ADAPTIVE ITERATIVE LINEARIZATION GALERKIN METHODS FOR NONLINEAR PROBLEMS

PASCAL HEID AND THOMAS P. WIHLER

Abstract. A wide variety of (fixed-point) iterative methods for the solution of nonlinear equations (in Hilbert spaces) exists. In many cases, such schemes can be interpreted as iterative *local linearization* methods, which, as will be shown, can be obtained by applying a suitable preconditioning operator to the original (nonlinear) equation. Based on this observation, we will derive a unified abstract framework which recovers some prominent iterative schemes. In particular, for Lipschitz continuous and strongly monotone operators, we derive a general convergence analysis. Furthermore, in the context of numerical solution schemes for nonlinear partial differential equations, we propose a combination of the iterative linearization approach and the classical Galerkin discretization method, thereby giving rise to the so-called *iterative linearization Galerkin (ILG)* methodology. Moreover, still on an abstract level, based on two different elliptic reconstruction techniques, we derive a posteriori error estimates which separately take into account the discretization and linearization errors. Furthermore, we propose an adaptive algorithm, which provides an efficient interplay between these two effects. In addition, the ILG approach will be applied to the specific context of finite element discretizations of quasilinear elliptic equations, and some numerical experiments will be performed.

1. Introduction

The aim of this paper is to establish a general (adaptive) *iterative linearization Galerkin (ILG)* framework for the numerical solution of nonlinear problems, with application to second-order partial differential equations (PDE) in divergence form. To set the stage, we consider a real Hilbert space X with inner product $(\cdot,\cdot)_X$ and induced norm denoted by $\|\cdot\|_X$. We remark that for most of our work it is sufficient for X to be a reflexive Banach space. Then, given a nonlinear operator $\mathsf{F}: X \to X'$, we focus on the equation

$$u \in X: \qquad \mathsf{F}(u) = 0 \quad \text{in } X', \tag{1}$$

where X' denotes the dual space of X. In weak form, this problem reads

$$u \in X: \qquad \langle \mathsf{F}(u), v\rangle = 0 \quad \text{for all } v \in X, \tag{2}$$

with $\langle\cdot,\cdot\rangle$ signifying the duality pairing in $X' \times X$.

Iterative linearization. The development of an *iterative linearization scheme* for (1) is based on applying suitable preconditioning operators. More precisely, for given $v \in X$, we introduce a linear and invertible *preconditioning operator*

$$\mathsf{A}(v): X \to X', \tag{3}$$

which allows to transform (1) into $\mathsf{A}(u)^{-1}\mathsf{F}(u) = 0$. This in turn gives rise to a fixed point iteration

$$u^{n+1} = u^n - \mathsf{A}(u^n)^{-1}\mathsf{F}(u^n), \qquad n \geq 0,$$

Mathematics Institute, University of Bern, CH-3012 Switzerland
E-mail address: pascal.heid@math.unibe.ch and wihler@math.unibe.ch.

2010 *Mathematics Subject Classification.* 35J62, 47J25, 47H05, 47H10, 49M15, 65J15, 65N30.
Key words and phrases. Numerical solution methods for nonlinear PDE, monotone problems, fixed point iterations, linearization schemes, Kačanov method, Newton method, Galerkin discretizations, adaptive finite element methods, a posteriori error estimation.
The authors acknowledge the financial support of the Swiss National Science Foundation under grant no. 200021_182524.

for an initial guess $u^0 \in X$, or equivalently,

$$u^{n+1} \in X : \qquad \mathsf{A}(u^n)u^{n+1} = \mathsf{A}(u^n)u^n - \mathsf{F}(u^n), \qquad n \geq 0. \tag{4}$$

Letting

$$f : X \to X', \qquad f(u) := \mathsf{A}(u)u - \mathsf{F}(u), \tag{5}$$

the fixed-point iteration (4) takes the form of the following *iterative linearization scheme*:

$$\mathsf{A}(u^n)u^{n+1} = f(u^n), \qquad n \geq 0. \tag{6}$$

We emphasize that, given $u^n \in X$, this is a *linear* problem for $u^{n+1} \in X$.

The general iteration scheme (6) recovers some of the widely used fixed-point iterations occurring in the literature. These include, for instance, the Zarantonello iteration, the Kačanov scheme, and the Newton method; see Section 2.3 for a detailed discussion. In the context of the Zarantonello iteration, the interested reader is referred to the original work [30] (cf. also [9] for a generalization), or the monographs [25, §3.3] and [33, §25.4]. Incidentally, the latter two references also deal with the Kačanov approach, see [25, §4.5] or [33, §25.14]. For the (damped and adaptive) Newton method we refer to [12] for an extensive overview, or the recent works on adaptive Newton schemes [3, 4, 21, 26, 28].

Iterative linearized Galerkin approach. The iteration (6) generates a sequence $\{u^n\}_{n\geq 0}$ which potentially converges to a solution $u^\star \in X$ of (1). In general, however, the computation of this sequence is not feasible if X is infinite- or high-dimensional. Therefore, in order to cast the iterative linearization approach described above into a computational framework, we will consider Galerkin discretizations of (6) in terms of finite-dimensional conforming subspaces $X_N \subset X$. Then, a discrete approximation, $u_N^{n+1} \in X_N$, based on a starting guess $u_N^0 \in X_N$, is obtained by solving the *linear* discrete system

$$u_N^{n+1} \in X_N : \qquad \big\langle \mathsf{A}(u_N^n)u_N^{n+1}, v\big\rangle = \langle f(u_N^n), v\rangle \quad \forall v \in X_N, \qquad n \geq 0. \tag{7}$$

We note that the discretization of the linearized problem (6) coincides with the linearization of the discretized problem (33), i.e. the discretization and linearization commute; see [14] for a related discussion. For the resulting sequence $\{u_N^n\}_{n\geq 0} \subset X_N$ of discrete solutions it is possible, under certain conditions, to obtain general a posteriori estimates for the difference to the exact solution, $u^\star \in X$, i.e. for $\|u^\star - u_N^{n+1}\|_X$, $n \geq 0$. The emphasis of such bounds is that they enable the individual identification of different sources of error in the approximation process, such as, e.g., the linearization and discretization errors (further errors, not to be considered here, may result, for instance, from a linear solver iteration, see, e.g., [15], or from quadrature). This can be accomplished by means of two conceptionally different techniques, both of which will be presented in this work:

(a) The first approach is based on the assumption that a computable bound for the residual of the *linear* Galerkin discretization of the form (7) is available. Then, applying an elliptic reconstruction technique (see, e.g., [22, 23]) yields a computable a posteriori error estimate for the error $\|u^\star - u_N^{n+1}\|_X$, which can be expressed in terms of a discretization and linearization contribution. In fact, these estimators can also be applied to appropriately enrich the space X_N, thereby leading to a new space X_{N+1}. We note that this approach has been applied previously in [11] in the specific context of the Zarantonello iteration scheme.

(b) Alternatively, we may consider, for $n \geq 0$, a *nonlinear* discrete problem which, on the one hand, features the *nonlinear* operator F from (1), and, on the other hand, possesses the same solution, $u_N^{n+1} \in X_N$, as the *linear* Galerkin formulation (7). Assuming that there exists a computable bound for the residual of the discrete solution to a suitably reconstructed nonlinear problem, our analysis will show that such a bound can be exploited for the purpose of deriving an a posteriori error estimator.

A posteriori error estimates as outlined above constitute an essential building block in the development of adaptive ILG schemes for nonlinear problems (1). Indeed, recalling that such bounds allow to distinguish the different sources of error in the approximation process, the key idea of the fully adaptive ILG methodology is to provide an appropriate *interplay* between the

fixed-point linearization iteration and possible Galerkin space enrichments (e.g., mesh refinements for finite elements) depending on whether the discretization error or the linearization error is dominant. In this way, the goal of the adaptive ILG approach is 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 simultaneous control of different sources of error in the context of adaptive finite element methods for monotone problems has been presented in a number of earlier papers. For instance, in the work [10], the authors have considered general linearizations of strongly monotone operators, and have derived computable a posteriori estimators for the total error (consisting of the linearization error and the Galerkin error) with identifiable components for each of the error sources. For even more sophisticated a posteriori error estimators in the specific context of the Newton linearization scheme for second-order monotone quasilinear diffusion problems we refer to [14, 15]. The a posteriori error analysis derived in those papers includes—in addition to the discretization and linearization errors—also the algebraic linear solver error; moreover, the authors have proposed an adaptive iterative procedure, which takes into account all components of the numerical scheme in each refinement step. For a further development of that research in the context of compositional two-phase flow with nonlinear complementarity constraints we refer to [7]. Furthermore, first a posteriori error estimates in the framework of the Kačanov iteration for quasilinear diffusion problems in divergence form have been presented in [19]. Later on, an adaptive iterative linearized Galerkin type approach has been introduced and discussed in [18]; indeed, the convergence of the Kačanov-Galerkin iteration is proved therein. Moreover, for semilinear second-order elliptic problems, two different linearization schemes of Kačanov type have been analyzed in [8]. Just recently, based on the ILG approach in [11], the convergence of an adaptive Zarantonello-Galerkin iterative scheme for monotone elliptic PDE has been proved in [17]. Finally, we point to the fact that the ILG methodology has been applied also to high-order (so-called *hp*) [2] and discontinuous Galerkin [21] finite element discretizations, as well as to nonlinear parabolic problems [5].

Outline of the paper. In Section 2 we state and prove a global convergence result for the unified iteration scheme (6). In particular, in order to provide a few examples, we apply our result to the Zarantonello, Kačanov, and (damped) Newton methods, thereby recovering some of the well-known convergence results from the literature. Furthermore, still on an abstract level, in Section 3 we discuss conforming Galerkin discretizations of (6), and present general a posteriori error estimates based on the two approaches outlined in (a) and (b) above. On that account, we propose in Section 4 a fully adaptive algorithm based on the a posteriori error estimates. More specifically, in Section 5, we derive computable error bounds for a second-order PDE in divergence form; finally, in Section 5.3, these theoretical estimates are employed within a series of numerical experiments in the framework of the fully adaptive ILG approach.

2. Iterative linearization

The goal of this section is to prove a general convergence result for the iterative linearization iteration (6) under the condition that F in (1) is a Lipschitz continuous and strongly monotone operator. Furthermore, we will review a few classical examples.

2.1. **Abstract framework.**

For the purpose of this work, we restrict ourselves to Lipschitz continuous, strongly monotone operators F:

(F1) The operator F is Lipschitz continuous, i.e. there exists a constant $L_{\mathsf{F}} > 0$ such that

$$|\langle \mathsf{F}(u) - \mathsf{F}(v), w\rangle| \leq L_{\mathsf{F}} \|u - v\|_X \|w\|_X ,$$

for all $u, v, w \in X$.

(F2) The operator F is strongly monotone, i.e. there exists a constant $\nu > 0$ such that

$$\nu \|u - v\|_X^2 \leq \langle \mathsf{F}(u) - \mathsf{F}(v), u - v\rangle ,$$

for all $u, v \in X$.

Under these conditions, the theory of strongly monotone operators implies that (1) possesses a unique solution $u^\star \in X$; see, e.g., [25, §3.3] or [33, §25.4].

Furthermore, for given $u \in X$, we introduce the bilinear form

$$a(u; v, w) := \langle \mathsf{A}(u)v, w\rangle\,, \qquad v, w \in X, \tag{8}$$

where $\mathsf{A}(\cdot)$ is the preconditioning operator from (3). Then, we can write (6) in weak form: given $u^n \in X$, find $u^{n+1} \in X$ such that

$$a(u^n; u^{n+1}, w) = \langle f(u^n), w\rangle \qquad \forall w \in X. \tag{9}$$

Throughout this paper, for any $u \in X$, we assume that the bilinear form $a(u; \cdot, \cdot)$ is uniformly coercive and bounded. Those assumptions refer to the fact that there are two constants $\alpha, \beta > 0$ independent of $u \in X$, such that

$$a(u; v, v) \geq \alpha \|v\|_X^2 \qquad \forall v \in X, \tag{10}$$

and

$$a(u; v, w) \leq \beta \|v\|_X \|w\|_X \qquad \forall v, w \in X, \tag{11}$$

respectively. In particular, owing to the Lax-Milgram Theorem, these properties imply the well-posedness of the solution $u^{n+1} \in X$ of the linear equation (6), for any given $u^n \in X$.

2.2. **A global convergence result.** Given the framework introduced in the previous Section 2.1, the ensuing proposition is an abstract global convergence result for the iteration scheme (6). We note that it can be extended readily to the case where X is a reflexive Banach space.

Proposition 2.1. *Suppose that* (F2) *(cf. Section 2.1),* (10) *and* (11) *are satisfied, and* $u \mapsto a(u; u, \cdot)$ *and* $u \mapsto \mathsf{F}(u)$ *are continuous mappings from* X *into its dual space* X' *with respect to the weak topology on* X'*. If the sequence* $\{u^n\}_{n\geq 0}$ *defined by* (6) *satisfies* $\|u^{n+1} - u^n\|_X \to 0$ *as* $n \to \infty$*, then it converges to the unique solution* $u^\star \in X$ *of* (1)*.*

Proof. We begin by showing that $\{u^n\}_{n\geq 0}$ is a Cauchy sequence. Indeed, by virtue of (F2) and (5), for any $m \geq n \geq 0$, it holds that

$$\begin{aligned}\nu \|u^m - u^n\|_X^2 &\leq \langle \mathsf{F}(u^m) - \mathsf{F}(u^n), u^m - u^n\rangle \\ &= \langle \mathsf{A}(u^m)u^m - f(u^m), u^m - u^n\rangle - \langle \mathsf{A}(u^n)u^n - f(u^n), u^m - u^n\rangle\,.\end{aligned}$$

Hence, involving (8) and (9) gives

$$\nu \|u^m - u^n\|_X^2 \leq a(u^m; u^m - u^{m+1}, u^m - u^n) - a(u^n; u^n - u^{n+1}, u^m - u^n).$$

Furthermore, (11) implies that

$$\|u^m - u^n\|_X \leq \frac{\beta}{\nu}\left(\left\|u^{m+1} - u^m\right\|_X + \left\|u^{n+1} - u^n\right\|_X\right) \to 0,$$

for $n, m \to \infty$. Hence, $\{u^n\}_{n\geq 0}$ is a Cauchy sequence, and, therefore, converges to some limit $u^\star \in X$. Next, we show that $u^\star$ is the unique solution of (1). Owing to (9), we notice the identity

$$a(u^n; u^n, v) - \langle f(u^n), v\rangle + a(u^n; u^{n+1} - u^n, v) = 0 \qquad \forall v \in X,$$

for all $n \geq 0$. Here, due to (11), and because $\left\|u^{n+1} - u^n\right\|_X$ is a vanishing sequence, we observe that

$$a(u^n; u^n, v) - \langle f(u^n), v\rangle \to 0 \qquad \forall v \in X,$$

for $n \to \infty$. Hence, by continuity of a and f, we deduce that

$$a(u^\star; u^\star, v) = \langle f(u^\star), v\rangle \qquad \forall v \in X,$$

i.e. $u^\star$ is a solution of (1); we note that $u \mapsto f(u) = a(u; u, \cdot) - \mathsf{F}(u)$ is continuous by the continuity of a and F. It remains to show that $u^\star$ is the only solution of (1). In fact, if $u^\square \in X$ is any other solution, then (F2) leads to

$$\nu \left\|u^\star - u^\square\right\|_X^2 \leq \left\langle \mathsf{F}(u^\star) - \mathsf{F}(u^\square), u^\star - u^\square\right\rangle = 0,$$

i.e. $u^\star = u^\square$. □

2.3. **Applications.** In the ensuing section we will discuss the general Proposition 2.1 in the context of the Zarantonello, Kačanov, and Newton iterations.

2.3.1. *Zarantonello iteration.* A most simple choice for the preconditioning operator from (3) is $\mathsf{A}(v)u := (\delta^{-1}u, \cdot)_X$, where $\delta > 0$ is a fixed constant; in particular, here, $\mathsf{A} = \mathsf{A}(v)$ is independent of v. In this case, the iterative linearization scheme (6) turns out to be

$$(u^{n+1}, \cdot)_X = (u^n, \cdot)_X - \delta \left\langle \mathsf{F}(u^n), \cdot \right\rangle . \tag{12}$$

Theorem 2.2 (Convergence of the Zarantonello iteration)**.** *Assuming* (F1) *and* (F2) *(cf. Section 2.1), the Zarantonello iteration* (12) *converges to the unique solution $u^\star$ of* (1) *for any $\delta \in \left]0, {}^{2\nu}/_{L_{\mathsf{F}}^2}\right[$.*

Proof. We verify the assumptions required for Proposition 2.1 to hold. For $a(u, v) = (\delta^{-1}u, v)_X$, $u, v \in X$, we note that (10) and (11) are satisfied with

$$\alpha = \beta = \delta^{-1} > 0. \tag{13}$$

Moreover, both $u \mapsto a(u, \cdot) = (\delta^{-1}u, \cdot)_X$ and $u \mapsto \mathsf{F}(u)$ are continuous on X. It remains to show that $\left\|u^{n+1} - u^n\right\|_X$ vanishes. For that purpose, we denote by $\mathsf{J} : X \to X'$ the Riesz-Fréchet isometry. The iteration (12) can then be written, in strong form, as $u^{n+1} = \mathsf{T}(u^n)$, where $\mathsf{T}(u) := u - \delta \mathsf{J}^{-1}\mathsf{F}(u)$. This leads to

$$\begin{aligned} \left\|u^{n+1} - u^n\right\|_X^2 &= \left\|\mathsf{T}(u^n) - \mathsf{T}(u^{n-1})\right\|_X^2 \\ &= \left\|u^n - u^{n-1}\right\|_X^2 - 2\delta \left\langle \mathsf{F}(u^n) - \mathsf{F}(u^{n-1}), u^n - u^{n-1} \right\rangle \\ &\quad + \delta^2 \left\|\mathsf{J}^{-1}(\mathsf{F}(u^n) - \mathsf{F}(u^{n-1}))\right\|_X^2 , \end{aligned}$$

where we have used the linearity of J^{-1}. Invoking (F1) and (F2), together with the fact that J^{-1} is isometric, we further get

$$\left\|u^{n+1} - u^n\right\|_X^2 \leq \left(1 - 2\delta\nu + \delta^2 L_{\mathsf{F}}^2\right) \left\|u^n - u^{n-1}\right\|_X^2 .$$

We note that

$$\gamma := \left(1 - 2\delta\nu + \delta^2 L_{\mathsf{F}}^2\right) < 1 \tag{14}$$

if and only if $\delta \in \left]0, {}^{2\nu}/_{L_{\mathsf{F}}^2}\right[$. Hence, by induction,

$$\left\|u^{n+1} - u^n\right\|_X^2 \leq \gamma^n \left\|u^1 - u^0\right\|_X^2 ,$$

which shows that $\|u^{n+1} - u^n\|_X \to 0$ as $n \to 0$. □

Remark 2.3. We notice that the contraction factor γ from (14) is minimal for the choice $\delta = {}^{\nu}/_{L_{\mathsf{F}}^2}$.

2.3.2. *Kačanov iteration.* Here we assume that the nonlinear operator F from (1) takes the form $\mathsf{F}(u) = \mathsf{A}(u)u - g$, where $\mathsf{A}(u) : X \to X'$ is linear (for given $u \in X$), and $g = -\mathsf{F}(0) \in X'$ is fixed. Then, the Kačanov iteration is defined by

$$\mathsf{A}(u^n)u^{n+1} = g, \qquad n \geq 0. \tag{15}$$

Note that this iteration can be cast into the setting of (6), where $\mathsf{A}(u^n)$ takes the role of the preconditioning operator, and $f(u^n) = \mathsf{A}(u^n)u^n - \mathsf{F}(u^n) = g$ is constant. We make the assumption that there exists a Gateaux differentiable functional $\mathsf{G} : X \to \mathbb{R}$ which satisfies the following properties:

(K1) $\mathsf{G}'(u) = a(u; u, \cdot)$ on X, and G' is continuous and strongly monotone, i.e. there exists a real number $c_0 > 0$ such that, for any $u, v \in X$, it holds

$$\langle \mathsf{G}'(u) - \mathsf{G}'(v), u - v \rangle \geq c_0 \left\|u - v\right\|_X^2 ; \tag{16}$$

(K2) For each $u, v \in X$ we have the bound $\mathsf{G}(u) - \mathsf{G}(v) \geq {}^1/_2 \, (a(u; u, u) - a(u; v, v))$.

In order to be able to apply Proposition 2.1, we need an auxiliary result, which will also be crucial in the analysis of the Newton method in Section 2.3.3 below.

Lemma 2.4. *If $\mathsf{H} : X \to \mathbb{R}$ is Gateaux differentiable with H' continuous and strongly monotone, then H is bounded from below.*

Proof. For fixed $v \in X$, and $t \in [0,1]$, we define the function $\varphi(t) := \mathsf{H}(tv)$. We note that $\varphi'(t) = \langle \mathsf{H}'(tv), v\rangle$, and, invoking the fundamental theorem of calculus, we find that

$$\mathsf{H}(v) - \mathsf{H}(0) = \int_0^1 \langle \mathsf{H}'(tv), v\rangle \, \mathrm{d}t = \int_0^1 \langle \mathsf{H}'(tv) - \mathsf{H}'(0), v\rangle \, \mathrm{d}t + \langle \mathsf{H}'(0), v\rangle \,. \tag{17}$$

Since H' is strongly monotone, there exists a constant $\gamma > 0$ such that

$$\langle \mathsf{H}'(tv) - \mathsf{H}'(0), v\rangle = \frac{1}{t} \langle \mathsf{H}'(tv) - \mathsf{H}'(0), tv\rangle \geq \gamma t \, \|v\|_X^2 \,,$$

for any $t \in]0,1]$. Inserting this bound into (17), integrating with respect to t, and using the submultiplicativity of the operator norm, yields

$$\mathsf{H}(v) \geq \frac{\gamma}{2} \|v\|_X^2 - \|\mathsf{H}'(0)\|_{X'} \|v\|_X + \mathsf{H}(0).$$

It is elementary to verify that the right-hand side is minimal for $\|v\|_X = \gamma^{-1} \|\mathsf{H}'(0)\|_{X'}$. With this choice we arrive at $\mathsf{H}(v) \geq \mathsf{H}(0) - 1/2\gamma \|\mathsf{H}'(0)\|_{X'}^2$ for all $v \in X$, i.e. H is bounded from below. □

Theorem 2.5 (Convergence of the Kačanov iteration)**.** *Suppose that* (K1) *and* (K2) *hold. Furthermore, assume that the bilinear form $a(u;\cdot,\cdot)$ induced by A satisfies* (10) *and* (11)*, and is symmetric for all $u \in X$. Then the sequence $\{u^n\}_{n\geq 0}$ defined by* (15) *converges to the unique solution $u^\star$ of* (1)*.*

Proof. Because of (K1) it follows that $u \mapsto a(u;u,\cdot) = \mathsf{G}'(u)$ is continuous. Moreover, $u \mapsto f(u)$ is constant, and thus continuous. Consequently, $u \mapsto \mathsf{F}(u) = a(u;u,\cdot) - f(u)$ is continuous as well. We show that $\|u^{n+1} - u^n\|_X$, $n \geq 0$, is a vanishing sequence. To this end, we follow closely along the lines of the proof of [33, Theorem 25.L]. Let us introduce the functional $\mathsf{H}(u) := \mathsf{G}(u) - \langle g, u\rangle$. We note that $\mathsf{H}'(u) = \mathsf{G}'(u) - g = \mathsf{A}(u)u - g = \mathsf{F}(u)$, i.e. H is the potential of F. Moreover, by virtue of (K1), the derivative $\mathsf{H}' = \mathsf{G}' - g$ is continuous and strongly monotone, and thus F satisfies (F2). In particular, with the aid of Lemma 2.4, we deduce that H is bounded from below. Next, we will verify that $\{\mathsf{H}(u^n)\}_{n\geq 0}$ is a monotone decreasing sequence. Indeed, noticing that $a(u^n; u^{n+1}, u^{n+1} - u^n) = \langle g, u^{n+1} - u^n\rangle$, and employing (K2), yields

$$\begin{aligned}
\mathsf{H}(u^n) - \mathsf{H}(u^{n+1}) &= \langle g, u^{n+1} - u^n\rangle + \mathsf{G}(u^n) - \mathsf{G}(u^{n+1}) \\
&\geq a(u^n; u^{n+1}, u^{n+1} - u^n) + \frac{1}{2} a(u^n; u^n, u^n) - \frac{1}{2} a(u^n; u^{n+1}, u^{n+1}) \\
&\geq \frac{1}{2} a(u^n; u^n, u^n) - a(u^n; u^{n+1}, u^n) + \frac{1}{2} a(u^n; u^{n+1}, u^{n+1}),
\end{aligned}$$

for any $n \geq 0$. Then, employing the symmetry of $a(u^n;\cdot,\cdot)$, and involving (10), we obtain

$$\mathsf{H}(u^n) - \mathsf{H}(u^{n+1}) \geq \frac{1}{2} a(u^n; u^{n+1} - u^n, u^{n+1} - u^n) \geq \frac{\alpha}{2} \|u^{n+1} - u^n\|_X^2 \geq 0, \tag{18}$$

which shows that $\{\mathsf{H}(u^n)\}_{n\geq 0}$ is monotone decreasing. Then, recalling the boundedness from below, we conclude that $\mathsf{H}(u^n) - \mathsf{H}(u^{n+1}) \to 0$ as $n \to \infty$. Hence, exploiting (18), it follows that $\|u^{n+1} - u^n\|_X$ vanishes, and the proof is complete. □

2.3.3. *Newton iteration.* For the Newton iteration the preconditioning operator in (4) is selected to be $\mathsf{A}(v) = \delta(v)^{-1}\mathsf{F}'(v)$, $v \in X$, where $\delta(v) > 0$ is a (damping) parameter, and F' signifies the Gateaux derivative of F. Then, the (damped) Newton iteration is given by

$$\mathsf{F}'(u^n)u^{n+1} = \mathsf{F}'(u^n)u^n - \delta(u^n)\mathsf{F}(u^n), \qquad n \geq 0. \tag{19}$$

For the purpose of applying Proposition 2.1, we make the following assumptions:

(N1) The operator F is Gateaux differentiable. Moreover, F' is coercive and bounded in the sense that, for any given $u \in X$, it holds

$$\langle \mathsf{F}'(u)v, v\rangle \geq \alpha_{\mathsf{F}'} \|v\|_X^2 \qquad \forall v \in X, \tag{20}$$

and

$$\langle \mathsf{F}'(u)v, w\rangle \leq \beta_{\mathsf{F}'} \|v\|_X \|w\|_X \qquad \forall v, w \in X, \tag{21}$$

where $\alpha_{\mathsf{F}'}, \beta_{\mathsf{F}'} > 0$ are independent of u.

(N2) It exists a Gateaux differentiable functional $\mathsf{G} : X \to \mathbb{R}$ such that $\mathsf{G}'(u) = \mathsf{F}'(u)u$ in X' for any $u \in X$, and G' is continuous when X' is endowed with the weak topology.

(N3) It exists a Gateaux differentiable functional $\mathsf{H} : X \to \mathbb{R}$ such that $\mathsf{H}' = \mathsf{F}$.

(N4) There are some constants $0 < \delta_{\min} \leq \delta_{\max} < \infty$ such that $\delta : X \to [\delta_{\min}, \delta_{\max}]$ is a continuous functional.

Theorem 2.6 (Convergence of the damped Newton iteration)**.** *Assume* (F1) *and* (F2) *(cf. Section 2.1), as well as* (N1)–(N4)*. Then, for* $\delta_{\max} < {}^{2\alpha_{\mathsf{F}'}}/_{L_{\mathsf{F}}}$ *in* (N4) *the damped Newton iteration* (19) *converges to the unique solution* $u^\star \in X$ *of* (1) *.*

Proof. We aim at employing Proposition 2.1 as before. By virtue of (20), (21), and (N4), we obtain

$$a(u; v, v) \geq \alpha_{\mathsf{F}'} \delta_{\max}^{-1} \|v\|_X^2 , \qquad u, v \in X,$$

and

$$a(u; v, w) \leq \beta_{\mathsf{F}'} \delta_{\min}^{-1} \|v\|_X \|w\|_X , \qquad u, v, w \in X,$$

which are the coercivity and boundedness conditions (10) and (11), with

$$\alpha = {}^{\alpha_{\mathsf{F}'}}/_{\delta_{\max}}, \qquad \beta = {}^{\beta_{\mathsf{F}'}}/_{\delta_{\min}}, \tag{22}$$

respectively. Next, we remark that the maps $u \mapsto a(u; u, \cdot) = \delta(u)^{-1}\mathsf{F}'(u)u$ and $u \mapsto \mathsf{F}(u)$ are both continuous, when X' is endowed with the weak topology, by (N2) and (N4), and by (F1), respectively. Therefore, by the same arguments as in the proof of Theorem 2.5, it suffices to show that there exists a constant $C > 0$ such that

$$\mathsf{H}(u^n) - \mathsf{H}(u^{n+1}) \geq C \left\|u^{n+1} - u^n\right\|_X^2 , \qquad n \geq 0. \tag{23}$$

To this end, we define the function $\varphi(t) := \mathsf{H}(u^n + t(u^{n+1} - u^n))$, $t \in [0, 1]$, and observe that

$$\varphi'(t) = \left\langle \mathsf{H}'(u^n + t(u^{n+1} - u^n)), u^{n+1} - u^n \right\rangle = \left\langle \mathsf{F}(u^n + t(u^{n+1} - u^n)), u^{n+1} - u^n \right\rangle .$$

Then, the fundamental theorem of calculus implies that

$$\begin{aligned}
\mathsf{H}(u^n) - \mathsf{H}(u^{n+1}) &= -\int_0^1 \left\langle \mathsf{F}(u^n + t(u^{n+1} - u^n)), u^{n+1} - u^n \right\rangle \mathrm{d}t \\
&= -\int_0^1 \left\langle \mathsf{F}(u^n + t(u^{n+1} - u^n)) - \mathsf{F}(u^n), u^{n+1} - u^n \right\rangle \mathrm{d}t \\
&\quad - \left\langle \mathsf{F}(u^n), u^{n+1} - u^n \right\rangle .
\end{aligned}$$

By the definition of the Newton iteration (19), it holds that $\mathsf{F}(u^n) = \delta(u^n)^{-1}\mathsf{F}'(u^n)(u^n - u^{n+1})$, $n \geq 0$. Thus, with the aid of (F1) and (20), it follows that

$$\begin{aligned}
\mathsf{H}(u^n) - \mathsf{H}(u^{n+1}) &\geq -L_{\mathsf{F}} \int_0^1 t \left\|u^{n+1} - u^n\right\|_X^2 \mathrm{d}t + \delta(u^n)^{-1} \left\langle \mathsf{F}'(u^n)(u^{n+1} - u^n), u^{n+1} - u^n \right\rangle \\
&\geq -\frac{L_{\mathsf{F}}}{2} \left\|u^{n+1} - u^n\right\|_X^2 + \alpha_{\mathsf{F}'} \delta(u^n)^{-1} \left\|u^{n+1} - u^n\right\|_X^2 .
\end{aligned}$$

If $\delta(u^n) \leq \delta_{\max} < {}^{2\alpha_{\mathsf{F}'}}/_{L_{\mathsf{F}}}$, then

$$\frac{\alpha_{\mathsf{F}'}}{\delta(u^n)} - \frac{L_{\mathsf{F}}}{2} \geq \frac{\alpha_{\mathsf{F}'}}{\delta_{\max}} - \frac{L_{\mathsf{F}}}{2} =: C > 0, \qquad n \geq 0. \tag{24}$$

We conclude that (23) is satisfied. □

Remark 2.7 (Classical Newton scheme)**.** Recalling (20) with $u = u^0$, and applying [25, Theorem 3.3.23], we deduce the bound $\left\|\mathsf{F}'(u^0)^{-1}v\right\|_X \leq \alpha_{\mathsf{F}'}^{-1} \|v\|_{X'}$ for all $v \in X'$. Furthermore, assume that F is Fréchet differentiable and F' is Lipschitz continuous, i.e. there exists a constant $L_{\mathsf{F}'} > 0$ such that

$$\|\mathsf{F}'(v) - \mathsf{F}'(w)\|_{X'} \leq L_{\mathsf{F}'} \|v - w\|_X \qquad \forall v, w \in X.$$

This leads to

$$\left\|\mathsf{F}'(u^0)^{-1}(\mathsf{F}'(u)-\mathsf{F}'(v))\right\|_X \leq \frac{L_{\mathsf{F}'}}{\alpha_{\mathsf{F}'}}\|u-v\|_X \qquad \forall u,v\in X.$$

Moreover, if the initial guess $u^0\in X$ in (19) is sufficiently close to the solution $u^\star\in X$ of (1) in the sense that $\left\|\mathsf{F}(u^0)\right\|_{X'} < \alpha_{\mathsf{F}'}^2/2L_{\mathsf{F}'}$, then we infer that

$$\left\|\mathsf{F}'(u^0)^{-1}\mathsf{F}(u^0)\right\|_X \leq \frac{1}{\alpha_{\mathsf{F}'}}\left\|\mathsf{F}(u^0)\right\|_{X'} < \frac{\alpha_{\mathsf{F}'}}{2L_{\mathsf{F}'}}.$$

Referring to [12, Theorem 2.1], it follows that the classical Newton iteration with $\delta(u^n)=1$ in (19) is well-defined, converges to a solution of (1), and converges *quadratically*.

Remark 2.8. The proof of Theorem 2.6 is crucially based on (23). We emphasize that this bound may be satisfied even if the damping parameter $\delta(u^n)$ in (19) is larger than $2\alpha_{\mathsf{F}'}/L_{\mathsf{F}}$. This is particularly important when $2\alpha_{\mathsf{F}'}/L_{\mathsf{F}}\leq 1$, and the choice $\delta(u^n)=1$ (leading to local quadratic convergence, cf. Remark 2.7) is not admissible *a priori*. In this case, we may fix $\epsilon>0$ small, and aim to *a posteriori* attain the bound, for $n\geq 0$,

$$\mathsf{H}(u^n)-\mathsf{H}(u^{n+1})\geq \epsilon\left\|u^{n+1}-u^n\right\|_X^2. \tag{25}$$

To this end, we may pursue, for instance, the adaptive damping parameter selection approach proposed in [12, §3.1]. More precisely, in each iterative step, we define an initial value for $\delta(u^n)$ by the following *prediction* strategy:

$$\delta^{n,0}=\begin{cases}\min\left(\delta(u^{n-1})/\kappa,1\right) & \text{if } \delta(u^{n-2})\leq\delta(u^{n-1}),\\ \delta(u^{n-1}) & \text{else.}\end{cases}$$

where $0<\kappa<1$ is a fixed (correction) factor. Here, we set $\delta(u^{-2})=\delta(u^{-1})=\delta^0$, with δ^0 an initial choice. If u^{n+1} is obtained by the damped Newton method with damping parameter $\delta^{n,i}$, for some $i\geq 0$, then we need to verify wether or not (25) is satisfied. If not, then we adjust the damping parameter according to the *correction* strategy

$$\delta^{n,i+1}=\max\left(\alpha_{\mathsf{F}'}(\epsilon+L_{\mathsf{F}}/2)^{-1},\kappa\delta^{n,i}\right), \qquad i\geq 0. \tag{26}$$

Subsequently, we will compute u^{n+1} for the new choice $\delta^{n,i+1}$. This process is repeated until (25) is true, say after i^n iterations of (26). At this point, we let $\delta(u^n):=\delta^{n,i^n}$. Evidently, in view of (24), we remark that (25) will certainly hold once $\delta^{n,i}\leq\alpha_{\mathsf{F}'}(\epsilon+L_{\mathsf{F}}/2)^{-1}$. Moreover, in the Galerkin setting, we note that (20) can be verified numerically at the cost of an eigenvalue problem. Finally, if the values of the constants $\alpha_{\mathsf{F}'}$ and L_{F} are not easily accessible, we can simply use the (possibly pessimistic) damping parameter $\delta^{n,i+1}:=\kappa\delta^{n,i}$.

Remark 2.9. We emphasize that the proof of Theorem 2.6 works for much more general preconditioning operators A. Assume that F satisfies (F1), (F2), and (N3), the mapping $u\mapsto\mathsf{A}(u)u=a(u;u,\cdot)$ is continuous w.r.t. the weak topology on X', and the bilinear form induced by A fulfills (10) and (11). If $\alpha>L_{\mathsf{F}}/2$, then the crucial property (23) holds with $C:=\alpha-L_{\mathsf{F}}/2>0$; indeed, this can be shown as in the proof of Theorem 2.6, and the convergence of the method can be proved similarly as before. In particular, our unified iteration scheme does also recover Newton-like methods, e.g., the case $\mathsf{A}(u):=\delta\mathsf{F}'(u_0)$ for some initial guess $u_0\in X$, with a small enough damping parameter $\delta>0$.

3. Galerkin approach and a posteriori error analysis

The numerical solution of (1) is based on a finite-dimensional subspace $X_N\subset X$, and on the iterative linearization Galerkin (ILG) formulation (7), with a given initial guess $u_N^0\in X_N$. Since $X_N\subset X$, the assumptions in Section 2.1 guarantee the existence of $u_N^{n+1}\in X_N$ in each iteration step.

In this section, we will pursue two different strategies for the derivation of a posteriori error estimates for $\left\|u^\star-u_N^{n+1}\right\|_X$, where $u^\star\in X$ is the unique solution of (1). In both approaches an *elliptic reconstruction* technique, cf. [22, 23], will be employed. In the first method we use an

elliptic reconstruction for the solution of the *linear problem* (7), and the second strategy is based on applying a similar idea for a nonlinear discrete problem equivalent to (7). We will refer to this methods as the *linear* and *nonlinear elliptic reconstruction*, respectively.

3.1. **A posteriori error analysis based on a *linear* elliptic reconstruction.** For the sake of a general a posteriori error analysis, using a linear elliptic reconstruction, we suppose that there exists a computable bound $\eta(u_N^{n+1}, u_N^n)$ for the residual

$$\sup_{\substack{v\in X\\ \|v\|_X=1}} \left\{a(u_N^n; u_N^{n+1}, v) - \langle f(u_N^n), v\rangle\right\} \leq \eta(u_N^{n+1}, u_N^n). \tag{27}$$

We remark that, in the context of finite element methods for linear elliptic problems, there is a large body of literature focusing on the development of such estimates; see, e.g., [1, 29].

Theorem 3.1. *Suppose that* (F1) *and* (F2)*, cf. Section 2.1, as well as* (10) *and* (11) *hold true. Then, we have the a posteriori error bound*

$$\left\|u^\star - u_N^{n+1}\right\|_X \leq \frac{\beta}{\alpha\nu}\eta(u_N^{n+1}, u_N^n) + \frac{\beta + L_{\mathsf{F}}}{\nu}\left\|u_N^{n+1} - u_N^n\right\|_X, \tag{28}$$

where $u^\star$ is the unique solution of (1)*.*

Proof. Due to (10) and (11) there exists a unique $\widetilde{u}^{n+1} \in X$ such that

$$a(u_N^n; \widetilde{u}^{n+1}, v) = \langle f(u_N^n), v\rangle \qquad \forall v \in X. \tag{29}$$

We note that $\widetilde{u}^{n+1}$ is a reconstruction in the sense that $u_N^{n+1} \in X_N$ is the Galerkin projection of $\widetilde{u}^{n+1}$. By using the assumption (F2), we find that

$$\nu\left\|u^\star - u_N^{n+1}\right\|_X^2 \leq \left\langle \mathsf{F}(u^\star) - \mathsf{F}(u_N^{n+1}), u^\star - u_N^{n+1}\right\rangle = -\left\langle \mathsf{F}(u_N^{n+1}), u^\star - u_N^{n+1}\right\rangle,$$

since $u^\star \in X$ is the solution of (1). Hence,

$$\begin{aligned}\nu\left\|u^\star - u_N^{n+1}\right\|_X^2 \leq\ & -a(u_N^n; u_N^{n+1}, u^\star - u_N^{n+1}) + \left\langle f(u_N^n), u^\star - u_N^{n+1}\right\rangle\\ &+ a(u_N^n; u_N^{n+1} - u_N^n, u^\star - u_N^{n+1})\\ &+ a(u_N^n; u_N^n, u^\star - u_N^{n+1}) - \left\langle f(u_N^n), u^\star - u_N^{n+1}\right\rangle\\ &- \left\langle \mathsf{F}(u_N^{n+1}), u^\star - u_N^{n+1}\right\rangle.\end{aligned}$$

Using (29) and (5), this estimate transforms into

$$\begin{aligned}\nu\left\|u^\star - u_N^{n+1}\right\|_X^2 \leq\ & a(u_N^n; \widetilde{u}^{n+1} - u_N^{n+1}, u^\star - u_N^{n+1}) + a(u_N^n; u_N^{n+1} - u_N^n, u^\star - u_N^{n+1})\\ &- \left\langle \mathsf{F}(u_N^{n+1}) - \mathsf{F}(u_N^n), u^\star - u_N^{n+1}\right\rangle.\end{aligned}$$

Applying (11) and (F1), we find that

$$\begin{aligned}\nu\left\|u^\star - u_N^{n+1}\right\|_X^2 \leq\ & \beta\left\|\widetilde{u}^{n+1} - u_N^{n+1}\right\|_X\left\|u^\star - u_N^{n+1}\right\|_X + \beta\left\|u_N^{n+1} - u_N^n\right\|_X\left\|u^\star - u_N^{n+1}\right\|_X\\ &+ L_{\mathsf{F}}\left\|u_N^{n+1} - u_N^n\right\|_X\left\|u^\star - u_N^{n+1}\right\|_X.\end{aligned}$$

Dividing by $\left\|u^\star - u_N^{n+1}\right\|_X$ yields

$$\nu\left\|u^\star - u_N^{n+1}\right\|_X \leq \beta\left\|\widetilde{u}^{n+1} - u_N^{n+1}\right\|_X + (\beta + L_{\mathsf{F}})\left\|u_N^{n+1} - u_N^n\right\|_X. \tag{30}$$

Moreover, by the coercivity property (10), for $u_N^{n+1} \neq \widetilde{u}^{n+1}$, we note that

$$\alpha\left\|u_N^{n+1} - \widetilde{u}^{n+1}\right\|_X \leq \frac{a(u_N^n; u_N^{n+1} - \widetilde{u}^{n+1}, u_N^{n+1} - \widetilde{u}^{n+1})}{\left\|u_N^{n+1} - \widetilde{u}^{n+1}\right\|_X} \leq \sup_{\substack{v\in X\\ \|v\|_X=1}} a(u_N^n; u_N^{n+1} - \widetilde{u}^{n+1}, v).$$

Involving (29) and (27), we arrive at

$$\alpha\left\|u_N^{n+1} - \widetilde{u}^{n+1}\right\|_X \leq \sup_{\substack{v\in X\\ \|v\|_X=1}} \left(a(u_N^n; u_N^{n+1}, v) - \langle f(u_N^n), v\rangle\right) \leq \eta(u_N^{n+1}, u_N^n). \tag{31}$$

Inserting this estimate into (30), finishes the proof. □

Remark 3.2. We emphasize that the estimator (28) permits to bound the error $\|u^\star - u_N^{n+1}\|_X$ separately in terms of the *discretization* error indicator, ${}^{\beta}/_{\alpha\nu}\eta(u_N^{n+1}, u_N^n)$, and of the *linearization* error indicator, ${}^{(\beta+L_{\mathsf{F}})}/_{\nu}\|u_N^{n+1} - u_N^n\|_X$. Let us discuss these error contributions in more detail: First, recall that u_N^{n+1} is the Galerkin projection of $\widetilde{u}^{n+1}$ in the sense of the Galerkin orthogonality property

$$a(u_N^n; \widetilde{u}^{n+1} - u_N^{n+1}, v) = 0 \qquad \forall v \in X_N.$$

Thus, the quantity $\|\widetilde{u}^{n+1} - u_N^{n+1}\|_X$ is an indicator for the quality of approximation of the Galerkin discretization. Here, invoking (31), we see that

$$\|\widetilde{u}^{n+1} - u_N^{n+1}\|_X \leq \alpha^{-1}\eta(u_N^{n+1}, u_N^n),$$

wherefore it is reasonable to interpret the η-term as the discretization error contribution in the total estimator. Secondly, it holds that $\|u_N^{n+1} - u_N^n\|_X \to 0$ for $n \to 0$, which underlines the convergence of the iterative linearization; thereby, this term can be seen to quantify the linearization effect in the ILG approximation.

Remark 3.3. The constants for the estimator in Theorem 3.1 for the Zarantonello iteration can be slightly improved. This is due to the fact that the preconditioning operator A is constant in this case; cf. [11, Proposition 2.2].

3.2. **A posteriori error analysis based on a *nonlinear* elliptic reconstruction.** In this section we devise an a posteriori error estimate for the *linear* Galerkin iteration (7) based on applying the reconstruction technique to a *nonlinear* discrete problem equivalent to (7). We underline that the nonlinear problem (34), exactly as in the case of the linear elliptic reconstruction from (29), is of purely theoretical relevance in the derivation of the estimator, and does not need to be solved in the actual computations.

We define an operator $\psi_N : X \to X_N$, where, for fixed $w \in X$, we let $\psi_N(w)$ to be the Riesz representative of $\mathsf{F}(w) \in X_N'$ with respect to the inner product in X, i.e.

$$(\psi_N(w), v)_X = \langle \mathsf{F}(w), v\rangle \qquad \forall v \in X_N. \tag{32}$$

Note that, if $\mathfrak{u}_N$ is the solution of the *nonlinear* Galerkin approximation of (2) with respect to the discrete space X_N, i.e.

$$\mathfrak{u}_N \in X_N : \qquad \langle \mathsf{F}(\mathfrak{u}_N), v\rangle = 0 \qquad \forall v \in X_N, \tag{33}$$

then it holds that $\psi_N(\mathfrak{u}_N) = 0$.

For each $n \geq 0$, we define the *nonlinear* elliptic reconstruction $\widetilde{u}^{n+1} \in X$ of the solution $u_N^{n+1} \in X_N$ of (7) by

$$\big\langle \mathsf{F}(\widetilde{u}^{n+1}), v\big\rangle = (\psi_N(u_N^{n+1}), v)_X \qquad \forall v \in X. \tag{34}$$

By construction of the operator ψ_N, it holds that u_N^{n+1} is the Galerkin approximation of (34), i.e.

$$\big\langle \mathsf{F}(\widetilde{u}^{n+1}) - \mathsf{F}(u_N^{n+1}), v\big\rangle = 0 \qquad \forall v \in X_N.$$

Then, with the aid of (F2), we infer that

$$\nu\|\widetilde{u}^{n+1} - u_N^{n+1}\|_X^2 \leq \big\langle \mathsf{F}(\widetilde{u}^{n+1}) - \mathsf{F}(u_N^{n+1}), \widetilde{u}^{n+1} - u_N^{n+1}\big\rangle.$$

Hence,

$$\begin{aligned}\nu\|\widetilde{u}^{n+1} - u_N^{n+1}\|_X &\leq \sup_{\substack{w\in X\\ \|w\|_X=1}} \big\langle \mathsf{F}(\widetilde{u}^{n+1}) - \mathsf{F}(u_N^{n+1}), w\big\rangle \\ &\leq \sup_{\substack{w\in X\\ \|w\|_X=1}} \big\{(\psi_N(u_N^{n+1}), w)_X - \big\langle \mathsf{F}(u_N^{n+1}), w\big\rangle\big\}.\end{aligned}$$

Now, suppose that there exists a computable bound $\eta(u_N^{n+1})$ such that

$$\sup_{\substack{w\in X\\ \|w\|_X=1}} \big\{(\psi_N(u_N^{n+1}), w)_X - \big\langle \mathsf{F}(u_N^{n+1}), w\big\rangle\big\} \leq \eta(u_N^{n+1}).$$

Then,

$$\left\|\widetilde{u}^{n+1} - u_N^{n+1}\right\|_X \leq \nu^{-1}\eta(u_N^{n+1}). \tag{35}$$

Similarly as in the linear case, in the specific context of the finite element method for elliptic PDE, such residual bounds can be obtained by standard techniques, see, e.g., [1,29]. This will be carried out in Section 5.2 for quasilinear elliptic PDE.

Theorem 3.4. *Given* (F1) *and* (F2) *(cf. Section 2.1), there holds the a posteriori error bound*

$$\left\|u^\star - u_N^{n+1}\right\|_X \leq \frac{1}{\nu}\left(\eta(u_N^{n+1}) + \left\|\psi_N(u_N^{n+1})\right\|_X\right),$$

where $u^\star$ is the exact solution of (1).

We note that, in the bound above, $\nu^{-1}\eta(u_N^{n+1})$ is an indicator for the discretization error by a similar argument as in Remark 3.2, and $\nu^{-1}\left\|\psi_N(u_N^{n+1})\right\|_H$ controls the linearization error. Indeed, since $\psi_N(\mathfrak{u}_N) = 0$ and $\{u_N^n\}_{n\geq 0}$ converges to the solution $\mathfrak{u}_N$ of (33) by our analysis in Section 2, we see that $\left\|\psi_N(u_N^{n+1})\right\|_X \to 0$ as $n \to \infty$.

Proof. By invoking the triangle inequality and (35), we find that

$$\left\|u^\star - u_N^{n+1}\right\|_X \leq \nu^{-1}\eta(u_N^{n+1}) + \left\|u^\star - \widetilde{u}^{n+1}\right\|_X.$$

Moreover, due to (F2), we observe that

$$\nu\left\|u^\star - \widetilde{u}^{n+1}\right\|_X^2 \leq \left\langle \mathsf{F}(u^\star) - \mathsf{F}(\widetilde{u}^{n+1}), u^\star - \widetilde{u}^{n+1}\right\rangle = \left\langle \mathsf{F}(\widetilde{u}^{n+1}), \widetilde{u}^{n+1} - u^\star\right\rangle.$$

By using (34), and upon applying the Cauchy-Schwarz inequality, this leads to

$$\nu\left\|u^\star - \widetilde{u}^{n+1}\right\|_X^2 \leq (\psi_N(u_N^{n+1}), \widetilde{u}^{n+1} - u^\star)_X \leq \left\|\psi_N(u_N^{n+1})\right\|_X \left\|\widetilde{u}^{n+1} - u^\star\right\|_X.$$

This yields the claim. □

Remark 3.5. We compare the a posteriori error estimators from Theorem 3.1 and Theorem 3.4: The proof of Theorem 3.1—and thus the constants in the bound (28)—strongly depend on the choice of the preconditioning operator A, i.e. on the specific iterative linearization method. Moreover, the same comment applies for the computable bound $\eta(u_N^{n+1}, u_N^n)$. In contrast, the estimator from Theorem 3.4, resulting from the nonlinear elliptic construction, as well as the computable bound $\eta(u_N^n)$ are completely independent of the iteration scheme, and merely relies on the underlying PDE problem. We note further that the a posteriori error estimator from Theorem 3.1 allows for more general reflexive Banach spaces X, whereas Theorem 3.4 requires a Hilbert space setting.

4. An abstract ILG procedure

The estimates from Theorems 3.1 and 3.4 allow to control the error between the solution of (1) and the discrete system (7) with respect to two individual terms, one of which expresses the error of the linearization, and will be denoted by $\mathcal{E}^n_{\mathrm{Linear},N}$, and the other, which we signify by $\mathcal{E}^n_{\mathrm{Galerkin},N}$, bounds the Galerkin discretization error. In a finite element context, the latter error will typically be composed of local contributions for each element; this, in turn, enables to refine the mesh locally. The algorithm, which will be presented below, uses an *adaptive interplay* between those two controlling terms. More precisely, on a given Galerkin space, we iterate as long as the linearization error dominates and, in addition, until it is, in a certain way, smaller than a given bound depending on the number of Galerkin space enrichments, N. Once the linearization error is small enough, and is up to a factor ϑ less than the one arising from the Galerkin method, we enrich the Galerkin space according to the local error indicators in order to attain a smaller discretization error. Subsequently, we will perform the linearization on the enriched space. In this way, the goal of the ILG algorithm is to compute an approximation of the solution of (1) which, on the one hand, is sufficiently accurate, and, on the other hand, is attained from a minimal number of iterations.

4.1. **Adaptive ILG algorithm.** For the purpose of this section, we assume that our ILG Algorithm 1, to be presented below, generates a sequence of hierarchically enriched Galerkin spaces, $X_0 \subset X_1 \subset X_2 \subset \dots$, on each of which we perform at least one iterative step. Furthermore, we will make use of a prescribed positive function $\sigma : \mathbb{N} \to (0, \infty)$ which satisfies

$$\sigma(N) \to 0 \quad \text{for} \quad N \to \infty. \tag{36}$$

Its role is to ensure that the linearization error tends to zero for an increasing number N of Galerkin space enrichments. For instance, in the context of the finite element method, a sensible choice is $\sigma(N) = \mathcal{O}(|\mathcal{T}_N|^{-s})$, where $|\mathcal{T}_N|$ signifies the number of elements in the mesh, and s is the expected convergence rate; cf. Section 5.3 below. Recall that, for any fixed $N \geq 0$, our theory in Section 2 guarantees, under certain conditions, that the difference $\left\|u_N^n - u_N^{n-1}\right\|_X$ tends to zero for increasing n; in particular, it can be made smaller than $\sigma(N)$ for n large enough.

An adaptive ILG procedure for the interactive reduction of discretization and linearization errors is proposed in Algorithm 1. We note that this algorithm can be performed with any of the iterative procedures from Section 2.3, and with either the error estimators obtained from the linear or nonlinear elliptic reconstructions from Section 3. The input and output arguments as well as the components of the implemented algorithm may, of course, depend on the error estimator and the specific iterative linearization scheme applied.

Algorithm 1 Adaptive ILG algorithm

1: Prescribe a tolerance $\epsilon_{\text{tol}} > 0$, and an adaptivity parameter $\vartheta > 0$. Set $N := 0$ and $n := 0$. Start with an initial Galerkin space $X_0 \subset X$, and an initial guess $u_0^0 \in X_0$.
2: **repeat**
3: Set $\mathcal{E}^n_{\text{Linear},N} := 1$ and $\mathcal{E}^n_{\text{Galerkin},N} := 0$.
4: **while** $\mathcal{E}^n_{\text{Galerkin},N} \leq \vartheta \mathcal{E}^n_{\text{Linear},N}$ or $\left\|u_N^n - u_N^{n-1}\right\|_X > \sigma(N)$ **do**
5: Perform a single iterative linearization step to obtain u_N^{n+1} from u_N^n; cf. (7).
6: Estimate the linearization error $\mathcal{E}^{n+1}_{\text{Linear},N}$ and the Galerkin error indicator $\mathcal{E}^{n+1}_{\text{Galerkin},N}$.
7: Update $n \leftarrow n + 1$.
8: **end while**
9: Let $u_N^{n^\star} := u_N^n \in X_N$, and enrich the Galerkin space X_N appropriately based on the error indicator $\mathcal{E}^n_{\text{Galerkin},N}$ in order to obtain X_{N+1}.
10: Set $\mathcal{E}_{\text{tot}} := \mathcal{E}^n_{\text{Linear},N} + \mathcal{E}^n_{\text{Galerkin},N}$.
11: Define $u_{N+1}^0 := u_N^{n^\star}$ by inclusion $X_{N+1} \hookleftarrow X_N$.
12: Update $N \leftarrow N + 1$, and set $n := 0$.
13: **until** $\mathcal{E}_{\text{tot}} < \epsilon_{\text{tol}}$.
14: **return** the sequence of discrete solutions $u_N^{n^\star} \in X_N$.

4.2. **A remark on convergence.** Given a Galerkin space X_N, we recall the solution $\mathfrak{u}_N \in X_N$ of (33). Furthermore, we let, as in the Algorithm 1, $u_N^{n^\star} \in X_N$ be the final approximation on X_N (i.e. before the Galerkin space is enriched). We establish the convergence of $u_N^{n^\star}$ to the unique solution $u^\star$ of (1) under the following assumption:

(AG) The hierarchically enriched Galerkin spaces $X_0 \subset X_1 \subset X_2 \subset \dots$ generated by Algorithm 1 are such that the iterative Galerkin approximations $\mathfrak{u}_N \in X_N$ from (33) converge to the exact solution $u^\star \in X$ of (1) for $N \to \infty$.

Proposition 4.1. *If* F *from* (1) *fulfills* (F1) *and* (F2) *(cf. Section 2.1), and the Galerkin method satisfies* (AG)*, then Algorithm 1based on an iterative linearization scheme* (7)*, with* $a(u;\cdot,\cdot)$ *satisfying the properties* (10) *and* (11)*, generates a sequence* $\{u_N^{n^\star}\}_{N\geq 0}$ *which converges to the unique solution* $u^\star \in X$ *of* (1)*.*

Proof. Using (F2) and involving (33), it holds that

$$\nu \left\|\mathfrak{u}_N - u_N^{n^\star-1}\right\|_X^2 \leq \left\langle \mathsf{F}(\mathfrak{u}_N) - \mathsf{F}(u_N^{n^\star-1}), \mathfrak{u}_N - u_N^{n^\star-1} \right\rangle = \left\langle \mathsf{F}(u_N^{n^\star-1}), u_N^{n^\star-1} - \mathfrak{u}_N \right\rangle.$$

Invoking (5), (7), and (11), we obtain

$$\nu \left\| \mathfrak{u}_N - u_N^{n^\star-1} \right\|_X^2 \leq a(u_N^{n^\star-1}; u_N^{n^\star-1} - u_N^{n^\star}, u_N^{n^\star-1} - \mathfrak{u}_N) \leq \beta \left\| u_N^{n^\star} - u_N^{n^\star-1} \right\|_X \left\| u_N^{n^\star-1} - \mathfrak{u}_N \right\|_X,$$

and thus

$$\left\| \mathfrak{u}_N - u_N^{n^\star-1} \right\|_X \leq \frac{\beta}{\nu} \left\| u_N^{n^\star} - u_N^{n^\star-1} \right\|_X.$$

By the triangle inequality, this leads to

$$\left\| \mathfrak{u}_N - u_N^{n^\star} \right\|_X \leq \left\| \mathfrak{u}_N - u_N^{n^\star-1} \right\|_X + \left\| u_N^{n^\star} - u_N^{n^\star-1} \right\|_X \leq \left(\frac{\beta}{\nu} + 1 \right) \left\| u_N^{n^\star} - u_N^{n^\star-1} \right\|_X. \tag{37}$$

Notice that the stopping criterion for the while loop in Algorithm 1 implies that

$$\left\| u_N^{n^\star} - u_N^{n^\star-1} \right\|_X \leq \sigma(N) \qquad \forall N \geq 0, \tag{38}$$

with σ satisfying (36). Then, invoking the triangle inequality, as well as (37) and (38), yields

$$\left\| u^\star - u_N^{n^\star} \right\|_X \leq \| u^\star - \mathfrak{u}_N \|_X + \left\| \mathfrak{u}_N - u_N^{n^\star} \right\|_X \leq \| u^\star - \mathfrak{u}_N \|_X + \left(\frac{\beta}{\nu} + 1 \right) \sigma(N).$$

The first term on the right-hand side tends to zero for $N \to \infty$ by virtue of (AG), and the same holds true for the second term due to (36). We deduce that $u_N^{n^\star} \to u^\star$ as $N \to \infty$. □

In our subsequent paper [20] we further analyze the convergence of the adaptive ILG algorithm. In fact, we establish the linear convergence rate of our algorithm, with a slightly different a posteriori error estimator, under reasonable assumptions. For instance, in the context of finite element discretizations of second-order PDE in divergence form, cf. Section 5, we state in [20] an a posteriori error estimator which guarantees the linear convergence regime.

5. Application to second-order PDE in divergence form

In this section, we will apply our analytical findings to the quasilinear elliptic PDE problem

$$u \in X: \qquad \mathsf{F}(u) := -\nabla \cdot \left\{ \mu \left(|\nabla u|^2 \right) \nabla u \right\} - g = 0 \qquad \text{in } X'. \tag{39}$$

Here, $\Omega \subset \mathbb{R}^d$, for $d \in \mathbb{N}$, is an open and bounded domain with Lipschitz boundary $\Gamma := \partial\Omega$, and $X := H_0^1(\Omega)$ is the standard Sobolev space of H^1-functions on Ω with zero trace along the boundary Γ; the inner product and norm on X are defined, respectively, by $(u, v)_X := (\nabla u, \nabla v)_{L^2(\Omega)}$ and $\|u\|_X := \|\nabla u\|_{L^2(\Omega)}$, for $u, v \in X$. Equations of the form (39) are widely used in mathematical models of physical applications including, for instance, hydro- and gas-dynamics, or plasticity; we refer to [32, §69.2–69.3] and [6, §1.1] for a discussion of the physical meaning. We suppose that $g \in X' = H^{-1}(\Omega)$ in (39) is given, and $\mu \in C^1([0, \infty))$ fulfills

$$m_\mu(t - s) \leq \mu(t^2)t - \mu(s^2)s \leq M_\mu(t - s), \qquad t \geq s \geq 0, \tag{40}$$

with constants $m_\mu, M_\mu > 0$. In particular, upon setting $s = 0$, we observe that

$$m_\mu \leq \mu(t^2) \leq M_\mu \qquad \forall t \geq 0. \tag{41}$$

Under condition (40) it can be shown that the nonlinear operator F from (39) satisfies the properties (F1) and (F2) with

$$\nu = m_\mu, \qquad L_{\mathsf{F}} = 3M_\mu; \tag{42}$$

see [33, Proposition 25.26].

We note the weak form of the boundary value problem (39) in X:

$$u \in X: \qquad \int_\Omega \mu \left(|\nabla u|^2 \right) \nabla u \cdot \nabla v \, \mathrm{d}\boldsymbol{x} = \langle g, v \rangle \qquad \forall v \in X. \tag{43}$$

5.1. **Convergence of iterative linearizations.** In the sequel, we will investigate the convergence of the various iteration schemes from Section 2.3 as applied to the PDE (39). The convergence of the Zarantonello iteration follows immediately from Theorem 2.2.

Proposition 5.1. *If μ satisfies* (40) *and* F *is given by* (39)*, then the Zarantonello iteration* (12)*, i.e.*

$$u^{n+1} \in X: \qquad -\Delta u^{n+1} = -\Delta u^n + \delta \nabla \cdot \left\{ \mu\left(|\nabla u^n|^2\right) \nabla u^n \right\} + \delta g, \qquad n \geq 0,$$

converges to the unique solution of (39) *for any* $\delta \in \left]0, {}^{2m_\mu}/_{9M_\mu^2}\right[$.

In order to study the Kačanov iteration method for (39), let us define, for $u \in X$, the linear preconditioning operator

$$\mathsf{A}(u)v := -\nabla \cdot \left\{ \mu\left(|\nabla u|^2\right) \nabla v \right\}, \qquad v \in X. \tag{44}$$

In addition to (40), we assume that μ is monotone decreasing, i.e.

$$\mu'(t) \leq 0 \qquad \forall t \geq 0. \tag{45}$$

Proposition 5.2. *Let μ satisfy* (40) *and* (45)*. Then, the Kačanov iteration* (15)*, i.e.*

$$u^{n+1} \in X: \qquad -\nabla \cdot \left\{ \mu\left(|\nabla u^n|^2\right) \nabla u^{n+1} \right\} = g, \qquad n \geq 0,$$

converges to the unique solution of (39)*.*

Proof. We will show that the assumptions of Theorem 2.5 are satisfied. To this end, for A from (44), and any $u \in X$, we define the symmetric bilinear form $a(u; v, w) := \langle \mathsf{A}(u)v, w\rangle$, for $v, w \in X$. Then, using (41) in combination with the Cauchy-Schwarz inequality shows the coercivity and continuity properties (10) and (11) with

$$\alpha = m_\mu, \qquad \beta = M_\mu, \tag{46}$$

respectively. Furthermore, we introduce the potential $\mathsf{G}: X \to \mathbb{R}$ by

$$\mathsf{G}(u) := \int_\Omega \psi\left(|\nabla u|^2\right) \,\mathsf{d}\boldsymbol{x}, \qquad \text{with} \quad \psi(s) := \frac{1}{2}\int_0^s \mu(t)\,\mathsf{d}t. \tag{47}$$

For $u \in X$, taking the Gateaux derivative of G, we find that

$$\langle \mathsf{G}'(u), v\rangle = \int_\Omega 2\psi'\left(|\nabla u|^2\right) \nabla u \cdot \nabla v \,\mathsf{d}\boldsymbol{x} = \int_\Omega \mu\left(|\nabla u|^2\right) \nabla u \cdot \nabla v \,\mathsf{d}\boldsymbol{x} = a(u; u, v),$$

for any $v \in X$. Thus, we infer that $\mathsf{G}'(u) = a(u; u, \cdot) = \mathsf{F}(u) + g$. Recalling (F2), this implies the strong monotonicity property (16) with $c_0 = m_\mu$, and we conclude that (K1) holds true. In addition, due to (45), for any $t \geq s \geq 0$, it holds that

$$\psi(t) - \psi(s) = \frac{1}{2}\int_s^t \mu(\tau)\,\mathsf{d}\tau \geq \frac{1}{2}(t-s)\mu(t),$$

and similarly for $s \geq t \geq 0$,

$$\psi(t) - \psi(s) = -\frac{1}{2}\int_t^s \mu(\tau)\,\mathsf{d}\tau \geq -\frac{1}{2}(s-t)\mu(t) = \frac{1}{2}(t-s)\mu(t).$$

Hence, for any $u, v \in X$, we have

$$\mathsf{G}(u) - \mathsf{G}(v) \geq \frac{1}{2}\int_\Omega \mu\left(|\nabla u|^2\right)\left(|\nabla u|^2 - |\nabla v|^2\right) \,\mathsf{d}\boldsymbol{x} = \frac{1}{2}\left(a(u; u, u) - a(u; v, v)\right),$$

which shows (K2). □

Finally, we turn our attention to the damped Newton iteration.

Proposition 5.3. *Let μ satisfy* (40) *and* (45)*. Moreover, suppose that the damping parameter $\delta: X \to [\delta_{\min}, \delta_{\max}]$ is a continuous functional, for some constants $\delta_{\min}, \delta_{\max}$, with $0 < \delta_{\min} \leq \delta_{\max} < {}^{2m_\mu}/_{3M_\mu}$. Then, the damped Newton iteration* (19) *for the nonlinear PDE* (39) *converges to its unique solution in X.*

We will prove this proposition by showing that the assumptions of Theorem 2.6 are satisfied. For this purpose we require the following auxiliary result.

Lemma 5.4. *If μ satisfies* (40)*, then the operator $u \mapsto \mathsf{F}'(u)u$ is continuous from X to X' with respect to the weak topology on X'.*

Proof. By taking the limit $s \nearrow t$ in (40), we infer that $m_\mu \leq \frac{\mathrm{d}}{\mathrm{d}t}\left(\mu(t^2)t\right) \leq M_\mu$, and, thereby,

$$m_\mu \leq 2\mu'(t^2)t^2 + \mu(t^2) \leq M_\mu \qquad \forall t \geq 0. \tag{48}$$

Moreover, a simple but lengthy calculation shows that

$$\langle \mathsf{F}'(u)v, w\rangle = \int_\Omega 2\mu'(|\nabla u|^2)(\nabla u \cdot \nabla v)(\nabla u \cdot \nabla w)\,\mathrm{d}\boldsymbol{x} + \int_\Omega \mu(|\nabla u|^2)\nabla v \cdot \nabla w\,\mathrm{d}\boldsymbol{x}, \tag{49}$$

for any $u, v, w \in X$. Consider a sequence $\{u^k\}_{k\geq 0} \subset X$ which converges to a limit $u \in X$, i.e.

$$\left\|u - u^k\right\|_X \to 0, \qquad k \to \infty. \tag{50}$$

Since $X = H_0^1(\Omega)$, we find that $\nabla u^k \to \nabla u$ in $L^2(\Omega)$ for $k \to \infty$. Thus, there is a subsequence such that

$$\nabla u^{k'} \to \nabla u \quad \text{a.e. in } \Omega \text{ for } k' \to \infty, \tag{51}$$

see, e.g., [27, Theorem 3.12]. Hence, defining the function $\omega(t) := 2\mu'(t)t + \mu(t)$, $t \geq 0$, it holds

$$\begin{aligned}\left\langle \mathsf{F}'(u)u - \mathsf{F}'(u^{k'})u^{k'}, w\right\rangle &= \int_\Omega \left(\omega(|\nabla u|^2) - \omega(|\nabla u^{k'}|^2)\right)\nabla u \cdot \nabla w\,\mathrm{d}\boldsymbol{x} \\ &\quad + \int_\Omega \omega(|\nabla u^{k'}|^2)\nabla(u - u^{k'}) \cdot \nabla w\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

We note that both terms on the right-hand side tend to 0 as $k' \to \infty$: Indeed, for the first integral this follows from the continuity of ω, (51), (48), and the dominated convergence theorem; for the second integral, we recall (48) and (50). Finally, referring to [31, Proposition 10.13(2)], we conclude the weak convergence of the entire sequence, i.e. $\mathsf{F}'(u^k)u^k \rightharpoonup \mathsf{F}'(u)u$ as $k \to \infty$. This finishes the proof. □

Proof of Proposition 5.3. For $v = w$ in (49) we have that

$$\langle \mathsf{F}'(u)v, v\rangle = \int_\Omega 2\mu'(|\nabla u|^2)|\nabla u \cdot \nabla v|^2 + \int_\Omega \mu(|\nabla u|^2)|\nabla v|^2\,\mathrm{d}\boldsymbol{x}.$$

Exploiting (45), and using the Cauchy-Schwarz inequality, we notice that

$$2\mu'(|\nabla u|^2)|\nabla u \cdot \nabla v|^2 \geq 2\mu'(|\nabla u|^2)|\nabla u|^2|\nabla v|^2.$$

It follows that

$$\langle \mathsf{F}'(u)v, v\rangle \geq \int_\Omega \left(2\mu'(|\nabla u|^2)|\nabla u|^2 + \mu\left(|\nabla u|^2\right)\right)|\nabla v|^2\,\mathrm{d}\boldsymbol{x}.$$

Applying (48) implies that $\langle \mathsf{F}'(u)v, v\rangle \geq m_\mu \|v\|_X^2$ for any $u, v \in X$; this shows (20) with

$$\alpha_{\mathsf{F}'} = m_\mu. \tag{52}$$

In addition, in view of (42), we observe that ${}^{2\alpha_{\mathsf{F}'}}/_{L_\mathsf{F}} = {}^{2m_\mu}/_{3M_\mu} > \delta_{\max}$, as required in Theorem 2.6. Furthermore, application of the Cauchy-Schwarz inequality, and involving (45), yields

$$\begin{aligned}\langle \mathsf{F}'(u)v, w\rangle &\leq \int_\Omega \left(\left|2\mu'(|\nabla u|^2)\right||\nabla u|^2 + \mu\left(|\nabla u|^2\right)\right)|\nabla v||\nabla w|\,\mathrm{d}\boldsymbol{x} \\ &= -\int_\Omega \left(2\mu'(|\nabla u|^2)\,|\nabla u|^2 + \mu\left(|\nabla u|^2\right)\right)|\nabla v||\nabla w|\,\mathrm{d}\boldsymbol{x} \\ &\quad + 2\int_\Omega \mu\left(|\nabla u|^2\right)|\nabla v||\nabla w|\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Employing (48) and (41), this leads to

$$\langle \mathsf{F}'(u)v, w\rangle \leq (2M_\mu - m_\mu)\int_\Omega |\nabla v||\nabla w|\,\mathrm{d}\boldsymbol{x} \leq (2M_\mu - m_\mu)\,\|v\|_X\,\|w\|_X,$$

which gives (21) with

$$\beta_{\mathsf{F}'} = 2M_\mu - m_\mu. \tag{53}$$

In order to prove (N3), let us define the functional $\mathsf{H} : X \to \mathbb{R}$ by

$$\mathsf{H}(u) := \int_\Omega \psi\left(|\nabla u|^2\right) \,\mathrm{d}\boldsymbol{x} - \langle g, u\rangle\,, \qquad u \in X,$$

with ψ as in (47). It holds that

$$\langle \mathsf{H}'(u), v\rangle = \int_\Omega \mu\left(|\nabla u|^2\right) \nabla u \cdot \nabla v \,\mathrm{d}\boldsymbol{x} - \langle g, v\rangle = \langle \mathsf{F}(u), v\rangle\,,$$

for all $v \in X$. Finally, to establish (N2), we introduce the functional $\mathsf{G} : X \to \mathbb{R}$ by $\mathsf{G}(u) := \mathcal{F}(u) - \mathsf{H}(u)$, where $\mathcal{F}(u) := \langle \mathsf{F}(u), u\rangle$, $u \in X$. For $u \in X$, the Gateaux derivative of $\mathcal{F}$ is given by

$$\langle \mathcal{F}'(u), v\rangle = \langle \mathsf{F}'(u)u, v\rangle + \langle \mathsf{F}(u), v\rangle \qquad \forall v \in X.$$

It follows that $\mathsf{G}'(u) = \mathcal{F}'(u) - \mathsf{H}'(u) = \mathsf{F}'(u)u + \mathsf{F}(u) - \mathsf{F}(u) = \mathsf{F}'(u)u$. Finally, due to Lemma 5.4 the mapping $u \mapsto \mathsf{G}'(u) = \mathsf{F}'(u)u$ is continuous with respect to the weak topology on X'. □

5.2. **Iterative linearized FEM.** For the sake of discretizing (43), and thereby, of obtaining an ILG formulation for (39), we will use a conforming finite element framework. To illustrate our approach we deal with a physical domain $\Omega \subset \mathbb{R}^2$; we remark, however, that the discussion below can, in principle, be generalized to higher dimensions. We consider regular and shape-regular meshes $\mathcal{T}_h$ that partition the domain Ω into open and disjoint triangles $K \in \mathcal{T}_h$ such that $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} \overline{K}$. We denote by $h_K := \mathrm{diam}(K)$ the diameter of $K \in \mathcal{T}_h$, and let $h := \max_{K\in\mathcal{T}_h} h_K$. Moreover, we consider the finite element space

$$X_h := \left\{v \in H^1_0(\Omega) : v|_K \in \mathcal{P}_p(K)\ \forall K \in \mathcal{T}_h\right\}, \tag{54}$$

where, for fixed $p \in \mathbb{N}$, we signify by $\mathcal{P}_p(K)$ the space of all polynomials of total degree at most $p \geq 1$ on $K \in \mathcal{T}_h$.

Within the adaptive ILG framework, we will consider a sequence of meshes $\{\mathcal{T}_N\}_{N\geq 0}$, whereby we start with an initial conforming triangulation $\mathcal{T}_0$ of Ω. All subsequent meshes are obtained by refinement, i.e. for $N \geq 0$, the mesh $\mathcal{T}_{N+1}$ is a hierarchical refinement of $\mathcal{T}_N$. Moreover, we will denote by X_N the finite element space associated to the mesh $\mathcal{T}_N$.

For an edge $e \subset \partial K^+ \cap \partial K^-$, which is the intersection of two neighbouring elements $K^\pm \in \mathcal{T}_N$, we signify by $[\![\boldsymbol{v}]\!]\,|_e = \boldsymbol{v}^+|_e \cdot \boldsymbol{n}_{K^+} + \boldsymbol{v}^-|_e \cdot \boldsymbol{n}_{K^-}$ the jump of a (vector-valued) function $\boldsymbol{v}$ along e, where $\boldsymbol{v}^\pm|_e$ 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.

5.2.1. *A posteriori error analysis via linear elliptic reconstruction.* In this section, we discuss the a posteriori error estimate from Theorem 3.1 in the specific context of the nonlinear PDE (39) and the finite element framework presented above. Introducing the residual

$$\mathsf{R}(u; v, w) := a(u; v, w) - \langle f(u), w\rangle\,, \qquad u, v \in X_N,\ w \in X,$$

it is fairly straightforward to verify that, for all of the three iterative linearization schemes from Section 5.1, and for $g \in L^2(\Omega)$ in (39), it holds the special form

$$\mathsf{R}(u_N^n; u_N^{n+1}, w) = -\int_\Omega \mathbf{q}_N^n \cdot \nabla w \,\mathrm{d}\boldsymbol{x} + \int_\Omega p_N^n w \,\mathrm{d}\boldsymbol{x} \qquad \forall w \in X,$$

with some $p_N^n \in L^2(\Omega)$ and $\mathbf{q}_N^n \in H^1(\Omega)^2$, which can be represented explicitly. Then, recalling (7), we may conclude that

$$\mathsf{R}(u_N^n; u_N^{n+1}, w) = \mathsf{R}(u_N^n; u_N^{n+1}, w - w_N) = -\int_\Omega \mathbf{q}_N^n \cdot \nabla(w - w_N) \,\mathrm{d}\boldsymbol{x} + \int_\Omega p_N^n (w - w_N) \,\mathrm{d}\boldsymbol{x},$$

for any $w_N \in X_N$. Therefore, choosing w_N to be a quasi-interpolant of w, and pursuing a standard residual-based a posteriori error analysis (see, e.g., [29]), we deduce the upper bound

$$\sup_{\substack{w\in X\\ \|w\|_X=1}} \mathsf{R}(u_N^n; u_N^{n+1}, w) \leq C_{\mathsf{I}} \left(\sum_{K\in\mathcal{T}_N} \eta_K^2 \right)^{1/2},$$

where $C_{\mathsf{I}} > 0$ is an interpolation constant (only depending on the polynomial degree p and on the shape-regularity of the mesh), and

$$\eta_K^2 = h_K^2 \left\| \nabla \cdot \mathbf{q}_N^n + p_N^n \right\|_{L^2(K)}^2 + \frac{1}{2} h_K \left\| [\![\mathbf{q}_N^n]\!] \right\|_{L^2(\partial K\setminus\Gamma)}^2, \qquad K \in \mathcal{T}_N, \tag{55}$$

is a computable error indicator.

Theorem 5.5. *Let F be defined by (39) with μ fulfilling (40) and (45), and let $X_N \subset H_0^1(\Omega)$ be a conforming finite element space as in (54) on a mesh $\mathcal{T}_N$. If $u^\star$ is the unique solution of (39), and $\{u_N^n\}_{n\geq 0}$ is a sequence of ILG solutions obtained by any of the iterative linearization procedures from Section 5.1 on X_N, then it holds the a posteriori estimate*

$$\left\| u^\star - u_N^{n+1} \right\|_X \leq \frac{\beta C_{\mathsf{I}}}{\alpha m_\mu} \left(\sum_{K\in\mathcal{T}_N} \eta_K^2 \right)^{1/2} + \frac{\beta + 3M_\mu}{m_\mu} \left\| u_N^{n+1} - u_N^n \right\|_X,$$

where $C_{\mathsf{I}} > 0$ is a constant, and

$$(\alpha, \beta) = \begin{cases} (\delta^{-1}, \delta^{-1}) & \text{for the Zarantonello iteration, cf. (13),} \\ (m_\mu, M_\mu) & \text{for the Kačanov iteration, cf. (46),} \\ (m_\mu/\delta_{\max}, (2M_\mu - m_\mu)/\delta_{\min}) & \text{for the Newton iteration, cf. (22), (52), and (53),} \end{cases}$$

and η_K, for $K \in \mathcal{T}_N$, is defined in (55).

Proof. The result follows from Theorem 3.1, whereby we replace the constants ν and L_{F} from (42), and insert the values of α and β from (10) and (11) for the respective iterative schemes from Section 5.1. □

5.2.2. *Error estimator via* nonlinear *elliptic reconstruction.* Following our abstract analysis in Section 3.2, we consider the residual

$$\mathsf{R}(u_N^{n+1}) := \sup_{\substack{w\in X\\ \|w\|_X=1}} \left\{ (\psi_N(u_N^{n+1}), w)_X - \left\langle \mathsf{F}(u_N^{n+1}), w \right\rangle \right\}.$$

Noticing (32), for any $w_N \in X_N$, we have

$$\mathsf{R}(u_N^{n+1}) := \sup_{\substack{w\in X\\ \|w\|_X=1}} \left\{ (\psi_N(u_N^{n+1}), w - w_N)_X - \left\langle \mathsf{F}(u_N^{n+1}), w - w_N \right\rangle \right\}.$$

Then, for $g \in L^2(\Omega)$ in (39), and $w_N \in X_N$ an appropriate quasi-interpolant of $w \in H_0^1(\Omega)$, we employ a standard residual-based a posteriori error analysis (see, e.g., [29]) to infer the upper bound

$$\mathsf{R}(u_N^{n+1}) \leq C_{\mathsf{I}} \left(\sum_{K\in\mathcal{T}_N} \eta_K^2 \right)^{1/2},$$

where C_{I} is a quasi-interpolation constant, and

$$\begin{aligned} \eta_K^2 = h_K^2 \left\| \Delta\psi_N(u_N^{n+1}) + g + \nabla \cdot \left\{ \mu \left(\left| \nabla u_N^{n+1} \right|^2 \right) \nabla u_N^{n+1} \right\} \right\|_{L^2(K)}^2 \\ + \frac{1}{2} h_K \left\| \left[\!\!\left[\nabla\psi_N(u_N^{n+1}) + \mu \left(\left| \nabla u_N^{n+1} \right|^2 \right) \nabla u_N^{n+1} \right]\!\!\right] \right\|_{L^2(\partial K\setminus\Gamma)}^2, \end{aligned} \tag{56}$$

for any $K \in \mathcal{T}_N$. Then, invoking Theorem 3.4 and recalling (42), we obtain the following result.

Theorem 5.6. *Given the same assumptions as in Theorem 5.5, then it holds the a posteriori error estimate*

$$\left\|u^\star - u_N^{n+1}\right\|_X \leq \frac{C_{\mathrm{I}}}{m_\mu}\left(\sum_{K\in\mathcal{T}_N}\eta_K^2\right)^{1/2} + \frac{1}{m_\mu}\left\|\psi_N(u_N^{n+1})\right\|_{L^2(\Omega)},$$

where $u^\star$ is the unique solution of (39), C_{I} is a constant, and η_K, for $K\in\mathcal{T}_N$, is given in (56).

5.3. **Numerical Experiments.** In this section, we test the adaptive ILG Algorithm 1 in the context of the iterative linearized FEM for second-order PDE in divergence form discussed in Section 5. We perform a series of numerical experiments to compare the various iterative linearization procedures from Section 2.3 and to validate the *a posteriori* error estimators from Section 5.2. For all our experiments, we consider the L-shaped domain $\Omega = (-1,1)^2\setminus([0,1]\times[-1,0])$, and an initial mesh consisting of 192 uniform triangles. Moreover, we will always choose the initial guess to be $u^0\equiv 0$, and run the algorithm until the number of elements exceeds 10^6. On a given mesh, we perform at least one iterative linearization step, and continue until the linearization error is at most half as large as the discretization error, i.e. we let $\vartheta = 2$ in Algorithm 1. Furthermore, for a given constant $\Upsilon > 0$, we let

$$\sigma(N) := \Upsilon\,|\mathcal{T}_N|^{-1/2}\left\|u_0^1\right\|_X, \qquad N\geq 0,$$

which relates to the expected convergence rate of $\mathcal{O}(|\mathcal{T}_N|^{-1/2})$; in our experiments below the choice $\Upsilon = 10$ has proved to be a sensible value. Moreover, we set the constant factors for the discretization and linearization estimators appearing in the right-hand sides of the a posteriori error bounds to 1 (cf. Theorems 5.5 and 5.6). In the adaptive process, we mark the elements for refinement by use of the Dörfler marking strategy, see [13], and process them by the newest vertex bisection method, see [24]. The true error $\|u^\star - u_N^n\|_X$ and the error estimator will be displayed each time before a mesh refinement is undertaken. Our implementation is based on the Matlab package [16], with the necessary modifications.

In the Experiments 5.3.1–5.3.3 below we consider the different iterative procedures discussed in Section 2.3. For the problems under consideration, our computations consistently indicate that, in the a posteriori error estimates from Theorem 5.5 and Theorem 5.6, the discretization part clearly dominantes the linearization contribution. Not surprisingly, after a brief initial mesh refinement phase, the algorithm only undertakes one iterative linearization step per space enrichment, i.e. our algorithm is highly efficient for the proposed examples. Moreover, both the discretization and linearization error indicators generally converge at the expected rate of $\mathcal{O}(|\mathcal{T}_N|^{-1/2})$. More precisely, this holds true for any iterative scheme except for the damped Newton method (in combination with the a posteriori error estimator from Theorem 5.6), where the linearization error estimator exhibits an even higher convergence rate; this may result from the local quadratic convergence property of the Newton iteration.

5.3.1. *Smooth solution.* We consider the nonlinear diffusion coefficient $\mu(t) = (t+1)^{-1} + 1/2$, for $t\geq 0$, and select g in (39) such that the analytical solution of (43) is given by the smooth function $u^\star(x,y) = \sin(\pi x)\sin(\pi y)$. It is straightforward to verify that μ fulfills the requirements (40) and (45) from Section 5, so that the convergence of the three iterative procedures from Section 2.3 is guaranteed. The parameter δ in the Zarantonello iteration (12) is chosen to be 0.85 as this seems to be close to optimal. The initial damping parameter on the initial mesh for the damped Newton method is chosen to be $\delta^0 = 1$ in Remark 2.8; moreover, throughout all our experiments, the factor κ for the correction and prediction strategy of the damping parameter is set to be $1/2$.

In Figure 1, for each of the three iterative linearization schemes presented in Section 5.1, we plot the error $\|u^\star - u_N^n\|_X$ and both error estimators from Theorems 5.5 and 5.6 against the number $|\mathcal{T}_N|$ of elements in the mesh. In addition, we display the effectivity indices for each experiment, i.e. the ratio of the error estimator and the true error; we see that they are roughly bounded between 2 and 4. Furthermore, we notice that (nearly) optimal convergence rates $\mathcal{O}\left(|\mathcal{T}_N|^{-1/2}\right)$ are achieved in all plots.

5.3.2. *Nonsmooth solution.* In our second experiment, we consider the nonlinear diffusion parameter $\mu(t) = 1+\mathrm{e}^{-t}$, for $t \geq 0$. Again, it is easily seen that μ satisfies the assumptions (40) and (45). We choose g in (39) such that the analytical solution is given by

$$u^\star(r,\varphi) = r^{2/3}\sin\left(2\varphi/3\right)(1 - r\cos(\varphi))(1 + r\cos(\varphi))(1 - r\sin(\varphi))(1 + r\sin(\varphi))\cos(\varphi), \qquad (57)$$

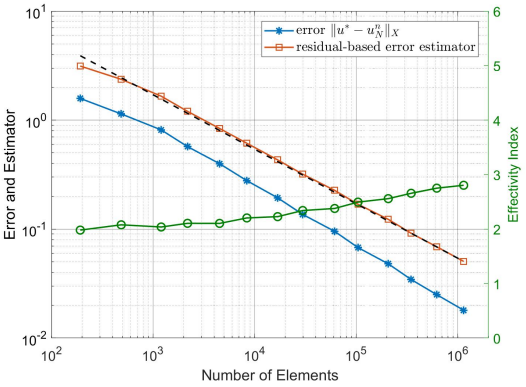
where r and φ are polar coordinates. This is the prototype singularity for (linear) second-order elliptic problems with homogeneous Dirichlet boundary conditions in the L-shaped domain; in particular, we note that the gradient of $u^\star$ is unbounded at the origin. As before, in Figure 2, we plot the error $\|u^\star - u_N^n\|_X$, the error estimators from Theorems 5.5 and 5.6, as well as the effectivity indices versus the number $|\mathcal{T}_N|$ of elements in the mesh for each of the three iterative linearization schemes from Section 5.2. We let $\delta = 0.5$ for the Zarantonello iteration, and use the initial damping parameter $\delta^0 = 1$ for the Newton method as in Experiment 5.3.1. As before, we observe that optimal rates of convergence are attained in all six cases.

5.3.3. *Nonsmooth solution with monotone increasing diffusion.* Finally, we consider the nonlinear diffusivity function $\mu(t) = 2 - \mathrm{e}^{-t}$, for $t \geq 0$. Again, we choose g in (39) such that the analytical solution is given by the nonsmooth function (57). Since μ is monotone increasing, it does not have the property (45), which is needed to guarantee the convergence of the Kačanov iteration and of the damped Newton method. It still fulfills, however, the assumption (40), which, in turn, is sufficient to guarantee the convergence of the Zarantonello method. In this experiment, we choose the damping parameter for the Zarantonello method to be $\delta = 0.4$, and the initial damping parameter in the Newton method to be $\delta^0 = 1$. We see from the plots in Figure 3 that the Kačanov and damped Newton methods converge, even with optimal order, which indicates that the property (45) does not seem to be necessary for the current example and the initial setup chosen here. We emphasize that this observation for the Kačanov method was already made in [18].

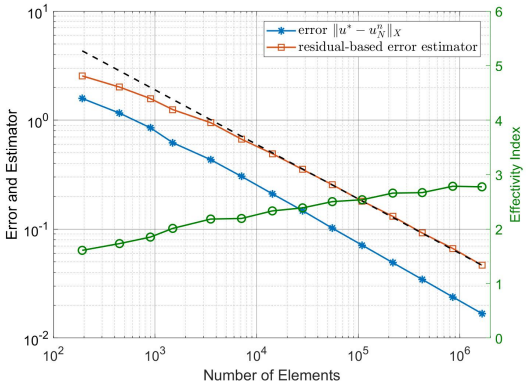
References

1. M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Series in Computational and Applied Mathematics, Elsevier, 1996.
2. M. Amrein, J. M. Melenk, and T. P. Wihler, *An hp-adaptive Newton-Galerkin finite element procedure for semilinear boundary value problems*, Math. Methods Appl. Sci. **40** (2017), no. 6, 1973–1985.
3. M. Amrein and T. P. Wihler, *An adaptive Newton-method based on a dynamical systems approach*, Commun. Nonlinear Sci. Numer. Simul. **19** (2014), no. 9, 2958–2973.
4. ———, *Fully adaptive Newton-Galerkin methods for semilinear elliptic partial differential equations*, SIAM J. Sci. Comput. **37** (2015), no. 4, A1637–A1657.
5. M. Amrein and T. P. Wihler, *An adaptive space-time Newton-Galerkin approach for semilinear singularly perturbed parabolic evolution equations*, IMA J. Numer. Anal. **37** (2017), no. 4, 2004–2019.
6. K. Astala, T. Iwaniec, and G. Martin, *Elliptic partial differential equations and quasiconformal mappings in the plane*, Princeton Mathematical Series, vol. 48, Princeton University Press, Princeton, NJ, 2009.
7. I. Ben Gharbia, J. Dabaghi, V. Martin, and M. Vohralík, *A posteriori error estimates and adaptive stopping criteria for a compositional two-phase flow with nonlinear complementarity constraints*, working paper or preprint, May 2019.
8. C. Bernardi, J. Dakroub, G. Mansour, and T. Sayah, *A posteriori analysis of iterative algorithms for a nonlinear problem*, J. Sci. Comput. **65** (2015), no. 2, 672–697.
9. F. E. Browder, *Remarks on nonlinear functional equations. II, III*, Illinois J. Math. 9 (1965) 608–616; ibid. **9** (1965), 617–622.
10. A. L. Chaillou and M. Suri, *A posteriori estimation of the linearization error for strongly monotone nonlinear operators*, J. Comput. Appl. Math. **205** (2007), no. 1, 72–87.
11. S. Congreve and T. P. Wihler, *Iterative Galerkin discretizations for strongly monotone problems*, Journal of Computational and Applied Mathematics **311** (2017), 457–472.
12. P. Deuflhard, *Newton methods for nonlinear problems*, Springer Series in Computational Mathematics, vol. 35, Springer-Verlag, Berlin, 2004, Affine invariance and adaptive algorithms.
13. W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SINUM **33** (1996), 1106–1124.
14. L. El Alaoui, A. Ern, and 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** (2011), no. 37-40, 2782–2795.
15. A. Ern and M. Vohralík, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*, SIAM J. Sci. Comput. **35** (2013), no. 4, A1761–A1791.

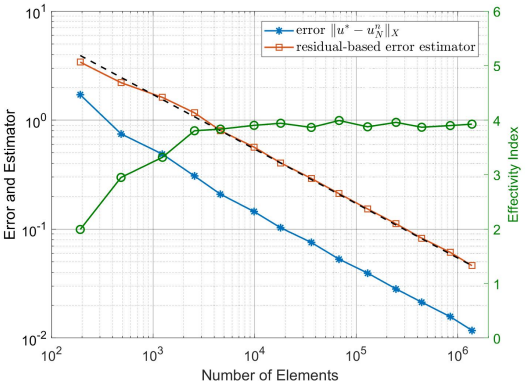
16. S. Funken, D. Praetorius, and P. Wissgott, *Efficient implementation of adaptive P1-FEM in Matlab*, Computational Methods in Applied Mathematics **11** (2011), no. 4, 460–490. MR 2875100
17. G. Gantner, A. Haberl, D. Praetorius, and B. Stiftner, *Rate optimal adaptive FEM with inexact solver for nonlinear operators*, IMA Journal of Numerical Analysis **38** (2018), no. 4, 1797–1831.
18. E. M. Garau, P. Morin, and C. Zuppa, *Convergence of an adaptive Kačanov FEM for quasi-linear problems*, Appl. Numer. Math. **61** (2011), no. 4, 512–529.
19. W. Han, S. Jensen, and I. Shimansky, *The Kačanov method for some nonlinear problems*, Appl. Numer. Meth. **24** (1997), 57–79.
20. P. Heid and T.P. Wihler, *On the convergence of adaptive iterative linearized Galerkin methods*, Tech. Report 1905.06682, arxiv.org, 2019.
21. P. Houston and T. P. Wihler, *An hp-adaptive newton-discontinuous-galerkin finite element approach for semilinear elliptic boundary value problems*, Math. Comp. **87** (2018), no. 314, 2641–2674.
22. O. Lakkis and C. Makridakis, *Elliptic reconstruction and a posteriori error estimates for fully discrete linear parabolic problems*, Math. Comp. **75** (2006), no. 256, 1627–1658.
23. C. Makridakis and R. H. Nochetto, *Elliptic reconstruction and a posteriori error estimates for parabolic problems*, SIAM Journal on Numerical Analysis **41** (2003), no. 4, 1585–1594.
24. W.F. Mitchell, *Adaptive refinement for arbitrary finite-element spaces with hierarchical basis*, J. Comput. Appl. Math. **36** (1991), 65–78.
25. J. Nečas, *Introduction to the theory of nonlinear elliptic equations*, John Wiley and Sons, 1986.
26. A. Potschka, *Backward step control for global newton-type methods*, SIAM J. Numer. Anal. **54** (2016), no. 1, 361–387.
27. W. Rudin, *Real and complex analysis*, third ed., McGraw-Hill Book Co., New York, 1987.
28. H. R. Schneebeli and T. P. Wihler, *The Newton-Raphson method and adaptive ODE solvers*, Fractals. Complex Geometry, Patterns, and Scaling in Nature and Society **19** (2011), no. 1, 87–99.
29. R. Verfürth, *A posteriori error estimation techniques for finite element methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.
30. E. H. Zarantonello, *Solving functional equations by contractive averaging*, Tech. Report 160, Mathematics Research Center, Madison, WI, 1960.
31. E. Zeidler, *Nonlinear functional analysis and its applications. I*, Springer-Verlag, New York, 1986, Fixed-point theorems.
32. ———, *Nonlinear functional analysis and its applications. IV*, Springer-Verlag, New York, 1988, Applications to mathematical physics, Translated from the German and with a preface by Juergen Quandt.
33. ———, *Nonlinear functional analysis and its applications. II/B*, Springer-Verlag, New York, 1990.



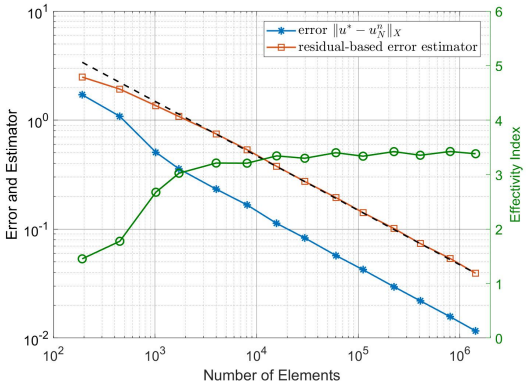
(A) Zarantonello iteration with the a posteriori error bound from Theorem 5.5.



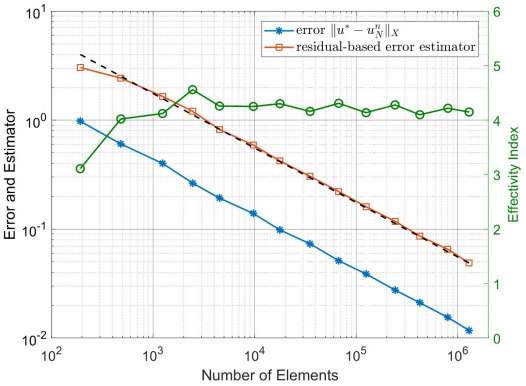
(B) Zarantonello iteration with the a posteriori error bound from Theorem 5.6.



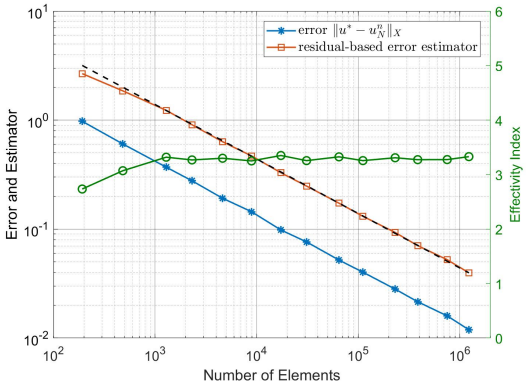
(C) Kačanov iteration with the a posteriori error bound from Theorem 5.5.



(D) Kačanov iteration with the a posteriori error bound from Theorem 5.6.

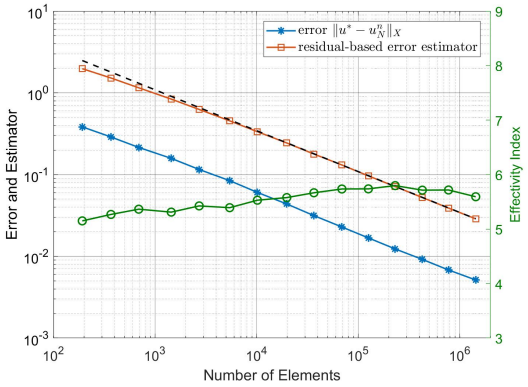


(E) Damped Newton iteration with the a posteriori error bound from Theorem 5.5.

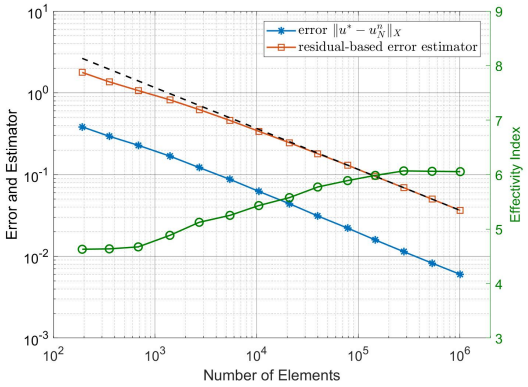


(F) Damped Newton iteration with the a posteriori error bound from Theorem 5.6.

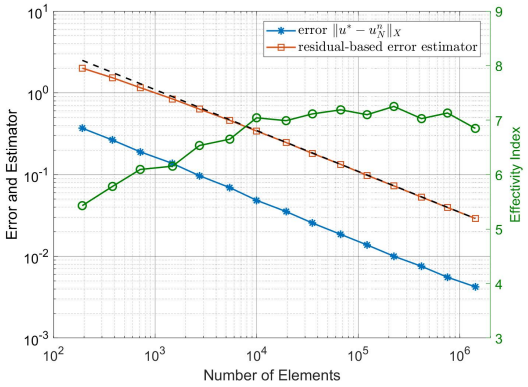
FIGURE 1. Experiment 5.3.1: Performance data for the error estimators from Theorem 5.5 (left) and Theorem 5.6 (right) for the Zarantonello, Kačanov and Newton iterations.



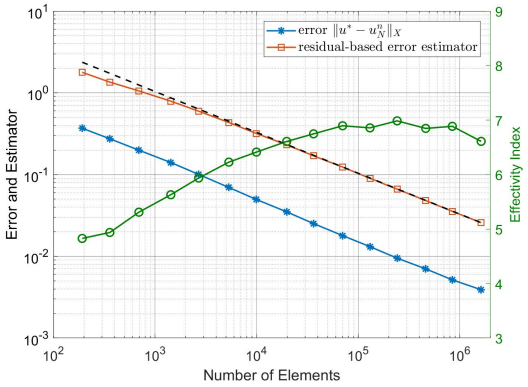
(A) Zarantonello iteration with the a posteriori error bound from Theorem 5.5.



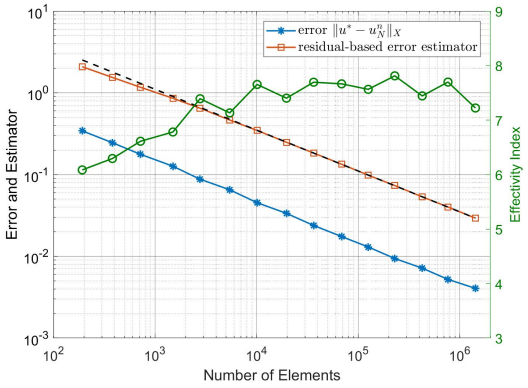
(B) Zarantonello iteration with the a posteriori error bound from Theorem 5.6.



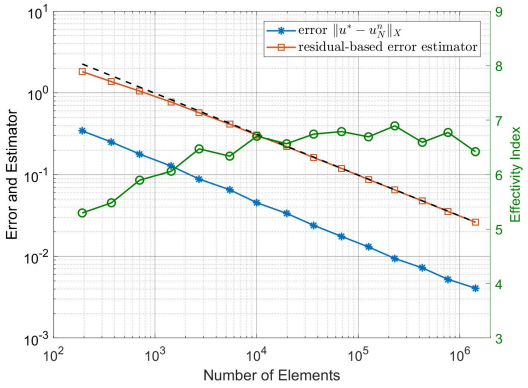
(C) Kačanov iteration with the a posteriori error bound from Theorem 5.5.



(D) Kačanov iteration with the a posteriori error bound from Theorem 5.6.

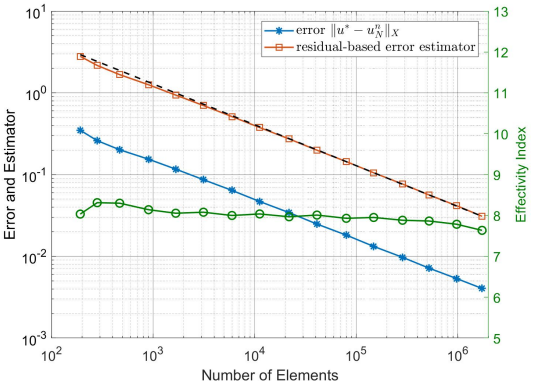


(E) Damped Newton iteration with the a posteriori error bound from Theorem 5.5.

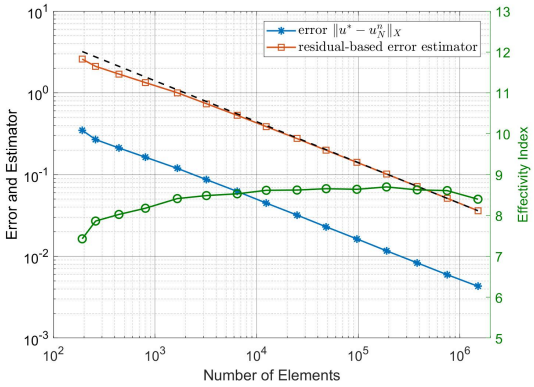


(F) Damped Newton iteration with the a posteriori error bound from Theorem 5.6.

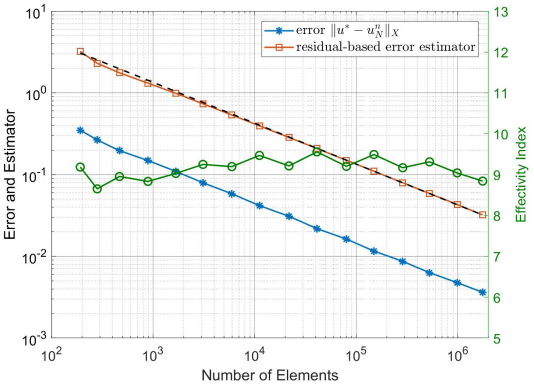
FIGURE 2. Experiment 5.3.2: Performance data for the error estimators from Theorem 5.5 (left) and Theorem 5.6 (right) for the Zarantonello, Kačanov and Newton iterations.



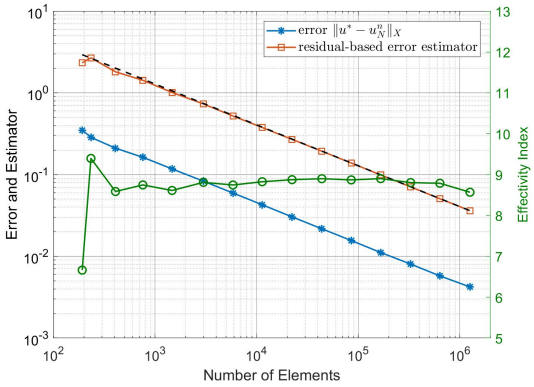
(A) Zarantonello iteration with the a posteriori error bound from Theorem 5.5.



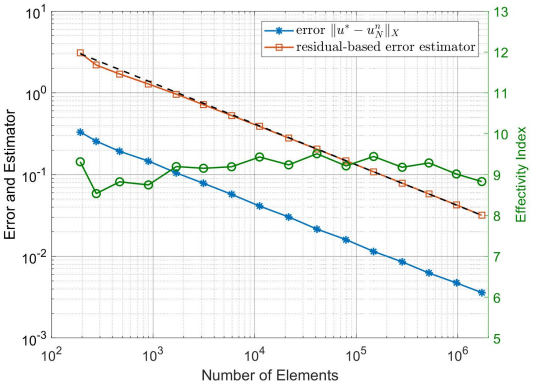
(B) Zarantonello iteration with the a posteriori error bound from Theorem 5.6.



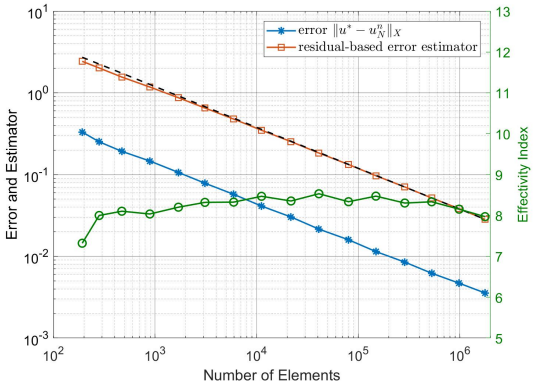
(C) Kačanov iteration with the a posteriori error bound from Theorem 5.5.



(D) Kačanov iteration with the a posteriori error bound from Theorem 5.6.



(E) Damped Newton iteration with the a posteriori error bound from Theorem 5.5.



(F) Damped Newton iteration with the a posteriori error bound from Theorem 5.6.

FIGURE 3. Experiment 5.3.3: Performance data for the error estimators from Theorem 5.5 (left) and Theorem 5.6 (right) for the Zarantonello, Kačanov and Newton iterations.