Journal of Computational and Applied Mathematics 311 (2017) 457–472

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Iterative Galerkin discretizations for strongly monotone problems

Scott Congreve[a,*], Thomas P. Wihler[b]

[a] *Fakultät für Mathematik, Universität Wien, Oskar-Morgenstern-Platz 1, A-1090 Wien, Austria*
[b] *Mathematisches Institut, Universität Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland*

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form 6 May 2016

MSC:
65N30

Keywords:
Banach fixed point methods
Finite element methods
Monotone problems
Quasilinear PDEs
Nonlinear elliptic PDE
Adaptive mesh refinement

ABSTRACT

In this article we investigate the use of fixed point iterations to solve the Galerkin approximation of strictly monotone problems. As opposed to Newton's method, which requires information from the previous iteration in order to linearize the iteration matrix (and thereby to recompute it) in each step, the alternative method used in this article exploits the monotonicity properties of the problem, and only needs the iteration matrix calculated once for all iterations of the fixed point method. We outline the abstract *a priori* and *a posteriori* analyses for the iteratively obtained solutions, and apply this to a finite element approximation of a second-order elliptic quasilinear boundary value problem. We show both theoretically, as well as numerically, how the number of iterations of the fixed point method can be restricted in dependence of the mesh size, or of the polynomial degree, to obtain optimal convergence. Using the *a posteriori* error analysis we also devise an adaptive algorithm for the generation of a sequence of Galerkin spaces (adaptively refined finite element meshes in the concrete example) to minimize the number of iterations on each space.

1. Introduction

In this paper we study Galerkin approximations of strictly monotone problems of the form:

$$\text{find } u \in X : A(u, v) = 0 \quad \forall v \in X. \tag{1.1}$$

Here, X is a real Hilbert space, with inner product denoted by $(\cdot, \cdot)_X$ and induced norm $\|x\| = \sqrt{(x, x)_X}$. Furthermore, $A : X \times X \to \mathbb{R}$ is a possibly nonlinear form such that, for any $u \in X$, the mapping $v \mapsto A(u, v)$ is linear and bounded. Moreover, we suppose that A satisfies

(P1) the *strong monotonicity* property

$$A(u, u - v) - A(v, u - v) \geq c_0 \|u - v\|_X^2 \quad \forall u, v \in X, \tag{P1}$$

for a constant $c_0 > 0$, and

(P2) the *Lipschitz continuity* condition

$$|A(u, w) - A(v, w)| \leq L \|u - v\|_X \|w\|_X \quad \forall u, v, w \in X, \tag{P2}$$

with a constant $L > 0$.

[*] Corresponding author.
E-mail addresses: scott.congreve@univie.ac.at (S. Congreve), wihler@math.unibe.ch (T.P. Wihler).

http://dx.doi.org/10.1016/j.cam.2016.08.014

Under these assumptions, there exists a unique solution $u \in X$ of the weak formulation (1.1); see, e.g., [1, Theorem 2.H] or [2, §3.3]. In addition, the solution can be obtained as limit of a sequence $u^0, u^1, u^2, \ldots \in X$ resulting from the fixed point iteration

$$(u^n, v)_X = (u^{n-1}, v)_X - \frac{c_0}{L^2} A(u^{n-1}, v) \quad \forall v \in X,\ n \geq 1, \tag{1.2}$$

for an arbitrary initial value $u^0 \in X$. Indeed, defining the contraction constant

$$k = \sqrt{1 - \left(\frac{c_0}{L}\right)^2}, \tag{1.3}$$

there holds the *a priori* bound

$$\|u - u^n\|_X \leq \frac{k^n}{1-k} \|u^0 - u^1\|_X, \quad n \geq 1, \tag{1.4}$$

for the iteration (1.2), i.e., $\|u - u^n\|_X \xrightarrow{n\to\infty} 0$; incidentally, under the given assumptions, this contraction constant is minimal; see, e.g., [2, Theorem 3.3.23].

Restricting the iteration (1.2) to a finite dimensional linear subspace $X_h \subseteq X$, leads to an iterative Galerkin approximation scheme for (1.1). More precisely, we consider, for an initial guess $u_h^0 \in X_h$ and $n \geq 1$, the iteration

$$u_h^n \in X_h : \ (u_h^n, v_h)_X = (u_h^{n-1}, v_h)_X - \frac{c_0}{L^2} A\left(u_h^{n-1}, v_h\right) \quad \forall v_h \in X_h, \tag{1.5}$$

where c_0 and L are the constants from (P1) and (P2), respectively; we note that (P1) and (P2) indeed hold as X_h is a conforming subspace of X. We emphasize that the problem of finding u_h^n from u_h^{n-1} in the iteration scheme (1.5) is *linear* and uniquely solvable. Similarly as for (1.1) and (1.2), the fixed point iteration (1.5) converges to the (unique) solution $u_h \in X_h$ of the Galerkin formulation

$$A(u_h, v_h) = 0 \quad \forall v_h \in X_h. \tag{1.6}$$

Furthermore, we note the *a priori* bound

$$\|u_h - u_h^n\|_X \leq \frac{k^n}{1-k} \|u_h^0 - u_h^1\|_X, \quad n \geq 1, \tag{1.7}$$

analogous to (1.4).

In solving nonlinear differential equations numerically two approaches are commonly employed. Either the nonlinear problem under consideration is discretized by means of a suitable numerical scheme thereby resulting in a (finite-dimensional) nonlinear algebraic system, or the differential equation problem is approximated by a sequence of (locally) linearized problems which are discretized subsequently. The latter approach is attractive from both a computational as well as an analytical view point; indeed, working with a sequence of linear problems allows the application of linear solvers as well as the use of a linear numerical analysis framework (e.g., in deriving error estimates). In the context of fixed point linearizations (1.5) yet another benefit comes into play: solving for u_h^n from u_h^{n-1} involves setting up and inverting a mass matrix on the left-hand side of (1.5). We emphasize that this matrix is the same for all iterations; hence, it only needs to be computed once (on a given Galerkin space).

The idea of approximating nonlinear problems within a *linear* Galerkin framework has been applied in a variety of works. For example, in the work [3], a Kačanov fixed-point iteration, whereby any nonlinear terms are expressed by means of a previously determined approximation, is employed. Furthermore, in the article [4], the authors have considered general linearizations of strongly monotone operators, and have derived computable estimators for the total error (consisting of the linearization error and the Galerkin approximation error), with identifiable components for each of the error sources. A more specific linearization approach for monotone problems, which is based on the Newton method, has been presented in [5]. In a related context linear finite element approximations resulting from adaptive Newton linearization techniques as applied to semilinear problems have been investigated in the papers [6,7]. Finally, we remark that the linear Galerkin approximation approach for nonlinear problems is not only employed for the purpose of obtaining linearized schemes, but also to address the issue of modeling errors in linearized models; see, e.g. [8,9].

The aim of the current paper is to derive *a priori* and *a posteriori* error bounds for the Galerkin iteration method (1.5). Our error estimates are expressed as the summation of the linearization error resulting from the fixed point formulation with the Galerkin approximation error. In particular, based on the *a posteriori* error analysis, we will develop an adaptive solution procedure for the numerical solution of (1.1) that features an appropriate interplay between the fixed point iterations and possible Galerkin space enrichments (e.g., mesh refinements for finite elements); specifically, our scheme selects between these two options depending on whichever constitutes the dominant part of the total error. In this way, we aim to keep the number of fixed point iterations at a minimum in the sense that no unnecessary iterations are performed if they are not expected to contribute a substantial reduction of the error on the actual Galerkin space.

The outline of the rest of this article is as follows. In Section 2 we derive an abstract analysis for the fixed point iteration (1.5), which includes the derivation of both *a priori* and *a posteriori* error bounds; in addition, we formulate an abstract adaptive procedure. The purpose of Section 3 is the application of our abstract theory to the finite element approximation of a second-order elliptic quasilinear elliptic diffusion reaction boundary value problem; in particular, we derive a fully adaptive algorithm based on suitable *a posteriori* error estimates, and provide a series of numerical experiments. Finally, in Section 4 we summarize the work presented and draw some conclusions.

2. Abstract analysis

2.1. Fixed point Galerkin approximation

As previously discussed, we let X_h be a finite dimensional linear subspace of a Hilbert space X. Then, in order to approximate the solution $u \in X$ of (1.1), we consider the Galerkin solution $u_h \in X_h$ defined in (1.6). For the purpose of calculating u_h we consider, in turn, the discrete and linear fixed point iteration scheme (1.5). Evidently, this is equivalent to a linear algebraic system of equations. More precisely, using basis functions $\phi_i \in X_h$, for $i = 1, \dots, m$, where $m = \dim(X_h)$ is the number of degrees of freedom, and letting

$$u_h^n = \sum_{i=1}^{m} \phi_i \alpha_i^n,$$

for some unknown coefficients $\boldsymbol{\alpha}^n = \{\alpha_i^n\}_{i=1}^m$, we obtain the linear system version of the fixed point iteration (1.5):

$$\sum_{i=1}^{m} \boldsymbol{M}_{ij}\alpha_i^n = \sum_{i=1}^{m} \boldsymbol{M}_{ij}\alpha_i^{n-1} - \frac{c_0}{L^2}\boldsymbol{A}(\boldsymbol{\alpha}^{n-1})_j, \quad \text{for } j = 1, \dots, m.$$

Here, $\boldsymbol{M}$ is the iteration matrix defined by $\boldsymbol{M}_{ij} = \left(\phi_i, \phi_j\right)_X$, and $\boldsymbol{A}(\boldsymbol{\alpha}^{n-1})$, with

$$\boldsymbol{A}(\boldsymbol{\alpha}^{n-1})_j = A\left(\sum_{i=1}^{m} \phi_i \alpha_i^{n-1}, \phi_j\right), \quad j = 1, \dots, m,$$

is the vector form of $A(u_h, v_h)$. We can see that the iteration matrix $\boldsymbol{M}$ does *not depend on the iteration number* n; hence, it only needs calculating once for all iterations of the fixed point method (on a given Galerkin space).

2.2. A priori error bound

Denoting by

$$e_h^n = u - u_h^n \tag{2.1}$$

the error between the solution u of (1.1) and u_h^n from (1.5), we employ the triangle inequality and (1.7) to obtain

$$\|e_h^n\|_X \leq \|u - u_h\|_X + \frac{k^n}{1-k}\|u_h^0 - u_h^1\|_X,$$

where $u_h \in X_h$ is the Galerkin solution defined in (1.6). Furthermore, employing the monotonicity property (P1) leads to

$$\begin{aligned} c_0\|u - u_h\|_X^2 &\leq A(u, u - u_h) - A(u_h, u - u_h) \\ &= A(u, u - v_h) - A(u_h, u - v_h), \end{aligned}$$

for any $v_h \in X_h$. Involving (P2), we conclude

$$c_0\|u - u_h\|_X^2 \leq L\|u - u_h\|_X\|u - v_h\|_X \quad \forall v_h \in X_h,$$

and thus

$$\|u - u_h\|_X \leq \frac{L}{c_0}\|u - v_h\|_X \quad \forall v_h \in X_h.$$

Combining these estimates we obtain the following result.

Proposition 2.1. *For the error between the solution $u \in X$ of* (1.1) *and its iterative Galerkin approximation $u_h^n \in X_h$ from* (1.5) *there holds the* a priori *error estimate*

$$\|u - u_h^n\|_X \leq \frac{L}{c_0}\inf_{v \in X_h}\|u - v\|_X + \frac{k^n}{1-k}\|u_h^0 - u_h^1\|_X,$$

for any $n \geq 1$.

2.3. A posteriori error analysis

In order to derive an *a posteriori* error analysis for (1.5) let us consider the auxiliary problem of finding $\widetilde{u}^n \in X$ such that

$$(\widetilde{u}^n, v)_X = (u_h^{n-1}, v)_X - \frac{c_0}{L^2} A(u_h^{n-1}, v) \quad \forall v \in X,\ n \geq 1. \tag{2.2}$$

We note that $\widetilde{u}^n \in X$ is a *reconstruction* (cf. [10]) in the sense that $u_h^n \in X_h$ from (1.5) is the Galerkin approximation of $\widetilde{u}^n$. We *assume* that we can bound the error between the solution $\widetilde{u}^n \in X$ of (2.2) and its Galerkin approximation $u_h^n \in X_h$ in terms of an *a posteriori* computable quantity $\eta(u_h^n, X_h)$, i.e.,

$$\|\widetilde{u}^n - u_h^n\|_X \leq \eta(u_h^n, X_h), \quad n \geq 1. \tag{2.3}$$

For instance, if (2.3) corresponds to a (partial) differential equation for which computable *a posteriori* error estimators (of preferably optimal order) for the error in the associated norm $\|\cdot\|_X$ are available, then $\eta(u_h^n, X_h)$ can be represented by *any* of such quantities. Involving the monotonicity property (P1), the error e_h^n from (2.1) satisfies

$$c_0\|e_h^n\|_X^2 \leq A(u, e_h^n) - A(u_h^n, e_h^n) = -A(u_h^n, e_h^n).$$

Furthermore, recalling (2.2), we write

$$\begin{aligned} c_0\|e_h^n\|_X^2 &\leq A(u_h^{n-1}, e_h^n) - A(u_h^n, e_h^n) + \frac{L^2}{c_0}(\widetilde{u}^n - u_h^{n-1}, e_h^n)_X \\ &= A(u_h^{n-1}, e_h^n) - A(u_h^n, e_h^n) + \frac{L^2}{c_0}(u_h^n - u_h^{n-1}, e_h^n)_X + \frac{L^2}{c_0}(\widetilde{u}^n - u_h^n, e_h^n)_X. \end{aligned}$$

Then, using (P2) and applying the Cauchy–Schwarz inequality, we infer that

$$c_0\|e_h^n\|_X^2 \leq L\|u_h^{n-1} - u_h^n\|_X\|e_h^n\|_X + \frac{L^2}{c_0}\|u_h^n - u_h^{n-1}\|_X\|e_h^n\|_X + \frac{L^2}{c_0}\|\widetilde{u}^n - u_h^n\|_X\|e_h^n\|_X,$$

and dividing by $c_0\|e_h^n\|_X$, we obtain

$$\|e_h^n\|_X \leq \frac{L}{c_0}\left(1 + \frac{L}{c_0}\right)\|u_h^n - u_h^{n-1}\|_X + \frac{L^2}{c_0^2}\|\widetilde{u}^n - u_h^n\|_X.$$

Hence, inserting (2.3), the following result can be deduced.

Proposition 2.2. *For the error between the solution $u \in X$ of* (1.1) *and its iterative Galerkin approximation $u_h^n \in X_h$ from* (1.5) *there holds the* a posteriori *error estimate*

$$\|u - u_h^n\|_X \leq \frac{L}{c_0}\left(1 + \frac{L}{c_0}\right)\|u_h^n - u_h^{n-1}\|_X + \frac{L^2}{c_0^2}\eta(u_h^n, X_h), \tag{2.4}$$

where $\eta(u_h^n, X_h)$ satisfies (2.3).

2.4. An abstract adaptive algorithm

The *a posteriori* error estimate (2.4) shows that the error e_h^n from (2.1) is controlled by two separate parts: a fixed point iteration error given by $Lc_0^{-1}\left(1 + Lc_0^{-1}\right)\|u_h^n - u_h^{n-1}\|_X$, and a Galerkin approximation term $L^2c_0^{-2}\eta(u_h^n, X_h)$. When performing the fixed point iteration (1.5) it is worth noting that once the fixed point error is less than the Galerkin error carrying out another iteration will not cause a substantial reduction of the error on the actual Galerkin space. Based on this observation we are able to develop an algorithm which generates a sequence of hierarchically enriched Galerkin spaces $X_{h,1} \subset X_{h,2} \subset X_{h,3} \subset \cdots \subset X$, with the aim of performing a minimal number of fixed point iterations at each enrichment step. Our algorithm will, therefore, feature an interplay between fixed point iterations and Galerkin space refinements.

On the Galerkin space $X_{h,i}$, $i \geq 1$, we define the *Galerkin approximation error* by

$$\mathcal{E}^n_{\mathsf{Galerkin},i} := \frac{L^2}{c_0^2}\eta(u^n_{h,i}, X_{h,i}),$$

and the *fixed point error*

$$\mathcal{E}^n_{\mathsf{FP},i} := \frac{L}{c_0}\left(1 + \frac{L}{c_0}\right)\|u^n_{h,i} - u^{n-1}_{h,i}\|_X.$$

This allows us to write the *a posteriori* error bound (2.4) as

$$\|u - u^n_{h,i}\|_X \leq \mathcal{E}^n_{\mathsf{Galerkin},i} + \mathcal{E}^n_{\mathsf{FP},i}.$$

Here, we denote by $u^n_{h,i} \in X_{h,i}$ the Galerkin solution obtained after n steps of the fixed point iteration (1.5) on the current space $X_{h,i}$; for $i > 0$, the initial guess $u^0_{h,i} \in X_{h,i}$ on the current Galerkin space $X_{h,i}$ is obtained as the natural inclusion (or a projection) of the solution $u^{n^\star}_{h,i-1} \in X_{h,i-1}$ of the last (namely, the $n^\star$th) iteration on the previous Galerkin space $X_{h,i-1}$ to the space $X_{h,i}$. In particular, the fixed point iteration index n is reinitialized in each space enrichment step.

Algorithm 2.3. Choose an initial starting space $X_{h,0}$, and an initial guess $u^0_{h,0} \in X_{h,0}$.

```
for $i \leftarrow 0, 1, 2, \dots$ do
    $n \leftarrow 0$
    repeat
        $n \leftarrow n + 1$
        Perform a single fixed point iteration (1.5) to calculate $u^n_{h,i} \in X_{h,i}$.
    until $\mathcal{E}^n_{\mathrm{FP},i} \leq \vartheta \mathcal{E}^n_{\mathrm{Galerkin},i}$
    Perform a hierarchical enrichment of $X_{h,i}$
      based on the error indicator $\eta(u^n_{h,i}, X_{h,i})$ from (2.3)
      to obtain a new Galerkin space $X_{h,i+1} \supset X_{h,i}$.
    Define $u^0_{h,i+1} \leftarrow u^n_{h,i}$ (by inclusion $X_{h,i+1} \hookleftarrow X_{h,i}$ or by projection)
end for
```

Here, $\vartheta > 0$ is a prescribed parameter. The algorithm is stopped if either the iteration number i reaches a given maximum, or if the right-hand side of (2.4) is found to be sufficiently small.

3. Application to quasilinear elliptic PDE

3.1. Problem formulation

In this section, we focus on the numerical approximation of second-order elliptic diffusion reaction boundary value problems of the form

$$-\nabla \cdot (\mu(\boldsymbol{x}, |\nabla u|)\nabla u) + f(\boldsymbol{x}, u) = 0 \quad \text{in } \Omega, \tag{3.1}$$

$$u = 0 \quad \text{on } \Gamma, \tag{3.2}$$

where Ω is a bounded, open, polygonal domain in $\mathbb{R}^2$, with boundary $\Gamma = \partial\Omega$. Here, we assume the following *monotonicity* conditions on the nonlinearities μ and f:

1. $\mu \in C^0(\overline{\Omega} \times [0, \infty))$, and there exist constants $\alpha_1 \geq \alpha_2 > 0$ such that the following property is satisfied:

$$\alpha_2(t - s) \leq \mu(\boldsymbol{x}, t)t - \mu(\boldsymbol{x}, s)s \leq \alpha_1(t - s), \qquad t \geq s \geq 0, \ \boldsymbol{x} \in \overline{\Omega}. \tag{3.3}$$

2. $f \in C^0(\overline{\Omega} \times \mathbb{R})$, and there exist constants $\beta_1 \geq \beta_2 \geq 0$ such that

$$\beta_2(t - s) \leq f(\boldsymbol{x}, t) - f(\boldsymbol{x}, s) \leq \beta_1(t - s), \qquad t \geq s, \ \boldsymbol{x} \in \overline{\Omega}. \tag{3.4}$$

From [11, Lemma 2.1] we note that, as μ satisfies (3.3), for all vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^2$ and all $\boldsymbol{x} \in \overline{\Omega}$, we have

$$|\mu(\boldsymbol{x}, |\boldsymbol{v}|)\boldsymbol{v} - \mu(\boldsymbol{x}, |\boldsymbol{w}|)\boldsymbol{w}| \leq \alpha_1|\boldsymbol{v} - \boldsymbol{w}|, \tag{3.5}$$

$$\alpha_2|\boldsymbol{v} - \boldsymbol{w}|^2 \leq (\mu(\boldsymbol{x}, |\boldsymbol{v}|)\boldsymbol{v} - \mu(\boldsymbol{x}, |\boldsymbol{w}|)\boldsymbol{w}) \cdot (\boldsymbol{v} - \boldsymbol{w}). \tag{3.6}$$

Similarly, as f satisfies (3.4), it holds that for all $s, t \in \mathbb{R}$ and all $\boldsymbol{x} \in \overline{\Omega}$,

$$|f(\boldsymbol{x}, t) - f(\boldsymbol{x}, s)| \leq \beta_1|t - s|, \tag{3.7}$$

$$\beta_2|t - s|^2 \leq (f(\boldsymbol{x}, t) - f(\boldsymbol{x}, s))(t - s). \tag{3.8}$$

For ease of notation we shall write $\mu(t)$ and $f(t)$ instead of $\mu(\boldsymbol{x}, t)$ and $f(\boldsymbol{x}, t)$, respectively; thereby, suppressing the dependence of μ and f on $\boldsymbol{x}$. Incidentally, we remark that the continuity of these functions with respect to the first variable, $\boldsymbol{x}$, can be weakened.

The weak formulation of the boundary value problem (3.1)–(3.2) is to find $u \in X := H^1_0(\Omega)$ such that

$$A(u, v) = 0 \quad \forall v \in H^1_0(\Omega), \tag{3.9}$$

where

$$A(u, v) = \int_\Omega \{\mu(|\nabla u|)\nabla u \cdot \nabla v + f(u)v\} \,\mathrm{d}\boldsymbol{x}, \qquad u, v \in H^1_0(\Omega). \tag{3.10}$$

Throughout this section, we use function spaces based on a polygonal Lipschitz domain $D \subset \mathbb{R}^2$. We denote by $H^k(D)$ the Sobolev space of order $k \in \mathbb{N}_0$ endowed with the norm $\|\cdot\|_{H^k(D)}$. In the case that $k = 0$, we set $H^k(D) = L_2(D)$ and denote the matching norm by $\|\cdot\|_{L_2(D)}$. Furthermore, we define $H_0^1(D)$ as the space of functions in $H^1(D)$ with zero trace on ∂D.

Introducing the inner product

$$(u, v)_\Omega := \int_\Omega \{\alpha_2 \nabla u \cdot \nabla v + \beta_2 uv\} \,\mathrm{d}\boldsymbol{x}, \quad u, v \in H_0^1(\Omega),$$

where α_2 and β_2 are the constants from (3.3) and (3.4), respectively, we note the induced norm

$$|\!|\!| v |\!|\!|_\Omega^2 := \alpha_2 \|\nabla v\|_{L_2(\Omega)}^2 + \beta_2 \|v\|_{L_2(\Omega)}^2$$

on $H_0^1(\Omega)$. In view of a robust error analysis, let us emphasize that the weights α_2, β_2 in the above norm play a crucial role, for example, in the singularly perturbed case $\beta_2/\alpha_2 \gg 1$; indeed, in such cases, using the standard H^1-norm may potentially lead to ill-scaled error estimates.

3.2. Basic properties

Under the conditions (3.3) and (3.4) we can show that the properties (P1) and (P2) are satisfied for $X = H_0^1(\Omega)$, and $\|\cdot\|_X := |\!|\!|\cdot|\!|\!|_\Omega$. Indeed, noting the Poincaré–Friedrichs inequality,

$$\|v\|_{L_2(\Omega)} \le C_P \|\nabla v\|_{L_2(\Omega)} \quad \forall v \in H_0^1(\Omega), \tag{3.11}$$

where $C_P > 0$ is a constant dependent only on Ω, there holds the ensuing result.

Proposition 3.1. *Provided that (3.3) and (3.4) hold, then the form A from (3.10) is both strongly monotone with constant $c_0 = 1$ in (P1), and Lipschitz continuous with constant*

$$1 \le L = \frac{\alpha_1 + \max(\beta_1, \alpha_1\beta_2/\alpha_2)\, C_P^2}{\alpha_2 + \beta_2 C_P^2} \tag{3.12}$$

in (P2) with respect to the norm $|\!|\!|\cdot|\!|\!|_\Omega$. Here, $C_P > 0$ is the Poincaré–Friedrichs constant from (3.11).

Proof. In order to show (P1) with $c_0 = 1$, we apply (3.6) and (3.8) to arrive at

$$\begin{aligned} A(u, u - v) - A(v, u - v) &= \int_\Omega (\mu(|\nabla u|)\nabla u - \mu(|\nabla v|)\nabla v) \cdot \nabla(u - v) \,\mathrm{d}\boldsymbol{x} + \int_\Omega (f(u) - f(v))\,(u - v) \,\mathrm{d}\boldsymbol{x} \\ &\ge \int_\Omega \left\{\alpha_2 |\nabla(u - v)|^2 + \beta_2 |u - v|^2\right\} \mathrm{d}\boldsymbol{x} = |\!|\!| u - v |\!|\!|_\Omega^2. \end{aligned}$$

Furthermore, to prove the Lipschitz continuity property (P2), we recall (3.5) and (3.7). In combination with the Cauchy–Schwarz inequality, this yields

$$\begin{aligned} |A(u, w) - A(v, w)| &\le \int_\Omega |\mu(|\nabla u|)\nabla u - \mu(|\nabla v|)\nabla v||\nabla w| \,\mathrm{d}\boldsymbol{x} + \int_\Omega |f(u) - f(v)||w| \,\mathrm{d}\boldsymbol{x} \\ &\le \alpha_1 \|\nabla(u - v)\|_{L_2(\Omega)} \|\nabla w\|_{L_2(\Omega)} + \beta_1 \|u - v\|_{L_2(\Omega)} \|w\|_{L_2(\Omega)}. \end{aligned}$$

We first consider the case when $\beta_2 = 0$; then, noting that $|\!|\!| v |\!|\!|_\Omega = \sqrt{\alpha_2}\|\nabla v\|_{L_2(\Omega)}$, we apply the Poincaré–Friedrichs inequality (3.11) to observe that

$$\begin{aligned} |A(u, w) - A(v, w)| &\le \alpha_1 \|\nabla(u - v)\|_{L_2(\Omega)} \|\nabla w\|_{L_2(\Omega)} + \beta_1 C_P^2 \|\nabla(u - v)\|_{L_2(\Omega)} \|\nabla w\|_{L_2(\Omega)} \\ &= \left(\frac{\alpha_1 + \beta_1 C_P^2}{\alpha_2}\right) |\!|\!| u - v |\!|\!|_\Omega |\!|\!| w |\!|\!|_\Omega. \end{aligned}$$

When $\beta_2 > 0$ we introduce a constant $0 \le \delta \le \beta_1$ and apply the Poincaré–Friedrichs inequality (3.11), to get that

$$\begin{aligned} |A(u, w) - A(v, w)| &\le \alpha_1 \|\nabla(u - v)\|_{L_2(\Omega)} \|\nabla w\|_{L_2(\Omega)} + (\beta_1 - \delta) \|u - v\|_{L_2(\Omega)} \|w\|_{L_2(\Omega)} \\ &\quad + \delta C_P^2 \|\nabla(u - v)\|_{L_2(\Omega)} \|\nabla w\|_{L_2(\Omega)} \\ &= \left(\frac{\alpha_1 + \delta C_P^2}{\alpha_2}\right) \sqrt{\alpha_2}\|\nabla(u - v)\|_{L_2(\Omega)} \sqrt{\alpha_2}\|\nabla w\|_{L_2(\Omega)} \\ &\quad + \left(\frac{\beta_1 - \delta}{\beta_2}\right) \sqrt{\beta_2}\|u - v\|_{L_2(\Omega)} \sqrt{\beta_2}\|w\|_{L_2(\Omega)}. \end{aligned}$$

Using the Cauchy–Schwarz inequality yields

$$|A(u,w)-A(v,w)| \le L(\delta)|\!|\!|u-v|\!|\!|_{\Omega}|\!|\!|w|\!|\!|_{\Omega},$$

where

$$L(\delta)=\max\left(\frac{\alpha_1+\delta C_P^2}{\alpha_2},\frac{\beta_1-\delta}{\beta_2}\right).$$

Minimizing $L(\delta)$ within the given range, $0\le\delta\le\beta_1$, depends on the constants $\alpha_1,\alpha_2,\beta_1,\beta_2$. More precisely, if ${}^{\alpha_1}/_{\alpha_2}\ge{}^{\beta_1}/_{\beta_2}$ then $\delta^\star=0$ is the optimal choice, and there holds $L(0)={}^{\alpha_1}/_{\alpha_2}\ge 1$; otherwise, we select the optimal choice

$$\delta^\star:=\frac{\beta_1\alpha_2-\beta_2\alpha_1}{\alpha_2+\beta_2C_P^2}\in(0,\beta_1),$$

by setting ${}^{(\alpha_1+\delta^*C_P^2)}/_{\alpha_2}={}^{(\beta_1-\delta^*)}/_{\beta_2}$, and thereby obtain the minimum

$$L(\delta^\star)=\frac{\alpha_1+\beta_1C_P^2}{\alpha_2+\beta_2C_P^2}\ge 1.$$

This completes the proof. □

Remark 3.2. Incidentally, referring to, e.g., [2, Theorem 3.3.23] or [1, Theorem 2.H], the above result, Proposition 3.1, guarantees the existence of a unique solution $u\in H_0^1(\Omega)$ of (3.9).

Remark 3.3. We note that the fixed point iteration (1.2) for the current problem (3.1)–(3.2) reads in strong form as

$$\begin{aligned}-\alpha_2\Delta u^n+\beta_2u^n&=-\alpha_2\Delta u^{n-1}+\beta_2u^{n-1}-L^{-2}\left(-\nabla\cdot(\mu(|\nabla u^{n-1}|)\nabla u^{n-1})+f(u^{n-1})\right) && \text{in } \Omega\\ u^n&=0 && \text{on } \partial\Omega,\end{aligned}$$

in $H^{-1}(\Omega)$, the dual space of $H_0^1(\Omega)$, for $n\ge 1$.

Remark 3.4. From (3.12) it becomes clear, unless $\beta_1=\beta_2=0$, that the constant L depends on the Poincaré–Friedrichs constant $C_P=C_P(\Omega)$ from (3.11). In practice, in order to compute L, upper bounds on C_P need to be determined; we refer, e.g., to the works [12,13] and the references therein for details on this matter.

3.3. *Finite element discretization*

In order to solve (3.9) by a fixed point Galerkin iteration, we will use a finite element framework.

3.3.1. *Meshes and spaces*

We consider regular and shape-regular meshes $\mathcal{T}_h$ that partition the domain $\Omega\subset\mathbb{R}^2$ into open disjoint triangles and/or parallelograms K, such that $\overline{\Omega}=\bigcup_{K\in\mathcal{T}_h}\overline{K}$. We denote by h_K the elemental diameter of $K\in\mathcal{T}_h$, and let $h=\max_{K\in\mathcal{T}_h}h_K$.

With this notation, for a fixed polynomial degree $p\ge 1$, we are ready to introduce the finite element space

$$V_{\mathsf{FEM}}:=\{v\in H_0^1(\Omega):v|_K\in\mathcal{S}_p(K)\ \forall K\in\mathcal{T}_h\}\subset H_0^1(\Omega),\tag{3.13}$$

where

$$\mathcal{S}_p(K)=\begin{cases}\mathcal{P}_p(K) & \text{if } K \text{ is a triangle,}\\ \mathcal{Q}_p(K) & \text{if } K \text{ is a parallelogram.}\end{cases}$$

Here, $\mathcal{P}_p(K)$ denotes the space of polynomials of total order at most p on K, while $\mathcal{Q}_p(K)$ is the tensored space of polynomials of order at most p in each variable on K. The assumptions (3.3) and (3.4) hold on this space as V_{FEM} is a conforming subspace of $H_0^1(\Omega)$.

3.3.2. *Iterative Galerkin FEM*

Based on the class of spaces V_{FEM} introduced before, we can now introduce the finite element formulation for a *linear* fixed point formulation (1.5) of (3.9): Given an initial guess $u_h^0\in V_{\mathsf{FEM}}$, we iterate for $n=1,2,3\dots,$

$$\left(u_h^n,v_h\right)_{\Omega}=\left(u_h^{n-1},v_h\right)_{\Omega}-L^{-2}A(u_h^{n-1},v_h)\quad\forall v_h\in V_{\mathsf{FEM}}.\tag{3.14}$$

Remark 3.5. Recalling (1.3) the contraction constant for this iteration is given by

$$k=\sqrt{1-L^{-2}}<1.$$

Here, we point out that, in the singularly perturbed case when $\alpha_2\approx\alpha_1=\mathcal{O}(\varepsilon)$, for $0<\varepsilon\ll 1$, and $\beta_2\approx\beta_1=\mathcal{O}(1)$, the contraction factor k does not deteriorate to 1. Indeed, this follows from the fact that the Lipschitz constant L from Proposition 3.1 remains robustly bounded from 0 in this situation.

3.4. Error analysis

We will now apply the abstract analysis derived in Section 2 to the iterative Galerkin method (3.14) for the numerical approximation of (3.1)–(3.2).

3.4.1. A priori error bound

Using our abstract *a priori* error analysis from Section 2, we are able to obtain a bound for the error between the numerical solution u_h^n obtained at the nth step of the fixed point iteration (3.14) and of the exact solution u from (3.9). For simplicity of presentation we assume a (quasi-) uniform diameter $h > 0$ of all elements.

Theorem 3.6. *Let $u \in H^{\kappa+1}(\Omega) \cap H_0^1(\Omega)$, with $\kappa \geq 1$, be the solution to the weak formulation (3.9), $u_h^0 \in V_{\mathsf{FEM}}$ any initial guess, and $u_h^n \in V_{\mathsf{FEM}}$ the numerical solution after n steps of the fixed point iteration (3.14); then, for $n \geq 1$, there holds the* a priori *error estimate*

$$|||u - u_h^n|||_\Omega \leq CL \frac{h^{\min(\kappa,p)}}{p^\kappa} \|u\|_{H^{\kappa+1}(\Omega)} + 2L^2 \left(1 - L^{-2}\right)^{n/2} |||u_h^0 - u_h^1|||_\Omega, \tag{3.15}$$

where $C > 0$ is a constant independent of h, p, and L from (3.12), but depends on α_2 and β_2 from (3.3) and (3.4), respectively.

Proof. This follows directly from Proposition 2.1, and by applying standard hp-approximation results (see, e.g., [14]).

Remark 3.7. The hp-approximation bounds yielding the first term on the right-hand side of (3.15) require H^k-regularity of u, with $k \geq 2$. We shall now discuss briefly how the parameters in (3.15) influence the (approximate) number of fixed point iterations required to obtain an optimal convergence rate in the linear finite element iteration (3.14). To this end, we ask for the second term on the right-hand side of (3.15) to converge at least at the rate of the first term. In order to discuss the resulting convergence behavior of the numerical solution u_h^n obtained from (3.14), we distinguish two different cases:

- *h-FEM*: We fix a low polynomial degree p and investigate the convergence properties with respect to the mesh size h as $h \to 0$ (mesh refinement). Here, for $\kappa \geq p$, we need $(1 - L^{-2})^{n/2} = \mathcal{O}(h^p)$, and hence, $n = \mathcal{O}(|\log h|)$ as $h \to 0$.
- *p-FEM*: We now fix the mesh, and suppose that the solution of (3.1)–(3.2) is analytic. Then, as $p \to \infty$, it can be shown that the FEM converges exponentially (see [15]), i.e., the error bound (3.15) reads

$$|||u - u_h^n|||_\Omega \leq \mathcal{O}\left(e^{-bp}\right) + 2L^2 \left(1 - L^{-2}\right)^{n/2} |||u_h^0 - u_h^1|||_\Omega,$$

for some constant $b > 0$. Again, balancing the two terms on the right, we require $n = \mathcal{O}(p)$ iterations as $p \to \infty$.

We will test these observations with some numerical experiments in Section 3.6.

3.4.2. A posteriori error analysis

In this section we obtain an *a posteriori* error bound for the error between the numerical solution u_h^n obtained at the nth step of the fixed point iteration (3.14) and of the exact solution u obtained from (3.1)–(3.2). According to our analysis in Section 2.3 the key is to derive an *a posteriori* error estimate between the reconstruction $\widetilde{u}^n \in H_0^1(\Omega)$ from (2.2) and the iterative Galerkin solution u_h^n from (1.5) (i.e., u_h^n from (3.14) in the present context); see (2.3).

To establish such a bound, we begin with a quasi-interpolation result.

Lemma 3.8. *Consider a finite element mesh $\mathcal{T}_h$, and a corresponding FEM space V_{FEM} as in (3.13). Moreover, let $\pi : H_0^1(\Omega) \to V_{\mathsf{FEM}}$ be the Clément interpolation operator [16]. Then,*

$$\sum_{K \in \mathcal{T}_h} \left(\gamma_K^{-1} \|v - \pi v\|_{L_2(K)}^2 + \alpha_2 \|\nabla (v - \pi v)\|_{L_2(K)}^2 + \frac{1}{2} \alpha_2^{1/2} \gamma_K^{-1/2} \|v - \pi v\|_{L_2(\partial K)}^2 \right) \leq C_I^2 |||v|||_\Omega^2$$

for all $v \in H_0^1(\Omega)$, with a constant $C_I > 0$ independent of the local element sizes, and

$$\gamma_K = \begin{cases} \min\left(\alpha_2^{-1} h_K^2, \beta_2^{-1}\right) & \text{if } \beta_2 \neq 0, \\ h_K^2 \alpha_2^{-1} & \text{otherwise,} \end{cases} \tag{3.16}$$

for any $K \in \mathcal{T}_h$. Here, α_2 and β_2 are the constants from (3.3) and (3.4), respectively.

Proof. Let $v \in H_0^1(\Omega)$. We begin by recalling the following well-known approximation properties of the Clément interpolant:

$$h_K^{-2} \|v - \pi v\|_{L_2(K)}^2 + \|\nabla (v - \pi v)\|_{L_2(K)}^2 \leq C \|\nabla v\|_{L_2(\omega_K)}^2, \qquad \|v - \pi v\|_{L_2(K)}^2 \leq C \|v\|_{L_2(\omega_K)}^2,$$

for any $K \in \mathcal{T}_h$, with a constant $C > 0$ independent of the local element sizes and of v; for $K \in \mathcal{T}_h$ we denote by ω_K the patch of all elements in $\mathcal{T}_h$ adjacent to K. In particular, following the approach in [17], this implies that

$$\|v - \pi v\|^2_{L_2(K)} \leq C\alpha_2^{-1} h_K^2 \left(\alpha_2 \|\nabla v\|^2_{L_2(\omega_K)} + \beta_2 \|v\|^2_{L_2(\omega_K)}\right),$$

and that, if $\beta_2 \neq 0$, then

$$\|v - \pi v\|^2_{L_2(K)} \leq C\beta_2^{-1} \left(\alpha_2 \|\nabla v\|^2_{L_2(\omega_K)} + \beta_2 \|v\|^2_{L_2(\omega_K)}\right).$$

Therefore,

$$\|v - \pi v\|^2_{L_2(K)} \leq C\gamma_K \left(\alpha_2 \|\nabla v\|^2_{L_2(\omega_K)} + \beta_2 \|v\|^2_{L_2(\omega_K)}\right),$$

and so

$$\gamma_K^{-1} \|v - \pi v\|^2_{L_2(K)} + \alpha_2 \|\nabla (v - \pi v)\|^2_{L_2(K)} \leq C \left(\alpha_2 \|\nabla v\|^2_{L_2(\omega_K)} + \beta_2 \|v\|^2_{L_2(\omega_K)}\right). \tag{3.17}$$

Moreover, using the multiplicative trace inequality, that is,

$$\|v - \pi v\|^2_{L_2(\partial K)} \leq C \left(h_K^{-1} \|v - \pi v\|^2_{L_2(K)} + \|v - \pi v\|_{L_2(K)} \|\nabla (v - \pi v)\|_{L_2(K)}\right) \quad \forall K \in \mathcal{T}_h,$$

we infer that

$$\|v - \pi v\|^2_{L_2(\partial K)} \leq C \left(h_K^{-1}\gamma_K + \gamma_K^{1/2} \alpha_2^{-1/2}\right) \left(\alpha_2 \|\nabla v\|^2_{L_2(\omega_K)} + \beta_2 \|v\|^2_{L_2(\omega_K)}\right).$$

Observing that

$$h_K^{-1}\gamma_K + \gamma_K^{1/2}\alpha_2^{-1/2} = \left(h_K^{-1}\alpha_2^{1/2}\gamma_K^{1/2} + 1\right)\gamma_K^{1/2}\alpha_2^{-1/2} \leq 2\gamma_K^{1/2}\alpha_2^{-1/2},$$

we now arrive at

$$\alpha_2^{1/2}\gamma_K^{-1/2} \|v - \pi v\|^2_{L_2(\partial K)} \leq C \left(\alpha_2 \|\nabla v\|^2_{L_2(\omega_K)} + \beta_2 \|v\|^2_{L_2(\omega_K)}\right). \tag{3.18}$$

Finally, combining (3.17) and (3.18), and summation over all $K \in \mathcal{T}_h$ concludes the argument. □

In order to formulate the following result, we consider a series of meshes, $\{\mathcal{T}_{h,i}\}_{i \geq 0}$; for each mesh $\mathcal{T}_{h,i}$ we denote the finite element space on that mesh as $\mathsf{V}_{\mathsf{FEM},i}$.

Theorem 3.9. *Let $u \in H_0^1(\Omega)$ be the exact solution to the boundary value problem (3.1)–(3.2), and $\mathcal{T}_{h,0}$ be an initial mesh with initial guess $u_{h,0}^0 \in \mathsf{V}_{\mathsf{FEM},0}$. Moreover, denote by $\mathcal{T}_{h,i}$ the mesh after i mesh refinements, and let $u_{h,i}^n \in \mathsf{V}_{\mathsf{FEM},i}$ be the FEM solution obtained after n steps of the fixed point iteration (3.14) on $\mathcal{T}_{h,i}$. Here, the initial guess $u_{h,i}^0 \in \mathsf{V}_{\mathsf{FEM},i}$ on the current mesh $\mathcal{T}_{h,i}$, $i > 0$, is obtained as an (appropriate) projection of the solution $u_{h,i-1}^{n^\star} \in \mathsf{V}_{\mathsf{FEM},i-1}$ of the last ($n^\star$th) iteration on the mesh $\mathcal{T}_{h,i-1}$ to the space $\mathsf{V}_{\mathsf{FEM},i}$. Then, for $n \geq 1$, there holds the* a posteriori *error estimate*

$$|\!|\!| u - u_{h,i}^n |\!|\!|_\Omega \leq C_I \left(\sum_{K \in \mathcal{T}_{h,i}} \eta_K^2\right)^{1/2} + L(1 + L)\, |\!|\!| u_{h,i}^n - u_{h,i}^{n-1} |\!|\!|_\Omega, \tag{3.19}$$

where C_I is the constant from Lemma 3.8, and

$$\eta_K^2 = \gamma_K \left\| f\left(u_{h,i}^{n-1}\right) - \nabla \cdot \left(\mu\left(|\nabla u_{h,i}^{n-1}|\right) \nabla u_{h,i}^{n-1}\right) + L^2 \mathsf{F}\left(u_{h,i}^n - u_{h,i}^{n-1}\right) \right\|^2_{L_2(K)}$$
$$+ \frac{1}{2}\alpha_2^{-1/2}\gamma_K{}^{1/2} \left\| \left[\!\left[\mu\left(|\nabla u_{h,i}^{n-1}|\right) \nabla u_{h,i}^{n-1} + L^2 \alpha_2 \nabla \left(u_{h,i}^n - u_{h,i}^{n-1}\right) \right]\!\right] \right\|^2_{L_2(\partial K \setminus \Gamma)},$$

for any $K \in \mathcal{T}_{h,i}$ and $n \geq 1$, are local error indicators. *Here, γ_K, $K \in \mathcal{T}_{h,i}$, is defined in (3.16) and*

$$\mathsf{F}(v) = -\alpha_2 \Delta v + \beta_2 v.$$

Moreover, for an edge $e \subset \partial K^+ \cap \partial K^-$ between two neighboring elements $K^\pm \in \mathcal{T}_{h,i}$, we signify by $[\![\boldsymbol{v}]\!]\big|_e = \boldsymbol{v}^+|_e \cdot \boldsymbol{n}_{K^+} + \boldsymbol{v}^-|_e \cdot \boldsymbol{n}_{K^-}$ the jump of a (vector-valued) function $\boldsymbol{v}$ along e, where $\boldsymbol{v}^\pm$ denote the traces of the function $\boldsymbol{v}$ on the edge e taken from the interior of $K^\pm$, respectively, and $\boldsymbol{n}_{K^\pm}$ are the unit outward normal vectors on $\partial K^\pm$, respectively.

Proof. Recalling our abstract result, Proposition 2.2, it is sufficient to derive a quantity $\eta(u_{h,i}^n, \mathsf{V}_{\mathsf{FEM},i})$ such that

$$|\!|\!| \widetilde{u}_i^n - u_{h,i}^n |\!|\!|_\Omega \leq \eta(u_{h,i}^n, \mathsf{V}_{\mathsf{FEM},i});$$

cf. (2.3). The reconstruction $\widetilde{u}_i^n \in H_0^1(\Omega)$ fulfills

$$(\widetilde{u}_i^n, v)_\Omega = (u_{h,i}^{n-1}, v)_\Omega - L^{-2} A(u_{h,i}^{n-1}, v) \quad \forall v \in H_0^1(\Omega) \supset \mathsf{V}_{\mathsf{FEM},i},$$

see (2.2), where $u_{h,i}^n \in \mathsf{V}_{\mathsf{FEM},i}$ is just the Galerkin approximation of $\widetilde{u}_i^n$ (cp. (3.14)).

Define the error $\widetilde{e}^n_{h,i} = \widetilde{u}^n_i - u^n_{h,i}$, and let $v_{h,i} = \pi \widetilde{e}^n_{h,i} \in V_{\mathsf{FEM},i}$, where π is the interpolation operator from Lemma 3.8. We notice that there holds

$$
\begin{aligned}
|||\widetilde{e}^n_{h,i}|||^2_{\Omega} &= (\widetilde{u}^n_i, \widetilde{e}^n_{h,i})_{\Omega} - (u^n_{h,i}, \widetilde{e}^n_{h,i})_{\Omega} \\
&= (\widetilde{u}^n_i, \widetilde{e}^n_{h,i} - v_{h,i})_{\Omega} - (u^n_{h,i}, \widetilde{e}^n_{h,i} - v_{h,i})_{\Omega} \\
&= -L^{-2} A(u^{n-1}_{h,i}, \widetilde{e}^n_{h,i} - v_{h,i}) - \left(u^n_{h,i} - u^{n-1}_{h,i}, \widetilde{e}^n_{h,i} - v_{h,i}\right)_{\Omega} .
\end{aligned}
$$

Integration by parts elementwise leads to

$$
\begin{aligned}
L^2 |||\widetilde{e}^n_{h,i}|||^2_{\Omega} = &- \sum_{K \in \mathcal{T}_{h,i}} \int_K \left(\mu\left(|\nabla(u^{n-1}_{h,i})|\right) \nabla u^{n-1}_{h,i} + L^2 \alpha_2 \nabla\left(u^n_{h,i} - u^{n-1}_{h,i}\right)\right) \cdot \nabla\left(\widetilde{e}^n_{h,i} - v_{h,i}\right) \,\mathrm{d}\boldsymbol{x} \\
&- \sum_{K \in \mathcal{T}_{h,i}} \int_K \left(f\left(u^{n-1}_{h,i}\right) + L^2 \beta_2 \left(u^n_{h,i} - u^{n-1}_{h,i}\right)\right)\left(\widetilde{e}^n_{h,i} - v_{h,i}\right) \,\mathrm{d}\boldsymbol{x} \\
= &\sum_{K \in \mathcal{T}_{h,i}} \int_K \nabla \cdot \left(\mu\left(|\nabla u^{n-1}_{h,i}|\right) \nabla u^{n-1}_{h,i} + L^2 \alpha_2 \nabla\left(u^n_{h,i} - u^{n-1}_{h,i}\right)\right)\left(\widetilde{e}^n_{h,i} - v_{h,i}\right) \,\mathrm{d}\boldsymbol{x} \\
&- \sum_{K \in \mathcal{T}_{h,i}} \int_K \left(f\left(u^{n-1}_{h,i}\right) + L^2 \beta_2 \left(u^n_{h,i} - u^{n-1}_{h,i}\right)\right)\left(\widetilde{e}^n_{h,i} - v_{h,i}\right) \,\mathrm{d}\boldsymbol{x} \\
&- \sum_{K \in \mathcal{T}_{h,i}} \int_{\partial K \setminus \Gamma} \left(\mu\left(|\nabla u^{n-1}_{h,i}|\right) \nabla u^{n-1}_{h,i} + L^2 \alpha_2 \nabla\left(u^n_{h,i} - u^{n-1}_{h,i}\right)\right) \cdot \boldsymbol{n}_K \left(\widetilde{e}^n_{h,i} - v_{h,i}\right) \,\mathrm{d}s.
\end{aligned}
$$

A few elementary calculations show that

$$
\begin{aligned}
&\sum_{K \in \mathcal{T}_{h,i}} \int_{\partial K \setminus \Gamma} \left(\mu\left(|\nabla u^{n-1}_{h,i}|\right) \nabla u^{n-1}_{h,i} + L^2 \alpha_2 \nabla\left(u^n_{h,i} - u^{n-1}_{h,i}\right)\right) \cdot \boldsymbol{n}_K \left(\widetilde{e}^n_{h,i} - v_{h,i}\right) \,\mathrm{d}s \\
&\qquad = \frac{1}{2} \sum_{K \in \mathcal{T}_{h,i}} \int_{\partial K \setminus \Gamma} [\![\mu\left(|\nabla u^{n-1}_{h,i}|\right) \nabla u^{n-1}_{h,i} + L^2 \alpha_2 \nabla\left(u^n_{h,i} - u^{n-1}_{h,i}\right)]\!] \left(\widetilde{e}^n_{h,i} - v_{h,i}\right) \,\mathrm{d}s,
\end{aligned}
$$

and thus, using the Cauchy–Schwarz inequality, implies

$$
\begin{aligned}
L^2 |||\widetilde{e}^n_{h,i}|||^2_{\Omega} \le &\sum_{K \in \mathcal{T}_{h,i}} \gamma_K^{1/2} \left\| f\left(u^{n-1}_{h,i}\right) - \nabla \cdot \left(\mu\left(|\nabla u^{n-1}_{h,i}|\right) \nabla u^{n-1}_{h,i}\right) + L^2 \mathsf{F}\left(u^n_{h,i} - u^{n-1}_{h,i}\right) \right\|_{L_2(K)} \gamma_K^{-1/2} \|\widetilde{e}^n_{h,i} - v_{h,i}\|_{L_2(K)} \\
&+ \frac{1}{2} \sum_{K \in \mathcal{T}_{h,i}} \alpha_2^{-1/4} \gamma_K^{1/4} \left\| [\![\mu\left(|\nabla u^{n-1}_{h,i}|\right) \nabla u^{n-1}_{h,i} + L^2 \alpha_2 \nabla\left(u^n_{h,i} - u^{n-1}_{h,i}\right)]\!] \right\|_{L_2(\partial K \setminus \Gamma)} \\
&\times \alpha_2^{1/4} \gamma_K^{-1/4} \|\widetilde{e}^n_{h,i} - v_{h,i}\|_{L_2(\partial K \setminus \Gamma)} \\
\le &\left(\sum_{K \in \mathcal{T}_{h,i}} \eta_K^2\right)^{1/2} \left(\sum_{K \in \mathcal{T}_{h,i}} \left(\gamma_K^{-1} \|\widetilde{e}^n_{h,i} - v_{h,i}\|^2_{L_2(K)} + \frac{\alpha_2^{1/2}}{2\gamma_K^{1/2}} \|\widetilde{e}^n_{h,i} - v_{h,i}\|^2_{L_2(\partial K \setminus \Gamma)}\right)\right)^{1/2} .
\end{aligned}
$$

Therefore, we infer that, employing Lemma 3.8,

$$
L^2 |||\widetilde{e}^n_{h,i}|||^2_{\Omega} \le C_I \left(\sum_{K \in \mathcal{T}_{h,i}} \eta_K^2\right)^{1/2} |||\widetilde{e}^n_{h,i}|||_{\Omega},
$$

which implies

$$
|||\widetilde{e}^n_{h,i}|||_{\Omega} \le C_I L^{-2} \left(\sum_{K \in \mathcal{T}_{h,i}} \eta_K^2\right)^{1/2} =: \eta(u^n_{h,i}, V_{\mathsf{FEM},i}).
$$

Inserting this bound into (2.4) with $c_0 = 1$ (cp. Proposition 3.1) completes the proof. □

3.5. Adaptive refinement algorithm

Proceeding along the lines of Section 2.4, we notice that the *a posteriori* error bound from Theorem 3.9 controls the error in terms of two contributions: The *finite element error*, defined as

$$\mathcal{E}^n_{\mathsf{FEM},i} = \left(\sum_{K \in \mathcal{T}_{h,i}} \eta_K^2 \right)^{1/2},$$

and the *fixed point error*

$$\mathcal{E}^n_{\mathsf{FP},i} = L(1+L) |\!|\!| u^n_{h,i} - u^{n-1}_{h,i} |\!|\!|_\Omega .$$

This allows us to write the error bound as

$$|\!|\!| u - u^n_{h,i} |\!|\!|_\Omega \leq C_I \mathcal{E}^n_{\mathsf{FEM},i} + \mathcal{E}^n_{\mathsf{FP},i}.$$

Based on this bound we can cast the abstract adaptive Algorithm 2.3 into the fixed point Galerkin iteration (3.14) for the solution of (3.1)–(3.2).

Algorithm 3.10. Choose an initial starting mesh $\mathcal{T}_{h,0}$, and an initial guess $u^0_{h,0} \in V_{\mathsf{FEM},0}$ in the associated finite element space $V_{\mathsf{FEM},0}$ (of fixed polynomial degree $p \geq 1$).

```
for $i \leftarrow 0, 1, 2, \dots$ do
    $n \leftarrow 0$
    repeat
        $n \leftarrow n + 1$
        Perform a single fixed point iteration (3.14) to calculate $u^n_{h,i}$.
    until $\mathcal{E}^n_{\mathsf{FP},i} \leq \vartheta \mathcal{E}^n_{\mathsf{FEM},i}$
    Perform mesh refinement (and/or derefinement) on $\mathcal{T}_{h,i}$
      based on the error indicators $\eta_K$ from Theorem 3.9
      together with a suitable marking strategy in order to obtain $\mathcal{T}_{h,i+1}$.
    $u^0_{h,i+1} \leftarrow \pi_{i,i+1} u^n_{h,i}$
end for
```

Here, $\pi_{i,i+1}$ is some projection from $V_{\mathsf{FEM},i}$ to $V_{\mathsf{FEM},i+1}$ (for instance, the $(.,.)_\Omega$-projection), and $\vartheta > 0$ is a (prescribed) parameter.

3.6. Numerical experiments

In this section we perform a series of numerical experiments to validate the *a priori* and *a posteriori* error bounds for the fixed point iteration (3.14) from Theorems 3.6 and 3.9, respectively. For the purpose of adaptive mesh refinements, the unspecified constant C_I from Lemma 3.8, which appears in (3.19), is set to $C_I = 1$ (evidently, in the context of error estimation a more accurate bound on C_I would be required). Furthermore, although numerical integration is not taken into account in our error analysis, we make use of Gauss quadrature to compute the nonlinear form $A(\cdot,\cdot)$ and inner product $(\cdot,\cdot)_\Omega$.

3.6.1. Validation of Remark 3.7

We consider the domain $\Omega = (0,1)^2 \subset \mathbb{R}^2$ with nonlinearity

$$\mu(|\nabla u|) = 2 + \frac{1}{1 + |\nabla u|^2},$$

and select f independent of u such that the analytical solution to (3.1)–(3.2) is given by

$$u(x,y) = x(1-x)y(1-y)(1-2y)\mathrm{e}^{-20(2x-1)^2}.$$

We note that $\beta_1 = \beta_2 = 0$, $\alpha_1 = 3$ and $\alpha_2 = {}^{15}/_{8}$.

Firstly, we consider the case when the mesh is fixed as a 16 × 16 uniform square mesh of quadrilaterals with $\mathcal{Q}_p$ basis functions and perform uniform refinement of the polynomial degree p from an initial guess $u^0_{h,0} \equiv 0$. In this situation we restrict the number of iterations of the fixed point iteration to $C_q \cdot p$, for $C_q = 1, 2, 3$ and plot in Fig. 1(a) the error $|\!|\!| u - u_h |\!|\!|_\Omega$ against the polynomial degree p. For comparison, we also perform the same experiment continuing the fixed point iteration until the residual $A(u^n_h, v_h)$ is below a given tolerance (10^{-14}) and, hence, the approximation is close to the best possible FEM approximation for the mesh. We clearly observe that by restricting the number of iterations we obtain exponential convergence of the error, and when $C_q = 3$ we gain the same convergence rate as allowing the iteration to continue until a tolerance is reached. Hence only performing the iteration $3p$ times gives an optimal convergence rate in the given example.

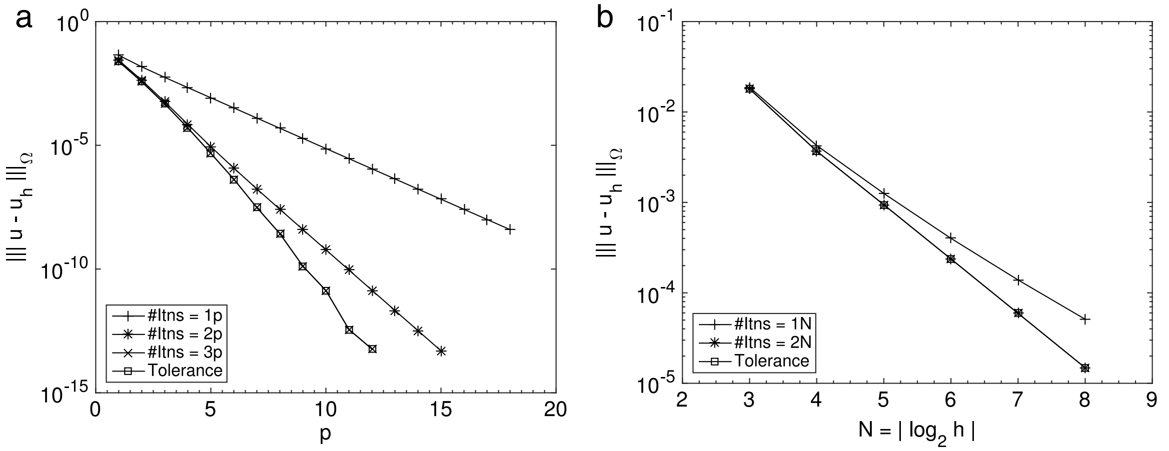

Fig. 1. Convergence with restricted number of fixed point iterations with uniform (a) p-refinement and (b) h-refinement.

Secondly, we consider fixed degree $\mathcal{Q}_p$ basis functions with $p = 2$ and perform h refinement to generate a sequence of $2^N \times 2^N$ uniform square meshes of quadrilaterals, for $N = 3, \ldots, 8$. We again perform both a restriction of the number of iterations of the fixed point method to $C_N \cdot N$, for $C_N = 1, 2$, as well as allowing the iteration to continue until a tolerance is reached, and plot in Fig. 1(b) the error $|\!|\!| u - u_h |\!|\!|_\Omega$ against $|\log_2 h|$. We obtain algebraic convergence, and already when $C_N = 2$ we achieve the optimal convergence rate $\mathcal{O}(h^2)$.

3.6.2. Validation of Theorem 3.9 and Remark 3.5

We now consider automatic h-adaptive mesh refinement, with linear ($p = 1$) $\mathcal{P}_p$ basis functions, using Algorithm 3.10 and the *a posteriori* error bound from Theorem 3.9. For the purpose of mesh refinement we use a fixed fraction refinement strategy, where the 25% of elements with the largest local error indicators η_K are marked for refinement, and the 5% of elements with the smallest local error indicators are marked for derefinement.

Example 1 (*Nonlinear Diffusion*). We first consider the case of a u-independent f with a nonlinear μ on the unit square $\Omega = (0, 1)^2 \subset \mathbb{R}^2$. To this end, we let

$$\mu(|\nabla u|) = 1 + \arctan(|\nabla u|^2),$$

and select f such that the analytical solution to (3.1)–(3.2) is given by

$$u(x, y) = x(1 - x)y(1 - y)(1 - 2y)\mathrm{e}^{-20(2x-1)^2}.$$

We note that $\beta_1 = \beta_2 = 0$, $\alpha_1 = 1 + \sqrt{3}/2 + \pi/3$ and $\alpha_2 = 1$. For this problem we set the steering parameter $\vartheta = 1/2$ in Algorithm 3.10.

We first plot, in Fig. 2(a), the relative true error $|\!|\!| u - u_{h,i}^{n^\star} |\!|\!|_\Omega / |\!|\!| u |\!|\!|_\Omega$ and the error bound, with $C_I = 1$, from Theorem 3.9 after the last iteration $n^\star$ on each mesh i against the number of degrees of freedom on that mesh. As can be seen, both the true error and the error bound converge at the same rate, and the error bound appears to overestimate the true error by a roughly constant amount. We also consider in Fig. 2(b) the effectivity index at each step of the fixed point iteration on each mesh, where the effectivity index is the error bound (calculated with $C_I = 1$) divided by the true error. As can be seen this is roughly constant (approximately 4) for all meshes and iterations, which indicates that the error bound overestimates the true error by roughly this amount, independent of mesh properties. We do note, however, that the effectivity rises slightly due to the fixed point iteration on each mesh, this is likely caused by setting $C_I = 1$. We also plot in Fig. 2(c) the number of fixed point iterations at each mesh step required to ensure that the fixed point error is less than the finite element error. We note this is fairly constant although minor variations exist. A few mesh refinements are made early on which is likely caused by the fact that the features of the solution are not accurately captured at the beginning.

In Fig. 2(d) we plot the mesh $\mathcal{T}_{h,7}$ after 7 h-adaptive mesh refinements. The areas of mesh refinement coincide with the hill and valley in the analytical solution, which is the location we would expect the greatest error to occur, and matches the sort of refinement that occurs when the nonlinear methods are computed to a minimal residual. This suggests that the mesh refinement algorithm behaves in the expected manner.

Example 2 (*Strong Nonlinear Reaction*). We now consider a fairly strong nonlinear f with a constant diffusion coefficient $\mu(|\nabla u|) = \varepsilon$, where ε is a small positive constant, on the unit square $\Omega = (0, 1)^2 \subset \mathbb{R}^2$. To this end, we consider $\varepsilon = 0.01$

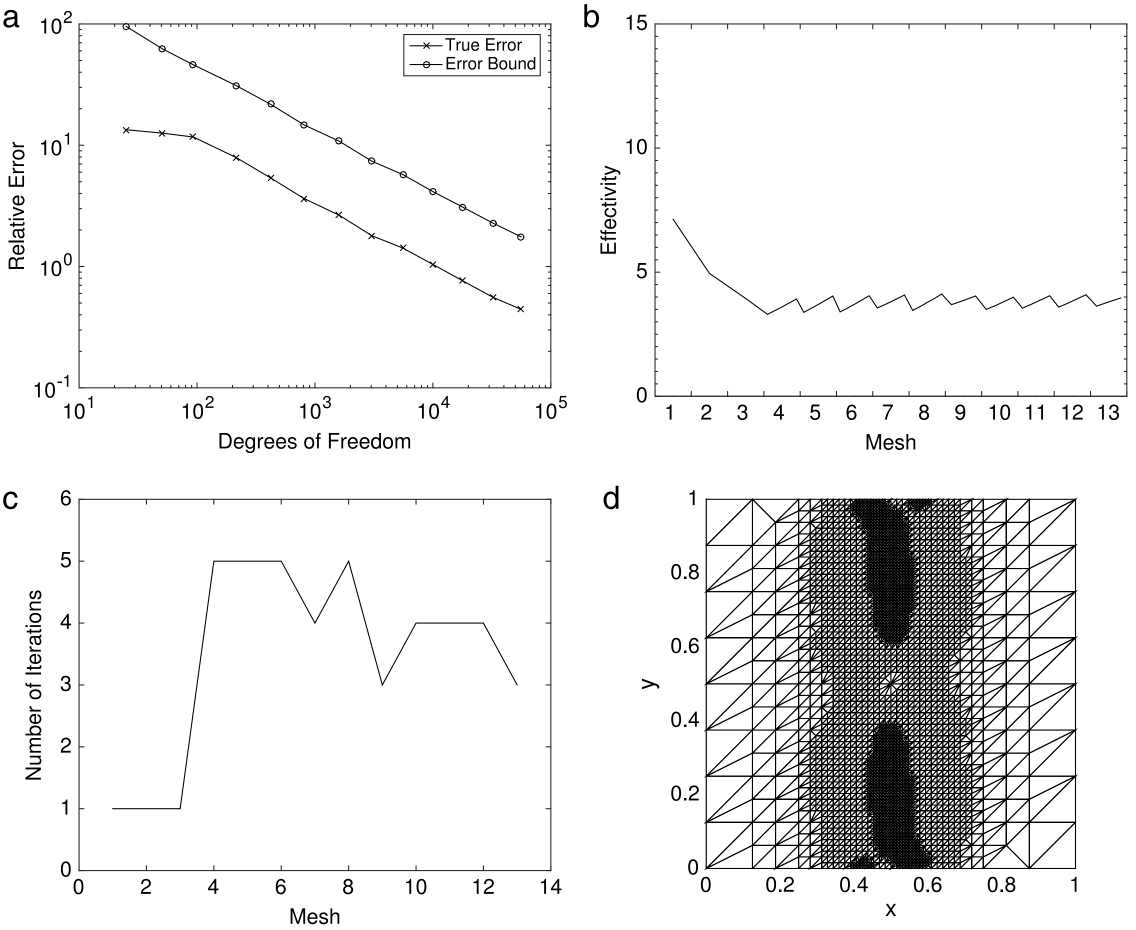

Fig. 2. Example 1. (a) Error in $|||\cdot|||_\Omega$-norm and error bound from Theorem 3.9 after the final fixed point iteration on each mesh compared to the number of degrees of freedom; (b) Effectivity at each fixed point iteration for all meshes; (c) Number of fixed point iterations on each mesh; (d) Mesh $\mathcal{T}_{h,7}$ after 7 h-refinements.

and let

$$f(u) = (0.2 + x^2 + y^2)\left(\frac{u^3}{u^2+1} + u\right) + c(x, y),$$

where $c(x, y)$ is a function dependent only on x and y selected such that the analytical solution to (3.1)–(3.2) is given by

$$u(x, y) = (1 - x)(1 - y)(e^{5x^2} - 1)(e^{5y^2} - 1). \tag{3.20}$$

We note that $\alpha_1 = \alpha_2 = \varepsilon$, $\beta_1 = {}^{187}/_{40}$, and $\beta_2 = {}^{1}/_{5}$. For this problem we set the steering parameter $\vartheta = 1$ in Algorithm 3.10.

We again plot, in Fig. 3(a), the relative true error $|||u-u_{h,i}^{n^\star}|||_\Omega / |||u|||_\Omega$ and error bound, with $C_I = 1$, from Theorem 3.9 after the last iteration $n^\star$ on each mesh i against the number of degrees of freedom on that mesh. Except for a few early meshes, where the mesh is unlikely to accurately capture the boundary layer in the analytical solution, both the true error and the error bound converge at the similar rate with the error bound overestimating the true error by a roughly constant amount. This is supported by the effectivity indices at each iteration, Fig. 3(b), which are roughly constant for all meshes and iterations, although slightly decreasing over the course of the mesh refinement. For this problem we note that the number of fixed point iterations at each mesh step is fairly high, caused by the stronger nonlinearity, but only once the boundary layer is captured accurately by the mesh (the 5th mesh onwards); cf. Fig. 3(c). Before this mesh the finite element error is considerably larger than the fixed point error due to the inaccurate capture of the boundary layer. The mesh after 7 h-adaptive mesh refinements, Fig. 3(d), demonstrates how the mesh captures the boundary layer. This demonstrates an important benefit of only refining the fixed point error while it is greater than the finite element error, as the algorithm has managed to reduce the number of iterations in the early meshes by a considerable number by not performing fixed point iterations until after the mesh has started to accurately capture the solution's features.

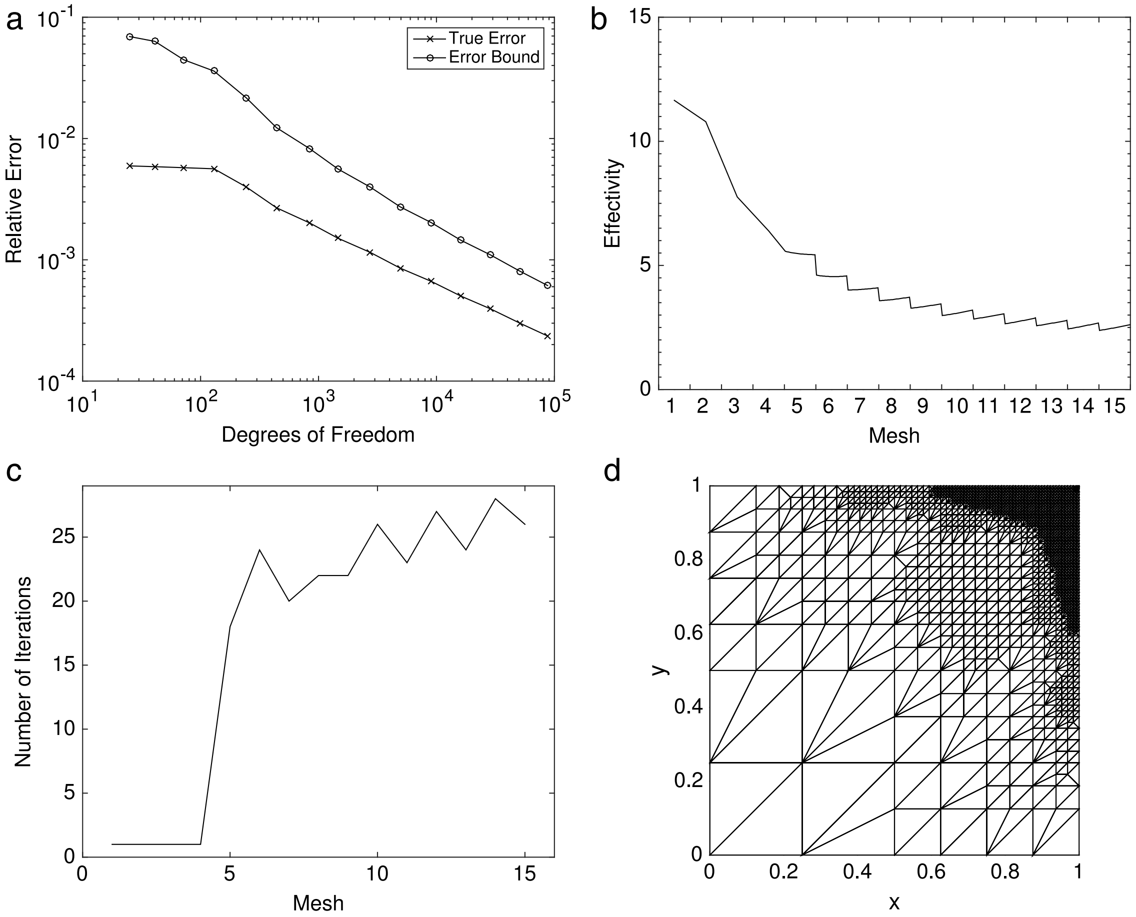

Fig. 3. Example 2. (a) Error in $|\!|\!| \cdot |\!|\!|_{\Omega}$-norm and error bound from Theorem 3.9 after the final fixed point iteration on each mesh compared to the number of degrees of freedom; (b) Effectivity at each fixed point iteration for all meshes; (c) Number of fixed point iterations on each mesh; (d) Mesh $\mathcal{T}_{h,7}$ after 7 h-refinements.

Example 3 (*Nonlinear Reaction*)**.** We now consider a weaker nonlinear f with a constant diffusion coefficient $\mu(|\nabla u|) = \varepsilon$, where ε is a small positive constant, such that $\beta_1 \approx \beta_2 = \mathcal{O}(1)$, on the unit square $\Omega = (0, 1)^2 \subset \mathbb{R}^2$. To this end, we consider $\varepsilon = 10^{-k}$, for $k = 0, \ldots, 6$, and let

$$f(u) = \frac{u^3}{10u^2 + 1} + u + c(x, y),$$

where $c(x, y)$ is a function dependent only on x and y selected such that the analytical solution to (3.1)–(3.2) is given by (3.20). We note that $\alpha_1 = \alpha_2 = \varepsilon$, $\beta_1 = 89/80$, and $\beta_2 = 1$. For this problem we set the steering parameter $\vartheta = 1$ in Algorithm 3.10.

We plot, in Fig. 4(a), the true error $|\!|\!| u - u_{h,i}^{n^\star} |\!|\!|_{\Omega}$ for $\varepsilon = 10^{-k}$, where $k = 0, \ldots, 6$, on each mesh i against the number of degrees of freedom on that mesh. We note that we appear to achieve a higher initial rate of convergence for smaller ε values, although they all appear to tend to similar convergence rates as refinement progresses. For each ε we also calculate the effectivity indices at each iteration, Fig. 4(b). We note that initially they are highly oscillatory for small values of ε, but as refinements progress they tend to smoothen towards a constant value, with the high values of ε converging at earlier mesh steps; in particular, the effectivity indices do not deteriorate as $\varepsilon \to 0^+$. We also plot in Fig. 4(c) the number of fixed point iterations at each mesh step required to ensure that the fixed point error is less than the finite element error. We note this is fairly constant and independent of the value of ε, which supports Remark 3.5.

4. Conclusion

In this article we have shown that it is possible, within a Galerkin framework, to use a simple fixed point iteration to solve strongly monotone problems requiring only the computation of the iteration matrix, opposed to a Newton's method requiring computation at each iteration; while we have only focused on finite element approximations of quasilinear PDEs

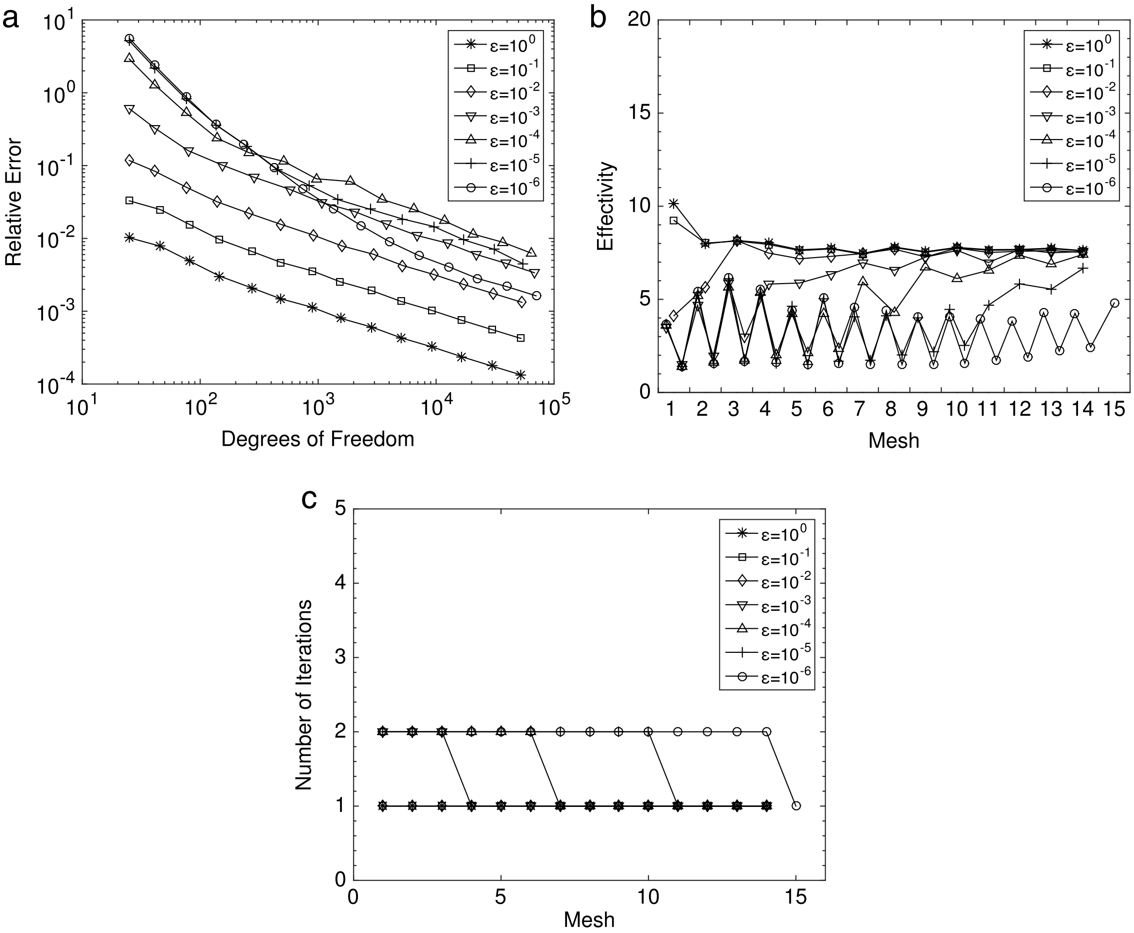

Fig. 4. Example 3. (a) Error in $|||\cdot|||_{\Omega}$-norm after the final fixed point iteration on each mesh compared to the number of degrees of freedom for various values of ε; (b) Effectivity at each fixed point iteration for all meshes; (c) Number of fixed point iterations on each mesh.

with (homogeneous) Dirichlet boundary conditions, our approach carries over to more general types of boundary conditions and differential equations. We have shown that an optimal *a priori* convergence rate can be obtained for a fixed number of iterations, dependent on the mesh size or polynomial degree. Moreover, we have demonstrated that it is possible to perform adaptive mesh refinement based on an *a posteriori* error analysis, where only a minimal number of fixed point iterations are required on each mesh to obtain a good approximation to the solution, without continuing the fixed point iteration until the fixed point error is insignificant.

Acknowledgment

Both authors acknowledge the support of the Swiss National Science Foundation (SNF) (Grant 200021_162990).

References

[1] E. Zeidler, Applied Functional Analysis: Applications to Mathematical Physics, in: Applied Mathematical Sciences, vol. 108, Springer-Verlag, New York, 1995.
[2] J. Nečas, Introduction to the Theory of Nonlinear Elliptic Equations, John Wiley and Sons, 1986.
[3] E.M. Garau, P. Morin, C. Zuppa, Convergence of an adaptive Kačanov FEM for quasi-linear problems, Appl. Numer. Math. 61 (4) (2011) 512–529. http://dx.doi.org/10.1016/j.apnum.2010.12.001.
[4] A. Chaillou, M. Suri, A posteriori estimation of the linearization error for strongly monotone nonlinear operators, J. Comput. Appl. Math. 205 (1) (2007) 72–87.
[5] L. El Alaoui, A. Ern, M. Vohralík, Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems, Comput. Methods Appl. Mech. Engrg. 200 (37–40) (2011) 2782–2795.
[6] M. Amrein, T.P. Wihler, An adaptive Newton-method based on a dynamical systems approach, Commun. Nonlinear Sci. 19 (9) (2014) 2958–2973.
[7] M. Amrein, T.P. Wihler, Fully adaptive Newton-Galerkin methods for semilinear elliptic partial differential equations, SIAM J. Sci. Comput. 37 (4) (2015) A1637–A1657.
[8] A. Chaillou, M. Suri, Computable error estimators for the approximation of nonlinear problems by linearized models, Comput. Methods Appl. Mech. Engrg. 196 (1–3) (2006) 210–224.

[9] W. Han, A posteriori error analysis for linearization of nonlinear elliptic problems and their discretizations, Math. Methods Appl. Sci. 17 (7) (1994) 487–508.

[10] C. Makridakis, R.H. Nochetto, Elliptic reconstruction and a posteriori error estimates for parabolic problems, SIAM J. Numer. Anal. 41 (4) (2003) 1585–1594.

[11] W.B. Liu, J.W. Barrett, Quasi-norm error bounds for the finite element approximation of some degenerate quasilinear elliptic equations and variational inequalities, RAIRO Modél. Math. Anal. Numér. 28 (6) (1994) 725–744.

[12] L. Boulton, Spectral pollution and eigenvalue bounds, Appl. Numer. Math. 99 (2016) 1–23.

[13] I. Šebestová, T. Vejchodský, Two-sided bounds for eigenvalues of differential operators with applications to friedrichs, poincaré, trace, and similar constants, SIAM J. Numer. Anal. 52 (1) (2014) 308–329.

[14] I. Babuška, M. Suri, The *hp*–version of the finite element method with quasiuniform meshes, RAIRO Anal. Numér. 21 (1987) 199–238.

[15] C. Schwab, *p*- and *hp*-FEM — Theory and Applications in Solid and Fluid Mechanics, in: Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 1998.

[16] P. Clément, Approximation by finite element functions using local regularization, RAIRO Anal. Numér. 9 (R-2) (1975) 77–84.

[17] R. Verfürth, Robust a posteriori error estimators for a singularly perturbed reaction–diffusion equation, Numer. Math. 78 (3) (1998) 479–493.