# Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs

Alexandre Ern, Martin Vohralík

# Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*

Alexandre Ern† Martin Vohralík‡

October 28, 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. We derive adaptive stopping criteria for both iterative solvers. Our 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 hinge on equilibrated flux reconstructions and yield a general framework. We show how to apply this framework to various discretization schemes like finite elements, nonconforming finite elements, discontinuous Galerkin, finite volumes, and mixed finite elements; to different linearizations like fixed point and Newton; and to arbitrary iterative linear solvers. Numerical experiments for the $p$-Laplacian illustrate the tight overall error control and important computational savings achieved in our approach.

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

## 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. A classical solution algorithm consists in forming a system of linear algebraic equations

$$\mathbb{A}^{k-1} U^k = F^{k-1} \tag{1.2}$$

by a given linearization on each iteration step $k \geq 1$. Then some iterative algebraic solver is applied to (1.2), yielding on step $i \geq 0$ an approximation $U^{k,i}$ to $U^k$ satisfying

$$\mathbb{A}^{k-1} U^{k,i} = F^{k-1} - R^{k,i}, \tag{1.3}$$

with $R^{k,i} \in \mathbb{R}^N$ the algebraic residual vector.

*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).

‡INRIA Paris-Rocquencourt, B.P. 105, 78153 Le Chesnay, France (martin.vohralik@inria.fr).

If the algebraic solve of (1.2) is done "exactly", i.e. $R^{k,i} = 0$ (typically up to computer working precision), an *exact iterative linearization* is obtained. Probably the most well-known example is the Newton method, where

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

Convergence and a priori error estimates for the Newton method have been obtained by Kantorovich [25] and Ortega [32]. 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 [21] and improved by Potra and Pták [33] and Yamamoto [44], see also references therein.

The Newton method can be computationally demanding because of the solve of the linear system (1.2). The *inexact Newton method* is a popular approach to speed it up. It has been used in practice for decades and studied theoretically in many papers. In particular, Eisenstat and Walker [15] have shown the convergence, a posteriori error estimates were proved by Moret [31], and adaptive algorithms were derived by Deuflhard [12, Section 1.2.3], see also 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* in inexact linearizations. Let $u$ be the solution of (1.5) and let $u_h^{k,i}$ be the approximation to $u$ obtained by the discretization scheme on the $k$-th nonlinear solver step and the $i$-th linear solver step, whose algebraic representation is the vector $U^{k,i}$ of (1.3). Our second goal is to obtain *guaranteed* (without undetermined constants) *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 our approach 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 derived in various specific situations. Verfürth [38] developed a general framework for reliable and efficient a posteriori estimates in the finite element setting. For the $p$-Laplacian, quite tight guaranteed upper bounds have been obtained by Carstensen and Klose [8], convergence of an adaptive finite element method was first proven by Veeser [37] for the energy norm and a quasi-optimal rate was recently obtained by Belenki *et al.* [3] for an error measure related to the quasi-norm of Barrett and Liu [1]. Other discretization schemes were also studied; let us mention, in particular, Creusé *et al.* [11] for mixed finite elements and the $p$-Laplacian, Houston *et al.* [23] for the discontinuous Galerkin method and quasi-linear diffusion, and Kim [26] for locally conservative methods and strongly monotone problems. Estimates and stopping criteria *independently* for linear and nonlinear solvers were proposed by Becker *et al.* [2], Chaillou and Suri [9], and, more closely to the present approach, in [24, 16], see also the references therein. Both linearization and algebraic errors are simultaneously addressed in the context of goal-oriented error estimation by Rannacher *et al.* [35] and Meidner *et al.* [30], see also the survey by Strakoš and Liesen [36].

We are not aware of estimates of the error between $u$ and $u_h^{k,i}$ which provide, at the same time, a *guaranteed 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 through three suitable *flux reconstructions* following the spirit of Prager and Synge [34], see [18, 22] and references therein for recent contributions. We describe a possible handling of the algebraic error in Section 4, leading to quasi-equilibrated fluxes. The distinction of error components 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.

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 [9, 16] 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 also locally around each mesh element for an easily computable upper bound of our error measure evaluating the $[L^q(\Omega)]^d$ *distance of the fluxes*. Overall, our estimates seem to give a very tight local control of the $[L^q(\Omega)]^d$ error in the fluxes. For Leray–Lions problems, our results thus complement those obtained in the quasi-norm setting. Convergence and optimality of our adaptive inexact Newton approach shall be addressed elsewhere.

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. These assumptions are verified in Section 6 for various discretization schemes, the Newton and fixed point nonlinear solvers, and an arbitrary iterative linear solver. In Section 7, we study numerically the behavior of our a posteriori estimates and the computational gains of our stopping criteria for the $p$-Laplacian, the Crouzeix–Raviart nonconforming finite element method, the Newton linearization, and the conjugate gradient algebraic solver. An example of application of the present framework to two-phase flow simulation can be found in [41]. Finally, we draw some conclusions in Section 8.

# 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 general form $\boldsymbol{\sigma}(\mathbf{x}, v, \boldsymbol{\xi}) = \underline{\mathbf{A}}(\mathbf{x}, v, \boldsymbol{\xi})\boldsymbol{\xi}$, for all $(\mathbf{x}, v, \boldsymbol{\xi}) \in \Omega\times\mathbb{R}\times\mathbb{R}^d$, 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

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

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

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

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 [27], 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 $\alpha_0$ and $\alpha_1$ 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 henceforth 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

$$(\boldsymbol{\sigma}(u, \nabla u), \nabla v) = (f, v) \qquad \forall v \in V. \tag{2.4}$$

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. We assume that there exists a unique weak solution to (2.4). Owing to the above assumptions and to (2.4), the flux $-\boldsymbol{\sigma}(u, \nabla u)$ is then in the space $\mathbf{H}^q(\mathrm{div}, \Omega)$ spanned by the functions in $[L^q(\Omega)]^d$ with weak divergence in $L^q(\Omega)$.

## 2.2 Discrete setting

Let $\mathcal{T}_h$ be a simplicial mesh of $\Omega$. 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, $0 \le l \le d-1$. A generic element of $\mathcal{T}_h$ is denoted $K$ 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 $e$ and its diameter by $h_e$. The faces of an element $K$ are collected in the set $\mathcal{E}_K$. For any $K \in \mathcal{T}_h$, $\mathfrak{T}_K$ collects the elements $K' \in \mathcal{T}_h$ which share at least a vertex with $K$. Similarly, $\mathfrak{E}_K$ collects the faces which share at least a vertex with $K$, and we set $\mathfrak{E}_K^{\mathrm{int}} := \mathfrak{E}_K \cap \mathcal{E}_h^{\mathrm{int}}$. For any $e \in \mathcal{E}_h$, $\mathbf{n}_e$ stands for the unit normal vector to $e$ (the orientation is irrelevant, but fixed, for all $e \in \mathcal{E}_h^{\mathrm{int}}$ and points outward $\Omega$ for all $e \in \mathcal{E}_h^{\mathrm{ext}}$) and, for any $K \in \mathcal{T}_h$, $\mathbf{n}_K$ stands for the outward unit normal vector to $K$.

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). Suppose that we are on step $k$, $k \ge 1$, of the nonlinear solver and on step $i$, $i \ge 0$, of the linear solver. This corresponds to problem (1.3). We denote $u_h^{k,i}$ 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

$$V(\mathcal{T}_h) := \{v \in L^p(\Omega),\, v|_K \in W^{1,p}(K) \quad \forall K \in \mathcal{T}_h\}. \tag{2.5}$$

In what follows, for a function $v \in V(\mathcal{T}_h)$, $\nabla v$ 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 (evaluated along $\mathbf{n}_e$) 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., [14, Lemma 1.23].

Separately from $u_h^{k,i}$, we also consider a discrete gradient $\mathbf{g}_h^{k,i} \in [L^p(\Omega)]^d$. This allows us to handle a wide class of discretization schemes in a unified setting. For conforming schemes, $\mathbf{g}_h^{k,i}$ is obtained by applying the usual gradient to $u_h^{k,i}$; for various nonconforming schemes, the broken gradient can be used instead, but some schemes employ a more elaborate construction of $\mathbf{g}_h^{k,i}$, taking into account, e.g., the jumps of $u_h^{k,i}$. In all cases, we require that whenever $u_h^{k,i} \in V$, there holds $\mathbf{g}_h^{k,i} = \nabla u_h^{k,i}$.

## 2.3 Error measure

The error between the exact solution $u$ of (2.4) and the approximate solution $u_h^{k,i}$ is measured as

$$\mathcal{J}_u(u_h^{k,i}, \mathbf{g}_h^{k,i}) := \mathcal{J}_{u,\mathrm{F}}(u_h^{k,i}, \mathbf{g}_h^{k,i}) + \mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}), \tag{2.6}$$

where

$$\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i}, \mathbf{g}_h^{k,i}) := \sup_{\varphi \in V;\, \|\nabla\varphi\|_p = 1} \big(\boldsymbol{\sigma}(u, \nabla u) - \boldsymbol{\sigma}(u_h^{k,i}, \mathbf{g}_h^{k,i}), \nabla\varphi\big), \tag{2.7a}$$

$$\mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}) := \left\{ \sum_{K \in \mathcal{T}_h} \sum_{e \in \mathcal{E}_K} \alpha_e^s h_e^{1-s} \|[\![u - u_h^{k,i}]\!]\|_{s,e}^s \right\}^{\frac{1}{q}}. \tag{2.7b}$$

The quantity $\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i},\mathbf{g}_h^{k,i})$ measures the error in the fluxes and represents the dual norm of the residual of (2.4). This error measure has been considered by Chaillou and Suri [9] and in [16] for conforming discretizations. Owing to the well-posedness of (2.4) and the above requirement on $\mathbf{g}_h^{k,i}$, whenever $u_h^{k,i}\in V$, $\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i},\mathbf{g}_h^{k,i})=0$ 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 $V$. A specific value for the weights $\alpha_e>0$ and the exponent $s\geq 1$ is only needed in Section 6.3.4 below; otherwise 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 $\mathcal{J}_u(u_h^{k,i},\mathbf{g}_h^{k,i})=0$ if and only if $u_h^{k,i}=u$ and $\mathbf{g}_h^{k,i}=\nabla u$.

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

$$\mathcal{J}_u(u_h^{k,i},\mathbf{g}_h^{k,i})\leq \mathcal{J}_u^{\mathrm{up}}(u_h^{k,i},\mathbf{g}_h^{k,i}):=\|\boldsymbol{\sigma}(u,\nabla u)-\boldsymbol{\sigma}(u_h^{k,i},\mathbf{g}_h^{k,i})\|_q+\mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}),\tag{2.8}$$

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

# 3 A posteriori error estimates and the adaptive inexact Newton algorithm

In this section, we present our a posteriori error estimates and the inexact Newton algorithm with adaptive stopping criteria. We proceed generally, with a given discrete potential $u_h^{k,i}\in V(\mathcal{T}_h)$ and the corresponding discrete gradient $\mathbf{g}_h^{k,i}\in[L^p(\Omega)]^d$, $k\geq 1$, $i\geq 0$, not linked to any particular discretization scheme or to any iterative nonlinear or linear solvers. Examples of application are given in Section 6. 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 [34], Luce and Wohlmuth [28], Braess and Schöberl [4], or the unified approaches in [18, 22] and the references therein. In practice, see Section 6, 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 [24, 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 [24, Section 7.2].

**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 whose elementwise mass balance misfit is merely* $\rho_h^{k,i}$.

### 3.1 Guaranteed a posteriori error estimate

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

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

where $\varphi_K$ denotes the mean value of $\varphi$ in $K$. Since simplices are convex, there holds $C_{\mathrm{P},p} = \pi^{-\frac{2}{p}} d^{\frac{1}{2}-\frac{1}{p}}$ for $p \ge 2$, see Verfürth [39], and $C_{\mathrm{P},p} = p^{\frac{1}{p}} 2^{\frac{(p-1)}{p}}$ for all $p \in (1, +\infty)$, see Chua and Wheeden [10]. The generalized Friedrichs inequality states that

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

In what follows, we denote our estimators in the form $\eta^{k,i}_{\cdot,K}$ where $k \ge 1$ stands for the nonlinear solver step, $i \ge 0$ for the linear solver step, and $K \in \mathcal{T}_h$ for the mesh element. We define global versions of these estimators as $\eta^{k,i}_{\cdot} := \left\{\sum_{K\in\mathcal{T}_h} \left(\eta^{k,i}_{\cdot,K}\right)^q\right\}^{1/q}$. Our main result on the a posteriori error estimate is:

**Theorem 3.4** (Guaranteed upper bound)**.** *Let $u \in V$ solve (2.4), let $u_h^{k,i} \in V(\mathcal{T}_h)$ and the corresponding $\mathbf{g}_h^{k,i} \in [L^p(\Omega)]^d$ be arbitrary, and let Assumption 3.1 hold. For any $K \in \mathcal{T}_h$, define respectively the* flux *and the* nonconformity estimators *as*

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

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

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

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

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

*Then,*

$$\mathcal{J}_u(u_h^{k,i}, \mathbf{g}_h^{k,i}) \le \eta^{k,i} := \eta^{k,i}_{\mathrm{F}} + \eta^{k,i}_{\mathrm{NC}} + \eta^{k,i}_{\mathrm{rem}} + \eta^{k,i}_{\mathrm{osc}}. \tag{3.6}$$

*Proof.* Taking into account that $[\![u]\!] = 0$ for all $e \in \mathcal{E}_h$, it is clear that $\mathcal{J}_{u,\mathrm{NC}}(u_h^{k,i}) = \eta^{k,i}_{\mathrm{NC}}$. We are thus left with bounding $\mathcal{J}_{u,\mathrm{F}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$. 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 (2.4) and adding and subtracting $(\mathbf{t}_h^{k,i}, \nabla\varphi)$, we infer

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

The Hölder inequality yields

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

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

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

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 error flux reconstructions)**.** *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 $-\boldsymbol{\sigma}(u_h^{k,i}, \mathbf{g}_h^{k,i})$, $\mathbf{l}_h^{k,i}$ represents the *linearization error*, and $\mathbf{a}_h^{k,i}$ 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 and nonlinear solver at hand, see Section 6.

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

**Theorem 3.6** (A posteriori error estimate distinguishing the error components)**.** *Let* $u \in V$ *solve* (2.4) *and let* $u_h^{k,i} \in V(\mathcal{T}_h)$ *and the corresponding* $\mathbf{g}_h^{k,i} \in [L^p(\Omega)]^d$ *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*

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

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

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

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

*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,*

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

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

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

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 Adaptive inexact Newton method

We are now ready to present our adaptive inexact Newton method with a posteriori stopping criteria for the linear and nonlinear solvers. 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. Owing to the presence of the function $\rho_h^{k,i}$, we introduce a third (balancing) requirement, namely that the algebraic remainder estimator is sufficiently small with respect to the three other estimators. The adaptive inexact Newton algorithm for (1.1) reads:

**Algorithm 3.7** (Adaptive inexact Newton method)**.** 1. *Choose an initial vector* $U^0 \in \mathbb{R}^N$*. Set* $k := 1$.

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

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

   *(b) Perform* $\nu > 0$ *steps of a chosen iterative linear solver for the solution of the linear system (1.2), starting from the vector* $U^{k,i}$. *This yields an approximation* $U^{k,i+\nu}$ *to* $U^k$ *which satisfies*

$$\mathbb{A}^{k-1}U^{k,i+\nu} = F^{k-1} - R^{k,i+\nu}, \tag{3.9}$$

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

$$\eta_{\mathrm{rem}}^{k,i} \le \gamma_{\mathrm{rem}} \max\{\eta_{\mathrm{disc}}^{k,i}, \eta_{\mathrm{lin}}^{k,i}, \eta_{\mathrm{alg}}^{k,i}\}. \tag{3.10}$$

   *(c) Check the convergence criterion for the linear solver in the form*

$$\eta_{\mathrm{alg}}^{k,i} \le \gamma_{\mathrm{alg}} \max\{\eta_{\mathrm{disc}}^{k,i}, \eta_{\mathrm{lin}}^{k,i}\}. \tag{3.11}$$

   *If satisfied, set* $U^k := U^{k,i}$. *If not, set* $i := i+\nu$ *and go back to step 3b.*

4. *Check the convergence criterion for the nonlinear solver in the form*

$$\eta_{\mathrm{lin}}^{k,i} \le \gamma_{\mathrm{lin}} \eta_{\mathrm{disc}}^{k,i}. \tag{3.12}$$

*If satisfied, finish. If not, set* $k := k+1$ *and go back to step 2.*

Above, $\gamma_{\mathrm{rem}}$, $\gamma_{\mathrm{alg}}$, and $\gamma_{\mathrm{lin}}$ are positive user-given weights, typically of order 0.1, representing the relative size (percentage) of the algebraic remainder, algebraic, and linearization errors. The balancing and *stopping criteria* (3.10)–(3.12) are *global* in the sense that they are evaluated over all mesh elements. They are sufficient to establish the *global efficiency* of our error estimators, see Theorem 5.4 below. Alternatively, *local stopping criteria* are elementwise equivalents in the form

$$\eta_{\mathrm{rem},K}^{k,i} \le \gamma_{\mathrm{rem},K} \max\{\eta_{\mathrm{disc},K}^{k,i}, \eta_{\mathrm{lin},K}^{k,i}, \eta_{\mathrm{alg},K}^{k,i}\} \qquad \forall K \in \mathcal{T}_h, \tag{3.13}$$

$$\eta_{\mathrm{alg},K}^{k,i} \le \gamma_{\mathrm{alg},K} \max\{\eta_{\mathrm{disc},K}^{k,i}, \eta_{\mathrm{lin},K}^{k,i}\} \qquad \forall K \in \mathcal{T}_h, \tag{3.14}$$

$$\eta_{\mathrm{lin},K}^{k,i} \le \gamma_{\mathrm{lin},K} \eta_{\mathrm{disc},K}^{k,i} \qquad \forall K \in \mathcal{T}_h, \tag{3.15}$$

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 below, 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 construct the algebraic remainder $\rho_h^{k,i}$ and the algebraic error flux reconstruction $\mathbf{a}_h^{k,i}$. To do so, we suppose that the sum $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$ of the flux reconstructions $\mathbf{d}_h^{k,i}$ and $\mathbf{l}_h^{k,i}$ satisfies:

**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 3.7 where the linear system (1.2) for $k \ge 1$ 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 Section 6 how the (piecewise polynomial) function $r_h^{k,i}$ of (4.1) can be constructed from the components of $R^{k,i}$ for various discretizations. We then define:

**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* $\nu$ *additional steps of the linear solver, yielding (3.9) and* $(\mathbf{d}_h^{k,i+\nu} + \mathbf{l}_h^{k,i+\nu})$, $r_h^{k,i+\nu}$ *satisfying (4.1) with* $i+\nu$ *in place of* $i$. *Set*

$$\mathbf{a}_h^{k,i} := (\mathbf{d}_h^{k,i+\nu} + \mathbf{l}_h^{k,i+\nu}) - (\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i}), \tag{4.2a}$$

$$\rho_h^{k,i} := r_h^{k,i+\nu}. \tag{4.2b}$$

In practice, the parameter $\nu$ can be determined adaptively by increasing its value until satisfying (3.10) or (3.13). We emphasize that this construction is independent of the actual linear solver. Importantly, the following 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.*

# 5 Local and global efficiency and robustness

We prove in this section 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 (3.10)–(3.12) or (3.13)–(3.15).

## 5.1 Local approximation property

To proceed generally, we make one last assumption on the discretization error flux reconstruction $\mathbf{d}_h^{k,i}$. Define

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

Let $\eta_{\cdot,\mathfrak{T}_K}^{k,i} := \left\{ \sum_{K' \in \mathfrak{T}_K} (\eta_{\cdot,K'}^{k,i})^q \right\}^{\frac{1}{q}}$ 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 $\Omega$, 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*

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

**Remark 5.2** (Tighter approximation property). *In most cases, it is actually possible to prove* $\| \overline{\boldsymbol{\sigma}}_h^{k,i} + \mathbf{d}_h^{k,i} \|_{q,K} \lesssim \eta_{\sharp,K}^{k,i}$. *The term $\eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}$ appears for conforming finite elements in the lowest-order setting $m = 1$ and $l = 0$ in Section 6.2.4, while $\eta_{\mathrm{NC},\mathfrak{T}_K}^{k,i}$ appears for interior penalty discontinuous Galerkin and quasi-linear diffusion, see Section 6.3.4.*

## 5.2 Local efficiency

Our local efficiency result is achieved with respect to the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$ defined by (2.8), which we localize around any $K \in \mathcal{T}_h$ as $\mathcal{J}_{u,\mathfrak{T}_K}^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i}) := \| \boldsymbol{\sigma}(u, \nabla u) - \boldsymbol{\sigma}(u_h^{k,i}, \mathbf{g}_h^{k,i}) \|_{q,\mathfrak{T}_K} + \eta_{\mathrm{NC},\mathfrak{T}_K}^{k,i}$.

**Theorem 5.3** (Local efficiency). *Let $u \in V$ solve (2.4) and let $u_h^{k,i} \in V(\mathcal{T}_h)$ and the corresponding $\mathbf{g}_h^{k,i} \in [L^p(\Omega)]^d$ be arbitrary. Let the local stopping criteria (3.13)–(3.15) be satisfied. Then, under Assumption 5.1, there holds, for all $K \in \mathcal{T}_h$,*

$$\eta_{\mathrm{disc},K}^{k,i} + \eta_{\mathrm{lin},K}^{k,i} + \eta_{\mathrm{alg},K}^{k,i} + \eta_{\mathrm{rem},K}^{k,i} \lesssim \mathcal{J}_{u,\mathfrak{T}_K}^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i}) + \eta_{\mathrm{quad},\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}. \tag{5.3}$$

*Proof.* Let $K \in \mathcal{T}_h$ be fixed. Owing to the local criteria (3.13)–(3.15), 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 $\eta_{\mathrm{disc},K}^{k,i} \lesssim \eta_{\sharp,K}^{k,i} + \eta_{\mathrm{NC},\mathfrak{T}_K}^{k,i} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}$, whence

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

Now, the inequalities (A.6) and (A.7) from [16, Proof of Lemma 4.3] together with the triangle inequality yield

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

whence the assertion of the theorem. ∎

### 5.3 Global efficiency and robustness

Proceeding as above (while relying on (A.10) and (A.11) from [16, Proof of Lemma 4.7]) yields our main result for global efficiency and robustness with respect to the original error measure $\mathcal{J}_u(u_h^{k,i}, \mathbf{g}_h^{k,i})$.

**Theorem 5.4** (Global efficiency and robustness)**.** *Let $u \in V$ solve (2.4) and let $u_h^{k,i} \in V(\mathcal{T}_h)$ and the corresponding $\mathbf{g}_h^{k,i} \in [L^p(\Omega)]^d$ be arbitrary. Let the global stopping criteria (3.10)–(3.12) be satisfied. Then, under Assumption 5.1, there holds*

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

**Remark 5.5** (Comparison with [16])**.** *In [16], the linearization stopping parameters $\gamma_{\mathrm{lin},K}$ (or $\gamma_{\mathrm{lin}}$) 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 Applications

We show here how the above developments apply to various discretizations and to Newton and fixed point linearizations (recall that any algebraic solver is admissible). This consists in specifying the approximate gradient $\mathbf{g}_h^{k,i}$, flux reconstructions $\mathbf{d}_h^{k,i}$ and $\mathbf{l}_h^{k,i}$, data approximation $f_h$, polynomial approximation $\overline{\boldsymbol{\sigma}}_h^{k,i}$, residual function $r_h^{k,i}$, and in verifying Assumptions 3.5(iii), 4.1, and 5.1.

We first recall some discrete subspaces of $\mathbf{H}^q(\mathrm{div},\Omega)$. For an integer $l \geq 0$, let $\mathbb{P}_l(\mathcal{T}_h)$ denote the broken polynomial space spanned by $v_h|_K \in \mathbb{P}_l(K)$ for all $K \in \mathcal{T}_h$. For $K \in \mathcal{T}_h$ and $l \geq 0$, let $\mathbf{RTN}_l(K) := [\mathbb{P}_l(K)]^d + \mathbf{x}\mathbb{P}_l(K)$ be the Raviart–Thomas–Nédélec finite element space of order $l$. We then set $\mathbf{RTN}_l^{-1}(\mathcal{T}_h) := \{\mathbf{v}_h \in [L^q(\Omega)]^d; \mathbf{v}_h|_K \in \mathbf{RTN}_l(K) \quad \forall K \in \mathcal{T}_h\}$ and $\mathbf{RTN}_l(\mathcal{T}_h) := \mathbf{RTN}_l^{-1}(\mathcal{T}_h) \cap \mathbf{H}^q(\mathrm{div},\Omega)$; we will use $\mathbf{RTN}_l(\mathcal{T}_h)$ to reconstruct the fluxes $\mathbf{d}_h^{k,i}$ and $\mathbf{l}_h^{k,i}$ (and consequently $\mathbf{a}_h^{k,i}$ by (4.2a)). Functions $\mathbf{v}_h \in \mathbf{RTN}_l(K)$ are such that, cf. Brezzi and Fortin [5], $\nabla\cdot\mathbf{v}_h \in \mathbb{P}_l(K)$ and $\mathbf{v}_h\cdot\mathbf{n}_e \in \mathbb{P}_l(e)$ for all $e \in \mathcal{E}_K$, and functions in $\mathbf{RTN}_l(\mathcal{T}_h)$ have a continuous normal component across interfaces. We use a similar notation for these spaces on various patches of elements.

Next, let $\mathbf{I}_l^{\mathrm{RTN}}$ stand for the broken Raviart–Thomas–Nédélec interpolation operator; for a smooth enough function $\mathbf{v}$, $\mathbf{I}_l^{\mathrm{RTN}}\mathbf{v} \in \mathbf{RTN}_l^{-1}(\mathcal{T}_h)$ is such that, for all $K \in \mathcal{T}_h$, letting $\langle w, v\rangle_e$ stand for $\int_e w(s)v(s)\,\mathrm{d}s$,

$$\langle(\mathbf{I}_l^{\mathrm{RTN}}\mathbf{v} - \mathbf{v})|_K\cdot\mathbf{n}_e, q_h\rangle_e = 0 \qquad \forall e \in \mathcal{E}_K,\ \forall q_h \in \mathbb{P}_l(e), \tag{6.1a}$$

$$(\mathbf{I}_l^{\mathrm{RTN}}\mathbf{v} - \mathbf{v}, \mathbf{r}_h)_K = 0 \qquad \forall \mathbf{r}_h \in [\mathbb{P}_{l-1}(K)]^d. \tag{6.1b}$$

Finally, for $\phi \in L^1(\Omega)$, $\Pi_l\phi \in \mathbb{P}_l(\mathcal{T}_h)$ is such that $(\phi - \Pi_l\phi, v_h) = 0$ for all $v_h \in \mathbb{P}_l(\mathcal{T}_h)$; $\mathbf{\Pi}_l$ is the operator acting componentwise as $\Pi_l$ on vector-valued functions.

### 6.1 Nonconforming finite elements

We treat here the discretization of problem (2.4) by lowest-order nonconforming finite elements.

#### 6.1.1 Discretization

The Crouzeix–Raviart finite element space $V_h$ is spanned by piecewise affine polynomials on $\mathcal{T}_h$ such that the interface jumps and boundary values have zero mean value over the corresponding face. The discretization of problem (2.4) reads, with $f_h := \Pi_0 f$: find $u_h \in V_h$ such that

$$(\boldsymbol{\sigma}(u_h, \nabla u_h), \nabla v_h) = (f_h, v_h) \qquad \forall v_h \in V_h. \tag{6.2}$$

The basis functions in $V_h$ are associated with the interfaces and are denoted $\{\psi_e\}_{e\in\mathcal{E}_h^{\mathrm{int}}}$. Testing (6.2) against these functions yields the nonlinear algebraic system (1.1).

### 6.1.2 Linearization

Let $u_h^0 \in V_h$, fixing the initial vector $U^0$ in Algorithm 3.7. The linearization of (6.2), for $k \geq 1$, reads: find $u_h^k \in V_h$ such that

$$(\boldsymbol{\sigma}^{k-1}(u_h^k, \nabla u_h^k), \nabla \psi_e) = (f_h, \psi_e) \qquad \forall e \in \mathcal{E}_h^{\mathrm{int}}, \tag{6.3}$$

which is the functional form of the algebraic system (1.2). Two common ways to define the flux function $\boldsymbol{\sigma}^{k-1}$ are the fixed point linearization where

$$\boldsymbol{\sigma}^{k-1}(v, \boldsymbol{\xi}) := \underline{\mathbf{A}}(u_h^{k-1}, \nabla u_h^{k-1})\boldsymbol{\xi}, \tag{6.4}$$

and the Newton linearization where

$$\begin{aligned}\boldsymbol{\sigma}^{k-1}(v, \boldsymbol{\xi}) := \underline{\mathbf{A}}(u_h^{k-1}, \nabla u_h^{k-1})\boldsymbol{\xi} + (v - u_h^{k-1})\partial_v \underline{\mathbf{A}}(u_h^{k-1}, \nabla u_h^{k-1})\nabla u_h^{k-1} \\ + (\partial_{\boldsymbol{\xi}} \underline{\mathbf{A}}(u_h^{k-1}, \nabla u_h^{k-1}) \cdot \nabla u_h^{k-1}) \cdot (\boldsymbol{\xi} - \nabla u_h^{k-1}).\end{aligned} \tag{6.5}$$

### 6.1.3 Algebraic solution

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

$$(\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla \psi_e) = (f_h, \psi_e) - R_e^{k,i} \qquad \forall e \in \mathcal{E}_h^{\mathrm{int}}. \tag{6.6}$$

### 6.1.4 Flux reconstruction

Let $K \in \mathcal{T}_h$. We define $\mathbf{f}_h(\mathbf{x})|_K := \frac{f_h|_K}{d}(\mathbf{x} - \mathbf{x}_K)$, with $\mathbf{x}_K$ the barycenter of $K$. For all $e \in \mathcal{E}_K$, let $\mathbf{a}_{K,e}$ be the vertex of $K$ opposite to the face $e$. Let $\mathcal{T}_e$ stand for the patch of elements sharing the face $e$.

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

$$(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})|_K := \big(-\boldsymbol{\Pi}_0 \boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{f}_h\big)|_K - \sum_{e \in \mathcal{E}_K} |\mathcal{T}_e|^{-1} \frac{R_e^{k,i}}{d}(\mathbf{x} - \mathbf{a}_{K,e}). \tag{6.7}$$

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

$$\bar{R}_e^{k,i} := (f_h, \psi_e) - (\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla \psi_e) \qquad \forall e \in \mathcal{E}_h^{\mathrm{int}}, \tag{6.8}$$

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 $K \in \mathcal{T}_h$,*

$$\mathbf{d}_h^{k,i}|_K := \big(-\boldsymbol{\Pi}_0 \boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{f}_h\big)|_K - \sum_{e \in \mathcal{E}_K} |\mathcal{T}_e|^{-1} \frac{\bar{R}_e^{k,i}}{d}(\mathbf{x} - \mathbf{a}_{K,e}). \tag{6.9}$$

**Definition 6.3** (Approximate gradient, data oscillation, quadrature, and algebraic remainder)**.** *Set $\mathbf{g}_h^{k,i} := \nabla u_h^{k,i}$, $f_h := \Pi_0 f$, $\overline{\boldsymbol{\sigma}}_h^{k,i} := \boldsymbol{\Pi}_0 \boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$, and $r_h^{k,i}|_K := \sum_{e \in \mathcal{E}_K} |\mathcal{T}_e|^{-1} R_e^{k,i}$ for all $K \in \mathcal{T}_h$.*

### 6.1.5 Assumptions verification

**Lemma 6.4** (Linearization error convergence)**.** *Assumption 3.5*(iii) *holds.*

*Proof.* The requirement is obvious from Definitions 6.1 and 6.2. □

**Lemma 6.5** (Quasi-equilibration)**.** *Assumption 4.1 holds.*

*Proof.* The proof exploits the link between nonconforming finite elements and mixed finite elements, cf. Marini [29]. For all $K \in \mathcal{T}_h$ and all $e \in \mathcal{E}_K$, we introduce the geometric weight $\omega_{e,K} := |K|/|\mathcal{T}_e|$. Note that $0 < \omega_{e,K} \le 1$ and $\omega_{e,K} = 1$ only on boundary faces. For any interface $e \in \mathcal{E}_h^{\rm int}$ such that $e = \partial K \cap \partial K'$, $K, K' \in \mathcal{T}_h$, observing that $\omega_{e,K} + \omega_{e,K'} = 1$, we define the weighted average of a piecewise polynomial function $\mathbf{v}_h$ at $e$ as $\{\!\{\mathbf{v}_h\}\!\}_\omega := \omega_{e,K'}(\mathbf{v}_h|_K)|_e + \omega_{e,K}(\mathbf{v}_h|_{K'})|_e$. On $e \in \mathcal{E}_h^{\rm ext}$, we set $\{\!\{\mathbf{v}_h\}\!\}_\omega := \mathbf{v}_h|_e$. We first show that, for all $K \in \mathcal{T}_h$ and all $e \in \mathcal{E}_K$,

$$(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})|_K \cdot \mathbf{n}_e = \{\!\{-\boldsymbol{\Pi}_0 \boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{f}_h\}\!\}_\omega \cdot \mathbf{n}_e. \tag{6.10}$$

This is obvious for $e \in \mathcal{E}_h^{\rm ext}$. Let now $e \in \mathcal{E}_h^{\rm int}$. Set $\mathbf{w}_h := -\boldsymbol{\Pi}_0 \boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i}) + \mathbf{f}_h$. It is readily seen that $(\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla \psi_e) = |e| [\![\boldsymbol{\Pi}_0 \boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i})]\!] \cdot \mathbf{n}_e$ and $(f_h, \psi_e) = |e| [\![\mathbf{f}_h]\!] \cdot \mathbf{n}_e$ (recall that $[\![\cdot]\!]$ denotes the jump across $e$ in the direction of $\mathbf{n}_e$). Hence, owing to (6.6), $[\![\mathbf{w}_h]\!] \cdot \mathbf{n}_e = |e|^{-1} R_e^{k,i}$. The result (6.10) then follows from

$$\mathbf{w}_h|_K \cdot \mathbf{n}_e = \{\!\{\mathbf{w}_h\}\!\}_\omega \cdot \mathbf{n}_e + \omega_{e,K} [\![\mathbf{w}_h]\!] \cdot \mathbf{n}_K \tag{6.11}$$

and (6.7). Now, (6.10) shows that $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$ has continuous normal component across interfaces, so that $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i}) \in \mathbf{RTN}_0(\mathcal{T}_h)$. Finally, the property (4.1) follows by taking the divergence of (6.7) and considering the definition of $r_h^{k,i}$. □

**Lemma 6.6** (Local approximation)**.** *Assumption 5.1 holds.*

*Proof.* Let $\mathbf{v}_h := \overline{\boldsymbol{\sigma}}_h^{k,i} + \mathbf{d}_h^{k,i} \in \mathbf{RTN}_0^{-1}(\mathcal{T}_h)$ and use, for all $K \in \mathcal{T}_h$, the estimate $\|\mathbf{v}_h\|_{q,K} \lesssim \{\sum_{e \in \mathcal{E}_K} h_e \|\mathbf{v}_h|_K \cdot \mathbf{n}_e\|_{q,e}^q\}^{\frac{1}{q}}$ shown in [16, Section A.4]. Let $e \in \mathcal{E}_K$. If $e \in \mathcal{E}_h^{\rm ext}$, using $\bar{R}_e^{k,i} := 0$ in (6.9), $|\mathbf{x} - \mathbf{x}_K| \le h_K$, a $q$-robust inverse inequality (see [16, Section A.1 and A.4]), the fact that $f_h$ is constant on $K$, and $\nabla \cdot \overline{\boldsymbol{\sigma}}_h^{k,i} = 0$ yields

$$\begin{aligned} h_e \|\mathbf{v}_h|_K \cdot \mathbf{n}_e\|_{q,e}^q &= h_e \|f_h|_K d^{-1}(\mathbf{x} - \mathbf{x}_K) \cdot \mathbf{n}_e\|_{q,e}^q \le h_K^{1+q} \|f_h|_K\|_{q,e}^q \\ &\lesssim h_K^q \|f_h\|_{q,K}^q = h_K^q \|f_h + \nabla \cdot \overline{\boldsymbol{\sigma}}_h^{k,i}\|_{q,K}^q. \end{aligned}$$

If $e \in \mathcal{E}_h^{\rm int}$, reasoning as in the proof of Lemma 6.5 yields $\mathbf{d}_h^{k,i} \cdot \mathbf{n}_e = \{\!\{-\overline{\boldsymbol{\sigma}}_h^{k,i} + \mathbf{f}_h\}\!\}_\omega \cdot \mathbf{n}_e$ (so that $\mathbf{d}_h^{k,i} \in \mathbf{RTN}_0(\mathcal{T}_h)$). Using this relation, (6.11) to evaluate $\mathbf{v}_h|_K \cdot \mathbf{n}_e$, and the continuity of the normal component of $\mathbf{d}_h^{k,i}$ yields $\mathbf{v}_h|_K \cdot \mathbf{n}_e = \{\!\{\mathbf{f}_h\}\!\}_\omega \cdot \mathbf{n}_e + \omega_{e,K} [\![\overline{\boldsymbol{\sigma}}_h^{k,i}]\!] \cdot \mathbf{n}_K$. We conclude by proceeding as in the first part of the proof. □

**Remark 6.7** (A tighter flux reconstruction using a dual mesh)**.** *A slightly tighter flux reconstruction can be devised using a dual mesh: for all $K \in \mathcal{T}_h$ and all $e \in \mathcal{E}_K$, let $K_e$ be the sub-simplex of $K$ formed by the face $e$ and the barycenter $\mathbf{x}_K$. Let $D_e$ regroup the sub-simplices which share $e$. Then replace in the last terms of (6.7) and (6.9) the vertex $\mathbf{a}_{K,e}$ by the barycenter $\mathbf{x}_K$ and $|\mathcal{T}_e|^{-1}$ by $|D_e|^{-1}$. Then, using local stopping criteria, elementwise efficiency (without neighbors) can be proven on each element of the dual mesh $\mathcal{D}_h = \{D_e\}_{e \in \mathcal{E}_h}$.*

## 6.2 Conforming finite elements

We treat here the discretization of problem (2.4) by conforming finite elements.

### 6.2.1 Discretization

Let $V_h := \mathbb{P}_m(\mathcal{T}_h) \cap V$, $m \ge 1$, be the usual finite element space of continuous, piecewise $m$-th order polynomial functions. The corresponding discretization of problem (2.4) reads: find $u_h \in V_h$ such that

$$(\boldsymbol{\sigma}(u_h, \nabla u_h), \nabla v_h) = (f_h, v_h) \qquad \forall v_h \in V_h. \tag{6.12}$$

Let $\psi_j \in V_h$, $j \in \mathcal{C} := \{1, \dots, \dim(V_h)\}$, denote the basis functions of $V_h$. Employing these functions in (6.12) gives rise to the nonlinear algebraic system (1.1).

### 6.2.2 Linearization

Let $u_h^0 \in V_h$, fixing the initial vector $U^0$ in Algorithm 3.7. The linearization of (6.12), for $k \geq 1$, reads: find $u_h^k \in V_h$ such that

$$(\boldsymbol{\sigma}^{k-1}(u_h^k, \nabla u_h^k), \nabla \psi_j) = (f, \psi_j) \qquad \forall j \in \mathcal{C}, \tag{6.13}$$

which is the functional form of the algebraic system (1.2). Two common linearizations are the fixed point (6.4) and the Newton one (6.5).

### 6.2.3 Algebraic solution

On $i$-th step, $i \geq 0$, of an iterative linear solver for the algebraic system (1.2), we obtain the algebraic residual vector $R^{k,i}$ in (1.3), with components associated with the set $\mathcal{C}$, $R^{k,i} = \{R_j^{k,i}\}_{j\in\mathcal{C}}$. The functional form of (1.3) is: find $u_h^{k,i} \in V_h$ such that

$$(\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla \psi_j) = (f, \psi_j) - R_j^{k,i} \qquad \forall j \in \mathcal{C}. \tag{6.14}$$

### 6.2.4 Flux reconstruction

We construct $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i}) \in \mathbf{RTN}_l(\mathcal{T}_h)$ with $l := m-1$ or $l := m$, using local homogeneous Neumann mixed finite element problems posed on patches around mesh vertices, in an equivalent reformulation of the approach of Braess and Schöberl [4]. Let $\mathcal{V}_h$ denote the set of mesh vertices with subsets $\mathcal{V}_h^{\mathrm{int}}$ for interior vertices and $\mathcal{V}_h^{\mathrm{ext}}$ for boundary ones. Let $\psi_{\mathbf{a}} \in \mathbb{P}_1(\mathcal{T}_h) \cap C^0(\Omega)$ stand for the hat basis function associated with vertex $\mathbf{a} \in \mathcal{V}_h$. To distribute the algebraic residual onto vertices, we set, for all $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$, $R_{\mathbf{a}}^{k,i} := \sum_{j\in\mathcal{C}} \beta_j R_j^{k,i}$, where the coefficients $\beta_j$ are such that $\psi_{\mathbf{a}} = \sum_{j\in\mathcal{C}} \beta_j \psi_j$, while, for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}$, we set $R_{\mathbf{a}}^{k,i} := 0$. Furthermore, for all $\mathbf{a} \in \mathcal{V}_h$, let $\mathcal{T}_{\mathbf{a}}$ be the patch of elements of $\mathcal{T}_h$ that share $\mathbf{a}$, and let $\mathbf{RTN}_l^{\mathrm{N},0}(\mathcal{T}_{\mathbf{a}})$ be the subspace of $\mathbf{RTN}_l(\mathcal{T}_{\mathbf{a}})$ with zero normal flux through $\partial\mathcal{T}_{\mathbf{a}}$ for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$ and through that part of $\partial\mathcal{T}_{\mathbf{a}}$ which lies inside $\Omega$ for $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}$. Let $\mathbb{P}_l^*(\mathcal{T}_{\mathbf{a}})$ be spanned by piecewise $l$-th order polynomials on $\mathcal{T}_{\mathbf{a}}$, with zero mean on $\mathcal{T}_{\mathbf{a}}$ when $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$.

**Definition 6.8** (Construction of $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$)**.** *For all vertices $\mathbf{a} \in \mathcal{V}_h$, define $(\mathbf{d}_{\mathbf{a}}^{k,i} + \mathbf{l}_{\mathbf{a}}^{k,i}) \in \mathbf{RTN}_l^{\mathrm{N},0}(\mathcal{T}_{\mathbf{a}})$ and $q_{\mathbf{a}} \in \mathbb{P}_l^*(\mathcal{T}_{\mathbf{a}})$ by*

$$(\mathbf{d}_{\mathbf{a}}^{k,i} + \mathbf{l}_{\mathbf{a}}^{k,i}, \mathbf{v}_h)_{\mathcal{T}_{\mathbf{a}}} - (q_{\mathbf{a}}, \nabla\cdot\mathbf{v}_h)_{\mathcal{T}_{\mathbf{a}}} = -(\mathbf{I}_l^{\mathrm{RTN}}(\psi_{\mathbf{a}}\boldsymbol{\Pi}_l\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i})), \mathbf{v}_h)_{\mathcal{T}_{\mathbf{a}}}, \tag{6.15a}$$

$$(\nabla\cdot(\mathbf{d}_{\mathbf{a}}^{k,i} + \mathbf{l}_{\mathbf{a}}^{k,i}), \phi_h)_{\mathcal{T}_{\mathbf{a}}} = (f\psi_{\mathbf{a}} - \boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i})\cdot\nabla\psi_{\mathbf{a}}, \phi_h)_{\mathcal{T}_{\mathbf{a}}} - (R_{\mathbf{a}}^{k,i}, \phi_h)_{\mathcal{T}_{\mathbf{a}}}|\mathcal{T}_{\mathbf{a}}|^{-1}, \tag{6.15b}$$

*for all $(\mathbf{v}_h, \phi_h) \in \mathbf{RTN}_l^{\mathrm{N},0}(\mathcal{T}_{\mathbf{a}}) \times \mathbb{P}_l^*(\mathcal{T}_{\mathbf{a}})$. Then, set $\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i} := \sum_{\mathbf{a}\in\mathcal{V}_h}(\mathbf{d}_{\mathbf{a}}^{k,i} + \mathbf{l}_{\mathbf{a}}^{k,i})$.*

In (6.15b), we can take $\phi_h \in \mathbb{P}_l(\mathcal{T}_{\mathbf{a}})$ since multiplying (6.14) by the coefficients $\beta_j$, summing over all $j \in \mathcal{C}$, and using the definition of $R_{\mathbf{a}}^{k,i}$, yields, for all $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$, the Neumann compatibility condition

$$(\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla\psi_{\mathbf{a}})_{\mathcal{T}_{\mathbf{a}}} = (f, \psi_{\mathbf{a}})_{\mathcal{T}_{\mathbf{a}}} - R_{\mathbf{a}}^{k,i}. \tag{6.16}$$

We proceed similarly for $\mathbf{d}_h^{k,i}$. Set $\bar{R}_{\mathbf{a}}^{k,i} := 0$ for any $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}$ and

$$\bar{R}_{\mathbf{a}}^{k,i} := (f, \psi_{\mathbf{a}})_{\mathcal{T}_{\mathbf{a}}} - (\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla\psi_{\mathbf{a}})_{\mathcal{T}_{\mathbf{a}}} \qquad \forall \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}. \tag{6.17}$$

**Definition 6.9** (Construction of $\mathbf{d}_h^{k,i}$)**.** *Define $\mathbf{d}_{\mathbf{a}}^{k,i} \in \mathbf{RTN}_l^{\mathrm{N},0}(\mathcal{T}_{\mathbf{a}})$ and $\bar{q}_{\mathbf{a}} \in \mathbb{P}_l^*(\mathcal{T}_{\mathbf{a}})$ by solving the mixed finite element problems (6.15) with $\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$ in place of $\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i})$ and $\bar{R}_{\mathbf{a}}^{k,i}$ in place of $R_{\mathbf{a}}^{k,i}$. Then, set $\mathbf{d}_h^{k,i} := \sum_{\mathbf{a}\in\mathcal{V}_h} \mathbf{d}_{\mathbf{a}}^{k,i}$.*

**Definition 6.10** (Approximate gradient, data oscillation, quadrature, and algebraic remainder)**.** *Set $\mathbf{g}_h^{k,i} := \nabla u_h^{k,i}$, $f_h := \Pi_l f$, $\overline{\boldsymbol{\sigma}}_h^{k,i} := \boldsymbol{\Pi}_l\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$, and $r_h^{k,i}|_K := \sum_{\mathbf{a}\in\mathcal{V}_K} |\mathcal{T}_{\mathbf{a}}|^{-1} R_{\mathbf{a}}^{k,i}$ for all $K \in \mathcal{T}_h$, where $\mathcal{V}_K$ collects the vertices of $K$.*

### 6.2.5 Assumptions verification

Definitions 6.8 and 6.9 readily imply:

**Lemma 6.11** (Linearization error convergence)**.** *Assumption* 3.5(iii) *holds.*

**Lemma 6.12** (Quasi-equilibration)**.** *Assumption* 4.1 *holds.*

*Proof.* Let $K \in \mathcal{T}_h$ and let $v_h \in \mathbb{P}_l(K)$ (and zero elsewhere) be fixed. For any $\mathbf{a} \in \mathcal{V}_K$, by (6.16), we can take $v_h$ as test function $\phi_h$ in (6.15b). Since $\sum_{\mathbf{a}\in\mathcal{V}_K} \psi_\mathbf{a}|_K = 1$ and $\sum_{\mathbf{a}\in\mathcal{V}_K} \nabla\psi_\mathbf{a}|_K = 0$ ($\psi_\mathbf{a}$ form a partition of unity on $K$), we infer

$$(\nabla\cdot(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i}), v_h)_K = \sum_{\mathbf{a}\in\mathcal{V}_K} (\nabla\cdot(\mathbf{d}_\mathbf{a}^{k,i} + \mathbf{l}_\mathbf{a}^{k,i}), v_h)_K = (f, v_h)_K - \sum_{\mathbf{a}\in\mathcal{V}_K} (R_\mathbf{a}^{k,i}, v_h)_K |\mathcal{T}_\mathbf{a}|^{-1},$$

whence the assertion of the lemma follows from the definition of $r_h^{k,i}$. □

**Lemma 6.13** (Local approximation)**.** *Assumption* 5.1 *holds.*

*Proof.* Let $K \in \mathcal{T}_h$. Since $\mathbf{I}_l^{\mathrm{RTN}}(\overline{\boldsymbol{\sigma}}_h^{k,i}) = \overline{\boldsymbol{\sigma}}_h^{k,i}$, by the partition of unity and linearity of the projection operator $\mathbf{I}_l^{\mathrm{RTN}}$, it follows that $(\mathbf{d}_h^{k,i} + \overline{\boldsymbol{\sigma}}_h^{k,i})|_K = (\mathbf{d}_h^{k,i} + \mathbf{I}_l^{\mathrm{RTN}}(\overline{\boldsymbol{\sigma}}_h^{k,i}))|_K = \sum_{\mathbf{a}\in\mathcal{V}_K}(\mathbf{d}_\mathbf{a}^{k,i} + \mathbf{I}_l^{\mathrm{RTN}}(\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i}))|_K$. We thus only work with $(\mathbf{d}_\mathbf{a}^{k,i} + \mathbf{I}_l^{\mathrm{RTN}}(\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i}))|_K$ for a vertex $\mathbf{a} \in \mathcal{V}_K$, or, more precisely, with $(\mathbf{d}_\mathbf{a}^{k,i} + \mathbf{I}_l^{\mathrm{RTN}}(\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i}))|_{\mathcal{T}_\mathbf{a}}$, in order to prove (5.2). Note that $(\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla\psi_\mathbf{a})_{\mathcal{T}_\mathbf{a}} = (\overline{\boldsymbol{\sigma}}_h^{k,i}, \nabla\psi_\mathbf{a})_{\mathcal{T}_\mathbf{a}}$ and, for all $\phi_h \in \mathbb{P}_l(\mathcal{T}_\mathbf{a})$, $(\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})\cdot\nabla\psi_\mathbf{a}, \phi_h)_{\mathcal{T}_\mathbf{a}} = (\overline{\boldsymbol{\sigma}}_h^{k,i}\cdot\nabla\psi_\mathbf{a}, \phi_h)_{\mathcal{T}_\mathbf{a}}$, so that we can replace $\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$ by $\overline{\boldsymbol{\sigma}}_h^{k,i}$ everywhere in Definition 6.9. We next proceed as in [16, Section A.4], cf. also [22, Proof of Lemmas 7.5 and 7.8]. Firstly, let $M(\mathcal{T}_\mathbf{a})$ denote the postprocessing space of piecewise (discontinuous) polynomials $m_h$ on $\mathcal{T}_\mathbf{a}$ such that

$$\langle [\![m_h]\!], v_h\rangle_e = 0 \qquad \forall e \in \mathcal{E}_\mathbf{a},\ \forall v_h \in \mathbb{P}_l(e), \tag{6.18}$$

where $\mathcal{E}_\mathbf{a}$ collects the faces to which $\mathbf{a}$ belongs. Moreover, the functions $m_h$ in $M(\mathcal{T}_\mathbf{a})$ satisfy $(m_h, 1)_{\mathcal{T}_\mathbf{a}} = 0$ for interior vertices $\mathbf{a}$. [40, Lemma 5.4] and [16, Section A.4] yield

$$\|\mathbf{d}_\mathbf{a}^{k,i} + \mathbf{I}_l^{\mathrm{RTN}}(\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i})\|_{q,\mathcal{T}_\mathbf{a}} \lesssim \sup_{m_h\in M(\mathcal{T}_\mathbf{a}),\, \|\nabla m_h\|_{p,\mathcal{T}_\mathbf{a}}=1} (\mathbf{d}_\mathbf{a}^{k,i} + \mathbf{I}_l^{\mathrm{RTN}}(\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i}), \nabla m_h)_{\mathcal{T}_\mathbf{a}}.$$

Let $m_h \in M(\mathcal{T}_\mathbf{a})$ with $\|\nabla m_h\|_{p,\mathcal{T}_\mathbf{a}} = 1$ be fixed and consider the right-hand side of the above inequality. The Green theorem, the fact that $\mathbf{d}_\mathbf{a}^{k,i} + \mathbf{I}_l^{\mathrm{RTN}}(\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i})$ has zero normal flux through (a part of) $\partial\mathcal{T}_\mathbf{a}$ together with (6.18) on $\partial\mathcal{T}_\mathbf{a} \cap \partial\Omega$ when $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}$, the fact that $\mathbf{d}_\mathbf{a}^{k,i} \in \mathbf{RTN}_l^{\mathrm{N},0}(\mathcal{T}_\mathbf{a})$, (6.18), and the properties (6.1) of $\mathbf{I}_l^{\mathrm{RTN}}$ yield

$$\begin{aligned}
&-\sum_{K'\in\mathcal{T}_\mathbf{a}} (\nabla\cdot(\mathbf{d}_\mathbf{a}^{k,i} + \mathbf{I}_l^{\mathrm{RTN}}(\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i})), m_h)_{K'} + \sum_{e\in\mathcal{E}_h^{\mathrm{int}},\, e\cap\mathbf{a}\neq\emptyset} \langle [\![\mathbf{I}_l^{\mathrm{RTN}}(\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i})\cdot\mathbf{n}_e]\!], m_h\rangle_e \\
&= -\sum_{K'\in\mathcal{T}_\mathbf{a}} (\nabla\cdot(\mathbf{d}_\mathbf{a}^{k,i} + \psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i}), \Pi_l(m_h))_{K'} + \sum_{e\in\mathcal{E}_h^{\mathrm{int}},\, e\cap\mathbf{a}\neq\emptyset} \langle [\![\psi_\mathbf{a}\overline{\boldsymbol{\sigma}}_h^{k,i}\cdot\mathbf{n}_e]\!], \Pi_l(m_h)\rangle_e
\end{aligned}$$

that we denote as $I + II$. Employing the second lines of the problems of Definition 6.9 (recall that we can take $\phi_h \in \mathbb{P}_l(\mathcal{T}_\mathbf{a})$), the first term $I$ above can be developed as

$$\begin{aligned}
&-\sum_{K'\in\mathcal{T}_\mathbf{a}} (\psi_\mathbf{a}(\nabla\cdot\overline{\boldsymbol{\sigma}}_h^{k,i} + f) - \bar{R}_\mathbf{a}^{k,i}|\mathcal{T}_\mathbf{a}|^{-1}, \Pi_l(m_h))_{K'} \\
&\le \left\{\sum_{K'\in\mathcal{T}_\mathbf{a}} h_{K'}^{-p}\|m_h\|_{p,K'}^p\right\}^{\frac{1}{p}} \left\{\sum_{K'\in\mathcal{T}_\mathbf{a}} h_{K'}^q(\|f + \nabla\cdot\overline{\boldsymbol{\sigma}}_h^{k,i}\|_{q,K'} + \|\bar{R}_\mathbf{a}^{k,i}|\mathcal{T}_\mathbf{a}|^{-1}\|_{q,K'})^q\right\}^{\frac{1}{q}} \\
&\lesssim h_{\mathcal{T}_\mathbf{a}}^{-1}\|m_h\|_{p,\mathcal{T}_\mathbf{a}} \left(\left\{\sum_{K'\in\mathcal{T}_\mathbf{a}} h_{K'}^q\|f + \nabla\cdot\overline{\boldsymbol{\sigma}}_h^{k,i}\|_{q,K'}^q\right\}^{\frac{1}{q}} + |\bar{R}_\mathbf{a}^{k,i}||\mathcal{T}_\mathbf{a}|^{-1+\frac{1}{q}} h_{\mathcal{T}_\mathbf{a}}\right),
\end{aligned}$$

where we have also used the Hölder inequality, the stability of the $\Pi_l$-projection, and the fact that $\|\psi_{\mathbf{a}}\|_{\infty,\mathcal{T}_{\mathbf{a}}} = 1$. Finally, for any interior vertex $\mathbf{a}$, we get from (6.17), the Green theorem, the Hölder inequality, and the $p$-robust inverse inequality $\|\psi_{\mathbf{a}}\|_{p,e} \lesssim h_e^{-\frac{1}{p}} \|\psi_{\mathbf{a}}\|_{p,K'}$, $e \in \mathcal{E}_{K'}$, see [16, Section A.4], that the term $\bar{R}_{\mathbf{a}}^{k,i}$ can be developed as

$$\begin{aligned}
&\sum_{K'\in\mathcal{T}_{\mathbf{a}}} (f + \nabla\cdot\overline{\boldsymbol{\sigma}}_h^{k,i}, \psi_{\mathbf{a}})_{K'} - \sum_{e\in\mathcal{E}_h^{\mathrm{int}},\, e\cap\mathbf{a}\neq\emptyset} \langle [\![\overline{\boldsymbol{\sigma}}_h^{k,i}\cdot\mathbf{n}_e]\!], \psi_{\mathbf{a}}\rangle_e \\
&\lesssim \left( \left\{ \sum_{K'\in\mathcal{T}_{\mathbf{a}}} h_{K'}^q \|f + \nabla\cdot\overline{\boldsymbol{\sigma}}_h^{k,i}\|_{q,K'}^q \right\}^{\frac{1}{q}} + \left\{ \sum_{e\in\mathcal{E}_h^{\mathrm{int}},\, e\cap\mathbf{a}\neq\emptyset} h_e \|[\![\overline{\boldsymbol{\sigma}}_h^{k,i}\cdot\mathbf{n}_e]\!]\|_{q,e}^q \right\}^{\frac{1}{q}} \right) h_{\mathcal{T}_{\mathbf{a}}}^{-1} |\mathcal{T}_{\mathbf{a}}|^{\frac{1}{p}}.
\end{aligned}$$

Using the $p$-robust discrete Poincaré/Friedrichs inequality $\|m_h\|_{p,\mathcal{T}_{\mathbf{a}}} \lesssim h_{\mathcal{T}_{\mathbf{a}}} \|\nabla m_h\|_{p,\mathcal{T}_{\mathbf{a}}}$ from [16, Section A.4] and the triangle inequality for separating the data oscillation terms $\eta_{\mathrm{osc},K}^{k,i}$, we conclude that $I \le \eta_{\sharp,K}^{k,i} + \eta_{\mathrm{osc},\mathfrak{T}_K}^{k,i}$. Proceeding similarly for the jump term $II$ (with the above treatment of $\psi_{\mathbf{a}}$ and $\Pi_l$) yields the desired result. □

## 6.3 Interior penalty discontinuous Galerkin (IPDG) for quasi-linear diffusion

We treat here the IPDG method applied in the quasi-linear setting (2.2).

### 6.3.1 Discretization

Let $V_h := \mathbb{P}_m(\mathcal{T}_h)$, $m \ge 1$. The IPDG discretization of problem (2.4) in the case (2.2) reads: find $u_h \in V_h$ such that, for all $v_h \in V_h$,

$$\begin{aligned}
&(\boldsymbol{\sigma}(u_h, \nabla u_h), \nabla v_h) - \sum_{e\in\mathcal{E}_h} \big\{ \langle \{\!\{\boldsymbol{\sigma}(u_h, \nabla u_h)\}\!\}\cdot\mathbf{n}_e, [\![v_h]\!]\rangle_e \\
&+ \theta \langle \{\!\{\underline{\mathbf{A}}(u_h)\nabla v_h\}\!\}\cdot\mathbf{n}_e, [\![u_h]\!]\rangle_e \big\} + \sum_{e\in\mathcal{E}_h} \langle \bar{\alpha}_e h_e^{-1} [\![u_h]\!], [\![v_h]\!]\rangle_e = (f, v_h),
\end{aligned} \tag{6.19}$$

with $\theta \in \{-1, 0, 1\}$ and $\bar{\alpha}_e := \|\underline{\mathbf{A}}\|_{L^\infty(\mathbb{R})}\chi_e$ where $\chi_e$ is a large enough positive parameter. The average operator $\{\!\{\cdot\}\!\}$ yields the mean value of the traces from adjacent mesh elements on interfaces and the actual trace on boundary faces. Testing (6.19) against the basis functions in $V_h$ gives rise to the nonlinear algebraic system (1.1); these basis functions are denoted $\psi_{K,j}$, for all $K \in \mathcal{T}_h$ and all $j \in \mathcal{C}_K := \{1, \dots, \dim(\mathbb{P}_m(K))\}$.

### 6.3.2 Linearization

Let $u_h^0 \in V_h$, fixing $U^0$ in Algorithm 3.7. The linearization of (6.19), for $k \ge 1$, is: find $u_h^k \in V_h$ such that, for all $K \in \mathcal{T}_h$ and all $j \in \mathcal{C}_K$,

$$\begin{aligned}
&(\boldsymbol{\sigma}^{k-1}(u_h^k, \nabla u_h^k), \nabla\psi_{K,j}) - \sum_{e\in\mathcal{E}_h} \big\{ \langle \{\!\{\boldsymbol{\sigma}^{k-1}(u_h^k, \nabla u_h^k)\}\!\}\cdot\mathbf{n}_e, [\![\psi_{K,j}]\!]\rangle_e \\
&+ \theta \langle \{\!\{\underline{\mathbf{A}}^{k-1}(u_h^k)\nabla\psi_{K,j}\}\!\}\cdot\mathbf{n}_e, [\![u_h^k]\!]\rangle_e \big\} + \sum_{e\in\mathcal{E}_h} \langle \bar{\alpha}_e h_e^{-1} [\![u_h^k]\!], [\![\psi_{K,j}]\!]\rangle_e = (f, \psi_{K,j}),
\end{aligned} \tag{6.20}$$

which is the functional form of (1.2). The fixed point linearization corresponds to $\boldsymbol{\sigma}^{k-1}(v, \boldsymbol{\xi}) := \underline{\mathbf{A}}(u_h^{k-1})\boldsymbol{\xi}$ and $\underline{\mathbf{A}}^{k-1}(v) := \underline{\mathbf{A}}(u_h^{k-1})$, and the Newton linearization to

$$\boldsymbol{\sigma}^{k-1}(v, \boldsymbol{\xi}) := \underline{\mathbf{A}}(u_h^{k-1})\boldsymbol{\xi} + (v - u_h^{k-1})\partial_v\underline{\mathbf{A}}(u_h^{k-1})\nabla u_h^{k-1}, \tag{6.21a}$$

$$\underline{\mathbf{A}}^{k-1}(v) := \underline{\mathbf{A}}(u_h^{k-1}) + \partial_v\underline{\mathbf{A}}(u_h^{k-1})(v - u_h^{k-1}). \tag{6.21b}$$

### 6.3.3 Algebraic solution

On $i$-th step, $i \geq 0$, of an iterative linear solver applied to (1.2), we obtain (1.3) with algebraic residual vector $R^{k,i} = \{R^{k,i}_{K,j}\}_{K\in\mathcal{T}_h, j\in\mathcal{C}_K}$. The functional form of (1.3) is: find $u^{k,i}_h \in V_h$ such that, for all $K \in \mathcal{T}_h$ and all $j \in \mathcal{C}_K$,

$$(\boldsymbol{\sigma}^{k-1}(u^{k,i}_h, \nabla u^{k,i}_h), \nabla\psi_{K,j}) - \sum_{e\in\mathcal{E}_h} \big\{\langle \{\!\!\{\boldsymbol{\sigma}^{k-1}(u^{k,i}_h, \nabla u^{k,i}_h)\}\!\!\}\cdot\mathbf{n}_e, [\![\psi_{K,j}]\!]\rangle_e$$
$$+\theta\langle \{\!\!\{\underline{\mathbf{A}}^{k-1}(u^{k,i}_h)\nabla\psi_{K,j}\}\!\!\}\cdot\mathbf{n}_e, [\![u^{k,i}_h]\!]\rangle_e\big\} + \sum_{e\in\mathcal{E}_h}\langle \bar\alpha_e h_e^{-1}[\![u^{k,i}_h]\!], [\![\psi_{K,j}]\!]\rangle_e = (f,\psi_{K,j}) - R^{k,i}_{K,j}.$$

### 6.3.4 Flux reconstruction

We construct $\mathbf{d}^{k,i}_h$ and $\mathbf{l}^{k,i}_h$ in the space $\mathbf{RTN}_l(\mathcal{T}_h)$ with $l := m-1$ or $l := m$, following Kim [26] and [17]. For all $e \in \mathcal{E}_h$, we set $w_e := \frac{1}{2}$ if $e \in \mathcal{E}_h^{\rm int}$ and $w_e := 1$ if $e \in \mathcal{E}_h^{\rm ext}$.

**Definition 6.14** (Construction of $(\mathbf{d}^{k,i}_h + \mathbf{l}^{k,i}_h)$)**.** *The function $(\mathbf{d}^{k,i}_h + \mathbf{l}^{k,i}_h)$ is defined in $\mathbf{RTN}_l(\mathcal{T}_h)$ such that, for all $K \in \mathcal{T}_h$ and all $e \in \mathcal{E}_K$,*

$$\langle(\mathbf{d}^{k,i}_h + \mathbf{l}^{k,i}_h)\cdot\mathbf{n}_e, q_h\rangle_e := \langle -\{\!\!\{\boldsymbol{\sigma}^{k-1}(u^{k,i}_h, \nabla u^{k,i}_h)\}\!\!\}\cdot\mathbf{n}_e + \bar\alpha_e h_e^{-1}[\![u^{k,i}_h]\!], q_h\rangle_e,$$
$$(\mathbf{d}^{k,i}_h + \mathbf{l}^{k,i}_h, \mathbf{r}_h)_K := -(\boldsymbol{\sigma}^{k-1}(u^{k,i}_h, \nabla u^{k,i}_h), \mathbf{r}_h)_K + \theta\sum_{e\in\mathcal{E}_K} w_e\langle \underline{\mathbf{A}}^{k-1}(u^{k,i}_h)\mathbf{r}_h\cdot\mathbf{n}_e, [\![u^{k,i}_h]\!]\rangle_e,$$

*for all $q_h \in \mathbb{P}_l(e)$ and all $\mathbf{r}_h \in [\mathbb{P}_{l-1}(K)]^d$.*

**Definition 6.15** (Construction of $\mathbf{d}^{k,i}_h$)**.** *The function $\mathbf{d}^{k,i}_h$ is in $\mathbf{RTN}_l(\mathcal{T}_h)$ and is defined using the prescription of Definition 6.14 with $\boldsymbol{\sigma}(u^{k,i}_h, \mathbf{g}^{k,i}_h)$ in place of $\boldsymbol{\sigma}^{k-1}(u^{k,i}_h, \nabla u^{k,i}_h)$ and $\underline{\mathbf{A}}(u^{k,i}_h)$ in place of $\underline{\mathbf{A}}^{k-1}(u^{k,i}_h)$.*

**Definition 6.16** (Approximate gradient, data oscillation, quadrature, and algebraic remainder)**.** *Set $\mathbf{g}^{k,i}_h := \nabla u^{k,i}_h$, $f_h := \Pi_l f$, $\overline{\boldsymbol{\sigma}}^{k,i}_h := \mathbf{I}^{\rm RTN}_l(\boldsymbol{\sigma}(u^{k,i}_h, \mathbf{g}^{k,i}_h))$, and $r^{k,i}_h \in \mathbb{P}_m(\mathcal{T}_h)$ with $(r^{k,i}_h, \psi_{K,j})_K = R^{k,i}_{K,j}$ for all $K \in \mathcal{T}_h$ and all $j \in \mathcal{C}_K$.*

### 6.3.5 Assumptions verification

As above, Definitions 6.14 and 6.15 yield:

**Lemma 6.17** (Linearization error convergence)**.** *Assumption 3.5*(iii) *holds.*

**Lemma 6.18** (Quasi-equilibration)**.** *Assumption 4.1 holds.*

*Proof.* Direct verification by proceeding as in [17, 26], see also [14, Section 5.5]. □

**Lemma 6.19** (Local approximation)**.** *Assumption 5.1 holds using weights $\alpha_e := \bar\alpha_e^2$ and exponent $s := p$ in the nonconformity estimator.*

*Proof.* We observe that, for all $K \in \mathcal{T}_h$ and all $e \in \mathcal{E}_K$, there holds

$$\langle(\mathbf{d}^{k,i}_h + \overline{\boldsymbol{\sigma}}^{k,i}_h)\cdot\mathbf{n}_e, q_h\rangle_e = (1-w_e)\langle[\![\overline{\boldsymbol{\sigma}}^{k,i}_h]\!]\cdot\mathbf{n}_e + \bar\alpha_e h_e^{-1}[\![u^{k,i}_h]\!], q_h\rangle_e, \tag{6.22a}$$
$$(\mathbf{d}^{k,i}_h + \overline{\boldsymbol{\sigma}}^{k,i}_h, \mathbf{r}_h)_K = \theta\sum_{e\in\mathcal{E}_K} w_e\langle\underline{\mathbf{A}}(u^{k,i}_h)\mathbf{r}_h\cdot\mathbf{n}_e, [\![u^{k,i}_h]\!]\rangle_e, \tag{6.22b}$$

for all $q_h \in \mathbb{P}_l(e)$ and all $\mathbf{r}_h \in [\mathbb{P}_{l-1}(K)]^d$. The assertion then follows from standard approximation properties in Raviart–Thomas–Nédélec spaces, see, e.g., [14, Section 5.5]. □

## 6.4 Discontinuous Galerkin with gradient reconstruction

We treat here the discretization of the full problem (2.4) by the discontinuous Galerkin method with a discrete gradient suitable especially for the Leray–Lions setting (2.3).

### 6.4.1 Discretization

Let $l \geq 0$ be an integer. For all $e \in \mathcal{E}_h$, we define the map $\ell_e : L^1(e) \to [\mathbb{P}_l(\mathcal{T}_h)]^d$ such that, for all $\phi \in L^1(e)$, $\ell_e(\phi)$ is the unique function in $[\mathbb{P}_l(\mathcal{T}_h)]^d$ such that, for all $\mathbf{v}_h \in [\mathbb{P}_l(\mathcal{T}_h)]^d$, $(\ell_e(\phi), \mathbf{v}_h) = \langle \{\!\{\mathbf{v}_h\}\!\}\cdot\mathbf{n}_e, \phi\rangle_e$. The vector-valued, piecewise polynomial function $\ell_e(\phi)$ is supported in $\mathcal{T}_e$ (the patch of elements sharing the face $e$) and is colinear to $\mathbf{n}_e$. Then, for a function $v \in V(\mathcal{T}_h)$, we define its discrete gradient $\nabla_h v \in [L^p(\Omega)]^d$ (see [14, Section 4.2] and references therein) as

$$\nabla_h v := \nabla v - \mathtt{L}_h([\![v]\!]), \qquad \mathtt{L}_h([\![v]\!]) := \sum_{e\in\mathcal{E}_h} \ell_e([\![v]\!]). \tag{6.23}$$

Observe that $\mathtt{L}_h([\![v]\!])$ is a (piecewise polynomial) correction to the broken gradient $\nabla v$ based on the jump liftings. The discrete gradient is an important tool in the design of discontinuous Galerkin methods for nonlinear problems, see Buffa and Ortner [6] and [7] for the $p$-Laplacian and [13] for the incompressible Navier–Stokes equations.

Let $V_h := \mathbb{P}_m(\mathcal{T}_h)$, $m \geq 1$. We consider here the following gradient reconstruction discontinuous Galerkin method: find $u_h \in V_h$ such that

$$(\boldsymbol{\sigma}(u_h, \nabla_h u_h), \nabla_h v_h) + \sum_{e\in\mathcal{E}_h} \langle s_e([\![u_h]\!]), [\![v_h]\!]\rangle_e = (f, v_h) \qquad \forall v_h \in V_h, \tag{6.24}$$

with the stabilization operator $s_e : L^p(e) \to L^q(e)$ for all $e \in \mathcal{E}_h$ such that, for all $v \in L^p(e)$, $s_e(v) = \bar\alpha_e h_e^{1-p}|v|^{p-2}v$ with a positive parameter $\bar\alpha_e$. Testing (6.24) against the basis functions in $V_h$ gives rise to the nonlinear algebraic system (1.1).

**Remark 6.20** (Stencil reduction and link with IPDG)**.** *The discretization stencil resulting from (6.24) includes neighbors and neighbors of neighbors in the sense of faces. This stencil can be reduced by adding to the left-hand side of (6.24) the form $-(\boldsymbol{\sigma}(u_h, \nabla_h u_h) - \boldsymbol{\sigma}(u_h, \nabla u_h), \nabla_h v_h - \nabla v_h)$. For quasi-linear diffusion and strong enough penalty, this leads to an IPDG formulation of type (6.19) (with $\theta = 1$).*

### 6.4.2 Linearization

Let $u_h^0 \in V_h$, fixing the initial vector $U^0$ in Algorithm 3.7. The linearization of (6.24), for $k \geq 1$, reads: find $u_h^k \in V_h$ such that, for all $K \in \mathcal{T}_h$ and all $j \in \mathcal{C}_K := \{1, \dots, \dim(\mathbb{P}_m(K))\}$,

$$(\boldsymbol{\sigma}^{k-1}(u_h^k, \nabla_h u_h^k), \nabla_h \psi_{K,j}) + \sum_{e\in\mathcal{E}_h} \langle s_e^{k-1}([\![u_h^k]\!]), [\![\psi_{K,j}]\!]\rangle_e = (f, \psi_{K,j}), \tag{6.25}$$

which is the functional form of (1.2). In the fixed-point linearization, $\boldsymbol{\sigma}^{k-1}(v, \boldsymbol{\xi})$ is defined by (6.4) with $\nabla_h u_h^{k-1}$ in place of $\nabla u_h^{k-1}$, while $s_e^{k-1}(v) := \bar\alpha_e h_e^{1-p}|[\![u_h^{k-1}]\!]|^{p-2}v$. In the Newton linearization, $\boldsymbol{\sigma}^{k-1}(v, \boldsymbol{\xi})$ is defined by (6.5) with $\nabla_h u_h^{k-1}$ in place of $\nabla u_h^{k-1}$, while $s_e^{k-1}(v) := \bar\alpha_e h_e^{1-p}|[\![u_h^{k-1}]\!]|^{p-2}$ $((p-1)v - (p-2)[\![u_h^{k-1}]\!])$.

### 6.4.3 Algebraic solution

On $i$-th step, $i \geq 0$, of a linear solver for (1.2), we obtain the system (1.3) with $R^{k,i} = \{R_{K,j}^{k,i}\}_{K\in\mathcal{T}_h, j\in\mathcal{C}_K}$. The functional form of (1.3) is: find $u_h^{k,i} \in V_h$ such that, for all $K \in \mathcal{T}_h$ and all $j \in \mathcal{C}_K$,

$$(\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla_h u_h^{k,i}), \nabla_h \psi_{K,j}) + \sum_{e\in\mathcal{E}_h} \langle s_e^{k-1}([\![u_h^{k,i}]\!]), [\![\psi_{K,j}]\!]\rangle_e = (f, \psi_{K,j}) - R_{K,j}^{k,i}. \tag{6.26}$$

### 6.4.4 Flux reconstruction

We proceed as in Section 6.2.4 hinging on the hat basis functions $\psi_{\mathbf{a}} \in \mathbb{P}_1(\mathcal{T}_h) \cap C^0(\Omega)$. This in particular allows us to eliminate the nonlinear jump terms in the local flux expressions, compare with (6.22). Since $m \geq 1$, there holds $\psi_{\mathbf{a}} \in V_h$, so that there are coefficients $\beta_{K,j}$ such that $\psi_{\mathbf{a}} = \sum_{K\in\mathcal{T}_{\mathbf{a}}} \sum_{j\in\mathcal{C}_K} \beta_{K,j}\psi_{K,j}$. We then distribute the components of $R^{k,i}$ onto vertices by setting $R_{\mathbf{a}}^{k,i} := \sum_{K\in\mathcal{T}_{\mathbf{a}}} \sum_{j\in\mathcal{C}_K} \beta_{K,j} R_{K,j}^{k,i}$ for all $\mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}$, and $R_{\mathbf{a}}^{k,i} := 0$ for all $\mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}$.

We construct $\mathbf{d}_h^{k,i}$ and $\mathbf{l}_h^{k,i}$ in the space $\mathbf{RTN}_l(\mathcal{T}_h)$ with $l := m-1$ or $l := m$. We use the notation from Section 6.2.4.

**Definition 6.21** (Construction of $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$)**.** *We define* $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i}) \in \mathbf{RTN}_l(\mathcal{T}_h)$ *using Definition 6.8 with* $\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla_h u_h^{k,i})$ *in place of* $\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla u_h^{k,i})$.

In the local mixed problems considered in Definition 6.8, we can take $\phi_h \in \mathbb{P}_l(\mathcal{T}_\mathbf{a})$ since multiplying (6.26) by the coefficients $\beta_{K,j}$, summing over all $K \in \mathcal{T}_\mathbf{a}$ and all $j \in \mathcal{C}_K$, using the definition of $R_\mathbf{a}^{k,i}$, and the fact that $[\![\psi_\mathbf{a}]\!] = 0$, yields, for all $\mathbf{a} \in \mathcal{V}_h^{\rm int}$, the Neumann compatibility condition $(\boldsymbol{\sigma}^{k-1}(u_h^{k,i}, \nabla_h u_h^{k,i}), \nabla\psi_\mathbf{a})_{\mathcal{T}_\mathbf{a}} = (f, \psi_\mathbf{a})_{\mathcal{T}_\mathbf{a}} - R_\mathbf{a}^{k,i}$. We proceed similarly for the construction of $\mathbf{d}_h^{k,i}$, setting, for all $\mathbf{a} \in \mathcal{V}_h^{\rm int}$, $\bar{R}_\mathbf{a}^{k,i} := (f, \psi_\mathbf{a})_{\mathcal{T}_\mathbf{a}} - (\boldsymbol{\sigma}(u_h^{k,i}, \nabla_h u_h^{k,i}), \nabla\psi_\mathbf{a})_{\mathcal{T}_\mathbf{a}}$ and, for all $\mathbf{a} \in \mathcal{V}_h^{\rm ext}$, $\bar{R}_\mathbf{a}^{k,i} := 0$. This yields:

**Definition 6.22** (Construction of $\mathbf{d}_h^{k,i}$)**.** *We define* $\mathbf{d}_h^{k,i} \in \mathbf{RTN}_l(\mathcal{T}_h)$ *using Definition 6.9 with* $\boldsymbol{\sigma}(u_h^{k,i}, \nabla_h u_h^{k,i})$ *in place of* $\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$.

**Definition 6.23** (Approximate gradient, data oscillation, quadrature, and algebraic remainder)**.** *Set* $\mathbf{g}_h^{k,i} := \nabla_h u_h^{k,i}$, $f_h := \Pi_l f$, $\overline{\boldsymbol{\sigma}}_h^{k,i} := \mathbf{\Pi}_l(\boldsymbol{\sigma}(u_h^{k,i}, \nabla_h u_h^{k,i}))$, *and* $r_h^{k,i}|_K := \sum_{\mathbf{a}\in\mathcal{V}_K} |\mathcal{T}_\mathbf{a}|^{-1} R_\mathbf{a}^{k,i}$ *for all* $K \in \mathcal{T}_h$.

### 6.4.5 Assumptions verification

The results of Section 6.2.5 apply here identically:

**Lemma 6.24** (Linearization error convergence)**.** *Assumption 3.5*(iii) *holds.*

**Lemma 6.25** (Quasi-equilibration)**.** *Assumption 4.1 holds.*

**Lemma 6.26** (Local approximation)**.** *Assumption 5.1 holds.*

## 6.5 Cell-centered finite volumes and lowest-order mixed finite elements

We apply here cell-centered finite volumes and closely related lowest-order mixed finite elements to the discretization of (2.4).

### 6.5.1 Discretization

Let $V_h := \mathbb{P}_0(\mathcal{T}_h)$. Fix an element $K \in \mathcal{T}_h$ and a face $e \in \mathcal{E}_K$. We denote $\sigma_{K,e} : V_h \to \mathbb{R}$ the finite volume flux function, which maps a piecewise constant function $\bar{v}_h \in V_h$ to the normal flux through $e$, $\sigma_{K,e}(\bar{v}_h)$. We do not need the specific form of the flux functions $\sigma_{K,e}$, except that conservativity be satisfied in the form $\sigma_{K,e}(\bar{v}_h) = -\sigma_{K',e}(\bar{v}_h)$ for any function $\bar{v}_h \in V_h$ and any interface $e \in \mathcal{E}_h^{\rm int}$ such that $e = \partial K \cap \partial K'$. A general cell-centered finite volume method for the problem (2.4), cf. Eymard *et al.* [19], reads: find $\bar{u}_h \in V_h$ such that

$$\sum_{e\in\mathcal{E}_K} \sigma_{K,e}(\bar{u}_h) = (f, 1)_K \qquad \forall K \in \mathcal{T}_h. \tag{6.27}$$

This gives rise to the nonlinear algebraic system (1.1).

### 6.5.2 Linearization

Let $\bar{u}_h^0 \in V_h$, fixing the initial vector $U^0$ in Algorithm 3.7. The linearization of (6.27), for $k \geq 1$, reads: find $\bar{u}_h^k \in V_h$ such that

$$\sum_{e\in\mathcal{E}_K} \sigma_{K,e}^{k-1}(\bar{u}_h^k) = (f, 1)_K \qquad \forall K \in \mathcal{T}_h, \tag{6.28}$$

which is the functional form of the algebraic system (1.2). Here, $\sigma_{K,e}^{k-1} : V_h \to \mathbb{R}$ is the finite volume flux function on the $k$-th linearization step. We again suppose conservativity, i.e., $\sigma_{K,e}^{k-1}(\bar{v}_h) = -\sigma_{K',e}^{k-1}(\bar{v}_h)$ for

any $\bar v_h \in V_h$ and $e = \partial K \cap \partial K' \in \mathcal{E}_h^{\rm int}$. It is not possible to specify the fixed point linearization directly from (6.27), as it depends on the actual form of $\sigma_{K,e}$. For the Newton linearization, $\sigma_{K,e}^{k-1}$ is such that

$$\sigma_{K,e}^{k-1}(\bar v_h) := \sigma_{K,e}(\bar u_h^{k-1}) + \sum_{K'\in\mathcal{T}_h} \frac{\partial \sigma_{K,e}}{\partial \bar u_h|_{K'}}(\bar u_h^{k-1})(\bar v_h|_{K'} - \bar u_h^{k-1}|_{K'}). \tag{6.29}$$

As an example, we detail the linearized flux function $\sigma_{K,e}^{k-1}$ for a two-point finite volume scheme. Let $d = 2$ and assume that $\mathcal{T}_h$ is strictly Delaunay, so that the circumcircle of each triangle does not contain any other triangle vertex, and each circumcenter of a boundary triangle is inside $\Omega$. Consider the quasi-linear diffusion setting (2.2) with a scalar-valued function $a(\mathbf{x}, v)$ (in place of the tensor-valued function $\underline{\mathbf{A}}(\mathbf{x}, v)$). Let $\mathbf{x}_K^\circ$ stand for the circumcenter of the triangle $K \in \mathcal{T}_h$ and $\mathbf{x}_e$ for the center of the edge $e \in \mathcal{E}_h^{\rm ext}$. We use the shorthand notation $a_K(\cdot)$ in place of $a(\mathbf{x}_K^\circ, \cdot)$ and $\bar v_K$ in place of $\bar v_h|_K$ for any function $\bar v_h \in V_h$. Then, a two-point finite volume scheme for the quasi-linear diffusion problem takes the form (6.27) with

$$\sigma_{K,e}(\bar u_h) := \frac{k_e}{2}\big\{a_K(\bar u_K) + a_{K'}(\bar u_{K'})\big\}(\bar u_K - \bar u_{K'}) \quad \forall e = \partial K \cap \partial K' \in \mathcal{E}_h^{\rm int}, \tag{6.30a}$$

$$\sigma_{K,e}(\bar u_h) := k_e a_K(\bar u_K)\bar u_K \qquad \forall e = \partial K \cap \partial\Omega \in \mathcal{E}_h^{\rm ext}, \tag{6.30b}$$

where $k_e := \frac{|e|}{|\mathbf{x}_K^\circ - \mathbf{x}_{K'}^\circ|}$ in (6.30a) and $k_e := \frac{|e|}{|\mathbf{x}_K^\circ - \mathbf{x}_e|}$ in (6.30b). The Newton linearization leads to, for all $K \in \mathcal{T}_h$ and all $e = \partial K \cap \partial K' \in \mathcal{E}_h^{\rm int}$,

$$\begin{aligned}\sigma_{K,e}^{k-1}(\bar v_h) := {}& \frac{k_e}{2}\big\{a_K(\bar u_K^{k-1}) + a_{K'}(\bar u_{K'}^{k-1})\big\}(\bar v_K - \bar v_{K'}) \\ &+ \frac{k_e}{2}\big\{a_K'(\bar u_K^{k-1})(\bar v_K - \bar u_K^{k-1}) + a_{K'}'(\bar u_{K'}^{k-1})(\bar v_{K'} - \bar u_{K'}^{k-1})\big\}(\bar u_K^{k-1} - \bar u_{K'}^{k-1}),\end{aligned} \tag{6.31}$$

and, for all $e = \partial K \cap \partial\Omega \in \mathcal{E}_h^{\rm ext}$,

$$\sigma_{K,e}^{k-1}(\bar v_h) := k_e a_K(\bar u_K^{k-1})\bar v_K + k_e a_K'(\bar u_K^{k-1})(\bar v_K - \bar u_K^{k-1})\bar u_K^{k-1}. \tag{6.32}$$

The fixed point linearization is derived from (6.31)–(6.32) by omitting the terms with the derivative of $a$.

### 6.5.3 Algebraic solution

On $i$-th step, $i \ge 0$, of an iterative linear solver for the algebraic system (1.2), we obtain the algebraic residual vector $R^{k,i}$ in (1.3) with $R^{k,i} = \{R_K^{k,i}\}_{K\in\mathcal{T}_h}$. The functional form of (1.3) is: find $\bar u_h^{k,i} \in V_h$ such that

$$\sum_{e\in\mathcal{E}_K} \sigma_{K,e}^{k-1}(\bar u_h^{k,i}) = (f,1)_K - R_K^{k,i} \qquad \forall K \in \mathcal{T}_h. \tag{6.33}$$

### 6.5.4 Flux reconstruction

We follow Eymard *et al.* [20] to define:

**Definition 6.27** (Construction of $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$)**.** *The function $(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})$ is defined in $\mathbf{RTN}_0(\mathcal{T}_h)$ such that, for all $K \in \mathcal{T}_h$ and all $e \in \mathcal{E}_K$,*

$$\langle(\mathbf{d}_h^{k,i} + \mathbf{l}_h^{k,i})\cdot\mathbf{n}_K, 1\rangle_e = \sigma_{K,e}^{k-1}(\bar u_h^{k,i}). \tag{6.34}$$

**Definition 6.28** (Construction of $\mathbf{d}_h^{k,i}$)**.** *The flux $\mathbf{d}_h^{k,i}$ is defined in $\mathbf{RTN}_0(\mathcal{T}_h)$ using Definition 6.27 with $\sigma_{K,e}(\bar u_h^{k,i})$ in place of $\sigma_{K,e}^{k-1}(\bar u_h^{k,i})$.*

The piecewise constant discrete potential $\bar u_h^{k,i} \in V_h$ has not enough regularity to be meaningful as an argument in the error measure (2.6), in particular regarding the size of its jumps. For this reason, following [40] and the references therein, we introduce an elementwise postprocessing of $\bar u_h^{k,i}$, leading to a new discrete potential $u_h^{k,i}$ sitting in the richer polynomial space $\mathbb{P}_2(\mathcal{T}_h)$. The first step is to determine $\nabla u_h^{k,i}$

from $\mathbf{d}_h^{k,i}$. For simplicity, we assume that the $\boldsymbol{\xi}$-dependency of $\boldsymbol{\sigma}$ can be inverted, i.e., there is a function $\underline{\mathbf{B}} : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d,d}$ such that, for all $(\mathbf{x}, v, \boldsymbol{\xi}, \boldsymbol{\tau}) \in \Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d$,

$$\boldsymbol{\tau} = \underline{\mathbf{A}}(\mathbf{x}, v, \boldsymbol{\xi})\boldsymbol{\xi} \iff \boldsymbol{\xi} = \underline{\mathbf{B}}(\mathbf{x}, v, \boldsymbol{\tau})\boldsymbol{\tau}. \tag{6.35}$$

For the quasi-linear diffusion problem, there holds $\underline{\mathbf{B}}(\mathbf{x}, v) = \underline{\mathbf{A}}(\mathbf{x}, v)^{-1}$, while for the Leray–Lions problem in the $p$-Laplace setting, $\underline{\mathbf{B}}(\boldsymbol{\tau}) = |\boldsymbol{\tau}|^{q-2}\underline{\mathbf{I}}$. Then, we set

$$\nabla u_h^{k,i}|_K := \underline{\mathbf{B}}(\mathbf{x}_K, \bar{u}_h^{k,i}|_K, \mathbf{d}_h^{k,i}(\mathbf{x}_K))\mathbf{d}_h^{k,i}|_K \qquad \forall K \in \mathcal{T}_h, \tag{6.36}$$

where $\mathbf{x}_K$ denotes the barycenter or the circumcenter of $K$. Once $\nabla u_h^{k,i}$ is known, the second step is to determine a suitable integration constant in each element $K \in \mathcal{T}_h$. Possible choices are (depending on the finite volume scheme at hand) $(u_h^{k,i}, 1)_K/|K| := \bar{u}_h^{k,i}|_K$ or $u_h^{k,i}(\mathbf{x}_K) := \bar{u}_h^{k,i}|_K$. This now fully defines $u_h^{k,i} \in \mathbb{P}_2(\mathcal{T}_h)$.

**Definition 6.29** (Approximate gradient, data oscillation, quadrature, and algebraic remainder)**.** *Set* $\mathbf{g}_h^{k,i} := \nabla u_h^{k,i}$, $f_h := \Pi_0 f$, $\overline{\boldsymbol{\sigma}}_h^{k,i} := \mathbf{d}_h^{k,i}$, *and* $r_h^{k,i}|_K := |K|^{-1} R_K^{k,i}$ *for all* $K \in \mathcal{T}_h$.

### 6.5.5 Assumptions verification

The above developments readily yield:

**Lemma 6.30** (Linearization error convergence)**.** *Assumption* 3.5(iii) *holds.*

**Lemma 6.31** (Quasi-equilibration)**.** *Assumption* 4.1 *holds.*

**Lemma 6.32** (Local approximation)**.** *Assumption* 5.1 *holds.*

### 6.5.6 Lowest-order mixed finite elements

We finally tackle the mixed finite element case. We assume that the $\boldsymbol{\xi}$-dependency of $\boldsymbol{\sigma}$ can be inverted, see (6.35), and, omitting the $\mathbf{x}$-dependency, we set $\boldsymbol{\gamma}(v, \boldsymbol{\tau}) := \underline{\mathbf{B}}(v, \boldsymbol{\tau})\boldsymbol{\tau}$ for all $(v, \boldsymbol{\tau}) \in \mathbb{R} \times \mathbb{R}^d$. Let $V_h := \mathbb{P}_0(\mathcal{T}_h)$ and $\mathbf{V}_h := \mathbf{RTN}_0(\mathcal{T}_h)$. The lowest-order Raviart–Thomas mixed method for (2.4) reads: find $(\boldsymbol{\sigma}_h, \bar{u}_h) \in \mathbf{V}_h \times V_h$ such that, for all $(\mathbf{v}_h, v_h) \in \mathbf{V}_h \times V_h$,

$$(\boldsymbol{\gamma}(\bar{u}_h, \boldsymbol{\sigma}_h), \mathbf{v}_h) - (\bar{u}_h, \nabla{\cdot}\mathbf{v}_h) = 0, \tag{6.37a}$$

$$(\nabla{\cdot}\boldsymbol{\sigma}_h, v_h) = (f, v_h). \tag{6.37b}$$

This gives rise to the nonlinear algebraic system (1.1).

Let $(\boldsymbol{\sigma}_h^0, \bar{u}_h^0) \in \mathbf{V}_h \times V_h$, fixing the initial vector $U^0$ in Algorithm 3.7. The linearization of (6.37), for $k \geq 1$, reads: find $(\boldsymbol{\sigma}_h^k, \bar{u}_h^k) \in \mathbf{V}_h \times V_h$ such that, for all $(\mathbf{v}_h, v_h) \in \mathbf{V}_h \times V_h$,

$$(\boldsymbol{\gamma}^{k-1}(\bar{u}_h^k, \boldsymbol{\sigma}_h^k), \mathbf{v}_h) - (\bar{u}_h^k, \nabla{\cdot}\mathbf{v}_h) = 0, \tag{6.38a}$$

$$(\nabla{\cdot}\boldsymbol{\sigma}_h^k, v_h) = (f, v_h), \tag{6.38b}$$

which is the functional form of the algebraic system (1.2). Two common ways to define the function $\boldsymbol{\gamma}^{k-1}(v, \boldsymbol{\tau})$ are the fixed point linearization where $\boldsymbol{\gamma}^{k-1}(v, \boldsymbol{\tau}) := \underline{\mathbf{B}}(\bar{u}_h^{k-1}, \boldsymbol{\sigma}_h^{k-1})\boldsymbol{\tau}$ and the Newton linearization where

$$\begin{aligned}\boldsymbol{\gamma}^{k-1}(v, \boldsymbol{\tau}) := \underline{\mathbf{B}}(\bar{u}_h^{k-1}, \boldsymbol{\sigma}_h^{k-1})\boldsymbol{\tau} + (v - \bar{u}_h^{k-1})\partial_v\underline{\mathbf{B}}(\bar{u}_h^{k-1}, \boldsymbol{\sigma}_h^{k-1})\boldsymbol{\sigma}_h^{k-1} \\ + (\partial_{\boldsymbol{\tau}}\underline{\mathbf{B}}(\bar{u}_h^{k-1}, \boldsymbol{\sigma}_h^{k-1}){\cdot}\boldsymbol{\sigma}_h^{k-1}){\cdot}(\boldsymbol{\tau} - \boldsymbol{\sigma}_h^{k-1}).\end{aligned} \tag{6.39}$$

Problem (6.38) gives rise to a linear system which is of a saddle-point form for a pair of vectors associated with $\bar{u}_h^k$ and $\boldsymbol{\sigma}_h^k$. As such, it is not suitable to the present framework. However, following [42] and the references therein, the resulting algebraic systems can be equivalently rewritten as (6.28), with the only unknowns the discrete potentials $\bar{u}_h^k$. Then, the approach of Section 6.5.3–Section 6.5.5 can be readily used.

| 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

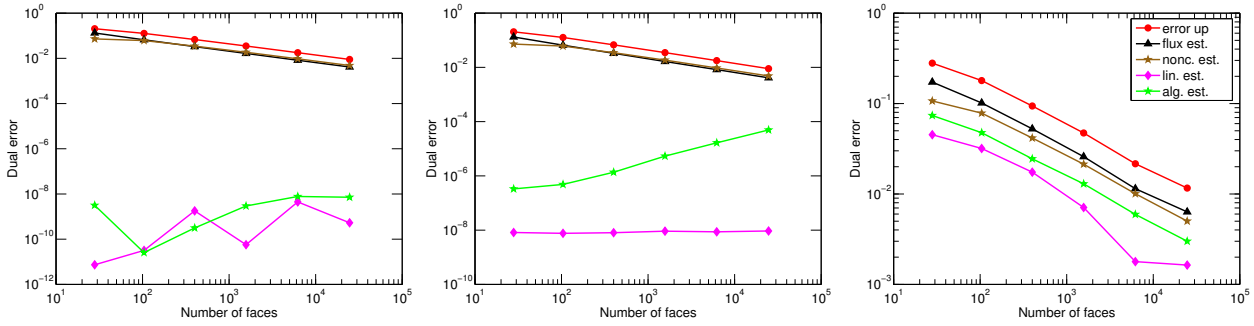


Figure 1: Error and estimators on uniformly refined meshes, case 1, $p = 10$. Exact Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

# 7 Numerical experiments

This section illustrates numerically our theoretical developments. We consider the $p$-Laplacian for $d = 2$ and two test cases with known analytical solution. We employ the Crouzeix–Raviart nonconforming finite element method (6.2), the Newton linearization (6.5), the conjugate gradient (CG) method with diagonal preconditioning, and use the flux reconstructions of Remark 6.7. In (2.7b), the coefficients $\alpha_e$ are set to one and $s := q$.

## 7.1 Test case 1

We set $\Omega := (0,1) \times (0,1)$, $f := 2$, and prescribe the Dirichlet boundary condition by the exact solution $u(\mathbf{x}) = q^{-1}((0.5)^q - |\mathbf{x} - (0.5, 0.5)|^q)$. This is a two-dimensional extension of a test case from Chaillou and Suri [9]. The error stemming from inhomogeneous boundary conditions is neglected. We consider six levels of uniform mesh refinement, together with the values $p \in \{1.5, 10\}$.

We test three approaches with three different stopping criteria in Algorithm 3.7. In the *Exact Newton (EN) method*, both the nonlinear and linear solvers are iterated to "almost" convergence: we impose $\eta_{\rm alg}^{k,i} \le 10^{-8}$ and $\eta_{\rm lin}^{k,i} \le 10^{-8}$. The criterion (3.10) is employed with $\gamma_{\rm rem} = 0.1$; this influences the precision of the calculation of the algebraic error component but not Algorithm 3.7. The *Inexact Newton (IN) method* is as EN except that a fixed number of preconditioned CG iterations is performed on each Newton step. These values were chosen respectively as 2, 3, 5, 8, 10, 15 on each level of mesh refinement. The *Adaptive Inexact Newton (AIN) method* of this work relies on the global stopping criteria (3.10)–(3.12) with $\gamma_{\rm lin} = \gamma_{\rm alg} = 0.3$ and $\gamma_{\rm rem} = 0.3$. This choice of $\gamma_{\rm rem}$ leads to values of $\nu$ increasing on average by 20% 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 $\mathcal{J}_u^{\rm up}(u_h^{k,i}, \mathbf{g}_h^{k,i})$, cf. (2.8), 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

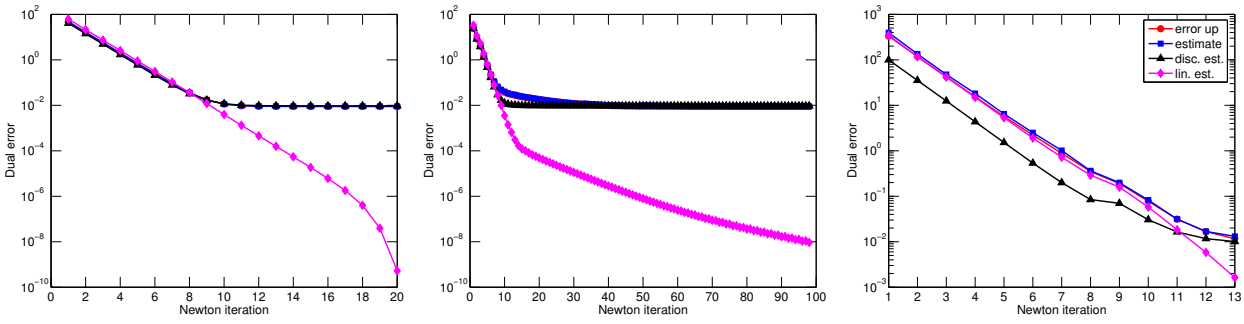


Figure 2: Error and estimators as a function of Newton iterations, case 1, $p = 10$, 6th level mesh. Exact Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

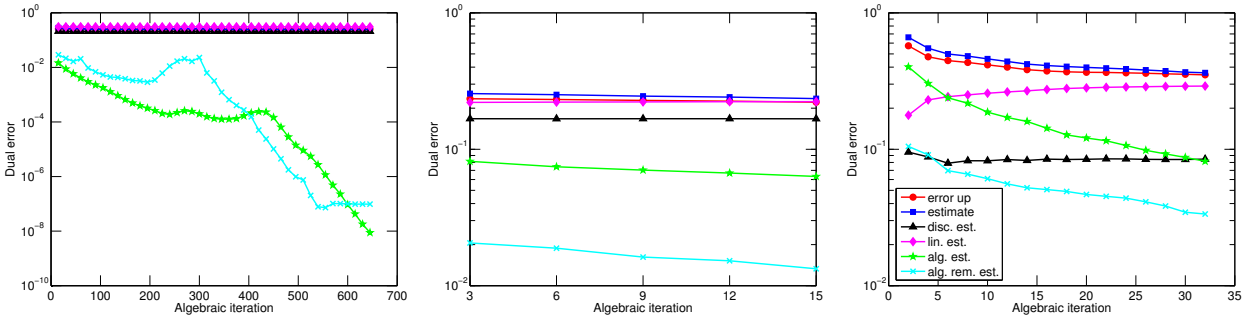


Figure 3: Error and estimators as a function of preconditioned CG iterations, case 1, $p = 10$, 6th level mesh. Exact Newton, 6th step (left), inexact Newton, 6th step (middle), and adaptive inexact Newton, 8th step (right)

three methods (EN, IN, and AIN) yield almost indistinguishable values for $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$, $\eta_{\mathrm{F}}^{k,i}$, and $\eta_{\mathrm{NC}}^{k,i}$, 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 $\eta_{\mathrm{lin}}^{k,i}$ and of the algebraic estimator $\eta_{\mathrm{alg}}^{k,i}$ 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 EN; $\eta_{\mathrm{alg}}^{k,i}$ takes larger values for IN; both $\eta_{\mathrm{alg}}^{k,i}$ and $\eta_{\mathrm{lin}}^{k,i}$ 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 $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$, the overall error estimator $\eta^{k,i}$ of Theorem 3.4, and the discretization and linearization estimators $\eta_{\mathrm{disc}}^{k,i}$ and $\eta_{\mathrm{lin}}^{k,i}$ of Theorem 3.6 on the Newton iterations. Typically, the error and all the estimators except $\eta_{\mathrm{lin}}^{k,i}$ 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 EN and IN perform many unnecessary additional iterations. We can also observe the appearance of quadratic convergence for EN 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 $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$, we have chosen the 6th iteration for EN and IN and the 8th iteration for AIN. We see that almost no decrease of the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$ can be observed during the almost 650 iterations of the preconditioned CG method in the EN case. The fixed 15 CG iterations in the IN case are, on the contrary, not completely sufficient to decrease significantly the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$. 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 EN. This observation is in agreement with the so-called asymptotic mesh independence of the number

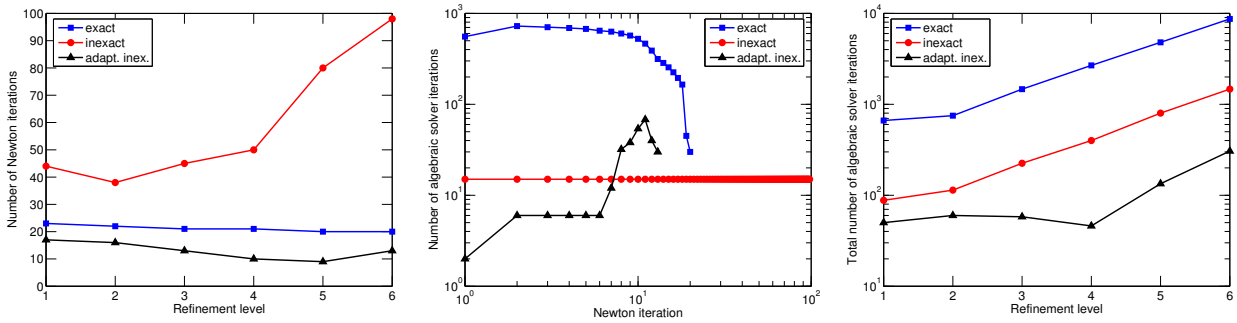


Figure 4: Number of Newton iterations per refinement level (left), number of linear solver iterations per Newton step on 6th level mesh (middle), and total number of linear solver iterations per refinement level (right). Case 1, $p = 10$

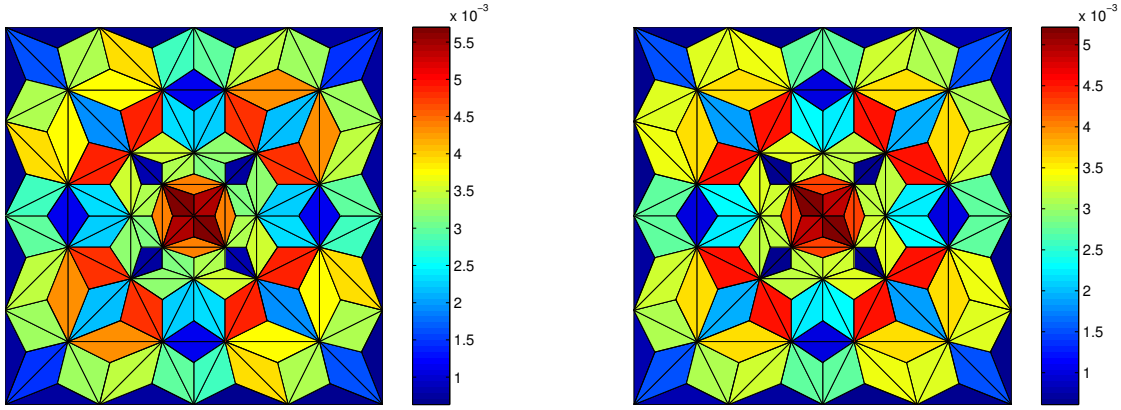


Figure 5: Estimated (left) and actual (right) error distribution, case 1, $p = 10$, 2nd level uniformly refined mesh, adaptive inexact Newton

of Newton iterations, cf., e.g., Weiser *et al.* [43] and references therein for theoretical results. It increases 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 EN 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 EN 8690 iterations. Thus, our approach yields an economy by a factor of roughly 5 with respect to IN and roughly 30 with respect to EN in terms of total algebraic solver iterations.

Figure 5 displays the distribution of the overall error estimator $\eta^{k,i}$ and of the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$ 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.

Figures 6–8 display similar results for the choice $p = 1.5$. The nature of the nonlinearity seems different here from the case $p = 10$, as the Newton-iteration dependence curves of Figure 6 illustrate. In particular, using our stopping criteria avoids the useless waiting before the plateau has been overcome in the classical approaches (EN and IN). As before, these criteria also allow one to invest the right amount of CG iterations in each Newton step, as Figure 7 shows. The computational gains of our approach are important here, with one Newton iteration per refinement up to the 5th level; we only require 122 total CG iterations on the 6th level mesh, in comparison to 3510 for EN and 7755 for IN, see Figure 8. The error and estimator

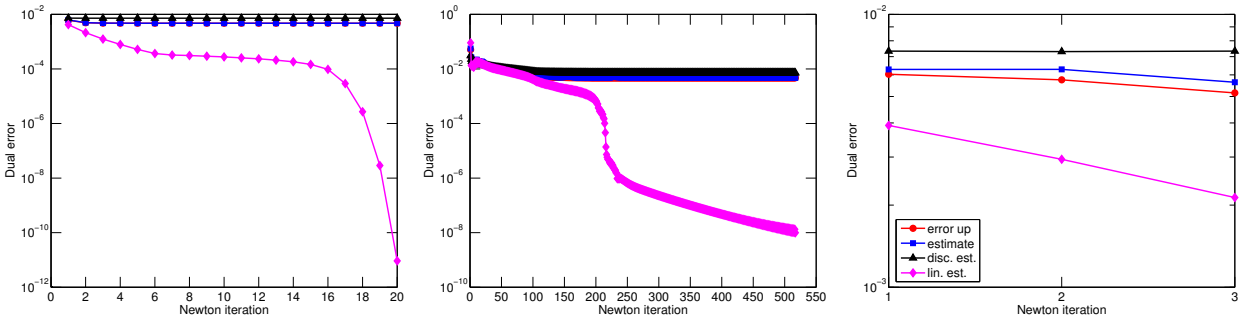


Figure 6: Error and estimators as a function of Newton iterations, case 1, $p = 1.5$, 6th level mesh. Exact Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

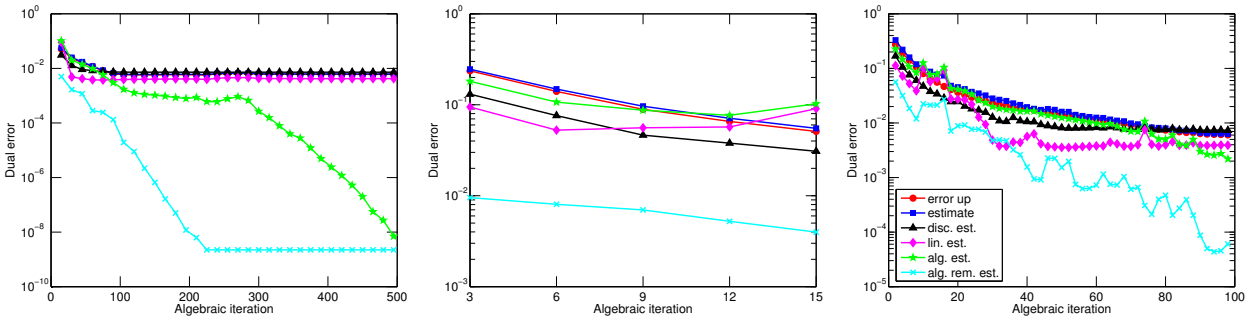


Figure 7: Error and estimators as a function of preconditioned CG iterations, case 1, $p = 1.5$, 6th level mesh, 1st Newton step. Exact Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

distributions are similar to those observed in Figure 5 (not shown).

Finally, we define the upper and lower effectivity indices respectively as $\mathcal{I}^{\mathrm{up}} := \eta^{k,i}/\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$ and $\mathcal{I}^{\mathrm{low}} := \eta^{k,i}/\mathcal{J}_u^{\mathrm{low}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$. Here, $\mathcal{J}_u^{\mathrm{low}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$ is a lower error bound obtained by estimating the supremum in (2.7a) 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 obtained by averaging of $u_h^{k,i}$ on interior vertices and by the Dirichlet condition on boundary vertices. Since $\mathcal{J}_u^{\mathrm{low}}(u_h^{k,i}, \mathbf{g}_h^{k,i}) \leq \mathcal{J}_u(u_h^{k,i}, \mathbf{g}_h^{k,i}) \leq \mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$, the effectivity index $\mathcal{I} := \eta/\mathcal{J}_u(u_h^{k,i}, \mathbf{g}_h^{k,i})$ lies between $\mathcal{I}^{\mathrm{up}}$ and $\mathcal{I}^{\mathrm{low}}$. All effectivity indices (especially $\mathcal{J}_u^{\mathrm{up}}$) are very close to the optimal value of one following Figure 9. This holds for both $p = 10$ and $p = 1.5$, from where we can experimentally confirm the robustness

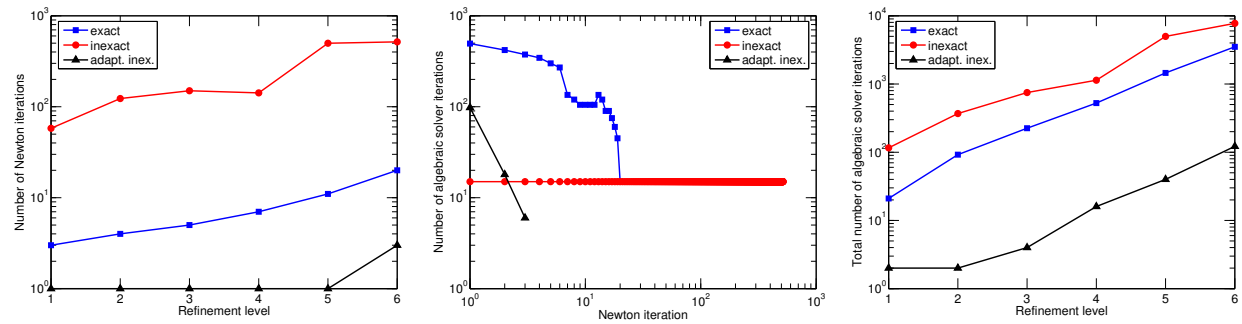


Figure 8: Number of Newton iterations per refinement level (left), number of linear solver iterations per Newton step on 6th level mesh (middle), and total number of linear solver iterations per refinement level (right). Case 1, $p = 1.5$

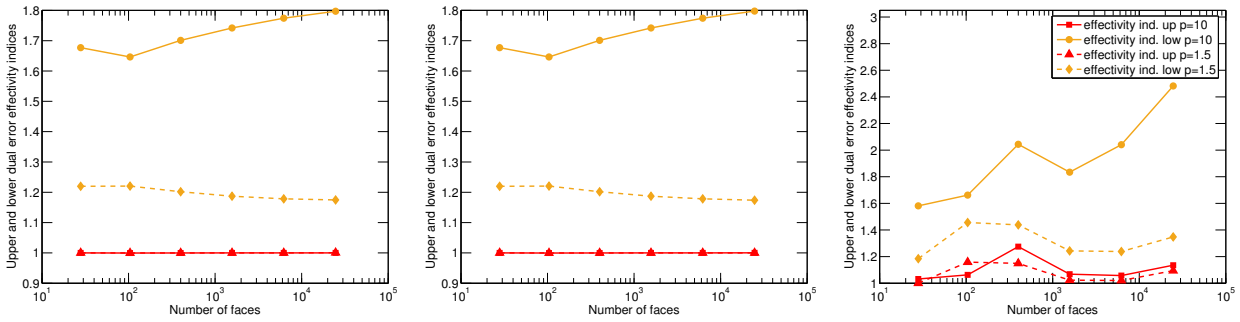


Figure 9: Upper and lower effectivity indices, case 1. Exact Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)

of our estimates with respect to the size of the nonlinearity, given here by the exponent $p$.

## 7.2 Test case 2

This test case is taken from Carstensen and Klose [8, 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, $(r,\theta)$ 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 method. We use the local criteria (3.14) and (3.15) (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 (3.13) 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 approximate 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-elements and the so-called longest edge refinement is used so as to recover a matching mesh (without hanging nodes).

Figure 10 plots the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$ and several estimators as before. In contrast to test case 1, 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 $\eta_{\mathrm{lin}}^{k,i}$ and $\eta_{\mathrm{alg}}^{k,i}$ 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 $\gamma_{\mathrm{lin},D_e} = 1$, the linearization estimator $\eta_{\mathrm{lin}}^{k,i}$ drops to values as low as $10^{-7}$, whereas this estimator would not be so small if the global linearization stopping criterion (3.12) was used.

Figure 11, 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 fast decrease does not appear if the global stopping criterion (3.12) is employed, as the meshes are not sufficiently graded. Figure 11, middle, illustrates that as few as 2 Newton iterations per refinement level are sufficient in our approach (except for initial meshes). The overall efficiency of the AIN combined with adaptive mesh refinement is best appreciated when evaluating the total number of linear solver iterations per refinement level in Figure 11, right: only a very mild increase is observed for adaptive mesh refinement case.

Finally, in Figure 12, we plot the distribution of the estimate $\eta^{k,i}$ and of the error measure $\mathcal{J}_u^{\mathrm{up}}(u_h^{k,i}, \mathbf{g}_h^{k,i})$ 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.

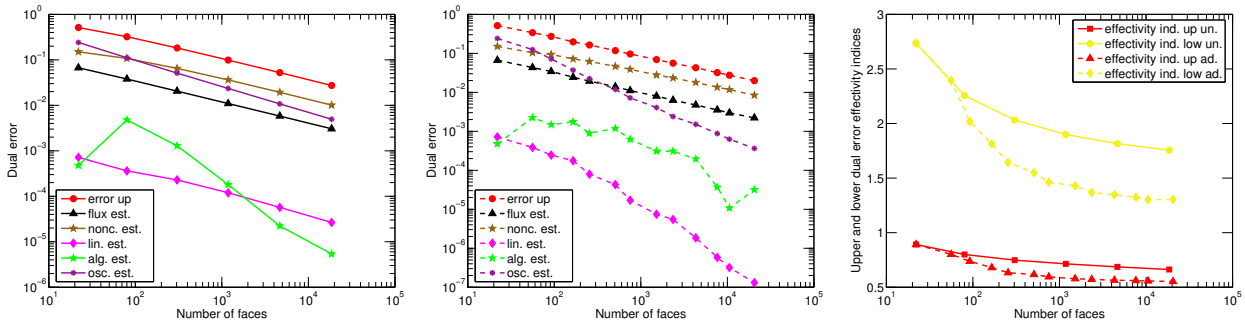


Figure 10: Error and estimators on uniformly (left) and adaptively (middle) refined meshes and upper and lower effectivity indices (right). Case 2, $p = 4$

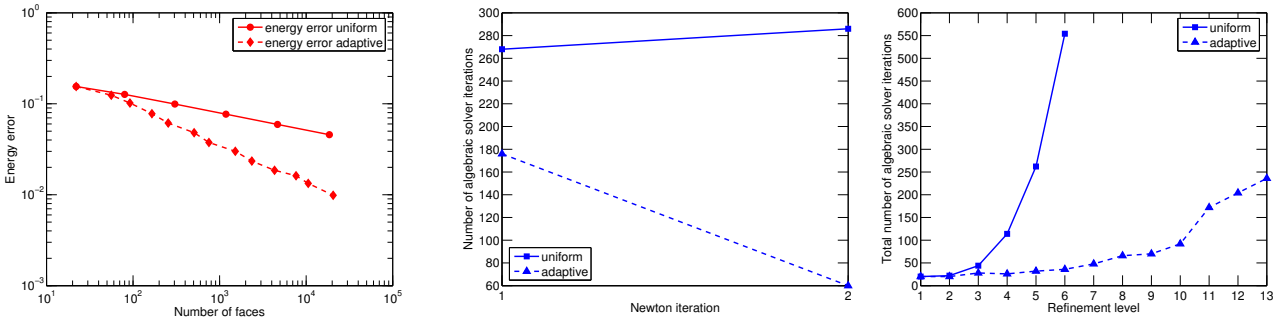


Figure 11: Energy error on uniformly and adaptively refined meshes (left), number of linear solver iterations per Newton step (6th level uniformly and 13th level adaptively refined mesh) (middle), and total number of linear solver iterations per refinement level (right). Case 2, $p = 4$

# 8 Conclusions

In this work, we have designed an inexact Newton method with adaptive stopping criteria for iterative nonlinear and linear solvers. These criteria are based on guaranteed and robust a posteriori error estimates. A complete adaptive strategy combined with adaptive mesh refinement has also been proposed. We have presented numerical experiments illustrating the computational gains achieved by our approach. Our error estimates are derived in an abstract unified framework using equilibrated flux reconstructions. These reconstructions must comply with a couple of assumptions which we have verified for a wide class of discretization schemes and linearizations. In some cases, local mixed finite element problems are to be solved. In practice, the corresponding local matrices can be assembled only once in a preprocessing stage. Additional computational savings are possible by evaluating the error estimators only periodically and not at each iteration of both solvers or by simplifying the estimators by employing quadrature formulas to evaluate the norms.

# References

[1] J. W. Barrett and W. B. Liu, *Finite element approximation of the p-Laplacian*, Math. Comp., 61 (1993), pp. 523–537.

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

[3] L. Belenki, L. Diening, and C. Kreuzer, *Optimality of an adaptive finite element method for the p-Laplacian equation*, IMA J. Numer. Anal., 32 (2012), pp. 484–510.

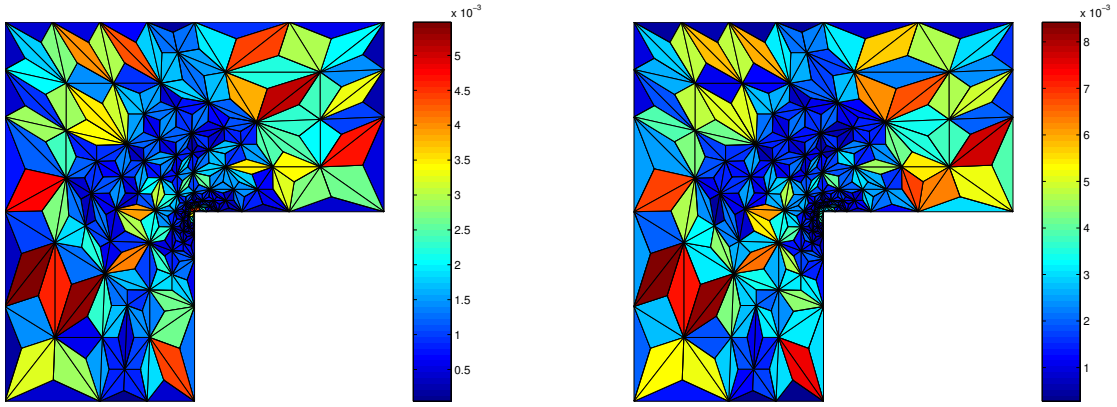


Figure 12: Estimated (left) and actual (right) error distribution, case 2, $p = 4$, 5th level adaptively refined mesh


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

[5] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.

[6] A. Buffa and C. Ortner, *Compact embeddings of broken Sobolev spaces and applications*, IMA J. Numer. Anal., 29 (2009), pp. 827–855.

[7] E. Burman and A. Ern, *Discontinuous Galerkin approximation with discrete variational principle for the nonlinear Laplacian*, C. R. Math. Acad. Sci. Paris, 346 (2008), pp. 1013–1016.

[8] 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.

[9] A. L. Chaillou and M. Suri, *A posteriori estimation of the linearization error for strongly monotone nonlinear operators*, J. Comput. Appl. Math., 205 (2007), pp. 72–87.

[10] 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.

[11] E. Creusé, M. Farhloul, and L. Paquet, *A posteriori error estimation for the dual mixed finite element method for the p-Laplacian in a polygonal domain*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 2570–2582.

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

[13] D. A. Di Pietro and A. Ern, *Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier-Stokes equations*, Math. Comp., 79 (2010), pp. 1303–1330.

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

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

[16] 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.

[17] A. Ern, S. Nicaise, and M. Vohralík, *An accurate* $\mathbf{H}(\mathrm{div})$ *flux reconstruction for discontinuous Galerkin approximations of elliptic problems*, C. R. Math. Acad. Sci. Paris, 345 (2007), pp. 709–712.

[18] 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.

[19] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, North-Holland, Amsterdam, 2000, pp. 713–1020.

[20] ——, *Finite volume approximation of elliptic problems and convergence of an approximate gradient*, Appl. Numer. Math., 37 (2001), pp. 31–53.

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

[22] A. Hannukainen, R. Stenberg, and M. Vohralík, *A unified framework for a posteriori error estimation for the Stokes problem*, Numer. Math., (2012). DOI 10.1007/s00211-012-0472-x.

[23] P. Houston, E. Süli, and T. P. Wihler, *A posteriori error analysis of hp-version discontinuous Galerkin finite-element methods for second-order quasi-linear elliptic PDEs*, IMA J. Numer. Anal., 28 (2008), pp. 245–273.

[24] 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.

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

[26] K. Y. Kim, *A posteriori error estimators for locally conservative methods of nonlinear elliptic problems*, Appl. Numer. Math., 57 (2007), pp. 1065–1080.

[27] 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.

[28] 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.

[29] L. D. Marini, *An inexpensive method for the evaluation of the solution of the lowest order Raviart–Thomas mixed method*, SIAM J. Numer. Anal., 22 (1985), pp. 493–496.

[30] 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.

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

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

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

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

[35] 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.

[36] Z. Strakoš and J. Liesen, *On numerical stability in large scale linear algebraic computations*, ZAMM Z. Angew. Math. Mech., 85 (2005), pp. 307–325.

[37] A. Veeser, *Convergent adaptive finite elements for the nonlinear Laplacian*, Numer. Math., 92 (2002), pp. 743–770.

[38] R. Verfürth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Math. Comp., 62 (1994), pp. 445–475.

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

[40] M. Vohralík, *Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods*, Math. Comp., 79 (2010), pp. 2001–2032.

[41] M. Vohralík and M. F. Wheeler, *A posteriori error estimates, stopping criteria, and adaptivity for two-phase flows.* Preprint R11031, Laboratoire Jacques-Louis Lions & HAL Preprint 00633594, submitted for publication, 2011.

[42] M. Vohralík and B. I. Wohlmuth, *Mixed finite element methods: implementation with one unknown per element, local flux expressions, positivity, polygonal meshes, and relations to other methods*, Math. Models Methods Appl. Sci., (2012). DOI 10.1142/S0218202512500613.

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

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