a_d(u_h, v_h) = L(v_h) \quad \forall\, v_h \in H_h^r, \tag{6.4}$$

where a_d and L are defined in (3.7) *and* (1.7)*. Let the quadrature formula on edges satisfy* (4.2)*. Then*

$$|u-u_h|_{1,2,\Omega} \leqslant \begin{cases} c|u|_{r+1,q,\Omega} h^{r+1-2/q}, & q \in [1,2), \\ c|u|_{r+1,q,\Omega} h^r, & q \in [2,\infty). \end{cases} \tag{6.5}$$

P r o o f. Neglecting the second term on the right-hand side of (4.3) gives us

$$|u_h - \pi_h u|_{1,2,\Omega}^2 \leqslant a_d(u_h, u_h - \pi_h u) - a_d(\pi_h u, u_h - \pi_h u).$$

The definitions of solutions u_h and u yield

$$a_d(u_h, u_h - \pi_h u) = L(u_h - \pi_h u) = a(u, u_h - \pi_h u).$$

Using the fact that u is zero on $\partial\Omega$ and thus the integral of $|u|^\alpha u(u_h - \pi_h u)$ on the boundary is evaluated exactly, we obtain

$$a(u, u_h - \pi_h u) = a_d(u, u_h - \pi_h u).$$

Taking again into account that $u|_{\partial\Omega} = 0$, by the above relations and the Hölder inequality we get

$$|u_h - \pi_h u|^2_{1,2,\Omega} \leqslant a_d(u, u_h - \pi_h u) - a_d(\pi_h u, u_h - \pi_h u)$$
$$= \int_\Omega \nabla(u - \pi_h u) \cdot \nabla(u_h - \pi_h u)\,\mathrm{d}x \leqslant |u - \pi_h u|_{1,2,\Omega} |u_h - \pi_h u|_{1,2,\Omega}.$$

Dividing this inequality by $|u_h - \pi_h u|_{1,2,\Omega}$ leads to the estimate

$$|u_h - \pi_h u|_{1,2,\Omega} \leqslant |u - \pi_h u|_{1,2,\Omega}.$$

The triangle inequality further gives us

$$|u - u_h|_{1,2,\Omega} \leqslant |u - \pi_h u|_{1,2,\Omega} + |\pi_h u - u_h|_{1,2,\Omega} \leqslant 2|u - \pi_h u|_{1,2,\Omega}.$$

This relation and Theorem 2.3 yield estimate (6.5). □

7. Numerical experiments

In this chapter we present two numerical examples computed using the FEniCS software [1]. We explore the reduction of the order of convergence caused by the nonlinearity, how it affects different norms, and whether this changes if the exact solution of problem (1.1)–(1.2) is zero on the whole boundary $\partial\Omega$. In both experiments we discretize the problem by the FEM. We use uniform triangular meshes with element diameters $h_l = h_0/2^l$, $l = 0, 1, \ldots, 5$. The amount of degrees of freedom (DOF) is therefore expected to increase about four times with each refinement. Denoting the error of the discrete solution by $e_h = u - u_h$, we compute the experimental order of convergence (EOC) by

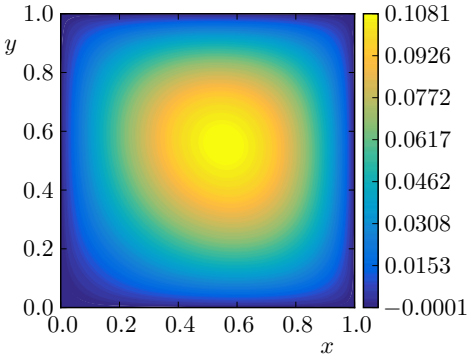
$$\mathrm{EOC} = \frac{\log e_{h_{l-1}} - \log e_{h_l}}{\log h_{l-1} - \log h_l}, \quad l = 1, 2, \ldots, 5. \tag{7.1}$$

The discrete problems (2.6), (4.1) represent nonlinear systems for $\alpha > 0$. We solved these problems by a dampened Newton method with tolerance on the residual 10^{-9}.

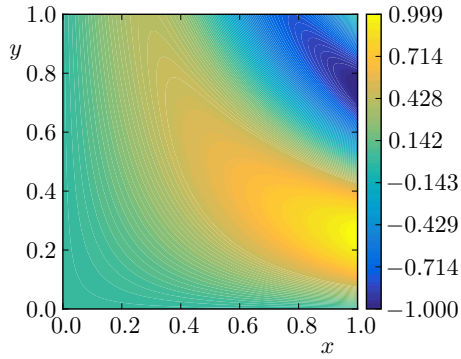
7.1. Example 1—solution is zero on the boundary. In the first experiment we consider problem (1.1)–(1.2) on a unit square domain $\Omega = (0, 1)^2$. The data f and φ are chosen so that the exact solution is

$$u(x_1, x_2) = x_1(1 - x_1)x_2(1 - x_2)(x_1^2 + x_2^2)^{1/4}. \tag{7.2}$$

This function belongs to $W^{4,q}(\Omega)$, $q \in (1, \frac{4}{3})$, or $H^{3.5-\delta}(\Omega)$, $\delta > 0$. Therefore, we expect $|e_h|_{1,2,\Omega} \approx O(h^{\min(2.5,r)})$ and $\|e_h\|_{0,2,\Omega} \approx O(h^{\min(2.5,r)/(\alpha+1)})$. This function is shown in Figure 1(a).



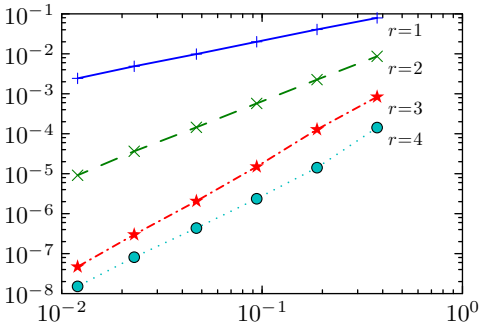
(a) Example 1—function, which is zero on the whole boundary



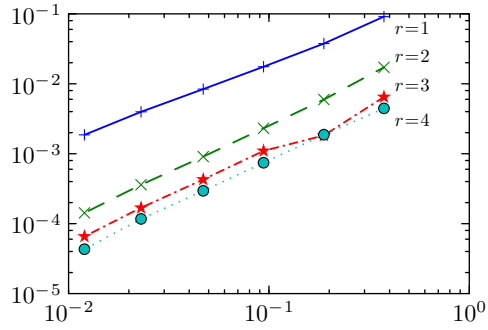
(b) Example 2—smooth function, which is nonzero on a part of the boundary

Figure 1. The exact weak solutions of the discretized problems.

We have discretized the problem by the FEM. For polynomials of degree $r = 2$ we have tried different values of nonlinearity parameter $\alpha = 0.5, 1.0, 1.5, 2.0$, and for parameter $\alpha = 1.5$ we have tried FEM with polynomials of degrees $r = 1, 2, 3, 4$. The results shown in Table 2 and Figures 2 and 3 also include the mesh element size $h = \max_{T \in \mathcal{T}_h} h_T$, the number of degrees of freedom and the number of Newton iterations.



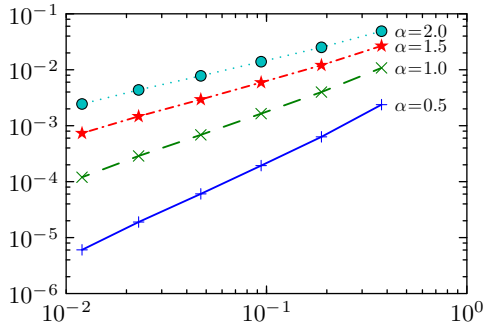
(a) error measured in $|\cdot|_{1,2,\Omega}$



(b) error measured in $\|\cdot\|_{0,2,\Omega}$

Figure 2. Example 1—EOC of FEM for $\alpha = 1.5$.

The H^1-seminorm seems to behave as expected, i.e. the order of convergence is $\min(2.5, r)$. The most significant part of the error measured in H^1-norm was its L^2-norm. Our estimates for the L^2-norm give us the order of convergence $\min(2.5, r)/(\alpha + 1)$, which would be $1/(\alpha + 1)$, $2/(\alpha + 1)$, $2.5/(\alpha + 1)$, $2.5/(\alpha + 1)$ for $r = 1, 2, 3, 4$, respectively. The EOC, however, suggests $2/(\alpha + 1)$, $2.5/(\alpha + 1)$, $2.5/(\alpha + 1)$, $2.5/(\alpha + 1)$ for $r = 1, 2, 3, 4$, respectively. The theoretical error estimate is therefore suboptimal.



(a) error of FEM

Figure 3. Example 1—EOC measured in $\|\cdot\|_{0,2,\Omega}$ for $r = 2$.

$\alpha = 1.5$, $r = 1$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	49	4	9.3448e-02	–	7.9119e-02	–	1.2244e-01	–
0.188	161	6	4.8018e-02	0.96	4.0634e-02	0.96	6.2904e-02	0.96
0.094	577	6	2.7109e-02	0.82	2.0042e-02	1.02	3.3713e-02	0.90
0.047	2177	6	1.5600e-02	0.80	9.8458e-03	1.03	1.8447e-02	0.87
0.023	8449	6	8.8992e-03	0.81	4.8780e-03	1.01	1.0148e-02	0.86
0.012	33281	6	5.0395e-03	0.82	2.4321e-03	1.00	5.5957e-03	0.86

$\alpha = 1.5$, $r = 2$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	161	3	2.6724e-02	–	8.6570e-03	–	2.8091e-02	–
0.188	577	6	1.2058e-02	1.15	2.2618e-03	1.94	1.2268e-02	1.20
0.094	2177	6	5.9243e-03	1.03	5.7373e-04	1.98	5.9520e-03	1.04
0.047	8449	6	2.9464e-03	1.01	1.4479e-04	1.99	2.9499e-03	1.01
0.023	33281	6	1.4700e-03	1.00	3.6421e-05	1.99	1.4704e-03	1.00
0.012	132097	6	7.3425e-04	1.00	9.1384e-06	1.99	7.3430e-04	1.00

$\alpha = 1.5$, $r = 3$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	337	3	1.2840e-02	–	8.3916e-04	–	1.2867e-02	–
0.188	1249	6	4.9724e-03	1.37	1.2809e-04	2.71	4.9741e-03	1.37
0.094	4801	5	3.3908e-03	0.55	1.5021e-05	3.09	3.3908e-03	0.55
0.047	18817	6	1.6746e-03	1.02	2.0634e-06	2.86	1.6746e-03	1.02
0.023	74497	6	8.3301e-04	1.01	2.9962e-07	2.78	8.3301e-04	1.01
0.012	296449	3	4.1014e-04	1.02	4.7016e-08	2.67	4.1014e-04	1.02

Table 2. Example 1—number of DOF and Newton iterations, discretization errors and convergence rates for $r = 1, 2, 3, 4$ and $\alpha = 0.5, 1.0, 1.5, 2.0$ in FEM.

$\alpha = 1.5$, $r = 4$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	577	3	9.6870e-03	–	1.4266e-04	–	9.6880e-03	–
0.188	2177	6	5.0551e-03	0.94	1.4161e-05	3.33	5.0551e-03	0.94
0.094	8449	6	2.5318e-03	1.00	2.3612e-06	2.58	2.5318e-03	1.00
0.047	33281	6	1.2653e-03	1.00	4.3600e-07	2.44	1.2653e-03	1.00
0.023	132097	6	6.3245e-04	1.00	8.1398e-08	2.42	6.3245e-04	1.00
0.012	526337	4	2.9917e-04	1.08	1.5154e-08	2.43	2.9917e-04	1.08

$\alpha = 0.5$, $r = 2$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	161	4	2.3779e-03	–	8.6544e-03	–	8.9752e-03	–
0.188	577	5	6.3232e-04	1.91	2.2617e-03	1.94	2.3485e-03	1.93
0.094	2177	4	1.9356e-04	1.71	5.7372e-04	1.98	6.0550e-04	1.96
0.047	8449	3	6.0476e-05	1.68	1.4479e-04	1.99	1.5691e-04	1.95
0.023	33281	3	1.8977e-05	1.67	3.6421e-05	1.99	4.1069e-05	1.93
0.012	132097	3	6.0396e-06	1.65	9.1384e-06	1.99	1.0954e-05	1.91

$\alpha = 1.0$, $r = 2$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	161	4	1.0793e-02	–	8.6566e-03	–	1.3835e-02	–
0.188	577	6	3.9942e-03	1.43	2.2618e-03	1.94	4.5901e-03	1.59
0.094	2177	6	1.6433e-03	1.28	5.7373e-04	1.98	1.7406e-03	1.40
0.047	8449	5	6.8640e-04	1.26	1.4479e-04	1.99	7.0150e-04	1.31
0.023	33281	4	2.8784e-04	1.25	3.6421e-05	1.99	2.9014e-04	1.27
0.012	132097	3	1.1988e-04	1.26	9.1384e-06	1.99	1.2023e-04	1.27

$\alpha = 2.0$, $r = 2$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	161	3	4.8888e-02	–	8.6572e-03	–	4.9648e-02	–
0.188	577	6	2.5182e-02	0.96	2.2618e-03	1.94	2.5284e-02	0.97
0.094	2177	6	1.3928e-02	0.85	5.7373e-04	1.98	1.3940e-02	0.86
0.047	8449	6	7.7818e-03	0.84	1.4479e-04	1.99	7.7831e-03	0.84
0.023	33281	6	4.3594e-03	0.84	3.6421e-05	1.99	4.3595e-03	0.84
0.012	132097	6	2.4446e-03	0.83	9.1384e-06	1.99	2.4446e-03	0.83

Table 2. Example 1—number of DOF and Newton iterations, discretization errors and convergence rates for $r = 1, 2, 3, 4$ and $\alpha = 0.5, 1.0, 1.5, 2.0$ in FEM (continuation).

7.2. Example 2—solution not identically zero on the boundary. In the second experiment, we again consider problem (1.1)–(1.2) on a unit square domain $\Omega = (0,1)^2$. We prescribe the data f and φ in such a way that the exact solution is

$$u(x_1, x_2) = \tfrac{1}{4}(1 + x_1)^2 \sin(2\pi x_1 x_2), \tag{7.3}$$

shown in Figure 1(b).

This function was used in [17]. It is smooth, zero on boundary segments going through points $[0,1]$, $[0,0]$, $[1,0]$ and nonzero on segments going through points $[1,0]$, $[1,1]$, $[0,1]$. The expected order of convergence is r in all norms and seminorms considered and should not depend on the nonlinearity parameter α. The computed results are presented in Table 3 and Figure 4.

$\alpha = 1.5$, $r = 1$								
h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	49	6	2.5883e-01	–	9.5881e-01	–	9.9314e-01	–
0.188	161	5	6.1723e-02	2.07	5.3381e-01	0.84	5.3736e-01	0.89
0.094	577	4	1.5381e-02	2.00	2.8145e-01	0.92	2.8187e-01	0.93
0.047	2177	4	3.9289e-03	1.97	1.4421e-01	0.96	1.4426e-01	0.97
0.023	8449	3	9.9584e-04	1.98	7.2704e-02	0.99	7.2711e-02	0.99
0.012	33281	3	2.4986e-04	1.99	3.6390e-02	1.00	3.6391e-02	1.00
$\alpha = 1.5$, $r = 2$								
h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	161	6	1.4730e-02	–	2.3514e-01	–	2.3560e-01	–
0.188	577	4	1.2493e-03	3.56	5.8813e-02	2.00	5.8826e-02	2.00
0.094	2177	3	1.3819e-04	3.18	1.5173e-02	1.95	1.5173e-02	1.95
0.047	8449	3	1.6986e-05	3.02	3.8676e-03	1.97	3.8676e-03	1.97
0.023	33281	2	2.1254e-06	3.00	9.7489e-04	1.99	9.7489e-04	1.99
0.012	132097	2	2.6587e-07	3.00	2.4425e-04	2.00	2.4425e-04	2.00
$\alpha = 1.5$, $r = 3$								
h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	337	6	4.5914e-03	–	2.3116e-02	–	2.3568e-02	–
0.188	1249	3	2.4182e-04	4.25	3.4931e-03	2.73	3.5015e-03	2.75
0.094	4801	3	1.3800e-05	4.13	4.7873e-04	2.87	4.7893e-04	2.87
0.047	18817	2	8.5542e-07	4.01	6.2363e-05	2.94	6.2369e-05	2.94
0.023	74497	2	5.4140e-08	3.98	7.9229e-06	2.98	7.9231e-06	2.98
0.012	296449	2	3.4211e-09	3.98	9.9474e-07	2.99	9.9474e-07	2.99

Table 3. Example 2—number of DOF and Newton iterations, discretization errors and convergence rates for $r = 1, 2, 3, 4$ and $\alpha = 1.5, 0.5$ in FEM.

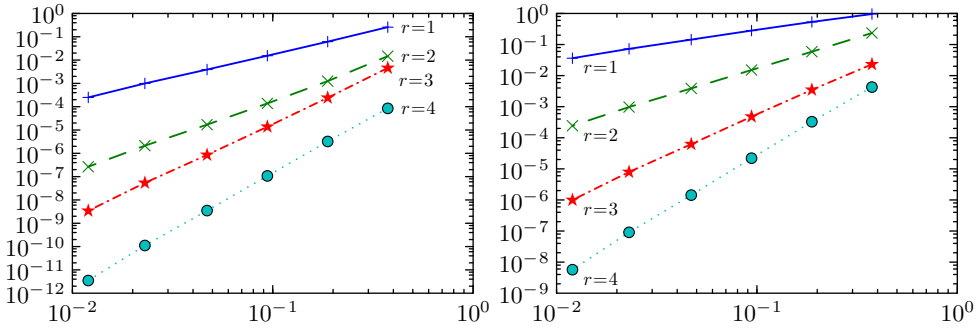
$\alpha = 1.5$, $r = 4$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	577	6	8.4789e-05	–	4.2824e-03	–	4.2832e-03	–
0.188	2177	3	3.2227e-06	4.72	3.2812e-04	3.71	3.2813e-04	3.71
0.094	8449	2	1.0740e-07	4.91	2.2035e-05	3.90	2.2036e-05	3.90
0.047	33281	2	3.4969e-09	4.94	1.4299e-06	3.95	1.4299e-06	3.95
0.023	132097	2	1.1140e-10	4.97	9.0809e-08	3.98	9.0809e-08	3.98
0.012	526337	2	3.5005e-12	4.99	5.6988e-09	3.99	5.6988e-09	3.99

$\alpha = 0.5$, $r = 2$

h	DOF	iter	$\|e\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC
0.375	161	6	1.4072e-02	–	2.3527e-01	–	2.3569e-01	–
0.188	577	4	1.2379e-03	3.51	5.8815e-02	2.00	5.8828e-02	2.00
0.094	2177	4	1.3806e-04	3.16	1.5173e-02	1.95	1.5173e-02	1.95
0.047	8449	3	1.6989e-05	3.02	3.8676e-03	1.97	3.8676e-03	1.97
0.023	33281	3	2.1256e-06	3.00	9.7489e-04	1.99	9.7489e-04	1.99
0.012	132097	2	2.6588e-07	3.00	2.4425e-04	2.00	2.4425e-04	2.00

Table 3. Example 2—number of DOF and Newton iterations, discretization errors and convergence rates for $r = 1, 2, 3, 4$ and $\alpha = 1.5, 0.5$ in FEM (continuation).



(a) error of FEM measured in $\|\cdot\|_{0,2,\Omega}$ (b) error of FEM measured in $|\cdot|_{1,2,\Omega}$

Figure 4. Example 2—EOC for $\alpha = 1.5$.

In the discretization of this problem, we have chosen $\alpha = 1.5$ and degrees of polynomials $r = 1, 2, 3$ for FEM. We have also tried $r = 4$ and $\alpha = 0.5$. The order of convergence is not affected by boundary nonlinearity parameter α, which is in agreement with theoretical results. The H^1-seminorm converges with the predicted order of convergence r, but the L^2-norm converges faster with order $r + 1$. The L^2-norm error estimate is again suboptimal, but in this case, the error is dominated by the H^1-seminorm. Therefore the resulting order of convergence in H^1-norm is still r in accordance with the theoretical results.

The numerical experiments confirmed that the theoretical error estimates in seminorms were optimal and that the order of convergence depends on whether the exact solution is zero on the whole boundary. The numerical results, however, suggest that the order of convergence in L^2-norm is suboptimal. The theoretical results give us the order of convergence r (or $r/(\alpha+1)$), but the EOC is $r+1$ (or $(r+1)/(\alpha+1)$). This improvement only appeared when the exact solution belonged to the space $H^{r+2}(\Omega)$.

8. Conclusion

We have shown theoretically that the use of numerical integration for evaluating forms in the definition of the approximate solution does not decrease the order of convergence, which was derived in Section 2. In the case of noninteger $\alpha > 0$ and the degree of used polynomials $r > \alpha+1$, it might be necessary to refine the triangulation $\mathcal{T}_h$ near the roots of the exact solution u on the boundary $\partial\Omega$. These refined triangles T_S, $S \in s_{h2}$, would require their size to be $h_{T_S} \leqslant ch^{r/(\lfloor\alpha\rfloor+1)}$. It is also possible to say that the estimates in Section 2 only require the regularity of the solution, but the estimates near the boundary edges are only possible under the regularity specified in Section 1. Numerical experiments did not require this refinement to converge with the derived order of convergence.

Combining Theorem 2.5, ϱ_1^{-1} given in (4.11), and Theorem 2.6 suggested that the Galerkin approximation given in (2.6) should always converge to the exact weak solution defined in (1.8) in the H^1-seminorm with a rate of convergence of r. The same conclusions can be drawn from Theorem 6.1, R^{-1} given in (4.11) and Theorem 6.2 in Section 6, which takes into account the effect of numerical integration. This theoretical result is in agreement with the numerical experiments.

References

[1] *M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, G. N. Wells*: The FEniCS Project Version 1.5. Archive of Numerical Software *3* (2015), 9–23. doi

[2] *R. Bialecki, A. J. Nowak*: Boundary value problems in heat conduction with nonlinear material and nonlinear boundary conditions. Appl. Math. Modelling *5* (1981), 417–421. zbl doi

[3] *P. G. Ciarlet*: The Finite Element Method for Elliptic Problems. Studies in Mathematics and Its Applications 4, North Holland, Amsterdam, 1978. zbl MR doi

[4] *V. Dolejší, M. Feistauer*: Discontinuous Galerkin Method. Analysis and Applications to Compressible Flow. Springer Series in Computational Mathematics 48, Springer, Cham, 2015. zbl MR doi

[5] *J. Douglas,* Jr., *T. Dupont*: Galerkin methods for parabolic equations with nonlinear boundary conditions. Numer. Math. *20* (1973), 213–237. zbl MR doi

[6] *L. C. Evans*: Partial Differential Equations. Graduate Studies in Mathematics 19, American Mathematical Society, Providence, 2010. zbl MR doi

[7] *M. Feistauer, H. Kalis, M. Rokyta*: Mathematical modelling of an electrolysis process. Commentat. Math. Univ. Carol. *30* (1989), 465–477. zbl MR
[8] *M. Feistauer, K. Najzar*: Finite element approximation of a problem with a nonlinear Newton boundary condition. Numer. Math. *78* (1998), 403–425. zbl MR doi
[9] *M. Feistauer, K. Najzar, V. Sobotíková*: Error estimates for the finite element solution of elliptic problems with nonlinear Newton boundary conditions. 20 (1999), 835–851. zbl MR doi
[10] *M. Feistauer, F. Roskovec, A.-M. Sändig*: Discontinuous Galerkin method for an elliptic problem with nonlinear Newton boundary conditions in a polygon. IMA J. Numer. Anal. *39* (423–453). MR doi
[11] *B. Fornberg*: A Practical Guide to Pseudospectral Methods. Cambridge Monographs on Applied and Computational Mathematics 1, Cambridge University Press, Cambridge, 1996. zbl MR doi
[12] *J. Franců*: Monotone operators. A survey directed to applications to differential equations. Appl. Mat. *35* (1990), 257–301. zbl MR
[13] *M. Ganesh, I. G. Graham, J. Sivaloganathan*: A pseudospectral three-dimensional boundary integral method applied to a nonlinear model problem from finite elasticity. SIAM J. Numer. Anal. *31* (1994), 1378–1414. zbl MR doi
[14] *M. Ganesh, O. Steinbach*: Nonlinear boundary integral equations for harmonic problems. J. Integral Equations Appl. *11* (1999), 437–459. zbl MR doi
[15] *M. Ganesh, O. Steinbach*: Boundary element methods for potential problems with nonlinear boundary conditions. Math. Comput. *70* (2001), 1031–1042. zbl MR doi
[16] *P. Grisvard*: Elliptic Problems in Nonsmooth Domains. Monographs and Studies in Mathematics 24, Pitman Publishing, Boston, 1985. zbl MR doi
[17] *K. Harriman, P. Houston, B. Senior, E. Süli*: *hp*-version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form. Recent Advances in Scientific Computing and Partial Differential Equations 2002 (S. Y. Cheng et al., eds.). Contemp. Math. 330; American Mathematical Society, Providence, 2003, pp. 89–119. zbl MR doi
[18] *J. S. Hesthaven*: From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex. SIAM J. Numer. Anal. *35* (1998), 655–676. zbl MR doi
[19] *M. Křížek, L. Liu, P. Neittaanmäki*: Finite element analysis of a nonlinear elliptic problem with a pure radiation condition. Applied Nonlinear Analysis (A. Sequeira et al., eds.). Kluwer Academic/Plenum Publishers, New York, 1999, pp. 271–280. zbl MR doi
[20] *L. Liu, M. Křížek*: Finite element analysis of a radiation heat transfer problem. J. Comput. Math. *16* (1998), 327–336. zbl MR
[21] *R. Moreau, J. W. Ewans*: An analysis of the hydrodynamics of aluminum reduction cells. J. Electrochem. Soc. *31* (1984), 2251–2259. doi
[22] *M. J. D. Powell*: Approximation Theory and Methods. Cambridge University Press, Cambridge, 1981. zbl MR doi
[23] *H.-J. Rack, R. Vajda*: Optimal cubic Lagrange interpolation: Extremal node systems with minimal Lebesgue constant. Stud. Univ. Babeş-Bolyai, Math. *60* (2015), 151–171. zbl MR
[24] *T. Roubíček*: A finite-element approximation of Stefan problems in heterogeneous media. Free Boundary Value Problems 1989. Int. Ser. Numer. Math. 95, Birkhäuser, Basel, 1990, pp. 267–275. zbl MR doi
[25] *W. Rudin*: Real and Complex Analysis. McGraw-Hill, New York, 1987. zbl MR

Authors' address: *Ondřej Bartoš*, *Miloslav Feistauer* (corresponding author), *Filip Roskovec*, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: `bartos@karlin.mff.cuni.cz`, `feist@karlin.mff.cuni.cz`, `roskovec@gmail.com`.