Adaptive inexact Newton methods for discretizations of nonlinear diffusion PDEs. I. General theory and a posteriori stopping criteria

Alexandre Ern, Martin Vohralík

Adaptive inexact Newton methods for discretizations of nonlinear diffusion PDEs. I. General theory and a posteriori stopping criteria*

Alexandre Ern* Martin Vohralík†

March 21, 2012

Abstract

We consider nonlinear algebraic systems resulting from numerical discretizations of nonlinear partial differential equations of diffusion type. To solve these systems, some iterative nonlinear solver, and, on each step of this solver, some iterative linear solver are used. In this first part, we derive adaptive stopping criteria for both iterative solvers. Both criteria are based on an a posteriori error estimate which distinguishes the different error components, namely the discretization error, the linearization error, and the algebraic error. We stop the iterations whenever the corresponding error does no longer affect the overall error significantly. Our estimates also yield a guaranteed upper bound on the overall error at each step of the nonlinear and linear solvers. We prove the (local) efficiency and robustness of the estimates with respect to the size of the nonlinearity owing, in particular, to the error measure involving the dual norm of the residual. Our developments are carried at an abstract level, yielding a general framework. We show how to apply this framework to the Crouzeix–Raviart nonconforming finite element discretization, Newton linearization, and conjugate gradient algebraic solution, and we illustrate on numerical experiments for the p-Laplacian the tight overall error control and important computational savings achieved in our approach. Part II is devoted to the application of our abstract framework to a broad class of discretization methods.

Key words: nonlinear diffusion PDE, nonlinear algebraic system, adaptive linearization, adaptive algebraic solution, adaptive mesh refinement, stopping criterion, a posteriori error estimate

1 Introduction

Consider a system of nonlinear algebraic equations written in the form: find a vector $U \in \mathbb{R}^N$, $N \geq 1$, such that

$$\mathcal{A}(U) = F, \tag{1.1}$$

where $\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^N$ is a discrete nonlinear operator and $F \in \mathbb{R}^N$ a given vector. We consider the following *inexact iterative linearization* of problem (1.1):

Algorithm 1.1 (Inexact iterative linearization of (1.1)).

1. *Choose an initial vector $U^0 \in \mathbb{R}^N$. Set $k := 1$.*

2. *From U^{k-1}, define a matrix $\mathbb{A}^k \in \mathbb{R}^{N,N}$ and a vector $F^k \in \mathbb{R}^N$. Consider the following system of linear algebraic equations:*
$$\mathbb{A}^k U^k = F^k. \tag{1.2}$$

3. *(a) Define $U^{k,0} := U^{k-1}$ and set $i := 1$.*

*This work was partly supported by the Groupement MoMaS (PACEN/CNRS, ANDRA, BRGM, CEA, EdF, IRSN) and by the ERT project "Enhanced oil recovery and geological sequestration of CO_2: mesh adaptivity, a posteriori error control, and other advanced techniques" (LJLL/IFPEN).

*Université Paris-Est, CERMICS, Ecole des Ponts ParisTech, 77455 Marne la Vallée cedex 2, France (ern@cermics.enpc.fr).

†UPMC Univ. Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, 75005, Paris, France & CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, 75005, Paris, France (vohralik@ann.jussieu.fr).

(b) *Perform a step of an iterative linear solver for the solution of the linear system (1.2), starting from the vector $U^{k,i-1}$. This yields an approximation $U^{k,i}$ to U^k which satisfies*

$$\mathbb{A}^k U^{k,i} = F^k - R^{k,i}, \tag{1.3}$$

where $R^{k,i} \in \mathbb{R}^N$ is the algebraic residual vector on step i.

(c) *Check the convergence criterion for the linear solver, i.e., check whether $U^{k,i}$ is sufficiently close to U^k, the solution of (1.2). If yes, set $U^k := U^{k,i}$. If not, set $i := i + 1$ and go back to step 3b.*

4. *Check the convergence criterion for the nonlinear solver, i.e., check whether U^k is sufficiently close to U, the solution of (1.1). If yes, finish. If not, set $k := k + 1$ and go back to step 2.*

If the criterion in Step 3c of Algorithm 1.1 is set to zero or "close" to zero, typically of the order of computer working precision, an *exact iterative linearization* is obtained. Probably the most well-known example is the Newton method; therein

$$\mathbb{A}^k_{ij} := \frac{\partial \mathcal{A}_i}{\partial U_j}(U^{k-1}), \qquad F^k := F - \mathcal{A}(U^{k-1}) + \mathbb{A}^k U^{k-1}. \tag{1.4}$$

Convergence and a priori error estimates for the Newton method have been obtained by Kantorovich [19], Kantorovich and Akilov [20], and Ortega [26]. A posteriori error estimates, that is, fully computable quantities yielding an upper bound on the error $\|U^k - U\|$ between U^k, the solution of (1.2), and U, the solution of (1.1), have been proved by Gragg and Tapia [16] and improved by Potra and Pták [27] and Yamamoto [32], see also the references therein.

The Newton method can be computationally demanding, since, on each step, the linear system (1.2) needs to be solved "exactly". The *inexact Newton method* is a popular approach to speed-up the original Newton method. It has been used in practice for decades and studied theoretically in many papers. In particular, Eisenstat and Walker [12, 13] have shown the convergence, a posteriori error estimates were proved by Moret [25], and adaptive algorithms were derived by Deuflhard [10, Section 1.2.3], see also the references therein. Furthermore, in the context of nonlinear inverse problems, the convergence of inexact Newton methods has been recently proved in a rather general framework by Lechleiter and Rieder [21], and a posteriori error estimates have been obtained, e.g., by Bauer and Hohage [2] and Bakushinsky and Smirnova [1], see also the references therein.

(Inexact) iterative linearization methods are typically understood and studied as methods for the solution of systems of *general nonlinear algebraic equations* of the form (1.1), without much (any) specification of their structure and origin. In this work, we pursue a conceptually different approach, in that we investigate nonlinear algebraic systems *originating* from a given *discretization* of a given *partial differential equation* (PDE). We write the PDE in the following abstract form: given a nonlinear operator A, find a function u such that

$$A(u) = f. \tag{1.5}$$

The nonlinear algebraic system (1.1) then stems from some discretization of (1.5).

Our first goal is to derive *stopping criteria* to be used in steps 3c and 4 of Algorithm 1.1. Let u be the solution of (1.5) and let $u_h^{k,i}$ be the approximation to the solution u obtained by the given discretization scheme on the k-th nonlinear solver step and the i-th linear solver step in Algorithm 1.1, whose algebraic representation is the vector $U^{k,i}$. Our second goal is to obtain *a posteriori estimates* for the error between u and $u_h^{k,i}$. We carry this task for a broad class of nonlinear PDEs of the form (1.5); details are given in Section 2. The iterative nonlinear and linear solvers need not be specified in our setting; for simplicity, we refer to Algorithm 1.1 as *adaptive inexact Newton method*.

A posteriori error estimates for the error between the exact solution u and an approximate solution u_h in the absence of errors stemming from the iterative nonlinear and linear solvers have been obtained in the literature for various discretization schemes; we postpone their discussion to Part II. We are not aware of estimates of the error between u and $u_h^{k,i}$ which provide, at the same time, a *guaranteed* (without undetermined constants) *upper bound* and a *distinction* among the *different error components*, namely *discretization*, *linearization*, and *algebraic* errors. We achieve such a result in Section 3 of this paper. In turn, this leads to stopping criteria expressing that there is no need to continue with the algebraic solver

iterations once the linearization or discretization error components start to dominate, and that there is no need to continue with the nonlinear solver iterations once the discretization error component starts to dominate. This approach is in line with those of Becker *et al.* [3] and Chaillou and Suri [6, 7], see also the references therein, and, more closely, with [17, 14]. In these works, estimates and stopping criteria are derived *independently* for linear and nonlinear solvers. Similar ideas have been developed in the context of goal-oriented error estimation by Rannacher *et al.* [29] and Meidner *et al.* [24].

The three error components are estimated through three suitable *flux reconstructions*. To handle algebraic solver errors, we consider *quasi-equilibration* of the total flux with the data, instead of the more usual exact equilibration. Specifically, flux equilibration holds up to an algebraic remainder whose size can be controlled adaptively. A simple and effective way to achieve such a control is described in Section 4. Following [17, Section 7.2], see also the references therein, it consists in performing, on a given algebraic solver step i, a few additional algebraic solver iterations.

A further important result is the *efficiency* of the estimators, answering the question whether the estimators are also a lower bound for the error, possibly up to a generic constant. Whenever such a constant is independent of the nonlinear operator at hand, the approximate and exact solutions, the mesh size, and the computational domain, we speak of *robustness*. We use an error measure based on the dual norm of the residual for conforming discretizations as in [6, 7, 14] which we augment by a jump seminorm in the nonconforming case. We show in Section 5 that, under the above-discussed stopping criteria and for this error measure, our estimates are efficient and robust. Moreover, when a *local*, elementwise version of the *stopping criteria* is used, we obtain this efficiency locally around each mesh element for a slightly different error measure.

The developments of Section 3, Section 4, and Section 5 constitute a general framework which is built on a couple of clearly identified assumptions on the flux reconstructions. We apply this framework to a broad class of discretization schemes and solvers by verifying these assumptions in Part II. We close Part I by exemplifying in Section 6 the flux reconstructions for the p-Laplacian, the Crouzeix–Raviart nonconforming finite element method, and either Newton or fixed point linearizations, while, in Section 7, we study numerically the behavior of our a posteriori estimates and the computational gains of our stopping criteria, in conjunction with the conjugate gradient algebraic solver.

2 Setting

This section describes the continuous problem, sets up the basic notation, and introduces the error measure.

2.1 Continuous problem

Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a polygonal (polyhedral) domain (open, bounded, and connected set). We consider the following model nonlinear diffusion problem: find $u : \Omega \to \mathbb{R}$ such that

$$-\nabla\cdot\boldsymbol{\sigma}(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x})) = f \qquad \text{in } \Omega, \tag{2.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{2.1b}$$

where $\boldsymbol{\sigma} : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is the nonlinear flux function and $f : \Omega \to \mathbb{R}$ the source term. The scalar-valued unknown function u is termed the *potential*, and, given a potential u, the vector-valued function $-\boldsymbol{\sigma}(\cdot, u, \nabla u) : \Omega \to \mathbb{R}^d$ is termed the *flux*.

The nonlinear flux function $\boldsymbol{\sigma}$ takes the form

$$\boldsymbol{\sigma}(\mathbf{x}, v, \boldsymbol{\xi}) = \underline{\mathbf{A}}(\mathbf{x}, v, \boldsymbol{\xi})\boldsymbol{\xi} \qquad \forall (\mathbf{x}, v, \boldsymbol{\xi}) \in \Omega \times \mathbb{R} \times \mathbb{R}^d, \tag{2.2}$$

where $\underline{\mathbf{A}} : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a Carathéodory (tensor-valued) function (measurable in $\mathbf{x}$ and continuous in v and $\boldsymbol{\xi}$). Two key examples are the *quasi-linear diffusion* problem in which $\underline{\mathbf{A}}$ is independent of $\boldsymbol{\xi}$ (so that $\boldsymbol{\sigma}$ depends linearly on $\boldsymbol{\xi}$) yielding

$$\boldsymbol{\sigma}(\mathbf{x}, v, \boldsymbol{\xi}) = \underline{\mathbf{A}}(\mathbf{x}, v)\boldsymbol{\xi} \qquad \forall (\mathbf{x}, v, \boldsymbol{\xi}) \in \Omega \times \mathbb{R} \times \mathbb{R}^d, \tag{2.3}$$

and the *Leray–Lions* problem in which $\underline{\mathbf{A}}$ depends on $\boldsymbol{\xi}$ (so that $\boldsymbol{\sigma}$ depends nonlinearly on $\boldsymbol{\xi}$), but is independent of v, yielding

$$\boldsymbol{\sigma}(\mathbf{x}, \boldsymbol{\xi}) = \underline{\mathbf{A}}(\mathbf{x}, \boldsymbol{\xi})\boldsymbol{\xi} \qquad \forall (\mathbf{x}, \boldsymbol{\xi}) \in \Omega \times \mathbb{R}^d. \tag{2.4}$$

For the quasi-linear diffusion problem, we assume that $\underline{\mathbf{A}}$ is bounded and that it takes symmetric values with minimal eigenvalue uniformly bounded away from zero. For the Leray–Lions problem, see [22], we assume that, for a real number $p > 1$, there holds, for all $\boldsymbol{\xi}, \boldsymbol{\zeta} \in \mathbb{R}^d$ and a.e. $\mathbf{x} \in \Omega$, $\boldsymbol{\sigma}(\mathbf{x}, \boldsymbol{\xi}) \cdot \boldsymbol{\xi} \geq \alpha_0 |\boldsymbol{\xi}|^p$, $(\boldsymbol{\sigma}(\mathbf{x}, \boldsymbol{\xi}) - \boldsymbol{\sigma}(\mathbf{x}, \boldsymbol{\zeta})) \cdot (\boldsymbol{\xi} - \boldsymbol{\zeta}) > 0$ for $\boldsymbol{\xi} \neq \boldsymbol{\zeta}$, and $|\boldsymbol{\sigma}(\mathbf{x}, \boldsymbol{\xi})| \leq g(\mathbf{x}) + \alpha_1 |\boldsymbol{\xi}|^{p-1}$ for positive real numbers α_0 and α_1 and a function $g \in L^q(\Omega)$ where $q := \frac{p}{p-1}$, so that $\frac{1}{p} + \frac{1}{q} = 1$. A typical Leray–Lions problem is the p-Laplacian where $\underline{\mathbf{A}}(\mathbf{x}, \boldsymbol{\xi}) = |\boldsymbol{\xi}|^{p-2} \underline{\mathbf{I}}$ and $\underline{\mathbf{I}}$ is the identity tensor.

To alleviate the notation, we leave the dependence on the space variable $\mathbf{x}$ implicit, so that we simply write $\boldsymbol{\sigma}(u, \nabla u)$. To allow for a unified presentation of the quasi-linear diffusion and Leray–Lions settings, we set $p := 2$ for the quasi-linear diffusion problem, while, for the Leray–Lions problem, the real number p results from the above assumptions. Then, we seek in both cases the potential u in the energy space $V := W_0^{1,p}(\Omega)$ (that is, the space of $L^p(\Omega)$ functions whose weak derivatives are in $L^p(\Omega)$ and with zero trace on $\partial\Omega$). Assuming $f \in L^q(\Omega)$, the model problem (2.1) can be written in the form (1.5) as follows: find $u \in V$ such that

$$(\boldsymbol{\sigma}(u, \nabla u), \nabla v) = (f, v) \qquad \forall v \in V. \tag{2.5}$$

For $w \in L^q(\Omega)$, $v \in L^p(\Omega)$, (w, v) stands for $\int_\Omega w(\mathbf{x}) v(\mathbf{x}) \, \mathrm{d}\mathbf{x}$ and similarly in the vector-valued case. Owing to the above assumptions and to (2.5), the flux $-\boldsymbol{\sigma}(u, \nabla u)$ is in the space $\mathbf{H}^q(\mathrm{div}, \Omega)$ spanned by the functions in $[L^q(\Omega)]^d$ with weak divergence in $L^q(\Omega)$. We assume that there exists a unique weak solution of (2.5).

2.2 Discrete setting

Let $\mathcal{T}_h$ be a simplicial mesh of Ω. For simplicity, we suppose that there are no hanging nodes in the sense that, for two distinct elements of $\mathcal{T}_h$, their intersection is either an empty set or a common l-dimensional face, $0 \leq l \leq d-1$. A generic element of $\mathcal{T}_h$ is denoted by K and its diameter by h_K. The $(d-1)$-dimensional faces of the mesh are collected in the set $\mathcal{E}_h$ such that $\mathcal{E}_h = \mathcal{E}_h^{\mathrm{int}} \cup \mathcal{E}_h^{\mathrm{ext}}$, with $\mathcal{E}_h^{\mathrm{int}}$ collecting interfaces and $\mathcal{E}_h^{\mathrm{ext}}$ boundary faces. A generic face is denoted by e and its diameter by h_e. The faces of an element K are collected in the set $\mathcal{E}_K$.

Discretizing problem (2.1) leads to a nonlinear algebraic system of the form (1.1). Let some nonlinear and linear solvers be applied to problem (1.1) via Algorithm 1.1. Suppose that we are on step k, $k \geq 1$, of the nonlinear solver and on step i, $i \geq 1$, of the linear solver. This corresponds to problem (1.3). We denote by $u_h^{k,i}$ the discrete potential associated with the vector $U^{k,i}$. Our framework covers both conforming schemes, where $u_h^{k,i} \in V$, and nonconforming schemes, where $u_h^{k,i} \notin V$. To proceed generally, we assume that $u_h^{k,i}$ is in the broken Sobolev space

$$V(\mathcal{T}_h) := \{v \in L^p(\Omega),\, v|_K \in W^{1,p}(K) \quad \forall K \in \mathcal{T}_h\}. \tag{2.6}$$

In what follows, for a function $v \in V(\mathcal{T}_h)$, ∇v denotes its so-called *broken gradient*, that is, the distributional gradient evaluated elementwise. As functions in $V(\mathcal{T}_h)$ are not necessarily single-valued at interfaces, we introduce the jump operator $[\![\cdot]\!]$ yielding the difference of (the traces of) the argument from the two mesh elements that share e on interfaces and the actual trace if e is a boundary face. Classically, $v \in V(\mathcal{T}_h)$ is in V if and only if $[\![v]\!] = 0$ for all $e \in \mathcal{E}_h$, see, e.g., [11, Lemma 1.23].

2.3 Error measure

The error between the exact solution u of (2.5) and the approximate solution $u_h^{k,i}$ is measured as

$$\mathcal{J}_u(u_h^{k,i}) = \mathcal{J}_{u,\mathrm{F}}(u_h^{k,i}) + \mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}), \tag{2.7}$$

where

$$\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i}) := \sup_{\varphi \in V;\, \|\nabla\varphi\|_p = 1} \left(\boldsymbol{\sigma}(u, \nabla u) - \boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla \varphi \right), \tag{2.8a}$$

$$\mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}) := \left\{ \sum_{K \in \mathcal{T}_h} \sum_{e \in \mathcal{E}_K} \alpha_e^s h_e^{1-s} \| [\![u - u_h^{k,i}]\!] \|_{s,e}^s \right\}^{\frac{1}{q}}. \tag{2.8b}$$

The quantity $\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i})$ measures the error in the fluxes and represents the dual norm of the residual of (2.5). This error measure has been considered by Chaillou and Suri [6, 7] and in [14] for conforming discretizations. Owing to the well-posedness of (2.5), whenever $u_h^{k,i} \in V$, $\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i}) = 0$ if and only if $u_h^{k,i} = u$. Furthermore, the quantity $\mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i})$ measures the nonconformity of the discrete potential, i.e., the departure of $u_h^{k,i}$ from the space V. The actual value of the weights $\alpha_e > 0$ and of the exponent $s \geq 1$ is irrelevant in Part I; we only use that $\mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}) = 0$ if and only if $u_h^{k,i} \in V$. All in all, we see that $\mathcal{J}_u(u_h^{k,i}) = 0$ if and only if $u_h^{k,i} = u$.

Although the quantity $\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i})$ is a dual norm and as such is not easily computable (assuming u known), the Hölder inequality yields

$$\mathcal{J}_u(u_h^{k,i}) \leq \mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}) := \|\boldsymbol{\sigma}(u,\nabla u) - \boldsymbol{\sigma}(u_h^{k,i},\nabla u_h^{k,i})\|_q + \mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}), \tag{2.9}$$

which features the $[L^q(\Omega)]^d$-*difference* of the *fluxes*. Our numerical experiments in Section 7 indicate that both error measures $\mathcal{J}_u(u_h^{k,i})$ and $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i})$ exhibit a very close behavior.

3 A posteriori error estimates and stopping criteria

In this section, we present our a posteriori error estimates and stopping criteria. We proceed generally, with a given discrete potential $u_h^{k,i} \in V(\mathcal{T}_h)$, $k \geq 1$, $i \geq 1$, not linked to any particular discretization scheme or to any iterative nonlinear or linear solvers. Examples of application can be found in Section 6 and, more extensively, in Part II. The starting point of our general framework is the following assumption:

Assumption 3.1 (Quasi-equilibrated flux reconstruction)**.** *There exist a vector-valued function* $\mathbf{t}_h^{k,i} \in \mathbf{H}^q(\mathrm{div},\Omega)$ *and a scalar-valued function* $\rho_h^{k,i} \in L^q(\Omega)$ *such that*

$$\nabla\cdot\mathbf{t}_h^{k,i} = f_h - \rho_h^{k,i}, \tag{3.1}$$

where f_h *is a piecewise polynomial approximation of the source term* f *verifying* $(f_h,1)_K = (f,1)_K$ *for all* $K \in \mathcal{T}_h$.

The function $\mathbf{t}_h^{k,i}$ plays the role of a *flux reconstruction* providing a discrete approximation of the exact flux $-\boldsymbol{\sigma}(u,\nabla u)$. Such a function is traditional in equilibrated flux estimates, see Prager and Synge [28], Luce and Wohlmuth [23], Braess and Schöberl [4], or the unified approach in [15]. In practice, see Part II, we construct $\mathbf{t}_h^{k,i}$ in Raviart–Thomas–Nédélec discrete subspaces of $\mathbf{H}^q(\mathrm{div},\Omega)$. Furthermore, the function $\rho_h^{k,i}$ plays the role of an *algebraic remainder*. This function is introduced to facilitate the practical construction of $\mathbf{t}_h^{k,i}$. Indeed, while using iterative linear solvers, it is usually difficult to achieve exact equilibration in the sense that (3.1) is satisfied with $\rho_h^{k,i} = 0$. An example for constructing $\mathbf{t}_h^{k,i}$ such that $\rho_h^{k,i} = 0$ is the algorithm of [17, Section 7.3] which requires an ordering of the mesh elements and then a run through all the elements with a local minimization problem inside each element. Herein, we consider instead a general nonzero $\rho_h^{k,i}$ with the only requirement that it can be made small enough (the precise requirement is stated in Section 3.3). A simple and practical way to devise the algebraic remainder $\rho_h^{k,i}$ is presented in Section 4, following [18].

Remark 3.2 (Function f_h)**.** *For lowest-order discretizations,* f_h *is generally the piecewise constant function given by the elementwise mean values of* f*. For higher-order discretizations, a more accurate approximation of* f *is considered.*

Remark 3.3 (Local mass conservation)**.** *Even if we work with not fully converged linear and nonlinear solvers, Assumption 3.1 means that* $\mathbf{t}_h^{k,i}$ *represents a flux with a continuous normal trace, the elementwise mass balance misfit being* $\rho_h^{k,i}$.

3.1 Guaranteed error upper bound

For any $K \in \mathcal{T}_h$, the generalized Poincaré inequality states that

$$\|\varphi - \varphi_K\|_{p,K} \leq C_{\mathrm{P},p} h_K \|\nabla\varphi\|_{p,K} \qquad \forall \varphi \in W^{1,p}(K), \tag{3.2}$$

where φ_K denotes the mean value of φ in K. Since simplices are convex, there holds $C_{\mathrm{P},p} = \pi^{-\frac{2}{p}} d^{\frac{1}{2}-\frac{1}{p}}$ for $p \geq 2$, see Verfürth [30], and $C_{\mathrm{P},p} = p^{\frac{1}{p}} 2^{\frac{(p-1)}{p}}$ for all $p \in (1, +\infty)$, see Chua and Wheeden [8]. The generalized Friedrichs inequality states that

$$\|\varphi\|_p \leq h_\Omega \|\nabla\varphi\|_p \qquad \forall \varphi \in V. \tag{3.3}$$

In what follows, we denote our estimators in the form $\eta_{\cdot,K}^{k,i}$ where $k \geq 1$ stands for the nonlinear solver step, $i \geq 1$ for the linear solver step, and $K \in \mathcal{T}_h$ for the mesh element. We define global versions of these estimators as $\eta^{k,i}_{\cdot} := \left\{\sum_{K\in\mathcal{T}_h} \left(\eta_{\cdot,K}^{k,i}\right)^q\right\}^{1/q}$. Our main result on the a posteriori error estimate is:

Theorem 3.4 (Guaranteed upper bound)**.** *Let $u \in V$ solve (2.5), let $u_h^{k,i} \in V(\mathcal{T}_h)$ be arbitrary, and let Assumption 3.1 hold. For any $K \in \mathcal{T}_h$, define respectively the* flux *and the* nonconformity estimators *as*

$$\eta_{\mathrm{F},K}^{k,i} := \|\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{t}_h^{k,i}\|_{q,K}, \tag{3.4a}$$

$$\eta_{\mathrm{NC},K}^{k,i} := \left\{\sum_{e\in\mathcal{E}_K} \alpha_e^s h_e^{1-s} \|[\![u_h^{k,i}]\!]\|_{s,e}^s\right\}^{\frac{1}{q}}, \tag{3.4b}$$

and the algebraic remainder *and* data oscillation estimators *as*

$$\eta_{\mathrm{rem},K}^{k,i} := h_\Omega \|\rho_h^{k,i}\|_{q,K}, \tag{3.5a}$$

$$\eta_{\mathrm{osc},K}^{k,i} := C_{\mathrm{P},p} h_K \|f - f_h\|_{q,K}. \tag{3.5b}$$

Then,

$$\mathcal{J}_u(u_h^{k,i}) \leq \eta^{k,i} := \eta_{\mathrm{F}}^{k,i} + \eta_{\mathrm{NC}}^{k,i} + \eta_{\mathrm{rem}}^{k,i} + \eta_{\mathrm{osc}}^{k,i}. \tag{3.6}$$

Proof. Taking into account that $[\![u]\!] = 0$ for all $e \in \mathcal{E}_h$, it is clear that $\mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}) = \eta_{\mathrm{NC}}^{k,i}$. We are thus left with bounding $\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i})$. Let $\varphi \in V$ with $\|\nabla\varphi\|_p = 1$ be fixed. Since $\mathbf{t}_h^{k,i} \in \mathbf{H}^q(\mathrm{div}, \Omega)$, the Green formula yields $(\mathbf{t}_h^{k,i}, \nabla\varphi) = -(\nabla{\cdot}\mathbf{t}_h^{k,i}, \varphi)$. Hence, using (2.5) and adding and subtracting $(\mathbf{t}_h^{k,i}, \nabla\varphi)$, we infer

$$(\boldsymbol{\sigma}(u, \nabla u) - \boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla\varphi) = (f - \nabla{\cdot}\mathbf{t}_h^{k,i}, \varphi) - (\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{t}_h^{k,i}, \nabla\varphi).$$

The Hölder inequality yields

$$|(\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{t}_h^{k,i}, \nabla\varphi)| \leq \sum_{K\in\mathcal{T}_h} \|\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{t}_h^{k,i}\|_{q,K} \|\nabla\varphi\|_{p,K} \leq \eta_{\mathrm{F}}^{k,i}.$$

Assumption 3.1, the Hölder inequality, the generalized Poincaré inequality (3.2), and the generalized Friedrichs inequality (3.3) lead to

$$\begin{aligned}
|(f - \nabla{\cdot}\mathbf{t}_h^{k,i}, \varphi)| &= \sum_{K\in\mathcal{T}_h} (f - \nabla{\cdot}\mathbf{t}_h^{k,i} - \rho_h^{k,i}, \varphi)_K + (\rho_h^{k,i}, \varphi) \\
&= \sum_{K\in\mathcal{T}_h} (f - f_h, \varphi - \varphi_K)_K + (\rho_h^{k,i}, \varphi) \\
&\leq \sum_{K\in\mathcal{T}_h} \|f - f_h\|_{q,K} C_{\mathrm{P},p} h_K \|\nabla\varphi\|_{p,K} + \|\rho_h^{k,i}\|_q h_\Omega \|\nabla\varphi\|_p \\
&\leq \eta_{\mathrm{osc}}^{k,i} + \eta_{\mathrm{rem}}^{k,i}.
\end{aligned}$$

Combining the above bounds yields (3.6). □

3.2 Distinguishing the different error components

We now identify and estimate separately the various error components. To proceed generally, we introduce the following assumption:

Assumption 3.5 (Discretization, linearization, and algebraic errors)**.** *There exist vector-valued functions* $\mathbf{d}_h^{k,i}, \mathbf{l}_h^{k,i}, \mathbf{a}_h^{k,i} \in [L^q(\Omega)]^d$ *such that*

(i) $\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i} + \mathbf{a}_h^{k,i} = \mathbf{t}_h^{k,i}$;

(ii) *as the linear solver converges,* $\|\mathbf{a}_h^{k,i}\|_q \to 0$;

(iii) *as the nonlinear solver converges,* $\|\mathbf{l}_h^{k,i}\|_q \to 0$.

The function $\mathbf{d}_h^{k,i}$ is meant to approximate the *discretization* flux $-\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$, the function $\mathbf{l}_h^{k,i}$ represents the *linearization error*, and the function $\mathbf{a}_h^{k,i}$ the *algebraic error*. A generic way to construct $\mathbf{a}_h^{k,i}$ is presented in Section 4; the construction of the functions $\mathbf{d}_h^{k,i}$ and $\mathbf{l}_h^{k,i}$ then depends on the discretization scheme at hand, see Section 6 for a first example and, more generally, Part II.

The last error component we distinguish is *quadrature*. Indeed, because of nonlinearities, $\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$ is not necessarily a piecewise polynomial even if the discrete potential $u_h^{k,i}$ is so. We introduce a piecewise polynomial vector-valued function $\overline{\boldsymbol{\sigma}}_h^{k,i}$ meant to approximate $\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$; the specific definition of $\overline{\boldsymbol{\sigma}}_h^{k,i}$ depends on the discretization scheme at hand, see Section 6 and Part II. The main result of this section is:

Theorem 3.6 (A posteriori error estimate distinguishing the error components)**.** *Let* $u \in V$ *solve* (2.5) *and let* $u_h^{k,i} \in V(\mathcal{T}_h)$ *be arbitrary. Let Assumptions* 3.1 *and* 3.5 *hold. For any* $K \in \mathcal{T}_h$*, define respectively the* discretization*,* linearization*,* algebraic*, and* quadrature estimators *as*

$$\eta_{\mathrm{disc},K}^{k,i} := 2^{1/p}\big(\|\overline{\boldsymbol{\sigma}}_h^{k,i} + \mathbf{d}_h^{k,i}\|_{q,K} + \eta_{\mathrm{NC},K}^{k,i}\big), \tag{3.7a}$$

$$\eta_{\mathrm{lin},K}^{k,i} := \|\mathbf{l}_h^{k,i}\|_{q,K}, \tag{3.7b}$$

$$\eta_{\mathrm{alg},K}^{k,i} := \|\mathbf{a}_h^{k,i}\|_{q,K}, \tag{3.7c}$$

$$\eta_{\mathrm{quad},K}^{k,i} := \|\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) - \overline{\boldsymbol{\sigma}}_h^{k,i}\|_{q,K}, \tag{3.7d}$$

with $\eta_{\mathrm{NC},K}^{k,i}$ *defined by* (3.4b)*. Let* $\eta_{\mathrm{rem},K}^{k,i}$ *and* $\eta_{\mathrm{osc},K}^{k,i}$ *be defined respectively by* (3.5a) *and* (3.5b)*. Then,*

$$\mathcal{J}_u(u_h^{k,i}) \le \eta_{\mathrm{disc}}^{k,i} + \eta_{\mathrm{lin}}^{k,i} + \eta_{\mathrm{alg}}^{k,i} + \eta_{\mathrm{rem}}^{k,i} + \eta_{\mathrm{quad}}^{k,i} + \eta_{\mathrm{osc}}^{k,i}. \tag{3.8}$$

Proof. The decomposition of Assumption 3.5 and the triangle inequality yield

$$\|\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{t}_h^{k,i}\|_{q,K} \le \|\overline{\boldsymbol{\sigma}}_h^{k,i} + \mathbf{d}_h^{k,i}\|_{q,K} + \|\mathbf{l}_h^{k,i}\|_{q,K} + \|\mathbf{a}_h^{k,i}\|_{q,K} + \eta_{\mathrm{quad},K}^{k,i}.$$

The assertion then follows from Theorem 3.4 combined with the triangle inequality, the Hölder inequality, and the inequality $a^q + b^q \le (a+b)^q$ for $a, b \ge 0$ used to regroup $\|\overline{\boldsymbol{\sigma}}_h^{k,i} + \mathbf{d}_h^{k,i}\|_{q,K}$ with $\eta_{\mathrm{NC},K}^{k,i}$. □

3.3 Stopping criteria

We can now devise stopping criteria for the linear solver in step 3c of Algorithm 1.1 and for the nonlinear solver in step 4 of Algorithm 1.1. The idea is to require the algebraic estimator to be sufficiently small with respect to the linearization or discretization estimators and the linearization estimator to be sufficiently small with respect to the discretization estimator. We introduce a third (balancing) requirement, namely that the algebraic remainder estimator be sufficiently small with respect to the three other estimators.

We consider two versions for the criteria, a global one and a local one. Our *global stopping criteria* are

$$\eta_{\mathrm{rem}}^{k,i} \le \gamma_{\mathrm{rem}} \max\big\{\eta_{\mathrm{disc}}^{k,i}, \eta_{\mathrm{lin}}^{k,i}, \eta_{\mathrm{alg}}^{k,i}\big\}, \tag{3.9a}$$

$$\eta_{\mathrm{alg}}^{k,i} \le \gamma_{\mathrm{alg}} \max\big\{\eta_{\mathrm{disc}}^{k,i}, \eta_{\mathrm{lin}}^{k,i}\big\}, \tag{3.9b}$$

$$\eta_{\mathrm{lin}}^{k,i} \le \gamma_{\mathrm{lin}} \eta_{\mathrm{disc}}^{k,i}, \tag{3.9c}$$

where γ_{rem}, γ_{alg}, and γ_{lin} are positive user-given weights, typically of order 0.1. These global criteria are used to establish the *global efficiency* of our error estimators, see Theorem 5.4. Our *local stopping criteria*

are elementwise equivalents of (3.9),

$$\eta_{\mathrm{rem},K}^{k,i} \le \gamma_{\mathrm{rem},K} \max\{\eta_{\mathrm{disc},K}^{k,i}, \eta_{\mathrm{lin},K}^{k,i}, \eta_{\mathrm{alg},K}^{k,i}\} \qquad \forall K \in \mathcal{T}_h, \tag{3.10a}$$

$$\eta_{\mathrm{alg},K}^{k,i} \le \gamma_{\mathrm{alg},K} \max\{\eta_{\mathrm{disc},K}^{k,i}, \eta_{\mathrm{lin},K}^{k,i}\} \qquad \forall K \in \mathcal{T}_h, \tag{3.10b}$$

$$\eta_{\mathrm{lin},K}^{k,i} \le \gamma_{\mathrm{lin},K} \eta_{\mathrm{disc},K}^{k,i} \qquad \forall K \in \mathcal{T}_h, \tag{3.10c}$$

where, for any $K \in \mathcal{T}_h$, $\gamma_{\mathrm{rem},K}$, $\gamma_{\mathrm{alg},K}$, and $\gamma_{\mathrm{lin},K}$ are positive user-given weights, typically of order 0.1. These local criteria are used to establish the *local efficiency* of our error estimators, see Theorem 5.3, and are essential for mesh adaptivity.

4 Algebraic remainder and algebraic error flux reconstruction

The goal of this section is to present a simple and practical way to devise the algebraic remainder $\rho_h^{k,i}$ and the algebraic error flux reconstruction $\mathbf{a}_h^{k,i}$ so that the stopping criteria (3.9a) or (3.10a) together with Assumption 3.5(i-ii) are satisfied. To this purpose, we suppose that we have at our disposal the flux reconstructions $\mathbf{d}_h^{k,i}$ and $\mathbf{l}_h^{k,i}$ such that the sum $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$ satisfies the following assumption (recall that f_h is the piecewise polynomial approximation of f introduced in Assumption 3.1):

Assumption 4.1 (Quasi-equilibration for $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$)**.** *The function $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$ is in $\mathbf{H}^q(\mathrm{div}, \Omega)$, and there exists a scalar-valued function $r_h^{k,i} \in L^q(\Omega)$ such that*

$$\nabla\cdot(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i}) = f_h - r_h^{k,i}. \tag{4.1}$$

Referring to Algorithm 1.1 where the linear system (1.2) for some fixed $k \ge 1$ is being solved iteratively, the i-th step of the linear solver yields the algebraic residual vector $R^{k,i}$ in (1.3). We will see in Part II how the (piecewise polynomial) function $r_h^{k,i}$ can be related in a simple manner to the components of the vector $R^{k,i}$ for various discretization schemes, see also Section 6 for a simple example. Now, the idea to construct $\rho_h^{k,i}$ and $\mathbf{a}_h^{k,i}$, following [17, Section 7.2] and [18], consists in performing some additional steps, say $\nu > 0$, of the iterative linear solver. We emphasize that this construction is independent of the actual form of this solver. Performing these additional ν steps yields the algebraic residual vector $R^{k,i+\nu}$ such that

$$\mathbb{A}^k U^{k,i+\nu} = F^k - R^{k,i+\nu}, \tag{4.2}$$

and from this vector, the function $r_h^{k,i+\nu}$ is constructed in the same way as $r_h^{k,i}$ from $R^{k,i}$. We can now construct $\rho_h^{k,i}$ and $\mathbf{a}_h^{k,i}$ as follows:

Definition 4.2 (Construction of $\rho_h^{k,i}$ and $\mathbf{a}_h^{k,i}$)**.** *Let the k-th step of the nonlinear solver and the i-th step of the linear solver be given, yielding $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$ and $r_h^{k,i}$ satisfying (4.1). Let $\nu > 0$ and perform ν additional steps of the linear solver, yielding $(\mathbf{d}_h^{k,i+\nu} + \mathbf{l}_h^{k,i+\nu})$ and $r_h^{k,i+\nu}$ satisfying (4.1) with i replaced by $i+\nu$. Then, set*

$$\mathbf{a}_h^{k,i} := (\mathbf{d}_h^{k,i+\nu} + \mathbf{l}_h^{k,i+\nu}) - (\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i}), \tag{4.3a}$$

$$\rho_h^{k,i} := r_h^{k,i+\nu}. \tag{4.3b}$$

The following important result can be easily verified:

Lemma 4.3 (Assumptions 3.1 and 3.5(i-ii))**.** *Under Assumption 4.1 and with the construction of Definition 4.2, define $\mathbf{t}_h^{k,i} := \mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i} + \mathbf{a}_h^{k,i}$. Then, Assumptions 3.1 and 3.5(i-ii) hold.*

In practice, the parameter ν can be determined adaptively by increasing progressively its value and checking (3.9a) or (3.10a).

5 Local and global efficiency and robustness

This section is devoted to the proof of the efficiency and robustness of our a posteriori error estimates. The specific construction of Section 4 is not needed; we just use the stopping criteria (3.9) or (3.10).

5.1 Local approximation property

To proceed generally, we make an additional assumption on the discretization error flux reconstruction $\mathbf{d}_h^{k,i}$. We verify this assumption for a large class of discretization schemes in Part II. Recall that $u_h^{k,i} \in V(\mathcal{T}_h)$ is arbitrary, that f_h is a piecewise polynomial approximation of f, and that $\overline{\boldsymbol{\sigma}}_h^{k,i}$ is a piecewise polynomial approximation of $\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$. For any $K \in \mathcal{T}_h$, let $\mathfrak{T}_K$ collect all elements $K' \in \mathcal{T}_h$ which share at least a vertex with K. Similarly, let $\mathfrak{E}_K$ collect all faces which share at least a vertex with K, and $\mathfrak{E}_K^{\mathrm{int}} := \mathfrak{E}_K \cap \mathcal{E}_h^{\mathrm{int}}$. Let $\mathbf{n}_e$ be the unit normal vector of the face e (its orientation is irrelevant). Define

$$\eta_{\sharp,\mathfrak{T}_K}^{k,i} := \left\{ \sum_{K' \in \mathfrak{T}_K} h_{K'}^q \| f_h + \nabla\cdot\overline{\boldsymbol{\sigma}}_h^{k,i} \|_{q,K'}^q + \sum_{e \in \mathfrak{E}_K^{\mathrm{int}}} h_e \| [\![\overline{\boldsymbol{\sigma}}_h^{k,i} \cdot \mathbf{n}_e]\!] \|_{q,e}^q \right\}^{\frac{1}{q}}. \tag{5.1a}$$

We also use the notation $\eta_{\cdot,\mathfrak{T}_K}^{k,i} := \left\{ \sum_{K' \in \mathfrak{T}_K} \left(\eta_{\cdot,K'}^{k,i} \right)^q \right\}^{\frac{1}{q}}$ for the estimators introduced in Section 3. Henceforth, $A \lesssim B$ stands for the inequality $A \le CB$ with a generic constant C independent of the mesh sizes h_K and h_e, the domain Ω, the nonlinear function $\boldsymbol{\sigma}$, and the Lebesgue exponent p, but that can depend on the shape regularity of the mesh family $\{\mathcal{T}_h\}_h$ and on the polynomial degrees of $\overline{\boldsymbol{\sigma}}_h^{k,i}$ and f_h.

Assumption 5.1 (Local approximation property)**.** *For all $K \in \mathcal{T}_h$, there holds*

$$\| \overline{\boldsymbol{\sigma}}_h^{k,i} + \mathbf{d}_h^{k,i} \|_{q,K} \lesssim \eta_{\sharp,\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{NC},\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}. \tag{5.2}$$

Remark 5.2 (Tighter approximation property)**.** *For most discretization schemes, see Part* II*, it is actually possible to prove $\| \overline{\boldsymbol{\sigma}}_h^{k,i} + \mathbf{d}_h^{k,i} \|_{q,K} \lesssim \eta_{\sharp,\mathfrak{T}_K}^{k,i}$.*

5.2 Local efficiency

Our local efficiency result is achieved with respect to the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i})$ defined by (2.9).

Theorem 5.3 (Local efficiency)**.** *Let $u \in V$ solve (2.5) and let $u_h^{k,i} \in V(\mathcal{T}_h)$ be arbitrary. Let the local stopping criteria (3.10) be satisfied. Then, under Assumption 5.1, there holds, for all $K \in \mathcal{T}_h$,*

$$\begin{aligned} &\eta_{\mathrm{disc},K}^{k,i} + \eta_{\mathrm{lin},K}^{k,i} + \eta_{\mathrm{alg},K}^{k,i} + \eta_{\mathrm{rem},K}^{k,i} \\ &\lesssim \| \boldsymbol{\sigma}(u, \nabla u) - \boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) \|_{q,\mathfrak{T}_K} + \eta_{\mathrm{NC},\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{quad},\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}. \end{aligned} \tag{5.3}$$

Proof. Let $K \in \mathcal{T}_h$ be fixed. Owing to the local criteria (3.10), we infer

$$\eta_{\mathrm{lin},K}^{k,i} + \eta_{\mathrm{alg},K}^{k,i} + \eta_{\mathrm{rem},K}^{k,i} \lesssim \eta_{\mathrm{disc},K}^{k,i}.$$

Combining the definition (3.7a) of $\eta_{\mathrm{disc},K}^{k,i}$ with Assumption 5.1 yields $\eta_{\mathrm{disc},K}^{k,i} \lesssim \eta_{\sharp,\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{NC},\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}$, whence

$$\eta_{\mathrm{disc},K}^{k,i} + \eta_{\mathrm{lin},K}^{k,i} + \eta_{\mathrm{alg},K}^{k,i} + \eta_{\mathrm{rem},K}^{k,i} \lesssim \eta_{\sharp,\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{NC},\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}.$$

Now, the inequalities (A.6) and (A.7) from [14, Proof of Lemma 4.3] together with the triangle inequality yield

$$\begin{aligned} \eta_{\sharp,\mathfrak{T}_K}^{k,i} &\lesssim \| \boldsymbol{\sigma}(u, \nabla u) - \overline{\boldsymbol{\sigma}}_h^{k,i} \|_{q,\mathfrak{T}_K} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i} \\ &\lesssim \| \boldsymbol{\sigma}(u, \nabla u) - \boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) \|_{q,\mathfrak{T}_K} + \eta_{\mathrm{quad},\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}, \end{aligned}$$

whence the assertion of the theorem. □

5.3 Global efficiency and robustness

Proceeding as above (while relying on (A.10) and (A.11) from [14, Proof of Lemma 4.7]) yields our main result for global efficiency and robustness with respect to the original error measure $\mathcal{J}_u(u_h^{k,i})$.

Theorem 5.4 (Global efficiency and robustness)**.** *Let $u \in V$ solve (2.5) and let $u_h^{k,i} \in V(\mathcal{T}_h)$ be arbitrary. Let the global stopping criteria (3.9) be satisfied. Then, under Assumption 5.1, there holds*

$$\eta_{\rm disc}^{k,i} + \eta_{\rm lin}^{k,i} + \eta_{\rm alg}^{k,i} + \eta_{\rm rem}^{k,i} \lesssim \mathcal{J}_u(u_h^{k,i}) + \eta_{\rm quad}^{k,i} + \eta_{\rm osc}^{k,i}. \tag{5.4}$$

Remark 5.5 (Comparison with [14])**.** *In [14], the linearization stopping parameters $\gamma_{{\rm lin},K}$ (or $\gamma_{\rm lin}$) had to be "small enough" in order that the equivalents of Theorems 5.3 and 5.4 hold. This is no longer necessary in the present setting owing to the decomposition introduced in Assumption 3.5 and the fact that Assumption 5.1 concerns the component $\mathbf{d}_h^{k,i}$ of the flux reconstruction.*

6 Example: p-Laplacian discretized by nonconforming finite elements

We show in this section how to apply the above abstract theory to a specific example: the p-Laplacian discretized by nonconforming finite elements and linearized either by the Newton or the fixed point method. We recall that the p-Laplacian fits the general form (2.4) with $\boldsymbol{\sigma}(\boldsymbol{\xi}) = |\boldsymbol{\xi}|^{p-2}\boldsymbol{\xi}$ for all $\boldsymbol{\xi} \in \mathbb{R}^d$ so that, in this example, the nonlinear flux function $\boldsymbol{\sigma}$ only depends on $\boldsymbol{\xi}$.

6.1 Discretization

Let f_h be the piecewise constant function given by the elementwise mean values of the source term f. Let V_h be the Crouzeix–Raviart finite element space [9] of piecewise affine polynomials on $\mathcal{T}_h$ such that the interface jumps and boundary face values have zero mean value over the corresponding face. The corresponding discretization of problem (2.5) reads: find $u_h \in V_h$ such that

$$(\boldsymbol{\sigma}(\nabla u_h), \nabla v_h) = (f_h, v_h) \qquad \forall v_h \in V_h. \tag{6.1}$$

Let $\psi_e \in V_h$ stand for the Crouzeix–Raviart basis function associated with the face $e \in \mathcal{E}_h^{\rm int}$, i.e., the function that takes the value 1 at the barycenter of e and the value 0 at the barycenter of other faces in $\mathcal{E}_h$. Testing (6.1) against these basis functions gives rise to the nonlinear algebraic system (1.1).

6.2 Linearization

Choosing $u_h^0 \in V_h$ yields the initial vector U^0 in Algorithm 1.1. The linearization of (6.1), for $k \geq 1$, reads: find $u_h^k \in V_h$ such that

$$(\boldsymbol{\sigma}^{k-1}(\nabla u_h^k), \nabla \psi_e) = (f_h, \psi_e) \qquad \forall e \in \mathcal{E}_h^{\rm int}, \tag{6.2}$$

which is a functional form of the algebraic system (1.2). Two common linearizations to define the flux function $\boldsymbol{\sigma}^{k-1}$ are the fixed point linearization where $\boldsymbol{\sigma}^{k-1}(\boldsymbol{\xi}) := |\nabla u_h^{k-1}|^{p-2}\boldsymbol{\xi}$, and the Newton linearization where

$$\boldsymbol{\sigma}^{k-1}(\boldsymbol{\xi}) := |\nabla u_h^{k-1}|^{p-2}\boldsymbol{\xi} + (p-2)|\nabla u_h^{k-1}|^{p-4}(\nabla u_h^{k-1} \otimes \nabla u_h^{k-1})(\boldsymbol{\xi} - \nabla u_h^{k-1}). \tag{6.3}$$

6.3 Algebraic solution

On the i-th step, $i \geq 1$, of an iterative linear solver for the algebraic system (1.2), we obtain the algebraic residual vector $R^{k,i}$ in (1.3). Here, this vector has components associated with the interior faces, $R^{k,i} = \{R_e^{k,i}\}_{e \in \mathcal{E}_h^{\rm int}}$. For convenience, we also set $R_e^{k,i} := 0$ for any $e \in \mathcal{E}_h^{\rm ext}$. The functional form of (1.3) is: find $u_h^{k,i} \in V_h$ such that

$$(\boldsymbol{\sigma}^{k-1}(\nabla u_h^{k,i}), \nabla \psi_e) = (f_h, \psi_e) - R_e^{k,i} \qquad \forall e \in \mathcal{E}_h^{\rm int}. \tag{6.4}$$

6.4 Flux reconstruction

We present here a simple approach where the flux reconstructions $\mathbf{d}_h^{k,i}$ and $\mathbf{l}_h^{k,i}$ are prescribed elementwise, while the flux reconstruction $\mathbf{a}_h^{k,i}$ results from Definition 4.2. For all $K \in \mathcal{T}_h$, let $\mathbf{f}_h(\mathbf{x})|_K := \frac{f_h|_K}{d}(\mathbf{x}-\mathbf{x}_K)$ where $\mathbf{x}_K$ is the barycenter of K. For all $e \in \mathcal{E}_K$, let K_e be the sub-simplex of K formed by the face e and the point $\mathbf{x}_K$. Let D_e regroup the two (or one for boundary faces) sub-simplices K_e which share e. We first prescribe the sum $(\mathbf{d}_h^{k,i}+\mathbf{l}_h^{k,i})$:

Definition 6.1 (Construction of $(\mathbf{d}_h^{k,i}+\mathbf{l}_h^{k,i})$)**.** *Set, for all $K \in \mathcal{T}_h$,*

$$(\mathbf{d}_h^{k,i}+\mathbf{l}_h^{k,i})|_K := \big(-\boldsymbol{\sigma}^{k-1}(\nabla u_h^{k,i})+\mathbf{f}_h\big)|_K - \sum_{e\in\mathcal{E}_K} |D_e|^{-1}\frac{R_e^{k,i}}{d}(\mathbf{x}-\mathbf{x}_K)|_{K_e}. \tag{6.5}$$

To comply with Assumption 3.5(iii), the construction of $\mathbf{d}_h^{k,i}$ mimics that of $(\mathbf{d}_h^{k,i}+\mathbf{l}_h^{k,i})$ with $\boldsymbol{\sigma}(\nabla u_h^{k,i})$ in place of $\boldsymbol{\sigma}^{k-1}(\nabla u_h^{k,i})$. Specifically, let

$$\bar{R}_e^{k,i} := (f_h,\psi_e) - (\boldsymbol{\sigma}(\nabla u_h^{k,i}),\nabla\psi_e) \qquad \forall e\in\mathcal{E}_h^{\mathrm{int}}, \tag{6.6}$$

and $\bar{R}_e^{k,i} := 0$ for all $e \in \mathcal{E}_h^{\mathrm{ext}}$. We prescribe $\mathbf{d}_h^{k,i}$ (and hence, also $\mathbf{l}_h^{k,i}$ by subtraction):

Definition 6.2 (Construction of $\mathbf{d}_h^{k,i}$)**.** *Set, for all $K \in \mathcal{T}_h$,*

$$\mathbf{d}_h^{k,i}|_K := \big(-\boldsymbol{\sigma}(\nabla u_h^{k,i})+\mathbf{f}_h\big)|_K - \sum_{e\in\mathcal{E}_K} |D_e|^{-1}\frac{\bar{R}_e^{k,i}}{d}(\mathbf{x}-\mathbf{x}_K)|_{K_e}. \tag{6.7}$$

We postpone the verification of Assumptions 4.1 and 5.1 to Part II. At this stage, we just observe that the function $r_h^{k,i}$ in Assumption 4.1 is piecewise constant on the dual mesh $\mathcal{D}_h = \{D_e\}_{e\in\mathcal{E}_h}$ with $r_h^{k,i}|_{D_e} = R_e^{k,i}|D_e|^{-1}$ for all $e \in \mathcal{E}_h$, and that, in Assumption 5.1, the function $\overline{\boldsymbol{\sigma}}_h^{k,i}$ is piecewise constant on $\mathcal{T}_h$ and simply equal to $\boldsymbol{\sigma}(\nabla u_h^{k,i})$ (so that $\eta_{\mathrm{quad}}^{k,i}$ is zero). Moreover, $\mathbf{d}_h^{k,i}$ and $\mathbf{l}_h^{k,i}$ turn out to have continuous normal component across interfaces, so that both functions sit in $\mathbf{H}^q(\mathrm{div},\Omega)$.

7 Numerical experiments

This section illustrates numerically our theoretical developments. We consider the p-Laplacian with $p \in \{1.5, 4, 10\}$, $d = 2$, and two test cases with known analytical solution. The first case only considers uniform mesh refinement, whereas the second one also includes adaptive mesh refinement. We employ the Crouzeix–Raviart nonconforming finite element method (6.1) and use the flux reconstructions of Section 6.4. We use the Newton linearization (6.3) together with the conjugate gradient (CG) method with diagonal preconditioning as linear solver. In (2.8b), the coefficients α_e are set to one and $s := q$.

7.1 A first test case

We set $\Omega := (0,1)\times(0,1)$, $f := 2$, and prescribe the Dirichlet boundary condition by the exact solution

$$u(\mathbf{x}) = -\frac{p-1}{p}|\mathbf{x}-(0.5,0.5)|^{p/(p-1)} + \frac{p-1}{p}\Big(\frac{1}{2}\Big)^{p/(p-1)}.$$

This is a two-dimensional extension of a test case from Chaillou and Suri [7]. The error stemming from inhomogeneous boundary conditions is not taken into account. We consider six levels of uniform mesh refinement, together with the values $p \in \{1.5, 10\}$.

We compare three different stopping criteria in Algorithm 1.1, leading to three different solution approaches:

- In the *Full Newton (FN) method*, both the nonlinear and linear solvers are iterated to "almost" convergence. More precisely, we use the global stopping criteria $\eta_{\mathrm{alg}}^{k,i} \le 10^{-8}$ and $\eta_{\mathrm{lin}}^{k,i} \le 10^{-8}$. The stopping criterion (3.9a) is employed with $\gamma_{\mathrm{rem}} = 0.1$; this criterion influences the precision of the calculation of the algebraic error component but not Algorithm 1.1.

Setting				Flux				Potential	
Case	p	Mesh	D. osc.	$[W^{s_q,q}(\Omega)]^d$	$\mathcal{J}_u^{\rm up}$	$\mathcal{J}_u^{\rm low}$	$\eta^{k,i}$	$W^{s_p,p}(\Omega)$	$\|\nabla(u-u_h^{k,i})\|_p$
1	1.5	unif.	—	$s_q = 1.67$	1.00	0.99	1.00	$s_p = 4.33$	1.00
1	10	unif.	—	$s_q = 2.80$	0.99	1.01	0.99	$s_p = 1.31$	0.31
2	4	unif.	1.13	$s_q = 1.13$	0.94	0.95	0.99	$s_p = 1.38$	0.38
2	4	adap.	1.64	$s_q = 1.13$	0.97	1.00	0.99	$s_p = 1.38$	0.89

Table 1: Flux and potential regularities and experimental orders of convergence

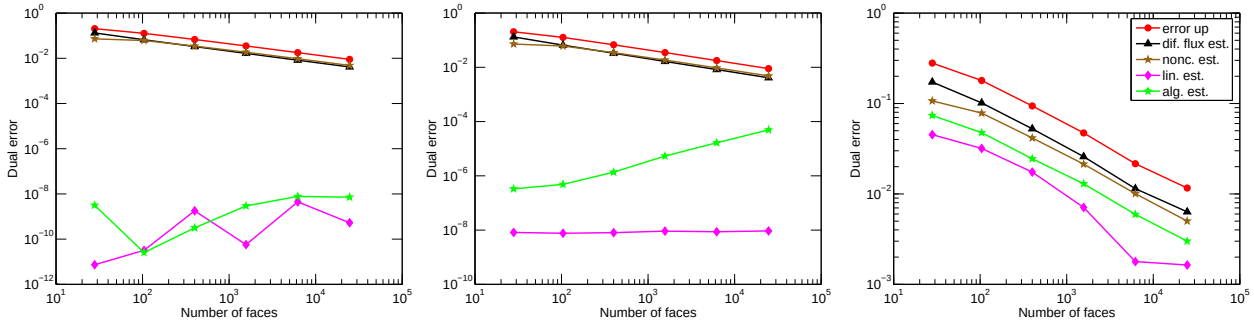

Figure 1: Error and estimators on uniformly refined meshes, case 1, $p = 10$. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

- In the *Inexact Newton (IN) method*, the only difference with FN is that a fixed number of preconditioned CG iterations is performed on each Newton linearization step. These values were chosen respectively as 2, 3, 5, 8, 10, 15 on each level of uniform mesh refinement.
- Finally, in the *Adaptive Inexact Newton (AIN) method* introduced in this work, we rely on the global stopping criteria of Section 3.3 with $\gamma_{\rm lin} = \gamma_{\rm alg} = 0.3$ and $\gamma_{\rm rem} = 0.3$, which leads to values of ν in the order of 20% of the number of algebraic solver iterations on each linearization step.

The initial linearization guess $u_h^0 \in V_h$ is defined, on every considered mesh, by perturbed punctual values of the exact solution u in the form $u_h^0(x,y) := u(x,y)(1 + \lambda(x-\mu)(y-\mu))$ with perturbation parameters $\lambda := 1$ and $\mu := 0.5$.

We begin with our results for $p = 10$. Figure 1 displays the curves of the error measure $\mathcal{J}_u^{\rm up}(u_h^{k,i})$, cf. (2.9), and of the estimators $\eta_{\rm F}^{k,i}$ and $\eta_{\rm NC}^{k,i}$ of Theorem 3.4 as a function of the number of mesh faces. In the present setting, the estimator $\eta_{\rm osc}^{k,i}$ is zero, and $\eta_{\rm rem}^{k,i}$ takes very small values. We observe that the three methods (FN, IN, and AIN) yield almost indistinguishable values for $\mathcal{J}_u^{\rm up}(u_h^{k,i})$, $\eta_{\rm F}^{k,i}$, and $\eta_{\rm NC}^{k,i}$, and these quantities exhibit optimal decrease with the number of mesh faces, see Table 1. Figure 1 also displays the curves of the linearization estimator $\eta_{\rm lin}^{k,i}$ and of the algebraic estimator $\eta_{\rm alg}^{k,i}$ of Theorem 3.6. The conceptual difference between the three methods lies in the size and behavior of these two estimators: both take values below 10^{-8} for FN; $\eta_{\rm alg}^{k,i}$ takes larger values for IN; both $\eta_{\rm alg}^{k,i}$ and $\eta_{\rm lin}^{k,i}$ take larger values that are just sufficiently small so as not to influence the error and estimators for AIN.

Figure 2 focuses more closely on the last, 6th level uniformly refined mesh, and tracks the dependence of the error measure $\mathcal{J}_u^{\rm up}(u_h^{k,i})$, the overall error estimator $\eta^{k,i}$ of Theorem 3.4, and the discretization and linearization estimators $\eta_{\rm disc}^{k,i}$ and $\eta_{\rm lin}^{k,i}$ of Theorem 3.6 on the Newton iterations. Typically, the error and all the estimators except $\eta_{\rm lin}^{k,i}$ start to stagnate after the linearization error ceases to dominate. This is precisely the point where the nonlinear iteration is stopped in AIN, whereas both FN and IN perform many unnecessary additional iterations. We can also observe the appearance of quadratic convergence for FN and a convergence slow-down for IN.

Figure 3 further analyzes the situation on one chosen Newton iteration from Figure 2. To be in a region with similar error measure $\mathcal{J}_u^{\rm up}(u_h^{k,i})$, we have chosen the 6th iteration for FN and IN and the 8th iteration

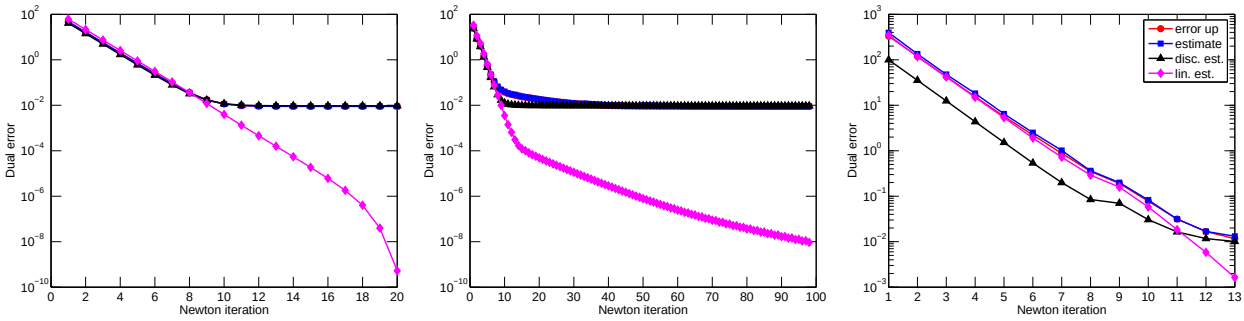

Figure 2: Error and estimators as a function of Newton iterations, case 1, $p = 10$, 6th level mesh. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

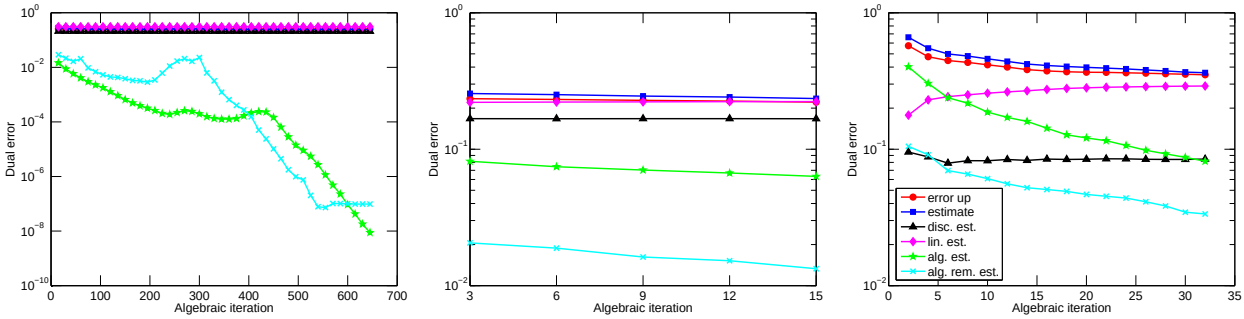

Figure 3: Error and estimators as a function of preconditioned CG iterations, case 1, $p = 10$, 6th level mesh. Newton, 6th step (left), inexact Newton, 6th step (middle), and adaptive inexact Newton, 8th step (right)

for AIN. We see that almost no decrease of the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i})$ can be observed during the almost 650 iterations of the preconditioned CG method in the FN case. The fixed 15 CG iterations in the IN case are, on the contrary, not completely sufficient to decrease significantly the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i})$. In our approach, just the sufficient, "online-decided" number of CG iterations is performed.

Figure 4 illustrates the overall performance of the three approaches. We can see that the number of Newton iterations (corresponding to the number of matrix assemblies) per refinement level is stable around 20 for FN. This observation is in agreement with the so-called asymptotic mesh independence of the number of Newton iterations, cf., e.g., Weiser *et al.* [31] and the references therein for theoretical results. It increases

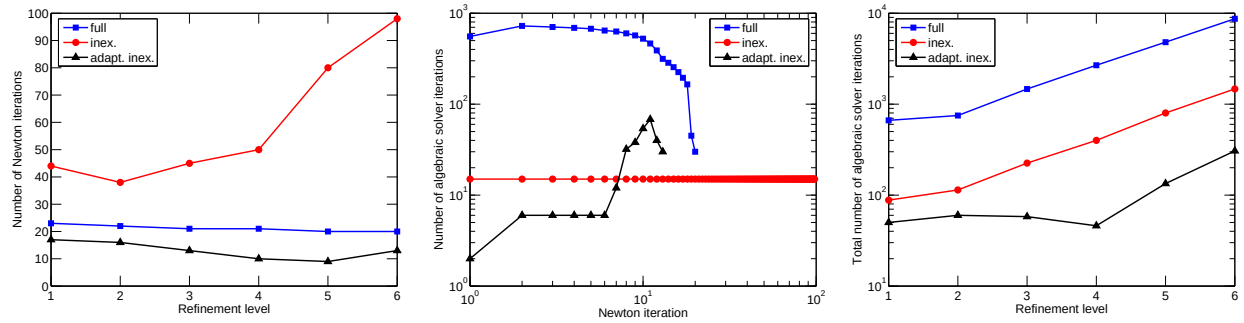

Figure 4: Number of Newton iterations per refinement level (left), number of linear solver iterations per Newton step on 6th level mesh (middle), and total number of linear solver iterations per refinement level (right). Case 1, $p = 10$

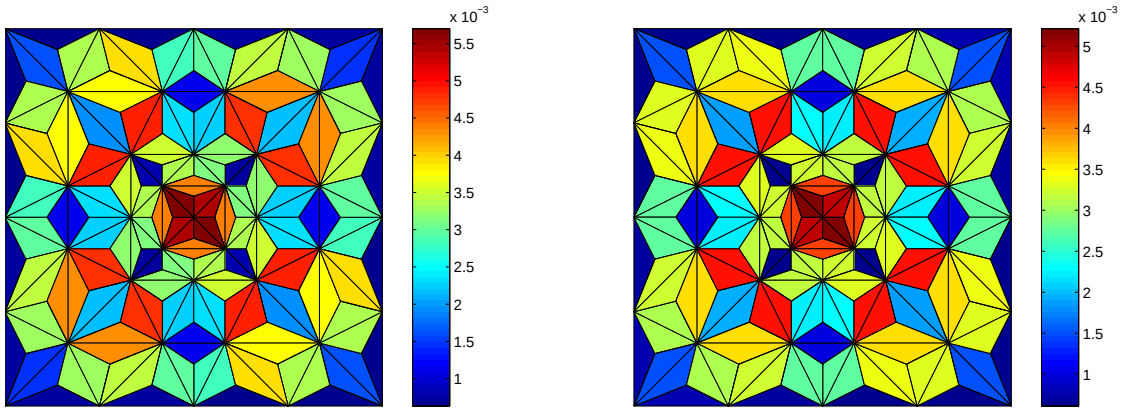

Figure 5: Estimated (left) and actual (right) error distribution, case 1, $p = 10$, 2nd level uniformly refined mesh, adaptive inexact Newton

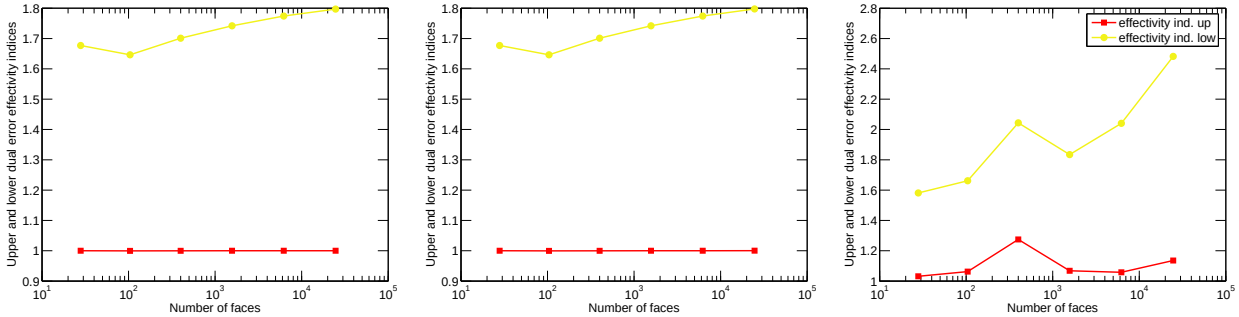

Figure 6: Upper and lower effectivity indices, case 1, $p = 10$. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

significantly for IN, whereas it is still reduced for AIN. On one Newton iteration (example for the 6th level refined mesh), the number of CG iterations also varies significantly between the three approaches. Many iterations are necessary in the FN case and fixed 15 iterations in the IN case, whereas AIN picks up the number that is "just necessary." Remark that this number is equal to two on the first Newton step; from here, the error is "lagged" as a function of Newton iterations in the AIN case, cf. Figure 2. The total number of necessary CG iterations per refinement level is displayed in the right part of Figure 4. On the last mesh, AIN only needs 306 total iterations, whereas IN needs 1470, and FN 8690 iterations. Thus, our approach yields an economy by a factor of roughly 5 with respect to IN and roughly 30 with respect to FN in terms of total iterations.

Figure 5 displays the distribution of the overall error estimator $\eta^{k,i}$ and of the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i})$ on the 2nd level uniformly refined mesh for AIN. We see that even in presence of algebraic and linearization errors, the overall error distribution is very well predicted. Finally, we construct a lower bound $\mathcal{J}_u^{\mathrm{low}}(u_h^{k,i})$ for the error measure by estimating the supremum in (2.8a) just with $\varphi = \mathcal{I}_{\mathrm{av}}(u_h^{k,i})$ where $\mathcal{I}_{\mathrm{av}}(u_h^{k,i})$ is the continuous, piecewise affine function on $\mathcal{T}_h$ given by the Dirichlet condition on boundary vertices and such that, on interior vertices, its value is equal to the average of the values taken by $u_h^{k,i}$ from the surrounding elements. We define the upper and lower effectivity indices respectively as $\mathcal{I}^{\mathrm{up}} := \eta^{k,i}/\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i})$ and $\mathcal{I}^{\mathrm{low}} := \eta^{k,i}/\mathcal{J}_u^{\mathrm{low}}(u_h^{k,i})$. Since $\mathcal{J}_u^{\mathrm{low}}(u_h^{k,i}) \le \mathcal{J}_u(u_h^{k,i}) \le \mathcal{J}_u^{\mathrm{up}}(u_h^{k,i})$, the effectivity index for the original error measure $\mathcal{J}_u(u_h^{k,i})$, defined as $\mathcal{I} := \eta/\mathcal{J}_u(u_h^{k,i})$, lies between $\mathcal{I}^{\mathrm{up}}$ and $\mathcal{I}^{\mathrm{low}}$. For the three methods (FN, IN, and AIN), $\mathcal{I}^{\mathrm{up}}$ takes values very close to 1, see Figure 6.

Figures 7–11 display similar results for the choice $p = 1.5$. Almost no influence of the early stopping

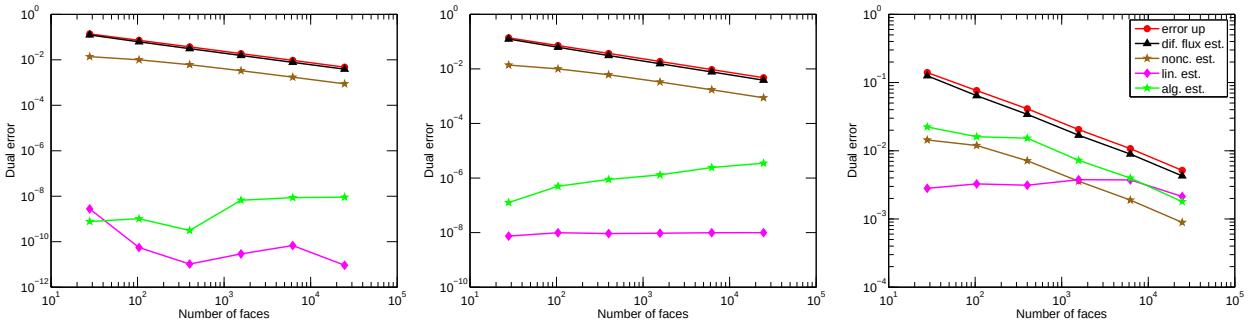

Figure 7: Error and estimators on uniformly refined meshes, case 1, $p = 1.5$. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

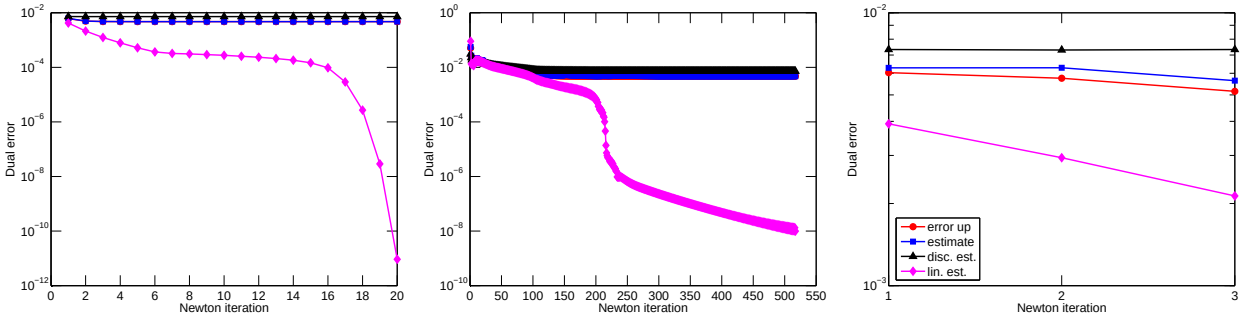

Figure 8: Error and estimators as a function of Newton iterations, case 1, $p = 1.5$, 6th level mesh. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

criteria can be seen in Figure 7. The nature of the nonlinearity seems different here from the case $p = 10$, as the Newton-iteration dependence curves of Figure 8 illustrate. In particular, using our stopping criteria avoids the useless waiting before the plateau has been overcome in the classical approaches (FN and IN). As before, these criteria also allow one to invest the right amount of CG iterations in each Newton step, as Figure 9 shows. The computational gains of our approach are important here, with one Newton iteration per refinement level up to the 5th level. Additionally, our approach only requires 122 total CG iterations on the 6th level mesh, in comparison to 3510 for FN and 7755 for IN, see Figure 10. Finally, the error and estimator distributions are similar to those observed in Figure 5, while the effectivity indices again take values very close to 1 for $\mathcal{I}^{\mathrm{up}}$, see Figure 11.

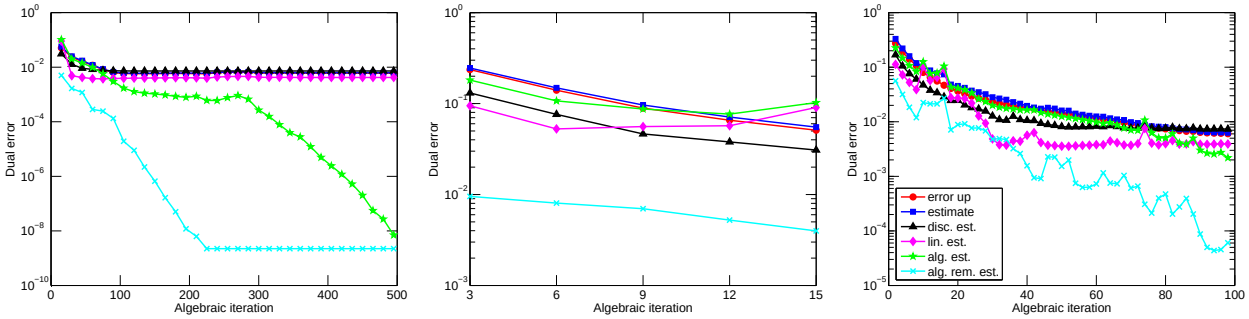

Figure 9: Error and estimators as a function of preconditioned CG iterations, case 1, $p = 1.5$, 6th level mesh, 1st Newton step. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

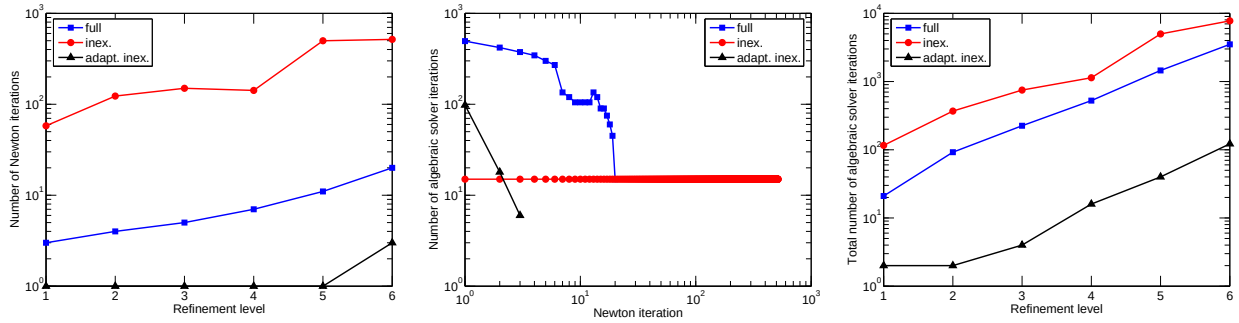

Figure 10: Number of Newton iterations per refinement level (left), number of linear solver iterations per Newton step on 6th level mesh (middle), and total number of linear solver iterations per refinement level (right). Case 1, $p = 1.5$

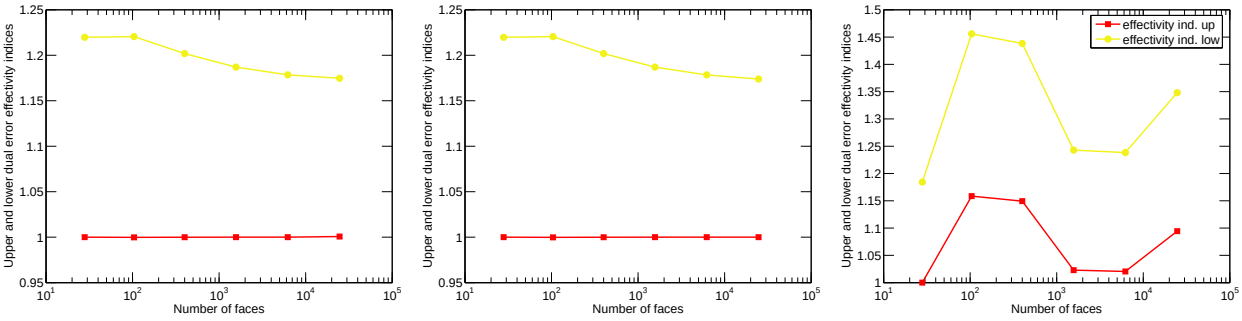

Figure 11: Upper and lower effectivity indices, case 1, $p = 1.5$. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

From the similarity of the results in the two settings of $p = 10$ and $p = 1.5$, compare Figure 1 with Figure 7 and Figure 6 with Figure 11, we can draw an experimental confirmation of the fact that our a posteriori error estimates are robust with respect to the size of the nonlinearity, represented by the exponent p.

7.2 A second test case

This test case is taken from Carstensen and Klose [5, Example 3]. We consider the L-shaped domain $\Omega := (-1,1)^2 \setminus [0,1] \times [-1,0]$ and prescribe the Dirichlet boundary condition and the source term f by the exact solution

$$u(r,\theta) = r^\delta \sin(\delta\theta).$$

Here, (r,θ) are the polar coordinates and $\delta := 7/8$. We consider the value $p = 4$ and, as in test case 1, we neglect the error stemming from inhomogeneous boundary conditions. The solution features a corner singularity with the regularity reported in Table 1. We only focus on our Adaptive Inexact Newton (AIN) method. We use the local criteria (3.10b) and (3.10c) (on the dual mesh $\mathcal{D}_h$) with $\gamma_{\mathrm{alg},D_e} = \gamma_{\mathrm{lin},D_e} = 1$ for all $e \in \mathcal{E}_h^{\mathrm{int}}$ and the local criterion (3.10a) with $\gamma_{\mathrm{rem},D_e} = 1$ for all $e \in \mathcal{E}_h^{\mathrm{int}}$. We perform both uniform and adaptive mesh refinement. The starting value u_h^0 is selected as above only on the coarsest mesh; on every subsequent refinement, this function is obtained from the converged solution $u_h^{k,i}$ on the previous mesh. Mesh adaptation is driven by our a posteriori error estimate $\eta^{k,i}$ of Theorem 3.4. All the elements where the estimate exceeds 50% of the maximal error are marked for refinement. Every marked element is refined regularly into four sub-simplices and the so-called longest edge refinement is used so as to recover a matching mesh (without hanging nodes).

Figure 12 plots the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i})$ and several estimators as before. In contrast to test case 1,

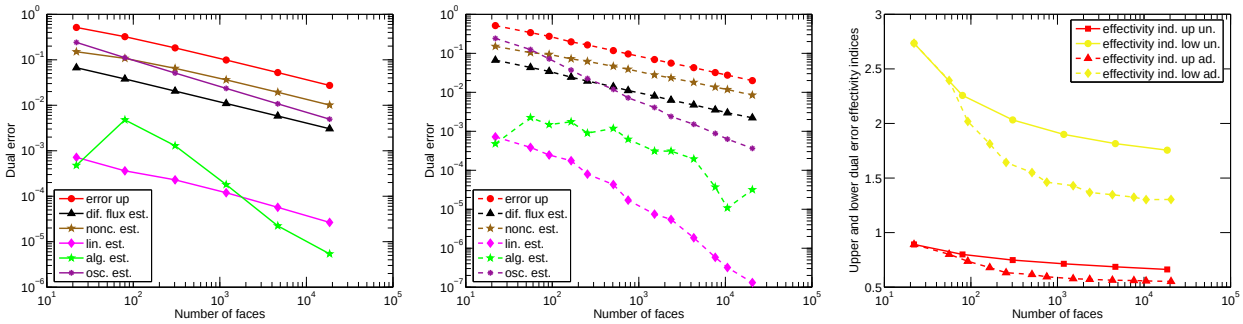

Figure 12: Error and estimators on uniformly (left) and adaptively (middle) refined meshes and upper and lower effectivity indices (right). Case 2, $p = 4$

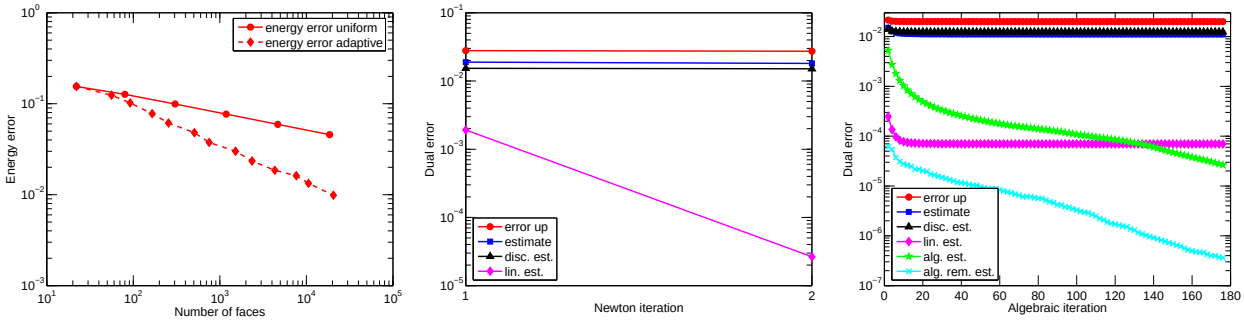

Figure 13: Energy error on uniformly and adaptively refined meshes (left), dual error and estimators as a function of Newton iterations, 13th level adaptively refined mesh (middle), and as a function of preconditioned CG iterations, same mesh, 1st Newton step (right). Case 2, $p = 4$

the data oscillation estimators (3.5b) are not zero and actually represent the most significant contribution to the overall error on the coarsest meshes. The linearization and algebraic estimators $\eta_{\text{lin}}^{k,i}$ and $\eta_{\text{alg}}^{k,i}$ are, as expected, only slightly below the other curves for uniform mesh refinement (a little more than in Section 7.1, as we employ here local and not global stopping criteria). An interesting phenomenon occurs for adaptive mesh refinement. Because of the corner singularity, the meshes are highly graded. Probably as a consequence, even if $\gamma_{\text{lin},D_e} = 1$, the linearization estimator $\eta_{\text{lin}}^{k,i}$ drops to values as low as 10^{-7}, whereas this estimator would not become so small if the global linearization stopping criterion (3.9c) was used.

Figure 13, left, traces the potential energy error $\|\nabla(u - u_h^{k,i})\|_p$ on both the uniformly and adaptively refined meshes. Here, we have observed that the usage of local stopping criteria (with the ensuing small values taken by the linearization estimator) is needed to achieve the quasi-optimal error decrease with adaptive mesh refinement, cf. Table 1. In particular, such a decrease does not appear if the global stopping criterion (3.9c) is employed, as the meshes are not sufficiently graded. As for the other graphs of Figure 13, similar conclusions as in Section 7.1 can be drawn.

Figure 14 illustrates the overall computational performance of AIN for this second test case. As few as 2 Newton iterations per refinement level are used except for the initial meshes. The efficiency of the AIN combined with adaptive mesh refinement is best appreciated when evaluating the total number of linear solver iterations per refinement level; only a very mild increase is observed for adaptive mesh refinement.

Finally, in Figure 15, we plot the distribution of the estimate $\eta^{k,i}$ and of the error measure $\mathcal{J}_u^{\text{up}}(u_h^{k,i})$ on the 5th level adaptively refined mesh. As before, even in the presence of linearization and algebraic errors, the overall error distribution is predicted very well, while the mesh has been refined around the corner singularity.

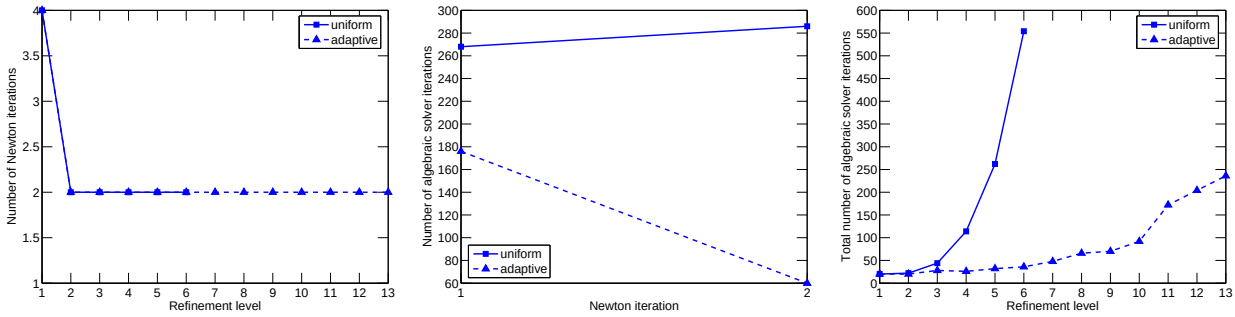

Figure 14: Number of Newton iterations per refinement level (left), number of linear solver iterations per Newton step (6th level uniformly refined mesh and 13th level adaptively refined mesh) (middle), and total number of linear solver iterations per refinement level (right). Case 2, $p = 4$

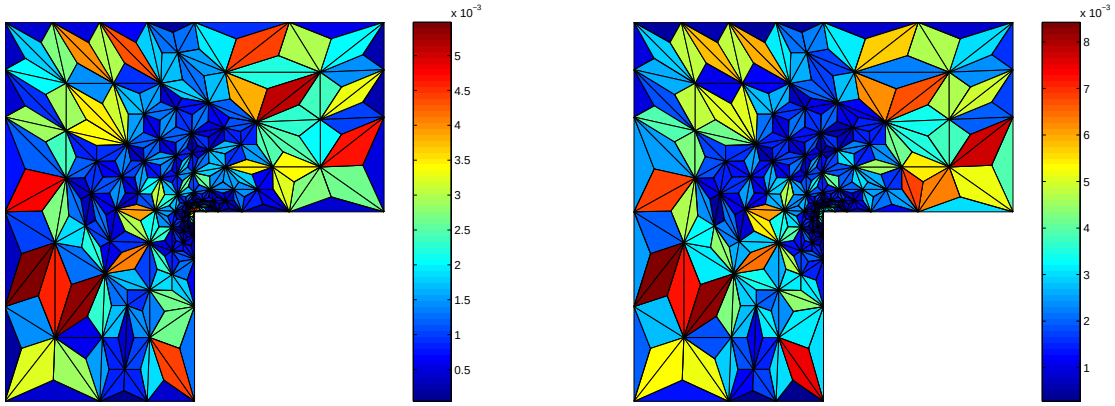

Figure 15: Estimated (left) and actual (right) error distribution, case 2, $p = 4$, 5th level adaptively refined mesh

References

[1] A. Bakushinsky and A. Smirnova, *A posteriori stopping rule for regularized fixed point iterations*, Nonlinear Anal., 64 (2006), pp. 1255–1261.

[2] F. Bauer and T. Hohage, *A Lepskij-type stopping rule for regularized Newton methods*, Inverse Problems, 21 (2005), pp. 1975–1991.

[3] R. Becker, C. Johnson, and R. Rannacher, *Adaptive error control for multigrid finite element methods*, Computing, 55 (1995), pp. 271–288.

[4] D. Braess and J. Schöberl, *Equilibrated residual error estimator for edge elements*, Math. Comp., 77 (2008), pp. 651–672.

[5] C. Carstensen and R. Klose, *A posteriori finite element error control for the p-Laplace problem*, SIAM J. Sci. Comput., 25 (2003), pp. 792–814.

[6] A. L. Chaillou and M. Suri, *Computable error estimators for the approximation of nonlinear problems by linearized models*, Comput. Methods Appl. Mech. Engrg., 196 (2006), pp. 210–224.

[7] ———, *A posteriori estimation of the linearization error for strongly monotone nonlinear operators*, J. Comput. Appl. Math., 205 (2007), pp. 72–87.

[8] S.-K. Chua and R. L. Wheeden, *Estimates of best constants for weighted Poincaré inequalities on convex domains*, Proc. London Math. Soc. (3), 93 (2006), pp. 197–226.

[9] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–75.

[10] P. Deuflhard, *Newton methods for nonlinear problems*, vol. 35 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2004. Affine invariance and adaptive algorithms.

[11] D. A. Di Pietro and A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*, vol. 69 of Mathématiques & Applications, Springer-Verlag, Berlin, 2011.

[12] S. C. Eisenstat and H. F. Walker, *Globally convergent inexact Newton methods*, SIAM J. Optim., 4 (1994), pp. 393–422.

[13] ——, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32. Special issue on iterative methods in numerical linear algebra (Breckenridge, CO, 1994).

[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), pp. 2782–2795.

[15] A. Ern and M. Vohralík, *A posteriori error estimation based on potential and flux reconstruction for the heat equation*, SIAM J. Numer. Anal., 48 (2010), pp. 198–223.

[16] W. B. Gragg and R. A. Tapia, *Optimal error bounds for the Newton-Kantorovich theorem*, SIAM J. Numer. Anal., 11 (1974), pp. 10–13.

[17] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590.

[18] ——, *Balancing discretization and algebraic errors in finite element computations.* In preparation, 2011.

[19] L. V. Kantorovich, *Functional analysis and applied mathematics*, Uspekhi Mat. Nauk, 3 (1948), pp. 89–185.

[20] L. V. Kantorovich and G. P. Akilov, *Functional analysis in normed spaces*, Translated from the Russian by D. E. Brown. Edited by A. P. Robertson. International Series of Monographs in Pure and Applied Mathematics, Vol. 46, The Macmillan Co., New York, 1964.

[21] A. Lechleiter and A. Rieder, *Towards a general convergence theory for inexact Newton regularizations*, Numer. Math., 114 (2010), pp. 521–548.

[22] J. Leray and J.-L. Lions, *Quelques résulatats de Višik sur les problèmes elliptiques nonlinéaires par les méthodes de Minty-Browder*, Bull. Soc. Math. France, 93 (1965), pp. 97–107.

[23] R. Luce and B. I. Wohlmuth, *A local a posteriori error estimator based on equilibrated fluxes*, SIAM J. Numer. Anal., 42 (2004), pp. 1394–1414.

[24] D. Meidner, R. Rannacher, and J. Vihharev, *Goal-oriented error control of the iterative solution of finite element equations*, J. Numer. Math., 17 (2009), pp. 143–172.

[25] I. Moret, *A Kantorovich-type theorem for inexact Newton methods*, Numer. Funct. Anal. Optim., 10 (1989), pp. 351–365.

[26] J. M. Ortega, *The Newton-Kantorovich theorem*, Amer. Math. Monthly, 75 (1968), pp. 658–660.

[27] F.-A. Potra and V. Pták, *Sharp error bounds for Newton's process*, Numer. Math., 34 (1980), pp. 63–72.

[28] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math., 5 (1947), pp. 241–269.

[29] R. Rannacher, A. Westenberger, and W. Wollner, *Adaptive finite element solution of eigenvalue problems: balancing of discretization and iteration error*, J. Numer. Math., 18 (2010), pp. 303–327.

[30] R. Verfürth, *A note on polynomial approximation in Sobolev spaces*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 715–719.

[31] M. Weiser, A. Schiela, and P. Deuflhard, *Asymptotic mesh independence of Newton's method revisited*, SIAM J. Numer. Anal., 42 (2005), pp. 1830–1845.

[32] T. Yamamoto, *A method for finding sharp error bounds for Newton's method under the Kantorovich assumptions*, Numer. Math., 49 (1986), pp. 203–220.